Evaluating the Perceived Quality of Binaural Technology

Christopher William Pike

Doctor of Philosophy

University of York
Electronic Engineering

January 2019

Abstract

This thesis studies binaural sound reproduction from both a technical and a perceptual perspective, with the aim of improving the headphone listening experience for entertainment media audiences. A detailed review is presented of the relevant binaural technology and of the concepts and methods for evaluating perceived quality. A pilot study assesses the application of state-of-the-art binaural rendering systems to existing broadcast programmes, finding no substantial improvements in quality over conventional stereo signals. A second study gives evidence that realistic binaural simulation can be achieved without personalised acoustic calibration, showing promise for the application of binaural technology.

Flexible technical apparatus is presented to allow further investigation of rendering techniques and content production processes. Two web-based studies show that appropriate combination of techniques can lead to improved experience for typical audience members, compared to stereo signals, even without personalised rendering or listener head-tracking. Recent developments in spatial audio applications are then discussed. These have made dynamic client-side binaural rendering with listener head-tracking feasible for mass audiences, but also present technical constraints. To limit distribution bandwidth and computational complexity during rendering, loudspeaker virtualisation is widely used.

The effects on perceived quality of these techniques are studied in depth for the first time. A descriptive analysis experiment demonstrates that loudspeaker virtualisation during binaural rendering causes degradations to a range of perceptual characteristics and that these vary across other system conditions. A final experiment makes novel use of the check-all-that-apply method to efficiently characterise the quality of seven spatial audio representations and associated dynamic binaural rendering techniques, using single sound sources and complex dramatic scenes. The perceived quality of these different representations varies significantly across a wide range of characteristics and with programme material. These methods and findings can be used to improve the quality of current binaural technology applications.

This thesis is dedicated to Team Human.

“We are here to help each other get through this thing, whatever it is.”
– Mark Vonnegut, in Timequake by Kurt Vonnegut.

Contents

List of Tables

List of Figures

Accompanying Materials

Accompanying materials are provided. This includes software source code, audio files, impulse response data, raw experiment results data, and documentation. The README.md Markdown file in the root directory gives a more detailed explanation of the contents.

Software
Several software libraries are included. These form major parts of the technical apparatus of this thesis.

- ***bbcat-cpp*** – C++ libraries for real-time spatial audio rendering, including the configurable host application *bbcrenderer*.
- ***bbcat-matlab*** – MATLAB functions for analysis and signal processing.
- ***bbcat-python*** – Python code for impulse response measurement and turntable control.

Other software libraries and scripts are included in subdirectories corresponding to part of the thesis for which they were used.

Listening Experiments
The consent forms and written instruction sheets from the listening experiments are included. Where possible, audio files of the test stimuli are included for reference and anonymised results data are included along with analysis scripts.

Associated Publications
PDF versions of the associated publications are included (see section 1.5).

Acknowledgements

I would like to thank my supervisor, Tony Tew, for his guidance, support and friendship during the last six years, as well as for his patience. From our early discussions about the topic to the final days before submission of this thesis, he has always shown dedication and enthusiasm towards this project. His wisdom has been hugely valuable. I am truly grateful.

I would also like to thank Frank Melchior, my industrial supervisor for the majority of this project. His belief in me and his passion to achieve the best have been inspirational. I would not be where I am today without his influence. Most importantly he taught me always to use my ears.

I have been very privileged to undertake my PhD research whilst in employment at BBC Research & Development. I would particularly like to thank Graham Thomas and Samantha Chadwick for enabling this to happen and for their support throughout the process. Many thanks to all of my colleagues at BBC R&D for creating such a uniquely enjoyable and stimulating environment in which to work. I would especially like to thank all members of the Audio Team, past and present, for this, for their support and collaboration, and for their generosity in proof-reading drafts of this thesis. Thanks too to Damian Murphy, for his support as my thesis advisor, and to other colleagues at the Audio Lab for making me feel so welcome when I have visited.

Many thanks to the following collaborators: Richard Day, Thomas Nixon, Richard Taylor and David Marston for contributions towards the software in the technical apparatus of this thesis, again to Thomas Nixon for implementing the web-based surveys, to Matthew Shotton for collaboration on the telescope mount, to Darius Satongar for collaboration during development of the technical apparatus, to Nick Zacharov for inspiring discussions on quality evaluation methods, and to Andrea Genovese and to Mark Ross-Smith for pursuing their MEng research projects in topics related to this thesis. I am very grateful to the sound designers, engineers and producers with whom I have collaborated during this work, particularly to Eloise Whitmore, Tom Parnell, Catherine Robinson, and Caleb Knightley. Besides the technical work of this thesis, some amazing creative works have been produced of which

I am very proud and I am in awe of your talents.

I am fortunate to be part of some vibrant communities of researchers and engineers, including the BBC Audio Research Partnership and the European Broadcasting Union Audio Systems group. I am grateful for the expertise and collaborative mindset of the individuals in these groups, which have benefited this work very much. More generally, I would like to give thanks to the academic community and to the open-source software community; this work would not have been possible without the amazing achievements of many before me and their willingness to share their knowledge with the world.

The intense process of writing a thesis has led me to a strong appreciation of some simple and important things in life: regular exercise, adequate sleep, eating well, strong coffee, and the joy of music. Most important of all, I would like to express my gratitude for the love, support and good company of friends and family. I am looking forward to spending more quality time with you all from now on.

Declaration

I declare that this thesis is entirely my own work and all contributions have been explicitly stated or referenced where appropriate. This work has not previously been presented for an award at this, or any other, University.

Parts of this research have been presented in the following publications:

> C. Pike and F. Melchior, "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio," in Proceedings of 134th AES Convention, 2013.
>
> M. Shotton, C. Pike, and F. Melchior, "A Motorised Telescope Mount as a Computer-Controlled Rotational Platform for Dummy Head Measurements," in Proceedings of 136th AES Convention, 2014.
>
> C. Pike, F. Melchior, and A. I. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," in AES 55th International Conference: Spatial Audio, 2014.
>
> C. Pike, F. Melchior, and A. I. Tew, "Descriptive analysis of binaural rendering with virtual loudspeakers using a rate-all-that-apply approach," in AES International Conference on Headphone Technology, 2016.
>
> C. Pike, and A. I. Tew, "Subjective Evaluation of HRTF Interpolation using Spherical Harmonics," in International Conference on Spatial Audio, 2017. (Abstract only.)
>
> C. Pike, and A. I. Tew, "Characterising the Quality of Dynamic Binaural Rendering of Spatial Audio Formats Using the Check-All-That-Apply Method," in Journal of the Audio Engineering Society, 2019. (Under revision.)

Chapter 1

Introduction

This thesis presents studies of binaural sound reproduction from both a technical and a perceptual perspective. The technology of binaural sound reproduction is aimed at producing convincing spatial impressions of sound scenes for listeners by controlling the sound pressure at the two eardrums (Møller, 1992). It is part of a wider collection of spatial audio technologies. Binaural sound has the advantage that it can be reproduced with an ordinary pair of headphones.

The research presented here was funded by the British Broadcasting Corporation (BBC). It was conducted during, and as part of, the author's employment at BBC Research & Development (R&D). The BBC is the world's oldest and largest public service broadcaster (BBC, 2018b,a). It has a mission "to act in the public interest, serving all audiences through the provision of impartial, high-quality and distinctive output and services which inform, educate and entertain" (BBC Charter, 2016). BBC R&D plays a key role in promoting technological innovation to support the fulfilment of this mission (BBC Charter, 2016, article 15).

Sound and audio technology are core to broadcasting and associated information and entertainment media. BBC engineers have been involved in the research and development of audio technology since the organisation first started radio broadcasting in 1922 (BBC R&D, 2018b). In recent years the Internet and mobile computing technology have transformed media distribution and consumption. Today broadcasters do not only distribute radio and television programmes; services also include a range of personalised, adaptive, interactive, and immersive experiences (BBC R&D, 2018a). Meanwhile audiences are consuming these programmes and services at home and on mobile devices in a wide range of environments.

In its role, BBC R&D must direct and support the organisation through these technological and cultural changes, so it can deliver on its public purpose to provide the most creative, high-quality and distinctive output to its audiences. It does this in large part through active cooperation with partners, particularly with academic researchers at universities. It is in this context that the work presented in this thesis has been carried out.

1.1 Motivation

Sound is perhaps the most important and powerful medium of communication. For the majority of people, hearing plays a major role in understanding the world around us. Through the sound reaching both of our ears, binaural hearing allows human beings to perceive accurate spatial information about our environment and the activity within it. Whilst our vision is dominant in this respect, we hear sounds from beyond our field of vision and our hearing never fully shuts down during sleep. Our hearing not only alerts us to potential dangers, it enables social interaction, exchange of ideas, and expression of emotions through spoken language. It also allows aesthetic appreciation, which leads to the joys of creating and listening to music. In short, hearing is a huge part of the human experience.

Through engineering innovation, audio technologies have allowed us to extend the reach of sound beyond its natural physical limits. We can talk with distant loved ones as if they're in the room or listen in to a collective conversation with millions of people. Audio technology allows us to be transported to different places and times, real or imagined; temporarily inhabiting the worlds of others, expanding our horizons, stimulating our minds. It has been an essential and defining part of the BBC since its first broadcasts.

Just five years after the telephone was first patented, efforts were first made to improve audio systems with spatial sound reproduction (du Moncel, 1881). Researchers around the world are now developing spatial audio technology in a variety of application domains, such as assistive devices (Courtois et al., 2018), architectural design (Katz, Poirier-Quinot, et al., 2018), and archeology (Murphy, Shelley, et al., 2017), with a common aim of providing listeners with more realistic auditory input. In broadcasting and entertainment media, it is hoped that spatial audio can provide new creative possibilities for programme makers and offer more compelling and enjoyable experiences for audiences.

A whole generation remembers the moment in 1963 when the otherworldly sounds of Doctor Who first appeared in their living rooms. Electronic music and sound effects pioneers like Daphne Oram, Delia Derbyshire, and Brian Hodgson of the BBC's Radiophonic Workshop have inspired many more generations of innovative music producers in the UK (Twells, 2013). What would they have done with 3D sound at their disposal? What can we create today?

Then there are moments that we share as a nation. Such as in the London 2012 Olympics, when Mo Farah came around the final bend to win the 10 000 m gold medal. He was lifted by the monumental roar of the crowd (Gibson, 2012). What if the audience at home could feel like they were right there in the stadium, cheering along with them?

David Attenborough's documentary programmes have filled audiences with awe at the wonders of the natural world. They have huge public value, inspiring people to protect wildlife and the environment for future generations (Blake, 2018). The soundtrack is vital to "convey the experience of 'being there' that is important for an immersive TV experience" (Honeyborne, 2018). What if the sound of the dawn chorus at the Great Barrier Reef came from all around the listeners, making them feel like they had really visited this beautiful and endangered underwater environment? Could it enhance our sense of connection with nature?

Some 22 years ago, engineers from BBC R&D published a report of production experiments in five-channel surround sound (Kirby et al., 1996). This format was first introduced on free-to-view broadcasting in the UK in 2006 (BBC, 2006), being already established for home-cinema uses. Since then there has been little change in broadcast audio services offered to listeners. In the following year, the BBC launched an Internet-based on-demand video service, the BBC iPlayer (BBC, 2007), and with growing use of mobile devices (i.e. smartphones and tablets) for media consumption, dedicated mobile applications were released in 2011 (BBC, 2011a,b). Indeed, 472 million smartphones were sold throughout the world in 2011 (Gartner, 2012). On mobile devices, sound is commonly reproduced via headphones. This presents a compelling motivation to improve the listening experience for a growing population of headphone-using audience members.

It has long been acknowledged that headphone-reproduced sound typically results in the impression that the sound sources are inside the listener's head (Plenge, 1974). This is because the natural psychoacoustic cues used in spatial hearing are not adequately provided, particularly binaural cues due to differences in the sound reaching each ear from external sound sources. In the late 1980s, researchers at NASA Ames and the University of Wisconsin-Madison developed techniques for creating three-dimensional auditory virtual environments for headphone listeners, applying digital signal processing of audio signals to introduce the correct binaural information (Wenzel, Wightman, et al., 1988). There has since been a large body of research into binaural technology and techniques are now highly advanced. Several studies have shown that in controlled conditions with careful calibration for the individual listener, binaural rendering can achieve virtual sound sources that are indistinguishable from real sound sources e.g. (Zahorik, Wightman, and Kistler, 1995; Langendijk and Bronkhorst, 2000). There have also been developments into practical systems for creative and commercial applications e.g. (Jot, Larcher, et al., 1995; Jot, Walsh, et al., 2006).

There is significant public interest in the effects that can be achieved with binaural sound. A well-known binaural recording "Virtual Barber Shop" has had over 29 million views on

YouTube (QSound Labs, 2007). Yet at the outset of the work towards this thesis in 2012, binaural technology was not in widespread use in entertainment media. Significant quality enhancement by application of binaural technology had not been clearly demonstrated, in fact studies showed that conventional stereo headphone signals were often significantly preferred (Lorho and Zacharov, 2004; Lorho, 2005a). However development of spatial audio and binaural technology continued (Rumsey, 2011). Meanwhile conceptual understanding of the quality of experiences with media technology was advancing (Le Callet et al., 2012). The position of this author, working within a major broadcaster, offered an opportunity to carry out academic research on the challenges facing successful adoption of binaural technology within this application context, with access to programme production expertise and feedback from a large audience of listeners.

Over the course of this project, trends in technology usage and media consumption have continued to shift. Despite many years of widespread availability, a recent survey shows that in the UK only 11.5 % of households have a surround sound reproduction system (Cieciura et al., 2018); meanwhile 82.5 % of the respondents own a smartphone. Mobile device sales are still growing (Gartner, 2018), as are the sales of headphones: in 2017, worldwide sales grew 4 % to 362 million units and retail value grew 29 % to $16.8 billion (Futuresource, 2018). Radio Joint Audio Research (RAJAR)[1] figures from Spring 2018 indicate that the number of UK adults listening to audio via a mobile phone or tablet increased from 9 % to 28 % between 2011 and 2018. Similarly, in 2018 figures showed 31 % of UK adults listen to radio programmes using headphones, compared with 18 % in 2011. Meanwhile the BBC iPlayer received 96 million audiovisual content requests from mobile devices in October 2018, a number which continues to grow (Bell, 2018). There have also been new commercial technology developments that are highly compatible with the applications of binaural sound, including codec systems for 3D spatial audio and virtual and augmented reality devices.

Despite disappointing earlier findings, there appears to be great potential for binaural technology to improve the listening experiences of large audiences of entertainment media, and to offer exciting new creative possibilities to programme makers.

1.2 Aims and objectives

The primary aim of this work is to improve the quality of headphone listening experiences for entertainment media audiences by developing and evaluating binaural technology. The

[1] `http://www.rajar.co.uk`

ambition is to understand how best to apply binaural techniques in these applications. This will be driven primarily through evaluation by human listeners of the perceived quality of binaural technology systems, as well as their application to presentation of audio or audiovisual entertainment media. It will also be driven by the context in which this applied project operates i.e. the technological and economic factors that influence the way in which binaural technology may be applied in practice.

This can be formulated in terms of the following set of *research questions*:

1. Is binaural rendering capable of producing a convincing spatial impression without calibration for the individual listener?
2. Can binaural rendering improve the perceived quality of the headphone listening experience in entertainment media applications?
3. How does the programme production process influence the perceived quality when binaural rendering is applied?

These aims will be pursued through the following *objectives*:

1. Review the state-of-the-art in binaural technology as used in entertainment media.
2. Review the state-of-the-art in relevant quality evaluation methods.
3. Develop and validate tools for comparing state-of-the-art binaural rendering techniques, as well as for applying them in programme production.
4. Improve understanding of the quality of binaural technology by applying appropriate quality evaluation methods.
5. Evaluate in-depth the quality of a range of production and delivery options for providing headphone-reproduced spatial audio to listeners in entertainment media applications.

1.3 Structure of this Thesis

Chapter 2 introduces binaural technology. It begins with an overview of the processes of spatial hearing, which form the basis for binaural technology. The historical development of binaural recording technology is then discussed, followed by the fundamentals of binaural rendering technology. A more detailed review is then given of the state-of-the-art in a number of specific topics in the field of binaural rendering: HRTF individualisation, HRTF interpolation, headphones, head tracking and auditory virtual environments. Besides the techniques themselves, studies of their perceptual effects are also discussed. This chapter informs later implementations of binaural rendering to ensure that state-of-the-art performance is achieved.

Chapter 3 reviews perceived quality evaluation. First the common conceptual understanding of perceived quality is discussed, as it relates to media technology: its definition, how it is formed, what influences it and what contributes towards it. Particular attention is then given to sound quality. A review of methods for evaluating quality is presented, with reference to past studies in the field of spatial audio. These are broadly categorised into integrative evaluation, discriminative analysis and descriptive analysis methods. Descriptive analysis methods allow the features of quality to be characterised; besides classical methods, more recent methods are reviewed that allow rapid quality characterisation with less experienced assessors. The role of assessor experience in perceptual evaluation of sound quality is then discussed. Reflections on this review are given and the outlook for applications of binaural technology is considered. This chapter informs the methods for evaluation of binaural rendering used in later chapters.

Chapter 4 presents a pilot study carried out at the beginning of the work towards this thesis in 2012. It evaluates the status quo at that time in terms of the quality that could be achieved by applying binaural rendering to existing broadcast programme content in the 5.1 format. Twelve commercial systems are compared to a stereo down-mix. A system that uses head-tracking and individual BRIR measurements is also included. This study asks if recent advancements in binaural rendering technology have provided quality improvements. Reflections on the findings are given and used to inform the further research strategy for the project.

Chapter 5 assesses the plausibility of binaural rendering that can be achieved using a non-individualised approach. Using insights from prior research on binaural rendering and its perceptual effects, a state-of-the-art dynamic (head-tracked) binaural rendering system is created. A listening experiment is presented, based on signal detection theory, to determine whether the binaural rendering gives simulation in agreement with listeners' expectations of real sound sources in the listening room. This chapter concludes with a discussion on the implications of the findings in relation to the quality of binaural rendering.

Chapter 6 discusses developments in spatial audio applications since the outset of this project. First, a summary is given of the application to programme production of the binaural rendering apparatus developed during this thesis. The results of two web-based quality evaluation experiments are presented, conducted with BBC audience members. The establishment of so-called next-generation audio (NGA) standards is discussed; these support coding and distribution of 3D audio formats. The discussion also includes the growth in virtual reality (VR) and augmented reality (AR) devices and supporting services. The prevalence of loudspeaker virtualisation techniques in these applications is discussed and relevant studies of this approach are reviewed. This informs the approach to further research in the final chapters of this thesis.

Chapter 7 presents a descriptive analysis of the perceptual effects of virtual amplitude panning in binaural rendering. The experiment explores the perceptual differences between binaural rendering with a single virtual sound source and by virtual amplitude panning. Comparison is made across multiple other system factors: use of head-tracking, use of room response, and source positions. A pre-defined set of quality features is used, though assessors selected their own subsets for rating, this is known as a rate-all-that-apply (RATA) approach. The results give a characterisation of the effects that virtualised amplitude panning has on perceived quality for simple single-source scenes and how they vary across other rendering conditions.

Chapter 8 presents an experiment to characterise the quality of binaural rendering using a range of spatial audio formats that are common in emerging applications. In all cases, head tracking is used during rendering, in acknowledgement that this is now feasible on a mass scale with mobile VR systems. Use of head tracking requires real-time client-side binaural rendering of the content. Efficiency therefore becomes of concern alongside quality, both of which will be influenced by the 3D spatial audio format used. Evaluation is performed both for single musical sources and complex dramatic scenes. Again a pre-defined set of quality features is used to characterise the quality of these systems. Rather than rate stimuli on attribute scales, assessors used a simple binary response format, known as the check-all-that-apply (CATA) method. The method is reviewed in detail and then the quality characterisation experiment is presented. This chapter concludes with a discussion of the findings, both in terms of the spatial audio techniques and the evaluation method.

Chapter 9 concludes the thesis. It summarises the findings of the work and considers prospects for further research.

Appendix A presents the implementation and verification of technical apparatus for investigating binaural system factors and how they influence perceived quality. This apparatus is used for the experiments of chapters 6 to 8 and elements of it are used in chapter 5 also. A perceptual evaluation of HRTF interpolation using spherical harmonics is also presented in order to validate the use of that technique in the study of chapter 8.

Appendix B presents the background to techniques for loudspeaker rendering, which are relevant to studying loudspeaker virtualisation in binaural rendering systems. Extensions to the apparatus of appendix A for investigating binaural rendering via intermediate channel-based and scene-based (ambisonics) formats are presented. This enabled the study presented in chapter 8. Methods for analysing such techniques are discussed and an initial analysis of the systems that were evaluated perceptually in chapter 8 is presented.

1.4 Contributions

The principal contributions of this thesis are:

Chapter 4 The first quality evaluation of headphone virtualisation of 5.1 surround sound to include accompanying video and a head-tracked individualised system, whilst also comparing the effect of the listening environment.

Chapter 5 The first criterion-free evaluation of the plausibility of non-individualised dynamic binaural rendering in a small room.

Chapter 6 Two web-based studies showing significant listener preferences for static non-individualised binaural versions of audio drama material over stereo versions.

Chapter 7 The first evaluation of the effects on perceived quality of virtual amplitude panning in binaural rendering, conducted across multiple other system factor variables. The first application of the RATA evaluation and analysis methods to characterise the quality of audio technology.

Chapter 8 The first experiment to compare and characterise the perceived quality of object-based, ambisonics-based and VBAP-based binaural rendering. The evaluation was performed both with single sound sources and professionally-produced complex dynamic 3D scenes, which appears also to be novel. Whilst another application of the CATA method to audio systems was published during this study, this is the first application to spatial audio systems and the first study to relate CATA results to overall quality in the field of audio.

Appendix A A novel system for comparing binaural rendering system factors in real-time and using standardised representations of 3D audio scenes. The first publicly-available set of binaural room impulse responses (BRIRs) measured for all loudspeaker layouts in Recommendation ITU-R BS.2051 (ITU-R, 2018) at multiple head orientations. The first perceptual validation that spherical harmonic interpolation of head-related transfer functions (HRTFs) can show no audible differences from real measurements.

1.5 Associated Publications

The following publications are related directly to the work presented in this thesis:

P. I C. Pike and F. Melchior, "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio," in Proceedings of 134th AES Convention, 2013.

P. II C. Pike, F. Melchior, and A. I. Tew, "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," in AES 55th International Conference: Spatial Audio, 2014.

P. III C. Pike, F. Melchior, and A. I. Tew, "Descriptive analysis of binaural rendering with virtual loudspeakers using a rate-all-that-apply approach," in AES International Conference on Headphone Technology, 2016.

P. IV C. Pike, and A. I. Tew, "Subjective Evaluation of HRTF Interpolation using Spherical Harmonics," in International Conference on Spatial Audio, 2017. (Abstract only.)

P. V C. Pike, and A. I. Tew, "Characterising the Quality of Dynamic Binaural Rendering of Spatial Audio Formats Using the Check-All-That-Apply Method," in Journal of the Audio Engineering Society, 2019. (Under revision.)

These co-authored publications are also somewhat related to the work in this thesis:

Co.P. I M. Shotton, C. Pike, and F. Melchior, "A Motorised Telescope Mount as a Computer-Controlled Rotational Platform for Dummy Head Measurements," in 136th AES Convention, 2014.

Co.P. II C. Pike, P. Taylour, and F. Melchior, “Delivering object-based 3D audio using the web audio API and the audio definition model,” Web Audio Conf., 2015.

Co.P. III D. Satongar, C. Pike, Y. W. Lam, and A. I. Tew, “The Influence of Headphones on the Localization of External Loudspeaker Sources,” J. Audio Eng. Soc., vol. 63, no. 10, pp. 799-810, 2015.

Co.P. IV W. Bailer, C. Pike, R. Bauwens, R. Grandl, M. Matton, and M. Thaler, “Multi-sensor concert recording dataset including professional and user-generated content,” in Proceedings of the 6th ACM Multimedia Systems Conference - MMSys ’15, 2015.

Co.P. V J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck, and A. Hilton, “Presenting the S3A Object-Based Audio Drama dataset,” in 140th AES Convention, 2016.

Co.P. VI C. Pike, R. Taylor, T. Parnell, and F. Melchior, “Object-based spatial audio production for virtual reality using the Audio Definition Model,” in AES International Conference on Audio for Augmented and Virtual Reality, 2016.

Co.P. VII N. Zacharov, C. Pike, F. Melchior, and T. Worch, “Next generation audio system assessment using the multiple stimulus ideal profile method,” in International Conference on Quality of Multimedia Experience, 2016.

Co.P. VIII N. Zacharov, T. Pedersen, and C. Pike, “A common lexicon for spatial sound quality assessment - latest developments,” in International Conference on Quality of Multimedia Experience, 2016.

Co.P. IX C. Pike and M. Romanov, “An impulse response dataset for dynamic data-based auralisation of advanced sound systems,” in 142nd AES Convention, 2017.

Co.P. X T. Parnell and C. Pike, “An efficient method for producing binaural mixes of classical music from a primary stereo mix,” in 144th AES Convention, 2018.

Co.P. XI P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. J. Hughes, D. Menzies, M. F. S. Galvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchior, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, “An Audio-Visual System for Object-Based Audio: From Recording to Listening,” IEEE Trans. Multimed., vol. 20, no. 8, pp. 1919-1931, Aug. 2018.

Co.P. XII A. Franck, G. Costantini, C. Pike, and F. M. Fazi, “An Open Realtime Binaural Synthesis Toolkit for Audio Research,” in 144th AES Convention, 2018.

Co.P. XIII J. Woodcock, J. Franombe, A. Franck, P. Coleman, R. Hughes, H. Kim, Q. Liu, D. Menzies, M. F. Simón Gálvez, Y. Tang, T. Brookes, W. J. Davies, B. M. Fazenda, R. Mason, T. J. Cox, F. M. Fazi, P. J. B. Jackson, C. Pike, and A. Hilton, "A Framework for Intelligent Metadata Adaptation in Object-Based Audio," in AES International Conference on Spatial Reproduction - Aesthetics and Science, July 2018.

Chapter 2

A Review of Binaural Technology

This chapter presents a detailed review of binaural technology; including an overview of the processes of spatial hearing, the historical development of binaural recording technology, and the fundamentals and the state-of-the-art of binaural rendering technology. In addition the technology itself, related perceptual studies are discussed. This chapter informs later implementations of binaural rendering to ensure that state-of-the-art performance is achieved.

2.1 Introduction

The basis of binaural technology can be viewed through the following assumption (Møller, 1992):

> Recreation of sound pressure signals at a listener's eardrums will result in the same auditory experience as created by the original signals.

In order to give the auditory impression of a sound-emitting object without that object being present, binaural technology must create the same signals at the listeners ears as would the real object. Similarly for complex sound fields e.g. created by multiple sound sources in a reverberant environment, if the sound pressure at the eardrums created by a binaural technology system is the same as it would be if the listener were present in that environment, it is assumed that the listener will perceive the scene in the same way. The aim therefore is to create auditory virtual reality.

The term *binaural* is used because the distinct signals at both ears are produced correctly, so invoking the normal binaural hearing processes. Reproduction of binaural sound signals is normally done using headphones, since this makes it easier to control the sound pressure at the ears independently. Binaural technology originates from the binaural recording technique, where microphones are placed in the ears of a human listener or those of an artificial head. Further developments have led to systems capable of simulating binaural signals without requiring the existence of an original sound field.

Binaural sound reproduction can be viewed as one of a family of spatial sound reproduction techniques. There are other techniques which have different goals to the binaural approach and primarily target loudspeaker reproduction. Sound-field synthesis techniques aim at physically accurate reconstruction of sound fields across an extended listening region. Amplitude panning techniques take a more practical approach, creating simplified approximations to auditory cues at the listening position. Spors, Wierstorf, et al. (2013) provide a review of such techniques.

There are many potential applications of binaural technology. For example, in auditory display, for more effective information streaming and aurally-guided navigation (Begault, Wenzel, Godfroy, et al., 2010), or in virtual environment simulation, for controlled comparison of concert hall acoustics (Maempel and Lindau, 2012). The primary focus of this thesis is in entertainment media applications where the aim is to create immersive and interactive experiences.

Clearly the above stated assumption is reductive. The auditory system does not operate in isolation, there are interactions with other sensory modalities such as vision and proprioception (Wozny et al., 2008). The assumption is also over-strict. It appears that sound pressure signals do not need to be physically identical to create convincing simulations of sound sources and scenes (Brinkmann, Lindau, and Weinzierl, 2014). Certain approximations can be made whilst achieving perceptually accurate reproduction. This makes implementations of binaural technology feasible. However, successful real-world applications are still challenging.

To study how best to apply binaural technology in entertainment media applications, a comprehensive understanding of prior work in this area is required. This chapter first introduces the processes of spatial hearing, particularly binaural hearing, which form the basis for study and development of binaural technology. A detailed review of binaural recording and rendering technology is then given, including the historical development, the basic techniques, and the current state-of-the art.

2.2 Spatial Hearing

The human auditory system gives us a remarkable ability to interpret complex and information-rich acoustic scenes from binaural sound signals. Hearing is of profound importance to communication and social interaction. It also enables situational awareness, perceiving information about our environment, orienting ourselves within it, and detecting the activities around us. We are able to localise sounds with relatively good accuracy, including sources beyond our field of vision. Our auditory system utilises both monaural and binaural cues in sound localisation, but binaural hearing provides significant advantages. It is through an understanding of the mechanisms of spatial hearing that effective and practical binaural technology can be engineered. A comprehensive review of spatial hearing mechanisms is given by Blauert (1997) and a recent summary is given by Pulkki and Karjalainen (2015). This section will attempt only to summarise the necessary background.

2.2.1 Basic Concepts

To differentiate between acoustic activity and the perception of sound, Blauert (1997) uses the term *auditory event* to describe anything that is perceived auditorily (heard) and *sound event* to describe acoustic activity. The totality of auditory events define our *auditory space*. The auditory system represents the external acoustic environment by an internal auditory scene (Bregman, 1999). *Localisation* is the process of associating the spatial character of an auditory event in the auditory scene with the character of a sound event in the acoustic environment, including its direction, distance and extent. Auditory events are localised with varying degrees of precision and may be located in the external environment or sometimes inside the listener's head.

Binaural hearing describes hearing processes where information due to differences in signals between the two ears is present and is taken into account. Whereas *monaural hearing* refers to situations where there is no interaural difference information, or this information is ignored. Binaural hearing gives significant advantages for localisation, yielding substantially different auditory space to that given by monaural hearing. Properties in the signals reaching the ears that inform localisation processes are often called *localisation cues*.

Without use of binaural recording or processing techniques, listening to sound on headphones typically results in *inside-the-head* localisation, since natural localisation cues for external sounds are unavailable. The term *lateralisation* is used to describe the localisation of those auditory events that are inside-the-head at some point between the left and right ears, and *internalisation* is used to describe localisation of sources anywhere inside-the-head. In natural listening scenarios, we typically localise external sounds to locations outside of the head, which is often called *externalisation*.

2.2.2 Coordinate Systems

In order to describe spatial relationships and particularly the spatial attributes of auditory events, it is important to define the coordinate systems to be used. There are a variety of ways with which three-dimensional (3D) space is described in relation to spatial hearing.

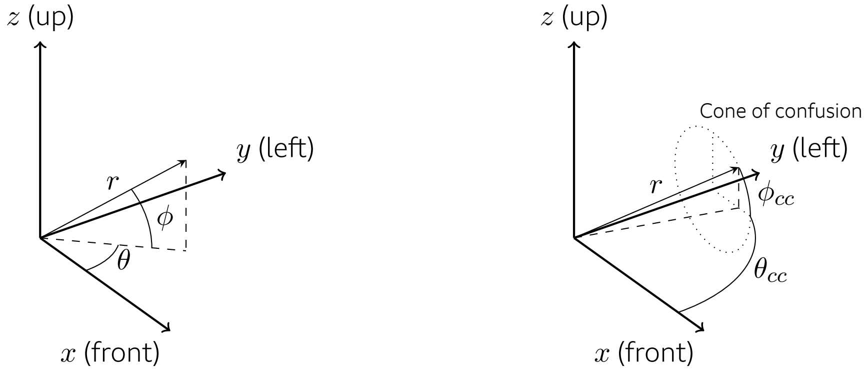
- *Cartesian coordinates* – Positions are defined according to location in the x, y, and z axes (see figure 2.1). This coordinate system is logical when viewing a scene and a listener from an external perspective e.g. in relation to the room or other environment. With the centre of the listener's interaural axis at the origin and the listener's head pointing towards the positive x-axis, y defines the inter-aural axis with positive coordinates to the left and z defines up-down with positive coordinates up.

- *Spherical coordinates* – Positions are defined according to azimuth angle θ, elevation angle ϕ and range/distance r (see figure 2.1a). This coordinate system is logical when viewing a scene in relation to a listener, with the origin at the centre of their head. The elevation angle has the value range ($-90° \leq \phi \leq 90°$). The direction directly in front of the listener has azimuth and elevation of 0°. Azimuth angle increases with counter-clockwise rotation about the z-axis. Elevation angle is positive above the horizontal plane and negative below it.

- *Interaural polar coordinates* – Positions are defined according to the lateral angle θ_{cc}, polar angle ϕ_{cc} and range/distance r (see figure 2.1b). These may also be called *cone of confusion coordinates*. For a given θ_{cc}, values of ϕ_{cc}, and r represent a conical surface comprising points at which the difference in distances to the positions of the two ears is constant. The direction directly in front of the listener has lateral and polar angles of 0°. The lateral angle has the value range $-90° \leq \theta_{cc} \leq 90°$ with positive values to the left of the median plane, along the y-axis. Positive values of the polar angle ϕ_{cc} correspond to positions above the horizontal plane, obtained by counter-clockwise rotation about the y-axis.

A 3D position **r** can be represented as a column vector when using Cartesian coordinates, with the following relationship to the spherical coordinates:

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\cos(\theta)\cos(\phi) \\ r\sin(\theta)\cos(\phi) \\ r\sin(\phi) \end{pmatrix}. \tag{2.1}$$



(a) Spherical coordinate system

(b) Interaural polar coordinate system

Figure 2.1: Coordinate systems

The spherical coordinates can be obtained from the Cartesian coordinates according to:

$$\theta = \text{arctan}\,\frac{y}{x}, \tag{2.2}$$

$$\phi = \text{arctan}\,\frac{z}{\sqrt{x^2+y^2}}, \tag{2.3}$$

$$r = \sqrt{x^2+y^2+z^2} = \|\mathbf{r}\|. \tag{2.4}$$

We also define here the anatomical planes used in subsequent discussion:

- *Frontal plane* – divides front and back positions, defined by $x = 0$.
- *Median plane* – divides the left-right sides of the listener along the x axis, defined by $y = 0$. It is also called the mid-sagittal plane, where other sagittal planes run in parallel with different y coordinate values.
- *Horizontal plane* – divides positions above and below ear height, defined by $z = 0$.

It is sometimes useful also to describe positions relative to one side of the head in the left-right axis: *ipsilateral* refers to positions on the near side of the head, whilst *contralateral* refers to positions on the far side.

Rotations of the head also have common terminology.

- *Yaw* (γ) – horizontal or azimuthal rotation about the vertical axis.
- *Pitch* (ψ) – vertical or elevational rotation about the inter-aural axis.
- *Roll* (ζ) – lateral rotation about the front-back axis.

Coordinates may be defined in terms of a local or global frame of reference. These head rotation coordinates are defined with reference to current orientation of the head, and applied in the order: yaw, pitch, roll. These are one convention of the intrinsic Tait-Bryan angles. There are many different ways to represent 3D rotations (Taubin, 2011), but this approach is most common and intuitive when considering head movements.

2.2.3 Historic Studies of Binaural Hearing

Binaural hearing mechanisms have been studied since at least the late 18th century, as reviewed by Wade and Detutsch (2008). Giovanni Battista Venturi suggested that sound localisation was based on amplitude differences between the ears as early as 1796. In the 19th century devices were invented to enable separate transmission of sounds to each ear, so allowing more controlled study of binaural hearing. This includes Somerville Scott Alison's stethophone (1858) and Silvanus Thompson's pseudophone (1879).

In 1907, Lord Rayleigh published a theory of binaural sound localisation (Lord Rayleigh, 1907), now known as his *duplex theory*. It stated that sound localisation is based on two types of cue, which are induced between the ear signals by our morphology. Dependent on the angle of incidence of the sound source and the frequency of the sound, the head acts as a baffle, inducing interaural intensity differences (IIDs), or interaural level differences (ILDs). Additionally, due to the spatial separation of the ears, interaural phase differences (IPDs), or interaural time differences (ITDs), are induced by path length differences to the sound source according to the source position. It was observed that interaural time differences (ITDs) are most important at low frequencies, where wavelengths mean that acoustic shadowing of the head is negligible and so interaural level difference (ILD) cues are lacking. Additionally at higher frequencies, phase differences become somewhat ambiguous cues to time of arrival (TOA), where the wavelength is short relative to the size of the head. It was also noted that in the median plane, these interaural differences tend to zero, and so localisation becomes challenging, resulting in front-back ambiguity.

2.2.4 Head-Related Acoustics

The scattering of sounds off the bodies of humans, especially the torso, head and ears, provides acoustic cues which the auditory system interprets in the perception of auditory events. This particularly informs the spatial characteristics of auditory events during localisation. Assuming a static arrangement of the listener and a sound source, this scattering process can be modelled as a linear time-invariant (LTI) transfer function, measured at the listener's eardrums (Mehrgardt and Mellert, 1977). These are commonly called head-related transfer

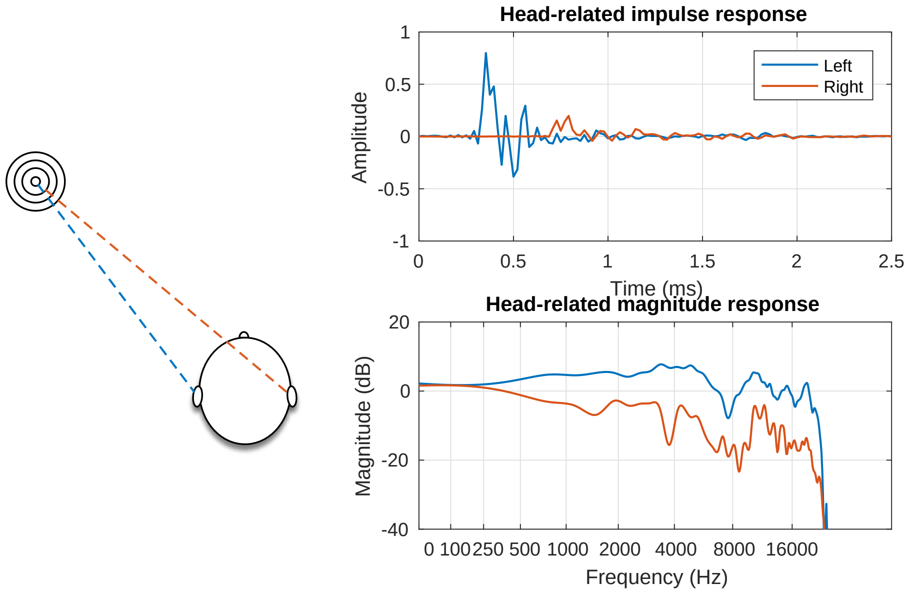


Figure 2.2: An example HRIR and HRTF magnitude response for a source at direction $\theta = 45°$, $\phi = 0°$. Using data measured on a Neumann KU100 dummy head microphone at a distance of 3.25 m by Bernschütz (2013).

functions (HRTFs), or in the time domain head-related impulse responses (HRIRs). The HRTF for a given source and listener configuration represents the temporal and spectral structure of the signals at the two ears, as well as frequency-dependent interaural intensity and phase differences. Importantly the HRTFs vary according to the spatial configuration of the sound source and the listener. A HRTF can be measured using a pair of microphones placed in the ears of a human subject or a dummy head[1] (see section 2.4.2). An example is plotted in figure 2.2 for a source at azimuth $\theta = 45°$ and elevation $\phi = 0°$.

A non-zero source azimuth angle leads to interaural time and level differences, due to the shape of the head and the spacing of the ears. The asymmetry of the head contributes to differences in ear signals with source direction, particularly in front-back and up-down axes. The torso also adds reflections that reach the ears, particularly from the shoulders, which are dependent on the polar angle of the source (Algazi, Avendano, et al., 2001a).

The pinna (illustrated in figure 2.3) contributes to signal differences at frequencies above approximately 4 kHz (Shaw, 2007), particularly concerning source positions varying in the front-back and up-down axes. The cavities of the pinnae resonate at different frequencies, dependent on the angle of incidence of the sound source, creating spectral peaks. Spectral

[1] A dummy head microphone is sometimes also called an artificial head or binaural microphone. When there is also a torso the term head and torso simulator (HATS) is often used.

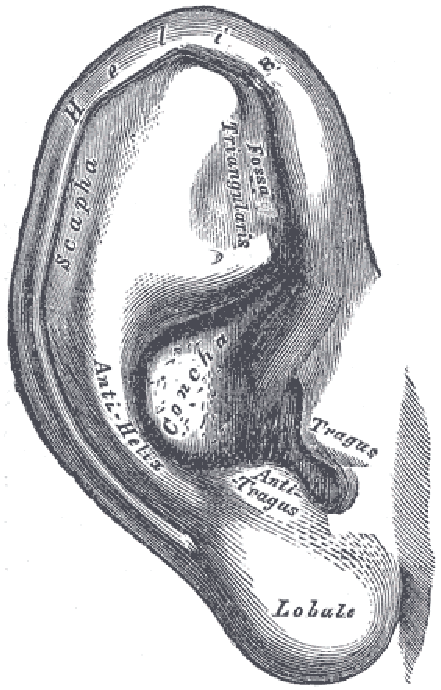


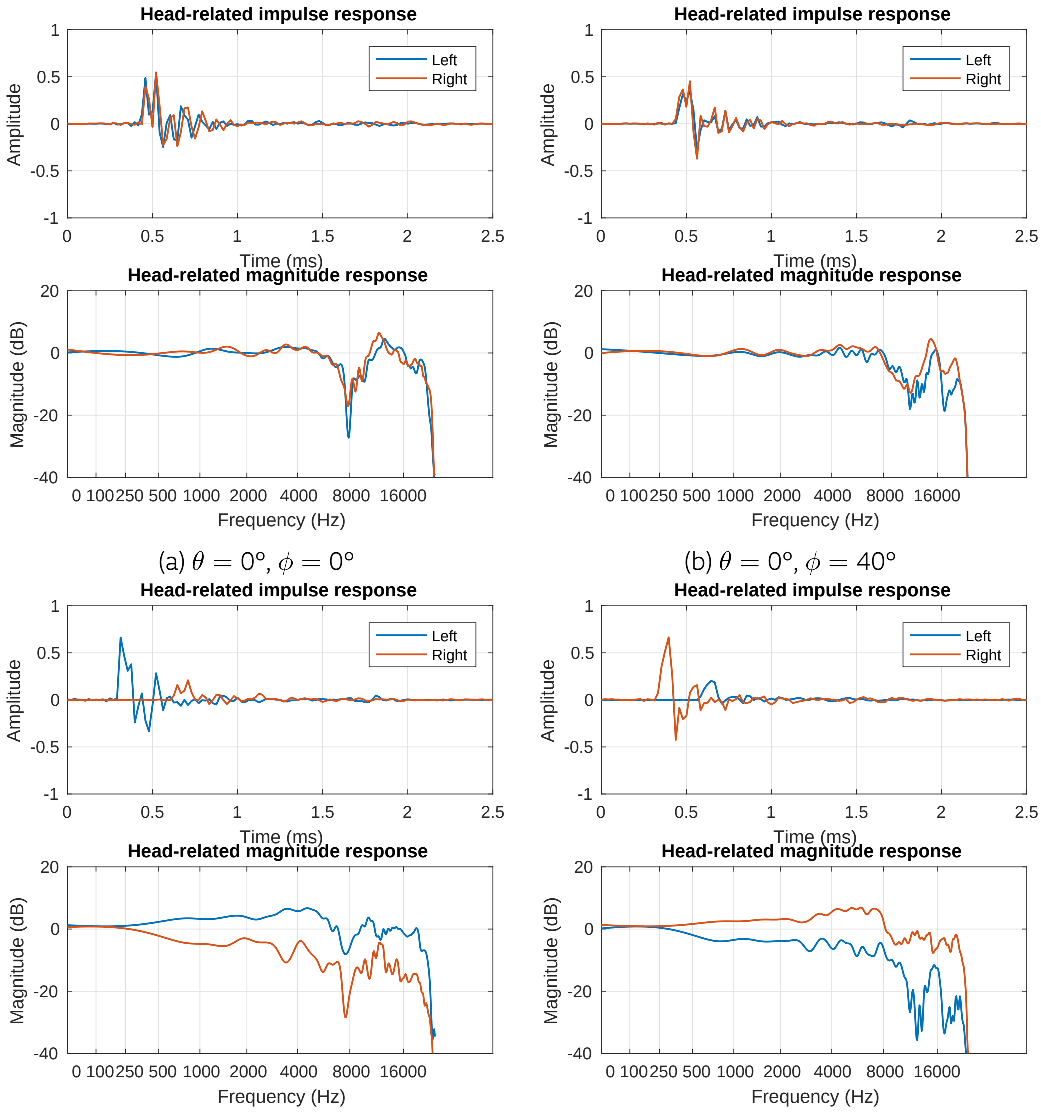
Figure 2.3: Anatomical features of the pinna (Gray, 1918, plate 904).

notches also occur, due to destructive interference caused by the reflection and diffraction of incident acoustic waves within the external ear (the pinna and the ear canal) (Lopez-Poveda and Meddis, 1996). These are typically called *pinna cues* or *spectral cues*. The direction-dependent effects on spectral content of a sound source are interpreted to obtain height information. For sound sources in the median plane, the interaural differences are minimal, and pinna spectral cues are particularly important for localisation (Hebrank and Wright, 1974). They are often thought of as monaural, though the pinnae are not symmetric and spectral cues at each ear differ (Searle et al., 1975).

A set of HRIRs and associated HRTFs is plotted in figure 2.4 for both ears with four different source directions. For a source in front in the horizontal plane, it can be seen that the HRIRs and HRTFs are similar for both ears. When the source is elevated within the median plane, the two ear signals are again similar, but the spectral structure above 4 kHz is different. With lateral displacement of the source, differences in TOA and level can clearly be seen between the ears.

To further illustrate the directional effects in head-related acoustics, the HRIRs and HRTFs can be plotted in two dimensions, according to varying source angles. Figure 2.5 shows measurements at the left ear in the horizontal plane as the source azimuth angle θ is varied. The time-domain onset can be seen to vary with θ, being earliest at around 90° when it is on the ipsilateral side, and latest on the contralateral side at around −90°. This is due to the lateral offset of the ear position from the coordinate system origin at the centre of the head. Amplitude difference can also be seen in the HRIR plot, and HRTF magnitude responses show the frequency-dependent head shadowing effect, with more attenuation at high frequencies on the contralateral side. Since sound reaches the contralateral ear not only by the shortest paths but through all other diffraction paths around the head, the bright spot at 1 ms in the time domain plot for $\theta \approx -90°$ represents the point where the wavefronts along multiple paths tend to reach the ear simultaneously and sum coherently, leading to greater energy than for adjacent directions. This also has effects in the frequency domain. Ripples can be seen in this azimuth region due to comb filtering, most prominently in the range 1 kHz–10 kHz, as the differences in TOA of the wavefronts along these multiple paths vary. At the azimuth of the bright spot in the time domain, the magnitude response is boosted in this frequency region.

Figure 2.6 shows measurements in the median plane. Here the maximum absolute time-domain value is quite consistently at around 0.9 ms, indicating that the centre of rotation is very close to the interaural axis. The grayscale level of the time-domain plot is clipped at ± 0.05 despite the largest absolute amplitude value being scaled to 1, so that the patterns in the later part of the impulse responses (IRs) are visible. Beyond about 1.25 ms several



(a) $\theta = 0°$, $\phi = 0°$

(b) $\theta = 0°$, $\phi = 40°$

(c) $\theta = 40°$, $\phi = 0°$

(d) $\theta = -40°$, $\phi = 40°$

Figure 2.4: HRIRs and HRTF magnitude responses for different source directions. Measured on a Neumann KU100 dummy head microphone at a distance of 3.25 m (Bernschütz, 2013).

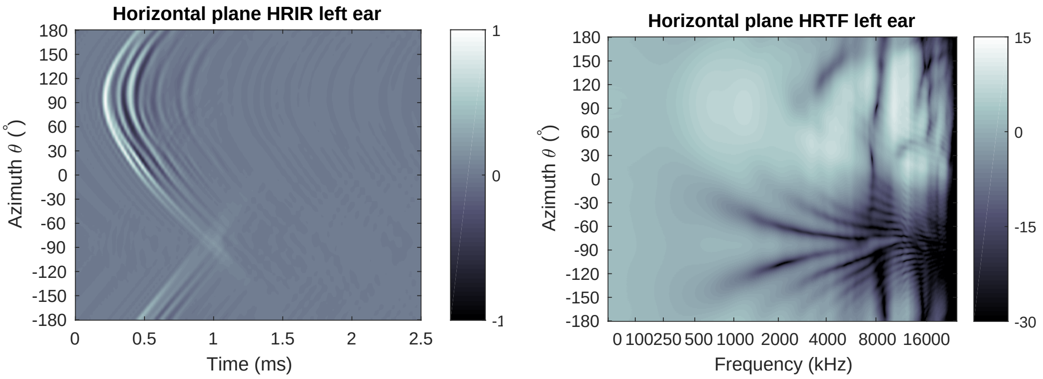


Figure 2.5: Horizontal plane HRIRs and HRTFs measured on a Neumann KU100 dummy head microphone at 1° increments of θ, using data from Bernschütz (2013).

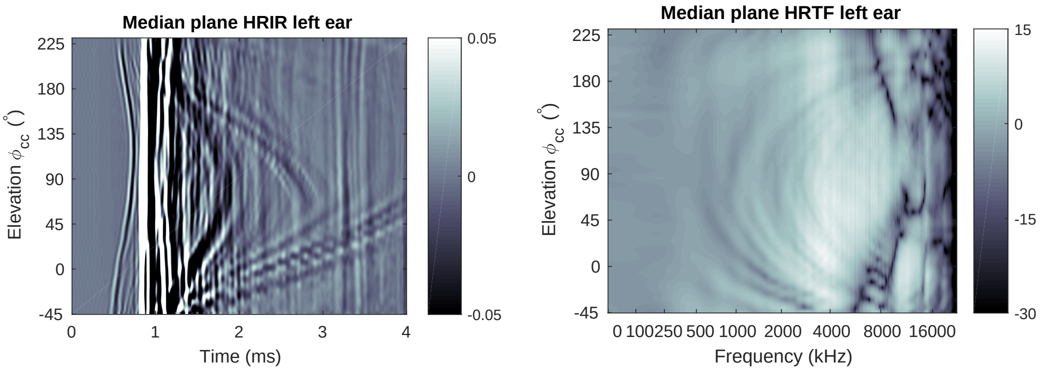


Figure 2.6: Median plane HRIRs and HRTFs measured on a human subject in 5.625° increments of ϕ_{cc}, using data from the CIPIC HRTF database (Algazi, Duda, Thompson, and Avendano, 2001).

arcs can be seen, varying by ϕ_{cc}, and meeting the direct path at angles of around −45° and 225°. The first of these arcs, with its apex at approximately 80° (source above the listener) and 1.9 ms, represents the reflection from the shoulder. The delay from the direct sound indicates a maximum path length difference of approximately 30 cm. As the source gets lower, the path length difference reduces and the shoulder reflection arrives earlier. In the frequency domain this can be seen as a varying comb filter effect in the range 500 Hz–4 kHz. At higher frequencies, the pinna features can be seen, with notches varying smoothly in centre frequency over ϕ_{cc} angle, for instance at around 8 kHz, and above 12 kHz occurring at more specific directions. Spectral peaks can also be seen e.g. at around 10 kHz, particularly for $\phi_{cc} < 0$, and there is another quite consistent peak at 15 kHz.

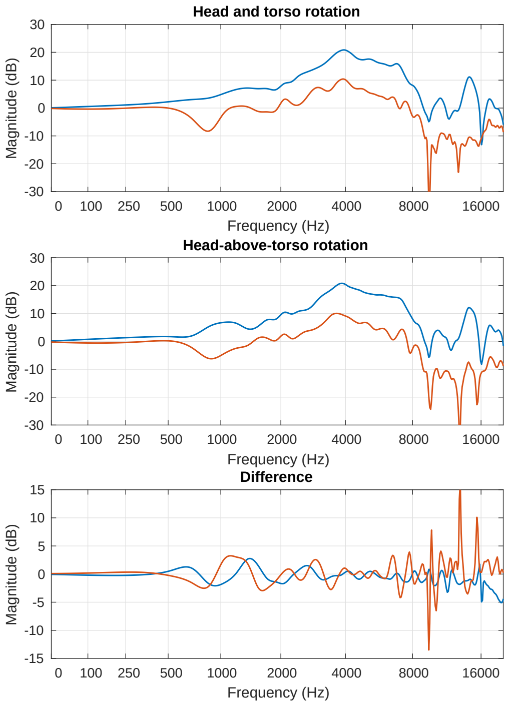


Figure 2.7: Comparison of HRTFs measured with head and torso rotation or head-above-torso rotation of $\theta_h = -40°$, for source position $\theta_s = 0°, \phi_s = 30°$, using data from Brinkmann, Lindau, Weinzierl, Geissler, van de Par, et al. (2017).

Typically HRTFs are measured with the head oriented forwards and only source position varying. Brinkmann, Roden, et al. (2014) showed that head-above-torso rotations yield audible differences to rotation of the head and torso, particularly due to variation in the effect of the shoulder reflection, and torso shadowing for sources at low elevations. An example of the shoulder-influenced comb-filter differences is shown in figure 2.7.

Figure 2.8 indicates the acoustic effects observed for near-field sources, as the source distance becomes less than 1.5 m. Increases in level differences between the two ears can be observed, whilst time differences only increase very slightly. Increased head shadowing is evident at high frequencies, but level difference increases even occur at low frequencies, where wavelengths are large relative to the head size, due simply to path length differences. Also clearly observable at distance 0.25 m is the effect of reflections between the head and the measurement loudspeaker. Brungart and Rabinowitz (1999) studied HRTFs for nearby sources in detail and additionally found that pinna effects are independent of distance until the source is within a few centimetres of the ear.

As might be expected, HRTFs vary individually due to the uniqueness of each person's morphology (Møller, Sorensen, et al., 1995). Common structure can be observed, espe-

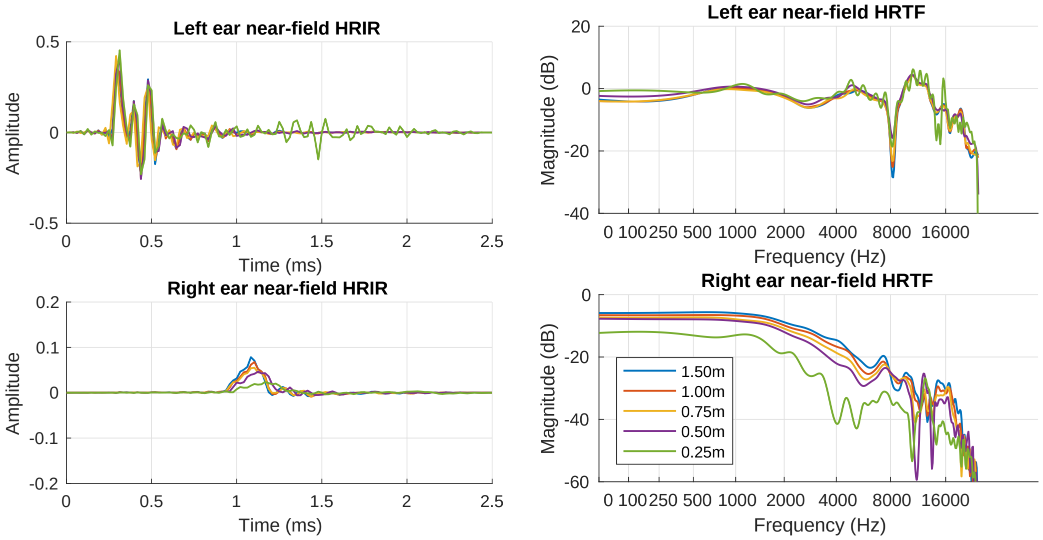


Figure 2.8: Near-field HRIRs and HRTFs measured on a Neumann KU100 dummy head microphone at different distances with direction $\theta = 90°$, $\phi = 0°$, using data from Arend et al. (2016).

cially below approximately 6 kHz, yet there are significant differences in the spectral peaks and notches in the range between 6 kHz and 15 kHz caused by variations in pinna shape. Middlebrooks (1999a) identified that the frequencies of spectral peaks and notches in the HRTFs vary systematically between individuals, and inter-individual differences can be reduced by frequency scaling. It was also shown that the optimal scaling factor between two individuals was highly correlated to head width and pinna size. Figure 2.9 presents the median plane HRTFs for four different individuals, measured at the blocked entrance to the ear canal. Whilst similar patterns can be seen in the spectral structure at high frequencies, there are also clear differences.

The ITDs and ILDs also vary between individuals, primarily due to differing head sizes, as shown in figure 2.10. Here broadband ITDs are estimated using the minimum-phase cross-correlation method (Nam et al., 2008) and broadband ILDs are estimated using the ratio of root mean square (RMS) level of the left and right HRIRs. The ear canal acoustics also vary significantly between individuals, particularly above 2 kHz (Hammershøi and Møller, 1996).

2.2.5 Directional Localisation Cues

Since the early pioneering work of Lord Rayleigh (1907), many studies have since confirmed the duplex theory e.g. (Macpherson and Middlebrooks, 2002). Localisation in the horizontal

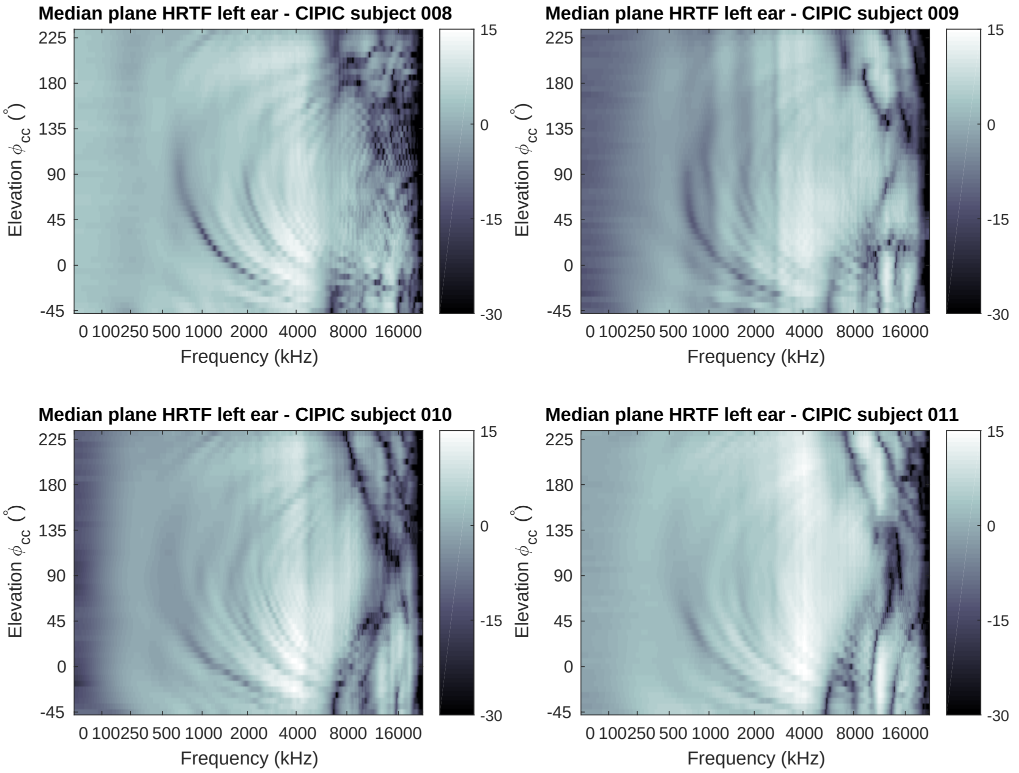


Figure 2.9: Median plane HRTFs measured on several human subjects in 5.625° increments of ϕ_{cc}, using data from the CIPIC HRTF database (Algazi, Duda, Thompson, and Avendano, 2001).

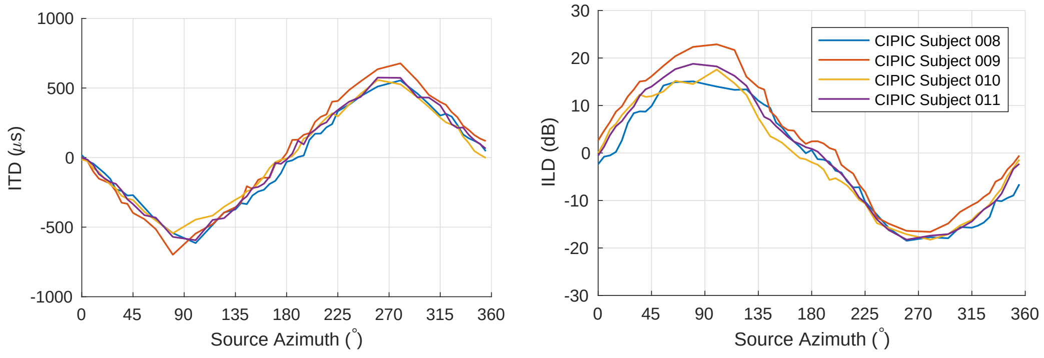
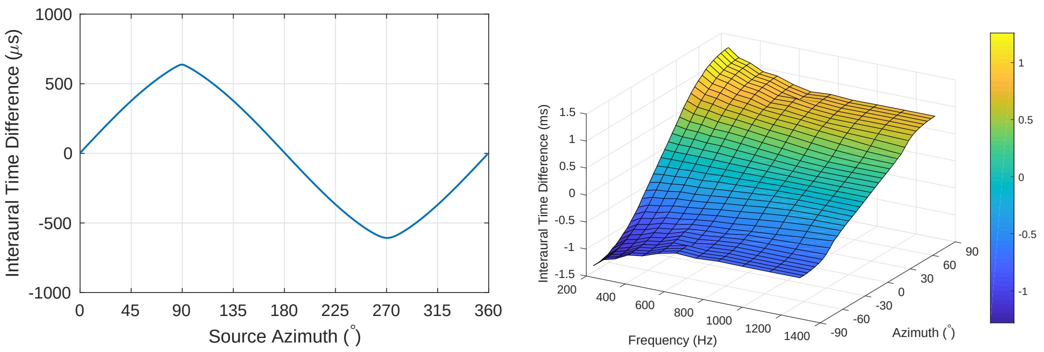


Figure 2.10: Horizontal plane ITDs and ILDs estimated from HRTF measurements for several human subjects in 5° increments of θ, using data from the CIPIC HRTF database (Algazi, Duda, Thompson, and Avendano, 2001).



(a) Based on modelled broadband TOA according to (Ziegelwanger and Majdak, 2014)

(b) Based on auditory model of IPD according to (Dietz, Ewert, et al., 2011)

Figure 2.11: Interaural time differences (ITDs) according to source azimuth angle θ, with positive values indicating arrival at the left ear first. Using HRTF data for the Neumann KU100 from (Bernschütz, 2013) with a source distance of 3.25 m.

plane is indeed informed by ITD and ILD cues in a complementary manner. The relative influence of each is dependent on the spectral and temporal character of the sound source.

The ITD is generally utilised below approximately 1500 Hz, where phase differences between the ears within auditory bands are detected, though at higher frequencies interaural differences in the temporal envelope also provide a cue to localisation (Blauert, 1997). ITDs vary from 0 μs to approximately 600 μs–700 μs. Kuhn (1977) proposed simple approximation formulae for the ITD in the horizontal plane, based on a spherical head model. More recently Ziegelwanger and Majdak (2014) presented an approach for fitting a spherical head model for TOAs to measured HRTF data. Figure 2.11a shows modelled broadband ITDs for the Neumann KU100 derived via the Ziegelwanger and Majdak off-axis model, with initial TOA estimation carried out according to Nam et al. (2008). Below 1500 Hz the auditory system detects ITD using interaural phase differences (IPDs) within auditory frequency bands. Figure 2.11b shows the frequency-dependent ITD determined via a model of the peripheral and binaural auditory processes (Dietz, Ewert, et al., 2011). It can be seen that the ITD increases below 700 Hz. The just noticeable difference (JND) level of the ITD is dependent on the nature of the stimulus and the ITD itself. For the most sensitive scenario, with a broadband source with strong temporal envelope near the median plane it is as low as 10 μs–20 μs (Hafter and De Maio, 1975).

The ILD is highly frequency-dependent, since it is primarily caused by the shadowing effect of the head, which only occurs when the wavelength is small enough relative to the head. At lower frequencies diffraction means that the ILD is negligible with a distant sound source. The azimuth- and frequency-dependence of the ILD can be seen clearly in figure 2.12, which

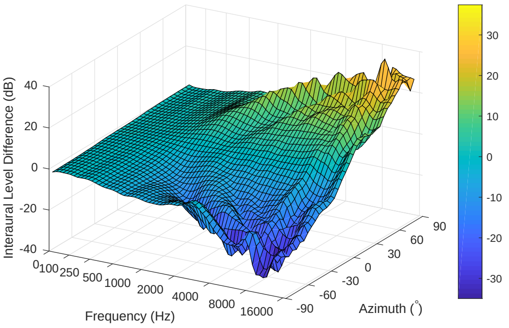


Figure 2.12: Interaural level differences (ILDs) according to source azimuth angle θ, with positive values indicating greater intensity at the left ear. Using HRTF data for the Neumann KU100 from Bernschütz (2013) and the auditory modelling of Dietz, Ewert, et al. (2011).

was derived using the auditory model of Dietz, Ewert, et al. (2011). Below about 400 Hz the ILD is close to zero, but the dependence on azimuth angle increases at higher frequencies, reaching greater than 20 dB above 3 kHz. The rate of change of ILD with changing source azimuth is greatest near the median plane and lowest in lateral directions. Weiping et al. (2010) measured JND of the ILD with sinusoids, showing that it is lowest (1 dB–3 dB) when ILD is close to zero and the frequency is below 2 kHz. At higher frequencies and higher ILDs, the JND is higher, in the range 3 dB–7 dB. As discussed in section 2.2.4, the ILD is increased for nearby sources, particularly with distances of less than 1 m, and this effect also occurs at low-frequencies. ILD cues are therefore utilised even at low frequency, primarily as a distance cue.

The cones of confusion are conical surfaces[2] originating at the ears and symmetric around the inter-aural axis i.e. with a constant θ_{cc}. For all points on these surfaces, the ITD and ILD cues are assumed to be constant, due to approximate symmetries of the head and ear geometry i.e. assuming a spherical head model with symmetric ear positions. Discrimination of source direction within these regions is challenging. Sources will often be localised near to the cone of confusion, but at the wrong location on it, most commonly mirrored in the frontal plane i.e. a *front-back confusion* (Makous and Middlebrooks, 1990). Other localisation cues besides the ITD and ILD are utilised to resolve source location within a cone of confusion, particularly the pinna spectral cues and the use of head motion. The cones of confusion only exist for sources of distance more than approximately 1 m. Due to the aforementioned near-field effects for ILD, closer sources have a reduced region of ambiguity that is approximately torus-shaped (Shinn-Cunningham, Santarelli, et al., 2000).

Interaural coherence (IC) is a measure of the similarity of signals at the ear canals, given

[2] More accurately they are hyperbolic surface contours of isometric ITD and ILD.

by the maximum of the interaural cross-correlation function (Faller and Merimaa, 2004). It provides a cue to the perceived extent of auditory sources and also the sense of envelopment in reverberant environments. IC is highest when listening to a single source sound in anechoic conditions, where point-like auditory events are perceived, and low IC occurs in a diffuse field or due to multiple sound sources coming from different directions. Faller and Merimaa (2004) utilise IC to indicate where there is insufficient information present in a given frequency band to estimate the direction of arrival (DOA) from ITD and ILD. In such situations, an auditory impression of spaciousness is given (Blauert and Lindemann, 1986). IC is frequency-dependent, below about 400 Hz it is high even in a diffuse field due to the long wavelength relative to head size. Low IC leads to random fluctuations in ILD and ITD over time and frequency (Goupell and Hartmann, 2007). It is not known whether the auditory system directly detects IC or instead these fluctuations affect localisation.

Since spectral cues occur at high frequencies (above 4 kHz), they are most effective with broadband signals. Elevation localisation is based on a learning process whereby the individual pinna spectral features are associated with source directions (Hofman et al., 1998). The relationship between pinna morphology and spectral features has been studied by means of numerical calculation of HRTFs under various morphological perturbations (Tew et al., 2012; Mokhtari et al., 2015, 2016).

2.2.6 Directional Localisation Acuity

With broadband sound sources in a free field, directional localisation is at its most accurate and precise. Variability in localisation responses is often called *localisation blur*. Blauert (1997) reviews the results of the many tests on localisation accuracy and blur, including results from a large-scale test by Haustein and Schirmer (1970) assessing localisation blur in the horizontal plane, with a white noise source and a static head. Results show the lowest mean localisation blur of ±3.6° in the frontal direction, slightly increased blur to the rear (±5.5°) and the highest (±10°) in lateral directions. Mills (1958) investigated localisation precision using JNDs in source position, otherwise known as minimum audible angles (MAAs). It was found that MAAs range from 1°–10° depending upon the direction of incidence, for tones with frequencies below 1 kHz. The lowest MAAs were shown for sources in the frontal region. At higher frequencies, localisation becomes highly ambiguous for lateral source positions.

Median-plane polar angle localisation is biased towards the horizontal plane in the frontal hemisphere and elevated in the rear (Damaske and Wagener, 1969). Localisation blur is greater for elevation than for horizontal localisation, though it is decreased when listeners are able to move their head (Perrett and Noble, 1997). Makous and Middlebrooks (1990) evaluated directional localisation in 3D using 250 ms noise-burst stimuli. Listeners rotated to report the perceived source location by pointing with their nose. In the frontal region, sources were localised with mean signed errors of less than 2° and standard deviations of approximately 3°. Standard deviation in responses (localisation blur) increased with source azimuth and were greatest for sources located outside the field of vision, though generally still less than 10°. Elevation errors were lower for sources located laterally, away from the median plane, whereas azimuth errors were greater for lateral directions. Middlebrooks (1992) showed, using a similar method, that with narrowband filtered noise bursts the perceived elevation is strongly biased by the frequency content, and that this is correlated with individual pinna spectral features. Acuity of elevation localisation is generally quite affected by the stimulus, including due to level, duration, and spectral structure (Macpherson and Middlebrooks, 2000, 2003; Macpherson and Sabin, 2013).

Localisation accuracy and precision varies individually and is influenced by the listening conditions, including nature of the source signal and the acoustic environment, as well as by the reporting method (Letowski and Letowski, 2011). Binaural hearing leads to significantly greater accuracy and precision than monaural hearing (Kohlrausch et al., 2014).

2.2.7 Head Movement During Localisation

Head movement has been shown to resolve front-back confusions, since the interaural cues will change in opposite directions depending whether a source is in front or behind (Wallach, 1939). With long stimuli, head movements can be used to locate the azimuth of a source accurately by pointing ones head towards it, therefore minimising the interaural differences and using the region of lowest localisation blur (Makous and Middlebrooks, 1990). Head movements also aid in resolving vertical position, improving pinna cues and making use of the rate of change in ITD due to head rotation (Perrett and Noble, 1997; Ashby et al., 2013). The results of McAnally and Martin (2014) suggest moderate head movements (16°–32°) are required to improve elevation localisation. If the source contains high frequencies then pinna cues dominate vertical localisation (Macpherson, 2013).

Blauert (1997, p.178) describes two classes of head movement during spatial hearing: reflexive movement towards the position of an auditory event and searching movements taken more or less consciously. Reflexive movements are made if the auditory event is already located fairly well. Searching movements are made to gather more information to establish the final position of an auditory event that is not initially well localised. Thurlow et al. (1967) found that head yaw rotations are much more common during sound localisation than pitch or roll, and are also larger in magnitude. Listeners often move towards a sound source initially during localisation, although not generally to the extent that they are directly facing it; back and forth changes in movement are also common, particularly in the case of yaw rotations. Kim, Mason, et al. (2013) showed that patterns of head movements are dependent on listening activity and vary between individuals. Listeners moved their heads to a greater extent when assessing source width and envelopment compared to source direction or timbre. Real life activities of watching movies and concerts and playing video games showed frontal focus compared with analytical spatial hearing tasks. As in earlier studies yaw rotations were larger in magnitude than pitch or roll. It is clear that frequent head movement occurs even when in a limited range. The upper quartile of individual maxima and lower quartile of individual minima of head rotations were approximately: ±60° for yaw and ±30° for pitch and roll. For head translation the equivalent results were: ±25 cm ±15 cm, ±10 cm in the x, y and z axes respectively. Small head movements of up to 5° have been shown to occur even when the listener is attempting to remain stationary (König and Sussmann, 1955). It appears that localisation accuracy, as measured using MAAs, is greater during head movements than during source movements (Brimijoin and Akeroyd, 2014). This suggests that a comparison occurs between auditory spatial cues and self-motion cues during auditory event localisation, though this motion compensation is not perfect (Freeman et al., 2017).

2.2.8 Multiple Sources and Reverberant Environments

Most real-world listening scenarios are not anechoic and involve multiple sound sources. Both of these scenarios tend to reduce our localisation acuity. Perrott (1984) showed that when there are concurrent sources, the MAA is often increased. Santala and Pulkki (2011) showed that our ability to determine the spatial distribution of (mutually incoherent) sources is quite limited compared to localising a single source.

Localisation acuity is degraded in reverberant environments (Hartmann, 1983), though with impulsive sounds it is not so affected. Within echoic acoustic environments, we receive not just the direct sound from sources but reflections and reverberation due to scattering off other surfaces. Despite this complex input, we are still able to perceive the position of sound sources with reasonable accuracy. This is due to the precedence effect (Litovsky et al., 1999), an auditory process which suppresses the effect of early reflections following the onset of direct sound in localisation.

For a sound arriving within approximately the first 1 ms after the first onset which is coherent, summing localisation occurs and a single auditory event occurs with the perceived direction of arrival shifted according to the relative level and time differences. When the time-lag is greater than 1 ms, the precedence effect occurs and the auditory event is generally localised in the direction of the leading sound source. Only when the time differences extend beyond a certain value are two separate auditory events perceived, this is known as the *echo threshold*. The echo threshold varies from about 1 ms for impulses, to 50 ms for speech and 80 ms for classical music (Kohlrausch et al., 2014). These auditory fusion effects are also dependent on level differences, as well as the nature of the source signal (Blauert, 1997, section 3.1). With reflections at levels lower than the direct sound, changes in tone colour are observed before separate echo detection. Reflections lead to reduced IC, and appear to influence sense of distance and source extent, as well as a sense of spaciousness (Rakerd and Hartmann, 2010; Blauert and Lindemann, 1986).

2.2.9 Distance Perception

Auditory distance perception has been less widely studied than directional localisation, but reviews are given by Zahorik, Brungart, et al. (2005) and Kolarik et al. (2016). Distance acuity is generally poorer than directional acuity. Typically sound source distance tends to be underestimated, especially for distances of more than 3 m, except for near-field sources where it is often overestimated. Distance perception is most accurate for distances at around 1 m. These systematic biases have been measured across a range of stimuli, acoustic environments and psychophysical measurement procedures, and can be well approximated by a compressive power function. Auditory distance perception is also imprecise, especially at large distances. Zahorik, Brungart, et al. (2005) reported standard deviations as high as 20 %–60 % of source distance. Visual and audiovisual distance perception are more precise and less biased (Anderson and Zahorik, 2014). Distance perception is more accurate for lateral sources (unlike direction) (Kopčo et al., 2011).

There are several different cues to the distance of auditory events, some are absolute cues, whereas others are relative, allowing discrimination of similar sounds at different distances. The role of cues differs depending on whether sources are nearby or distant. Accurate distance localisation is particularly important for near-field sources in the peripersonal space, where a rapid motor response may be required.

The primary cues are the intensity of the sound event at the ears and the direct-to-reverberant ratio (Mershon and King, 1975). Intensity at the ears is a relative distance cue since intensity at the source is unknown. Sources at different distances with same intensity can be perceived with constant loudness, which Zahorik and Wightman (2001) suggest may be due to constant reverberant energy. The direct-to-reverberant sound energy ratio is another important distance cue, it decreases with increasing source distance. The direct-to-reverberant ratio (DRR) enables absolute distance judgements (Mershon and King, 1975). When cues are isolated, intensity generally enables more accurate distance localisation than DRR, though accuracy is typically higher when both cues are available (Zahorik, Brungart, et al., 2005). The JND for DRR depends upon the DRR itself and the source bandwidth (Larsen et al., 2008). We are most sensitive (2 dB–3 dB) to changes in DRR with a broadband noise stimulus around the critical distance, where DRR is 0 dB, and less sensitive for nearer or further source positions which have higher and lower DRR.

A number of additional cues are utilised. Spectral structure provides a distance cue for sources greater than 15 m away and in the near-field. For distant sources, air-absorption leads to high frequency attenuation, and such spectral structure leads to increased perceived distance. For nearby sources, spectral content at low frequencies provide a useful distance cue, both due to increased ILDs (Brungart, 1999a) and frequency-dependent DRR (Kopčo et al., 2011). Temporal fluctuations in ILDs at mid- to high-frequencies also provide a cue to distance, being greater for distant sources (Catic et al., 2013). Listener movement is thought to provide auditory distance cues through the rate of change in level (known as acoustic tau) and motion parallax (Ashmead et al., 1995), though these cues are likely fairly weak. Familiarity with the stimulus and the acoustic environment allows improved distance perception, since a listener is able to interpret relative cues of spectral structure and intensity against an internal reference informed by their previous experience. This is particularly evident for speech signals where perceived speech effort (e.g. of whispering versus shouting) influences distance estimations for sounds of the same intensity (Brungart and Scott, 2001). Learning processes have been observed for improving distance estimation in reverberant environements (Shinn-Cunningham, 2000).

2.2.10 The Influence of Vision

Vision also influences localisation of sound events, a process commonly called *visual capture*, which has been shown to affect localisation of both direction (Jack and Thurlow, 1973) and distance (Mershon, Desaulniers, et al., 1980). Alais (2004) show that audiovisual localisation is integrated, with capture working both ways depending on which sense provides stronger cues, and that bimodal localisation usually gives greater accuracy than with either sense alone.

2.2.11 Other Benefits of Binaural Hearing

Binaural hearing provides benefits in segregating auditory streams, with concurrent multiple sound sources, a process called *binaural unmasking*. With competing speech signals this is often called the *cocktail party effect* (Bronkhorst, 2000). The effects of binaural unmasking are measured using the binaural intelligibility level difference (BILD), where equivalent intelligibility has been shown for 2 dB–8 dB increase in the level of distractor speech signals, depending on their number and position. Binaural hearing also leads to reduced perceptual effects of colouration (Brüggen, 2001) and reverberance (Tsilfidis et al., 2013) due to environmental reflections.

2.2.12 Summary

The acoustic scattering off the torso, head and external ear affects the sound that reaches the ear canals in ways that are dependent on the direction and distance of the sound source. These effects can be captured in the head-related transfer functions (HRTFs) and provide cues used by the auditory system in localisation. These cues, particularly interaural differences and spectral features, allow an internal representation of the external acoustic environment to be constructed i.e. the auditory scene. This process is influenced by reverberation, dynamic movements and vision. Generally it has a remarkable accuracy and binaural hearing provides significant benefits to understanding the world around us.

2.3 Binaural Recording

Binaural recording is the process of recording the sound pressure at the two ears of a listener using a pair of microphones. This will capture the scattering of external sound sources on the body, head, and ears of the listener, and so encoding the binaural and monaural cues that are interpreted by the auditory system in spatial hearing. In early binaural technology, artificial listeners were used to make recordings. Microphones were placed at the ears of a manikin, aiming to produce life-like recordings. So-called dummy head microphones or head and torso simulators (HATSs) are still often used in binaural technology today, both for artistic recording[3] and in research. Binaural recording in the ears of a human listener is also common.

[3] An example from the BBC at the start of the present project is available here: http://www.bbc.co.uk/programmes/p015njlg

2.3.1 Historical Development

According to Wade and Detutsch (2008), the term *binaural* was first coined by Alison in 1861 to describe the use of two ears in hearing. Until the 1970s binaural was often used to describe techniques involving recording and reproduction of two signals destined for the two ears (so including what we now know as stereo), as opposed to the modern definition of recording and reproducing the correct signals in the ear canals. The history of binaural recording technology is reviewed by Paul (2009). As audio technology and understanding of spatial hearing developed over time, obtaining binaural recordings in line with this modern definition became more feasible.

The origins of binaural recording can be linked to the *théâtrophone* (du Moncel, 1881). In 1881, just five years after the invention of the telephone, it was presented to the public at the International Exhibition of Electricity in Paris. Concerts from the Paris Opera were transmitted to listeners at the Exhibition Palace via two-channel telephone signals, using pairs of microphones widely spaced on the stage and pairs of telephone receivers at the remote listening site. The experience, providing interaural differences to the listeners, was novel to audiences and was received positively. A report from the time in Scientific American and republished in (Hertz, 1981) stated: “As soon as the experiment commences the singers place themselves, in the mind of the listener, at a fixed distance, some to the right and others to the left. It is easy to follow their movements, and to indicate exactly, each time that they change their position, the imaginary distance at which they appear to be.”

Later, in the early 1930s, Harvey Fletcher and colleagues at Bell Laboratories utilised understanding of binaural hearing to develop stereophonic recording and loudspeaker reproduction techniques. They also experimented with the use of a dummy manikin made from wax, with microphones placed in its cheeks and accompanying reproduction by headphones (Hammer and Snow, 1932). These experiments included musical recordings of the Philadelphia Symphony Orchestra. The dummy, named “Oscar”, was presented to the public at the 1933 World Fair in Chicago, where it was placed in a glass-fronted box with a demonstrator walking around it whilst speaking, meanwhile the public listened to the microphone signals using headphones. A convincing sense of direction and distance was achieved for the listener, although it was acknowledged that the visual cues were needed to achieve best effect. At approximately the same time in the United Kingdom, Blumlein (1931) developed stereophonic recording and reproduction techniques based on the principles of binaural hearing.

Over the subsequent decades, researchers gradually developed more representative systems and, as microphone technologies improved in fidelity and reduced in size, accurate binaural recording and measurement became more viable e.g. (Mills, 1958). A significant step forward was made with the introduction of the Knowles Electronic Manikin for Acoustic Research (KEMAR) in 1972[4], which was intended to provide a reproducible design based on representative anthropometric measurements (including the pinnae) and even modelling the acoustics of the ear canal (Burkhard and Sachs, 1975). The KEMAR was initially designed to enable in-situ evaluation of hearing aids, but was also used for measurement of headphones and hearing protection (Burkhard, 1978). It was later standardised for measurement of hearing aids in IEC TR 60959 (1990).

In 1973, the Neumann microphone company in Germany released the KU80 dummy head microphone, based on earlier work by Kürer et al. (1969). It was showcased at the IFA (International Broadcasting Fair) in Berlin that year. In conjunction, the first radio drama to be created using binaural recording was broadcast: "Demolition"[5], which was produced as a special commission by RIAS station. The first binaural radio drama broadcast in the United Kingdom was "The Revenge" (1978)[6], a 24 min long play without dialogue, written and performed by Andrew Sachs of Fawlty Towers fame, which involved a range of outdoor location recordings.

The KU80 was equalised for a free-field frequency response. Later the KU81 model was developed in collaboration with the Institut für Rundfunktechnik, the research centre for German-language broadcasters, and was equalised for a diffuse-field frequency response (Peus, 1985). This made it better suited to artistic recording and also more compatible with loudspeaker reproduction.

2.3.2 BBC Evaluation of the Neumann KU81

An internal British Broadcasting Corporation (BBC) Research technical report from this time evaluated the KU81 microphone (Meares and Taylor, 1982). It noted excellent quality except for the lack of frontal sound images for any sources located within ±30° of the front, though it was stated that it was a significant improvement over the previous KU80 model. They concluded that it "gives excellent binaural effects and at the same time entirely acceptable sound quality with loudspeaker reproduction." Møller (1992) cites lack of mono compatibility as a limitation in broadcast applications, though Meares and Taylor (1982) reported that sound quality for mono loudspeaker reproduction with the KU81 was "entirely acceptable".

[4] As stated at `http://kemar.us/`

[5] *Demolition* was recently rebroadcast by Bayerischer Rundfunk: `https://www.br.de/radio/bayern2/programmkalender/sendung-2111916.html`

[6] *The Revenge* was last aired in February 2016 by BBC Radio 4 Extra: `https://www.bbc.co.uk/mediacentre/proginfo/2016/07/the-revenge`

Meares and Taylor (1982) reported that test recordings were made in BBC studios of both musical and dramatic performances. Challenges with the binaural recording technique in general were noted. Production techniques by this time, especially for music, involved balancing multiple microphones and artificial reverberation for aesthetic control, yet this was not possible with the dummy head without corrupting the spatial image. The only aesthetic control available was the choice of placement of the microphone, and this enabled only a naturalistic perspective. For drama production, a different challenge was noted, use of sound effects required them also to be binaurally recorded.

2.3.3 Current Binaural Recording Systems

Currently there are a number of dummy head microphones and HATS commercially available. They target a variety of purposes, from measurement and research to artistic recording. The Neumann KU100 superseded the KU81 in the late 1990s as a diffuse-field equalised recording microphone. A number of measurements systems are available for research and development, following standardised specifications e.g. (IEC TS 60318, 2017), although still with different morphology. The KEMAR has been updated and is sold by G.R.A.S., whilst Brüel & Kjær offer the 4128 HATS, and HEAD Acoustics offer the HMS IV head-and-shoulder measurement system. Further developments of binaural microphone systems have been made by researchers e.g. (Christensen, Jensen, et al., 2000), including those with the capability to rotate the head above the torso e.g. (Lindau and Weinzierl, 2006).

A number of in-ear binaural microphone systems are available for recording on humans, both for measurement purposes (e.g. HEAD acoustics BHM III or Brüel & Kjær 4101) and for artistic recording (e.g. Soundman OKM II). Recently, systems designed for use with mobile phones have been emerging, targeted at the consumer market, such as the Sennheiser Ambeo Smart Headset. This provides earphones with external microphones, and offers binaural recording, as well as active noise cancellation and “transparent hearing” mode, where external sound is passed through. This potentially enables augmented reality applications, as described by Härmä, Jakka, et al. (2004).

2.3.4 Pros and Cons of Binaural Recording

With binaural recording, natural sound scenes can be recorded accurately, and the complexity of the scene has no impact on the complexity of the recording or reproduction process. This includes the environmental reverberation, with reflections from complex geometry. Dynamic sound source movements can also be captured and the changes in spatial cues will be physically correct.

However, one of the main limitations of binaural recording is that it can only capture naturally occurring scenes. This is particularly of concern in broadcast and media entertainment scenarios where sound scenes are regularly created which do not exist, or even cannot exist. Even for naturally occurring scenes, to be completely accurate for an individual, they must be present for the recording, with microphones in their own ears. Even then, the binaural recording method also prevents the reproduced sound from responding appropriately to the listener's head movement, which is an important aspect of spatial hearing.

One additional aspect worth noting briefly is that there have been a number of studies into binaural recording techniques that allow the listener to rotate their head during playback, with appropriate dynamic updates to the audio signals (Algazi, Duda, and Thompson, 2004; Algazi, Dalton, et al., 2005; Hom et al., 2006; Lindau and Roos, 2010; Meier et al., 2011). These techniques, often called *motion-tracked binaural recording*, use spherical arrays of similar diameter to a human head, with an array of microphones placed on the equator. In reproduction, dynamic interpolation is performed between the recorded microphone signals according to listener head orientation. Another more common means of creating dynamic motion-tracked binaural signals is discussed in the following section: binaural rendering.

2.4 Binaural Rendering

Binaural rendering, also known as binaural synthesis, is the processing of an input audio signal to give a realistic auditory spatial impression by simulating the binaural signals at the ears. The source signal is processed with filters describing the acoustic transfer function from the desired source position to the ears. The auditory event triggered by this rendering is often called a *virtual sound source*, since it is a simulation of a real sound source at target position. An *auditory virtual environment (AVE)* can be constructed, by simulating multiple sound sources at various directions and distances, and modelling environmental acoustic effects or otherwise using transfer functions measured in a reverberant space. Binaural rendering systems can be made dynamic and interactive, updating in real-time according to movement of both sources and the listener.

Binaural rendering provides advantages over binaural recording in that auditory scenes can be produced that are not based on a real acoustic scene, which is useful for creative applications. Binaural recordings may also be augmented with binaural rendering. In addition, binaural rendering provides a powerful means for investigating the mechanisms of spatial hearing, precisely controlling the acoustic input to the auditory system.

Whilst earlier research had used binaural recordings to explore binaural hearing mechanisms e.g. (Plenge, 1974), the effects of the headphone reproduction were not corrected.

It was in the late 1980s that digital signal processing (DSP) techniques were first applied in order to achieve accurate binaural rendering. An interactive binaural rendering system was developed as a collaboration between the NASA Ames Research Centre and Crystal River Engineering (Wenzel, Wightman, et al., 1988). As part of this project, researchers at the University of Wisconsin-Madison were contracted to make HRTF measurements and conduct psychoacoustic validation experiments[7]. This led to two seminal publications in the field of binaural rendering (Wightman and Kistler, 1989a,b).

The development of interactive binaural rendering systems has been enabled by a number of key technologies and techniques: advancements in DSP units, efficient low-latency convolution algorithms (Gardner, 1995), accurate impulse response measurement techniques (Müller and Massarani, 2001), motion tracking technologies (reviewed by Hess (2012)), environmental acoustic simulation algorithms (Kendall and Martens, 1984), and HATSs with automated rotary control (Lindau and Weinzierl, 2006). This section reviews the fundamentals of binaural rendering technology, whilst the following sections explore the state-of-the-art of various aspects of the technology.

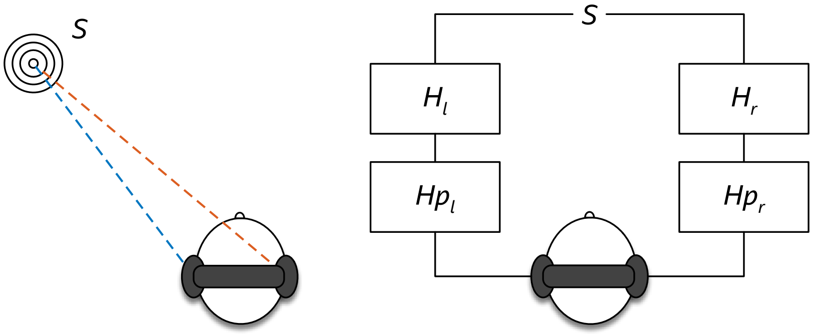
2.4.1 Free-Field Rendering

The simplest rendering scenario is free-field listening, where the HRTFs represent the anechoic transfer function from a source to a listener's eardrums for a given spatial configuration (section 2.2.4). For rendering of a single virtual sound source, a monophonic source signal is convolved with the two-channel HRTF for the desired virtual source position, the resulting two-channel signal represents the acoustic pressure signals that would be present at the ears for such a sound source in a free field. Multiple virtual sources can be rendered in this way, and the binaural signals combined by summation. Reproduction of the binaural signals over headphones, with the appropriate equalisation to correct for the headphone-to-ear transfer function (HpTF), can produce appropriate auditory events corresponding to the intended virtual scene. Figure 2.13 illustrates this process for a single source and figure 2.14 for two sources.

The HRTFs can be defined according to the three-dimensional position of an ideal point source relative to the origin at the centre of the head, and dependent upon the three-dimensional rotation of the head above the torso. The complex-valued sound source pressure s generates the sound pressure p at each ear via the transfer function h according to:

$$p_{lr}(\gamma, \psi, \zeta, f) = h_{lr}(\theta, \phi, r, \gamma, \psi, \zeta, f).s(\theta, \phi, r, f). \tag{2.5}$$

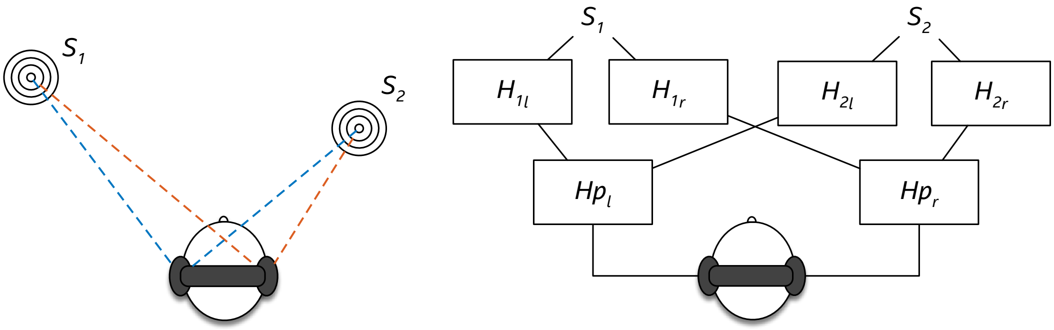
[7] As reported by Begault, Wenzel, Godfroy, et al. (2010)



(a) A virtual sound source

(b) Rendering using a HRTF and HpCF.

Figure 2.13: Binaural rendering of a virtual sound source in the free-field for headphones



(a) Two virtual sound sources

(b) Rendering using two HRTFs and a HpCF.

Figure 2.14: Binaural rendering of two virtual sound sources in the free-field for headphones

where θ is the source azimuth, ϕ is the source elevation, ρ is the source distance, γ is the head yaw, ψ is the head pitch, ζ is the head roll, ω is the angular temporal frequency, and subscripts $_l$ and $_r$ indicate the left and right ear respectively. Clearly even with these six spatial dimensions this is a simplification of natural dynamic hearing, with complex sound sources and natural listener movement. But simpler representations are often used. It is commonly assumed that distance can be represented purely by frequency-independent intensity and time delay changes, which is valid for source positions beyond approximately 1 m, and so HRTFs measured at a fixed distance in the far-field are used. It is also often assumed that head rotation is equivalent to the opposite source rotation, so measurements with constant head-above-torso rotation are used. This leads to HRTFs within the space $h_{lr}(\theta, \phi, f)$.

The headphone-to-ear correction filter (HpCF) is given by Hp_{lr}. It is a two-channel filter designed to correct for the HpTF, which is a LTI representation of the coupling of the headphone to the ear during playback. This allows more accurate reconstruction of the target pressure signals p_{lr} at the ears. More detailed discussion of measurement and equalisation for binaural rendering is given in subsequent sections.

The principles of the free-field binaural rendering technique were first detailed and validated by Wightman and Kistler. Wightman and Kistler (1989a) presented the technique of measuring the HRTF from the source to the eardrum using miniature electret probe microphones. HRTFs were measured at 144 source directions for 10 human subjects, with their heads always orientated forwards. The HpTFs were also measured whilst the subjects wore headphones. Impulse responses were measured using a noise-like source signal, and HRTFs were corrected for the HpTF by frequency-domain division, which will have also corrected for the measurement microphone response. The HpTF-equalised headphone playback of the HRIR was recorded at the ear using the probe microphone and then compared in the frequency-domain to the measured HRTF. Results showed magnitude errors under 2 dB and phase errors of less than 10° up to 14 kHz.

Wightman and Kistler (1989b) then presented a psychoacoustic validation of this binaural rendering process in terms of directional localisation. Eight listeners were presented with broadband noise bursts from both real loudspeakers and headphones via binaural rendering. The listeners were instructed to keep their heads still. After the stimulus they indicated localisation by calling out numerical estimates of the azimuth and elevation angles. The localisation of binaurally rendered stimuli was very similar to that of real loudspeaker stimuli, in terms of both accuracy and precision. However, an increased number of front-back confusions was observed and bigger differences were present in elevation judgements than azimuth judgements.

2.4.2 HRTF Measurement

To measure an HRTF, a pair of microphones is placed in each ear. A loudspeaker is placed at the desired position, normally in the far field, and IRs are measured from the loudspeaker to each microphone. To avoid substantial environmental reflections these measurements are normally carried out in an anechoic chamber, giving an approximation of a free-field. The exponential sine sweep technique for IR measurement is commonly used, since it can provide a high signal-to-noise ratio (SNR) and allows separation of non-linear distortion components (Müller and Massarani, 2001; Farina, 2007).

Møller (1992) gives a detailed discussion of the measurement and equalisation techniques for accurate simulation of the pressure signals at the eardrums. Probe microphones can be used to record signals within the ear canals without significantly disturbing the sound field or risking harm to the subject. The sound pressure signals within the ear canal are not dependent upon the direction of the sound source, and so any point along the canal can justifiably be used for measurement. This was further studied and verified by Hammershøi and Møller (1996). However, the ear canal will exhibit standing waves which can interfere with measurements at higher frequencies if the microphones are placed more than a few millimetres from the ear drum. Therefore measurements are generally made close to the ear drum or at the entrance to the ear canal. Measurements made at the ear canal entrance with the ear canal blocked have been shown to reduce individual variations due to the acoustic effects of the ear canal (Møller, Sorensen, et al., 1995). Directional information is retained even if recording at the blocked entrance to the ear canal, which presents practical benefits, enabling use of larger microphones and reducing the risk to the human subject. It also means that dummy head microphones can justifiably be made without simulated ear canals. The HRTF measurements include the effects of the measurement system and should be appropriately equalised, as discussed in section 2.4.3.

Hiipakka et al. (2012) present an alternative method for measuring the HRTF at the ear drums. Instead of probe microphones, pressure-velocity sensors are used at the entrance to the ear canal. It is shown that this can yield more accurate binaural rendering of the ear drum pressure signals than blocked ear canal measurements, and Takanen et al. (2012) show that this gives lower audible colouration.

HRTFs are typically measured from a large number of directions around the listener, to allow high-quality binaural rendering of spatial scenes or else detailed study of the spatial-dependence of head-related acoustics. Efficient methods are desired when making large numbers of measurements, especially with human subjects. Often multiple loudspeakers are mounted on an arc with mechanical rotary control e.g. Masiero, Dietrich, et al. (2012).

Since the sweeps are generally much longer than the resulting impulse responses, measurement time can be reduced by overlapping sweep measurements without significantly reducing SNR (Majdak, Balazs, et al., 2007; Dietrich et al., 2013). Recently dynamic adaptive system identification methods have been investigated for more rapid HRTF measurement with continuous movement of loudspeakers or listeners (Rothbucher et al., 2013; Fallahi et al., 2015; He et al., 2018).

An example modern measurement system using the exponential swept sine method captures HRTFs for an individual on a grid of 1680 directions in about 90 min (Carpentier, Bahu, et al., 2014). Utilising recent techniques can reduce measurement time significantly, for example with multiple overlapping exponential sine sweeps 1550 directions were measured in 20 min (Majdak, Goupell, et al., 2010). With a normalised least-mean-squares adaptive system identification technique, continuous-azimuth measurement could be made with sufficient SNR at 20 s per elevation position. State-of-the-art measurement systems can obtain sufficiently accurate HRTFs at directional-resolution approximating JNDs in a matter of tens of minutes.

There exist a number of freely-available HRTF datasets measured both on human subjects and HATSs e.g. (Algazi, Duda, Thompson, and Avendano, 2001; Watanabe et al., 2014; Kearney and Doyle, 2015a; Bomhardt et al., 2016; Carpentier, Bahu, et al., 2014; Bernschütz, 2013; Arend et al., 2016; Brinkmann, Lindau, Weinzierl, Geissler, and Par, 2013; Yu et al., 2018; Andreopoulou, Begault, et al., 2015). The spatially-oriented format for acoustics (SOFA) was defined to provide a common format for exchanging HRTF data (Majdak, Iwaya, et al., 2013). It has been standardised by the Audio Engineering Society (AES) (AES69:2015, 2015).

2.4.3 HRTF Equalisation

A true HRTF includes the direction-dependent transfer function from the source to the ear canal entrance and the direction-independent transfer function from there, via the ear canal, to the ear drum. A measured HRTF will include the transfer function of the measurement loudspeaker and microphones, and will be as seen at the measurement point, which may not be at the ear drum. Møller (1992) discussed considerations for equalisation in binaural technology in detail. If only analysis of the HRTF measurements is required, then an adequate equalisation filter may be calculated by inversion of a transfer function measured with the loudspeaker and measurement microphones in the free-field. However, for binaural rendering with headphones, the headphone-to-ear transfer function (HpTF) must be corrected for. Measurements of the HpTF will also incorporate the microphone responses and may not be measured at the eardrums. Ideally binaural rendering will account only for the true HRTF with all other effects in both the measurement and reproduction equalised. Equalisation of the HpTF will be covered in more detail in section 2.7, but this section introduces the general ideas behind equalisation of HRTFs and binaural rendering systems.

2.4.3.1 In Situ Equalisation

Physically accurate equalisation of binaural rendering systems is only feasible when the measurement microphones in the ear canals are not moved throughout the measurement and rendering processes. The equalisation filter should correct for the loudspeaker and microphones used in HRTF measurement, as well as the HpTF used for reproduction of the rendering. To avoid equalising for the in-ear measurement microphones twice, the loudspeaker response should be measured in the free-field using a high quality reference microphone. The HpTF should be measured in-situ, incorporating the measurement microphone response, at the same reference point in the ear canals. An equalisation filter can then be derived by inverting the frequency-domain product of these two transfer functions.

When HRTF measurement and HpTF do not occur at the same time, the positioning of the microphones within the ear canals cannot be guaranteed to be the same, and so physically accurate equalisation is not possible. However, if this were necessary, it would severely limit the usability of binaural rendering.

2.4.3.2 Decoupled Equalisation

There are different equalisation approaches taken if the rendering does not occur at the same time as the HRTF measurements (Larcher, Jot, and Vandernoot, 1998). Equalisation is performed relative to a reference sound field measurement, which also contains the measurement system effects and is measured at the same reference point in the ear canals. If this reference sound field can be reproduced in the reproduction scenario, the appropriate equalisation can be applied to produce correct ear signals. The equalisation filter should also compensate for the reproduction system effect, i.e. the HpTF, and the transducers used in reproduction-end measurements. This allows decoupling the equalisation of the HRTF measurement system and the reproduction system. Such equalisation is likely to give less accurate results than the in-situ scenario, since replication of the reference sound field is challenging and the equalisation of two separate measurement systems is required.

Free-field equalisation involves dividing the measured HRTFs by the HRTF measured for a reference direction, typically the frontal direction. The reference sound field is then considered to be a plane wave coming from this reference direction, for which the free-field equalised HRTF would have a flat frequency response. This approach eliminates the measurement system effects and potentially some direction independent aspects found at the reference direction, such as ear canal resonances. In this case the rendering system must be equalised to have the transfer function of a frontal sound source i.e. the frontal HRTF. It should also be corrected for the loudspeaker response and the HpTF measurement.

Diffuse-field equalisation instead involves dividing the measured HRTFs by the transfer functions that would be measured in a diffuse field. This can be estimated by the power-average of the HRTFs over all directions. Diffuse-field equalisation would give a flat frequency response for a recording in a diffuse field, with sound waves impinging from all directions, as can be approximated in a reverberation chamber. The diffuse-field-equalised HRTFs are often called the directional transfer functions (DTFs), since the direction-independent component has been removed. The direction-independent component includes the measurement transducer effects but also direction-independent aspects specific to the individual. In this case, the rendering system should then be equalised with the diffuse field response, whilst also compensating for the HpTF. This would require calculation of the diffuse-field response again, independent of the original HRTF measurements. Besides remeasuring the HRTFs, accurate methods require dedicated facilities and equipment, so only approximations are practical (Larcher, Jot, and Vandernoot, 1998).

Despite this limitation, diffuse-field equalisation is popular in practical applications, where measurement and equalisation at the rendering stage is often not possible. It is also thought to be the best approach when rendering with HRTFs measured on an individual other that the listener, because it removes the direction-independent individual effects. It can be said to reduce variability of HRTFs between individuals below about 5 kHz (Larcher, Jot, and Vandernoot, 1998). Diffuse-field equalisation is the recommended target for headphones in several current standards (ITU-R, 2002; IEC 60268-7, 2010). If headphones are diffuse-field equalised then diffuse-field equalised HRTFs should yield the most accurate results when no equalisation is applied. Although it would be far from perfect, due to individual differences and manufacturing tolerances. In reality however, most headphones are not equalised to the diffuse field (Møller, Hammershøi, et al., 1995). It is nonetheless a common assumption for practical binaural rendering applications that diffuse-field equalisation leads to the best results for uncontrolled reproduction scenarios, as recommended by Larcher, Jot, and Vandernoot (1998).

2.4.4 Real-Time Signal Processing Considerations

The creation of interactive auditory virtual environments (IAVEs) with binaural technology requires time-variant real-time rendering. Rendering of free-field sources requires a dataset of HRTFs filters corresponding to a set of source directions. Dependent on the desired source direction, an appropriate filter should be determined, either by selection from the dataset or by interpolation between measurements. The source signal is then convolved with the HRTF to obtain a binaural signal. When the source or listener position changes, a new HRTF must be obtained and the filter updated without artefacts.

Jot, Larcher, et al. (1995) discussed the DSP techniques used in design of early real-time binaural rendering systems. Clearly the capabilities of DSP units have advanced significantly since this time, yet many aspects discussed in this study are still relevant today.

Finite impulse response (FIR) representations of HRTF filters are often used and convolution is commonly carried out with the fast frequency-domain method. However, a wide variety of approaches exist. For filter updates, time-domain FIR filter coefficients can be smoothly interpolated. In the frequency-domain a block-wise crossfade is needed, applying both the previous and the new filter and crossfading between the two outputs (Franck, 2014).

Equalisation of the measurement transducer responses will enable reduced filter lengths, particularly correcting for low-frequency group delay rises due to loudspeaker response (Bernschütz, 2013). Diffuse-field equalisation of the HRTFs also has benefits in reducing filter length, since it removes strong ear canal resonances and generally flattens spectra (Jot, Larcher, et al., 1995). Another common technique is decomposition of the HRTFs into separate minimum-phase and excess-phase components, which has multiple benefits for efficiency and quality of rendering (see section 2.6.1).

2.4.5 Dynamic Systems with Head-Tracking

Head movements have been shown to be an important part of auditory perception, as reviewed in section 2.2. Changes in head orientation strongly change the signals at the eardrums. Allowing natural head movement in binaural rendering will clearly improve spatial impression. Whilst not suitable for all applications, it is clearly essential in some, such as virtual and augmented reality systems. Binaural rendering systems can be made to track the position of the listener's head and at regular intervals update the filters used in rendering of a sound source. Given the head position and the sound source position, an appropriate binaural filter is selected from a database in order to synthesise the sound source at the correct position. The aim is to keep the virtual auditory space fixed while the head is moving, rather than it moving with the head in an unrealistic way. Blauert (1997, p.383) describes the use of head tracking in binaural rendering as providing “perceptual space-constancy” which gives the listener “an improved sense of involvement, as they now perceive themselves as moving in an otherwise fixed scenario”. Since such systems update in real-time according to head movement they will be termed *dynamic* binaural rendering systems.

An early manifestation of such techniques was implemented by Boerger et al. (1977), using analogue circuitry to implement dynamic filters to control headphone signals and stabilise source direction during head movement. The earliest digital binaural rendering systems incorporated use of head tracking (Wenzel, Wightman, et al., 1988). Section 2.8 gives a detailed review of the perceptual considerations of head tracking in binaural systems, as well as system designs and specifications.

2.4.6 Environment Simulation

Real acoustic environments are almost never anechoic, reflection and diffraction occurs, giving additional auditory cues to the nature of sound sources and the environment itself. Therefore to achieve convincing binaural rendering of virtual environments, rendering of the virtual environmental acoustics is required. Approaches can be broadly categorised as data-based and model-based. Data-based AVEs are derived from acoustic measurements made in a real acoustic environment, whilst model-based AVEs are derived from computational simulations of the acoustics of an environment.

A room impulse response (RIR) is often described by several component stages, as illustrated in figure 2.15. The sound travelling along the direct path from the source to the listener arrives first, after some propagation delay T_0 due to the source-to-listener distance. Following the *direct sound*, a set of *early reflections* arrives from nearby surfaces and objects, at different directions to the direct sound. Higher order reflections occur as the initial reflections themselves reflect off other surfaces before reaching the listener.

Beyond a certain point, the echo density becomes such that patterns of individual reflections become indistinguishable and the room response is perceptually diffuse. This *late reverberation* can be modelled by a decaying stochastic signal, the properties of which are dependent on the room volume and absorption of its surfaces. The decay of this reverberation is described by the *reverberation time*, generally measured in seconds per 60 dB decay and represented by T_{60}. The point of transition to late reverberation is often called the *mixing time* (Lindau, Kosanke, et al., 2012). Beyond this point in the RIR, perceptual effects of changing source or listener position and orientation are largely insignificant. ISO 3382-1:2009 (2009) describes objective metrics of room acoustics that can be used to indicate

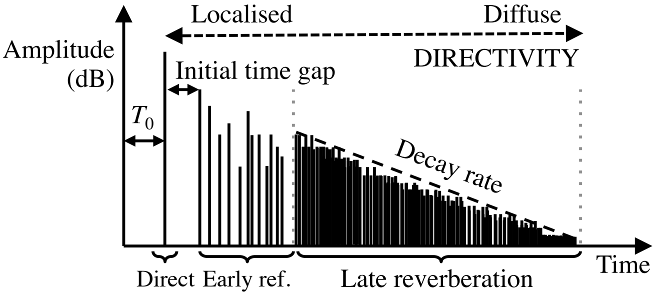


Figure 2.15: Simplified model of a generic room impulse response, showing the logarithm of the square amplitude. From Coleman, Franck, Jackson, et al. (2017) (used with permission).

perceptual character, such as reverberation time and clarity.

Binaural impulse responses can be measured in reverberant environments, capturing not only the head-related acoustic effects on the direct sound but also the environmental reflections and late reverberation. These are commonly called binaural room impulse responses (BRIRs), since they incorporate the HRIRs and the RIR at a given position. BRIRs are generally much longer than anechoic HRIRs, though the length is dependent on the reverberation time. Data-based AVEs typically use a set of measured BRIRs, for a range of source and listener positions, convolving the appropriate measurements with source signals.

By contrast, a model-based AVE typically makes use of anechoic HRIRs and simulates the environmental effects as a series of delayed and filtered versions of source sounds arriving from different directions, representing discrete reflections. In addition, the perceptually diffuse late reverberation that occurs in enclosed spaces is often modelled separately.

The techniques for creating AVEs are reviewed in more detail in section 2.9.

2.4.7 Perceptual Accuracy of Binaural Rendering

Since the early studies of Wightman and Kistler (1989a,b), the perceptual accuracy of binaural rendering technology has been evaluated many times. Subsequent sections will explore such evaluations in more detail, as they relate to specific aspects of binaural rendering system design. However, a few key studies that validate the techniques will be mentioned here.

Whilst the results of Wightman and Kistler were highly promising, slight degradations in localisation performance were observed, particularly increased front-back reversals. Other studies have found similar issues (Bronkhorst, 1995; Middlebrooks, 1999b). However, with careful calibration of individualised measurements, made in situ, Zahorik, Wightman, and Kistler (1995) and Langendijk and Bronkhorst (2000) both demonstrated that binaural rendering could be made indistinguishable from free field loudspeaker sources. Both experiments used probe microphones to measure pressure signals close to the eardrums and Langendijk and Bronkhorst used extra-aural mounting for the headphones (see section 2.7). In terms of localisation acuity, Martin, McAnally, and Senova (2001) demonstrated free-field equivalent localisation errors, including front-back confusions, though only with three listeners.

Oberem, Fels, et al. (2013) used less precise but more practical techniques, with open-backed circum-aural headphones and measurements at the entrance to the ear canal. A three-alternative forced choice method was used to identify differences between loudspeaker and headphone rendering. Results demonstrated that rendering was indistinguishable from a real-loudspeaker source with both speech and music signals. For noise signals, differences in high frequency content allowed listeners to differentiate the headphone rendered stimuli. Without comparing directly between the signals however, listeners were not able to correctly identify from which reproduction system (loudspeakers or headphones) the noise signals were rendered. No significant differences between blocked and open ear canal measurements were found.

Whilst previous studies maintained a static listener, Romigh, Brungart, and Simpson (2015) introduced head-tracking and demonstrated free-field equivalent localisation. The system used individual blocked-ear measurements and extra-aural headphones. In a first experiment full-phase HRTF measurements were used for rendering without head tracking, during short 250 ms noise bursts, and listeners reported that they thought these stimuli came from real loudspeakers for approximately 85% of these signals, which was not significantly different to the responses for real loudspeaker stimuli. When head tracking was applied, requiring dynamic switching between minimum-phase filter representations, localisation accuracy was equivalent to that for free-field loudspeakers, including polar angle errors and front-back reversals. This was both for long 10 s noise stimuli that allowed exploratory head movements, and short 250 ms noise bursts, where head movements would have little use.

Brinkmann, Lindau, and Weinzierl (2014) presented an evaluation of a data-based dynamic binaural rendering system that used in-situ measurement of BRIRs and correction filters for an extra-aural headphone system. The stated criterion for quality of the system was *authenticity* i.e. “perceptual identity with an explicitly presented real event”, after Pellegrini (2001). A two-alternative forced choice (2AFC) test paradigm was used to directly compare binaural rendering to real loudspeakers in a medium-sized room with $T_{60}^{1\,\text{kHz}} = 0.65\,\text{s}$. Dynamic changes due to head tracking only accounted for head yaw rotations (in 2° steps), and frontal and lateral loudspeakers were used. For pulsed pink noise signals, all subjects could reliably discriminate between the simulation and the real loudspeaker. Whereas for a speech signal, only about half of subjects showed detection rates above chance and on average performance was at the level of the detection threshold. Detection rates were higher than in previous studies, reflecting the more demanding simulation scenario of dynamic rendering. But the listeners were also permitted to freely switch between stimuli until making a decision, which is likely more sensitive than previous methods where typically only one presentation was made.

In summary, when binaural rendering technology is carefully calibrated for an individual listener in a controlled laboratory environment, a high degree of perceptual accuracy can be achieved. However, this is not practical for many applications. Binaural technology can still provide effective results in less controlled settings, though the strict criterion of perceptual authenticity might not be met. The following sections review specific aspects of binaural rendering and reproduction technology in more detail, with consideration given to applications where such precise control is not possible.

2.5 Individualisation of HRTFs

HRTFs differ significantly between individuals, as was shown in the early measurements made by Wightman and Kistler (1989a) and investigated in detail by Møller, Sorensen, et al. (1995). As introduced in section 2.2.4, this variation is due to the uniqueness of each individual's morphology, particularly the shape of the head and the pinnae, as well as the effects of the ear canals, where included. It seems intuitive that, given these acoustic differences, binaural rendering will be more effective with HRTFs measured on the individual listener than with those measured on another listener.

Individualisation[8] of binaural technology is conventionally achieved by making acoustic measurements of the HRTFs of the individual. This generally requires expensive specialist equipment and facilities, and can be both time consuming and invasive. For mass-market applications of binaural technology, such measurements are impractical. Hence there has been a lot of research into whether individual HRTFs are required and finding efficient and practical means of individualising binaural technology.

[8] A note on terminology: It is common in the literature to use the term *individualisation* to describe the use of measurements specific to an individual or the adaptation of systems to better approximate an individual's specifications. Whilst *personalisation* is synonymous and is more common in wider contexts, *individualisation* is used here for compatibility with other literature in this field. *Individual* HRTFs are those measured on the listener, and *non-individual* HRTFs are those measured on another person or a HATS.

2.5.1 Perception of Rendering with Non-Individual HRTFs

Whilst there are significant differences between in HRTFs between listeners, there is also commonality, in particular below 6 kHz (Møller, Sorensen, et al., 1995). Section 2.3 also introduced dummy head microphones and HATS, which are based on representative anthropometric data and have been used successfully for binaural recordings and communication systems testing for many years. This section reviews studies on the perceptual effects of binaural rendering without individual HRTFs.

2.5.1.1 Localisation Accuracy

It has been shown that localisation accuracy is reduced when using non-individual HRTFs in binaural rendering, particularly for elevated sources, however horizontal-plane localisation is fairly robust. Wenzel, Arruda, et al. (1993b) assessed localisation in binaural rendering systems with non-individual HRTFs, using noise bursts. The methods of Wightman and Kistler (1989b) were followed and the chosen HRTF set belonged to an individual with good localisation acuity in that study. A non-individualised HpCF was used, calculated from measurements made on the same individual as the HRTFs. Whilst most listeners localised sources on the correct cone-of-confusion, higher rates of front-back confusions and up-down confusions were observed compared with results for individualised rendering from Wightman and Kistler (1989b). Begault and Wenzel (1993) performed a similar study with speech stimuli, showing comparable directional localisation acuity results. Azimuth localisation was often biased towards medial or lateral directions, though this was subject-dependent. Front-back confusions were high, particularly 47 % of frontal sources were localised to the back, and although the target directions were in the horizontal plane, perceived elevation was dispersed and there was a tendency to perceive sources above the horizon. Whilst all HRTFs were measured with a source distance of 1.38 m and the sound pressure level was representative for speech at this distance, the perceived distance was much lower, generally less than 20 cm. Perceived distance was also lower in the median plane, and the proportion of stimuli heard inside-the-head by listeners was 15 %–46 %.

Middlebrooks (1999b) compared localisation accuracy in rendering with both individual and non-individual diffuse-field equalised HRTFs (DTFs). Localisation errors were lowest with individual HRTFs. The magnitude of errors with non-individual HRTFs was shown to vary in proportion to their degree of difference from the individual's own HRTFs. If spectral features were systematically at higher frequencies than in the listener's own HRTFs then elevation judgements showed an upward bias, whereas if they were at lower frequencies, then front-back reversals were common: low sources moved from front to back and high sources moved from back to front. The systematic scale difference also led to over-/underestimation of lateral angle, though the degradations were smaller in this respect. Adaptation of the non-individual HRTFs to best match the individual's by frequency scaling more than halved the degradation in localisation performance with the majority of listeners.

Hartmann and Wittenberg (1996) studied the signal characteristics that are responsible for causing externalisation i.e. localisation of auditory events outside of the listener's head. The experimental apparatus allowed direct comparison of loudspeaker-rendered (free-field) and headphone-rendered sources in an anechoic chamber. The binaural rendering used in-situ individual measurements, no head movement was permitted. A complex harmonic stimulus signal was used, imitating the vowel /a/. The authors demonstrated that, for a single source at 0° elevation, −37° azimuth, and 1.5 m radius, a binaurally-rendered source could be indistinguishable from the real free-field source. A 2AFC method was used, assessors reported whether the experienced stimulus was either real or virtual, as well as reporting externalisation using a four-point scale ranging from "in my head" to "externalised, compact and located in the right direction and at the right distance". A series of experiments was used to explore the cues that lead to authenticity and externalisation of auditory events. Externalisation could be controlled on a continuum from inside-the-head to the position of the loudspeaker, for example by removing inter-aural phase differences in a frequency-dependent manner. The majority of disruptions to the baseline rendering led to auditory events inside or very close to the head. However, by setting constant inter-aural phase differences, it was found that authenticity is not dependent on signal phases above 1.5 kHz. Additionally it was found that use of a frequency-independent inter-aural time difference could produce authentic auditory events, indistinguishable from the real loudspeaker. Correct inter-aural level differences across the full frequency range appear to be required for authentic rendering, however correct monaural cues at each ear are also required. Sources with correct inter-aural spectral level differences only were not perceived as authentic. These findings imply that individual HRTF measurements are likely important to achieve externalisation, at least when considering static free-field rendering.

Møller, Sørensen, et al. (1996) investigated localisation accuracy when using non-individual binaural recordings made in a "normal room", i.e. not an anechoic chamber but a room with natural reverberation. 14 loudspeakers were placed around the listener at a range of distances and directions. Participants also reported localisation for the real loudspeaker sources and individual binaural recordings. An individual HpCF was used in all headphone reproduction cases. Localisation with individual recordings was equivalent to that for real loudspeakers. With non-individual recordings, frontal sounds were often perceived behind and there were increased median-plane localisation errors. There were also increased distance errors, but no systematic reduction in perceived distance and no observed in-head localisation. Localisation errors in lateral θ_{cc} and polar ϕ_{cc} angle were rare outside of the median-plane. However, the experiment task was to indicate from which loudspeaker the source came, and the speaker separation angles were larger than 45°. Therefore small localisation errors could have occurred, but would not have been captured by the reporting method.

The previous study used a randomly selected non-individual HRTF, Møller, Jensen, et al. (1996) compared this approach to use of the HRTFs that gave best localisation performance across a panel of listeners, using a similar methodology. Median-plane localisation errors and front-back confusions were significantly reduced with this selected non-individual HRTF set, though still higher than that for real life localisation. It appears that some individuals more than others have HRTFs that exhibit reliable cues to a range of listeners.

Localisation acuity was also evaluated when using dummy head microphones (Møller and Hammershøi, 1999). When compared with errors for a randomly selected human head from (Møller, Jensen, et al., 1996), localisation acuity for these dummy heads was often worse. Minnaar, Olesen, et al. (2001) then compared real loudspeaker localisation to dummy head recordings and those for two non-individual human subjects that gave low errors in (Møller, Jensen, et al., 1996). There were significant differences in localisation acuity between the different dummy heads, whilst the selected non-individual human recordings gave significantly better localisation than all dummy heads and real loudspeaker localisation was substantially more accurate. The Neumann KU100, which is without shoulders, gave distinctly higher localisation errors than HATSs, particularly within the cones-of-confusion outside of the median plane.

Begault (1992) investigated the effects of synthetic environmental reverberation with non-individualised rendering. A ray-tracing technique was used for spatialised early reflections and a statistical model was used for the diffuse reverb tail. A significant reduction in in-head localisation was found when the room effect was applied (from 25 % to 3 % of stimuli). Zahorik (2000) showed that when natural reverberation is available to the listener, in BRIR measurements, the sense of distance and degree of externalisation do not depend upon individualisation of binaural filters.

Begault, Wenzel, and Anderson (2001) compared rendering with individual HRTFs to use of dummy head HRTFs, whilst also varying head tracking and room response system factors, in an experiment using speech stimuli at various target directions in the horizontal plane. Results generally showed little benefit of individualisation on localisation, front-back reversals or externalisation. When non-individual HRTFs were used, introduction of head tracking gave lower azimuth errors, since listeners could utilise dynamic localisation cues.

Romigh and Simpson (2014) decomposed the HRTFs into mean, lateral (θ_{cc}) and intra-conic (ϕ_{cc}) magnitude response components with minimum-phase, plus an ITD. These cues were then recombined for binaural rendering in a localisation experiment, using a mix of individual and non-individual components to investigate which components play an important role in individualisation. The KEMAR HATS was used for the non-individual HRTFs. The localisation errors were viewed separately in terms of θ_{cc} and ϕ_{cc}, as well as total angular error. The results suggest that individual localisation cues are primarily "intra-conic", meaning that they allow distinction of the polar angle ϕ_{cc} within a cone-of-confusion. When individual intra-conic components were used, with all other aspects not individualised, localisation errors were found equivalent to fully individualised rendering for long sounds (10 s), and for short sounds (250 ms) only lateral angle estimation was significantly worse. Conversely, de-individualising only the intra-conic component of the HRTFs caused significant degradation in ϕ_{cc} localisation accuracy, whilst de-individualising other components did not cause significant degradations, at least for long sound sources. The rendering system updated according to head tracking, so for long sounds listeners could utilise dynamic changes in interaural cues, to hone in on the sound source with head movements. For short sounds, ITD individualisation showed a significant improvement in lateral localisation acuity, though only by 1°–2°.

The ITD has been found to be the dominant cue in lateral localisation of broadband sources, at least for static scenarios (Wightman and Kistler, 1992), and it varies with head size. If the head used in non-individual binaural rendering is too small for the listener, the range of perceived azimuths will be reduced, and if it is too large, giving ITDs significantly greater than experienced naturally, a listener will be unable to localise some sources (Shinn-Cunningham, Durlach, et al., 1998a). With head-tracked rendering, if the ITD is incorrect, then the source position will be unstable with head movement (Lindau, Estrella, et al., 2010). This may not prevent listeners from ultimately localising a sound correctly but it will make the task more challenging. Head motion is also thought to aid vertical localisation, potentially allowing the listener to improve the pinna cues (Perrett and Noble, 1997). Mehra, Nicholls, et al. (2016) show benefits to using individual HRTFs over non-individual HRTFs even in head-tracked binaural rendering, in terms of localisation accuracy.

2.5.1.2 Alternative Measures of Performance

Whilst binaural technology is often evaluated in terms of localisation acuity. This is not the only aspect important to the quality of binaural systems, as will be discussed further in chapter 3. But additionally, even if accuracy of localisation is good once a judgement is reached, the localisation task may be more difficult for the listener. More recently the perceptual effects of individualisation in binaural rendering have been explored with other task performance aspects.

Wisniewski et al. (2016) studied sound localisation using performance in a detection task. The event concerned was a change in source elevation. Runs of noise bursts were played to listeners with random changes in elevation at random intervals, with 5–10 bursts at each elevation. Listeners were tasked with pressing a button each time elevation changed. The detection of elevation changes was significantly improved with individual HRTFs. Alongside the task performance, the event-related potential (ERP) was studied, which is the electrophysiological brain response to a stimulus measured by electroencephalography (EEG). The study looked for correlates between the ERP and the use of individual or non-individual HRTFs. The ERP responses during successful detection were analysed. ERP responses in the time window 300 ms–500 ms after the stimulus onset (elevation change) showed most significant differences, with enhanced amplitude when individual HRTFs were used. These amplitude differences were correlated to the individual performance difference between the two HRTF conditions. Responses in this time region are associated with domain-general cognitive processing, at the point where memory may be involved in stimulus detection (Polich, 2007). The authors stated that “part of the individualization benefit reflects post-sensory processes”, indicating that localisation processes in non-auditory brain regions respond differently to individualised binaural rendering.

Poirier-Quinot and Katz (2018b) evaluated task performance in a virtual reality (VR) shooter game, particularly inspecting game progress, reaction time and movement efficiency. Twenty participants used the best and worst performing non-individual HRTFs from a database, selected using a subjective rating procedure (Andreopoulou and Katz, 2016). Significant learning effects were observed in the task across six sessions. The two HRTF sets were alternated between sessions and no significant improvement was observed in the performance metrics overall due to the HRTF condition. For targets that approached from directly above the listener, however, the best HRTF set gave a significant reduction in angular distance travelled to locate the sources. A potential effect of presentation order of the two HRTF sets was observed, with those using the best HRTF set second performing better at the task after the first session. However, since the HRTF presentation order was varied between subjects and four of the strongest performers overall were in one group, this could have been due to uneven distribution of participant ability amongst the groups.

Lindau and Weinzierl (2012) evaluated a non-individual dynamic binaural rendering system that used BRIRs measurements from a HATS and demonstrated that it was capable of “a simulation in agreement with the listener’s expectation towards a corresponding real event". The experimental method is reviewed in chapter 5. It was shown that auditory simulations of loudspeakers in the test environment were highly plausible. The binaural signals with supporting dynamic, reverberation and visual cues were adequate and any deviations from reality were not sufficient to affect this plausibility.

Väljamäe et al. (2004) studied the role of HRTF individualisation on the sense of presence in virtual environments, which was defined as "a sensation of being actually present in the virtual world". A small but significant effect of individualisation increasing presence was found, though the stimuli were anechoic, only up to three sound sources were used, and no head tracking was used, so the overall level of presence given was likely low.

2.5.1.3 Learning Effects

There is also evidence that learning effects (plasticity) can occur, allowing improved effects with non-individual binaural rendering over time. Experiments often involve localisation tests before and after exposure to an adaptation phase that gives some form of feedback to localisation performance. Longitudinal study of localisation performance is sometimes also carried out.

Shinn-Cunningham, Durlach, et al. (1998b,a) found that, given trial-by-trial feedback on correct localisation responses, listeners can adapt to some extent to incorrect non-individual horizontal localisation cues. Hofman et al. (1998) found that listeners can adapt over time to modifications to their own pinnae, improving elevation localisation after an initial degradation. Zahorik, Bangayan, et al. (2006) found that training with feedback reduced front-back reversals in non-individual binaural rendering, even after relatively short training periods (two 30 min sessions), and the benefits persisted for several months after training.

Parseihian and Katz (2012) observed reduction also in polar angle localisation errors of 10°, after only three 12 min sessions. Rendering used approximately individual ITDs using an anthropometric model based on measured head size, and then using either best or worst selected non-individual minimum-phase HRTFs from a presented set of options. These improvements are relative to a control group that used individual HRTFs, so accounting for task learning processes. Adaptation/training sessions were within an auditory virtual environment, listeners located the target sound by pointing, with an audio indicator when correctly identified, thus giving auditory-kinaesthetic feedback on localisation. Listeners adapted better when the non-individual HRTF set was pre-selected based on good subjective localisation.

Berger et al. (2018) showed that visual feedback also improves learning processes with a non-individual HRTF set. A moving sound source was presented in VR environment with head-tracking. When this was paired with a spatiotemporally aligned visual source during an adaptation phase, localisation with non-individual HRTFs was significantly improved. No improvement was found either with the auditory stimulus alone or an asynchronous visual stimulus.

There appears to be a high-degree of cross-modal learning and adaptation in sensory processing. It seems that we can adapt to incorrect cues over a relatively short space of time when we have adequate feedback. This adaptation appears more effective when the cues are less distinct from our own. What is not clear is whether the adaptation occurs during the perception of auditory events or during reporting of their location.

2.5.1.4 Summary

In summary, localisation performance is worse when using non-individual HRTFs. Azimuth localisation is broadly correct, but elevation accuracy is significantly reduced and increases in front-back reversals occur. In addition to reduced accuracy, the localisation process seems to become more difficult. When head tracking is used, front-back reversals are avoided, but localisation is still better with individual HRTFs. In-head localisation is common for anechoic signals, but with a room response, distance and externalisation are not affected. Lack of individualisation also negatively affects other quality aspects such as timbral fidelity and naturalness, as reviewed in section 3.2.5.4. Despite some drawbacks, non-individualised rendering still provides a reasonably good spatial impression, especially with dynamic motion and environmental reverberation cues. Degradations are reduced when the HRTFs of a particular individual are chosen based on good general localisation performance across a population of listeners. There is also evidence that localisation processes are plastic and with adequate feedback listeners' localisation acuity will improve with non-individual HRTFs over time. However, it is clear that individualised binaural rendering will give better perceptual results.

2.5.2 Methods for HRTF Individualisation

There has been significant research into methods for individualisation of binaural systems, particularly considering the practical needs of applications outside of the laboratory. Besides the short review here, a more detailed review is given by Xu and Li (2007), and more recently by Guezenoc and Séguier (2018).

As introduced in section 2.4.2, the conventional method of obtaining HRTFs is by acoustic measurement. Whilst state-of-the-art techniques have reduced required measurement times significantly, specialised facilities and equipment are still required. Errors can also occur in measurements, due to listener movements (Hirahara et al., 2010) or measurement environment, process and equipment (Andreopoulou, Begault, et al., 2015).

2.5.2.1 Numerical Simulation

Numerical simulation of head-related acoustics is an alternative approach to generating an individual HRTF set. This is derived from detailed measurement of listener morphology i.e. a 3D scan. It was first investigated using the boundary element method (BEM) by Katz (2001a,b) and has since been advanced by the fast multipole accelerated boundary element method (FM-BEM) (Kreuzer et al., 2009)[9], allowing simulation in the full audible frequency range within a few hours. Alternative techniques such as a finite difference time domain (FDTD) approach have been used (Xiao and Huo Liu, 2003). The challenge for individualisation is obtaining accurate head and pinna scans in a practical manner. Accurate scanning of head and pinna morphology is feasible with various specialised techniques such as magnetic resonance imaging (MRI) (Jin, Tew, et al., 2013) or computed tomography (CT) (Ziegelwanger, Majdak, and Kreuzer, 2015), methods such as simultaneous photography (photogrammetry) and laser or structured light scanning have also been used (Huttunen, Vanne, et al., 2014). Significant mesh pre-processing is often required. As yet there has been little perceptual validation of these simulation techniques. In a study with high-resolution scans of three human subjects, Ziegelwanger, Majdak, and Kreuzer (2015) found that objective results were comparable to acoustic measurements and localisation accuracy was equivalent when using a 1 mm–2 mm mesh resolution. Polar angle localisation error increased with coarser mesh resolution. This suggests that less accurate scanning techniques such as photogrammetry with consumer cameras are unlikely to give suitable resolution to be equivalent to individual acoustic measurements. Yet professional 3D scanning services are now quite common, and Huttunen and Vanne (2017) describe an end-to-end process for HRTF personalisation by simulation, which is being developed commercially. Numerical simulation has also been used to study the relationship between listener morphology and acoustic features of the HRTFs, by systematically exploring the effects of surface perturbations (Tew et al., 2012; Mokhtari et al., 2010).

2.5.2.2 Anthropometric Input

There are a wide variety of methods for individualisation that involve selecting or adapting non-individual HRTFs from an available set of measurements. Selection can be based on

[9] An open-source implementation of this method is available (Ziegelwanger, Kreuzer, et al., 2015).

limited anthropometric information about the listener, or guided by perceptual feedback. These methods are typically designed considering low-cost consumer applications.

Regarding anthropometric techniques, rather than use detailed 3D morphology scans, simple measurements of head and ear shape are often used to perform indirect individualisation, starting from HRTFs measured on other individuals.

As previously introduced, Middlebrooks (1999b) took the approach that the primary inter-individual difference is overall size and so applied a frequency-scaling to a set of non-individual HRTFs to minimise spectral differences. The scaling reduced vertical localisation errors compared to the unscaled non-individual filters. A subsequent study then showed that the scaling factor could also be set either by perceptual adjustment (in a 20 min procedure) or based on measurements of head width and pinna height, achieving equivalent localisation accuracy in listening tests to the spectral error minimisation (Middlebrooks et al., 2000).

Zotkin, Hwang, et al. (2003) presented a method to find a nearest neighbour within a database of HRTF sets, based on similarity of a limited set of morphological measurements. The individualisation technique was based on a simplistic model of head and torso acoustics for which an analytical model was obtained (Duda, Algazi, et al., 2002). The high-frequency spectral magnitude cues were then introduced from a selected human HRTF from the CIPIC database (Algazi, Duda, Thompson, and Avendano, 2001), with a transition region of 500 Hz–3000 Hz. Head and torso model was based on "torso radius" (half shoulder width), "head radius" (half interaural distance) and neck length, whilst pinna cue personalisation was based on seven pinna feature measurements. Results suggested that head and torso adaptation improved azimuth localisation, whereas the anthropometrically-selected pinna cues gave mixed results across listeners, with no clear improvement in elevation localisation.

A number of further studies have used the CIPIC database, which provides anthropometric data alongside the HRTFs. Mohan et al. (2003) investigated using computer vision techniques to derive the anthropometric data directly from photographs. This was further developed in a more recent study (Torres-Gallegos et al., 2015). Often multiple linear regression techniques are used combined with statistical dimensionality reduction of the HRTF data e.g. using principal component analysis (PCA) (Xu, Li, and Salvendy, 2008). Other machine learning techniques such as artificial neural networks, which have grown in popularity in recent years, have been applied in similar ways (Hu et al., 2008; Li and Huang, 2013; Chun et al., 2017; Fayek et al., 2017). The studies rarely provide evidence of perceptual validation. Most often measures of average spectral magnitude deviation between estimated and measured HRTFs are used, which gives little indication of perceptual effects. Whilst the state-of-the-art machine learning techniques yield mean spectral distortion of less than

3 dB across the frequency range, Fayek et al. (2017) found that equivalent distortion was given by a model with no anthropometric input, which could only generate a generic HRTF set. Hu et al. (2008) performed a localisation experiment but only in the horizontal plane, neglecting more challenging elevation localisation.

Some way between anthropometric selection and simulation are structural modelling techniques where the HRTF is decomposed into a model of contributing features, often based on simplified approximations of head and pinna morphology which can be well modelled acoustically. Brown and Duda (1998) and Algazi, Avendano, et al. (2001b) introduced such models for the head and torso, whilst Spagnol, Rocchesso, et al. (2013) and Spagnol, Geronazzo, et al. (2013) extend the approach to the pinna. Only very limited perceptual evaluation has been performed with such models. Brown and Duda (1998) evaluated localisation with three subjects, limited to the frontal hemisphere and showed only a comparison to individual HRTFs, which gave much more accurate localisation. However, Geronazzo et al. (2018) utilised an auditory model for vertical localisation (Baumgartner, Majdak, and Laback, 2014) to provide estimates of perceptual quality of the non-individual HRTF match, this suggests that a non-individual HRTF set can be found that gives equivalent vertical localisation accuracy to the individual's own, at least in large HRTF datasets of more than 40 individuals. It was then found that pinna anthropometric data could be used effectively to define a subset of candidate HRTF sets that are likely to be suitable for the individual, via the structural pinna model.

2.5.2.3 Perceptual Feedback

By contrast, there are also methods for selection of non-individual HRTFs based directly on perceptual feedback. Seeber and Fastl (2003) used this approach with a two-step selection method, listeners considered the HRTFs in terms of a number of localisation attributes (horizontal movement, elevation, front-back confusions, externalisation, distance) with a virtual sound source which moved continuously between ±40° azimuth. The selection process, which took approximately 10 minutes, primarily minimised variance in localisation response and reduced in-head localisation, but also showed low localisation errors. However, only frontal sources in the horizontal plane were used in the listening test for validation and only twelve sets of HRTF were used in the study. Selection from larger scale datasets using the same procedure would take significantly longer, but it is unlikely that selection from such a small dataset could provide suitable spectral cues for most individuals.

Katz and Parseihian (2012) extended the perceptual selection approach by identifying a subset of perceptually-distinct HRTF sets. 45 participants each rated a set of 46 different

HRTFs, including their own individual HRTFs. Two moving sound source trajectories were used for evaluation, one circling the listener in the horizontal plane and one in the median plane, and a three-point quality scale was used to rate each HRTF set. Most individuals rated their own HRTF as excellent but there was no clear reciprocity between pairs of individuals i.e. they rated each others HRTFs quite differently. A subset of seven HRTFs was obtained which ensured that at least one was rated excellent by each individual. This subset was then used to perform a selection of best and worst HRTFs with 20 new individuals. Localisation experiments with a between-subjects HRTF condition demonstrated that the selected best HRTFs set gave significantly lower front-back and up-down reversals than the worst HRTFs. However, a control group with individual HRTFs gave clearly better results.

Xie et al. (2015) also reduced a large set of HRTFs to a small subset to enable easier selection, performing clustering based on magnitude response correlation instead of using subjective rating. HRTF selection based on perceptual feedback must be designed considering variability in responses, which has been shown to be significant across replicates (Schönstein and Katz, 2012) and due to source stimulus (Roginska et al., 2010). The process of subjective selection has been investigated, considering consumer applications, for example in a game (Härmä, van Dinther, et al., 2012), though with limited effectiveness.

PCA is a popular tool for dimensionality reduction of HRTF databases, with the aim of making individualisation processes more efficient e.g. (Jin, Leong, et al., 2003). This has been used directly in individualisation with perceptual feedback. Fink and Ray (2012) evaluated a method where the individual adjusts component weights from PCA of an HRTF database in order to generate personalised filters that improved perceived quality of the rendering. This process required tuning for each direction independently, six weights were tuned to control spectral features at each position, which seems impractical for consumer applications. Improvements were found in localisation in the horizontal plane, but polar angle localisation was not assessed.

Another approach to selecting a non-individual HRTF set is to use a sparse set of individual HRTF measurements to perform matching (Guillon et al., 2008). Andreopoulou, Roginska, et al. (2013) used linear discriminant analysis to identify the limited subset of source directions which best discriminate HRTFs between individuals. This then enables better selection of best-fitting HRTF sets from a database using a small set of measurements.

2.5.2.4 ITD Personalisation

Often, the ITD cues are modelled separately to spectral cues (see section 2.6.1), and separate individualisation techniques are also often applied. Whilst errors with non-individual HRTFs are larger in the vertical polar angle, Algazi, Avendano, et al. (2001b) describe how incorrect ITD can be particularly problematic in head-tracked dynamic rendering, where instability can occur with head movements. Lindau, Estrella, et al. (2010) describes this effect as “very annoying” and reports that learning/adaptation effects do not seem to occur. This suggests that correct individualisation of ITD is important. It appears also easier to achieve than full individualisation, including accurate pinna cues.

It is well established that ITD cues can be reasonably well modelled using simplified geometry e.g. a sphere (Kuhn, 1977) or an ellipsoid (Duda, Avendano, et al., 1999). Jot (1998) described individualisation of ITD cues based on a spherical head model, using the interaural distance as the sphere radius, the estimation was refined by Algazi, Avendano, et al. (2001b), using head width, height and length measurements. Individual asymmetries occur in ITDs due to asymmetric head shapes and ear positions, which are not accounted for in a simple spherical head model. Ziegelwanger and Majdak (2014) extend the spherical model to also incorporate ear position, though modelling TOA for each ear rather than ITD directly.

Another approach is to adapt ITDs from measured HRTF data. Katz and Parseihian (2012) utilised a method for modelling ITDs based on measured HRTF data rather than a geometric model, PCA weights derived from estimated ITD in the IRCAM LISTEN database were linked by linear regression to measured head circumference. Lindau, Estrella, et al. (2010) used a simpler approach to adapt the estimated ITDs from a specific non-individual BRIR set (from a HATS) to better match those of the individual listener, applying a single scaling factor. Individuals adjusted this single scaling factor in a psychophysical adjustment test. Linear regression was again applied to relate the perceptually identified scaling factor to intertragal distance[10].

2.5.2.5 Summary

Many methods for HRTF individualisation without acoustic measurement have been explored, particularly considering consumer applications. Numerical simulation from detailed head scans is a promising approach, although the acquisition of such scans in a consumer environment may be challenging for some time yet. Very often studies that use anthropometric data input are carried out without adequate perceptual validation, though it seems likely that at least some methods do give benefit. Methods based on perceptual feedback show promise in improving results over random or generic HRTFs, but challenges remain in

[10] The intertragal distance is a specific measure of the interaural distance measured between the left and right tragii.

how to perform such tasks with consumers, who may not have the inclination or the skill required. ITD adaptation seems an important first step towards individualisation, especially for dynamic head-tracked applications, to stabilise sources during head movement.

2.5.3 Do we need Individualisation?

From the literature it seems clear that HRTF individualisation can give improved binaural rendering. There are also many techniques proposed to achieve individualisation, with varying degrees of success. Yet despite this, current applications outside of the laboratory do not perform individualisation. This is partly because methods are either not yet practical for mainstream consumer applications or they are not yet good enough. Binaural technology can also be quite effective with generic binaural filters, particularly when other factors such as environmental reverberation and dynamic cues are presented adequately. It appears also that plasticity in our auditory system allows us to adapt to some extent to imperfect spatial cues reasonably quickly. For some applications generic HRTFs might already be good enough, but careful selection and processing of the HRTF set is still clearly required, and this must be considered alongside other aspects of system design. Significant research effort is being focussed on this topic and it seems likely that effective individualisation will become more widely available in future.

2.6 Interpolation of HRTFs

Jot, Larcher, et al. (1995) discuss considerations for time-variant real-time binaural rendering with moving sources and head tracking, which would enable IAVEs. The problem is separated into *interpolation* and *update* of binaural filters. Filter updates must be made whilst the rendering is running, in a manner which avoids signal discontinuities; Jot, Larcher, et al. call this “commutation”. Interpolation involves generating a filter at a target position that is between measured positions.

The required resolution for HRTF measurements should be determined considering localisation blur and MAAs, as introduced in section 2.2.6. MAAs range between 1°–10° dependent on the source signal and direction (Mills, 1958), so a high density of measurements may be required to adequately provide lifelike directional resolution. Thus researchers have investigated to what extent interpolation of measured HRTFs can reduce the required resolution of measurements.

HRTF interpolation strategies can be broadly divided into local and global approaches, both are discussed briefly by Jot, Larcher, et al. (1995). Local interpolation strategies typically involve weighted linear combination of the nearest measurements to the target position. For interpolation in both azimuth and elevation (or polar and lateral angle) the nearest three or four measurements are used. Global interpolation methods utilise the HRTF measurements at all available positions, by projection onto a set of basis functions, the coefficients of which are then used to derive new filters for the desired target position. Jot (1998) discusses global interpolation methods that utilise spatial basis functions, projecting HRTF coefficients onto the spherical harmonics, or spectral basis functions, using PCA of the HRTFs. Jot, Larcher, et al. (1995) state that interpolation is not required if the resolution of the measurements is less than the minimum-audible angle. In that situation, a nearest-neighbour (1-NN) search of the measurement positions will yield a suitable filter.

This section reviews both local and global interpolation methods, but first a common technique for decomposing HRTFs is introduced.

2.6.1 Minimum-phase plus Delay HRIR Model

The decomposition of the HRTF into separate minimum-phase and excess-phase components is often performed using the Hilbert transform. The excess-phase can be further sub-divided into linear phase (pure delay) and all-pass sections. Jot, Larcher, et al. (1995) approximated the excess-phase component by a pure delay, simply discarding the all-pass section.

The differences between this approach and the use of full-phase HRTFs are often inaudible, although with some exceptions, particularly where the source is in a lateral position and is arriving at the contralateral side from multiple diffraction paths around the head. In these cases, Plogsties et al. (2000) suggest that the all-pass component can be replaced by an additional broadband delay corresponding to the group delay of the all-pass component at 0 Hz. The underlying assumption is that if the correct broadband ITD is applied, minimum-phase HRTFs can be used without audible artefacts. Kulkarni, Isabelle, et al. (1999) further investigated this and found that it is the low-frequency ITD cue that must be preserved.

There are many different methods for the estimation of ITD, or TOA at each ear, and they show greatest variation in the region around the interaural axis (Katz and Noisternig, 2014). Andreopoulou and Katz (2017) performed an extensive study to evaluate the perceptual relevance of a large range of ITD estimation methods when combined with the minimum-phase transfer functions. Perceptual experiments were performed where listeners manually adjusted broadband interaural delays so that rendering through minimum-phase individual HRTFs achieved co-location with the full-phase HRTFs. The performance of different methods varied considerably, the best performing method utilised an onset threshold (−30 dB) for a low-pass filtered (3 kHz cutoff) version of the HRIR to estimate monaural TOAs.

There are practical advantages to the minimum-phase plus delay decomposition of HRTFs. The energy of the minimum-phase representation in the time domain is concentrated towards time $t = 0$ (Oppenheim and Schafer, 1998) and so shorter filters may be used, increasing rendering efficiency. Meanwhile a delay-line can be used for the linear-phase component. This also enables straightforward personalisation of ITD cues, scaling these onset delays appropriately to adapt for individual differences. Interpolation of minimum-phase HRTF representations avoids the comb-filtering artefacts seen with interpolation of the original mixed-phase representations due to linearly combining functions with different onset times. Separate delay interpolation is performed. Variable fractional delay lines are used to update delays over time to give accurate and artefact-free rendering with dynamic source or listener movements (Jot, 1998).

2.6.2 Local Interpolation Methods

Local interpolation methods involve the weighted combination of a small number of measured HRTFs around the target direction. Wenzel and Foster (1993) and Jot, Larcher, et al. (1995) describe piecewise linear interpolation of minimum-phase transfer functions. The weights for each of the nearest 3–4 filters are determined based on the azimuth and elevation angles or an inverse distance measure. Separate delay interpolation using variable delay lines avoids comb-filter artefacts due to misaligned onsets between the transfer functions to be combined. To ensure a minimum-phase result, the interpolation is performed on the log-magnitude spectra and the phase separately, or otherwise time-domain interpolation of the minimum-phase HRIRs can be performed (Jot, Larcher, et al., 1995).

Minnaar, Plogsties, et al. (2005) evaluated time-domain linear interpolation of minimum-phase HRIRs by comparing measurements at 2° resolution on a HATS to coarser measurement resolutions that were then interpolated to the same resolution. The ITD estimated from the original fine-resolution measurements was re-inserted after interpolation. A listening experiment was conducted using a 2AFC detection paradigm with a pink noise source signal to investigate the required directional-resolution of HRTF measurements. Comparisons were made for stationary and moving sound sources in the horizontal, median and frontal planes. For stationary sound sources in the horizontal plane, a resolution of 8° or less was sufficient for indistinguishable results and it was shown that greatest sensitivity was in lateral directions. Considering elevation in the median and frontal planes, sensitivity is lowest above the head (indistinguishable at 24° measurement resolution) but higher at horizontal elevations and below, where a resolution of 4° or less is required for indistinguishable results. When considering the most sensitive regions, it appears that measurement resolution needs to be higher in elevation than in azimuth.

For moving sources, with an angular velocity of 60 °/s, the resolution was found less critical, though similar patterns of sensitivity were observed. Sources were moved in 90° arcs in the horizontal, frontal, and median planes. A measurement resolution of 24° was undetectable in all but the movement below the listener in the median plane, where 8° resolution was required for undetectable results. The mean spectral magnitude error across frequency and both ears was found to predict the threshold of audibility when above 1 dB. The audibility of interpolation errors could then be estimated across the full sphere. It appears that a measurement resolution of 4° in azimuth and elevation would yield inaudible differences to measured filters for most directions, except below the listener. Concluding remarks suggest that 1130 HRTF measurements are required for rendering without audible errors when using local interpolation of minimum-phase filters in the time domain, compared to 11 975 original measurements.

Langendijk and Bronkhorst (2000) performed a similar evaluation though with individual HRTFs. Interpolation was performed in the frequency domain, separately operating on linear magnitude and phase. Target positions were always half-way between a pair of interpolated points and interpolation either occurred in horizontal, vertical or diagonal direction. A detection task was used, the four interval oddball paradigm (see section 3.3.4), with noise stimuli, to identify differences due to interpolation. At the highest resolution evaluated, 5.6° angular separation, results were indistinguishable from measurements for all interpolation directions. When the magnitude spectrum of the source signal was scrambled between presentations rather than flat, sensitivity was lower and 11.6° resolution also appeared inaudible. This suggests that sensitivity to interpolation is influenced by the stimulus, with natural time-varying sound sources likely being less sensitive than stationary noise stimuli.

By contrast, results of Wenzel and Foster (1993) suggested that measurement resolutions as low as 60° in azimuth and 36° in elevation could be used. However, this is likely because experiments only assessed directional localisation without assessing other differences, and used non-individual HRTFs. Martin and McAnally (2007) also analysed effects in terms of localisation acuity but with individual HRTF measurements, using an interpolation technique based in inter-aural polar coordinates. The four nearest measurements to the target position were interpolated in the frequency domain, with log-magnitude and unwrapped phase components interpolated separately. Localisation performance was equivalent to measured data when interpolated from resolutions of 20° in θ_{cc} and ϕ_{cc}. If interpolation only operated in one dimension then a resolution of 30° was equivalent.

Gamper (2013) proposed an extension of local interpolation to incorporate also distance.

3D Delaunay triangulation yields a tetrahedral mesh, grouping measurement points into four. When the target position falls within a certain tetrahedron, barycentric weights are derived to perform linear interpolation. Efficient identification of the appropriate tetrahedron is enabled through use of an octree structure (Samet, 1989) to identify the tetrahedra lying closest to the target and an adjacency walk search (Sundareswara and Schrater, 2003) is then used to find the appropriate one.

Brinkmann, Roden, et al. (2015) demonstrated that HRTFs with head-above-torso rotation are audibly different from those with head-and-torso rotation i.e. fixed orientation of the head relative to the torso, as is typically used in binaural systems. This suggests that HRTFs with variable head-above-torso rotation are required for perceptually transparent rendering with head tracking. In this case source rotation and head rotation are no longer reciprocal, the number of required HRTF measurements is greatly increased. The study investigated methods for interpolation when using variable head-above-torso rotations. It was found that interpolating between measurements with different torso-to-source orientations but identical head-to-source orientation was more robust against lower measurement resolution than using measurements with identical torso-to-source orientation and varying head-to-source orientation. This appears logical since changes in head orientation were shown to have greater physical effects than changes in shoulder orientation. The median threshold of audibility for the head-above-torso orientation resolution was about 25° with the torso-varying interpolation technique, when using noise signals. With this resolution, covering only head yaw rotation in the range ±75°, the authors estimate that between 8,000 and 16,000 HRTF measurements would be required to allow perceptually transparent representation of a point source position and head-above-torso orientation.

Lindau and Weinzierl (2009) investigated the resolution needed for measurements of BRIRs in dynamic rendering with head tracking. It was considered that this scenario may differ from the required resolution of anechoic HRTF measurements with static rendering. Using the FABIAN HATS with head-above-torso rotation (Lindau and Weinzierl, 2006), BRIRs were measured with 1° resolution for horizontal (yaw), vertical (pitch) and lateral (roll) head rotations, in anechoic conditions and in three reverberant environments with different reverberation times and source distances. An adaptive 2AFC method was used to find just noticeable discretisation thresholds, in comparison to the 1° resolution, for a frontal sound source in the horizontal plane with each kind of head movement. Sensitivity was greater with a pink noise stimulus than an acoustic guitar and was quite consistent across the acoustic environment conditions (including anechoic). Listeners were slightly more sensitive to head pitch rotations with the noise stimuli and head yaw rotations with the guitar stimuli. Head roll rotations showed much lower sensitivity for frontal sources but with vertical sound incidence this sensitivity was greatly increased since head roll rotations then introduced interaural cues. The thresholds for just noticeable discretisation for at least 95 % of listeners were 2° or above across conditions.

It should be noted that Lindau and Weinzierl (2009) did not perform interpolation between BRIRs, the measurement grid was only discretised to coarser resolutions. Interpolation of BRIRs is significantly more challenging than for HRIRs, since it includes many reflections arriving at different times from different directions. Kearney, Masterson, et al. (2009) presented a method for interpolating room impulse responses using a dynamic time warping algorithm. However, the BRIRs were only approximated in their study, by a HRIR convolved with a separate room impulse response measurement. Additionally dynamic time warping is not suitable for real-time applications. Recently Garcia-Gomez and Lopez (2018) presented an algorithm to operate directly on measured BRIRs, which is better suited to real-time operation, only warping detected peaks. The algorithm operated separately on the two ear signals. A preliminary perceptual test showed that there may be some improvements over basic time-domain interpolation in terms of both localisation and timbre.

2.6.3 Global Interpolation Methods

Early global interpolation methods were presented using decomposition of the HRTFs into spectral functions via PCA, with associated reconstruction weights which can be interpolated for resynthesis at unmeasured positions (Martens, 1987). Kistler and Wightman (1992) found that localisation accuracy could be largely maintained whilst reducing required data to approximately 3 % of that measured (265 positions), using only the first five principal components. Various developments of such techniques have been proposed and Larcher, Jot, Guyard, et al. (2000) give an overview of these approaches. Carlile et al. (2000) presented the use of the spherical thin-plate splines to allow smooth interpolation of the PCA weights over the sphere, concluding that localisation accuracy can be maintained using as few as 150 HRTF measurements, with 15° angular spacing between measurements. These decomposition techniques simultaneously derive the set of basis functions and set of weights for synthesising filters, however both are then dependent on the individual HRTF dataset analysed.

An alternative approach is to project the HRTF data onto a pre-defined set of spatial basis functions, to obtain coefficients used to construct filters for any given direction. This has the advantage of decoupling the individual HRTF data from the filter synthesis functions. Jot, Walsh, et al. (2006) present a generalised description of HRTF interpolation by projection onto spatial functions, such as spherical harmonics, and then discuss the use of “discrete” panning functions, which are spatial basis functions where only a few (three or less) functions have non-zero weight for any given direction, such as given by vector base amplitude panning (VBAP) or techniques such as independent component analysis (ICA). This work highlights the close relationship between HRTF interpolation and virtualised loudspeaker panning techniques. However, in most practical implementations, virtualisation techniques use a lower spatial fidelity than often considered for HRTF interpolation.

Interpolation using spherical harmonics was first described in detail by Evans, Angus, et al. (1998), who made measurements on a 648-point Gauss-Legendre quadrature grid and investigated time-domain and frequency-domain methods, as well as using separate onset or phase processing. Subsequent works have extended the methods, including a theoretical basis derived from the wave equation, using regularisation to allow measurement grids that exhibit incomplete or irregular sampling over the sphere, and also allowing extrapolation to different source distances (Duraiswaini et al., 2004; Zotkin, Duraiswami, et al., 2009; Zhang et al., 2011; Pollow et al., 2012). The required measurement grid resolution is based on numerical considerations for accurate acoustic reconstruction using a wavefield expansion, which for a bandwidth of 20 kHz results in more than 2000 measurements (Zhang et al., 2011; Bernschütz et al., 2014). Interpolation at this resolution is therefore intended to give physically accurate HRTFs for arbitrary directions. These studies all involved the analysis of acoustic measurements of HRTF data; none performed any psychoacoustic validation of the interpolation techniques. Bernschütz (2016) presented work on recording and dynamic binaural reproduction of sound fields. This work incorporated the use of complex spherical harmonic HRTF interpolation with spherical harmonic series of degree 35, based on 2702 measurements. The study assumed that there is no perceptible difference between measured and interpolated HRTFs, since only small objective errors were observed. This assumption was verified by the listening experiment reported in appendix A.8.

Many studies use complex spherical harmonics to expand the complex HRTFs, yet since the excess phase response can be adequately modelled with pure delay (section 2.6.1), Romigh, Brungart, Stern, et al. (2015) investigate the use of real spherical harmonics to model the log-magnitude responses and separately the ITDs. This allowed for a more efficient representation whilst also giving better modelling of HRTFs in low-level contralateral directions. Analysis was performed using HRTF measurements at 277 directions, and spherical harmonics analysis up to 14$^{\text{th}}$-order (i.e. a maximum series degree of 14). Localisation experiments were carried out using individual HRTF measurements. Results suggest that only very small degradations occur with spherical harmonic series truncation order as low as 4, compared to free-field localisation. Interpolation at this order would in theory only require 26 HRTF measurements. It is concluded that measurement resolution requirements derived from theoretical acoustic considerations are likely gross overestimates from a perceptual standpoint. Whilst 4th-order representations are very likely to cause timbral changes, a 14th-order analysis based on the measurements showed mean magnitude modelling errors of only 1 dB up to 14 kHz. The experiments did not indicate whether this resolution gave detectable differences from measured data or real loudspeaker sources however.

2.6.4 Summary

Studies show that interpolation techniques can be used effectively to significantly reduce the required measurement resolution from that suggested by localisation blur in psychoacoustics studies. Both global and local interpolation methods have been applied, and time-domain and frequency-domain techniques exist. Decomposition of the HRTFs into minimum-phase and linear-phase components is beneficial for improving interpolation results and is applied widely. For inaudible differences using densely measured data, the number of measurements required appears to be in the order of 1000. When only equivalent localisation acuity is required, the resolution can be reduced such that only a few hundred measured directions are needed. Besides directional interpolation, techniques for interpolation or extrapolation of distance have been studied, as has interpolation of head-above-torso rotations. Interpolation of BRIRs presents significant additional challenges and the few available algorithms have high complexity.

2.7 Headphones

Headphones are clearly an important part of binaural technology. They enable reproduction of audio signals at the listener's ears with minimal crosstalk between the left and right channels. Of course headphones have been in use for many years and are more commonly used to reproduce stereophonic content. Headphone usage is increasing in prevalence. In 2017, worldwide sales grew 4 % to 362M units and retail value grew 29 % to $16.8B (Futuresource, 2018). This is thanks to the popularity of mobile media-playing devices. These factors are a key driver for development of and interest in binaural technology.

2.7.1 Headphone Design Goals

Given that many applications of binaural technology are likely to be delivered to audiences who are using a wide range of headphones, it is worth reviewing their characteristics.

Many varieties of headphone design exist. They can be categorised according to their fitting:

- *Circum-aural* – Over-ear headphones, which completely enclose the external ear.
- *Supra-aural* – On-ear headphones, which rest on the external ear.
- *Intra-aural* – In-ear headphones or ear-buds, which are placed in the cavum-conchae or the entrance to the ear canal.
- *Extra-aural* – A less common design type where drivers are not in contact with the ear but are suspended near to the ears.

Circum-aural headphones are generally categorised into closed-back and open-back designs, dependent on whether or not the enclosure is sealed to external sound. Most often headphones use dynamic electromagnetic transducers, though other types are used, such as planar magnetic or electrostatic transducers, often in more expensive models. Other design features that have become popular in recent years include active noise cancellation (for ambient noise reduction) and wireless signal connection via technologies such as Bluetooth.

The design goal for the transfer function of headphones is not standardised in practice. Commercially-available headphones differ greatly in their response (Møller, Hammershøi, et al., 1995; Breebaart, 2017). Headphones are often used to reproduce signals that were produced for loudspeakers, so the general aim is to compensate to some extent for the missing transmission paths, at least in magnitude response, to reproduce these signals through the headphones with the correct timbre. This results in a target headphone frequency response which is not flat.

Some approaches to standardisation of the headphone transfer function have been proposed. A free-field equalisation will produce a flat frequency response at the receiver, i.e. the listener's ears, for a sound source in free-field conditions from a reference direction, normally frontal. This was first proposed by Sakamoto et al. (1978). Alternatively a diffuse-field equalisation will produce a flat frequency response at the receiver in a diffuse sound field, with equal energy arriving from all directions. The two transfer functions will differ for frequencies with wavelength shorter than size of the head i.e. above approximately 2 kHz. The true free-field and diffuse-field transfer functions will also be specific to individuals. Theile (1986) recommended that diffuse-field equalisation is used, since free-field equalisation is only correct for a single frontal source and natural sound fields tend to feature diffuse components and a wide range of source directions. Diffuse-field equalisation of headphones also improves compatibility with stereo recordings and diffuse-field equalisation of binaural recordings improves compatibility with loudspeaker rendering.

Current standards from the IEC 60268-7 (2010) and the ITU specify that professional headphones for critical listening should be equalised according to the diffuse-field. Recommendation ITU-R BS.708 (ITU-R, 2002) gives details of a procedure for measurement of the diffuse-field transfer function for a listener in a reverberation room, with given tolerances for the headphone response. Møller, Jensen, et al. (1995) presented a method based on power-averaging of HRTF measurements. This was included in the ISO 11904 standard (DIN EN, 2002), which also specifies a target diffuse-field equalised response curve with a gain of approximately 10 dB between frequencies of 4 kHz and 5 kHz. Lorho (2009) assessed different target responses in a listening experiment and found that listeners preferred a 3 kHz peak of lower amplitude than in the ISO standard for both music and speech.

Headphone design targets have been revisited over the last 6 years, with extensive research, particularly by Sean Olive and Todd Welti at Harmann International. Listener preference for headphones seems strongly influenced by the magnitude response curve (Olive, Welti, and McMullin, 2013b), whilst non-linear distortions and the excess phase response do not have a significant impact at moderate listening levels (Olive, Welti, and McMullin, 2013a; Temme et al., 2014). Olive, Welti, and McMullin (2013a) and Welti et al. (2016) show that a reference headphone can be used effectively to simulate other headphone models by simulating their magnitude response. Preferences for headphone response appear to be listener, content, and headphone-design dependent (Olive and Welti, 2015; Olive, Welti, and Khonsaripour, 2016). However, when a common target response is used, one based on the response of a calibrated loudspeaker within a listening room gives superior perceptual performance to either of these traditional approaches for both diffuse-field equalised binaural signals (Fleischmann, Silzle, et al., 2012) and stereo signals (Olive, Welti, and McMullin, 2013b). This appears to make sense, since loudspeaker listening is generally somewhere between the two theoretical extremes of diffuse-field and free-field conditions.

Breebaart (2017) performed a survey of 283 headphones and found almost no correlation between the headphone price and the measured frequency response. Target curves have been developed and validated for different headphone types, which lead to optimal listener preferences, along with statistical models for predicting listener preference of the headphone response, based on deviations from these targets (Olive, Welti, and Khonsaripour, 2017b,a, 2018). Olive, Khonsaripour, et al. (2018) performed a survey of 156 models of headphone with a wide range of designs and prices varying from $6 to $4000. They were evaluated using the predicted preference scores. Little correlation with price ($r = 0.23$) was found for circum-aural headphones and no correlation for supra-aural or intra-aural designs. Circum-aural headphones gave the highest preference on average, followed by intra-aural then supra-aural. Open-back headphones gave higher predicted preference than closed-back headphones, with closed-back headphones typically having too much upper-bass energy. When professional and consumer classes of circum-aural headphones were compared, it was found that the consumer-target ones had similar responses but with higher bass below 300 Hz and professional headphones had lower variance and higher preference for lower prices. Predicted headphone sound quality preference had fairly low correlation with third-party review scores, although clearly factors other than sound quality influence consumer reviews. The model also predicted low preference scores for the diffuse-field and free-field target responses, especially for the intra-aural type. The authors conclude that these standards are outdated and largely ignored by the industry.

2.7.2 Headphone-to-Ear Transfer Function

The headphone-to-ear transfer function (HpTF) is influenced by the response of the headphone and its coupling to the ear. The ideal headphone system for binaural rendering would give a HpTF with a flat frequency magnitude response and linear phase, allowing high-fidelity production of the rendered binaural signals. In reality this is far from the case, perhaps unsurprisingly the HpTF has been shown to be highly variable across headphone models and individuals. This makes a one-size-fits-all correction filter unlikely to add benefit. Whilst accurate equalisation requires in situ measurement, benefits can still be obtained by approximate equalisation and so study of the variability of the HpTF should enable design of suitably robust methods for real world applications outside of laboratory environments.

Møller, Hammershøi, et al. (1995) measured the HpTFs of 14 different headphones with varied designs on 40 human subjects. At low frequencies the HpTFs are quite smooth, but at higher frequencies (above 8 kHz) they are characterised by narrow peaks and notches. Large deviations in responses were observed between different headphone models. Considerable inter-individual variation was also observed, with deviations as high as ±20 dB at some frequencies on certain headphone models. Pralong and Carlile (1996) also observed large inter-individual differences in HpTF and they recommended that individual headphone HpCFs are used even when rendering with non-individual HRTFs, since a non-individual HpCF caused errors in rendering of up to 10 dB at some frequencies. The study of Christensen, Hess, et al. (2013) confirms that individual HpCFs are needed for accurate correction, though noting inter-individual similarity up to 4 kHz.

The method of measurement will also introduce differences in the obtained HpTF. In (Møller, Hammershøi, et al., 1995), for open ear canal measurements significant variation was observed above 2 kHz whereas blocked-ear canal measurements only varied significantly above 7 kHz, with common patterns visible up to 12 kHz. This is because blocked ear canal measurements exclude the individual effects of the transmission along the ear canal and the termination impedance at the ear drum. The positioning of the microphone still has an influence here, Lindau and Brinkmann (2012) demonstrate that using moulded silicon ear inserts instead of foam leads to better repeatability.

Whilst blocked ear measurements also present practical advantages, the effect of what Møller calls the pressure division ratio (PDR) must be compensated for accurate equalisation. This is caused by a change in impedance due to the coupling of the headphones, which alters the pressure signals at the entrance to the ear canal and so at the eardrum. The PDR cannot be measured with the blocked ear canal measurements alone, measurements with the ear canal open must also be obtained. Møller (1992) explains that the headphones should exhibit *free-air-equivalent coupling (FEC)* for the PDR to be unity, and Møller, Hammershøi, et al. (1995) show that the effect of the PDR is non-negligible for most headphones, even those with an open design. Most headphones, except extra-aural designs, have non-flat PDR above 2 kHz, although up to the highest measured frequency of 7 kHz for many open-backed models the mean showed less than 2 dB deviation, which the authors suggest might be acceptable. Masiero and Fels (2011) measured the PDR for two open-backed headphone models on a HATS with ear canal simulators and found it to be less than 2 dB below 10 kHz. However, Fleischmann, Silzle, et al. (2012) reports informal findings that the PDR has a significant perceptual effect for many headphones and so reports that it should be compensated. In practice, the pressure-division-ratio appears often to be ignored.

Lindau and Brinkmann (2012) measured HpTFs on 25 individuals with a single pair of headphones (Stax Lambda electrostatic circum-aural type) at the blocked ear canal entrance. They observed four different regions of the magnitude response. Below 200 Hz they observed a small amount of variance due to leakage effects, approximately ±3 dB. From 200 Hz up to 2 kHz the observed variation was less than 1 dB, but above this the variance increases to ±3 dB. Above 5 kHz pinna effects cause much larger variations; they are asymmetric, varying between 7 dB and −11 dB.

Intra-individual variation is also an issue, since the HpTF varies due to reseating of the headphones, as demonstrated by Møller, Sørensen, et al. (1996) and Kulkarni and Colburn (2000). Møller (1992) described a Thévenin equivalent model for the external human ear and presented the problem of headphone equalisation in these terms. For circum-aural headphones, this model is valid up to approximately 4 kHz, the headphone and ear coupling acts like an acoustic cavity, besides variance due to leakage caused by the fitting. At higher frequencies standing waves start to build up inside the cavity (Schmidt, 2009). This explains why the pressure at the eardrums becomes very dependent on headphone positioning at higher frequencies, and therefore the Thévenin equivalent model is not valid at this frequency range. The variation seems to be greater for supra-aural type headphones than circum-aural type (Pralong and Carlile, 1996; Kulkarni and Colburn, 2000). Paquier and

Koehl (2015) demonstrated that this variation due to positioning is audible. This presents challenges for accurate equalisation with a microphone at the blocked ear canal entrance, since the headphones must be temporarily removed to retrieve the microphones. However, Møller, Sørensen, et al. (1996) demonstrate that variance across multiple headphone placements is significantly reduced when the subject is able to adjust the headphones themselves according to comfort. Kulkarni and Colburn (2000) suggest that a HpCF for an individual based on the average of HpTFs for several headphone placements will give better results than that of a single measurement. Christensen, Hess, et al. (2013) noted that all-pass components were present in their measured HpTFs. This will have consequences for the temporal performance if not equalised, however such effects are dependent on the individual and the precise headphone positioning. Incorrect equalisation of such effects may cause more errors than discarding them.

Schärer and Lindau (2009) made 10 HpTF measurements with re-positioning for seven sets of professional-grade headphones on a HATS. Again, large variation was shown in the response of different models, suggesting that standardised design goals are not followed. It seems clear that compensation of the specific headphone is required. Large variations were observed in the notch depth and frequencies at approximately 8 kHz–9 kHz and 14 kHz, which are likely to be due to destructive interference in the pinna cavities. The supra-aural headphone model did not show these notches. Low frequency variation with repositioning was also common in many headphones. The extra-aural headphone model showed a stable response across frequency, which suggests it may be advantageous in a binaural reference system. This model of headphone has been discontinued, but Erbes et al. (2012) presented a custom extra-aural headphone designed for binaural reproduction.

2.7.3 Equalisation Filter Design Techniques

The goal of equalisation in binaural systems is to correct the transfer function of one or more components of the measurement or reproduction system, so that the correct binaural signals can be precisely produced at the eardrums. This target is represented by:

$$C(f).Hp(f) = 1 \tag{2.6}$$

where C is the measured transfer function and Hp is the resulting correction filter, and so the direct inversion of the measured transfer function is given by:

$$Hp(f) = \frac{1}{C(f)}. \tag{2.7}$$

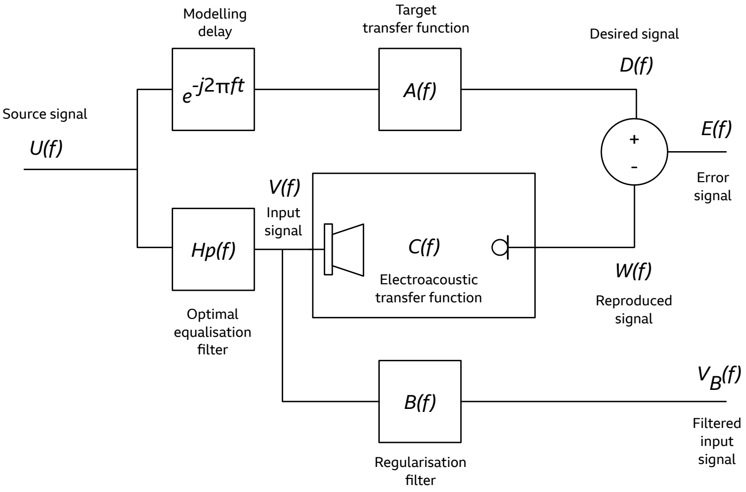


Figure 2.16: Filter design for single-channel electroacoustic frequency-domain inversion problems, after Kirkeby and Nelson (1999)

Such equalisation is an electroacoustic inversion problem, which in practice cannot be performed ideally. The ideal filter will often be non-causal, and can be ill-conditioned at certain frequencies, resulting in excessive amplification, which could cause distortions in the signal chain. Additionally measurement noise and non-linearities in the system prevent perfect equalisation using a LTI filter.

The problem is more sensibly formulated as a least-squares error minimisation problem, with a cost function that trades-off the accuracy of the inversion and the 'effort' required in equalisation, which is known as regularisation. Figure 2.16 provides an illustration of the problem, a detailed overview is given by Kirkeby, Nelson, et al. (1998).

The least-squares problem aims to minimise the frequency-domain cost function:

$$J(f) = \frac{1}{2}E(f)E^*(f) + \beta\frac{1}{2}V_B(f)V_B^*(f) \tag{2.8}$$

The following equation gives the analytical solution to this problem:

$$Hp(f) = \frac{C^*(f)A(f)}{C(f)C^*(f) + \beta B(f)B^*(f)} \tag{2.9}$$

where A is the target function, C is the measured transfer function, and B is the frequency-dependent regularisation function, which together with the parameter β limits the effort in the equalisation filter. The symbol * denotes the Hermitian operator. When no regularisation is applied $\beta = 0$ and when frequency-independent regularisation is applied $B(f) = 1$ for all values of f. Similar formulations can be made in the time domain.

2.7.4 Headphone-to-Ear Correction Filter Design

The design of the target function and the regularisation parameters should be considered with reference to the particular inversion problem. There are several factors that must be considered during HpCF design.

- Direct inversion of a mixed-phase system yields an acausal impulse response.
- Equalisation at very low and high frequencies (where measurement transducers have a poor response) leads to a high-energy filter.
- Accurate compensation is limited to a specific positioning of the specific headphones on the listener.
- Peaks in the residual transfer function will be more perceptible than notches.
- HpTFs have varying presence of all-pass components.
- The equalisation should not alter the loudness of the signals.

A mixed-phase system has in it some delay which when inverted causes the resulting filter to be acausal. The phase response of the target can be set to avoid acausality (Norcross et al., 2006), either incorporating modelling delay (i.e. a linear-phase target) or specifying minimum-phase and so not equalising all-pass components of the HpTF.

To avoid excessive gain at frequencies outside of the reproduction range for the headphones, a band-pass target can be used (Schärer and Lindau, 2009), instead of a unit impulse. The band limitation at high frequency is outside of the audible frequency range. It is the low frequency behaviour of equalisation filters that is of particular interest, especially since high energy at low frequencies results in long filters. The measurements in Schärer and Lindau (2009) suggest that 50 Hz is a sensible low-frequency limit for equalisation, since there is little energy in the HpTF of most headphones below this frequency. Alternatively Masiero and Fels (2011) processed the HpTF to ensure a flat magnitude response below the first maximum in the measured HpTF prior to inversion. This prevents applying excessive low-frequency attenuation. Regularisation may also be used at very high-frequencies to prevent excessive gain in the HpCF.

The presence of all-pass components in HpTFs has been shown to vary. Minnaar, Christensen, et al. (1999) investigated the audibility of all-pass components in binaural rendering in terms of a Q-factor threshold. If an all-pass component is present in the HpTF and has a high Q-factor, above a signal-dependent threshold of audibility, then ringing will be heard in the residual if it is not equalised. If an all-pass component is not present but the equalisation compensates for an all-pass component then the residual will be acausal, resulting in pre-ringing effects. Pre-ringing effects were shown to have lower detection thresholds than post-ringing effects due to causal high Q-factor all-pass components. Minnaar, Christensen, et al. (1999) state “that it will be more safe not to equalize for an all-pass that is there than to equalize for an all-pass that is not there”.

In (Schärer and Lindau, 2009), some subjects noted the transient response being a factor differentiating the real and virtual stimuli, which was thought to do with pre-ringing in the headphone equalisation filter. A minimum-phase target function may be used to avoid pre-ringing in the resulting impulse response, at the expense of non-constant group delay, i.e. having to change the structure of the signal phase. However, the short length of most headphone equalisation filters means that the effect of pre-ringing in a linear-phase design may be masked. Moore (2003, p.110-114) reveals that more backward masking occurs than forward masking at very short time intervals.

In order to minimise the effect of the headphone equalisation on loudness, normalisation is required. However, the impression of loudness will vary, both according to the input signals and the nature of the errors of imperfect equalisation. Masiero (2012, p.72) recommends normalising according to the RMS value of the equalisation filter in the frequency range of lower variability (below 4 kHz).

Robust correction of the HpTF when not measured in-situ is challenging due to the large variability previously described. The HpCF should be designed to achieve acceptable equalisation by minimising the perceptual effects of the inevitable inaccuracies. Bücklein (1981) showed that listeners are more sensitive to spectral peaks than notches when listening to speech, music, and noise stimuli. Therefore headphone equalisation filters should avoid introducing peaks in the residual, which can be challenging when notches of varying position and depth are seen in HpTFs. Averaging of multiple HpTFs has been recommended to avoid this issue e.g. (Møller, Jensen, et al., 1995; Kulkarni and Colburn, 2000). It is also often recommended that spectral smoothing be applied to the HpTFs before inversion. Another approach is to use frequency-dependent regularisation to minimise the effort in equalising deep notches. The following section reviews previous studies on the use of HpCFs in situations where direct measurement of the HpTF to be corrected is not feasible.

2.7.5 Headphone-to-Ear Correction Filters in Uncontrolled Scenarios

For applications outside of the laboratory, precise in situ equalisation of the HpTF is not possible. From the various sources of variability observed in HpTF, e.g. headphone model, individual, and headphone positioning, it must be considered which are the most significant and important aspects to correct for. This prioritisation informs the degree to which HpCF approaches can be generalised in real-world applications.

Hammershøi and Møller (2005) state that “generally speaking, the transfer function of the headphones used for the reproduction of binaural signals must be known and compensated for”. So the HpCF should normally be determined for the specific headphone model. Further they state that it is feasible to create effective individual HpCFs, due to the low variation in the repeated measurements. However, the large variation in HpTFs between individuals means that the HpCF “probably should be designed individually, at least for critical purposes.” Whilst it seems likely that an individual HpCF will give better results, Kulkarni and Colburn (2000) consider that an average filter based on measurements across multiple individuals may still be useful. Such generic HpCFs are unlikely to provide great improvements to the rendering of spatial cues, but as stated by Martens (2003) may achieve “general correction for tone colour”.

When individual HpCFs are applied, studies show benefits to the spatial reproduction, in terms of better externalisation and reduced front-back confusions. Kim and Choi (2005) used a Weiner filter algorithm to generate an individual HpCF from probe microphone measurements. The microphone was however removed before rendering. A distance localisation experiment with white noise stimuli was conducted with five subjects. It showed improved out-of-head localisation with the use of equalisation, both with individual and non-individual HRTFs. Guru et al. (2010) investigated the role of individualised HpCF in binaural rendering with individual HRTFs. The inverse filter was based on the average of multiple HpTFs measurements with re-seating, which were first smoothed in the time domain with linear predictive coding. Listeners who under unequalised conditions show front-back discrimination performance at better than chance levels achieved significantly better front-back discrimination for binaural rendering with individual headphone equalisation filters. The spectral centroid of the rear-presented sources was adjusted so that the timbre was more similar to frontal sources, thus requiring the listener to utilise spatial cues for discrimination.

Masiero and Fels (2011) presented a “perceptually robust” method for generating individual HpCFs, with the aim of minimising the impact of equalisation errors due to headphone positioning changes. The design was based on the research of Bücklein (1981), which suggests that spectral peaks are more audible than notches, so the equalised signal should aim not to create spectral peaks. The mean μ and standard deviation σ of the magnitude responses measured for multiple headphone fittings were used, and the response given by $\mu + 2\sigma$ was inverted. 1/6-octave smoothing was applied to the HpTFs to avoid errors introduced by inconsistent sharp spectral notches in the measurements, which could otherwise yield inappropriate peaks in the equalised response.

A notch-smoothing approach is described by Masiero (2012, p.72). A smoothed version of the magnitude response is compared to the original and notches are indicated where the smoothed version is greater than the original by more than a certain threshold. Within these regions, an interpolation between the original and the smoothed response is performed to reduce the notches.

As discussed in section 2.4.3, HRTFs are often diffuse-field equalised, assuming that this improves results in uncontrolled scenarios, since diffuse-field equalisation of headphones is recommended by standards bodies. For example, in Middlebrooks (1999b) no HpCF was used, and the authors reported that the headphone manufacturer's specifications indicated a flat diffuse-field response. However, recent studies of headphones reveal that such a response is rarely exhibited (see section 2.7.1).

There is evidence to suggest that generic HpCFs that are specific to the headphone but not to the individual can still provide benefit. Martens (2003) investigated a HpCF based on the average of measurements across individuals, with smoothing according to auditory critical bandwidths. The generalised HpCF was modified with various frequency scalings and a discrimination test was performed. Within a limited range, deemed equivalent to inter-individual differences in HpTF, binaurally rendered speech stimuli could not be discriminated. It was concluded that using a generalised HpCF for a given headphone model is justifiable when also rendering with non-individual HRTFs. Møller, Jensen, et al. (1996) compared individual and generic HpCFs in a localisation experiment with non-individual HRTFs and showed only very minor degradations when using a generic filter.

Schärer and Lindau (2009) assessed many methods of HpTF equalisation in a subjective test where listeners compared binaural rendering, with BRIRs in a head-tracked real-time system, to a real loudspeaker reference. The HpCFs were based on non-individual measurements on the same dummy head as used for the BRIR measurements. Various smoothing and regularisation methods were employed to derive the HpCF, attempting to avoid excessive peaks in the equalised response due to variability in the HpTF, again citing the work of Bücklein (1981). A clear increase in similarity to the reference stimulus was shown due to equalisation methods over no equalisation. The equalisation methods showed broadly similar results, though slightly better performance was seen with high-pass regularisation functions, which would have prevented large gain in the HpCF in the high-frequency region. Assessors reported attributes that contributed most to the differences from the real loudspeaker. High-frequency boost and ringing were cited most often, with general timbral differences also noted.

Lindau and Brinkmann (2012) carried out similar listening experiments to Schärer and Lindau (2009), evaluating HpCFs based on "generic" (averaged across subjects) individual and non-individual HpTF measurements. When the non-individual HpCF was measured on the same HATS as the BRIRs, a significant improvement over the individual HpCF was measured, which was unexpected. After further objective investigation into the effects, the authors suggested that the non-individual HpCF may be applying a "kind of de-individualisation", compensating high-frequency pinna-related cues especially. The generic HpCF was not significantly different from the individual HpCF. A second listening test comparing the individual HpCF with a non-individual HpCF not from the HATS used in the rendering showed that the individual equalisation was significantly better. Several regularisation functions were evaluated, including a high-pass filter (since notches are most prevalent above 4 kHz) and the smoothed inverse of the transfer function to be equalised, which reduced the adaptation effort in specific regions where notches occur. It was observed initially that the high-pass filter failed to correct for general damping at high-frequencies, which can lead to a muffled sound. However, there were no significant differences in results between filter designs with different regularisation functions. The phase of the target function also showed no significant effect, indicating that a minimum-phase target might be acceptable. Pink noise and drum kit signals were used in these studies.

It is clear that headphones exhibit a wide range of HpTFs and it is better to correct for the specific headphone model, even if generic to individuals. But in many applications it is not possible even to know this information. A relevant question then is whether this variability can be effectively modelled, such that even more generic equalisation approaches can be applied.

Fleischmann, Silzle, et al. (2012) found that the headphone response target of a calibrated frontal loudspeaker in a listening room was preferred by listeners when listening to diffuse-field equalised non-individual binaural signals. For high quality reference headphones (Stax SR-404 and Sennheiser HD-600), this equalisation did not show large improvements over the unequalised response. However, for consumer supra-aural headphones a huge improvement was observed. Fleischmann, Plogsties, et al. (2013) then obtained HpCFs for 13 consumer headphone models using expert listeners, who manually adjusted an equaliser control to match a frontal loudspeaker target. The headphones were first compensated to a flat frequency response before this adjustment by measurement on a HATS. Once these adjustments had been made for each of the headphones, principal component analysis was performed to obtain a headphone equalisation control based on only the mean and first principal component filters. This allowed correction of the headphone responses so that the deviation from the target function was within the same range as for the high quality reference headphones.

2.7.6 Summary

Headphones vary greatly in design and in their acoustic characteristics on the ears of listeners, and existing standards are largely ignored by the industry. The headphone-to-ear transfer function (HpTF) varies not only by headphone model but also by individual listener and to a lesser extent due to the positioning of the headphones. Headphone-to-ear correction filters (HpCFs) are designed to reduce the errors introduced to binaural signals during headphone reproduction.

Whilst in situ equalisation is possible in a laboratory, thus creating a new HpCF each time the headphones are used, this is not practical for wider application. With open headphones that do not significantly modify the impedance at the ear canal, measurements at the blocked ear canal entrance can be used. A HpCF for a specific individual and headphone model can be created from an average HpTF obtained by means of multiple headphone fittings. Designs avoid creating narrow spectral peaks in the equalised response that may arise from fitting differences; notches are made more likely but these will be less audible. Individual HpCFs have been shown to improve spatial aspects of binaural rendering, giving better out-of-head localisation and fewer front-back reversals.

Whilst generic HpCFs will likely only give fairly broad tonal correction, this seems still to provide benefits in binaural rendering, giving closer approximations to the target sound source. Evidence shows that the preferred target response for headphone listening approximates that of a calibrated frontal loudspeaker in a listening room, including when reproducing diffuse-field-equalised binaural signals. Existing consumer headphones can be adapted to match this target with reasonable accuracy, yet currently this would still require a skilled user. The headphone industry may converge on standard targets in future, but new developments such as self-calibrating headphones (Backman et al., 2017) also show promise for binaural applications.

2.8 Head Tracking

The use of head-tracking in binaural rendering systems was introduced in section 2.4.5. This section describes the perceptual effects of adapting to head movement in binaural rendering. It then reviews the performance requirements for head-tracking systems and discusses the range of head tracking technologies available.

2.8.1 Perceptual Effects of Head Movement in Binaural Systems

The effects of head movement in natural hearing also operate in dynamic binaural systems. In an early experiment, Koenig (1950) reported that when a dummy head microphone in another room was rotated in sympathy with the listener, a human speaker in the same room as the dummy head could be well localised and the listener could home in on the speakers. Without rotation, the speaker was always localised to the rear of the listener.

Front-back reversals are significantly reduced in dynamic binaural rendering. Wightman and Kistler (1999) investigated the rate of front-back reversals when using individualised binaural rendering with head tracking, as well as with real loudspeaker sources. The measurements and testing were carried out in anechoic conditions and the stimuli were 2.5 s bursts of white Gaussian noise. When the listeners were allowed to move their head during the stimulus, front-back reversals were greatly reduced with both real and virtual sources. A second test was performed without head tracking, but source position could be controlled. When the listener could control the position of the source, front-back reversals were resolved, but not when the experimenter controlled the movement. This implies that there may not be adequate information in the dynamic auditory cues alone, additional sensory input is required to indicate the direction of movement of either the source or the listener's head. The use of head movements was similar with real and virtual sources, but movement patterns were highly individual.

This reduction in front-back reversals has also been shown when using non-individual binaural systems. Horbach et al. (1999) demonstrated this using a HATS attached to a motorised rotary mount that rotated in concert with the listener's head, with the listener in an isolated room to the microphone, similar to Koenig (1950). With tests carried out in anechoic and reverberant environments, head movements gave significantly better azimuth localisation accuracy by resolving front-back reversals.

Wenzel (1995) investigated the importance of dynamic ITDs and ILDs for localisation using a head-tracked non-individual binaural rendering. A non-individual HRTF set was decomposed into minimum-phase filters and frequency-independent delays, such that broadband ITD and ILD could be independently varied, as introduced in (Kistler and Wightman, 1992). Three conditions were used, either ITDs and ILDs were both correctly correlated to head movement and source position, or ILDs or ITDs were fixed to the value for 0°. These conditions were also presented statically, without head tracking. When correct variation of both interaural cues was used the frequency of reversals was comparable with that found for individual binaural synthesis in previous experiments. However, the accuracy of the ILD cue seemed to be more important than the accuracy of the ITD cue for head movements to provide a benefit. This indicated that the role of spectral cues and ILDs may be more significant when head movements are involved than in the static scenario, where ITDs have been found dominant (Wightman and Kistler, 1992). The authors summarised the results by saying that "the data suggest that localization tends to be dominated by the cue that is most reliable or consistent, when reliability is defined by consistency over time as well as across frequency". Other studies have shown the benefit of head tracking for reducing front-back reversals in binaural rendering, for example Begault, Wenzel, and Anderson (2001) showed a consistent reduction of approximately 50 % across anechoic or reverberant rendering and individual or non-individual HRTFs.

Exploratory head movements in dynamic binaural rendering enable accurate localisation beyond just resolving reversals. With closed-loop localisation, using head movements to home in on a continuous sound source, localisation accuracy in binaural rendering can be comparable to real sources. Bronkhorst (1995) used a closed-loop localisation test with dynamic binaural rendering and compared results to that for real sources. It was shown that with individual HRTFs, localisation accuracy of virtual sources was equivalent to that for real sources. However, when non-individual HRTFs were used, the accuracy was generally worse, although it varied according to which non-individual HRTF set was used.

Romigh, Brungart, and Simpson (2015) showed that an individualised binaural rendering system with head tracking could achieve localisation accuracy equivalent to free-field localisation of real loudspeaker sources. This was found for continuous Gaussian white noise sources and for short noise bursts of 250 ms, where effective use of head movements is not possible. Polar angle error was significantly lower for the continuous stimuli than for the short bursts. This system utilised accurate in-situ measurement and equalisation. The head-tracked rendering was based on processed HRTFs, whereby the measured transfer functions were windowed, minimum-phase versions obtained and then truncated, so enabling separate interpolation of onset delays. The experiment used ear-bud style intra-aural headphones but mounted extra-aurally using a semi-rigid wire. This ensured FEC and allowed microphones to be inserted and removed from the ear canals without moving the headphones.

In an initial experiment, the assessors wore these headphones during the HRTF measurements and real and virtual stimuli were interleaved during the experiment. The presence of the headphones will have introduced some distortions to both the real and the virtual sounds. It appears that this caused negative effects in the minimum-phase processing for head-tracked rendering. In this experiment, when short noise bursts of 250 ms were used, localisation accuracy was worse than for free-field stimuli. The angular great circle errors were 3° larger on average, and this was particularly due to front-back and up-down (polar angle) errors rather than left-right (lateral) error. For the short burst stimuli, full-phase HRIRs were also used, with no head tracking, and these were found to give better localisation accuracy than the head-tracked minimum-phase condition. When continuous stimuli were used, so allowing effective use of head movements, the errors for head-tracked rendering were equivalent to that for real loudspeakers.

Head tracking also results in improved distance localisation and externalisation i.e. out-of-the-head localisation. Durlach, Rigopulos, et al. (1992) conjectured that head movement might be a factor in the limited externalisation often observed with binaural rendering. However, in early research on the topic, there was mixed evidence regarding the effects. Wenzel (1995) showed a small increase in reported perceived distance when head tracking was introduced to non-individual binaural rendering. This was presented as an increase in percentage of externalised events, defined as those with a distance of more than 4 inch. In (Wightman and Kistler, 1999), although not the main focus of the work, the distance of auditory events in individual binaural rendering was consistently reported as 1 m–2 m from the listener, whether dynamic head tracking was used or not. It was concluded that “the dynamic cues provided by head movements are not required for externalization of a virtual source”. Pellegrini (2001) reported that when using a full AVE with a parameterised room model, head tracking improved distance localisation, yet no evidence is provided to support this claim. Loomis et al. (1990) reported that head tracking can provide externalisation, but only gave informal findings, whereas Begault, Wenzel, and Anderson (2001) found, in a direct comparison of individualisation, reverberation and head tracking, that head tracking did not have a significant effect on perceived distance.

Brimijoin, Boyd, et al. (2013) addressed the effect of head tracking on externalisation experimentally. In one experiment with loudspeaker sources in an anechoic chamber, head tracking was used to try to cause in-head localisation. Loudspeaker stimuli were kept in front of the listener’s head by panning around a loudspeaker ring according to head orientation. A loudspeaker stimulus that moved with the listener’s head was more likely to be localised in-head. In a second experiment, headphone stimuli were created with individual BRIRs and head tracking was used in an attempt to increase externalisation. Measurements were made in the range ±25°. Measurements were also made without a head present, using a spaced microphone pair. These were mixed with the individual BRIRs in different amounts, the ratio of head-present to head-absent filters ranged from 0 to 1, to investigate whether degraded binaural impulse responses still afford externalisation. Participants gave a binary response to the question: “Did the signal sound like it was coming from out in the world or from inside your head?” Four conditions were used: with and without head tracking, and with and without head movements. Head tracking in binaural rendering afforded an increase in externalisation over the static head conditions when the BRIRs were degraded. With unimpaired BRIRs externalisation was high even without head movement. When head movement was allowed and the rendering did not update dynamically, externalisation was poor. Loudspeakers were visible during the test, so a visual capture effect may well have positively influenced the externalisation percept.

Hendrickx et al. (2017) investigated the effects of dynamic rendering on externalisation when using non-individual HRTFs. Recordings of speech were made in a reverberant environment with a six-channel equal segment microphone array (Williams, 1991) and post-processed with non-individual binaural rendering. It was shown that head tracking gave a substantial improvement in externalisation, and that the sensation persisted after head movements stopped. This improvement was most significant for sources located in front or behind; lateral sound sources are often externalised even without head tracking e.g. (Begault and Wenzel, 1993).

Pörschmann, Arend, et al. (2017) found that head tracking had no significant effect on distance estimation for near-field sounds. Interestingly, Wersenyi (2009) showed that simulated small random movements in binaural rendering enhanced the out-of-head impression, without use of a head tracking system. The random motion did not enhance front-back discrimination or localisation however.

2.8.2 Perceptually-Relevant Aspects of Head Tracking Systems

There are a number of factors relating to head tracking systems themselves which have perceptual implications when applied to dynamic binaural rendering:

- *Degrees of freedom* - The number of orthogonal axes of positional data reported by the device, which can be up to 6 for full position and orientation tracking of a single rigid body.

- *Accuracy* - The degree to which the position data reported by the system corresponds to real-world position.

- *Precision* - The fineness with which position data is reported by the system.

- *Range* - The limits of movement that can be detected in each available dimension.

- *Stability* - The smallness of drift or jitter away from the real-world position, both when the listener is stationary and during movements. Related to accuracy.

- *Absolute/relative* - Some systems are only capable of providing information about relative movements from an arbitrary starting state, whereas others provide absolute position data in a coordinate system relative to the receiver(s).

- *Latency* - The delay between movement of the listener and the corresponding change in data reported by the system at the output interface. Latency is also considered at the total system level, including the audio rendering delay, i.e. the delay between movement of the listener and the corresponding change in audio output.

The ideal head tracking system would have specifications whereby any limitations in these factors are below perceptual thresholds. Several studies have explored criteria for the use of tracking systems in dynamic binaural technology. They will be reviewed in the following sections.

Laitinen, Pihlajamäki, et al. (2012) investigated the influence of head tracking specifications on the impression of naturalness, focusing particularly on spatial aspects. First-order ambisonics signals were rendered binaurally using head tracking and parametric spatial audio coding. The head tracking data from an optical tracking system were treated as a reference and modified in several ways to simulate common errors in consumer-level head tracking technologies; limits to degrees of freedom, angular range, stability, accuracy and update rate were simulated. This study will be discussed in the following sections where relevant.

2.8.2.1 Degrees of freedom of head movement

The number of degrees of freedom (DoF) of head position that can be measured depends upon the head tracking technology used. The requirements are very much application-dependent, since natural listener movement is dependent upon activity. For example applications involving a seated listener such as a virtual concert or cinema experience might only need tracking of orientation within a limited range around the frontal region, whereas an interactive virtual environment might require also tracking head translation within a room-scale volume. The latter scenario typically requires tracking translation of the head position in 3 DoF and also changes in the head orientation in 3 DoF, which is often called 6 DoF tracking. Such systems might potentially require additional DoF, such as tracking the hands for natural gestural interaction.

Tracking only orientation, assuming no head translation, can give good perceptual results for a seated listener. Dynamic binaural rendering systems have been shown to give highly realistic simulations in such scenarios (Lindau and Weinzierl, 2012; Brinkmann, Lindau, and Weinzierl, 2014; Romigh, Brungart, and Simpson, 2015).

In many binaural systems, anechoic HRIRs are measured on a static head at a range of source positions covering the surface of a sphere. Head rotation is then approximated by source rotation in the opposite direction. Distance is simply represented by frequency-independent intensity and time delay changes. This does not account for near-field effects or audible effects of relative movement of different body parts. Brinkmann, Roden, et al. (2015) have demonstrated that head-above-torso rotation has audible differences to head-and-torso rotation. Suggesting that the tracking of the torso separately from the head might be required for high fidelity rendering, though this is not yet common in applications.

Studies reviewed in section 2.2.7 show that head yaw rotation is often dominant over pitch or roll movements. It appears most important to incorporate yaw rotation into binaural rendering. In (Laitinen, Pihlajamäki, et al., 2012) the three degrees of rotation were enabled/disabled in various combinations. The rendering used anechoic HRTFs, allowing head rotation in each of the three DoFs to be approximated by source rotation in the opposite direction or fixed at 0°. The results indicated that enabling yaw tracking has a much greater effect on naturalness than on pitch and roll. This is unsurprising given observed head movements in natural listening, although in the test signals there were no sources above or below the horizontal plane, so the importance of yaw movements may be overemphasised. There was no significant improvement from adding pitch and roll tracking, although there was a slight trend for increased naturalness ratings, also with narrower confidence intervals.

Mackensen, Fruhmann, et al. (2000) found that inclusion of pitch rotations in dynamic rendering with non-individual BRIRs gave a slight improvement in azimuth localisation but did not greatly reduce vertical localisation errors. Pellegrini et al. (2007) took this to show that "vertical head tracking is of minor importance for a plausible reproduction in reference room situations". However, it is expected that the plausibility of an AVE will be negatively affected if head pitch or roll rotations are made and not handled in the rendering, since it appears that listeners are sensitive to changes when making such movements. As discussed in the following section, Lindau and Weinzierl (2009) investigated discretisation of BRIR measurements in dynamic rendering, finding that listeners can detect differences at resolutions of a few degrees for pitch and roll movements.

2.8.2.2 Accuracy and precision

The accuracy and precision of a tracking system for binaural rendering should ideally be greater than that of auditory localisation processes, so it is relevant to refer to studies of localisation in natural listening (section 2.2.6). MAAs as low as 1° have been measured (Mills, 1958). Since some head tracking technologies are capable of very accurate measurement, the resolution of the dynamic binaural system is often limited by the HRIR/BRIR measurement resolution. Interpolation techniques are often used to increase this resolution artificially (see section 2.6).

Sandvad (1996) investigated the resolution required in dynamic binaural systems. The resolution was ranged from 1° to 13° by varying the grid of available HRTF measurements. The head tracker used in the test reportedly gave accuracy to approximately 1°. Localisation accuracy and response times were measured. No significant effect of resolution was observed, although a trend in increasing response times could be seen. It should be noted that the fine resolution of HRTF measurements was obtained by interpolation of measurements made at relatively low resolution (11.25° steps in θ and 22.5° in ϕ).

As reported in section 2.6.2, Lindau and Weinzierl (2009) investigated the spatial resolution of measured BRIRs required in dynamic binaural rendering. For separate head movements in yaw, pitch, and roll angles, listeners were asked to discriminate BRIR datasets of reduced resolution from a 1° reference resolution. For pink noise stimuli, yaw and pitch resolutions of 4° and 3° respectively could not be detected by 95 % of listeners. Two of the 21 listeners were able to detect a pitch resolution of 2°. For a frontal source, rolling head movements were much less critical, likely due to there being little change in inter-aural differences. In a second test, Schultz, Lindau, et al. (2009) showed that, for a source directly above the listener, roll resolution is much more critical, with a detection threshold of 2° for 95 % of listeners.

2.8.2.3 Range

In (Laitinen, Pihlajamäki, et al., 2012) angular range restriction caused degradation of naturalness under all tested values, so the angular range should be greater than ±30°. Degradation due to range restriction was content dependent, a more complex orchestral scene appeared to require greater range than a single violin. The range of head movements in natural listening was studied by Kim, Mason, et al. (2013) (see section 2.2.7), and it has been shown that in some scenarios head movements are focussed around the frontal region, so full 360° range may not be required always. Lindau and Weinzierl (2009) used ranges of ±80°, ±35° and ±30° for yaw, pitch and roll respectively, considering comfortable rotation ranges for a seated listener.

2.8.2.4 Stability and Absoluteness

Cheap inertial tracking systems are likely to exhibit drift and instability, and in general this should be avoided, especially for systems used in research experiments (unless that behaviour is the subject of the research). In (Laitinen, Pihlajamäki, et al., 2012), assessors were relatively insensitive to random bias that gradually changed over time, within limited bounds of ±20°. Results for these instability conditions were not significantly different from the reference tracking condition, though confidence intervals were quite large over this variable. Results may be influenced by the particular algorithm used to introduce artificial bias and drift.

The idea of absolute or relative tracking is linked to stability. Some motion detection devices, such as gyroscopes and accelerometers, are able to indicate changes in motion but not absolute position. Errors can accumulate over time which lead to drift. Other systems will retain a stable calibrated reference point. The requirement for a fixed frame of reference depends on the application. Where an external reference exists, such as related visual elements in the environment, absolute tracking is necessary. Sometimes this may not be required or is even undesirable, for example in an audio-only mobile experience. Algazi, Dalton, et al. (2005) describe the situation where a listener is walking and turns a corner and then keeps walking, it would not be appropriate for the previously frontal scene to then be held to the side of the listener.

2.8.2.5 Update rate and latency

Update rate and latency are related. Pellegrini (1999) states that they correspond to the perceptual counterparts "responsiveness (latency) and smoothness (update rate)". Update rate is due to temporal sampling of the scene, this occurs both in updates of the renderer (e.g. audio processing block size) and the head tracker. In (Laitinen, Pihlajamäki, et al., 2012) the effect of limited update rate varied between individuals, some were tolerant of update rates as low as 4 Hz, but others could detect differences in naturalness between the maximum rate of 18 Hz and the reference condition. The reference system in their experiment provided 100 Hz update rate, which equates roughly to the rendering update rate when a block size of 512 samples is used at a sampling rate of 48 kHz.

By contrast, latency is the time delay between an event, such as a head movement, and the corresponding response in the output of the binaural rendering system. The latency in a binaural system is created by several different component parts. Depending upon the system design these may include the head tracking device, the network connection or serial port connection to the binaural system, head tracker signal processing in software, audio output block size, and delay caused by the binaural signal processing. Total system latency (TSL) may not always be constant, it will often be distributed around a mean. Miller et al. (2003) presented a method for empirical measurement of binaural system latency.

In earlier studies the effects of system latency were investigated by observing effects on localisation accuracy or response time. Sandvad (1996) found that 96 ms of latency did show a significant degradation in azimuth localisation accuracy and response time over the lowest latency of 29 ms, while Wenzel (2001) found that localisation was not much degraded for latencies as high as 250 ms.

When the goal is plausible high-quality binaural rendering, detectability is the relevant issue and localisation does not seem to be a good predictor of this. Latency mainly increases response times in localisation tests. Wenzel (1997) suggested that this detection threshold can be determined from psychoacoustic data for minimum audible movement angles (MAMAs), which is the threshold of angular distance that must be covered for a moving source to be differentiated from a static source. MAMAs have been shown to vary with source velocity (Perrott and Musicant, 1977) and source bandwidth (Chandler and Grantham, 1992). Assuming that the same phenomenon occurs when the source is static and the head is moving, the MAMA can be divided by the respective head or source velocity to give the minimum detectable latency. Values from Perrott and Musicant (1977) suggest minimum system latency as low as 26.7 ms for a head velocity of 360 °/s, 45.6 ms for a velocity of 180 °/s. However, a recent study suggests that in fact auditory compensation for head rotation occurs. Brimijoin and Akeroyd (2014) found that the MAAs are smaller during self motion than during source motion (Brimijoin and Akeroyd, 2014), so estimates of latency detection thresholds from source MAMAs may not be accurate.

Detection thresholds of TSL have been investigated by Mackensen (2004), Brungart, Simpson, et al. (2005), and Yairi et al. (2007), finding minimum observed latency thresholds of 60 ms, 38 ms, and 85 ms respectively, using differing stimuli and test methods. Mackensen (2004) performed a test for detectable differences in system latency using a motorised dummy head system. The minimum possible system latency (50 ms) was compared with a range of other longer latencies (up to 150 ms) in a paired comparison. The method of determining the minimum system latency was not clear, it may have been calculated theoretically rather than measured. The listeners were asked if they could detect a difference between the stimuli and asked to identify the stimulus with greater latency. A range of signals were used, with castanets being the most critical. A difference was not perceptible for delays below 85 ms, however this only shows a detectable difference to the minimum latency condition, not an overall perception of latency in the system. It was claimed that the minimum latency of 50 ms was not perceptible, but it is not clear how this was determined. When long latencies were present “tracing effects” were observed where the source position lagged behind the head movement and then caught up after the head movement ceased. The authors decided that the maximum allowable system latency was 85 ms. Latency below this level reportedly resulted in “an aural sensation of *being there*”.

Lindau (2009) also investigated the minimum detectable system latency in dynamic binaural rendering. A three alternative forced choice design was used where the system latency was adaptively increased from the minimum system latency, measured at a mean of 43 ms, using a maximum likelihood adaptation rule. Pink noise and male speech stimuli were used, and anechoic and reverberant impulse responses were used resulting in four conditions. The minimum detected threshold was 53 ms and the mean of listeners’ thresholds was 107.63 ms with standard deviation of 30.39 ms. Latencies less than 64 ms were only detected 3 times out of 88, so this would seem to be a sensible strict target for TSL. No effect was observed for the stimulus or impulse response type conditions. As with Mackensen (2004) the discrimination is between minimum system latency and a delayed version, so it is not clear that the minimum system latency is below a just noticeable threshold compared to real listening or whether the relative difference between minimum and just noticeable system latency relates to an absolute latency detection threshold.

Brungart, Simpson, et al. (2005) had possibly the lowest baseline system latency of available studies, estimated as 11.7 ms with standard deviation of 1.5 ms. Whilst the minimum detected system latency was 60 ms, typical listeners could not detect latencies below 80 ms. However, when a second pilot source was included with the minimum possible latency, at the same position as the higher latency target, the latency detection thresholds were decreased by approximately 25 ms. This suggests that augmented reality (AR) systems have stricter latency requirements, because comparison can be made to real-world signals with no latency.

2.8.2.6 Summary

For head tracking in a high quality binaural rendering system, the system should allow tracking of natural head movements and give perceptual performance below detectable thresholds. The head rotation and position translation should be measured, giving 6 DoF, and this should be fixed to a calibrated frame of reference. Based on human localisation acuity in optimal conditions, rotation measurement should have a precision of less than 1°, and jitter in the measurement should be negligible at this order of magnitude. The system should be capable of measuring the full range of angular rotations and also position translations. The tracked volume is dependent on the activity, for seated applications it can be less than 1 m^3. The system latency should be less than 60 ms and the update rate should be in the order of 100 Hz. When the tracking system is also required for rendering of visual signals, the requirements for accuracy and latency become stricter (Welch and Foxlin, 2002).

2.8.3 Head Tracking Systems and Technologies

Head tracking technologies are reviewed in detail by Welch and Foxlin (2002) and Hess (2012) gives a review considering application to binaural systems specifically. There are many systems available for tracking head movements, with a variety of different approaches, meeting the requirements of the above factors to differing degrees. Hess (2012) presented a categorisation of available head tracking systems based on the configuration of measurement apparatus, it is summarised here:

- *Inside-out systems* - Use both transmitter(s) and receiver(s). Receivers are attached to and move with the listener. Transmitters are stationary and external to the listener.
  Example: an ultrasonic system with microphones on the listener's head and ultrasonic loudspeakers in front of the listener.
- *Outside-in systems* - Use both receiver(s) and transmitter(s). Transmitters are attached to the listener and affected by their movement. Receivers are stationary and external to the listener.
  Example: optical tracking with LED markers attached to the listener's head and camera(s) monitoring the movement of these markers.
- *Inside systems* - Use only a single device, attached to the listener.
  Example: accelerometer attached to the listener's headphones.
- *Outside systems* - Use only receiver(s), stationary and external to the listener.
  Example: Marker-free camera-based tracking.

Hess also reviewed many commercially available systems and published/patented tracking techniques. The different types of tracking technologies explored are listed here with some reference to systems used and similar additional systems:

- Ultrasonic systems – based on time-of-arrival differences of transmitted reference signals (inside-out: Logitech (2013) and outside-in: Beyerdynamic (2018))
- Sonic systems – within the frequency range of human hearing, based on time-of-arrival differences of environmental sound (inside-out: Lacouture-Parodi and Habets (2012) and Azizi and Munch (2008))
- Electro-magnetic systems – based on emitting and receiving a reference electromagnetic field (inside-out: Polhemus (2018), Ascension Technology Corporation (2013), and Munch et al. (2009))

- Optical analogue systems – based on infra-red LED transmitters and position sensitive receiver devices (inside-out: Hess and Mayer (2012) and Smyth Research (2018a))

- Optical digital systems – with camera(s) and LED marker(s), detecting infra-red/visible light points and using their relative positions (inside-out: Lee (2008), or outside-in: Natural Point (2018b) and FreeTrack (2013))

- Inertial systems – using combination of accelerometers, gyroscopes, and magnetometers (inside: Bartz (2012) and InterSense (2013))

- Inertial systems with optical reference – as above with added optical device for setting absolute reference direction (inside plus inside-out: Hoffman, Hess, et al. (2009))

- Single-camera-based systems – using a camera and image processing of the observed scene (inside: Simon, Fitzgibbon, et al. (2000) or outside: Seeingmachines (2018)). An extensive review of computer vision techniques for head pose estimation is available in (Murphy-Chutorian and Trivedi, 2009).

- Depth-camera-based systems – using a camera and a transmitted optical signal to resolve depth, then used to track 3D objects, often combined with an RGB camera (could be termed outside-out: Microsoft (2018a))

- Marker-based multiple-camera systems - optical digital systems that use multiple cameras and passive (reflective) or active (LED) markers to track objects, normally in 6 DoF (outside-in (active) and outside-out (passive): VICON (2013), Natural Point (2018a), and ART (2013))

The advantages and drawbacks of these systems are well summarised by Hess in relation to the requirements for dynamic binaural rendering. Some technologies cannot provide the accuracy and precision required for research work, some have stability issues. Latency is also a critical issue with many systems.

For use in a research environment, optical systems offer long-term stability which is desirable and camera-based systems are advantageous because they can deliver full 6 DoF tracking with reference to real-world coordinates, though low latency must be ensured. Electromagnetic systems also potentially offer accurate 6 DoF tracking but are susceptible to distortion when conductive objects are in the electromagnetic field of the device (LaScalza et al., 2003).

For consumer applications of binaural systems, e.g. in a mobile entertainment system, cost and size become more important. Inertial systems are appealing from this perspective but they suffer from drift and the lack of an absolute reference. Sensor fusion of a number of different components can be performed to obtain a reliable estimate with adequate filtering, such as in (Mahony et al., 2008). Most mobile phones now contain orientation and heading sensors, and these devices have become cheaper and of better quality in recent years. Some commercial headphone products have been released with integrated orientation tracking e.g. (3D Sound Labs, 2018). Camera-based inside and outside systems can offer full 6 DoF tracking with reference to real-world coordinates, but the main limitation with these systems is update rate and latency. A combination of inertial and optical sensors is common, with inertial sensors providing the required speed and optical sensors giving a stable frame of reference.

The recent popularity of virtual and augmented reality headsets has required advances in consumer tracking technology. Mobile VR headsets use high-quality gyroscope sensors, either in the headset (Samsung, 2018) or in an inserted mobile phone (Google, 2018a). In order to provide inside tracking[11] in 6 DoF, augmented reality headsets combine inertial sensors with cameras that track visual features in the environment (Aaron et al., 2017). The HTC Vive is a consumer VR system, which uses external laser emitters and photodiodes on the headset and handheld controllers, as well as inertial measurement, providing 6 DoF tracking of each object over an area of up to 4 m × 4 m. Niehorster et al. (2017) studied the suitability of the Vive for scientific research, finding suitably high precision and low latency (22 ms), yet it showed inaccuracy with reference to the physical ground plane, which changed significantly during momentary loss of tracking.

2.9 Auditory Virtual Environments

In most real-world environments, besides anechoic chambers, there will be acoustic reflections. Section 2.2 introduced the important role that environmental acoustics play in spatial hearing, particularly in terms of distance localisation. For binaural rendering to produce realistic AVEs, there must be a simulation of environmental acoustics. Novo (2005) gives a comprehensive overview of AVEs. This section presents only a short review.

2.9.1 Perceptual Effects of Environmental Acoustic Simulation

Including the effect of a reverberant space in the binaural impulse responses can improve the realism of generated sound sources, especially in terms of perceived distance and ex-

[11] *Inside* according to the definitions of Hess (2012), other studies also refer to such systems as *inside-out* e.g. (Welch and Foxlin, 2002).

ternalisation. Begault (1992) evaluated the effects of environmental reverb in binaural rendering, using a simple model-based approach, whereby early reflections were simulated by a ray-tracing technique and late reverberation was modelled using exponentially decaying noise to synthesise an artificial BRIR. This was then convolved with the dry source sound. Listening experiment results showed a benefit for the perceived distance and externalisation of auditory events. In a small study (5 listeners), Zahorik, Kistler, et al. (1994) found that synthetic room models did not change localisation of direction or distance compared with anechoic rendering, even when reflection directions were randomised. It was suggested that the precise reflection pattern is not important, but also that it might not be necessary to have environmental reflections at all. However, Begault, Wenzel, and Anderson (2001) found that model-based reverberation significantly increased the perceived source distance over anechoic rendering, enabling out-of-head localisation. Völk et al. (2008) also found that reverberation significantly enhanced perceived source distance when measured non-individual BRIRs were used, compared with anechoic recordings. The externalisation was better with human BRIRs than those measured on a dummy head (Neumann KU80).

In Begault, Wenzel, and Anderson (2001), it was found that synthetic early reflections up to 80 ms after the direct sound were sufficient, a full auralisation of the late reverberation (up to 2.2 s) was not required to achieve externalisation. Völk (2009) performed an investigation of externalisation with measured BRIRs, varying the impulse response length. It was found that reverberation after approximately 100 ms did not enhance the sense of source distance. It was also noted that the pattern of early reflections has an influence on externalisation. Begault, Wenzel, and Anderson and Völk both found that late reverberation decreased acuity of localisation in the vertical axis.

Zahorik (2000) found that when echoic BRIRs were used, there was no difference between individual and non-individual measurements in terms of distance perception. In both cases the distance judgements led the authors to conclude that sources were well externalised. Zahorik (2002a) used individual BRIRs measured at a range of distances (0.3 m–12.79 m) in a small auditorium. The listening test was conducted in a small booth and the listeners were blindfolded during measurements. Perceived distances tended to vary in correlation with the distance in the measurement. The source distance was never reported as zero, which denoted in-head-localisation. This was taken to indicate that sources were well externalised. In a subsequent test the intensity and direct-to-reverberant ratio were varied for BRIRs measured at 1.2 m to investigate their relative influence on distance judgements. The perceptual weighting of the two cues were more strongly dependent on source type and direction than the distance. Zahorik (2002b) described a subsequent experiment with the same six participants, in which they were all presented with non-individual BRIRs. There was no biasing of distance judgements by using non-individual measurements.

Zahorik (2009) investigated the perceptually relevant factors for differentiating reverberation, in the context of binaural rendering. Simulations were performed for 15 different small virtual rooms. Twelve were model-based, created with varying room size and surface absorption properties, whilst three were BRIRs measured in a real room, one individual and two non-individual. Three of the models were intended to correspond to the measured room, using the correct HRTF sets to match, whilst the other models all used individual HRTFs. A multi-dimensional scaling analysis examined relationships between acoustic properties of the BRIRs and perceptual ratings of similarity between the virtual rooms. Models of the measurement room were not identical in character to the measured BRIRs. The results were objectively similar in the frequency range 400 Hz–6000 Hz, but high and particularly low-frequency simulation was less accurate. There were perceptual differences from the multidimensional scaling analysis, but they were reasonably similar in relation to the range of room models evaluated. Differences caused by use of non-individual BRIRs were much smaller than differences between virtual rooms. T_{60} and interaural cross-correlation (IACC) were found to be the only two objective room acoustic metrics strongly correlated with listener's similarity ratings, which were often described as reverberation time and spaciousness in perceptual terms. Reverberation time was the primary perceived difference.

Timbre, as indicated by spectral centroid, was weakly correlated to the first perceptual dimension in (Zahorik, 2009). However, previous studies have shown that room acoustics can have a significant effect on timbre (Bech, 1995; Bech, 1996). A review of such aspects is given by Rubak and Johansen (2003). It is likely that environmental reverberation and reflections will have a significant impact on the perceived colouration in binaural rendering systems.

2.9.2 Data-based AVEs

Data-based approaches utilising convolution with measured BRIRs were proposed by (McKeeg and McGrath, 1997; Horbach et al., 1999). In such systems, head tracking controls real-time switching between measurements at different head orientations to create dynamic rendering of virtual loudspeaker sources. Partitioned convolution algorithms allow efficient processing of long impulse responses with low latency (Wefers, 2014). Lindau, Kosanke, et al. (2012) demonstrated that only the early part of the BRIRs needs to be dynamically switched, the late reverberation after the perceptual mixing time can be taken from a single BRIR measurement, which improves rendering efficiency. Dynamic data-based binaural rendering systems are sometimes termed *binaural room scanning* after Mackensen, Felderhof, et al. (1999).

These systems can be used to simulate loudspeaker reproduction systems in reverberant environments with high perceptual accuracy when using individual BRIR measurements (Brinkmann, Lindau, and Weinzierl, 2014), but also when using non-individual measurements (Lindau and Weinzierl, 2012). The techniques have been validated and used a number of times to simulate loudspeaker rendering systems in repeatable double-blind experiments (Olive and Welti, 2009; Gedemer and Welti, 2013; Wierstorf, Raake, Geier, et al., 2013).

A drawback with the data-based approach is that the source characteristics and spatial configuration are fixed in the measurement. It can no longer be assumed, as is often done with HRIRs, that rotation of the source is reciprocal to rotation of the listener's head. The environmental reflections reaching the listener's ears are dependent on the positioning of the listener and the sound source within the room, not just relative to each other, so a different set of impulse responses is needed for each source position to be rendered. This is also true for the listener morphology, which prevents later individualisation of the measurements. Data-based IAVEs therefore often require a high number of measurements, which can be impractical when considering more than the simplest forms of interaction.

A number of studies have explored methods to synthesise BRIRs from RIRs made without a head present, by combination with a set of HRIRs. Such techniques thus allow for individualisation without the listener having been present. Physical approaches have been taken, performing modal sound field analysis and binaural re-synthesis from very high-density microphone array measurements (Duraiswami et al., 2005; Melchior, Thiergart, et al., 2009; Schultz and Spors, 2013; Bernschütz, 2016). A recent study has shown that such techniques can give equivalent timbre and spaciousness to BRIR measurements (Ahrens, Helmholz, et al., 2018). Perceptual approaches using lower resolution microphone systems have also been taken, for example from first-order ambisonics ("B-format") IRs (Menzer, Faller, and Lissek, 2011; Zaunschirm, Frank, et al., 2018) and also from single omnidirectional IRs (Pörschmann, Stade, et al., 2017).

Techniques have also been investigated for modifying measured BRIRs to change the impression of distance and the congruence with the reproduction environment (Catic et al., 2013; Albrecht and Lokki, 2013; Werner and Liebetrau, 2013). Parametric control of measured spatial impulse responses has also been proposed, based on an analysis-modification-resynthesis approach (Melchior, Sladeczek, and de Vries, 2008; Carpentier, Szpruch, et al., 2013; Coleman, Franck, Jackson, et al., 2017).

2.9.3 Model-based AVEs

The major advantage of model-based AVEs is their flexibility. The characteristics of the source, room, and listener can all be freely defined in such a simulated environment, for example by selecting the HRTF set for the listener, the geometry of rooms and absorption characteristics of materials, and the directivity patterns of the sound sources. Environments can be simulated without accessing them physically for measurement, provided they can be adequately modelled, and environments that do not exist can also be simulated. Dynamic variation of properties is also more feasible, including free movement of sources and the listener in 6 DoF. This makes them better suited to interactive creative applications such as computer games, VR and AR, as well as auralisation for acoustic design.

Sound propagation modelling techniques can be divided into wave-based and ray-based (geometric) approaches (Välimäki, Parker, et al., 2012). Wave-based techniques, such as BEM and FDTD methods, can accurately model acoustic propagation at all frequencies, including diffraction and interference effects, yet they are more computationally intensive and challenging in real-time systems. Ray-based techniques are efficient and can handle dynamic and directional source and listener behaviour in large environments, but are only accurate for higher frequencies and cannot model wave effects like diffraction and interference.

The idea of model-based reverberation was introduced by Schroeder (1970). Since this time, many environmental acoustic modelling applications have targeted physical accuracy, and are capable of running at real-time rates. Commercial systems exist for acoustic design, such as ODEON and CATT. Offline simulation can be performed with these tools to generate BRIRs for data-based rendering.

The first IAVE systems used the image-source method to simulate a small number of early reflections in real-time (Foster et al., 1991). This was based on the image-source method of Allen and Berkley (1979), which was later extended to work for arbitrary geometry (Borish, 1984). Heinz (1993) introduced the use of a statistical model of the diffuse reverberant tail, alongside specular early reflections generated with the image-source method. This approach was applied by Begault (1992), but the BRIRs were generated offline.

Since the late reverberant decay of a room response has Gaussian properties, it can be modelled using decaying noise signals and processed with convolution (Begault, 1992). However, efficient signal processing techniques can be used to model the late reverberation. Jot and Chaigne (1991) presented a method for reverberation design based on feedback delay networks (FDNs), later control of inter-aural coherence was added (Jot, Larcher, et al., 1995). Jot, Cerveau, et al. (1997) demonstrated analysis of measured room reverberation for resynthesis with such FDNs and Menzer and Faller (2009) extended this to allow frequency-dependent inter-aural coherence to be matched to a reference BRIR measurement. Vilkamo, Neugebauer, et al. (2010) presented a sparse QMF-transform-domain reverberator that can efficiently model and synthesise the reverberant tail of a measured BRIR. This can be readily applied in MPEG spatial audio decoding and rendering (Herre, Purnhagen, et al., 2012).

Savioja et al. (1999) developed a real-time IAVE system for simulating enclosed environments, with rendering of direct sound and early reflections as discrete point sources. Reflections were calculated from room geometry using the image-source method and a FDN was used for late reverberant decay (Jot, 1992). Jot, Larcher, et al. (1995) took a simplified approach for rendering early reflections, recognising that the precise spatial and temporal patterns of early reflections are not as perceptually significant as the direct sound component. Instead, a cluster of reflections is efficiently modelled with high-level control of their distribution. This system also presents high-level perceptual control of the room model, through a parametric mapping layer. It was patented by Jot, Jullien, et al. (1998) and was described in detail by Jot (1999). Developments from this system are available as a low-cost package of externals for the Max/MSP environment (Carpentier, Noisternig, et al., 2015).

A number of other model-based AVE systems have been proposed. Pellegrini (2002) presented a model-based AVE with perceptual control of source distance, room size and reverberation time, which controlled lower-level model parameters. Silzle, Novo, et al. (2004) and Musil et al. (2005) presented comprehensive tools for designing IAVEs, still based on relatively simple image-source early reflection models and a separate reverberation decay using a FDN. The image-source method becomes too complex for high-order reflections to run in real-time, so is often used to simulate reflections up to only the first few orders (e.g. 2 or 3). Lentz et al. (2007) presented a more sophisticated IAVE system, utilising stochastic ray tracing techniques to allow simulation of higher-order specular reflections, and also importantly diffuse reflections. This avoided the a priori assumption of diffuse field properties made by using a statistical late reverberation model, so allowing simulation also of outdoor environments. The system also allowed scene partitioning with simulation of multiple coupled rooms.

The disadvantage of model-based approaches is that perceptual accuracy is often not achieved (Bork, 2000), particularly because modelling is sensitive to estimates of surface absorption properties and the lack of diffraction modelling causes low frequency errors. A recent evaluation of state-of-the-art offline simulation algorithms has shown perceptual characteristics comparable to real measurements only for some approaches (Katz, Poirier-Quinot, et al., 2018). Wave-based solutions performed worse than ray-based approaches, it appears numerical artefacts can accumulate in the simulation of large spaces.

Modern model-based IAVE systems are capable of simulating complex large open environments with dynamic sources and listener in real-time, with increased realism. Ray-based methods are now also capable of incorporating models of diffraction and diffuse reflections (Schröder and Vorländer, 2011; Schissler, Mehra, et al., 2014). Wave-based approaches can also be used (Mehra, Rungta, et al., 2015), pre-computation techniques are combined with graphics processing unit (GPU)-based run-time rendering.

Härmä, Jakka, et al. (2004) discussed the need to match parametric room models to the external environment in the context of augmented reality, and considered the effects of defining the parameters with offline or online estimation. With recent developments in consumer AR systems, such approaches have seen renewed activity. An interesting recent system utilised an iterative solver to estimate surface absorption properties in a ray-based model by comparison to a single measured RIR (Schissler, Loftin, et al., 2018). Geometry was also estimated from camera images. Jot and Lee (2016) and Murgai et al. (2017) discuss methods of estimating the character of the diffuse reverberation, which is important for perceptual plausibility.

When adapting parameters of a relatively simple model to match a measured BRIR data, Wendt et al. (2014) showed that similar perceptual characteristics can be obtained. However, such an approach does not have the flexibility of a full environmental acoustic model for interactive applications.

2.9.4 Summary

Adding environmental acoustic simulation to binaural rendering systems clearly has perceptual benefits. More accurate distance localisation is afforded and in-head localisation is largely avoided. Sense of distance and externalisation also seem possible without individual measurements when environmental reflections are provided.

Whilst the flexibility offered by model-based systems is attractive, a model will inevitably introduce approximations when compared to a real environment. Commercial solutions do now exist for GPU-based ray-tracing approaches (NVIDIA, 2017) and for wave-based approaches with pre-computation services running in the cloud (Microsoft, 2018b), open-source alternatives are also available (Poirier-Quinot, Katz, and Noisternig, 2017). Even if such systems might be capable of achieving a high degree of perceptual accuracy when simulating real spaces, this would require careful construction of a detailed model.

Data-based approaches are less flexible, but real environments can be simulated with a high degree of accuracy for a specific configuration of source and listener. There is more compelling evidence of the ability of data-based systems to create auditory events with a plausible spatial impression than for model-based systems. Commercial and open-source systems exist for performing dynamic data-based binaural rendering (Smyth et al., 2008; Ahrens, Geier, et al., 2008). Data-based binaural rendering is also more straightforward to implement than a model-based algorithm.

The objective of an AVE should also be considered when choosing the design approach, it is dependent on the intended application. This is discussed by Novo (2005), where three different goals are given:

- *Authentic* approach – authentic reproduction of an existing, real environment where the same percepts as the real environment are evoked
- *Plausible* approach – evoking of auditory events that a listener perceives as having occurred in a real environment
- *Creational* approach – evoking of auditory events where no authenticity or plausibility constraints are imposed

It appears that data-based approaches are currently better suited to applications targeting authenticity, at the cost of reduced flexibility, whilst model-based approaches are more suited to creational approaches. Methods for evaluating AVEs against such criteria will be reviewed in chapter 3.

Techniques continue to develop and the gaps between approaches are narrowing, enabling more accurate model-based approaches and more flexible data-based approaches. Hybrid methods are also used, either simulating early reflections and using measured late reverberation (Pellegrini, 1999) or using modelled late reverberation along with measured early reflections (Vilkamo, Neugebauer, et al., 2010).

2.10 Further Considerations

This chapter cannot give a complete review of binaural technology. It is a large field with many different aspects. However, this section presents a few additional topics which are of interest in the design of practical binaural rendering systems.

2.10.1 Efficient Filter Models

Alternate representations of HRTF filters have been discussed, including infinite impulse response (IIR) filter designs. These are often approximations of the original FIR HRTF filters, that offer greater efficiency for low-resource rendering scenarios. Minimum-phase

FIR approaches with significant frequency-domain smoothing have also been considered. Huopaniemi et al. (1999) reviewed such approaches and evaluated them with an auditory model, with the aim of finding perceptually accurate simplifications. It was found that frequency-domain smoothing could be performed with an auditory model without negative effects on localisation performance, but only azimuth direction was explored. Kulkarni and Colburn (1998) showed that HRTFs could be significantly smoothed in the linear frequency domain without causing audible differences, from 512 FIR coefficients down to 32, though again using only positions in the horizontal plane. Hobden and Tew (2015) used an auditory localisation model to show that further spectral resolution can be discarded whilst retaining vertical localisation acuity when using auditory bandwidth resolution. Andreopoulou and Katz (2018) evaluated such processing techniques subjectively for non-individual HRTF sets using pre-defined source trajectories and a rigorous process of pre-screening assessors based on HRTF rating ability. A truncated 64-point minimum-phase FIR representation (as in Kulkarni and Colburn (1998)) and a 12th-order IIR filter representation were compared to full-phase HRTFs. Both techniques caused some significant rating differences for the various non-individual HRTF sets evaluated. The correspondence of the minimum-phase FIR filters was closer to the full-phase HRTFs than the IIR filter, but not with significant differences. Both horizontal plane and median plane source rotations were affected in similar manner by such processing. The perceived quality of different HRTF sets was greatly affected by such processing i.e. a highly rated HRTF set without processing was then rated poorly after processing or vice versa. Such techniques should therefore be used with great care.

2.10.2 Reducing Colouration or Enhancing Localisation

Timbral colouration of the source is a common effect of binaural rendering, since HRTFs are generally far from flat in frequency response, which can be problematic in some applications e.g. music reproduction. Merimaa (2009, 2010) investigated reducing timbral colouration in non-individual and individual HRTFs by compressing the RMS spectral sum of the left and right magnitude responses, which effectively flattens the transfer functions, whilst preserving ITD and frequency-dependent ILD. This was effective in reducing timbral colouration, at the expense of increased polar angle errors. Silzle (2002) investigated a manual tuning procedure by experts to reduce colouration of HRTFs for use in a non-individualised rendering application. By contrast, Brungart and Romigh (2009) emphasised spectral differences across polar angles ϕ_{cc} with given lateral angle θ_{cc}. This gave improved localisation in ϕ_{cc} without negatively affecting θ_{cc} for both non-individual and individual HRTFs, though for individual measurements the benefits only occurred for relatively small modifications. It was suggested that this method might overcome some limitations of the measurement and equalisation procedure for individual measurements and for generic measurements it might enhance some cues that are common across listeners.

2.10.3 Near-Field Source Modelling

Section 2.2.9 introduced cues influencing distance localisation, highlighting that in the near-field (<1 m) frequency-dependent increases in ILDs are utilised. Brungart (1999b) also demonstrated that, in the near-field, the direction of incidence seen at each ear is quite different to the commonly used angle relative to the centre of the head. Romblom and Cook (2008) presented a method for modelling near-field effects during point-source rendering, without requiring near-field HRTF measurements. It was proposed to apply a filter to far-field HRTFs to modify the magnitude differences due to near-field effects and also to apply correct geometry look-up for each ear. Kan et al. (2009) applied a similar approach with a filter to simulate distance variation, applying appropriate frequency-dependent ILD changes, based on a spherical head model. A listening experiment showed improved distance estimation over a basic intensity-variation distance model, for distances of up to 60 cm whilst preserving directional localisation accuracy. Intensity is still the dominant cue however, when this was removed from the filter model distance estimation was poor. Jot, Walsh, et al. (2006) propose a simpler model that uses a frequency-independent ILD for near-field sources.

2.10.4 Extended Source Modelling

Binaural rendering often assumes that sound sources are simple point sources i.e. omnidirectional sound sources of infinitesimal size. In reality sound sources have complex directivity and have finite size, for example a violin or a waterfall. In more sophisticated AVE these simple source models may be inadequate. Menzies (2010) discussed soundfield expansion techniques for modelling complex source directivity using spherical harmonic representations. Similarly Schissler, Nicholls, et al. (2016) utilise spherical harmonic expansions of the soundfield to represent extended sound sources, as efficient projection of the shape onto a sphere around the listener, represented in spherical harmonics. With HRTF data also represented in the same basis, simple rendering can then be obtained. Results show that this gives more efficient and more accurate rendering than sampling the scene as a series of point sources. A drawback with this method however is that a single-channel audio signal is used in the rendering. A real-world volumetric sound source will likely involve decorrelated sound from different regions, particularly with a distributed stochastic source like a waterfall.

Extended sources can be rendered by spatially rendering multiple decorrelated versions of an input source, using techniques such as phase randomisation (Kendall, 1995; Jot, Walsh, et al., 2006), different frequency-dependent delays (Zotter and Frank, 2013), or frequency-dependent spatialisation (Pihlajamäki et al., 2014; Su et al., 2017). However, these techniques are susceptible to artefacts such as colouration and time-domain smearing (Franck, Fazi, et al., 2015).

2.11 Summary

Binaural technology is a large field of research and development activity, covering many disciplines, including: psychoacoustics and auditory sciences, cognitive psychology, audio engineering, digital signal processing, numerical acoustics and machine learning. The technology is based on our understanding of the processes of spatial hearing, particularly binaural hearing.

Binaural recording enables capture and reproduction of natural sound scenes with remarkable simplicity, but lacks the flexibility required for entertainment media production and interactive applications. Binaural rendering simulates the naturally-occurring psychoacoustic cues that are used in spatial hearing through digital signal processing. To make binaural rendering practicable in real-time (often interactive) systems, a number of signal processing techniques have been developed, including the measurement, equalisation, interpolation and online update of head-related transfer functions (HRTFs). Simulation of environmental acoustics is often used in binaural rendering systems and both data-based and model-based approaches exist. Also tracking technology is commonly applied to create interactive systems with dynamic adaptation to listener head movements. It has been shown that, when carefully implemented and calibrated in a controlled environment, binaural rendering can produce auditory events that are indistinguishable from those produced by a real loudspeaker. Binaural technology can be said to create authentic auditory virtual reality.

Development and evaluation of these techniques has furthered our understanding of spatial hearing processes themselves. This then enables perceptually-motivated engineering decisions to be taken in binaural rendering system design. Research in this field has been active for at least 30 years and continues to be popular. There have also been efforts to develop commercial applications of binaural technology, including in entertainment media. For mass-market applications, acoustic measurement of individual HRTFs and headphone-to-ear transfer functions (HpTFs) is not feasible. Many studies have explored practical methods for HRTF individualisation and headphone-to-ear correction filter (HpCF) design that are more suitable for uncontrolled environments. Individualisation is a popular topic of research and whilst advancements are being made, it appears that effective practical methods are not yet readily available.

Individual calibration can improve localisation acuity, but non-individualised binaural rendering can be quite effective already. Authentic auditory virtual reality may not be possible, due to substantial variations in HRTFs and HpTFs between individuals. However, when appropriate environmental acoustic cues and dynamic updates to listener head movement are also provided, it has been demonstrated that binaural rendering can produce auditory events that are in agreement with listeners' expectations of a corresponding real sound event. This can be described as plausible auditory virtual reality, meaning that the resulting spatial impression is convincing enough for the listener to accept that the sound sources could be real.

An important question to consider is how accurate does binaural rendering need to be? The majority of existing studies on binaural technology have evaluated perceptual effects solely in terms of localisation acuity (in terms of direction and/or distance). Demonstration of equivalent localisation performance to natural hearing scenarios provides some validation of the founding idea behind binaural technology, which was introduced at the start of this chapter. More so the small number of experiments that show auditory impressions are indistinguishable from reality. But as stated in the introduction, for many applications this criterion is overly strict. In entertainment media, a creational approach to auditory virtual environments is more common. Lack of creative control is a major reason why binaural recording did not reach more widespread use in the 1980s. Flexibility may be more important than realism in this context. If we have a system which can precisely render realistic sound sources in only a single virtual environment, then can we only make binaural programmes set in that specific environment?

The primary aim of this thesis is to improve the quality of headphone listening experiences for entertainment media audiences through use of binaural technology. The implicit assumption is that spatial enhancement given by binaural technology leads to an overall improvement in the quality of the listening experience. It is not yet clear what degree of realism is required to provide such an improvement. Binaural rendering has perceptual effects other than the spatial impression. A clearer understanding of what forms the quality of a listening experience is needed. This is the subject of chapter 3.

Chapter 3

Evaluating Perceived Quality

This chapter reviews the concepts of and methods for evaluating perceived quality. Particular attention is given to sound quality and previous studies of the quality of binaural technology. This chapter informs the methods used for evaluation of binaural technology in later chapters.

3.1 Introduction

A core aim of this project is to improve the quality of audience experience when listening to sound on headphones. Achieving this requires proper understanding of the concepts of quality, the factors that can influence it and methods for evaluating it, particularly in relation to the field of application. This chapter reviews literature on quality and quality evaluation of media systems, sound reproduction and binaural technology.

3.2 What is Quality?

The definition of quality in audio and acoustics research has in the past not been entirely clear. Blauert and Jekosch (2003) state that sound quality is “a mental construct which is often insufficiently defined and, consequently, not understood properly by many”. It is therefore important to establish a definition in the context of this study, and one that is in agreement with other researchers.

In engineering, the term *quality of service (QoS)* has often been used to refer to a measure in the physical domain of the characteristics of a system[1]. QoS has been referred to as *produced quality* and, in contrast, the term *perceived quality* refers to a measure in the perceptual domain, involving an individual making a judgement of an experience (Jumisko-Pyykkö, 2011). It is clear that perceived quality is paramount; technology is developed to serve human users. This is very apparent for media systems and human-computer-interaction. Over the years, the understanding and definition of quality in the perceptual domain has improved.

Martens and Martens (2001) outlined some common definitions of quality. Two of which are "the inherent characteristics of an entity", properties which are described objectively, and "the degree of excellence of an entity", which is a more subjective definition. These definitions are somewhat at odds to one another. ISO 9000, 2005 defines quality as:

> The ability of a set of inherent characteristics of a product, system or process to fulfil requirements of customers and other interested parties.

This can be seen as bringing together the first two definitions in that it relates the character of the entity to requirements in the perceptual domain, which define the character of excellence.

Blauert and Jekosch (1997) follow this approach in discussing a theoretical framework of evaluation of product-sound quality. The process of quality evaluation first requires determination of the character of the sound. The character of the sound is obtained by auditory perception, as well as input from other sensory modalities. This perception is influenced by the cognition, action, and emotion of the listener in their current state. Quality is then determined as distance between the character of the sound and a reference, the distance between perceived character and desired character. The reference is internal, based on the listener's expectations (the character of excellence). A quality judgement is formed from the distance of many characteristics of the product to those of the reference, this multivariate process is "non-linear, time-variant and loaded with a huge amount of memory" (Blauert and Jekosch, 1997). Blauert and Jekosch (1997) provide a formal definition:

> *Product-sound quality* is a descriptor of the adequacy of the sound attached to a product. It results from judgements upon the totality of auditory characteristics of the said sound - the judgements being performed with reference to the set of those desired features of the product which are apparent to the users in their actual cognitive, actional and emotional situation.

Lorho (2010) follows this idea of quality, as the distance from a desired reference.

> *Quality*: A measure of the distance between the character of an entity under study and the character of a target associated with this entity.

[1] Quality of system is similarly used as a measure in the physical domain of the characteristics of a system.

Lorho explains that the reference can be internal (the character desired by the listener) or external (presented explicitly as another stimulus). This reference may be known prior to evaluation, or it may need to be established during the process of evaluation. If external, the definition of this reference is critical in the evaluation process.

ITU-R Recommendation BS.1116 (ITU-R, 2015e) contains a similar definition of quality for use in perceptual assessment of audio systems:

> *Basic audio quality*: This single, global attribute is used to judge any and all detected differences between the reference and the object.

However both of these definitions do not refer to the role of the user / listener in the evaluation process, which is clearly important. The process of determining the distance of differences between the object/entity and the reference is not explicit. Blauert and Jekosch refer to the judgement by the individual. Typically quality evaluation studies look at a representative group of listeners not the judgement of a single individual, however every response is subjective and influenced by the expectation and knowledge, and emotional and cognitive state of the individual.

The European Network on Quality of Experience in Multimedia Systems and Services (Qualinet – COST Action IC 1003) published a White Paper reviewing definitions of quality in multimedia communications engineering and provided an updated framework of definitions (Le Callet et al., 2012). Here the aspects of judgement by the individual are further addressed:

> *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. By 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".

Quality is described as a degree of excellence as experienced and evaluated by an individual during a specific event. Also the term "adequacy" in the definition of Blauert and Jekosch (1997) is reflected in the concept of need fulfilment.

The text book by Möller and Raake (2014) was developed from the outcomes of the Qualinet initiative, with revised definitions and extensive discussion of applications of the concepts.

> *Quality (based on experiencing)* results from the "judgment of the perceived composition of an entity with respect to its desired composition".

This quotes the definition of quality by Jekosch (2005), but makes it explicitly based on experiencing, which is defined as follows.

> *Experiencing* is the individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference.

Möller and Raake distinguish *quality based on experiencing* from "*assumed quality*", which is perhaps a more accurate term for the typical QoS approach, previously described as *produced quality* by Jumisko-Pyykkö (2011).

> *Assumed quality* corresponds to the quality and quality features that users, developers, manufacturers or service providers assume regarding a system, service or product that they intend to be using, or will be producing, without however grounding these assumptions on an explicit assessment of *quality based on experiencing*.

The above definition of quality based on experiencing refers to an *entity*. Regarding experiences involving media technology, an entity could be an application, service or system for example. This gives a technology-oriented focus to quality judgements. It is often considered more appropriate to consider the *experience* rather than the entity, giving a more person-centric focus. The term experience refers to the perception of one or more events, at a specific moment in time or over a specific time period (Raake and Egger, 2014). In development of communications and multimedia systems, the term *quality of experience (QoE)* is often used to reflect this different focus. The main objective of Qualinet was to give a clear definition of QoE (Le Callet et al., 2012). This definition was updated by Raake and Egger (2014) and is presented here with minor modification[2].

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

This definition incorporates the influence of the system but acknowledges that there are many other influences, often specific to the individual. It also incorporates affect, in contrast to earlier definitions of quality, by focussing on the degree of delight or annoyance of the person.

[2] The gendered possessive pronouns "his or her" have been replaced by "their".

QoE is now a widely adopted term. Besides the above definition, it refers to a framework for understanding the formation of QoE judgements, including the many features of QoE and the factors influencing QoE. These will be presented in the following sections, with a focus on aspects relevant to applications of binaural technology.

This thesis will focus on *perceived quality* of binaural technology, rather than assumed quality based on technical characteristics, in line with the stated aims of this work (see section 1.2). Where the word *quality* is used without prefix, perceived quality is implied. For clarity, this author defines perceived quality as follows.

> *Perceived quality* results from the judgment of the perceived composition of an entity with respect to its desired composition in relation to a particular experience or set of experiences.

This is based on the definitions of *quality* by Jekosch (2005) and of *quality (based on experiencing)* by Möller and Raake (2014). However experiencing is here restricted to a particular situation of reference, as occurs during quality evaluation experiments. As Möller and Raake (2014) state “a snapshot is taken, resulting in the exchange of experiencing by experience”.

The conceptual framework of QoE is used to guide this research and the design of the studies herein. However evaluation of QoE is not the only relevant approach during development of technology. QoE is an affective judgement of perceived quality that is made within an applied context by representative people; engineers must acknowledge its importance. Yet, as discussed in subsequent sections of this chapter, evaluation of perceived quality may also focus on more objective evaluation of the sensory-domain character of an entity. This author is cautious about using the term QoE simply because it is a “buzzword” (Schatz and Reichl, 2011) and attempts only to use it where the above quoted definition of Möller and Raake (2014) is fitting.

3.2.1 Processes of Quality Formation

It is now widely accepted that quality is formed by the judgement of the perceived character of an entity with respect to its desired character. It is important to understand the mechanisms of this process, so that appropriate methods of measuring the perceived quality of a system or service can be used. This will allow appropriate design and validation of media technology.

From the perspective of the person experiencing events, quality formation involves the processes of perception and those of forming and retrieving perceptual references. It also involves processes of comparison and judgement based on the characteristics of the experience and the reference. These are parallel and interactive processes. A conceptual model of the formation of quality judgements is given by Raake and Egger (2014), developed from Jekosch (2005) and Le Callet et al. (2012). This model is illustrated in figure 3.1 and components are discussed in the following sections.

When considering quality judgements, it is common to define separately the set of *influence factors*, which affect perceived quality, and the set of *quality features*, which describe the experienced and desired characteristics considered during quality judgement (Le Callet et al., 2012). These will be specific to the nature of the experience and its evaluation. It is hoped that considering this conceptual framework and specific aspects relevant to binaural technology will facilitate the design of appropriate perceived quality evaluation experiments.

3.2.1.1 Perception and Experiencing

Perception begins with the incidence of stimuli at the sensory organs, e.g. the ears. Physical representations of the stimuli are converted to neural representations, with electrical signals that are characteristic of the stimuli. This information is conveyed to the relevant brain region(s) through neural transmission for further processing e.g. the auditory cortex. Throughout this transmission, the representations of stimuli are transformed into increasingly abstract symbolic representations. For hearing this occurs through the auditory pathway, which is reviewed briefly in the following section.

In general, sensory processing in the peripheral nervous system leads to a multidimensional neural representation of stimuli, covering aspects of time, space, frequency and activity. From this representation, *perceptual event formation* occurs in specific sensory modalities, e.g. auditory events. In higher-level brain regions there is some early-segmentation of stimulus features, where neural patterns likely to belong to the same event are associated. This leads to symbolic representations for event and object. Such Gestalt-analysis in the processes of auditory scene understanding is described by Bregman (1999). Event formation is influenced by memory and feedback-based adaptation of sensory processing. There is also cross-modal sensory integration e.g. influence of visual activity on auditory event formation.

Hypotheses are formed based on internal knowledge and rules, i.e. top-down processes are influenced by the cortical region of the brain. These are verified against the bottom-up perceptual evidence. This process is labelled *anticipation and matching* in figure 3.1. Experiencing, a stream of perception processes occurring due to a particular stimulus, leads to recognised perceptual objects with a *perceived character*. *Exploratory action* may result from certain stimuli, e.g. the turn-to reflex due to a sudden impulsive sound, or exploratory head movements made during localisation, altering the sensory and subsequent neural informa-

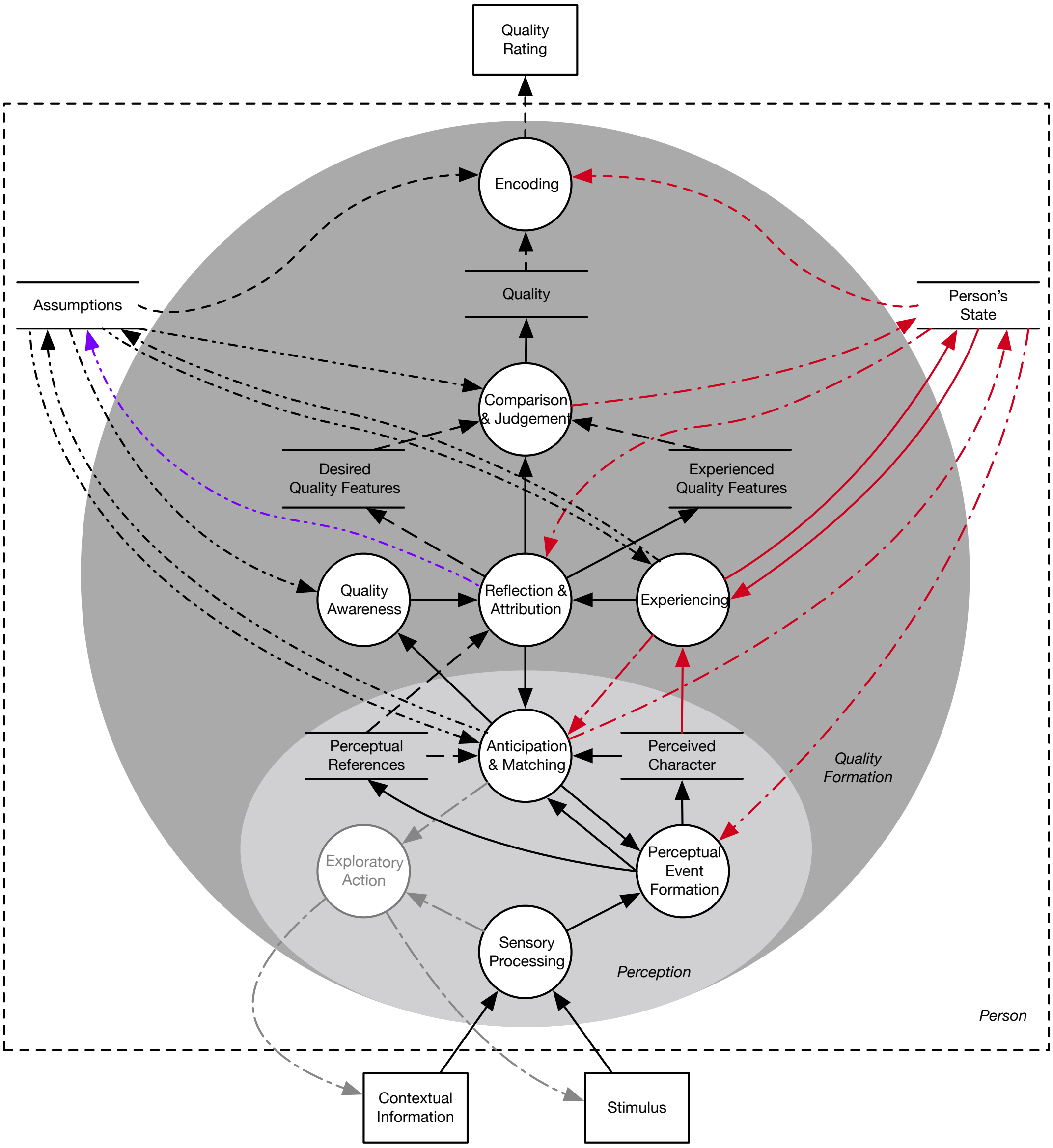


Figure 3.1: Quality formation process, after Raake and Egger (2014)

tion input.

In addition to the sensory stimulus of focus, contextual and task-related information are subject to sensory and subsequent cognitive processing. This might occur through other sensory modalities. This information may directly affect perception or do so through stored higher-level concepts termed *assumptions* in figure 3.1. The *person's (current) state* also determines perception. There will be situational and temporal changes to a person's feelings, thought-processes and behaviours.

3.2.1.2 The Auditory System

The physiology of the auditory system is reviewed in detail by Moore (2003) and more recently by Kohlrausch et al. (2014), only a brief overview is given here. Figure 3.2A shows the anatomy of the human ear.

Airborne sound vibrations reach the eardrum, also known as the tympanic membrane, via the ear canal. The three ossicles (incus, malleus and stapes) transmit vibrations from the tympanic membrane to the oval window at the base of the spiral-shaped cochlea. The basilar membrane within the cochlea has mechanical properties that mean high-frequencies will cause it to vibrate most near the base of the cochlea and low-frequencies will cause it to vibrate most near the apex of the cochlea. Each point along the basilar membrane has a characteristic frequency at which it is most sensitive and a sinusoidal tone will only excite a rather narrow part of the membrane. The cochlea is lined with approximately 30 000 motion-sensitive cells, known as the inner hair cells. The local frequency-dependent displacements of the basilar membrane stimulate nerve firings from the inner hair cells. The information is carried from these receptors via the auditory nerve fibres to the cochlear nucleus and on through the midbrain to the auditory cortex. The auditory system is tonotopic i.e. it operates using frequency-domain representations, with similar frequencies being processed in topologically-neighbouring regions. This tonotopic representation is maintained from the cochlea right through to the primary auditory cortex, as shown in figure 3.2.

The auditory pathway between the cochlea and the auditory cortex is illustrated in figure 3.3. In the cochlear nucleus, neurons with different response types are connected to the auditory nerve fibres. These provide neural representations of various signal features from incoming auditory nerve activity, e.g. representations of the temporal envelope of the signal. The superior olivary complex receives input from the ipsilateral and contralateral cochlear nuclei. Interaural cues are processed by neurons in the superior olivary complex, interaural phase differences (IPDs) in the medial superior olive and interaural intensity differences

[3] Image from Chittka and Brockmann (2005), under Creative Commons Attribution 2.5 Generic licence.

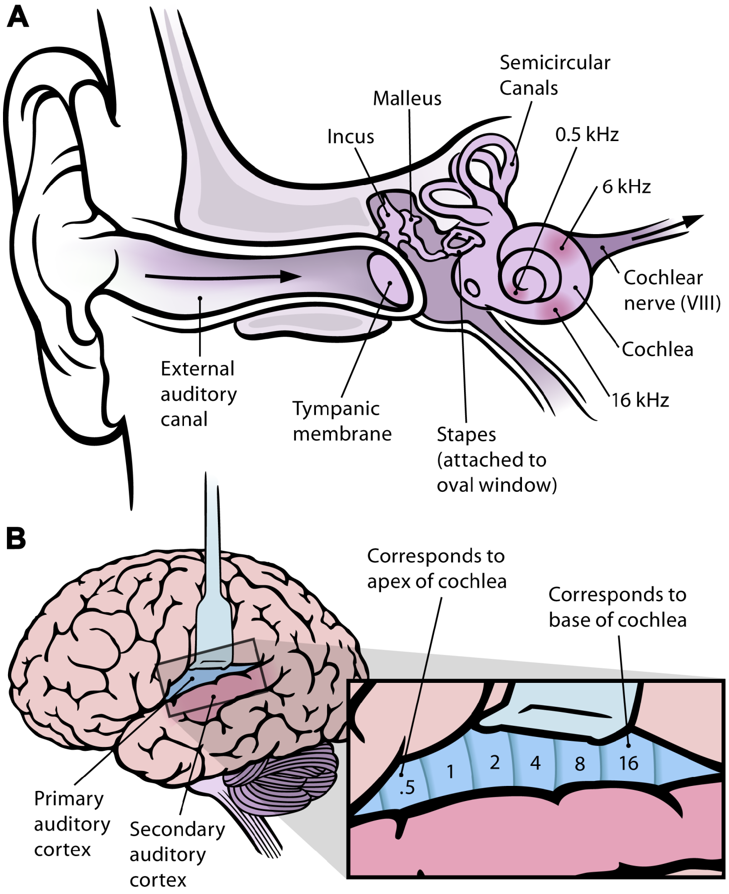


Figure 3.2: Frequency-based topological (tonotopic) representation in the auditory system, from the cochlea in the inner ear to the primary auditory cortex (Chittka and Brockmann, 2005)[3].

(IIDs) in the lateral superior olive.

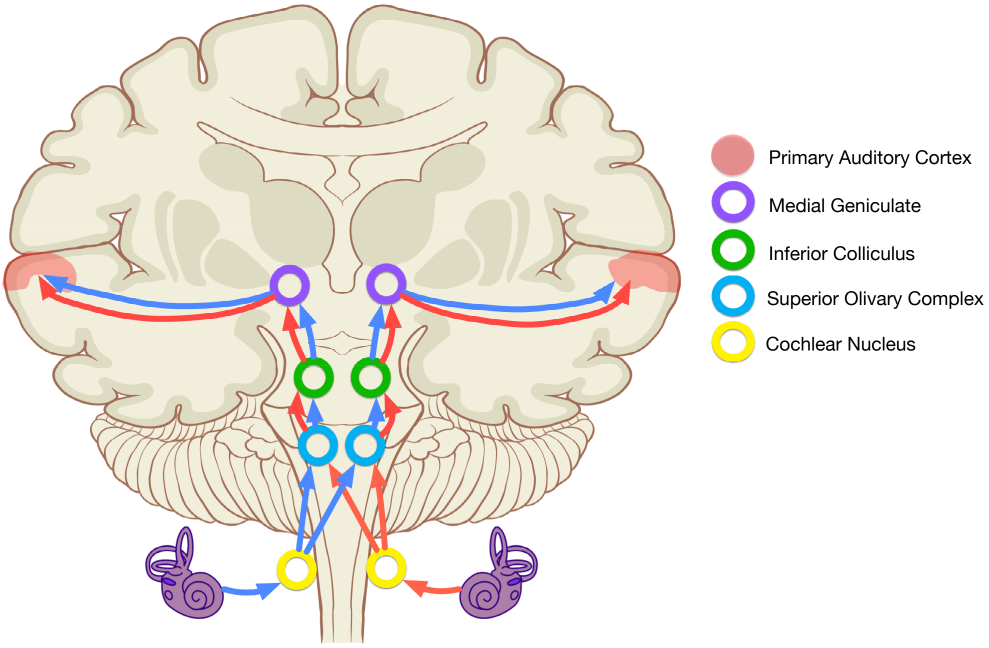


Figure 3.3: The auditory pathway from the cochlea through the brainstem and midbrain to the auditory cortex[4].

Whilst the functional mechanisms of the auditory periphery are well understood and agreed models exist (Meddis et al., 2013), consensus models for higher-level auditory processing are not available. The inferior colliculus, in the midbrain region, combines information from the cochlear nuclei and the superior olives. It is thought to play a role in preprocessing for sound event detection and localisation, e.g. (Skottun et al., 2001). It also exhibits multi-sensory integration, with a link to the superior colliculus responsible for visual processing. The medial geniculate body, in the thalamus, connects the auditory cortex and the inferior colliculus. It exhibits significant top-down processing, acting somewhat like a multiplexer or relay, to filter and combine lower-level inputs and so may direct auditory attention.

The auditory cortex in the temporal lobe of the brain processes auditory information received through the auditory pathway. It is clear that this region exhibits significant plasticity, and there is evidence of neurons that are sensitive to abstract properties of complex

[4] Human brain coronal section image by Patrick J. Lynch, medical illustrator; C. Carl Jaffe, MD, cardiologist. Cochlea image from Chittka and Brockmann (2005). Both are under the Creative Commons Attribution 2.5 Generic licence.

sounds, effectively providing object recognition (Chechik and Nelken, 2012). There are many descending projections in the auditory pathway, meaning that higher level regions of the pathway can influence processing at lower levels, which allows enhancement or suppression of certain sounds and their features (a kind of dynamic adaptive filtering). This descending pathway goes down as far as the cochlea, where the outer hair cells can be made to modify the vibratory response of the basilar membrane.

3.2.1.3 Memory and Perceptual References

In figure 3.1, stores of information are represented by components with two parallel horizontal lines. Different levels of memory can be identified, such as sensory memory, working memory and long-term memory, storing representations of sensory stimuli on timescales ranging from 150 ms, to tens of seconds, to a lifetime. Perceptual references can be present at different levels of memory. Long-term memory is used for event and object identification, whilst working memory allows perceptual integration of the scene. Sensory memory is a peripheral process that stores stimulus representations for short durations (up to 2 s), for access by higher-level processing stages.

Cowan (1984) discussed evidence of auditory storage working on short timescales (up to 300 ms) and used for object recognition, as well as in working memory of approximately 10 s–20 s. Working memory of timbre has been shown to be weak and subject to interference from other similar stimuli (Starr, 1997). Further discussions about the boundaries between sensory and categorical long-term memory are given by Winkler and Cowan (2005). Where sounds have a close link to an individual's real-world context, longer term memory is achieved, such as memory for the timbre of a friend's voice.

The perceived character of an event or flow of events can be stored in working memory, or even long-term memory e.g. due to verbal re-coding of speech signals. The learning of perceptual (or conceptual) references is associated with expertise. Learning is integral to the processes of perception. More acute perceptual performance is achieved through learning, with more detailed and relevant references available in long-term memory. For example, a skilled musician can associate auditory percepts, such as pitch, with associated actions e.g. to improve intonation in performance. As with perception of stimuli, references are also influenced by the personality and current state of the person, as well as contextual factors and they can change over time.

3.2.1.4 Quality Judgement

The process of quality formation is seen to operate in parallel to that of experiencing, involving higher-level brain function. Experiencing leads to formation of the perceived character of a stimulus, and formation and retrieval of perceptual references. In evaluating quality of experience, a process of *reflection and attribution* occurs. This might occur during or after the experience. The triggering of this process is represented by *quality awareness* in figure 3.1, which is said by Raake and Egger (2014) to operate "like a cognitive gate" that focusses attention on evaluation of quality. It could be triggered by an external task, such as in a quality evaluation experiment, or due to unexpected events, where the experience deviates from assumptions. Reflection leads to attribution of quality features to the experience, as well as the respective desired features, based on the perceived character and perceptual references (see section 3.2.3). The formation of quality results from *comparison and judgement* of these two sets of quality features, and finally when reporting the perceived quality externally an *encoding* process is required to generate a descriptive or quantitative response.

Regarding media technology systems, aspects of the system design will influence the experience, determining the perceived character of the stimulus. During reflection and attribution, the causes for this perceived character will be reflected on and may be attributed to the system. The resulting quality judgement may therefore relate to quality of the system, based on the experience, as well as the quality of the experience itself.

Assumptions about the context of the experience will influence reflection and attribution. This incorporates abstract conceptual expectations and attitudes of the person. In formal quality evaluation experiments, this will incorporate the defined task for the assessor. In this thesis the focus is often on the perceived quality of a system, based on experiences with that system, rather than the QoE. As indicated in figure 3.1, the *person's state* plays a key role in influencing quality formation processes.

3.2.2 Factors Influencing Perceived Quality

Reiter et al. (2014) discuss the range of factors that may influence QoE in the context of media applications, this follows earlier discussion in Le Callet et al. (2012) and also by Jumisko-Pyykkö (2011). These influence factors (IFs) can be broadly categorised into human, system and context categories. As illustrated in figure 3.4 the factors are often interrelated and overlapping. This is an important notion in the development of media technology and the work of this thesis. Specifically, the way that factors of the system influence QoE is likely dependent upon the human and context factors.

The three categories of IF are defined in (Le Callet et al., 2012):

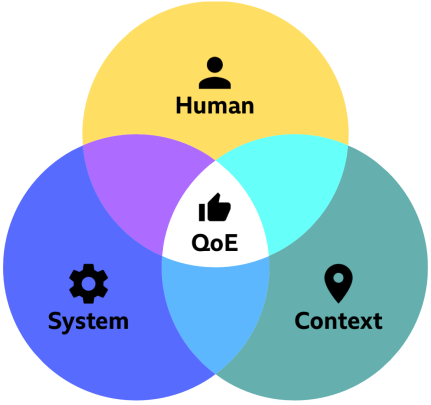


Figure 3.4: Factors influencing the quality of experience (QoE) can be grouped into system, context, and human. These often overlap and have a mutual effect on the QoE, after Reiter et al. (2014).

- A *human IF* is 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.
- *System IFs* refer to properties and characteristics that determine the technically produced quality of an application or service.
- *Context IFs* are factors that embrace any situational property to describe the user's environment in terms of physical, temporal, social, economic, task, and technical characteristics.

Human IFs can be categorised into those that influence lower- and higher-level processes. The physical and mental condition of the individual, and their emotional state can affect perception at the low-level early sensory stages. This can be dispositional such as age, sensorial acuity, gender, etc. or more dynamic such as mood, attention and emotions. Interpretation and judgement by the listener are involved at higher-level cognitive stages, the influencing factors at this level can also be variant (expectations, needs, knowledge, experience, emotions) or relatively stable (socio-economic situation, educational background, attitudes, values). In quality evaluation experiments the experience of assessors often has to be carefully considered, as it could bias judgements or lead to responses that lack detailed information. It is common to distinguish between expert and inexperienced assessors in the requirements of an experimental design (ITU-R, 2014b). More detailed consideration of human IFs is less common in experimental design. Though for example Olive, Welti, and Mcmullin (2014) investigated the impact of listener experience, age, and cultural background on preferences for different headphone frequency responses.

System IFs determine the technically produced quality of an application or service. In a multimedia streaming or broadcasting context they could related to any stage in the production chain; including capture, coding, transmission, storage, rendering, and reproduction. The System IFs are subdivided into content-, media-, network-, and device-related factors in Le Callet et al. (2012) and Reiter et al. (2014), but the examples are focussed around video streaming systems. Content-related factors will include signal capture and production techniques, and the signal content itself. Media-related factors refer to settings such as picture resolution and sample rate or coded bitrate. Network-related factors refer to issues with transmission, such as available bandwidth, delay, signal loss, error rate etc. Device-related factors refer to user-facing systems or devices in the delivery chain, for instance display screen resolution or headphone characteristics. There are many potential system IFs in each of these categories, they should be considered for each specific application domain.

In terms of binaural technology systems, the choice of the system components described in chapter 2 will influence quality. These factors cover content, media and device categories, dependent on the intended application. At the content level, audio signal characteristics will be important influences on perceived quality, including spectral and temporal structure and dynamic range. With spatial audio systems, the spatial scene complexity will also be a factor. Detailed analysis of audio and video production methods could be applied at the content level.

Context IFs can relate to the physical environment, including spatial location, movement and activity, transition between places, the functional role of place, and sensed environmental attributes. Temporal aspects include time (of day/week/year) and content duration and frequency of use of a system. Social context can influence quality of experience, any inter-personal relationships during an experience should be considered. The costs of the service/system and other such associated aspects, such as branding of the system, are economic context IFs. The task context will also have an influence, this is determined by the nature of the experience. It may affect the focus of attention, for example if the experience is part of a multitasking situation or there are interruptions. Technical and information context refers to the relationship between a system under test and other related systems, for example considering network connectivity. Context IFs are related to ecological validity; the context of a quality evaluation task should be appropriate to ensure that the results reflect that real-world target situation (Rumsey, 2002).

Modelling context IFs can allow system design or service delivery with multiple configurations, to achieve appropriate quality given the constraints of the present context. Lately there have been significant trends towards context-aware systems which adapt in an automated manner to the user and their environment. Examples include dynamic adaptive streaming of audiovisual media based on network conditions, adjustment of display brightness based on lightning conditions, and personalised recommendation or targeted advertising based on aggregated personal and social data. In the audio domain, Walton, Evans, Kirk, et al. (2016b) investigated the influence of environmental noise conditions on listeners' preferred mix between foreground and background layers of audio content. Adaptation to context IFs is a core concept in *object-based media* systems (Armstrong et al., 2014; Evans, Ferne, et al., 2017), and to a more limited extent in object-based audio as represented in *next-generation audio* systems, see section 6.3.

When carrying out quality evaluation experiments, IFs should be carefully considered. Where appropriate they should be controlled to ensure that the results have ecological and external validity, avoiding the introduction of unwanted bias or random error. System IFs will be the primary focus in experiments presented in this thesis, as is often the case in the development of media technology. Awareness of the human and context IFs is also very important in proper experimental design and the interpretation of results.

3.2.3 Features of Perceived Quality

Möller, Wältermann, et al. (2014) discuss how an understanding of the features[5] of perceived quality or QoE can be important when trying to understand why a specific experience has a certain level of quality. The most commonly used definition comes from Jekosch (2005):

> *Quality feature (QF)*: A perceivable, recognized and namable characteristic of the individual's experience of a service which contributes to its quality.

A feature can be seen as a dimension of a multi-dimensional perceptual space defining an experience. The features that can be perceived and named will be context-specific. A feature is considered a quality feature (QF) when it contributes to quality. The reference in quality formation influences whether a perceptual characteristic is deemed to contribute to quality or not. QFs will not necessarily be independent of one another. A number of methods exist for identifying and analysing QFs, these will be reviewed in 3.3.3. Möller, Wältermann, et al. (2014) discuss the relation of QFs to perceived quality or QoE, considering *vector-model* features, with a monotonic contribution to quality, or *ideal-point* features, which have an ideal value (that of the reference). Respective examples are signal noise (the more, the

[5] Characteristics, attributes and descriptors are commonly used as synonyms for features with respect to perceived quality.

worse) and loudness (not too loud, not too quiet).[6] More complex models of the contribution of QFs to quality exist as will be discussed in section 3.3.

Möller, Wältermann, et al. (2014) define five levels of QFs related to experiences involving media technology applications, services or systems:

- *Direct perception*, features that are created as an immediate response to sensory information, without much abstraction.
- *Action*, relates to the human perception of their own actions, including possibly immersion or presence.
- *Interaction*, refers to human-to-human and human-to-computer interactions, including features such as responsiveness and naturalness of interaction.
- *Usage situation*, refers to features specific to a single usage of a product or service, such as accessibility and stability.
- *Service*, refers to features that extend over many usages, such as usability, aesthetic feeling, and long-term stability.

QFs are specific to the particular type of system or service and can be defined in more detail for a given domain.

3.2.4 The Formation of Sound Quality

Blauert and Jekosch (2012) present a detailed discussion of the formation of quality judgements in the domain of sound quality. They give a layered model of quality features and judgement processes based on the level of conceptual abstraction. Taking a perceptionist perspective[7], the layers in increasing level of abstraction are: auditive quality, aural-scene quality, acoustic quality, and aural communication quality.

Auditive quality refers to individual auditory events, which occur at a specific time and location in space, with specific characteristics such as loudness, pitch, timbre, roughness, position, and spatial extent. This is the domain of classical psychoacoustics experiments, listeners are required to analyse a specific characteristic of the sound by focusing their attention. Judgements at this level require little cognitive interpretation.

[6] Some QFs can be multi-dimensional. This could be a grouping of several other QFs e.g. spatial impression, or it may not be possible to separate and name the dimensions.

[7] Approaching the layers the perspective of scientific realism, acoustic quality would be the lowest level of abstraction, as in Blauert and Jekosch (1997).

Aural-scene quality refers to perhaps a more natural, less analytical, listening state. Auditory events are combined into auditory entities or objects, which together form an auditory scene. The formation of auditory objects and their structuring into scenes is a perceptual process requiring much greater abstraction than perception of auditory events. Some aspects of aural-scene perception are thought to be automatic and follow common rules, such as the *Gestalt* rules (Bregman, 1999). Cues from other sensory modes will already have an influence at this level. Processes such as the precedence effect, auditory-stream separation, and room constancy occur at this level of abstraction. This is also said to be the primary domain of the tonmeister in sound recording, ensuring that the relevant characteristics of the aural scene are well captured and reproduced. Factors like balance of timbre, clear identification and localisation of sound sources, and spatial impression are likely to be taken into account. The decisions made by the tonmeister in the recording process will take into account the content. The content will also have an influence on audience members at this level, Blauert and Jekosch describe “quality building features” such as immersion, sense of presence and perceptual plausibility operating at the level of aural-scene quality. In contrast to auditive quality, aural-scene quality is seen to be highly context-dependent. Blauert and Jekosch state that it is aural-scene quality which is of primary concern when evaluating the quality of systems such as audio codecs, loudspeakers and spatial audio systems.

Acoustic quality is placed as the next layer since associating auditory events with physical measurements requires a high degree of abstraction. Blauert and Jekosch (2012) highlight the problem of validity when associating acoustic measures with perceptual constructs. It is important to consider the extent to which any physical measurements actually relate to the characteristics and quality of auditory activity of interest. However acoustic signals (or electric signals representing acoustic ones) can be useful in quality evaluation, especially when they represent perception well. The example of measuring distortion in an electrical audio transmission path is given by the authors. In relation to spatial audio, acoustic quality aspects may involve inferring the physical characteristics of the environment from the experience. This environment may be real, reproduced or simulated.

Aural-communication quality is at the highest level of abstraction, it is the assignment of meaning to auditory events. Blauert and Jekosch (2012) explain that “the perceptual organization of our world at large is based on the formation of signs, their conception, interpretation, and the subsequent assignment of meaning”. This follows the ideas of semiotics. An example is given in the context of product sound design: an auditory event provides a sign and the concept of the associated product is the cognitive referent, against which the interpretation of the auditory event is performed by the listener and meaning is assigned. Assignment of meaning is influenced by prior experience and learning, so is time-varying.

This layer of quality is very important, since the majority of listeners' responses to auditory stimuli are based on what the auditory events mean to them in their context, rather than a direct reaction to the characteristics or structure of auditory events.

Aural-communication quality relates to media experiences in terms of the understanding and impact of the creative ideas. Such considerations are discussed by Raake and Egger (2014): "Artists or content producers create entities (carriers, signs) that can be experienced, and thereby may attempt to deliberately provoke or achieve specific experiencing." This is the primary domain of sound designers, who aim to communicate information through sound. It is noted by Raake and Egger that technology is involved at various stages in this process of conveying intended meaning from the creator to the audience. However the experiences and their meaning are distinct: "in terms of semiosis, 'meaning' is associated with the creator's intentions ('sender'), while at the 'receiving' end, 'meaning' results from interpreting the content during experiencing" (Raake and Egger, 2014). User experience studies are highly relevant in the domain of aural-communication quality (Wechsung and De Moor, 2014).

Besides the layer of acoustic quality, some alignment can be seen with the levels of QFs given by Möller, Wältermann, et al. (2014). The level of direct perception relates clearly to auditive quality. Aural-scene quality appears to cross at least the levels of direct perception and action. Aural-communication quality relates more to interaction, usage situation and service QFs.

A core purpose of the model of Blauert and Jekosch (2012) is to provide a framework for design of sound quality evaluation experiments. The aim is to ensure that researchers ask the relevant questions for their given problem, and therefore use appropriate evaluation methods to answer these questions. Since the focus in this thesis is on the design of binaural technology systems and the impact that has on perceived quality, evaluation at the levels of auditive and aural-scene quality is most relevant here. Evaluation methods are discussed in more detail in section 3.3.

3.2.5 Sound Quality Features

A number of studies have attempted to define and categorise features of sound quality, some with respect to specific application domains and research questions, and some general purpose. Most often features are at the auditive and aural-scene levels and are descriptive in nature. They are frequently grouped into families of timbral and spatial characteristics, though technical artefacts are often included as a separate group. Sometimes researchers have defined attributes and scales based on their own experience and what are seen as commonly accepted concepts, particularly in earlier work. Formal experimental methods for defining QFs with a panel of assessors have also been used. These are discussed in section 3.3.3.

3.2.5.1 Early Attempts to Model Sound Quality

In an early attempt to identify the features of sound quality, Gabrielsson (1979) performed a set of experiments using loudspeakers, headphones, and hearing aids. Eight dimensions of quality were identified from these experiments: *Clearness/Distinctness*, *Sharpness/Hardness-Softness*, *Brightness-Darkness*, *Fullness-Thinness*, *Feeling of space*, *Nearness*, *Disturbing sounds*, and *Loudness*. These are primarily focussed on timbral aspects.

Toole (1985) applied these perceptual dimensions as attribute rating scales in the evaluation of loudspeaker reproduction. Additional attributes were introduced to better describe spatial features of stereo reproduction, based on comments by assessors in preliminary studies. There were: *definition of the sound images*, *continuity of the sound stage*, *width of the sound stage*, *impression of distance or depth*, *abnormal effects*, and *perspective*. In addition to spatial characteristics, *abnormal effects* captures technical artefacts of reproduction. The attributes were separately grouped into spatial quality and sound quality categories, with additional overall ratings of pleasantness and fidelity. The sound quality category related to timbral attributes. Pleasantness relates to emotion and preference, whilst fidelity represents "the extent to which the reproduced sound resembles an ideal", in line with the definition of quality by Jekosch (2005).

Letowski (1989) presented an early attempt to provide a hierarchical model of the features of sound quality. The model was called the Multi-Level Auditory Assessment Language or MURAL, it is shown in figure 3.5. At the highest level sound quality was split into two families of features, timbre and spaciousness. It featured some related concepts to previous studies, as well as some new ones. This model is not specific to any particular application domain. It was intended as a practical tool for researchers considering sound quality.

3.2.5.2 Timbral Quality of Sound

Whilst there have been a variety of specific timbral attributes identified in early and subsequent studies. It is worth noting some key definitions. Timbre is defined by the American Standards Association (1951):

> "*Timbre* is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar."

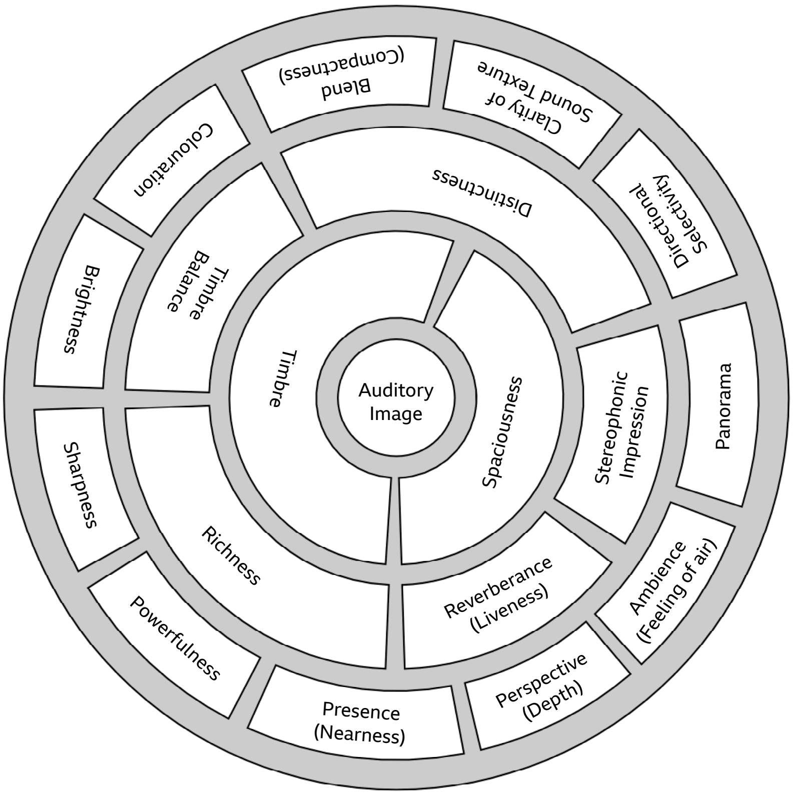


Figure 3.5: The Multi-Level Auditory Assessment Language model, redrawn from Letowski (1989)

Salomons (1995) later provided a definition of the colour of an auditory stimulus, incorporating also pitch and rhythmic character:

> "The *color* of a sound signal is that attribute of cochlear sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness are dissimilar; it thus comprises the timbre, rhythm sensation as well as the pitch of the signal."

Colouration is then defined as a differential measure:

> "The *coloration* of a signal is the audible distortion, which alters the (natural) colour of the sound."

An important aspect of this definition is that colouration is best judged by direct comparison of stimuli. Whilst timbre is defined as a subset of sound colour here, often colouration is categorised as a timbral attribute, see e.g. figure 3.5.

3.2.5.3 Features of Spatial Quality of Sound

The ITU-R standards for perceptual evaluation in a broadcast context present a limited set of rating scales for describing spatial attributes. Besides *basic audio quality*, an overall measure of quality taking into account all aspects, the following attributes are given in Recommendation ITU-R BS.1116 (ITU-R, 2015e).

- *Stereophonic image quality* – This attribute is related to differences between the reference and the object in terms of sound image locations and sensations of depth and reality of the audio event.
- *Front image quality* – This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition.
- *Impression of surround quality* – This attribute is related to spatial impression, ambience, or special directional surround effects.

The first is intended for use with two-channel stereo reproduction, whilst the latter two are intended for use with five-channel surround. These are broad multi-dimensional features which may be susceptible to differences in interpretation. Rumsey (1999) reported that assessors were not consistent in their use of *impression of surround quality* and found it difficult to understand, despite training.

Berg and Rumsey investigated the attributes of spatial quality of sound in-depth through a series of publications (Berg and Rumsey, 1999, 2000b,a, 2001, 2003, 2006). In (Berg and

Rumsey, 2003), three top-level categories of sound quality features are discussed: timbral, spatial, technical. It is the family of spatial attributes of sound quality that are explored in detail in their work, with a specific focus on reproduction of multichannel recordings using a five-channel surround sound loudspeaker array (ITU-R, 2012a).

A structured process was used for elicitation of attributes from individual assessors, based on the differences between stimuli under evaluation. Assessors also rated stimuli using the attributes. A refinement method was used to obtain a set of obtain common attributes. Part of this process involved verbal protocol analysis, where features were categorised as either descriptive or attitudinal.

Common descriptive attributes were extracted from the data, with use of clustering techniques based on the ratings given by assessors. The terms *localisation*, *width*, *envelopment*, *distance/depth*, and *room perception* were used to describe these common groups (Berg and Rumsey, 2006), but these terms were not given accompanying expanded definitions to aid in their interpretation. It was noted that localisation referred to ease of localising an auditory event rather than the position of the event, a percept that Blauert (1997) called "locatedness" and is elsewhere often termed *localisability*. The attitudinal features were also split into subgroups, those referring to *naturalness* (or conversely artificiality) and sense of *presence*, and those that were emotional/evaluative, which were summarised as *preference*.

Koivuniemi and Zacharov (2001) developed a set of 12 attributes for describing various spatial audio systems through a consensus-based descriptive analysis method (see section 3.3.3.1). The attributes covered a set of spatial and timbral features, with spatial features: *sense of direction*, *sense of depth*, *sense of space*, *sense of movement*, *penetration*, *distance to events*, *broadness*, *naturalness*, and timbral features: *richness*, *hardness*, *emphasis*, and *tone colour*. Each attribute was given a definition and positive and negative scale endpoints.

These studies have the advantage that the attributes were defined by the assessors through experience of the stimuli, rather than by the experimenters. This increases the likelihood that the assessors will be able to make effective use of the attributes in rating tasks. Considering such matters, Rumsey (2002) notes that: "Spatial attributes should be identified that are meaningful, in order of priority; 1) to individual subjects; 2) to a well-defined group of expert subjects forming a listening panel, and that agree upon a set of attributes to be graded; 3) to expert listeners not associated with that listening panel; 4) to independent observers or readers of the results." The last two points would seem to benefit from standardised terminology across studies, if this could be achieved without jeopardising the former two points.

Berg and Rumsey (2006) compared the results of their study to attributes resulting from previous studies and noted that similar concepts were found in each case. The authors state that this “suggests the existence of a more generalizable set, even though the stimulus sets were different”. Similarly Koivuniemi and Zacharov (2001) noted that “Whilst a different methodology, sample set and reproduction configuration was employed, it was encouraging to note that many of the attributes developed in this study are similar to those evolved in the Berg and Rumsey studies.” Particularly noting naturalness, width and localisation as corresponding well.

Whilst commonality exists between studies, the entities of evaluation were not always explicit, which can leave some ambiguity in interpretation. Rumsey (2002) presented a scene-based paradigm for descriptive spatial quality attributes. Features are described according to their role within the scene. Target entities can be individual objects or ensembles of objects, there are also features describing the environment, as well as for the scene as a whole. Spatial features, similar to those in Berg and Rumsey (2006), are described in detail at different scene levels. The concepts of distance and depth are disambiguated and applied to various entities, whilst still acknowledging their relation. Envelopment is also linked to a related but distinct concept of presence, which here is defined as the “sense of being inside an (enclosed) space”. This scene-based paradigm is a valuable step in clarifying the definition of spatial attributes.

3.2.5.4 Quality Features of Binaural Sound

Several studies have defined a set of quality features for experiences of binaural technology systems. These often relate to specific application contexts or system factors.

In (Lorho, 2005a), a set of 16 attributes was developed by an expert panel, describing the characteristics of spatial enhancement systems for headphone reproduction of stereo music. Three main categories of attributes were identified, features relating to localisation, space, and timbre (which included artefacts). Localisation features referred to the spatial characteristics of individual auditory objects, whereas features of space referred to environment- or scene-level characteristics. The attributes, their scale end-points, and definitions are listed below.

“Attributes relating to localisation:

- *Sense of distance* (not definable / well definable): This attribute describes how well the distance between the sound source(s) and the listener can be defined.
- *Sense of direction* (not definable / well definable): This attribute describes how well the direction of the sound source(s) can be defined.

- *Sense of movement* (not definable / well definable): This attribute describes how well the movement of the sound source(s) can be defined.
- *Ratio of localizability* (none / all): localizability describes how well the direction and the distance of a sound source(s) can be defined. The attribute ratio of localizability describes how many sound events can be localized from those present in the audio sample.

Attributes relating to space:

- *Quality of echo* (unpleasant / pleasant): This attribute describes how well the echoes relate to their sound source(s) in a qualitative way.
- *Amount of echo* (no echo / adequate echo / excessive echo): This attribute describes how the listener experiences the amount of echo in relation to the sound sources.
- *Sense of space* (not definable / well definable): This attribute describes how well the space represented in the audio sample can be defined.
- *Balance of space* (out of balance / in balance): This attribute relates to the space represented by the audio sample in relation to the listener's inner reference. A negative value means that the space is weighted in some direction. If no space is perceived, the space is out of balance.
- *Broadness* (inside head / broad): This attribute describes the perceived extent of the soundscape relative to the listener's head.

Attributes relating to timbre:

- *Separability* (none / all): This attribute describes how well the sound events can be separated out in the audio sample.
- *Tone Color* (lower sounds emphasized / higher sounds emphasized): This attribute describes the spectral content of the audio sample.
- *Richness* (flat / neutral / rich): This attribute describes how rich and nuanced the audio sample is overall, and relates to a combination of harmonics and dynamics perceived in the sample.
- *Distortion* (distorted / not distorted): This attribute describes the possible metallic, machine-like, electrical-like artifacts in the audio sample.
- *Disruption* (disrupted / not disrupted): This attribute describes how much hiss, snap/crackle/pop is perceived in the audio sample.
- *Clarity* (muffled / clear): This attribute describes if the sound sample appears clear of muffled, for example if the sound source is perceived as covered by something.

- *Balance of Sounds* (out of balanced / well balanced): This attribute describes the possible difference in loudness between the sound sources present in the audio sample. The sound sample is well balanced if it contains only one sound source." (Lorho, 2005a)

These attributes cover an extensive range of characteristics, with some similarities to previous studies. Again there are some concepts relating to specific entities within the scene, ranging from individual objects to the overall scene level. Some overlap appears to exist between attributes, for example *ratio of localizability* appears to comprise the *sense of direction* and *sense of distance* attributes. *Broadness* appears to relate to the common concept of *externalisation*, where one end of the scale is denoted by "inside head". The term broadness, to this author, appears more related to the width of sources or scene. The definition given in (Lorho, 2005a) appears similar to *depth* as used in other studies e.g. (Lindau, Erbes, et al., 2014). The attributes were defined in Finnish and translated to English for publication. Language translation is often noted as problematic in relation to quality features, it may introduce bias and invite different interpretation. This is discussed for example by Berg and Rumsey (2006).

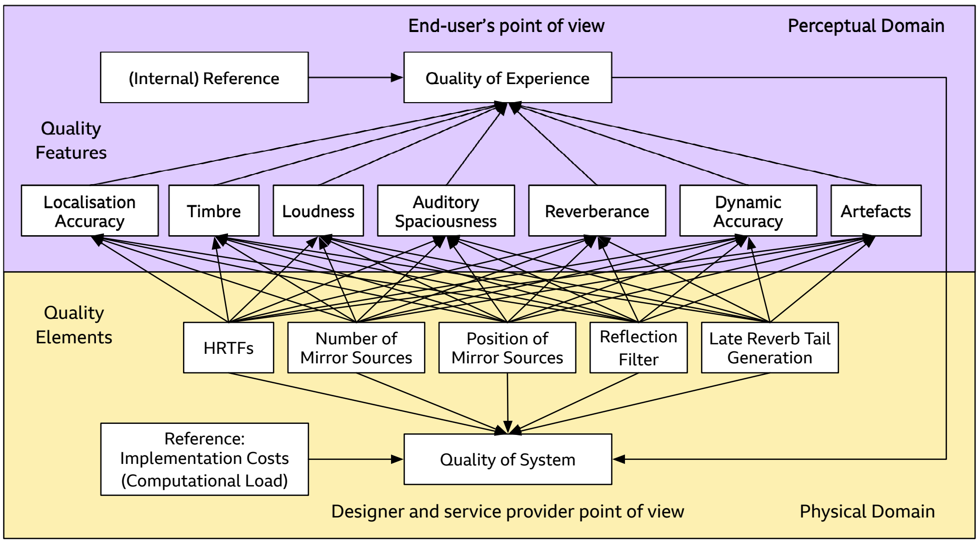


Figure 3.6: A quality taxonomy for interactive auditory virtual environments, redrawn from Silzle (2007)

In Silzle (2007) a list of relevant quality features was created by the author to characterise the quality of auditory virtual environments. This was presented as part of a taxonomy of the quality of interactive auditory virtual environments (IAVEs), which is shown in figure 3.6.

This work was based on developments in the understanding of QoE, and concepts of QFs and quality elements (IFs) defined by Jekosch (2004). It was intended as a tool to understand the quality requirements of IAVEs, as generated by systems such as IKA-SIM (Silzle, Novo, et al., 2004), particularly with respect to system constraints and trade-offs between different system factors, and different application contexts. The attributes were defined as follows:

- "*Localisation accuracy* – Discrepancy between the desired and the perceived location of the auditory event in the 3-D space
- *Timbre* – Plausible timbre, no unwanted sound colouration
- *Loudness balance* – Plausible balance of loudness between different sound sources, plausible loudness balance between direct sound part, early reflections, and reverberance
- *Auditory spaciousness* – Plausible auditory spaciousness or width of the auditory event
- *Reverberance* – Plausible reverberance (room perception)
- *Dynamic accuracy* – A sound source is moving smoothly and without perceived time delay regarding the interactive input movement
- *Audibility of Artefacts* – Prevention of audibility of artefacts (errors) which are unwanted auditory events like clicks"

These attributes were selected in order to be familiar to the participants in a survey, without listening to examples. The list was reviewed by domain-experts during the survey and the attribute *intelligibility* was reportedly added to the list, specifically considering a virtual chat room application. The results of the survey are discussed further in section 3.3.3.9. This list of attributes does not appear exhaustive compared to the outcomes of (Lorho, 2005a). Many of the concepts are related, though dynamic accuracy is an additional consideration for interactive systems. Also localisation accuracy in this instance relates to the location of auditory events, whilst the features from (Lorho, 2005a) relate instead to how well-defined the location is i.e. localizability or locatedness. The attributes defined by Silzle (2007) refer to *plausibility* several times. As mentioned in section 2.9, Novo (2005) discusses the different design goals of auditory virtual environments (AVEs), one of which is plausibility. This refers to "evoking auditory events which the user perceives as having occurred in a real environment".

In more recent work, the spatial audio quality inventory (SAQI) was developed through focus groups with domain-experts (Lindau, Erbes, et al., 2014). The SAQI provides an extensive set of 48 sound quality descriptors for evaluation of the perceptual character of virtual acoustic environments. These were categorised by timbre, tonalness, geometry, room, time behaviour, dynamics, and artefacts. There is also a general category that includes higher-level QFs such as *naturalness*, *presence*, *speech intelligibility* and *degree of liking*. The attribute definitions and scales are differential in nature, intended for paired comparison of a stimulus to a reference (internal or external) or between a pair of stimuli. Additionally, structured concepts of *assessment entities* and *modifications* are presented. The entities reflect the scene-based paradigm of Rumsey (2002), whilst modifications allow specification of time-varying behaviours, including those due to auditory scene events and user interactions.

Lindau, Erbes, et al. (2014) also presented an experimental methodology for applying the SAQI. This method has previously been used to give a sensory profile of individual and non-individual dynamic data-based binaural systems. Lindau, Brinkmann, and Weinzierl (2014) performed a comparison of the quality of a dynamic binaural rendering system compared to a real loudspeaker in a room, using both individual and non-individual binaural room impulse response (BRIR) measurements. Sound quality was evaluated in terms of many different attributes. Besides directional localisation errors, non-individual rendering showed degradations in source localisability, distance estimation, and externalisation, as well as more colouration. The experiential aspects naturalness and liking were also negatively affected.

Simon, Zacharov, et al. (2016) performed a descriptive analysis (DA) process to define perceptual attributes specifically for comparison of non-individual head-related transfer functions (HRTFs). Stimuli were developed considering ecological validity. Three virtual scenes were created by different sound engineers, each with static sound sources. These were

- a documentary with speech and background activity sounds in a kitchen environment, mixed in five-channel horizontal-surround format,
- an electronic music track with five sources at varied azimuth and elevation angles,
- a 13-track audio fiction with sound effects and ambience at varied azimuth and elevation angles.

Seven different HRTF sets were used to render these scenes. These were the perceptually-optimised reduced set identified in (Katz and Parseihian, 2012), which should give one very good match for a typical listener. Interaural time differences (ITDs) were replaced by a common average set.

A set of attributes was derived through a process of elicitation by individual assessors, based on differences between pairs of stimuli, followed by semi-automatic semantic reduction and then consensus reduction in several group discussions. The authors state that “these attributes can explain the primary differences listeners could meaningfully perceive as a function of the HRTF set employed for binauralization”. The resulting set of attributes

(QFs) are given in table 3.1.

Attribute	End-points	Definition
Position–lateral	More toward the left More toward the right	*Self-explanatory*
Position–front/back	Front Back	*Self-explanatory*
Elevation	More toward the top More toward the bottom	*Self-explanatory*
Externalization	Inside the head Outside the head	Perception of sounds located outside the head
Immersion	Immersive Non-immersive	Feeling of being located in the middle of the audio scene
Realism	Realistic Non-realistic	Sounds seem to come from real sources located around you
Relief	Compact Spread out	Distance between the closest sound objects and the farthest
Coloration	More high frequency content More low frequency content	Feeling of a sound richer in high/medium/low frequencies

Table 3.1: Attributes for describing perceived differences between HRTF sets (Simon, Zacharov, et al., 2016)

Three attributes resulting from the consensus forming sessions were discarded after a validation experiment, because they were found not to differentiate the stimuli. These were: *position–precision*, *reverberation*, and *distance*. The authors commented that panellists came with pre-conceptions of what differences they should hear, with reverberation being an example that persisted until the final validation stage. The attribute *sound level* was also generated and it was found to show differences between stimuli in the validation experiment. However it was not regarded as an important differentiator between HRTF sets, more an indication of the difficulty in equalising the perceived loudness of complex binaural signals.

Compared to previous studies timbral attributes are less prominent and detailed in this instance, indicating lower importance of timbre when comparing HRTF sets to other applications. It should be noted that all measurements were from the same dataset, so avoiding large differences due to measurement systems (Andreopoulou, Begault, et al., 2015), which may have placed more emphasis on timbral differences. The authors state that most of the attributes from their study have identical names in the SAQI, though apart from *externalization* this appears not to be the case, at least in the English version. However there are clearly related concepts for each, particularly from the general and geometry categories of the SAQI. For example, the SAQI attribute *presence* is related to the *immersion* attribute in table 3.1. Assessors reportedly found the *position–lateral* attribute challenging in the validation experiment, since in a complex scene it was not clear to which entity this should relate.

These results appear to suggest that the SAQI is a suitable superset for the QFs of experiences of binaural technology and AVEs. Clearly QFs are context-specific and should be considered for each application domain. But a common vocabulary for describing experiences within a broader field is beneficial, both for assessors in experiments and for interpretation of results by readers.

3.2.5.5 A Common Lexicon of Sound Quality Features

Acknowledging the commonality observed in earlier work, several studies have attempted to define a common set of sound quality features. These studies can be said to be working towards a lexicon of sound quality features. Lawless and Civille (2013) state that "*lexicons* are standardized vocabularies that objectively describe the sensory properties of consumer products." Rather than exposing assessors to a set of stimuli related to the specific study, these studies have been based on reviews of previous work and utilise the knowledge and experience of domain-experts.

It is worth considering the SAQI in this respect. Although it was focussed on the domain of virtual acoustic environments, it is one of the most extensive vocabularies of sound quality attributes presented thus far. Many of the concepts appear applicable to other applications of reproduced sound. However this would need verifying with multiple experiments.

An earlier study by Le Bagousse, Paquier, and Colomes (2010) collected 28 sound quality attributes from the literature and presented them to assessors for categorisation. No sound stimuli were presented during the process. Two methods were used; multidimensional scaling and free sorting with subsequent cluster analysis. The results of both processes led to three categories:

- "*Defects*[8]: are interfering elements or nuisances present in a sound, e.g. noise, distortion, background noise, hum, hiss, disruption...
- *Space*: refers to spatial impression-related characteristics, e.g. depth, width, localization, spatial distribution, reverberation, spatialization, distance, envelopment, immersion...
- *Timbre*: this family is split into 2 subfamilies: The first one deals with the sound color, e.g. brightness, tone color, coloration, clarity, hardness, equalization, richness... The second one composed by homogeneity, stability, sharpness, realism, fidelity and dynamics describes the timbre but can also be related to other characteristics of sound."

[8] The authors use "defaults" in these publications, which is assumed to be a mistranslation since this is later changed to "defects" in Le Bagousse, Paquier, and Colomes (2012).

These are broad QFs that each aggregate multiple concepts, which contrasts with the SAQI where the aim was to obtain more specific and less ambiguous QFs. The motivation of the authors was to generate a small set of attributes that could be used practically in listening experiments. They were subsequently used as rating scales in a listening experiment to evaluate the quality of binaural signals processed through various codecs (Le Bagousse, Paquier, Colomes, and Moulin, 2011).

More recently, Pedersen and Zacharov (2015) conducted a survey of literature that included the development and discussion of quality features relating to reproduced sound. A set of hundreds of sound quality features was collected. The aim was to identify a structured lexicon of sensory QFs for use in sound quality evaluation. Affective attributes were initially removed, limiting the study to descriptive sensory attributes. Analysis was carried out using expert panels to perform projective mapping (see section 3.3.3.5) with subsequent clustering. This was supported by additional attribute elicitation through listening experiments. A hierarchical categorisation of attributes was obtained and each was also given a detailed definition. Pedersen and Zacharov created the Sound Wheel as a hierarchical visual representation of this set of perceptual attributes of sound quality. This was inspired by previous work on wine aromas and that of Letowski (1989). A subset of the attributes was validated in subsequent listening experiments with expert assessors.

This initial work focussed on attributes relating to dynamics, timbral features, and transparency (e.g. naturalness and clarity). Zacharov and Pedersen extended the sound wheel to include spatial attributes by analysing relevant literature. The studies reviewed covered topics of spatial audio recording, processing, and reproduction techniques, including most of the studies reviewed in this section 3.2.5, as well as several studies of the QFs in concert hall acoustics. Analysis was first carried out using semantic text mining techniques to categorise the attributes (Zacharov and Pedersen, 2015). This was followed by further sorting and refinement in expert focus groups (Zacharov, Pedersen, and Pike, 2016), which this author contributed to running. Figure 3.7 shows the resulting Sound Wheel model. In efforts to move the industry towards a standardised lexicon, a version of the Sound Wheel is also given in an ITU-R Report BS.2399 (ITU-R, 2017c). However further development and validation is required.

3.3 Methods for Evaluating Quality

Perceived quality evaluation is a valuable tool in research and development of systems, applications and services. It can be used to inform or validate design decisions, improve production techniques, or identify audience/market groups. The aim is often to understand the

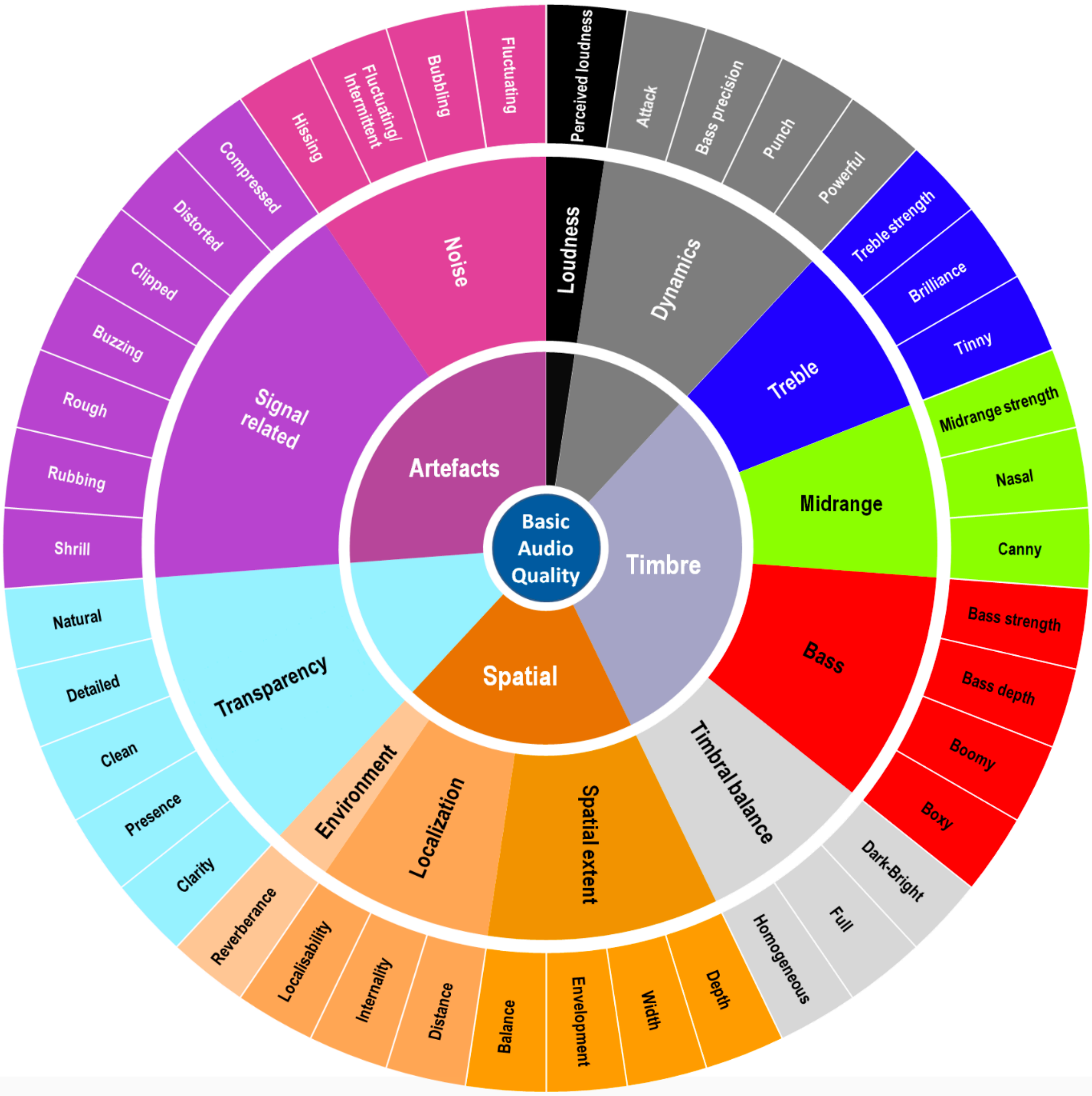


Figure 3.7: The Sound Wheel – A common lexicon of sound quality features (ITU-R, 2017c).

relationship between system factors and perceived quality, and further to optimise the perceived quality within the contextual constraints, such as financial and computational costs, environmental context, or application/task context.

Evaluation can also be carried out in the physical-domain, through analysis of acoustic or electrical signal measurements. Such techniques have a role to play in the development and understanding of audio systems, however they do not give direct information about how such signals will be perceived. Since the goal of audio and media technology is ultimately to provide services and experiences for people, perceived quality evaluation plays a vital role.

There are many techniques available for evaluating quality. Broadly they can be classified as quantitative and qualitative methods, but there are also mixed methods which combine the two approaches. In audio engineering, quantitative methods are more prevalent. One or more factors may be varied and the effect on quality or quality features quantified. However qualitative and mixed methods can offer descriptive insights into the dimensions of quality, exploring the interaction of influence factors, quality features and quality of experience.

It is clear that quality judgements are subjective. Yet experimental methods are most often designed to allow inferences about the quality of (or afforded by) an entity in an objective and externally-valid manner. This is achieved through collection of quality judgements from a group of assessors, with appropriate statistical design and experimental controls. Standards play a key role in this respect, where common, agreed-upon and verified methods are defined for specific application domains.

Bech and Zacharov (2006) give a thorough overview of methods for evaluating the perceived quality of audio systems. This includes considerations of experimental design and practical considerations for running such experiments. The aim of this section is not to repeat this work, but to review how the processes of quality formation presented in section 3.2 apply to practical experimental methods. This provides a basis for the choices of experimental methods used subsequently in this thesis.

3.3.1 The Nature of Judgements

Whilst section 3.2 presents a detailed conceptual model of quality formation, there is still some ambiguity in the processes discussed. In particular, the higher-level cognitive processes of reflection, attribution, comparison and judgement are not well understood. That section stated that the quality formation process could apply to QoE or the quality of a system based on an experience. The entities being considered and the nature of the judgements are not yet well defined; they are specific to the situation in which a judgement occurs.

Bech and Zacharov (2006) describe how quality evaluation can occur in *sensory* and *af-*

fective domains. Other researchers have used related terms. Nunnally and Bernstein (1994) differentiate between responses in terms of "judgements" and "sentiments", whilst Berg and Rumsey (2006) used *descriptive* and *attitudinal* in verbal protocol analysis. In the discussion of sound quality formation by Blauert and Jekosch (2012), the role of affect is not considered part of the quality judgement, but is seen as an influencing factor. Judgements are based on the degree to which expectations are fulfilled. Whilst the QoE definition of Raake and Egger (2014) is clearly affective. Lorho (2010) describes how affective-domain judgements are considered more subjective than those in the sensory domain, incorporating human IFs to a greater extent[9].

Bech and Zacharov (2006) describe *integrative* judgements (one overall measure) and *analytic* judgements (assessments of individual quality features), which are respectively somewhat linked to the affective and sensory domains in nature. Commonly, affective evaluation is made as a single global value, whereas sensory evaluation obtains a multidimensional profile of quality features. Additionally Raake and Egger (2014) describe *utilitarian* evaluation, i.e. how good or bad something is, which is an integrative judgement but on a different basis to an affective judgement. Both perspectives have a role in QoE. Möller, Wältermann, et al. (2014) describe quality features ranging from hedonic (affective) to pragmatic (utilitarian), and Raake and Egger (2014) explicitly state that their definition of experiencing incorporates both aspects, in terms of feelings and concepts respectively. Instead of utilitarian judgements, Bech and Zacharov (2006) describe integrative sensory-domain judgements. Whilst some judgements may be more closely linked to sensory perception, there will always be higher-level processes (and therefore influences) involved.

The nature of the quality judgement in an evaluation experiment will be influenced by the task presented to the assessor. It is likely that a task focussing on the quality of a system will lead to more utilitarian considerations, whereas a task focussing on the QoE will lead to more affective considerations. However many other factors will influence the process, including those internal to the individual, as indicated in figure 3.1.

Integrative evaluation methods are discussed in more detail in section 3.3.2, whilst descriptive analytic evaluation methods are discussed in section 3.3.3. A distinction between analytic sensory and integrative affective evaluation has long been established in the field of food science (Lawless and Heymann, 2010). Often the goal is to optimise the affective quality, as judged through consumer preferences, through study of a food's sensory characteristics, as judged by a panel of sensory experts. There are techniques for mapping between the sensory and affective domains, aiding understanding and informing product development.

[9] Affective judgements will involve greater influence of those human IFs that are involved in higher-level processes.

The expertise of assessors is an important consideration in experimental design. It is often assumed that experts may exhibit biases that mean they are not representative of typical consumers in affective evaluation tasks. Meanwhile untrained consumers are traditionally not deemed to exhibit sufficient acuity and objectivity in sensory evaluation tasks. These matters are discussed further in section 3.3.5.

3.3.2 Integrative Evaluation of Perceived Quality

When quality is treated as an integrative entity, Raake and Egger (2014, p.29) explains that "the influence of the impression for the individual attributes, the context, the mood, the expectations, the previous experience, traditions and so on, are all combined into one single-valued rating." Following the discussion in the previous section, such measures might be utilitarian or affective, depending upon the context of an evaluation. Utilitarian quality is based on the *degree of excellence* with which a system fulfils its intended role i.e. how good or bad it is. This is often the approach taken in standards for evaluating telecommunications systems (ITU-T, 1996, 1998), where quality is measured using mean opinion scores (MOSs). Affective quality might be based on the *degree of liking* of an experience, or *preference* for one experience relative to another. It is not clear that this distinction will have a substantial effect on quality ratings. The differences between opinions and preferences are semantically subtle, as noted by Rumsey, Zieliński, Kassier, et al. (2005b). It is likely that the prior experience of the assessors and the nature of the entities of evaluation will have bigger effects.

In the broadcasting domain, integrative evaluation tasks are often considered in terms of *fidelity* to a given reference stimulus. The desired quality features used for comparison in the quality judgement are explicitly informed by this given external reference, rather than being informed by memory and expectation alone i.e. an internal reference. This approach is taken in standards for evaluating media coding systems (ITU-R, 2015e,b, 2012b). Recommendation ITU-R BS.1534 (ITU-R, 2015b) also includes anchor stimuli, based on defined filtering applied to the reference, to make ratings by different assessors and laboratories more comparable.

A number of different presentation methods exist in established standards. For example, Recommendation ITU-R BS.1284 (ITU-R, 2003a) presents "general methods" for audio system evaluation in the broadcast domain, which includes single-stimulus presentation, paired-comparison and multiple-stimulus presentation methods, either with or without a given reference of target character. The choice of appropriate presentation method is dependent upon the nature of the response variable and the range of different quality levels expected, as well as considerations of task complexity and test duration. Single stimulus methods can be used when large quality differences are observable between the stimuli, whereas paired comparisons allow distinguishing of smaller quality differences. Multiple stimulus comparisons are deemed suitable for intermediate quality differences, and have the benefit of reducing rating duration over paired comparisons. Some methods allow the participant to repeat stimuli as many times as required, whilst others provide a fixed number of presentations. When comparison methods are used, the participant can sometimes freely switch between stimuli, other times this is fixed.

The ITU-R standards for audio evaluation use the response attribute *basic audio quality*, as introduced in section 3.2. This "single, global attribute is used to judge any and all detected differences between the reference and the object" (ITU-R, 2015e,b). In these standards it is recommended that the listener is able to freely control switching between and repetition of stimuli. Such comparison is an effective analytical listening method for experts to identify technical quality issues, avoiding problems of short-term auditory memory. However it may not be an appropriate method for assessing the affective quality of the experience, since it does not reflect the typical context of experiencing broadcast content.

Zacharov, Volk, et al. (2017) performed a comparison of different integrative response attributes and rating scales in the evaluation of coded audio signals by expert assessors. With both high and intermediate quality levels, a large degree of similarity between stimulus ratings was found using the ITU-R continuous quality scale, with the attribute *basic audio quality*, and three varieties of preference scale. The exception was in the introduction of an explicit reference stimulus, which significantly altered ratings for stimuli with higher quality levels. The use of anchor stimuli did not significantly alter scale usage.

Schoeffler (2013) presented a method for evaluating the quality of *overall listening experience (OLE)*. It is considered that the quality of a system evaluated in an analytical manner might not be an accurate measure of the acceptability or desirability of a system by target users in the context of its application, i.e. it has low ecological validity. Following considerations of QoE, a rating of the OLE is intended to reflect "how much someone enjoys a listening experience involving all aspects which are important to him or her" (Schoeffler and Herre, 2016). In these OLE evaluations, single-stimulus presentation is used to obtain item ratings on a 5-point Likert scale. This is considered more representative of real-world listening scenarios, particularly avoiding biasing the listener's internal reference through short-term auditory memory. Ratings are also obtained first for each audio item, without comparing processed versions, to identify how much they like the content without influence of processing.

It has been found that listeners show different preferences in their OLE ratings, with some more strongly influenced by the content and others by technical audio quality aspects such as signal bandwidth (Schoeffler, Edler, et al., 2013) or reproduction format (mono, stereo or 5.1) (Schoeffler, Conrad, et al., 2014). The quality of OLE has been shown to be influenced by individual human factors (Schoeffler and Herre, 2014). By modelling self-reported listener preferences, regression models for predicting OLE were improved. From the analysis, continuous variables describing listener tendencies were deemed more appropriate than distinct clustering of listeners.

Rather than direct scaling of quality, other studies have asked listeners to give their relative preferences for systems, for example by paired-comparison scales (Zacharov and Koivuniemi, 2001b) or stimulus ranking (Lorho, Isherwood, et al., 2002). Wickelmaier et al. (2009) compared both methods, additionally with a technique called "ranking by elimination", where the least desirable of multiple stimuli is successively eliminated. With these indirect rating approaches, the latent quality scale must be inferred by fitting a probabilistic model. It was found that the methods gave similar results when evaluating the quality of coded audio material. Ranking methods are faster on average than paired comparisons, though they suffer from some loss in discrimination.

Some studies have considered different criteria, besides assessment of excellence or preference. Jumisko-Pyykkö et al. (2008) used a binary measure of *acceptance* alongside preference. This acknowledges that contextual constraints may exist and aims to identify thresholds where experiences meet user expectations. Another measure, *appropriateness*, is intended to explicitly focus quality judgement on a specific usage context (Lawless and Heymann, 2010, p.341).

3.3.3 Characterising Quality with Descriptive Analysis Techniques

There are established techniques for evaluating the multidimensional character of an experience or an entity. Many of these techniques were developed for the evaluation of food products (Lawless and Heymann, 2010). Often the aim is to better understand why consumers might prefer one product over another. Collectively the techniques are called *sensory profiling* or *descriptive analysis (DA)* methods. They are traditionally applied to analytic measurement of descriptive QFs and result in a profile of the sensory character of the experience or entity under evaluation. Since the techniques are generally used to relate the QFs to integrative measures of quality, they are also considered methods for *characterising quality*. This section reviews DA techniques, first presenting the classical methods intended for use with expert assessors. It then considers methods better suited to use with assessors with less expertise.

3.3.3.1 Classical Descriptive Analysis – Eliciting Consensus Vocabularies

Classical DA methods are reviewed by Lawless and Heymann (2010, chapter 10). These methods start from the point of eliciting the QFs from assessors, avoiding a priori assumptions of what features exist and are relevant to quality. A panel of assessors derives a common set of descriptive terms and associated scales which can be used to characterise and differentiate a large set of stimuli relevant to the field of application. The assessor panel is then trained to use these attribute scales and rates stimuli in subsequent experiments. These techniques are often known as *consensus vocabulary* methods, the panel must be guided in forming shared concepts and terminology, and trained to use them as a reliable measurement instrument. Such a panel would typically include 8–12 assessors, selected based on experience and sensory acuity (Lawless and Heymann, 2010). The intent is that such panels of experts are maintained over extended periods of time and multiple studies. Examples of such techniques developed for the sensory evaluation of food are the Flavor Profile (Cairncross and Sjostrom, 1950), Quantitative Descriptive Analysis (Stone et al., 1974), and the Sensory Spectrum (Civille and Liska, 1975) methods.

The process of attribute development differs between methods. Typically it will involve many sessions, sometimes of several hours duration, over the course of weeks. Some sessions will be individual and some will be run as facilitated group panel discussions to ensure sufficient alignment of concepts amongst the panel. The key steps to the process are: stimulus familiarisation, attribute elicitation, attribute list reduction and refinement, development of rating scales, and panel training for the rating task. Bech and Zacharov (2006) state that the number of attributes resulting from such a process is typically between 8 and 15. A subsequent rating experiment is used to validate the attributes.

Lawless and Heymann (2010, p.229) define the desirable properties of an attribute for use in DA. These will be considered during validation. In order of importance, an attribute should:

- discriminate between stimuli,
- not exhibit redundancy with respect to other attributes,
- relate to target-user acceptance or preference,
- relate to physical measurements,
- be singular (unidimensional),
- be precise and reliable,
- show consensus amongst assessors,
- be unambiguous,
- be specified by an easily-obtained reference,

- be communicable without jargon,
- relate to reality.

Attributes will also be descriptive rather than attitudinal, so relating to sensory character rather than affect.

Sensory profiles obtained through consensus vocabulary DA methods can be analysed with factor analysis techniques, such as principal component analysis (PCA), to explore the structure of the data, i.e. visualising the relationships between attributes and systems under test in a low-dimensional space defined by the data variance.

Consensus vocabulary techniques with an expert panel were applied several times to audio research questions by researchers at Nokia Labs, to evaluate speech quality in mobile telecommunications (Mattila, 2001), spatial audio techniques (Koivuniemi and Zacharov, 2001; Zacharov and Koivuniemi, 2001a), and stereo enhancement systems (Lorho, 2005a). These studies involved elicitation of verbal descriptors by individuals, followed by formation of a consensus in a guided group discussion. Other examples of consensus vocabulary techniques applied to audio systems are in evaluation of loudspeaker-based spatial audio recording and reproduction (Guastavino and Katz, 2004; Choisel and Wickelmaier, 2007; Francombe, Brookes, et al., 2017), and evaluation of non-individual HRTF sets (Simon, Zacharov, et al., 2016).

Whilst consensus vocabulary elicitation has often been used, and has been shown to produce valid attributes, the required time commitments and associated costs are often a barrier in industrial or applied settings. For example, Lorho (2005a) stated that the training and development process to form a consensus vocabulary took approximately 20 h of contact time with assessors. This is prior to the rating experiments themselves.

3.3.3.2 Preference Mapping

Preference mapping refers to a family of techniques involving statistical analysis of preference patterns. It can be used to gain better insight, allowing optimisation of products/systems (McEwan, 1996). Internal preference mapping looks for underlying patterns in the preference ratings themselves. External preference mapping relates quantitative ratings of QFs to the preferences. These techniques can be a powerful tool in interpreting the importance of various quality features in the formation of preferences.

Lorho (2005a) applied PCA to preference ratings of virtual surround systems to explore patterns of preference for the systems under test and listener preferences in a principal components space. This is an example of internal preference mapping.

External preference mapping uses regression techniques to map between objective ratings of features and preferences. It is a key tool in understanding the importance of quality features on the overall quality. Techniques such as partial least squares regression (PLS-R) can be used. Rumsey, Zieliński, Kassier, et al. (2005a) used PLS-R to map from attributes timbral fidelity, frontal spatial fidelity, and surround spatial fidelity to basic audio quality as rated by expert listeners for five-channel surround sound. Timbral fidelity was found to have a much larger influence on basic audio quality than spatial fidelity, with a resulting regression model equation of BAQ $= 0.80$ Timbral $+ 0.30$ Frontal $+ 0.09$ Surround $- 18.7$. PLS-R methods were also applied to map larger numbers of features to preferences for specific application areas of spatial audio reproduction (Zacharov and Koivuniemi, 2001a), concert hall acoustics (Lokki, Pätynen, et al., 2012), and mobile 3D TV (Strohmeier, Jumisko-Pyykkö, Kunze, and Bici, 2011). Such analyses can lead to complex models of how attributes and their interactions explain the variance in preference ratings amongst stimuli. This can be useful for deriving predictive models of quality from sensory characteristics, though the model parameters can be challenging to interpret directly in terms of required technology developments.

Some preference mapping tools can use a variety of models of the features, such as a vector quantity (the more, the better) or an ideal point (not too much or too little). The PREFMAP algorithm (Schlich, 1995) is one example[10]. Strohmeier, Jumisko-Pyykkö, Kunze, and Bici (2011) reports that this method develops the model by decomposing only the sensory space and mapping preference onto it, whereas PLS-R performs simultaneous decomposition of both affective and sensory datasets. Therefore these two techniques will give different information about the weightings of quality features.

3.3.3.3 Rapid Characterisation

A number of methods for sensory evaluation have been developed that require less expertise and time commitment from assessors (Valentin et al., 2012; Varela and Ares, 2016). These have received a great deal of attention in sensory science fields. They are particularly appealing in industrial applications and it is considered that they may be more relevant to consumer preferences. These techniques often go beyond the sensory-domain, also incorporating descriptors that capture the attitudes and emotions of assessors.

The methods can be broadly categorised into: verbal methods and similarity-based methods. Verbal methods include individual vocabulary DA, where assessors are not re-

[10] Available in XLSTAT `http://www.xlstat.com/en/learning-center/tutorials/running-a-preference-mapping-in-xlstat.html`

quired to adopt shared concepts and terminology. Similarity-based methods instead reduce the requirements on assessors to translate their experience into verbal descriptors, focussing primarily on identifying sensory differences. Whilst qualitative approaches also exist, such as open interview questions and free sorting tasks, the focus here is on mixed-methods.

3.3.3.4 Descriptive Analysis with Individual Vocabularies

Individual vocabulary profiling methods allow each assessor to use their own set of QFs to describe their sensory experience. As stated by Strohmeier, Jumisko-Pyykkö, and Kunze (2010) this allows individuals to use their "own vocabulary, differing sensitivities and idiosyncrasies". Lorho (2005b) stated that individual vocabulary methods are not a replacement for consensus vocabulary methods but a complementary tool: "Such technique could be considered for instance as a quick exploration of the sensory space of a stimulus set, prior to a more thorough descriptive analysis process." However individual vocabulary techniques also allow sensory profiling to be performed with non-expert assessors, who can be considered more representative of target users. This is because assessors are not required to translate their experience onto externally defined concepts. It can be guaranteed that the assessor understands the rating scales.

In free-choice profiling (Williams and Langron, 1984; Jack and Piggott, 1991), assessors define their own attributes and scale end-points. No training is carried out before assessors rate the stimuli. It is therefore a much quicker process than the consensus vocabulary techniques discussed in section 3.3.3.1. Free-choice profiling involves single stimulus presentation. The Flash Profile method (Delarue and Sieffermann, 2004) introduced simultaneous comparison of multiple stimuli during the elicitation stage, to ease the process of attribute elicitation and reduce the time required for familiarisation and training. Multiple stimulus presentation is common in audio evaluation e.g. (ITU-R, 2015b). Lorho (2005b) developed a method based on the Flash Profile to evaluate spatial enhancement systems for headphone reproduction.

The repertory grid technique can also be used to support attribute elicitation for individual methods. It focusses explicitly on those characteristics that differentiate between pairs or triads of stimuli (Thomson and McEwan, 1988). Berg and Rumsey (2006) used this method in the context of surround sound recording techniques. Individual attribute sets were then reduced to a common set by verbal protocol analysis and automatic clustering.

Analysis of rating data from individual vocabulary profiling methods presents additional challenges over consensus methods. Statistical analysis techniques such as multiple factor analysis (MFA) (Abdi et al., 2013), Generalised Procrustes Analysis (Gower, 1975), and hierarchical cluster analysis can be used to identify patterns in the attributes elicited by many individuals. However such analysis is often very complex and still places some reliance on the interpretation of the set of individual terms by the experimenter.

Strohmeier, Jumisko-Pyykkö, and Kunze (2010) and Strohmeier, Jumisko-Pyykkö, Kunze, and Bici (2011) used individual vocabulary techniques in the evaluation of mobile 3D video systems by inexperienced assessors. This was combined with external preference mapping to relate the QFs to overall preferences. The author called this framework of methods Open Profiling of Quality. More recently individual vocabulary profiling and preference mapping techniques have been applied to the evaluation of concert hall acoustics (Lokki, Patynen, et al., 2011; Lokki, Pätynen, et al., 2012) and sound bars (Walton, Evans, Kirk, et al., 2016a). Kaplanis et al. (2017) also used individual vocabulary profiling based on the Flash Profile method to evaluate reproduced sound in car cabin acoustics.

3.3.3.5 Similarity-Based Methods

Classical similarity-based methods for defining perceptual dimensions require multiple pairwise comparisons between stimuli. The data can be analysed to uncover latent perceptual dimensions using techniques such as multidimensional scaling (Borg and Groenen, 2005). This then typically requires the experimenters to interpret the meaning of these dimensions, which may not be intuitive. Performing the required pairwise comparisons between multiple stimuli can also be very time-consuming.

Similarity-based methods using multiple paired comparisons between stimuli have been applied to audio systems, e.g. in evaluation of musical timbres (Grey, 1977), speech transmission (Wältermann et al., 2010) and headphone reproduction (Volk et al., 2016). In (Wältermann et al., 2012), a subsequent listening experiment was used to directly quantify samples along the three latent dimensions obtained in previous tests.

Projective mapping (Risvik et al., 1994) takes a different approach. Individual assessors are tasked with organising the stimuli into a two-dimensional perceptual map that represents their similarities and differences. In other DA methods, such perceptual maps are generated by statistical analysis of attribute ratings or similarity ratings. This has the advantage of explicitly defining the dimensions most important to individuals, rather than those explaining most variance in the data. Individual descriptors can also be elicited within the defined space following mapping of stimuli. This has been called Ultra Flash Profiling (Perrin et al., 2008). Individual perceptual maps are aggregated using similar techniques to individual vocabulary profiling, such as hierarchical MFA. These methods still present the challenge that the experimenter must interpret the meaning of the perceptual dimensions and individual terms.

Projective mapping was applied to DA of loudspeakers by Giacalone, Nitkiewicz, et al. (2017). It was found to provide a rapid overview of the main perceptual differences between models. The authors conclude that projective mapping can be a useful method for DA when time and budget constraints preclude other methods, and also as a means of eliciting attributes for more conventional DA methods. Expert and untrained assessors were used. Interestingly it appears that experts showed better descriptive abilities, but not necessarily better sensory discrimination.

3.3.3.6 Using Common Pre-Defined Attributes

Classical DA methods can be very time consuming and require regular access to a panel of trained expert assessors, which is often problematic in industrial contexts. Quicker methods that use individual attributes or similarity measures require aggregation and interpretation of individual perceptual spaces. The derived terms in many different audio-related studies often have a large degree of overlap, despite their differing application domains (Pedersen and Zacharov, 2015; Zacharov and Pedersen, 2015). It is evident that, to some extent, there is a shared scientific language for describing the perceptual dimensions of sound (see section 3.2.5).

It is common for a small expert panel, often the experimenters, to predefine the attributes to be rated. Often experimenters will select the attribute scales to be used from a previously defined set. For example, Cobos et al. (2015) used attributes defined in ITU-R Recommendations BS.1116 and BS.1286 relating to spatial impression as well as audiovisual correlation (ITU-R, 1997, 2015e). Whilst Millns and Lee (2018) selected four attributes from the work of Rumsey (2002) relating to source/ensemble spatial characteristics, and the width and depth of the environment. Zacharov, Pike, et al. (2016a) (co-authored publication Co.P. VII) used six of the attributes defined in (Zacharov and Pedersen, 2015) in an evaluation of loudspeaker-based spatial audio systems.

Attempts have been made to define comprehensive attribute sets for more general use in audio-related studies. The SAQI was defined in (Lindau, Erbes, et al., 2014), giving a set of 48 attributes and scales for the evaluation of virtual acoustic environments. This used a focus-group approach with domain-experts. It relied on their existing knowledge and experience, rather than exposure to a set of stimuli as in classical DA. Analysis of reproduced sound quality literature was used to obtain the Sound Wheel, with most recent extensions for spatial attributes based on using semi-automatic clustering, followed by expert focus groups (Zacharov, Pedersen, and Pike, 2016) (co-authored publication Co.P. VIII). A version of this is given in an ITU-R Report BS.2399 (ITU-R, 2017c).

Selection from a common set of attributes has advantages in terms of consistency of interpretation across studies. However it must be acknowledged that with novel experiences, shared understanding of attributes and how to use scales may be limited, as discussed by Berg and Rumsey (2006). Even though spatial audio technologies are becoming more widely available, terminology could still be quite specific to a particular application, set of systems or the audio material, or used differently in different contexts. Imposing attribute scales on assessors without their input may make rating more challenging, if they do not understand or agree with the chosen attributes. This directly relates to the experience of the assessors and places importance on assessor training.

Whilst this approach also makes the process more efficient, it can be reductive. In DA, the aim is to obtain a profile of the QFs experienced by the assessors. It is an exploratory technique to characterise quality. This aspect is omitted when experimenters define the attributes to be rated.

Another approach is to first present the assessors with the stimuli to be rated and a pre-defined attribute lexicon, asking them to indicate the attributes that apply to the stimuli. These responses could then be used to create a subset of the lexicon based on consensus of the assessors, for example identifying those attributes that are identified by all assessors or those that allow differentiation of the stimuli under test using statistics for categorical data. The former would still require assessors to rate attributes they perceived as not present and the later removes some attributes that the individual assessor perceives to be relevant. In the quantitative DA study by Moulin et al. (2016), assessors contributed to definition of the common attributes to be rated, but from a pre-defined superset. In an initial elicitation phase, candidate attributes were selected by individual assessors from the Sound Wheel attributes defined in Pedersen and Zacharov (2015). The most frequently selected attributes were used in the subsequent rating experiment. Francombe, Mason, et al. (2014) used a simplified ranking process to reduce an attribute set for DA, identifying those attributes that assessors found most relevant to describe the stimuli.

3.3.3.7 Descriptive Analysis with Individual Attribute Subsets

Another family of verbal techniques for rapid characterisation of quality allows assessors to use individual subsets from a pre-defined vocabulary. These methods have the advantage that common terminology is used to define the sensory space, whilst allowing assessors to respond only with terms they find relevant. This is likely to reduce experiment duration as well as assessor fatigue.

To enable use of untrained target-users with such approaches, simple response formats are adopted. The check-all-that-apply (CATA) method (Adams et al., 2007) requires assessors to indicate with a binary yes/no response whether each of a set of attributes can be used to characterise a stimulus. CATA has been widely applied in product evaluation studies because it allows quick collection and analysis of data from a large number of assessors in a structured format. The CATA method has also been applied to evaluation with domain-experts (Campo, Do, et al., 2008). The rate-all-that-apply (RATA) method (Ares, Bruzzone, et al., 2014) introduces a 3- or 5-point intensity scale for applicable attributes to improve sample discrimination.

Given the commonality in attribute sets developed in audio-related studies, it appears worth exploring such methods that allow assessors to select from a large vocabulary of predefined attributes. This could provide broad coverage of characteristics that an individual assessor might want to use, rather than imposing a limited set of predefined scales.

3.3.3.8 Attribute Rating Procedures

Regarding the procedures for obtaining ratings of QFs from participants, presentation methods are similar to those for integrative evaluation. Considerations of task complexity and duration, as well as the nature of the response attribute apply, as discussed in detail by Bech and Zacharov (2006). Zacharov and Koivuniemi (2001a) used a paired-comparison procedure with a fixed reference. Lorho (2005a) used multiple stimulus presentation to collect ratings for multiple systems on a single attribute. When reflecting on the tests of Lorho (2005a), the author stated that “assessors felt more confident in attribute scaling with the comparative approach than with a single stimulus method” (Lorho, 2005b). Lokki, Patynen, et al. (2011) also used multiple stimulus presentation for attribute ratings. Whereas Strohmeier, Jumisko-Pyykkö, and Kunze (2010) used a single stimulus presentation with absolute scales.

3.3.3.9 Other Methods

An additional tool in sensory evaluation with target users is to collect attribute data for an envisioned ideal stimulus, in addition to rating the presented stimuli (Worch et al., 2013). This can enable researchers to more easily interpret how a sensory profile compares to the assessors’ expectations. This was applied to spatial audio systems in co-authored publication Co.P. VII (Zacharov, Pike, et al., 2016b).

It is worth also noting also that non-verbal descriptive methods exist. This potentially removes barriers to communication of percepts in certain scenarios. Mason et al. (2001) investigated drawing methods in the context of evaluating spatial audio scenes and concluded that for certain attributes that are difficult to describe verbally, non-verbal elicitation can be useful and accurate if the experimental methods are carefully designed.

Silzle (2007) presented a method for exploring the quality of IAVEs by means of a survey of experts. The participants independently evaluated the importance of system IFs and QFs from Silzle's quality taxonomy (see section 3.2.5.4) in relation to QoE. This was performed for several specific applications: localisation test, virtual chat room, edutainment scenario. No listening examples were involved in the process, it relied solely on the experience of the participants. The average of responses was formed and in a second round participants had the opportunity to disagree with this average, which only happened in a few instances. The results presented a complex multi-dimensional relationship between system IFs and QFs, which differed between applications. A comparison was made to the results from a listening test, where assessors rated overall quality as well as the set of QFs without a reference signal. The author concluded that "expert survey is not only a much faster method to get a good overview about a specific application as compared to the listening tests, but it also reveals more information about it." However only a few of the assessors in the listening test were considered as experts. Whilst the efficiency of this method is appealing, it risks bias from survey participants which controlled perceptual experiments can avoid. Also, the relative importance of QFs has been found to differ with assessor expertise, as discussed in section 3.3.5.

3.3.3.10 Summary

There is a wide range of established techniques for characterising quality. Classical methods involve development of domain-specific consensus vocabularies with trained expert assessors. These can be very time consuming and therefore expensive processes. The sensory profiles resulting from these methods are often used to explain trends in preferences by target users. Many methods also exist for characterising quality with untrained assessors who are considered more representative of target users. These methods are also designed to reduce the time required from assessors.

Evaluation methods that utilise individual descriptors or similarity measures to define a sensory space present challenges of interpretation for experimenters. When a shared vocabulary is used to characterise quality, the results can be more directly interpreted. The process is then reliant on assessors understanding the concepts of these attributes and applying them consistently. However, there is evidence to suggest that a shared scientific language exists for describing the features of a listening experience. Provided that assessors have adequate domain-specific knowledge and experience, a common lexicon of quality features may be appropriate.

3.3.4 Discrimination Experiments

Discrimination experiments are another category of evaluation method. These methods typically ask assessors whether stimuli differ in any way, though sometimes differences are considered in terms of a specific QF. As such the required judgements are sensory-domain rather than affective. Discrimination experiments are normally applied to evaluate small differences between stimuli.

A common discrimination test paradigm is two-alternative forced choice (2AFC) where, as the name suggests, the listener has two response options following stimuli presentation and must make a decision either way; if they cannot distinguish then they must give an arbitrary response. There are a number of variants to 2AFC procedures. There may be one presentation of a single stimulus (sometimes called single-interval 2AFC), as used by Hartmann and Wittenberg (1996). In other cases, an assessor is presented with a reference stimulus and two other stimuli (often labelled A and B), one of which is identical to the reference (X). The assessor must identify which is different to the reference. Presentation may be in a fixed sequence, e.g. the oddball paradigm (four-interval: X, A, B, X) used by Moore et al. (2010), or the assessor may be allowed to freely compare between the three stimuli (the ABX test, many-interval), as used by Brinkmann, Lindau, and Weinzierl (2014).

The statistical design of such experiments is based on the binomial distribution, creating a null hypothesis that assumes assessor responses are random (i.e. when no differences can be heard). An alternative minimum-effect hypothesis is then set at a detection level above chance, which is considered minimal in the context of the experiment. If the empirical evidence of the experiment leads to rejection of this alternative hypothesis then it is said to support the null hypothesis, that there are no audible differences (Murphy and Myors, 1999). Considering these hypotheses, experimental parameters are set to balance error probabilities between type-1 and type-2 errors (Leventhal, 1986). A three-alternative forced choice (3AFC) paradigm is also common, as used by Oberem, Fels, et al. (2013). This has different chance probability (0.33 rather than 0.5), but the principals are similar.

Discrimination experiments have often been used in evaluation of binaural technology. Where engineering techniques are developed to make implementation more practicable, the aim is to confirm that these do not cause perceptible changes, for example in evaluating

HRTF interpolation (Minnaar, Plogsties, et al., 2005). They also have a role in validating AVEs, where reference is made to physical external events (real-world events). Several such studies are reviewed in section 2.4.7.

3.3.5 Assessor Expertise

The criteria for recruiting participants in an experiment must be carefully considered. As has been discussed, the role of experience is very influential in forming quality judgements. ISO 8586 (2012) describes five categories of assessor for sensory analysis, from naïve (a person who does not meet any particular criterion) to selected expert (someone with high sensory acuity, experience in the experimental methods, and expertise in the specific type of entity under evaluation).

The required level of experience depends upon the nature of judgements solicited. Assessments in the affective domain are very much subjective and inexperienced assessors are preferred, since they are considered representative of typical users. In this instance, expert assessors may exhibit biases due to their experience, meaning they are less typical of the target population. By contrast, sensory domain assessments generally require expert listeners, with solid understanding of attributes and rating procedures, and the ability to give objective analytical judgements. Characterisation of the sensory profile of a stimulus is expected to be relatively consistent between individuals. In affective measurements similarity between individuals is not necessarily expected, preferences can differ, as shown in (Silzle, Neugebauer, et al., 2009) for binaural technology. ITU-R standards for audio evaluation require expert assessors to make integrative assessments of *basic audio quality* (ITU-R, 2015e). The recommendation is that assessors are selected based on audiometric tests, previous experience, and performance in previous experiments. Post-experiment screening is also recommended, through analysis of the ability of assessors to correctly identify degradations.

Bech and Zacharov (2006) describe the process of assessor development, which includes training in methods and principals of evaluation, as well as evaluation and monitoring of performance in experiments. Report ITU-R BS.2300 (ITU-R, 2014b) provides guidelines for analysis of experimental data to gauge assessor expertise, based on (Lorho, Le Ray, et al., 2010). The method provides indicators of discrimination, reliability and agreement with the panel of assessors from ratings. The metrics are based on analysis of variance (ANOVA) with random permutation tests. This can be used to guide needs for further training and/or screening of assessors. It should be noted that expertise is specific to certain domains. Bech and Zacharov (2006) state that “an expert in audio coding may not (yet) be an expert in another field such as 3D sound evaluation." This highlights the importance of assessor training.

3.3.5.1 Comparing Ratings of Expert and Non-Expert Assessors

Some studies have investigated to what extent the quality ratings of expert assessors relate to those of inexperienced assessors. Schinkel-Bielefeld et al. (2013) demonstrated that experts provide ratings that are broadly similar but more detailed and consistent than inexperienced assessors when using the MUSHRA method to evaluate basic audio quality (BAQ) (ITU-R, 2015b). Similar observations were made by Olive (2003) when assessors rated preference for different loudspeakers.

Rumsey, Zieliński, Kassier, et al. (2005b) compared preference judgements of inexperienced listeners to BAQ ratings by expert listeners and analysed the relationship to timbral, surround spatial and frontal spatial fidelity ratings of the expert assessors. Stimuli taken from audio material in the five-channel surround format were degraded by bandwidth limitation and down-mix processes. The expert ratings were given for this set of stimuli in an earlier experiment (Zieliński, Rumsey, Kassier, et al., 2005). Inexperienced assessors gave preference ratings on a continuous Likert scale. The preference and BAQ scores, when averaged across assessors from each group, had a Pearson correlation coefficient of 0.82. A linear regression model explained 67% of the variance in preferences from BAQ ratings. The standard deviation of error residuals was 10.4 points on the 100-point scale range. In comparison, confidence intervals for the preference ratings ranged from 4 to 8 points. It was concluded that there was relatively large similarity between the two sets of ratings, and preference ratings by inexperienced assessors could, to a limited extent, be predicted by expert BAQ ratings.

Slightly improved prediction of preference ratings was obtained from the experts' attribute ratings through multiple linear regression; 73.6% of variance was explained, and the standard deviation of the residuals was 9.3 points on the 100-point scale. The preference scores were dependent upon the timbral fidelity (unstandardised model coefficient 0.633) and to a lesser extent surround spatial fidelity (0.249), whereas frontal spatial fidelity did not significantly contribute. By contrast, a model of the BAQ ratings showed more emphasis on frontal spatial fidelity (0.356) than surround spatial fidelity (0.052), but again ratings were primarily influenced by timbral fidelity (0.792). This appears to confirm that experts and inexperienced listeners can exhibit different weightings of quality features in their quality judgements.

Schoeffler and Herre (2016) performed an investigation to compare BAQ with OLE ratings. Both expert and non-expert assessors performed both kinds of ratings. Music excerpts in the five-channel surround format were processed with down-mixing and low-pass filtering to introduce degradations. Following earlier findings, experts gave more reliable and critical BAQ ratings. Meanwhile no significant differences were found between OLE ratings by expert and non-expert assessors. BAQ ratings were shown to be more reliable than OLE ratings in general, partly due to the use of multiple- versus single-stimulus presentation (Beresford et al., 2006). Whilst BAQ was again more heavily influenced by timbral quality than spatial quality, OLE showed a roughly even influence of both aspects. The dominance of timbre in BAQ was present for both experts and non-experts.

3.3.5.2 References and Expectations

The desired features of an experience are formed from perceptual references, taking into account the individual's assumptions, which include task, environment context, and past experiences. These desired features form the basis of judgement of the quality of an experience. This was introduced in section 3.2.1.

When unfamiliar (novel) experiences occur, Raake and Egger (2014) state that "a disequilibrium with available references and thus the anticipated event may result." Such situations are relevant to assessor expertise and training, particularly when evaluating new types of technology such as 3D spatial audio systems. Processes of adaptation have been described in these cases (Jekosch, 2005, chapter 5). *Accommodation* refers to the modification of the expected/desired features (changing the reference), and *assimilation* refers to the modification of the perceived character to fit existing references. In the short-term it is thought that both processes occur in part to establish a new equilibrium. Over time, through multiple similar experiences, a process of learning leads to new stable references.

Regarding DA, Lawless and Heymann (2010, p.229) state that a desired property of a QF is that it can be specified by an easily-obtained reference. This is beneficial to clarify the meaning and polarity of attribute scales, particularly when used with assessors not involved in defining them. In food evaluation such references are often obtained by chemical construction. The authors acknowledge that it is not always possible to obtain a suitable reference. Pedersen and Zacharov (2015) describe creation of timbral references by spectral modification of stimuli and Dick et al. (2017) presented methods for generating isolated coding artefacts. Regarding 3D spatial audio evaluation, providing reference stimuli for QF training is challenging. The physical causes for spatial QFs are difficult to isolate and control independently. Simon, Zacharov, et al. (2016) discuss the added challenges when considering binaural technology, individual differences in HRTFs will lead to varied perceptual results.

3.3.6 Predictive Models of Perceived Quality

Given the commitment of resources required to train assessor panels and run listening experiments, predictive models of perceived quality have received significant attention. A model for prediction of the perceived quality of speech in telephony systems was standardised by the ITU as Recommendation P.862 (ITU-T, 2001). Similarly Recommendation ITU-R BS.1387 (ITU-R, 2001) provides a model for perceived quality of audio as processed by low-bitrate audio codecs. These systems extract features of the input signals using psychoacoustic models, including aspects such as spectral masking, and apply machine learning techniques to model the responses obtained through real listening experiments from these feature representations. Rumsey, Zieliński, Jackson, et al. (2008) developed a predictive model for evaluating the quality of spatial audio transmission and reproduction. These systems require access to the unimpaired reference signal for comparison. Single-ended models without access to a reference signal present more significant challenges for predictive modelling.

Besides overall quality prediction, many models of lower-level auditory processes exist. For example, models for directional localisation of binaural stimuli have been published by Faller and Merimaa (2004) and Dietz, Ewert, et al. (2011). These approaches often directly model the neurophysiological processes of the early stages of the auditory system and use a reference HRTF set for matching the extracted cues to obtain a prediction of source direction. Søndergaard and Majdak (2013) present the Auditory Modelling Toolbox, a MATLAB toolbox for constructing such auditory models, containing validated implementations of a number of existing models. This is accompanied by a text book discussing the state-of-the-art in binaural modelling (Blauert, 2014).

Considerations for creating comprehensive reference-free models of sound quality are discussed by Blauert, Kolossa, et al. (2013). This would require extensions to existing auditory modelling approaches to incorporate higher-level cognitive processes. Acknowledging the large degree of context-dependence in quality formation, such models would need to capture the influences of task context and specific individuals in the references provided to the model. Current models tend to be application specific and limited in scope. Many aspects of the quality of listening experiences involving binaural technology cannot yet be modelled.

3.3.7 Psychophysiological Methods

Raake and Egger (2014, p.30) describe a “Schrödinger’s cat problem of QoE research”, where quality evaluation methods that require conscious judgements of an experience may interfere with the experience itself. There is a body of research on the physiological correlates to psychological phenomena and their use in quality evaluation. Engelke et al. (2017) presents a review from a QoE perspective. This includes measurement techniques such as electroencephalography (EEG), electrodermal activity and pupillometry. EEG measurements can be linked to memory and object recognition processes. Wisniewski et al. (2016) examined the event-related potential in EEG measurements when investigating HRTF individualisation (see section 2.5.1.2). Electrodermal activity correlates with emotional arousal and may be used to measure higher-level QFs, such as immersion in virtual environments (Egan et al., 2016). Pupillometry meanwhile has been correlated to cognitive load in the context of speech intelligibility issues (Kuchinsky et al., 2013). Reaction time has also been used as an indirect measure of low-level QFs such as audio-visual coherence (Pike and Stenzel, 2017) and high-level QFs such as immersion (Hinde, 2017). Such methods appear valuable tools in evaluating perceived quality, particularly concerning higher-level cognitive aspects.

3.4 Outlook

This chapter has reviewed the contemporary scientific understanding of perceived quality in the context of audio and media technology, as well as experimental methods for its evaluation. It must now be considered how this applies to the development and application of binaural technology.

Chapter 2 presented the development and state-of-the-art of binaural technology, including a review of several experiments that evaluated the perceptual aspects. Predominantly these studies focussed on whether the binaural rendering could accurately simulate natural external sound sources, which was introduced in section 2.1 as a foundational principle of binaural technology. Evaluation has been either based on task performance, particularly localisation, or discrimination between real and virtual sound events. In certain circumstances, the desired perceptual equivalence has been observed. However for binaural technology to be adopted into systems, applications and services, a beneficial change in the quality of experience should be demonstrated. The introduction of new technology is often associated with additional costs and sufficient benefit must be demonstrated to merit the changes.

Most relevant to the aims of this thesis is the quality of binaural sound in entertainment media applications, such as are relevant to broadcasters. Here benefits to the quality of experience must be observed in comparison to existing stereo headphone listening. The associated costs of technical implementation will be considered alongside further required changes to workflows and staff training. Quality evaluation should provide evidence to support strategic decision making. Quality should be evaluated within specific applications to assess potential benefits to the experience of audiences. Additionally evaluation methods should allow exploration of the effects of system design decisions within constraints such as compatibility with existing systems, or complexity and cost of implementation.

There appears to be little existing evidence that binaural technology provides significant benefits to the overall quality of experience in entertainment media applications. Chapter 4 presents an assessment of the state-of-the-art in applications of binaural technology to broadcasting at the outset of work on this thesis, with a review of previous studies and an experiment to evaluate the perceived quality of the latest systems on offer. At that time, in 2012, virtual surround sound was the most common and relevant application of binaural technology for broadcasters. More recent trends in 3D spatial audio applications, such as next-generation audio (NGA) codecs, and virtual reality (VR) and augmented reality (AR) systems will be discussed in chapter 6.

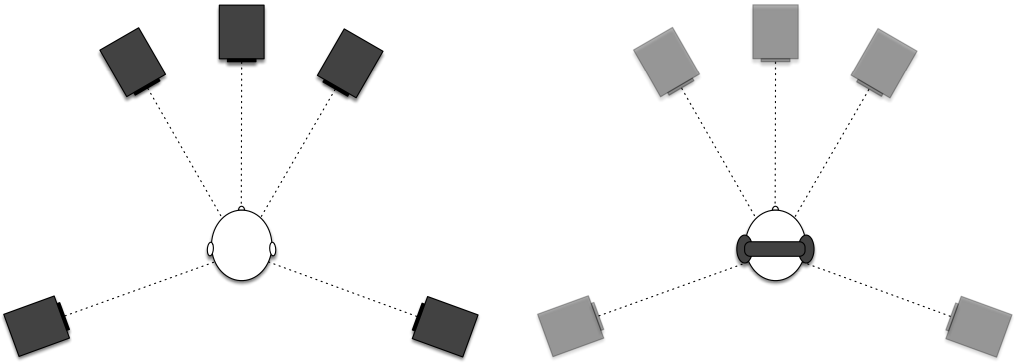
Chapter 4

Perceived Quality of Headphone Surround Sound Processing – A Pilot Study

This chapter presents a pilot study of the potential role of binaural technology in broadcasting, carried out at the beginning of this thesis project in 2012. The perceived quality of a range of commercially-available systems that perform virtual surround sound processing for headphones is evaluated using existing broadcast material in the 5.1 surround format.

4.1 Introduction

Many broadcasters currently distribute television programmes with 5.1 surround sound, a discrete multichannel stereophonic format intended for reproduction over a 3/2 loudspeaker configuration (ITU-R, 2012a), and a large body of programme material exists in this format. However, in recent years there has been a significant growth in the proportion of the audiences of these programmes who are watching on portable devices and listening to the programme sound using headphones. Most devices will use a stereo mix for headphone output, whether a separate artistic mix or a down-mix from the 5.1 signal. This gives an altered spatial image when compared with the 5.1 loudspeaker reproduction. Binaural processing can be used to create an alternative signal for headphone listening from the 5.1 signal, attempting to give an improved spatial listening experience. This kind of processing can be termed *virtual surround sound* or alternatively *headphone surround sound processing (HSSP)*. Figure 4.1 illustrates the idea.



(a) 3/2 surround sound on loudspeakers.

(b) Virtual 3/2 surround sound for headphones.

Figure 4.1: The headphone surround sound processing (HSSP) concept.

There are many commercially available products that offer virtual surround for headphones. This chapter presents an evaluation of the state-of-the-art systems for creating this experience which were available at the start of the project in 2012. Similar investigations have been carried out previously and are reviewed in section 4.2. At that time, systems showed little significant improvement over a stereo down-mix signal. The study reported here was conducted to review progress in this technology.

4.2 Background

4.2.1 Short Review of Binaural Rendering

The method for rendering virtual sound sources to headphones using head-related impulse responses (HRIRs) was presented and validated by Wightman and Kistler (1989b). Experimental evidence has suggested that rendering with anechoic HRIRs requires personalised measurements for plausible virtual sound sources located outside of the head with good directional accuracy (Hartmann and Wittenberg, 1996; Minnaar, Olesen, et al., 2001). However, when binaural room impulse responses (BRIRs) are used in combination with head-tracking to compensate for head motion, plausible rendering of virtual sound sources can be achieved with high localisation accuracy (Lindau and Weinzierl, 2012; Horbach et al., 1999). The influence of system factors has been addressed in terms of localisation (Begault, Wenzel, and Anderson, 2001), but the effects on overall sound quality are unclear. Many quality features contribute to the overall quality of experience provided by a binaural rendering system (Silzle, 2007). Binaural processing is known often to introduce timbral colouration. Equalisation can be applied to reduce colouration, although this may degrade localisation acuity (Merimaa, 2010). Commercial systems will use a variety of techniques to aim to achieve good overall quality. Chapter 2 gives a detailed review of binaural technology.

4.2.2 Previous Evaluations of Headphone Surround Sound Processing

Previous quality evaluation experiments have not found significantly improved quality for virtual surround systems compared with a conventional stereo down-mix. All of these experiments evaluated systems that use non-personalised head-related transfer functions (HRTFs) and did not use head tracking.

Lorho (2005a) assessed commercial stereo enhancement systems for headphone listening, some of which were said to use binaural rendering techniques. Relative preference ratings were given in comparison to a common stereo reference, using a range of stereo music recordings as stimuli. No systems were rated significantly higher than unprocessed stereo across the music items. Attribute ratings were also obtained using a consensus vocabulary descriptive analysis (DA) process. This indicated that those systems that were most preferred, i.e. rated closest to the stereo reference, were also closest in character to the original stereo signal. Those systems that were rated poorly showed distinctive differences in the sensory space however: in terms of tone colour and clarity, as well as spatial characteristics. The spatial differences appear dominated by the differences between the stereo signals and the mono down-mix. Many systems showed more low-frequency emphasis and less clarity compared with the unprocessed stereo signal. The level of reverberation (echo) and associated broadness of image was negatively correlated with the quality of reverberation and distortion (defined as metallic/machine-like artefacts).

In an earlier experiment, assessors gave preference ratings for headphone surround sound processing systems (Lorho and Zacharov, 2004). The authors listed several of the systems evaluated, but the results were anonymised. The audio source material was from movies, music and a computer game in the 5.1 surround format. The experiment did not present accompanying video to the listeners. It was found that no systems were rated significantly better than a basic stereo down-mix. Significant variations in system preferences were found according to content item and assessor.

Silzle, Neugebauer, et al. (2009) performed an experiment similar to Lorho and Zacharov (2004) using a wide range of 5.1 surround content, including music, movies, and a documentary. Again there were no accompanying videos shown to assessors. Two headphone surround sound processing systems were compared again to a stereo down-mix. No details of these systems were given. Overall no significant increase in quality was observed, however when the listeners were split into two groups, by performing a clustering on their stimulus ratings, a group of "binaural lovers" was clearly identified. This group corresponded to 38% of the participants and the virtual surround algorithms were both rated significantly better than stereo by this group. The other larger group clearly preferred the stereo down-mix and was described as "downmix lovers". The authors argued that this shows the importance of offering control of the processing to the listeners in consumer applications.

The virtual surround systems in the studies reviewed above were all commercial products or prototypes. Precise details of the binaural processing in these systems is not available. These systems are limited by the constraints of uncontrolled environments in order to create market-ready products. They lack HRTF individualisation, headphone equalisation, and head tracking. However, it is expected that the manufacturers will have spent significant effort in tuning the systems to achieve high sound quality, see e.g. (Silzle, 2002; Supper, 2010).

It was decided to perform a listening experiment to assess the latest developments in this technology and also to explore some additional system factors, such as the roles of individualisation, head tracking and the listening environment. Since virtual surround sound processing is likely to be used with accompanying video, it was deemed appropriate also to evaluate quality in this context.

4.3 Application

The experiment was designed considering broadcast and media distribution applications, where the processing is applied on the distributer-side rather than on a listener's own device. The virtual surround systems are considered as a *black box* process, taking a completed 5.1 programme and converting it for headphones using fixed parameters, as illustrated in figure 4.2. This study targets the scenario where the system is placed in the distribution chain of a broadcaster, so that all content is passed through this process to provide a dedicated headphone surround service. This represents the lowest cost approach for distribution of enhanced headphone sound, as it would require no additional production effort. However, it is also the most challenging scenario for the systems under evaluation, since there is no opportunity for manual adaptation of system parameters according to input signals. In particular, no processing that is specific to the listener and their environment context is possible. The aim is to establish whether in this context virtual surround systems can provide improved quality over existing stereo down-mix techniques.

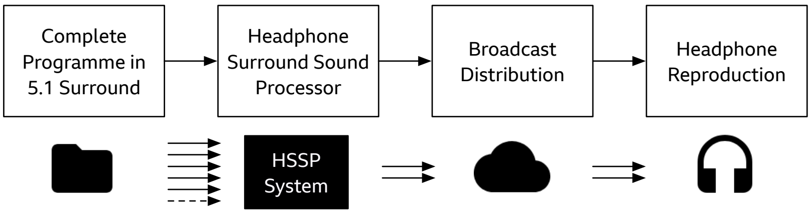

Complete Programme in 5.1 Surround
Headphone Surround Sound Processor
Broadcast Distribution
Headphone Reproduction
HSSP System

Figure 4.2: Application scenario – black box headphone surround sound processing (HSSP) of 5.1 multichannel input signal.

4.4 Experiment Variables

4.4.1 Headphone Surround Sound Processing Systems

A survey of available market products was conducted and many organisations were approached to provide systems for inclusion in this study. Twelve virtual surround systems were provided by commercial companies for evaluation, some of which were systems in development and not available for purchase[1]. Each of the systems under investigation was configured using parameter values provided by the system supplier.

Besides the systems provided by external organisations, two additional virtual surround system configurations were included for comparison. These used measured BRIRs to render virtual loudspeakers. The BRIRs were measured in the same controlled listening room used in the evaluation (see section 4.4.3). A commercial product was used for impulse response measurement and convolution, the Smyth Realiser A8 (Smyth Research, 2018b). It also allowed for equalisation of the headphone-to-ear response. One of the additional systems used BRIR measurements made on a KEMAR head and torso simulator (HATS), whilst the other used individual measurements and head tracking.

KEMAR system: The KEMAR HATS BRIRs were measured with it placed at the central listening position in the controlled listening room. No headphone equalisation or head tracking was applied during rendering for this system, because this is not feasible in the intended application. The KEMAR HATS has often been used in binaural research studies. This system represents a straightforward, repeatable approach to virtual surround rendering, equivalent to recording the loudspeaker reproduction with the HATS. This is referred to from herein as the *KEMAR* system.

Individualised system: Individual BRIRs were measured for each participant at three head yaw orientations (0° and ±30°), using the Smyth Realiser system. The system performs an

[1] Due to commercial sensitivities, the systems assessed cannot be identified. The sample covered the market well at the time of investigation, with many major organisations in the field represented.

automated measurement sequence by instructing the listener to point their head directly at the front centre, front left, and front right loudspeakers; at each head orientation exponential sine sweeps are used to measure BRIRs from each loudspeaker. Miniature microphones, provided with the system, are placed in foam inserts at the entrance to the ear canals.

A head tracking device allowed the BRIR filters in the convolver to be interpolated and updated in real-time by the Smyth Realiser system according to the tracked head yaw angle, preventing the virtual loudspeakers from moving with the listener (Smyth, 2005). An individual headphone equalisation function was also calculated for each assessor using the Smyth Realiser system and applied for this test system configuration only. This head-tracked individualised rendering approach could not be applied in the intended application, it was included as a point of comparison. The expectation was that it would provide high spatial quality for the listener, since the auditory impression created should be close to those given by the real 5.1 loudspeaker system in the listening room. This system will be referred to as the *individualised* system.

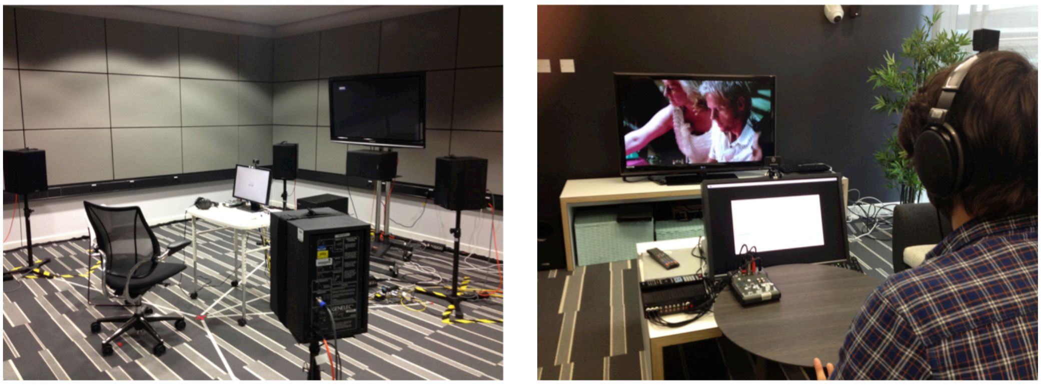
4.4.2 Audio Source Material

Ten audio items were used in the quality evaluation experiment. Seven were taken from broadcast programmes and had accompanying video. Broadcast sound is varied and often contains complex dynamic scenes. Test items were selected by the author and an experienced colleague to represent a range of scenes and genres found in broadcast sound and to reveal critical behaviours of the systems. In addition to the broadcast items, three items were chosen to present simpler sound scenes that may allow further insight. These items did not have accompanying video. All items were approximately 12 s long.

The chosen audio items are described in table 4.1, with an indication of spatial scene characteristics based on (Zieliński, Rumsey, and Bech, 2002). The symbol F indicates that signals contain foreground sounds that are clearly perceived and significant to the action, and B indicates that signals contain background sounds that primarily contain ambience. The scene characteristics in the frontal and rear surround channels are indicated with these symbols. For example, F-B refers to a scene where the frontal channels contain foreground activity and the rear surround channels contain background activity. The labels C and LFE are added when significant content exists in the front-centre and low-frequency effect channels respectively. The majority of the test items had a frontally dominant sound scene and contained broadband ambience, which reflects the conventions of broadcast television production.

ID	Name	Type	Scene	Video	Description
0	HDTrail	Sci-fi	F-F	Y	BBC HD channel trailer, many synthetic sound sources and effects, complex scene
1	PianoVox	Live music	F-B	Y	Voice, piano and cymbals. Audible concert hall reverberation
2	RockBand	Live music	F-B	Y	Live rock band, compressed dynamic range, distorted guitars, vocals
3	Explosion	Drama	F-B+LFE	Y	Cinematic explosion, with sound effects moving rapidly from front to surround. Background music (F-B)
4	Conversation	Drama	F-B+C	Y	On-screen speech in centre channel with strong ambience, occasional background voices
5	NHNarration	Wildlife	B-B+C	Y	Narration in centre channel with jungle ambience and low frequency rumble
6	NHMusic	Wildlife	F-B	Y	Orchestral music and transient sound effects
7	PannedNoise	Test signal	F-F	N	Amplitude-modulated white noise panned around the listener
8	Speech	Test signal	C	N	Dry female speech signal in centre channel
9	Applause	Test signal	B-F	N	Applause in surround channels, ambience in the front

Table 4.1: Items of test material used in the evaluation.



(a) Controlled listening room

(b) Usability laboratory

Figure 4.3: Listening environments used in the experiment.

4.4.3 Listening Environments

All assessors carried out the evaluation in two different rooms, to investigate whether a change in listening environment caused a significant difference in perceived quality of the systems. A controlled listening room environment was used with a volume of approximately 99 m^3, background noise level below NR10 (ISO 1996-1:2016, 2016) and a mean reverberation time from 200 Hz to 10 kHz of $T_{60} = 0.21$ s. This environment was used to measure BRIRs for the *Individualised* and *KEMAR* systems (section 4.4.1). Genelec 1031A loudspeakers, arranged according to (ITU-R, 2012a), were used for this measurement process, with a Genelec 1031A subwoofer. These loudspeakers were left in place during the evaluation, providing visual cues that may potentially influence the spatial impression of virtual sound sources in this environment (Cote et al., 2012). The other room was a usability laboratory, which is designed to be similar to a domestic living room, and has significantly higher background noise level and a different reverberation character to the controlled listening room. Both rooms are shown in figure 4.3 and are at the BBC Research & Development laboratory in MediaCityUK, Salford.

4.5 Experiment Description

4.5.1 Methodology

The experiment used a multiple-stimulus presentation method. This method was chosen because it allows quick and reliable evaluation of multiple systems by presenting them simultaneously in the test interface. Test duration was an important consideration in the design given the large number of test cases.

The method was based on Recommendation ITU-R BS.1534-1 (ITU-R, 2003b), which is a recommendation for testing of coding systems of intermediate quality. In such tests there is a known target quality, that of the uncoded signal, which is provided as an explicit reference used to determine a quality rating for each stimulus. In the case of virtual surround systems there is no known target. Given the application under investigation, the systems were assessed in comparison to a stereo down-mix of the 5.1 signal, created according to Annexe 4 of Recommendation ITU-R BS.775 (ITU-R, 2012a) and labelled as *Stereo* herein. This represents a common method of presenting 5.1 surround signals to two channel receivers, such as devices with headphones. This signal is not a high quality reference, but an alternative to be compared against, so a bipolar comparison rating scale was used.

Assessors were asked to rate the *sound quality* of the stimuli, "taking into account all aspects of the sound". The scale, shown in figure 4.4, was continuous and had a numeric range from +5 to -5 with text labels indicating the meaning of positive and negative ratings, and that zero implies equal quality to the reference. This is similar to the comparison scale given in Recommendation ITU-R BS.1284 (ITU-R, 2003a). Assessors were asked to rate each stimulus with a score of overall sound quality compared to the stereo down-mix.

Another version of the stereo down-mix signal was included as a hidden reference stimulus in each trial to assess a listener's ability to detect differences between systems, similar to the hidden reference in (ITU-R, 2003b). A mono down-mix of the 5.1 signal was also included as an anchor signal without spatial information, again created according to Annexe 4 of ITU-R (2012a). To prevent cognitive overload it has been recommended that an assessor is presented with no more than 15 stimuli in a single trial (ITU-R, 2003b), including the explicit reference, hidden reference and anchors. However, rating of this many stimuli was found challenging in pilot trials. In order to limit the duration of the test, each assessor was assigned a subset of 6 of the 12 systems to evaluate, chosen at random. The KEMAR and individual systems were included for every assessor, alongside the mono and stereo downmixes.

4.5.2 Preparation of test material

The systems were used to process each of the test items to create a set of stimuli for headphone listening. The audio source material was uncompressed, with 6 discrete channels, sampled at 48 kHz with 24-bit resolution, except for items 2, 3 and 4, which were originally encoded with Dolby E at 2250 kbps and were then decoded to 24-bit linear PCM. Video material was sourced from broadcast archives and encoded in a variety of high-quality formats.

All items were gain-adjusted to equalise the perceived loudness. The algorithm specified in Recommendation ITU-R BS.1770 (ITU-R, 2011) was used to provide an objective measure of the item loudness. The programme level was then adjusted to obtain a loudness value of −23 LUFS (EBU R128, 2014). This algorithm was designed considering loudspeaker reproduction. It includes a filter to model the effects of the head on incoming signals, which should not be needed in headphone reproduction. Stimuli were loudness normalised using a modified algorithm without the head-shadowing filter. Listening by this author and an experienced colleague suggested that this gave a closer loudness matching between stimuli than the complete algorithm.

Loudness equalisation of the dynamic individualised BRIR system involved an additional step, since it was processed in real-time using listener-specific measurements. Four expert assessors were twice asked to adjust a gain control on the output of this system (individualised for them) until it had equal loudness to the stereo reference signal, first starting from −6 dB and then again starting from +6 dB. The mean gain adjustment across all eight trials was used for all test participants, responses had a standard deviation of 1.24 dB.

Binaural processing systems introduce a delay to the source signals. This is partly due to the acoustic propagation that is modelled in the binaural impulse responses, but also due to system processing latency. In order to allow seamless switching between the stimuli, the delays were calculated using cross-correlation with the stereo reference and removed to the nearest sample. This could not be achieved for the real-time head-tracked rendering, but it is clear when this system is active anyway since it is the only one that responds to head movement.

4.5.3 Assessors

A sample of 41 assessors took part in the experiment, with a median age of 29 and an interquartile range of 11. All assessors had previous experience in listening critically to audio material, either in a professional context or as an experienced musician. The number of assessors required was based on an estimate of test power made using the G*Power software (Faul et al., 2007)[2].

[2] This calculation assumed test power of 95 % for detecting a small effect size ($f = 0.1$ with $\alpha = 0.05$) and considering the two-way interaction between fixed-effects *System* and *Item*, which had the largest degrees of freedom (135). This resulted in a minimum sample size of 6529 which required a minimum of 33 assessors. Though this makes a number of assumptions that may not hold with the obtained data.

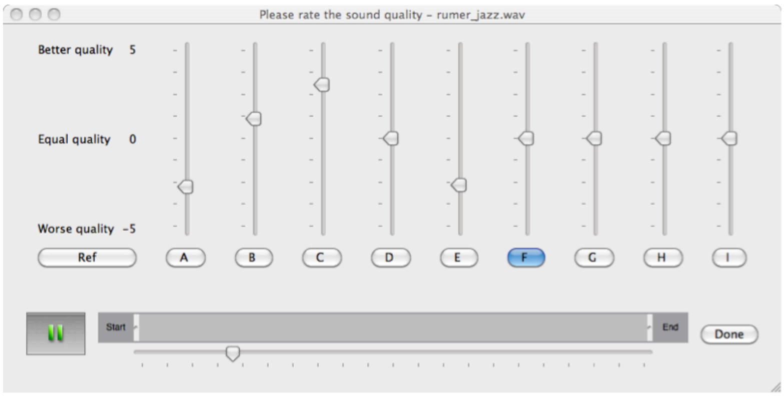


Figure 4.4: User interface for auditioning and grading of stimuli.

4.5.4 Experiment administration

A trial was performed for each of the 10 audio items in both listening environments, with 6 commercial system stimuli plus the hidden reference and anchors (10 stimuli per trial). Trials were presented in a randomised order, which differed between the two rooms. To prevent fatigue the experiment was split into four sessions of five trials. Sessions lasted approximately 30 minutes but durations varied between assessors. Assessors were given time to rest between sessions. In total the experiment took 2–3 hours per assessor. The assessors performed two sessions in one listening environment before doing two sessions in the other environment. The choice of first environment was randomised to reduce any systematic effect on the results.

High-quality circumaural diffuse-field equalised headphones (Sennheiser HD 650) were used for sound reproduction. In both listening environments the assessor sat at a small desk with a mouse and desktop monitor. Custom software was written to allow the assessor to audition and rate the test stimuli, using methods specified in (ITU-R, 2003b). The software recorded ratings to two decimal places. The graphical user interface for the software is shown in figure 4.4. Where audio items had accompanying video, it was displayed on a 47 inch flat panel television placed 2 m in front of the listener.

Individual BRIRs and headphone equalisation measurements were first made for each assessor in the controlled listening room. Each assessor then went through a familiarisation process before the grading sessions. This involved first listening to the stereo down-mix of each audio item, to introduce the programme material. Subsequently the assessors used the rating interface to evaluate stimuli for two training items, which were not in the main experiment. This latter step allowed familiarisation with both the interface and the range of quality of the systems.

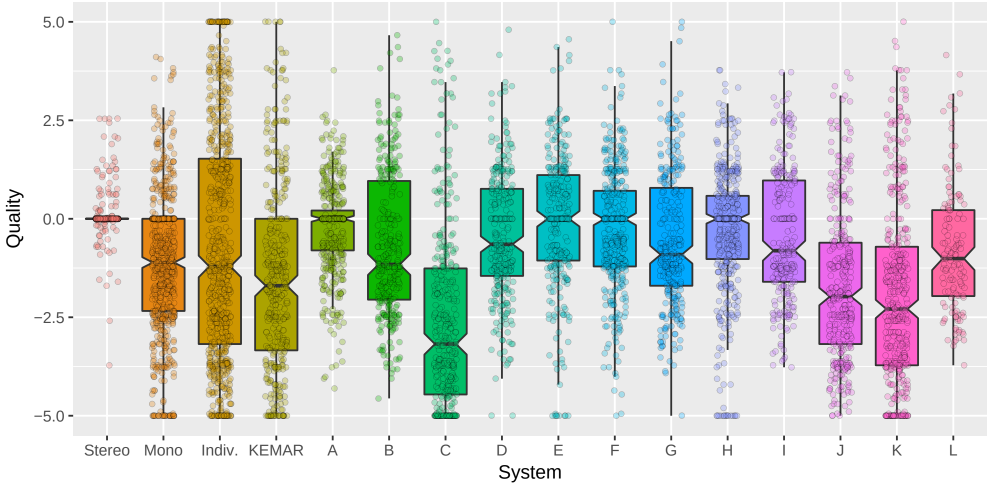


Figure 4.5: Combined scatter and box plots of quality ratings over all audio items

4.6 Results

The quality ratings can be seen in figure 4.5 for each system, grouped across attributes and rooms. The overall distribution of results showed a strong peak at a rating of zero, implying that often participants found it hard to differentiate between the stimuli and the stereo reference. There was also a skew towards negative ratings.

4.6.1 Post-Screening of Assessors

Before analysing the grading data a post-screening of assessors was carried out, rejecting responses of some assessors who were unable to make reliable discriminations. The assumption was that a reliable assessor should rate the hidden reference with a score of zero, corresponding to equal quality to the explicit reference.

A coefficient of variation (CV) was calculated for each assessor according to the equation:

$$CV = \frac{\sigma}{(x_{max} - x_{min})} \times 100\% \tag{4.1}$$

where σ is the standard deviation of scores for the hidden reference given by the assessor and x_{max} and x_{min} are the maximum and minimum values on the rating scale respectively. The CV was calculated for each assessor over all audio items and both rooms. Zieliński, Rumsey, and Bech (2002) state that a CV of 10 % or lower is acceptable. There were four assessors with a CV over 10 %, after removing their data there were 37 assessors remaining.

4.6.2 Broadcast Video Items

Results are analysed first for the seven broadcast programme items. The three audio-only items are discussed separately.

4.6.2.1 Omnibus Tests of Experiment Factors

Often a repeated-measures analysis of variance (RM-ANOVA) is used as an omnibus test of factorial design experiments such as this. However, since assessors rated different sets of systems, this test is not appropriate. Instead a *multilevel linear model* was fitted to the data. Multilevel models are hierarchical linear regression models that represent hierarchical structure in the data, explaining sources of variance at each of the different levels. For within-subjects experimental designs, this can be used to represent correlations between the responses of an individual. Furthermore, when different systems are evaluated under different conditions (e.g. rendering different audio items) then the ratings across items are likely to be correlated by system. This hierarchical representation allows modelling of the variance between assessors, the variance between systems within assessors, and the variance between items within systems within assessors. They can be used where there is dependence between observations at multiple levels. For further details of the method, the reader is referred to (Quen and Bergh, 2004; Hoffman and Rovine, 2007), only an overview can be provided here.

Multilevel models have the advantage that they do not make the assumptions of homogeneity of variances or sphericity that exist with RM-ANOVA (Quen and Bergh, 2004; Hoffman and Rovine, 2007). These models are also able to handle imbalanced or missing data, which is important for this dataset, since assessors rated different randomly-assigned subsets of systems. Multilevel models are estimated using maximum likelihood methods. In order to test whether experimental variables have a significant effect, models with and without that variable as a predictor are compared using a χ^2 likelihood ratio test, i.e. the change in log-likelihood (logLik) between models and the associated change in degrees of freedom (df) are tested against the corresponding χ^2 distribution.

$$\chi^2_{\text{test}} = 2\text{logLik}_{\text{new}} - 2\text{logLik}_{\text{old}} \tag{4.2}$$

$$\text{df}_{\text{test}} = \text{df}_{\text{new}} - \text{df}_{\text{old}} \tag{4.3}$$

The quality ratings were analysed with a multilevel model, fitted using maximum likelihood estimation, using the `nlme` R package (Pinheiro et al., 2018). The nested random

	Fixed Effects Model	df	AIC	BIC	logLik	Test	χ^2_{test}	p-value
1	Intercept only	6	16481.84	16520.52	-8234.920			
2	+***System***	21	16308.08	16443.47	-8133.038	1 vs 2	203.764	<.001***
3	+***Item***	27	16250.30	16424.37	-8098.148	2 vs 3	69.779	<.001***
4	+***Room***	28	16247.91	16428.43	-8095.956	3 vs 4	4.385	0.036*
5	+***Assessor***	64	16213.90	16626.52	-8042.952	4 vs 5	106.008	<.001***
6	+***System:Item***	154	15841.23	16834.10	-7766.615	5 vs 6	552.673	<.001***
7	+*System:Room*	169	15848.70	16938.27	-7755.348	6 vs 7	22.535	0.094

Table 4.2: Likelihood ratio tests for multi-level linear model fitting of quality rating data

effects model was: ~1 | *Assessor* / *System* / *Item* / *Room*. Table 4.2 shows the results of introducing successive fixed-effect model components, with bold font indicating significant effects and asterisks indicating significance level ($p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$). The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also shown for each model.

All main factors were found to significantly affect quality ratings, although *Room* to a lesser extent than others. Particularly noteworthy is that the systems were rated differently overall. There was a significant interaction between *System* and *Item*, indicating that the systems were rated differently between audio items. The interaction between *System* and *Room* was not significant, suggesting that systems were rated similarly in the two listening environments. That the main effect of *Assessor* was significant indicates that assessors used the scale differently. It is not appropriate to apply a centring and scaling transformation to individual ratings, because each rated a different subset of systems. This also means that the interaction between *System* and *Assessor* effects cannot be modelled. However, on inspection of the data, it appears that there were also large differences in system ratings between assessors. This is examined in more detail in section 4.6.4.

4.6.2.2 Effect of Listening Environment

The effect of the room was investigated further on the individualised system data. The BRIRs used by this system incorporate the room response of one of the listening environments (the controlled listening room) and it is expected that assessors will be most sensitive to the room effect in this case. In the controlled listening room, ratings for this system had mean and standard error of $\mu = -0.503, \sigma_\mu = 0.153$, whilst in the usability laboratory it had ratings of $\mu = -0.822, \sigma_\mu = 0.153$. A two-tailed paired-samples t-test was applied with significant result $t(369) = 2.8981, p = 0.0039$, though this only represents a small effect ($r = 0.149$). When viewed across all systems and items, ratings in the usability laboratory showed a very small decrease, with mean difference of 0.063. It should be recalled that the interaction

System	Pseudo-median	95 % C.I.		Z	p-value	r
		Low	High			
Stereo	0.42	0.18	0.66	2033.5	0.001^{**}	0.153
Mono	-1.23	-1.36	-1.11	17802.5	$<.001^{***}$	0.568
Indiv.	-1.13	-1.40	-0.84	40624.5	$<.001^{***}$	0.329
KEMAR	-1.77	-2.07	-1.48	5933.5	$<.001^{***}$	0.590
A	-0.07	-0.27	0.08	8779.5	0.387	0.048
B	-1.01	-1.28	-0.71	5193.0	$<.001^{***}$	0.469
C	-3.10	-3.35	-2.86	1077.5	$<.001^{***}$	0.803
D	-0.52	-0.77	-0.28	5571.0	$<.001^{***}$	0.306
E	0.42	0.18	0.66	8440.0	0.001^{**}	0.235
F	-0.60	-0.82	-0.37	9070.5	$<.001^{***}$	0.313
G	-0.91	-1.16	-0.64	5211.5	$<.001^{***}$	0.434
H	0.37	0.20	0.54	14834.5	$<.001^{***}$	0.226
I	-0.77	-1.06	-0.44	3726.5	$<.001^{***}$	0.370
J	-1.89	-2.12	-1.67	1981.0	$<.001^{***}$	0.725
K	-2.34	-2.59	-2.09	3578.0	$<.001^{***}$	0.719
L	-1.33	-1.63	-1.06	556.5	$<.001^{***}$	0.630

Table 4.3: Wilcoxon signed-rank tests of rating distributions for each system across broadcast video items, with pseudo-medians and associated 95 % confidence intervals, and estimated effect size r.

between *System* and *Room* factors was not significant in the multilevel model. Based on this finding, all descriptive statistical plots presented herein are for data from both listening rooms combined.

4.6.2.3 System Ratings

The ratings for each system are not all normally distributed[3]. Non-parametric Wilcoxon signed-rank tests were applied to the ratings for each system to test whether they are symmetric about zero. The results are shown in table 4.3 along with associated pseudomedians and their 95 % confidence intervals. The pseudomedian is the median of all midpoints of pairs of data, it is used as a measure of the centrality of a distribution. Based on this metric, only system A was not rated significantly different to the stereo down-mix reference at a 95 % confidence level. It can be seen that systems E and H had distributions significantly skewed towards positive ratings, though these are small effects. Both systems had a median rating of zero. All other systems had ratings skewed negatively, including the dynamic individualised BRIR system. Four systems (C, J, K, and KEMAR) had median ratings significantly below the mono down-mix anchor, as indicated by Wilcoxon signed-rank tests.

[3] An earlier publication of these results presented means and 95 % confidence intervals. Box plots are used here to give better insight into the underlying data. However, very similar interpretations can be made from both presentations.

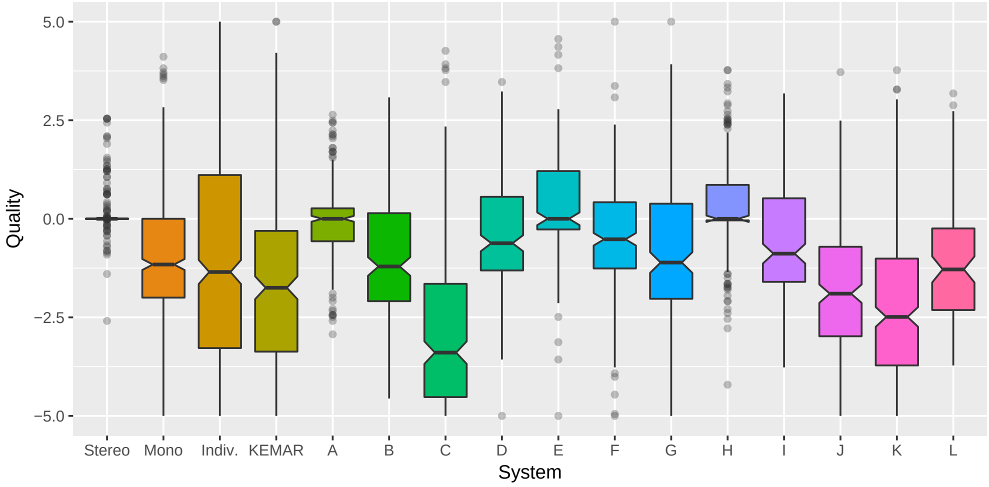
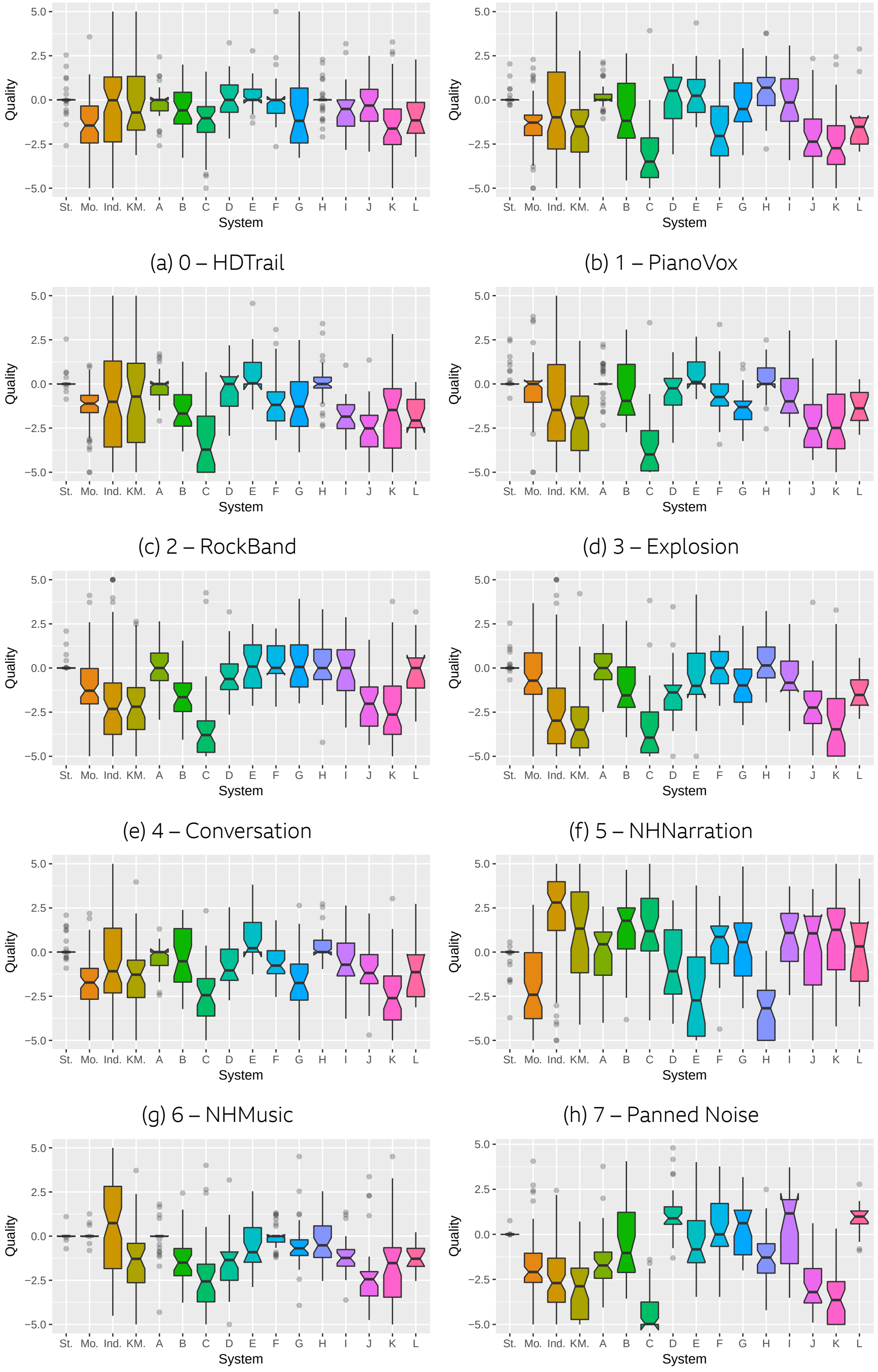


Figure 4.6: System quality ratings over all broadcast video items.

Boxplots for each system across broadcast programme items are shown in figure 4.6. Boxes extend between the first and third quartiles of the data distribution, whilst the whiskers extend to the value no further than 1.5 times the interquartile range from the end of the box. Data beyond the end of the whiskers are considered outliers and are plotted individually. The notches extend 1.58 IQR/$\sqrt{n}$, where IQR is the interquartile range, which gives an approximate 95 % confidence interval for comparing medians, as described by McGill et al. (1978).

Boxplots for the system ratings are shown individually for each audio item in figure 4.7, with the broadcast video items shown in figures 4.7a to 4.7g. It is clear that the performance of systems varied according to audio signal content, which is in agreement with the significant *System*:*Item* interaction obtained. Since there are similar patterns to the ratings for broadcast video items across systems, this justifies also presenting the results aggregated across items. It can be seen that item 0 (HDTrail) shows smaller differences between systems compared to other items. Many participants commented that this item was challenging to grade, it involved a dense arrangement of sound effects. Another notable result is for item 5 (NHNarration), where the systems with strong room response characteristic were rated particularly low, including the individualised system. This clip prominently featured the voice of a well-known narrator. Timbral colouration of this voice was clearly audible for these systems.

The similarities and differences between the systems across the range of broadcast audio items are of interest. An agglomerative hierarchical cluster analysis (HCA) was performed (Husson, Josse, et al., 2010), representing each system by a vector of its mean rating for



(a) 0 – HDTrail

(b) 1 – PianoVox

(c) 2 – RockBand

(d) 3 – Explosion

(e) 4 – Conversation

(f) 5 – NHNarration

(g) 6 – NHMusic

(h) 7 – Panned Noise

(i) 8 – Speech

(j) 9 – Applause

Figure 4.7: Boxplots of system quality ratings for each audio item.

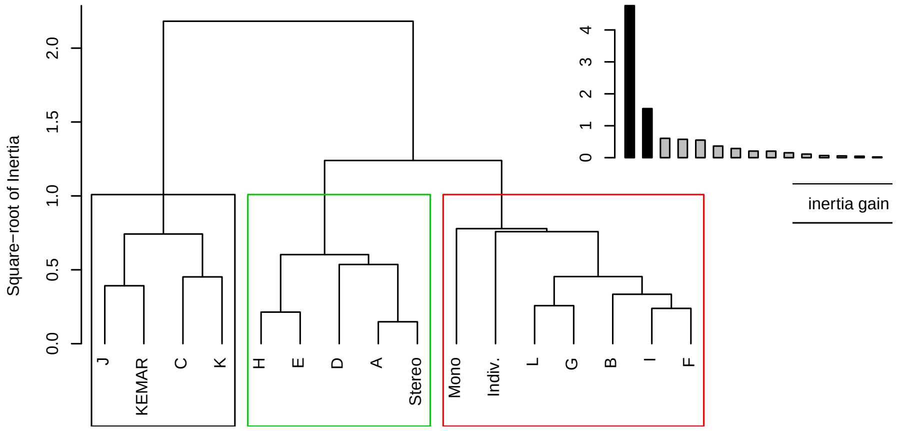


Figure 4.8: Hierarchical clustering of systems based on mean rating for each audio item.

each audio item. This clustering is shown in a dendrogram in figure 4.8. A Euclidean distance was used as a similarity measure and Ward's minimum variance criterion was used for aggregation of clusters (Ward, 1963). This aims to minimise the total within-cluster variance by iteratively merging the clusters that have the lowest squared Euclidean distance between their centroids. This process was performed using the `HCPC` method of the FactoMineR package in R (Lê et al., 2008). Three clusters of systems were defined, introducing additional clusters resulted in only a small reduction of within-cluster variance (inertia gain).

4.6.2.4 Discussion

The individualised head-tracked BRIR system was rated lower than the ITU stereo down-mix, and not significantly higher than the mono down-mix. This indicates that spatial image quality is not a dominant factor in the quality judgements of the assessors for broadcast video items.

The poorest performing systems overall apply a clearly audible room response which colours the sound. This can be seen in the system clustering, where the commercial systems C, J, and K are associated with the KEMAR BRIR rendering approach. Informal listening shows that these systems do provide reasonable spatial impression, with a sense of externalisation. It has been shown previously that timbral attributes are more dominant than spatial attributes in listener quality judgement when using the MUSHRA method (Rumsey, Zieliński, Kassier, et al., 2005a). It appears that the timbral effects of BRIR rendering often have a negative effect on overall quality.

The systems with the highest ratings over all broadcast video items were clustered with the stereo down-mix signal. Whilst system A appears very similar to stereo across items, the others vary more. They appear to perform well on music signals, but D and E do less well on those items with prominent speech. Informal[4] listening showed that these systems sounded similar to the stereo down-mix, with only subtle spatial image enhancements. Systems E and H both appeared to introduce an ambience enhancement. Interestingly these systems are rated quite poorly for the Panned Noise item, as discussed in the following section.

The final cluster contained the remaining systems, which have intermediate grades when viewed across all items. This cluster clearly includes a range of characteristics since it contains both the mono down-mix and individualised binaural systems.

4.6.3 Audio-only items

Three audio items without accompanying video were used in the experiment and the results are discussed separately here. Boxplots of the system ratings for each of these three items are shown in figures 4.7h to 4.7j. Alongside discussion of the results, observations from informal critical listening to the stimuli are discussed.

4.6.3.1 Panned Noise item

The movement of the noise source in item 7 was designed to make clear the spatial imaging effect of virtual surround systems. Whilst changes to tone colour between stimuli were obvious with this signal, listeners are likely not to have a strong internal reference for the timbre of this artificial signal, compared with a musical instrument or speech signal. Listeners commented that stimuli showed more obvious differences in spatial image for this item.

The ratings for this item are shown in figure 4.7h. The dynamic individualised BRIR system was rated significantly higher than the reference for this item, as were systems B, C, I, F, K, and KEMAR. The mono down-mix was rated lower than for other items; the loss of spatial information was particularly noticeable in this instance. Most systems had significantly higher ratings for this item than across the broadcast programme items. However, systems E and H did not. Both of these showed unpleasant artefacts of reverberation/ambience enhancement. These artefacts were not audible in denser more complex scenes, commonly used in broadcast material.

It is worth noting that the source movement in the stereo reference is different to the original 5.1 input signal, and that of the virtual surround stimuli. In the stereo down-mix, the

[4] Informal in the sense that it is not evaluated based on a formal controlled and statistically-designed experiment.

noise source will likely move along a line approximately between the listener's ears. Instead virtual surround sound processing aims to create a source moving around the listener. The wide confidence intervals indicate that there was a lack of agreement across the listeners regarding quality in this instance.

4.6.3.2 Speech item

Item 8 was chosen because dialogue and narration in the front centre channel are important in broadcast programmes. It allows assessment of this aspect in isolation. The ratings for this item can be seen in figure 4.7i. No systems were rated significantly higher than the stereo reference. The dynamic individualised BRIR system has a median rating that is higher but has wide confidence intervals.

A change in source distance was observed between systems but also changes in timbre and speech intelligibility. The ITU stereo down-mix will lead to in-head localisation and minimal colouration in this case. Some systems seem to compensate for the challenges of frontal localisation by reducing the binaural processing on the front centre channel to get lower colouration, these systems were rated similar to the stereo down-mix on this item. Systems that add strong reverberation character again performed poorly, except the individualised system.

Listening during item selection showed that only the individualised head-tracked system was able to create stable frontal externalisation for this author and several colleagues. Frontal sources are often challenging in binaural rendering, where interaural differences are small. Front-back reversal (Wightman and Kistler, 1989b) and in-head localisation (Hartmann and Wittenberg, 1996) problems are common. This may be influenced by increased reliance on pinna spectral cues for localisation in this frontal region (Searle et al., 1975). It has been shown that externalisation is greatly improved by use of room impulse responses (Völk et al., 2008), though these should correspond with the listener's real-world environment for best effect (Neidhardt and Knoop, 2015). Head tracking (Hendrickx et al., 2017) and correspondence of visual cues also have a strong influence on externalisation (Udesen et al., 2014). Differences in listener preferences may exist when considering internalised or externalised dialogue, this appears evident from the range in ratings for the individualised system on this item. Preference will likely also depend on narrative context of programme content, which is lacking for this simple test item. This becomes an issue at the communication sciences level of quality judgement, where context is an important influence (Blauert and Jekosch, 2012).

4.6.3.3 Applause item

Item 9 was included to assess the ability of systems to create foreground content behind the listener, which cannot be reliably achieved with a stereo down-mix. The applause signal is also useful for exposing the transient response of systems and ability to render a rather diffuse scene. It is thought that timbral and temporal quality features may have dominated the quality judgement in this case. Colouration was found to be critical for this item during test material selection.

Figure 4.7j shows the ratings for this item. Systems with strong reverberation were again rated low; particularly the cluster of C, J, K, and KEMAR, but also the individualised system. The individualised system was rated low for this item compared with its rating for other items. The mono down-mix was also rated low on this item relative to others. This is partly due to the loss of all spatial information but also comb-filter colouration appeared to be introduced in the down-mixing process. Systems D and L were rated higher than the reference for this item. Informal listening showed that these systems introduced relatively little colouration whilst providing a rear-surround spatial effect.

4.6.4 Clustering of assessors

To further investigate the variation in assessor's opinions of quality, clustering of the assessors was performed (Silzle, Neugebauer, et al., 2009). A k-means clustering algorithm was applied, representing each assessor by a vector of rating data for the dynamic individualised BRIR system. Clustering based on data for all systems was not possible due to the unbalanced test design, each assessor rated only a sub-set of the systems. The process was performed with $k = 2$ clusters. Since the algorithm uses a random initial partitioning, 100 runs were performed and the clustering with lowest sum of squared Euclidean distances from points to cluster centres was used. The algorithm of Hartigan and Wong (1979) was used via the R function `kmeans`.

The distribution of ratings over all broadcast items for each assessor cluster are shown in figure 4.9. Systems D, J, and L show significant improvement in median quality for assessors in cluster 2 when compared with the distribution of all assessors' ratings. Similarly a significant degradation in quality is seen for assessors in cluster 1 on these systems. The KEMAR and the dynamic individualised BRIR systems also show this pattern. For assessors in cluster 2, the individualised system is rated much higher than the stereo reference with a median of 1.50, indicating that a proportion of the population may have a preference for virtual surround processing with a strong spatial impression. Still, no system besides the individualised one had ratings significantly greater than the stereo reference for cluster 2.

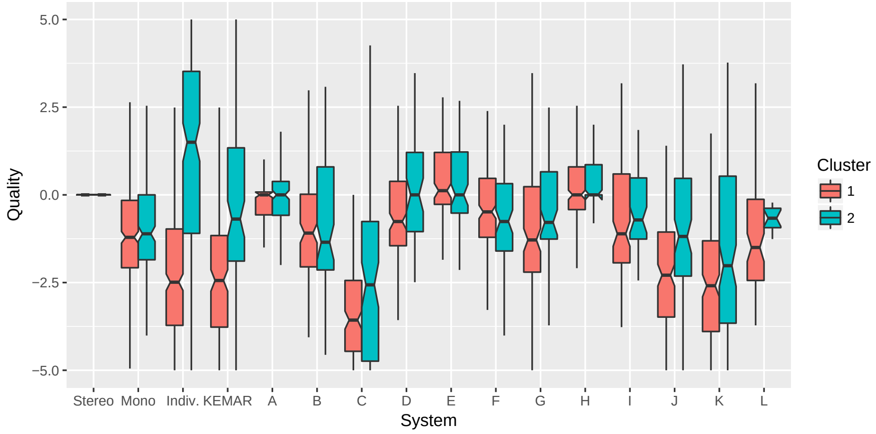


Figure 4.9: Separate distributions of system quality ratings for two assessor groups, resulting from k-means clustering, using broadcast video items.

Cluster 2 represents 35.1 % of the sample population of listeners, which is similar to the proportion of "binaural lovers" found in (Silzle, Neugebauer, et al., 2009).

A clustering with $k = 3$ was also performed, which yielded a primary mid-range group of 51.3 % and then two smaller groups of "binaural lovers" and "timbre lovers", who rated the dynamic individual BRIR system high and low respectively. However, this resulted in a small cluster of only four assessors with higher ratings for the individual system, so the cluster rating distributions are not plotted here.

4.7 Discussion

In general, the headphone virtual surround sound systems did not show a clear improvement over a stereo down-mix when using 5.1 broadcast material. Many of the systems showed a significant degradation of quality. This is in line with similar earlier experiments, it appears that there was no significant advancement in the state-of-the-art during this time. It is important to understand why this binaural processing does not improve quality.

4.7.1 Why does Headphone Surround Sound Processing not Improve Quality?

From the literature review in chapter 2, it can be expected that the spatial effect of these systems is limited without use of individual HRTFs or head tracking. It was noted in informal listening that systems with poor externalisation were felt to narrow the spatial image compared with the stereo down-mix. In stereo headphone listening, the spatial image is spread laterally between the ears, whereas some binaural systems appeared to have reduced width

(±30°) without much frontal imaging. However, even the individualised head-tracked BRIR system was graded lower than the stereo down-mix. This was somewhat unexpected since in informal listening it was found to create a realistic simulation of the loudspeaker reproduction. However, compared to direct stereo rendering, there was a noticeable change in timbre. This result suggests that without the reference of the real loudspeaker system, the perceived quality of this rendering is quite different, and that emulating loudspeaker reproduction in the current listening environment might not lead to optimal quality of headphone reproduction.

It was reported in section 3.3 that Rumsey, Zieliński, Kassier, et al. (2005b) found listener quality ratings to be dominated by timbral character over spatial character. Timbral colouration is a common problem in binaural rendering (Merimaa, 2009). It appears that the quality ratings given in this experiment reflect this and may have been dominated by the timbral characteristics of the stimuli. This is highlighted by the fact that several systems were graded lower than the mono down-mix and the individualised system was not rated highly. Clearly, spatial characteristics also had an influence. The mono down-mix was rated lower than stereo for example. The results for the Panned Noise item also show that binaural processing can have a positive influence on quality, when spatial movement is important and the internal timbral reference of the sound is weak.

It is assumed that the surprisingly low quality of the individualised head-tracked system is due mainly to timbral quality degradations. Undesirable characteristics could be introduced by a number of factors, including the measurement room, the measurement microphones (part of the Smyth Realiser system), the loudspeakers, and imperfect headphone equalisation. Some of the systems under investigation will likely have been developed to minimise such factors, to preserve timbral fidelity. This may sometimes have been at the expense of spatial quality. The use of non-individual HRTFs will also have introduced some colouration in these systems. The individualised head-tracked system was graded significantly higher than the KEMAR system, demonstrating that individual measurements and head-tracking provided better quality.

If it is assumed that overall grades across broadcast video clips are dominated by timbre and grades for the Panned Noise clip are dominated by spatial impression, then there appears to be a negative correlation between colouration and spatial enhancement. This may in part be attributed to use of a room response in the algorithms. Environmental reflections and reverberation appear important to create a convincing spatial impression (Begault, Wenzel, and Anderson, 2001), but it can also impact timbre heavily (Rubak and Johansen, 2003). The systems graded lower than the mono down-mix overall (systems C, J, K and KEMAR) show strong colouration with audible reverberation, but these performed well on item 7

(Panned Noise). Some systems showed audible reverberation but were not graded as low (systems B, D, E and H). The colouration effects of the specific reverberation used must be carefully considered it seems.

Systems E and H both appeared to apply ambience enhancement techniques on the input signals, or else added only a diffuse reverb characteristic, seemingly without the early reflections that often cause pronounced timbral changes. An example of such processing is described by Faller and Breebaart (2011). These systems showed relatively good performance on broadcast material. Though median ratings were zero, the distribution was positively skewed. They both gave only a subtle spatial enhancement, with an increased sense of envelopment but lacking a strong spatialisation of foreground sources. In contrast, other systems appeared to use room auralisation during the loudspeaker virtualisation, e.g. via measured BRIRs. Early reflections have been shown to help to create well localised and externalised sounds (Begault, Wenzel, and Anderson, 2001). These ambience enhancement algorithms showed artefacts on the Panned Noise item (item 7), resulting in poor grades, but not on the isolated speech signal (item 8). Whether artefacts can occur with more typical broadcast signals for these systems is a question worthy of further investigation.

Source material has a strong effect on the quality of these systems. Adaptation of the rendering techniques to the signal content might enable improved overall quality, though this is not compatible with the stated *black box* approach for the target broadcast distribution application of this study. In certain combinations of system and item, quality was rated positively over the down-mix. For the simple audio-only items, significant quality enhancements were achieved. It may be that different rendering techniques are required for different scene components of content to achieve good results. It also seems likely that spatial enhancements are more difficult to perceive in a complex scene.

Further insight into the characteristics of each system that listeners experienced would be beneficial in understanding the factors influencing quality and how the quality of systems might be improved. Interpretation from the obtained results can only be limited. Discussion in terms of spatial and timbral quality is a simplification of the multidimensional nature of the quality of these systems. Section 3.3.3 introduced descriptive analysis methods for characterising quality. Lorho (2005a) used this approach to gain insights into the quality of stereo enhancement systems. That study found timbral artefacts such as lack of clarity and dark tone colour, and a trade-off between spatial depth offered by reverberation and the unnatural colouration introduced. Similar effects were observed by the author when listening to the range of systems in this study. Descriptive analysis can provide further insights, though ultimately it is the overall quality of experience for the target audience in the appropriate context of use that matters most.

It is worth considering that listeners will have much more experience on headphones of stereo sound than of binaural signals. Expectations of sound character when listening on headphones will be influenced by this and it may partially explain why accurate simulation of external loudspeakers is not necessarily preferred on headphones. This cannot be avoided, especially with typical audience members who might have less experience of binaural signals than spatial audio researchers. However, the use of the stereo down-mix as a declared point of reference or comparison in this experiment may have over-emphasised bias towards the character of the stereo signals in terms of the quality judgements.

Shortly after this experiment was carried out, Iljazovic et al. (2012) published an investigation of the quality of headphone surround sound processing and its interaction with use of 2D or 3D video formats. The video condition was evaluated using a between-subjects design. Multiple stimulus presentation was used as in this study. A key difference between their study and this study was the use of a given reference. Iljazovic et al. did not present a reference and quality ratings were on an absolute scale. The stereo down-mix was instead included as any other stimulus, allowing comparison still to be made. A range of different entertainment content was used, including movies, music, and documentaries. When pooled over both video conditions and an audio-only condition, two of the three different virtual surround algorithms had mean ratings significantly higher than a stereo down-mix. However, the differences were small, less than 10 points on the 100-point ITU continuous quality scale. All systems were rated in the region of *good* quality.

The interaction effect of an accompanying picture has not been investigated in the present study. Interestingly, Iljazovic et al. (2012) found that, compared with the case with no video, adding 2D video improved perceived audio quality whilst 3D video degraded it. The presence of associated pictures will have an effect on interpretation of the auditory scene. The coherence between auditory and visual scenes will likely impact quality (Pike and Stenzel, 2017)[5]. A challenging aspect in this regard is the impact of visual display size. In this study a wide screen television display was used, however it is more likely that headphone listeners will be viewing on mobile devices with smaller displays.

The duration of the source material used in the present study was quite short (about 12 s) and the multiple stimulus presentation method meant that assessors were able to switch frequently between stimuli. Recently, Schoeffler and Herre (2016) found that expert assessors showed more influence of timbral quality when rating quality using a multiple-stimulus basic audio quality (BAQ) rating than with a single-stimulus overall listening experience (OLE) rating. The response attribute used in the experiment presented in this chapter (*overall sound quality*) may also lead to different weighting of quality features to an alter-

[5] Note that the first author of this study is not the present author but Cleopatra Pike.

native such as *OLE*. The presentation method and response attribute should be carefully considered for quality evaluation of spatial audio; more study of their influence on system ratings is needed. The current test method may have biased listener judgements towards timbral aspects. However, multiple stimulus presentation has been shown to give more reliable ratings (Beresford et al., 2006; Schoeffler, Silzle, et al., 2017).

Judgements varied significantly between individuals. There seem to be different preferences within the listener population. Clustering suggests that some listeners may prefer the spatial enhancement offered by some of the virtual surround systems, whilst others may dislike the timbral effects introduced. The audience could be given a choice between virtual surround and stereo versions to account for this. However, even in the listener group that favours spatial enhancement, most systems are graded no better than the reference. To justify additional configuration options in an application, clear benefits to at least a subset of users must be demonstrated.

It was shown that the listening room environment had no significant effect on the quality of systems, this follows earlier findings of Scott and Roginska (2008). Only small effects were found for the individualised system, where BRIRs measured in the controlled listening environment were used in both rooms during evaluation. Recent research has shown that so-called *room divergence*, i.e. incongruence between the virtual acoustic environment (VAE) and the physical environment, does have an influence on externalisation (Werner, Klein, et al., 2016). It appears that in this case there was not much influence on quality judgements, however the rooms were quite similar to one another compared with those typically in such studies. The influence of room divergence on quality needs further research.

4.7.2 Different Applications of Binaural Technology to Media Distribution

This study specifically evaluated the application of binaural techniques to virtual surround sound processing of 5.1 surround signals for headphones. This presents a method of applying binaural technology to the broadcast distribution process with minimal changes to infrastructure and workflows. It is however quite limited, only providing horizontal spatial information and with low resolution, particularly outside of the frontal region.

Virtual surround systems that offer client-side rendering with adaptable parameters did exist in 2012, such as the binaural decoding in the MPEG Surround codec system (Breebaart et al., 2006). However, such systems were not widely availability and their use would have required transmission of the content using a new codec, which have involved a substantial change to the prevailing practices. Near-term use of virtual surround for headphones requires a distributer-side rendering. The effects of distribution codecs on sound quality are not considered here and would require subsequent testing. Previous studies have investigated this aspect, using HRIR rendering of the uncompressed multichannel signal as a reference, and shown degradations to timbral and spatial quality (Breebaart, 2007; Le Bagousse, Paquier, Colomes, and Moulin, 2011).

The requirement for the investigation in this chapter was that systems have a fixed parameter configuration, without the freedom to make adjustments according to the audio signal. Some of the systems under investigation have many adjustable settings. In-depth investigation into settings of each system may give improved results, though this is expected to have been done by the manufacturers who provided the settings used here. There may be a system that can deliver the flexibility needed, which can be adapted to achieve appropriate rendering of a range of audio material, but further investigation of system parameters would be required. Any manual parameter adaptation would need to be straightforward for users.

Another step to obtain improved results would be to modify the original mix in the 5.1 format, as well as adapting virtualisation parameters. However, with this degree of manual tuning, more benefit is likely to be achieved from dedicated mixing in a binaural format. That the quality of systems varied between different audio items suggests that different rendering approaches may be appropriate for different scenes or for components within a single scene. A 3D binaural mixing system might permit this degree of flexibility e.g. the *Spatialisateur* (Carpentier, Noisternig, et al., 2015). For example, using only anechoic filtering on narration and music but applying virtual room acoustics to sound effects and dialogue in the scene. Such an approach would also allow the use of stereo panning where deemed more appropriate. This also allows for creative use of the distinction between in-head and out-of-head space.

These approaches vary in terms of the effort required in production, but could all be performed today with existing technology. The proposals are constrained by the fact that the end result would be a two-channel binaural signal, and so no individualisation or head tracking could be applied for the listener during rendering. It is not clear how much the quality can be improved without these elements of the system in place, further investigation is required. Since this study, formats for representing and distributing 3D audio scenes have become more widely established. These would allow such client-side rendering to be performed. Such formats are discussed in more detail in chapter 6.

4.8 Conclusion

An evaluation of 12 state-of-the-art commercial virtual surround systems showed no clear improvement over a stereo down-mix of 5.1 surround broadcast audio, with many systems performing significantly worse. An individualised head-tracked BRIR rendering approach was also evaluated and rated lower in quality than the stereo down-mix. It appears that realistic simulation of loudspeakers in the listener's environment may not be the optimal target for rendering 5.1 surround content to headphones. Although from this experiment, it cannot be said how realistic the spatial impression was to listeners.

The quality of headphone virtual surround sound processing has a significant dependence upon source material. A clustering analysis of systems revealed that a group of four systems was rated similarly to the stereo down-mix across the broadcast audio items. Another group of four systems was rated poorly, lower than a mono down-mix overall. From informal listening, the lower quality systems all introduced a clearly audible room effect. Whilst these appeared to give strong spatial impression, it is thought that timbral colouration might have dominated judgements.

There was a lack of consistency amongst assessors. Another clustering analysis suggested that 35 % of the assessors valued the spatial enhancement offered by virtual surround processing more than the remaining assessors. Such a trend may exist within the wider population of listeners. No significant differences in the quality of systems was found between the two listening environments used in this experiment.

Different applications of binaural rendering to the production and distribution of entertainment media may show greater benefits in terms of quality. Setting rendering parameters according to the audio signal content and using 3D audio formats will likely lead to improvements. Descriptive analysis experiments could give insight into the characteristics that are relevant for listeners' quality judgements. This could aid understanding and drive improvements to the design and application of binaural technology. An improvement in overall quality for listeners over current stereo services is ultimately still required to validate use of these techniques however.

Chapter 5

Evaluating The Plausibility of Non-Individualised Dynamic Binaural Rendering

This chapter presents an assessment of the plausibility of binaural rendering that can be achieved using a non-individualised approach. This includes the presentation of a state-of-the-art dynamic binaural rendering system, followed by a review of methods for validating the spatial impression given by an auditory virtual environment. A listening experiment is presented, based on signal detection theory, to determine whether the binaural rendering gives simulation in agreement with listeners' expectations of real sound sources in the listening room.

5.1 Introduction

With the prevalence of mobile computing devices today and an accompanying increase in headphone listening, the application of binaural techniques to entertainment media systems is popular. This is demonstrated by the large number of commercial systems that were available for evaluation in chapter 4. The intention behind the application of binaural technology is to provide an enhanced spatial impression for the headphone listener. Evidence shows that binaural rendering can be used to create auditory events at locations outside of the head with well defined direction (Wightman and Kistler, 1989b) and distance (Zahorik, 2002a). In comparison, conventional headphone reproduction of two-channel stereo is known to create auditory events inside the listener's head (Plenge, 1974). Binaural processing therefore has the potential to create realistic three-dimensional (3D) auditory scenes. It is assumed that this can lead to more immersive and enjoyable listening experiences, which needs to be verified.

Despite huge research and development effort, binaural technology is not yet in everyday use for mainstream audiences of entertainment media[1]. There is currently limited evidence that binaural processing can give an improvement in the quality of the listening experience in entertainment media applications, when compared with stereo signals. At the outset of work towards this thesis, the most common approach in this application domain was headphone surround sound processing (HSSP), where 5.1 surround signals are rendered for headphones with binaural techniques. In recent evaluations, Iljazovic et al. (2012) and the study reported in chapter 4 both found that some HSSP systems gave improved sound quality over a stereo down-mix, when averaged across a range of audio content. It is difficult to interpret the results in absolute terms, since the use of the rating scale is rather specific to each study. In both cases however, the difference between the best-performing HSSP systems and a stereo down-mix appears small. In chapter 4 many of the evaluated systems were rated significantly worse than stereo in terms of sound quality.

It is important to understand why this is. The discussion in section 4.7, considers a number of possible reasons.

- Poor spatial impression given by the rendering.
- Negative impact on timbral aspects introduced by the rendering.
- Limited spatial information in the input audio format.
- Expectations of headphone sound character due to stereo listening experience.
- The influence of the experimental design.
- Variations in listener preferences.

Arguably the primary aim of binaural rendering is to create realistic externally-localised auditory events. An important first step to understanding why existing systems do not offer high quality is to identify whether sufficient spatial impression is afforded. In chapter 4, the system considered to be of highest spatial quality used head-tracking and individual binaural room impulse response (BRIR) measurements. It was discussed in section 2.5 that simple and effective methods for individualisation, suited to use with a mass audience, are not yet widely available. Therefore it is of interest whether a non-individualised system can provide sufficient spatial impression for entertainment media applications. A drawback of evaluating commercial systems is that the specific details of the rendering approach cannot be known and reported, which limits interpretation of the results.

[1] Though there has been significant growth in its popularity since the start of this project, see chapter 6.

In this chapter, a custom rendering system is first presented in section 5.2. It is based on existing open-source software with documented extensions, to achieve specifications in-line with the state-of-the-art. Section 5.3 then presents a review of methods for evaluating the spatial impression of binaural rendering systems, considering appropriate target measures of performance for entertainment media applications. Subsequently a listening experiment is presented in section 5.4 to validate the performance of the rendering system described in section 5.2. Section 5.5 presents conclusions on the study and then considers the outlook for further research of binaural systems in light of these findings and those of earlier work.

5.2 A Non-Individualised Dynamic Binaural Rendering System

A non-individualised dynamic binaural rendering system was created using BRIRs measured with a dummy head microphone at multiple orientations, as introduced by Horbach et al. (1999). Besides the constraint of using non-individual impulse response data, the system was designed considering the perceptual tolerances and techniques detailed in the literature, with the aim of achieving a state-of-the-art system. The system is described in brief here, a number of the system elements are described in more detail in appendix A. Methods for evaluating the spatial impression given by this system are then discussed in section 5.3 and the plausibility of the system is evaluated in the listening experiment described in section 5.4.

5.2.1 Equipment and Facilities

The listening room at BBC Research & Development was used to make the BRIR measurements, where $V = 99\text{m}^3$, $T_{60} = 0.21\text{s}$, and $r_{crit} = 1.22\text{m}$. The critical distance r_{crit} is the distance from a sound source at which direct and reverberant fields have equal energy. Small active two-way loudspeakers (Genelec 8030A) were used during BRIR measurements. Electrostatic headphones (STAX SR-202 with SRM-252II driver unit) were used to reproduce the binaural signals. These were chosen primarily due to their open design, with relatively little physical barrier to external sounds. An earlier collaborative study involving this author found this headphone model to have the lowest effect on localisation cues amongst several tested (Satongar, Pike, et al., 2015) (co-authored publication Co.P. III).

5.2.2 BRIR Dataset

A Neumann KU100 dummy head microphone was used to measure the dataset of BRIR measurements. The headphones were placed on the dummy head during measurement to ensure that the headphone effects on external sounds were captured. This is because the headphones were worn whilst real loudspeaker sounds were presented in the plausibility evaluation experiment. The same approach was taken by Lindau and Brinkmann (2012).

The dummy head microphone was mounted on a custom-made rotary mount (Shotton et al., 2014) (co-authored publication Co.P. I)., with a rotational accuracy of 0.01°. BRIRs were measured from multiple loudspeakers placed in the room using the exponential swept sine technique (Farina, 2007) with a sweep length of 2^{18} samples. The dataset was measured for the full range of horizontal head rotations [0°, 359°] with an angular resolution of 1°.

5.2.3 Rendering Software

A modified version of the SoundScape Renderer (Ahrens, Geier, et al., 2008) software was used to perform the dynamic data-based binaural rendering. This allows real-time multi-threaded processing of spatial audio signals, including an efficient uniform-partitioned fast frequency-domain convolution engine. The convolution engine allows dynamic filter updates with cross-fading to reduce switching artefacts. The built-in binaural room scanning renderer allows dynamic head-tracked rendering of an auditory virtual environment (AVE) using BRIR datasets. This renderer was modified so that only the early part of the BRIRs was dynamically updated according to head orientation and the late reverb tail was kept static, as illustrated in figure 5.1. This was done to reduce the memory requirements of the system, with knowledge that listeners are not sensitive to lack of changes in this diffuse tail (Lindau, Kosanke, et al., 2012). These modifications were implemented using components from the Audio Processing Framework (Geier, Hohn, et al., 2012), an open-source C++ library providing core digital signal processing (DSP) components for the SoundScape Renderer.

The mixing time between the dynamic early part of the impulse response and the static late part was set at 60 ms, with a 20 ms cross-fade duration. This was chosen based on a signal-based predictor of perceptual mixing time (Lindau, Kosanke, et al., 2012). The predictor used the echo density estimation algorithm of Abel and Huang (2006). Lindau, Kosanke, et al. used linear regression to predict the perceptual mixing time, as measured in listening experiments, from this echo density estimation. The used metric t_{mp95} represents the time at which the perceptual mixing time is reached for 95 % of listeners, assuming a normal distribution. The authors provide MATLAB reference code for predicting t_{mp95}, which was applied to a BRIR measured for the frontal loudspeaker.

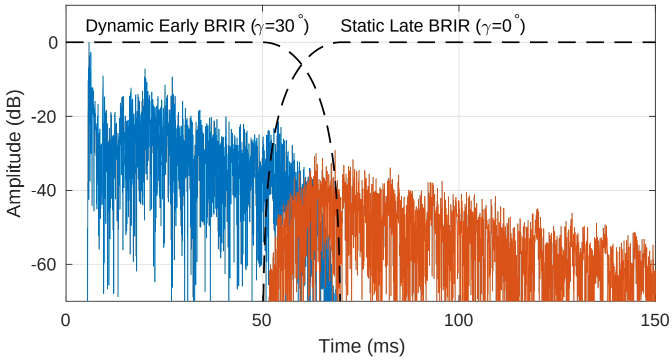


Figure 5.1: An example of splitting BRIR into dynamic early and static late regions by cross-fading after the perceptual mixing time. Here the early region corresponds to head orientation $\gamma = 30°$ and the late region, containing the diffuse reverberation tail, is fixed to use the measurement for $\gamma = 0°$.

5.2.4 Tracking and Latency

An optical tracking system (VICON Bonita) was used to detect head movements, with four cameras placed in the corners of the room and a rigid configuration of retroreflective markers attached to the listener's headphones. The tracking system is capable of measuring head rotation to an accuracy of 0.1° and was chosen due to its fast update rate (250 Hz), low processing latency (2.5 ms), long-term stability, and absolute coordinate reference. The tracking data was sent via TCP on a local area network to the rendering software, which was running with an audio buffer size of 256 samples.

The total system latency (TSL) was measured using the method of Lindau, 2009. 20 TSL measurements were made for this binaural rendering system. The mean TSL was $t_{\text{TSL}} =$ 41.2 ms ($\sigma_{t_{\text{TSL}}} =$ 2.6 ms). The maximum measured value was 47.9 ms, which is below the lowest subjectively detected TSL of 53 ms measured in Lindau, 2009.

5.2.5 Headphone Correction

A non-individual headphone-to-ear correction filter (HpCF) was applied to the BRIR dataset before rendering. Headphone-to-ear transfer function (HpTF) measurements were made on the same KU100 dummy head used in the BRIR measurement process. It was found by Lindau and Brinkmann (2012) that non-individual HpCF, when measured on the same head as used in the BRIR measurements, enabled a more realistic binaural simulation than individual HpCF measured on the listener, when comparing to a real loudspeaker source. Whilst this finding seems surprising and warrants further investigation, the requirement in this study is to use a non-individualised system. Lindau and Brinkmann (2012) showed that this approach is likely to give better results than a generic HpCF created from an average of multiple listeners' HpTF measurements.

The HpCF was calculated using the method described in Masiero and Fels, 2011, which attempts, in a perceptually robust manner, to account for variability in the HpTF due to different positioning of the headphones. Ten HpTFs were measured on the dummy head, with removal and replacement of the headphones each time. The impulse responses were truncated to a length of 2048-samples. The magnitude responses of the HpTFs were smoothed with 1/6$^{\text{th}}$ octave resolution. The mean plus two standard deviations of these magnitude responses was then used for inversion, in order to avoid overcompensating for notches that might not be present in the HpTF at the time of reproduction. This approach was chosen over the methods described by Lindau and Brinkmann, 2012 because it does not depend upon the tuning of a regularisation parameter by expert listeners.

5.2.6 Scaling Inter-aural Time Differences

Whilst the BRIR dataset used in the binaural rendering system was not measured for the individual listener, the impulse responses were post-processed to allow scaling of interaural time differences (ITDs). The purpose of this process is to reduce instability in the azimuthal direction of auditory events during horizontal head rotation, since the ITD is a primary cue for localisation of broadband sounds in this plane (Busson et al., 2005). When the ITDs are incorrect, the source position tends to drift either in the same direction as the listener's head moves or in the opposite direction.

A logarithmic-threshold method was used to detect and remove the onset delays in the impulse responses, as described in Lindau, Estrella, et al., 2010, with a threshold of 15 dB below the maximum value of the impulse response. These onset delays could then be scaled and reinserted at the start of the impulse responses to yield a change in the ITDs; subsample accuracy was achieved with oversampling by a factor of 10. The dataset was modified offline, using a range of different ITD scaling factors (range: [0.75,1.25], step size: 0.025), with the BRIRs for each scaling stored separately in the SoundScape Renderer's required 720 channel WAV file format. A control interface was built to allow a listener to select one of the available pre-processed ITD scalings, to find the value that best stabilised the direction of a sound source.

5.3 Evaluating the Spatial Impression Given by Binaural Rendering Systems

Whilst the aim of this thesis is to improve the overall listening experience, evaluation of listener preference, or an integrative rating of the quality of the experience, cannot directly explain why the experience is not improved. The perceived character of the experience with a binaural rendering system is not described, and so the quality of the spatial impression cannot be separated from the quality of the overall experience.

5.3.1 Rating Spatial Impression

Descriptive analysis (DA) and preference mapping techniques may be used to characterise quality (see section 3.3.3). This has led to assessment of quality features such as *sense of space* (Lorho, 2010), or *auditory spaciousness* (Silzle, 2007). These quality features appear to be multi-dimensional in nature and rather loosely defined. Also, in DA experiments such as these, continuous rating scales are often used, with verbal descriptors only at the scale end-points. Attribute interpretation and scale usage are likely to be subjective.

When evaluating virtual environments, either auditory or audiovisual, often assessors are asked to rate the sense of *presence*, and/or *realism* of the experience (Hendrix and Barfield, 1996; Väljamäe et al., 2004). These are high-level quality features, which are influenced by many factors, not only the reproduction technology. The level of interactivity offered within the environments and the complexity of the stimuli will play a part too.

Slater (2004) discussed that commonly-used *presence* questionnaires give no real evidence of the experienced phenomena. The questionnaire data would need to be verified with identifiable brain activity or behavioural responses that are known to relate to presence. This is demonstrated by an invented phenomenon called "colorfulness of an experience", where questionnaire participants were asked to rate the *colorfulness* of the previous day. From a set of supplementary questions, influencing factors were related to *colorfulness* through a regression model, as is often done with *presence*. The author states that "in this process the degree of experienced 'colorfulness' was brought into being only by asking about it—having no predictive or explanatory power, and no utility in itself." Questionnaire respondents had attributed meaning to this attribute although there is no evidence that it existed as a feature of their experience. Slater concludes by arguing that it is not the concept of presence that is flawed, but methodologies for evaluating it by eliciting direct judgements.

5.3.2 Externalisation

In terms of localisation of sound events, experiments have shown that free-field equivalent directional localisation can be achieved with individual measurements (Martin, McAnally, and Senova, 2001; Romigh, Brungart, and Simpson, 2015), but non-individualised rendering leads to increased errors (Middlebrooks, 1999b). Often, however, localisation studies have not considered the perceived distance of auditory events, or evaluated the associated concept of externalisation. It seems highly important to the success of binaural rendering that it can trigger auditory events outside of the listener's head (Begault, 1991).

Durlach, Rigopulos, et al. (1992) presented a discussion on externalisation. It is clearly a subjective process, and also a matter of degree: events can range from well inside the head, to the boundary between inside and outside the head, to well outside the head. Durlach, Rigopulos, et al. (1992) described it as a "crude representation of subjective distance." It is related to the distance of auditory events, though with some degree of quantisation. Determining when an event is outside of the head is a challenging and unusual task, for which listeners will develop subjective decision criteria. Durlach, Rigopulos, et al. also discussed that externalisation relates to our expectations of how natural external events behave. This leads to suggestions that reverberation, head movements and visual cues are all likely important for externalisation, alongside correct pinna cues. Furthermore, it is suggested that the process is influenced by prior experience, so learning effects are likely.

Hartmann and Wittenberg (1996) found that correct spectral magnitude cues are required for externalisation during static free-field rendering. The positive effects of head tracking and environmental reflections on externalisation have been confirmed experimentally, e.g. by Hendrickx et al. (2017) and Zahorik (2000, 2002a), respectively. These studies are reviewed in more detail in chapter 2. Mendonça, Campos, et al. (2013) also demonstrated increases in reported externalisation with non-individual head-related transfer functions (HRTFs) through increased exposure over time.

Werner, Klein, et al. (2016) presented two experiments that investigated the factors influencing externalisation. Both individual and KEMAR BRIRs were used in a static (non-head-tracked) binaural rendering system. It was shown that the degree of externalisation is increased by visual cues of the room and dummy loudspeakers at the target location of virtual sound sources, compared with when listeners were blindfolded. The experiments also demonstrated the influence of room divergence, where the virtual acoustic environment does not match the real environment in which the listener is situated. The degree of externalisation is negatively affected by room divergence, but experience with room-divergent rendering leads to improvements in externalisation with time. Additionally, use of individual

BRIR measurements showed increased externalisation over the KEMAR BRIRs, particularly in the region near or in the median plane.

In evaluating externalisation, assessors have been asked to report perceived distance, with the low-end of the scale taken to indicate in-head localisation (Zahorik, 2000; Begault, Wenzel, and Anderson, 2001). The conflation of externalisation and distance is problematic. Hartmann and Wittenberg (1996) point out that the direct-to-reverberant energy cue can be used to estimate source distance in a signal, even though the auditory event may not be localised externally, e.g. by applying artificial reverberation to a single-channel recording and presenting it diotically. In other studies, externalisation is reported using an ordinal scale, with explicit reference to in-head or out-of-head localisation and states in-between, e.g. Hendrickx et al. (2017). The degree of precision with which the event can be localised is also sometimes incorporated; Hartmann and Wittenberg (1996) and Werner, Klein, et al. (2016) both incorporate the notion of diffuseness at some levels of their externalisation scales. This appears to make the percept multi-dimensional. As with the sense of presence, it is difficult to know what each listener has experienced when externalisation is reported. But it appears from prior experiments that, as stated by Durlach and Colburn (1978), “it increases as the stimulation approximates more closely a stimulation that is natural”.

5.3.3 Discrimination and Authenticity

The above described methods of evaluation have limitations. An individual will establish criteria for judgment, which will be subjective and dependent on context, experience, and listener state (see section 3.2). Concrete interpretation of the meaning of individual responses, in terms of the phenomena experienced, is not possible.

The experiments of Hartmann and Wittenberg (1996) required listeners to judge whether sounds were emitted by a real loudspeaker or a pair of headphones, using individualised anechoic binaural rendering. After each sound event the listener had to identify whether it was “real” or “virtual”. This process avoids the requirement for listeners to directly judge the externalisation percept in the process of forming subjective decision criteria. If the listener believes the sound came from the loudspeaker then it is assumed that the auditory event must have been externalised. Since the listener is exposed to the real loudspeaker during the experiment, including in a training process, their judgement is informed by reference to recent experience. This relates to the goal of *authenticity*, requiring the simulation to be perceptually identical to reality, as discussed by Blauert (1997) and Novo (2005). However, in analysis of the results, Hartmann and Wittenberg only compared the mean percentage of correct answers across listeners to a 75 % detection threshold. The statistical significance of findings was not analysed and it is not possible to identify the propensity for listeners to answer one way or the other, i.e. to determine their subjective response bias.

Authenticity has also been evaluated by asking assessors to directly discriminate between stimuli, identifying the odd one out in a set of three or four presentations (Langendijk and Bronkhorst, 2000; Moore et al., 2010; Brinkmann, Lindau, and Weinzierl, 2014; Oberem, Masiero, et al., 2016). Langendijk and Bronkhorst (2000) compared three methods of assessing the authenticity of binaural rendering: a single-interval two-alternative forced choice (2AFC) method (real (R)/virtual (V)) in line with Hartmann and Wittenberg (1996), a two-interval 2AFC method (either R-then-V or V-then-R), and a four-interval 2AFC "oddball" method (RVRR, VRVV, RRVR or VVRV). In the oddball method the listener must identify whether the second or third stimulus was different to the others. The experiment results were compared using the distribution of listeners' percentage of correct answers. For the oddball method the distribution was significantly above the 50 % correct (pure chance) level, though only slightly (mean of 53 %). For the other two methods the responses were not significantly different from chance. This suggests that the oddball method possesses greater sensitivity.

The goal of *authenticity* is overly strict for many applications, where there is no external reference for direct comparison. At the time of preparing the experiment described later in this chapter, previous experiments evaluating authenticity had been in anechoic conditions with no head movement and in-situ measurement of individual HRTFs and headphone equalisation. Individualisation is not yet feasible in mass-audience applications. Dynamic virtual environment simulation, with reverberation and head tracking, has been shown to improve spatial impression; these techniques are likely important in many entertainment media applications and are also more practicable. Without individualisation, it appears unlikely that authenticity can be achieved. A more appropriate goal is required, one that validates the spatial impression afforded whilst also better suited to the application context.

5.3.4 Assessing Plausibility

Lindau and Weinzierl (2012) used an alternative goal of *plausibility*, defined as

> …a simulation in agreement with the listener's expectation towards a corresponding real event.

This may be judged by a "real or virtual?" style of question, as in previous studies. However, listeners compare to an internal reference for the character of a real event, rather than direct comparison to a real event as in *authenticity* judgements. This internal reference, e.g. the listener's expectations of the character of a real sound source, is informed by the listener's memory of past experience and their current state. For more detail see chapter 3.

This goal is better suited to evaluation of non-individualised rendering systems. Differences in perceptual character to a real event may exist, provided that the expectations of the listener are still met. As with authenticity, there is little ambiguity in the task, unlike ratings of *immersion* or *spaciousness*. If a binaural rendering system is capable of creating plausible auditory events, by the above definition, then it can be assumed that the spatial impression is adequate for many entertainment media applications, and clearly enhanced over stereo headphone reproduction.

Lindau and Weinzierl (2012) introduced a method for assessing the plausibility of binaural rendering in a subjective evaluation. Participants were presented with a stimulus which could either be a real loudspeaker or a binaurally-simulated loudspeaker, and asked whether the stimulus was simulated, giving a Yes/No answer. Using a signal detection theory (SDT) approach to analysis, the sensory difference between reality and the simulation could be separated from the participants' response bias. This is discussed further in the following section.

Lindau and Weinzierl evaluated two head-tracked binaural rendering systems using BRIRs measured using a head and torso simulator (HATS). Although the rendering was not individualised, the more advanced of the two systems was shown to have a very high degree of plausibility.

These AVEs were simulating loudspeaker sources in a large auditorium with a minimum source distance of 9.5m. HSSP systems used in entertainment media commonly simulate a small listening environment, similar to a domestic living room. It is thought that plausible simulation of sound sources in a smaller listening environment may be more challenging. This is because sound sources will be closer to the listener and there will be less reverberation, so spatial and timbral distortions may be more easily detected.

5.3.5 Signal Detection Theory Methods

SDT enables separation of sensory differences from the subjective response bias. It was originally applied to perceptual studies for detecting a signal in noise (Green and Swets, 1966). A good overview of methods for applying SDT in perceptual experiments is given in Stanislaw and Todorov, 1999. SDT has been used previously in the evaluation of total system latency (TSL) detection thresholds in dynamic binaural rendering systems (Yairi et al., 2007). Langendijk and Bronkhorst (2000) also discussed it briefly as a means of comparing results of authenticity tests that have used different methods.

		Signal	
		Present	Absent
Response	Yes	Hit	False Alarm
	No	Miss	Correct Rejection

Figure 5.2: Trial outcomes for a yes/no paradigm

For a yes/no task, a trial outcome may be in one of four categories, as shown in figure 5.2. In this context, the *signal* is the binaural rendering system or simulation and the assessor gives a yes/no *response* to the question: did the sound come from the headphones? When the stimulus is binaurally-rendered via the headphones (signal is present), the assessor either detects it correctly and responds *yes* (hit) or fails to detect it and responds *no* (miss). When the stimulus comes from a loudspeaker (signal is absent), either the assessor correctly identifies this to be the case and responds *no* (correct rejection) or incorrectly perceives it to come from the headphones and responds *yes* (false alarm). The assessor must base their response on a *decision variable*, which is the level of plausibility, i.e. the level of agreement with their expectations of the experience of listening to a real loudspeaker in the room. If this is sufficiently high then they will respond *yes*, otherwise they will respond *no*. This threshold is subjective and is called the *criterion* in SDT.

To analyse responses, a simple observer model is used, with equal-variance Gaussian distributions representing the responses to real stimulus (signal absent) and simulated stimulus (signal present) conditions. The sensory difference between reality and the simulation is described by the distance between the two distributions. This distance is the sensitivity parameter d', which can be estimated for individual assessor i from the hit-rate and false-alarm-rate in their responses:

$$\hat{d}'_i = Z(p_{\mathsf{Hit}_i}) - Z(p_{\mathsf{FA}_i}) \tag{5.1}$$

Here ^ indicates an estimated variable and $Z(p)$ is the inverse cumulative normal distribution. The hit-rate p_{Hit_i} is the proportion of presented binaural simulations that were correctly identified as such and the false-alarm-rate p_{FA_i} is the proportion of real loudspeaker stimuli that were incorrectly identified as a simulation.

The individual response criterion λ_i can be estimated from the false-alarm-rate, and

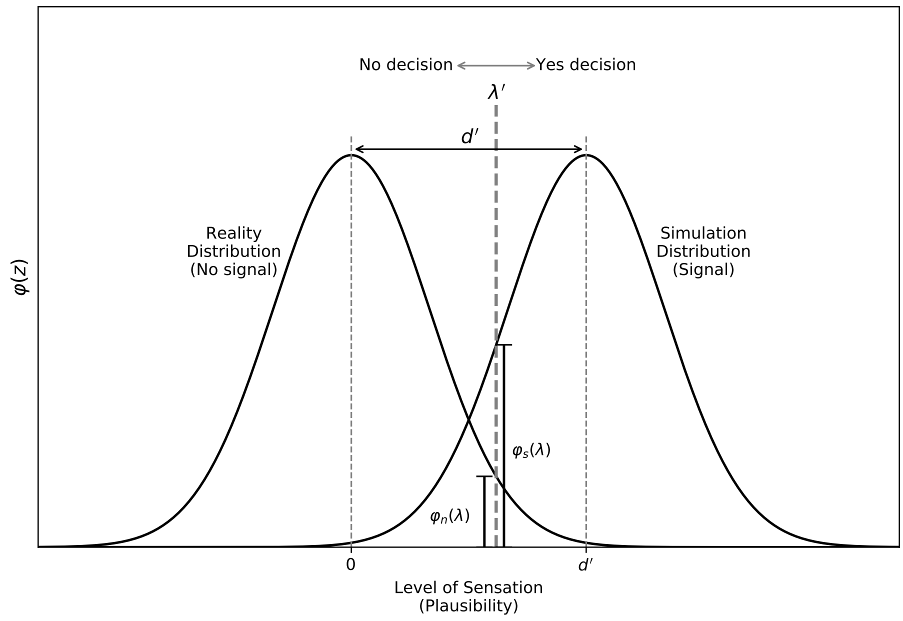

No decision ⟷ Yes decision
λ′
d′
Reality Distribution (No signal)
Simulation Distribution (Signal)
φ(z)
φ_s(λ)
φ_n(λ)
0
d′
Level of Sensation (Plausibility)

Figure 5.3: An illustration of the equal-variance Gaussian signal detection theory (SDT) observer model, with key parameters d' and λ' indicated.

provides an indication of bias in the participant's responses:

$$\hat{\lambda}_i = Z(1 - p_{\mathsf{FA}_i}) \tag{5.2}$$

An alternative measure of bias $\hat{\beta}_i$ can be obtained from the ratio of the values of the normalised probability density $\varphi(x)$ of the real and simulated distributions at the position of the response criterion:

$$\hat{\beta}_i = \frac{\varphi(\hat{\lambda}_i - \hat{d}'_i)}{\varphi(\hat{\lambda}_i)} \tag{5.3}$$

This represents response bias as a likelihood ratio. The numerator is the likelihood of obtaining plausibility at the criterion level $\hat{\lambda}_i$ in the case of a headphone simulation and the denominator is the likelihood of obtaining plausibility $\hat{\lambda}_i$ in the case of loudspeaker reproduction. This allows the bias to be interpreted independently from the sensory difference (d'_i) between the two cases. Since, according to this model, the sensory difference (d'_i) between the two conditions is independent of the response criterion λ_i, SDT approaches are often said to perform criterion-free measurement of sensory differences.

5.3.5.1 Minimum Effect Hypothesis

A sensitivity of $d' = 0$ would indicate complete plausibility of the binaural simulation. Since this null hypothesis cannot be proven with inferential statistics, an alternative hypothesis should be set, according to an effect that is perceptually irrelevant i.e. a minimal increase in the observed sensitivity over perfectly-random guessing. If this alternative hypothesis ($d' \geq d'_{\text{min}}$) can be rejected then the binaural simulation can be said to be plausible. Lindau and Weinzierl (2012) related sensitivity values to detection rates in the 2AFC paradigm in order to set a meaningful minimum effect hypothesis. A sensitivity $d'_{\text{min}} = 0.1777$ was chosen as the minimum effect, which is equivalent to a detection rate $P_c = 0.55$, according to:

$$d' = \sqrt{2}\, Z(P_c) \tag{5.4}$$

This represents quite a strict test. Lindau and Weinzierl (2012) report that $P_c = 0.75$ is commonly used as the detection threshold in 2AFC experiments, as was used by Hartmann and Wittenberg (1996).

To test the minimum effect hypothesis with given type-I and type-II error levels, an optimal sample size can be approximated:

$$N_{\text{opt}} = (z_\alpha + z_\beta)^2 \frac{2\pi}{\hat{d}'^2_{\text{min}}}, \tag{5.5}$$

where z_α and z_β are the z values for type-I and type-II error respectively. This estimation assumes perfectly unbiased participants and equal variance between noise and signal conditions.

5.4 A Subjective Assessment of Plausibility

A listening test was carried out to assess the plausibility of the previously described binaural rendering system, following the method of Lindau and Weinzierl, 2012.

5.4.1 Hypothesis

In this experiment the null hypothesis is:

H 0 $\hat{d}'_{avg} = 0$

This would indicate that headphone reproduction using the binaural rendering system gives simulation in perfect agreement with listeners' expectations of the sound of a real loudspeaker i.e. the rendering is entirely plausible.

The same stringent minimum effect hypothesis is used in this assessment as in Lindau and Weinzierl, 2012, $d' = 0.1777$, to allow comparison to this previous study. Therefore the alternative hypothesis is:

H 1 $\hat{d}'_{avg} < 0.1777$

If the sensitivity level observed across the population of listeners is significantly below the value of this minimum effect, then this alternative hypothesis can be rejected. This would indicate a highly plausible binaural rendering system, with evidence supporting the null hypothesis.

By applying this value of d'_{min} to equation (5.5) with type-I/type-II error level of 0.25/0.05, the optimal sample number $N_{\text{opt}} = 1071$ was calculated i.e. to be able to reject the alternative hypothesis that the binaural simulation is implausible with 95% test power, a minimum of 1071 samples should be taken.

5.4.2 Procedure

Each participant was presented with 100 stimuli, each one either from a real loudspeaker or simulated with binaural rendering. After each stimulus presentation, the participant had to decide if the stimulus was created by a headphone simulation, giving a “Yes” or “No” answer. Stimuli were randomly varied according to content item and source location, and the selection of real or simulated presentation was determined randomly for each presentation. Each combination of source location and content item only featured once in the test, whether real or simulated, in order to avoid memory effects that might bias results.

20 monophonic audio items were used, including male and female speech in native and foreign languages, popular and orchestral music ensembles, and individual instrument recordings. Loudness differences between the source items were compensated using Rec. ITU-R BS.1770 (ITU-R, 2011). Five source positions were used in the assessment. The loudspeakers were positioned as listed in Table 5.1, no attempt was made to ensure precise symmetry in the configuration. These loudspeakers were used for measurement of the BRIR dataset and during the assessment. They were visible to the participants during the test session.

Participants wore the headphones throughout the assessment. As mentioned in section 5.2.2, the headphones were also placed on the dummy head during BRIR measurement in an attempt to maintain a constant headphone effect between the two conditions. Participants were told that they may rotate their head, but only in the horizontal plane, avoiding tilting or rolling, and that they must keep their torso still. This was because BRIR measurements

Loudspeaker	Distance	Azimuth	Elevation
1	1.87 m	0.5°	−1.9°
2	1.87 m	33.2°	−2.0°
3	2.20 m	−32.4°	32.0°
4	2.29 m	−113.1°	30.8°
5	1.93 m	111.0°	−0.4°

Table 5.1: Positions of loudspeakers used in the assessment.

were only available for such head rotations, due to limitations of the rotational mount used (see appendix A.5.4). Head position was tracked in six degrees of freedom throughout the experiment, where movements substantially deviated from this requirement, assessors were reminded of it. The loudness of the binaural system was set before the assessment by two experienced listeners, this author and a colleague, to match that of the real stimuli across the range of content and source positions.

Prior to the test session, participants were asked to find the ITD scaling factor which best stabilised a frontal sound source. They were guided through this process by the experiment administrator. A speech stimulus was presented using the BRIRs for a frontal source and the listener could adjust the scaling factor in small steps ($\pm$0.025), resulting in typical changes to ITDs in the order of 10 μs. The chosen ITD scaling factor was then used throughout the rest of the test. The median of the chosen ITD scaling factors was 0.85 (minimum 0.75, maximum 1.0). This was followed by a familiarisation stage, in which the listener was first presented with each of the audio items to be used in the test, each time from a randomly chosen real loudspeaker. Subsequently they were presented each audio item from randomly chosen binaurally rendered virtual loudspeakers.

5.4.3 Participants

To achieve the required sample size, 11 listeners were recruited for the assessment. The participants were all staff at BBC R&D, they all had experience of critical listening and an awareness of binaural technology.

5.4.4 Results

Estimates of the individual sensitivity $\hat{d}'_i$ and bias $\hat{\beta}_i$ were calculated for each participant. The mean of the individual sensitivities across the participant group $\hat{d}'_{\text{avg}}$ was above zero, meaning that there was a sensory difference between the binaural simulation and reality, with the simulations being identified as such more often than the real stimuli. The mean of individual response biases indicates that the participant group did not show a bias towards reporting the stimuli as real or simulated ($\hat{\beta}_{\text{avg}} \simeq 1$), independent of the actual sensory differences. The mean and standard deviation of estimated individual sensitivities and biases is given in table 5.2 and plotted in figure 5.4 using 90 % confidence intervals. The figure uses 90 % confidence intervals after Lindau and Brinkmann (2012) to enable direct comparison.

$\hat{d}'_{\text{avg}}$	$\hat{\sigma}_{d'}$	$\hat{\beta}_{\text{avg}}$	$\hat{\sigma}_{\beta}$
0.1954	0.2927	1.0037	0.0873

Table 5.2: Mean and standard deviation of estimated individual sensitivity and bias values.

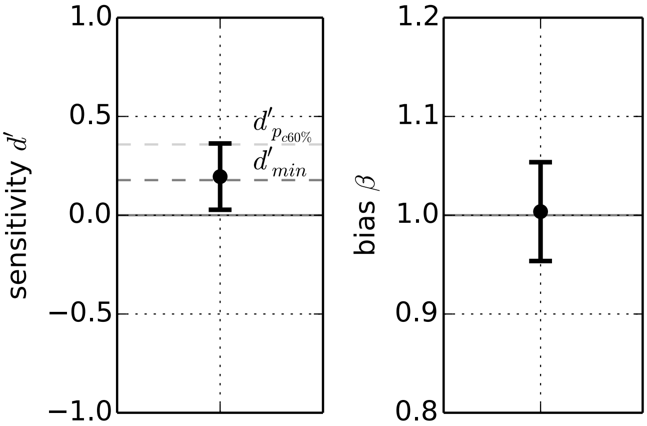


Figure 5.4: Mean of individual sensitivities d'_i and biases β_i with 90% confidence intervals

The participant group's mean sensitivity $\hat{d}'_{\text{avg}}$ is close to but greater than d'_{min} (0.1777), chosen to represent a meaningful effect i.e. a meaningful sensory difference between the real and simulated cases was observed. The alternative hypothesis 1 cannot therefore be rejected. The sensory difference is however quite small. All listeners stated that it was difficult to identify the simulation and that only in some cases could they do so with any confidence. Figure 5.4 also shows the value of d' equivalent to a detection rate $P_c = 0.6$ in a 2AFC test, $d'_{p_{c60\%}} = 0.3583$. A t-test was performed to assess the significance of the difference to this value, but there was no significant difference at a 95 % level (one-sided test, $t = -1.7602, p = 0.0544, r = 0.4864$). A Shapiro-Wilk test was run before this to confirm that the individual sensitivity values are normally distributed ($W = 0.9149, p = 0.2784$).

To assess whether the loudspeaker position had an effect on plausibility, sensitivity was calculated for each loudspeaker separately. Due to small sample sizes, sensitivity was calculated across the entire group ($\hat{d}'_{\text{grp}}$) rather than individually. This means that inter-individual variance cannot be assessed and so neither can the significance of differences of the distribution of individual sensitivities from a threshold. The results are shown in Table 5.3. The sensitivity was greater for elevated loudspeakers, particularly loudspeaker 4, which was to the rear of the listeners. Sensitivity was lower for horizontal sources located off the median plane.

Loudspeaker	$\hat{d}'_{\text{grp}}$
1	0.2047
2	-0.0945
3	0.2848
4	0.5279
5	0.0413

Table 5.3: Estimated group sensitivity values for each loudspeaker

A key aspect of this test design is that direct comparison between simulated and loudspeaker stimuli is not possible, assuming that memory effects are not strong enough. In case a measurable learning effect could be observed over the course of the test, a related samples t-test was carried out on the individual sensitivities for the first and second halves of the test sessions. There was no significant difference between the means of the two distributions ($t = -0.3403, p = 0.3703$).

5.4.5 Discussion

It is important to consider what level of plausibility is actually required for high quality immersive audio entertainment. Whilst this rendering system cannot be considered plausible in relation to the strict minimum-effect hypothesis, all participants said that it was challenging to identify the simulation and many said they felt that they were often simply guessing. If this system were rated poorly by listeners in a preference test, it is unlikely that it would be due to a failure to create a convincing spatial impression.

Figure 5.5 represents the result by plotting the probability density distributions of participant responses, as represented by the equal-variance Gaussian SDT observer model. The two distributions have a high degree of overlap. The mean sensitivity $d'_{\text{avg}} = 0.1954$ observed in this study is equivalent to an average detection rate of 55.49 % in a 2AFC test, where 50 % would indicate perfectly-random guessing.

Participant comments revealed the stimulus characteristics that were used to make their decisions, although it cannot be known whether these were used correctly or not. Several

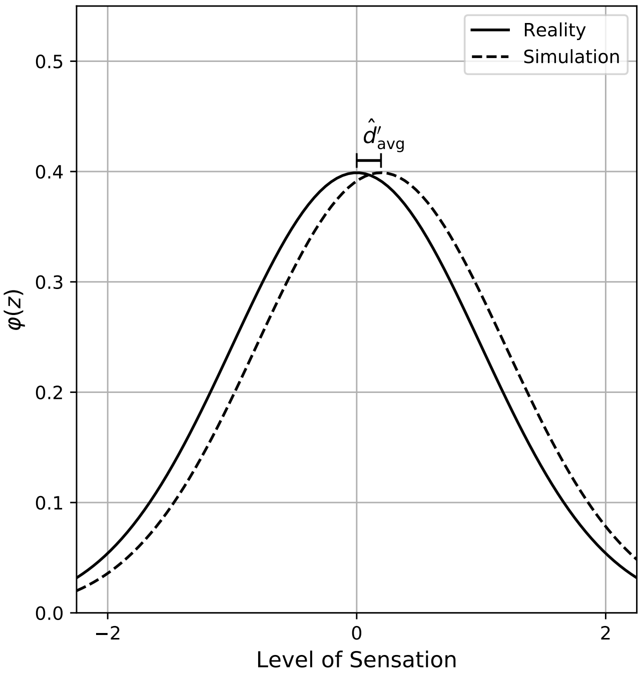


Figure 5.5: Probability density distributions of the equal-variance Gaussian SDT model of the participant group's performance in detecting the binaural simulation from reality

participants said that they made use of the dynamic behaviour of the system as they rotated their head, where instability in source position was often assigned to the simulation. However, some participants noted that they were unsure as to whether this instability and confusion was caused by the headphone's influence on the real loudspeaker sound field, as observed during the familiarisation process. It is also possible that some participants did not find the correct ITD scaling factor to maintain stationary source positions. Some participants found the initial ITD scaling procedure challenging and took several minutes to identify an appropriate value.

It was noted by a few participants that the simulated stimuli seemed to have less precisely defined position, which could be described as a larger apparent source width or increased localisation blur. Also, for the elevated source positions, it was noted that in some instances the source location appeared lower than the visible loudspeaker. This might explain the greater sensitivity observed for elevated loudspeakers. Two participants also commented that the source direction was sometimes initially ambiguous, before moving their head, and that the perceived direction sometimes reversed after head movement. Such issues have previously been observed in the literature when using non-individualised binaural rendering (Wenzel, Arruda, et al., 1993b). No participants reported experiencing inside-the-head localisation or changes to perceived source distance, which follows findings of Zahorik, 2002a.

It is worthwhile comparing this result to others in the literature. In the study reported by Lindau and Weinzierl (2012), the "*improved simulation*" was highly plausible ($d'_{\text{avg}} = 0.0512$), with participants "almost perfectly guessing", indicating that this system had a higher degree of plausibility than the system evaluated here. However, in that same study, the "*basic simulation*" showed greater sensory differences ($d'_{\text{avg}} = 0.2956$) than the system evaluated in this chapter. More recently, Bailey and Fazenda (2018) applied the same experimental method to evaluate an audiovisual virtual environment where virtual sound sources were created using B-format impulse responses and binaural rendering of first-order ambisonics, using the Google Resonance Audio SDK (Google, 2018b). In that study individual sensitivities were much higher, the majority of assessors had sensitivity $d' > 1$ and several had $d' > 5$. Note also that the sensitivity value corresponding to a detection rate of 75 % in a 2AFC test is $d'_{p_{c75\%}} = 0.9539$. Therefore that system was clearly not plausible. The system evaluated in this chapter showed a higher degree of plausibility than all but the "*improved simulation*" of Lindau and Weinzierl (2012).

There are large differences in size of the environments and loudspeaker distances used in these studies. Lindau and Weinzierl carried out their experiment in a large auditorium with longer reverberation time ($T_{60} = 2.0$s) and loudspeakers at a distance of 3–5 times the critical distance r_{crit}, compared to 1.5–2 times in this test. The differences between binaural microphones used in measurements may also have been significant. It has been shown that commercially available dummy head microphones, including the one used here, typically yield poorer localisation performance than using a non-individual human head (Minnaar, Olesen, et al., 2001). The FABIAN HATS used by Lindau and Weinzierl was constructed using a mould of the head of a human individual (Lindau and Weinzierl, 2006). FABIAN also has a torso, with the head rotating independently above it, whilst the Neumann KU100 has no torso. It has been shown that the reflection from the torso has an effect on perception of sound source elevation, although this is secondary to pinna cues in importance (Gardner, 1973). This may have led to decreased plausibility, particularly for elevated sources.

Aside from the different implementations of the systems, the decision to allow a familiarisation stage in this test may have affected results. This is a departure from the procedure presented by Lindau and Weinzierl (2012). During the formal testing of headphone effects on external sounds presented in co-authored publication Co.P. III and informal listening during this study, it was observed that the headphones do have an effect on external sounds. Whilst the headphones used here are more transparent than most models available, they can cause confusions to source localisation and the effects appear to change with head rotation. The familiarisation process was included to reduce the chance that participants wrongly interpreted these effects as artefacts in the simulation, which would increase the level of observed plausibility artificially. Extra-aural headphones (see section 2.7.1) may avoid these issues in similar future work, as have been used in subsequent studies by Brinkmann, Lindau, and Weinzierl (2014) and Romigh, Brungart, and Simpson (2015).

Following this experiment, this author advised on the work of Genovese (2014), where this plausibility evaluation method was applied to static binaural rendering with individualised binaural impulse response measurements in both anechoic and reverberant environments. Through the SDT analysis, it was found that assessors showed differing response biases in each environment. In anechoic environment there was a bias towards thinking that the sounds were simulated using the headphones, while in the reverberant environment there was a bias towards thinking that the sounds were real. There was a corresponding shift in sensitivity, with $d' < 0$ in anechoic conditions. The SDT model appears not to fit the assessors perfectly. This suggests that level of plausibility is influenced by the environmental conditions as well as the binaural rendering. In both environment conditions, higher absolute sensitivity was observed by Genovese (2014) than with the dynamic non-individual system used in this study.

Research question 1 asks if binaural technology is capable of producing a convincing spatial impression without calibration for the individual listener. The experiment presented here provides evidence that a convincing spatial impression of single loudspeaker sources can be achieved. More complex scenes have not been evaluated, which could be the subject of further work. Whilst the system does not require individualised measurements, the BRIRs were measured in the reproduction environment. Such acoustic measurements at the reproduction side are not feasible for mass audience applications. It would be of interest to study the sensitivity of plausibility further, but it seems highly unlikely that measurements from one room can be used to provide plausible rendering in a wide range of environments.

The experimental method is limited to assessing simulations of existing environments. In media production, realism is often not the creative aim. It may sometimes be more appropriate to simulate environments that do not exist or are not easily accessible. It would be interesting to apply this methodology to validation of a binaural system that aims to create plausible artificial environments through modelling rather than using measured data. Additionally, alternative experimental methods for evaluating plausibility in different application scenarios are needed.

An interesting recent study by Bergstrom et al. (2017), explores plausibility in an audio-visual virtual environment i.e. a virtual reality (VR) application. Plausibility is described as the illusion that events in the virtual environment are really happening, and is distinguished from *place illusion* or "being there". Both are said to be components of *presence*. Their experiment compared the importance of several factors of the virtual environment in achieving a sense of plausibility. Assessors were able to successively introduce features to the environment and were required to maximise the sense of plausibility as quickly as possible. The scene involved a string quartet rehearsing in a domestic environment. Features included: sound spatialisation, reverberation, environmental ambience sound effects, and performers returning the user's gaze. The latter two factors were selected earlier and more often than spatial audio factors. This method indicates the contribution of factors towards plausibility, but does not measure how plausible the experience itself is.

Brandenburg et al. (2018) presented a system based on measured BRIRs for augmented reality (AR) applications. Though measurements are still made in the reproduction environment, more flexibility in listener movement is permitted, using simple BRIR adaptation techniques for interpolating between measured positions. The paper cites the target of plausible reproduction and presents listening experiments demonstrating that externalisation can be preserved with these methods.

5.5 Conclusion

This chapter has considered the verification of the spatial impression given by binaural rendering. After considering several approaches, the plausibility of a non-individualised dynamic binaural rendering system was evaluated. A listening experiment was designed using SDT, which tested the null hypothesis that the system was capable of creating an entirely plausible simulation of real loudspeakers in the listening room. A strict alternative hypothesis was set, corresponding to a correct detection rate of 55 % in an equivalent 2AFC test. The results showed that listeners could detect the binaural system with a sensitivity too great to reject this minimum-effect hypothesis. However, the sensory differences were still small by common standards and all participants said they had difficulty detecting the simulation. This suggests that even without individualisation, binaural rendering can create listening experiences largely in accordance with expectations of real events, though occasionally deviating from those expectations.

5.5.1 Outlook

In the pilot study of chapter 4, a commercial system was included that had similar characteristics to the system evaluated in this chapter. It combined individualised HpCF and BRIR measurements made in the playback environment, with head tracking, to create a dynamic HSSP system. For this system, listeners rated the perceived sound quality significantly lower than a simple conventional stereo down-mix. The plausibility of that system has not been evaluated formally, but informal listening by this author shows similar degrees of realism to the system evaluated in this chapter. These impressions are shared by other listeners who were consulted. It appears that perceived sound quality is influenced by factors beyond just the level of plausibility of binaural rendering.

The listeners' expectations of sound character will be affected by the context of listening. When a listener wears headphones to watch a television programme, a different sound character may be expected to that heard on loudspeakers in the room. Yet in an augmented reality system, such correspondence to sounds in the external environment is expected. The effects of the content production decisions and the spatial audio format used to represent the content will have an influence. The 5.1 surround sound format does not enable optimal use of the auditory space, since the corresponding 3/2 loudspeaker layout is horizontal-only and the loudspeakers are sparse outside of the frontal region. Binaural technology is likely to provide more benefit when a full 3D scene can be presented. Since the beginning of work towards this thesis, there have been many developments in applications of spatial audio technology, which make this more feasible. These are discussed in chapter 6. The choice of binaural rendering techniques used in the system will also have an influence on quality and these should be clearly identified and controlled to gain deeper understanding. With the systems and apparatus used so far it is challenging to formally investigate these system and content format aspects together in detail. Therefore the experimental apparatus described in this chapter was extended, as described in appendix A. This has been used in the studies reported in subsequent chapters of this thesis.

Chapter 6

3D Spatial Audio Applications and the Role of Loudspeaker Virtualisation for Headphone Reproduction

This chapter first presents applications of a custom binaural rendering system to content production, including two web-based evaluation studies. It then discusses developments in 3D spatial audio standards and applications that have occurred since the beginning of this project. The widespread use of loudspeaker virtualisation techniques for headphone reproduction is noted. Existing work to study the effects of these techniques on quality are therefore reviewed.

6.1 Introduction

In this applied work, the aim is to use binaural technology in entertainment media applications to improve the headphone listening experience. The techniques and components used in a binaural rendering system will have an important influence on quality. A review of the state-of-the-art was given in chapter 2 and this informed the development of a rendering system which was evaluated in chapter 5. Appendix A presents extensions of this work to establish experimental apparatus for comparison of state-of-the-art binaural rendering techniques and study of their influence on quality.

The format in which spatial audio programme content is presented to the binaural rendering process will also likely have an impact on quality; as of course will the audio signals themselves and the production techniques used to create them. It is therefore important to consider the practicable options for producing and delivering binaural sound in media applications and the associated production formats. The developed apparatus can also be applied to programme production, allowing exploration of these aspects with professional content producers.

The practical work towards this thesis began, in chapter 4, by considering the distribution of a pre-rendered binaural signal and making use of existing programme content in the 5.1 surround format as the input. In this chapter, section 6.2 discusses work carried out to explore different approaches to the production of pre-rendered binaural signals. This includes two web-based studies to evaluate the outcomes of some of this work with target audience members. Section 6.3 then discusses recent developments in 3D spatial audio technologies, formats and standards. This presents new opportunities for mass-audience entertainment media applications of binaural technology and for spatial audio more widely.

It is common in these technologies to make use of loudspeaker virtualisation techniques, as found in earlier headphone surround sound processing (HSSP) systems. Often this processing will occur on the receiving-end in audience devices and so programme content must be produced in intermediate formats. Both the loudspeaker virtualisation process and the rendering of spatial audio scenes into the intermediate spatial audio formats are expected to influence the quality of the listening experience.

Section 6.5 in this chapter reviews prior studies of the perceptual effects of reproducing virtual sound sources using loudspeakers and those related to the perceived quality of loudspeaker virtualisation over headphones. Section 6.6 summarises the topics covered in this chapter and considers the outlook for the subsequent studies in this thesis.

Appendix B gives supporting information about techniques for rendering scenes to intermediate formats. It describes in detail the apparatus developed and subsequently applied in this thesis for investigating the rendering of these formats to headphones by virtualisation. Appendix B.7 also gives a short review of some more advanced techniques for loudspeaker virtualisation.

6.2 Web-based Evaluation of Binaural Programme Content

The apparatus presented in appendix A has been used in a number of programme productions at the BBC to generate a two-channel binaural signal for headphones. This has allowed exploration of creative techniques for the effective use of binaural rendering, working with professional sound engineers and programme producers. Detailed discussion of this work is not within the scope of this thesis, but a short summary of the approach is given in this section. Two online listening experiments were run to evaluate the output from early production work, shortly after the production process was established in October 2014 (Pike and Nixon, 2014). These are presented briefly in this section, since they help to explain the direction of further research in this thesis.

6.2.1 Distribution of Pre-Rendered Binaural Programmes

Distribution of binaural programme content as pre-rendered two-channel signals is appealing because it requires no modification of the receiving-end system. All that is required is a receiving device that can reproduce a two-channel signal over headphones. In this scenario though, the reproduction cannot incorporate head-tracking or binaural filter individualisation. Neither can an appropriate headphone-to-ear correction filter (HpCF) be applied without dedicated signal processing on the receiving-end, it is likely also to require measurement. The review in chapter 2 suggests that this will lead to somewhat limited perceptual accuracy, making well-externalised and precisely-localised auditory events less likely.

The experiment reported in chapter 4 explored this type of application, though with the added constraint that the programme production process should not be changed. Binaural rendering was applied to 5.1 programme content as a post-process. It seems likely that using the 5.1 surround format as the input is a limiting factor on the spatial impression that can be achieved. Binaural rendering can reproduce full 3D scenes with many virtual sound sources. Provided that the resources can be committed to content production, it seems likely that higher quality results can be achieved even with static non-individualised binaural rendering.

In applications where pre-rendered binaural signals are to be distributed, the complexity of the scene representation and of the rendering techniques is not heavily constrained. The end product is still a two-channel audio signal, which is efficient to distribute and can be reproduced on mobile devices without signal processing beyond that needed for normal reproduction of stereo signals. The main constraint is that the rendering can be reliably performed in real-time on the production system.

By contrast, applications that require rendering on the receiving-end are faced with constraints on the distribution bandwidth and the computational complexity of processing on the receiving device. Binaural technology applications are often associated with mobile systems, where there are technical limits to both of these factors, i.e. limited network bandwidth and processing power, as well as associated effects such as service cost and battery life. Technologies and techniques associated with these applications are discussed further in section 6.3.

6.2.2 Production of Binaural Programme Content

The software-based spatial audio rendering system described in appendix A.2 has been configured to enable use of multiple parallel rendering approaches, which can be applied to different elements within a scene. This is illustrated in figure 6.1. A remote control interface was established to a professional digital audio workstation through use of a third-party spatial audio authoring plug-in, the IOSONO Spatial Audio Workstation (Barco Audio Technologies, 2017). This plug-in sends UDP messages over the local network to the renderer application. It can control the spatial parameters of the scene elements, as well as the choice of rendering method for each element. The rendering includes three options, including two of the renderers described in appendix A.3: the basic binaural renderer using head-related impulse responses (HRIRs), the virtual loudspeaker renderer using measured binaural room impulse responses (BRIRs), and a simple stereo-panning renderer. Binaural impulse responses were all measured with the Neumann KU100 dummy head microphone as described in appendix A.5.

The HRIR renderer is used most often in practice. It allows positioning of sources at any target direction. The virtual loudspeaker renderer may sometimes be used where pronounced spatial effect is required, provided that the colouration caused by the room response is not problematic. Care was taken in processing the BRIR measurements, applying equalisation of loudspeaker responses, incorporating low-frequency response extension, and ensuring precise onset alignment with the HRIR used in the basic binaural renderer, as well as loudness alignment. The stereo rendering option is important in many cases. It allows creation of scene elements that bypass the binaural rendering and are clearly localised in-head. Controlled distinction between the internal and external auditory space has creative applications, such as separating narration from a dramatic scene. The IRCAM Spat 3D reverberation engine was also used via a link to the Max/MSP environment (Carpentier, Noisternig, et al., 2015). This provided a parametric room reverberation model that could receive source positions and be controlled by auxiliary sends from the audio workstation. It was configured to provide a return feed in a 16-channel 3D array of source positions, which are then rendered by the HRIR binaural renderer. This model-based auditory virtual environment (AVE) approach was found informally to give a less plausible spatial impression than the data-based approach with measured BRIRs, but is clearly more flexible, allowing the use of environmental reverberation that is appropriate for the content. Overall, this system provides a flexible range of options which may be combined in binaural programme production. It allows the sound engineer to select the most appropriate method for each scene element in an attempt to achieve high quality output.

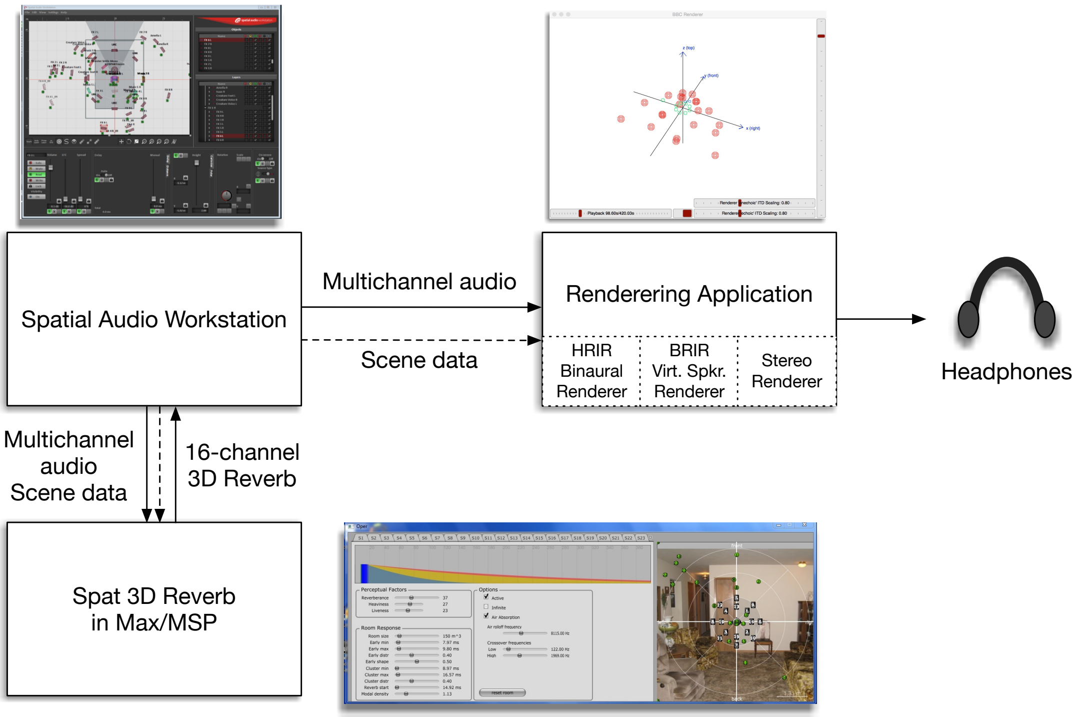
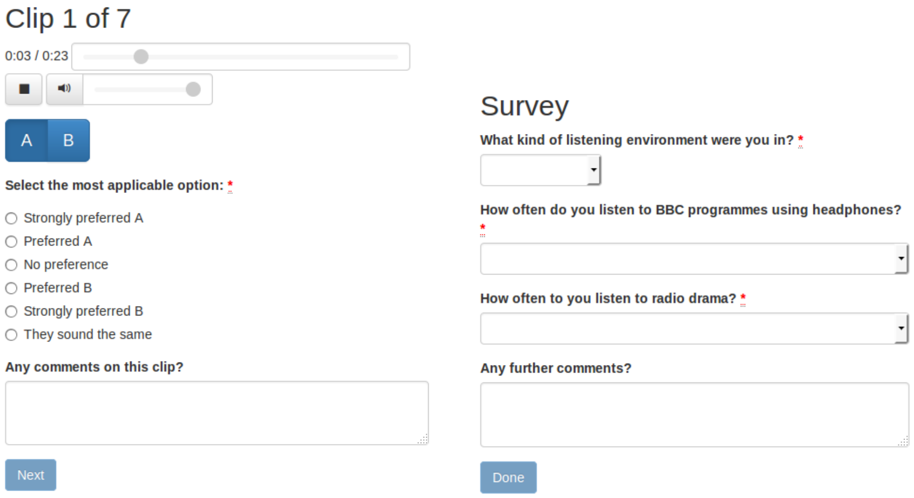


Figure 6.1: System for production of binaural programme content



Clip 1 of 7

0:03 / 0:23

Select the most applicable option: *

○ Strongly preferred A
○ Preferred A
○ No preference
○ Preferred B
○ Strongly preferred B
○ They sound the same

Any comments on this clip?

Survey

What kind of listening environment were you in? *

How often do you listen to BBC programmes using headphones? *

How often to you listen to radio drama? *

Any further comments?

(a) Comparison rating interface (b) Post-rating survey

Figure 6.2: Web-based comparison experiment interface

6.2.3 Web-based Evaluation of a Binaural Audio Drama

In October 2014, a binaural version was produced of an episode of a BBC audio drama called Tommies. It was made available on the BBC Radio 4 website from the time that the stereo version was broadcast on the radio service (Pike and Melchior, 2014). Tommies is a serial audio drama that follows the lives of soldiers in World War I. The system described in section 6.2.2 was used to produce the binaural version. This means that the sound engineer was able to choose from a variety of rendering methods to apply to each element within the scene.

An online (web-based) listening experiment and survey was conducted in addition to making the full programme available. The development of the test interface and data collection software was performed by a colleague. Audience members were invited to participate through the BBC website. In this experiment, assessors compared short excerpts of the stereo and binaural programme versions and give their preferences on a comparative scale, as shown in figure 6.2a. The comparison was blind, the assessor was not told which version was assigned to *A* or *B*, and the presentation order was randomised for each trial. An option was given to report that the versions sounded the same to the assessor, to distinguish this from the case where they sounded different but the assessor had no preference for either. Seven programme excerpts (*items*) were chosen from the episode to cover a range of different scenes. These are described in table 6.1.

Item	Description	Scene Layers	
Ceasefire	Outdoor scene. Dense gunfire and soldiers shouting at a range of distances and directions, with a distinct sniper rifle shot at the end.	Stereo	Dialogue and distant shots.
		HRIR	Foreground shouting (lateral), background shouting (wide and in-front), wind atmos. (elevated and lateral), gun shots (at a range of directions).
EngineMovement	Outdoor scene. Narration and music, with soldiers shouting, horses neighing, distant gunfire, and a rattling motor car driving away from the scene.	Stereo	Narration, music (strings), background shouting, horse neighing.
		HRIR	Background chatter and shouting (rear), wind and rain atmos. (lateral and rear), distant guns (lateral).
		BRIR	Car engine (moving across the front and round to the rear-right).
HallBriefingShort	Indoor scene. Two soldiers speaking quietly nearby and another giving a briefing at a distance. Occasional activity of other soldiers in the room e.g. laughing and standing to attention.	Stereo	Muttering, feet stamping.
		HRIR	Close dialogue (wide rear).
		BRIR	Briefing speech (front), laughing, chairs moving, and feet stamping (lateral and rear), room tone (wide, rear and elevated).
Horse	Outdoor scene. A soldier arrives on a horse. There is a conversation with speakers at varying distances, with distant gunfire in the background. The soldier then leaves on the horse.	Stereo	Dialogue, wind atmos.
		HRIR	Wind atmos. (lateral and elevated), shelling and distant gunfire (lateral and rear). Horse (moving from the right in to the front and then back to the right).
Running	Outdoor scene. A band of soldiers running and shouting. Another solider walking and rustling through leaves. There is shellfire and shouting in the distance.	Stereo	Narration, dialogue, soldiers panting, distant gunfire, wind atmos.
		HRIR	Wind and rain atmos. (lateral, rear and elevated), rustling foliage (rear and below). Soldiers shouting and running (wide front to rear right, footsteps at low elevation). Distant character shout (front-right-up).
ShellSound	Outdoor scene. The sound of shell fire travelling overhead and then exploding around the listener, with narration describing the sounds.	Stereo	Narration, dialogue, sub-bass explosion. Shrapnel falling (panned hard right).
		HRIR	Wind and rain atmos. (lateral, rear and elevated), shell shot, flight and overhead explosion (moving in an overhead arc from rear-left to up-front).
		BRIR	Whistle of falling shrapnel (moving up-front-right to down-front-right).
Writing	Indoor scene. Narration, piano music, and the sound of handwriting.	Stereo	Narration and piano.
		BRIR	Handwriting (in front).

Table 6.1: Items used in the Tommies experiment

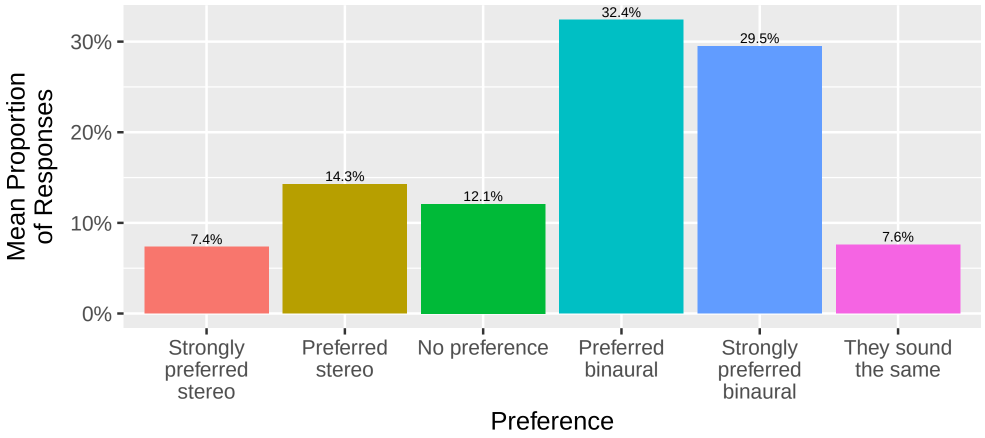


Figure 6.3: Preferences for binaural or stereo versions of Tommies programme excerpts. The proportion of responses in each category was first calculated for each excerpt, to account for the different number of ratings for each excerpt. The mean proportion of responses in each response category over all excerpts is presented.

Before rating, participants were asked what type of headphones they were using; although they were encouraged to use headphones, a "not using headphones" option was provided to prevent contamination of the data from users who wanted to try the experiment but did not want to use headphones. Following the rating, participants were invited to fill in a short survey, shown in figure 6.2b.

A total of 408 preference ratings were received from 74 participants. The results were filtered to remove responses from those who reported not to have been using headphones and those who did not listen to the full length of each version of the programme item. After this process, 213 ratings from 59 participants remained. The number of ratings differed between programme items. The proportion of responses in each category was first calculated for each item. Figure 6.3 shows the proportion of preference ratings in each category over all items, after this adjustment.

The 13 responses where the assessor heard no differences were then removed, and the preference ratings were treated as a 5-point ordinal scale ranging from -2 ("strongly preferred stereo") to 2 ("strongly preferred binaural"). A Wilcoxon signed rank test shows that there was a significant preference for the binaural versions of the programme excerpts ($Z = 12320, p < .001, r = 0.488$).

Figure 6.4 shows the preference ratings separately for each programme item. It is clear that there were different preferences amongst these excerpts. Table 6.2 shows the results of Wilcoxon signed-rank tests for the preference ratings of each item. There was a significant

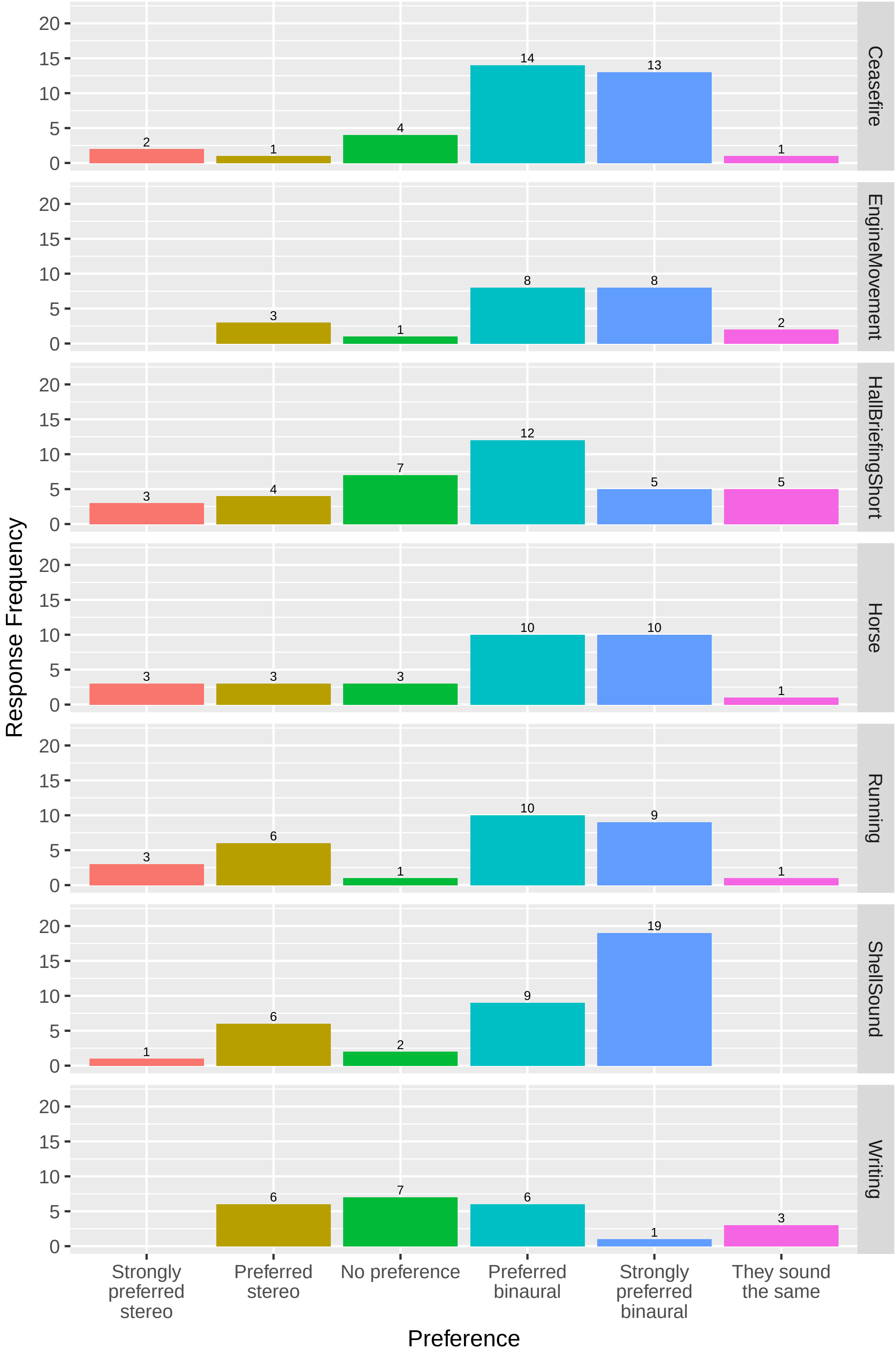


Figure 6.4: Preferences for binaural or stereo versions of Tommies by programme item.

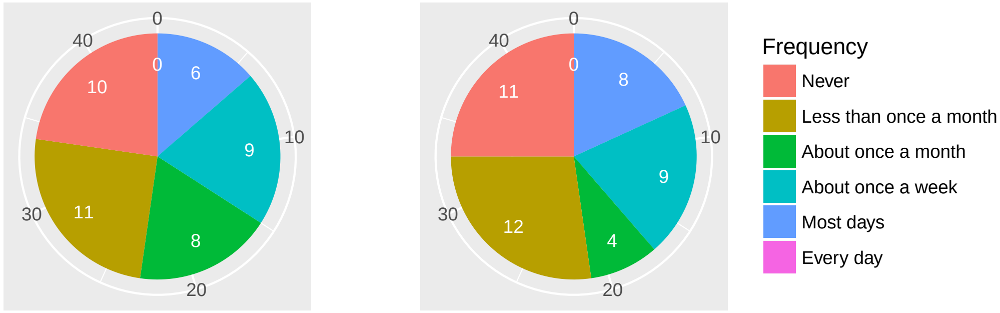
Item	Pseudo-	95 % C.I.		Z	N	p-value	r
	median	Low	High				
Ceasefire	1.50	1.00	1.50	411.0	35	<.001***	0.638
EngineMovement	1.50	0.50	1.50	172.0	22	0.001**	0.678
HallBriefingShort	1.00	-0.00	1.00	204.5	36	0.108	0.268
Horse	1.00	0.00	1.50	270.0	30	0.014*	0.449
Running	0.50	-0.00	1.50	287.5	30	0.048*	0.361
ShellSound	1.50	0.50	1.50	556.5	37	<.001***	0.671
Writing	0.00	-1.00	1.00	52.0	23	0.644	0.096

Table 6.2: Wilcoxon signed-rank tests of preference rating distributions for each programme excerpt, with pseudo-medians and associated 95 % confidence intervals, and estimated effect size r.

preference for the binaural version on 5 of the 7 items. For the other two items there was no significant preference. Both of these items also showed higher frequencies of “they sound the same” responses. These two scenes were less complex than the other items, with dominant scene elements using the stereo panning option, so it is unsurprising that preferences were not so apparent.

Overall these results suggest that improvements in the headphone listening experience can be achieved with binaural rendering, when compared with conventional stereo signals. This is despite the rendering and reproduction making no use of individualisation, head tracking or headphone correction filtering. It appears that enabling full 3D scene authoring and adaptation of the rendering approach for each scene element during production might provide benefits over the 5.1 surround virtualisation approach evaluated in chapter 4. It should be noted, however, that this approach is highly specific to the content and production approach used and these findings are for scenes from a single audio drama. Walton (2017) presented the results of another web-based study, to evaluate the overall listening experience (OLE) of binaural music productions. These were created using the same production system as used for the experiment presented here. Compared to stereo versions of the same music, the binaural production showed a small but significant negative effect on OLE.

Due to the nature of this web-based study, there was no control over who participated, the conditions they listened in, or their approach to performing the test, other than the written instructions provided and post-hoc analysis of their activity during the rating. The participants were a self-selecting sample, presumably interested in aspects of the study, whether binaural sound or the Tommies drama. These people may be more inclined to prefer binaural audio than the general population. A more controlled recruitment process would be expected to give a more representative sample of audience members. Survey responses



(a) Audio drama

(b) BBC programmes with headphones

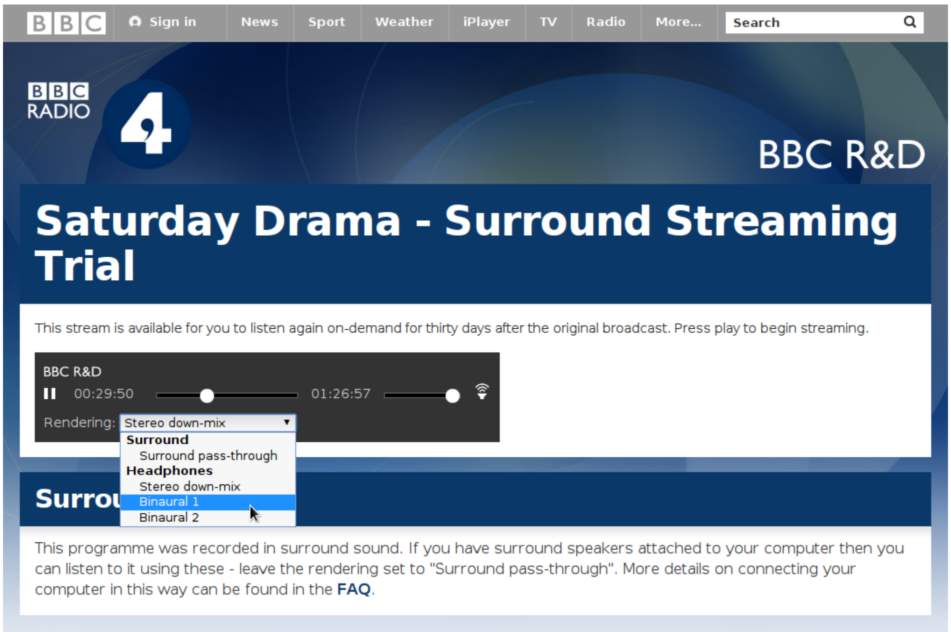
Figure 6.5: Listening habits of survey respondents, showing the number of people who listen with a given frequency to audio drama and to BBC programmes using headphones.

suggest that 51 % of headphone-wearing participants used over-ear headphones, whilst 25 % used in-ear and 24 % used on-ear types. In response to a question about their listening environment, 43 people said that they were in a quiet indoor environment, whilst one other said that they were in a noisy indoor environment. No participants reported being outdoors or in some form of transport, which were the other options. The remaining participants did not answer this question. It is likely that binaural services would be listened to in a wider range of environments, especially considering that headphones are often used with mobile devices.

Figure 6.5 shows the participants' reported listening habits, indicating how frequently they listen to audio drama and use headphones to listen to BBC programmes. At least in respect of these two factors, there was variation amongst the sample of listeners. However, there was a skew towards more regular listening in both respects, though none were in the "every day" category.

6.2.4 Web-based Evaluation of a Headphone Virtual Surround Sound Audio Drama

A second web-based study was run to evaluate a virtual surround sound reproduction of an audio drama. A 5.1 surround sound mix of the play Under Milk Wood by Dylan Thomas was produced by the BBC in 2003 and made available at the time as an audio download. Developments in open standards for streaming multichannel audio on the web enabled a trial of surround sound streaming to be run when the play was rebroadcast in 2014. Listeners with a 5.1 surround system which was connected to a device with a compatible web-browser



Saturday Drama - Surround Streaming Trial

This stream is available for you to listen again on-demand for thirty days after the original broadcast. Press play to begin streaming.

BBC R&D

00:29:50 01:26:57

Rendering: Stereo down-mix

Surround
Surround pass-through
Headphones
Stereo down-mix
Binaural 1
Binaural 2

Surrou

This programme was recorded in surround sound. If you have surround speakers attached to your computer then you can listen to it using these - leave the rendering set to "Surround pass-through". More details on connecting your computer in this way can be found in the **FAQ**.

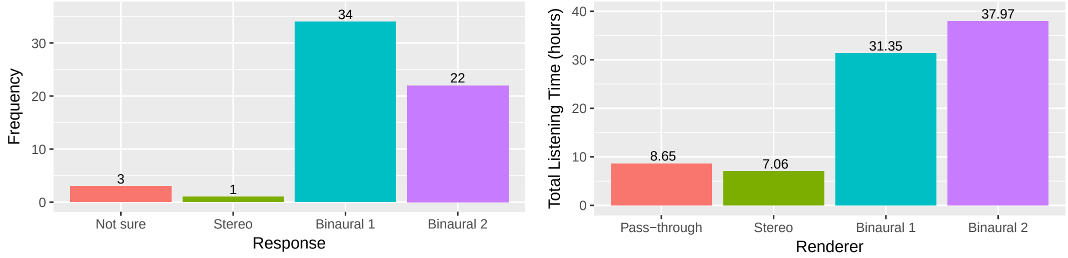
Figure 6.6: Web-based player for streaming the audio play Under Milk Wood and applying optional headphone surround sound processing.

could stream the programme and listen to the 5.1 surround mix directly.

Implementations of the Web Audio API (W3C, 2018) were becoming more common at this time. This enabled real-time client-side binaural audio to be rendered by impulse response convolution. A custom web player was built to optionally apply HSSP to the 5.1 signal in the browser client, to generate a headphone surround sound signal for the listener. Two different options were provided, HRIR rendering and BRIR rendering, using Neumann KU100 impulse response measurements. The development of the survey interface and analytics software was performed by a colleague.

This production of Under Milk Wood is distinctive. The dialogue of various characters is placed in single loudspeaker channels, rather than being panned between loudspeaker pairs as is common. A lot of foreground material is also placed in the rear surround channels, which is different to conventions of 5.1 surround sound production for television. Informal listening showed that the HSSP gave pronounced effect with this production.

The web player (shown in figure 6.6) was configured to offer the listener multiple headphone processing options: stereo down-mix, binaural 1 (HRIR), and binaural 2 (BRIR). Each of these options was applied using the Web Audio API audio processing graph nodes. Additionally, there was a default pass-through option which bypassed the Web Audio API and



(a) Preferences of survey respondents

(b) Time spent listening to each version

Figure 6.7: Results of Under Milk Wood study for those who listened to both binaural versions and at least one of the pass-through or explicit down-mix options.

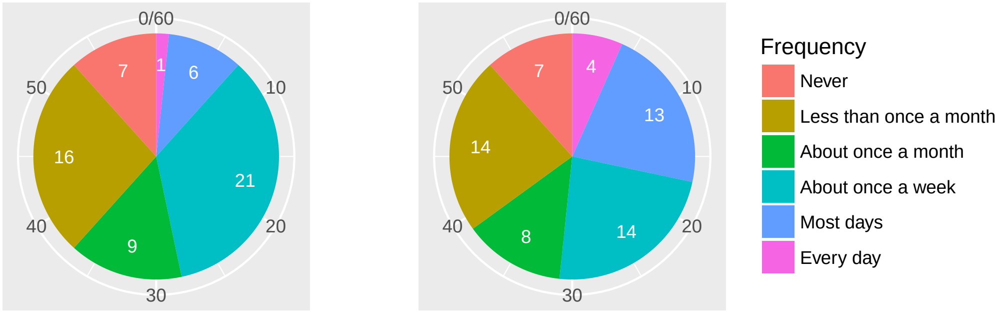
played the 5.1 signal directly from the audio element in the web page, allowing the browser to handle interpretation of this signal. Where a 5.1 surround reproduction system was not available the web browser will have generated a stereo down-mix.

A survey form was presented on the player page to solicit feedback from the audience and player analytics were recorded to monitor how the player was used. Participants were asked which their preferred rendering option was, along with the same questions described in section 6.2.3. The trial was promoted on the BBC Radio 4 homepage as well as through blog posts and social media. Over the course of the trial, 1385 page visitors played at least 5 min of audio[1]. There were 240 survey responses. Of these, 60 reported that they had used headphones and had listened to both binaural rendering options and either the pass-through or stereo down-mix options.

Figure 6.7a shows the preferences for each rendering option from the survey responses of those 60 listeners: 56 preferred either one of the binaural versions, whilst three did not know and only one preferred stereo. Whilst there are significant differences between rendering options (χ^2 goodness-of-fit test, $\chi^2(3) = 50, p < .001$), the differences between binaural options are not significant (binomial exact test, $p = 0.141$). As an additional indicator of which option may have been preferred, Figure 6.7b shows the total amount of time in hours that was spent listening to each rendering option by those 60 listeners. The binaural versions were listened to for much longer, with the BRIR version listened to for the longest duration. Of course, since the stereo version of the programme could be heard on the radio, most listeners would be visiting this page to listen to the binaural options.

As with the previous web-based study, there was no control over who participated and it is likely that many will have done so due to interest in binaural audio and/or the programme

[1] There were 15 111 page visits in total.



(a) Audio drama

(b) BBC programmes with headphones

Figure 6.8: Listening habits of survey respondents, showing the number of people who listen with a given frequency to audio drama and to BBC programmes using headphones.

content. Of the 60 participants who listened to both binaural versions and either the pass-through or explicit stereo down-mix, 40 reported using over-ear headphones, whilst 10 used on-ear and 10 used in-ear headphones. 56 reported listening in a quiet indoor environment and 4 in a noisy indoor environment. Figure 6.8 shows the participants' reported listening habits, indicating how frequently they listen to audio drama and use headphones to listen to BBC programmes. Again, as with the previous study, there seems to be a range of listening habits in the sample of listeners.

It should be noted that, in this case, the participants knew which version they were listening to. The rendering options were not anonymised. This may have created a bias in the responses.

The experiment reported in chapter 4 showed that HSSP was not considered to have substantially better quality than a stereo down-mix. The results are much more positive in this case. The methods of the two studies are substantially different, in this study assessors were rating preference, not quality, which may have contributed to the differing outcomes. However, with this programme content, the HSSP effect appeared very effective to this author and colleagues. It is proposed that this is not due to the specific binaural processing applied[2], but the production approach used for this programme. The lack of amplitude panning of sources appears to give better spatial impression, and more frequent use of the surround channels for prominent sound elements draws attention to the binaural effects.

[2] This author is making no claim of improving the state-of-the-art in this respect.

6.2.5 Summary

The two web-based studies reported in this section indicate that binaural programme content may be viewed positively by broadcast audiences. In both studies there was no control over who participated or what conditions they listened in. It is likely that, since they participated, the listeners had an interest in binaural audio. Despite this lack of control, these web-based studies give valuable insights. They have higher external validity than laboratory-based quality evaluations, since the listeners are in representative environments and are themselves representative members of the target audience.

In the first study reported, listeners rated their preference for either a binaural or stereo version of an audio drama. The binaural version was created through a dedicated production; a professional sound engineer applied a range of processing options to different elements within the scenes. Significant preferences for the binaural versions were found.

In the second study, listeners showed significant preferences for HSSP of an audio drama in the 5.1 surround sound format, compared with a stereo down-mix. Listeners were divided over whether use of HRIRs or BRIRs was preferred. This programme was distinctive in that the dialogue was frequently routed to a single loudspeaker channel, rather than making use of amplitude panning, and the surround channels were used more heavily for foreground scene elements than is common in broadcast production.

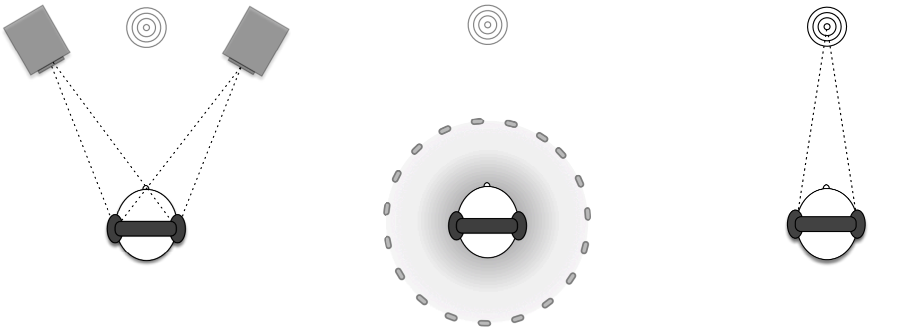
More generally, it appears that adapting production techniques to create binaural audio is likely to give improved results, compared with the post-processing of 5.1 surround sound evaluated in chapter 4. This approach has been taken a number of times at the BBC, using binaural rendering to produce alternate mixes of programmes specifically for headphone listening. This output has been received well by audiences. However, the challenge with this approach is that it requires significant additional production effort. It also requires distribution of multiple versions of the same programme to different audience members, which adds complexity. Approaches that allow simultaneous creation of programme output for headphones and for loudspeakers are therefore appealing. Section 6.3 introduces recent developments in formats and standards for 3D spatial audio production and delivery that might provide these advantages. As will be discussed, it is common for loudspeaker virtualisation approaches to be used in these 3D spatial audio technologies. The experience with the Under Milk Wood production suggests that the effects on perceived quality of amplitude panning to virtual loudspeakers should be investigated further.

6.3 3D Spatial Audio Formats and Standards

Since the initial experiments of this project, standards for representing, distributing, and reproducing 3D spatial audio content have been established, and applications have grown in popularity. In 2014, the ITU-R defined an *advanced sound system* for broadcasting (ITU-R, 2014a). It is capable of representing content with audio signals plus static or dynamic metadata. The content representation can comprise constituent elements in channel-based, scene-based and/or object-based formats.

- *Channel-based audio* – The scene is represented by a set of signals which correspond to loudspeaker positions.
- *Scene-based audio* – The scene is represented by a set of coefficient signals which together describe a spatial scene, but which are not associated with loudspeaker positions, e.g. ambisonics.
- *Object-based audio* – The scene is represented as separate sound sources with spatial parameters that can vary over time.

These will be explained in more detail in the following sections. An illustration of the use of each of these formats to represent a single virtual source for a headphone listener is given in figure 6.9.



(a) Channel-based (b) Scene-based (c) Object-based

Figure 6.9: Illustration of the use different spatial audio representations to create a virtual sound source for a headphone listener.

Such systems and content representations allow adaptation at the reproduction stage, e.g. accounting for the available reproduction system or adjusting the balance between elements. From a user perspective, advanced sound systems provide 3D spatial audio, but also offer personalisation of the audio presentation, such as dialogue enhancement. The adaptation capability avoids duplication in production effort, while delivering to a range of reproduction systems. Various loudspeaker configurations are supported, including 3D with-height configurations. The most complex configuration specified is the 22.2 layout (Hamasaki et al., 2005), termed system H or 9+10+3 by the ITU-R.

In wider industry use, the term *next-generation audio (NGA)* is used to describe such advanced sound systems and the features they offer to users. Several standardised NGA systems are available, including MPEG-H 3D Audio (ISO/IEC 23008-3:2015, 2015), Dolby AC-4 (ETSI TS 103 190, 2014), and DTS-UHD Audio (ETSI TS 103 491, 2017). These provide capabilities in line with the ITU-R's advanced sound system, with associated audio coding technologies. These NGA systems are included in standards for broadcast- and Internet-delivery of media by the DVB organisation (ETSI TS 101 154, 2018) and HbbTV Association (2018), respectively.

Recommendation ITU-R BS.2076 (ITU-R, 2017b) defines the Audio Definition Model (ADM), which is a data model capable of describing audio content and formats associated with NGA systems. The ADM enables interoperability throughout production, archive and distribution processes. The EBU ADM Renderer specification was published in EBU Tech 3388 (2018). It defines algorithms for rendering content specified using the ADM to the loudspeaker layouts in Recommendation ITU-R BS.2051. This is currently being considered for standardisation by the ITU-R, to provide a common interpretation of the ADM parameters for loudspeaker monitoring in broadcast applications.

NGA standards are beginning to be used in media services. In 2017 South Korea launched an ultra-high definition television service with MPEG-H 3D audio (Business Wire, 2017). Netflix, an Internet-streaming media service, presents content in *Dolby Atmos*[3], which is delivered by producers in an ADM Broadcast Wave Format (BWF) file (Netflix, 2018), and to consumers using E-AC-3 (ETSI TS 102 366, 2017). Many broadcasters and other media service providers are considering the adoption of NGA, including the DTG (2018) in the UK.

Virtual reality (VR) and augmented reality (AR) applications have become very popular in recent years, with huge investment (VRVCA, 2018) and a great deal of consumer adoption (YouGov, 2017). The increasingly widespread availability of these head-mounted display (HMD) devices, which provide real-time head tracking data and use headphones for sound reproduction, enables mass-audience applications of dynamic binaural rendering. Many of the devices currently available are relatively low-power mobile/untethered systems, so pro-

[3] Dolby Atmos is a branding term to describe 3D spatial audio services, but is not specifically linked to a codec. It could be delivered with E-AC-3 or AC-4, though there are no known consumer implementations of AC-4 to date.

cessing complexity must be seriously considered.

The ITU-R definition of an advanced sound system arguably also applies to VR and AR, where the same content representations and codec systems are often used. The 3GPP has recently adopted MPEG-H 3D Audio for use in VR streaming (3GPP TS 26.118, 2018). Ambisonics is now used alongside omnidirectional video (often known as 360° video) on major social media platforms such as Facebook (Facebook, 2018b) and YouTube (YouTube, 2018) to deliver VR. Because many VR and AR applications are delivered as standalone software packages and the experiences are often interactive games, codec systems for streaming are not always necessary or appropriate. However, most of the common software development kits (SDKs) for audio in these applications support at least object-based and scene-based audio formats, such as the Google Resonance SDK (Google, 2018b), the Oculus Audio SDK (Oculus, 2018) and the Steam Audio SDK (Steam, 2018).

6.3.1 Channel-based Audio

The term *channel-based audio* simply means that the audio content has been mixed into a pre-determined number of channels that correspond to pre-determined loudspeaker positions. This includes the existing two-channel stereo and five-channel surround sound formats. It is how current broadcast programmes are made. For 3D spatial audio, more channels are transmitted, corresponding to additional loudspeakers above, and sometimes below, the listener's head height.

Recommendation ITU-R BS.2051 (ITU-R, 2018) defines several loudspeaker layouts that should be supported by advanced sound systems. This ranges from two-channel stereo (0+2+0 or system A) to 22.2 (9+10+3 or system H). There are other 3D with-height configurations, such as system J (4+7+0), which has four loudspeakers above head-height and seven at head-height. NGA systems utilise metadata to indicate the loudspeaker configuration that channel-based audio signals correspond to. Where this does not match the available configuration of speakers during reproduction, a conversion process may be applied e.g. down-mixing.

Channel-based signals may be generated by amplitude panning of sound sources, as well as microphone array recordings. The long-established techniques for two-channel stereo reproduction have been extended to surround and 3D surround-with-height loudspeaker arrays. Appendix B.3 gives details on amplitude panning, including the extension to 3D loudspeaker arrays using vector base amplitude panning (VBAP).

6.3.2 Scene-based Audio

Scene-based audio refers to formats where the spatial audio content is represented by a set of coefficient signals, which is independent of the loudspeaker configuration used for reproduction. Such representations are also sometimes described as transform-based, as a mathematical transformation is required to obtain loudspeaker signals, or other directional signals considered as plane-wave- or point-sources, for reproduction.

In practice, scene-based audio refers to ambisonics, where the signals are coefficients of the spherical harmonics basis functions. Ambisonics was first developed by Gerzon (1973), using the four-channel first-order approach, which is associated with the sound field microphone (Gerzon, 1975). It was later expanded to higher-order representations e.g. by Daniel (2000), which use a greater number of channels to represent the scene, dependent upon the maximum order used.

Ambisonic signals have to be "decoded" by a transformation, to obtain signals for the reproduction loudspeaker configuration. The advantage of this process is that ambisonics is somewhat independent of the reproduction layout. NGA systems utilise metadata to describe the set of ambisonics signals, including the order and normalisation. This allows for appropriate decoding to take place for reproduction. A more detailed description of ambisonics methods is given in appendix B.4.

6.3.3 Object-based Audio

In *object-based audio*, the audio scene is represented as separate sound sources. The source audio signals are accompanied by static or dynamic metadata. The metadata normally define spatial position, but may also include spatial extent, diffuseness, and interactivity-related parameters, as in the ADM (ITU-R, 2017b). Object-based audio requires rendering for reproduction, but keeping sound sources (*objects*) separate gives greater flexibility than with other formats, making adaptation to the reproduction layout or personalisation of presentation more straightforward. Modelling the individual objects can be an efficient means of representing sound scenes when compared to the channel-based approach in cases where many loudspeakers are used, such as wave-field synthesis (Melchior, Sladeczek, Partzsch, et al., 2010). In consumer delivery applications, for efficiency, often a small number of objects are used in conjunction with a *bed* element, containing most of the sound scene, which might be channel-based or scene-based (Herre, Hilpert, et al., 2015).

Sometimes object-based audio is used synonymously with next-generation audio. In this case, channel-based and scene-based elements are considered to be objects, alongside the separate sound source objects. Within the ADM a piece of content comprises one or more `audioObjects` and these reference format definitions of various types, including channel-based, scene-based and object-based scene representations. Unfortunately the industry has not yet converged on clear common terminology. In this chapter, *element* refers to a component part of the content which could be of any format, and *object* refers specifically to a single sound source within the scene with defined (and potentially time-varying) spatial parameters, i.e. the object-based format type.

Object-based audio is often used in the domain of broadcast applications to differentiate from traditional channel-based approaches. In more interactive applications such as gaming, the term is not as common, but the approach of representing separate sources within the scene is the common practice. In interactive virtual acoustic environment (IVAE) applications, the object-based representation is essential, to allow exploration of and interaction with the scene. Whilst sharing the elements described here, IVAE systems often incorporate more complex models, including source directivity and simulation of room acoustics, as well as more sophisticated user interactions (Savioja et al., 1999). The MPEG-4 standard includes a scene description format with high-level parameterisation of environmental reverberation and associated rendering techniques for IVAEs (Scheirer et al., 1999).

6.3.4 Headphone Delivery of 3D Spatial Audio Formats

Recommendation ITU-R BS.2051 was revised in 2018 (ITU-R, 2018), adding headphones to the list of reproduction configurations. This was to reflect the rise in consumption of audio on mobile devices, and the associated increase in use of headphones for reproduction. The ITU-R also acknowledges an increase in production by broadcasters of spatial audio content specifically for headphones using binaural techniques, including in VR and AR. As yet, there is no initiative to standardise a common approach to headphone rendering of NGA formats defined with the ADM, unlike the ongoing initiative to standardise loudspeaker rendering in the ITU-R (ITU-R, 2015a).

To create an immersive sound scene, 3D spatial audio content often includes many different sound sources with distinct spatial positions. Considering headphone reproduction, each object can be separately rendered to a binaural signal, as far as possible according to the spatial characteristics defined in the object metadata. By summing the resulting signals a binaural output signal can be delivered to headphones to create an auditory scene for the listener. Dynamic rendering with listener tracking allows for interactive auditory virtual environments (IAVEs) to be constructed[4]. Where the object parameters include only spatial

[4] Though of course for a truly interactive environment, the environment also needs to respond to the actions of the user.

position, the basic binaural rendering system described in appendix A.3 can be used. Each object is rendered separately by convolution with a binaural impulse response. In this case the rendering complexity is highly variable, as is the data required for the scene representation. Both are dependent on the scene complexity. For applications such as broadcast or streaming to mobile devices this is undesirable.

Alternatively, the scene can be pre-rendered to a binaural signal, as investigated in section 6.2 and chapter 4. This approach has the same distribution bandwidth requirements as two-channel stereo. It also requires no rendering process at the consumer-end, besides perhaps headphone correction filtering which is rarely used in practice in consumer applications. However, this approach precludes dynamic rendering according to listener movement and any further interaction with the content, as well as head-related transfer function (HRTF) individualisation[5]. Also, the content can then only be used for binaural delivery. The adaptation and personalisation benefits of NGA systems cannot be utilised with this approach.

There are established methods for rendering a spatial audio scene to an array of loudspeakers. Applying these techniques prior to distribution of the content limits the required data, using a fixed number of audio channels. For complex content this channel-based representation may use less data than when representing the scene as component objects. The loudspeaker array can then be virtualised on the receiving-end for headphone playback, i.e. each loudspeaker is rendered as a virtual sound source using binaural techniques. The binaural rendering complexity is then limited to processing these channels, irrespective of scene complexity.

The experiment presented in chapter 4 found that reproduction of 5.1 surround signals by loudspeaker virtualisation did not give substantial quality enhancement over a stereo down-mix. It is likely that virtualisation of 3D loudspeaker layouts will allow better quality than with the 3/2 horizontal layout, known as system B (0+5+0) in (ITU-R, 2018). The spatial impression could be expected to improve when 3D loudspeaker layouts with more channels are used, since more spatial information will be available in the input format.

In current NGA systems, the virtual loudspeaker rendering approach is used for headphone output (Herre, Hilpert, et al., 2015; Dolby, 2015). Although object-based signals can be delivered to the end device, headphone rendering of spatial audio is still applied as a post-processing step. Sound sources are first rendered to a 3D loudspeaker configuration using amplitude panning and then the loudspeaker array is virtualised using binaural processing. The approach is appropriate because, besides limiting the complexity of binaural

[5] The term *individualisation* is commonly used in relation to binaural rendering, though it is synonymous with personalisation.

rendering, the majority of existing media content is still channel-based and the same binaural rendering structure can then be used for both channel- and object-based signals. As mentioned in section 6.3.3, a channel-based element, containing the majority of scene content, is often combined with a small number of objects (Herre, Hilpert, et al., 2015). This enables efficient delivery of complex scenes, reducing signal bandwidth, whilst still permitting some personalisation of the audio presentation. Loudspeaker virtualisation techniques also permit efficient binaural rendering of more complex sound source models, including aspects such as directionality, non-zero extent and diffuseness (Jot, Walsh, et al., 2006).

Appendix B.2 describes a binaural renderer that has been developed as part of the experimental apparatus of this thesis. It first converts an object-based scene representation to a channel-based format by using a loudspeaker rendering technique and then generates a headphone output signal by binaural rendering of this virtual loudspeaker array.

Similar approaches can be applied with scene-based ambisonics representations. An object-based representation can be rendered to an intermediate scene-based ambisonics representation, either before distribution or within the reproduction renderer. Often the ambisonics signals are then decoded to an array of virtual loudspeakers for binaural rendering, as described in 3GPP TS 26.118 (2018, Annex B). Ambisonics is used as an intermediate virtualisation format to reduce rendering complexity in mobile VR and AR systems, such as the Google Resonance Audio SDK (Google, 2018b) and the Facebook Audio360 SDK (Facebook, 2018a). This approach is attractive because efficient scene rotation is possible in the ambisonics domain, enabling dynamic listener tracking. Some NGA systems, such as MPEG-H 3D Audio, also support use of ambisonics beds (Herre, Hilpert, et al., 2015; 3GPP TS 26.118, 2018).

A dynamic renderer to convert ambisonics into a binaural signal has also been implemented in the experimental apparatus, it is described in appendix B.4.6. In contrast to the previously mentioned binaural renderers, head rotation is applied in the ambisonics-domain rather than during binaural rendering.

Distribution of content in channel-based or scene-based formats largely limits the application of listener tracking to orientation only (3 DoF). If an object-based scene representation is rendered to a virtual loudspeaker layout on the reproduction device, then distance effects may be applied before mixing to the intermediate format to be virtualised. This allows virtualisation to be used even in IAVEs with scene interaction and 6 DoF listener movement, whilst limiting the rendering complexity.

Since loudspeaker virtualisation techniques are widely used, the background to this approach is next discussed in more detail.

6.4 Loudspeaker Virtualisation for Headphones

Loudspeaker virtualisation is a common application of binaural rendering. It enables reproduction on headphones of multichannel audio signals that were authored for loudspeakers. The aim is to improve the listening experience compared to a conventional stereo down-mix by reproducing the spatial imaging that would be present in the loudspeaker reproduction. The most common and straightforward approach is to render each loudspeaker signal as a virtual sound source at the intended loudspeaker position relative to the listener, as illustrated in figure 4.1 (page 191) and defined here mathematically as a summation of time-domain convolution operations:

$$p_{lr}(t) = \sum_{l=1}^{\mathsf{L}} x_l(t) * h_{lr}(t, \theta_l, \phi_l), \tag{6.1}$$

where x_l is the signal for the l^{th} loudspeaker, $h_{lr}(t, \theta_l, \phi_l)$ is a HRIR for the direction of the l^{th} loudspeaker and p_{lr} is the rendered binaural signal.

The virtualisation approach was proposed as early as 1961 by Bauer (1961b). It was acknowledged that two-channel stereo signals are designed for reproduction over spaced loudspeakers in front of the listener, and that direct reproduction of these signals over headphones leads to distorted imaging due to inappropriate interaural time and level differences. Electronic circuits were proposed for headphone reproduction of stereo recordings, introducing inter-channel cross-talk with delay and filtering in an approximation of the relevant HRTFs.

The use of DSP techniques to convert content in the well-established 5.1 multichannel surround format to a binaural signal was introduced in the 1990s, e.g. (McKeeg and McGrath, 1997). These are often called virtual surround sound or headphone surround sound processing (HSSP) systems. This approach has been popular in computer game systems and for movie playback on mobile devices for some time. Many broadcasters currently distribute television programmes with 5.1 surround, and a significant proportion of audiences are watching these programmes on portable devices and listening using headphones. Virtual surround sound allows spatial audio to be offered to headphone listeners without changing production workflows or remixing archive content specifically for headphones. However, most devices currently use a stereo mix for headphone output, be it a separate artistic mix or a down-mix from the 5.1 signal. As introduced in section 6.3, loudspeaker virtualisation is also used widely with 3D spatial audio formats. Appendix B gives more background to these techniques and the experimental apparatus for investigating loudspeaker virtualisation.

6.5 A Review of the Perceived Quality of Amplitude Panning and Loudspeaker Virtualisation for Headphones

Loudspeaker virtualisation is used widely for headphone delivery of spatial audio, including in new and emerging 3D audio systems; yet there has been relatively little detailed study of the effects of this approach on perceived quality, particularly in comparison to direct binaural rendering of each source.

This section presents a review of related studies that give insights into what the effects might be. Section 6.5.1 reviews existing research into the perceptual effects of amplitude panning over loudspeakers. Several studies have used binaural simulations to study loudspeaker rendering. To validate the methods, the perceptual effects observed with real loudspeakers have been compared to those observed with the binaurally-rendered simulations, with a focus on localisation and colouration aspects. These are described in section 6.5.2. Ambisonics rendering appears to have been studied more often than rendering of 3D channel-based formats to headphones. Section 6.5.3 reviews studies to evaluate binaural rendering of ambisonics signals, particularly comparing performance at different ambisonics orders. Chapters 7 and 8 will present experiments carried out to evaluate the perceptual effects of loudspeaker virtualisation techniques in more detail, with comparison to direct binaural rendering of sound sources and scenes.

6.5.1 Perceptual Effects of Amplitude-Panning over Loudspeakers

Pulkki and Karjalainen (2001) investigated the localisation of amplitude-panned virtual sound sources on a stereo loudspeaker setup using a series of listening tests, as well as with an auditory model. They found that the perceived location of an amplitude-panned sound source is dependent on the spectral and temporal structure of the source signal. The method of adjustment was used, with assessors changing the target position of amplitude panning in 1° steps. Three types of source signal were used: broad-band pink noise, filtered pink noise and filtered impulse trains, using one-third-octave and two-octave band-pass filters at a range of centre frequencies. Real loudspeakers were placed at $\pm 15°$ and initial virtual source target positions were randomised. The auditory model provided estimates of the perceived localisation angle separately for interaural time difference (ITD) and interaural level difference (ILD) cues. Results showed that low-frequency ITD and high-frequency ILD cues approximately match in direction, explaining why sound sources are localised fairly consistently across the frequency range. In the mid-frequency range (1 kHz–2 kHz), however, the directions suggested by these cues differ significantly. Both cues indicate angles greater than the target panning position but the angle implied by the ILD cues is significantly larger than that for the ITD cues. The perceived direction in this frequency range is influenced by both ILD and ITD cues and their relative prominence depends on the nature of the source signal, with ITD cues more dominant for the impulse trains. The deviations in this mid-frequency range were only found to affect localisation for narrow-band sources but Pulkki and Karjalainen suggest that this could lead to a spread in the perceived width of the virtual source for broadband signals.

Toole (2008, p. 121) suggests another reason for an increase in auditory source width with a phantom centre image produced by a stereo pair of loudspeakers. When in a reverberant listening room, amplitude panning of speakers at ±30° causes an increase in lateral early reflections occurring in the listening room, which are known to increase apparent source width. Toole refers to Choisel and Wickelmaier (2007) who, as part of a wider evaluation of the quality of multichannel sound systems, reported that an increase in width and spaciousness was observed when comparing a phantom source to a real loudspeaker.

Pulkki (2001b) applied a similar methodology to a previous study (Pulkki and Karjalainen, 2001) to investigate the effects of amplitude panning in the median plane with a vertically separated pair of loudspeakers, and in three-dimensions with triplets of loudspeakers using VBAP. These experiments used pink noise and octave-band-filtered noise signals. For vertical panning in the median plane, elevation localisation was found to vary significantly amongst individuals, yet the median responses were reasonably close to the real source elevations. Modelling of spectral cues using individual HRTFs suggested that amplitude panning significantly distorts spectral peaks and notches that might form cues for elevation localisation, yet some assessors were still able to localise the source elevation consistently. It was hypothesised that changes in the loudness of a pinna mode frequency relative to adjacent frequencies can be used by assessors to estimate the source elevation. The tripletwise panning experiments showed that assessors were consistently able to estimate source azimuth, though biased towards the median plane compared to the real source. Results showed again that median elevation angles were quite close to the target real source elevations, with high inter-subject variance but relatively low intra-subject variance. Auditory modelling was also used to investigate the effects on spatial hearing cues for a larger range of loudspeaker triplets and pairs at different directions around the listener. For both pairwise and tripletwise amplitude panning of sound sources at lateral directions, the perceived source direction is biased towards the median plane. This appears mainly influenced by low-frequency ITD cues, whilst ILD cues are heavily distorted.

Earlier work by Theile and Plenge (1977) evaluated the localisation of phantom sources from pairwise amplitude panning with two laterally-located loudspeakers. Pulsed white noise signals were used with inter-channel level differences ranging between ±18 dB for each loudspeaker pair. The lateral displacement of the loudspeaker base centre angle ranged between 0° and 90°. Localisation was reported by adjusting an acoustic indicator loudspeaker on a movable arm to match the perceived position. It was also found here that source localisation was biased towards the median plane. As lateral displacement of the loudspeakers approached 90° the variability of localisation responses increased. The authors recommend that a loudspeaker arrangement designed for "all round effect" should use loudspeakers at ±90°.

Perceived elevation of sources generated using VBAP was further investigated by Baumgartner and Majdak (2015) using an auditory model. Specifically, localisation of virtual sources in sagittal planes was assessed using monaural spectral localisation cues. The effects of panning were examined with vertically-spaced loudspeaker pairs with varied target panning angle and loudspeaker span, as well as using the VBAP algorithm for a set of with-height loudspeaker arrays across a large range of lateral θ_{cc} (±45°) and polar ϕ_{cc} (0°–180°) angles. The spectral cues used to discriminate source elevation are not well produced during amplitude panning with vertically spaced loudspeakers and so localisation errors are common. It was found that localisation of elevated sources varies significantly between individuals, due to large variability between the HRTFs of individuals in the relevant frequency range ($>$700 Hz). In addition, when averaged across individuals, there were not consistent patterns in localisation of sources from different loudspeaker configurations and panning ratios. It was concluded that systems with fewer loudspeakers and larger vertical spans between loudspeaker layers yield poorer accuracy in localisation of polar angles. Loudspeaker arrays with a vertical span of 30° gave lower mean errors than those with a vertical span of 45°.

Pulkki (2001a) also studied the effect of amplitude panning on colouration, using both auditory modelling and listening tests. For a frontal virtual source with loudspeakers symmetrically placed either side (±30°) there is a difference in time of arrival (TOA) between the two loudspeaker signals at each ear, which creates timbral colouration with a pronounced spectral dip between 1 and 2 kHz. This frequency range is directly related to the inter-aural distance of the listener. Without the acoustic effect of the listener's head, this would create a comb filter at the position of the ears, with the frequency of the first notch lying in this range. Higher-order notches of the comb filter are not so apparent, however, because of attenuation of the ipsilateral loudspeaker signal due to head shadowing and other effects of the head-related transfer function.

In listening tests, participants adjusted the level of amplitude-panned virtual sources to match that of the corresponding source from a real-loudspeaker. Narrow-band stimuli were used at equivalent rectangular bandwidth (ERB) spacings in the range 0.2 kHz–8.5 kHz for a stereo pair at ±30° and a virtual source target position at 0°. In anechoic conditions participants applied a consistent gain of approximately 6 dB at 1.7 kHz. Variability amongst individuals was greater at higher frequencies. When a reverberant listening room was used, the 1 kHz–2 kHz dip was reduced due to the influence of the diffuse field, but there were larger changes at low frequencies, possibly due to differing excitation of room modes. The amount of colouration introduced by amplitude panning in a reverberant environment is dependent on the room itself. Since the auditory model corresponded well to listening test results, it was then used to assess colouration with a wider variety of loudspeaker setups and panning angles in anechoic conditions. Spectral changes are most significant when the target position is at the centroid of the loudspeaker positions, where they have equal panning gains. The colouration effect will vary with sound source direction and head orientation. As the listener is oriented further away from the virtual source direction, the path length difference from each loudspeaker to the ears is reduced and so the frequency of the spectral notch is increased and the magnitude of the notches is reduced slightly. The colouration effects due to amplitude panning may also change the perceived elevation of sound sources by altering spectral cues in the range 4 kHz–12 kHz. When triplet-wise panning is used, it was shown that colouration effects are more pronounced.

Shirley et al. (2007) observed a loss of speech intelligibility as a result of using a phantom source. The study involved a test of speech perception amongst background multi-talker babble, with listeners tasked with identifying keywords at the end of sentences. Babble was played over a stereo loudspeaker pair at ±30° azimuth whilst the main speech was either reproduced with a real loudspeaker at 0° or as a phantom centre image through the stereo pair. There was a significant 4.1 % increase in number of keywords correctly identified when using the real centre source. It appears that the characteristic dip in the magnitude response due to amplitude panning is in a frequency range important for speech intelligibility.

In summary, amplitude panning appears to induce errors in localisation, particularly for lateral and elevated target directions and there is considerable variability in the magnitude of localisation errors between individuals. It is also expected that there will be increases in source extent and colouration effects (predominantly due to a mid-frequency spectral dip), which in turn may cause speech intelligibility issues.

6.5.2 Evaluation of Loudspeaker Rendering Techniques using Non-Individualised Binaural Simulation

Binaural simulation has been used to evaluate the effects of loudspeaker rendering, allowing the experimenter to vary factors such as listener position and loudspeaker array during the experiment. Some studies have attempted to validate the use of non-individualised binaural simulation for such investigations, which may reveal insights into the differences between panning on loudspeakers and virtual loudspeakers over headphones.

Wierstorf, Raake, and Spors (2013) evaluated perception of virtual sources rendered using wave-field synthesis (WFS) in terms of localisation, with a non-individualised dynamic binaural rendering system. The use of binaural simulation was first validated by asserting that the localisation blur was equivalent to that for real sources (Wierstorf, Spors, and Raake, 2012), using a head pointing task during sound presentation. The standard deviation of localisation responses was equivalent when rendering using BRIRs ($\sim$2.2°), whereas for HRIRs it was 1.8° higher. Task completion time and the number of head turning points were also equivalent with BRIRs and significantly higher with HRIRs, indicating that the auditory and sensorimotor cues were not as clear in the latter case. The same binaural system was used to evaluate colouration in (Wierstorf, Hohnerlein, et al., 2014), though without head tracking. The authors acknowledged that a non-individualised system will cause some timbral effects, but if these can be limited and made constant across tested virtual loudspeaker systems, then it was assumed that relative differences in colouration can be attributed to the systems under test. In Wierstorf, 2014, section 4.4 this was investigated objectively by means of two different head and torso simulators (HATSs). One was used to measure the binaural impulse respones used in the simulation, and a different HATS was used for recording the reproduction both by real loudspeaker systems and the binaural simulation of those. Deviations in magnitude response were ±5 dB up to 5 kHz and as high as 15 dB above this frequency, but there was no systematic change due to the loudspeaker system. This was taken to validate the use of non-individualised binaural simulation for investigating differences in colouration between loudspeaker systems.

Satongar (2016) studied the perceived effects of amplitude panning techniques for horizontal loudspeaker arrays, with a particular focus on azimuthal localisation and colouration. The work assesses the validity of using a non-individualised dynamic binaural rendering system in listening tests to understand the effects of loudspeaker rendering. Specifically, it investigates whether the use of non-individualised binaural rendering to simulate loudspeaker reproduction has a significant impact on the measurement of localisation and colouration effects of various panning techniques.

Two separate validation studies were carried out. Satongar (2016, Chapter 6) compared localisation errors during amplitude panning of horizontal loudspeaker arrays at central and off-centre listening positions using both real loudspeakers and the binaural rendering system. Tests were carried out in a reverberant environment and so BRIRs were used for dynamic binaural rendering. As in (Wierstorf, Spors, and Raake, 2012), a head pointing task was used during sound presentation. This allows users to home in on the sound location by minimising interaural differences. Additionally, task completion duration and number of head turning points were used to indicate the ease of localisation. Taken over 8 different reproduction scenarios (including single speakers, VBAP and ambisonics), the average absolute localisation error was 4.0° at the central listening position and 5.7° off-centre. The use of binaural simulation did not show significant effects on localisation error overall, but some differences were present for specific conditions. Equivalence boundaries were defined at ±7°, based on typical minimum audible angles (MAAs) at azimuths of about 50°, and the localisation errors between the simulation and real loudspeakers were equivalent for 15 of 20 conditions. Scenarios where deviations in absolute localisation errors were high were those with low loudspeaker density and therefore wide base angles in panning. Localisation precision is low with such systems (Pulkki, 2001c, p.29), which will have contributed to these observations, but it appears also that systematic differences occurred. The binaural simulation may not have adequately reproduced the cues used by the listeners in this case, such as changes during small head translation (only orientation tracking was supported) and monaural spectral cues (due to non-individualised rendering).

Satongar (2016, Chapter 8) investigated differences in colouration detection thresholds (CDTs). Results did not show consistent reduction or increase in colouration acuity when using a non-individualised dynamic binaural rendering system compared to real loudspeakers. A feed-forward comb filter was used to introduce timbral changes. CDTs measured for the binaural simulation were within ±4 dB of those for real loudspeakers, which is significantly lower than the inter-subject range in CDTs and small relative to typical CDT levels.

It appears from these studies that the use of non-individualised binaural rendering will introduce relatively small additional colouration and localisation errors over real loudspeaker reproduction. These errors appear not to be consistent across individuals, as might be expected due to individual deviations from the HRTFs used in rendering. It appears that differences between loudspeaker techniques are not changed systematically by the use of binaural virtualisation, except for very sparse loudspeaker arrangements.

6.5.3 Evaluation of Binaural Rendering of Ambisonics

There have been a number of studies to evaluate the rendering of ambisonics signals to headphones with binaural techniques. Thresh et al. (2017) compared localisation acuity for 1st, 3rd, and 5th order ambisonics on both loudspeakers and headphones. Binaural loudspeaker virtualisation was performed using BRIRs measured with the Neumann KU100 dummy head, head tracking was applied in the ambisonics domain. Target source directions covered a large range of azimuth and elevation angles. There was a large reduction in angular error when going from 1st to 3rd order but no significant reduction in errors between 3rd and 5th order. Increased variance was observed in localisation responses for the binaural rendering compared with loudspeaker rendering, which was largely influenced by front-back and up-down confusions. The authors attribute this to the lack of personalisation of the binaural filters.

Kearney and Doyle (2015b) investigated the reconstruction of elevation cues in binaural rendering of ambisonics using objective analysis with KEMAR HRTFs. Spectral errors were observed in the median plane and the sagittal plane localisation model of Baumgartner and Majdak (2015) was used to indicate probability of localisation errors. The analysis showed spectral errors with all tested ambisonics decoders, though these errors reduced with increasing ambisonics order. At 5th-order, the auditory model suggested that rendering could adequately synthesise elevation cues.

Kearney, Gorzel, et al. (2012) investigated distance perception in binaural rendering of ambisonics signals, with comparison to distance estimation accuracy for real sound sources. 1st-, 2nd- and 3rd-order ambisonics representations of room impulse responses were used in rendering, synthesised from 1st-order microphone measurements made at source distances in the range 2 m to 8 m, using the methods of Merimaa and Pulkki (2006). Individual HRTF measurements were used for rendering. With 1st-order rendering, distance estimation was already equivalent to that for real sound sources. This study used sources in front of the listener and head tracking was utilised with ambisonics-domain sound field rotation.

Kearney, Liu, et al. (2015) investigated distance perception in binaural rendering further. Both static and dynamic rendering scenarios were evaluated. An initial study with real sound sources demonstrated that distance estimation did not vary significantly with source angle. The binaural rendering experiments used frontal sources and non-individual HRTFs. For static binaural rendering, BRIRs were measured with a KEMAR HATS. For dynamic rendering, 1st-order ambisonics room impulse responses (RIRs) were measured, and ambisonics-to-binaural rendering was performed, with head tracking again applied as a sound field rotation. In both cases, listeners' distance estimation was equivalent to that for real sources, although with static binaural rendering front-back reversals were common.

These studies suggest that ambisonics-based binaural systems are likely to achieve reasonably good localisation quality when using higher orders ($N \geq 3$). Little comparison to direct binaural rendering of sources has been performed, however, and insights into the impacts on other quality features, such as timbral aspects, are not available.

6.6 Summary and Outlook

This chapter has discussed the influences of content production techniques and spatial audio signal representations on the quality of binaural technology applications. Following the pilot study of chapter 4, the approach of distributing pre-rendered, non-individual binaural signals was further investigated. Section 6.2 has shown that with appropriate use of binaural rendering in programme production, the headphone listening experience can be improved, even without head-tracking and individualisation.

There have been developments in spatial audio technology and its application since the beginning of this project, in particular NGA codecs for efficient content delivery and mass-market VR and AR devices. This makes 3D spatial audio services feasible for large audiences, including the provision of interactive client-side binaural rendering on mobile devices.

Applications of spatial audio to entertainment media are still developing. A range of spatial audio content representations are supported in these technologies, broadly categorised as object-based, channel-based and scene-based formats. Given the importance of headphone delivery for providing audiences with spatial audio experiences, insight into the quality of binaural rendering of these different formats is required. Loudspeaker virtualisation is used widely in these applications, yet there has been relatively little detailed study of the effects of this approach on perceived quality, or comparison of the quality offered by the different representations available.

In many scenarios, there are also constraints on distribution bandwidth and computational complexity that will influence the choice of format. For a broadcaster, or other media distributor, the delivery bandwidth has a significant impact in terms of cost and is constrained by overall service capacity and competing demands from other service aspects. When rendering is performed on low-power mobile devices, the computational complexity must also be minimised. To make informed decisions about the best approach for a given scenario, the constraints must be considered alongside an understanding of the quality offered by different options. A better understanding of the quality of different techniques and formats available could help both the providers and users of the technologies to achieve better results.

For the remainder of this thesis, the relationship between representations of spatial audio content and binaural rendering techniques will be studied in terms of perceptual quality and considering the applications that have been described in this section.

Chapter 7

Characterising the Perceptual Effects of Binaural Rendering with Virtual Loudspeakers using a Rate-All-That-Apply Approach

The purpose of this chapter is to evaluate the perceptual effects of the virtual loudspeaker approach to providing spatial sound for headphone listeners. Single noise sources were rendered through two virtual loudspeakers using pairwise amplitude panning and compared with the same sources directly binaurally rendered. Three other rendering system factors were varied: use of head-tracking, use of room impulse responses, and source positions. A listening experiment was conducted using descriptive analysis to characterise the effects with a pre-defined vocabulary of quality features and a rate-all-that-apply (RATA) method.

7.1 Introduction

In chapter 6 it was observed that loudspeaker virtualisation is commonly used to spatialise sound through headphones. This practice is set to continue in next-generation audio (NGA) codec systems with 3D loudspeaker arrays. The basic technique of loudspeaker virtualisation is described in section 6.4, where it is explained that each loudspeaker signal is treated as an independent virtual sound source and is rendered using a binaural filter that corresponds to the intended loudspeaker position (relative to the listener). Audio sources are first rendered (or mixed) into a channel-based format, to produce a signal corresponding to each virtual loudspeaker position. Amplitude panning techniques are used to position sources within this channel-based format. A scene-based format (e.g. ambisonics) may also be used and then subsequently converted to a channel-based format for rendering.

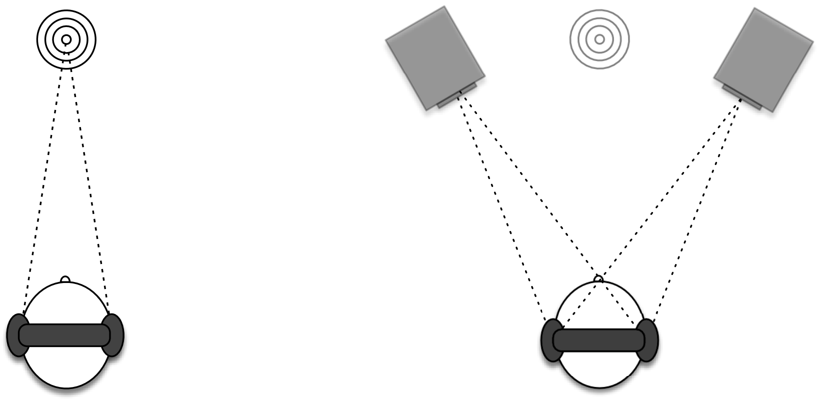
There are obvious advantages to this approach when compared with an object-based representation of the scene, in terms of limiting distribution bandwidth and rendering complexity. However, it is important to understand the impact that this design choice has on the quality of the listening experience. Evaluations of headphone surround sound processing (HSSP) have found little or no enhancement in quality over a stereo down-mix of 5.1 surround programme material, as discussed in chapter 4. This is despite plausible simulation of real loudspeakers being achievable (see chapter 5). The web-based studies reported in chapter 6 showed that different approaches to the production of binaural signals can lead to improvements over stereo, even without head-tracking and individualisation.

There are several potential reasons why HSSP does not lead to high quality, as discussed in section 4.7. It is likely that virtualisation of 3D loudspeaker layouts will allow better quality than use of the five-channel horizontal layout. More spatial information will be available in the input signal, so it is supposed that a better spatial impression can be provided to the headphone listener. One aspect that was not considered in that discussion is the interaction between amplitude panning and the binaural rendering process. In section 6.2, a distinctive audio drama production in 5.1 surround was found to give very effective results with HSSP applied. It made little use of amplitude panning, instead routing dialogue to single loudspeaker channels. It appears that the perceptual effects of headphone-virtualised amplitude panning, as compared to direct binaural rendering of a source, have not been studied in detail previously. Given the prevalence of these approaches, this warrants investigation.

In the study presented in chapter 4, the systems evaluated were from commercial suppliers and the precise details of signal processing performed were not available. Experimental apparatus has been developed in this work and is described in detail in appendix A. This allows the effects of virtual loudspeaker rendering to be investigated in more detail, with full knowledge of the signal processing being applied.

Also, in chapter 4 it was found that the programme items had a significant influence on the relative quality of different systems. With complex scenes, as are common in real programme content, it can be difficult to untangle the interactions between the many layers of audio material, the methods used in its production, and the rendering system factors, in terms of their influence on perceived quality. In this chapter, the effects of the virtual loudspeaker approach will be investigated using a single audio source at defined static target positions.

From the review presented in section 6.5, similar perceptual effects are expected with



(a) Direct binaural rendering

(b) Virtual loudspeaker rendering

Figure 7.1: Illustration of the two rendering approaches compared in this study.

headphone rendering as during amplitude panning over real loudspeakers, where spatial and timbral artefacts have been observed previously. However, further issues may arise due to imperfect simulation of a real loudspeaker array by the binaural rendering; for example due to inaccuracies in equalisation, limited adaptation to listener movements, and lack of individualisation. These are known to cause perceptual artefacts for simulating a single loudspeaker source (see chapter 2), but there may also be interaction between these inaccuracies and the effects of amplitude panning.

This chapter presents a listening experiment conducted to study the effect of loudspeaker virtualisation in binaural rendering. Assessors compared (a) *direct binaural* rendering of a single source produced by convolution with a binaural impulse response (BIR) measured at the target position with (b) *virtual loudspeaker* rendering, whereby the source is first distributed to two virtual loudspeakers by amplitude panning and then these are each convolved with a binaural impulse response corresponding to their positions. These two approaches are illustrated in figure 7.1. To assess the impact of the virtual loudspeaker approach across a range of common scenarios, rendering was performed with and without head tracking and using anechoic head-related impulse responses (HRIRs) and binaural room impulse responses (BRIRs) measured in a small room. Descriptive analysis (DA) methods were applied to explore the characteristics of the perceptual effects experienced by assessors.

Section 7.2 briefly revisits DA methods and discusses an approach suitable for this study. Section 7.3 then presents the apparatus and experimental design for the study. The results are presented in section 7.4, followed by discussion in section 7.5 and conclusions in section 7.6

7.2 Consideration of Descriptive Analysis Methods

Given that we can expect there to be differences between direct binaural rendering and the virtual loudspeaker approach, a subjective evaluation should aim to reveal the ways in which they differ and not just how much they differ. Relevant existing studies focus mainly on specific quality features, particularly localisation and colouration. DA techniques can be applied to elicit the perceived quality features of the stimuli from listeners. This allows a broader characterisation of the effects of headphone-virtualised amplitude panning, which might then highlight important quality features that could be improved with further research into underlying rendering techniques.

DA methods are reviewed in detail in section 3.3.3. Classical DA methods make use of panels of expert assessors, who, through an extensive development process, define a set of attribute scales and then use them to rate the stimuli, e.g. Stone et al. (1974). Other methods have been proposed to allow the user to describe stimuli using their own individual vocabularies, e.g. (Williams and Langron, 1984), which is more suitable for less experienced assessors. Such techniques have been applied successfully to sound reproduction (Lorho, 2010) and room acoustics (Lokki, 2014). A shared vocabulary has benefits when evaluating the impact of technical system design parameters on perceived quality. The results can be directly interpreted without requiring the interpretation of individual terminology, which often involves complex projection of individual responses into a common factor space. Several experienced critical listeners were available to participate in this study. It was anticipated that they would have the domain-specific knowledge and experience required to use shared terminology effectively.

7.2.1 The Spatial Audio Quality Inventory

The spatial audio quality inventory (SAQI) was discussed in section 3.2.5.4. It provides an extensive set of 48 descriptors of the perceptual character of virtual acoustic environments, along with a methodology for their evaluation (Lindau, Erbes, et al., 2014). These attributes describe differences between two stimuli, as in semantic differential scales (Osgood et al., 1957), rather than intensities on an absolute scale. The SAQI is well suited to assessing the effect of changing a specific parameter or design choice in a binaural rendering system. The large number of attributes allows for an exploratory assessment of differences, whilst retaining a common vocabulary of attributes amongst assessors. The method has previously been used to give sensory profiles of individualised and non-individualised dynamic data-based binaural systems (Lindau, Brinkmann, and Weinzierl, 2014), as described in section 3.2.5.4.

The concern with this method is the duration of the test. With paired comparisons over the full set of attributes, the number of ratings required of assessors can be very high. In Lindau, Brinkmann, and Weinzierl (2014), two binaural conditions were compared with a real loudspeaker for a single source position over 45 of the 48 attributes, with a full repetition. This resulted in 180 ratings per assessor. This approach will become unmanageable as soon as more system configurations are to be assessed.

7.2.2 Reducing the Number of Attributes

There is a risk that listeners will become fatigued or frustrated if the test duration is unduly long, and the quality of their responses may decline as a result. Lawless and Heymann, 2010, section 9.5.3 report over-partitioning effects, where presenting an excessive number of attributes results in reduced rating intensities. The authors state that: “It is obviously important to be inclusive and exhaustive, but also not to waste the panelists’ time with irrelevant attributes.” The SAQI test manual (Lindau, 2015) outlines ways in which the method can be reduced for specific experimental needs, including omitting irrelevant attributes or aggregating several attributes.

It is common in audio applications for the experimenters (or a small panel of experts) to preselect the attributes to be rated, typically resulting in use of only 4–6 attributes, e.g. (Cobos et al., 2015; Moulin et al., 2016; Millns and Lee, 2018; Reardon et al., 2018). Whilst quicker, this approach may be reductive. In exploratory rather than confirmatory evaluations, the aim is to obtain a sensory profile of the character of systems as perceived by the assessors. Imposing a limited set of attributes risks omitting some information about the assessors’ experience and may introduce dumping effects on ratings of the given attributes (Lawless and Heymann, 2010, section 9.5.2).

Another approach is to allow the assessors themselves to contribute to selection of a common subset of attributes from a larger lexicon. In an initial session, assessors could be presented with the stimuli to be rated and the attribute lexicon, and asked to indicate all attributes that apply to the stimuli. These responses could then be used to create a subset of the lexicon, based on the consensus opinion of the assessors, for use in a subsequent main rating session. The subset could be defined based on several approaches, for example, those attributes that are identified by all assessors, or those that allow differentiation of the stimuli under test, using statistics for categorical data as in (Francombe, Mason, et al., 2014). These methods will likely reduce the number of attributes presented in the main rating experiment, but may still require assessors to respond to attributes they do not perceive, or remove attributes that they do perceive.

In (Lindau, Brinkmann, and Weinzierl, 2014) when no difference was perceived, the assessor could skip the rating and the response was treated as a zero rating. It was found that no difference was ever perceived on several of the attributes. For many others, the approach resulted in non-normal skewed rating distributions with a large number of zeros, which presents challenges for statistical analysis. However, such an approach appears to retain a comprehensive and exploratory approach to capturing the individuals' sensations in terms of a common attribute set, whilst also improving efficiency. This still requires presentation of each attribute and system condition combination though. By allowing a listener to first identify only the attributes that they consider useful for differentiating the stimuli under evaluation, test duration and listener fatigue may be further reduced. This is particularly important when several system conditions are to be evaluated.

7.2.3 The Rate-All-That-Apply Method

The approach taken by Lindau, Brinkmann, and Weinzierl (2014) has similarities to the rate-all-that-apply (RATA) method (Ares, Bruzzone, et al., 2014), which was recently proposed for efficient sensory profiling of food products by untrained assessors. RATA is an extension of the check-all-that-apply (CATA) method, in which consumers simply indicate the attributes that apply to a product from a pre-defined lexicon of potentially relevant attributes with a binary response. CATA has been widely applied in sensory evaluation because it allows quick collection and analysis of data from many assessors in a simple structured format. RATA introduces a 3- or 5-point intensity scale for applicable attributes to improve sample discrimination, particularly when conditions have similar character.

Since RATA was designed for evaluation of food products by untrained assessors, it typically uses single stimulus presentation and simple low-resolution intensity scales. The SAQI is instead designed for use with trained expert assessors and provides continuous differential scales. Despite these differences, insights into the analysis of responses to an individual subset of the presented attributes can be gained from the RATA literature.

7.3 Methods

To better understand the implications of virtual loudspeaker rendering techniques on sound quality, a listening experiment was carried out to perform paired comparison ratings between direct binaural rendering and virtual loudspeaker rendering of a simple monophonic input source.

Direct binaural rendering, in which an audio source is convolved with a BIR measured at the target position, was compared to a *virtual loudspeaker rendering*. In the virtual loudspeaker rendering condition the source is first distributed to two virtual loudspeakers by means of amplitude panning and then these are each convolved with a measured BIR corresponding to their target positions. The two approaches are illustrated in figure 7.1.

To assess the effects of the virtualisation technique across a range of common binaural rendering system configurations, the evaluation was performed both with and without head-tracking and using anechoic HRIRs and BRIRs measured in a reverberant room. The binaural rendering was not individualised. Whilst it has been shown that individualised binaural rendering gives a better simulation of a real sound source across a wide range of attributes (Lindau, Brinkmann, and Weinzierl, 2014), current commercial systems are not individualised and so in the context of this research a non-individual system is more relevant.

Inspired by the RATA method of Ares, Bruzzone, et al. (2014), assessors were able to select an individual subset of the available attributes with which to perform the rating. These were those attributes that they found relevant to describe the perceived differences.

7.3.1 Apparatus

The development of the experimental apparatus used in this study is described in detail in appendices A and B. This section gives a short summary of the most relevant aspects.

7.3.1.1 Impulse Response Data

Both sets of impulse responses (IRs) were measured using a Neumann KU100 dummy head microphone. Anechoic rendering used a freely-available dataset of far-field HRIR measurements on a 2° Gauss-Legendre grid, as described by Bernschütz (2013). The HRIRs had a length of 128 samples. The BRIRs were measured as described in appendix A.5.5, in the listening room at BBC R&D. The subset of speakers measured is shown in figure 7.2. This room has a mean reverberation time of 0.21s in the frequency range 125 Hz to 8 kHz (Nixon et al., 2015). Genelec 8030B loudspeakers were used and the dummy head microphone was rotated about the vertical axis in 2° steps. Excess onset delay common to all measurements

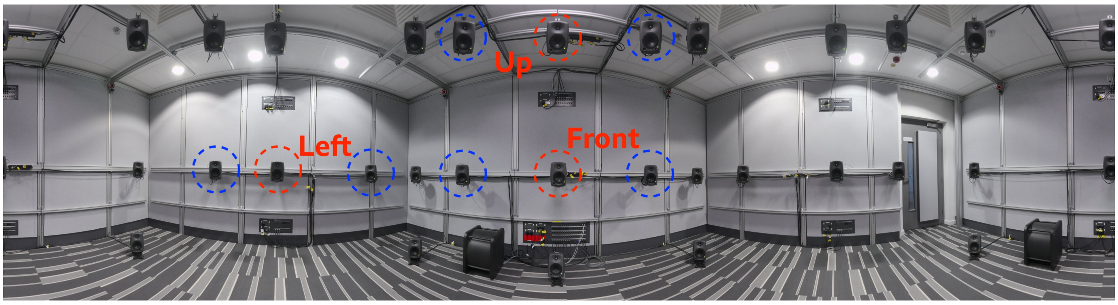


Figure 7.2: Panoramic photograph of the Recommendation ITU-R BS.1116-compliant listening room at BBC Research & Development in Salford, UK. Loudspeakers used in the experiment are circled (red for direct rendering, blue for virtual panning) and position names labelled, see table 7.1

was removed and the IRs were truncated to 2^{14} samples using a half-Hann window fade out. The BRIRs were level adjusted to match the HRIRs using the mean magnitude response between 200 Hz and 4 kHz.

To enable real-time adaptation of the interaural time difference (ITD), onsets were modelled and stored separately. The log-threshold method of onset estimation was used with a threshold of −20 dB and 10 times oversampling (Lindau, Estrella, et al., 2010). A parametric model for the time of arrival (TOA) was then fitted to the estimates to achieve a smooth direction-continuous function (Ziegelwanger and Majdak, 2014). This uses a simplified geometric model of the head, but allows for variations in the positioning of the head and the ears within the coordinate system of the measurement positions.

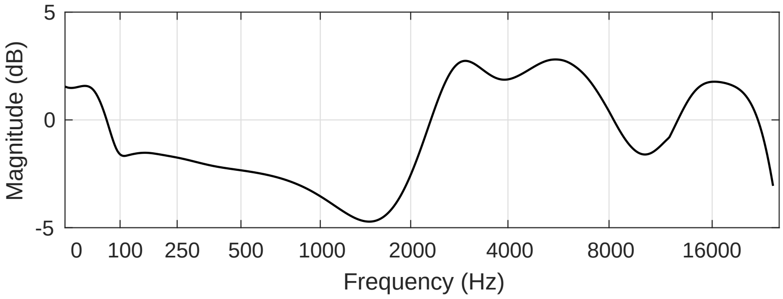


Figure 7.3: Magnitude response of headphone correction filter for Stax SR-207, based on 20 headphone transfer function measurements made on Neumann KU100 dummy head microphone.

Stax SR-207 headphones were used in the experiment. A non-individual headphone correction filter was generated for these headphones, shown in figure 7.3. For headphone correction with non-individualised measurements, results in Lindau and Brinkmann, 2012 suggest that the same dummy head microphone should be used as for the head-related transfer function (HRTF) measurements in the rendering. Twenty headphone transfer function measurements were made on the KU100. The headphones were reseated before each measurement to capture variance due to headphone positioning. Based on the mean magnitude response of these measurements across both ears, a minimum-phase correction filter was calculated using least mean squares inversion, following methods similar to Lindau and Brinkmann (2012). Prior to inversion, the mean magnitude response was pre-processed using approaches described by Masiero (2012). A “compare and squeeze” notch reduction algorithm was used, comparing to a $\frac{1}{3}$-octave smoothed version of the response with a mix factor of 0.5. This was followed by $\frac{2}{3}$-octave smoothing. A minimum-phase transfer function was derived from this smoothed magnitude response for inversion. The frequency-dependent regularisation function was set at −1.5 dB above 12 kHz and below 50 Hz, and at −20 dB between 100 Hz and 8 kHz. The target magnitude response was also flattened below 50 Hz to further prevent attempts at excessive boosting at low frequencies that are not reproduced by the headphones. These parameters were found by perceptual adjustment by this author and an expert colleague. The resulting 2048-sample filter was normalised based on the magnitude response between 200 Hz and 4 kHz (Masiero, 2012). It was then applied to the binaural IRs offline, prior to real-term rendering in the experiment. This is discussed further in appendix A.6.4.

7.3.1.2 Real-Time Rendering

The system described in appendix A.2 was used to render stimuli in real-time. Since BRIR measurements were only made for horizontal rotations of the dummy head, only the horizontal head rotation was used during this condition. For HRIR rendering, 3 degrees of freedom (DoF) tracking was applied using the head orientation data. A filter partition size of 128 samples (also the HRIR length) was used and the software used the JACK audio server with a 128-sample I/O buffer size. As reported in appendix A.4.3, the total system latency (TSL) for this configuration was measured as $\mu_{\text{TSL}} = 47.2\,\text{ms}, \sigma_{\text{TSL}} = 3.7\,\text{ms}$, which is below the detection threshold for all listeners in Lindau, 2009.

7.3.1.3 Test Administration Software

A separate graphical user interface (GUI) application was used to administer the test. The rendering software was controlled over a local user datagram protocol (UDP) socket using Open Sound Control (OSC) messages. This made it possible to select appropriate system settings, set source positions and control playback according to the test design. The test administration software performed appropriate test scheduling including randomisation, as well as logging of responses and timing information.

7.3.2 Experiment Design

Listeners performed paired comparisons over every combination of the tracking (**T** = on, **N** = off) and IR set (**H** = HRIR, **B** = BRIR) conditions for each of three target source positions. For each target position, virtual loudspeaker rendering was performed using a pair of virtual loudspeakers with equal elevation angle. The choice of positions of target sources and virtual loudspeakers was informed by the review of perception of panning on loudspeakers given in section 6.5.1, as well as the standardised loudspeaker positions in Recommendation ITU-R BS.2051 ITU-R, 2017a and considering directions that are important in 3D spatial audio programme material.

Two target positions were at 0° azimuth and employed virtual loudspeakers located symmetrically about the median plane. Stereo amplitude panning techniques were developed with such left-right symmetry in mind (see appendix B.3 for more background). One target position was directly in front at an elevation of 0° (**F**) and one was at an elevation of 40° (**U**). The frontal position is of great importance in broadcast material, since that is where the video display is located, and it is also known to be a region challenging for externalisation during binaural rendering (Kim and Choi, 2005). Channel-based 3D spatial audio formats include elevated loudspeaker channels, but the perceptual character of sources at elevated target positions may be altered by the use of amplitude panning. A spatial audio system needs to be able to render lateral sources as well as those to the front, but the review in section 6.5.1 showed that this is challenging when using amplitude panning over loudspeakers. The third target position was to the left of the listener (**L**) with virtual loudspeakers located asymmetrically about this target direction (due primarily to practical aspects of the loudspeaker installation in the test environment). Table 7.1 gives the spherical polar angles for these target source positions as well as the corresponding virtual loudspeaker positions used to render them, using the spherical coordinate system defined in section 2.2.2. The table also shows the corresponding panning gains, which were derived using the tangent panning law (see appendix B.3.1 for definition).

A subset of 19 of the SAQI attributes was identified by the author and another expert listener as being potentially relevant to the set of stimuli (Lindau, Erbes, et al., 2014). These attributes, to be presented to the assessors for selection, are listed in table 7.2. Attributes representing affective or aesthetic aspects and those that were clearly not applicable to

	Target	Virtual Loudspeakers	Panning Gains
Front (F)	$(0°, 0°)$	$(30°, 0°), (-30°, 0°)$	$1/\sqrt{2}, 1/\sqrt{2}$
Left (L)	$(90°, 0°)$	$(110°, 0°), (60°, 0°)$	$0.825, 0.565$
Up (U)	$(0°, 40°)$	$(30°, 40°), (-30°, 40°)$	$1/\sqrt{2}, 1/\sqrt{2}$

Table 7.1: Source directions used in the evaluation, showing target source directions, virtual loudspeaker directions, and corresponding panning gains. See figure 7.2 for visualisation.

the stimuli were removed. Some attributes were aggregated, where further distinction was thought to be challenging. The attribute *tone colour* is an aggregate of multiple SAQI attributes within the timbral category. It was defined as: “timbral change in any frequency range” and used a unipolar difference scale with end labels “no colouration” and “very large colouration”. The attribute *extent* is an aggregate of the three SAQI extent attributes in each dimension of width, depth, and height. It was defined as: “perceived extent of a sound source in any direction or all directions” and used a bipolar difference scale with end labels “smaller” and “bigger”.

An amplitude-modulated white noise signal was used throughout the test, modulated with a 30 Hz sinusoid, in 300 ms bursts with 200 ms gaps. This stimulus was found to be revealing of differences in characteristics between the test conditions, in terms of changes in both timbral and spatial aspects.

Table 7.2: The 19 attributes used in the listening experiment, taken from the SAQI (Lindau, Erbes, et al., 2014).

Attribute	**Definition**	**Scale Labels**	**Range**
Overall difference	Existence of a noticeable difference.	None – Very large	0 to 1
Tone colour	Timbral change in any frequency range.	No colouration – Very large colouration	0 to 1
Comb filter colouration	Often perceived as tonal coloration. ‘Hollow’ sound. Example: speaking through a tube.	Less pronounced – More pronounced	-1 to 1
Horizontal direction	Direction of a sound source in the horizontal plane.	Shifted anti-clockwise – Shifted clockwise	−180° to 180°
Vertical direction	Direction of a sound source in the vertical plane.	Shifted down – Shifted up	−90° to 90°
Front-back position	Refers to the position of a sound source before or behind the listener only. Impression of a position difference of a sound source caused by ‘reflecting’ its position on the frontal plane going through the listener.	Not reversed – Reversed	False or True

Continued on next page

Table 7.2 – *Continued from previous page*

Attribute	**Definition**	**Scale Labels**	**Range**
Distance	Perceived distance of a sound source.	Closer – More distant	-1 to 1
Extent	Perceived extent of a sound source in any direction or all directions.	Smaller – Bigger	-1 to 1
Externalisation	Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance. Terminologically often enclosed between the phenomena of in-head localisation and out-of-head localisation. Examples: Poorly/not externalised = perceived position of sound sources at diotic sound presentation via headphones, good/strongly externalized = perceived position of a natural source in reverberant environment and when allowing for movements of the listener.	More internalised – More externalised	-1 to 1
Localisability	If localisability is low, spatial extent and location of a sound source are difficult to estimate, or appear diffuse, resp. if localisability is high, a sound source is clearly delimited. Low/high localisability is often associated with high/low perceived extent of a sound source. Examples: sound sources in highly diffuse sound field are poorly localisable.	More difficult – Easier	-1 to 1
Spatial disintegration	Sound sources, which - by experience - should have a united spatial shape, appear spatially separated. Possible cause: Parts of the sound source have been synthesized/simulated using separated algorithms/simulation methods and between those exists an unwanted offset in spatial parameters. Examples: fingering noise and playing tones of an instrument appear at different positions; spirant and voiced phonemes of speech are synthesized separately and then reproduced with an unwanted spatial separation.	More coherent – More disjointed	-1 to 1
Reverberation level	Perception of a strong reverberant sound field, caused by a high ratio of reflected to direct sound energy. Leads to the impressoin of high difussivity in case of stationary excitation (in the sense of a low direct/revererant-ratio). Example: The perceived intensity of reverberation differs significantly between rather small and very large spaces, such as living rooms and churches.	Less – More	-1 to 1

Continued on next page

Table 7.2 – *Continued from previous page*

Attribute	**Definition**	**Scale Labels**	**Range**
Reverberation time	Duration of the reverberant decay. Well audible at the end of signals.	Shorter – Longer	-1 to 1
Envelopment (by reverberation)	Sensation of being spatially surrounded by the reverberation. With more pronounced envelopment of reverberation, it is increasingly difficult to assign a specific position, a limited extension or a preferred direction to the reverberation. Impressions of either low or high reverberation envelopment arise with either diotic or dichotic (i.e. uncorrelated) presentation of reverberant audio material.	Less pronounced – More pronounced	-1 to 1
Pre-echos	Copies of a sound with mostly lower loudness prior to the actually intended starting point of a sound.	Less intense – More intense	-1 to 1
Post-echos	Copies of a sound with mostly descreasing loudness after the actually intended starting point of a sound. Example: repetition of one's own voice through reflection on mountain walls.	Less intense – More intense	-1 to 1
Crispness	Characteristic which is affected by the impulse fidelity of systems. Perception of the reproduction of transients. Transients can either be more soft/more smoothed/less precise, or - as opposed - be quicker/more precise/more exact. Example for 'smoothed' transients: A transmission system that exhibits strong group delay distortions. Counter-example: Result of an equalization aiming at phase linearization.	Less pronounced – More pronounced	-1 to 1
Loudness	Perceived loudness of a sound source. Disappearance of a sound source can be stated by a loudness equaling zero. Example of a loudness contrast: Whispering vs Screaming	Quieter – Louder	-1 to 1

Continued on next page

Table 7.2 – *Continued from previous page*

Attribute	Definition	Scale Labels	Range
Ghost Source	Spatially separated, nearly simultaneous and not necessarily identical image of a sound source. A kind of a spatial copy of a signal: a sound source appears at one or more additional positions in the scene. Examples: two sound sources which are erroneously playing back the same audio content; double images when down-mixing main and spot microphone recordings; spatial aliasing in wave field synthesis (WFS): sound sources are perceived as ambivalent in direction.	Less intense – More intense	-1 to 1
Distortion	Percept as a result of non-linear distortions as caused e.g. by clipping. Scratchy or 'broken' sound. Often dependent on signal amplitude. Perceptual quality can vary widely depending on the type of distortion. Example: clipping of digital input stages.	Less intense – More intense	-1 to 1

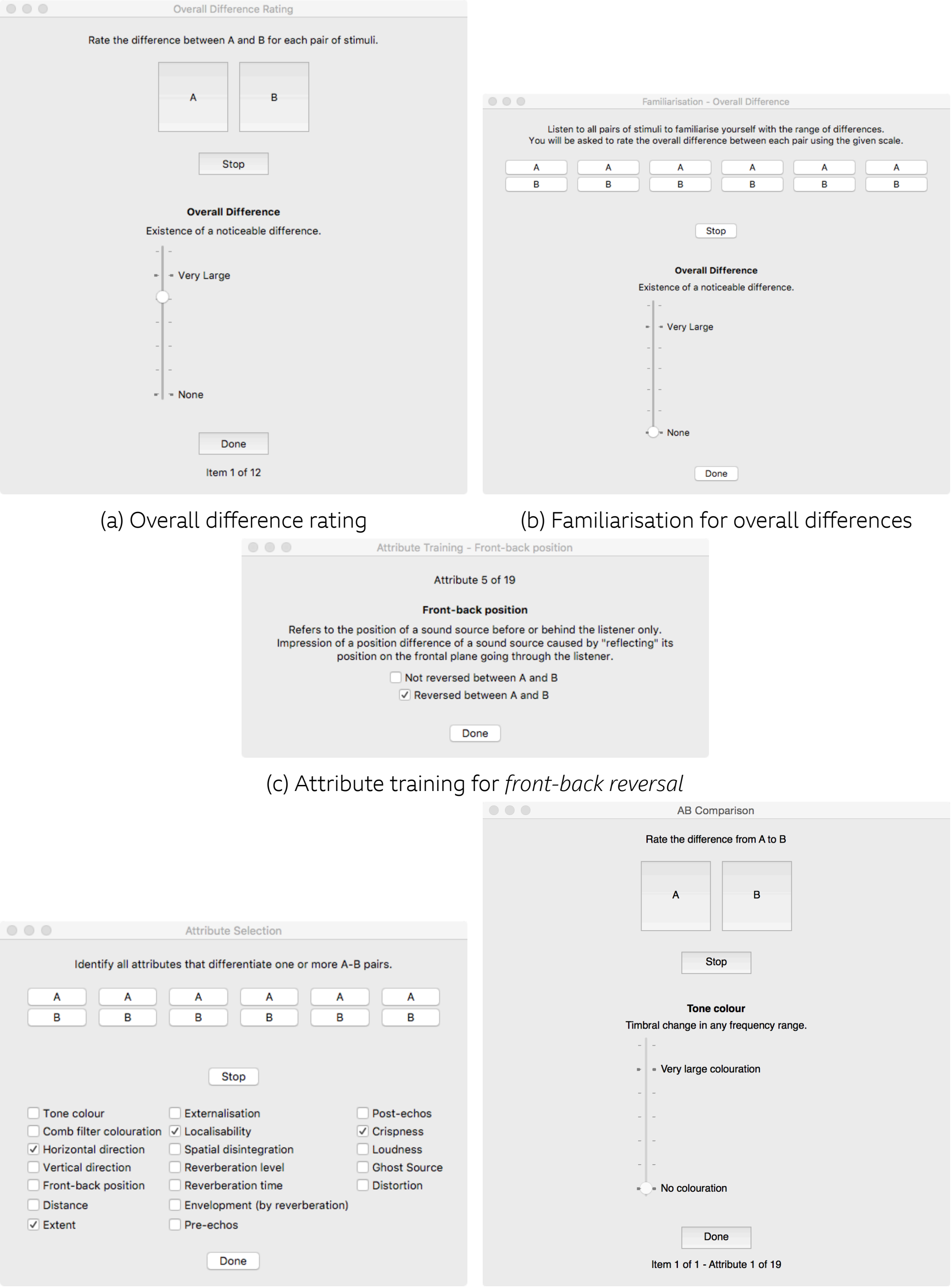
7.3.2.1 Procedure

The evaluation consisted of five stages:

- ITD scaling
- Overall difference rating
- Attribute training
- Attribute selection
- Attribute rating

The initial stage allowed the listener to adjust the real-time scaling of ITDs to better match the non-individual cues to their own. The scale was adjusted to maximise stability of a frontal source during head movements. This process was guided carefully by the experimenter, since it was found to be challenging for assessors in the study of chapter 5. Separate scaling values were found for each IR dataset. The adjustment process was performed only once in each case, starting from the original ITD values of the dummy head.

The rating interface for overall differences is shown in figure 7.4a. Two practice ratings were performed to ensure that the assessor was familiar with the procedure. The rendering conditions were randomly selected for these. Prior to each rating task (both for overall



Overall Difference Rating

Rate the difference between A and B for each pair of stimuli.

A B

Stop

Overall Difference

Existence of a noticeable difference.

Very Large

None

Done

Item 1 of 12

(a) Overall difference rating

Familiarisation - Overall Difference

Listen to all pairs of stimuli to familiarise yourself with the range of differences.
You will be asked to rate the overall difference between each pair using the given scale.

A A A A A A
B B B B B B

Stop

Overall Difference

Existence of a noticeable difference.

Very Large

None

Done

(b) Familiarisation for overall differences

Attribute Training - Front-back position

Attribute 5 of 19

Front-back position

Refers to the position of a sound source before or behind the listener only.
Impression of a position difference of a sound source caused by "reflecting" its position on the frontal plane going through the listener.

□ Not reversed between A and B
☑ Reversed between A and B

Done

(c) Attribute training for *front-back reversal*

Attribute Selection

Identify all attributes that differentiate one or more A-B pairs.

A A A A A A
B B B B B B

Stop

□ Tone colour
□ Comb filter colouration
☑ Horizontal direction
□ Vertical direction
□ Front-back position
□ Distance
☑ Extent
□ Externalisation
☑ Localisability
□ Spatial disintegration
□ Reverberation level
□ Reverberation time
□ Envelopment (by reverberation)
□ Pre-echos
□ Post-echos
☑ Crispness
□ Loudness
□ Ghost Source
□ Distortion

Done

(d) Attribute selection

AB Comparison

Rate the difference from A to B

A B

Stop

Tone colour

Timbral change in any frequency range.

Very large colouration

No colouration

Done

Item 1 of 1 - Attribute 1 of 19

(e) Attribute rating for *tone colour*

Figure 7.4: User interfaces presented during the experiment.

differences and attributes), a familiarisation page was displayed. Assessors were presented with each scale to be used, alongside all pairs of stimuli in the experiment (e.g. figure 7.4b). The pairs were the direct binaural and virtual loudspeaker rendering treated with the different combinations of system configuration conditions for impulse response and target position. These pages were intended to allow the assessor to become familiar with the range of differences amongst the test conditions and consider how they could be mapped onto the scale.

For attribute training, the assessor was introduced to the definition and rating scale for each attribute in turn (e.g. figure 7.4c). The meaning of the attribute was discussed and for most attributes a listening example was given for clarification. During attribute selection, assessors were instructed to listen carefully to all of the pairs of stimuli to be evaluated (as in the familiarisation pages) and consider each attribute in turn (figure 7.4d). An attribute should be selected if it could be used to describe a perceived difference within one or more of the pairs. Attribute definitions and scale end-points were made available to assessors during this process. Finally, attribute rating (figure 7.4e) was carried out, with all conditions being evaluated for a single attribute before moving onto the next. If no difference was observed, assessors were instructed to leave the rating scale untouched.

The head tracking conditions were split into separate sessions for each test stage, since they invited different listening behaviour; the order of these sessions was also randomised. The attribute selection process was performed separately for each head tracking condition. Listeners were instructed to investigate the effect of horizontal head rotations, avoiding systematic usage of tilting and rolling movements. Throughout the experiment, the test stimuli could be played as many times as required. After the evaluation, assessors were interviewed to gain insight into what they perceived and how they approached the test procedure.

A full replication was used for the overall difference ratings, with the order of assignment to the A and B buttons reversed, to permit the reliability of the listeners' ratings to be analysed. For the attribute ratings, two of the six conditions were replicated (**B**-**F** and **H**-**L**). The order of presentation of stimuli was randomised within each session and the order of presentation of attributes was also randomised. All pairs were rated in terms of an attribute before moving on to the next attribute. A familiarisation page was provided each time a new attribute was presented.

Assessors were given rest breaks between sessions approximately every 30 minutes to reduce fatigue.

7.3.2.2 Presentation Method

The test was carried out in the same listening room used for the BRIR measurements. Loudspeakers were screened from view during the test to avoid the influence of visual cues. The level of the noise signal was set at −23 LUFS on input to the renderer. The headphone level was aligned to match the loudness of a loudspeaker generating 72 dBA at the listening position using a pink noise signal at −18 dBFS RMS.

7.4 Results

Ten assessors participated in the experiment, all had prior experience in perceptual audio evaluation experiments and worked at the BBC. The assessment took 90 min–180 min, excluding breaks. Assessors were encouraged to take breaks at least every 30 min. The large range was, in part, due to the varying number of attributes selected by assessors. For bipolar scales, data were adjusted for the presentation ordering, so that values always indicate the difference in the attribute from the virtual loudspeaker rendering to the direct binaural rendering.

7.4.1 Overall Differences

The overall difference ratings are summarised in combined box and scatter plots in figure 7.5, separated across the different conditions and including both replicates. Assessors identified clear differences between direct binaural and virtual loudspeaker rendering under all conditions. Using Shapiro-Wilk tests, it was confirmed that in each rendering condition the rating distributions were approximately normal. One-sided one-sample t-tests then confirmed that the mean was significantly above zero.

The expertise gauge analysis (ITU-R, 2014b), shown in figure 7.6, was applied to assessors' overall difference ratings. This revealed that there was poor agreement between assessors. The conditions were well discriminated in terms of the overall difference ratings by some assessors, but not all. It is not a requirement that assessors should detect different levels of overall difference across the different binaural rendering system conditions, however. Based on replicated ratings, one assessor showed very low reliability, but analysis with their data removed led to the same conclusions, so they were not removed from the analysis herein. Further analysis is performed with only the first rating for each condition, replicated ratings are removed.

Shapiro-Wilk tests were performed to confirm that the distributions of ratings for each condition can be assumed normal and Levene's test suggested that homogeneity of vari-

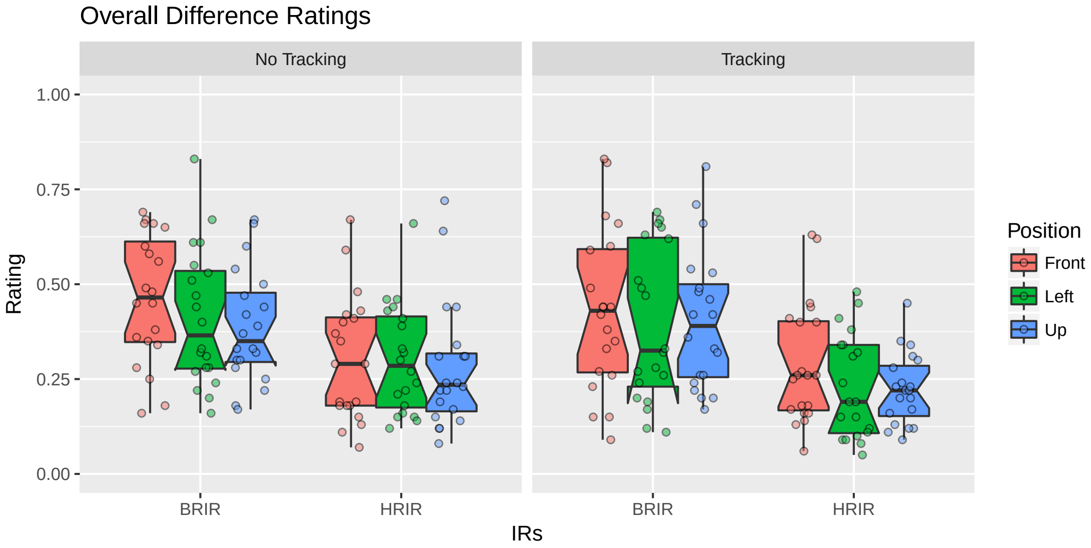
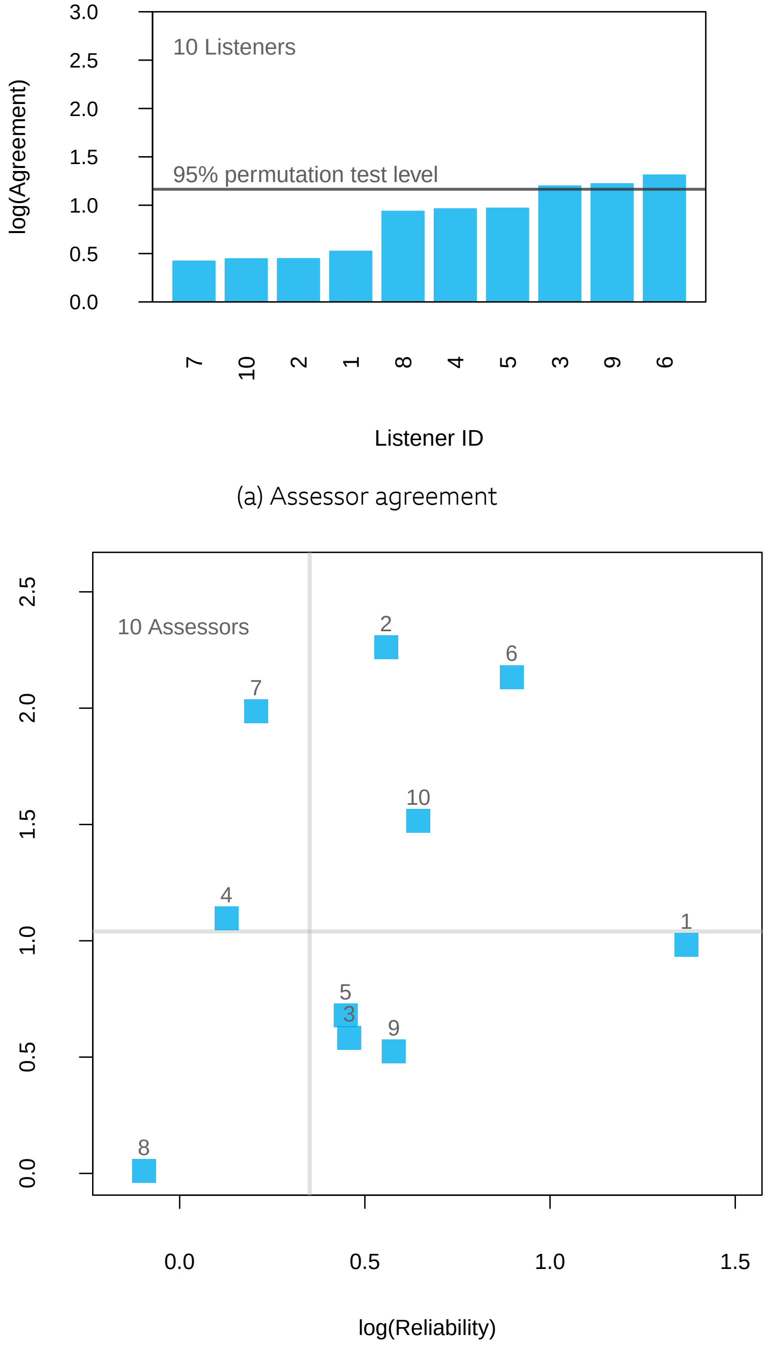


Figure 7.5: Combined box and scatter plots of overall difference ratings across condition variables. Box notches indicate 95% confidence intervals.

ance can be assumed. A repeated-measures analysis of variance (RM-ANOVA) was then performed on overall difference ratings with independent variables *Tracking*, *IRs*, and *Position* and their interactions, using a type-III sum-of-squares model. Mauchly's test indicated that the assumption of sphericity was met, and a Shapriro-Wilk test was used to check that the residuals of the linear model appear normally distributed.

The RM-ANOVA results are shown in table 7.3. The table shows degrees-of-freedom (DF) and sum-of-squares (SS) for the numerator and denominator, as well as the F statistic and the corresponding p-value. Significance is indicated with * for $p < 0.05$, ** for $p < 0.01$ and *** for $p < 0.001$. The generalised eta-squared effect size (η^2) is also indicated. All three main effects are shown to be significant, whilst the interactions are not. The influence of IR set (BRIR/HRIR) shows the most significant effect, which is medium-sized (Fritz et al., 2012). The effects for *Tracking* and *Position* are both small and p is close to the significance threshold ($\alpha = 0.05$). Combined box and scatter plots are shown for each of the main effects in figure 7.7, whilst figure 7.8 shows the means and confidence intervals.

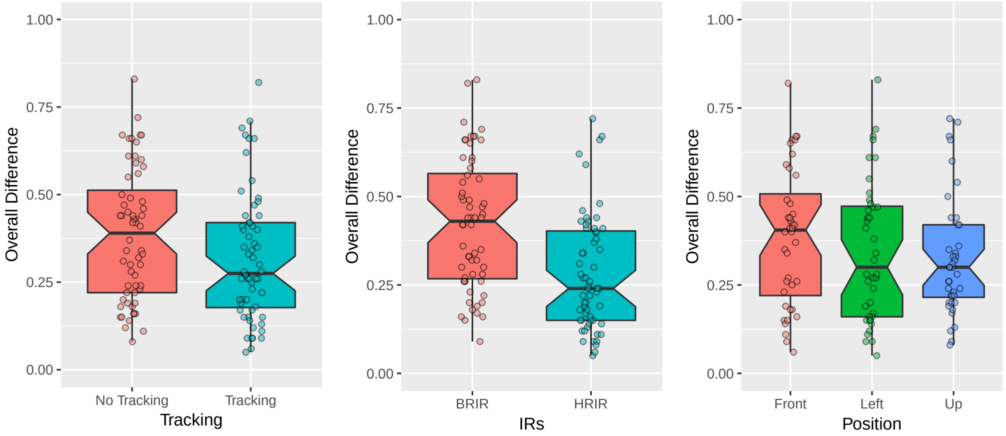
An alternative analysis was performed using a multi-level model with nested random effects (*Assessor* $\rightarrow$ *Tracking* $\rightarrow$ *IRs* $\rightarrow$ *Position*) using maximum likelihood estimation. Multi-level models are introduced in section 4.6.2.1, where they are also used. The random error variance in the data is modelled following this hierarchy. The results are shown in table 7.4. For this test, only the *IRs* effect was significant ($\chi^2(1) = 17.22, p < 0.001$), with BRIRs showing slightly but significantly greater differences than HRIRs ($b = 0.071, t(18) = 4.59, p <$



(a) Assessor agreement

(b) Assessor discrimination and reliability

Figure 7.6: Expertise gauge analysis on overall difference ratings.



(a) Tracking (b) IRs (c) Position

Figure 7.7: Combined box and scatter plots of *overall difference* ratings for main effects with box notches showing 95% confidence intervals.

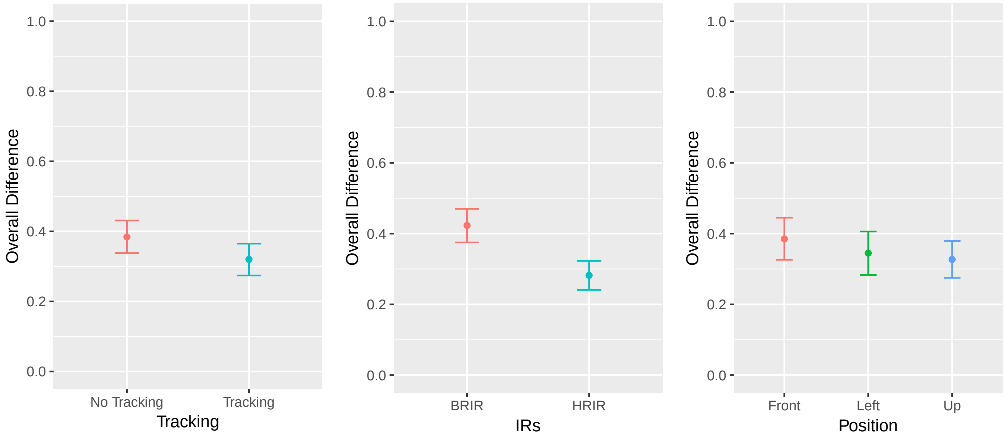
$0.001, r = 0.734$).

Effect	DF$_n$	DF$_d$	SS$_n$	SS$_d$	F	p	η^2
(Intercept)	1	9	14.904	1.437	93.312	0.000***	0.817
Tracking	1	9	0.124	0.210	5.283	0.047*	0.036
IRs	1	9	0.601	0.363	14.898	0.004**	0.152
Position	2	18	0.071	0.154	4.138	0.033*	0.021
Tracking:IRs	1	9	0.005	0.196	0.215	0.654	0.001
Tracking:Position	2	18	0.009	0.476	0.175	0.841	0.003
IRs:Position	2	18	0.008	0.350	0.200	0.821	0.002
Tracking:IRs:Position	2	18	0.036	0.162	2.034	0.160	0.011

Table 7.3: Repeated measures ANOVA of overall difference rating data.

7.4.2 Attribute Data

The number of attributes selected by each assessor varied widely, see table 7.5. Viewed across both sessions it ranged from 5–15 with median 7. The frequency of selection for each attribute across assessors is given in table 7.6, with and without head tracking and in either condition. It is clear that some attributes are more popular than others for describing the differences. The most commonly selected attributes were tone colour, horizontal direction and vertical direction. Seven attributes were selected by more than half of assessors in either tracking condition. Temporal attributes and those related to reverberation were selected infrequently and, when they were, the effects were small.



(a) Tracking (b) IRs (c) Position

Figure 7.8: Mean *overall difference* ratings for main effects with bootstrapped 95% confidence intervals.

Formula	Model	df	AIC	BIC	logLik	Test	L.Ratio	p
Baseline	1	6	-88.45	-71.73	50.23			
+*Tracking*	2	7	-89.03	-69.52	51.52	1 vs 2	2.58	0.108
+***IRs***	3	8	-104.26	-81.96	60.13	2 vs 3	17.22	<0.001***
+*Position*	4	10	-104.86	-76.99	62.43	3 vs 4	4.60	0.100
+*Tracking:IRs*	5	11	-103.04	-72.38	62.52	4 vs 5	0.18	0.670
+*Tracking:Position*	6	13	-99.66	-63.43	62.83	5 vs 6	0.62	0.733
+*IRs:Position*	7	15	-96.19	-54.38	63.10	6 vs 7	0.53	0.769
+*Tracking:IRs:Position*	8	17	-94.71	-47.32	64.35	7 vs 8	2.52	0.284

Table 7.4: Multi-level model fitting of overall difference rating data.

Assessor		**Selection Frequency**	
	Tracking	**No Tracking**	**Either**
1	4	5	6
2	6	6	7
3	12	8	13
4	11	7	13
5	6	4	7
6	6	6	6
7	8	12	13
8	14	13	15
9	3	4	5
10	7	7	7

Table 7.5: Attribute selection frequencies by assessor.

Attribute	Selection Frequency			
	Tracking	No Tracking	Either Condition	Total
Tone colour	10	10	10	20
Horizontal direction	8	9	9	17
Vertical direction	7	8	9	15
Externalisation	6	6	7	12
Distance	6	5	7	11
Comb-filter colouration	6	5	6	11
Extent	5	5	7	10
Localisability	4	4	4	8
Spatial disintegration	5	2	5	7
Crispness	2	4	4	6
Pre-echos	1	4	4	5
Distortion	2	2	4	4
Front-back position	2	2	3	4
Post-echos	2	2	2	4
Loudness	2	2	2	4
Ghost Source	3	0	3	3
Reverberation level	2	1	2	3
Envelopment (by reverberation)	2	1	2	3
Reverberation time	2	0	2	2

Table 7.6: The 19 attributes from the SAQI (Lindau, Erbes, et al., 2014) used in the listening test, with the number of times they were selected by a listener ($N = 10$) to differentiate between the direct binaural and virtual loudspeaker rendering.

7.4.2.1 Univariate Analysis of Attribute Ratings

Detailed analysis of attribute ratings is challenging, since the dataset is sparse or else, when data points for unselected attributes are populated with zeros, it is highly non-normal in distribution. Shapiro-Wilk tests for each attribute and condition combination showed that the assumption of a normal distribution is valid in fewer than 10% of cases and so parametric analysis is generally inappropriate. The only attribute for which all conditions had distributions that can be assumed normal is *tone colour*.

For bipolar attribute scales, Wilcoxon signed rank tests were used to test the null hypothesis that rating distributions are symmetric about zero. Figure 7.9 shows combined box and scatter plots of the distributions for each attribute over all rendering conditions. Where the Wilcoxon test suggests deviation from zero, the box is shaded blue and with full opacity, indicating a potentially systematic effect in one direction. For front-back reversals, a one-sided sign test showed that reversals were not significant overall. For the attribute *tone colour*, a one-sided t-test showed that the mean rating was significantly different to zero. One-sided sign tests were also applied to the absolute horizontal and vertical position change data, where t-tests and Wilcoxon tests are invalid due to the distribution. These showed that

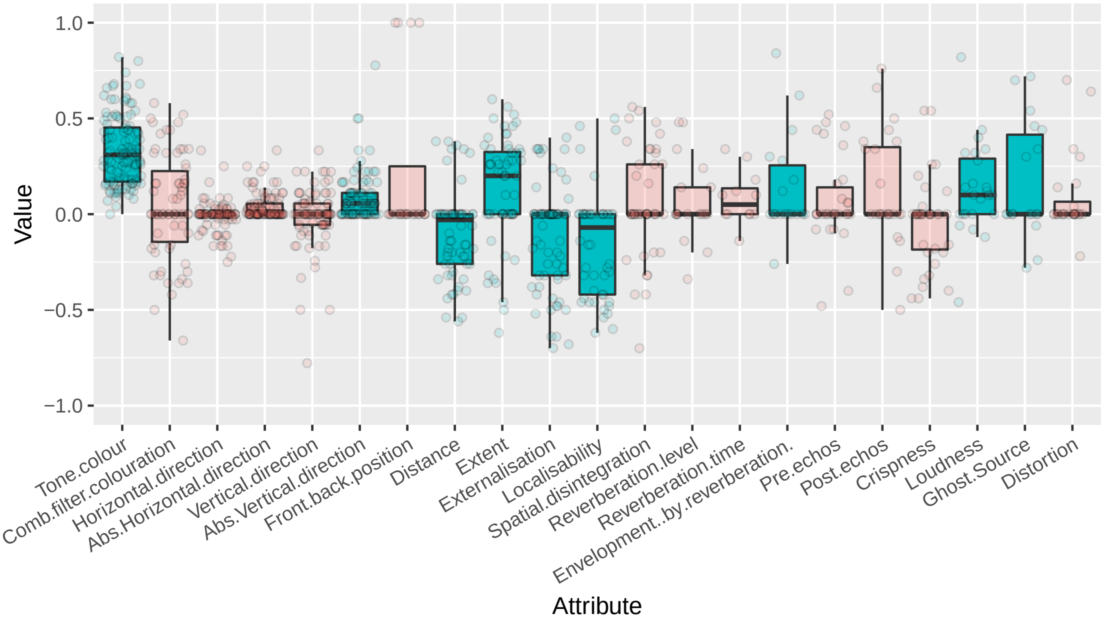


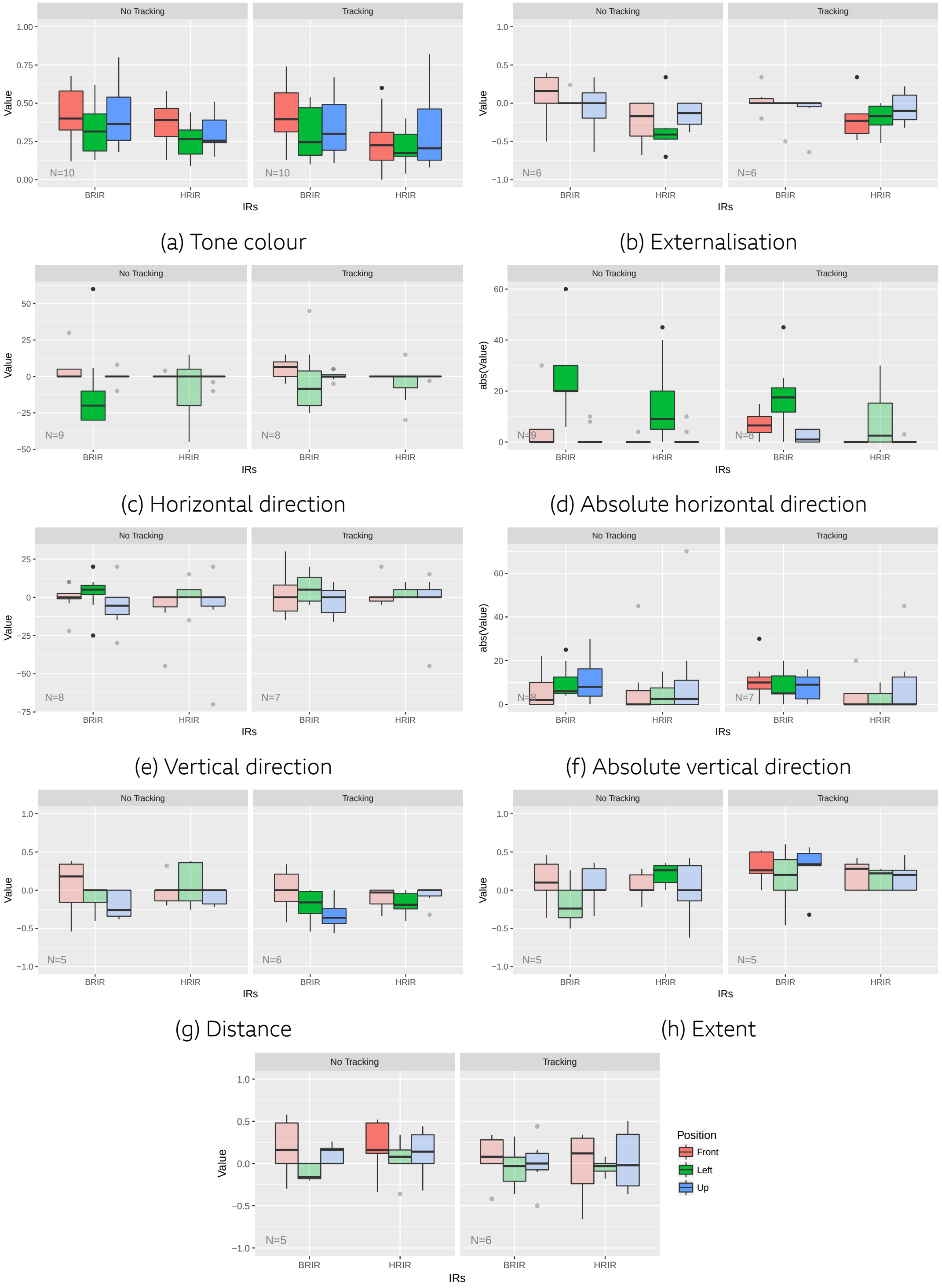
Figure 7.9: Combined box and scatter plots of ratings for each attribute over all rendering conditions by assessors who selected the attribute[1]. Blue boxes indicate significant deviation from zero.

changes in vertical position were significant overall, but horizontal position changes were not (when only considering assessors who selected the attributes).

Figure 7.10 shows box plots for the attributes that were selected by at least 50 % of assessors, with separate distributions for each rendering condition. Only data from those assessors who selected the attribute are used i.e. no zeros are inserted for the other assessors. Boxes are shaded with full opacity if the inter-quartile range does not intersect zero, indicating potentially systematic effects, following Lindau, Brinkmann, and Weinzierl, 2014. For horizontal and vertical direction ratings, the absolute values are also plotted, representing localisation errors in either direction.

For the *tone colour* data, a RM-ANOVA was performed. It showed small significant effects for *IRs* ($F(1,9) = 7.061, p = 0.026, \eta^2 = 0.067$) and *Position* ($F(2,18) = 4.036, p = 0.036, \eta^2 = 0.059$). A multilevel model for this data gave the same significant effects: *IRs* ($\chi^2(1) = 11.24, p < 0.001$) and *Position* ($\chi^2(2) = 11.15, p = 0.004$). Virtual loudspeaker rendering added significantly more colouration with BRIRs than with HRIRs ($b = 0.045, t(18) = 3.53, p = 0.002, r = 0.639$). The front position showed significantly more differences than

[1] Direction attributes have been normalised to the range ±1. The distributions of absolute values are also plotted.



(a) Tone colour

(b) Externalisation

(c) Horizontal direction

(d) Absolute horizontal direction

(e) Vertical direction

(f) Absolute vertical direction

(g) Distance

(h) Extent

(i) Comb-filter colouration

Figure 7.10: Box plots of attribute rating distributions, showing differences between direct binaural and virtual loudspeaker rendering. Data is shown only for the number of listeners (N) who selected the attribute. Transparency used when inter-quartile range intersects zero.

the other positions ($b = 0.021, t(72) = 2.33, p = 0.023, r = 0.264$), and the left position showed smaller differences than the elevated position (*up*) ($b = -0.037, t(72) = -2.34, p = 0.022, r = 0.266$).

A multilevel model was fitted for each of the attributes selected by more than 50 % of assessors, omitting the missing data. The results are summarised in table 7.7. Only a few effects are significant.

Attribute	*Tracking*	*IRs*	*Position*	*Tracking: IRs*	*Tracking: Position*	*IRs: Position*	*Tracking: IRs: Position*
Tone colour		***	**				
Horizontal direction			*				
Vertical direction			.				
Externalisation		**				.	
Distance	.		*			*	
Comb-filter colouration							
Extent	.						

Table 7.7: Multilevel model effect significance for commonly selected attributes.
Symbol *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$, . indicates $p < 0.1$.

7.4.2.2 Exploring Condition Differences with Factor Analysis

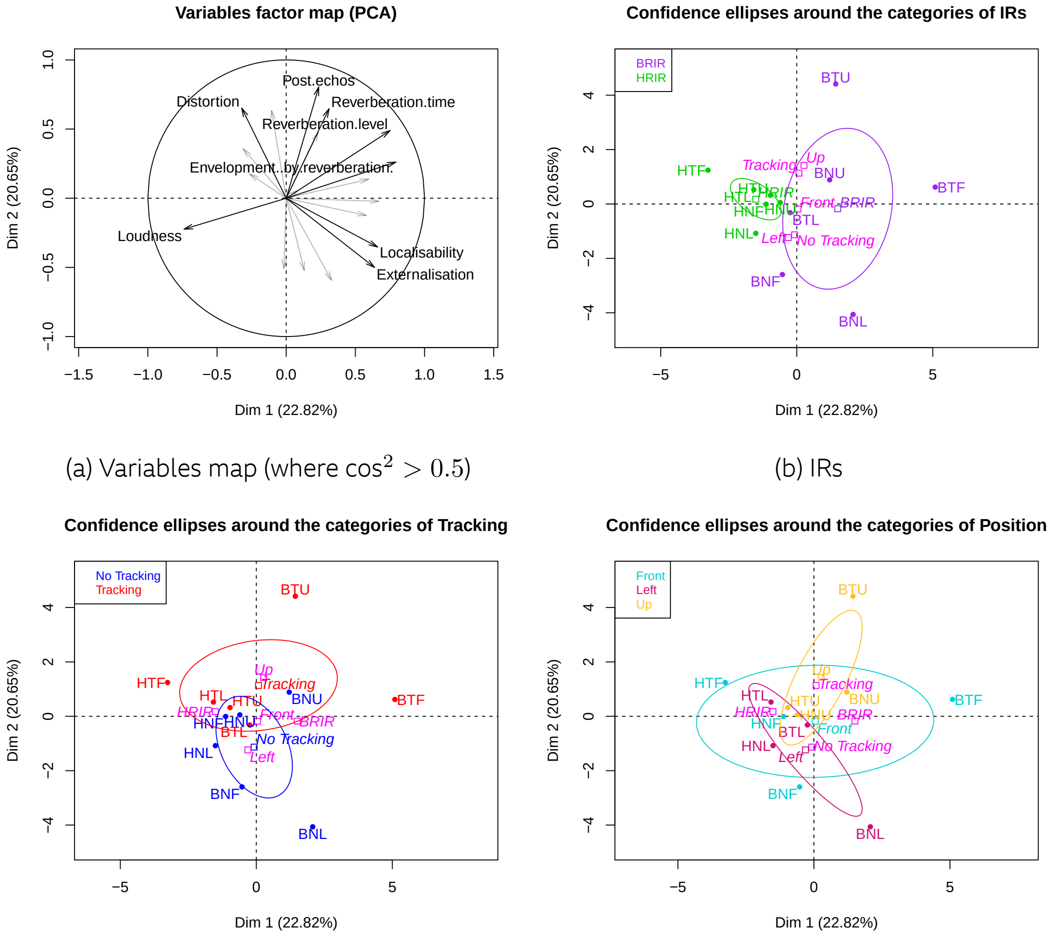
A detailed study on analysis of RATA data was carried out by Meyners, Jaeger, et al. (2016). Lack of selection of an attribute was defined as a score of zero, yielding 4- or 6-point intensity scales. With comparison to “gold standard” data-based random permutation tests, it was found that common parametric tests such as F- and t-tests gave valid inference, despite skewed distributions.

Response data for a set of multiple attributes are often explored with factor analysis techniques, to identify the relationships between the test conditions and the attributes in a lower dimensional space (Husson, Lê, and Pagès, 2010). An approach to principal component analysis (PCA) of RATA data is described by Meyners, Jaeger, et al. (2016), after Dravnieks (1982). The mean ratings are taken for each combination of system condition and attribute, ignoring cases where assessors did not give a rating (because they felt the attribute did not apply). These are then weighted by the selection rate across assessors for that combination of attribute and system condition. This balances the analysis between the actual rating scores and the selection rates. Alternative approaches would be to insert zero ratings when an attribute was not selected and perform a standard PCA, or to use only selection frequency data and perform correspondence analysis (CA). The Dravnieks method of using weighted means incorporates both aspects and can be seen as an intermediate approach between the other two alternatives.

This selection-rate-weighted PCA was carried out on the attribute data, with the conditions *IRs*, *Tracking* and *Position* used as qualitative supplementary variables. For this analysis, horizontal and vertical direction data were converted to absolute values, to capture the localisation error within the mean. The `FactoMineR` R package was used to perform the analysis. It allows 95 % confidence ellipses to be added to plots, indicating significant differences between conditions in the factor space. The function `dimdesc` also indicates which attributes and conditions define the dimensions well. Since ratings in this experiment were indicating differences between virtual loudspeaker and direct binaural rendering, the PCA indicates variations in these differences according to the various system conditions evaluated.

The first four dimensions together explain 69.3 % of variance. The first two dimensions are shown in figure 7.11. The first dimension distinguishes between the *IRs* conditions, being characterised by the reverberation level and envelopment, loudness, and externalisation. Externalisation was particularly negatively affected by virtual loudspeaker rendering when using HRIRs, and reverberation level and envelopment were increased when using BRIRs. Loudness also appears to be increased by loudspeaker virtualisation more when using HRIRs for the rendering. The third and fourth dimensions are shown in figure 7.12. The third dimension characterises the left position, for which virtual speaker rendering appears to have less effect on tone colour, comb-filter colouration and distance changes. The fourth dimension characterises tracking most strongly, relating to comb filter colouration and front-back position. The analysis suggests that without tracking, comb filter colouration becomes more noticeable with virtual loudspeaker rendering. It also suggests that front-back reversals between the direct binaural and virtual loudspeaker rendering methods are more frequent when head tracking is used. However, there is little data to suggest this, only two assessors selected front-back position for each tracking condition, and for each, three reversals were indicated. It is wise to avoid over-interpreting the analysis for attributes when there is so little data available.

The analysis was run again with only those attributes selected by at least 50 % of assessors (those plotted in figure 7.10), to avoid those infrequently used from dominating the analysis. In this case, the first four dimensions explained 92.2 % of variance. The first two dimensions are shown in figure 7.13. The left position was distinguished from others, most strongly in the second dimension, which is associated with horizontal direction and comb filter colouration. Though the effects are not significant, the *IRs* and *Tracking* conditions are



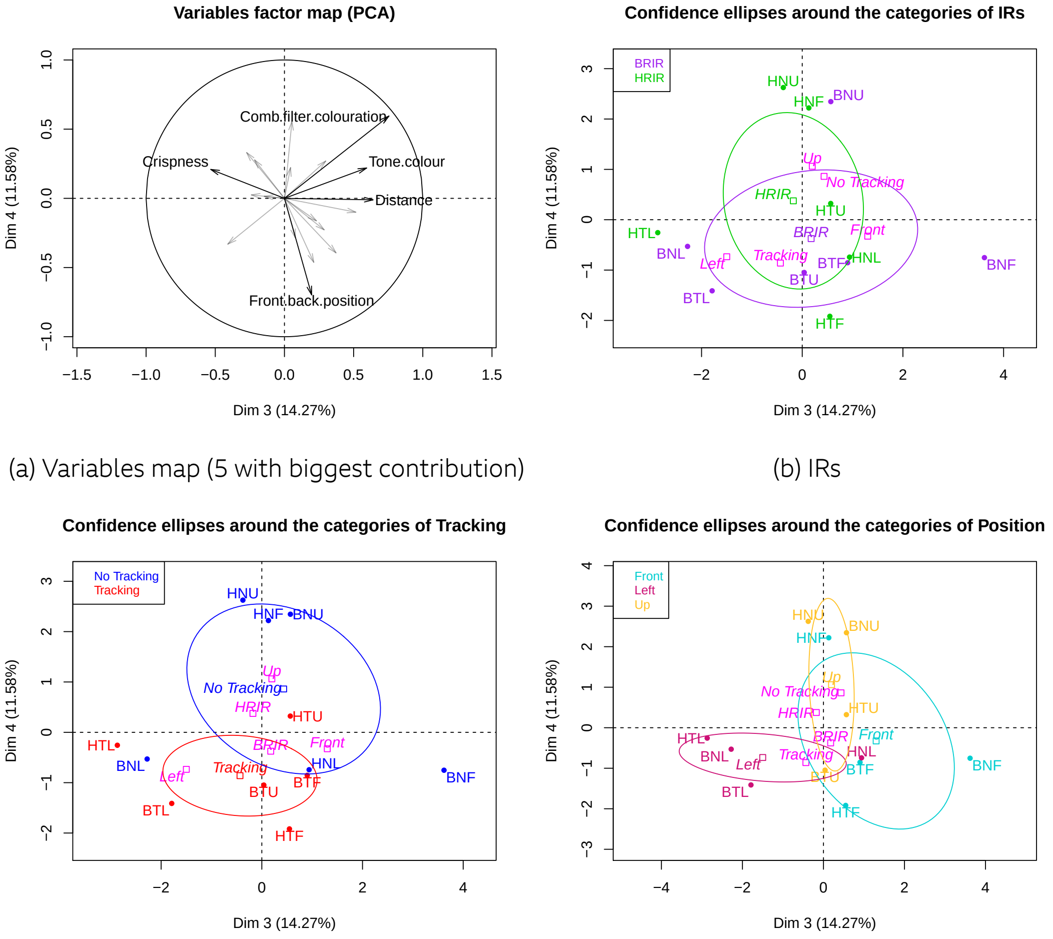
(a) Variables map (where cos$^2 > 0.5$)

(b) IRs

(c) Tracking

(d) Position

Figure 7.11: Dimensions 1 and 2 of PCA of selection-frequency-weighted mean ratings using all attributes.



(a) Variables map (5 with biggest contribution)

(b) IRs

(c) Tracking

(d) Position

Figure 7.12: Dimensions 3 and 4 of PCA of selection-frequency-weighted mean ratings using all attributes.

somewhat separated along the first dimension. The positive direction of which is defined by increased changes in tone colour and comb filter colouration, as well as vertical direction errors and externalisation. Rendering with BRIRs and without tracking are associated in this direction, with HRIRs and tracking associated with the negative direction of this axis. The third and fourth dimensions are shown in figure 7.14. The third dimension separates position *up* from the other positions, distance is more negatively affected for this position. Both *IRs* and *Tracking* conditions are distinguished along the fourth dimension, which is defined by extent and externalisation. Externalisation appears more negatively affected when HRIRs are used, whereas extent appears to increase more when tracking is used.

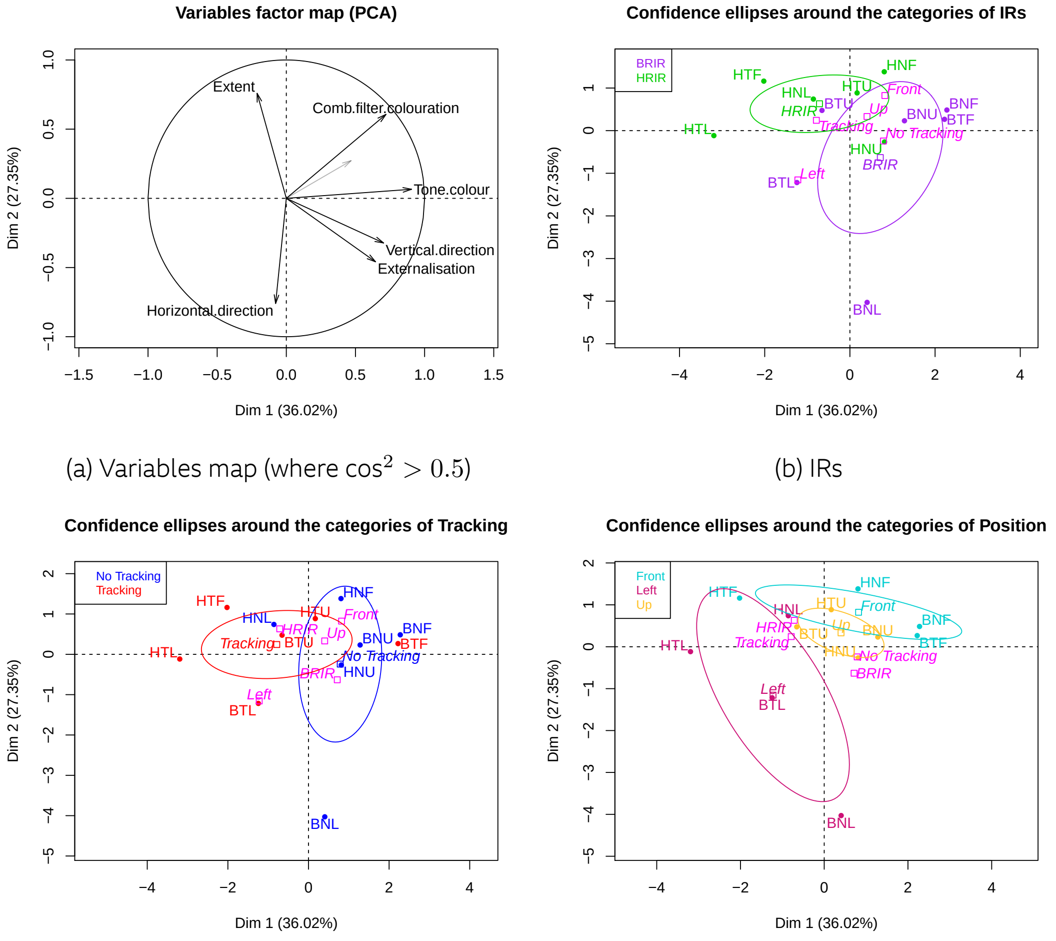
In Figures 7.11a and 7.13a, the labelled variables are those for which the angle from the variable vector to the plane is $< 45°$, since others are not well represented by the axes. For axes 3 and 4, only one or two variables were well represented. For this reason, the five variables which contributed most strongly to the construction of the axes are shown in Figures 7.12a and 7.14a. The quality of representation in the plane can be seen by the vector length.

7.5 Discussion

7.5.1 Experiment Findings

The virtual loudspeaker rendering approach used in this test leads to perceptible changes when compared with direct binaural rendering. The changes are larger with BRIRs than with HRIRs. Both timbral and spatial characteristics are affected. In general, tone colour is modified, extent increased, and distance, externalisation, and localisability are decreased. Position changes are also common, though the direction of changes is not consistent.

Tone colour changes are dominant, as shown by the frequency of selection and the magnitude of the perceived effect. It seems that the tone colour differences were larger with BRIR rendering. This differs from the findings of Pulkki (2001a) for loudspeaker listening, where colouration was larger in anechoic conditions. This may be due to the influence of non-individual binaural filters and general inaccuracies of simulation in the binaural rendering interacting with the room response. The lack of small head translation effects that would naturally occur in loudspeaker listening may have an impact on the perceived colouration, since any comb filter effects will be unnaturally constant. It could also be due to imperfect time alignment of the loudspeaker signals to the central listening position in the BRIR measurements. The loudspeakers were time-aligned to within 20 μs at the central listening position before BRIR measurement.



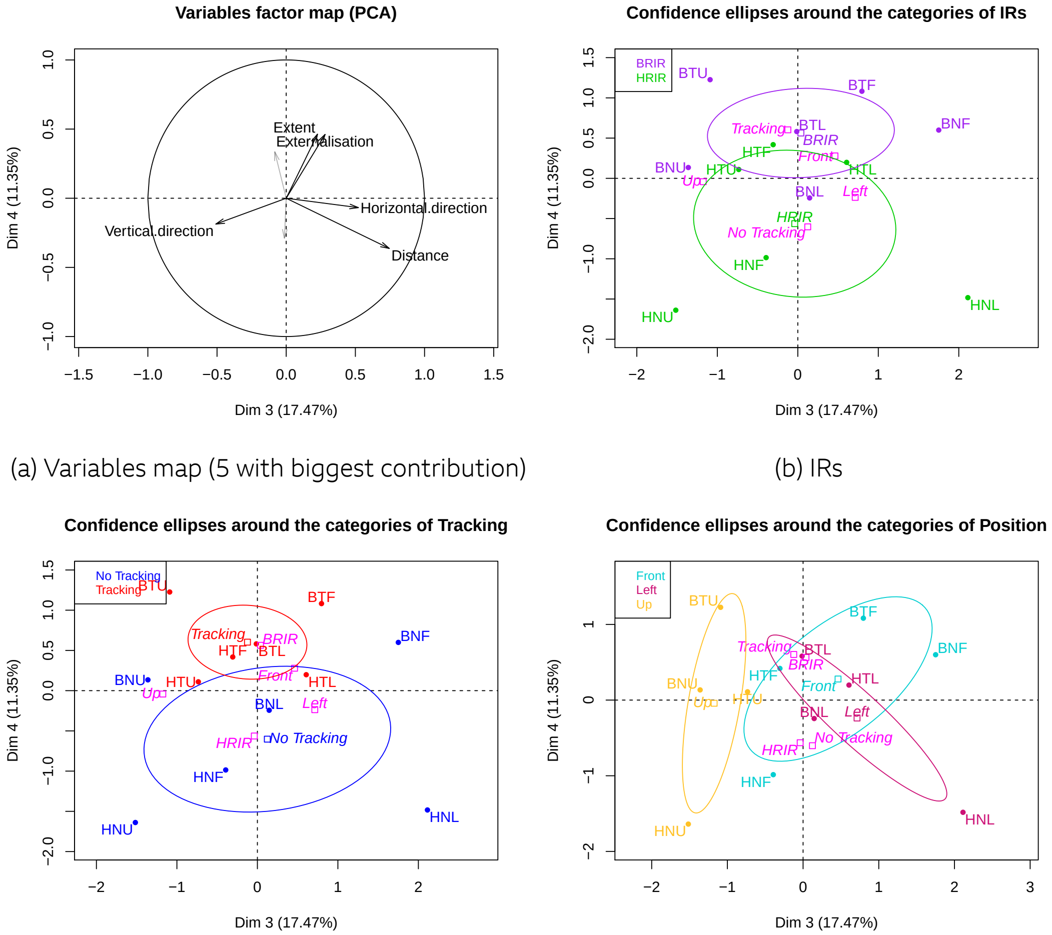
(a) Variables map (where $\cos^2 > 0.5$)

(b) IRs

(c) Tracking

(d) Position

Figure 7.13: Dimensions 1 and 2 of PCA of selection-frequency-weighted mean ratings using only attributes selected by at least 50 % of assessors.



(a) Variables map (5 with biggest contribution)

(b) IRs

(c) Tracking

(d) Position

Figure 7.14: Dimensions 3 and 4 of PCA of selection-frequency-weighted mean ratings using only attributes selected by at least 50 % of assessors.

The level of tone colour changes also appears to vary with target source position, with the front source having the largest changes and the lateral source having the smallest. This is likely influenced by the magnitude of TOA differences at the ears, which is lowest for the lateral loudspeaker pair used for the left position and highest for the loudspeaker pair for the front position at 0° elevation. When estimated using the Neumann KU100 HRIRs for frontal head orientation, the TOA difference between the pair of virtual loudspeakers at the left ear is 260 µs for position *front*, 200 µs for position *up* and 14 µs for position *left*.

The nature of the tone colour differences is complex. Comb filter colouration was often also identified. This can be seen as a specific sub-component of tone colour. Some listeners may have specifically identified comb filter artefacts while others may have assigned similar perception to tone colour. The SAQI defines several timbral attributes, including those for changes specific to high, mid, and low frequency regions. The tone colour attribute used here was an aggregate. It was chosen because it was felt that the colouration effects present would be difficult for assessors to further sub-divide. However, given the relatively large effect observed for this attribute, it would be worth investigating more specific timbral attributes in future.

The comb filter colouration scale was bipolar, as opposed to the unipolar tone colour difference, and it seems listeners were often divided over whether it had been increased or decreased. Colouration is generally considered a difference measure, e.g. (Salomons, 1995). With two binaurally rendered noise stimuli and no declaration of which is the original signal, it is perhaps understandable that whilst hearing differences, assessors could not consistently identify the direction of increased comb-filter colouration. In general, it cannot be said from the data which rendering approach sounds best, assessors were just asked to rate differences.

Changes in source direction were also common. In terms of azimuth (horizontal direction), there were large changes in the lateral source position. This aligns with earlier studies of loudspeaker rendering. The errors for BRIR rendering of the front source position show a slight trend towards clockwise movement, which suggests that there might have been a small error in the time alignment of the loudspeaker signals. For the vertical direction, the change was largest with head-tracked BRIR rendering, possibly because in this combination it is easiest to identify source elevation and so detect changes. It was often commented by listeners that, with head-tracking, the source position changed during head movement more in one case than the other (within the A-B comparison). This was difficult to capture in the rating procedure and individual strategies for reporting this will have varied. So-called 'modifications' reporting, proposed in the SAQI method, could have been used to capture this, whereby aspects of an attribute that vary temporally or due to interaction are indicated.

This would have increased the length of the evaluation however. Alternatively a distinct attribute, such as *stability* could be defined to capture this.

The source extent was mostly increased with the virtual loudspeaker approach, apart from in the lateral BRIR rendering without tracking. Extent increases also appeared slightly greater with head tracking on. It may be that increased extent was assigned to other characteristics without head tracking, such as lower localisability, since head movement cannot be used to resolve the source position outside of the head. The source distance tended to be reduced, but there was an interaction between IRs and tracking conditions. The elevated position during BRIR rendering exhibited the biggest distance reduction due to virtualised amplitude panning, whilst for the frontal BRIR rendering it seems there was a spread, with distance sometimes increased and sometimes reduced. It is not clear why this occurred. Externalisation was clearly decreased by using virtual loudspeakers for HRIR rendering but little effect was seen for BRIR rendering. It seems that the room response provided adequate cues for out-of-the-head localisation, even when virtual loudspeaker rendering negatively affected other attributes. Interestingly it seems that some participants perceived an increase in externalisation for frontal sources in this case, as well as an increase in source distance.

When analysing the differences it is worth considering that, if the percept were weak or confused in the direct binaural rendering scenario then, it is likely that differences will be small. As an example, for an anechoic rendering without head tracking, distance will likely already be low, especially considering this non-individualised rendering, and so it perhaps cannot be changed a great deal by introduced impairments.

In addition to those attributes plotted, localisability was more clearly reduced in the cases with head tracking and a ghost source was only identified in the cases with tracking, but given the small number of listeners involved this would need further investigation.

When the direct binaural rendering method was used with head tracking, crossfading was used to switch between selected IRs during head movements. With a 2° resolution of measurements this will have happened frequently during listening and so a process similar to virtual loudspeaker rendering will have been performed, however with much finer angular separation. The onset delays were also interpolated separately in the direct binaural rendering to reduce comb filtering effects.

The differences caused by amplitude panning in binaural rendering seem similar to those observed on real loudspeakers in the literature. The effects of amplitude panning on loudspeakers have not been evaluated in terms of this set of attributes previously and a comparison would be needed to indicate the degree of this similarity. In-head localisation is not a frequent occurrence for amplitude-panned loudspeaker-reproduced sources, so there are clearly some differences.

Where content is produced while monitoring on a loudspeaker array, the sound engineer will likely account for the behaviour of the system in the process. In this case, virtual loudspeaker rendering for headphones may be more faithful to the intended character than a direct binaural rendering of the object-based scene, despite better spatial and timbral quality characteristics being achievable with direct rendering. This would need to be verified through further investigations.

It is clear that virtualised panning has some undesirable effects on quality features, but in many applications of binaural technology this approach is used. The findings of this study indicate quality features that should be evaluated when trying to improve loudspeaker virtualisation techniques for headphone rendering. Further analysis and modelling of the causes of these observed perceptual effects is necessary to direct such work. Appendix B.5 discusses some relevant techniques for this purpose.

This study has evaluated a small set of source and loudspeaker positions, with a single audio test signal as the input. It has allowed controlled investigation of the low-level perceptual effects, but investigation with a wider range of source positions and audio material would give further insights. Comparison between loudspeaker virtualisation and direct binaural rendering using more representative programme material, with more complex and varied scenes, is likely to give better indication of how the techniques will impact on the overall listening experience in applications. With multiple sources at a range of directions, the way in which loudspeaker virtualisation is used must be considered, including the number and positions of the virtual loudspeakers. There are infinite potential approaches and a number of different configurations are common in practice. This will be the subject of further investigation in chapter 8.

7.5.2 Experimental Method

Most participants commented that the attributes are not orthogonal. In general the pattern of changes introduced by the virtual loudspeaker rendering has common characteristics across the range of binaural rendering configurations. Listeners often reported relationships between attributes such as distance, extent, localisability, externalisation, and spatial disintegration. The PCA results indicate where attribute variables were correlated across the experiment conditions. Sometimes, in order to capture the effect in such situations, participants reported a subset of the attributes which were influenced. More in-depth training with further examples may help listeners to differentiate, but it seems clear that there are related percepts within the SAQI, especially in the context of the stimuli presented in this evaluation.

Listeners commented that the test procedure itself was not overly complex or fatiguing. However, the test signal was fatiguing; this should be reconsidered in future. Reporting direction changes in degrees was challenging without an external reference, which may have increased variance. A visual guide or an alternate reporting method could be introduced. Some of the SAQI attribute definitions could be revised to more natural phrasing for a native English speaker and some assessors found the language confusing. Also, certain examples given are strongly focussed towards engineering knowledge which is not always appropriate. Further development and standardisation of listening examples for attribute training would allow more confidence in the listener's ability to identify and discriminate in terms of the perceived sound character, rather than relying on written descriptions.

The range in number of attributes selected is of concern, since those listeners who selected the most attributes found the test tiring. Some listeners reported feeling that some selected attributes were not that useful during rating, whilst others wished they had selected additional attributes with hindsight. The procedure should perhaps be adapted to allow more time for review of the selection before rating. With hindsight, attribute selection should be performed by presenting one attribute at a time, to encourage proper consideration of each attribute.

An additional challenge is that listeners may become aware that the attribute selection will alter the duration of the test and become influenced by this. Meyners, Jaeger, et al. (2016) term this response strategy *satisficing* and acknowledge that there might be a risk of this with the RATA approach, but also suggest that rating every attribute presents a risk of tedium and fatigue, which could also reduce the accuracy of assessor evaluations. The RATA approach clearly reduces the number of ratings required in the SAQI test method, allowing multiple factors to be compared in a manageable time-frame while still leaving listeners free to identify the attributes that they perceive to be relevant. It may be a useful approach for research questions that require an exploratory descriptive analysis and need to make efficient use of an expert's time.

The ITD scaling procedure was challenging for assessors and required a lot of guidance from the experimenter. It is an unusual task for most listeners. The process should be refined for future experiments.

This experiment involved only ten assessors. This made analysis challenging for attributes that were not selected by most assessors. A larger scale study could be performed, however the results presented here already give some insights into the effects of the loudspeaker virtualisation approach.

RATA was developed as a method for comparing multiple systems (or products) with target users (consumers). Here only differences between pairs were rated, giving a profile of those differences. It would be of interest to apply the RATA method to compare multiple systems in a future study, to obtain a characterisation of each system. However, with multiple systems and many attributes, as in SAQI, the evaluation process would likely become unmanageably long. Therefore, such studies would require selection of a reduced number of attributes by experimenters or a small-scale pilot experiment.

Recently Ares, Picallo, et al. (2018) compared use of RATA with inexperienced assessors to conventional DA with expert assessors, to evaluate complex and similar stimuli. Results showed that, in such studies, the conventional DA provided better discrimination and different characterisation of differences. When expert assessors are available and experiment duration is manageable, conventional DA can be expected to provide more detailed information. In this study, the RATA approach has allowed exploration of a large set of potential quality features. A consensus-based approach to reducing the initial attribute set could provide similar information, allowing assessors still to indicate the attributes that they feel are relevant, but only rating attributes that were commonly identified, for example.

Alternatively, if the CATA method can provide adequate discrimination between multiple system conditions then this offers a simpler, and probably more efficient, response format. This would enable consideration of a large set of attributes within a reasonable time frame. CATA is used in chapter 8 to compare binaural rendering of different spatial audio signal representations.

7.6 Conclusions

The perceptual effects of the virtual loudspeaker approach to providing spatial sound for headphone listeners were evaluated in a listening experiment, by comparison to direct binaural rendering of a target source. This was to better understand the implications that use of this common approach will have on perceived quality. A descriptive analysis (DA) method was used, based on the quality features of the spatial audio quality inventory (SAQI), with a rate-all-that-apply (RATA) approach. Assessors were able to select the subset of presented attributes that they felt differentiated the pairs of stimuli under evaluation, thereby only rating the characteristics that they perceived to be relevant.

The results showed that tonal colouration is the most prominent effect, with changes in source direction also common. Often listeners also perceived a decrease in distance and externalisation and an increase in source extent. Interactions with system rendering factors were observed, as well as the effect of target source position.

Chapter 8

Characterising the Quality of Dynamic Binaural Rendering of Spatial Audio Formats using a Check-All-That-Apply Approach

This chapter presents an experiment to characterise the quality of dynamic binaural rendering using seven different spatial audio formats that are common in emerging applications. Original object-based scene representations were rendered for headphones using either per-object HRTF convolution or via intermediate loudspeaker virtualisation approaches based on either ambisonics or vector base amplitude panning (VBAP). Evaluation was performed both for single musical sources and complex dramatic scenes using a pre-defined set of quality features to characterise the quality of the systems. The check-all-that-apply (CATA) method was used to allow characterisation across a wide range of features.

8.1 Introduction

Loudspeaker virtualisation techniques are widely used in applications of binaural technology in entertainment media. This includes relatively recent developments: next-generation audio (NGA) codec systems for efficient delivery of 3D spatial audio signals and the spatial audio rendering systems in virtual reality (VR) and augmented reality (AR) applications (see section 6.3). The growth of these technologies makes it feasible to provide head-tracked binaural rendering for mass audiences, so offering interactive spatial sound experiences for headphone listeners.

In chapter 7 the perceptual effects of loudspeaker virtualisation were investigated under different binaural rendering system conditions. Using a noise burst signal, pairwise amplitude panning to virtual loudspeakers was compared to direct binaural rendering of the source for three different target source positions. This study showed that virtualised amplitude panning has detrimental effects on several timbral and spatial quality features. Despite these issues, loudspeaker virtualisation is used for practical reasons, due to constraints within the application context.

Section 6.3 discusses the different formats for representing 3D spatial audio programme material that are widely used in spatial audio technologies and applications: object-based, channel-based, and scene-based. An object-based representation uses separate audio signals for each element within the scene. Accompanying data, which may be time-varying, describes the spatial parameters for these signals. The bandwidth of the representation therefore scales with the scene complexity. When the direct binaural rendering approach is applied to an object-based scene representation, a binaural filter convolution operation is required for each sound source and so the rendering complexity also increases with more elements in the scene.

By contrast, when using either channel-based or scene-based formats, audio programme material can be represented with a fixed number of signals. This limits the data bandwidth required for distribution and storage of the signals. It also means that the computational complexity of the binaural rendering process can be kept constant and independent of the complexity of the programme material. The loudspeaker virtualisation process, introduced in section 6.4, can be applied to render these formats to a binaural headphone signal. For scene-based ambisonics signals, a decoding step must first be applied to generate the loudspeaker signals.

An object-based scene representation, or parts of it, may be rendered to an intermediate channel-based or scene-based format prior to distribution; there are well-known techniques for doing so. Less interactivity and adaptation are possible with channel-based or scene-based representations, since the elements of the scene are combined. For example, adaptation to listener movement (tracking) is then largely limited to orientation only. In interactive auditory virtual environment (IAVE) systems, particularly for VR and AR on low-power mobile devices, rendering is often performed by virtualisation of an intermediate loudspeaker array, despite an object-based representation being available from the game engine (Google, 2018b; Facebook, 2018a). Similarly, NGA delivery codecs for spatial audio produce headphone signals using loudspeaker virtualisation, even when an object-based representation is available (Herre, Hilpert, et al., 2015; Dolby, 2015). Limiting the computational requirements for the binaural rendering is important in these systems.

There are many potential approaches to representing and rendering 3D spatial audio programme material for delivery to a mass audience. Technology and media service providers need to make informed decisions about the best approaches to take within available resources. Better understanding of the quality of different techniques and formats should help to improve quality of experience for audiences. From the review in section 6.5, it seems that there has previously been little comparison of the quality of object-based, channel-based and scene-based representations of 3D spatial audio and associated rendering techniques for headphone delivery.

This chapter presents a study which characterises the quality of a representative set of seven approaches. These have different levels of complexity in terms of required signal bandwidth and rendering computations. To ensure that the evaluation is representative of the applications under consideration, complex audio scenes are used, but a set of single sound sources are also evaluated, to obtain more controlled assessment of perceptual characteristics of the approaches. To obtain an efficient characterisation of the quality of the different methods, in terms of a wide range of quality features, a listening experiment was designed using the check-all-that-apply (CATA) method.

Section 8.2 gives the background to the CATA method and considers its suitability for evaluating spatial audio systems. Section 8.3 then describes the experimental apparatus used to compare different spatial audio formats and binaural rendering techniques in a listening experiment. The experimental design is described in section 8.4 and the analysis of results is presented in section 8.5. Section 8.6 discusses the findings, and conclusions are drawn in section 8.7.

8.2 Check-all-that-apply

Check-all-that-apply (CATA) is a method for rapidly obtaining sensory profiles of products or systems using a panel of human assessors. The assessor is presented with a list of attributes from which they should indicate those that best describe their experience with a stimulus, for example with a check- or tick-box. An excellent overview is given by Meyners and Castura (2014).

CATA was first developed in market research and was applied to sensory evaluation by Adams et al. (2007). Although earlier studies had used the question format to indicate the presence or absence of sensory characteristics, the novelty in this study was to use the technique with consumers, who also provided preference ratings. The attributes often include sensory characteristics, as well as hedonic and emotional responses. CATA results are often linked to consumer acceptance with accompanying *liking* questions or collecting CATA data for a hypothetical ideal product. The method has been applied many times to rapid sensory evaluation with consumers e.g. in evaluation of milk desserts (Barreiro et al., 2010), ice creams (Dooley et al., 2010), antiaging creams (Parente et al., 2011), fruit varieties (Ares and Jaeger, 2013), orange juices (Lee et al., 2013) and beers (Reinbach et al., 2014).

CATA questions give multivariate binary data. There is some ambiguity in the meaning of a single response. Whilst a checked attribute indicates that the assessor believes the stimulus to be characterised by the attribute, an unchecked term could either mean that the assessor did not perceive the attribute, or that it was perceived but the assessor considered that the attribute was not appropriate for characterising the stimulus. The binary response clearly gives no indication of the intensity of the sensation, and when relating the CATA data to overall quality or liking, no direct indication of the valence is given, i.e. whether this characteristic attribute has a positive or negative effect overall.

8.2.1 Comparison with other methods

CATA differs from free elicitation methods, which require subsequent classification of responses prior to analysis, e.g. (Francombe, Brookes, et al., 2017). Instead, assessors use a pre-defined list of attributes, which removes the requirement for the experimenter to interpret responses. The task of interpretation is instead on the assessor when considering the attributes. In conventional descriptive analysis (DA), expert assessors need significant training to use the provided scales in a consistent manner. Individual vocabulary techniques avoid this, allowing the assessor to define and use their own attribute scales, but instead demand greater analysis effort and experimenter interpretation. In CATA the task is simpler, with no scaling required, so it may be less affected by limited experience, although assessors may still interpret the terms differently. The simpler task compared to scale ratings may ease the load on the assessors and enable more rapid evaluation, whilst also allowing a greater number of characteristics to be considered (Meyners and Castura, 2014).

CATA evaluation with untrained non-expert assessors (commonly referred to as consumers in the literature) has been found to yield product characterisations similar to those obtained from conventional DA with expert assessors (Barreiro et al., 2010; Dooley et al., 2010). The CATA approach, when used with experienced assessors, was compared to conventional DA by Campo, Ballester, et al. (2010). It should be noted that assessors were trained extensively in using and refining the list of CATA attributes for this study, including provision of odour references. In this case, CATA was shown to yield finer characterisation of wine aromas, with additional practical benefits that the training was more generally applicable to different sets of wine products than conventional DA. This was partly due to one product having significantly different characteristics to the others, which in the principal component analysis (PCA) based on DA ratings explained a large proportion of the variance and meant that other products were less well differentiated. A greater number of descriptors were also used by assessors with the CATA method.

Reinbach et al. (2014) compared CATA with rate-all-that-apply (RATA) and projective mapping. CATA and RATA gave very similar results in this study, suggesting that the extra intensity information in RATA did not offer a significant advantage in this case. It was noted that there were rather large sensory differences between products in these experiments and future research might investigate the discriminative power of CATA in less heterogeneous sample sets. Ares, Deliza, et al. (2010) compared CATA with projective mapping and found that they yielded very similar sensory profiles. Vidal et al. (2018) performed an extensive comparison of CATA and RATA approaches across seven experiments with between-subjects designs. It was found that they gave very similar outcomes, with minor differences that suggest the best choice of approach is dependent on the nature of the experiment. The RATA format is recommended when it is expected that the salient attributes will be present in most stimuli but at different intensities.

8.2.2 CATA Experiment Design Considerations

In response to the great interest in the CATA method in the field of sensory science, there have been a number of investigations into factors of the experimental design. Through a number of studies, Jaeger, Chheang, et al. (2013) demonstrated that the CATA method provides stable product profiles and attribute usage with untrained assessors, representative of target users. It was found that assessors were able to reproduce their product characterisations, though with varying levels of accuracy.

The number of assessors required is high in sensory evaluation with consumers. For example, Ares, Deliza, et al. (2010) used 50 consumers to evaluate milk desserts, Parente et al. (2011) used 69 consumers to evaluate six anti-aging creams, Dooley et al. (2010) used 80 consumers to evaluate vanilla ice-creams, and Plaehn (2012) used 120 consumers to evaluate citrus-flavoured sodas. Ares, Tárrega, et al. (2014) analysed data from 13 CATA studies and found that 60-80 consumers are needed to obtain stable profiles of samples. However, with assessors who have domain-specific experience, it is expected that useful results can be obtained with fewer. For example, Campo, Ballester, et al. (2010) and Campo, Do, et al. (2008) used 33 and 36 trained assessors to evaluate wine aromas, respectively.

The number of attributes used and the way they are presented to assessors varies across studies, and these factors can have an effect on results. Krosnick (1999) describes how, in a questionnaire, fatigue can lead to suboptimal responses that are nonetheless believe to be adequate; a process termed *satisficing* (also discussed in section 7.5.2). When satisficing occurs, the order of items in a questionnaire can bias responses. With a CATA question, or other visually presented questionnaire, *primacy* is the most significant aspect of satisficing, where assessors are more likely to use the first attributes in the list.

Jaeger, Beresford, et al. (2015) compared studies with long (20–28) and short (10–17) lists of CATA attributes. Longer lists may increase the likelihood of satisficing effects. They noted some evidence of dilution in attribute selections when using long lists containing synonyms and antonyms. They recommend experimenters not to use an excessive number of terms, but still to include a range of related terms to account for assessor heterogeneity.

Positional biases cannot be removed during visual presentation of attributes, but they must be balanced in the experimental design. Ares and Jaeger (2013) found evidence of primacy effects on attribute salience in CATA studies. They also found that entirely random ordering led to reduced total frequency of usage of attributes compared with grouping attributes with similar terms. Ares, Etchemendy, et al. (2014) used eye-tracking technology to explore the visual attention of assessors during rating. Over the course of the experiment assessors used fewer and shorter eye movements to explore the list of attributes. They recommend that attributes are randomised for each trial, since this led to greater visual attention to the attributes throughout the test, which was taken to indicate deeper engagement with the task. However, Meyners and Castura (2016) found that randomisation of attributes per-assessor rather than per-trial gave better operational power. The authors suggested that consistent positioning of attributes for the assessor may reduce cognitive load and allow more effort to be assigned to the response itself. They also state that “random allocation of attribute list orders is essential for CATA studies; both ‘to sample’ and ‘to assessor’ allocation are clearly preferable to using the same fixed order for all assessors and all evaluations.”

Studies differ in whether terms of different categories are presented simultaneously or in separate questions. Parente et al. (2011) presented assessors with 42 terms in a single CATA question, covering sensory properties, emotions, pricing and applications, whereas Lee et al. (2013) presented three separate CATA questions for attributes in separate categories (appearance, flavour and texture).

8.2.3 Analysis of CATA Data

Meyners and Castura (2014) give an overview of techniques for analysis of CATA data. The basic components are:

- Significance tests for the binary attribute responses, to check whether each attribute can be used to differentiate the systems under evaluation.
- Multivariate analysis of attribute selection frequencies across the panel of assessors, to characterise the systems across the set of attributes and their differences.
- Analysing relationships between the attribute data and overall quality or preference ratings.

Often Cochran's Q test (Cochran, 1950) is used to assess significance of attributes, and contingency tables of attribute selection frequencies are analysed with correspondence analysis (CA) to identify patterns in system and attribute profiles. However, a range of different techniques have been used in the literature. More detail of the analysis techniques used in this study are given in the results analysis, section 8.5.

8.2.4 Considering Spatial Audio Evaluation

Experiences with 3D spatial audio systems have many different quality features and, at present, are novel to many listeners. Exploratory evaluation methods are often used to characterise the quality of them. There are multiple reasons why the CATA method appears suited to the evaluation of spatial audio systems.

8.2.4.1 Assessor Expertise

Standardised methods for sound quality evaluation within the ITU-R require expert assessors, with a high degree of sensitivity and reliability (ITU-R, 2014b). Expertise generally requires significant experience in the field of application and in using the evaluation methods.

Regarding 3D spatial audio evaluation, this is currently a challenge. There is limited example material for listener training compared with traditional formats, such as two-channel stereo. Also, whilst training items for coding artefacts have been generated to demonstrate isolated artefacts (Dick et al., 2017). Such a resource does not currently exist for spatial audio systems and may be difficult to generate. Spatial attributes appear to be interrelated and at a higher level of abstraction from technical implementation. The simpler response format of the CATA method, which makes it suitable for sensory evaluation using consumers, may also have advantages therefore in complex spatial audio evaluation.

There is still value in recruiting experienced assessors for perceptual evaluations, they are more likely to show sensitivity and reliability in evaluating stimuli, with a better understanding of the methods and terminology. However, the collected data should be analysed to assess the listener expertise for the specific experiment and post-hoc screening of assessors may be necessary.

8.2.4.2 Common Attributes

Conventional DA methods are reviewed in section 3.3.3. They can be very time consuming and require regular access to a panel of trained expert assessors, which can be problematic in industrial contexts. The derived terms in many different audio-related studies often have a large degree of overlap (Zacharov and Pedersen, 2015), despite their differing application domains. As an alternative to deriving attribute by panel elicitation and consensus, experimenters will often select attribute scales from a previously defined set, e.g. Cobos et al. (2015), Moulin et al. (2016), and Millns and Lee (2018), who used attributes from ITU-R (2015b), Pedersen and Zacharov (2015), and Rumsey (2002), respectively.

As discussed in section 3.2.5.5, attempts have been made to define comprehensive sets of quality features for more general use in audio-related studies. It is evident that, to some extent, there is a shared scientific language for describing the perceptual dimensions of sound quality. The spatial audio quality inventory (SAQI), presented by Lindau, Erbes, et al. (2014), was used in chapter 7. It gives a set of 48 attributes and scales for the evaluation of virtual acoustic environments. These were obtained through focus groups with domain-experts. The Sound Wheel attempts to provide a common lexicon of quality features for evaluation of reproduced sound. It was most recently described in ITU-R Report BS.2399 (ITU-R, 2017c). Co-authored publication Co.P. VIII presented the development of spatial attributes for the Sound Wheel (Zacharov, Pedersen, and Pike, 2016), which incorporates those of the spatial audio quality inventory (SAQI).

A common set of attributes has advantages in terms of consistency of interpretation across studies, but, as discussed by Berg and Rumsey (2006), it must be acknowledged that with novel experiences, shared understanding of attributes and how to use scales may be limited. They state: “in new contexts where the experience is limited, clearly defined attributes or scales may not exist”. Even though spatial audio technologies are becoming more widely available, the experiences they provide are still often novel to listeners. Terminology could be quite specific to a particular application, a set of systems or the audio material, or used differently in different contexts. Imposing attribute scales on assessors without their input may make rating more challenging if they do not understand or agree with the chosen attributes. Training is clearly important in this case.

8.2.4.3 Interrelated Scales

Lawless and Heymann (2010, p.229–230) describe the importance of lack of redundancy between attribute scales in DA, to reduce mental load on assessors. When referring to descriptive analysis of complex odours, Lawless states that “the use of simple and apparently independent intensity scales may produce the illusion that the odor experience is a collection of independent analyzable notes when it is not” (Lawless, 1999). Campo, Ballester, et al. (2010) discuss how a frequency of citation method (i.e. CATA) might be better suited to evaluating complex stimuli (particularly wine odours), where assessors may struggle to identify the intensity of a particular attribute across stimuli amongst complex mixtures. This might also be a relevant issue in spatial audio, where spatial attributes are often related (e.g. source extent, localisability and diffuseness).

8.2.4.4 Prior Application to Audio Systems

The CATA method was recently applied to evaluation of high-end loudspeakers by 26 inexperienced assessors (Hicks et al., 2018). This study was published whilst the present work was being carried out. The authors reported that this method provided only a coarse characterisation of the loudspeakers. This is partly due to the coarse nature of the reporting task itself and the level of experience of the assessors. However, many attributes did reveal significant differences between the loudspeakers and there were also significant differences in the CA factor maps, so the method did provide useful insights into the differing characters of the loudspeakers.

Hicks et al. (2018) suggest that the rate-all-that-apply (RATA) method, similar to CATA but with intensity rating for each relevant attribute, might provide more distinction between similar attributes and systems. RATA was applied in the study reported in chapter 7. Whilst further experience with RATA is required, it was found to be labour-intensive for assessors when many attributes are considered. It also requires a higher level of assessor expertise to adequately use the given scales. If CATA can provide adequate characterisation when used with experienced listeners, it presents a more efficient alternative.

8.3 Apparatus

A system was built to compare different approaches to real-time binaural rendering for headphones. Functionality included dynamic adaptation to listener head rotation, but not translation i.e. 3 degrees of freedom (DoF) tracking. The rendering approach was varied according to a set of different content representations, including object-based and various intermediate virtual loudspeaker formats. Comparison of these methods was enabled while controlling other system factors, including the head-related transfer functions (HRTFs), head tracking system, system latency, and headphone correction filter. This experimental apparatus is discussed in more detail in appendices A and B, but a summary of important aspects is given here.

8.3.1 Impulse Response Data

A set of far-field head-related impulse responses (HRIRs), measured with a Neumann KU100 dummy head microphone are described by Bernschütz (2013). These HRIRs are available in several full-sphere quadrature grids and are 128 samples in length; they were adopted for this study. Whilst it has been shown that personalised binaural rendering gives a better simulation of a real sound source across a wide range of attributes (Lindau, Brinkmann, and Weinzierl, 2014), current commercial systems are not personalised and so in the context of this research a non-individual system is relevant.

To enable real-time adaptation of the interaural time difference (ITD), broadband onsets were modelled and stored separately. The minimum-phase cross-correlation method of onset estimation was used and a parametric model for the time of arrival (TOA) (Ziegelwanger and Majdak, 2014) was then fitted to the estimates to achieve a smooth direction-continuous function, as described in appendix A.5.6. Modelled onsets were extracted with 100 times oversampling, giving a precision of 0.21 μs, with a 5-sample safety margin to preserve the onset envelope.

The headphone correction filter for the Stax SR-207 headphones described in section 7.3.1 was used in this study also (see figure 7.3 on page 281). The filter was applied to all HRIRs offline, rather than during real-time rendering. HRIRs were then converted to complex frequency-domain HRTF data at runtime and stored in memory.

8.3.2 Real-time Rendering

A software system to allow comparison of dynamic binaural rendering approaches was written in the C++ language. The core signal processing component of the system was a uniform-partitioned fast convolution engine (Wefers, 2014). When changing impulse responses due to source or listener rotation, cross-fading was applied over the duration of a single partition. Subsample onset delays were inserted in real-time using a variable fractional delay line (VFDL) (Välimäki and Laakso, 2000), to reduce comb filtering artefacts when cross-fading between HRTF filters and also to allow for personalised scaling of the ITDs (Jot, Larcher, et al., 1995). A marker-based optical tracking system provided listener head orientation data at a rate of 250 Hz and with a reported angular accuracy of 0.5°. A filter partition size of 128 samples was used, which was also the original HRIR length. The system was configured with a 256-sample input/output buffer size. The total system latency (TSL) from head movement to the corresponding change in headphone output was measured as $\mu_{\mathsf{TSL}} = 55.7\,\mathrm{ms}$, $\sigma_{\mathsf{TSL}} = 1.54\,\mathrm{ms}$, below the detection threshold for all listeners (64 ms) observed by Lindau (2009). This was similar to the apparatus used in chapter 7, though with increased buffer size to accommodate the greater computational processing requirements when rendering complex scenes.

8.3.3 Binaural Rendering Approaches

The dynamic binaural rendering of seven different content representations was evaluated. This was an attempt to cover typical approaches relevant to current applications, whilst controlling for other aspects of rendering as described above. The seven representations are outlined below.

8.3.3.1 Object-Based Binaural

Object-based rendering was performed using a binaural filter for every sound object in the scene. A binaural filter was selected for the target source direction using a nearest neighbour search within a 16 020-point Gaussian grid with approximately 2° spacing in both azimuth and elevation. To achieve dynamic rendering, target source directions were rotated in opposition to the tracked listener head orientation, and binaural filters were updated according to source or listener movement. Therefore, a change in head orientation caused filter updates for all sources. This rendering approach is described in more detail in appendix A.3, as a *basic binaural renderer*. In this chapter, this approach is indicated by the label *OB*.

8.3.3.2 Scene-Based Binaural using Ambisonics

Three ambisonics-based virtual loudspeaker rendering approaches were used, at ambisonics truncation orders 1, 3, and 5, herein labelled *A1*, *A3*, and *A5*. Ambisonics is described in detail in appendix B.4 and the common approach to binaural rendering of ambisonics with loudspeaker virtualisation is discussed in appendix B.4.6.

Table 8.1 gives a set of parameters used for each of the three systems and figures 8.2a, 8.2c and 8.2e plot the virtual loudspeaker positions using a Hammer projection. Each of the ambisonics approaches used a spherical t-design for the virtual loudspeaker positions, i.e. the sampling points of the spherical harmonic coefficient signals, where $t = 2N + 1$

Label	N	t	L	f_x
A1	1	3	8	748 Hz
A3	3	7	24	2805 Hz
A5	5	11	70	5701 Hz

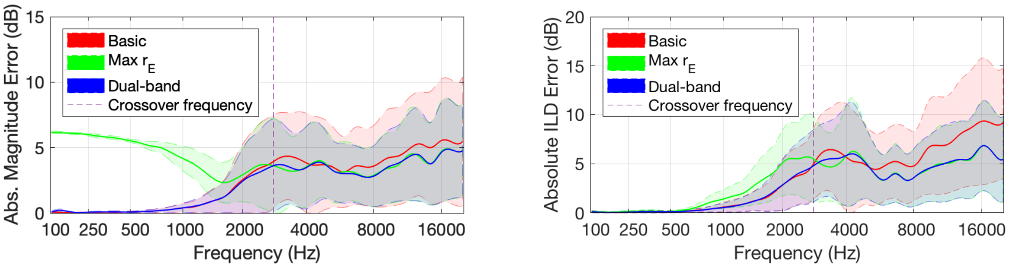
Table 8.1: Parameters for ambisonics-based binaural renderers. N is the ambisonics truncation order, L is the number of virtual loudspeakers used, t is the degree of the spherical grid, and f_x is the crossover frequency for dual-band decoding.

and N is the ambisonics truncation order. A spherical t-design can sample an arbitrary spherical polynomial of limited degree $n \leq t$, such that the discrete sum and the equivalent continuous integral are equal. For spherical designs with $t \geq 2N + 1$, panning invariant energy and spread is achieved, and basic sampling decoding gives identical results to mode-matching or energy-preserving solutions (Zotter and Frank, 2012).

Complex spherical harmonic interpolation of a 2702-point Lebedev grid of HRTFs was used to generate filters at the precise target positions of the sampling grid (Duraiswaini et al., 2004), using a series truncation order of 35. Appendix A.7 presents the details of this technique and appendix A.8 presents listening experiments showing that HRTFs can be generated for this Neumann KU100 dataset that are indistinguishable from real measurements.

The ambisonics decoding matrices were applied offline to generate a binaural filter for each ambisonics coefficient channel. Each filter is a weighted linear combination of the HRTFs for the target virtual loudspeaker position. This has been described previously, for example by Politis and Poirier-Quinot (2016). Where the number of loudspeakers exceeds the number of ambisonics signals this approach is more efficient than first decoding to virtual loudspeaker signals and then convolving those with HRTFs.

A dual-band decoding approach was used, with the max-$\mathbf{r}_E$ weighting applied at higher frequencies to maximise the energy concentration towards the target source direction at frequencies where the sound field reconstruction is less accurate (Daniel, Rault, et al., 1998). The crossover frequency was chosen based on an analysis similar to that of Daniel, Rault, et al. (1998), comparing the head-related impulse response simulated by panning through the virtual ambisonics system to the HRTF measured at the target source direction, at each of the 2702-point Lebedev grid positions. The crossover frequencies, given in table 8.1, were chosen to minimise the quadrature-weighted-mean absolute log-magnitude error. Figure 8.1 shows the mean and standard deviation of the magnitude response errors and the estimated inter-aural level difference errors over the sphere for the basic sampling decoder, the max-$\mathbf{r}_E$ weighted decoder, and the dual-band approach for system *A3*. Magnitude responses were first smoothed to approximate auditory frequency resolution, using a gamma-



(a) Magnitude response errors

(b) Inter-aural level difference errors

Figure 8.1: Absolute magnitude response and inter-aural level difference errors over the positions of a 2702-point Lebedev grid, for the basic sampling decoder (red), the max-$\mathbf{r}_E$ weighted decoder (green) and the dual-band decoding (blue) approach for the third-order ambisonics rendering (*A3*) when compared to measured HRTFs. Transfer functions were first smoothed by a Gammatone filterbank with ERB spacing before calculation. Quadrature-weighted mean (solid lines) and standard deviations around the mean (filled shapes) are shown, with the cross-over frequency 2805 Hz (purple line).

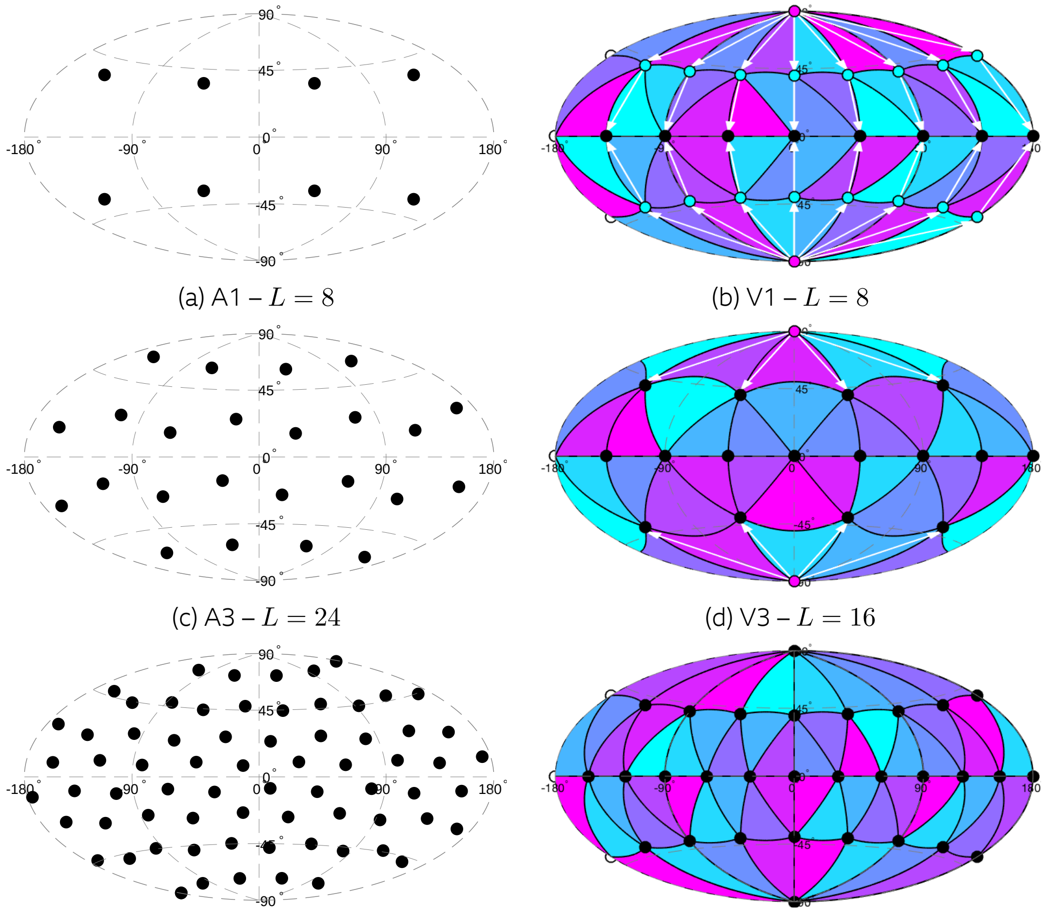
tone filterbank with equivalent rectangular bandwidths (ERBs) (Søndergaard and Majdak, 2013).

Source objects were rendered into a set of intermediate ambisonics signals using conventional ambisonics encoding (Daniel, Nicol, et al., 2003). This would allow the object-based representation to be rendered to ambisonics prior to distribution, reducing the data rate and processing complexity on the rendering device. The head-tracker orientation was used to derive a real-valued spherical harmonics rotation matrix of the appropriate order (Ivanic, 1996), which was applied to the ambisonics signal. On a change of source position or head orientation, the relevant gain vector or rotation matrix was linearly interpolated to the new values over 128 samples. No filter exchange was required during rendering in these approaches. These methods are discussed in more detail in appendix B.4.

8.3.3.3 Channel-Based Binaural using VBAP

Three systems used virtual loudspeaker rendering of a channel-based format, each with a different number and configuration of virtual loudspeakers. Vector base amplitude panning (VBAP) was used to render objects into the channel-based formats (Pulkki, 1997). This flexible amplitude panning technique is capable of rendering sources to arbitrary loudspeaker layouts and is widely used, including in the MPEG-H NGA system (Herre, Hilpert, et al., 2015) and the EBU ADM Renderer specification (EBU Tech 3388, 2018).

The three approaches evaluated were designed to be roughly equivalent in complexity



(a) A1 – $L = 8$

(b) V1 – $L = 8$

(c) A3 – $L = 24$

(d) V3 – $L = 16$

(e) A5 – $L = 70$

(f) V5 – $L = 32$

Figure 8.2: Virtual loudspeaker layouts used. Black circles indicate virtual loudspeakers. For VBAP systems, the Delaunay triangulation is shown and white arrows indicate redistribution of points after the VBAP algorithm: magenta circles represent points redistributed evenly to multiple adjacent points and cyan circles indicate points remapped directly to another point (see main text for further detail).

Name	L	$\theta(°)$		$\phi(°)$	Nearest ITU Layout
		0	[a]	*-90*	
		{0, 45, 90, …, 315}	[b]	*-40*	
V1	8	{0, 45, 90, . . ., 315}		0	I (0+7+0)
		{0, 45, 90, …, 315}	[b]	*40*	
		0	[a]	*90*	
		0	[a]	*-90*	
		{45, 135, 225, 315}		-40	
V3	16	{0, 45, 90, . . ., 315}		0	J (4+7+0)
		{45, 135, 225, 315}		40	
		0	[a]	*90*	
		0		-90	
		{0, 45, 90, . . ., 315}		-40	
V5	32	{0, 30, 60, . . ., 330}		0	H (9+10+3)
		{0, 45, 90, . . ., 315}		40	
		0		90	

Table 8.2: Virtual loudspeaker configurations used for VBAP-based binaural renderers, where L is the number of virtual loudspeakers, θ is the loudspeaker azimuth angle, and ϕ is the loudspeaker elevation angle. Grey text indicates dummy virtual loudspeakers (see main text). The closest system from Rec. ITU-R BS.2051 (ITU-R, 2017a) to each of these systems is also indicated.

to the three ambisonics systems[1], and so are labelled *V1*, *V3* and *V5* herein. Configurations were chosen to relate more closely to real loudspeaker layouts used in programme production than those used in the ambisonics decoding, with rings of speakers at different elevation layers. Table 8.2 gives the loudspeaker configurations used and they are plotted in figures 8.2b, 8.2d and 8.2f. The most similar system to each from Recommendation ITU-R BS.2051 (ITU-R, 2017a) is also indicated. The coordinate system is as defined in section 2.2.2, which matches that of Recommendation ITU-R BS.2076 (ITU-R, 2017b), where for azimuth and elevation angles (θ,ϕ): (0°,0°) is in-front, (0°,90°) is above, and (90°,0°) is left. The similarity to real loudspeaker layouts is seen as an advantage because programmes could be produced in this intermediate representation and then processed for reproduction on common loudspeaker layouts with a simple down-mix, as well as being rendered for headphones by virtualisation. Such an approach is used by Radio France for their spatial audio productions (Nicol et al., 2016; Dejardin, 2018). Regular azimuthal arrangements were used, rather than directly using the ITU layouts, which have higher resolution in the frontal region, where the television screen is located.

Extensions to the VBAP algorithm were made to avoid undesirable behaviours with

[1] Since the ambisonics-to-binaural rendering applies directly to the ambisonics channels, the number of virtual loudspeakers L in the vector base amplitude panning (VBAP) systems should be similar to the number of ambisonics channels $M = (N+1)^2$ for approximately equivalent complexity.

sparse loudspeaker arrangements. Dummy virtual loudspeakers were inserted before the VBAP algorithm, their signal was then redistributed to adjacent virtual speakers before the binaural rendering step, instead of being processed with a binaural filter at that position. Two down-mix approaches were used: equal redistribution to all adjacent speakers and direct remapping to one adjacent speaker channel (indicated by ^{a} and ^{b} respectively in table 8.2). The former approach was used at the poles of the *V1* and *V3* systems, giving even spreading when a source was panned to the target position of those dummy loudspeakers. This smoothed the transitions through the poles, avoiding perceived discontinuities during source movement in these regions. The *V1* system also included rerouted dummy virtual speakers, positioned at the elevation layers used in the other systems (±40°) and routed to the speaker on the horizontal layer at the same azimuth. Redistributed pole speakers were processed prior to the rerouted ones. This approach prevented sources with non-zero elevation angle from being excessively spread over the 8 virtualised speakers in the horizontal plane (when just using redistribution from pole dummy speakers), whilst also preventing a sudden jump in position as a source moved across the poles (when using no dummy speakers). White arrows are shown in figures 8.2b, 8.2d and 8.2f to indicate the redistribution of energy from dummy virtual loudspeakers. Those that are distributed to multiple other virtual loudspeakers are shown by magenta circles and those that are directly routed to a single other virtual loudspeaker are shown by cyan circles. These techniques are similar to those in the EBU ADM Renderer (EBU Tech 3388, 2018).

For these VBAP-based approaches, sources in the object-based representation were panned into the intermediate virtual loudspeaker signals using the modified VBAP algorithm. A change in source position did not require binaural filter updates, instead VBAP panning gains were updated, the relevant gain vector was linearly interpolated to the new target values over 128 samples. With a change in listener head orientation, binaural filters corresponding to the virtual loudspeaker positions were updated in the same way as for sources in the object-based rendering approach. This would allow the object-based representation to be rendered (via modified VBAP) to the virtual loudspeaker layout prior to distribution, reducing the data rate and processing complexity on the rendering device. By contrast, for the head-tracking to be applied to the source positions prior to panning, the object-based scene representation would need to be available on the rendering system.

8.3.4 Control and Playback Engine

A custom software application incorporated the real-time rendering algorithms into a playback engine capable of reading object-based audio scenes from BW64 files with Audio Definition Model (ADM) metadata (ITU-R, 2015c,d). The application presented a control interface, listening on a user datagram protocol (UDP) port for Open Sound Control (OSC) messages. Incoming messages could control playback (e.g. start and stop, and set loop points), select an audio item from a playlist, manually control source positions (overriding the ADM parameters) and control output levels of the multiple running rendering algorithms. A separate software application handled the listening experiment user interface and data collection, sending the required OSC messages to control the rendering engine application appropriately. Head tracking data for both orientation and position were also recorded by the application, potentially allowing later analysis of the assessor's head movements during listening and rating. More details on this software apparatus are given in appendix A.2.

8.4 Experiment Design

A listening experiment was designed to evaluate the perceived quality of the different binaural rendering approaches. This involved multiple stimulus presentation with rating of overall sound quality and indicating relevant characteristics using a CATA approach. A within-subjects factorial design was used to compare systems across multiple audio items using a panel of assessors. A range of audio items was used, both single sources and complex object-based scenes. The methods are described in detail in this section.

The experiment aimed to test the following hypotheses:

H 1 *The rendering approaches will show differences in overall sound quality ratings.*

H 2 *The CATA method can be used to identify different characterisations of the rendering approaches.*

H 3 *The CATA method can be used to gain better understanding of the overall sound quality rating results.*

H 4 *Single sources and complex scenes will lead to use of different attributes to distinguish between systems.*

The experiment consisted of several stages: an initial ITD calibration stage, followed by two main rating sessions (one for each audio item type); these each involved familiarisation and training steps before the formal rating.

8.4.1 Audio Items

Considering the discussion of sound quality formation processes in chapter 3, the evaluation involved two categories of audio item: single sources and complex scenes. It was hypothesised that use of single sources would allow a more controlled low-level analytical assessment of spatial and timbral characteristics of rendering at a specific set of target source positions. On the contrary, evaluation of complex scenes, with many various sound sources, would allow broader but less controlled exploration of the rendering capabilities of the systems, whilst also being more representative of the intended application. The audio items were stored in ADM BW64 files.

8.4.1.1 Single Sources

Three source positions were used and are given here with their identifying name and direction angles in terms of azimuth (θ) and elevation (ϕ): *Frontal* (2°,0°), *Lateral* (−100°,0°), and *Elevated* (30°,30°). A rhythmic acoustic guitar recording[2] was used for all positions, chosen to excite both timbral and spatial characteristics, whilst also being ecologically valid, i.e. representative of entertainment media use and not an abstract signal such as noise. These source positions were selected so as not to coincide with the loudspeaker positions on any of the virtual loudspeaker grids used. However, they were chosen to show a range in performance impact particularly on VBAP systems, where, as a source coincides with a virtual loudspeaker, the rendering becomes identical to object-based binaural. The frontal position was very close to the frontal virtual loudspeaker in all of the VBAP systems, whereas the lateral position resulted in pair-wise panning for each system and the elevated position resulted in triplet-wise panning. The positions were also chosen in regions of importance to 3D spatial audio entertainment, and that show different challenges for non-personalised binaural rendering (Wenzel, Arruda, et al., 1993a) and amplitude panning (Pulkki, 2001b).

8.4.1.2 Complex Scenes

Excerpts from three object-based audio drama scenes were used, taken from the S3A object-based audio drama dataset (Woodcock, Pike, Coleman, et al., 2016). These are spatial audio drama scenes, available in the ADM BW64 format. All scene components are defined as audio objects with 3D position meta-data, many of which are time varying, and sources are placed at directions all around the listener including at elevated positions. They were originally mixed at BBC R&D on a 32-loudspeaker system with two subwoofers, using a

[2] An excerpt from EBU SQAM track 58 (EBU TECH 3253, 2008)

VBAP-based 3D panning system. They were designed as material suitable for demonstration and testing of object-based spatial audio systems. For further details on the creation of these items see co-authored publication Co.P. V (Woodcock, Pike, Melchior, et al., 2016).

The *Family* clip features a domestic interior scene with three family members talking around the listener across a wide frontal stage, from front-left to the right. There is some movement, with accompanying footsteps. There is also another character directly overhead, with the impression of muffled shouting and movement of heavy objects coming through the ceiling. The scene features subtle room ambience as well as quiet exterior road noise and a distant washing machine sound. It also uses a 16-channel 3D reverb with a medium room effect. The clip was 25 s long and contained 30 active object tracks.

The *Forest* clip features a diffuse-sounding forest ambience, comprising a 3D 16-channel field recording with widely spaced cardioid microphones in two height layers, as well as several wildlife sound effects (including birds in the trees). A stereo location recording of a child running and laughing is automated to pan from rear-right to rear-left and then around to front-left. There is also incidental music in the clip, with instruments (flute, harp, strings, mallets and percussion) spread around the scene. The clip was 22 s long and contained 31 active object tracks.

The *Protest* item features a dense crowd scene, set outdoors, with many voices spread around the listener at a range of elevations. There is distinct foreground dialogue as well as background crowd voices. A helicopter flies through the scene towards the end of the item, from rear-left to front-right and over the listener's head. It also features a 16-channel 3D reverb bus, containing mostly sparse reflections (as if from nearby buildings). The clip was 25 s long and contained 53 active object tracks.

The original items from the dataset were checked by a BBC sound engineer with significant experience working with binaural and spatial audio systems. Monitoring was performed using the *OB* renderer and some small adjustments were made to levels and positions of objects.

8.4.2 Inter-aural Time Difference Scaling

Prior to the main experiment, assessors determined an individual ITD scaling factor by the method of adjustment, following Lindau, Estrella, et al. (2010). A low-pass filtered white noise signal (transition-band 1 kHz to 1.5 kHz with −80 dB stop-band attenuation) was convolved with binaural room impulse responses (BRIRs) measured with the Neumann KU100 in the listening room used for the experiment (see appendix A.5.5). Assessors adjusted the scaling of modelled ITDs in steps of 1 % until they were satisfied that the virtual source remained stationary in front of them during head yaw rotations. Ten repetitions were performed, with the initial scaling value taken randomly from a uniform distribution between 0.5 and 1.5. Assessors could switch to a real loudspeaker at the target position for comparison. Prior to the adjustment trials, listeners were shown the behaviour at the two extremes of the start-point distribution and then guided once through the process of adjustment to find a suitable scaling value. The assessors' mean scaling value from the 10 repetitions was used for the main experiment.

8.4.3 Attributes

Attributes were based on those in the SAQI (Lindau, Erbes, et al., 2014), since it was deemed the most relevant available lexicon to the domain of study. A conservative pre-filtering process was applied by the authors to the SAQI scales to remove those attributes that were not deemed relevant to the experiment e.g. *tactile vibration*, *speed* and *sequence of events*.

The SAQI attributes are defined in terms of differential scales, so were adapted to reflect absolute characteristics for the CATA method. For example, the attribute *tone colour bright-dark* with a scale ranging from *darker* to *brighter*, became two CATA attributes *dark* and *bright*. The attribute definitions were modified accordingly. In total, 48 CATA attributes were presented to the assessors, the names of which can be seen in table 8.8. The full list of these attributes is given in appendix C, with the accompanying definitions.

8.4.4 Presentation Method

A multiple-stimulus presentation method was used, allowing assessors to directly compare the rendering of the different systems whilst rating. Assessors were asked to rate overall sound quality and indicate attributes that contributed to it within the same presentation interface, so they could be considered simultaneously. Figure 8.3 shows the user interface developed for the experiment. Assessors could freely switch between stimuli during playback and adjust loop points in the audio item to focus on certain sections for comparison, though they were encouraged to listen to the whole item.

Assessors were asked to rate *overall sound quality*, taking in all aspects, using a 100-point scale with equally spaced descriptive labels (*bad*, *poor*, *fair*, *good*, and *excellent*). They were also asked to check up to five attributes that most contributed to the given quality rating for each system. The link between attributes and the perceived overall sound quality was therefore made the explicit focus of the CATA process. The attribute list reflected the selections for the currently playing stimulus, so the status of check boxes was updated each time the assessors changed between stimuli.

The rating was split into two sessions, according to the audio item category (single source or complex scene). For each session, one item was rated twice to facilitate assessment of assessor expertise. The repeated items were *Elevated* and *Forest*. Assessors were encouraged to take rest breaks whenever they felt it necessary, but at least every 30 minutes.

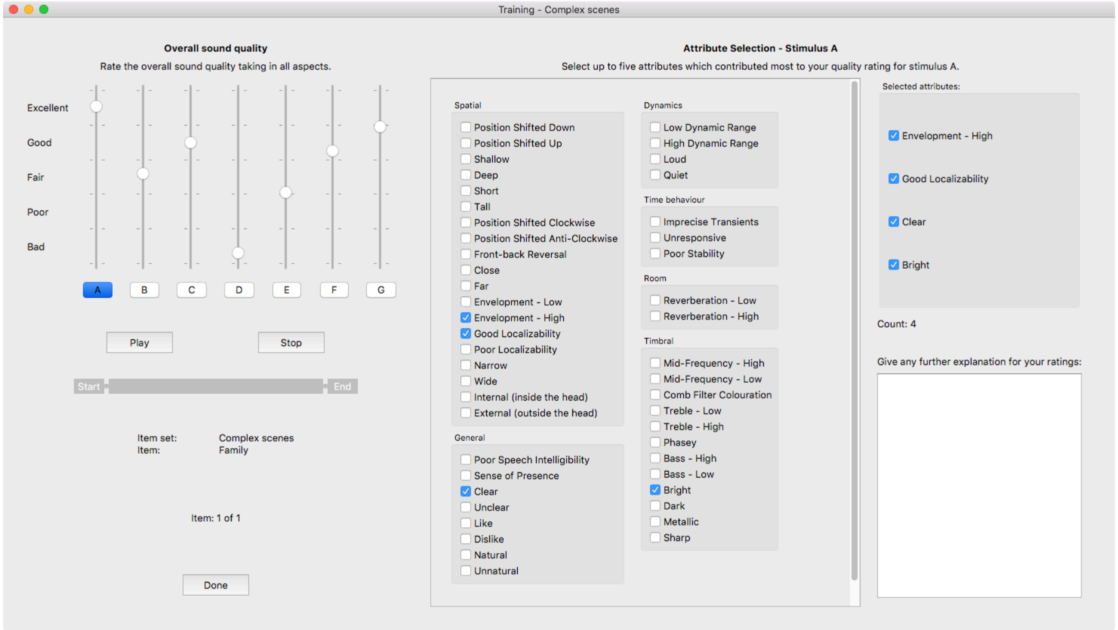
In both sessions there was no reference rendering given, since the ideal rendering approach is not known. This is particularly the case since we are using non-personalised HRTF data, but even personalised rendering has approximations at many points such as HRTF measurement, headphone reproduction and equalisation, source directivity, room effect, tracking accuracy and latency. Providing a reference based on some system that is not known to be ideal may bias the assessors' attention to features of the reference that are not necessarily important to the overall quality of experience.

For complex audio drama scenes, it was hypothesised that the listener expectations will be informed by the narrative context and prior experience with related environments in real life and in media entertainment. No guidance was given to inform the listeners' internal reference. However, due to the more abstract scenario of rendering of a single source, guidance was given to inform the expectations of the assessor. A visual indication of the target source position relative to the listener was given on screen, using projection onto the frontal and horizontal planes. The assessor could also listen to the unprocessed mono source sound, alongside the rendered stimuli. The assessors were told that the ideal system should preserve the timbral character of the source sound, whilst achieving the pictured spatial rendering of the source. As discussed in Rummukainen, Robotham, et al. (2018), the listeners' expectations will likely be influenced by other sensory modalities as well, e.g. visual input and motor actions.

Prior to the experiment, assessors were presented with all stimuli for familiarisation and performed a practice rating to ensure they were comfortable with the process. Within the rating sessions the order of presentation of audio items was randomised, as was the order of systems on the rating page. The order of item sessions was balanced across assessors. Following the discussion in section 8.2.2, the order of the attributes and attribute categories was randomised per assessor, but kept constant across items. Attribute definitions were provided on paper and also as pop-up hints on the user interface when the assessor held the mouse pointer over the attribute name.

8.4.5 Questionnaire

Following each evaluation session, assessors were asked to respond on a Likert scale to three statements, relating to the *ease* and *tediousness* of the evaluation task, as well as


Training - Complex scenes
Overall sound quality
Rate the overall sound quality taking in all aspects.
Excellent
Good
Fair
Poor
Bad
A
B
C
D
E
F
G
Play
Stop
Start
End
Item set:
Complex scenes
Item:
Family
Item: 1 of 1
Done
Attribute Selection - Stimulus A
Select up to five attributes which contributed most to your quality rating for stimulus A.
Selected attributes:
Count: 4
Give any further explanation for your ratings:

Spatial

☐ Position Shifted Down
☐ Position Shifted Up
☐ Shallow
☐ Deep
☐ Short
☐ Tall
☐ Position Shifted Clockwise
☐ Position Shifted Anti-Clockwise
☐ Front-back Reversal
☐ Close
☐ Far
☐ Envelopment - Low
☑ Envelopment - High
☑ Good Localizability
☐ Poor Localizability
☐ Narrow
☐ Wide
☐ Internal (inside the head)
☐ External (outside the head)

General

☐ Poor Speech Intelligibility
☐ Sense of Presence
☑ Clear
☐ Unclear
☐ Like
☐ Dislike
☐ Natural
☐ Unnatural

Dynamics

☐ Low Dynamic Range
☐ High Dynamic Range
☐ Loud
☐ Quiet

Time behaviour

☐ Imprecise Transients
☐ Unresponsive
☐ Poor Stability

Room

☐ Reverberation - Low
☐ Reverberation - High

Timbral

☐ Mid-Frequency - High
☐ Mid-Frequency - Low
☐ Comb Filter Colouration
☐ Treble - Low
☐ Treble - High
☐ Phasey
☐ Bass - High
☐ Bass - Low
☑ Bright
☐ Dark
☐ Metallic
☐ Sharp

Selected attributes:

☑ Envelopment - High
☑ Good Localizability
☑ Clear
☑ Bright

Figure 8.3: User interface for evaluation

the *appropriateness* of the set of attributes. This is based on the approach taken in Ares, Bruzzone, et al. (2014). The statements presented were:

- *It was easy to answer the questions about these samples.*
- *It was tedious to answer the questions about these samples.*
- *The attributes described the quality of the samples well.*

Assessors responded using the following 7-point Likert scale:

1. *Strongly disagree*
2. *Disagree*
3. *Somewhat disagree*
4. *Neither agree nor disagree*
5. *Somewhat agree*
6. *Agree*
7. *Strongly agree*

There were also free-text response fields for any further comments that the assessors had about each session and the experimental design in general. The text in the questionnaire was: *Please give any further comments about the experiment: (i) for the single sources session, (ii) for the complex scenes session, (iii) about the experimental design*.

8.4.6 Pilot Test

Prior to the main experiment, four assessors performed a pilot test. The same stimuli used in the main test were rated in terms of overall sound quality and the assessor selected up to five attributes that could be used to differentiate between the presented stimuli, i.e. an attribute relevant across all stimuli rather than a characteristic of one particular stimulus, as in the conventional CATA method. For this stage, attributes from the SAQI (Lindau, Erbes, et al., 2014) were used directly, since they indicated a scale for differentiating stimuli rather than characteristics of a specific stimulus. Assessors were also given the option to define a new attribute if they felt it necessary for describing a difference. In the pilot, a full replication for rating sessions was used, with each audio item rated twice for all systems. The pilot test also involved the ITD scaling calibration step described in section 8.4.2.

Participants took two sessions of two hours duration to complete this (including short breaks). Training sessions, which involved a practice rating for each item, took approximately 45 minutes, whereas the rating sessions took approximately 30 minutes each.

8.4.6.1 ITD Scaling

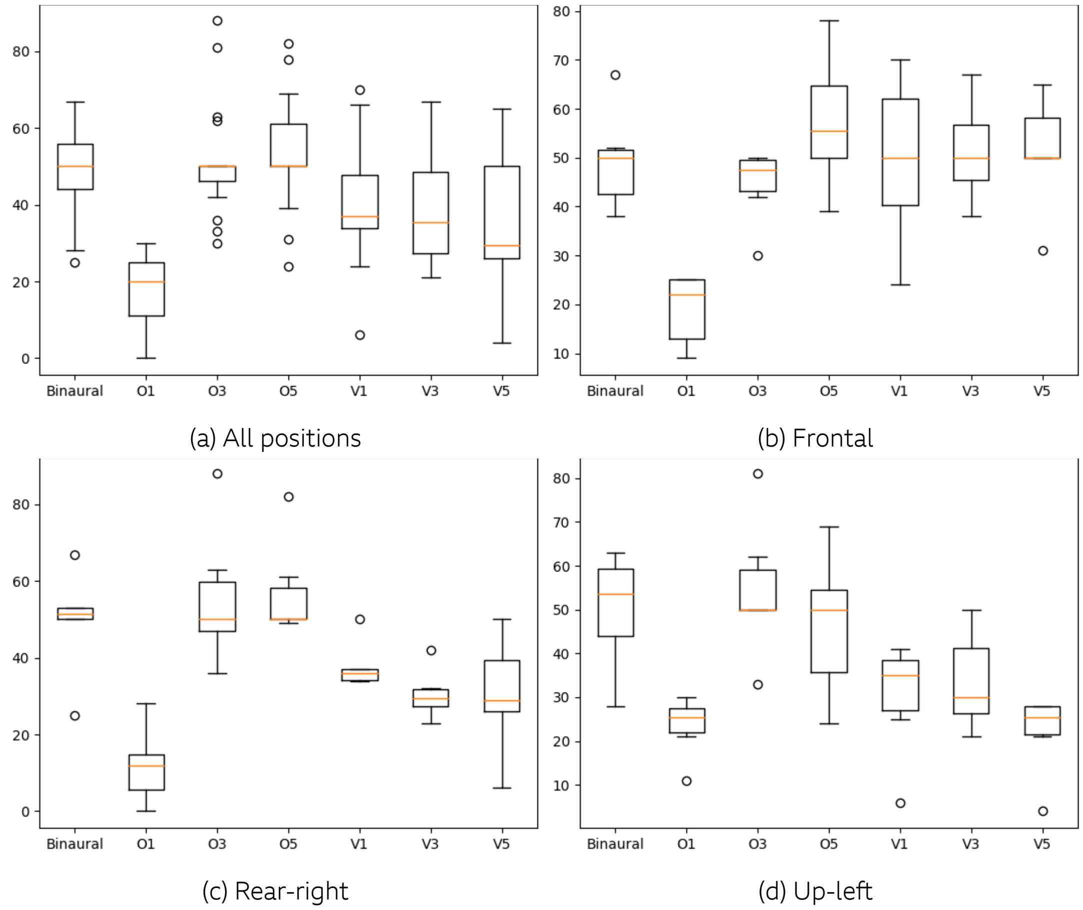
The results of the pilot test ITD scaling are summarised in table 8.3. The range of results for three of the four participants was rather high. It suggests that they were not able to reliably set the ITD scaling value. Assessor P2 reported having had otoplasty several years ago which affected their auditory localisation and could explain the very low ITD scaling value.

Listener ID	μ	σ	min	max
P1	0.98	0.06	0.90	1.10
P2	0.62	0.16	0.32	0.88
P3	0.83	0.12	0.62	1.01
P4	0.82	0.12	0.67	1.04

Table 8.3: Pilot test: ITD scaling results, presenting the mean (μ), standard deviation (σ), minimum and maximum ITD scaling values.

8.4.6.2 Overall Rating Box Plots

Figures 8.4 and 8.5 show box plots of the overall sound quality ratings from the pilot test for single sources and complex scenes, respectively. These show initial trends that were



(a) All positions

(b) Frontal

(c) Rear-right

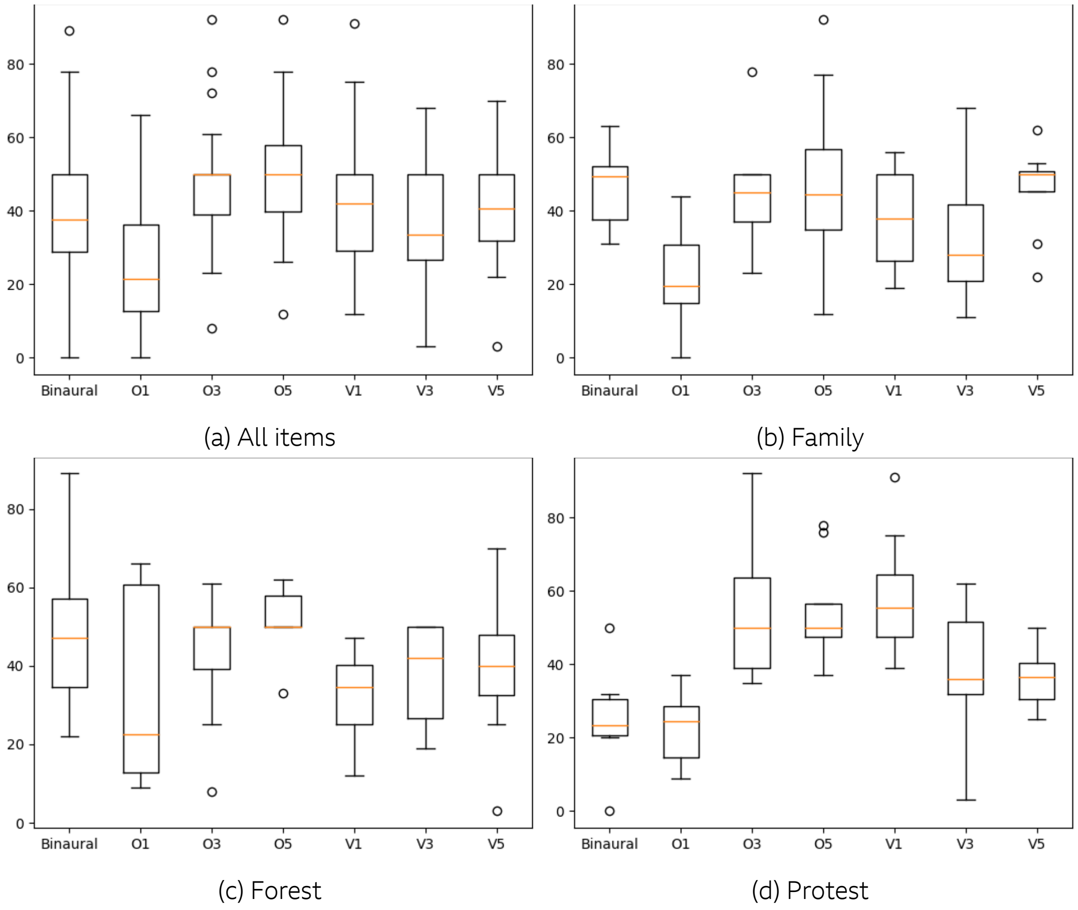
(d) Up-left

Figure 8.4: Pilot test: Box plots showing overall quality ratings for rendering of single sources

largely as expected, with *A1* consistently rated the lowest. For single-sources, the other systems were rated much closer to each other for the frontal position than for the other two positions, where VBAP-based systems were rated lower. *A3* and *A5* were rated similarly to the HRTF-per-source object-based binaural renderer *OB*. For the complex scenes, the differences between systems were smaller or rather the variance is larger. Some patterns were not common between the scenes. *OB* was rated low in *Protest*, which was unexpected.

8.4.6.3 Attribute Selections

Figures 8.6 and 8.7 show the attribute selection frequencies from the pilot tests. Note that each assessor performed six trials per session, so the maximum possible selection count is 24. Between the four assessors, 18 attributes were used in single source rating and 23 for complex scenes. Some attributes clearly occurred more frequently: externalisation, lo-



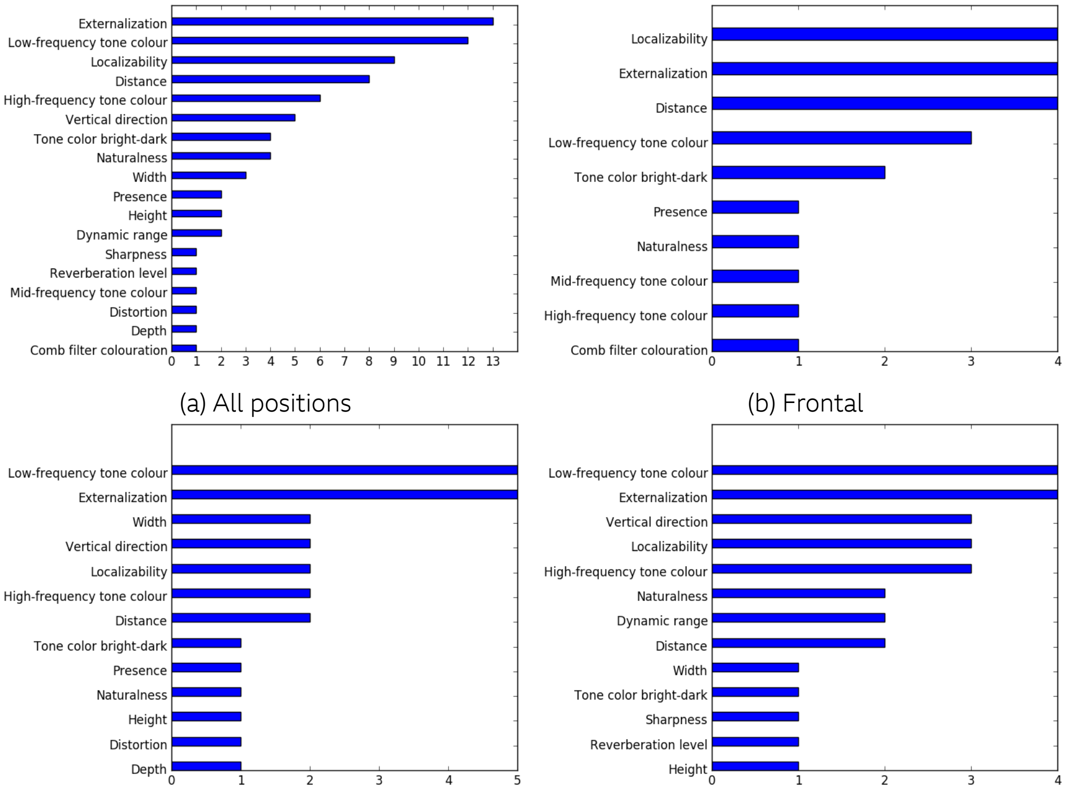
(a) All items

(b) Family

(c) Forest

(d) Protest

Figure 8.5: Pilot test: Box plots showing overall quality ratings for rendering of complex scenes



(a) All positions

(b) Frontal

(c) Rear-right

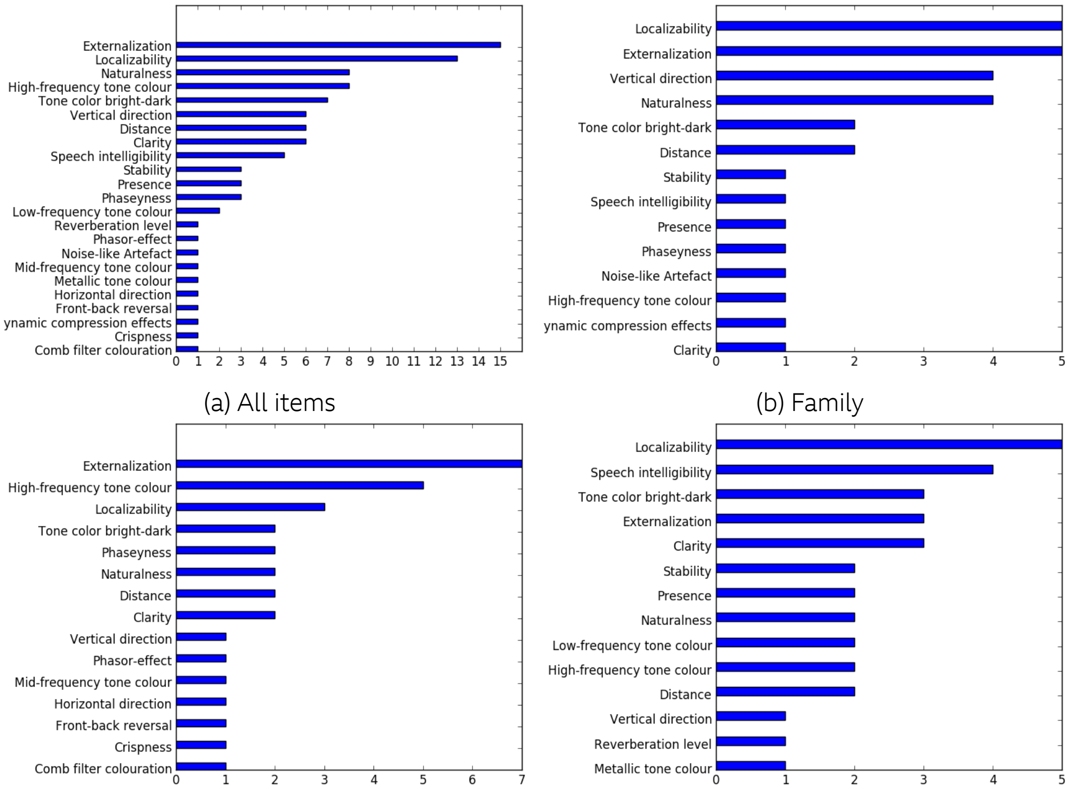
(d) Up-left

Figure 8.6: Attribute selection counts for single source rendering of a Spanish guitar recording

calisability, distance, vertical direction, tone colour (low-, high-frequency, bright-dark), and naturalness. Twice assessors added attributes related to "phasiness" i.e. time varying comb-filtering. There was also an attribute added for tracking stability i.e. spatial stability with head movements, which was mentioned by others in verbal comments. It is worth noting that these attributes could both be captured with the SAQI using its "modifications" concept, since they relate to time-varying comb-filter colouration and spatial characteristics respectively.

8.4.6.4 Pilot Findings

Running the pilot test led to several modifications to the design for the main experiment. In response to the wide ranging ITD scaling values, a more careful training process was adopted for the main experiment. In the pilot test assessors simply performed one practice trial. In the main experiment they were first shown the effect of head movements at the minimum



(a) All items

(b) Family

(c) Forest

(d) Protest

Figure 8.7: Attribute selection counts for rendering of drama scenes

and maximum ends of the ITD adjustment scale. Care was taken to ensure that each assessor felt that they understood the task required and could identify the stable point of the virtual sound source.

The surprisingly low grading of *OB* on *Protest* prompted further inspection. Stimuli were not loudness aligned before the pilot. It was assumed that the energy normalisation steps in the ambisonics and VBAP algorithms would mean there were no major loudness variations. It was apparent though that the *OB* system was noticeably quieter for the *Protest* item and that other loudness variations also existed. Algorithmic measurements showed loudness differences of up to 3.5 LUFS. Therefore the loudness alignment method outlined in section 8.4.7, which takes into account the effect of the audio item and head rotation, was established for the main experiment.

The initial intention was that collection of CATA data across systems would be used also for the main experiment, to reveal a subset of attributes for which more conventional per-stimulus CATA reporting could then be requested in a subsequent session. However, assessors' comments revealed that they were naturally thinking about the attributes which characterised the quality of each individual system during the rating. They then had to re-focus their attention on deciding which attributes were important for differentiating across the set of stimuli, which took some effort. It was therefore decided that CATA responses would be collected per stimulus in the main experiment, across the range of potential attributes, rather than using this pre-selection step.

A wide range of the available attributes were selected in responses, although some were used much more frequently. It was decided to continue to use a broad range in the main experiment to ensure assessors were able to express the perceived characteristics. Two additional attributes *phasiness* and *stability* were incorporated into the main experiment based on assessors' proposals.

The original plan had been to perform the test both with and without head tracking, as in the experiment of chapter 7. It quickly became apparent in setting up the pilot test that this would take too long and so the decision was made to focus only on the head-tracked case. To reduce the duration of the experiment to a more manageable level, it was also decided to use only partial replication of ratings, i.e. to repeat one of the three items for each session, and to reduce the training to just one practice rating. Some modifications were also made to the user interface following the pilot.

8.4.7 Loudness Alignment

Loudness alignment with interactive systems and complex spatially-varying stimuli is challenging, for example the loudness is likely to change dependent on the listener's head orientation. There are several possible approaches. A small panel of listeners could adjust the levels of each renderer to align loudnesses subjectively. Whilst likely to achieve good subjective alignment (at least for those listeners), this process is not precisely repeatable and relies solely on the subjective judgement of those listeners and particularly on their head movement patterns whilst performing the task. Another alternative is to use a repeatable algorithmic alignment of the rendered output when the listener's head is assumed static and front-facing. However, this clearly doesn't account for changes in loudness due to head rotation. To account for this, measurements can be made for a range of head orientations and the average loudness of these for each renderer can be aligned. It should also be checked that the range of values is not too large.

Loudness was aligned algorithmically by rendering all audio items through each renderer at simulated head orientations between [−90°,90°] in 15° steps and applying the objective measure of Recommendation ITU-R BS.1770 (ITU-R, 2011). The mean loudness for each system-item combination across head orientations was used to align to −26 LUFS. System-item combinations had a median loudness range of 1.89 LU across the head orientations, with maximum range of 5.83 LU of any single combination. The alignment was then verified subjectively by the author, confirming that the renderers had the same perceived loudness.

8.5 Analysis and Results

Twenty-two assessors completed the main experiment (15 male, 7 female). All assessors had considerable experience as musicians, audio engineers or audio producers. However, only eleven can be described as *experienced* according to the formal definition in Report ITU-R BS.2300 (ITU-R, 2014b), having also participated in previous listening tests. Five had considerable professional experience using or evaluating spatial audio systems. Data were analysed using the R language and environment (version 3.5.0) (R Core Team, 2018), with use of the SensoMineR (Husson, Lê, and Cadoret, 2017) and FactoMineR (Lê et al., 2008) packages in particular. The significance level α was set at 0.05 throughout the analysis.

The median total test duration was 2 h 27 min 13 s, which included ITD scaling, familiarisation and training, as well as the rating. The maximum total test duration for any one assessor was 6 h 55 min 37 s, which was a large outlier, the next longest was 3 h 40 min 7 s. The minimum total test duration was 1 h 17 min 34 s. The median duration of a rating session was 39 min 35 s, i.e. rating all stimuli on four trial pages for either single sources or complex scenes. A Wilcoxon signed-rank test indicates that the duration was not significantly different between single source or complex scene sessions ($p = 0.203$). The median duration of the ITD scaling stage was 19 min 8 s. These durations include rest breaks taken during a rating session, but not between sessions. Assessors were requested to take a break every 30 min.

8.5.1 Interaural Time Difference Scaling

The data from ITD scaling sessions are represented in table 8.4. Based on Shapiro-Wilk tests of the distributions of ITD scaling values, there is no evidence of major deviations from normality ($p < \alpha$ for only three listeners), so the assumption made in using sample means seems justifiable. The median scaling amongst assessors' mean values was 0.91 (range: 0.38). The highest standard deviation for individual ITD scaling values was 0.41 (range: $[0.25, 1.45]$) and for nine listeners the standard deviation was more than 0.2. All assessors were asked to check the stability of a rendered source during head movement before undertaking the main test, and none reported any issues.

8.5.2 Assessor Post-Screening

Post-hoc analysis was performed to determine assessor expertise within the context of this experiment.

8.5.2.1 Post-Screening of Overall Sound Quality Ratings

Individual analysis of variance (ANOVA) models were applied to the overall sound quality ratings for repeated items to obtain measures of assessor discrimination and reliability. The system and item factors were unfolded into a single stimulus factor. A random permutation test was used to set a threshold of acceptability in responses. For each of 150 permutations, the responses were randomly shuffled per assessor across stimuli, separately for each replicate. Conceptually, this defines a noise floor for assessor reliability and discrimination. This method is described in Report ITU-R BS.2300 (ITU-R, 2014b).

The results are shown in figure 8.8. From this subset of replicated data, analysis suggests that not all assessors can be considered as experts. Ten assessors were below the 95% permutation test level for reliability and nine for discrimination. However, only the four assessors who were well below the thresholds were removed (assessors 0, 2, 8, and 15). Other assessors were close to the thresholds, and it should be considered that this analysis

Table 8.4: ITD scaling data for each assessor, presenting the mean (μ), standard deviation (σ), minimum and maximum ITD scaling values. Shapiro-Wilk test statistics (W) and p-values are also shown, $p < 0.05$ indicates data are not normally distributed (indicated by bold font and a *).

Listener	μ	σ	min	max	W	p
0	0.92	0.37	0.36	1.41	0.914	0.313
1	0.80	0.41	0.25	1.45	0.936	0.509
2	1.08	0.09	0.99	1.23	0.878	0.124
3	0.76	0.24	0.44	1.20	0.956	0.734
4	0.94	0.11	0.77	1.18	0.962	0.808
5	1.03	0.12	0.77	1.18	0.946	0.621
6	0.86	0.11	0.62	0.99	0.895	0.193
7	1.04	0.25	0.69	1.39	0.901	0.223
8	0.78	0.05	0.71	0.85	0.931	0.455
9	0.84	0.16	0.54	1.16	0.960	0.788
10	0.87	0.19	0.62	1.18	0.938	0.530
11	1.04	0.44	0.25	1.50	0.834	0.037^*
12	0.96	0.08	0.87	1.17	0.788	0.010^*
13	0.76	0.26	0.25	1.04	0.718	0.001^*
14	0.85	0.25	0.25	1.15	0.881	0.135
15	0.79	0.27	0.30	1.21	0.962	0.809
16	0.91	0.41	0.30	1.47	0.891	0.174
17	1.03	0.11	0.85	1.17	0.902	0.228
18	0.93	0.04	0.87	0.99	0.966	0.856
19	0.88	0.13	0.66	1.13	0.981	0.971
20	1.05	0.09	0.85	1.17	0.910	0.284
21	1.14	0.19	0.67	1.38	0.870	0.101

was only based on partial replication of the test design (two of six items). The outcomes of subsequent analysis were affected little by removing more assessors. Following this post-screening of assessors, the data analysis is based on the first rating of each item only i.e. replicates are removed.

Figure 8.9 shows the agreement scores, indicating how closely the system ratings for each assessor relate to the ratings across the panel. Those assessors below the 95% permutation test line differ significantly from the panel's system ratings. It can be seen that these assessors are not necessarily those with low reliability and discrimination scores. There is some heterogeneity amongst assessors.

The SensoMineR package in R (Husson, Lê, and Cadoret, 2017) provides two additional methods for assessing the performance of an assessor panel in a quantitative evaluation: `panelperf` and `paneliperf`. Using `panelperf`, an ANOVA of the replicated data with the assessor as a random effect showed significant effects for the assessor and the two-way interaction between assessor and system. This indicates that there wasn't agreement between the assessors in their use of the scale or the quality of a certain system or item. The `paneliperf` method gives measures of assessor agreement, reliability, and discrimination much like the ITU expertise gauge method, though using different models. The four assessors removed previously also had low reliability and discrimination methods using this approach.

8.5.2.2 Post-screening of CATA data

Attribute data were also examined for the replicated stimuli. Repeatability was measured as the percentage of attributes that were checked both times when they were checked in either.

$$R_i = \frac{\sum_{a,s} \mathrm{attr}_{i,a,s,r=1} \& \mathrm{attr}_{i,a,s,r=2}}{\sum_{a,s} \mathrm{attr}_{i,a,s,r=1} \parallel \mathrm{attr}_{i,a,s,r=2}} \tag{8.1}$$

where $\mathrm{attr}_{i,a,s,r}$ is the binary CATA response for a^{th} attribute, by the i^{th} assessor to the s^{th} stimulus for the r^{th} replicate. The results are shown for each listener in figure 8.10.

Conflicts were also inspected, for attributes that related to two ends of a single scale (e.g. *bright* and *dark*). If an assessor checked both attributes for the same stimulus it may indicate that they have not understood the scale, though it could also mean that they simultaneously perceived both attributes within the stimulus. Only one assessor (assessor 2) made conflicts within the same presentation, three times selecting both *inside-the-head* and *outside-the-head*, each time for a single source stimulus. Most conflicts were between the two replicates. Figure 8.11 shows the percentage of attribute scale conflicts between replicates.

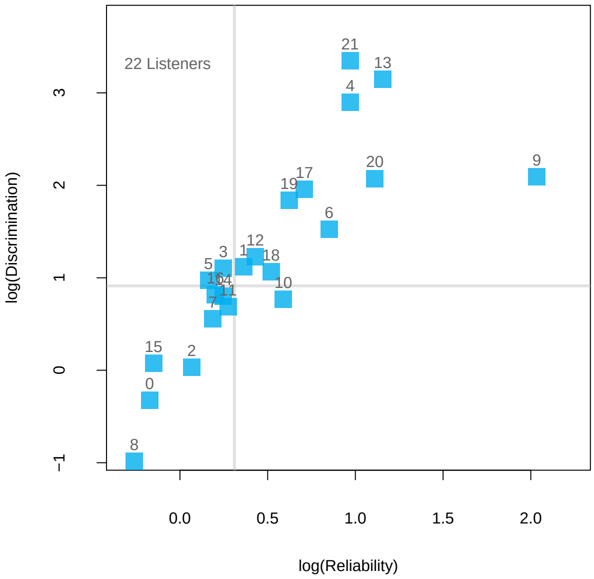


Figure 8.8: Scatter plot of the natural logarithm of discrimination and reliability scores for each assessor, obtained from individual ANOVA models according to the expertise gauge (ITU-R, 2014b). Grey lines represent 95% thresholds derived using a random permutation test.

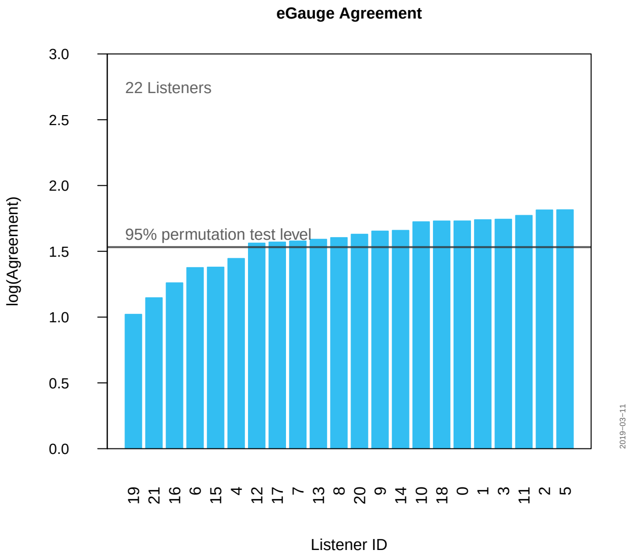


Figure 8.9: Natural logarithm of agreement scores for each assessor, obtained from individual ANOVA models according to the expertise gauge (ITU-R, 2014b). Grey line represents 95% threshold derived using a random permutation test.

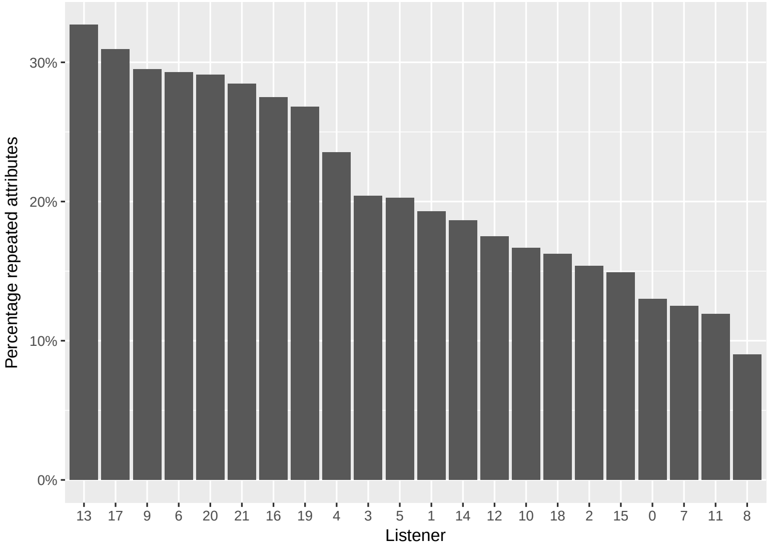


Figure 8.10: Percentage of repeated attribute responses between replicates

The assessors with low repeatability and many conflicts in attribute selections tended to also be those with lower reliability and discrimination scores for overall sound quality. Attribute conflicts show negative correlation with log-discrimination scores from the expertise gauge test on overall sound quality ratings, Spearman's $r = -0.69$. Attribute repeatability is positively correlated with log-reliability scores, $r = 0.74$. The four assessors removed during post-screening of overall sound quality ratings were in the five highest in terms of attribute conflicts and the six lowest in terms of attribute repeatability. This supports removing the CATA data for the post-screened assessors as well as the overall quality ratings.

Note that Campo, Ballester, et al. (2010) proposed an alternative measure of repeatability with CATA responses:

$$R_i = (1/n_s) \sum_s \frac{2 \sum_a \text{attr}_{i,a,s,r=1} \& \text{attr}_{i,a,s,r=2}}{\sum_a \text{attr}_{i,a,s,r=1} + \sum_a \text{attr}_{i,a,s,r=2}} \tag{8.2}$$

where n_s is the number of unique stimuli. This measure is shown in figure 8.12, the ranking of assessors is broadly similar to that of figure 8.10 (values correlated with $r = 0.93$). Campo, Ballester, et al. (2010) suggest a threshold of 20% for screening of assessor grades using this measure. These repeatability measures suggest that CATA data for assessors 7 and 11 should also be considered for removal. However, the decision was made to only remove those that ranked low in the overall sound quality expertise gauge test, since removing additional assessors appeared to reduce the clarity in subsequent analysis.

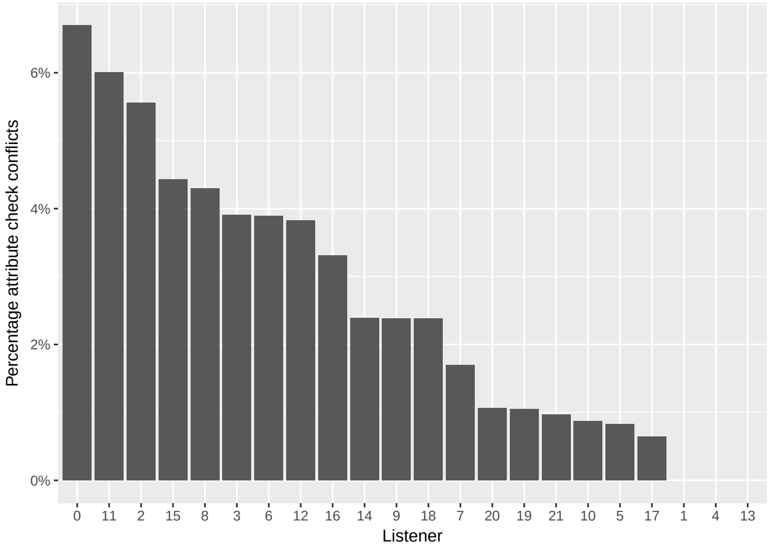


Figure 8.11: Percentage of attribute scale conflicts between replicates for each assessor

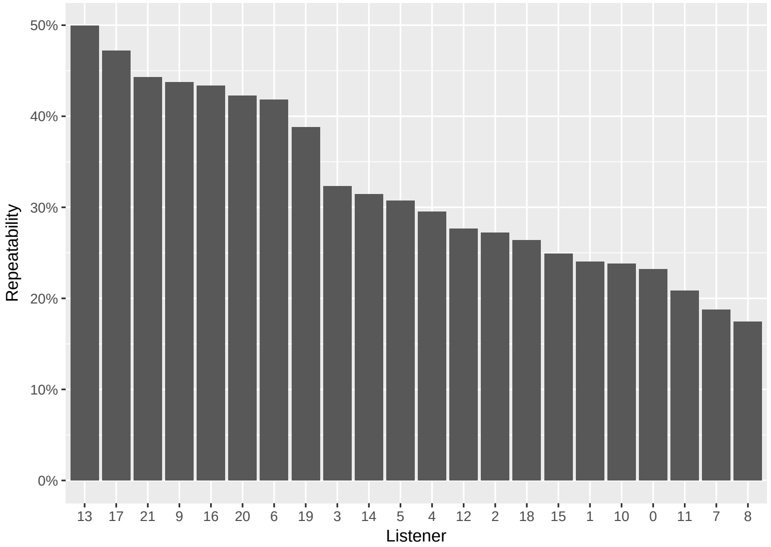


Figure 8.12: Attribute repeatability measure, according to Campo, Ballester, et al. (2010)

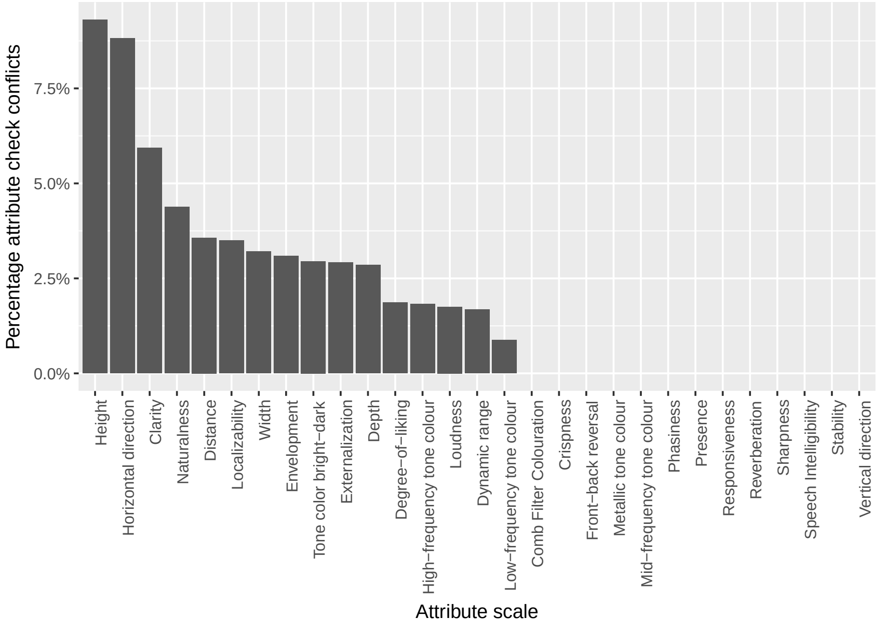


Figure 8.13: Percentage of attribute scale conflicts by attribute scale

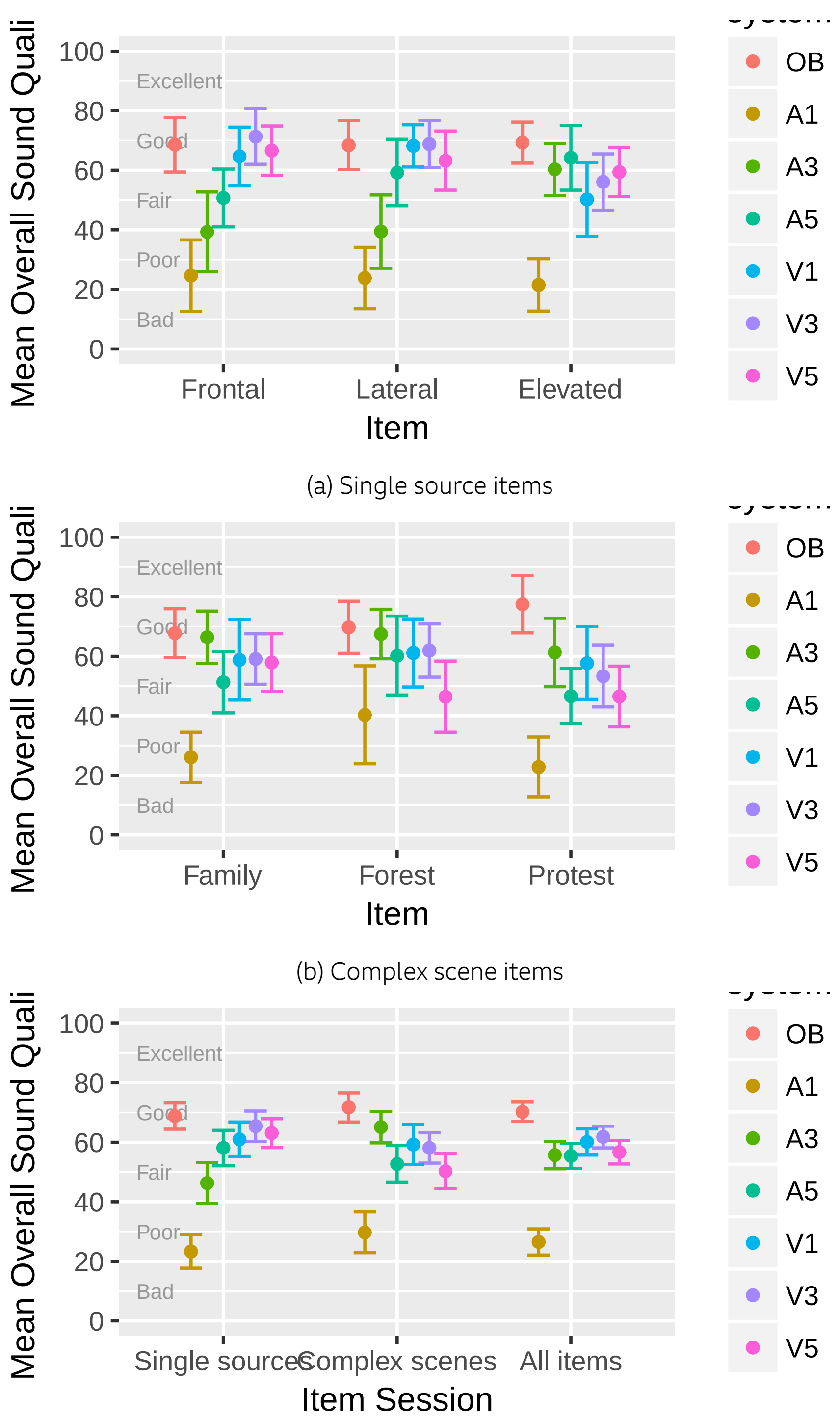
Figure 8.13 shows the conflicts by attribute scale as a percentage of uses of the two attributes relating to that scale, which may be an indication of how well that attribute scale was understood. It should be noted though that it may also be that the assessors heard both characteristics within the stimuli, perhaps focussing their attention on different aspects of the stimulus each time.

8.5.3 Overall Sound Quality

Figure 8.14 shows the mean overall sound quality ratings for each system by individual items, grouped by item session and across all items. No systems were graded as excellent on average. The first-order ambisonics system was graded as poor in most cases, and the other systems were graded as good or fair in almost all cases.

An omnibus test was performed on the overall sound quality ratings to assess whether the experimental variables had significant effects overall. Primarily of interest is the influence of the dynamic binaural rendering method on quality i.e. the *system* effect, though the influence of the content items on the system quality (the *item-system* interaction effect) is also of interest.

Planned contrasts were set to investigate effects amongst systems. The object-based binaural rendering was compared against all virtual loudspeaker methods. Ambisonics



(a) Single source items

(b) Complex scene items

(c) Grouped across sessions and overall

Figure 8.14: Mean overall sound quality ratings for each system. Error bars represent parametric 95% confidence intervals.

methods were compared to VBAP methods. Within each virtualisation approach, the lowest complexity method was compared to the other two i.e. first-order (*A1*) versus higher-order ambisonics methods (*A3*, *A5*), and 2D (*V1*) versus 3D VBAP methods (*V3*, *V5*). Finally the two higher-complexity methods in each approach were compared to each other. These contrasts are illustrated in figure 8.15a.

For the item effect, the single sources and complex scenes were compared. For single sources, horizontal versus elevated items were compared, and then the distinction between frontal and lateral. For complex scenes, the sparser indoor scene (*Family*) was compared to the two denser outdoor scenes (*Forest* and *Protest*), and then these were also compared to one another. These contrasts are illustrated in figure 8.15b.

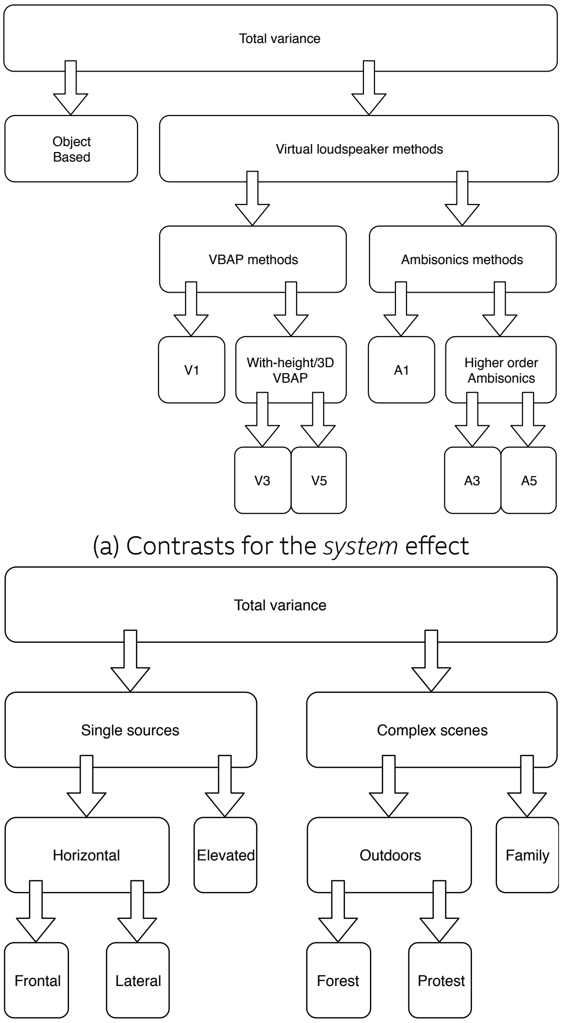
8.5.3.1 Test Assumptions

Shapiro-Wilk tests of the distributions of ratings for each item-system combination suggest a small number (4 of 42) are not normally distributed. Levene's test at the item-system level showed $F = 1.39, p = 0.057$, which casts doubt on the homogeneity of variance. This implies that ANOVA test assumptions may be invalid. It has also been discussed that the five-label ITU-R quality scale may not be considered as an interval scale by assessors and also that observations cannot be considered independent since they are gathered simultaneously (Mendonça and Delikaris-Manias, 2018). Independence of observations (within-assessor for repeated measures) is an assumption of an ANOVA performed with a general linear model and violation of this assumption can lead to inflated type-I errors, i.e. suggesting an effect is significant when it is not. Although these issues exist, ANOVA is widely used. It often provides valid conclusions despite invalid assumptions, as discussed by Schoeffler, Silzle, et al. (2017). A Friedman test presents a non-parametric alternative to the one-way repeated-measures analysis of variance (RM-ANOVA) and versions of ANOVA with trimmed-means are also available. However, equivalents to multi-way tests to allow for exploring multiple factors and their interactions are not commonly available in statistics packages.

8.5.3.2 Multilevel Modelling

Multilevel models are introduced in section 4.6.2.1. They are hierarchical linear regression models that represent hierarchical structure in the data, explaining sources of variance at each of the different levels. Multilevel models have some advantages over RM-ANOVA for analysis of within-subjects experiments (Quen and Bergh, 2004; Hoffman and Rovine, 2007).

The overall sound quality ratings were analysed with a multilevel model, fitted using maximum likelihood estimation. The `nlme` R package was used (Pinheiro et al., 2018). The



(a) Contrasts for the *system* effect

(b) Contrasts for the *item* effect

Figure 8.15: Planned contrasts in analysis of overall sound quality ratings.

Fixed Effects Model		df	AIC	BIC	logLik	Test	χ^2_{test}	p-value
1	Intercept only	5	6827.676	6850.816	-3408.838			
2	+***System***	11	6714.962	6765.870	-3346.481	1 vs 2	124.714	<.001***
3	+*Item*	16	6718.634	6792.682	-3343.317	2 vs 3	6.328	0.276
4	+***System*:*Item***	46	6669.473	6882.363	-3288.737	3 vs 4	109.161	<.001***

Table 8.5: Likelihood ratio tests for multi-level linear model fitting of overall quality rating data

nested random effects model was: ~1 | *assessor* / *system* / *item*. Table 8.5 shows the results of introducing successive fixed-effect model components, with bold font indicating significant effects and asterisks indicating significance level ($p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$). In order to test whether experimental variables have a significant effect, models with and without that variable as a predictor are compared using a χ^2 likelihood ratio test, i.e. the change in log-likelihood (logLik) between models and the associated change in degrees of freedom (df) are tested against the corresponding χ^2 distribution. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also shown for each model. The type of binaural rendering approach (*system*) had a significant effect on overall sound quality, whilst the audio *item* did not. The *system:item* interaction was significant, indicating that the audio item significantly influenced the quality scores for systems.

The planned contrasts (figure 8.15) were used to investigate the variance within the model effects, giving further insights into the results. For each contrast the model parameter b, the t-value with associated degrees of freedom (df), and probability value p are reported. We also present the effect size r, which is calculated from t using the equation (Fritz et al., 2012):

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{8.3}$$

For point biserial r, values of 0.5, 0.3, and 0.1 have been said to represent large, medium and small effect sizes respectively (Cohen, 1988).

The model estimates of fixed-effects parameters are shown in table 8.6. Object-based rendering (*OB*) showed significantly better quality than the virtual loudspeaker techniques (*OB* vs virt.). Overall the ambisonics techniques gave lower quality than VBAP techniques (ambi. vs VBAP), with *A1* significantly worse than higher-order ambisonics (*A1* vs HOA). In terms of system-item interactions, the relative quality of ambisonics against VBAP systems was lower on single items than on full-scenes (Ambi. vs VBAP : Source vs scene). The difference in quality between *A3* and *A5* ambisonics was strongly affected by the use of single sources or complex scenes, with *A3* having higher quality relatively when using complex scenes (*A5* vs *A3* : Source vs scene). The relative quality of ambisonics against VBAP systems was better for the elevated single source than the horizontal sources (Ambi. vs VBAP : Elevated vs horizontal). Interestingly, the horizontal VBAP system (*V1*) was not significantly lower in quality than the 3D VBAP systems (*V1* vs 3D VBAP), including for the distinction between elevated and horizontal single source items (*V1* vs 3D VBAP : *Elevated* vs horizontal).

8.5.3.3 Repeated Measures ANOVA

RM-ANOVA was also applied to the overall sound quality ratings, for comparison to other studies, since it is a more commonly used tool. A generalised linear model was fitted to the data, modelling the system and item effects and their interaction, with a random assessor effect. The generalised eta-squared effect sizes may be interpreted as follows, an η_g^2 of .02 is considered as small, one of .13 as medium, and one of .26 as large (Fritz et al., 2012). The results of the RM-ANOVA are shown in table 8.7. The system had a large and significant effect on overall quality and there was also a smaller significant system-item interaction effect. The item effect was not significant. These results correspond well to those of the multi-level model in section 8.5.3.2.

8.5.3.4 Pairwise Comparisons

Post-hoc pairwise comparisons were performed using paired-samples t-tests between all system-item interaction levels, using Benjamini-Hochberg false discovery rate correction (Benjamini and Hochberg, 1995). Additionally, pairwise tests were carried out between systems for ratings across each item group and across all items. In this case the more conservative Bonferroni correction was used, in acknowledgement of the inflation of family-wise error rate by performing these additional tests. Point biserial correlation was used to estimate effect sizes, following equation 8.3.

When results are grouped across all items and across the item groups, *A1* had significantly different quality ratings to all other systems, showing a large effect in all cases ($r > 0.5$). It can be seen in figure 8.14c that this system was graded in the region corresponding to *poor* quality in each case, much lower than the other systems. Grouped across all items, the object-based binaural system had significantly different quality ratings to all other systems. This effect was large for each of the ambisonics systems and the V5 system ($r \geq 0.48$) and medium-sized for the V1 and V3 systems ($r = 0.37$ in both cases). There were no other significant differences between systems when grades were grouped over all items.

	b	σ_μ	df	t	p	r
(Intercept)	55.19	2.05	595	26.89	<.001***	0.74
OB **vs virt.**	2.50	0.35	102	7.23	<.001***	0.58
Ambi. vs VBAP	-6.82	0.92	102	-7.45	<.001***	0.59
A1 **vs HOA**	-9.67	0.92	102	-10.55	<.001***	0.72
A5 vs *A3*	-0.16	1.59	102	-0.10	0.921	0.01
V1 vs 3D VBAP	0.29	0.92	102	0.32	0.750	0.03
V3 vs *V5*	2.54	1.59	102	1.60	0.113	0.16
Source vs scene	-0.06	0.66	595	-0.09	0.931	0.00
Elevated vs horizontal	-0.36	0.66	595	-0.54	0.587	0.02
Frontal vs *Lateral*	-0.37	1.14	595	-0.32	0.746	0.01
Interior vs outdoor	0.04	0.66	595	0.06	0.955	0.00
Forest **vs** ***Protest***	2.96	1.14	595	2.59	0.010*	0.11
OB vs virt. : Source vs scene	-0.23	0.27	595	-0.87	0.387	0.04
Ambi. vs VBAP : Source vs scene	-3.48	0.71	595	-4.88	<.001***	0.20
A1 vs HOA : Source vs scene	0.04	0.71	595	0.06	0.952	0.00
A5 **vs** ***A3*** **: Source vs scene**	6.02	1.23	595	4.88	<.001***	0.20
V1 vs 3D VBAP : Source vs scene	-1.37	0.71	595	-1.92	0.055	0.08
V3 vs *V5* : Source vs scene	-1.39	1.23	595	-1.13	0.260	0.05
OB vs virt. : *Elevated* vs horizontal	0.10	0.27	595	0.38	0.702	0.02
Ambi. vs VBAP : ***Elevated*** **vs horizontal**	3.51	0.71	595	4.93	<.001***	0.20
A1 **vs HOA :** ***Elevated*** **vs horizontal**	-1.98	0.71	595	-2.78	0.006**	0.11
A5 vs *A3* : *Elevated* vs horizontal	-1.94	1.23	595	-1.58	0.115	0.07
V1 vs 3D VBAP : *Elevated* vs horizontal	-0.73	0.71	595	-1.02	0.307	0.04
V3 vs *V5* : *Elevated* vs horizontal	-1.42	1.23	595	-1.15	0.249	0.05
OB vs virt. : *Frontal* vs *Lateral*	0.07	0.47	595	0.15	0.879	0.01
Ambi. vs VBAP : *Frontal* vs *Lateral*	-0.86	1.23	595	-0.69	0.487	0.03
A1 vs HOA : *Frontal* vs *Lateral*	0.84	1.23	595	0.68	0.494	0.03
A5 vs *A3* : *Frontal* vs *Lateral*	-2.11	2.13	595	-0.99	0.323	0.04
V1 vs 3D VBAP : *Frontal* vs *Lateral*	-1.07	1.23	595	-0.87	0.386	0.04
V3 vs *V5* : *Frontal* vs *Lateral*	-0.21	2.13	595	-0.10	0.922	0.00
OB vs virt. : Interior vs outdoor	-0.33	0.27	595	-1.23	0.220	0.05
Ambi. vs VBAP : Interior vs outdoor	-1.00	0.71	595	-1.40	0.162	0.06
A1 vs HOA : Interior vs outdoor	-0.61	0.71	595	-0.85	0.394	0.04
A5 vs *A3* : Interior vs outdoor	-0.69	1.23	595	-0.56	0.576	0.02
V1 vs 3D VBAP : Interior vs outdoor	-0.78	0.71	595	-1.10	0.271	0.05
V3 vs *V5* : Interior vs outdoor	-1.66	1.23	595	-1.34	0.179	0.06
OB **vs virt. :** ***Forest*** **vs** ***Protest***	-1.14	0.47	595	-2.45	0.015*	0.10
Ambi. vs VBAP : *Forest* vs *Protest*	2.12	1.23	595	1.72	0.086	0.07
A1 vs HOA : *Forest* vs *Protest*	1.26	1.23	595	1.03	0.306	0.04
A5 vs *A3* : *Forest* vs *Protest*	1.85	2.13	595	0.87	0.387	0.04
V1 vs 3D VBAP : *Forest* vs *Protest*	-0.16	1.23	595	-0.13	0.898	0.01
V3 vs *V5* : *Forest* vs *Protest*	2.17	2.13	595	1.02	0.311	0.04

Table 8.6: Multi-level model estimates of fixed-effects parameters, using planned contrasts for *system* and *item*. Model parameter estimates b and given with standard error σ_μ and associated t-test results are given with probability p and effect size r. Significant effects are shown in bold.

Effect	DF$_n$	DF$_d$	SS$_n$	SS$_d$	F	p	η_g^2
System	6	102	120638.13	55510.25	36.945	<.001***	0.325
Item	5	85	2343.35	48173.25	0.827	0.534	0.009
System* : *Item	30	510	36929.78	147028.13	4.270	<.001***	0.128

Table 8.7: Repeated measures ANOVA of overall sound quality ratings.

For complex scenes, *OB* was significantly different to all other systems, except *A3*. For those with significant differences, the effects were large, except for the V1 system where a medium effect was observed ($p = 0.047, r = 0.40$). The *A3* system was graded significantly differently to *A5* ($p = 0.009, r = 0.46$) and to the *V5* system ($p = 0.001, r = 0.52$). There were no other significant differences between systems when grades were grouped over complex scene items.

For single sources, *OB* was rated significantly different to each of the ambisonics systems, with a large effect for *A1* ($p < 0.001, r = 0.90$) and *A3* ($p < 0.001, r = 0.68$) and a medium-sized effect for *A5* ($p = 0.044, r = 0.41$). However, it was not significantly different from any of the VBAP systems. *A3* was graded significantly differently to each of the three VBAP systems (r was 0.44, 0.53, and 0.49 for *V1*, *V3* and *V5* respectively). There were no other significant differences between systems when grades were grouped over single source items.

It can be seen in figure 8.14b that *A1* was rated higher on *Forest* than it was for other items, where it was not significantly different from *A5*, *V3*, and *V5*. *OB* is rated particularly highly on *Protest*, where it is significantly better than all but *A3*. On other complex scenes it is only rated significantly higher than *A1* (and *V5* for *Forest*). *V5* was rated significantly lower than the other VBAP systems on *Forest*.

V1 and *V3* were rated significantly worse than *OB* on *Elevated*, but not for the two horizontal positions. Similarly, the pairwise tests within the system but across items, show significant differences between the *Elevated* position and each of the two horizontal positions for both *V1* and *V3*. These results can be seen in figure 8.14a.

8.5.4 CATA Data

The mean number of attributes checked per stimulus ranged amongst assessors, from 1.83 to 4.98, as shown in figure 8.16. Overall, assessors used a median of 29.5 of the 48 CATA attributes, with the lowest number of attributes used by an assessor being 18 and the maximum being 43. Viewed by system (figure 8.17), assessors used significantly more attributes to describe *A1* than for other systems, with a mean of 3.97 attributes per stimulus. The mean number of attributes checked per stimulus, averaged across all assessors and systems, was

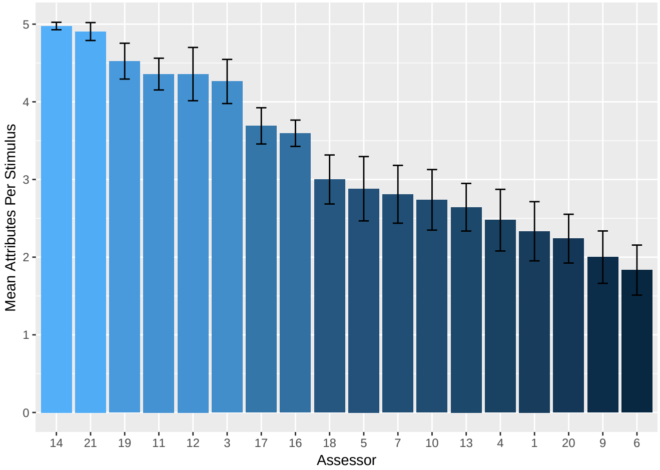


Figure 8.16: Mean attribute selection frequency per stimulus for each assessor. Error bars show 95 % confidence intervals.

3.31. Over all stimuli and assessors, between 42 and 46 attributes were used to describe the systems, with a median value of 45. Figure 8.18 shows the attribute selection frequencies across all stimuli and assessors. Over all stimuli, the most frequently selected attributes were *Good Localizability*, *Natural*, *Clear*, *Dark*, *External* and *Close*. The least frequently selected attribute was *Front-Back Reversal*, which was used only once.

8.5.4.1 Significance Tests

Cochran's Q test is a non-parametric test to check whether treatments have identical effects in a two-way randomised block design where there is a binary response variable. It is often used to check whether the panel of assessors revealed significant differences between systems using each of the CATA attributes. Under the null hypothesis, Cochran's Q test statistic is asymptotically χ^2-distributed. The test does not make the assumption of independent observations of Pearson's χ^2 test. The effective sample size is the number of systems multiplied by the number of assessors who showed some rating differences between systems (i.e. neither all systems checked nor none checked). Tate and Brown (1970) suggest that under the χ^2 approximation, the required effective sample size is ≥ 24.

An omnibus test was used to assess whether the attributes differentiated systems overall, before examining the significance of each attribute in turn. An omnibus test based on the χ^2-distribution would assume independence between attributes, which is clearly not the case. To avoid this assumption, a randomisation test was used (Meyners, Castura, and Carr, 2013). CATA data were randomly permuted between systems but within assessor, item, and

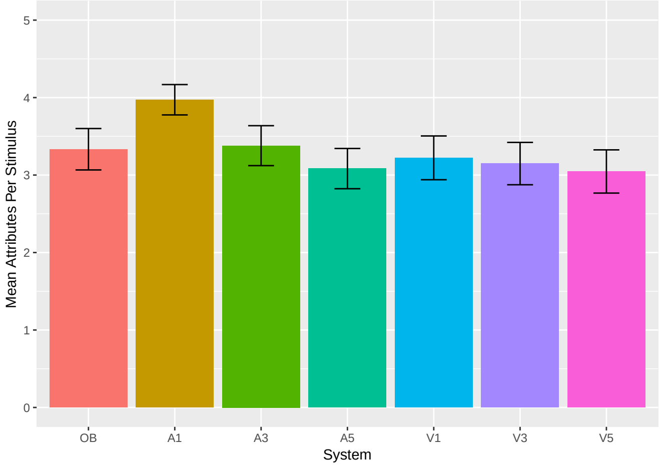


Figure 8.17: Mean attribute selection frequency per stimulus for each system. Error bars show 95 % confidence intervals.

attribute. For each of 1000 randomisations, Cochran's Q test statistic was compared to that for the original data. The p-value was calculated as the proportion of randomisations where Q was greater than for the original data. The omnibus test was based on the sum of Q test statistics for each attribute. For each item group (single sources and complex scenes) and across all items, the system-item combinations were represented separately. Tests at this level therefore assessed the significance of attributes over the system and item factors and their interaction. Meyners, Castura, and Carr (2013) suggest that the randomisation test is also valid in cases when the effective sample size is too low for assumption of the χ^2-distribution; they give a minimum effective sample size of 24. Only two significant cases had effective sample sizes that fell below this threshold; these were: *Elevated–Position Shifted Anti-Clockwise* and *Family–Phasey*. In both cases the effective sample size was 21.

The results, shown in table 8.8 in the "Overall" row, suggest that the systems are described differently by the set of CATA attributes for every item and across item groups. Across all system-item combinations ("All" column), 28 of the 48 attributes were significant. This is also indicated alongside the attribute names in figure 8.19. For individual content items the number of significant attributes ranges from 7 (*Forest*) to 16 (*Lateral* and *Family*). 20 of the attributes were not significant for any item. These were: *Mid-Frequency Low*, *Sharp*, *Position Shifted Clockwise*, *Position Shifted Up*, *Position Shifted Down*, *Close*, *Front-back Reversal*, *Deep*, *Shallow*, *Tall*, *Short*, *Envelopment–High*, *Reverberation–Low*, *Imprecise Transients*, *Unresponsive*, *Poor Stability*, *Loud*, *Quiet* and *Low Dynamic Range*.

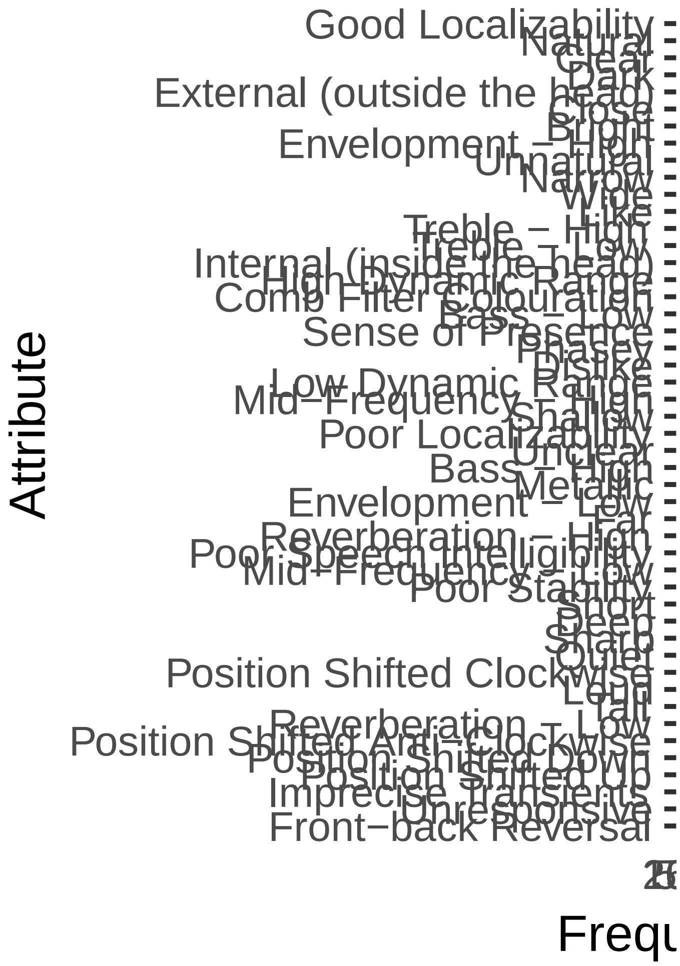


Figure 8.18: Attribute selection frequency across all stimuli and assessors

	Frontal	Lateral	Elevated	Family	Forest	Protest	Single	Scenes	All
Dark	**	**	**	**	**	*	**	**	**
Bright	*	**	**	**	**		**	**	**
Treble–High			**	**	**	**	**	**	**
Treble–Low	*	.	.				**	*	**
Mid-Freq.–High		**	.				**		*
Mid-Freq.–Low									
Bass–High	na	**	.				**	*	**
Bass–Low		*	**			**	**	**	**
Sharp									
Comb Filter Col.	**	*	**	**			**	**	**
Phasey	.			*				*	*
Metallic		**	**	**		*	**	**	**
Pos. Shift. ACW			*		na				.
Pos. Shift. CW				.					
Pos. Shift. Up	na				na	na			
Pos. Shift. Down									
Close									
Far		**					*		
Front-back Rev.	na	na	na	na		na	na		
Wide				**	*			**	**
Narrow				*					
Deep		na	na						*
Shallow									
Tall	na	na							
Short		na							
Internal		**				.	**	*	**
External	.	**		*			*	.	*
Good Localiz.	**	*	*			**	**	*	**
Poor Localiz.		**				*	**	.	**
Envel.–High						.		*	**
Envel.–Low					*				*
Reverb.–High								**	.
Reverb.–Low			na						*
Imprecise Trans.				na		na			
Unresponsive	na	na				na			
Poor Stability									
Loud									*
Quiet		na							
High Dyn. Range						*		.	.
Low Dyn. Range					.			.	**
Clear	**		**	**		**	**	**	**
Unclear						*		**	**
Natural	**	**	.	**		*	**	*	**
Unnatural	**	**	**	**		**	**	**	**
Sense of Presence						*		*	*
Poor Speech Int.	na	na	na	.	na	**	na	**	**
Like			*	*			*	*	**
Dislike	**	*	.	**	*		**	**	**
Overall	**	**	**	**	**	**	**	**	**

Table 8.8: Results of randomisation test based on Cochran's Q for all attributes and overall, and for each item, item group and over all items. ** indicates significant differences at p <0.01, * indicates significant differences at p <0.05, . indicates p <0.1, an empty cell indicates p >0.1 and na indicates zero selections in that case.

8.5.4.2 System-Item Interaction for Attributes

It is useful to test whether the attributes are used to describe the systems differently between items. Omnibus tests can be performed, treating each attribute as a single dependent variable, in order to test the interaction between the system and item factors. For a continuous dependent variable, a RM-ANOVA could be used, but CATA attribute responses are binary. Jaeger (2008) describes issues with applying ANOVA to categorical data. The assumption of homogeneity of variances does not hold if conditions have different mean values (i.e. different proportions of stimuli are checked for the attribute). Confidence intervals for percentages can also extend to proportions above 1 or below 0, which are not interpretable. The correct approach is to fit a generalized linear model with a binomial distribution function and a logit link function, also known as a logit model, which is fitted to the data using logistic regression. A mixed logit model allows for modelling the random assessor effect in a repeated measures design, using mixed-effects logistic regression.

The logit transform is defined as:

$$g(p) = \text{ln}(\frac{p}{1-p}) \tag{8.4}$$

where p is the proportion of checked stimuli for the attribute and the logit $g(p)$ represents the log-odds of a checked attribute.

The `lme4` package in R provides mixed-effects logistic regression via the generalized linear mixed-effects model function `glmer` with the `binomial` family option (uses the binomial distribution function and the logit link function). Attempts were made to perform such logistic regression on the CATA data, e.g. using the equation:

$$\text{Dark} \sim 1+\text{system}+\text{item}+\text{system:item}+(1+\text{system}+\text{item}+\text{system:item} \mid \text{assessor}) \tag{8.5}$$

However, most of the models failed to converge, so the parameters cannot be reliably interpreted.

Ordinary logistic regression does not account for repeated measures, i.e. the lack of independence of observations, but it may be used to give an indication of interaction effects. It appears that this approach was used by Hicks et al. (2018). Table 8.9 shows the results of analysis using ordinary logistic regression. Successive models were fitted to the data (see equations (8.6) to (8.8), for example with *Dark*) and the significance of each effect was tested using a likelihood ratio test.

$$\text{Dark} \sim 1 + \text{system} \tag{8.6}$$

$$\text{Dark} \sim 1 + \text{system} + \text{item} \tag{8.7}$$

$$\text{Dark} \sim 1 + \text{system} + \text{item} + \text{system:item} \tag{8.8}$$

It appears from the results in table 8.9 that the use of some attributes varied according to audio *item* and in some cases there was a *system*-*item* interaction, e.g. *Natural* and *Clear*. Such interaction implies that the systems that showed *Clear* characteristics varied according to audio item. Though the limitations of the model in not acknowledging the dependence of repeated measures from assessors should be acknowledged when interpreting these results.

	system	item	system:item
Dark	**	**	.
Bright	**	**	
Treble - High	**		*
Treble - Low	**		.
Mid-Frequency - High	**		
Mid-Frequency - Low		**	
Bass - High	**	**	
Bass - Low	**	**	
Sharp			
Comb Filter Colouration	**		
Phasey	*	*	
Metallic	**		
Position Shifted Anti-Clockwise			
Position Shifted Clockwise			
Position Shifted Up		.	
Position Shifted Down			
Close			
Far			.
Front-back Reversal			na
Wide	**		.
Narrow	*		
Deep		**	
Shallow		*	
Tall		*	
Short		**	
Internal (inside the head)	**	*	*
External (outside the head)	.		.
Good Localizability	**	**	.
Poor Localizability	**		*
Envelopment - High	*	**	
Envelopment - Low	.	*	
Reverberation - High		*	.
Reverberation - Low		.	
Imprecise Transients			
Unresponsive			
Poor Stability			
Loud	**	.	
Quiet		**	
High Dynamic Range	*	.	
Low Dynamic Range	.	**	.
Clear	**		**
Unclear	**	**	
Natural	**		**
Unnatural	**	**	
Sense of Presence	*	*	
Poor Speech Intelligibility	**	**	
Like	**		
Dislike	**		

Table 8.9: Analysis of variance of ordinary logistic regression models, using maximum likelihood estimation to test significance of predictor variables *system* and *item* and their interaction. ** indicates significant differences at $p < 0.01$, * indicates significant differences at $p < 0.05$, . indicates $p < 0.1$, an empty cell indicates $p > 0.1$ and `na` indicates insufficient data.

8.5.4.3 Analysing Contingency Tables

Contingency tables show the frequency of selection of each attribute for each system by the assessors. Frequency data can be visualised directly, using methods such as bar plots, matrices, and word clouds, see e.g. figures 8.19 to 8.21. Figure 8.19 presents the contingency table with data from all items as a colour matrix. The column corresponding to system *A1* is distinctly different to other systems. The most frequently and rarely used attributes are indicated by light and dark shading, respectively. Careful inspection allows relative frequency of each attribute amongst systems to be determined. Bar plots for individual attributes can be used to show these patterns more clearly, as in figure 8.20. Two of the most frequently used opposing pairs of attributes are shown (*Dark/Bright* and *Good/Poor Localizability*). The attributes *Sense of Presence* and *Poor Speech Intelligibility* are also presented, these were less frequently used but the selection frequency varies markedly across systems in both cases. Figure 8.21 shows word clouds for two example systems, where the size of the attribute names is based on their selection frequency across all audio items. This reveals the most frequently selected attributes for each system from amongst the 48 attributes available.

These methods can highlight the most frequently selected attributes for systems, and some patterns can be interpreted between systems for a given attribute or pair of attributes. However, with many attributes and systems, the process soon becomes unmanageable. It can be difficult to observe deviations from the average frequency of selection for a given attribute or system for all but the most obvious of cases.

8.5.4.4 Correspondence Analysis

Correspondence analysis (CA) is a technique for analysing the relationship between two categorical variables in a contingency table (Greenacre, 2007). It does this by modelling the deviations from the independence model and using singular value decomposition to reduce these data to a small set of orthogonal factors. It is similar to PCA, but it is applicable to categorical data.

The frequency data of the contingency table is converted to a probability table by dividing by the sum of all cells, i.e. the total number of checked attributes over all systems. Deviations from the probability that the systems and attributes are independent are analysed for each cell, giving the strength of relationship between a system-attribute pair. These form the contributions to the χ^2-statistic for the overall table.

For a contingency table with I rows (systems) and J columns (attributes) where x_{ij} is the frequency in the i^{th} row and the j^{th} column, the probability table has values $s_{ij} = x_{ij}/n$. Where $n = \sum_{i,j} x_{ij}$, the sum of attribute checks across all systems and attributes. The

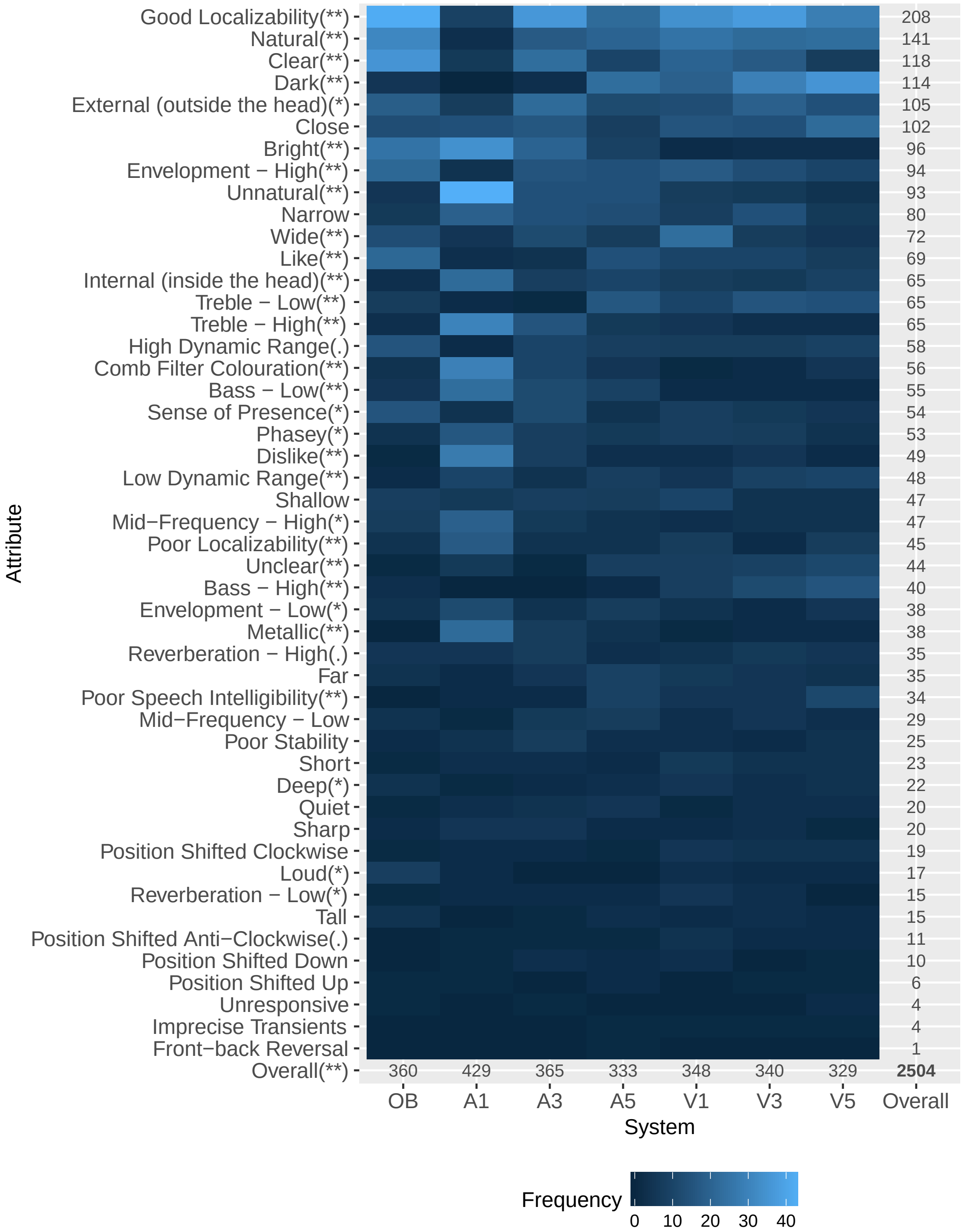


Figure 8.19: Contingency table for CATA attributes over all items. Asterisks indicate the significance level for each attribute using Cochran's Q statistic: $p < 0.05^{*}$, $p < 0.01^{**}$, $p < 0.001^{***}$.

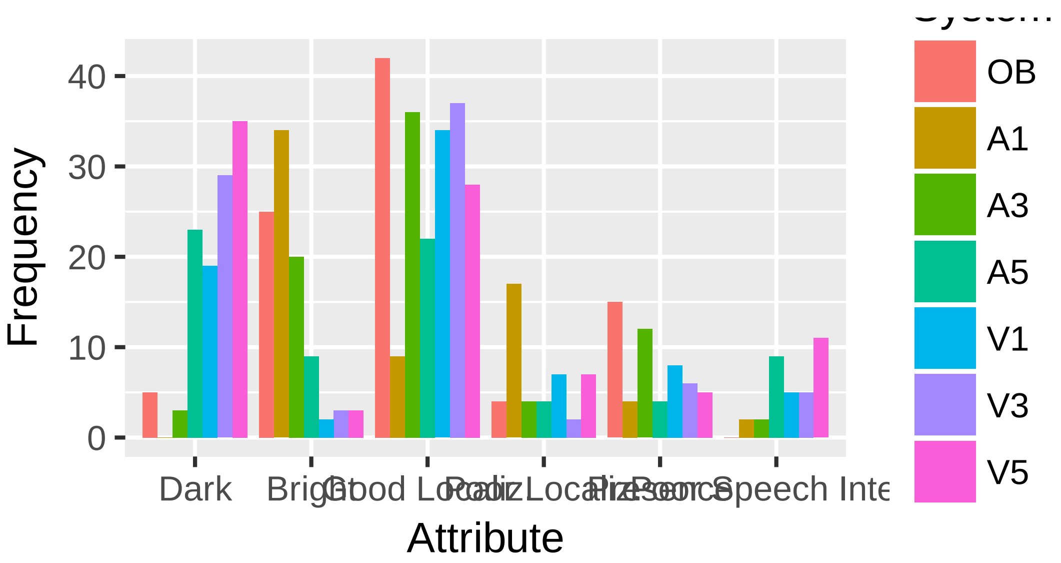
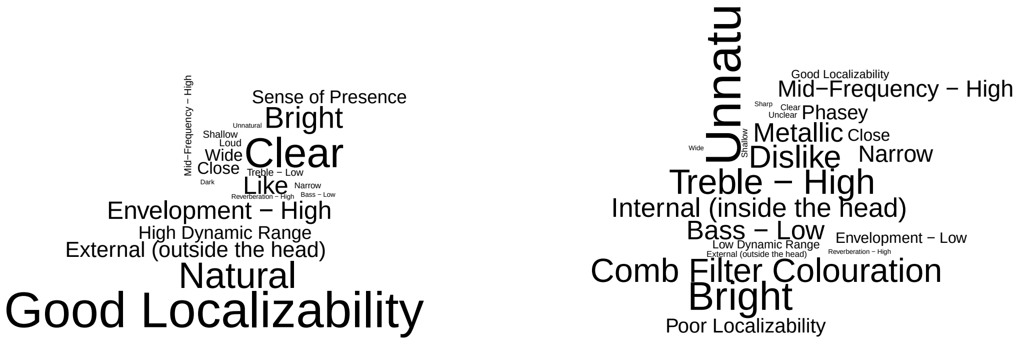


Figure 8.20: Bar plots of frequency of selection per system for example CATA attributes across all items.



(a) Object-based (*OB*)

(b) First-order ambisonics (*A1*)

Figure 8.21: Wordclouds based on attribute frequencies across all items for two example systems

marginal probabilities are then calculated as:

$$s_{i\cdot} = \sum_{j=1}^{J} s_{ij} \qquad s_{\cdot j} = \sum_{i=1}^{I} s_{ij} \tag{8.9}$$

These margins represent either the probability of a check occurring for the system on any attribute ($s_{i\cdot}$) or the probability of a check for the attribute occurring for any system ($s_{\cdot j}$).

The independence model can be expressed as:

$$\forall i, j \quad s_{ij} = s_{i\cdot} s_{\cdot j} \tag{8.10}$$

i.e. the probability of a specific attribute occurring for a given system is equal to the product of the probability of this attribute occurring on any system multiplied by the probability of any attribute occurring for this system. The χ^2 statistic for the table indicates the overall deviation from the independence model and is calculated by summing the contributions of each cell.

$$\chi^2 = n \sum_{i,j} \frac{(s_{ij} - s_{i\cdot} s_{\cdot j})^2}{s_{i\cdot} s_{\cdot j}} = n\Phi^2 \tag{8.11}$$

This χ^2 value is used to test the null hypothesis that the system and attribute categories are independent, by inspecting the probability that it came from a χ^2 distribution with degrees of freedom $df = (I-1)(J-1)$. If the null hypothesis can be rejected, then there is an overall relationship between the systems and the attributes. Φ^2 is the total inertia of the table, which is independent of the sample size and indicates the magnitude of the relationship between systems and attributes.

The independence model can also be expressed as:

$$\forall i, j \quad \frac{s_{ij}}{s_{i\cdot}} = s_{\cdot j} \tag{8.12}$$

where $s_{ij}/s_{i\cdot}$ is the conditional probability of a check for attribute j when it is for system i. Independence arises when the probability for attribute j is not dependent on the system, meaning that the conditional probability is equal to the marginal probability $s_{\cdot j}$. The model can be expressed in symmetrical terms for the conditional probability of a check for system i when it is for attribute j:

$$\forall i, j \quad \frac{s_{ij}}{s_{\cdot j}} = s_{i\cdot} \tag{8.13}$$

In CA, the row profiles $\{s_{ij}/s_{i\cdot}\,;\, j = 1, J\}$ and the column profiles $\{s_{ij}/s_{\cdot j}\,;\, i = 1, I\}$ characterise each system across all attributes and each attribute across all systems, respectively. In a similar complementary manner, the average profiles define the average system $\{s_{\cdot j}\,;\,j=1,J\}$ and attribute $\{s_{i\cdot}\,;\,i=1,I\}$ across the attributes and systems, respectively.

Distances are then calculated between pairs of systems and pairs of attributes, using their row/column profiles and the χ^2-distance. A system has a point i in a J-dimensional space, with coordinate $s_{ij}/s_{i\cdot}$ for the j^{th}-dimension. The χ^2 distance is a weighted Euclidean distance, defined here between systems i and l:

$$d^2_{\chi^2}(i,l)=\sum_{j=1}^{J}\frac{1}{s_{\cdot j}}(\frac{s_{ij}}{s_{i\cdot}}-\frac{s_{lj}}{s_{l\cdot}})^2 \tag{8.14}$$

Each point i is also weighted by $s_{i\cdot}$, so the influence of a system in the analysis increases with its frequency of attribute selection. The centre of mass of the cloud of system points is the average system profile $G_I=\{s_{\cdot j}\,;\,j=1,J\}$. The inertia[3] of point i relative to G_I is:

$$\text{Inertia}(i/G_I)=s_{i\cdot}\,d^2_{\chi^2}(i,G_I)=\sum_{j=1}^{J}\frac{(s_{ij}-s_{i\cdot}s_{\cdot j})^2}{s_{i\cdot}s_{\cdot j}} \tag{8.15}$$

When viewed with respect to equation 8.11, it can be seen that this represents the contribution of the row/system to the χ^2 statistic besides the multiplicative factor n, which is why it is called the χ^2 distance. It actually describes the contribution to the overall inertia Φ^2.

The columns/attributes can be analysed in the same manner as in Equations (8.14) and (8.15):

$$d^2_{\chi^2}(j,k)=\sum_{i=1}^{I}\frac{1}{s_{i\cdot}}(\frac{s_{ij}}{s_{\cdot j}}-\frac{s_{ik}}{s_{\cdot k}})^2 \tag{8.16}$$

In the factor analysis, each point j is weighted by $s_{\cdot j}$, so the influence of an attribute in the analysis increases with its frequency of selection. The centre of mass of the cloud of attribute points is the average attribute profile $G_J=\{s_{i\cdot}\,;\,i=1,I\}$ and the inertia is given by:

$$\text{Inertia}(j/G_J)=s_{\cdot j}\,d^2_{\chi^2}(j,G_J)=\sum_{i=1}^{I}\frac{(s_{ij}-s_{i\cdot}s_{\cdot j})^2}{s_{i\cdot}s_{\cdot j}} \tag{8.17}$$

A generalised singular value decomposition is performed using the χ^2 distances and the marginal probabilities as row and column weights to obtain a set of orthogonal axes which maximises inertia. The average system and attribute profiles are both located at the origin of these axes, and the individual system and attribute profiles can be simultaneously represented on the axes (unlike in PCA). The results can be viewed on a two-dimensional graph,

[3] The use of the term inertia comes from analogy to the physical concept of the moment of inertia.

	Dim.1	Dim.2	Dim.3	χ^2	df	p-value
Frontal	42.6	24.5	11.9	340.2	246	<0.001
Lateral	53.1	17.0	11.8	357.9	240	<0.001
Elevated	43.7	21.5	14.7	379.0	258	<0.001
Family	41.3	22.0	12.6	453.3	270	<0.001
Forest	30.5	22.9	15.1	314.9	264	0.017
Protest	34.6	31.2	11.5	394.7	258	<0.001
Single	59.9	14.6	9.9	561.1	270	<0.001
Complex	44.2	29.7	8.4	696.5	282	<0.001
All items	59.4	22.3	6.2	972.2	282	<0.001

Table 8.10: Correspondence analysis (CA) results with percentage of variance explained by each of the first three dimensions, χ^2 statistic with associated degrees-of-freedom (df) and p-value, for each item plot shown in figure 8.22.

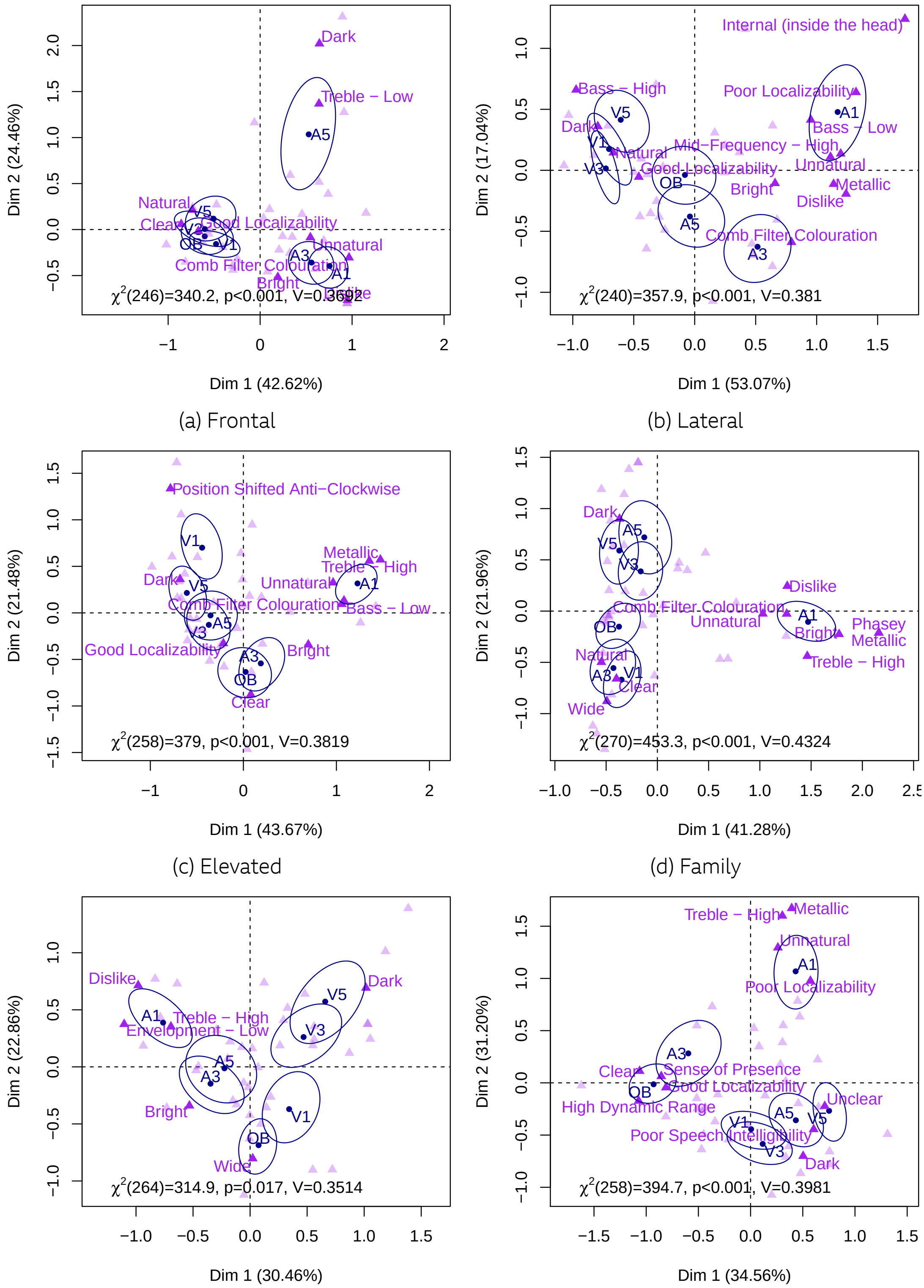
normally using the dimensions that explain the largest amount of inertia. Points on the graph represent the profiles of attributes and systems within this common factorial space, as projected onto the two displayed dimensions. These profiles are relative to the average profiles at the origin, so the distance from the origin represents the deviation from independence within the two presented dimensions.

8.5.4.5 Correspondence Analysis Results

Contingency tables were created for the CATA attribute data corresponding to each audio item. CA was applied to each contingency table, using χ^2-distances, and the resulting bi-dimensional factor maps are presented in figure 8.22. The profiles of both the attributes and systems are represented, along with 95 % confidence ellipses for the systems. Table 8.10 shows the percentage of the variance explained by the first three dimensions (eigenvectors) for each CA, along with the χ^2-statistic associated with the contingency tables on which the CA is based. The CA was carried out using the FactoMineR package in R (Lê et al., 2008).

To properly interpret the factor maps, one should consider both the contributions of the attributes to the construction of the dimensions and the quality of representation of the attributes by the axis (given by the angle between the attribute profile and the axis). When interpreting the strength of relationships, the eigenvalues for each axis should be taken into account. The only attributes appearing in figures 8.22 and 8.24 are those which show significance in the permutation tests (table 8.8), make a strong contribution to the axes ($\geq$ 5 %), and are well represented by the axes (cos^2 ≥ 0.5).

Another tool is to inspect the angle between a system and the attributes in the multi-dimensional space, Meyners, Castura, and Carr (2013) call this multidimensional alignment. It establishes whether the relationship between a system and an attribute is genuine and



(a) Frontal

(b) Lateral

(c) Elevated

(d) Family

(e) Forest

(f) Protest

Figure 8.22: Correspondence analysis for each item, using χ^2-distances. Systems shown with 95% confidence ellipses. Labelled attributes (purple) were found to be significant, made a strong contribution to the axes ($\geq 5\%$), and are well represented by the axes ($\cos^2 \geq 0.5$). Other attributes are unlabelled and semi-transparent.

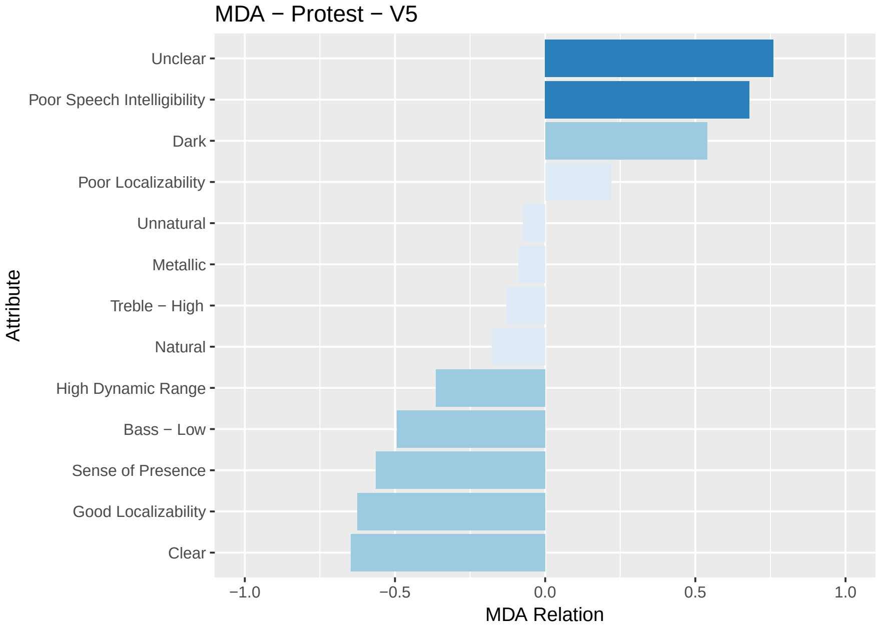


Figure 8.23: Example of multidimensional alignment of attributes to system *V5* for *Protest* item. 1 indicates perfect positive relation, -1 indicates perfect negative relation and 0 indicates no relation.

not simply a result of the projection onto the first two dimensions. Smaller angles indicate stronger characterisation of a system by an attribute, with respect to the character of the other systems. These angles were used during the following interpretation of the CA maps. An example is given in figure 8.23 for system *V5* and item *Protest*. Here the angle ψ_{mda} has been transformed by: $(\pi - 2\psi_{\text{mda}})/\pi$ so that 1 indicates a perfect positive relation, -1 indicates a perfect negative relation and 0 indicates no relation. Dark blue bars have an relation value of more than 0.66 i.e. an angle of less than 30°. This shows that *Unclear* and *Poor Speech Intelligibility* are closely related to system *V5* in this case.

For the frontal source, the *A1* system is most closely aligned with the first dimension (which explains 42.62 % of the inertia) and makes the largest contribution to it, but all systems except *A5* are fairly well represented (cos$^2 > 0.4$, corresponding to an angle of approximately 50°). The most strongly associated attributes are naturalness, clarity, localizability and comb filter colouration. Thus, the first dimension separates the ambisonics systems from the *OB* and VBAP systems, and the ambisonics systems have unnatural characteristics relating to comb filter artefacts. *Dark* makes by far the largest contribution to the second dimension. This dimension is also related to brightness or spectral balance; it separates the ambisonics systems, indicating that *A5* is particularly *Dark*. Since the second dimension explains 24.42 % of the inertia, this is a noteworthy effect. The confidence ellipses indicate that the VBAP and *OB* systems are poorly separated in these two dimensions, as are the systems *A1* and *A3*. Inspecting the third dimension (representing 12.99 % of the inertia) reveals separation between *A1* and *A3*, where *A3* is aligned with the low mid-frequency attribute and *A1* with high mid-frequency, although less strongly.

For the lateral source, *OB* is close to the average system profile in the first two dimensions. Again *A1* contributes strongly to the first dimension (representing 53.07 % of the inertia), which separates it from the VBAP systems. This dimension appears to relate most strongly to tone colour attributes (e.g. *Metallic*, *Dark*) and localizability, with *A1* displaying *Poor Localizability* and *Bright*, *Metallic* tone colour, whilst the VBAP systems show more *Dark*/*Bass–High* tone colour, but also good localizability. *OB* is also weakly characterised by *Good Localizability* and *External*. The second dimension (representing 17.03 % of the inertia) best represents *A3*, associating it with *Treble–High* and *Comb Filter Colouration* most strongly, but this dimension also has a large contribution from *Internal* and appears related to distance/externalisation. *A5* is strongly represented by the attribute *Far*, and *A1* by inside the head. *A3* is well separated from *A5* and *OB* in the third dimension (representing 11.21 % of the inertia), which seems characterised by distance/depth and also by liking, with *OB* most strongly aligned with *Like*.

For the elevated source, again the *A1* system is strongly represented in the first dimension (representing 43.73 % of the inertia), separating it from other systems. This dimension relates to colouration (*Bright*-*Dark*, *Metallic*), naturalness and localizability. *V1* and *V5* are related to *Dark*. The second dimension (representing 20.79 % of the inertia) strongly relates to clarity, but also localizability. It separates *A3* and *OB* from the other systems. This dimension is also influenced by *Position Shifted Anti-Clockwise*, by which the *V1* system is strongly represented. The third dimension (representing 14.57 % of the inertia) best represents *Position Shifted Up*, which was not significant in differentiating the systems, but is worth mentioning given the source material. It is also related to *Position Shifted Anti-Clockwise* and *Wide*. This dimension most strongly relates to systems *A5* and *V1*, with the source perceived shifted up for *A5* and shifted wider for *V1*.

For *Family*, *A1* is strongly represented by the first dimension (representing 45.16 % of the inertia) and represented by colouration attributes (most strongly *Metallic*), and *Unnatural* and *Dislike*. The negative side of the axis represents *Natural* well, with *OB* and *A3* most closely aligned. The second dimension (representing 20.51 % of the inertia) describes *Dark* versus *Clear* and *Wide*, with *V5* and *A5* having particularly unclear dark character and *V1* being particularly *Wide* and *V1* and *A3* being *Clear*. *OB* was most strongly represented by *Like*, but also *Natural* and *Outside the head*. *Like* is represented in the third dimension (representing 10.88 % of the inertia), which clearly separates the *OB* system from others. *High dynamic range* is also well represented in this axis.

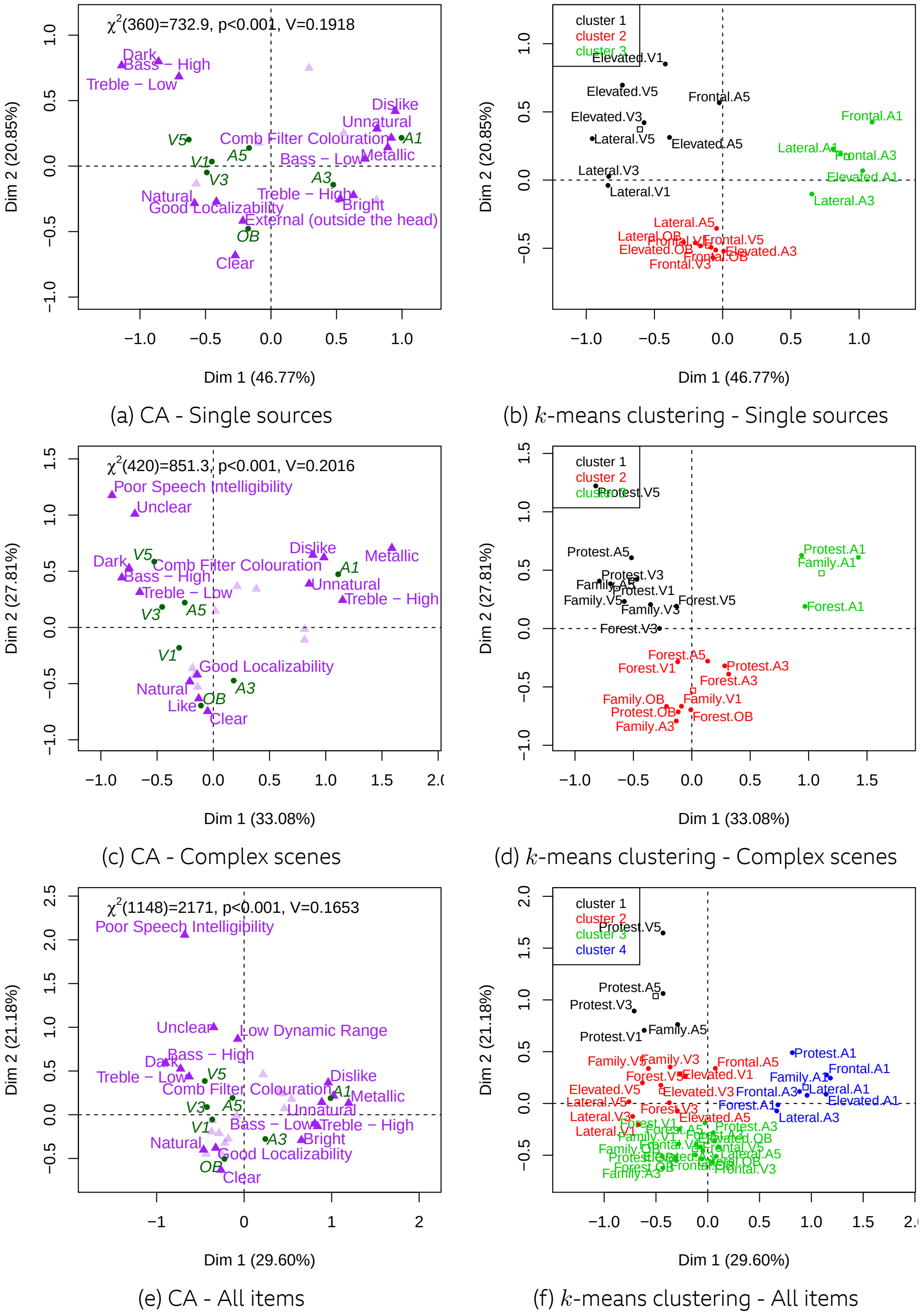
For the *Forest* scene, the first dimension (representing 34.60 % of the inertia) is most strongly characterised by envelopment, spectral balance and liking, with *A1* being well represented by *Dislike*, *Envelopment–Low* and *Treble–High*. The *V3* and *V5* systems are represented by *Dark*. The second dimension (representing 21.37 % of the inertia) separates *Dark* from *Wide*, with *OB* and *V1* characterised by width.

For *Protest*, the first dimension (representing 34.71 % of the inertia) appears to describe clarity, localizability and the sense of presence, with *OB* particularly characterised by *Clear*, *Good Localizability* and *Sense of Presence*, and *A3* similarly, but to a lesser extent. System *V5* is well represented in the opposing direction, associated with *Unclear* and *Poor Speech Intelligibility*. The second dimension (representing 31.20 % of the inertia), relates primarily to spectral balance and colouration. System *A1* is clearly separated from the others in this dimension and is associated with *Metallic* and *Unnatural*. *Poor Localizability* is also related to this dimension and system *A1*. The third dimension (representing 11.55 % of the inertia) separates *OB* from *A3*, and *V1* from *V3*. *V3* and *A3* are characterised more by *Bass–Low*, *Sharp* and *Narrow*, and *OB* more by *Natural*.

For the item groups, contingency tables were constructed from CATA data frequencies for each system-item combination, essentially concatenating tables for each individual item, using only attributes that were found to be significant across the item group. A CA was performed and the sums of attribute selection frequencies for each system across items were used as supplementary variables, projected into the resulting latent space. The left-hand plots of figure 8.24 show the systems averaged across items, alongside the attributes. Similar patterns can be seen in the factor maps for each item group. The first dimension separates systems according to tone colour, bright systems with comb filter artefacts having positive coordinates (mainly *A1*). The second dimension separates systems with dark tone colour and lack of clarity (particularly *A5* and *V5*) from those that are clear and natural, with good localizability (particularly *OB*). *A3* is more closely aligned with *A1* for single sources and with *OB* for complex scenes.

8.5.4.6 Hierarchical Cluster Analysis

Subsequently, hierarchical cluster analysis (HCA) was performed in the CA factor space using Ward's method (Husson, Josse, et al., 2010) on the contingency table rows, i.e. on each combination of system and audio item. Clustering was done for each audio item group and across all items. The number of clusters in each case was chosen automatically to minimise



(a) CA - Single sources

(b) k-means clustering - Single sources

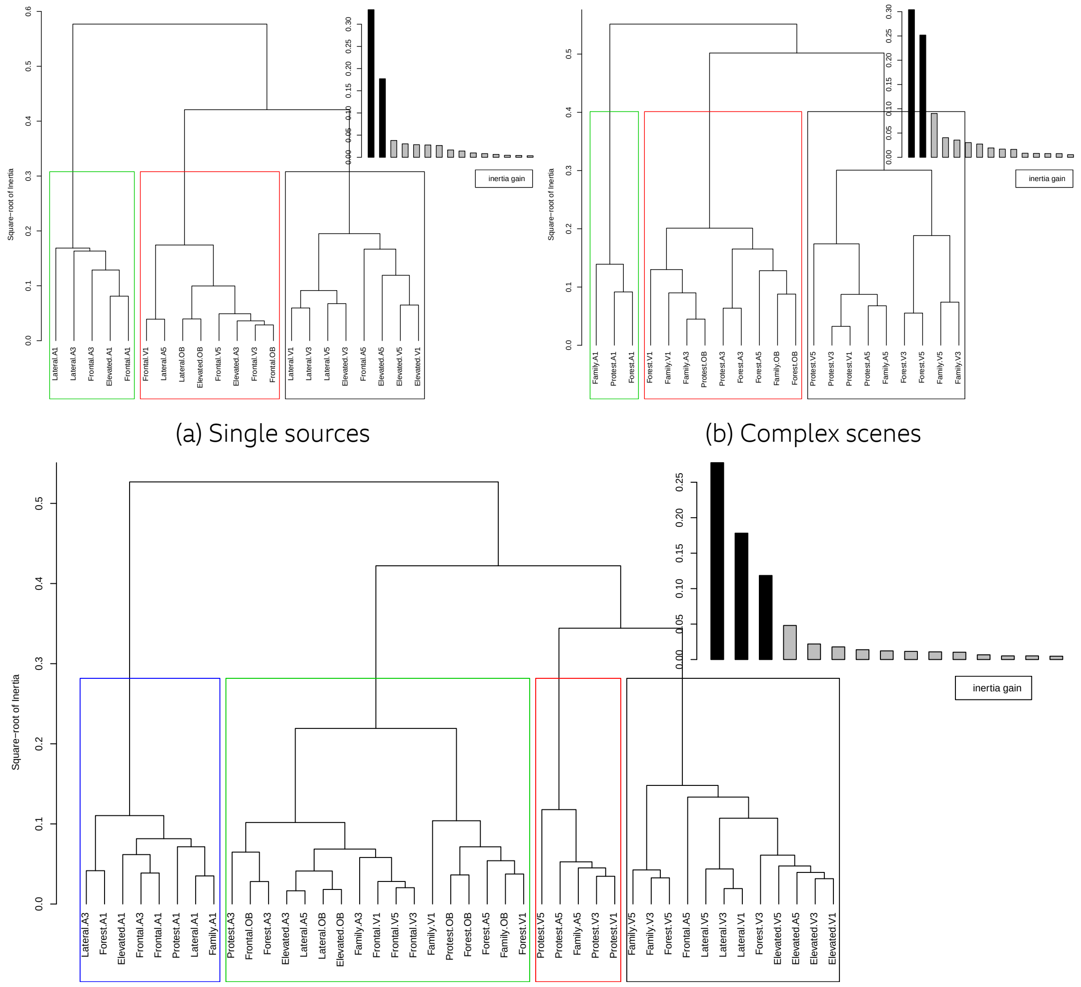
(c) CA - Complex scenes

(d) k-means clustering - Complex scenes

(e) CA - All items

(f) k-means clustering - All items

Figure 8.24: Factor maps resulting from correspondence analysis (CA) of all system-item combinations, using only significant attributes (table 8.8). Left-hand plots: systems grouped across items (dark green) projected as supplementary variables. Labelled attributes made a significant contribution to the axes, as in figure 8.22. Right-hand plots: system-item combinations represented in clusters, obtained using hierarchical cluster analysis (HCA) with k-means consolidation. Number of clusters chosen automatically to minimise loss of inertia.



(a) Single sources

(b) Complex scenes

(c) All items

Figure 8.25: Hierarchical cluster analysis (HCA) of system profiles for individual items.

Cluster 1		Cluster 2		Cluster 3	
Attribute	Test value	Attribute	Test value	Attribute	Test value
Dark	9.09	Clear	4.85	Comb Filtering	5.16
Treble–Low	5.81	Good Localizability	3.31	Unnatural	5.15
Bass–High	4.89	Natural	3.14	Dislike	4.69
Natural	2.11	External	2.78	Metallic	3.83
				Bright	3.44
				Bass–Low	3.16
				Mid–High	3.1
				Treble–High	3
				Internal	2.69
Bright	-5.19	Dark	-4.96	Natural	-7.22
Mid–High	-3.73	Treble–Low	-3.48	Dark	-5.55
Treble–High	-3.38	Poor Localizability	-2.84	Good Localizability	-5.5
Comb Filtering	-3.13	Dislike	-2.69	Like	-3.81
Bass–Low	-2.91	Bass–High	-2.65	Clear	-3.7
Unnatural	-2.73	Unnatural	-2.39	Treble–Low	-2.92
Metallic	-2.27	Comb Filtering	-2.12	Bass–High	-2.67
Dislike	-2.03			External	-2.08

Table 8.11: v-test statistics for the contribution of attributes to each cluster for single sources (significant contributions only)

loss of inertia, the clusters were then consolidated with the k-means algorithm. This process was performed using the HCPC method of the FactoMineR package in R (Lê et al., 2008). Figure 8.25 shows the hierarchical clustering of the system profiles for each audio item, indicating the optimum segmentation, whilst the right-hand plots of figure 8.24 present the system-item clusters resulting from the k-means consolidation step. Tables 8.11 to 8.13 show all attributes that significantly contributed to the formation of the construction of the clusters given in Figures 8.24b, 8.24d and 8.24f, respectively. The v-test statistics are shown, which result from a hypergeometric test. The attributes with positive v-test statistics can be said to characterise the cluster, whilst those with negative v-test statistics are opposed to the character of the cluster.

The clustering for single source items in shown in figure 8.24b, alongside the corresponding attribute map in figure 8.24a. Table 8.11 shows the v-test statistics for all attributes that significantly contribute to the construction of the clusters for single source items. Cluster 2 (*Clear*, *Good localizability*) is dominated by the *OB* system, and cluster 3 (*Comb-filter colouration*, *Unnatural*) is dominated by the *A1* system. It can be seen that the VBAP systems are grouped with *OB* for the frontal source. Cluster 1 is dominated by the characteristically *Dark* system-item combinations, including non-frontal VBAP stimuli and *A5* for the frontal and elevated sources. For the *Lateral* source, *A5* is grouped with the *OB* stimuli.

Cluster 1		Cluster 2		Cluster 3	
Attribute	Test value	Attribute	Test value	Attribute	Test value
Dark	7.56	Clear	5.99	Treble–High	5.08
Poor Intelligibility	6.04	Wide	3.28	Metallic	4.84
Unclear	4.77	Natural	3.03	Unnatural	4.71
Bass–High	4.24	Like	2.97	Comb Filtering	4.16
Treble–Low	3.42	Envelopment–High	2.68	Dislike	4.14
		Sense of Presence	2.65	Bright	3.14
		Good Localizability	2.41		
Clear	-4.14	Dark	-5.07	Dark	-3.92
Bright	-3.86	Poor Intelligibility	-4.92	Envelopment–High	-3.14
Treble–High	-3.86	Unclear	-4.51	Treble–Low	-2.87
Bass–Low	-2.73	Dislike	-3.05	Clear	-2.84
Metallic	-2.28	Comb Filtering	-2.59	Good Localizability	-2.68
Unnatural	-2.01	Bass–High	-2.56	Natural	-2.53
		Metallic	-2.12	Bass–High	-2.26
		Unnatural	-2.05	Wide	-2.12
				Sense of Presence	-1.99

Table 8.12: v-test statistics for the contribution of attributes to each cluster for complex scenes (significant contributions only)

The clustering for complex scene items in shown in figure 8.24d, alongside the corresponding attribute map in figure 8.24c. Table 8.12 shows the v-test statistics for all attributes that significantly contribute to the construction of the clusters for complex scenes. Cluster 3 (*Treble–High*, *Metallic*) contains only *A1* stimuli, cluster 2 (*Clear*) includes all *OB* and *A3* stimuli, and cluster 1 (*Dark*) includes all *V3* and *V5* stimuli.

The clustering for all items in shown in figure 8.24f, alongside the corresponding attribute map in figure 8.24e. Table 8.13 shows the v-test statistics for all attributes that significantly contribute to the construction of the clusters over all items. There were four clusters, cluster 1 (*Poor speech intelligibility*, *Unclear*) particularly included *Protest* stimuli, cluster 3 (*Clear*) was dominated by *OB* and *A3* with complex scenes, cluster 4 (*Unnatural*) was dominated by *A1*, with cluster 2 (*Dark*) containing many of the VBAP stimuli, those that were not so strongly characterised by poor speech intelligibility.

hierarchical cluster analysis (HCA) was also applied to the attributes, as shown in figure 8.26. The HCA produced a tree (or dendrogram) identifying the hierarchical relationships between attributes (left-hand side: Figures 8.26a, 8.26c and 8.26e). The trees are shown using the square-root of inertia to make diagrams more readable, since most groupings gave low inertia gain. Again the number of clusters was chosen to minimise loss of inertia and clusters were consolidated with the k-means algorithm. The attribute clusters resulting from k-means are shown in CA maps (right-hand side: Figures 8.26b, 8.26d and 8.26f).

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Attribute	Test value	Attribute	Test value	Attribute	Test value	Attribute	Test value
Poor Intelligibility	10.19	Dark	9.82	Clear	7.98	Unnatural	7.12
Unclear	4.49	Treble–Low	4.92	Natural	4.44	Comb Filtering	6.82
Low Dynamic Range	4.2	Bass–High	4.7	Good Localizability	4.06	Dislike	6.24
Dark	2.78	Good Localizability	2.39	Wide	3.57	Metallic	6.17
Treble–Low	2.51			Envelopment–High	3.19	Treble–High	5.59
Bass–High	2.48			External	3.15	Bright	4.92
				Like	2.95	Bass–Low	3.85
				Sense of Presence	2.21	Mid–High	3.46
						Poor Localizability	3.05
						Internal	2.96
						Phasey	2.88
Bright	-3.5	Bright	-5.03	Dark	-7.06	Natural	-7.22
Treble–High	-3.22	Treble–High	-3.68	Unclear	-4.44	Dark	-7.01
Good Localizability	-2.93	Poor Intelligibility	-3.38	Poor Intelligibility	-4.34	Good Localizability	-5.85
Clear	-2.71	Bass–Low	-3.04	Dislike	-4.14	Clear	-5.03
Unnatural	-2.41	Mid–High	-2.88	Low Dynamic Range	-4.05	Treble–Low	-4.38
Metallic	-2.19	Clear	-2.84	Bass–High	-3.65	Envelopment–High	-4.28
Bass–Low	-2.15	Comb Filtering	-2.37	Comb Filtering	-3.44	Like	-4.15
		Unnatural	-2.26	Treble–Low	-3.34	Bass–High	-4.07
		Metallic	-2.19	Unnatural	-3.27	Wide	-3.16
				Metallic	-2.67	External	-2.71
				Poor Localizability	-2.37	Poor Intelligibility	-2.43

Table 8.13: v-test statistics for the contribution of attributes to each cluster for all items (significant contributions only)

For single sources, cluster 1 groups *Dark* with the related bass and treble effects, cluster 2 groups attributes related to good spatial quality, and cluster 3 has all other significant attributes, including colouration related to brightness and comb filtering, and poor spatial quality. Cluster 2 also contains *Like* whilst cluster 3 contains *Dislike*. Note that *Far* was reassigned from cluster 3 to cluster 2 through the k-means consolidation step, which makes more sense conceptually. For complex scenes, cluster 1 features attributes that relate to dark timbre and lack of clarity, cluster 2 features attributes related to convincing spatial reproduction of a scene and liking, and cluster 3 features colouration attributes relating to brightness and comb filtering, as well as inside-the-head localisation and disliking. Note that here *Reverberation–High* has been moved from cluster 3 to cluster 2 during k-means consolidation. Over all items, the clustering is largely similar to that for complex scenes.

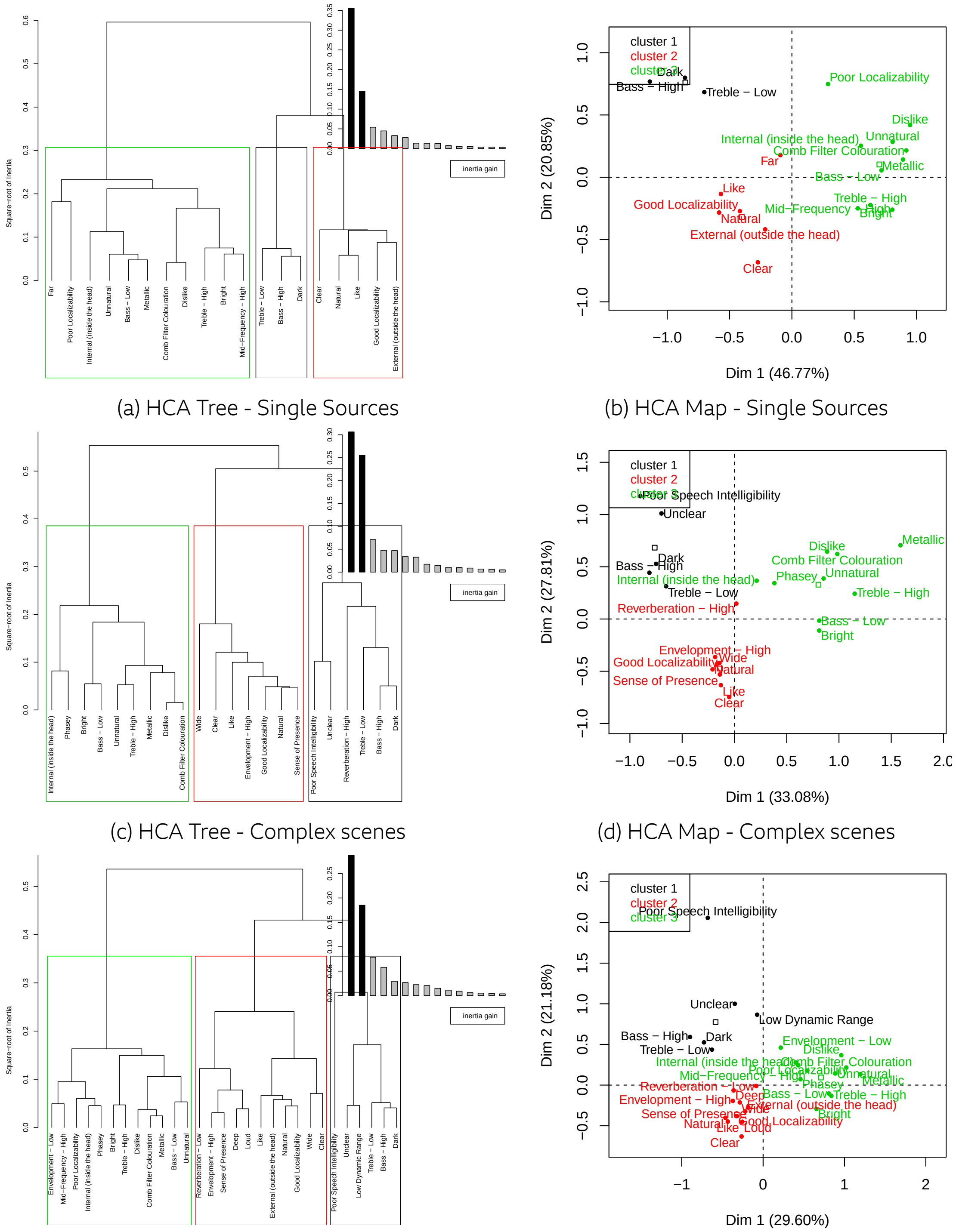
8.5.4.7 Penalty-Lift Analysis

To get some indication of the influence of characteristics on the perceived overall sound quality, the penalty-lift analysis described by Meyners, Castura, and Carr (2013) was applied. This takes the mean of overall quality ratings when the attribute is used and subtracts the mean of ratings when the attribute is not used, suggesting a penalty or lift in overall quality due to the presence of an attribute. Penalty-lift analysis was performed on all significant attributes for each item group and across all items, shown in figure 8.27.

8.5.5 Head Rotations

The head tracking data collected during the rating trials were analysed. Quaternions representing the orientation of the listener's head were used to rotate a front-facing unit vector to obtain a *look vector*, i.e. a unit vector pointing in the direction that the listener's head was facing. Directional statistics could then be used to summarise the distribution of look vectors during the rating sessions.

For each trial, the concentration parameter κ of a von Mises-Fisher distribution was estimated from this set of vectors. The von Mises-Fisher distribution is a generalisation of the von Mises distribution to a unit hypersphere in p-dimensions, where here $p = 3$, i.e. the unit 2-sphere S^2. The von Mises distribution itself is a simplified close approximation to the wrapped normal distribution on the circle. It is a unimodal and rotationally symmetric distribution. The concentration parameter κ is zero for a uniform distribution of look vectors on the sphere and tends to infinity as the distribution becomes tightly focussed around the mean. It is a reciprocal measure of dispersion, so κ^{-1} is analogous to the standard deviation



(a) HCA Tree - Single Sources

(b) HCA Map - Single Sources

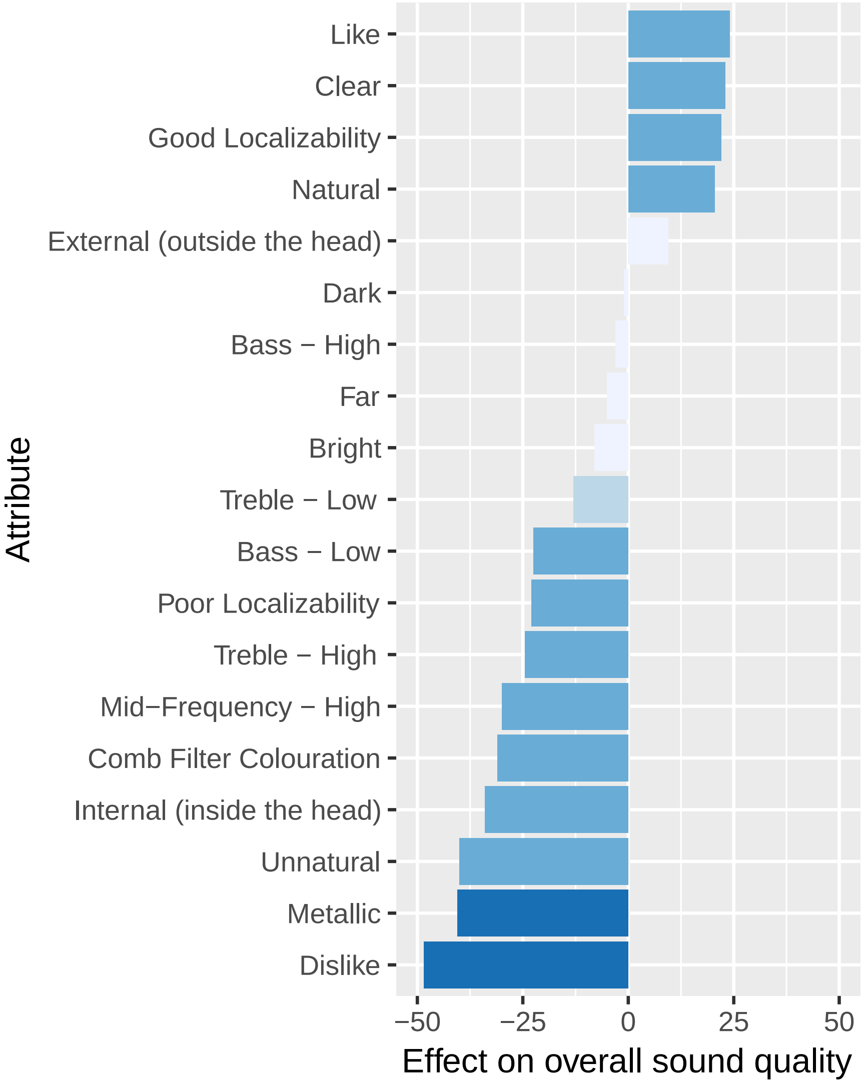
(c) HCA Tree - Complex scenes

(d) HCA Map - Complex scenes

(e) HCA Tree - All items

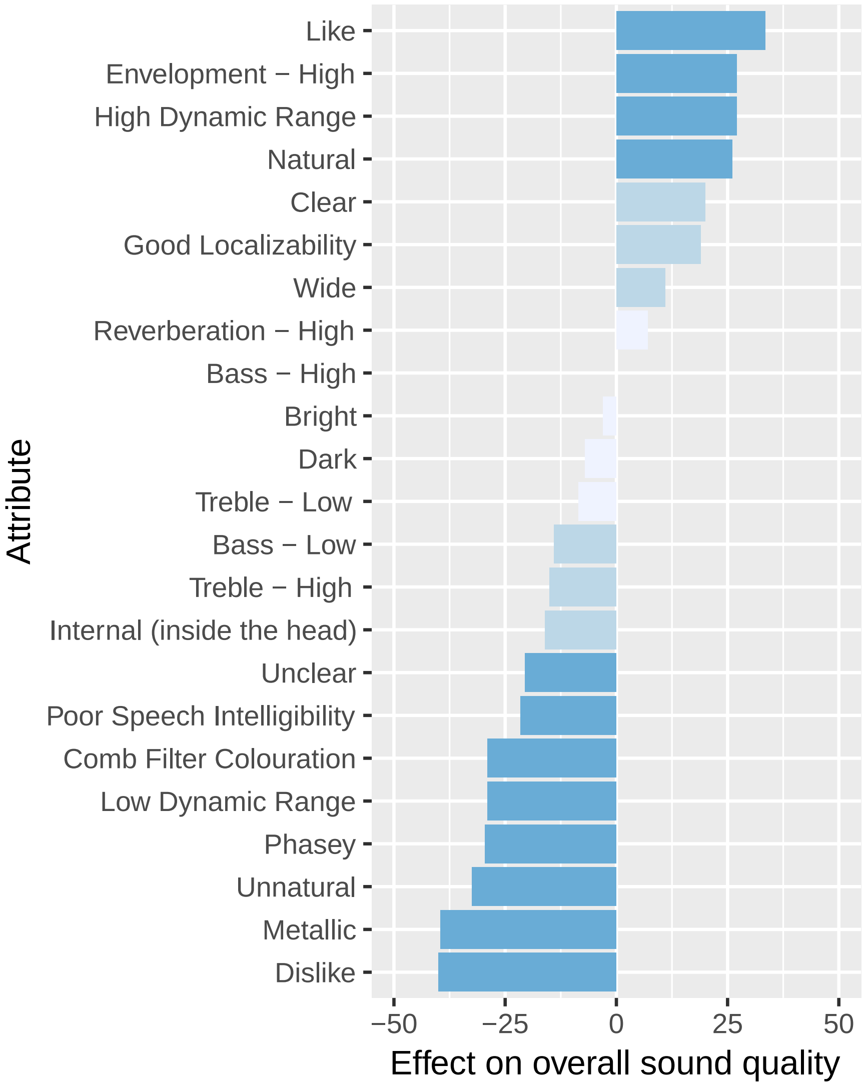
(f) HCA Map - All items

Figure 8.26: HCA of attributes for each item group, based on the CAs shown in figure 8.24. Left-hand plots show the hierarchical clustering analysis in a tree, with the y-axis showing the square-root of inertia to aid visualisation of the relationships between attributes. The number of clusters was chosen automatically, to minimise loss of inertia. Right-hand plots show the clusters after k-means consolidation.



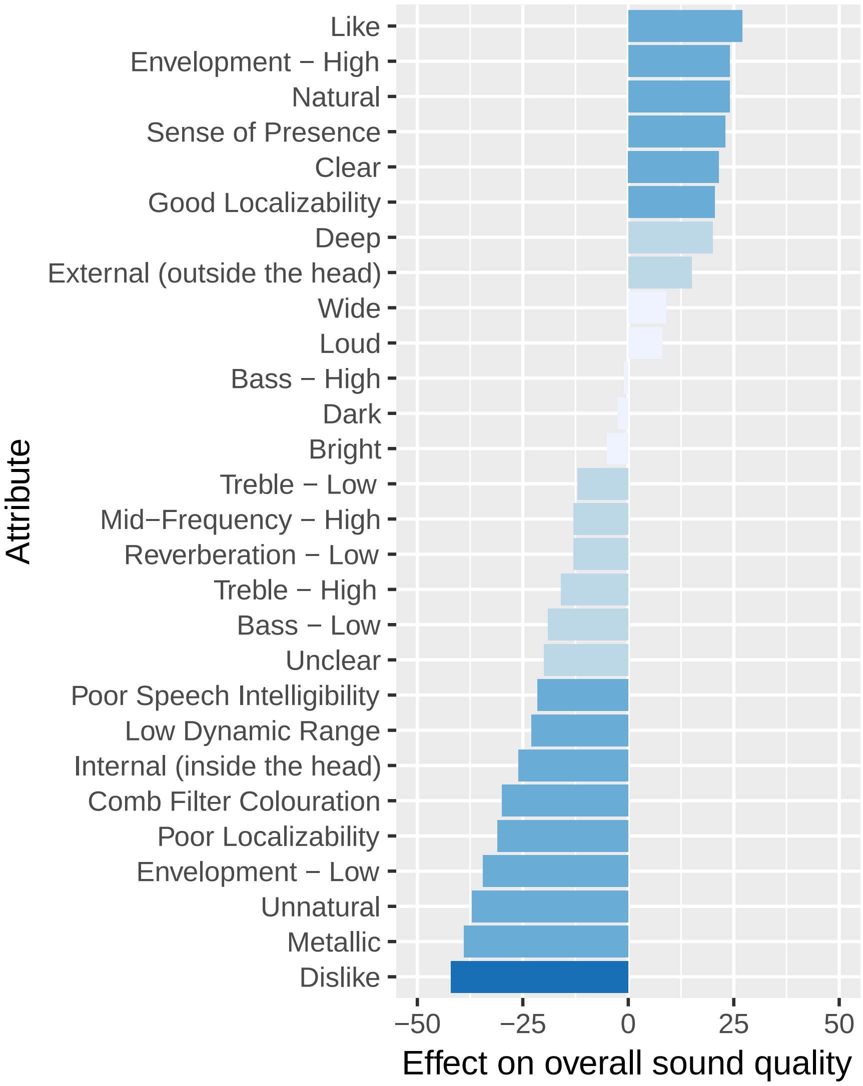
(a) Single Sources

Figure 8.27: Penalty-lift analysis showing the difference in mean overall sound quality values for when each significant attribute is used compared to when it is not used, for item groups and all items.



(b) Complex scenes

Figure 8.27: Penalty-lift analysis showing the difference in mean overall sound quality values for when each significant attribute is used compared to when it is not used, for item groups and all items.



(c) All items

Figure 8.27: Penalty-lift analysis showing the difference in mean overall sound quality values for when each significant attribute is used compared to when it is not used, for item groups and all items.

σ. A simple approximation of the dispersion parameter is given by (Sra, 2012):

$$\hat{\kappa} = \frac{\bar{R}(p - \bar{R}^2)}{1 - \bar{R}^2} \tag{8.18}$$

where $\bar{R}$ is the length of the averaged vector

$$\bar{R} = \frac{\| \sum_{i=1}^{N} x^i \|}{N} \tag{8.19}$$

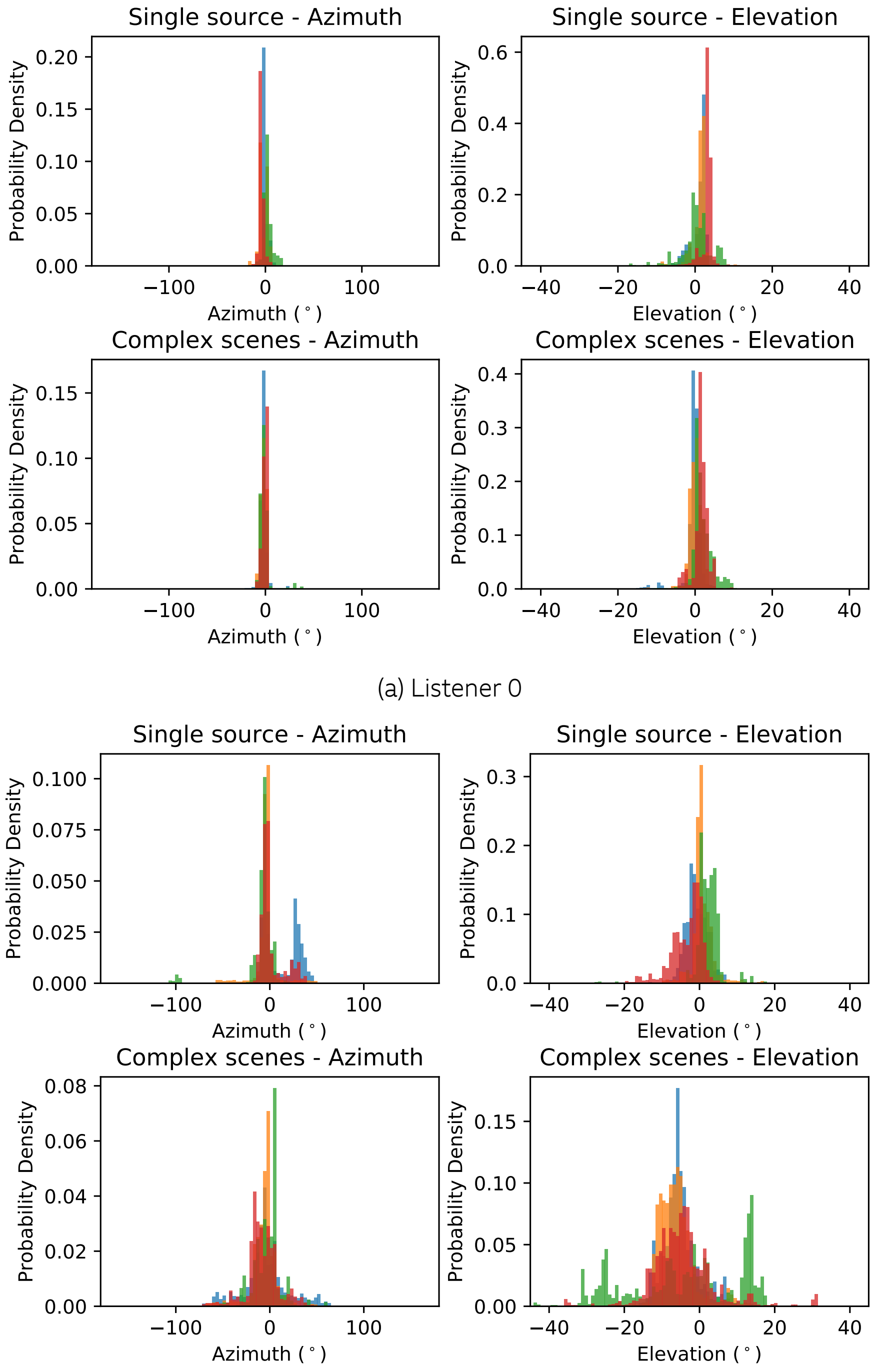
with N being the number of orientation measurements and x^i being the i^{th} orientation vector. Wightman and Kistler (1989b) use estimates of κ^{-1} as a measure of dispersion of localisation data. In that study $\kappa^{-1} = 0.01$ is representative of small dispersions and $\kappa^{-1} = 0.18$ is deemed representative of large dispersions. Clearly the nature of localisation is different to the distribution of look vectors, but it provides a point of reference. A two-way ANOVA of κ data with explanatory variables for *session* and *assessor* indicates that the assessor effect was highly significant ($p < 0.001$) but the session effect was not ($p = 0.252$).

For ease of interpretation, the azimuth (θ) and elevation (ϕ) angles of the look vectors were also calculated, and the interquartile range of the deviation from the median values was determined for each trial. The median interquartile range of look azimuth deviations across all trials and assessors was 8.21° and for elevation it was 3.04°. The amount of head rotation, rather than just the direction of the look vector, was also obtained via the axis-angle representation of the quaternions (this angle is represented with ψ). This allows roll rotations to be incorporated too, which were smaller but still present. The median of interquartile range for ψ across all trials and assessors was 7.50°.

Table 8.14 summarises the data by assessor and figure 8.28 shows histograms of the look vector azimuth and elevation angles for two example assessors. The range of head orientations is low in general, with a strong focus towards the front, which is where the screen with the experiment interface was located.

8.5.6 Questionnaire Responses

This section provides analysis of the responses to the post-experiment questionnaire described in section 8.4.5. First the quantitative results are presented, followed by a discussion of the free-text responses.



(a) Listener 0

(b) Listener 1

Figure 8.28: Histograms of the azimuth and elevation of assessor look vectors from head tracking data for each trial, split by audio item group.

Assessor	IQR(θ)	IQR(ϕ)	κ^{-1}	IQR(ψ)
0	3.652	1.585	0.003	5.092
1	13.186	4.653	0.048	12.148
2	14.312	4.217	0.059	18.800
3	3.436	4.967	0.009	5.510
5	10.893	1.873	0.013	3.475
6	8.576	4.023	0.045	5.568
7	3.843	4.369	0.012	6.376
8	10.296	2.768	0.042	6.360
9	11.127	2.050	0.030	18.411
10	5.507	2.095	0.036	8.840
11	9.111	4.294	0.043	4.880
12	5.557	2.361	0.008	3.975
13	15.063	7.490	0.040	9.421
14	9.489	3.362	0.072	9.160
15	6.742	1.992	0.053	4.655
16	4.676	2.913	0.009	3.885
17	11.271	2.255	0.119	28.107
18	6.609	1.708	0.046	7.656
19	7.909	5.082	0.022	8.530
20	11.777	2.848	0.099	14.713
21	13.207	4.547	0.052	12.603

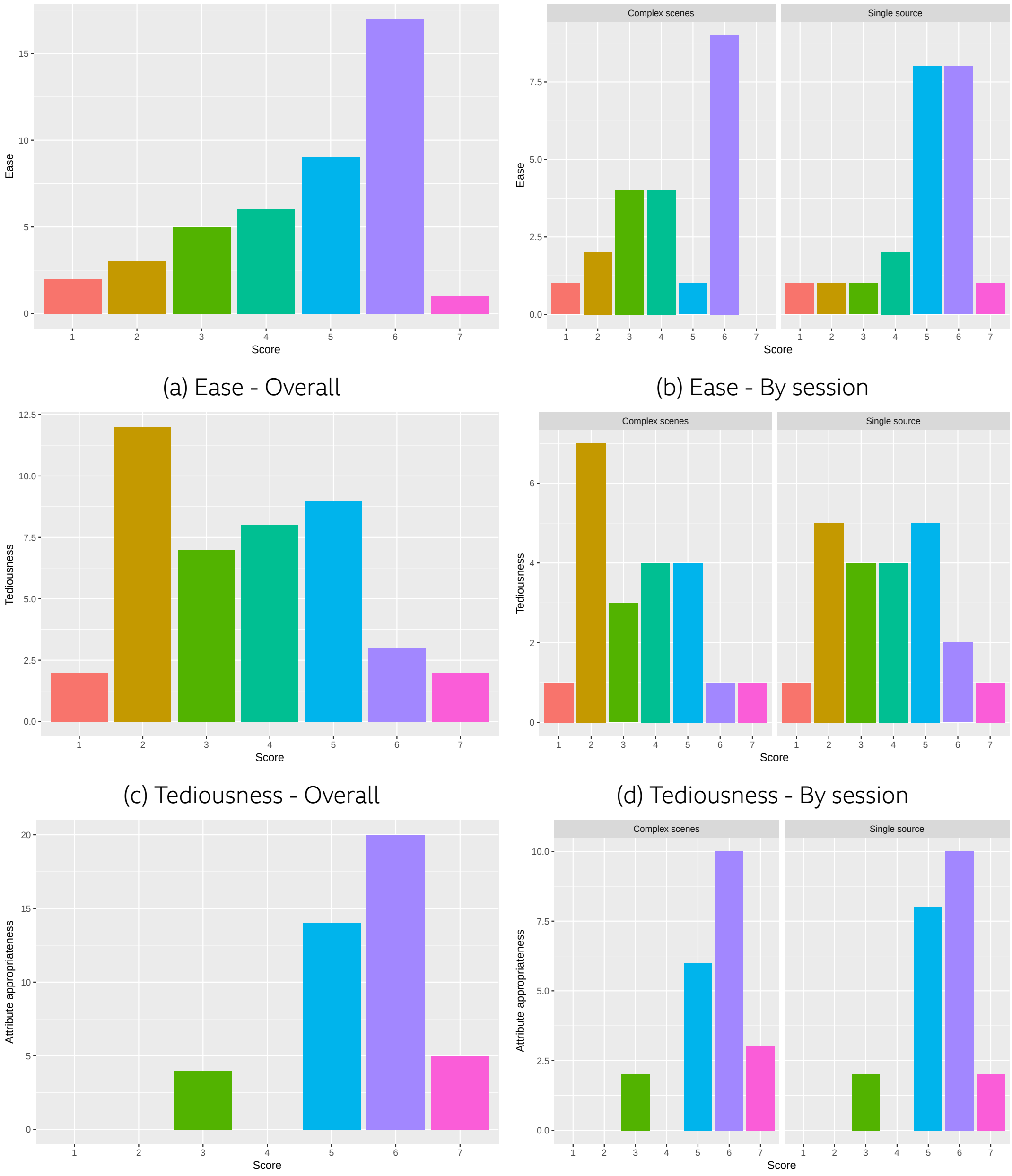
Table 8.14: Median values of head orientation data across trials for each assessor during rating.

8.5.6.1 Quantitative Results

The median responses to the questionnaire statements on the 7-point Likert scale, averaged across both item group sessions and all assessors, were: 5 - "*somewhat agree*" for *ease*, 4 - "*neither agree nor disagree*" for *tediousness*, and 6 - "*agree*" for the *appropriateness* of the attributes. Figure 8.29 presents histograms of the questionnaire responses, overall and split by item group session. There was a wide range of responses given to each statement. Friedman tests were carried out to test the effect of the session on each answer, but it was not significant for any statement (*ease*: $p = 0.285$, *tediousness*: $p = 0.467$, *appropriateness*: $p = 0.317$).

8.5.6.2 Free text responses

The assessors' free text responses were reviewed; the identified topics were (in order of popularity): complexity of the task, attribute relevance, choice/nature of audio items, lack of differences between stimuli, task context, relationship between overall sound quality and attribute rating, overall rendering quality, the role of a reference, fatigue, head movement, and the ITD calibration.



(a) Ease - Overall

(b) Ease - By session

(c) Tediousness - Overall

(d) Tediousness - By session

(e) Attribute appropriateness - Overall

(f) Attribute appropriateness - By session

Figure 8.29: Histograms of questionnaire responses.

Whilst assessors on average agreed somewhat that the task was easy, some assessors clearly found it difficult. It was reported by several assessors that differences between many of the stimuli were small, which made rating difficult, but the task was complex also due the nature of the stimuli and the detailed response format. Some assessors reported that the complex scenes were difficult because they found it hard to know what to focus on. For complex scenes, assessor 9 reported "there's so much going on, with multiple moving sources and head movement, that it's very overwhelming and difficult to make consistent judgements". This was reflected in comments by other assessors. However, it was also commented that these scenes presented more varied elements, better highlighting differences and allowing easier use of attributes. Amongst the full scenes, it was felt by some that simpler scenes were easier to evaluate than the most complex *Protest* scene. Two assessors explicitly stated that the single sources session was much easier, whilst another said they found the complex scenes easier. The large number of attributes was seen by several assessors as overwhelming and it was felt that they weren't all relevant.

From the Likert scale responses, it seems that the attributes were appropriate for characterising the stimuli for most assessors. Further insights can be garnered from free text comments of those who disagreed. Three assessors (1, 2, and 3) pointed out that some attributes were more relevant to a particular session (either single sources or complex scenes). Assessor 16 reported that the list of attributes was quite technical in nature and challenging for them to interpret. Assessor 7 mentioned that there was no "just right" option and that high and low (used for tone colour attributes) indicates a comparison to them. Assessor 9 stated that they found it difficult to describe the stimuli with the available attributes, particularly timbral differences between stimuli.

It was also commented that quality judgements are specific to the context, which was not always clear in the experiment. Assessors were not told what the artistic intent behind the audio scenes was, nor the envisaged situation in which they would be using such systems, which some found difficult when determining quality ratings. The use of relatively short clips also made it challenging for some assessors to evaluate their experience. Regarding the role of a reference, one assessor requested an explicit reference to make judgements easier. For the single source reference, one assessor thought it made them focus more on timbral aspects, whilst another commented that no stimuli had the same timbre as the reference. It was also mentioned that interpreting the visual position reference was difficult.

Regarding tediousness, opinion was varied. Those who found the single sources session tedious commented on the lack of variety in source material. With hindsight, using different items for each single source position would have given more variety for the assessors and allowed testing of a broader range of material. Fewer comments focussed specifically on tediousness, compared to ease/difficulty. Spearman's rank correlation test showed only a weak correlation between ease and tediousness ($\rho = -0.41$).

Several comments were made regarding the link between the attributes and the overall sound quality. Three assessors commented that they found the simultaneous collection of CATA data and overall sound quality ratings helpful, since the attribute characterisation could be used to guide the formation of their overall quality judgements. However, one assessor commented that switching attention between the two rating processes was difficult. One assessor found it useful to have attributes to explain, for example, that they felt two stimuli were as good as each other, but for different reasons. Another commented that the attributes could have positive or negative impact on quality, depending on the situation and that it wasn't possible to capture this within the rating process.

There were a few comments regarding the quality across all systems, e.g. that spatialisation at positions in front and behind was not particularly good in any of the systems compared with the lateral positions. Several assessors commented that no systems created a convincing impression of height. One assessor also mentioned that they felt they rotated their head more in the complex scenes session.

8.6 Discussion

This section discusses the results. The overall sound quality ratings are considered before then discussing the insights gained from the attribute data. Reflections are also given on the experimental methods.

8.6.1 Interaural Time Difference Scaling

The variance within the individual ITD scaling results suggests that this process was difficult for assessors and may not be suitable in future studies, or at least significantly more training would be required. Future studies should consider adaptive psychophysics methods such as a staircase procedure. Despite this difficulty, no assessors reported issues with source stability when checking after calibration (for the *OB* renderer), when all reported that they could hear unstable sources during training for ITD scaling.

8.6.2 Overall Sound Quality

It is clear from the results that the choice of content representation and associated binaural rendering technique can have a significant impact on the overall sound quality. The object-based binaural rendering method (*OB*), using HRTF convolution per sound object, provided the best quality overall. The first-order ambisonics rendering (*A1*) had poor quality, significantly lower than all other systems.

Adding more virtual loudspeakers does not necessarily increase overall sound quality. An increase in quality with ambisonics order was observed for single sources, but *A5* had lower quality than *A3* for complex scenes. The *V1* system was never rated significantly lower than any other virtual loudspeaker system, despite having eight virtual loudspeakers on the horizontal plane only. By contrast, system *V5*, with 32 virtual loudspeakers covering the full sphere, was rated lower than the other VBAP systems for *Forest*.

Results for single sound sources highlight that VBAP-based virtualisation techniques can have equivalent quality to per-object binaural rendering when a source is close to one of the virtual loudspeakers, but between loudspeakers degradations in quality can occur. This may often be the case. *V5* had lower quality for complex scenes than for single sources.

In terms of overall quality, there is no clear difference overall between VBAP-based and ambisonics-based systems. At the lowest complexity level, the *V1* system performs much better than the *A1* system. It should be noted, however, that the *V1* system requires eight real-time binaural filters whereas the *A1* system requires only four. This is because the decoder matrix can be applied to the filters rather than to the audio signals. Therefore *V1* might be more appropriately compared with second-order ambisonics, which would have nine real-time binaural filters.

The differences in system quality between the items highlight the importance of evaluating systems with a range of content. The stark difference in system rankings between single source and complex scene item groups indicates that there was value in evaluating systems with representative complex content, as well as more controlled stimuli. The character of the output of these dynamic binaural rendering systems will vary according to the spectral content of the input signals, the spatial location of the sources, the interaction between the sources and the listener movement. This large parameter space cannot be exhaustively probed in a controlled evaluation experiment. The use of complex sound scenes has allowed a wide range of content types and source positions to be assessed simultaneously, along with their interaction in terms of scene-level quality features.

Differences between item sessions may not be solely due to the scene complexity, as they also differed in terms of content type and use of a reference. The unprocessed timbral reference for single source items may have allowed assessors to be more critical of colouration in this case. For the complex scenes, assessors may not have perceived timbral differences as incorrect, having no audible reference timbre. If complex musical scenes were used, assessors may have been more sensitive to timbral colouration.

No rendering approach was rated as excellent. It may be that the system and content production could be improved upon, for example using HRTF personalisation, improving equalisation, higher-quality 3D reverb simulation, etc. However, the lack of a given reference may also have caused upper-end scale compression, as observed by Silzle, George, et al. (2011).

It should be noted that the results are somewhat specific to the rendering methods used in this test. These were chosen to reflect common practice, but will still be somewhat specific to the implementation used. There are also more advanced techniques for virtualisation, including time-frequency analysis-synthesis methods, which exist both for ambisonics (Laitinen and Pulkki, 2009; Berge and Barrett, 2010) and channel-based representations (Faller and Baumgarte, 2003; Goodwin and Jot, 2007). These rendering techniques may offer improved quality over the time-invariant virtualisation approaches used here. Other techniques have been proposed to improve binaural rendering with loudspeaker virtualisation. Jot and Noh (2017) describe the use of inter-channel decorrelation for improved reproduction phantom sources. For ambisonics rendering, Zaunschirm, Schörkhuber, et al. (2018) recently presented a technique based on frequency-dependent time-alignment of HRTFs followed by an optimisation approach to retain the diffuse-field response and the interaural coherence of the original HRTFs set. A listening experiment confirmed that this allowed closer approximation of the original HRTFs for ambisonics orders 1, 3 and 5, than the virtual loudspeaker approach[4]. Further investigation of the quality of these techniques is needed.

8.6.3 Quality Attributes

Whilst several of the virtual loudspeaker approaches have equivalent overall quality, they have different characteristics. The most frequently selected attributes over all stimuli (figure 8.18) indicate that, in general, systems were perceived with good spatial characteristics, but there were often colouration issues.

For some audio items, only a relatively small number of the 48 given attributes were significant. Since 20 of the 48 attributes were not significant on any item, it may have been possible to reduce the number of attributes presented and so ease the task of performing the test. Looking at the data, attributes *Front-Back Reversal*, *Imprecise Transients*, and *Unresponsive* could have been removed, as well as possibly vertical and radial extent and loudness attributes. All attributes were used at least once, however, and the significance of attributes varied by audio item. The CATA approach of providing a large set of attributes

[4] The *spatial resampling* approach described by Zaunschirm, Schörkhuber, et al. (2018) is equivalent to a virtual loudspeaker rendering with single-band ambisonics decoding.

allows the assessors to indicate what is relevant in each case, but this should be balanced against increasing task complexity.

Information is also gained from lack of significance. Position-related attributes were often not useful for discriminating the systems and were rarely used, suggesting that without a clear spatial reference or a task requiring it, localisation accuracy is not greatly important. On the other hand, good localizability (i.e. localisation precision) was the most commonly used attribute. Interestingly *Close* was frequently used but did not discriminate the systems, suggesting that the systems all created sources perceived close to the listener. Loudness differences were not a significant issue, even though the related attributes were used occasionally, which suggests that the loudness alignment process was effective.

Many attributes were significant for both audio item sessions, but there were some notable differences. Complex scenes allowed differentiation of systems in terms of scene-level characteristics such as width, envelopment and reverberance, as well as sense of presence, which were not significant for single sources. Single sources showed better distinction of systems in terms of externalisation and localisability.

The correspondence analyses showed that the first-order ambisonics system was perceived very differently to the others, with strong comb-filter colouration, bright tonal balance and poor localizability. The per-object binaural rendering was consistently characterised by clear, natural sound with particularly good localizability. The virtual loudspeaker systems were sometimes characterised by dark tone colour and lack of clarity, particularly those with more virtual loudspeakers and particularly for more complex scenes. However, at other times, virtualisation approaches were of similar character to the per-object binaural rendering. The effects are clearly dependent on audio content. For example, for the *Frontal* source system *A5* was most distinctly separated from other systems, whilst for many other items system *A1* was most distinctly separated.

The system *V1* showed remarkably good overall quality given its lack of complexity compared with other approaches. For complex scenes, particularly *Family* and *Forest*, it was characterised similarly to the *OB* system. The dummy loudspeaker re-routing approach may have been beneficial in this system; the panning will often have led to only two virtual loudspeakers being active for a given source. This system provides no elevated virtual loudspeakers. If individualised HRIRs were used in the rendering, it seems likely that there would be more substantial differences between *V1* and the *OB* approach.

The hierarchical cluster analysis on attributes showed that some semantically-related attributes were used similarly. This was particularly true for tone colour attributes. *Dark* was grouped with *Bass–High* and *Treble–Low*, and similarly *Bright* was grouped with *Bass–Low* and/or *Treble–High*. Since these are overlapping concepts (i.e. *Dark* = *Bass–High* + *Treble–*

Low) this is perhaps not surprising. If trying to reduce the number of attributes presented in future studies, experimenters may wish to avoid such closely related terms unless they are deemed particularly relevant to the study. Hicks et al. (2018) analysed the effect of grouping such attributes by summing the frequencies into a single column of the contingency table and repeating the CA. It was found that findings were largely unchanged, with similar locations of system profiles and levels of overlap of confidence ellipses, though for two items differences were slightly reduced. The authors conclude that semantically-similar attributes should be avoided and suggest that CATA might work best with a smaller coarser set of attributes. It should be noted though that their study was with inexperienced assessors unlike this study.

Attributes most strongly associated with liking could also be seen from the HCA due to the presence of *Like* and *Dislike* terms. Comb-filter colouration influenced liking negatively and natural sounding scenes with good localisability influenced it positively.

The penalty-lift analysis shows the relationship of attributes to overall sound quality, which appears quite intuitive. It can be seen clearly that liking has a strong relationship with overall sound quality. Unnatural colouration (*Unnatural*, *Metallic*, *Comb-filter colouration*) and poor spatial quality (*Internal*, *Poor Localizability*, *Envelopment–Low*) negatively influenced overall quality and high-level characteristics relating to a convincing spatial scene positively influenced overall quality (*Envelopment–High*, *Sense of Presence*, *Natural*). Tonal balance attributes did not have a strong effect, despite *Dark* and *Bright* being used frequently. This may be because they can be both good and bad in different contexts.

The attribute data gives insight into why the higher complexity virtual loudspeaker renderers did not perform better in terms of overall sound quality. This was due mainly to dark timbre and also the effect on speech intelligibility. It seems that most of the virtual loudspeaker systems were perceived similarly in terms of spatial character, but timbral character was an important differentiator. The causes for these colouration effects should be the subject of further investigation. It is likely due to the presence of coherent signals with different time-of-arrival at the ears caused by the virtualised panning. The nature of the timbral effects caused by panning appear be more noticeable with more closely spaced loudspeakers. A denser loudspeaker configuration will also lead to more rapid changes in panning gains during source movement. Appendix B.6 presents an initial analysis of the loudspeaker virtualisation systems used in this experiment in terms of representation of auditory cues, but further investigation is needed.

The argument could be made that the filter exchange (by cross-fade) in binaural rendering means that virtualised panning is still being used in the *OB* rendering case to some extent. The per-object rendering and the virtual loudspeaker approaches would converge as loudspeaker density increases, except that the former uses separate delay interpolation, and this reduces comb filter colouration by aligning the onsets. Where the object-based representation is to be rendered at the user-end, and where computation and memory resources allow, there may be benefits in inserting the onset delays separately and then performing virtualisation independently for the left and right ear signals. This approach was presented by Jot, Larcher, et al. (1995) and is discussed further in appendix B.7. Since the time-of-arrival for each ear is dependent on head orientation as well as source position, the delay insertion cannot be rendered prior to transmission if head tracking is used. This approach would mean that the system could not take advantage of the transmission/storage efficiency of pre-rendering objects to an intermediate bed signal.

When intermediate channel-based and scene-based formats are rendered as in this study, by loudspeaker virtualisation without separate binaural delay processing or any other more sophisticated signal processing, quality does not increase monotonically with loudspeaker density. There may be an optimum loudspeaker density, above which colouration gets worse, and so the required increase in computational resources and distribution bandwidth would be wasted. This is an important finding when considering which spatial audio signal representation should be used in an application. It is important to acknowledge that other technical factors within the application context will influence quality and so design decisions. For example, Narbutt et al. (2017) evaluated the quality of binaurally-rendered ambisonics after bit-rate reduction coding and found that bit-rate-per-channel was a bigger influence on quality than ambisonics order. First-order ambisonics achieved better overall quality than third-order ambisonics with lower total bitrate.

8.6.4 Method

The CATA method has proven an effective technique for characterising complex systems, when using experienced assessors. Insights into the nature of the rendering systems have been gained for a large number of attributes, showing significant differences between systems and interaction with content items. Experience suggests that a similar breadth of evaluation using traditional quantitative DA techniques would have taken significantly longer.

Despite the relatively long total duration of the experiment for assessors, it was felt to be easy and not tedious by most. However, it is clear that work should be done in future to reduce the complexity and duration of the test. One obvious way to achieve this is to reduce the number of attributes presented to assessors.

Hicks et al. (2018) used a method with some similarities to the pilot test (section 8.4.6) to assess the relevance of attributes to the stimuli under evaluation. It is noteworthy that their study provided a clearer, simpler interface and also collected scores of attribute understanding, as well as relevance, from assessors. This was used to select a subset of attributes from the initial set. Such a method may be advisable to assure data quality and reduce the effort required in the main experiment. But this pre-selection task itself can be laborious; Hicks et al. (2018) required assessors to make 28 attribute judgements across 5 stimuli for each of 5 audio items (700 judgements). Pre-selection may, however, be performed with a smaller panel of assessors than the main experiment. For well-understood study domains, the experimenters may be better able to narrow down the set of relevant attributes, though it seems advisable to use the input of representative assessors over experimenter judgement, especially with inexperienced assessors.

Another way to simplify the experiment task could be to change the overall quality rating process. Assigning a score on a continuous quality scale to a set of multiple complex stimuli is a challenging task in itself. A simplified response format such as a five-point quality scale may be sufficient and ease the load on the assessor. Quantising the 100-point overall sound quality data in this experiment to a five-point scale gave very similar statistical outcomes in analysis. Schoeffler, Silzle, et al. (2017) used a five-point scale to present the single stimulus rating of quality of overall listening experience for spatial audio systems. Whilst it appears likely that single stimulus rating would be less sensitive than a multiple stimulus comparison (without a reference), Schoeffler, Silzle, et al. (2017) found the results to be similar. Paired comparisons potentially present the assessor with a simpler task than absolute quality rating, particularly with multiple stimulus presentation, but this would be very time consuming with seven systems. Wickelmaier et al. (2009) proposed a method of ranking by elimination, where multiple systems are compared and the assessor successively eliminates the lowest quality system. This method was found to be much fast than paired comparisons, with equivalent accuracy.

One of the potential advantages of the method used in the present study is that assessors could simultaneously consider the CATA attributes and the overall quality ratings. Some questionnaire responses indicated that this approach was helpful and it would appear to make the link between characteristics and quality ratings explicit. However, separating these two steps, as in conventional CATA, might be another way to simplify the task for the assessors.

The lack of declared reference in this study presented a challenge for some assessors. With relatively short sample lengths and a lack of visual input, the assessors had little context for the dramatic content. Forming a judgement of how it *should* sound was difficult for some. For this context, the assessors would need more information about the material, given in advance of the experiment or using longer excerpts, during the evaluation. With longer excerpts the present method, requiring analytical comparison of multiple stimuli, is likely to become more challenging. One of the alternative presentation paradigms described above might be more appropriate in this case.

Since the production of the complex scene items was conducted using the *OB* system, this will have biased results in favour of that system. If a sound engineer were monitoring the output on other systems, it is likely that at least some of the deficiencies could be compensated for during production. An object-based workflow is, though, based on the assumption that the reproduction might differ from the production system, e.g. due to a different reproduction layout or rendering algorithm; in such cases the reproduction system should adapt the signals appropriately. The benefits of such a workflow only come when the content creator can trust that the original experience can be reproduced faithfully by the system, without manual checks and adjustments. However, when the range of target systems is distinctly different, the precise perceptual character cannot be preserved and instead the system should somehow aim to represent the creator's intention. This concept presents new challenges for content creators and system designers, as well as for researchers trying to evaluate system performance.

As discussed in section 8.6.3, the results of this study show that, whilst the impression of overall quality may not be different between systems, the character of the systems can be significantly different. This highlights the value of the descriptive analysis method for complex systems. New methodological challenges arise when the systems under evaluation are all distinctly different from the original production system, e.g. testing headphone rendering of an original production for a 3D cinema sound system. In this case, a declared reference would not be appropriate and the original system could not be easily included as a test stimulus. Reference-free evaluation measures against the assessors' own expectations, with no information about the original production and the creator's intentions. It could be that no system created an impression precisely in line with what the content producer intended, or that some did and some did not; the method provides no such insights in that case.

Whilst the statistical techniques used here appear to have drawn sufficient detail from the data, there may be benefits to exploring further analysis methods. Parente et al. (2011) applied multiple factor analysis (MFA) to reflect the categorisation of attributes, providing equal weight to each. This may be beneficial here to balance the weight between timbral and spatial characteristics, though it may also put undue emphasis on the smaller categories such as *room* and *time behaviour*. Parente et al. (2011) also used external preference mapping techniques to relate CATA data to consumer preferences. It would be interesting to compare partial least-squares regression results to those of the penalty-lift analysis; though the simplicity and clarity of the penalty-lift analysis seems to present an advantage.

There could be some benefit in exploring human factors within the quality evaluation. Giacalone, Bredie, et al. (2013) presented the "all-in-one" method, which combined CATA sensory profiling with hedonic ratings, assessor demographics and psychographics, as well as ratings for a perceived ideal product. A partial least squares regression model was used to analyse correlations between the multiple different data sources. These techniques are likely to be most useful for large-scale exploratory analyses with inexperienced assessors, where greater heterogeneity in responses can be expected.

Walton and Evans (2018) demonstrated correlation between listener psychographic variables and ratings of the overall listening experience (OLE) for different versions of music signals. This included binaural, stereo and mono versions, as well as low-pass filtered stereo. The binaural signals were dedicated binaural mixes from multitrack source material, using the system described in appendix A of this thesis. Following Schoeffler and Herre (2014), correlation in ratings was used to determine the influences of technical quality and content preferences on OLE and these were subsequently correlated with psychographic variables. A number of psychographic variables were significantly correlated to these influences, particularly a self-reported measure of technical competence related to the influence of technical quality on OLE. Further development of such methods could be highly valuable in understanding listener needs and desires, and adapting audio services to meet them.

Recent work has carried out subjective evaluations of spatial audio rendering directly in audiovisual VR environments (Poirier-Quinot and Katz, 2018a,b; Rummukainen, Robotham, et al., 2018). When specifically targeting such applications, it appears important to evaluate audio systems within VR applications. Rummukainen, Robotham, et al. (2018) compared rating within audiovisual VR viewed in a headset with 6DoF tracking, to rating on a 2D visual display screen with a predefined motion path, and found the VR presentation to give higher discriminability between audio systems. It is also hypothesised in this paper that evaluation of spatial audio rendering in a different usage context from that of the intended application may bias the listener's attention to features that are not relevant in the target context, namely a multimodal audiovisual experience which responds to motor actions.

Rummukainen, Wang, et al. (2018) observed significant differences in quality ratings in a virtual environment with 6DoF tracking, dependent on whether the visual scene information was displayed or not. The importance of different attributes is likely to be different again in AR applications, where congruence with cues from the real-world environment is also required. This topic is as yet little researched, but recent work has begun to explore the requirements, e.g. Brandenburg et al. (2018). The method used in this chapter, multiple-stimulus quality rating with CATA characterisation, is likely to be practically-challenging in audiovisual rendering environments.

In the recent paper by Vidal et al. (2018) which compared the RATA and CATA methods, the authors concluded that RATA is more effective when the stimuli all share characteristics but at different intensities, and that CATA should be used where the attribute either applies or does not apply: "that might provide a stronger argument in favor of using CATA than just that RATA is more labour-intense, which only becomes a convincing justification when otherwise fatigue might hit in and decrease data quality." Many of the attributes considered in this study are really a matter of degree/intensity rather than binary. It seems obvious that rating using ordinal or ratio scales will give better distinction, but in exploratory studies with a large set of potential attributes, test duration and fatigue are an important consideration. When working with experienced assessors, the CATA method may be most suitable for an initial broad characterisation, to indicate relevant quality features on which to investigate systems in more detail using continuous rating scales. However, in this case, it has already revealed useful insights into the systems. Whilst the methods may each be better suited to different types of attributes, Vidal et al. (2018) concluded that RATA does not necessarily give better sample discrimination or attribute intensity measurement than CATA, and that the outcomes of the two methods were found to be "very similar" across the seven studies that they conducted.

To summarise, when evaluating specific techniques with a well-defined objective, e.g. rendering of extended sources, having experienced assessors perform ratings on quantitative scales for a select few attributes that are known in advance will likely give detailed insights. CATA responses could provide only coarse information in terms of these specific attributes. In other cases, simpler evaluation tasks that are more representative of target applications and listening contexts, using inexperienced assessors representative of target users, may give more insight into audience benefits and/or acceptance of new technology. But, in situations where exploratory analytical evaluation needs to cover a wide parameter space to give insights into a set of technologies, the CATA method shows strong promise.

8.7 Conclusions

Loudspeaker virtualisation is often used in binaural rendering applications as a practical means of limiting transmission data and rendering complexity. A listening experiment was conducted to evaluate the sound quality of such techniques compared to per-object rendering of a scene. A custom software system allowed direct comparison of seven dynamic binaural rendering approaches relating to different spatial audio representations. Virtualisation approaches based on ambisonics and VBAP were used, with varying and approximately equivalent levels of complexity. Experienced assessors rated the overall sound quality of these methods for three single guitar sources at different positions and three complex audio drama scenes. Assessors simultaneously gave a characterisation of the quality of the systems using the check-all-that-apply (CATA) method, considering a pre-defined set of 48 attributes with a simple binary “applies” or “does not apply” response.

The choice of method for representing and rendering spatial audio scenes can significantly affect sound quality. The CATA data showed significant differences amongst systems, including cases where the overall sound quality was equivalent. Single sources and complex scenes gave different results in terms of overall quality. Additionally, the use of both types of material allowed appropriate evaluation of low-level perceptual characteristics of sources as well as scene-level and experiential characteristics. The influence of these characteristics on overall sound quality was also analysed.

Overall, the per-object HRTF convolution approach had better quality than all virtualisation techniques. First-order ambisonics rendering had significantly lower overall quality than other approaches. The highest-quality, per-object rendering approach gave clear, natural impression with good spatial characteristics and was most liked. First-order ambisonics had unnatural bright tone colour and poor localisability and was most disliked. The other virtualisation techniques were equivalent in terms of overall quality when averaged across all audio items, but there was a strong variation across the audio items. The character of these systems was content-dependent. Interestingly, the ambisonics and VBAP systems with a denser configuration of loudspeakers showed quality degradations on complex scenes, compared with sparser, less complex options. The CATA data showed that they sometimes had dark tone colour and lacked clarity. It appears that, with the rendering approaches evaluated, the quality may not increase monotonically with the detail of the scene representation. This is an important finding when considering the design of applications. Further investigation of the nature and causes of these effects is required. Initial objective analyses are given in appendix B.6.

The method of multiple stimulus quality rating with CATA questions enabled exploratory characterisation of a range of complex systems, using a range of audio material. On average assessors did not find the task too difficult or tedious, but it is clear that it could be simplified further. The associated analysis techniques such as correspondence analysis (CA) and penalty-lift analysis provide a relatively straightforward means for extracting detailed interpretable information from the data. It is hoped that CATA methods will be used more widely for efficient characterisation of audio systems in future, perhaps to guide towards specific aspects that need further in-depth investigation.

Besides technical and cost constraints, choice of spatial audio signal representations and binaural rendering techniques should be based on an understanding of quality impact. This should be evaluated using appropriate content and within the appropriate context of use. Future work, beyond the scope of this thesis, will explore these challenges further (see section 9.3).

Chapter 9

Conclusions

The aim of this thesis is “to improve the quality of headphone listening experiences for entertainment media audiences by developing and evaluating binaural technology” (section 1.2). After an initial study to assess the status quo, work focussed on the implementation of technical apparatus and the development of experimental methods for investigating the perceived quality of binaural technology. A review of the scientific understanding of perceived quality (section 3.2) highlighted that there are many different factors that can influence the quality of an experience. In fulfilment of the aim, this work has attempted to identify and study specific aspects that are of strong relevance to the domain of application. This work has included:

- A review of the state-of-the-art in binaural technology as used in entertainment media.
- A review of the state-of-the-art in relevant quality evaluation methods.
- The development of tools for comparing state-of-the-art binaural rendering techniques and for applying them in media production.
- A criterion-free evaluation of the plausibility that can be achieved when using non-individualised dynamic binaural-rendering in a small listening environment.
- A web-based evaluation of the preferences of the target audience between static non-individualised binaural signals and stereo signals for headphone listening.
- Characterisation of the effects on quality of virtual loudspeaker rendering across different binaural rendering approaches using a descriptive analysis method.
- Characterisation, using the check-all-that-apply (CATA) method, of the quality of head-tracked binaural rendering using a range of formats for representing 3D spatial audio scenes that are relevant to mass-audience services.

To conclude this thesis, the approach, results, and contributions are first discussed in section 9.1. Section 9.2 then answers the research questions posed in section 1.2 by reflecting on these outcomes. Finally, section 9.3 describes possible routes for further research.

9.1 Summary of Findings

This section summarises the findings from the preceding chapters of this thesis.

9.1.1 Review

In chapter 2 a detailed review of binaural technology was presented. Binaural sound systems aim to create convincing auditory events by precisely controlling the acoustic pressure signals that reach the eardrums. Based on an understanding of the processes of spatial hearing, binaural rendering techniques have been developed and are capable of creating simulations that are indistinguishable from real sound events. This degree of realism requires highly controlled conditions however, including *in situ* acoustic measurements specific to the individual listener.

In creative applications targeting a mass audience, practicality and flexibility are important, and a lesser-degree of realism may be sufficient. Many studies have explored the robustness of binaural rendering in less controlled scenarios and have developed engineering techniques to improve performance. The review suggested that, with careful design, a non-individualised binaural rendering system might still be able to provide a convincing spatial impression. But there was not clear evidence to suggest that this can improve the overall headphone listening experience in entertainment media applications compared with established stereo services.

Chapter 3 presented a discussion of perceived quality and the experimental methods for evaluating it, with a particular focus on the quality of spatial sound reproduction and binaural technology. There are a range of different methods. Descriptive analysis methods aim at analytic evaluation of the perceptual characteristics of technology systems, often making use of expert listeners. Other methods are suited to evaluating the quality of experiences of target users with applications of the technology, relating to their delight or annoyance in light of their expectations and needs. Discrimination methods explore small perceptual differences close to the threshold of detection and are useful for validating engineering approximations. In combination, these methods can be used to gather in-depth understanding of the perceived quality of binaural technology and of how to apply it appropriately to improve headphone listening experiences.

9.1.2 Checking the Status Quo

Chapter 4 assessed the status quo at the outset of the project. A quality evaluation experiment was carried out to investigate whether commercially-available binaural rendering systems were capable of improving the headphone listening experience, using existing broadcast programme material in the 5.1 surround format. This primarily considered application of headphone surround sound processing (HSSP) prior to distribution of the headphone signal, rather than use of client-side processing which would require changes to consumer technology.

Previous studies have found that such HSSP systems do not provide substantial benefits over a conventional stereo down-mix, often showing significantly worse quality e.g. Lorho and Zacharov (2004). This experiment extended previous work by presenting accompanying video, making it more representative of the target use case. The effect of the listening environment on quality was also studied, using two different test facilities. The results of this pilot study confirmed earlier findings, indicating that there have not been major advancements in technology.

Included in the study was a system that made use of binaural room impulse response (BRIR) measurements made on the individual listener in the experiment environment, as well as using head tracking. Earlier evaluations had only used static non-individualised systems. Despite this system being unsuited to the primary target application scenario, it was included to provide a point of reference. According to the review of chapter 2, this system should provide convincing spatial impression. Informal listening confirmed this, yet it did not give an enhancement of quality in this experiment.

These results suggest that realistic simulation of loudspeakers in the listener's environment may not be the optimal target for rendering 5.1 surround content to headphones. Although, from this experiment, it cannot be said how realistic the spatial impression was to listeners. The results for simple test signals and from a cluster-analysis showed that, in some cases, large improvements can be made, at least for a subset of people. A significant dependence on audio source material was also observed. This study prompted further investigation of spatial impression afforded by binaural rendering systems and the various influences on the quality of the listening experience.

9.1.3 Spatial Impression and Plausibility

In chapter 5 the realism of binaural simulation was directly addressed. A non-individualised dynamic binaural rendering system was constructed so that the signal processing techniques were precisely known. Since individualised rendering is not yet feasible for mass-audience applications, BRIRs were measured using a dummy head microphone in the experiment environment. An experiment was run to evaluate the plausibility of the binaural rendering system, following the study by Lindau and Weinzierl (2012). A sound was played to the assessors from either a real loudspeaker or the headphones and they were asked "did the sound come from the headphones?", giving a forced-choice *yes/no* answer.

It was found that this system gave simulations that were largely in accordance with listeners' expectations of real events, though occasionally it deviated from those expectations. Use of signal detection theory (SDT) models allowed separate analysis of sensitivity to perceived differences and the listener response biases. There was no significant bias towards either "yes" or "no" responses. The sensitivity was equivalent to an average detection rate of 55.49 % in a two-interval two-alternative forced choice (2AFC) test, where 50 % would indicate perfectly-random guessing.

Listeners confirmed in comments that they found it very challenging to identify the headphone rendering. Slightly increased sensitivity was observed for elevated source positions and the overall observed sensitivity was slightly higher than that observed by Lindau and Weinzierl (2012) for their "improved simulation", but lower than their other system. That study was conducted in a much larger environment with loudspeakers more distant; it also used a head and torso simulator (HATS), whilst here only a dummy head microphone was used.

Despite the lack of individualisation, this experiment provides evidence of highly realistic rendering of external sound sources for headphone listeners. The individualised rendering system used in the experiment of chapter 4 was found informally to be at least as realistic, yet for reproducing 5.1 surround content it was not considered to be of good quality overall. This prompted further investigation of the influences of system and content factors on perceived quality.

9.1.4 3D Spatial Audio Applications

In chapter 6 the influences of the content production techniques and associated formats for representing spatial audio scenes were discussed. First, the approach of distributing prerendered non-individual binaural signals was investigated further. Two web-based studies were reported. They both showed significant listener preferences for binaural signals over stereo versions of programmes.

In one study, a dedicated binaural mix of an audio drama was created, using apparatus constructed during the work of this thesis. Professional sound engineers used a mixture of head-related impulse response (HRIR) and BRIR rendering, along with a 3D parameteric reverb effect and a stereo panning option. Components of each scene were processed differently to achieve the desired effects and maintain high quality. This production approach has now been applied many times in programme production at the BBC.

In a second study, HSSP was applied to an audio drama in the 5.1 surround sound format. This programme was distinctive in that the dialogue was frequently routed to a single loudspeaker channel, rather than making use of amplitude panning, and the surround channels were used more heavily for foreground scene elements than is common in broadcast production.

These results indicate that through the appropriate combination of content and rendering techniques, binaural technology can improve the listening experience, even without head-tracking and individualisation. Whilst these studies lacked control over participant selection and listening conditions, they are more representative of the target application than laboratory studies.

Chapter 6 continued by discussing recent developments in spatial audio applications and associated technologies. Use of 3D spatial audio is becoming more widespread, due to standardisation of so-called “next-generation audio (NGA)” systems for coding and distribution of 3D spatial audio content, and the rapid growth in popularity and maturity of virtual reality (VR) and augmented reality (AR) systems. These technologies make it feasible to provide interactive client-side binaural rendering of 3D scenes to large audiences. However, this also applies constraints to the system design. Content delivery bandwidth and the computational requirements of rendering are both limited, especially considering reproduction on mobile devices.

Three common approaches to representing spatial audio content are used in these applications: object-based, channel-based and scene-based formats; they are sometimes used in combination. The format in which the spatial audio content is represented should ideally give optimal quality within available resources. Therefore, the influence of the content production techniques and formats and their interaction with the binaural rendering approaches on quality must be considered. Loudspeaker virtualisation is widely used, yet there is little published scientific evaluation of the effects it has on the quality of headphone listening. The earlier findings in the thesis suggested that this aspect was worth further investigation.

9.1.5 Investigating Loudspeaker Virtualisation

A listening experiment was conducted to characterise the effects of virtualised amplitude panning on perceived quality and this was presented in chapter 7. Direct binaural rendering of an audio signal at a target source position was compared to a virtualised pair-wise panning, where the signal was distributed to two virtual loudspeaker sources for rendering. This was explored over other system factors: use of head tracking, use of environmental acoustics in the rendering, and source positions. The experiment made use of a pre-existing set of sound quality features for auditory virtual environments (AVEs), the spatial audio quality inventory (SAQI) (Lindau, Erbes, et al., 2014). A rate-all-that-apply (RATA) approach was used, whereby assessors only rated attributes that they perceived relevant (Ares, Bruzzone, et al., 2014).

Significant overall differences between the direct rendering and virtual loudspeaker approach were observed for all system configurations. Analysis of the quality features indicated that colouration was the most prominent effect of using loudspeaker virtualisation. However, a wide range of spatial features were also affected: distance, externalisation and localisability of auditory events were all decreased, whilst the extent was increased. The horizontal and vertical direction of events was also modified.

Interactions with the binaural rendering approach were observed. Distance and externalisation were more affected when using HRIRs, whilst the added colouration was more apparent when BRIRs were used. The perceived horizontal direction was changed when the target source direction and virtual loudspeakers were in lateral positions. Changes in vertical direction were more apparent with BRIRs.

Whilst using only 10 assessors, the results of this study suggest that amplitude panning of sources might cause significant quality issues in loudspeaker virtualisation applications. Even if plausible impression is achievable for non-individualised direct binaural rendering, it seems likely that the observed issues will negatively affect this plausibility when amplitude-panning and loudspeaker virtualisation are used together.

9.1.6 Characterising the Quality of Spatial Audio Formats

Despite these apparent issues, loudspeaker virtualisation techniques are widely used in binaural rendering applications. Characterising the quality of different options allows more informed decisions to be made regarding the design and use of services and technologies. Chapter 8 presented a listening experiment to characterise the quality of seven representative approaches to delivering binaurally-rendered spatial audio signals in consumer applications with head-tracking. An object-based approach to representing scenes as component sound sources was compared with various channel-based and scene-based representations of different complexity levels. Novel experimental apparatus was constructed to allow direct comparison of these approaches. To explore different features of the quality of experience when listening to rendering of these formats, the study made use of both single musical sound sources and complex 3D dramatic scenes. This programme material was defined using the Audio Definition Model (ADM) standard (ITU-R, 2017b).

The original object-based scenes were converted to channel-based and scene-based representations using vector base amplitude panning (VBAP) and ambisonics techniques, respectively. First-, third- and fifth-order ambisonics representations were used, these were rendered by a dual-band decoding technique with virtual loudspeaker positions at points on a spherical t-design suited to correct decoding at the respective ambisonics order. Head tracking rotations were applied in the ambisonics domain as is common in applications of this rendering method. The channel-based representations were designed to be of approximately equivalent complexity levels, whilst also using virtual loudspeaker positions that corresponded better to standardised layouts. Channel-based representations are used in NGA systems to provide audiences with 3D spatial audio for both loudspeaker and headphone reproduction. For these representations, head tracking rotations were applied by dynamic updates to the head-related transfer functions (HRTFs) used for loudspeaker virtualisation.

To obtain an efficient characterisation of this wide range of options, the CATA method was used (Meyners and Castura, 2014). Listeners were presented with a pre-defined set of sound quality features, again based on the SAQI, and used a simple binary response format to indicate those that influenced the overall quality rating for each stimulus. This allowed for an exploratory approach, assessors were able to describe the stimuli with a wide range of potentially relevant terms, rather than take a reductive approach of pre-defining a small set of attribute scales. There was no explicit reference of target quality given in this experiment, assessors rated the quality of stimuli by considering their expectations of the desired characteristics.

The choice of representation format and associated binaural rendering approach had significant effects on overall quality. Binaural rendering of the object-based scene representation by separate HRTF convolution for each source, was found to have significantly higher quality than all of the loudspeaker virtualisation methods. The first-order ambisonics approach showed particularly poor quality. The other approaches were not well distinguished in terms of overall quality when averaged over all content items, despite using very different scene representations. As with earlier studies, the audio programme material was found to be a significant influence on quality, the system ranking differed between items.

The CATA characterisation data showed significant differences amongst systems, even where their overall quality was rated similarly. The object-based approach gave clear, natural impression with good spatial characteristics. First-order ambisonics had unnatural tone colour and poor localisability. The character of other virtualisation systems was content-dependent, but higher-complexity systems with denser loudspeaker arrays sometimes had dark tone colour which resulted in lack of clarity. It appears that, with the rendering approaches evaluated, the quality may not increase monotonically with the detail of the scene representation, e.g. with the number of loudspeakers used in the virtualisation. This is an important finding when considering the design of applications.

The use of both single musical sources and full drama scenes allowed appropriate evaluation of low-level perceptual characteristics of sources, as well as of scene-level and experiential characteristics. The relationship of these characteristics to overall sound quality was also analysed, giving intuitive results. Characteristics relating to unnatural sound colour and poor spatial impression had negative impact on quality, while attributes relating to a convincing experience of the overall scene led to high quality. Liking attributes (*like*/*dislike*) showed a strong relationship to quality ratings, though for proper indication of the impact of these formats on quality of experience, evaluation with target users in a representative application should be performed.

The CATA method allowed a wide range of quality features and influencing factors to be explored in a practical manner, whilst also revealing significant differences in the characteristics of systems. A post-experiment survey was conducted and, on average, the assessors reported that they did not find the process too difficult or tedious. However, it is expected that the process could be further simplified and still achieve useful results. The CATA method, when combined with objective analysis of system behaviours, appears to be a valuable tool for directing improvements to the design and application of spatial audio and binaural rendering systems.

9.2 Discussion of Contributions and Research Questions

This thesis has shown that the perceived quality of binaural technology has many influencing factors and is formed by judgement of many characteristic features. The application of binaural technology to entertainment media has become increasingly popular over the course of this work and can take a wide variety of forms. The studies herein have primarily explored system-related influences on quality, considering the interaction between binaural rendering methods and programme content representations. The studies have been shaped by the contextual factors of entertainment media applications and the constraints that they impose on the use of binaural technology. Though not the main focus of study, consideration has also been given to the central role of the listener in evaluating perceived quality and the influence of experience and expectation on their judgements.

The technical apparatus created during this work has provided the flexibility needed to investigate the interaction of state-of-the-art binaural rendering techniques with spatial audio content representations, including the use of representative complex and dynamic programme content. A range of experimental methods for evaluating perceived quality has been carefully applied to deliver an improved understanding of the applications of binaural technology. This work has also led to the creation of tools suitable for investigating these complex challenges further.

In section 1.2 three *research questions* were posed:

Research question 1

> *Is binaural rendering capable of producing a convincing spatial impression without calibration for the individual listener?*

Chapter 5 demonstrated that a non-individualised dynamic binaural rendering system can provide simulation of loudspeakers in the listening environment that is largely in accordance with listeners' expectations of real events.

Research question 2

> *Can binaural rendering improve the perceived quality of the headphone listening experience in entertainment media applications?*

Chapter 4 found that binaural rendering did not provide substantial improvements in quality over stereo signals when applied to typical broadcast programme material in the 5.1 surround format. Chapter 6 showed that different approaches to content production *can* lead to improved quality, resulting in preference for binaural signals by target audience members.

Research question 3

> *How does the programme production process influence the perceived quality when binaural rendering is applied?*

The programme producer often has a choice of the spatial audio representation and rendering processes that are applied. For distribution of pre-rendered binaural signals, where no head tracking is applied, a wide range of intermediate formats and processes can be used, since they are all rendered to binaural in the production process. Through the development and application of the production system described in section 6.2.2, it was found that the best choice of binaural rendering method is often content-dependent. 3D binaural production with choice of the binaural rendering approach used for different scene components gave good results, as found in the web-based preference test presented in section 6.2.3.

For client-side binaural rendering applications, the spatial audio representation used for distribution may be determined by the platform or application being used, but often, as with NGA systems, there is some flexibility in the representation used. The choices of representation and rendering are therefore seen as a part of the programme production process. These choices have been shown to create significant variations in perceived quality. Chapter 6 demonstrated the prevalence of loudspeaker virtualisation in entertainment media applications of binaural technology. The influence of loudspeaker virtualisation on quality was explored in depth in chapters 7 and 8, where it was found to have negative effects when compared with direct binaural rendering of sound sources in the scene. In many practical situations, however, virtualisation techniques are required due to technological and economic constraints. Chapter 8 showed that perceived quality evaluation techniques developed in other fields can be used to optimise design decisions in light of such constraints.

9.3 Further Work

There are several directions for further work following on from this thesis, to which the apparatus and techniques presented may be applied:

9.3.1 Improving the Quality of Loudspeaker Virtualisation

The studies in chapters 7 and 8 showed that many quality features were negatively impacted by the loudspeaker virtualisation approaches investigated. Further objective analysis of these approaches, using techniques such as those described in appendix B.5 and in light of the results of these perceptual studies, should indicate aspects of the virtualisation process that could be improved.

Appendix B.7 presents some more advanced techniques for virtualisation which attempt to improve quality. These include methods to improve time-alignment, use of decorrelation, and time-frequency analysis-synthesis approaches. There is little available evidence on the quality of these techniques, though it seems likely that they will provide improvements. Implementation of these approaches and integrating them within the experimental apparatus developed in this thesis will allow comparative evaluation of the quality they provide.

9.3.2 Study of Additional Content and Rendering Factors using the CATA method

Chapter 8 investigated the formats for representing spatial audio and the associated rendering techniques. There are other factors in content production that can influence quality. An important aspect to consider is methods for recording the spatial information in real-world scenes. During the course of this thesis, this author has performed several investigations of microphone techniques for spatial audio production, which have led to the open dataset described in co-authored publication Co.P. IV (Bailer et al., 2015) and a spatial audio drama that was produced simultaneously for headphones and surround-with-height loudspeaker reproduction (BBC Radio 4 - Drama, 2016).

Millns and Lee (2018) recently investigated microphone techniques suited to dynamic binaural rendering in VR applications. A pre-defined set of four quality features were used for rating scales. The CATA method is well-suited to characterising the quality of recording techniques, allowing exploration of an expanded range of perceptual effects that different techniques may have. This author was recently industrial supervisor for a MSc research project by Sproule (2018), where the experimental apparatus for chapter 8 was applied to evaluation of a range of microphone arrays. Again, single sound sources and complex scenes were both evaluated: measured room impulse responses were used to create the single source stimuli, and recordings of natural sound scenes (a park soundscape and an orchestral performance) were made simultaneously with the microphone arrays. The results of this small-scale pilot test give initial insights into the different characteristics of several 3D spatial audio recording techniques when applied to binaural rendering applications, but the study could be further expanded.

9.3.3 Study of Plausibility in Virtual and Augmented Reality Applications

The method of evaluating the plausibility presented by Lindau and Weinzierl (2012) and used in chapter 5 shows great promise for validating AVEs. However, it is limited to indirect comparison to a real loudspeaker source in the listening environment. The listener is only permitted to rotate their head horizontally and the evaluated systems have used a dataset of BRIR measurements for the specific listening room. In AR applications, more flexibility in listener interactivity and environmental conditions is required, so different rendering approaches must be taken. Recent studies have shown promising developments in such techniques (Brandenburg et al., 2018). The “real or virtual” test, using criterion-free analysis, is still a highly suitable validation method for AR applications.

For virtual reality systems, the same method cannot be applied. No real-world stimulus can be presented as the world is virtual. In this situation, plausibility is an appropriate success goal. Chapter 5 discussed problems with asking listeners to directly judge the plausibility of an experience, or other experiential factors such as sense of presence and immersion. It cannot be known what internal criteria are used for this judgement, there is no evidence of the actual brain activity that occurs during the experience (Slater, 2004). New experimental methods are needed to evaluate plausibility and associated factors, such as immersion and presence, without requiring a corresponding real world event and without requiring listeners to directly report these aspects on rating scales.

9.3.4 Studying the Overall Listening Experience of Binaural Technology Applications

Besides the web-based studies reported in chapter 6, there remains a surprising lack of evidence of the positive influence of binaural rendering technology on overall listening experience (OLE). Walton and Evans (2018) investigated the OLE of static non-individual binaural mixes of music, created using the apparatus described in this thesis, in comparison to stereo and mono mixes. Binaural rendering had a negative impact on OLE. The role of listener psychographic and attitudinal factors was studied to gain deeper insights. These methods could be applied to a wider range of programme material. The apparatus constructed during this thesis has been used often in productions at the BBC. As described in co-authored publications Co.P. VI and Co.P. X for example.

Besides the work of Walton and Evans (2018) and the studies in chapter 6, there has been no further comparative scientific evaluation of this material with the standard stereo versions. Additional studies of the OLE of binaural rendering applications should focus on ecological validity, considering human and context influence factors carefully in experimental design. Further evidence is required to provide a case for introducing binaural technology into services.

Alongside OLE evaluation, it would be of interest also to apply the rapid characterisation methods developed for sensory evaluation of food by consumers, which are described in section 3.3.3 and include the CATA approach. A potential drawback of OLE evaluation is that it does not indicate the features of the experience directly, development of efficient quality characterisation methods for use with target audience members is of great interest.

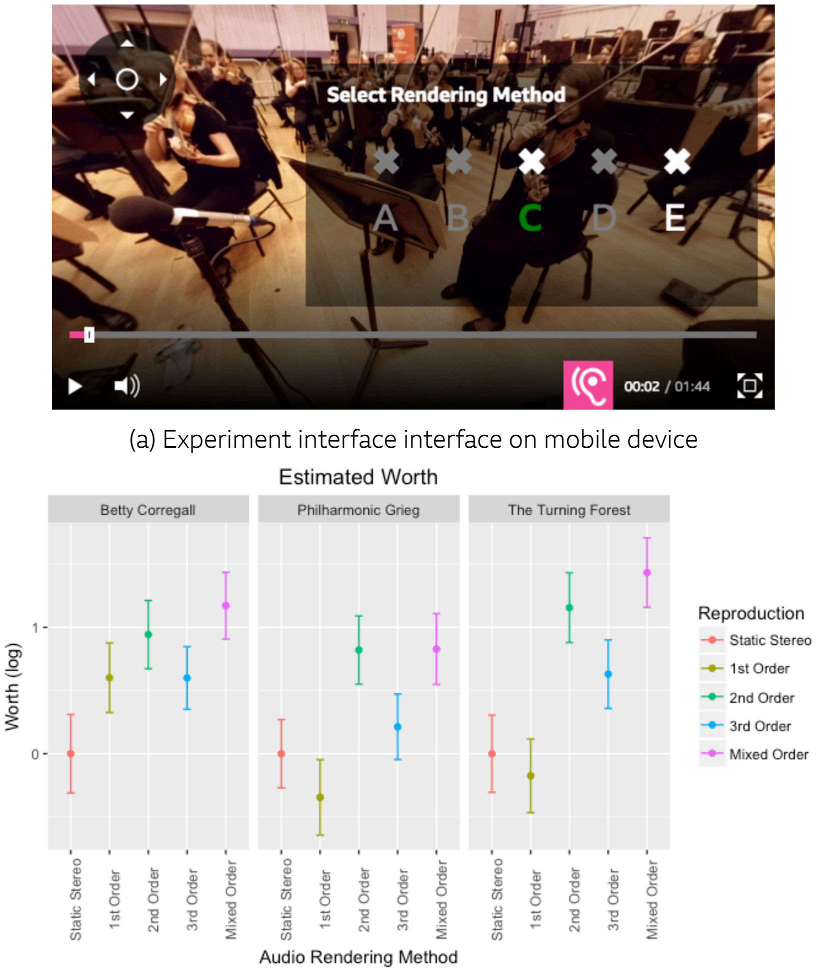
It is important to evaluate overall experience in the appropriate application context. Chapter 8 used dynamic binaural rendering without accompanying visual signals, whereas virtual reality systems most often present audiovisual content. Rummukainen, Robotham, et al. (2018) published a study that clearly demonstrates the benefit of dynamic binaural rendering over static stereo in audiovisual virtual reality systems with 6DoF tracking. A recent unpublished study by this author and a colleague has implemented ambisonics-based dynamic binaural rendering in a web-based 360° video player for mobile devices, in order to study the influence of spatial audio on the overall experience in this context (Pardoe and

Pike, 2018). This uses the same rendering techniques developed for the study of chapter 8 and follows the experimental method of Rummukainen, Robotham, et al. (2018), using a ranking-by-elimination procedure first introduced by Wickelmaier et al. (2009). The interface and initial results are shown in figure 9.1. Further analysis is required, but this already shows promising results for binaural technology and, as with chapter 8, it is helping to inform design decisions in the presence of many options.

9.3.5 Industry Standardisation of Techniques

Through this author's role in industry bodies, there may be a route to standardisation of some of the techniques described in this thesis, though further investigation and validation is clearly needed first. For example, this author was involved in creating the EBU ADM Renderer (EBU Tech 3388, 2018). This specification is currently under consideration for standardisation in the ITU-R Working Party 6C, for production and monitoring of NGA programme material. The specification currently only supports rendering to loudspeakers and will likely be extended to specify rendering to headphones in future. The knowledge gained of state-of-the-art binaural rendering techniques for this thesis and the apparatus developed to evaluate these techniques will prove valuable in this process.

This author is also at present a co-chair of a Rapporteur Group within ITU-R Working Party 6C on the topic of "subjective audio evaluation methods" and is an Advisor to the Chairman of the EBU Audio Systems group. In both of these forums, there is a requirement for improved methods to evaluate the quality of NGA technologies for use in the production and distribution of broadcast programmes. The insights gained during this thesis will be shared and may inform the direction of future standards.



(a) Experiment interface interface on mobile device

(b) Initial results showing Plackett-Luce estimated worth of rendering techniques (error bars represent standard error of estimates)

Figure 9.1: Work-in-progress on evaluating dynamic binaural rendering in an audiovisual mobile application

Appendix A

Apparatus for Investigating Binaural System Factors

A.1 Introduction

The two preliminary studies of chapters 4 and 5 together demonstrate challenges for successful application of binaural rendering technology. Commercial binaural systems cannot deliver sufficient quality enhancements over existing stereo headphone reproduction, despite it being feasible to provide a plausible simulation of real sound sources. Chapters 2 and 3 highlight the many aspects to be considered when designing binaural technology systems and investigating the perceived quality that they provide. To understand these challenges in more depth and potentially find ways to improve the headphone listening experience, controlled exploration of these influencing factors is required. When using commercial systems, it cannot be known precisely what signal processing is being applied, which limits interpretation and external validity of results.

A set of tools has been developed and compiled to allow control of content and system factors associated with the use of binaural technology in entertainment media. Together these form apparatus to investigate the influence of these factors on perceived quality, as demonstrated in experiments reported in this thesis. Additionally, this apparatus enables experimental production of 3D spatial audio programme content. This chapter presents the design, implementation and validation of the components of this apparatus, which includes:

- A software system for the rendering of spatial audio scenes, with remote control to allow use in listening experiments.
- A set of real-time binaural renderers and associated signal processing components.

- Evaluation and integration of head tracking systems.
- Evaluation of system latency.
- Impulse response measurement and processing.
- Headphone-to-ear correction filter (HpCF) design.
- Head-related transfer function (HRTF) interpolation.

The process also involved the selection and procurement of hardware equipment, including headphones, microphones, and a tracking system. The tools described in this chapter were used in the studies of chapters 5 to 8.

A.2 A Flexible Spatial Audio Rendering Software System

This section describes a flexible software system for rendering spatial audio scenes. It was designed to overcome limitations of available systems, whilst also taking advantage of new open standards for representing spatial audio. It was developed with other colleagues at BBC R&D for conducting research experiments, and also to support experimental content production.

A.2.1 The Need for a New System

The effects of audio content production techniques and the formats used to represent the content clearly have an influence on the quality of the listening experience, as of course does the binaural technology used to produce the listening experience. The systems and apparatus used in chapters 4 and 5 of this thesis have some limitations for further investigation of factors influencing the perceived quality of binaural technology.

The systems evaluated in chapter 4 take a fixed input format (5.1 multichannel surround). They are also *black box* systems, it is difficult to analyse and control the signal processing techniques that they apply with sufficient accuracy. The SoundScape Renderer (SSR) system (Ahrens, Geier, et al., 2008), which was modified for use in the study of chapter 5, gives advantages in this respect. It is an open-source software project, therefore the applied signal processing techniques can be known and modified. It also allows for definition of a spatial scene, with variable source positions, as well as listener tracking interfaces. The network control interface also provides a means to perform listening tests using real-time rendering processes and controlling parameters from a separate test interface application.

However the SSR system also has some limitations. The representation of source positions is two-dimensional (2D), limited to the horizontal plane[1]. Additionally, it is not possible

[1] For the study in chapter 5, source elevation was captured in measured binaural room impulse responses

to represent dynamic scenes with moving sources. It uses a static scene file format (Geier, Ahrens, et al., 2008) and provides a network interface for varying source positions from a remote application, but it would require development of an external application to control repeatable time-varying scenes. Furthermore, the binaural renderers in the SSR are quite limited. They require filters with fixed spatial resolution, 1° in azimuth or head yaw. The listener tracking only supports head yaw rotations, not pitch or roll. The rendering uses dynamic convolution to allow updating filter coefficients without discontinuities, but separate processing of onset delays is not possible. No capability for online filter interpolation is provided, it is only possible to switch between available measurements.

Therefore, for the reasons stated above, a new apparatus for using and comparing state-of-the-art binaural rendering techniques was deemed necessary. Its design, implementation and validation are documented in the rest of this appendix.

A.2.2 New Standards for Representing Spatial Audio

At the time of these developments, standard formats for representing object-based audio scenes and binaural filter data were under development. This author participated in the standardisation of these formats:

- The Audio Definition Model (ADM) and its use within Broadcast Wave Format (BWF) files, developed within the European Broadcasting Union (EBU) and International Telecommunication Union (ITU) (EBU TECH 3285, 2011; EBU TECH 3364, 2014; ITU-R, 2015c,d)
- The spatially-oriented format for acoustics (SOFA), standardised by the Audio Engineering Society (AES) (AES69:2015, 2015).

These formats enable standardised representation of audio content formats and binaural filter data in 3D. ADM BWF files allow storage of dynamic 3D content scenes with many sound sources in a single file, containing both the meta-data and the audio samples. Whilst SOFA files allow storage of large datasets of associated filter data with accompanying metadata, which define the measurement geometry.

A.2.3 A Configurable Software Application

A set of software libraries was developed in C++ to allow real-time rendering of spatial audio content, including a highly configurable real-time spatial audio rendering application.

(BRIRs), so this limitation did not apply.

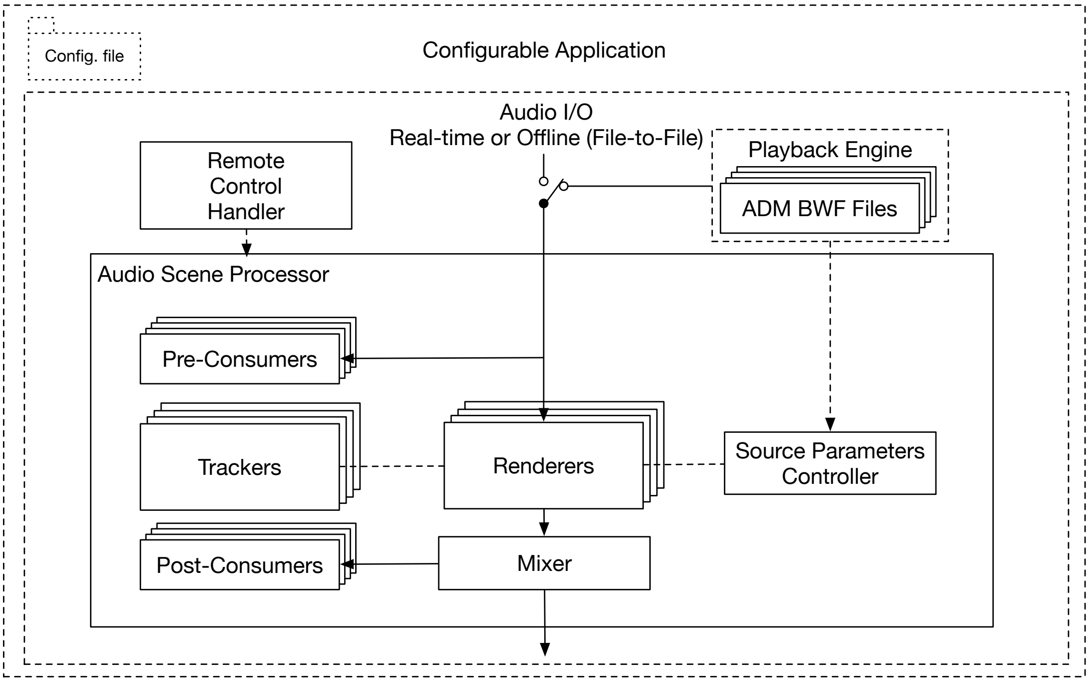


Figure A.1: Application structure

Figure A.1 represents the components of the application. The application is configured at run-time, using a configuration script format to specify the components to instantiate, as well as their initial parameter settings. Rendering can be performed in real-time, processing input signals from an audio interface or reading from audio files. Output signals can be sent to an audio interface and/or written to file. Offline operation can also be configured, for example, for file-to-file processing; this may run faster than real-time. The application host can maintain multiple instances of renderers and trackers, controlling which tracker influences which renderer(s). The renderers are processed simultaneously, which allows dynamically switching between them or mixing of their outputs, depending on the intended application. A playlist of audio files can be represented and controls for playback include play/pause, seeking and looping. When ADM BWF files are read, the parameters of the sound sources in the scene (e.g. position) are updated in each of the renderers. When real-time audio input is used instead, other source parameter generator components can be utilised.

A remote-control component may be used, exposing control parameters to external interfaces, such as remote control over the network via UDP with the Open Sound Control (OSC) protocol. Examples of such controls are renderer mute and level controls, master output level, playback transport controls and source parameters. Analyser components can also be used for signal metering, as pre-consumers (input metering) or post-consumers (output metering). Similarly, the file writer component may be used pre- or post-rendering. Dynamic scenes can be written to ADM BWF files if captured pre-render, meaning that time-varying source movements can be captured and later reproduced. The number of input and output channels that can be handled is dependent upon the audio hardware or file interfaces that the application is configured with. Although, for real-time output, constraints on channel count also apply due to computational processing limits. Within the application, control updates and audio signal processing occur on different threads, to avoid control changes causing audio buffer under-runs during real-time operation.

A.2.4 A Spatial Audio Renderer

The structure of an individual renderer is shown in figure A.2. Within the renderer, mute and level controls on the input can operate on a per-channel basis or across all channels. Chains of pre- and post-processing components may be added, for example for equalisation or dynamic range control. The core rendering process converts from the input signals to output signals via a defined processing algorithm. On the control thread, the renderer interprets changes to control parameters for these components, including source and listener positions received from the scene position controller (e.g. the ADM BWF playback engine) and the head tracker. It is possible for remote control to override updates from the scene controller if desired.

Parameter calculations should occur in the control thread to avoid causing audio drop outs during real-time operation, particularly with complex operations such as filter calculation. Asynchronous control changes should be smoothly updated in the signal processing components on the audio thread.

This generic renderer is agnostic to the reproduction transducers, so could be applied to spatial audio rendering for loudspeakers or headphones. Class inheritance is used to create specific renderer implementations, defining the rendering algorithm to be used, and often targeted specifically at either headphones or loudspeakers. A renderer will be instantiated with a set of parameters, which might include the target reproduction system (i.e. loudspeaker layout), as well as required parameter data such as filter coefficients.

A.3 Real-time Binaural Rendering

Several binaural rendering approaches have been implemented for the above described system, each with flexible parameter configuration. Figure A.3 presents the structure of a basic

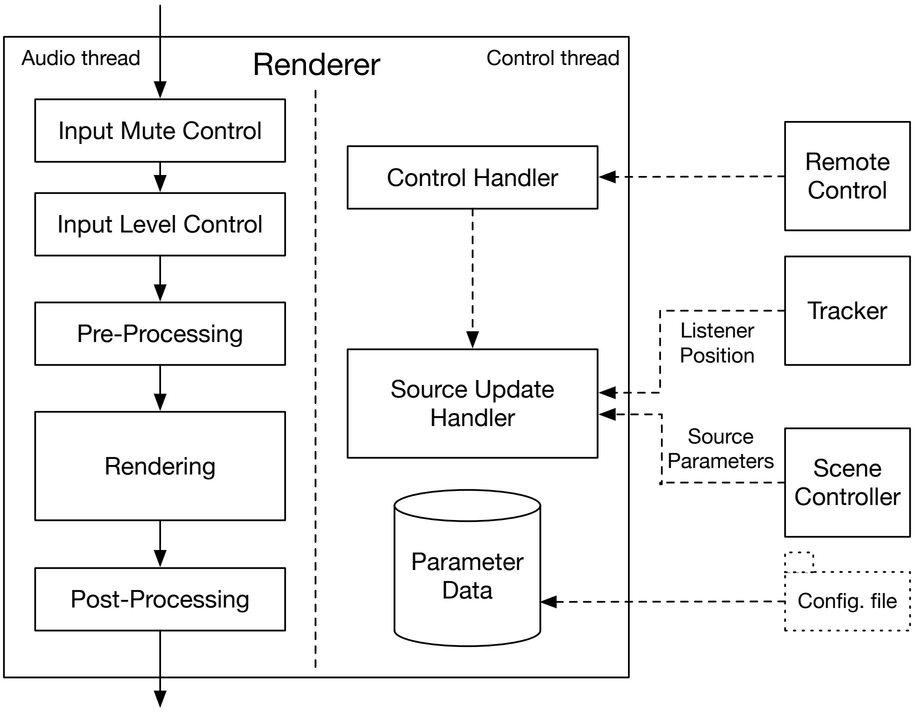


Figure A.2: Renderer structure

binaural renderer. A SOFA file is loaded and the data is prepared for use in a filter generator, for example by creating data structures to allow look up of filters and delays by measured source position, and by converting time-domain impulse responses to complex-frequency domain coefficients. Updates to source position or head tracker data lead to generation of new binaural filters. The source position and the head tracker data are combined in the filter generator to provide an appropriate filter. The filter generator provides new filter coefficients and potentially also separate onset delays. This happens in the control thread.

In the audio processing thread, convolution and delay-line processes occur for each source-to-ear path (i.e. two per source). When new filter parameters are available, they are smoothly updated over the course of a short time transition (e.g. one processing block of 128 samples). Besides delays and filters, a source-to-ear level may be controlled in the same way. This is not shown in the diagram, but could allow, for example, simple broadband modelling of near-field interaural level difference (ILD) changes. A multi-threaded audio processing engine is used to handle the convolution and delay interpolation. Processing tasks are allocated across multiple threads, to take advantage of multicore processors common to modern computers.

Figure A.3 gives a high-level structure for binaural rendering implementations, but many variations exist. Four main classes of binaural renderer were implemented in the software framework.

- *Basic binaural renderer* – Best suited to rendering with short anechoic HRTFs. During filter generation, the source position is rotated by the inverse of the head tracker orientation to give a head-relative source position. This assumes that source rotation and head rotation are reciprocal and ignores head-above-torso orientation.

- *Binaural room scanning renderer* – When using BRIRs, rotation of the listener is quite different to opposite rotation of the source, due to room effects. Sources can only be rendered at the positions of measured loudspeakers within the room (no BRIR interpolation is performed) and head orientation is used for filter look up. As introduced in section 5.2.3, this renderer also uses a static late reverb tail after the perceptual mixing time. Only the early part of the BRIRs is updated according to head orientation, with the tail from a single measurement used. An additional partitioned convolver and delay line are used per source, though with a static filter for the tail response and fixed delay related to the mixing time. The reverb tail impulse responses include a short fade-in (with a fade-out in the early parts) thus giving a cross-fade around the mixing time. This has benefits for reducing memory requirements, because the long late tail only needs to be stored for a single head orientation per source position. It also

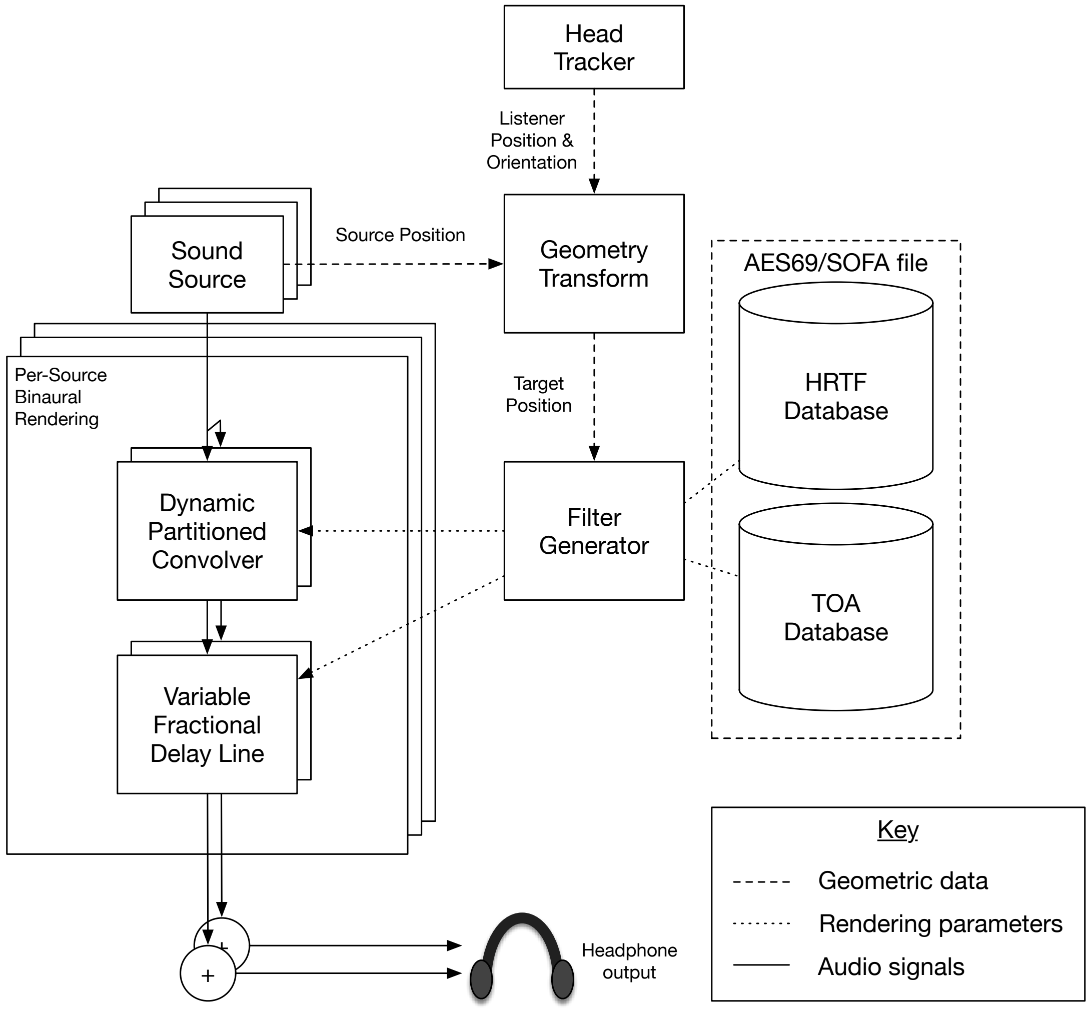


Figure A.3: Basic binaural renderer structure

reduces real-time processing requirements in dynamic conditions, because switching between filters using two convolutions and a cross-fade only has to be performed for the early part of the BRIRs.

- *Virtual loudspeaker binaural renderer* – This is an extension of the above two binaural rendering systems, which sets the sound source positions to be fixed to those of an array of virtual loudspeakers, but adds a loudspeaker rendering stage as a pre-processor. The loudspeaker renderer will take the source positions in the target scene and render them (e.g. by panning) to the given virtual loudspeaker layout. This can be done with either the basic binaural renderer or the binaural room scanning renderer.

- *Ambisonics binaural renderer* – Whilst the virtual loudspeaker renderer could be used with ambisonics-based panning to render a scene, a renderer has also been defined which directly converts an ambisonics signal to a headphone output using binaural processing. Here the inputs are not position-related, ambisonics signals are directly processed.

Techniques for virtual loudspeaker and ambisonics rendering are discussed in appendix B.

A.3.1 Nearest Filter Selection

The filter generator block in figure A.3 provides a simple interface which returns a binaural filter pair, potentially with a pair of onset delays, when given a target position. Interpolation may be used in real-time to generate a filter at the precise position given. The implementation of a real-time interpolation technique is discussed in appendix A.7. However, if the data is measured at a resolution finer than just noticeable differences, or interpolated offline to such a resolution, then the filter generator can simply select the filter corresponding to the nearest available position.

The AES69/SOFA files loaded can contain filters at arbitrary measurement positions. Datasets at the appropriate resolution will also include a large number of measurements (Minnaar, Plogsties, et al. (2005) used more than 10 000). An efficient algorithm is required for selecting the filters. Given a target vector for the source position relative to the listener, a nearest-neighbour search must be performed over the source positions in the set of available measurements. For this a three-dimensional binary search tree is used (Bentley, 1975). The multidimensional binary search tree is commonly named a k-d tree. A k-d tree provides $\mathcal{O}(\log n)$ complexity on average, where n is the number of measurement points, and $\mathcal{O}(n)$ in the worst case. Since a k-d tree works in Cartesian coordinates with a Euclidean distance measure, it can also be used with filter datasets that vary in source distance as well as direction. An open-source implementation of the k-d tree was used to perform nearest neighbour search for head-related impulse response (HRIR) measurements in the binaural renderer (Muja and Lowe, 2009; Blanco and Rai, 2014). A set of unit tests was written to verify the operation of the k-d tree.

A.3.2 Dynamic Convolution

One of the core components to the real-time binaural renderers is a uniform-partitioned fast frequency-domain convolution engine (Wefers, 2014). The overlap-save convolution method is utilised, and the filter and input signal are partitioned into short blocks of uniform size. A frequency-domain delay line is used to store partitions of the input signal in the complex frequency domain for convolution with the later parts of the filter in subsequent processing blocks, thus reducing the number of forward fast Fourier transforms (FFTs) required. This gives large efficiency savings over non-partitioned convolution when used with filters of more than a few hundred coefficients in length. The greater efficiency means that the size of processing blocks for the input signal can be made smaller for real-time applications, so reducing latency.

This reduced latency has particular advantages in dynamic rendering scenarios, where filter updates are required due to changes in source or listener position. When a filter is updated, convolution is performed with both the old and the new filter, and a cross-fade is performed in the time domain between the two output signals to provide a smooth transition without discontinuities (Jot, Larcher, et al., 1995). With partitioned convolution, this cross-fading happens on blocks of the size of a single partition.

In a real implementation on a standard operating system, the algorithm efficiency is greatly dependent on the implementation of the FFT and complex multiplication. An open-source and highly efficient library for performing the FFT is used (Frigo and Johnson, 2005). It offers a function to find the optimal method for calculating the FFT with the required parameters by measuring the run-time of a range of options. The complex multiplication is implemented using single instruction multiple data (SIMD) instructions for greater efficiency: SSE3 intrinsic functions are used for x86 processor architectures, which are available in most modern C++ compilers. This leads to an efficient low-latency dynamic convolution engine capable of handling long impulse responses such as measured BRIRs. A set of unit tests have been implemented to verify the operation of the implementation.

A.3.3 Onset Insertion

Time of arrival (TOA) onset delays must be inserted into the signals when they are stored separately to the filters. These should be inserted with sub-sample accuracy in order to preserve interaural time differences (ITDs), and when updates occur due to changes in source or listener positions, the delays should be interpolated to avoid signal discontinuities.

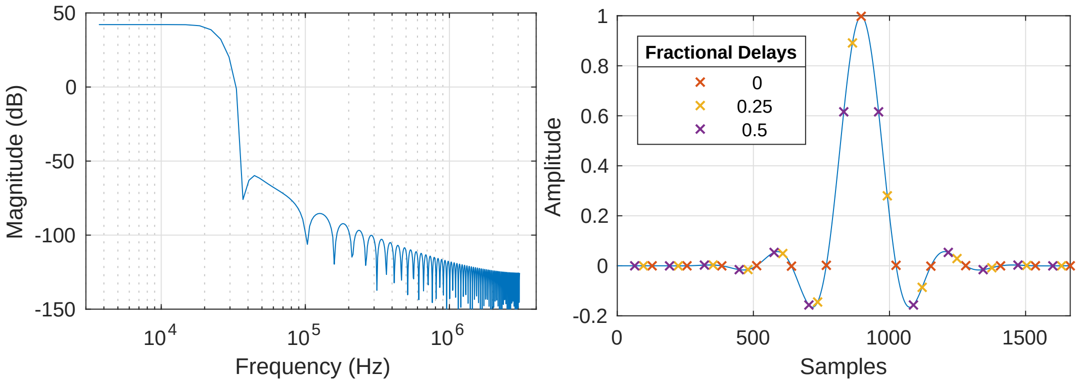
Band-limited delay interpolation was performed using a set of 12th-order low-pass finite impulse response (FIR) filters, designed using a windowed sinc function (Laakso et al., 1996)[2].

$$h_\delta(n) = \begin{cases} \alpha w(n-\delta)\text{sinc}(\alpha(n-\delta)), & 0 \leq n \leq L \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

where n is the sample index, L is the filter order, δ is the delay, α is the normalised cut-off frequency, and w is a window function. This window function is used to reduce the Gibbs phenomenon ripple effect that would be caused by truncation of the sinc function. Since such filters are most accurate when $\delta \approx (L+1)/2$, the provided delay values ranged over $6 \leq \delta < 7$. A set of filters was created to correspond to discretised delays in steps of 1/128 samples, which could be stored in a lookup table. This process is shown in figure A.5. The filter design was performed with 128 times oversampling, using a Kaiser window with $\beta = 10.056$, giving side-lobe attenuation of approximately 100 dB, and a cut-off frequency of 23 945 Hz (see figure A.4). The resulting filter was then downsampled to a rate of 48 kHz for each delay value. This led to group delay variation and magnitude response characteristics dependent on the target delay. But the lowest cut-off frequency of the filter set is 19.75 kHz (for $\delta = 6.5$), and the group delay is flat up to at least 13 kHz, as shown in figures A.5b and A.5c.

During real-time operation, the appropriate filter is selected from the lookup table and filter coefficients can be updated on a per-sample basis to interpolate delays between two values over a block of samples. A standard delay line is used to provide the additional integer sample delays. Figure A.6 shows a comparison of an original measured HRIR, prior to onset removal, and the impulse response generated by the real-time binaural rendering software, with onset reinsertion using these fractional delay filters. The source position was $(\theta, \phi) =$ (30°, 0°). When the ITD was estimated from the rendered impulse response using the log-threshold method, it produced the same value as estimated from the original measured HRIR. More detail on TOA onset estimation and associated estimation of the ITD is given in appendix A.5.6.

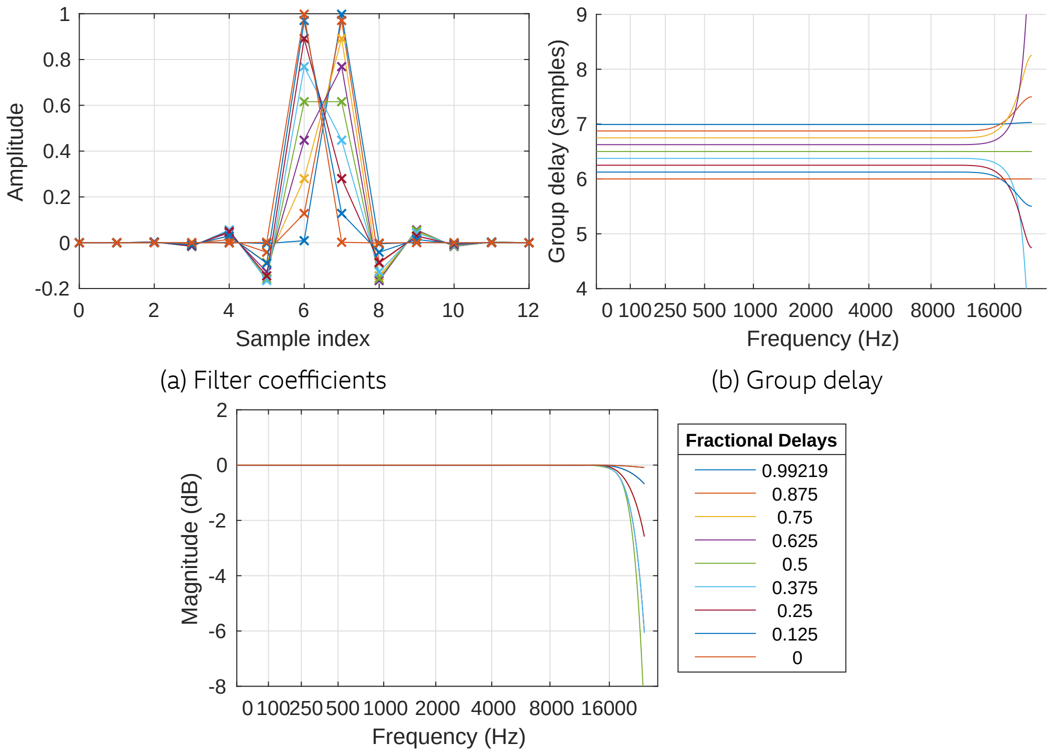
[2] This functionality is provided by the `fir1` function in MATLAB.



(a) Magnitude response of FIR filter designed with 128 times oversampling

(b) Downsampling of FIR filter with varying fractional delay

Figure A.4: Design of fractional delay filters by downsampling a band-limited windowed sinc-function FIR, initially designed by oversampling.

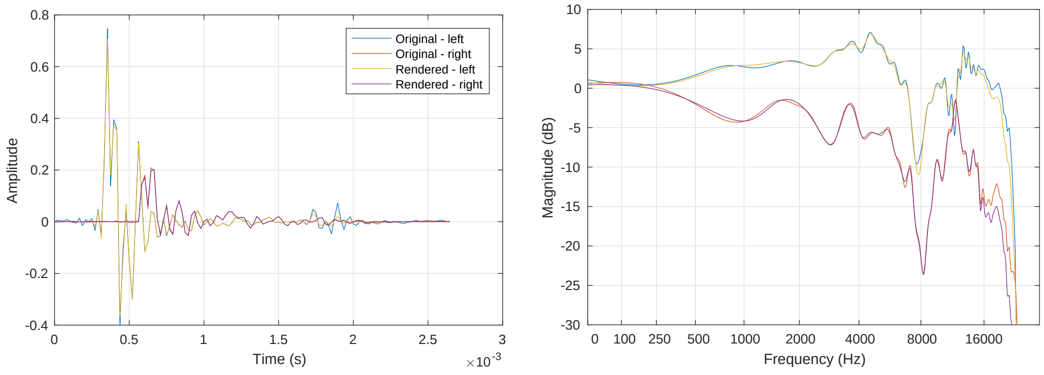


(a) Filter coefficients

(b) Group delay

(c) Magnitude responses

Figure A.5: 13th-order FIR filters with varying fractional delay, obtained by downsampling a band-limited windowed sinc-function filter that was created with oversampling by a factor of 128.



(a) HRIRs

(b) HRTFs

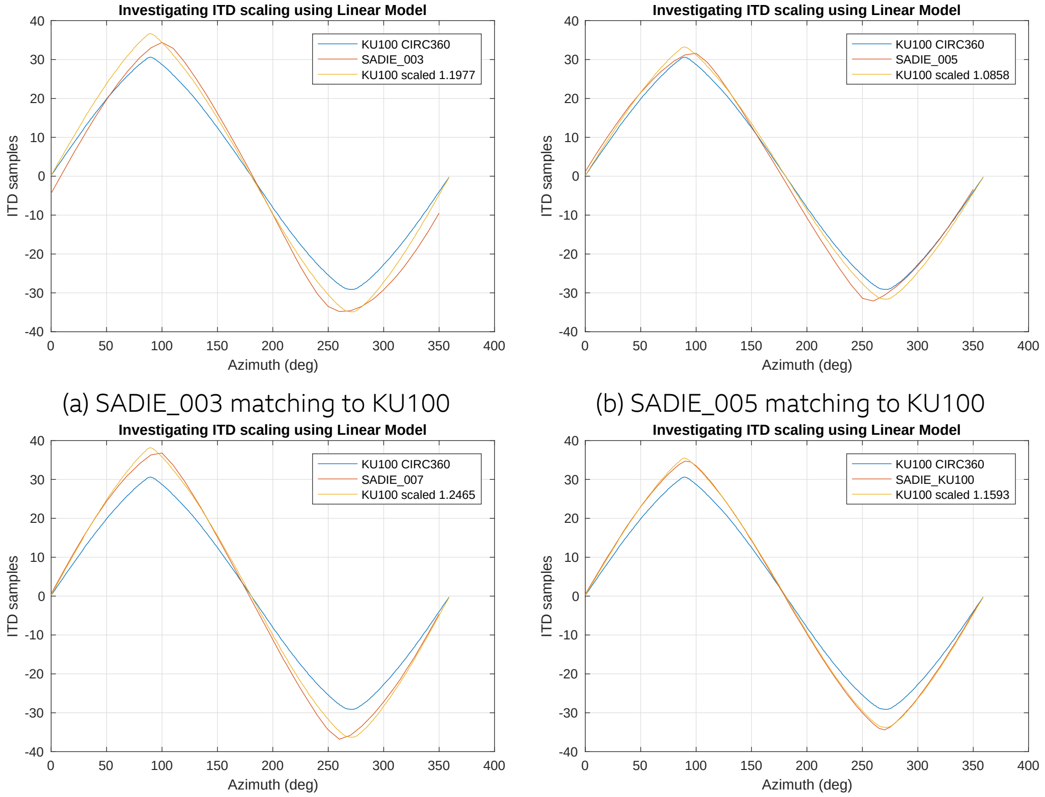
Figure A.6: Onset re-insertion with fractional-delay filtering using the real-time renderer application, compared to original measurements, for a source at (30°,0°).

A.3.4 Interaural Time Difference Adaptation

When the rendering system uses non-individual binaural filters, by default those from the Neumann KU100 dummy head, it is possible to adapt the ITDs by scaling the onset delays for each ear in real-time. The aim is to adjust the broadband TOA estimated from the measured data so that the ITDs better match that of the individual. This approach is taken by Lindau, Estrella, et al., 2010. This is a means of individualising to the listener to some extent. It is particularly useful to stabilise the position of a source during head-tracked dynamic rendering.

Figure A.7 demonstrates the approach by adjusting ITDs estimated from the Neumann KU100 HRIR data of Bernschütz (2013) to match individuals in the SADIE HRTF database (Kearney and Doyle, 2015a). ITD estimation is performed by modelling with the Ziegelwanger and Majdak, 2014 off-axis spherical head model, fitted to initial minimum-phase cross-correlation TOA estimates. A linear model fitting function was used to find the best scaling value that minimised ITD errors in a least-squares sense. Figure A.7d shows that this is sensitive to the measurement and TOA estimation process, since the ITDs for the KU100 as measured in the SADIE database are not the same as those estimated from the data of Bernschütz (2013). After scaling, the data is well aligned. For the human subjects in Figures A.7a to A.7c, the effect of different angular position of the ears can be observed, leading to incomplete matching, particularly at the extremes of the ITD function.

The ITD scaling is presented as a control parameter that can be adjusted in real-time whilst rendering, therefore allowing perceptual adjustment by a listener. It should be noted that the renderer implementation does not scale the ITDs directly, but modifies the TOA at



(a) SADIE_003 matching to KU100

(b) SADIE_005 matching to KU100

(c) SADIE_007 matching to KU100

(d) SADIE_KU100 matching to KU100

Figure A.7: Scaling modelled ITDs of the Neumann KU100 HRIRs to best fit those of SADIE HRTF database subjects, using least-mean squares error fitting. Using sources at varying azimuth in the horizontal plane.

each ear independently, by scaling the difference to the mean overall TOA for the dataset. This better preserves the phase relationships between multiple coherent sources e.g. during virtual loudspeaker panning.

By running an impulse signal through the real-time renderer application, the rendered HRIRs can be obtained. Figure A.8 demonstrates the results of this process, with two different source positions and two different target ITD scaling values. This scaling process can be performed on HRIRs and BRIRs in the renderer, provided that the loaded SOFA file includes separate delay data.

A.4 Head Tracking System

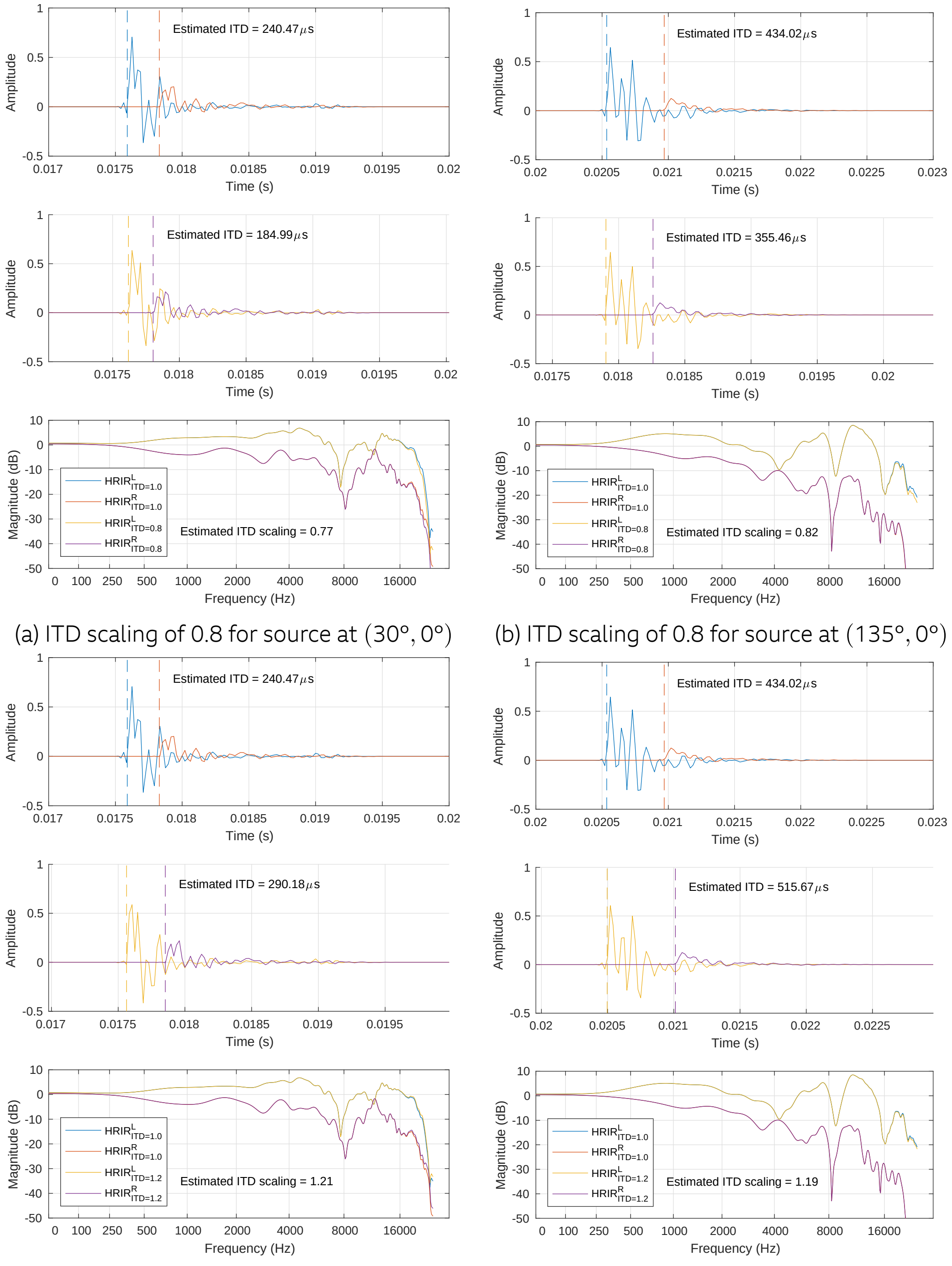
A state-of-the-art binaural rendering system requires a head tracking system in order to achieve dynamic rendering. This should meet the perceptual requirements for plausible synthesis in dynamic environments, as outlined in section 2.8. These include latency, accuracy, stability, range, and degrees of freedom (DoF) of tracking. A number of tracking devices were tested on loan from manufacturers in order to assess their suitability for this purpose. The most suitable system was purchased and is presented in more detail.

A.4.1 Tracking Systems Investigated

The following tracking devices were tested in 2013, considering their use in a dynamic binaural rendering system:

- Razor IMU - a cheap inertial device with motion filtering software which gives 3 DoF orientation tracking.
- Microsoft Kinect - a consumer depth-camera with open-source software for face tracking in 6 DoF.
- InterSense InertiaCube 4 - an inertial device with motion filtering which gives 3 DoF orientation tracking.
- Polhemus Fastrak - an electro-magnetic tracking system with transmitter and receiver.
- ART SmartTrack - a desktop mounted optical marker-based tracking system with two cameras.
- Vicon Bonita - a multiple camera optical marker-based tracking system.

These devices were tested with the SoundScape Renderer (Ahrens, Geier, et al., 2008), since the evaluation took place before the new rendering system was implemented. Where software interfaces were not already available, they were written, to allow testing in context of the dynamic binaural rendering application.



(a) ITD scaling of 0.8 for source at (30°, 0°)

(b) ITD scaling of 0.8 for source at (135°, 0°)

(c) ITD scaling of 1.2 for source at (30°, 0°)

(d) ITD scaling of 1.2 for source at (135°, 0°)

Figure A.8: ITDs of the Neumann KU100 HRIRs using the real-time renderer, scaled to a given target value. For each subfigure there are three plots: top - renderer HRIR with scaling of 1.0, middle - renderer HRIR with target scaling, bottom - comparison of HRTF responses. Text labels indicate the ITDs estimated from the resulting HRIRs and the corresponding ITD scaling value.

The Razor IMU device provides responsive orientation tracking but has issues with stability, especially since careful calibration of the multiple sensors is required. It also is unable to provide an absolute reference of orientation. The face pose tracking software for consumer depth cameras such as the Microsoft Kinect was not reliable enough for high quality rendering and the latency was clearly audible when one's head was turned. However these two systems are useful cheap solutions for desktop testing and informal demonstrations. The InertiaCube device also showed issues with stability, the output data drifted over time and this drift was audible and annoying during binaural rendering. Such errors are common with inertial systems. Filtering can be applied to reduce the effects but it is still undesirable in a high quality reference system. The responsiveness of the tracker is reduced by this filtering. The Fastrak electromagnetic system gave much better performance. Tracking in 6 DoF was highly accurate and responsive when the receiver was close to the transmitter. However the data became less accurate with distance and audible jitter was present at distances of greater than 1.5 m. The accuracy was also severely degraded by the presence of metal objects within the electromagnetic field. This is problematic in the BBC R&D laboratory, since the listening room has metal in the walls and floor.

Optical tracking systems were identified by Hess (2012) as being optimal, due to their long term stability and maintaining an absolute reference, provided that latency requirements can be met. Unfortunately the ART SmartTrack's software system has a minimum processing latency of 130 ms and the camera unit has an update rate of only 60 Hz. In comparison, the Vicon Bonita has an update rate of up to 250 Hz and its software reports a processing latency of 2.5 ms.

A.4.2 The Vicon Bonita System

Vicon Bonita is a multiple camera optical tracking system that uses passive reflective markers to detect objects. It is designed for motion tracking applications including biomedical sciences and animation. The system is modular and scalable, in this instance four cameras were used. These cameras have an array of LEDs which emit infra-red light that is reflected by the markers (see figure A.9a). The cameras then apply filters to detect the regions of reflected infra-red light. The cameras send a processed grayscale image over an IP network to a PC running the Vicon Tracker software (see figure A.9c), which locates the markers within each image and triangulates their three-dimensional (3D) position. Markers must be in view of at least two cameras for their position to be determined. The triangulation is supported by a calibration procedure which identifies the relative positions and orientations of the cameras and then locates them with reference to a defined coordinate system origin. Tracking techniques are used to identify markers between frames and maintain estimated position and movement during occlusions. A set of markers can be identified as a rigid body (known as an *object*) which is then tracked in 6 DoF within the volume of the camera view.

In this scenario the object is the head of the listener, using a rigid body of markers mounted on the headphones. A set of small asymmetric structures upon which to mount the markers was designed and 3D-printed, as pictured in figure A.9b. A software development kit is provided to allow access to a real-time stream of object tracking data over the network. The tracking was integrated into the SoundScape Renderer for early testing and then subsequently the new spatial audio rendering software described in appendix A.2.3.

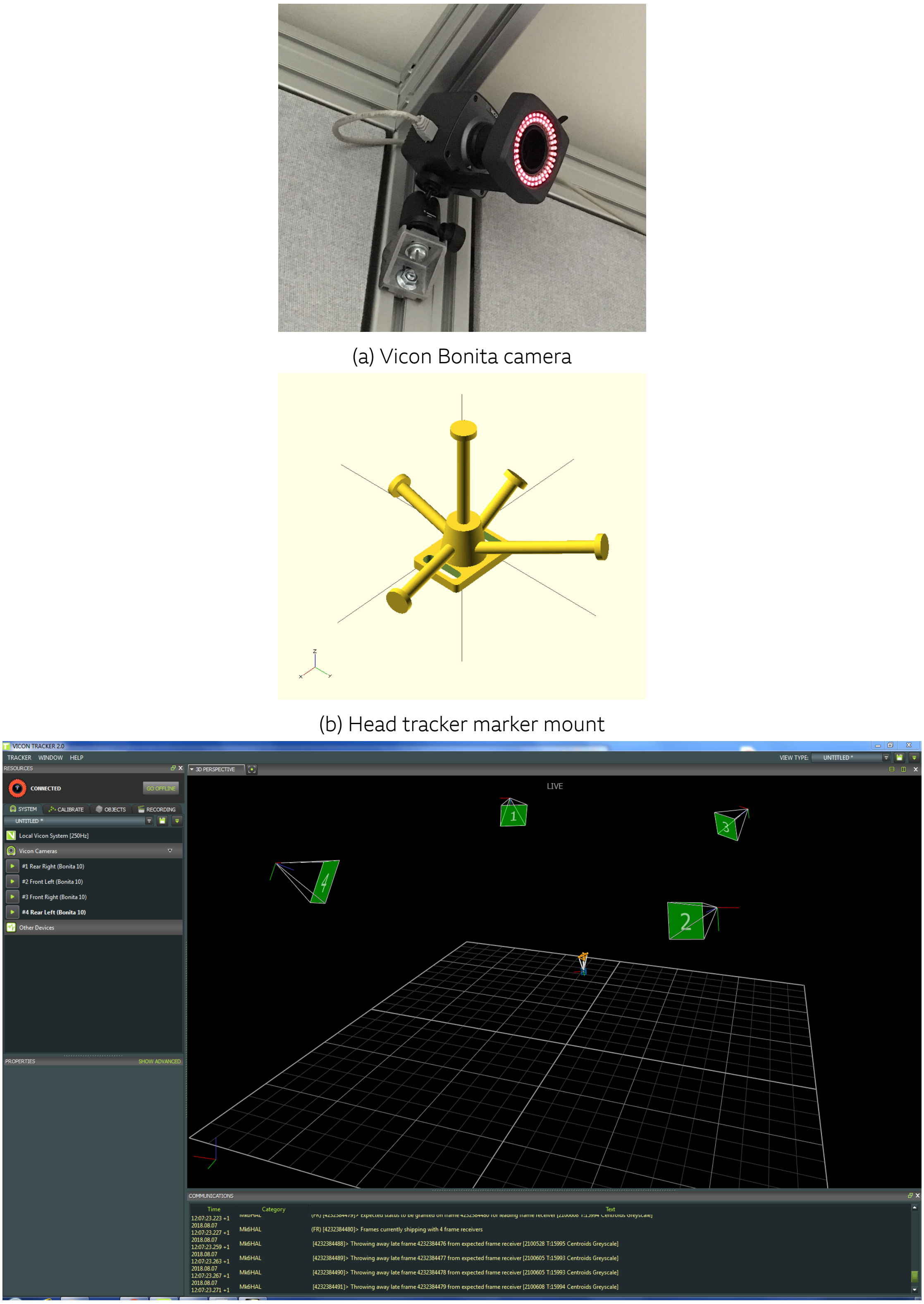
Objective verification of the performance of tracking systems can be challenging due to the lack of ground truth data. Experience with the Vicon Bonita system and listening with a binaural renderer showed highly accurate and stable tracking of objects. The system has specified tracking precision of 0.5 mm and 0.5°. Latency is known to be the major issue with many optical camera-based tracking systems (Hess, 2012). The maximum update rate of the Vicon system is 250 Hz and the Tracker software has a reported processing latency of 2.5 ms. Yet, other factors such as network latency and audio processing latency also contribute to the total system latency. To determine if latency is below acceptable thresholds when using the tracker, the total system latency must be measured.

A.4.3 Measuring Total System Latency

Apparatus for measuring the total system latency was set up, inspired by the techniques of Wenzel, 1997 and Lindau, 2009. Headphones were attached to a camera tripod, an accelerometer was attached to the handle and a measurement microphone was placed at the headphones (figure A.10a). Both were connected to the same audio ADC interface for recording on a PC.

The binaural impulse response data was adjusted so that the measurement at (0°,0°) contained only zeros. The renderer output was therefore silent when the headphones were oriented in this position. The BRIRs data was at 1° resolution, so when the measured angle moved to ±1°, the renderer emitted sound. A white noise source was used as input to the renderer. The accelerometer and microphone output signals were recorded at 48 kHz. Onsets were detected using the log-threshold method, with a threshold of −10 dB below the peak value, to estimate the latency between the movement and the onset of sound output from the headphones. This process is shown in figure A.10b.

The tripod was oriented so that the tracker data reported a value of 0.0° and then the arm was moved suddenly by hand. The latency of the SoundScape Renderer system used



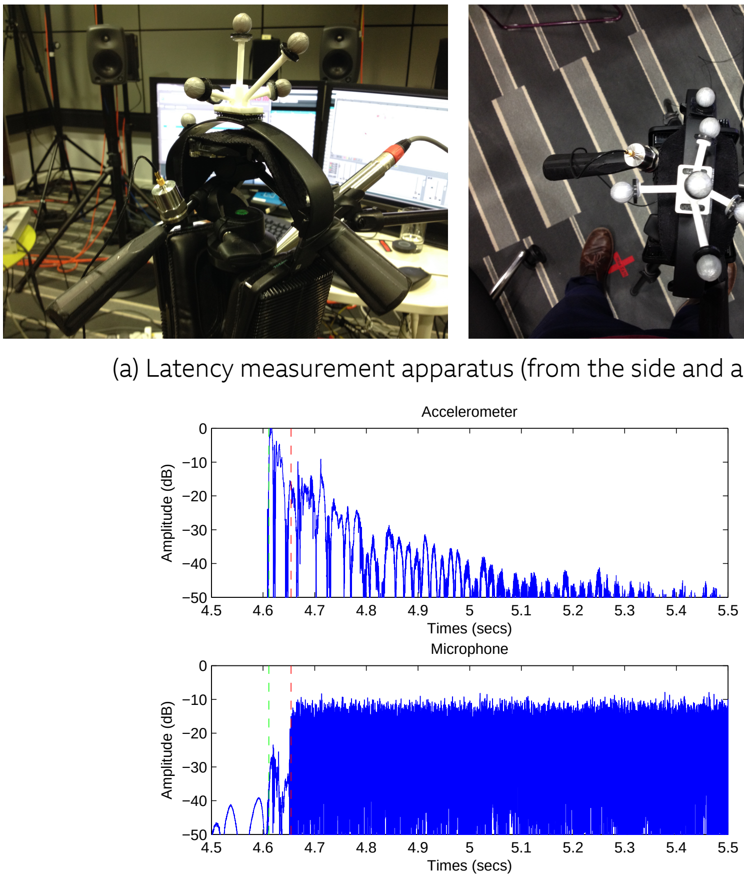
(a) Vicon Bonita camera

(b) Head tracker marker mount

```
Time              Category    Text
12:07:23.223 +1   MK6HAL      (FR) [4232384479]> Expected status to be granted on frame 4232384480 for leading frame receiver [2100608 T:15994 Centroids Greyscale]
2018.08.07
12:07:23.227 +1   Mk6HAL      (FR) [4232384480]> Frames currently shipping with 4 frame receivers
2018.08.07
12:07:23.259 +1   Mk6HAL       [4232384488]> Throwing away late frame 4232384476 from expected frame receiver [2100528 T:15995 Centroids Greyscale]
2018.08.07
12:07:23.263 +1   Mk6HAL       [4232384489]> Throwing away late frame 4232384477 from expected frame receiver [2100605 T:15993 Centroids Greyscale]
2018.08.07
12:07:23.267 +1   Mk6HAL       [4232384490]> Throwing away late frame 4232384478 from expected frame receiver [2100605 T:15993 Centroids Greyscale]
2018.08.07
12:07:23.271 +1   Mk6HAL       [4232384491]> Throwing away late frame 4232384479 from expected frame receiver [2100608 T:15994 Centroids Greyscale]
```

(c) Vicon Tracker user interface, showing BBC R&D listening room system

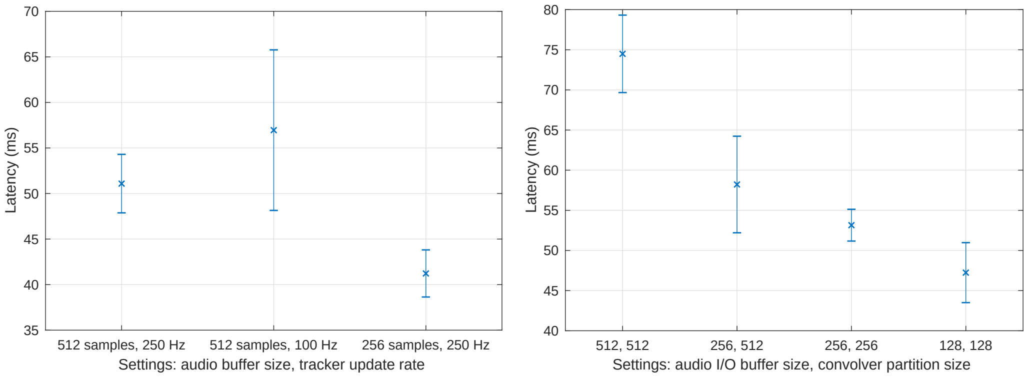
Figure A.9: Vicon Bonita Tracking System



(a) Latency measurement apparatus (from the side and above)

(b) Estimating onsets of accelerometer and headphone signals

Figure A.10: System latency measurement apparatus



(a) Using the SoundScape Renderer with varied audio buffer sizes and tracker update rates.

(b) Using the new rendering system with varied audio buffer sizes and filter partition sizes, tracker update rate fixed at 250 Hz.

Figure A.11: Total system latency measurements. The mean latency is indicated, with error bars showing ± one standard deviation.

in chapter 5 was evaluated. Twenty repetitions were made for each setting, using audio input/output buffer sizes of 256 or 512 samples and setting the tracker update rate to either 250 Hz or 100 Hz. The results are plotted in figure A.11a. With an update rate of 250 Hz and an audio buffer size of 256 samples, the mean latency was 41.2 ms, with standard deviation of 2.6 ms and maximum observed value of 47.9 ms. This is lower than most just detectible system latency values found in the literature (section 2.8.2.5).

Subsequently latency was measured with the new system described in appendix A.2.3. The tracker update rate was fixed at 250 Hz, but the audio input/output buffer size and the convolver partition size were varied[3]. With the lowest measured buffer and partition sizes, both at 128 samples, the mean measured system latency of the new system was 47.2 ms, with standard deviation 3.7 ms and maximum measured value of 53.1 ms. This is below the lowest measured detection threshold for all but one of the listeners in the study of Lindau (2009), so is deemed adequate for experimental use.

A.5 Binaural Impulse Responses

Core to the quality of binaural rendering is the set of binaural filter data used. Since the rendering software can be configured at runtime to load a SOFA file with arbitrary measurement data and geometry, the system is flexible to experimentation with various filter data options. However the source of the data must be considered. As introduced in section 2.4.2,

[3] The SoundScape Renderer always uses a convolver partition size equal to the audio input/output buffer size.

a large number of HRTF measurement databases are publicly available. The SOFA database makes many of these available as SOFA files (SOFA, 2018).

A.5.1 The Neumann KU100

In the course of this project, a Neumann KU100 dummy head microphone was purchased. This commercial microphone is designed for professional recording purposes. It has a more practical form than a full head and torso simulator (HATS) for location recording and has a diffuse-field equalised response, which is acceptable to professional sound engineers (see section 2.3.2, for a review of an earlier model by BBC engineers). It must be acknowledged that the localisation quality afforded by this microphone system may be more limited than other options, particularly in terms of vertical localisation, as observed by Minnaar, Olesen, et al. (2001). But this should be balanced against the good tonal quality it offers.

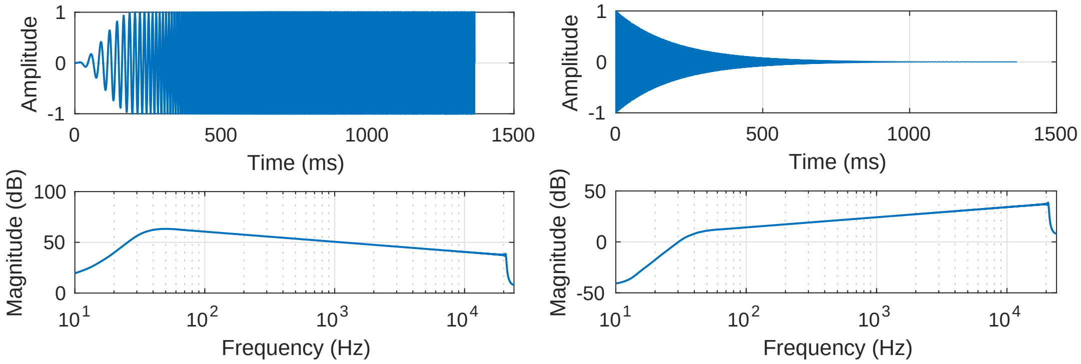
The Google VR Audio tools use HRTFs measured on the KU100. Google (2018c) describe listening tests carried out using a range of HRTF sets in a mobile VR application, where measurements on the KU100 microphone were rated highest quality for most participants.

A.5.2 Head-Related Impulse Responses

The HRIRs used by default in the binaural rendering system of this project are described by Bernschütz, 2013. They were measured with a Neumann KU100 in the anechoic chamber at the Cologne University of Applied Sciences. The measurements are available for a number of full-sphere measurement grids (i.e. with no gaps), measured with the source in the far field, at a distance of 3.25 m. A precise computer-controlled rotational measurement system was used to automatically rotate the microphone relative to the loudspeaker.

The measurements were post-processed with an adaptive low-frequency extension algorithm, which corrects for deficiencies in the measurement system (excessive group delay, low energy) and assumes that the head behaves like a sphere at low frequencies. The magnitude and phase responses of the measurement loudspeaker were also equalised. As a result of this processing reducing the filter group delay, the resulting filters are short, only 128 samples in length at 48 kHz sampling rate. This has benefits both for sound quality, due to reduced phase distortion, and rendering efficiency.

The same research group subsequently made near-field HRIR measurements at several distances (Arend et al., 2016).



(a) Exponential swept sine signal

(b) Inverse sweep filter

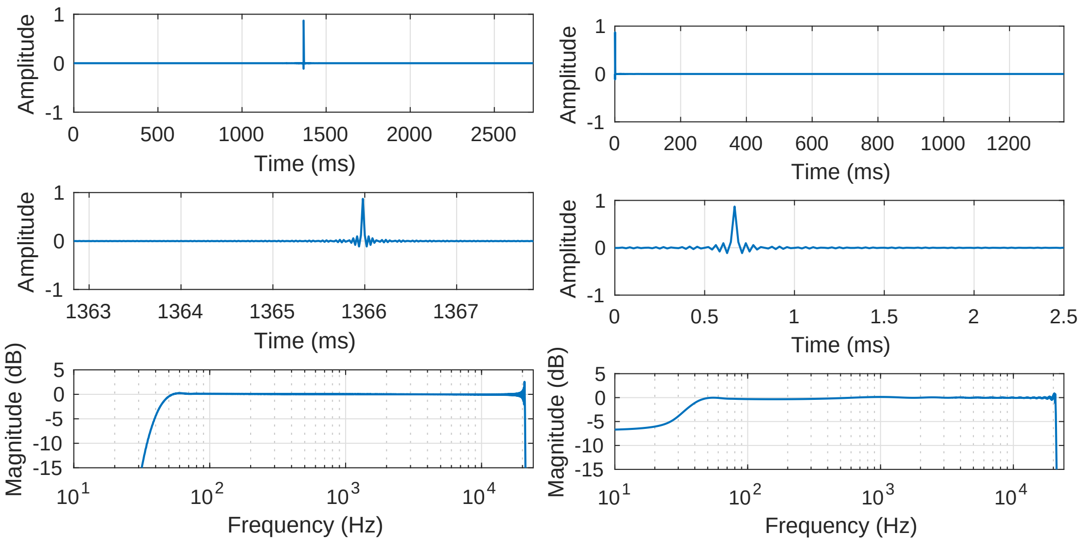
Figure A.12: Exponential swept sinusoid signals for impulse response measurement

A.5.3 Impulse Response Measurement

The exponential swept sine technique was used for impulse response measurement, using the techniques described by Farina (2007). This includes applying a half-cosine window fade-in at low frequencies and ensuring that the sweep ends at a zero crossing. The inverse filter is the time-reversed sweep signal with an amplitude envelope decaying at 6 dB/octave. The sweep and inverse filter signals ranging from 20 Hz to 21 kHz are plotted in figure A.12. The impulse response resulting from convolution of these two signals is shown in figure A.13, a simulated measurement system latency of 32 samples is included. A small amount of ringing can be observed in the time domain due to the limited bandwidth.

The swept sine measurement technique allows separation of non-linear harmonic distortion components from the linear impulse response when they are present in the measurement. Figure A.14 demonstrates these effects with clipping distortion, where the sweep measurement was limited to absolute amplitudes of 0.9 or less. The linear impulse response can be separated from the harmonic distortion components by time windowing of the response.

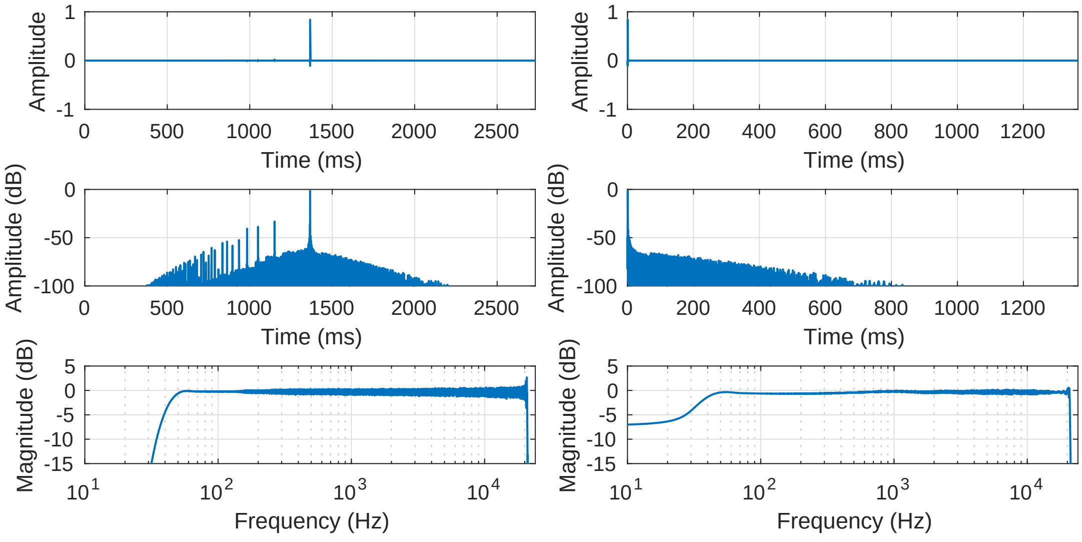
As part of this project, software tools were written for sweep generation and deconvolution of measurements with the inverse signal, in several programming languages (MATLAB, Python, C++). Each system was capable of making scripted sequences of measurements from a number of output channels using multiple input channels. It performed simultaneous playback and recording of signals using a hardware audio interface.



(a) Full response

(b) Cropped linear reponse

Figure A.13: Impulse response generated by convolution of the sweep and inverse filter signals. Top - full impulse response, middle - impulse response, zoomed around peak, bottom - magnitude response.



(a) Full response

(b) Cropped linear reponse

Figure A.14: Effects of non-linear clipping distortion on impulse response from swept-sine measurement. Top - impulse response (linear amplitude), middle - impulse response (log amplitude), bottom - magnitude response.

A.5.4 A Computer-Controlled Rotational Mount for Dummy Head Measurements

For a dynamic data-based auditory virtual environment (AVE), BRIRs should be measured at many head orientations. When measurements are made using a dummy-head microphone, a computer-controllable rotational mount can be used to make these measurements efficient and precise.

There are some commercially available systems, but these often cost many thousands of pounds, e.g. (Brüel & Kjær, 2018; Four Audio, 2018). They are also typically designed to handle much heavier loads than the Neumann KU100, which weighs 3.5 kg, e.g. large loudspeakers. In collaboration with a colleague at BBC R&D, a motorised telescope mount was adapted for this role. The control system of the telescope mount was replaced, to provide remote control of the dummy head's orientation from the measurement software tool, as reported in co-authored publication Co.P. I (Shotton et al., 2014). This provided an affordable tool for making the required measurements.

A.5.4.1 Design Requirements

The system should be fully controllable from a computer, ideally with a simple protocol, which would be easy to implement in a variety of programming languages. This would allow the system to be set up and left running unattended, with minimal human intervention, as well as allowing integration with a range of existing software systems. The connection interface should also be commonly supported on modern computers, allowing flexibility.

The system should, at a minimum, be capable of azimuthal/yaw rotations. As reviewed in section 2.2.7, these are the most commonly used rotation axes in natural listening conditions. The ability to control the elevation angle of the platform is also desirable, since head tilt and roll are also used during sound localisation and natural listening tasks, although to a lesser extent. It is critical that azimuthal rotation be around the vertical line passing through the mid-point of the interaural axis. Changes in elevation should rotate the dummy head about the interaural axis.

The accuracy and precision of such a system should be designed considering localisation acuity and experiments on the required resolution of impulse response measurements of binaural systems, as reviewed in chapter 2. Minimum audible angles as low as 1° have been observed (Mills, 1958), whilst BRIR discretisation at 2° has been shown to be audible in dynamic binaural rendering. The measurement system should operate with an accuracy at least one order of magnitude below these perceptual thresholds, i.e. 0.1°, to ensure that reliable data can be obtained.

The measurement system should be capable of moving the weight of the dummy head microphone in both axes. When the system is not rotating, it should make no noise, so as not to interfere with the measurements. Since the dummy head must be static during measurements, the system noise whilst moving is not of concern.

A.5.4.2 Implementation

The Celestron Nexstar SE Computerized Mount cost only a few hundred pounds and the product specifications report precision of 0.26″ (arcsecond i.e. 1/3600°). It is rated for weights up to 5 kg. The telescope mount is in the "alt-azimuth" style, providing rotation in two degrees of freedom about the azimuth and elevation axes. The rotation about each axis is controlled using a set of gears connected to an electric motor, with a rotary encoder attached to the spindle of the motor to provide a feedback mechanism.

After evaluating the built-in control system a number of issues were discovered. Firstly, the serial interface used an outdated standard (RS-232). Secondly, it was required that the mount be calibrated using a multi-step procedure, specific to astronomy applications, each time it was powered on. The serial control protocol was also targeted specifically at astronomy applications and not for general positioning about two axes. Due to these limitations it was decided to replace the control electronics with a system more appropriate for dummy head measurements.

The new control system was based on the open hardware Arduino Mega2560 micro-controller platform, with a motor controller shield. Implementation details are described in (Shotton et al., 2014). The system was made computer controllable, providing a simple control protocol over a serial interface, which appears as a standard serial port using USB serial emulation implemented on the micro-controller board. A Python library was written to provide portable control software that can be run on many computer operating systems. Most modern programming languages support communication through a serial port, so it is feasible to create a library for controlling the mount for other languages.

The control software provides two basic messages: `move_to`, which performs rotation first by azimuth then elevation in the specified number of degrees and `zero`, which sets the orientation reference, such that any subsequent rotations are relative to the current mount orientation. The initial orientation of the mount therefore needs to be externally verified. The control microprocessor responds to commands with an 'ok' message once the operation is complete.

A mounting plate was produced to allow the adjustment of the position of the dummy head, to achieve correct rotation around its vertical axis at the intersection with the mid-point of the interaural axis. Due to mechanical limitations it was not possible to mount the head so that it rotated around its interaural axis, meaning the system is only valid for azimuthal/yaw rotation.

The completed rotational platform is shown in Figure A.15 with the dummy head microphone mounted.

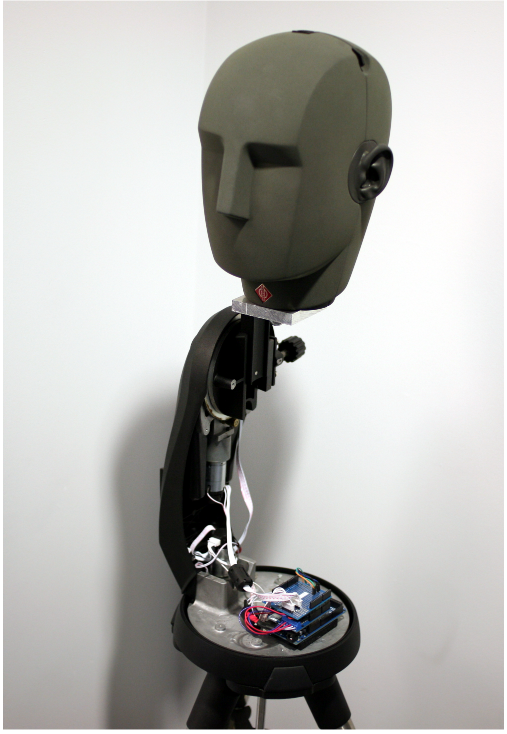


Figure A.15: Computer-controlled rotational platform for automated binaural impulse response measurements, with custom control system.

A.5.4.3 Validation

The resolution of the mount control is 0.90″ in azimuth and 1.56″ in elevation. The control system reported an error of at worst 6.48″ or 0.0018° after movement ceased. It is challenging to validate angular rotations with this degree of accuracy. A digital angle measurement tool with a laser pointer was available, but this only provided precision to 0.1°. A laser pointer was attached to the dummy head, whilst mounted on the platform. It pointed at a wall at a distance of 5.28 m. The mount was rotated in azimuth from 0° to 360° and back five times; the laser mark on the wall was a maximum of 1 mm from its original position (to the nearest mm). Ten different series of sequential and random movements were made in both azimuth

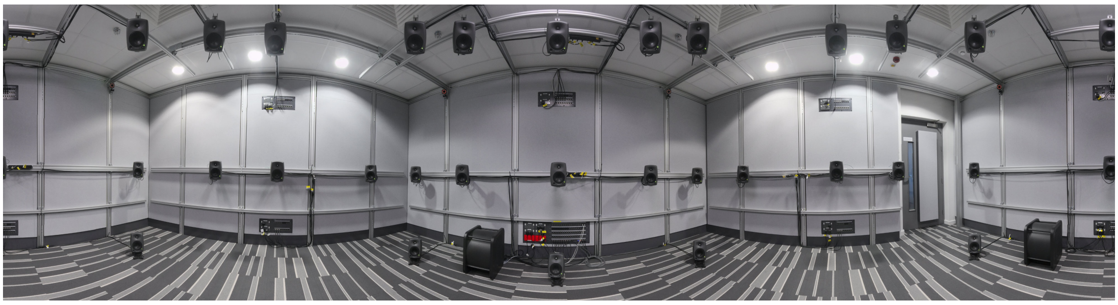


Figure A.16: Panoramic photograph of the Recommendation ITU-R BS.1116-compliant listening room at BBC Research & Development in Salford, UK.

and elevation, before returning to the original position. No deviation greater than 1 mm was observed, which is equivalent to an angular error of 0.011°. The system appears to provide more than sufficient angular accuracy.

The system can safely support the weight of the dummy head. No audible noise is made by the system when it is not rotating. Since the system is currently only capable of relative movements, it must be manually aligned to the frontal direction. This can be achieved using laser pointers and a spirit level, as well as checking the measured ITD on the dummy head.

A.5.4.4 Summary

A highly accurate computer-controllable rotational mount for a dummy head microphone has been constructed. It is suitable for use in automated azimuthal rotation of a dummy head microphone during impulse response measurements. Further work would be needed to allow automatic alignment of the system and rotation about the interaural axis.

A.5.5 An Example Binaural Room Impulse Response Dataset

The tools presented in the previous sections can be used to perform automated measurement of large datasets of impulse response data. The BRIR dataset used in chapter 5 was measured in this way, as were the impulse responses presented in (Satongar, Lam, et al., 2014). Following that study, the listening room at BBC R&D was refurbished, with new acoustic treatment and loudspeaker mounting system (Nixon et al., 2015). A BRIR dataset was measured in this refurbished facility, using an array of 32 loudspeakers, which include all layouts described in Rec. ITU-R BS.2051 (ITU-R, 2014a).

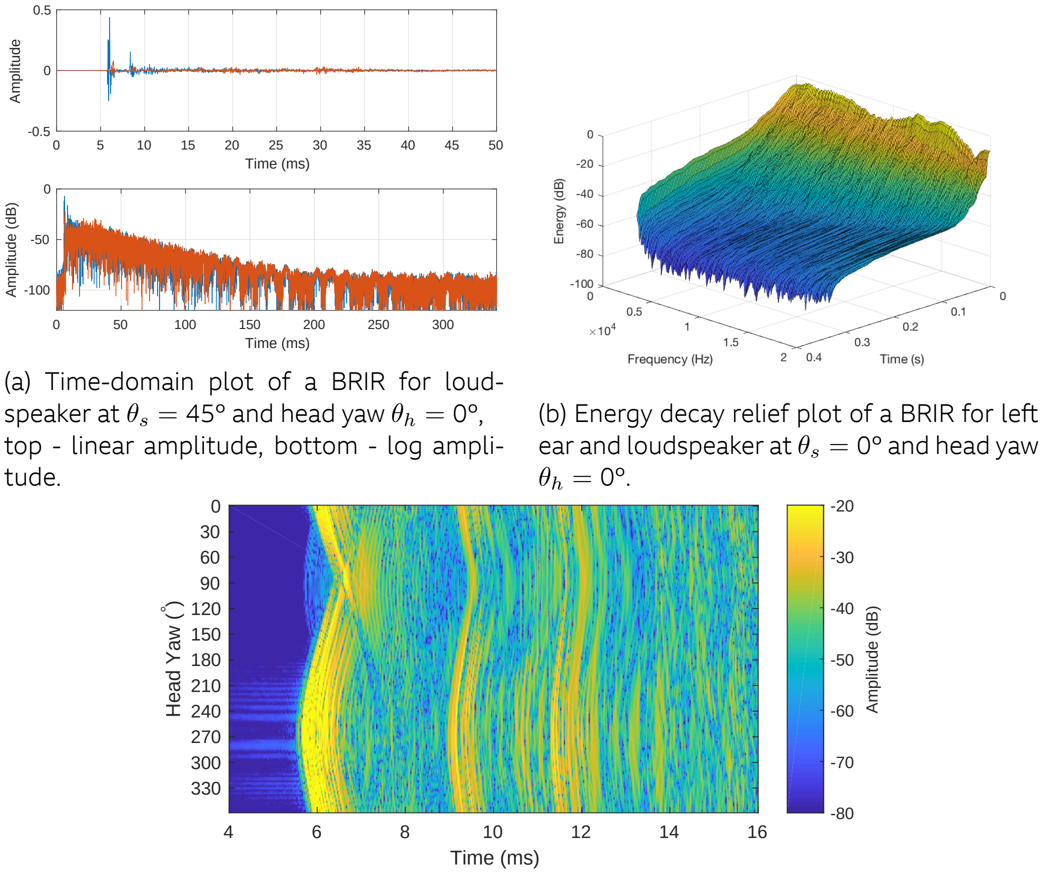
The room was measured and found to be compliant with Recommendation ITU-R BS.1116 (ITU-R, 2015e) and is pictured in figure A.16. This room has a mean reverberation time of 0.21s in the frequency range 125 Hz to 8 kHz. Genelec 8030B loudspeakers were used and the dummy head microphone was rotated about the vertical axis in 2° steps using the motorised rotary mount (appendix A.5.4). Loudspeaker magnitude responses were pre-equalised using an IOSONO CORE system to ensure magnitude response variation within 1 dB between 250 Hz and 2 kHz and time-aligned to within 20 µs at the central listening position. The loudspeaker sound pressure levels were aligned at the central listening position to 70 dBA within ±0.1 dB using a band-limited pink noise signal (20 Hz–20 kHz). BRIR measurements were made using an exponential sine sweep of length 2^{18} samples at 48 kHz sampling rate. Excess onset delay common to all measurements was removed and the IRs were truncated to 2^{14} samples.

Figures A.17a and A.17b show the time-domain impulse responses and energy decay relief respectively for example BRIR measurements. The achieved signal-to-noise ratio is approximately 60 dB to 85 dB in the frequency range 200 Hz to 18 kHz. Below 200 Hz there is more background noise present. Figure A.17c shows the first 20 ms of the BRIRs for the front-centre loudspeaker at each rotation of the dummy head microphone, from which the smoothly changing time-of-arrival of the direct sound can be observed, as well as the first two reflections, which are likely from the floor and ceiling. Whilst the peak amplitude in the measurements was −2.29 dB, the colour scale is clipped at −20 dB to make the reflections more clearly visible.

This dataset was reported in co-authored publication Co.P. IX (Pike and Romanov, 2017a) and has been released under a Creative Commons licence (CC-BY-SA 4.0) on GitHub (Pike and Romanov, 2017b). The data were made available in a number of formats, including configuration files suitable for use in the SoundScape Renderer and AmbiX plug-ins, to allow dynamic auralisation of 3D loudspeaker layouts in applications. The BRIRs were also published as SOFA files. These use the *MultiSpeakerBRIR* convention, which was developed through discussion with Piotr Majdak, one of the leading developers of the SOFA format.

A.5.6 Broadband Onset Modelling

To enable filter updates without comb-filtering artefacts and to allow real-time adaptation of the ITD, onsets are modelled and stored separately. The SOFA provides a field for separate delays. Many methods have been proposed to estimate the TOA and ITD in binaural impulse response data, as reviewed by Katz and Noisternig, 2014. Andreopoulou and Katz (2017) investigated the most perceptually accurate methods for ITD estimation when combined with minimum-phase HRTF representations. With the HRTF measurement data used, it was found through localisation experiments that using a −30 dB threshold below the peak value, applied to a 3 kHz low-pass filtered signal, gave best correspondence with the original



(a) Time-domain plot of a BRIR for loudspeaker at $\theta_s = 45°$ and head yaw $\theta_h = 0°$, top - linear amplitude, bottom - log amplitude.

(b) Energy decay relief plot of a BRIR for left ear and loudspeaker at $\theta_s = 0°$ and head yaw $\theta_h = 0°$.

(c) First 20 ms of BRIRs for left ear and loudspeaker at $\theta_s = 0°$, with varying head yaw θ_h.

Figure A.17: BRIR measurements from Pike and Romanov (2017a).

HRTFs.

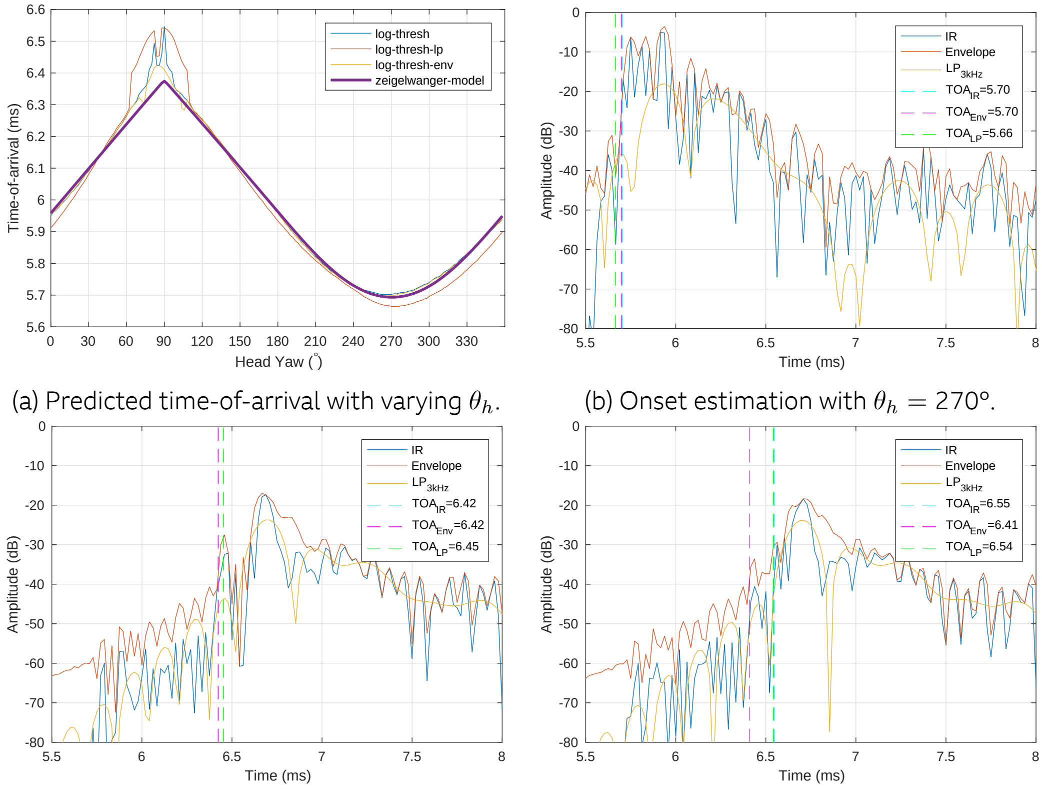
It is also possible to remove broadband onset delays from the filters, simply aligning the onsets, rather than using minimum-phase decomposition. This is particularly necessary when working with BRIRs, which are clearly far from minimum-phase since they incorporate many reflections. Lindau, Estrella, et al. (2010) used the log-threshold method of onset estimation with a threshold of −20 dB and 10 times oversampling. This method finds the first point at which the impulse response exceeds a log-amplitude value that is a given number of decibels below the peak absolute value. Oversampling is used to give sub-sample resolution to the TOA data.

Figure A.18 shows the results of TOA estimation from the BRIRs measured as described in appendix A.5.5. The estimates are for the frontal loudspeaker in the horizontal plane, which had the directional coordinates $(\theta_s, \phi_s) = (0°, 0°)$. The log threshold method was applied to a 3 kHz low-pass filtered version of the impulse response[4], with a threshold of −20 dB and 10 times oversampling, as well as the analytic envelope given by the magnitude of the Hilbert transformed signal. Figure A.18a shows the TOA with varying head yaw. It can be seen that the estimates are quite erratic around the contralateral side where the signal level is low and sound arrives from multiple paths; these effects can be observed in figure A.17c. Figures A.18b to A.18d present the onset estimation process for three different head yaw angles. The sources of these discontinuities can be observed, the time-domain signals fluctuate before the main peak, and the detected onset time will be influenced by whether these exceed the threshold. The low-pass filtered approach leads to larger estimated TOA variation when the loudspeaker is lateral to the head, whilst the envelope leads to a smoother spatial function.

Ziegelwanger and Majdak (2014) present a parametric model of a spherical head which can be fitted to signal-based TOA estimates to give a smooth direction-continuous TOA function. This presents potential benefits in dynamic head tracked rendering, since it reduces the likelihood of rapid changes in delay. Such rapid changes were noted to cause audible artefacts during rendering. Figure A.18a shows the output of the model when fitted to the envelope-based estimates.

The modelled TOA onsets can then be extracted from the BRIR measurements with 10 times oversampling to preserve sub-sample accuracy. The onset delays are reduced by 5 samples at the base sampling rate to preserve the onset slope of the impulse responses. A short fade in of 4 samples is applied to the start of the impulse responses to avoid discontinuities. Figure A.19 shows the process of alignment by onset removal. It can be seen in figure A.19b that the spherical head model leads to underestimates at the contralateral po-

[4] With this data, a threshold of −30 dB gave poor estimates of TOA for the low-pass filtered signal.



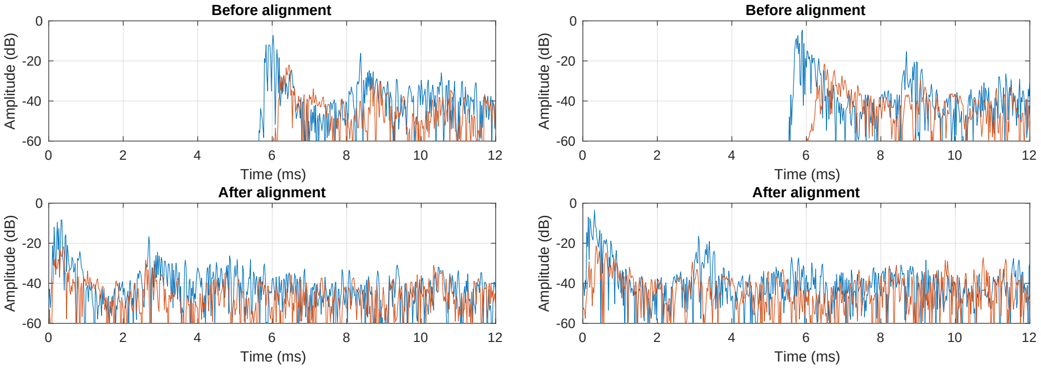
(a) Predicted time-of-arrival with varying θ_h.

(b) Onset estimation with $\theta_h = 270°$.

(c) Onset estimation with $\theta_h = 84°$.

(d) Onset estimation with $\theta_h = 90°$.

Figure A.18: Onset estimation for BRIRs measured at the left ear of KU100 dummy head microphone with a frontal loudspeaker $\theta_s = 0°$ and varying head yaw θ_h, using the log-threshold method with a threshold of −20 dB and 10 times upsampling. Two pre-processing methods are compared to the use of the unprocessed IR; using a 3 kHz low-pass filter or using the analytic signal envelope. The Ziegelwanger and Majdak (2014) model is also fitted to estimations from the envelope.



(a) BRIR alignment for $\theta_s = 45°$ (b) BRIR alignment for $\theta_s = 110°$

Figure A.19: Onset alignment of BRIR measurements using modelled TOAs with 5-sample onset preservation margin.

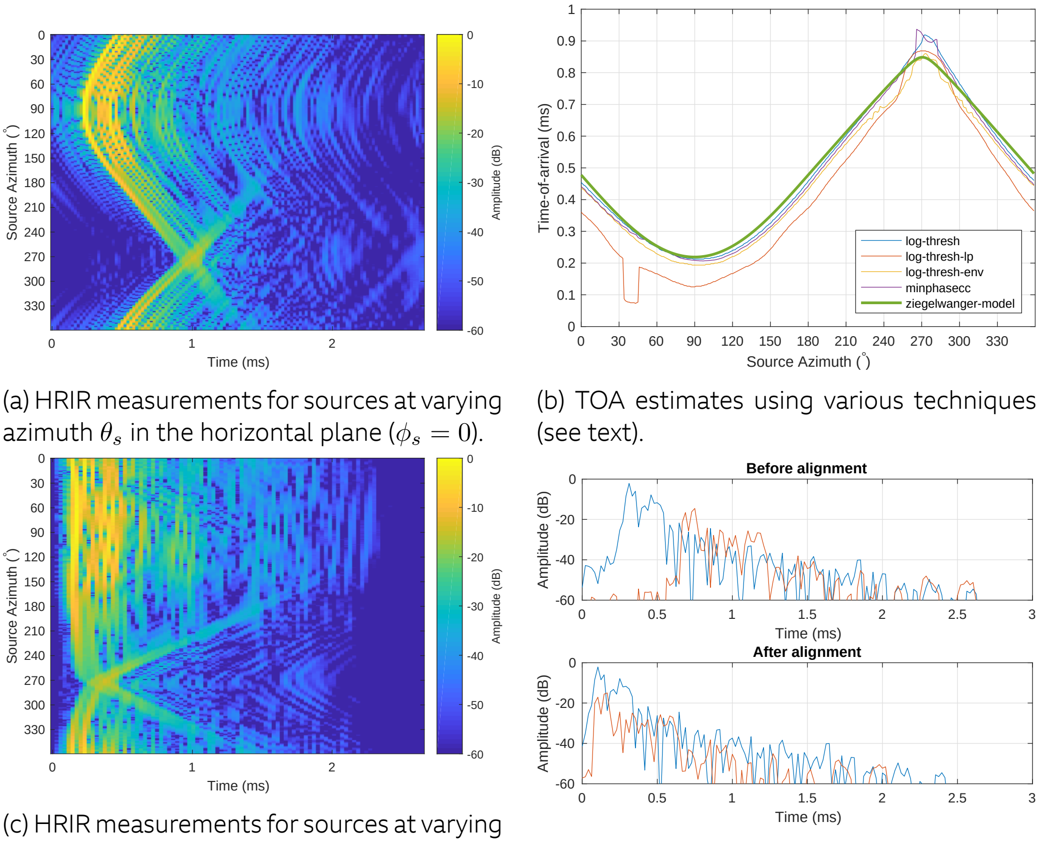
sition, which causes slight misalignment of the peak values with the ipsilateral measurement. However, because BRIRs will be cross-faded with adjacent measurements for the same ear, a smooth function is the primary concern. Erratic estimates in TOA would cause misaligned peaks in adjacent measurements, as well as large jumps in delay which would need to be interpolated rapidly during rendering.

For HRIRs, a similar approach can be taken to ensure a smooth TOA function. Figure A.20 illustrates the estimation of onsets for HRIRs measured on the Neumann KU100 dummy head for sources at different azimuths in the horizontal plane, using data from (Bernschütz, 2013). The same log-threshold method was applied, with low-pass filter and envelope pre-processing, however a threshold of −10 dB was needed for stable results with this data (avoiding large erratic jumps in TOA over azimuth). The commonly used minimum-phase cross-correlation method is also shown (Nam et al., 2008), as well as the output of the Ziegelwanger and Majdak (2014) model when using this as the initial estimate.

The benefit of re-inserting extracted TOAs from HRIRs, without using minimum-phase functions, is that the original full-phase HRIR can be reconstructed. This reduces the chance of errors where the estimated TOA might not correspond well to the perceptually correct value.

A.6 Headphone Correction

The goal of headphone correction filtering in binaural systems is to compensate for the linear effects of the headphone and its coupling to the ear, so that the cues within the binaural signals are precisely recreated (see section 2.7). Ideally the correction of the headphone-



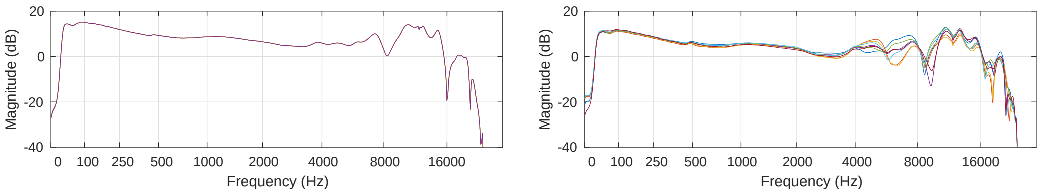
(a) HRIR measurements for sources at varying azimuth θ_s in the horizontal plane ($\phi_s = 0$).

(b) TOA estimates using various techniques (see text).

(c) HRIR measurements for sources at varying azimuth after alignment.

(d) HRIR alignment for $\theta_s = 45°$.

Figure A.20: TOA estimation and alignment for HRIR measurements of the Neumann KU100.



(a) Four HpTFs measurements without headphone repositioning.

(b) Seven HpTFs measurements with headphone repositioning.

Figure A.21: The effect of headphone repositioning on the HpTF. Measured on the right ear of the Neuman KU100 using Beyerdynamic DT990 circumaural headphones.

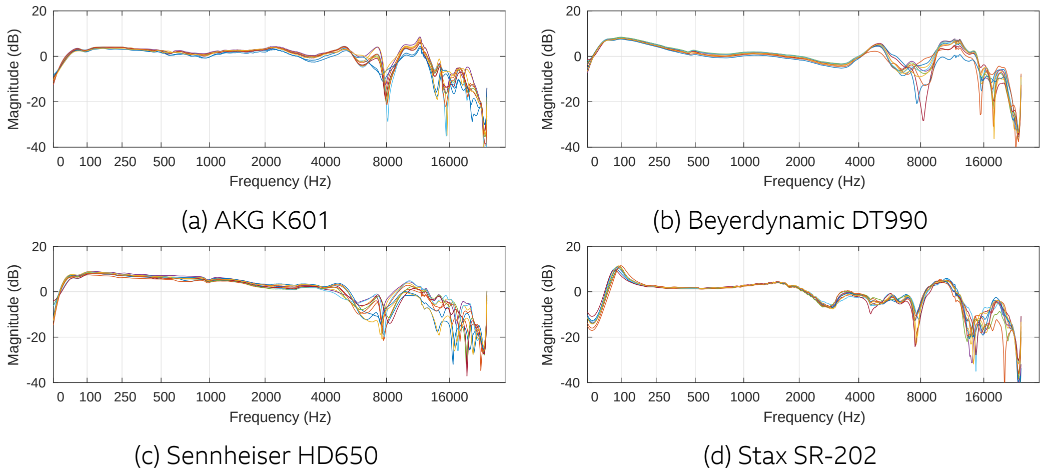
to-ear transfer function (HpTF) would result in a magnitude response of one and phase response of zero at all frequencies i.e. a unit impulse.

This section presents investigations into headphone correction filter design algorithms for use in uncontrolled environments, which were reviewed in section 2.7.5. A set of HpTFs were measured and various state-of-the-art algorithms have been implemented and evaluated in MATLAB. A headphone equalisation filter can be generated using these tools and applied to a database of HRIRs or BRIRs before use in the real-time binaural rendering system.

A.6.1 Headphone-to-Ear Transfer Functions

The variability in HpTF due to headphone positioning can be seen in figure A.21 for measurements on the Neumann KU100 dummy head microphone. With repositioning between measurements, large variations can be observed above 4 kHz, particularly the position and depth of spectral notches. When measurements are made without repositioning the headphones, the measurement variation is minimal. If a human subject were used there might be more variance observed without headphone repositioning. However, it is highly unlikely to be as large as the variance observed with repositioning.

Subsequently HpTFs were measured for four different headphone models. Ten measurements were made for each headphone model on the KU100, with removal and replacement of the headphone between measurements. Figure A.22 presents the magnitude responses of these measurements. Clear differences can be seen between models, despite the positioning variability.



(a) AKG K601

(b) Beyerdynamic DT990

(c) Sennheiser HD650

(d) Stax SR-202

Figure A.22: The effect of headphone model on the HpTF. Measured on the right ear of the Neuman KU100, 10 repetitions with removal and replacement of the headphone each time.

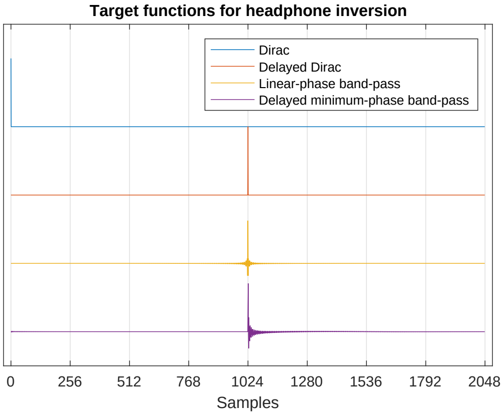
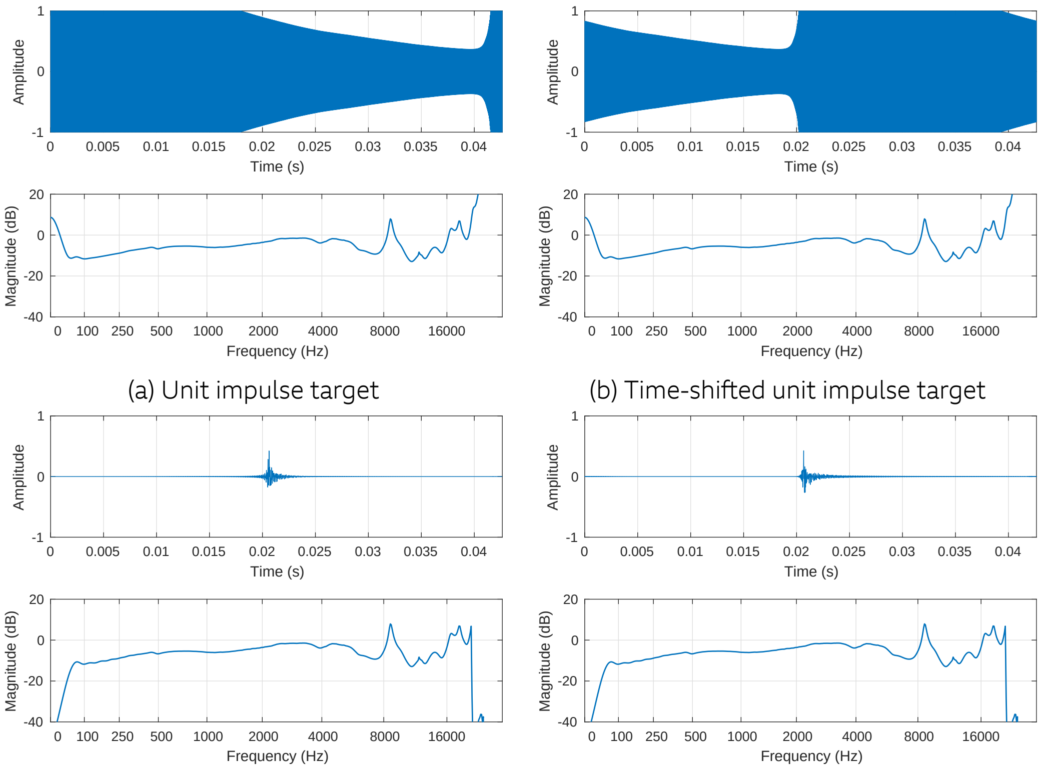


Figure A.23: Headphone correction filter target functions

A.6.2 Basic Least-Squares Inversion

To illustrate the requirements for more sophisticated filter design approaches, basic least-squares inversion of a single HpTF measurement is shown in figure A.24. The measurement was made with the Beyerdynamic DT990 headphones and truncated to 2048 samples prior to inversion. Four different target functions have been used (shown in figure A.23), but no additional regularisation or pre-processing has been applied.

The lack of energy at very low and very high frequencies in the HpTF leads to high energy in the HpCFs when using a unit impulse target (Figures A.24a and A.24b). The range of the vertical axes in figure A.24 is common across the different subfigures, so the response is clipped for these HpCFs. Time shifting the target function prevents the acausal pre-ringing



(a) Unit impulse target

(b) Time-shifted unit impulse target

(c) Linear-phase band-pass target

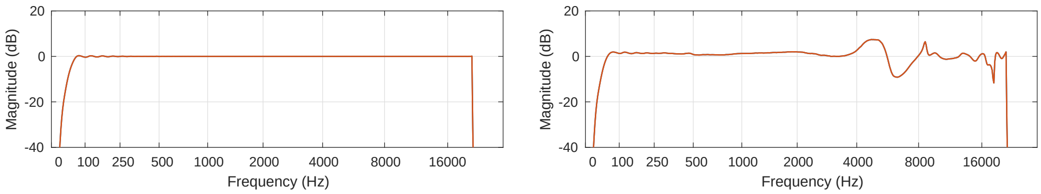
(d) Minimum-phase band-pass target

Figure A.24: Headphone correction filters obtained by least-squares inversion of a single HpTF measurement with different target functions.

effects that have wrapped around at the end of the filter to some extent. Due to this high energy, however, the filter impulse response is still too long, so significant pre-ringing effects occur with time-shifting. Note that time shifting before the inversion gives identical results to time-shifting after the inversion.

The band-pass target functions have a pass-band from 50 Hz to 21 kHz with 60 dB stop-band attenuation, as in Schärer and Lindau, 2009. The filters generated with linear-phase and minimum-phase band-pass targets are shown in Figures A.24c and A.24d respectively. These filters are much more suitable for applying headphone correction. The energy and length of the HpCFs is reduced because the excessive gain at low and high frequencies is not required. Both of these filters preserve the equalisation within the target pass-band. The filter design with a minimum-phase target shows reduced pre-ringing compared to the linear phase target.

If the headphones are moved, sharp notches in the HpTF may be changed, as can be seen in figure A.21. The problem with the direct inversion of a single HpTF is that sharp peaks in



(a) Without repositioning.

(b) After repositioning.

Figure A.25: Residual error after headphone correction filtering on the magnitude response of the HpTF, before and after positioning. Using a basic least-squares inverse with linear-phase band-pass target function.

the pass-band, such as the one at 8.7 kHz, will no longer align with a notch of the same shape and magnitude and therefore will cause a peak in the residual error function. Figure A.25a shows the residual magnitude response after applying the HpCF shown in figure A.24c to the HpTF on which it was based, whereas figure A.25b shows the result of applying the same filter to another HpTF measured after repositioning the headphones. Significant errors are introduced, including a sharp peak at 8.7 kHz.

A.6.3 State-of-the-art Techniques

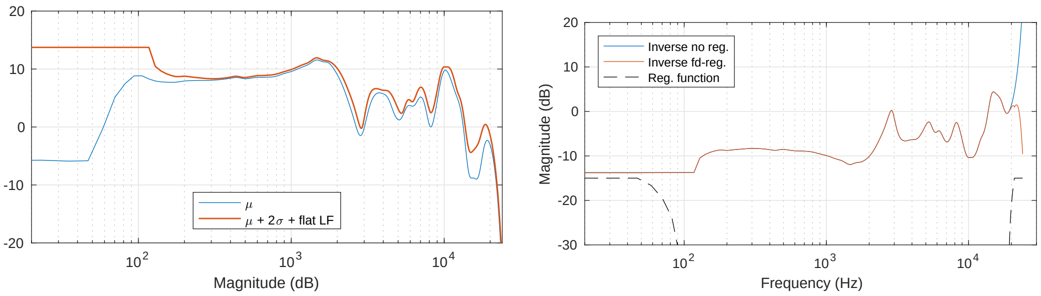
Several techniques have been presented in the literature to provide suitable correction of the HpTF in spite of the variability found, as reviewed in section 2.7.5. A number of these were implemented in MATLAB and are evaluated in this section. For this comparison, 20 HpTFs were measured on the Neumann KU100 dummy head microphone with Stax SR-207 headphones. The headphones were removed and replaced on the head each time. The impulse responses were truncated to 4096 samples before analysis.

Two methods presented by Masiero were implemented, which both use pre-processing techniques to reduce notches in the magnitude response to be inverted. Additionally two methods presented by Lindau and Brinkmann (2012) were also implemented, which use a band-pass target function and use regularisation to reduce peaks in the resulting filter.

A.6.3.1 Magnitude Response Pre-Processing

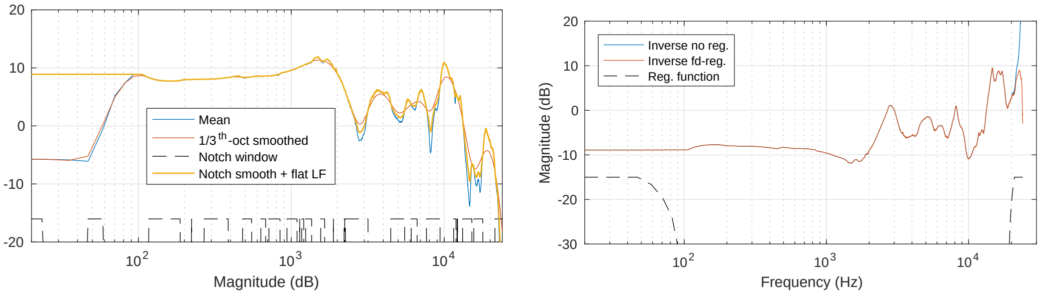
Masiero presented "perceptually robust" methods for HpCF design, which are based on pre-processing of the function to be inverted. In each case, a magnitude response is derived with the aim of avoiding residual peaks after equalisation, acknowledging the variability of the HpTFs and perceptual sensitivity to resonant peaks.

In Masiero and Fels, 2011, the magnitude responses of the multiple measured HpTFs are



(a) Magnitude response pre-processing (b) Inversion with regularisation effect

Figure A.26: HpCF design using mean plus two standard deviations of the measured magnitude responses with 1/6$^{\text{th}}$-octave smoothing (Masiero and Fels, 2011).



(a) Magnitude response pre-processing (b) Inversion with regularisation effect

Figure A.27: HpCF design by locally smoothing notches in the mean magnitude response (Masiero, 2012).

smoothed with a 1/6$^{\text{th}}$-octave filterbank. From this set of smoothed magnitude responses, the mean plus two standard deviations is used for inversion, so reducing the notches found in the mean which are not common across measurements. Figure A.26 shows the design process for this filter, with the magnitude response pre-processing and the regularised inversion steps. This is the HpCF design approach that was used in chapter 5.

In Masiero (2012, p.72), instead a local notch-smoothing technique is used to process the mean magnitude response. First the mean magnitude response is smoothed with a 1/3$^{\text{rd}}$-octave filterbank. This smoothed function is compared to the original, in regions where it exceeds the unsmoothed response, a notch is detected and the two functions are mixed by a given factor, thereby locally smoothing the original function. The mix factor was set at 0.5 for this filter design. Figure A.27 shows the notch-smoothing process and the effect of the regularisation during filter inversion.

Both of these techniques were implemented in MATLAB and used to design HpCFs. Following Masiero (2012, p.72), the low frequency response was flattened below the first peak to avoid excessive amplification at frequencies below which the headphone is incapable of reproduction. A minimum-phase transfer function was obtained from the derived magnitude responses for inversion. Least-squares inversion was performed with a small amount of regularisation at frequencies above 21 kHz and frequencies below 50 Hz, in order to reduce the risk of excessive gain outside of the range of reproduction. The regularisation function was set at −15 dB in these regions. The target function was a Kronecker delta unit impulse at time zero, because the function being inverted was already minimum-phase.

The resulting HpCFs were circularly shifted in time to ensure any pre-ringing was not present at the end of the filter and then temporally windowed. Finally, a normalisation was performed in the frequency range 150 Hz to 4 kHz.

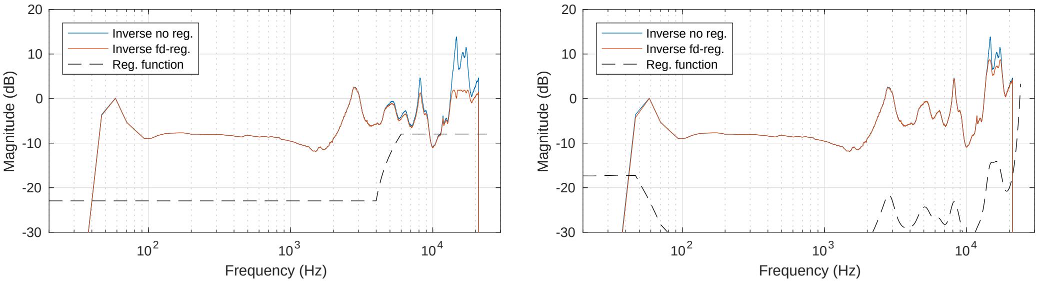
A.6.3.2 Regularisation Approaches

Lindau and Brinkmann (2012) evaluated several filter design methods that used a frequency-dependent regularisation function to reduce the compensation effort in regions where sharp notches occur. This was combined with a band-pass target function, such that the HpCF removed very low and high frequency content, assuming that the headphones cannot reproduce these frequencies well anyway.

The mean HpTF was obtained from the measurements for inversion. A high-shelf regularisation function was used, with 15 dB gain above 6 kHz, based on the knowledge that most sharp notches occur in the high frequency range. As an alternative, a 1/6$^{\text{th}}$-octave smoothed version of the mean magnitude response was used as the regularisation function. The regularisation parameter β was set to 0.4 and 0.07 for the high-shelf and smoothed-inverse approaches respectively, after Schärer and Lindau (2009). Figure A.28 shows the two regularisation functions and their effect on the magnitude response resulting from the inversion. For both filter designs, a minimum-phase band-pass target function was used with cut-off frequencies at 50 Hz and 21 kHz and −60 dB stop-band attenuation. Lindau and Brinkmann (2012) found that listeners were not sensitive to the phase response of the filter design. As with the magnitude pre-processing methods, the resulting HpCFs were circularly shifted in time, temporally windowed and then normalised in the frequency range 150 Hz to 4 kHz.

A.6.3.3 Auditory Filterbank Evaluation

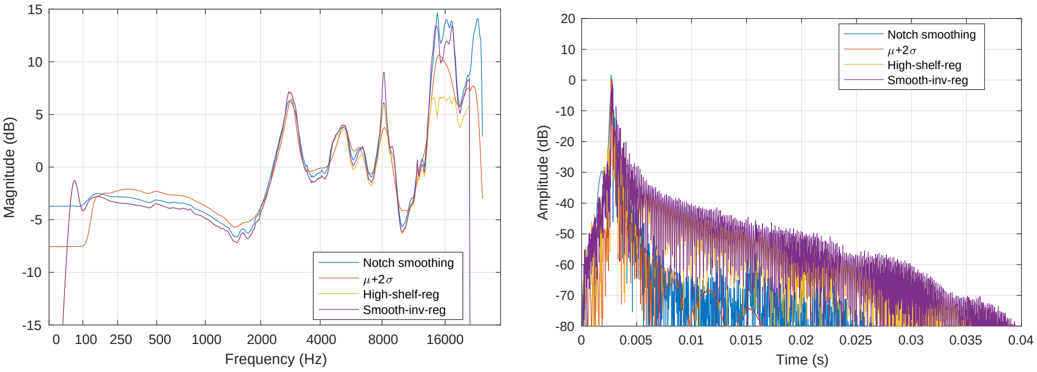
The HpCFs designs described above are plotted together in figure A.29. They were evaluated using the original HpTFs measured on the Neumann KU100, from which the HpCF designs were based. Following the approach of Lindau and Brinkmann (2012), each HpTF was convolved with the HpCF and subsequently processed with a filterbank of 44 equivalent



(a) High-shelf regularisation function

(b) Smoothed inverse regularisation

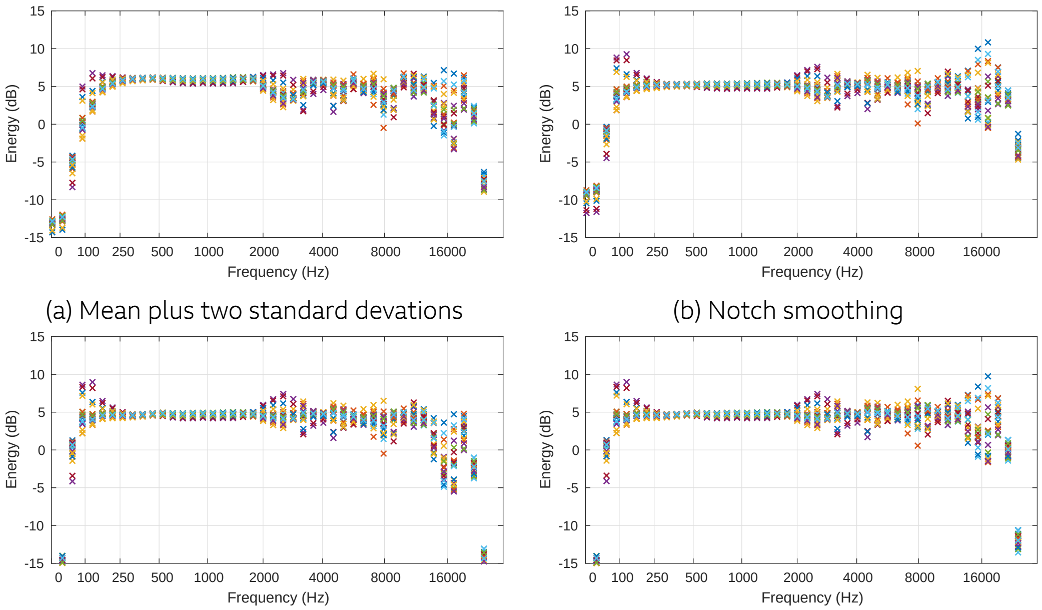
Figure A.28: HpCF design by inversion of the mean HpTF magnitude response with frequency-dependent regularisation function and minimum-phase band pass target function (Lindau and Brinkmann, 2012).



(a) HpCF magnitude responses

(b) HpCF IRs

Figure A.29: Comparison of HpCF designs.



(a) Mean plus two standard devations

(b) Notch smoothing

(c) High-shelf regularisation

(d) Smoothed-inverse regularisation

Figure A.30: Auditory filterbank analysis of equalisation accuracy for each HpCF design, using the original twenty measured HpTFs on the Neumann KU100.

rectangular bandwidth (ERB) gammatone filters, which represents the auditory frequency resolution. The energy of the output of these filters was analysed and the distribution of results is plotted for each HpCF design in figure A.30.

It is clear that none of these filters will produce perfect correction. All perform well in the range 250 Hz to 2 kHz where there is low variability. The mean plus two standard deviations approach appears effective at avoiding peaks in the error function, but besides that it is difficult to differentiate the filters from this analysis. Indeed the perceptual evaluation by Lindau and Brinkmann (2012) showed little discrimination between different non-individual HpCFs using such designs.

Figure A.29b indicates that the band-pass target function yields longer filter impulse responses. Looking at the filterbank analysis, this target function seems not to provide added benefit over the frequency-dependent regularisation that was used with the magnitude response pre-processing designs.

A.6.4 Perceptually-Adjusted Filter Design

The filter designs presented in appendix A.6.3 may not be well differentiated. Informal listening confirmed this. With all of these HpCFs, dynamic binaural rendering with measured

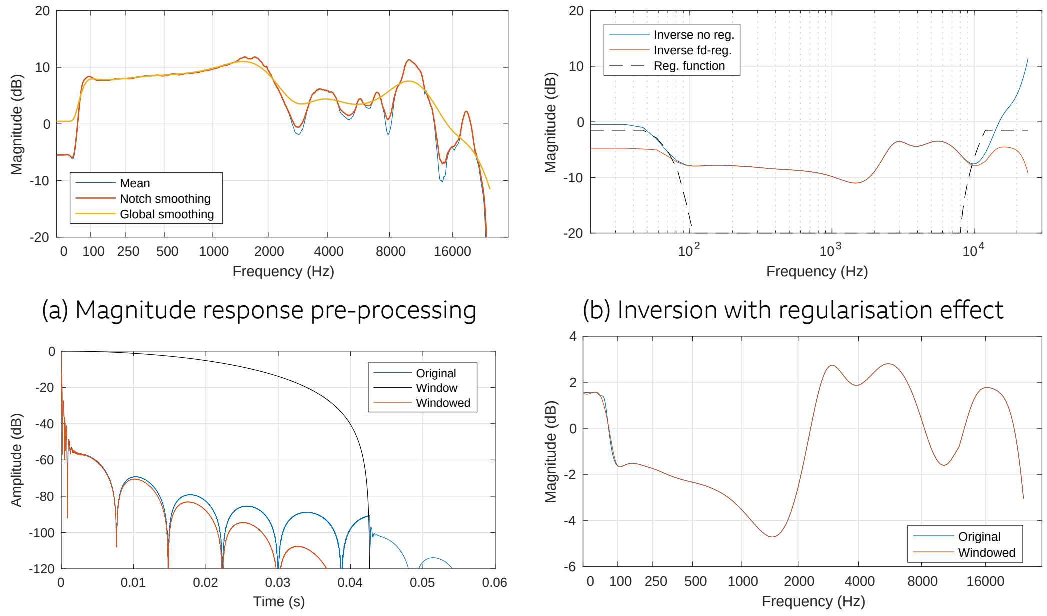
BRIRs showed clear timbral differences when compared directly with a real loudspeaker. Rather than carry out a perceptual evaluation to compare these filters, a HpCF was obtained by perceptual tuning of parameters. Two experienced critical listeners, this author and a colleague, adjusted the parameters of the design manually. The filter was assessed by comparison between a real loudspeaker, placed directly in front of the listeners, and the simulation of this loudspeaker by dynamic binaural rendering with measured BRIRs, equalised with the HpCF. This process was performed in the BBC R&D listening room. The Stax SR-207 headphones were worn during binaural rendering but removed when listening to the loudspeaker. This approach was chosen rather than using BRIRs measured with the headphones on the dummy head as in (Schärer and Lindau, 2009), because this was found to make the adjustment task more challenging and the binaural rendering generally less convincing.

The notch-smoothing and high-shelf regularisation techniques were combined in the filter design, as shown in figure A.31. Following notch-smoothing, the mean magnitude response was significantly smoothed overall, using a 2/3rd-octave smoothing. The notch-smoothing did still have a small effect. A minimum-phase transfer function was derived from this smoothed magnitude response for inversion. The regularisation function was set at −1.5 dB, both above 12 kHz and below 50 Hz, and was set at −20 dB level between 100 Hz and 8 kHz. A temporal window was applied to truncate the filter length from 4096 to 2048 samples without significantly affecting the magnitude response (see Figures A.31c and A.31d). The filter gain was finally adjusted to normalise the root mean square (RMS) level in the frequency range 150 Hz to 4 kHz.

Informal subjective evaluation during the tuning process suggested that this filter design gave a better approximation of the real loudspeaker when compared to the previously presented HpCF designs. Those all appeared to over emphasise energy in the high-mid range (approximately 2 kHz–5 kHz) and at high frequencies (above 10 kHz), as well as often lacking bass. The perceptually-tuned HpCF also better approximated the real loudspeaker compared to the case with no HpCF, particularly in the lower-mid and mid-frequency range (i.e. 250 Hz to 2 kHz). The virtual loudspeaker was also more easily localised with the HpCF, with a narrower source extent, closer to that of the real loudspeaker. However there were still noticeable differences in timbre to the real loudspeaker source.

A.6.5 Individual HpTF Measurements

To understand these differences further, HpTF measurements were made on a single human subject (this author) and subsequently analysed. This individual was also involved in the perceptual adjustment of the Neumann KU100 HpCF design.



(a) Magnitude response pre-processing

(b) Inversion with regularisation effect

(c) Temporal windowing of HpCF

(d) Windowing effect on magnitude response

Figure A.31: Perceptually-tuned HpCF design, using 2/3$^{\text{rd}}$-octave smoothed mean HpTF magnitude response and high-shelf regularisation.

Small omnidirectional electret microphones (Sennheiser KE-411-2) were mounted in silicon ear inserts, after Brinkmann (2011). This enables binaural measurements to be made with more repeatability than when using foam ear inserts, since the fitting is more consistent. The mould was designed using morphological data of the ear canals of a large number of individuals, obtained from a hearing aid manufacturer. Brinkmann kindly shared the 3D model of the negative mould, and it was 3D printed at the University of York.

Figure A.32 shows one of the microphones in place at the entrance to the ear canal of a human subject. Correction filters were created for each microphone in its silicon insert, to account for small deviations from a flat response. Impulse responses were measured in the anechoic chamber at the University of York with a Genelec 8040A loudspeaker. The measurements were first corrected for the response of this loudspeaker, since it was also measured with an Earthworks M30 reference microphone. These correction filters were applied to the HpTF measurements before analysis.

Seven HpTF measurements were made on the human subject, with reseating of the headphones each time. A set of seven new measurements was also made on the Neumann KU100. The magnitude responses of these measurements are plotted in figure A.33. There are clear differences between the two sets of measurements. The broad spectral differences

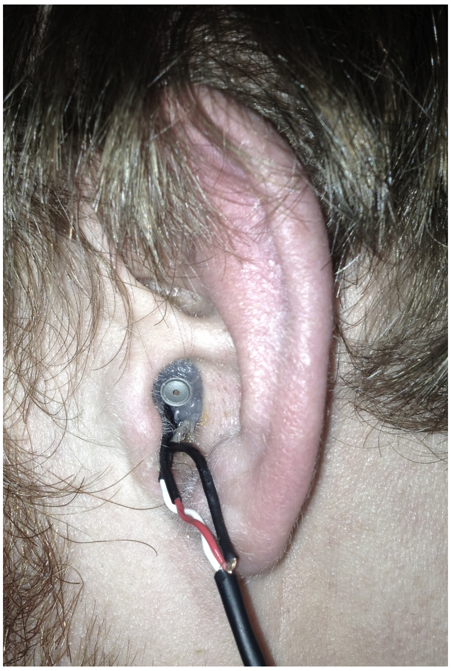


Figure A.32: Sennheiser KE-411-2 Silicon Microphone Ear Insert

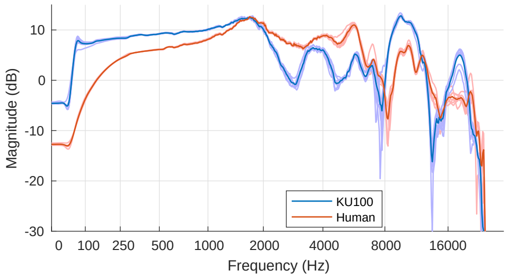


Figure A.33: Comparison of seven right-ear HpTF measurements on the Neumann KU100 and a human subject, with the mean magnitude response shown in colour.

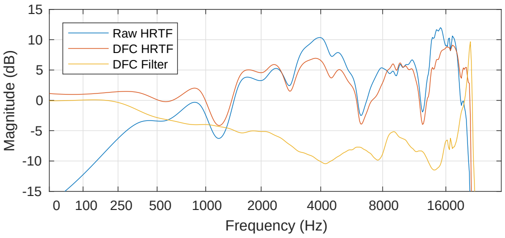
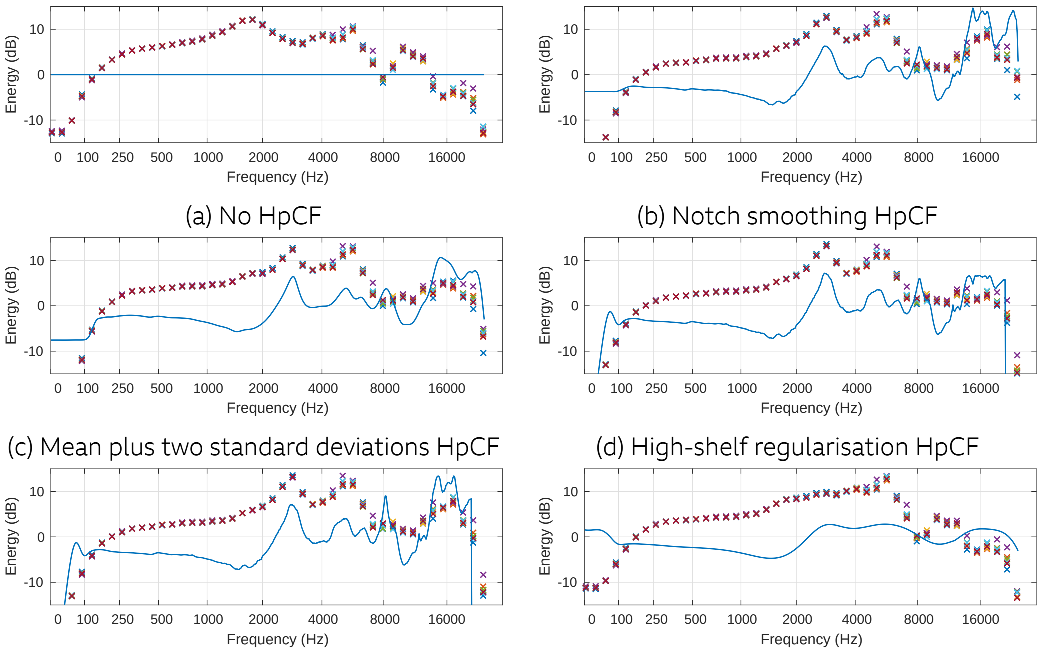


Figure A.34: Example of diffuse-field compensation (DFC) for measurements on human subject 003 from the SADIE database (Kearney and Doyle, 2015a)

are likely due to the diffuse-field equalisation of the Neumann KU100. The KU100 HpTFs are flatter in the mid-frequency range. The Stax SR headphones have previously been shown to have a magnitude response that approximates the typical diffuse field HRTF (Christensen, Hess, et al., 2013). To give further insight into this, figure A.34 shows an example diffuse-field equalisation filter for a human subject from the SADIE database (Kearney and Doyle, 2015a) and a frontal HRTF before and after diffuse-field equalisation. The differences between these magnitude responses have similarities to the differences between HpTFs in figure A.33. Despite these broad differences, similar patterns can be observed in the spectral shape between the KU100 and the human HpTF measurements. This includes notches at around 8 kHz and 14 kHz–15 kHz in both measurement sets, though these are lower in frequency for the KU100 data.

The auditory filterbank analysis was repeated with these human HpTF measurements, as shown in figure A.35. The HpCFs evaluated were all created using the Neumann KU100 measurements, as described in previous sections. The broad spectral differences due to diffuse-field equalisation of the KU100 can be seen in the errors for all HpCFs, leading to a boost at around 3 kHz. However the perceptually tuned design leads to a smoother residual response, avoiding narrowband boosts in the high frequency range which can be seen with the other filters.

Considering that headphones equalised to give a diffuse-field HRTF response are desirable when the binaural recordings/filters are diffuse-field equalised (Jot, Larcher, et al., 1995), the general shape of the errors observed in figure A.35 should not be flat. It should approximate the diffuse-field response of the human listener. This can be challenging to measure, either requiring a reverberation chamber or many HRTF measurements in an anechoic chamber (see section 2.4.3.2). Møller, Jensen, et al., 1995 presented a diffuse-field headphone design goal based on the average diffuse field HRTF of 40 human subjects, which was incorporated into ISO 11904-1 (DIN EN, 2002). This curve, when measured at the blocked ear canal entrance, has a steady rise from 0 dB at 0 Hz to a peak of 10 dB at around



(a) No HpCF

(b) Notch smoothing HpCF

(c) Mean plus two standard deviations HpCF

(d) High-shelf regularisation HpCF

(e) Smoothed-inverse regularisation HpCF

(f) Perceptually-tuned HpCF

Figure A.35: HpCF designs, based on Neumann KU100 HpTF measurements, evaluated with an auditory filterbank after application to HpTFs measured on a human individual. The HpCFs are shown by the blue lines and the crosses indicate the energy of output from ERB filters during analysis.

4 kHz–5 kHz and a steady decay down to 0 dB at 20 kHz. The auditory filterbank energy distribution of the perceptually tuned HpCF most closely resembles this target function, of all of the evaluated HpCFs. Although the notches at 8 kHz and 15 kHz are not adequately corrected for.

Further refinement of the perceptual adjustment procedure might be achieved using manual equalisation by several skilled listeners. For example by using parametric equalisation or adjusting the level in 1/3$^{\text{rd}}$-octave bands, to better match the binaural rendering to a real loudspeaker, as in (Fleischmann, Silzle, et al., 2012). In contrast the present approach involved only tuning the HpCF filter design parameters such as regularisation and smoothing.

During this investigation, an individual HpCF was also designed from the human HpTF measurements. Notch reduction, smoothing and high-shelf regularisation were used as with the perceptually tuned non-individual filter, though only 1/6$^{\text{th}}$-octave smoothing was used and the high-shelf regularisation function was lower, at −10 dB, in this case. Also flattening of the magnitude response was performed below 150 Hz and above 21 kHz before inversion.

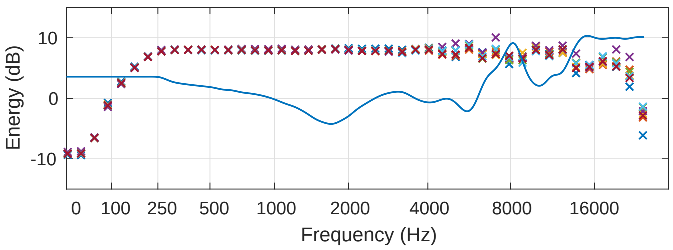


Figure A.36: Individual HpCF evaluated with an auditory filterbank using HpTFs measured on the same individual. The HpCF is shown by the blue line and the crosses indicate the energy of output from ERB filters during analysis.

The HpCF and results of auditory filterbank analysis are shown in figure A.36. This filter was found by this author to give closer approximation to the real loudspeaker during listening than the non-individual HpCFs, particularly in the low- and mid-frequency range. However it lacked energy in the upper mid-frequency range compared with the real loudspeaker, where the diffuse field target function has its peak.

From this rather informal evaluation, it seems likely that individual headphone equalisation will give better results for realistic binaural simulation, as has been found in the literature. But this also highlights that headphone correction to a flat frequency response is not necessarily appropriate. The target response curve should be designed considering the equalisation of the binaural recording or the filters used in rendering.

A.7 Interpolation of HRTFs with Spherical Harmonics

Besides the nearest-neighbour filter selection technique described in appendix A.3.1, it is also desirable to have a means of generating HRTFs by interpolation for arbitrary directions. Section 2.6 reviewed HRTF interpolation techniques, showing that they can be used to significantly reduce the required measurement resolution. Also, in applications such as sound field analysis and rendering HRTFs may be required on a precise grid of positions for which there are no available measurements.

The Neumann KU100 HRTF measurement sets presented by Bernschütz (2013) provide several high-density gapless spherical sampling grids, using quadratures that allow precise integration of the spherical harmonics up to a certain order. This provides a valuable dataset for analysis of global interpolation using spherical harmonics, as introduced in section 2.6.3.

This section first presents the mathematical formulation of spherical harmonic interpolation of HRTFs, then presents an analysis of the KU100 data. Finally listening experiments are presented to validate the interpolation method.

A.7.1 Mathematical Formulation

A.7.1.1 Spherical Wave Spectrum

When a body scatters sound from a source at position (r_s, θ_s, ϕ_s) the complex pressure at any point (r, θ, ϕ) is known to satisfy the Helmholtz equation (Duraiswaini et al., 2004)

$$(\triangle + k^2)\, p(r, \theta, \phi, k) = 0 \tag{A.2}$$

where $\triangle$ is the Laplace operator, the wavenumber is $k = 2\pi f/c$ and c is the speed of sound.

Outside of the surface of the scattering body, which contains all of the acoustic sources within the environment, the pressure field can be expanded as a series of multipoles or spatial modes which can be called the spherical wave spectrum. This assumes that inward travelling waves are negligible.

The spherical wave spectrum is defined as

$$p(r, \theta, \phi, k) = \sum_{n=0}^{+\infty} \sum_{m=-n}^{n} a_{nm}(r, k) Y_n^m(\theta, \phi) \tag{A.3}$$

$$a_{nm}(r, k) = b_{nm}(k) h_n(kr) \tag{A.4}$$

Here h_n is the spherical Hankel function of the first kind, associated with the outgoing radial component of the sound field, Y_n^m are the spherical harmonics, associated with the directional component of the sound field, and a_{nm} are the spherical expansion coefficients, with order n and degree m. These indices characterise each multipole, which is the product of the spherical harmonic of order n and degree m and the corresponding Hankel function of order n. b_{nm} denotes the spherical harmonic coefficients, which are the directional component of the spherical expansion.

The orthonormal complex-valued spherical harmonic basis functions are defined by

$$Y_n^m(\theta, \phi) = (-1)^m \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-|m|)!}{(n+|m|)!}}\, P_n^{|m|}(\cos\phi)\, e^{im\theta} \tag{A.5}$$

where P_n^m are the associated Legendre functions

$$P_n^m(x) = \frac{(1-x^2)^{\frac{m}{2}}}{2^n\, n!} \frac{d^{n+m}}{dx^{n+m}} (x^2 - 1)^n \quad \text{for } m \geq 0. \tag{A.6}$$

A.7.1.2 Modelling of HRTF data

The assumption in forming equation (A.3) is that the region outside a given surface contains no acoustic sources, this then allows the pressure at any point in the pressure field to be obtained. With HRTF data this is not the case, instead we have a (virtual) array of loudspeakers around the head and are measuring the pressure on the surface of the head, at e.g. the blocked entrance to the ear canals. In order to calculate the spherical wave spectrum of the HRTF, the reciprocity principal is used (Duraiswaini et al., 2004), i.e. if an acoustic source at point A creates pressure p at point B, then the same acoustic source placed at point B will create the same pressure at point A. This means that an identical HRTF measurement would be obtained if the loudspeaker were placed at the point of the microphone in the ear canal and the microphone were placed at the location of the loudspeaker, assuming ideal transducers.

The projection of measured HRTF data onto the spherical harmonic expansion a_{nm} is computed using integration over the sphere[5]

$$\tilde{a}_{nm}(r_0, k) = \int_{\mathbb{S}^2} p'(r_0, \theta, \phi, k) Y_n^{m*}(\theta, \phi) \mathrm{d}\Omega \tag{A.7}$$

where p' denotes the measured pressure response on the surface of the sphere with radius r_0, the symbol ˜ indicates an estimate of a parameter from the projection of measured data, and * denotes the complex conjugate. This is known as the spatial Fourier transform (SFT). The spherical expansion is performed separately for the left and right ears.

In practice the expansion is limited to series orders $n \leq N$, which corresponds to a spatial bandwidth limitation. Pressure fields that exceed this order can therefore not be adequately represented and spatial aliasing occurs. Equation A.7 also requires a continuous measurement of the pressure on the surface of the sphere. In practice this must normally be discretised to a set of measured points. For this we use the discrete spatial Fourier transform (DSFT). With a discrete set of L measurement points and defined sample weighting, i.e. a *quadrature*, which is known to precisely integrate the spherical harmonics up to order N, the integral is evaluated as

$$\tilde{a}_{nm}(r_0, k) = \sum_{l=1}^{L} w_l p'(r_0, \theta_l, \phi_l, k) Y_n^{m*}(\theta_l, \phi_l) \tag{A.8}$$

where w_l indicates the sample weight for the l^{th} measurement point. The discrete pressure

[5] Note that $\int_{\mathbb{S}^2} \mathrm{d}\Omega = \int_0^{2\pi} \int_0^{\pi} \cos\phi d\phi d\theta$ indicates the integration of the solid angle Ω over the surface of the 2-sphere $\mathbb{S}^2$.

field measurements p' are the HRTF measurements.

The synthesis of HRTFs at arbitrary points can be obtained using the (truncated) inverse spatial Fourier transform (ISFT):

$$\tilde{p}(r,\theta,\phi,k) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \tilde{a}_{nm}(r,k) Y_n^m(\theta,\phi) \tag{A.9}$$

Note that this is similar to equation (A.3) but with truncated series order.

A.7.1.3 HRTF Range Extrapolation

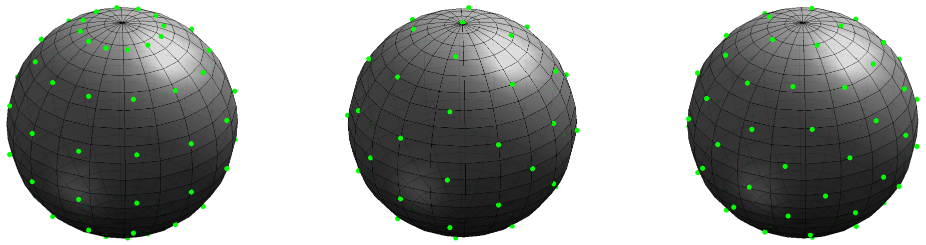
To obtain HRTFs at a distance/range other than that of the measured HRTFs, $\tilde{a}_{nm}(r_0,k)$ is first calculated by the DSFT, followed by a range extrapolation process, and then projecting the spectrum back onto the target direction with the ISFT. The range extrapolation is performed using a ratio of radial propagation functions (spherical Hankel functions of the first kind).

$$\tilde{a}_{nm}(r_1,k) = \tilde{a}_{nm}(r_0,k)\frac{h_n(kr_1)}{h_n(kr_0)}. \tag{A.10}$$

The practical use of this range extrapolation process is somewhat limited by the large oscillations of the high-order spherical Hankel functions at low frequencies, which can cause numerical instability. For this reason it is often recommended to use a frequency-dependent series truncation order (Pollow et al., 2012), especially since the energy at low frequencies is often concentrated in the lower orders. When no range extrapolation is required and simple directional interpolation is used, the radial terms are not required and equation (A.9) can be used directly with $\tilde{a}_{nm}(r_0,k)$.

A.7.1.4 Spatial Sampling Schemes - Quadrature

The DSFT (equation (A.8)) requires a sampling scheme that properly integrates the spherical harmonics up to modal order N. Since the spatial structure of the spherical harmonics gets increasingly detailed at higher orders, a given sampling of the sphere will have a limited modal resolution, above which the spherical harmonics cannot be uniquely resolved. Rafaely (2015, chapter 3) provides an excellent introduction to sampling schemes for spherical functions and Zotter (2009, chapter 4.2) provides a detailed analysis of the properties of various sampling schemes. Bernschütz (2016, chapter 3.2.2) also discusses quadrature on the sphere for analysing the spherical wave spectrum.



(a) Gauss grid ($L = 72$) (b) Lebedev grid ($L = 50$) (c) Spherical t-design ($L = 70$)

Figure A.37: Spherical sampling schemes for $N = 5$, plotted on the sphere surface

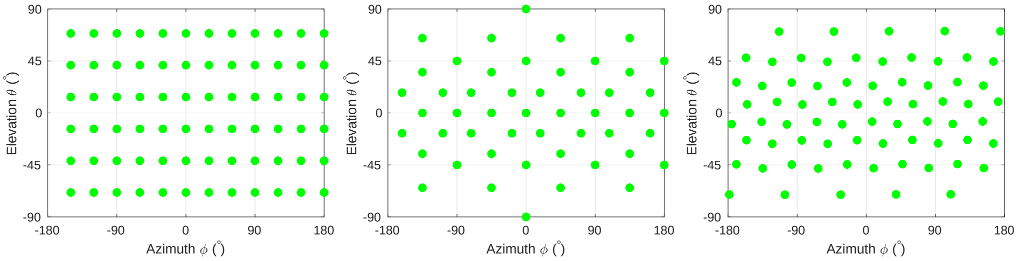
Quadratures provide a grid of sampling points and associated sampling weights and can be directly applied in the DSFT. Quadratures provide orthogonal sampling of spherical polynomials up to order N and so lead to efficient implementation in the DSFT (i.e. they do not require matrix inversion). The sampling weights have the following summation:

$$\sum_{l=1}^{L} w_l = 4\pi \tag{A.11}$$

There are various quadrature schemes, Rafaely (2015) describes equal-angle, Gaussian, and Lebedev quadratures, Zotter (2009) also discusses the Fliege quadratures. Gaussian quadratures are uniform in azimuth and nearly uniform in elevation (sampling at the roots of the Legendre polynomial at order $N + 1$), which means that the sampling is not uniform over the sphere surface itself. Lebedev grids are close to uniformly spaced over the surface of the sphere, though sampled more densely around the Cartesian axes. Uniform and nearly-uniform sampling schemes of the sphere are also available, which have equal weights $w_l = 4\pi/L$. Spherical t-designs are nearly-uniform sampling schemes that can yield a valid integration of a polynomial on the sphere up to order t, with equal sample weights (Hardin and Sloane, 1996). Given that the DSFT involves the product of the spherical expansion coefficients and the spherical harmonics, to preserve the orthogonality of the spherical harmonics, the t-design must be valid for $t \geq 2N$. Zotter and Frank (2012) state that to achieve direction-constant energy $t \geq 2N + 1$.

Figures A.37 and A.38 show sampling grids suitable for integrating functions on the sphere up to order $N = 5$, plotted on a spherical surface and plotted by azimuth and elevation angles, respectively.

The relationship between the number of sampling points L and the maximum valid



(a) Gauss grid ($L = 72$) (b) Lebedev grid ($L = 50$) (c) Spherical t-design ($L = 70$)

Figure A.38: Spherical sampling schemes for $N = 5$, plotted by polar angles θ, ϕ

modal order N_g that can be resolved by the measurement grid is given by

$$L = \eta_q(N_g + 1)^2 \tag{A.12}$$

where η_q indicates the degree of overdeterminacy of the specific sampling scheme (Bernschütz, 2016, chapter 3.2.2). $\eta_q = 1$ is the best possible sampling efficiency, whereas $\eta_q > 1$ indicates some inefficiency. Very few sampling schemes have $\eta_q = 1$ (such as hyperinterpolation (Zotter, 2009)), the Gauss-Legendre quadratures have $\eta_q = 2$, whereas Lebedev quadratures have $\eta_q \approx 1.3$.

A.7.1.5 Sources of Error in Spatially Band-Limited Discrete Systems

If the spatially sampled signal, i.e. the measured HRTF data, has energy in spherical harmonic orders $n > N_g$ then that will not be adequately measured and spatial aliasing will occur, with the signal energy appearing in the lower orders. Rafaely (2015, chapter 3.7) discusses spatial aliasing in spherical arrays. Bernschütz (2016, subsection 3.5) states that “higher orders arise due to air-path delays and complex physical scattering effects around the head”. When the transform order is limited below N_g, additional truncation error occurs, whereby the energy in the spherical harmonic orders above the truncation order is lost.

When range extrapolation is considered, ill-conditioning of the matrix of $a_{nm}(k)$ is also an issue, due to large oscillations in the spherical Hankel functions (Ben Hagai et al., 2011). It is therefore often recommended that the expansion order is limited to $\lfloor kr_h \rfloor$ (related to equation (A.17)), but this would result in truncation error and so a compromise between the two issues is needed.

A.7.1.6 Acoustic Centring

Richter et al. (2014) discuss an approach to reduce the required truncation order by centring the calculations on the receiver, i.e. on the ear, as opposed to the centre of the head, using a geometric adjustment, including range extrapolation, to move the centre point of the measurement grid.

Ben Hagai et al. (2011) presented metrics for assessing the off-centre misalignment of a sound source within a microphone array, based on the spherical harmonic spectrum and Richter et al. (2014) used their “centre of power” metric to optimise the acoustical centre of the measurement grid at each frequency. This has the effect of reducing the energy in the higher-order spherical harmonics and so reduces truncation error or allows for a lower truncation order to be used in synthesis with equivalent error.

Richter et al. (2014) also present considerations for real-time rendering. The mean acoustic centre across the frequency range is used so that only one set of spherical harmonic coefficients is required per synthesis position. A look-up table of discretised Hankel functions with fast linear interpolation is used to reduce synthesis complexity during range extrapolation. In the far-field a simple broadband $1/r$ distance model is used for greater simplicity.

A.7.1.7 Fitting Irregular or Incomplete Measurement Sets with Regularisation

If a set of measurements on a suitable quadrature grid is not available, then a spherical harmonics expansion can be carried out via approximate inversion of a matrix of the spherical harmonics sampling points $Y_n^m(\theta_l, \phi_l)$.

In matrix notation, the ISFT is given by:

$$\mathbf{p} = \mathbf{Y}\mathbf{a} \tag{A.13}$$

and therefore the spherical expansion coefficient matrix can be estimated by:

$$\hat{\mathbf{a}} = \mathbf{Y}^{\dagger}\mathbf{p} \tag{A.14}$$

where † indicates the pseudo-inverse.

The use of Tikhonov regularisation with order-dependent regularisation weights is recommended (Zotkin, Duraiswami, et al., 2009; Pollow et al., 2012):

$$\mathbf{a}_{\text{reg}} = (\mathbf{Y}^{H}\mathbf{W}\mathbf{Y} + \varepsilon\mathbf{D})^{-1}\mathbf{Y}^{H}\mathbf{W}\mathbf{p} \tag{A.15}$$

where ^{H} is the Hermitian (conjugate transpose), **W** is the diagonal matrix composed from the sampling weights vector and **D** is the applied regularisation matrix,

$$\mathbf{D} = (1 + \mathbf{n}(\mathbf{n} + 1))\mathbf{I} \tag{A.16}$$

where **I** is the identity matrix and **n** is a vector of the order of the columns of matrix **Y**. Ahrens, Thomas, et al. (2012) proposes an alternate method for partial-sphere HRTF sets that avoids the issue of regularisation reducing accuracy within regions with known measurements. Although Neumann KU100 measurements are available at suitable quadratures, use of small amounts of regularisation may still improve the spherical expansion results since there will be measurement noise present.

A.7.2 Analysis of the Neumann KU100 Far-field HRTF Dataset

Bernschütz (2013) provides HRTF measurements of the Neumann KU100 dummy head microphone on Lebedev quadratures with 2354 and 2702 points, as well as a Gauss-Legendre quadrature with 16 020 points, corresponding to a 2° spacing in azimuth and a similar resolution in elevation. The 2354- and 2702-point Lebedev quadratures have $N_g = 41$ and $N_g = 44$ respectively, whilst the 16 020-point Gauss-Legendre quadrature has $N_g = 89$.

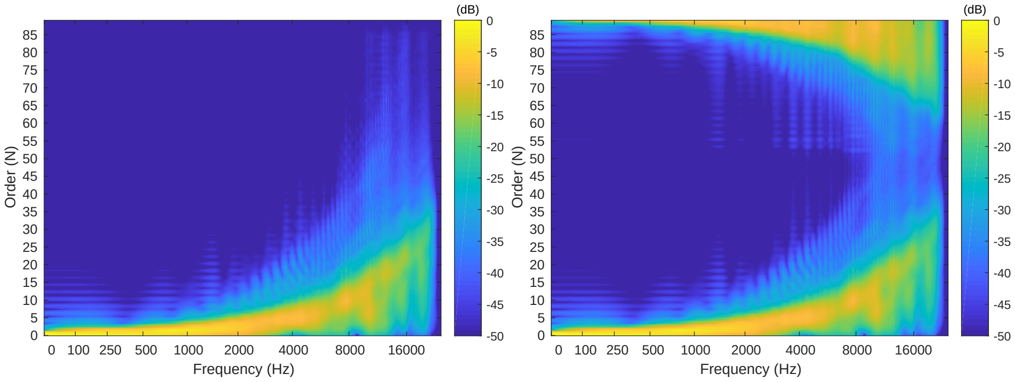
Bernschütz (2016, chapter 3.12.4) provides the following formula for estimating maximum required spherical harmonic series order

$$N \approx \frac{\omega}{c} r_h \tag{A.17}$$

where ω is the maximum angular frequency of interest, c is the speed of sound, and r_h is the radius of a virtual sphere located at the centre of the head and which spans the pinnae. For $r_h = 9.5$cm (valid for Neumann KU100) and $\omega = 2\pi.20$ kHz, $N \approx 34.8$, so a truncation order of 35 is appropriate. Therefore, all of the quadratures measured in Bernschütz (2013) should, according to equation A.17, give valid modelling of the HRTFs up to 20 kHz.

Elsewhere a similar formula is used as the basis for frequency-dependent truncation orders (Duraiswaini et al., 2004), with r_h being the minimum radius that spans the scatterer (e.g. the dummy head). This limit is set to allow for range extrapolation where $r < r_h$ i.e. very near to the head, avoiding numerical instability due to large growth of the radial functions. When $r > r_h$, Duraiswaini et al. (2004) state that the truncation order should be roughly equivalent to the wavenumber, stating that higher truncation limits will lead to overfitting.

The DSFT (equation A.8) was applied to the HRTF datasets for each ear separately up to a chosen truncation order to generate the spherical expansion coefficients for the HRTFs.



(a) Gauss grid $N_g = 89$

(b) Lebedev grid $N_g = 44$

Figure A.39: Modal energy distribution of the spherical wave spectrum of Neumann KU100 HRTFs measured on different grids

The results are analysed in the following sections.

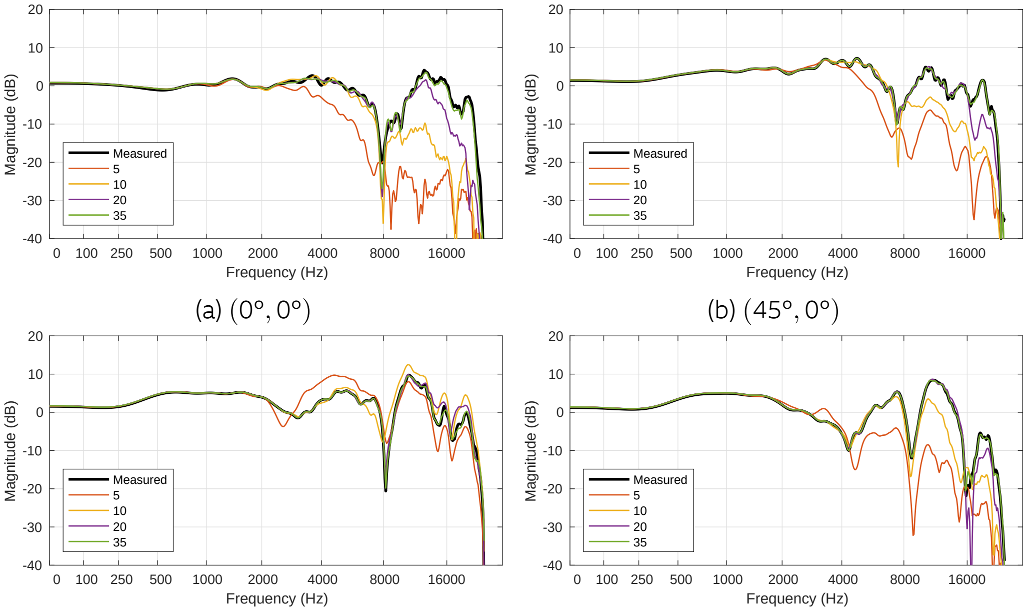
A.7.2.1 Measurement Grid Order

Figure A.39 shows the effect of sampling the HRTFs with two grids of different orders. These plots are calculated by summing the energy at each frequency bin for all of the modes at a given order and then normalising the result on a log scale.

Figure A.39a shows the modal energy distribution for the Gauss grid with $N_g = 89$. It can clearly be seen that the majority of the energy is indeed at orders $N \leq 35$, however at high frequencies there is still some energy in modes above order 50. Figure A.39b shows the modal energy distribution for the Lebedev grid with $N_g = 44$. Here the effect of aliasing can be observed, whereby the modal energy distribution is reflected about the Nyquist frequency. Therefore the energy above order 44 in figure A.39a will be introduced into the lower modal orders ($N \leq 44$) for the Lebedev grid.

A.7.2.2 Synthesis Truncation Order

When the synthesis order in the ISFT (equation A.9) is lower than the grid order, the energy in the higher order spatial modes is lost. Figure A.40 shows the magnitude response of left ear HRTFs synthesised with varying truncation order and at different azimuth angles in the horizontal plane. The spherical harmonic expansion was calculated from the 2702-point Lebedev grid ($N_g = 44$). It can be seen that the temporal frequency above which synthesis is inaccurate increases with truncation order. The dominant feature of the series order truncation is a low-pass filtering effect. This is greatest for the frontal HRTF and



(a) (0°, 0°)

(b) (45°, 0°)

(c) (90°, 0°)

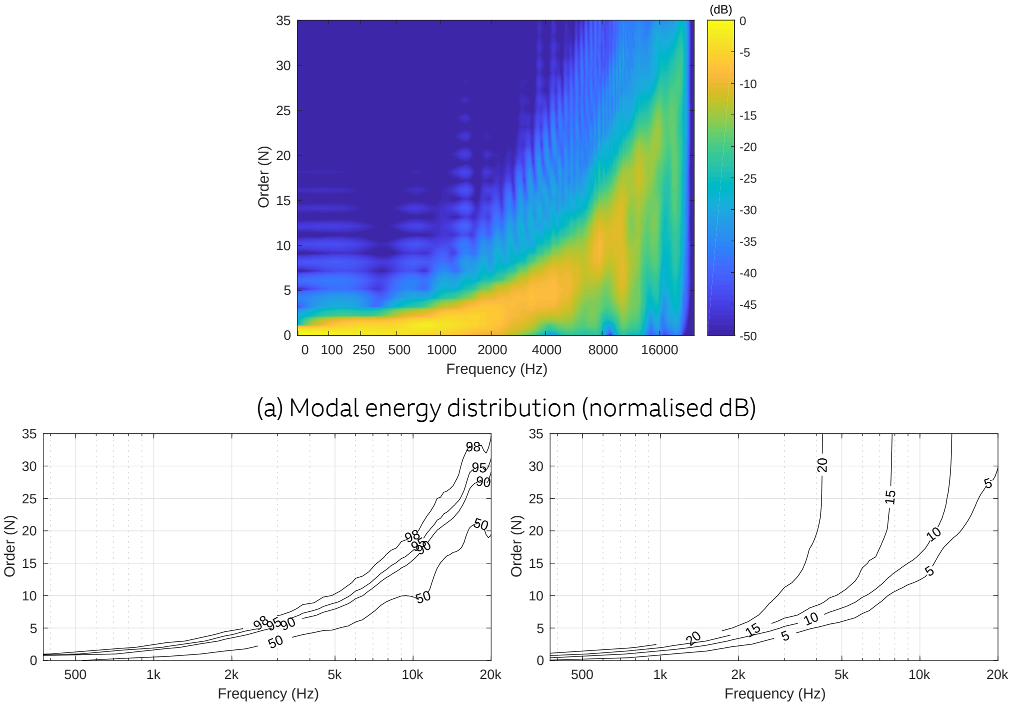
(d) (135°, 0°)

Figure A.40: Effect of truncation order N in HRTF resynthesis from the spherical wave spectrum of HRTFs of measurements on a 2702-point Lebedev grid at various directions. Shown for the left ear and compared to measured HRTFs taken from a different measurement session (2534-point Lebedev grid).

the effect is small for the lateral HRTF. This effect is due to path length differences from the contributing HRTF measurement positions to the ear. Bernschütz et al. (2014) give the analogy of an array of isophasic (equal phase) monopoles on a spherical cap about the target direction, which is wider with lower orders. Path length differences, and so phase differences at the ears, will therefore be greater at the frontal orientation than the lateral one. At low frequencies the phase differences will be much smaller and so destructive interference does not occur.

Figure A.41a shows the modal energy distribution truncated up to order 35. It can be seen that there is little energy in higher order modes at low temporal frequencies and that the amount of energy in higher modal orders increases at higher temporal frequencies, which explains the tendency towards a low-pass effect at lower truncation orders. Figure A.41b is a contour plot of the percentage of energy at each temporal frequency, indicating the truncation order required to retain a given proportion of energy. With truncation order $N = 35$, 98% of the total energy can be resynthesised up to 20 kHz.

As an alternate view, we can look at synthesis errors compared to measured data. Figure A.41c gives contours for the “logarithmic error distance” (Richter et al., 2014), which is



(a) Modal energy distribution (normalised dB)

(b) Modal energy ratio contours (% of total energy)

(c) Signal-to-error ratio contours (dB)

Figure A.41: Modal energy of the spherical wave spectrum of Neumann KU100 HRTFs measured on a 2702-point Lebedev grid

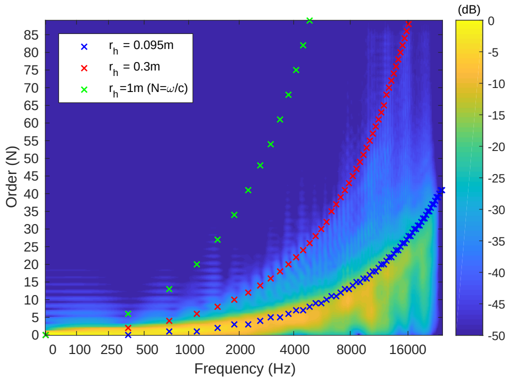


Figure A.42: Modal energy distribution of KU100 HRTF on Gauss grid with $N = 89$, with various frequency-dependent truncation orders calculated according to equation A.17

the squared difference between the original measured and reconstructed filters, integrated over the sphere. The resulting error energy is then normalised to the energy of the original signal. It can be seen as a signal-to-error ratio, based on the weighted means of error and signal energy. For this analysis, the measured HRTFs are taken from the KU100 HRTF dataset measured on the 2354-point Lebedev grid, whereas the spherical expansion coefficients are derived from the 2702-point Lebedev grid dataset and then HRTFs are synthesised from this at the set of 2354 points.

It was stated previously that Duraiswaini et al. (2004) use a frequency-dependent truncation order and various studies have followed this convention (Zotkin, Duraiswami, et al., 2009; Ben Hagai et al., 2011; Pollow et al., 2012). Figure A.42 shows the modal energy distribution for the 16 020-point Gauss grid, with frequency-dependent truncation orders overlaid, calculated according to equation A.17 at different values of r_h. Whilst the truncation limits for $r_h =$ 9.75 cm do seem to follow the upper bound of the region of strongest energy in the distribution, there would be significant energy lost by this approach, particularly at lower frequencies. Larger values of r_h lead to very high truncation orders ($N = 366$ for $r_h =$ 1 m at 20 kHz), which would not be practicable for real-time systems. With respect to Figure A.41b, the frequency-dependent truncation orders given by $r_h =$ 9.75 cm result in less than 90 % of total energy being retained in the region below 5 kHz. Figure A.43 shows the error introduced by this for an example HRIR.

Bernschütz (2016) successfully uses a frequency-independent truncation order and, as shown in Figure A.40, this appears to achieve accurate synthesis results when $N = 35$. Therefore a frequency-independent truncation order will be used in the ISFT herein.

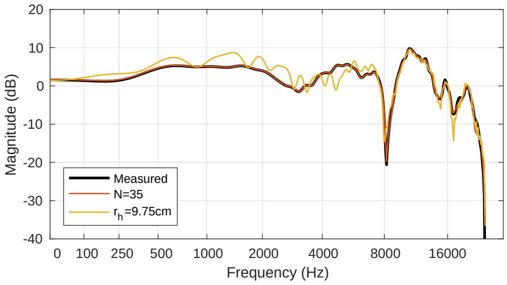


Figure A.43: Effect of using frequency-dependent truncation order according to equation A.17 with $r_h = 9.75$ cm compared with frequency-independent truncation order $N = 35$

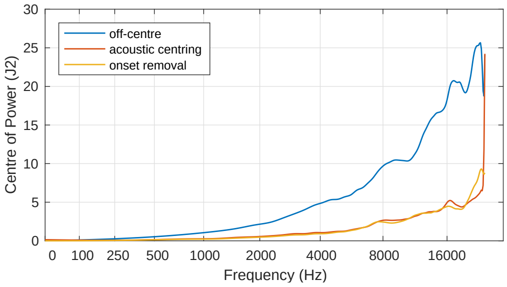


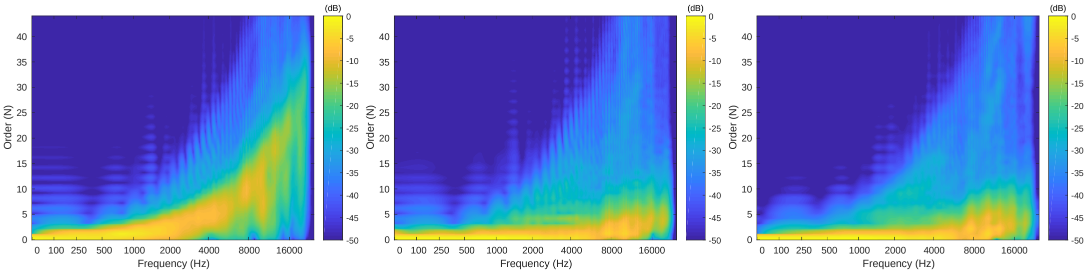
Figure A.44: Centre of power of the spherical wave spectrum of KU100 HRTFs from 2702-point Lebedev grid with and without acoustic centring

A.7.2.3 Acoustic Centring

The acoustic centring process described by Richter et al. (2014) was applied to the spherical expansion coefficients of the 2702-point Lebedev grid. This involves translating the measurement grid positions relative to the estimated acoustical centre (at the microphone position for higher frequencies) and then performing a modified DSFT, which includes a measurement-dependent range extrapolation. Richter et al. (2014) performed an optimisation routine to find the frequency-dependent acoustic centre. Since the energy in higher-order modes is at higher temporal frequencies, here a simple frequency-independent y-offset of 9.75 cm was used, to place the centre at the microphone position.

Figure A.44 shows the effect of the acoustic centring on the centre of power of the spherical wave spectrum. This can be seen as the centroid of the modal energy distribution. The energy is concentrated in lower orders as a result of the acoustic centring, which will reduce the energy loss for truncation at a given order (truncation error). The process removes the effect of path length differences, which must be represented by higher order modes.

Synthesis from the spherical wave spectrum that results from acoustic centring will be more computationally expensive, even with a frequency-independent translation, since the



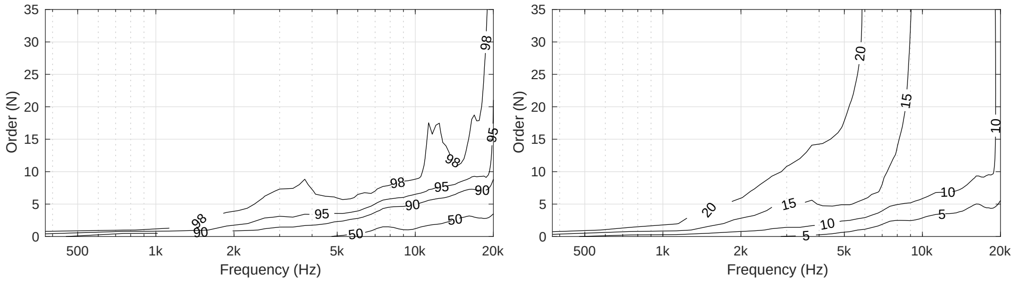
(a) Original (b) Acoustic Centring Transform (c) Onset Extraction

Figure A.45: Effect of acoustic centring on modal energy distributions of the spherical wave spectrum of Neumann KU100 HRTFs measured on a 2702-point Lebedev grid

transform will be different for the two ears and will require order-dependent radial filtering. Binaural synthesis often involves separate processing of broadband onset delays and frequency-dependent filters (Jot, Larcher, et al., 1995), which has the benefit of avoiding comb-filter artefacts when interpolating between measurements. This also allows for individualisation of the inter-aural time differences. Figure A.44 also shows the centre of power for the spherical wave spectrum obtained after first extracting broadband onset delays from the HRIRs as described in section A.5.6. It can be observed that the result is close to that obtained by the acoustic centring transform. This approach should provide the benefits of reduced truncation error whilst being more efficient and also allowing ITD individualisation during real-time synthesis.

Figure A.45 shows the modal energy distribution for the spherical wave spectra obtained by the acoustic centring transform and the onset extraction. Whilst the centroid of the spectrum is moved to lower orders, there is still energy present at higher orders and so truncation will still introduce errors. Figure A.46 shows the modal energy ratio and signal-to-error ratio contour plots when onset extraction is used as a pre-processing step. In comparison to figure A.41, a greater percentage of modal energy is found in the lower modal orders and the same signal-to-error ratio could be achieved with lower truncation order, at least up to around 19 kHz. Alternatively there will be lower error for a given truncation order. To show this, figure A.47 compares the mean signal-to-error ratio over all frequencies up to 20 kHz for synthesis at $N = 35$ from the spherical wave expansion calculated from HRIRs with and without onsets. This suggests that with onset extraction better performance is achieved at $N = 10$ than at $N = 35$ without onset extraction.

Figure A.48 shows the synthesis of HRTFs from the spherical wave spectrum with varying truncation where onsets were extracted prior to the DSFT. The magnitude error at lower orders is greatly reduced in comparison to figure A.40 where onsets were not extracted. At



(a) Modal energy ratio contours (% of total energy)

(b) Signal-to-error ratio contours (dB)

Figure A.46: Modal energy of the spherical wave spectrum of Neumann KU100 HRTFs measured on a 2702-point Lebedev grid after onset extraction

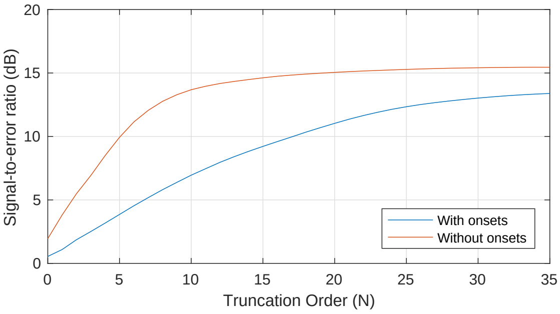
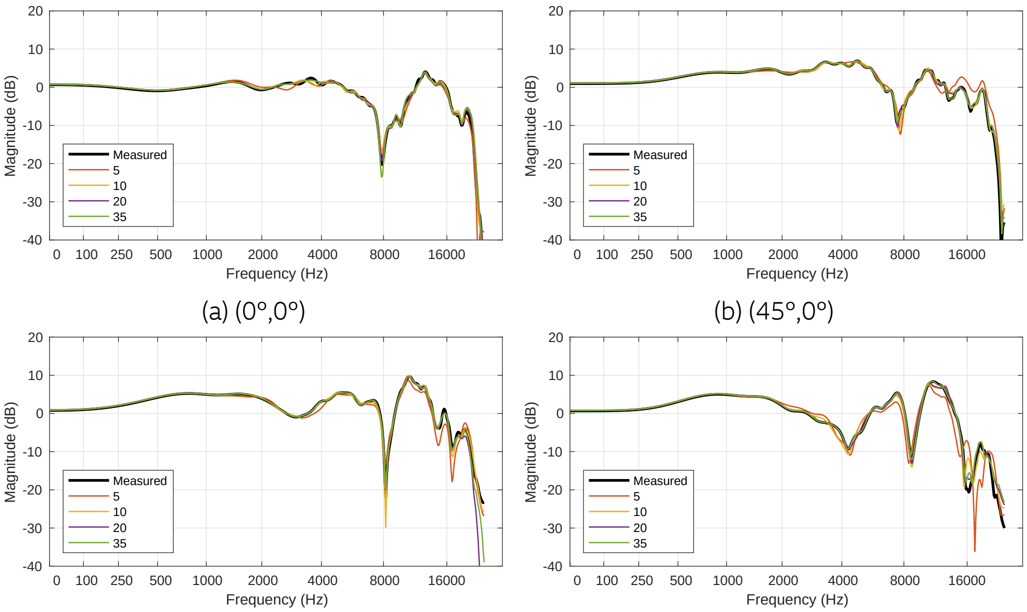


Figure A.47: Mean signal-to-error ratio (dB) of synthesis of HRTFs on 2354-point Lebedev grid from a spherical wave spectrum of HRTFs measured on a 2702-point Lebedev grid with varying truncation order. Comparison of results with and without broadband onset extraction.



(a) (0°,0°)

(b) (45°,0°)

(c) (90°,0°)

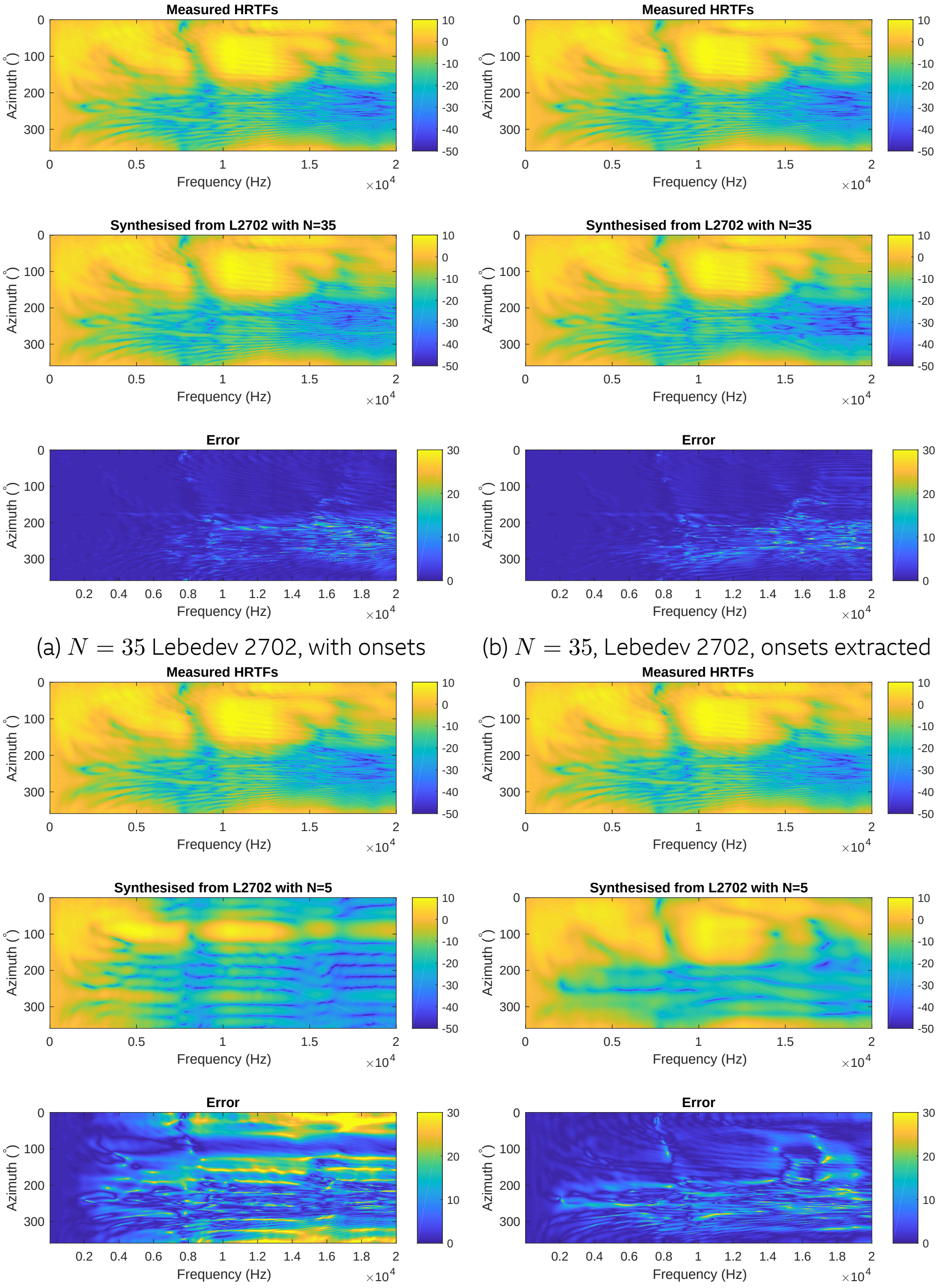
(d) (135°,0°)

Figure A.48: Effect of truncation order N in HRTF resynthesis at various directions for the left ear, measured HRTFs are taken from a different measurement session (2534-point Lebedev grid), onsets removed before DSFT

higher truncation orders the synthesised HRTFs are close to the measured HRTF but small differences remain.

To observe the synthesis error magnitude across frequency and azimuth angle, the resynthesis is compared to HRTFs measured at 1° increments of azimuth in the horizontal plane, which are also available from Bernschütz (2013). This allows easier comparison of HRTF features between the synthesised and measured data, observing the differences with source direction, particularly between ipsilateral and contralateral sides.

Figure A.49 shows the synthesised and measured magnitude responses, and the error, for truncation orders 35 and 5, both with and without broadband onsets. For $N = 35$, the largest errors occur on the contralateral side at frequencies above 8 kHz, where the HRTF energy is low. Even with a synthesis order of 89 using the 16 020-point Gauss grid (not plotted) errors still exist in this region, including large errors at notch frequencies. For $N = 5$ the errors are much larger. They are lowest at the ipsilateral side where the energy is greatest in the measured data. An angular ripple effect can be seen in the synthesised magnitude response data. With onsets extracted there is a significant reduction in errors. However on the contralateral side, the errors are still quite large and they extend much lower in frequency than the errors for $N = 35$.



(a) $N = 35$ Lebedev 2702, with onsets

(b) $N = 35$, Lebedev 2702, onsets extracted

(c) $N = 5$ Lebedev 2702, with onsets

(d) $N = 5$, Lebedev 2702, onsets extracted

Figure A.49: Synthesis of horizontal plane HRTFs compared to measured data, the error plot is clipped at 30 dB although errors do exceed this value

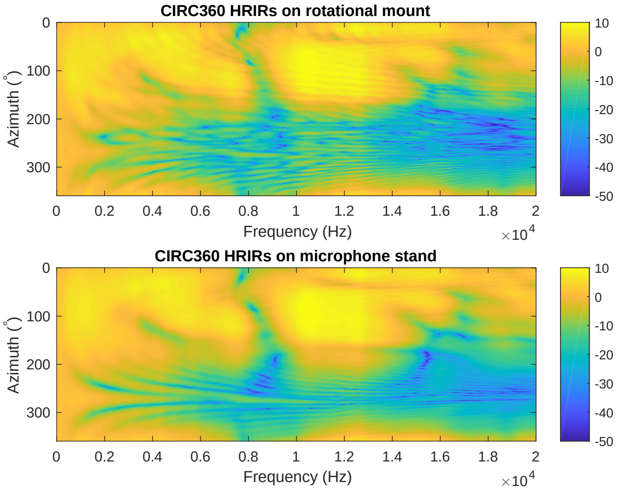


Figure A.50: Measured horizontal plane HRTFs measured with computer-controlled rotational mount and mic stand (Bernschütz, 2013)

The effect of the rotational platform on which the KU100 microphone was mounted has an observable effect on the measured HRTFs. Figure A.50 shows horizontal plane measurements made on the rotational mount and also on a normal microphone stand (Bernschütz, 2013). The scatterer is larger than the 9.75 cm radius used to estimate the required truncation order $N = 35$. This will contribute to the errors in the representation. There will also be some measurement noise. Whether the objective errors are perceptually significant should be determined by a listening experiment.

A.7.3 Software Implementations

The spherical harmonic interpolation techniques described in previous sections were implemented both in MATLAB and in C++. The MATLAB implementation was used for the analysis presented in appendix A.7.2. Regarding the C++ implementation, a new filter generator block was created to enable use of the spherical harmonic interpolation in real-time binaural rendering, see figure A.3. Since this representation of HRTF data is not yet supported in the AES69 standard, the $\tilde{a}_{nm}$ coefficients are loaded from a configuration file in a simple custom format. For $N = 35$ and a half-complex HRTF representation of 65 frequency points, filter synthesis using equation (A.9) requires multiplication of a 65x1296 matrix with a 1296x1 vector. Real-time operation is feasible for multiple sources, but when head tracking is used and complex scenes are represented, the computational requirements for filter updates lead to buffer under-runs on a typical computer. With further optimisation work this may be resolved, but this has not been done in this project.

A.8 Perceptual Evaluation of Interpolation with Spherical Harmonics

Two listening experiments were performed to investigate the perceptual effects of these spherical harmonic interpolation techniques. Both experiments used the 2702-point Lebedev grid for spherical harmonic analysis and compared the synthesised HRTFs to measured HRTFs from the 16 020-point Gauss-Legendre measurement set, at points where no measurement was available in the Lebedev set. An ABX experiment was performed to assess whether interpolation up to order 35 can generate HRTFs that are inaudible from an HRTF measurement. A multiple-stimulus rating experiment was performed with a hidden reference to evaluate the level of perceived differences to an HRTF measurement for several different interpolation options.

For both of these experiments, three source positions were used: frontal (2°,0°), rear-right (−100°,0°), and up-left (30°,30°). The nearest measurements in the 2702-point Lebedev grid are at angular separations of 1.44°, 2.41°, and 1.18°, respectively. The sound stimulus was a repeated pink noise burst of 750 ms duration with 20 ms half-cosine window fade-in and fade-out and 1 s of silence between bursts. The listening tests were carried out in the BBC R&D listening room, with Stax SR-202 headphones. The headphone level was adjusted manually to match the loudness of a loudspeaker which was calibrated at a level of 70 dBA at the listening position.

A.8.1 ABX Experiment

The ABX experiment tested the following hypothesis:

> Sounds rendered using HRTFs synthesised from a 35$^{\text{th}}$-order complex spherical harmonic representation of the 2702-point Lebedev grid (with no prior onset removal) are perceptually indistinguishable from sounds rendered using HRTFs measured at the target position.

The null hypothesis H_0 can be posed as the assumption that the group detection rate is 50%. Differences are said to be audible if we obtain a group detection rate significantly above this level. The experiment is designed with an alternative hypothesis H_1, that the group detection rate is 65% or above, which would be sufficient to say that the differences are audible. When the group detection rate is significantly below 65%, differences are said to be inaudible. The long-term group detection rate of the population of participants cannot be measured directly, the test design provides estimates based on these hypotheses, with defined error probabilities. If the number of correct trial responses is greater than or equal to the *critical number*, H_0 must be rejected in favour of H_1, i.e. differences are audible. If the critical number of correct responses is not obtained then H_1 is rejected in favour of H_0, i.e. differences are inaudible.

A.8.1.1 Balancing Errors

The number of trials N_{trials} and the critical number N_{crit} must be designed considering the appropriate probability of errors. Leventhal (1986) describes the need to assess the probability of both type-I and type-II errors during design of experiments on audibility of differences. Type-I errors occur when concluding that inaudible differences are audible, whereas type-II errors occur when concluding that audible differences are inaudible. The experiment design balanced type-I and type-II error levels, keeping them below 5% after accounting for repeated tests with Sidak correction (one test for each source position). Error probabilities for a given number of trials were calculated using the binomial cumulative distribution function.

The binomial cumulative distribution function is given by

$$F(x|n,p) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{(n-i)}, \tag{A.18}$$

This is the probability of observing x successes in n independent trials, where the probability of success in any given trial is p. The number of successes x can only adopt values of $0, 1, ..., n$.

The inverse binomial cumulative distribution function is given by

$$F^{-1}(y|n,p) = x, \tag{A.19}$$

where x is the smallest integer such that

$$y \leq \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{(n-i)}. \tag{A.20}$$

The critical number N_{crit} is the number of correct answers for which the null hypothesis can be rejected. Given the total number of trials N_{trials} and the target type-I error probability α_{target}, with p_{H_0} the probability of a random guess, N_{crit} is calculated as follows

$$N_{\text{crit}} = F^{-1}((1-\alpha_{target})|N_{\text{trials}}, p_{H_0}) + 1. \tag{A.21}$$

Due to the underlying binomial distribution, error levels are not continuous but change in discrete steps. The term $F^{-1}((1-\alpha_{target})|N_{\text{trials}}, p_{H_0})$ calculates the minimum number of correct responses for which type-I error probability is 5% or less, given N_{trials}. Therefore N_{crit} is one correct answer higher, where we have less than 95% confidence that H_0 is correct.

The actual type-I error probability α is given by

$$\alpha = 1 - F((N_{\text{crit}} - 1)|N_{\text{trials}}, p_{H_0}) \tag{A.22}$$

The type-II error probability β is

$$\beta = F((N_{\text{crit}} - 1)|N_{\text{trials}}, p_{H_1}). \tag{A.23}$$

Leventhal (1986) defines the fairness coefficient for use in experiment design

$$c_{\text{fair}} = \frac{\text{min}(\alpha, \beta)}{\text{max}(\alpha, \beta)}. \tag{A.24}$$

For balanced type-I and type-II errors $c_{\text{fair}} = 1$. Figure A.51 shows the fairness coefficient for a range of N_{trials}. The test design was chosen to maximise c_{fair} whilst keeping both α and β below 5%. This resulted in a critical number of 125 or more correct answers out of 216 trials i.e. $N_{\text{crit}} = 125$ and $N_{\text{trials}} = 216$. This gives $\alpha = 0.0363$ and $\beta = 0.0369$ i.e. a test power of 96.31%.

A.8.1.2 Method

In order to obtain the 216 trial responses, a panel of 12 assessors each performed 18 repetitions for each of the 3 position conditions. The assignment of sounds rendered using measured and synthesised HRTFs to A, B, and X was randomised for each trial. Assessors had to decide which of A or B was perceptually different to X.

The assessors could repeat the sound as many times as desired to make a decision and freely switch between the three stimuli, as in Brinkmann, Lindau, and Weinzierl (2014). Assessors were first given a training session where they were asked to perform four trials for each position (one for each variation of A, B, X assignment). This was performed first using interpolation at 5$^{\text{th}}$-order so that differences were obvious, allowing the assessor to become familiar with the experiment process. Following this the training was performed with 35$^{\text{th}}$-order spherical harmonic interpolation, as used in the main experiment.

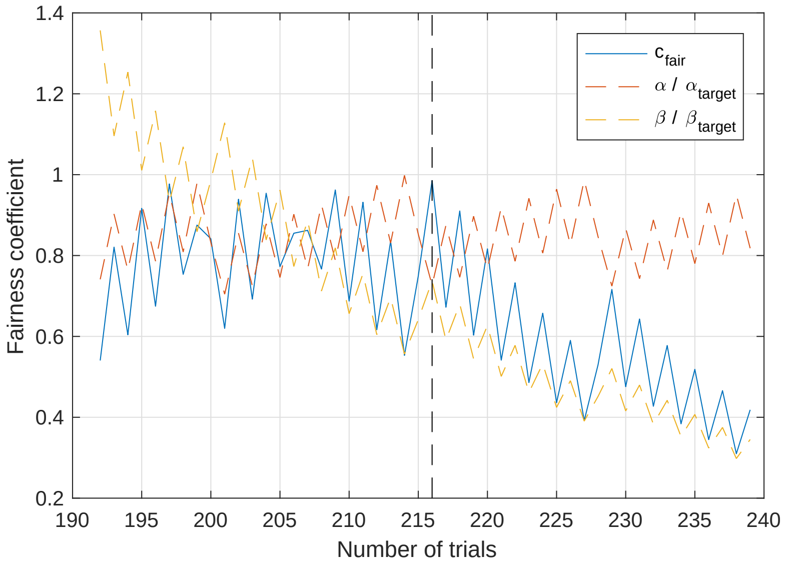


Figure A.51: Fairness coefficient (ratio of type-I and type-II error levels) based on ABX experiment designs of differing length. Also shown are the ratio of actual type-I (α) and type-II (β) error levels compared to the target of 0.05. The black dashed line indicates the chosen number of trials.

Source Direction	Number of correct answers
Frontal $(2°, 0°)$	118
Rear-right $(-100°, 0°)$	118
Up-left $(30°, 30°)$	115

Table A.1: Total correct answers across all assessors in ABX experiment

A.8.1.3 Results

The results for each of the three positions are shown in Table A.1. For each position, the number of correct answers was lower than the critical number, so no significant pattern of detection is observed. Figure A.52 plots the results as a percentage of correct answers, both for the group and for each individual assessor.

These results suggest that the 2702-point Lebedev grid dataset can be used with spherical harmonic interpolation at order 35 to obtain HRTFs that are indistinguishable from real measurements, at least for this measurement set made on the Neumann KU100 dummy head.

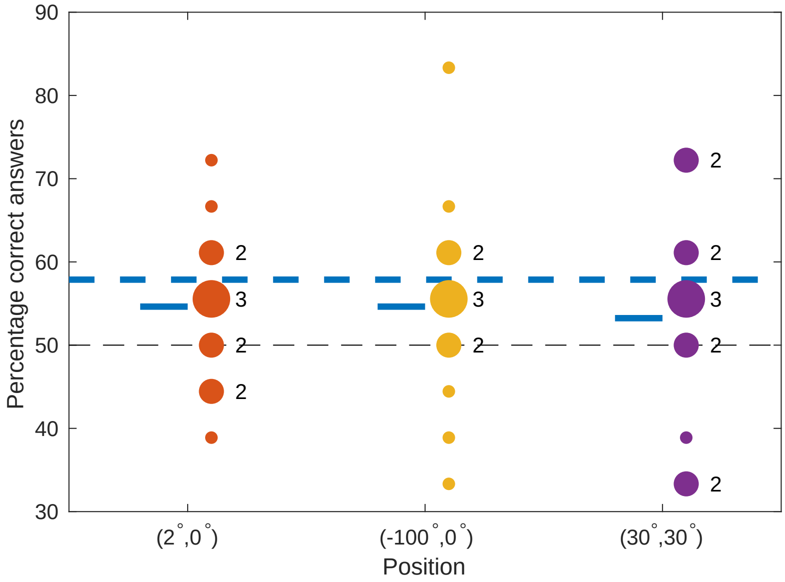


Figure A.52: ABX experiment results shown as a percentage of correct answers, both for the group (blue bars) and individual assessors (coloured circles). Where multiple assessors had the same result, the circle radius is larger and the number of assessors is labelled. The dashed blue line represents the critical number derived from the null hypothesis.

A.8.2 Multiple-Stimulus Experiment

The multiple-stimulus experiment investigated how differences to a target measurement compare between different methods of HRTF interpolation. Subjects were asked to rate the overall difference of stimuli to the reference, taking into account all aspects of the sounds. The experiment assessed spherical harmonic interpolation at orders 35 and 5, both with onsets included and with onsets removed and separately processed. It also included nearest-neighbour selection of measurements in the 2702-point Lebedev set for comparison.

The hypotheses for the experiment were as follows.

- There will be no perceived differences for 35^{th}-order spherical harmonic interpolation, with or without separate onset processing.
- Perceived differences will be greater for nearest-neighbour filter selection than spherical harmonic interpolation at 35^{th}-order.
- Perceived differences to the reference measurement will be greatest for 5^{th}-order spherical harmonic interpolation without separate onset processing.
- Perceived differences to the reference measurement with 5^{th}-order spherical harmonic interpolation will be significantly reduced when using separate onset process-

ing.

A.8.2.1 Method

The assessors were given a multiple stimulus presentation with a hidden reference and asked to rate the perceived differences to the reference on a 100-point scale, taking into account all aspects of the stimuli. Assessors performed ratings for each source position twice. There were 15 assessors in total, including the 12 assessors who participated in the ABX experiment.

A.8.2.2 Results

Figure A.53 plots the distribution of difference ratings for each system and source position.

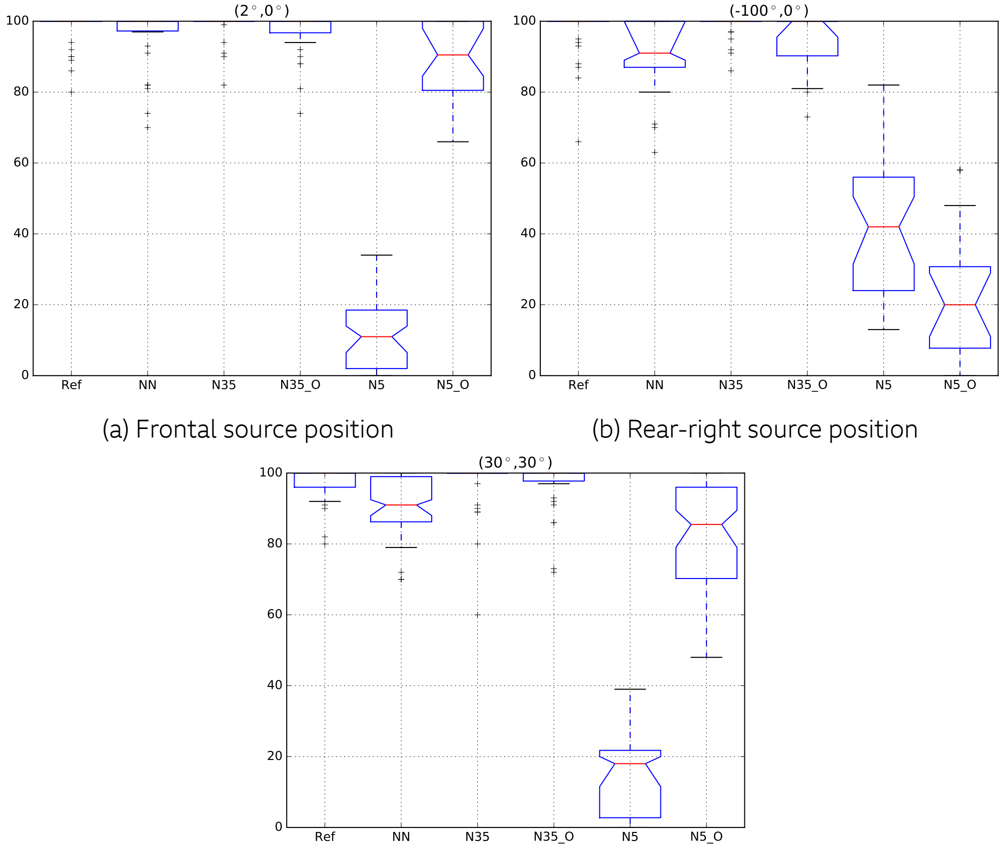
For the frontal source position (figure A.53a), interpolation at order $N = 5$ with onsets included was very different to the reference measurement. At $N = 5$ with onsets processed separately, it was perceived as very close to the reference but with significant differences. All other systems had no significant differences, when considering the bootstrapped median distribution with 95% confidence intervals.

For the rear-right source position (figure A.53b), both versions at order $N = 5$ were very different to the reference. The case with separate onset processing was perceived as more different than with onsets included. Results suggest that nearest-neighbour and spherical harmonic interpolation at $N = 35$ with separate onset processing may be perceivably different to the target measurement, though the results appear not to be significant.

For the up-left source position (figure A.53c), the nearest-neighbour selection was very close to the reference but showed perceivable differences. Spherical harmonic interpolation with $N = 5$ and onsets included was perceived as very different to the reference measurement. At $N = 5$ with separate onset processing, stimuli were close to the reference but with significant perceivable differences. For both versions at $N = 35$ no significant differences were observed.

A.8.2.3 Discussion

This experiment confirms that the use of spherical harmonic HRTF interpolation at order $N = 35$ with onsets included cannot be distinguished from the use of a real HRTF measurement at that position, at least for the positions tested and the HRTF measurement set used. With separate onset processing, interpolation at $N = 35$ appears also not audibly-different from measurements. Nearest neighbour selection from the 2702-point Lebedev grid is also



(a) Frontal source position

(b) Rear-right source position

(c) Up-left source position

Figure A.53: Box-plots of multiple-stimulus rating results. Red lines show the median, notches show the bootstrapped-95% confidence intervals of the median, boxes show the inter-quartile range, whiskers are 1.5 times the inter-quartile range, and crosses indicate outliers.

perceived as very close to the target measurements (which are not within this set of positions), however differences are sometimes audible. For spherical harmonic interpolation at $N = 5$, differences are clear, but separate onset processing makes the differences much smaller, except for the lateral position. Figure A.49d shows broadband errors in generating contralateral HRTFs with $N = 5$, even with onsets removed.

A.8.3 Conclusions

Spherical harmonic interpolation allows the generation of HRTFs at arbitrary field points that appear to be indistinguishable from real measurements for this Neumann KU100 dataset, when used in binaural rendering of pink noise sources. Separate onset processing improves performance at lower-orders, except for low-energy contralateral HRTFs, and does not negatively affect performance at 35th order. This approach allows personalisation of interaural time differences.

Romigh, Brungart, Stern, et al. (2015) evaluated interpolation of HRTF log-magnitudes using real spherical harmonics, taking the minimum-phase representation and using separate ITD insertion. The use of log-magnitudes is likely to better retain the features of the low-energy contralateral HRTFs at low spherical harmonic orders. Their study found that localisation accuracy was preserved with spherical harmonic series as low as 4th-order. However the evaluations described here take into account all differences, rather than pure localisation changes. Informal listening suggests that other features such as source distance, width, localisability and externalisation, as well as tone colour, are affected by reducing interpolation order.

For low-order spherical harmonic representations of the HRTF, Bernschütz et al. (2014) showed that resampling to a grid of the appropriate order reduces errors, avoiding the loss of energy due to truncation. Since we now know that 35th-order interpolation produces perceptually indistinuishable results from real measurements, this approach can be followed. This is used when considering binaural rendering of ambisonics in appendix B and chapter 8.

A.9 Summary

This chapter has presented apparatus for investigating the perceived quality of binaural technology and its applications, particularly allowing study of the influence of content and system factors. The design and validation of the components of the apparatus have been presented. This process was informed by the review of binaural technology reported in chapter 2 in order to achieve state-of-the-art performance.

The apparatus is based on a flexible software system for representing 3D spatial audio scenes and, by extension, programme content. The system allows real-time comparison of multiple rendering configurations and providing a remote control interface. A number of binaural rendering techniques have been implemented in this real-time software system; these make use of core signal processing blocks such as partitioned fast frequency-domain convolution and variable fractional delay lines.

A number of tracking systems were evaluated, to enable dynamic head-tracked binaural rendering, which required implementation of interfaces to the rendering software. These were judged in terms of criteria for good perceptual results in binaural rendering. An optical tracking system was found to be most suitable. It uses four wall-mounted cameras and retroreflective markers worn on the headphones. With the tracking system integrated, the total system latency was evaluated to ensure that it is below detection thresholds observed in the literature.

A Neumann KU100 microphone was procured for the apparatus. A high-resolution set of full-sphere HRTF measured on this microphone is freely-available. Additional tools for measurement and post-processing of binaural impulse responses were created and presented. This includes software for impulse response measurement using the exponential swept sinusoid technique and a rotational platform for automating measurements at multiple precisely-controlled orientations of the dummy head microphone. Techniques for estimating and extracting the broadband time-of-arrival were also presented to enable separate onset delay interpolation during real-time rendering.

Implementations were made of four HpCF design techniques from the literature that are suitable for use in uncontrolled environments, where precise in-situ headphone correction is not feasible. These were based on HpTF measurements made on the Neumann KU100 dummy head microphone and evaluated by use of an auditory filterbank. After this analysis, a new HpCF created by expert users perceptually tuning parameters was described. The set of five HpCFs were analysed again with an auditory filterbank, though this time using HpTFs measured on a human individual. The residual errors of the perceptually-tuned filter most closely resembled the standard diffuse-field equalised headphone target function.

Finally, spherical harmonic interpolation methods for obtaining HRTFs at any required direction were presented, with an analysis of the KU100 HRTF dataset. It was observed that removal of broadband onsets allows more accurate representation when using reduced spherical harmonic series truncation order. A listening experiment was presented, using an ABX paradigm to verify that interpolation of the KU100 HRTFs with order 35 yields binaurally-rendered pink noise sources with inaudible differences to rendering using real measurements. Another listening test demonstrated that this approach gives more accurate results than the nearest-filter selection method for a 2702-point Lebedev grid, with or without separate onsets. It also showed that onset removal improves interpolation accuracy at order 5, except in the lateral direction where the low-energy contralateral response is inaccurate. Following this validation, this interpolation technique will be used to calculate HRTFs at precise required directions in ambisonics-to-binaural decoding designs presented in chapter 8 and appendix B.

The components of the apparatus have been evaluated objectively and are expected to give state-of-the-art performance. The object-oriented design of the software system will allow straightforward extension of the system in future, with additional signal processing blocks and rendering algorithms, or interfaces to new tracking devices and remote control applications. The software system makes use of recent standards for representing 3D spatial audio programme material (ITU-R, 2017b) and binaural filter data (AES69:2015, 2015), which enhances compatibility with other systems and datasets. For example, individualised HRTFs can be used for rendering, provided that the measurements can be made available in the AES69 (SOFA) standard format. This also means that experimental comparison of binaural rendering using a range of HRTF sets is straightforward, allowing for perceptual evaluation of individualisation techniques or inter-aural time difference estimation methods for example.

Most importantly, this apparatus can be used for research experiments that compare and evaluate binaural rendering system designs, using a range of audio programme material and formats, including representative complex sound scenes with moving sources. It is used in the listening experiments presented in chapters 7 and 8. It has also been used to create datasets of audio programme material in the ADM BWF format and BRIRs in the SOFA format, which have been made available to the research community, as described in co-authored publications Co.P. V and Co.P. IX (Woodcock, Pike, Melchior, et al., 2016; Pike and Romanov, 2017a).

Other researchers have used this apparatus or components of it in additional research studies. McArthur (2016) investigated the effects of audio-visual spatial congruence on sense of presence in virtual reality (VR), using this system to perform dynamic BRIR rendering and control sound source position. Hughes et al. (2016) studied the effect of interaction between *target* room reverberation in recorded multichannel audio signals and the *reproduction* room reverberation used during dynamic binaural rendering of virtual loudspeakers with BRIRs. Listeners rated the perceived room size. It was found that the reproduction room only affects judgements when the target room has lower reverberation time. This relates to work by Werner, Klein, et al. (2016) on room-divergence effects on externalisation. The binaural rendering tool also forms a component of the end-to-end object-based audio system described in co-authored publication Co.P. XI (Coleman, Franck, Francombe, et al., 2018). This paper describes the role of the binaural renderer in estimating the speech intelligibility and loudness of loudspeaker reproduction, as well as reproduction of object-based audio signals to headphones. It is expected that the apparatus will prove useful for future research studies beyond the scope of this thesis.

Besides controlled scientific study of system and content factors, the apparatus functions as a rendering engine for use in content production. This has been applied frequently at the BBC for experimental production of programmes, enabling to explore the use of binaural rendering creatively with production professionals and to present the results to audience members for feedback.

It has been mentioned that the binaural rendering apparatus described in this chapter allows for study of content-related factors that may influence quality. Content factors can include a wide range of aspects. At the high level this relates to the meaning of the audio signals, the programme creators' intentions and the interpretation by listeners. There are also a number of technical factors in content production that, while not directly binaural rendering technology, will interact with the binaural rendering and have an impact on quality. This includes the source audio signal characteristics and related aspects such as microphone techniques for recording, but also the representations of the spatial audio scenes to be rendered. Through use of the ADM the apparatus is flexible in the ways that scenes can be represented. The format used to represent the content scenes will likely have a big impact on the quality of the experience provided. When considering applications of media distribution and reproduction on consumer devices, the spatial audio format will also interact with context influence factors, particularly economic and technical factors (see chapter 3).

Appendix B

Apparatus for Investigating Loudspeaker Virtualisation and Spatial Audio Formats

B.1 Introduction

This chapter presents techniques for rendering a 3D sound scene to headphones via a virtual loudspeaker array. Some common techniques for loudspeaker rendering of spatial audio are presented, along with the associated approaches to binaural post-processing.

Section 6.4 presents the basic technique of loudspeaker virtualisation rendering. Appendix B.2 gives details of the virtual loudspeaker binaural renderer that forms part of the apparatus introduced in appendix A. Amplitude panning techniques for rendering sound sources to loudspeakers are then discussed in appendix B.3, covering classical pair-wise panning and its extension to 3D loudspeaker arrays. Appendix B.4 reviews the methods of a popular alternative approach called ambisonics. Appendix B.5 then presents common methods for analysing the performance of loudspeaker rendering. These can be useful supplementary tools alongside perceived quality evaluation experiments, such as those performed in chapters 7 and 8.

B.2 A Binaural Renderer using Loudspeaker Virtualisation

Figure B.1 shows the structure of a binaural renderer that uses an intermediate virtual loudspeaker array to render an object-based input scene to headphones. This renderer has been

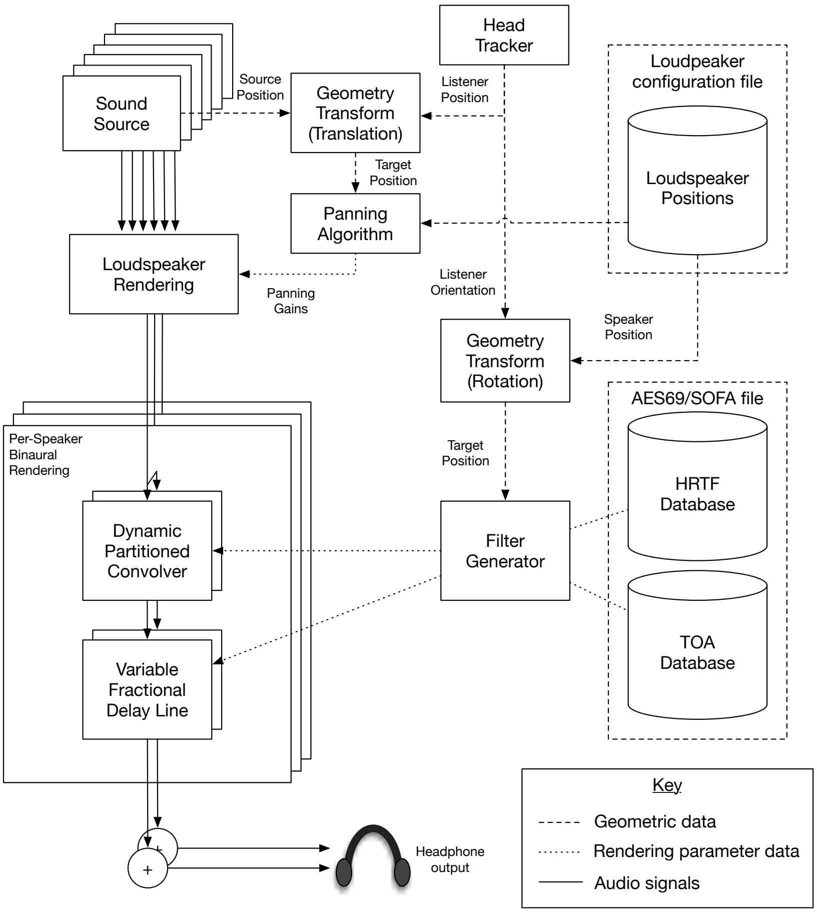


Figure B.1: Structure of a binaural renderer that uses an intermediate virtual loudspeaker array. Multiple sound sources are first rendered to the virtual loudspeaker array using a loudspeaker rendering process, driven by a panning algorithm. These virtual loudspeaker signals are then binaurally-rendered for headphone output.

implemented into the apparatus described in appendix A. The sound source positions are first rendered to the loudspeakers, represented here by a generic loudspeaker rendering block. The loudspeaker positions are loaded from a configuration file and used to initialise this process. These positions are also used for loudspeaker virtualisation. A binaural rendering process is performed for each virtual loudspeaker signal. This stage is therefore independent of the number and position of the input sources. Binaural filters are generated according to the virtual loudspeaker positions and the head tracker orientation data.

The two stages, rendering to loudspeakers and virtualisation by binaural rendering, can be decoupled. Loudspeaker rendering can be performed first, as a separate process. For example a channel-based signal may be distributed and then virtualised on the receiver. For virtualisation of a channel-based input signal, the loudspeaker rendering stage may be removed from figure B.1. It is worth noting that channel-based production might not only involve rendering of single-channel sound sources to a loudspeaker array, but also multichannel microphone arrays, which may be directly routed to the loudspeaker array e.g. Theile and Wittek (2011).

As indicated in figure B.1, the loudspeaker rendering step may also incorporate translation according to listener position to allow 6 degrees of freedom (DoF) tracking. Listener orientation could also be applied at this stage instead of by update to binaural filters. This is only feasible in scenarios where an object-based input is available to the rendering system. Many applications will use a scene-based or channel-based input, perhaps with a small number of additional objects.

It is common for next-generation audio (NGA) and virtual reality (VR) audio systems to follow an approach similar to that in figure B.1 to render object-based input, as introduced in section 6.3.4. The binaural rendering process generally requires significantly more processing per source than the loudspeaker rendering. By keeping the number of binaurally-rendered sources fixed, the processing requirements are more stable, and for complex scenes, where the number of sources is greater than the number of virtual loudspeakers, it is also more efficient. The following section gives more detail on commonly-used loudspeaker rendering algorithms.

B.3 Rendering to Loudspeakers using Amplitude Panning

To render an object-based scene to headphones via loudspeaker virtualisation, first requires techniques for rendering the scene to a loudspeaker array. There are many such techniques available. Most common are *amplitude panning* techniques which create a virtual source position by distributing the source energy between multiple loudspeakers, potentially with differing amplitudes. This can be defined mathematically:

$$x_l(t) = g_l\ s(t), \quad l = 1, \ldots, \mathsf{L}\ , \tag{B.1}$$

for L loudspeakers, where g_l is the gain applied to the l^{th} loudspeaker and $x_l(t)$ is the signal for the l^{th} loudspeaker.

Through summing localisation processes (see section 2.2.8), the coherent signals from the multiple loudspeakers are combined in the listener's auditory system, and the listener will often perceive a single auditory event. The perceived direction of the source depends on the panning gains that are applied to distribute the source to the loudspeakers. An algorithm that defines panning gains based on a target source direction is called a *panning law*, or a panning algorithm. Most commonly, amplitude panning only controls the direction of the sound source, though additional techniques to simulate distance and extent can be considered. Amplitude panning can be applied to multiple input sources, mixing them into the multiple loudspeaker output signals via panning gains relating to each target position. The scene is then represented in a channel-based format. Amplitude panning can be considered analogous to coincident microphone techniques, since in principle both only capture amplitude differences. Panning techniques that incorporate time-delay differences also exist, but these are less common.

B.3.1 Panning for Two-Channel Stereo

Amplitude panning was developed for positioning virtual sources in two-channel stereophonic reproduction. A source is distributed between the two loudspeakers, placed at ±30° azimuth at equal distance from the listening position. The sound arrives from both loudspeakers at the two ears. For each loudspeaker the sound arrives at the ipsilateral ear before it arrives at the contralateral ear and it is louder at the ipsilateral ear particularly at high frequencies. At low frequencies, where head shadowing is not prominent, the level differences at the loudspeakers are converted to phase differences at the listener's ears (Bauer, 1961a). At higher frequencies, level differences are created at the ears due to head shadowing. These effects provide an approximation of the natural interaural time differences (ITDs) and interaural level differences (ILDs). Blumlein, 1931 originally presented a stereo panning law, now known as the sine law, which approximately estimates the perceived direction θ from the panning gains. It was reformulated in phasor form by (Bauer, 1961a):

$$\frac{\sin\,\theta}{\sin\,\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{B.2}$$

where 0° $< \theta_0 <$ 90°, $-\theta_0 \leq \theta \leq \theta_0$, and $g_1, g_2 \in [0, 1]$.. Here θ_0 is the angle between the loudspeakers and the frontal x-axis, and g_1 and g_2 are the gains for the left and right loudspeakers respectively. The tangent law was later proposed, which approximately modelled the curved path around the head from the contralateral loudspeaker to the ear and was deemed more accurate when the listener turns to face the direction of the source (Bennett et al., 1985; Bernfeld, 1973). It is given by:

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2}. \tag{B.3}$$

A constant-power panning constraint is applied to either of the above equations. Solving equation (B.4) and equation (B.3) or equation (B.2) provides the panning gains for the specified direction:

$$g_1^2 + g_2^2 = 1. \tag{B.4}$$

In practice stereo panning is only effective for θ_0 of no more than about 30°, and virtual sources can be positioned between the loudspeakers only for a listener positioned at a point within a small region of a few tens of centimetres around the central listening position.

Extensions of traditional stereo amplitude panning to 2D and 3D loudspeaker arrays have been made. The following sections introduce two of the most common techniques for rendering sound sources to 3D loudspeaker arrays. Vector base amplitude panning (VBAP) was developed as a method to allow amplitude panning onto any number of loudspeakers placed arbitrarily. Besides amplitude panning, ambisonics techniques exist, which can be seen to split the panning process into two decoupled stages.

B.3.2 Vector-base Amplitude Panning

VBAP (Pulkki, 1997) is a technique for panning a sound source to a given direction on an arbitrary 2D or 3D configuration of loudspeakers. It is a reformulation of the tangent panning law with vector bases. When using 2D vectors, it provides pairwise panning in the horizontal plane, as shown in figure B.2 for the five-channel surround layout of Recommendation ITU-R BS.775 (ITU-R, 2012a). When using 3D vectors this extends panning to 3D loudspeaker layouts, using up to three active loudspeakers to pan a single virtual source.

When the target source direction lies within a spherical triangular region formed by the positions of three available loudspeakers, the VBAP algorithm is defined as follows. The direction vectors pointing towards the three loudspeaker positions are denoted by $\mathbf{l}_1$, $\mathbf{l}_2$ and $\mathbf{l}_3$, and the target source direction by $\mathbf{d}$. These are 3D unit-vectors which have their origin at the centre of the unit-sphere, as indicated by $\mathbf{d} = [d_x\ d_y\ d_z]^\mathsf{T}$. With the formation of a

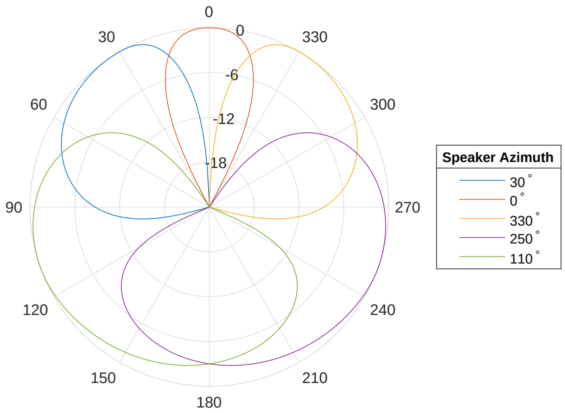


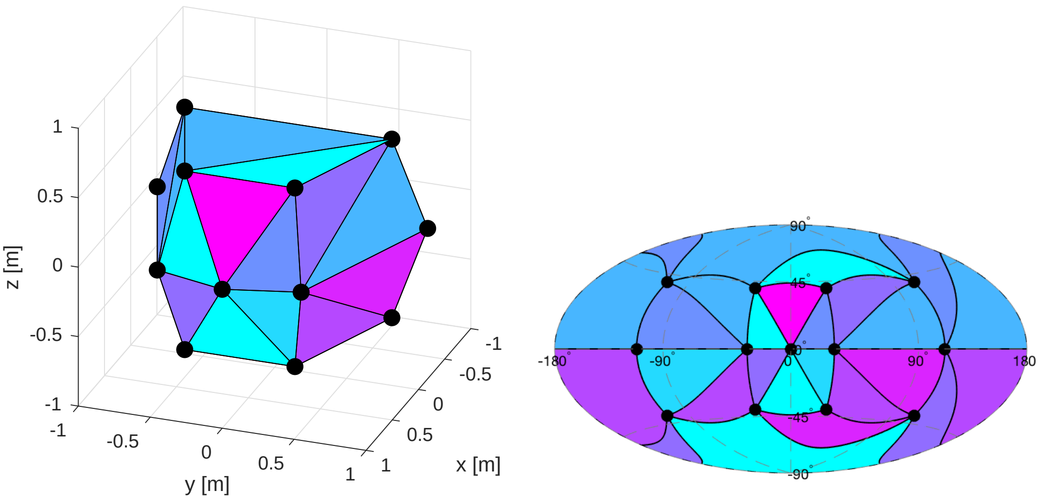
Figure B.2: 2D VBAP panning functions for the five-channel surround layout of Recommendation ITU-R BS.775. Panning gains are shown in decibels in the range −24 dB–0 dB with varying target source angle θ_s.

3×3 matrix of the loudspeaker direction vectors $\mathbf{L} = [\mathbf{l}_1\ \mathbf{l}_2\ \mathbf{l}_3]^{\mathsf{T}}$, the VBAP formulation gives corresponding panning gains $\mathbf{g} = [g_1\ g_2\ g_3]$ by:

$$\mathbf{g} \cdot \mathbf{L} = a\mathbf{d} \tag{B.5}$$

with scale factor $a > 0$, $g_l \geq 0 \ \forall \ l \ni \{1, 2, 3\}$, and $\|\mathbf{g}\| = 1$. A solution, within a small numerical tolerance, is possible provided that $\mathbf{L}^{-1}$ exists, which it does when the loudspeaker positions span a 3D space i.e. the set of points defined by the loudspeaker vectors are not co-linear or co-planar with the origin, and so rank($\mathbf{L}$) $= 3$. The scale factor a performs normalisation to ensure that the gain vector has unit norm.

To apply the VBAP algorithm to panning over a 3D array of loudspeakers, a triangular mesh formed from the loudspeaker positions is required, such that a triplet of loudspeakers can be found that encloses any given target panning direction $\mathbf{d}$. For an arbitrary array of loudspeaker positions, this triangular mesh can be obtained algorithmically. The convex hull of the loudspeaker positions is obtained, normally first projecting the points onto the unit sphere, and from this the Delaunay triangulation is obtained. The quickhull algorithm provides an efficient means for computing the convex hull and Delaunay triangulation (Barber et al., 1996). An implementation is given in the Qhull open-source software library. Figure B.3 shows the triangulation of an example array of loudspeakers obtained in this manner. Figure B.3a shows the triangulation of the convex hull of loudspeaker positions, and figure B.3b shows the same triangulation viewed in terms of azimuth and elevation using the Hammer projection.



(a) Triangulation of the convex hull of loudspeaker positions.

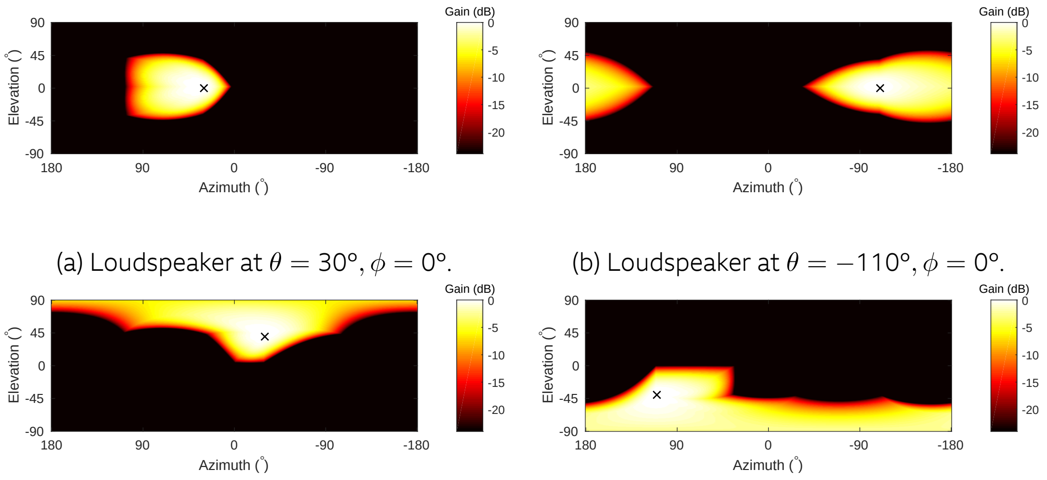
(b) Triangulation viewed in terms of azimuth and elevation using the Hammer projection.

Figure B.3: Loudspeaker triangulation for VBAP algorithm on a 4+5+4 layout.

Whether a triangle of loudspeakers encloses **d** can be ascertained by checking that valid panning gains are obtained from equation (B.5). For most arrays, the number of loudspeaker triangles is low enough that an exhaustive search can be performed without efficiency concerns. The VBAP algorithm was implemented in MATLAB and C++. Figure B.4 shows the panning gains for example loudspeakers from the 4+5+4 layout (see figure B.3) for target source directions covering the full sphere.

Several extensions have been made to the original VBAP method presented by Pulkki (1997) to make panning more robust in practical applications. Without modification, the VBAP algorithm presents issues for many of the loudspeaker layouts defined in Recommendation ITU-R BS.2051. Besides system H, these layouts do not include loudspeakers below the horizontal plane, so no valid triangles of loudspeakers can be found for source directions where $d_z < 0$. Often these layouts also have large apertures between loudspeakers, which will likely yield unstable virtual sources (Pulkki, 2001c, p.29).

Extensions have been made to incorporate additional “imaginary” loudspeaker points, to enable use of VBAP where the loudspeaker array does not cover the sphere well. Zotter and Frank, 2012 described omitting triangles with aperture $\geq$ 90° from the convex hull, and then using an imaginary loudspeaker point in the direction opposing the sum of surface normals of the admissible triangles. This provides a complete convex hull surrounding the listener, but the signal of the imaginary loudspeaker is then discarded, resulting in a region of no output energy. Instead Borß et al. (2016) describe redistribution of the energy of the



(a) Loudspeaker at $\theta = 30°, \phi = 0°$.

(b) Loudspeaker at $\theta = -110°, \phi = 0°$.

(c) Loudspeaker at $\theta = -30°, \phi = 40°$.

(d) Loudspeaker at $\theta = 110°, \phi = -40°$.

Figure B.4: Panning gains over the full sphere for certain loudspeakers given by the VBAP algorithm on a 4+5+4 layout. Loudspeaker positions marked by black cross. Gains shown by colour in decibels, limited to range −24 dB–0 dB.

imaginary loudspeakers to adjacent real loudspeaker points after the VBAP algorithm, which effectively results in panning to more than 3 loudspeakers, i.e. the imaginary loudspeaker is down-mixed to create an N-gon in the convex hull. This technique has also been applied where other deficiencies have been observed, such as perceived asymmetric effects when spherical quadrilaterals of loudspeakers are split into two triangles. Further extensions have been made to incorporate atomic quadrilaterals into the panning algorithm in these cases (EBU Tech 3388, 2018).

These techniques yield broad panning apertures in N-gons where $N > 3$, which can result in unstable localisation. In the case where no loudspeakers are present above or below the horizontal plane (e.g. 0+5+0), N-gon down-mix techniques will lead to broad spreading for sources that have $\phi \neq 0°$. The EBU ADM Renderer places imaginary loudspeakers directly above/below the horizontal-plane loudspeakers where there are no real loudspeakers in that region and these are directly routed to the real horizontal plane loudspeakers, thus avoiding excessive spreading at the cost of incorrect panning direction (EBU Tech 3388, 2018). Similar techniques are applied in chapter 8.

Franck, Wang, et al. (2017) present amplitude panning as a minimum-ℓ_1-norm optimisation problem. When this incorporates a non-negativity constraint to panning gains, the solution is identical to VBAP with Delaunay triangulation. VBAP minimises the number of loudspeakers used in panning, which differentiates it from another common technique

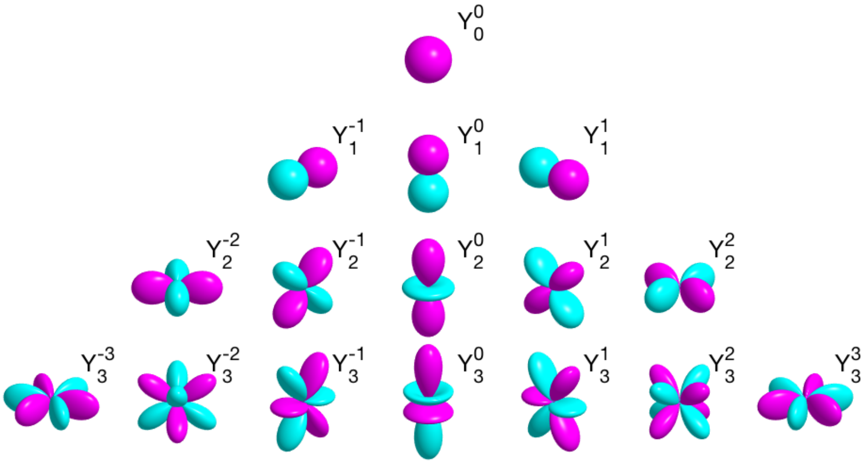


Figure B.5: Real-valued spherical harmonics series up to order $N = 3$. The shape extension indicates the magnitude of the function and colour indicates the sign.

called ambisonics.

B.4 Ambisonics

Ambisonics is a spatial audio technique based on a series of spherical harmonic coefficient signals. The spherical harmonics are visualised in figure B.5 up to the third order and defined mathematically in appendix B.4.2.2. Gerzon (1973) introduced the concept of ambisonics as a theoretical framework for 3D spatial audio or "periphony". It originated from the extension of coincident microphone recording techniques to 3D and it was developed into a system encompassing recording, transmission and reproduction.

B.4.1 Background

Early applications used what is now known as first-order ambisonics. Here, the sound field is represented by four signals, an omnidirectional sound pressure signal and three coincident dipole sound pressure gradient signals corresponding to the Cartesian axes in 3D space. These are the first two rows in figure B.5. This can be considered a spherical Fourier series decomposition of the sound field at a certain point in space, truncated at the first order, and so comprising coefficient signals of the zeroth- and first-order spherical harmonics. It can also be thought of as an extension of the stereo mid-side recording technique to 3D.

A tetrahedral array of cardioid microphones was designed to record first-order ambisonics (Gerzon, 1975; Craven and Gerzon, 1975), which was and still is sold commercially as a SoundField microphone. This simulates the ambisonics signals from the nearly-coincident array. The set of microphone signals from this array is commonly called *A-format* and the spherical harmonic coefficient signals obtained from conversion are called *B-format*. Sound sources can also be placed artificially in the ambisonics sound field representation by projection of the target position onto the spherical harmonic functions, a process often known as *encoding*.

The ambisonics sound field representation is independent of the reproduction loudspeaker configuration, or the recording microphone configuration. Although in practice these configurations do have an impact on signal quality. Ambisonics *decoding* is the process by which the ambisonic sound field representation is converted into loudspeaker signals. This is normally performed by computing a weighted linear combination of the ambisonics signals, which may vary with frequency. This independence from loudspeaker topology presents benefits for applications, since a single set of ambisonics signals can be reproduced on a range of reproduction systems.

Ambisonics can be viewed as splitting amplitude panning into two decoupled stages, encoding and decoding. Encoding gives a set of intermediate ambisonics signals. These are commonly distributed to receivers. Then decoding to loudspeaker signals occurs in a receiver, where the loudspeaker arrangement is known. However, the two stages can also be combined into one process to perform ambisonics-equivalent panning. In contrast to VBAP which uses up to three active loudspeakers, ambisonics provides no limit on the number of loudspeakers used in panning. This is dependent upon the ambisonics order, the arrangement of loudspeakers and the decoding approach used.

In the 1990s, researchers began to generalise the application of ambisonics techniques to higher-orders (Poletti, 1996; Daniel, Rault, et al., 1998). When the order is greater than one, the term higher-order ambisonics (HOA) is frequently used. The theoretical formulation of ambisonics was also further developed, relating it to the solution of the wave equation in spherical coordinates, and representing sound sources and the reproduction loudspeakers as spherical wave sources (i.e. sources with finite distance) rather than plane wave sources (i.e. sources with infinite distance) (Daniel, 2000; Zotter, Pomberger, and Frank, 2009).

Whilst ambisonics offers a compact representation of full 3D scenes, there are errors introduced by this approach (Zotter, Pomberger, and Frank, 2009). Truncating the spherical harmonics series acts as a spatial low-pass filter, with lower ambisonics orders giving lower spatial resolution. This also reduces the region of accurate sound field reconstruction within the loudspeaker array (the *sweet spot*). Another source of error is angular spatial aliasing, caused by the sampling of the spherical harmonic signals by the loudspeaker positions. During reproduction with real loudspeakers, deviations from the model of an ideal point sources and interaction with the room will also have significant effects.

Theoretically, accurate decoding of ambisonics requires loudspeaker positions that preserve the orthonormality of the spherical harmonics functions. Decoding to real 3D loudspeaker arrays presents challenges, since it is rarely practicable to use a set of loudspeaker positions with this property. There has been much research into robust methods for decoding ambisonics for practical loudspeaker configurations which can provide good localisation performance, e.g. Zotter and Frank, 2012. This is less of an issue when considering virtual loudspeaker positions for headphone rendering, provided that head-related transfer functions (HRTFs) are available at the desired directions.

B.4.2 Ambisonic Theory

Zotter, Pomberger, and Frank (2009) presented a theoretical formulation of ambisonics. The sound field reproduction problem is approached by aiming to recreate the sound pressure within a spherical volume, using a distribution of sound sources on the boundary of a sphere. Loudspeakers are modelled as full-bandwidth, ideal point sources, which emit spherical waves. This formulation is reviewed here, as a means for then explaining the practical application of ambisonics.

B.4.2.1 Coordinate representations

The ambisonics formulation is considered in spherical coordinates with radius r, azimuth angle θ and elevation angle ϕ. The angular component of the position can be expressed as a Cartesian unit vector $\|\boldsymbol{\vartheta}\| = 1$:

$$\boldsymbol{\vartheta} = \begin{pmatrix} \cos(\theta)\cos(\phi) \\ \sin(\theta)\cos(\phi) \\ \sin(\phi) \end{pmatrix}, \tag{B.6}$$

so that position is given by the vector $\mathbf{r} = r\boldsymbol{\vartheta}$.

B.4.2.2 Spherical harmonics

The angular dependency of a signal can be described as a spherical Fourier series by weighting an infinite series of the spherical harmonic basis functions. The real-valued spherical harmonics functions, as used in ambisonics, are given by the following equations (Nachbar et al., 2011).

$$Y_n^m(\boldsymbol{\vartheta}) = Y_n^m(\theta,\phi) = N_n^{|m|} P_n^{|m|}(\sin\theta) \begin{cases} \sqrt{2}\cos|m|\theta & \text{for } m > 0 \\ 1 & \text{for } m = 0 \;, \\ \sqrt{2}\sin|m|\theta & \text{for } m < 0 \end{cases} \tag{B.7}$$

where P_n^m are the associated Legendre functions and N_n^m is a normalisation term. The series of spherical harmonics is indexed by symbols for the order and degree of the harmonic, respectively n and m, where $n \in 0, \dots, \infty$ (i.e. $n \in \mathbb{Z}^{\geq}$) and $-n \leq m \leq n$. The Legendre functions used in ambisonics are defined by

$$P_n^m(x) = (1-x^2)^{\frac{m}{2}} \frac{d^m}{dx^m} P_n(x), \quad m \geq 0, \tag{B.8}$$

with the Legendre polynomial $P_n(x)$ and without the Condon-Shortley phase $(-1)^m$ (ITU-R, 2017b).

Various normalisation schemes are used in ambisonics. The orthonormalised spherical harmonics are given by

$$N_{\text{ON}}{}_n^{|m|} = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}. \tag{B.9}$$

Orthonormalisation indicates the following integral over $\mathbb{S}^2$:

$$\int_0^{2\pi}\int_{-\pi}^{\pi} Y_n^m(\theta,\phi) Y_{n'}^{m'}(\theta,\phi) \cos\phi \, d\phi \, d\theta = \delta_{nn'}\delta_{mm'} \tag{B.10}$$

where δ denotes the Kronecker delta function. Figure B.5 shows the orthonormalised spherical harmonic functions up to the third order i.e. $n \leq 3$. Commonly the constant $\sqrt{\frac{1}{4\pi}}$ term is neglected, giving the N3D normalisation:

$$N_{\text{N3D}}{}_n^{|m|} = \sqrt{(2n+1)\frac{(n-|m|)!}{(n+|m|)!}}. \tag{B.11}$$

With N3D normalisation, the higher order components can have energy greater than that of the $n = 0$ component, which risks clipping distortion in some application scenarios. The SN3D normalisation is given by

$$N_{\text{SN3D}}{}_n^{|m|} = \sqrt{\frac{(n-|m|)!}{(n+|m|)!}}, \tag{B.12}$$

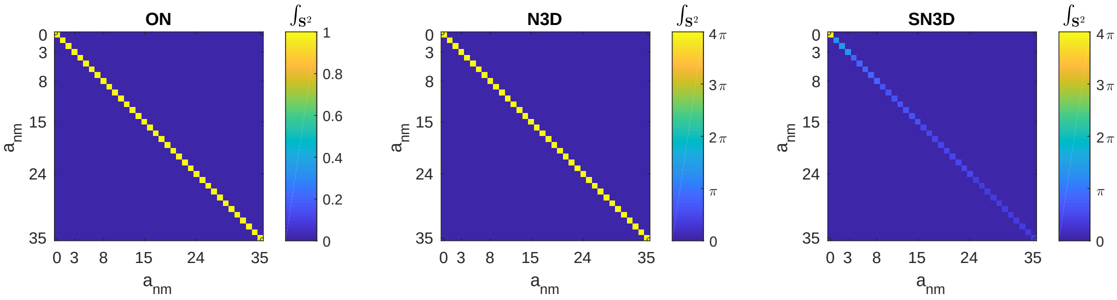


Figure B.6: Gram matrix for spherical harmonics up to 5th order using different normalisation schemes, where $a_{nm} = n^2 + n + m$ is the mode index.

which applies an order-dependent weighting to avoid the energy of components with $n > 0$ exceeding that of the $n = 0$ component.

These normalisation schemes can be visualised using the Gram matrix resulting from integration of the outer product of the spherical harmonics series over $\mathbb{S}^2$, as in equation (B.10). Figure B.6 shows the Gram matrix for spherical harmonics up to 5th order for each normalisation scheme. Here the mode index is given by $a_{nm} = n^2 + n + m$.

B.4.2.3 Physical Formulation

The acoustic input to the system is modelled as a theoretical, continuous sound source surrounding the spherical listening volume. The problem is described by the Helmholtz equation, with a "continuous spherical source strength distribution" (Zotter, Pomberger, and Frank, 2009) on the surface of a sphere of radius r_l

$$(\Delta + k^2)p = -\frac{\delta(r - r_l)}{r^2} f(\boldsymbol{\vartheta}) \tag{B.13}$$

wherein p is sound pressure, $k = 2\pi f/c$ is the acoustic wave number, Δ is the Laplace-operator, $f(\boldsymbol{\vartheta})$ is the continuous sound source strength distribution, and $\delta(r - r_l)$ is the Dirac delta function. Note that the area within this sphere is considered to be free of acoustic sources and scatterers, which is invalid in real-world applications where a listener is present.

B.4.2.4 Separation of variables

When there is a sound excitation, i.e. $f(\boldsymbol{\vartheta}) \neq 0$, we have the inhomogeneous Helmholtz equation, which is solved by separation of radial and angular variables (Morse and Ingard, 1987).

$$p(k,\mathbf{r}) = -ik\sum_{n=0}^{\infty}\sum_{m=-n}^{n} Y_n^m(\boldsymbol{\vartheta})\varphi_{nm}\cdot\begin{cases} h_n(kr_l)j_n(kr), & \text{for } r < r_l \\ j_n(kr_l)h_n(kr), & \text{for } r > r_l \end{cases} \tag{B.14}$$

where $i = \sqrt{-1}$, $Y_n^m(\boldsymbol{\vartheta})$ is a spherical harmonic of order n and degree m, and j_n and h_n are the order n spherical Bessel and Hankel functions of the first kind respectively.

The representation of the soundfield in equation (B.14) is valid for all values of **r**, both inside and outside of the loudspeaker sphere. Ambisonics considers only reproduction of sound pressure within the sphere of radius r_l. This leads to

$$p(k,\mathbf{r}) = -ik\sum_{n=0}^{\infty} j_n(kr)h_n(kr_l)\sum_{m=-n}^{n} Y_n^m(\boldsymbol{\vartheta})\varphi_{nm} \quad \forall \quad r < r_l. \tag{B.15}$$

Here φ_{nm} is the exact representation of the signal $f(\boldsymbol{\vartheta})$ in the spherical Fourier domain by a series of real-valued spherical harmonic coefficients. These two functions are related by the real spherical Fourier transform pair

$$\varphi_{nm} = \int_{S^2} f(\boldsymbol{\vartheta})Y_n^m(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}, \quad \forall \quad \begin{matrix} 0 \leq n \leq \infty \\ -n \leq m \leq n \end{matrix} \tag{B.16}$$

$$f(\boldsymbol{\vartheta}) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} \varphi_{nm}Y_n^m(\boldsymbol{\vartheta}). \tag{B.17}$$

B.4.2.5 Spherical expansion of a point source

A single point source at direction $\boldsymbol{\vartheta}_0$ is represented as an angular Dirac delta function in the formulation of equation (B.13):

$$f(\boldsymbol{\vartheta}) = \delta(1 - \boldsymbol{\vartheta}^{\mathsf{T}}\boldsymbol{\vartheta}_0), \tag{B.18}$$

which in the spherical harmonics domain is

$$\varphi_{nm} = \int_{S^2} \delta(1 - \boldsymbol{\vartheta}^{\mathsf{T}}\boldsymbol{\vartheta}_0)Y_n^m(\boldsymbol{\vartheta})d\boldsymbol{\vartheta} = Y_n^m(\boldsymbol{\vartheta}_0). \tag{B.19}$$

Inserting this into equation (B.15) gives the series expansion for a point source

$$p(k,\mathbf{r},\mathbf{r_0}) = -ik\sum_{n=0}^{\infty}\sum_{m=-n}^{n} j_n(kr)h_n(kr_0)Y_n^m(\boldsymbol{\vartheta})Y_n^m(\boldsymbol{\vartheta}_0) \quad \forall \quad r < r_0. \tag{B.20}$$

B.4.2.6 Angular band limitation

In practice the infinite series of spherical harmonics coefficients φ_{nm} is truncated to a maximum order, resulting in a finite set of coefficients. The truncation order N corresponds to the maximum order of the spherical harmonics used. The truncated spatial Fourier transform pair is

$$\varphi_{nm} = \int_{S^2} f(\boldsymbol{\vartheta}) Y_n^m(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}, \quad \forall \quad \begin{matrix} 0 \leq n \leq N \\ -n \leq m \leq n \end{matrix} \tag{B.21}$$

$$f(\boldsymbol{\vartheta}) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \varphi_{nm} Y_n^m(\boldsymbol{\vartheta}). \tag{B.22}$$

The resulting series of spherical harmonic coefficients has length $(N+1)^2$. The series truncation results in reduced angular resolution i.e. angular band-limitation.

The spherical wave approximation using a truncated Fourier-Bessel series is therefore (see equation (B.20))

$$p(k, \mathbf{r}, \mathbf{r_0}) = -ik \sum_{n=0}^{N} \sum_{m=-n}^{n} j_n(kr) h_n(kr_0) Y_n^m(\boldsymbol{\vartheta}) Y_n^m(\boldsymbol{\vartheta}_0) \quad \forall \quad r < r_0. \tag{B.23}$$

Zotter, Pomberger, and Frank (2009) derive the following equation for calculating the normalised squared error associated with a point source representation of finite order

$$\epsilon^2 = 1 - \frac{2kr_0kr}{\ln(\frac{r+r_0}{r-r0})} \sum_{n=0}^{N} (2n+1) |j_n(kr) h_n(kr_0)|^2. \tag{B.24}$$

This can be used to identify a radius in wavelengths r/λ for the central listening position within which resynthesis error does not exceed a given limit. Note that this still assumes a continuous spherical source distribution around the reproduction volume.

B.4.2.7 Discrete angular sampling

To apply ambisonics to the reproduction of sound fields on loudspeakers, the continuous spherical source must be discretised. A finite set of loudspeakers $l = 1, ..., L$ is arranged on the surface of the sphere with radius r_l at discrete angles $\boldsymbol{\vartheta}_l$, and driven using gains g_l. Modelling the loudspeakers as point sources gives the discrete source strength function

$$\hat{f}(\boldsymbol{\vartheta}) = \sum_{l=1}^{L} g_l \delta(1 - \boldsymbol{\vartheta}^{\mathsf{T}} \boldsymbol{\vartheta}_l), \tag{B.25}$$

resulting in the following inhomogeneous Helmholtz equation (from equation (B.13))

$$(\Delta + k^2)\hat{p} = -\frac{\delta(r - r_l)}{r^2} \sum_{l=1}^{L} g_l \delta(1 - \boldsymbol{\vartheta}^\mathsf{T} \boldsymbol{\vartheta}_l). \tag{B.26}$$

This formulation models loudspeakers as ideal point sources, emitting spherical waves at all frequencies. It also imposes the restriction that all loudspeakers are at distance r_l from the centre of the reproduction volume.

B.4.2.8 Modal source strength matching

The aim of sound field reproduction is to match the continuous and discrete source strength distributions[1]

$$f(\boldsymbol{\vartheta}) \overset{!}{\approx} \hat{f}(\boldsymbol{\vartheta}), \tag{B.27}$$

however it is not possible to control the values of the distribution between the sample points (loudspeaker positions) $\boldsymbol{r}_l$ because angular information is lost in the discretisation. Assuming that both the continuous and discrete source strength distributions are ideally band limited, modal source strength matching is achieved.

$$\hat{\varphi}_{nm} = \sum_{l=1}^{L} g_l Y_n^m(\boldsymbol{\vartheta}_l), \tag{B.28}$$

$$\varphi_{nm} \underset{g_l}{\overset{!}{=}} \hat{\varphi}_{nm}, \tag{B.29}$$

$$\forall \begin{array}{c} 0 \leq n \leq N \\ -n \leq m \leq n \end{array}.$$

This means that the gains g_l should be found such that the weighted sum of spherical harmonic projection of the loudspeaker directions equals the spatial Fourier expansion coefficients φ_{nm} for the target angularly-bandlimited continuous source strength distribution.

The problem can be reformulated in matrix notation. The following conventions are used

[1] The symbol $\overset{!}{=}$ denotes that the equality is the desired outcome and ˆ indicates a discretised function.

in the following sections (note the use of bold typeface to indicate matrices/vectors):

$$\boldsymbol{\varphi}_N \underset{\mathbf{g}}{\overset{!}{=}} \hat{\boldsymbol{\varphi}}_N \tag{B.30}$$

$$\hat{\boldsymbol{\varphi}}_N = \mathbf{Y}_{\mathrm{N}}\mathbf{g}, \tag{B.31}$$

$$\begin{aligned}
\text{with} \quad \mathbf{Y}_{\mathrm{N}} &= [\mathbf{y}_N(\boldsymbol{\vartheta}_1), \mathbf{y}_N(\boldsymbol{\vartheta}_2), \ldots, \mathbf{y}_N(\boldsymbol{\vartheta}_L)], \\
\mathbf{g} &= (g_1, g_2, \ldots, g_L)^{\mathsf{T}}, \\
\mathbf{y}_N(\boldsymbol{\vartheta}) &= [Y_0^0(\boldsymbol{\vartheta}), Y_1^{-1}(\boldsymbol{\vartheta}), Y_1^0(\boldsymbol{\vartheta}), Y_1^1(\boldsymbol{\vartheta}), \ldots, Y_N^N(\boldsymbol{\vartheta})]^{\mathsf{T}}, \\
\boldsymbol{\varphi}_N &= [\varphi_{0,0}, \varphi_{1,-1}, \varphi_{1,0}, \varphi_{1,1}, \ldots, \varphi_{N,N}]^{\mathsf{T}}, \\
\mathrm{diag}_N\{a_{nm}\} &= \mathrm{diag}\{\mathbf{a}_N\},
\end{aligned}$$

where $^{\mathsf{T}}$ is the matrix transpose operator. The operator diag forms a diagonal matrix from the generic vector $\mathbf{a}_N$ represented by the index operator $a_{nm} = n^2 + n + m$ $(\forall\ 0 \leq n \leq N, -n \leq m \leq n)$.

B.4.3 Ambisonics Decoding

To tackle the *modal source strength matching* problem of equation (B.30), a decoding matrix $\mathbf{D}_N$ is used. This matrix determines the decoding gains from the smoothed continuous modal source strength distribution

$$\mathbf{g} \overset{!}{=} \mathbf{D}_N \boldsymbol{\varphi}_N, \tag{B.32}$$

which, using equation (B.31), gives the smoothed reproduced modal source strength distribution

$$\hat{\boldsymbol{\varphi}}_N \overset{!}{=} \mathbf{Y}_{\mathrm{N}}\, \mathbf{D}_N\, \boldsymbol{\varphi}_N\ . \tag{B.33}$$

The matching is exact if $\mathbf{D}_N$ is the right inverse of $\mathbf{Y}_{\mathrm{N}}$. Decoding by modal source strength matching is well-conditioned if the right-inverse of $\mathbf{Y}_{\mathrm{N}}$ is well-conditioned. If this is not the case then an approximate matching must be found. For $\mathbf{Y}_{\mathrm{N}}$ to be well-conditioned, the sampling must preserve the orthogonality of the spherical harmonics. This requires a certain-kind of regular spacing over the sphere, which is often challenging to achieve in practice with real loudspeaker arrays.

For binaural rendering, decoding requires an HRTF for each desired sampling point. Appendix A.7 presented and verified a continuous HRTF model and so the virtual loudspeaker positions can be chosen freely. Therefore sampling grids that lead to well-conditioned $\mathbf{Y}_{\mathrm{N}}$ can be used. The following sections describe how to assess conditioning and find the decoding matrix.

B.4.3.1 Decoding as an inverse problem

The decoding task can be seen as a set of M equations with L unknown variables, where L is the number of loudspeakers and $M = (N+1)^2$ is the number of truncated spherical harmonic series coefficients. The conditioning of the problem can be inspected via the singular value decomposition (SVD) of $\mathbf{Y}_\mathrm{N}$. The compact SVD representation is

$$\mathbf{Y}_\mathrm{N} = \hat{\mathbf{U}}\,\hat{\mathbf{S}}\,\hat{\mathbf{V}}^\mathrm{T}, \tag{B.34}$$

where $\hat{\mathbf{S}} = \mathrm{diag}(\hat{s}_1, \ldots, \hat{s}_\mathrm{K}) \in \mathbb{R}^{\mathrm{KxK}}$ is a diagonal matrix containing the singular values in decreasing order ($\hat{s}_1 \geq \hat{s}_2 \geq \cdots \geq \hat{s}_\mathrm{K} \geq 0$). $\hat{\mathbf{U}} \in \mathbb{R}^{\mathrm{MxK}}$ and $\hat{\mathbf{V}} \in \mathbb{R}^{\mathrm{LxK}}$ are matrices with orthogonal columns, containing the left and right singular vectors respectively,

$$\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \ldots, \hat{\mathbf{u}}_\mathrm{M}) \tag{B.35}$$

$$\hat{\mathbf{V}} = (\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \ldots, \hat{\mathbf{v}}_L) \tag{B.36}$$

The left and right singular vectors have length K, and $K = \min\{L, M\}$.

The singular values $\hat{s}_1, \ldots, \hat{s}_\mathrm{K}$ reveal the conditioning of the decoding problem. Condition number and rank are commonly used metrics for the conditioning of a problem. The *condition number* indicates the numerical stability of the solution of the system of equations i.e. the decoding process. It is given by

$$\kappa = \hat{s}_1 / \hat{s}_\mathrm{K}. \tag{B.37}$$

The *rank* is the number of linearly independent row/column vectors in a matrix. It can be computed from the singular values of a matrix as the number of non-zero singular values. A system with a low condition number will have full rank and is termed *well-conditioned*, this means that a stable solution can be found. A system with a high condition number will be rank deficient and is termed *ill-conditioned*, this means that the solution is unstable in the presence of changes to the input. When the problem is well-conditioned $\hat{s}_\mathrm{K}$, the smallest singular value, is non-zero. In this case the SVD can be used to provide the inverse of $\mathbf{Y}_\mathrm{N}$,

$$\mathbf{D}_N = \hat{\mathbf{V}}\,\hat{\mathbf{S}}^{-1}\,\hat{\mathbf{U}}^\mathrm{T}. \tag{B.38}$$

B.4.3.2 Exact solutions

If $L = M$, and $\mathbf{Y}_\mathrm{N}$ has full rank, only one solution exists:

$$\mathbf{D}_N = \mathbf{Y}_\mathrm{N}^{-1} \tag{B.39}$$

If $L > M$, and $\mathbf{Y}_\mathrm{N}$ has full rank, the problem is underdetermined and many solutions exist. The right-inverse

$$\mathbf{D}_N = \mathbf{Y}_\mathrm{N}^\mathsf{T} \, (\mathbf{Y}_\mathrm{N} \, \mathbf{Y}_\mathrm{N}^\mathsf{T})^{-1} \tag{B.40}$$

gives the minimum-norm solution:

$$\text{minimise} \quad \|\mathbf{g}\|_2 \tag{B.41}$$

$$\text{subject to} \quad \boldsymbol{\varphi}_N = \mathbf{Y}_\mathrm{N} \, \mathbf{g} \tag{B.42}$$

This is the solution that minimises the loudspeaker gains.

When the sampling scheme is highly regular over the surface of the sphere, preserving the orthonormality of the spherical harmonics up to order N, requiring $L \geq M$, direct sampling of the modal source strength is possible. In a real system this is useful since it does not require matrix inversion or SVD. The sampling is performed using the transpose of the spherical harmonic coefficients, with a weighting factor dependent on the number of loudspeakers

$$\mathbf{D}_N = \frac{4\pi}{L} \mathbf{Y}_\mathrm{N}^\mathsf{T}. \tag{B.43}$$

For a defined quadrature grid that also preserves the orthonormality of the spherical harmonics up to order N, with sample weights characterised by $\sum_{l=1}^{L} w_l = 4\pi$ (equation (A.11)), the decoding matrix is defined by

$$\mathbf{D}_N = \mathrm{diag}\{\boldsymbol{w}_L\} \mathbf{Y}_\mathrm{N}^\mathsf{T}. \tag{B.44}$$

B.4.3.3 Sampling Schemes

An overview of sampling schemes for the integration of functions on the sphere is given in section A.7.1.4. Quadrature grids and uniform or nearly-uniform sampling schemes are available that will preserve the orthonormality of the spherical harmonics up to a given order. If such a grid is available then exact decoding/resynthesis up to order N is possible.

Zotter and Frank (2012) discuss the use of spherical t-designs, which are able to discretise an arbitrary spherical polynomial up to a given order t, and explain that for panning-invariant energy E and energy-spread $||\mathbf{r}_E||$, $t \geq 2\mathrm{N} + 1$. Published spherical designs are available[2] up to $t = 21$. These grids have equal sample weights.

B.4.3.4 Approximate Solutions

The formulation of the inverse discrete modal source strength in equation (B.38) can be used to find approximate pseudo-inverse solutions, when the problem is ill-conditioned. When $\hat{s}_{\mathsf{K}} = 0$ the condition number is infinite and $\mathbf{Y}_{\mathsf{N}}$ is rank-deficient, so an inverse matrix cannot be found. In practice singular values are often very small but non-zero, which can make inversion numerically unstable. Therefore a tolerance threshold is used, below which a singular value is considered to be zero. In MATLAB's `rank` function, the default tolerance value is given by

$$\alpha = \hat{s}_1 \, \epsilon \, \mathsf{max}\{L, M\}, \tag{B.45}$$

where ϵ is a small value indicating the precision of the floating-point number representation.

A numerically stable implementation of the pseudo-inverse can be obtained by performing truncated-SVD inversion, setting singular values $\hat{s}_k < \alpha$ to zero in equation (B.38). This is found, for example, in MATLAB's `pinv` function. The pseudo-inverse minimises $\|\mathbf{Y}_{\mathsf{N}}\mathbf{g} - \boldsymbol{\varphi}_{\mathsf{N}}\|_2$, i.e. a best-fit approximation of modal source strength matching is found in a least-squares sense.

Many other decoding approaches have been proposed for handling the irregularly-distributed loudspeaker layouts that are often used in practice. These use a range of different design goals, such as projecting local panning functions (e.g. VBAP) on to the spherical harmonics (Furse, 2010; Zotter and Frank, 2012), preserving decoded sound-field energy (Zotter, Pomberger, and Noisternig, 2012), non-linear optimisation of heuristic quality measures (Wiggins et al., 2003; Scaini and Arteaga, 2014).

B.4.3.5 Synthesis Equation

The reproduced modal source strength distribution has maximum degree Q and the target modal source strength distribution has maximum degree N. The reproduced source strength is unsmoothed, Q → ∞, since an array of real loudspeakers is used to create a discrete source strength distribution which is not angularly band-limited. Based on a vector form of equation (B.15), the synthesis of the sound field can then be expressed using

$$\hat{p}(kr, \boldsymbol{\vartheta}) = -ik \, \mathbf{y}_{\mathsf{Q}}^{\mathsf{T}}(\boldsymbol{\vartheta}) \, \mathsf{diag}_{\mathsf{Q}}\{j_n(kr)h_n(kr_l)\} \, \hat{\boldsymbol{\varphi}}_{\mathsf{Q}} \, . \tag{B.46}$$

[2] `http://neilsloane.com/sphdesigns/`

with the unsmoothed reproduced modal source strength

$$\hat{\boldsymbol{\varphi}}_{\mathsf{Q}} = \mathbf{Y}_{\mathsf{Q}}\, \mathbf{D}_{\mathsf{N}}\, \boldsymbol{\varphi}_{\mathsf{N}}\;. \tag{B.47}$$

The spherical wave spectrum of this resynthesis equation is obtained by the spatial Fourier transform (see equation (B.16)):

$$\hat{\boldsymbol{\psi}}_{\mathsf{Q}}(kr) = \int_{\mathbb{S}_2} \mathbf{y}_{\mathsf{Q}}(\boldsymbol{\vartheta})\hat{p}(kr, \boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta}, \tag{B.48}$$

and, since the spherical harmonics are orthonormal:

$$\int_{\mathbb{S}_2} \mathbf{y}_{\mathsf{Q}}(\boldsymbol{\vartheta})\mathbf{y}_{\mathsf{Q}}^{\mathsf{T}}(\boldsymbol{\vartheta})\mathrm{d}\boldsymbol{\vartheta} = \mathbf{I}, \tag{B.49}$$

this yields the following resynthesis equation in the spherical Fourier domain:

$$\hat{\boldsymbol{\psi}}_{\mathsf{Q}}(kr) = -ik\;\mathrm{diag}_{\mathsf{Q}}\{j_n(kr)h_n(kr_l)\}\;\mathbf{Y}_{\mathsf{Q}}\mathbf{D}_{\mathsf{N}}\boldsymbol{\varphi}_{\mathsf{N}}. \tag{B.50}$$

This can be used more easily in analysis of the squared resynthesis error.

B.4.3.6 Distance correction in reproduction

When aiming to accurately synthesise an encoded sound field using equation (B.46), correct reproduction of the radial function should be considered as well as the angular function. If the loudspeaker distance does not match the source encoding distance, the effect of loudspeaker distance on the generated sound field $h_n(kr_l)$ is compensated for by adding a distance coding factor into the source strength distribution

$$\boldsymbol{\varphi}_{\mathsf{N}} = \mathrm{diag}_{\mathsf{N}}\left\{\frac{h_n(kr_s)}{h_n(kr_l)}\right\}\tilde{\boldsymbol{\varphi}}_{\mathsf{N}}. \tag{B.51}$$

With an exact decoding solution and when $\mathsf{Q} = \mathsf{N}$ (only theoretically possible), $\mathbf{Y}_{\mathsf{Q}}\mathbf{D}_{\mathsf{N}}$ is equal to the identity matrix. This achieves a replacement of the radial function for the loudspeakers by the radial function for the source in the spherical wave spectrum

$$\boldsymbol{\psi}_{\mathsf{N}}(kr) = -ik\;\mathrm{diag}_{\mathsf{N}}\{j_n(kr)h_n(kr_s)\}\;\tilde{\boldsymbol{\varphi}}_{\mathsf{N}}\;. \tag{B.52}$$

That $\mathsf{Q} \neq \mathsf{N}$ means that this expression becomes an approximation.

B.4.3.7 Resynthesis error

The resynthesis error is the difference between the synthesised and target fields

$$\mathbf{e}_{\mathrm{Q}} = \boldsymbol{\psi}_{\mathrm{Q}}(kr) - \hat{\boldsymbol{\psi}}_{\mathrm{Q}}(kr), \tag{B.53}$$

for an angularly band-limited point source at direction $\boldsymbol{\vartheta}_0$ it is defined as

$$\mathbf{e}_{\mathrm{Q}} = ik\,\mathrm{diag}_{\mathrm{Q}}\{j_n(kr)h_n(kr_l)\}\left[\mathbf{I} - \mathbf{Y}_{\mathrm{Q}}(\mathbf{D}_{\mathrm{N}}, \mathbf{0})\right]\mathbf{y}_{\mathrm{Q}}(\boldsymbol{\vartheta}_0). \tag{B.54}$$

wherein $(\mathbf{D}_{\mathrm{N}}, \mathbf{0})$ indicates that $\mathbf{D}_{\mathrm{N}}$ is zero-padded to the right up to $(\mathrm{Q}+1)^2$ columns. The error depends on the direction of the virtual source $\boldsymbol{\vartheta}_0$, the radius of the loudspeaker array r_l, and the radius of observation r. The squared error, as in equation (B.24), is $\epsilon^2 = \|\mathbf{e}_{\mathrm{Q}}\|^2$.

This equation is useful to analyse the spatial aliasing contributions of the spherical harmonics above the limited encoding order. The decoding can be exact for orders $0 \leq n \leq \mathrm{N}$, however there is implicit spatial aliasing at higher orders $\mathrm{N} < n \leq \mathrm{Q}$. Analysis is often considered with truncated resynthesis order $\mathrm{Q} = \mathrm{N}$, however in reality it is not feasible to reproduce a band-limited section of the spherical wave spectrum in isolation. Loudspeaker sources have a full bandwidth spherical wave spectrum ($\mathrm{Q} \to \infty$).

Finally, Zotter, Pomberger, and Frank (2009) provide a resynthesis error equation for when $\mathrm{Q} \to \infty$ with source radius r_0 not limited to speaker radius r_l:

$$\epsilon^2 = \frac{8\pi k r_0 k r}{\ln\left(\frac{r+r_0}{r-r_0}\right)} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} |e_{nm}|^2, \tag{B.55}$$

with

$$e_{nm} = j_n(kr)\left[h_n(kr_0)Y_n^m(\boldsymbol{\vartheta}_0) - h_n(kr_l)\sum_{l=1}^{\mathrm{L}} Y_n^m(\boldsymbol{\vartheta}_l)\mathbf{d}_l^{\mathsf{T}}\mathrm{diag}_{\mathrm{N}}\left\{\frac{h_{n'}(kr_0)}{h_{n'}(kr_l)}\right\}\mathbf{y}_{\mathrm{N}}(\boldsymbol{\vartheta}_0)\right], \tag{B.56}$$

where $\mathbf{d}_l^{\mathsf{T}}$ represents the l^{th} row of the decoder matrix $\mathbf{D}_{\mathrm{N}}$. This formulation captures the following aspects of sound-field synthesis using ambisonics:

- Angularly band-limited encoding of the angular $\boldsymbol{\vartheta}_0$ and radial r_0 positions of a point source at order N.

- Decoding to a discrete array of L loudspeakers at directions $\boldsymbol{\vartheta}_l$ and fixed distance r_l.

- Reproduction by these loudspeakers, represented as ideal point-sources.

The sources of error therefore represented in this resynthesis error equation are:

- Angular band-limitation errors due to truncation of spherical wave spectrum of the target continuous source strength distribution.
- Any decoding errors that might be present in reproducing the target angularly-band-limited sound field with loudspeakers considered as angularly-band-limited point sources.
- Spatial aliasing errors due to reproduction of the angularly-band-limited sound field representation by point sources of infinite angular bandwidth.

Beyond these theoretical considerations there will be other errors introduced into the sound field synthesis process. Real loudspeakers do not act as point sources and there will be room effects. The above analysis considers encoded virtual sources that can be ideally angularly-band-limited. When real sound fields are recorded, e.g. with spherical microphone arrays, additional errors will be introduced (Rafaely, 2005; Moreau et al., 2006).

B.4.4 Ambisonics Encoding

A signal s is encoded into an ambisonics representation by projection onto the spherical harmonics basis functions to obtain the desired modal source strength distribution $\boldsymbol{\varphi}_{\mathsf{N},s}$. With reference to equation (B.19), an arbitrary source signal s is encoded as an angularly band-limited virtual point source at direction $\boldsymbol{\vartheta}_s$ according to

$$\boldsymbol{\varphi}_{\mathsf{N},s} = \mathbf{y}_{\mathsf{N}}(\boldsymbol{\vartheta}_s)s\ . \tag{B.57}$$

If no distance correction is carried out in reproduction then the source will be rendered at the distance of the loudspeakers. Distance encoding can be used to encode a source at arbitrary distance, with respect to the target loudspeaker array radius or, if this is not known, a reference distance could be used.

$$\boldsymbol{\varphi}_{\mathsf{N},s,r_{\mathsf{ref}}} = \mathsf{diag}_{\mathsf{N}}\left\{\frac{h_n(kr_s)}{h_n(kr_{\mathsf{ref}})}\right\}\mathbf{y}_{\mathsf{N}}(\boldsymbol{\vartheta}_s)s\ . \tag{B.58}$$

The reference distance r_{ref} would then be substituted for r_s in equation (B.51) for correct adjustment during reproduction. This kind of distance coding was introduced by Daniel (2003). It results in complex-valued frequency-dependent encoding coefficients, which adds complexity to implementations.

B.4.5 Modal Weighting

Angular band limitation leads to a point source representation which is a rotationally-symmetric sinc-like function, with a series of side-lobes, the largest of which is in the opposite direction to the target position. These side-lobes can cause mis-localisation, particularly in loudspeaker playback with an off-centre listening position. A modal-weighting can be applied to reduce the energy of higher-order ambisonics channels. This performs angular smoothing and reduces these side lobes; it can be described as a spherical convolution operation (Zotter, Pomberger, and Noisternig, 2010).

An order-dependent weighting vector $\mathbf{a}_{\mathrm{N}}$ can be applied in virtual point source encoding of signal s as follows

$$\boldsymbol{\varphi}_{\mathrm{N},s} = \mathrm{diag}\{\mathbf{a}_{\mathrm{N}}\}\mathbf{y}_{\mathrm{N}}(\boldsymbol{\vartheta}_s)s. \tag{B.59}$$

It can alternatively be applied during decoding

$$\boldsymbol{g} = \mathbf{D}_{\mathrm{N}}\mathrm{diag}\{\mathbf{a}_{\mathrm{N}}\}\boldsymbol{\varphi}_{\mathrm{N}}. \tag{B.60}$$

B.4.5.1 Max $\mathbf{r}_E$ weights

Daniel (2000, section A.4.2) defines a solution for the “max $\mathbf{r}_E$” weights, which minimises the energy-spread around the target direction by maximising $||\mathbf{r}_E||$. The maximum attainable value of $||\mathbf{r}_E||$, named $\mathbf{r}_{E\mathrm{max}}$, is the largest root of the Legendre polynomial of order $\mathrm{N}+1$, i.e. the maximum of the $\mathrm{N}+1$ solutions to:

$$P_{\mathrm{N}+1}(\mathbf{r}_{E\mathrm{max}}) = 0 \tag{B.61}$$

and the modal weights are then defined as:

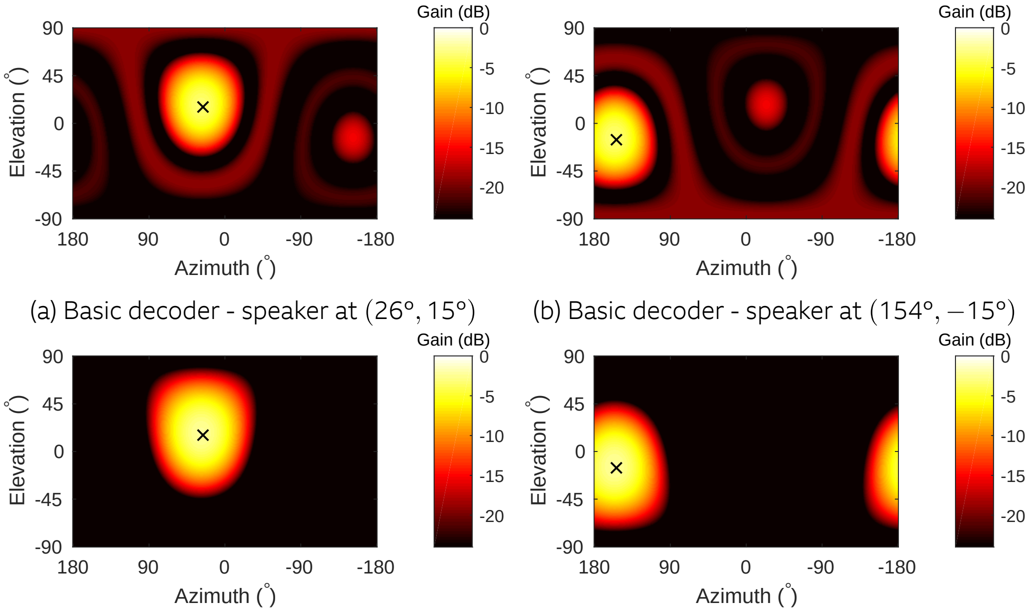
$$a_n = P_n(\mathbf{r}_{E\mathrm{max}}) \qquad \text{for} \quad n = 0, 1, \ldots, \mathrm{N} \tag{B.62}$$

Zotter and Frank (2012) state that these modal weights can be sufficiently approximated by

$$a_n = P_n\left(\cos\left(\frac{137.9°}{\mathrm{N}+1.51}\right)\right) \tag{B.63}$$

Figure B.7 shows the effect of this modal weighting on the resulting panning functions obtained with 3rd-order ambisonics for a 24-loudspeaker spherical t-design. The side-lobes are reduced, but at the expense of a wider main lobe.

Daniel (2000, p. 312) provides an analytical expression for the energy vector norm in the



(a) Basic decoder - speaker at (26°, 15°)

(b) Basic decoder - speaker at (154°, −15°)

(c) Max $\mathbf{r}_E$ decoder - speaker at (26°, 15°)

(d) Max $\mathbf{r}_E$ decoder - speaker at (154°, −15°)

Figure B.7: Comparison of panning gains given by basic and max $\mathbf{r}_E$ decoding of 3rd-order ambisonics for a 24-loudspeaker spherical t-design. Panning gains are shown for two loudspeakers with varying panning angles (θ, ϕ). Loudspeaker positions are marked by a black cross. Gains are shown by colour in decibels, limited to range the −24 dB–0 dB.

ideal case of a continuous array of loudspeakers based on the modal weights

$$||\mathbf{r}_E|| = \frac{2\sum_{n=1}^{N} n a_n a_{n-1}}{\sum_{n=0}^{N}(2n+1)a_n^2}. \tag{B.64}$$

This can be compared to the result for the discretised version to understand the influence of the loudspeaker array.

B.4.5.2 Scaling max $\mathbf{r}_E$ modal weights

The reduction of modal-weights to minimize energy-spread (max $\mathbf{r}_E$), whilst preserving the amplitude at the central point of the reproduction array, will decrease the reproduction energy over the sphere. Daniel (2000, p.183) provides a formula to calculate this energy reduction:

$$E_{\{a_n\}} = \sum_{n=0}^{\mathrm{N}}(2n+1)a_n^2 \tag{B.65}$$

where a_n are the modal weights at each spherical harmonic order n and $2n+1$ is the number of ambisonics signals at that order. To preserve the total energy over the sphere for max $\mathbf{r}_E$ weighting, the modal weights a_n are multiplied by a scaling factor

$$a_0 = \sqrt{\mathrm{L}/E_{\{a_n\}}} \tag{B.66}$$

This will be termed "energy-preserving" weighting from here on, whereas "amplitude-preserving" implies $a_0 = 1$.

B.4.5.3 Dual-band decoding

It is common for the decoding to be divided into low and high frequency solutions. Decoding without modal-weighting, often called *basic* decoding, is generally used at low frequencies because this preserves the $\|\mathbf{r}_V\|$, and the max $\mathbf{r}_E$ weighted decoder is used at high frequencies (Daniel, Rault, et al., 1998). A phase-aligned crossover filter, as described by Linkwitz (1976), is used to combine the outputs of the two decoders. For binaural processing, this crossover can be applied when generating the filters, rather than in real-time decoding. The crossover frequency at which this split occurs is another parameter in the design of decoder; it should be set around the frequency above which spatial aliasing error becomes too large, where an energy based decoding is more useful.

B.4.6 Binaural Rendering of Ambisonics with Loudspeaker Virtualisation

Headphone reproduction of ambisonics signals can be achieved by decoding to a virtual loudspeaker array, with subsequent virtualisation by binaural rendering. An appealing aspect for dynamic binaural rendering is that head tracking can be applied rather efficiently in the ambisonics domain, by rotation of the spherical harmonic coefficients. Travis (1996) presented an early discussion of the use of ambisonics as an intermediate virtualisation format in a VR audio system. By performing rotation on the ambisonics signals, the HRTF filters for virtualisation can be kept static, i.e. time-invariant. The ambisonics decoding process is linear and time-invariant. It is therefore usually more efficient to apply the decoding matrix to the HRTF filters, generating a new set of binaural filters that are applied directly to the ambisonics signals. Herein these will be called ambisonics-to-binaural filters. This approach is described, for example, by Wiggins (2004, p.104) for first-order ambisonics, and by Politis and Poirier-Quinot (2016) for higher-order ambisonics.

Ambisonics-to-binaural rendering can be described mathematically in the temporal-frequency-domain:

$$p_{lr}(f) = \mathbf{H}_{lr}(f)\mathbf{D}_{\mathsf{N}}\mathrm{diag}\{\mathbf{a}_{\mathsf{N}}\}\boldsymbol{\varphi}_{\mathsf{N}}(f) \tag{B.67}$$

where p_{lr} is the rendered binaural signal and $\mathbf{H}_{lr}$ is a matrix of HRTFs for each virtual loudspeaker

$$\mathbf{H}_{lr}(f) = \begin{bmatrix} h_l(f,\boldsymbol{\vartheta}_1) & h_r(f,\boldsymbol{\vartheta}_1) \\ \dots & \dots \\ h_l(f,\boldsymbol{\vartheta}_{\mathsf{L}}) & h_r(f,\boldsymbol{\vartheta}_{\mathsf{L}}) \end{bmatrix}^{\mathsf{T}}. \tag{B.68}$$

The ambisonics-to-binaural filters are obtained according to:

$$\boldsymbol{\Phi}_{\mathsf{N}lr}(f) = \mathbf{H}_{lr}(f)\mathbf{D}_{\mathsf{N}}\mathrm{diag}\{\mathbf{a}_{\mathsf{N}}\}. \tag{B.69}$$

However, this formulation does not incorporate rotation to allow for head tracking. The spherical harmonics rotation matrix for rotation angles (α, β, γ) is denoted by $\mathbf{R}_{\mathsf{N}}$. For the real spherical harmonics it can be efficiently obtained by recursion relations (Ivanic, 1996). The rotated sound field is given by

$$\boldsymbol{\varphi}_{\mathsf{N}}^{\mathsf{rot}} = \mathbf{R}_{\mathsf{N}}(\alpha,\beta,\gamma)\boldsymbol{\varphi}_{\mathsf{N}} \tag{B.70}$$

which leads to the following ambisonics-to-binaural rendering equation

$$p_{lr}^{\mathsf{rot}} = \boldsymbol{\Phi}_{\mathsf{N}lr}\mathbf{R}_{\mathsf{N}}(\alpha,\beta,\gamma)\boldsymbol{\varphi}_{\mathsf{N}}. \tag{B.71}$$

Figure B.8 presents the structure of a binaural renderer that uses these techniques to process an object-based scene via an intermediate ambisonics format. This involves ambisonics encoding of source positions, which mixes all sources into a single ambisonics-format multichannel signal bus. The ambisonics signals can then be rotated to account for changes in listener head orientation received from the head tracker. The encoding step may also incorporate translation according to listener position to allow 6 DoF tracking. The ambisonics signals are then rendered to a binaural signal for headphone playback using static convolution, with pre-calculated ambisonics-to-binaural filters. This rendering approach has been implemented into the apparatus described in appendix A to permit investigation of the binaural rendering of ambisonics.

B.5 Analysing Loudspeaker Rendering

From the literature there are various methods for objective analysis of loudspeaker rendering. Whilst not a substitute for perceptual evaluation, these can be useful to gain insights into the behaviour of techniques and to help in explaining the perceptual effects observed.

B.5.1 Gerzon Vector Analysis

Much of the ambisonics literature uses the so-called *Gerzon vectors* to analyse the performance of an ambisonic panning system, after Gerzon's "meta-theory of auditory localisation" (Gerzon, 1992a). The linear and squared summation of loudspeaker amplitudes leads to relative measures of the playback magnitude of a panned source in terms of pressure and energy, respectively.

$$P = \sum_{l=1}^{L} g_l \quad E = \sum_{l=1}^{L} |g_l|^2 \tag{B.72}$$

The centroids of the loudspeaker amplitudes and their squares are then used as a measure of the directional concentration

$$\mathbf{r}_V = \frac{\sum_{l=1}^{L} g_l \boldsymbol{\vartheta}_l}{P} \tag{B.73}$$

$$\mathbf{r}_E = \frac{\sum_{l=1}^{L} |g_l|^2 \boldsymbol{\vartheta}_l}{E} \tag{B.74}$$

where $\boldsymbol{\vartheta}_l$ refers to the unit-length loudspeaker direction vector. They point in the direction where the loudspeaker amplitude or their squares appear strongest and their length relates to the directional concentration. These are known in the ambisonics literature respectively as the velocity and energy vectors, see e.g. (Heller et al., 2008). $\mathbf{r}_V$ involves summation

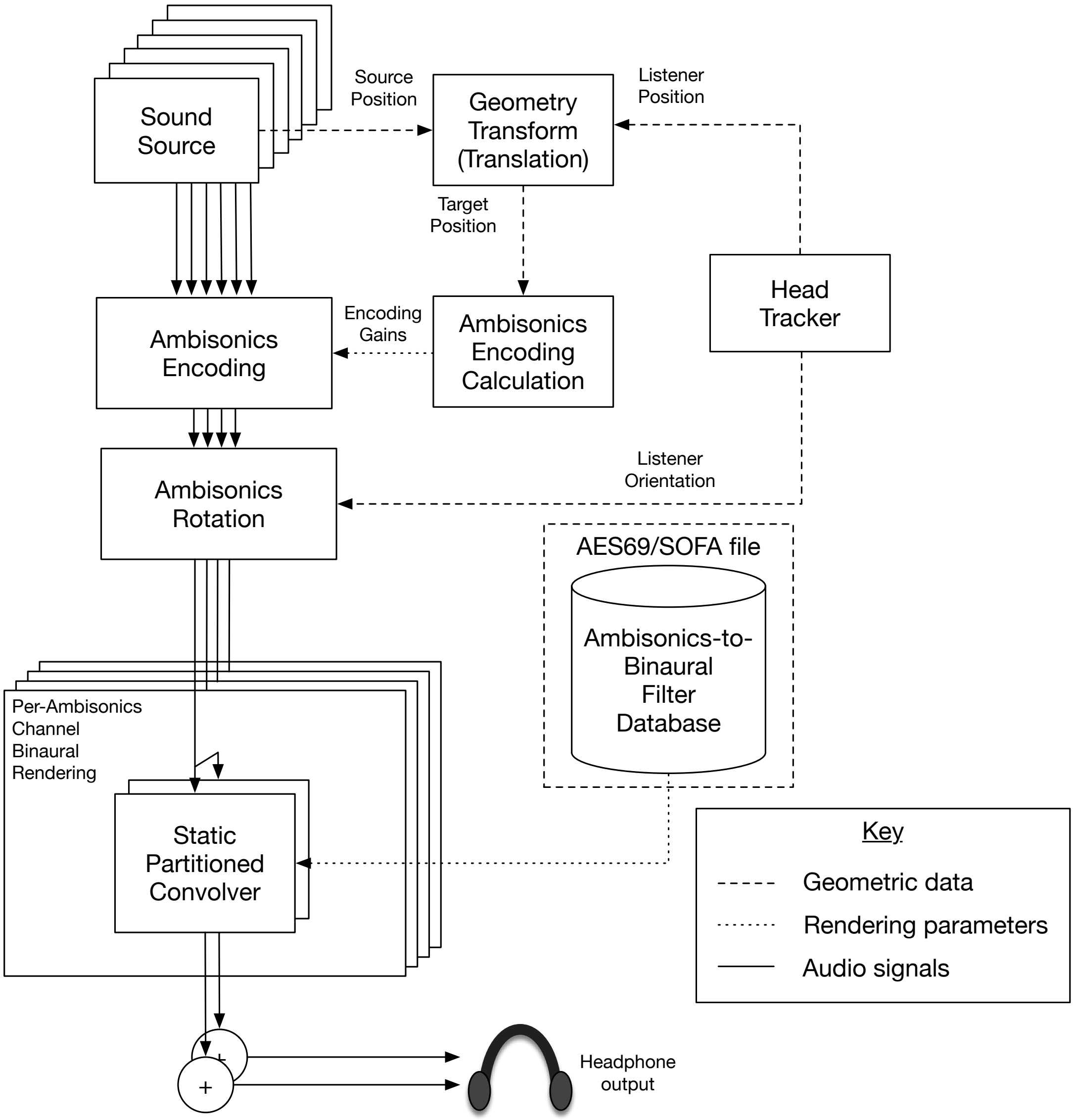


Figure B.8: Structure of a binaural renderer that uses an intermediate ambisonics format. Multiple sound sources are first encoded into ambisonics. Rotation is performed in the ambisonics domain before binaural rendering by static convolution with pre-calculated ambisonics-to-binaural filters.

with fixed phase relation and as such is valid for low frequencies. In the expression for $\mathbf{r}_E$ the squared amplitudes are proportional to the energy radiated by the loudspeakers. This is valid at high frequencies where wavelengths are smaller than the head and so the phase relationships are not fixed, but vary according to frequency. The velocity and energy vectors are therefore equated to low and high frequency sound localisation cues according to ITD and ILD respectively, though this is only a simple approximation.

The measures of playback magnitude (P and E) are often equated to subjective loudness whilst the centroid norms ($||\mathbf{r}_V||$ and $||\mathbf{r}_E||$) are equated to perceived source width. Whilst this analysis is quite primitive in relation to auditory processes, it does lead to some simple high level design goals for the system behaviour. These have been used by Gerzon to define ambisonic decoders (Gerzon, 1992c) and panning laws (Gerzon, 1992b).

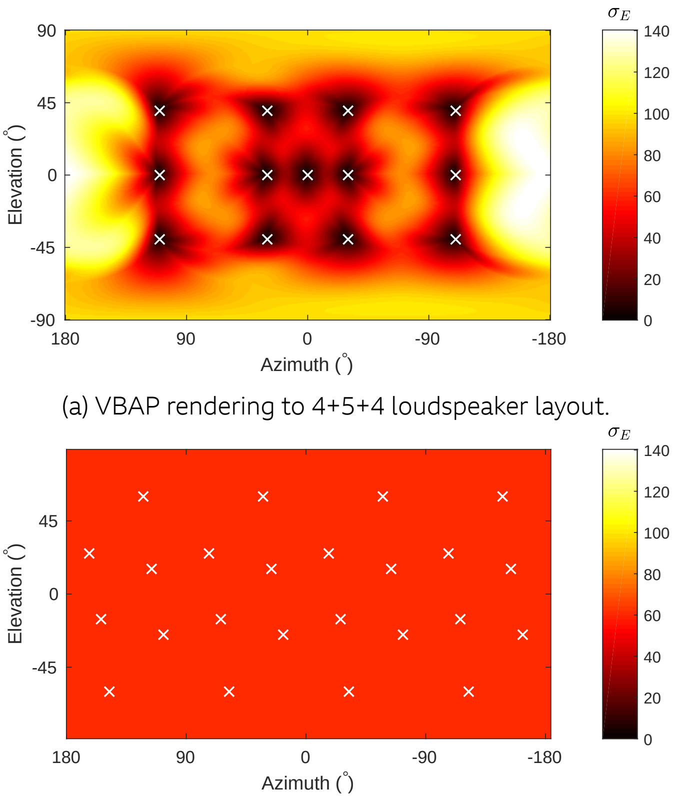
Heller et al. (2008) present metrics based on the velocity and energy vectors for the design of ambisonic decoders, following the work of Gerzon and Barton (1992). Scaini and Arteaga (2014) use heuristic search algorithms to generate decoders for irregular layouts based on these metrics also. Daniel, Rault, et al. (1998) and Zotter, Pomberger, and Noisternig (2012) also describe objective quality metrics based on the Gerzon analysis. The objectives can be summarised as:

- The panning pressure P and energy E should be equal to 1 at low and high frequencies, respectively, to preserve loudness.
- The direction of the vectors $\mathbf{r}_V$ and $\mathbf{r}_E$ should match the target source direction.
- $||\mathbf{r}_V||$ should equal 1 to correctly reconstruct ITDs (Daniel, Rault, et al., 1998).
- The value of $||\mathbf{r}_E||$ should be maximised to be as close as possible to 1 to minimise energy spread and hence to reduce localisation blur.
- The panning energy E and energy spread $||\mathbf{r}_E||$ should ideally be constant with direction.

These simple metrics can be used to gain some insights into the behaviour of panning algorithms and ambisonics decoders. However, in-depth analysis requires binaural auditory modelling and perceptual evaluation.

Frank (2013) demonstrated strong correlation between $||\mathbf{r}_E||$ and perceived source width, as measured in listening test data. The angular spread can be determined from the energy vector norm as follows

$$\sigma_E = 2\text{arccos}\|\mathbf{r}_E\|. \tag{B.75}$$



(a) VBAP rendering to 4+5+4 loudspeaker layout.

(b) 3rd-order ambisonics rendering to 24-speaker spherical t-design.

Figure B.9: Angular spread σ_E estimated from energy vector norm $\|\mathbf{r}_E\|$. Loudspeakers are indicated by white crosses.

As an example of this type of analysis, figure B.9a shows the estimated angular spread for VBAP to the 4+5+4 loudspeaker layout shown in figure B.3. It can be seen that spread is low at the position of the loudspeakers, where $\|\mathbf{r}_E\| = 1$. Much greater spreading can be seen in the rear region where the aperture between loudspeakers is 140° in azimuth. By constrast, figure B.9b shows the same analysis for ambisonics-based panning to a 24-loudspeaker array based on a spherical t-design, with max $\mathbf{r}_E$ decoding. Here $\|\mathbf{r}_E\|$ is constant across the sphere with value 0.86, yielding $\sigma_E = 61.1°$. This sampling scheme has other desirable properties in terms of the Gerzon vector analysis: the direction of $\mathbf{r}_V$ and $\mathbf{r}_E$ always matches the target source direction, P, E, $||\mathbf{r}_V||$ and $||\mathbf{r}_E||$ are constant with direction, P and $||\mathbf{r}_V||$ are unity for the modal source strength matching decoder, and E is unity when max $\mathbf{r}_E$ modal weighting is applied. These properties exist for all spherical t-designs where $t \geq 2\text{N} + 1$.

B.5.2 Sound Field Analysis

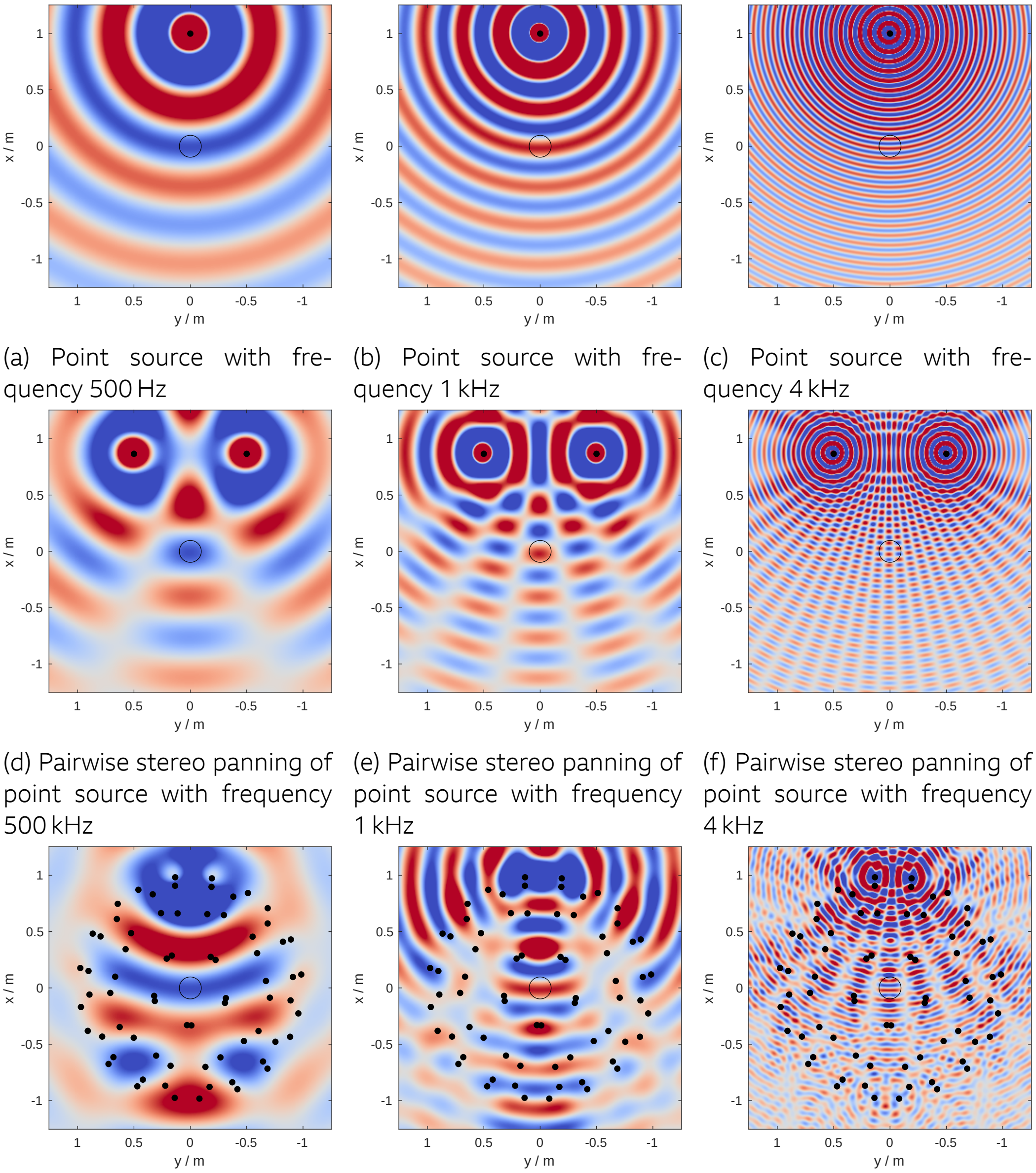
Another method of analysis is to plot the reconstruction of the sound field over the listening area, assuming a free field and ideal monopole sources. This illustrates the physical process of sound field reconstruction rather than the perceptual effects, but can be very useful for gaining insights into the processes involved and the effect of certain parameters in the design of ambisonic systems.

The error across the soundfield is compared with the target, e.g. a plane wave or spherical wave at a certain frequency. This can be useful for demonstrating and separating the effects of order truncation, discretisation and near-field correction (or lack thereof). It can also help in understanding the effects perceived at off-centre listening positions. For binaural rendering of ambisonics, it has more limited use, since the listener is fixed at the centre of the virtual array, although it can be used to assess the correct reconstruction around the volume of the head at different frequencies.

Such analysis has been used frequently in the literature. Daniel, Rault, et al. (1998) used this approach to indicate valid frequencies of reconstruction for a given truncation order, with an ideal continuous driving function. Spors and Ahrens (2008) demonstrate changes in spatial reconstruction error in HOA with temporal frequency. Ahrens, Wierstorf, et al. (2010) demonstrate time-domain artefacts in HOA such as pre-ringing. Zotter, Pomberger, and Frank (2009) break down the effects of series truncation and discretisation, discussing the effects of spatial aliasing in particular.

Figure B.10 gives an example of sound field analysis in the horizontal plane for a steady-state monochromatic (single-frequency) frontal point source at three frequencies, as well as the rendering achieved by two-channel stereo panning and 5th-order ambisonics rendering using a 70-point spherical t-design of virtual loudspeakers. This analysis was conducted using the Soundfield Synthesis Toolbox in MATLAB (Wierstorf and Spors, 2012). The decoder was obtained using the direct sampling method (equation (B.43)). It can be seen that the region of accurate reproduction reduces as temporal frequency increases. Reproduction is more accurate using 5th-order ambisonics compared to stereo panning, but at 4 kHz the region of accurate reproduction becomes smaller than head size.

In the time-domain, Figure B.11 shows the impulse response of a broadband point source, and the rendering of that source with pair-wise stereo panning and 5th-order ambisonics rendering. The two ambisonics rendering cases use a modal source strength matching decoder and a max $\mathbf{r}_E$ weighted decoder, respectively. The impulse response is shown at 2.915 ms, the time at which the wave fronts should meet the origin for $c = 343$ m/s, and at 1.458 ms, at which time the wave fronts should be half-way to the origin. With 5th-order



(a) Point source with frequency 500 Hz

(b) Point source with frequency 1 kHz

(c) Point source with frequency 4 kHz

(d) Pairwise stereo panning of point source with frequency 500 kHz

(e) Pairwise stereo panning of point source with frequency 1 kHz

(f) Pairwise stereo panning of point source with frequency 4 kHz

(g) 5$^{\text{th}}$-order ambisonics rendering of point source with frequency 500 kHz

(h) 5$^{\text{th}}$-order ambisonics rendering of point source with frequency 1 kHz

(i) 5$^{\text{th}}$-order ambisonics rendering of point source with frequency 4 kHz

Figure B.10: Soundfield analysis of a monochromatic point source at position (0°, 0°, 1 m) at three frequencies, and its approximation by stereo panning and by 5$^{\text{th}}$-order ambisonics rendering with a 70-loudspeaker t-design with distance $r_l = 1$ m. Pressure fields are presented on a linear scale, normalised to the pressure at the central listening position ($r = 0$) and the colour axis is clipped to the range [-1 1]. The circle has a radius of 9.75 cm corresponding to the spacing of the microphone capsules in the Neumann KU100 dummy head.

ambisonics rendering there are many more contributing loudspeaker signals. It can be seen how max $\mathbf{r}_E$ decoding reduces the out-of-phase contributions (blue lines), but broadens the region of high energy around the target direction.

B.5.3 Auditory Analysis

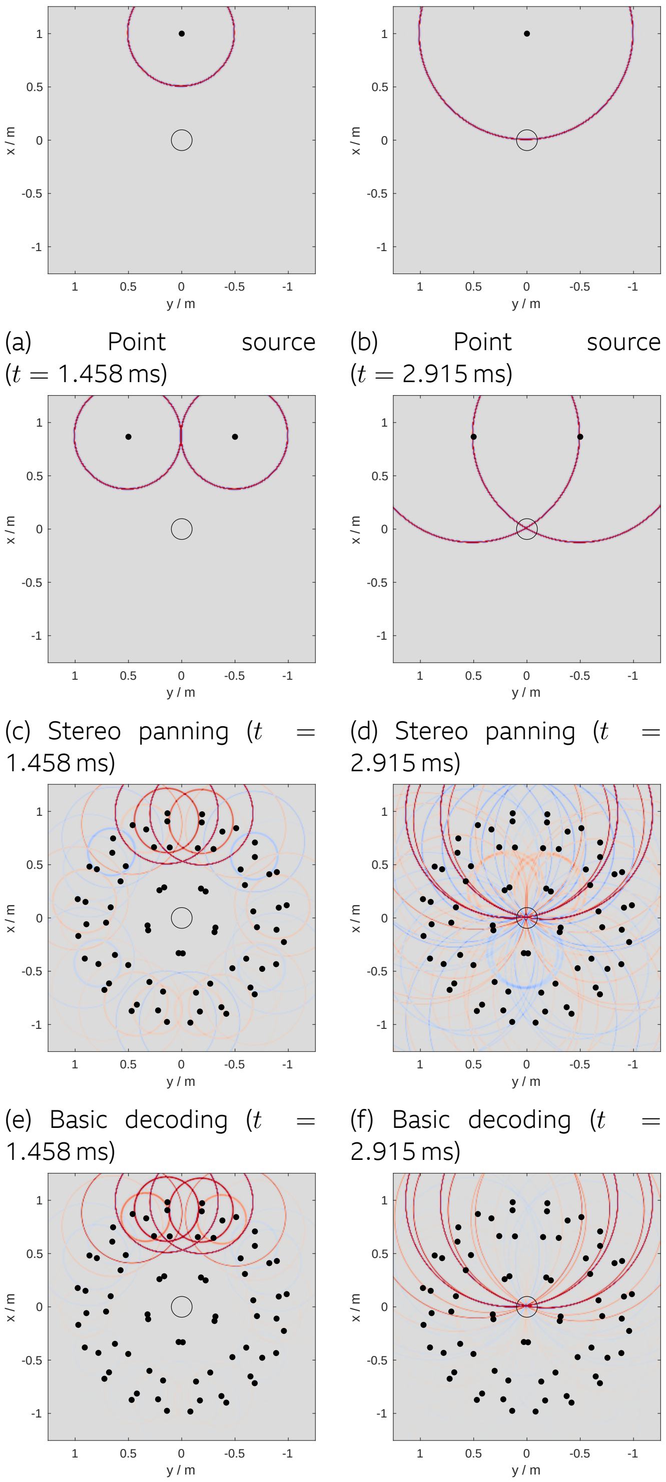
Loudspeaker rendering and panning techniques can also be analysed in terms of the auditory cues provided in a binaural signal. This analysis can be made by binaural recording of a real loudspeaker system, or by simulation of the loudspeaker array with HRTFs or binaural room impulse responses (BRIRs) i.e. binaural rendering by loudspeaker virtualisation. There are assumptions in this approach, the analysis is specific to the listener, or the HRTF set, and is only valid for the listening position at which the binaural signal is measured/simulated. But the advantage is that it allows direct assessment of the cues created at the listener's ears.

Using HRTFs to represent the transfer function from each loudspeaker to the ears means assuming anechoic reproduction, which may be a limitation when evaluating real loudspeaker reproduction. However, for loudspeaker virtualisation applications, this is often exactly the scenario used in reproduction, therefore the effects can be directly interpreted.

The binaural transfer functions generated by rendering through a loudspeaker virtualisation approach can be obtained by combining panning and loudspeaker virtualisation, or by ambisonics encoding and ambisonics-to-binaural decoding. These transfer functions can then be compared to measured HRTFs in terms of the auditory cues they generate, such as ILDs and ITDs, as well as by their spectral magnitude responses.

Figure B.12 shows an example of such an analysis for virtualised VBAP, using the 4+5+4 array introduced in appendix B.3.2. The transfer function of the renderer for a specific target source direction can be compared to a measured HRTF for the same direction, as in figure B.12a. There are distinct differences in the temporal and magnitude responses at the tested source direction $(90°, 0°)$. To gain an overview of the differences across all directions, transfer functions were synthesised at all points in the 16 020-point Gauss-Legendre quadrature grid, for which HRTF measurements are available from Bernschütz (2013). Both transfer functions were smoothed with a gammatone filterbank spaced at equivalent rectangular bandwidths (ERBs), to approximate the frequency-resolution of the auditory system. From this, mean magnitude errors can be viewed across direction and frequency, shown in figures B.12b and B.12d respectively.

Figure B.12b shows the quadrature-weighted-mean of the log-magntiude error between the HRTF and the virtual-VBAP transfer functions at each auditory filter frequency, averaged across all directions. The grey region shows one standard deviation above and below the



(a) Point source ($t = 1.458$ ms)

(b) Point source ($t = 2.915$ ms)

(c) Stereo panning ($t = 1.458$ ms)

(d) Stereo panning ($t = 2.915$ ms)

(e) Basic decoding ($t = 1.458$ ms)

(f) Basic decoding ($t = 2.915$ ms)

(g) Max $\mathbf{r}_E$ decoding ($t = 1.458$ ms)

(h) Max $\mathbf{r}_E$ decoding ($t = 2.915$ ms)

Figure B.11: Soundfield analysis of an impulse from a broadband point source at position (0°, 0°, 1 m), and its approximation by stereo panning and by 5th-order ambisonics rendering with a 70-loudspeaker t-design with distance $r_l = 1$ m. Pressure fields are presented on a linear scale, normalised to the maximum value, and colour is clipped to [-0.25 0.25] for better visibility.

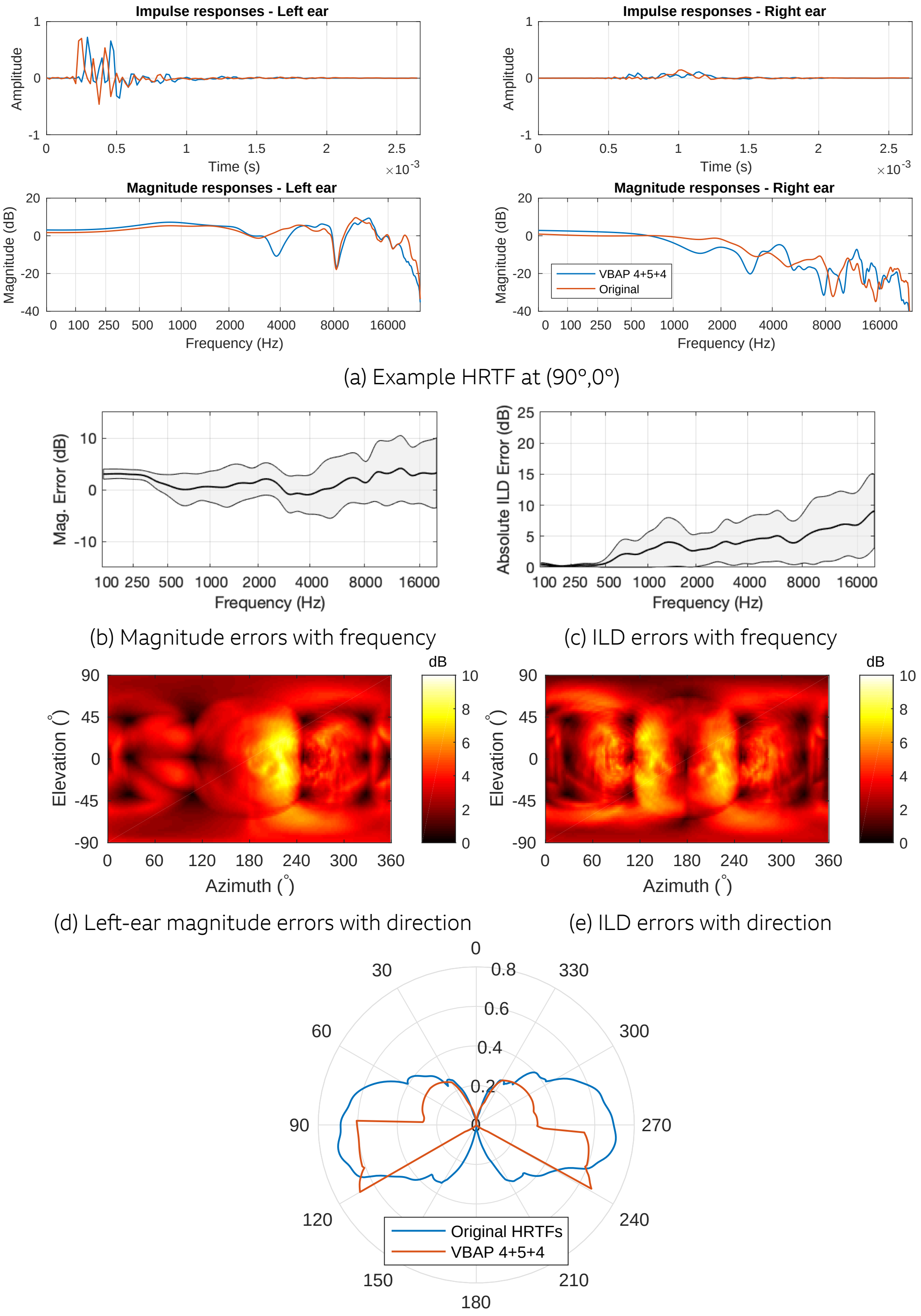
mean, to indicate the directional-variance at different frequencies. The variance increases at higher frequencies, which is expected since the HRTFs show more variance in that region anyway. Also notable is the large error at low frequencies, which is due to coherent summation. VBAP is a constant-power panning law, not constant-amplitude.

Figure B.12d shows the mean absolute log-magntiude error for each direction, averaged across the auditory frequency scale. When the source direction approaches that of a loudspeaker, the error tends to zero. When panning to the position of a loudspeaker, VBAP uses only this loudspeaker and so the transfer function is identical to the HRTF. The largest errors can be seen in the rear region, where the loudspeakers are sparse, but particularly on the contralateral side. The ILDs can also be estimated in each auditory frequency band from the filterbank analysis, as shown in figures B.12c and B.12e.

The broadband ITDs were estimated from the head-related impulse responses (HRIRs) using the maximum of the minimum-phase cross-correlation after a 3 kHz low-pass filter, as described in Katz and Noisternig, 2014. The log-threshold method with a 3 kHz low-pass filter was found the most perceptually-accurate method by Andreopoulou and Katz (2017), in terms of rendering with minimum-phase transfer functions. However this gave erratic results for this data. Figure B.12f shows the estimated ITDs for the HRIR measurements and the virtual VBAP impulse responses for source directions in the horizontal plane. Whilst the ITDs are approximated quite well in the frontal region, there are large errors at lateral and rear directions, except near to the rear-surround loudspeakers at ±110°.

Figure B.13 shows the same analysis for 3rd-order ambisonics with a 24-loudspeaker spherical t-design and the basic modal source strength matching decoder i.e. single-band decoding without modal-weighting. This system is more accurate up to about 1.5 kHz and results in much better approximation of the ITDs, though they are underestimated at lateral directions. The spectral errors show smoother variation with direction and are greater on the contralateral side. The ILD errors are greater in the high-frequency range. Neither virtualisation approach gives accurate simulation of the high-frequency spectral cues.

Beyond direct-inspection of the auditory cues, auditory models can be applied to estimate the resulting perceptual effects. Such models are reviewed in section 3.3.6. As an example, the sagittal-plane localisation model of Baumgartner and Majdak (2015) was applied to the median-plane HRTFs generated by virtual loudspeaker rendering, using the measured KU100 HRTFs as the template of the model listener. This model was developed to evaluate localisation of amplitude-panned virtual sources in the mid-sagittal or median plane. Here the modelling analysis was performed in the range 700 Hz–18 kHz with a sensitivity value $S = 0.5$, indicating the model listener's ability to detect spectral similarity between the input and the template. The model gives a discrete probability distribution of the polar



(a) Example HRTF at (90°,0°)

(b) Magnitude errors with frequency

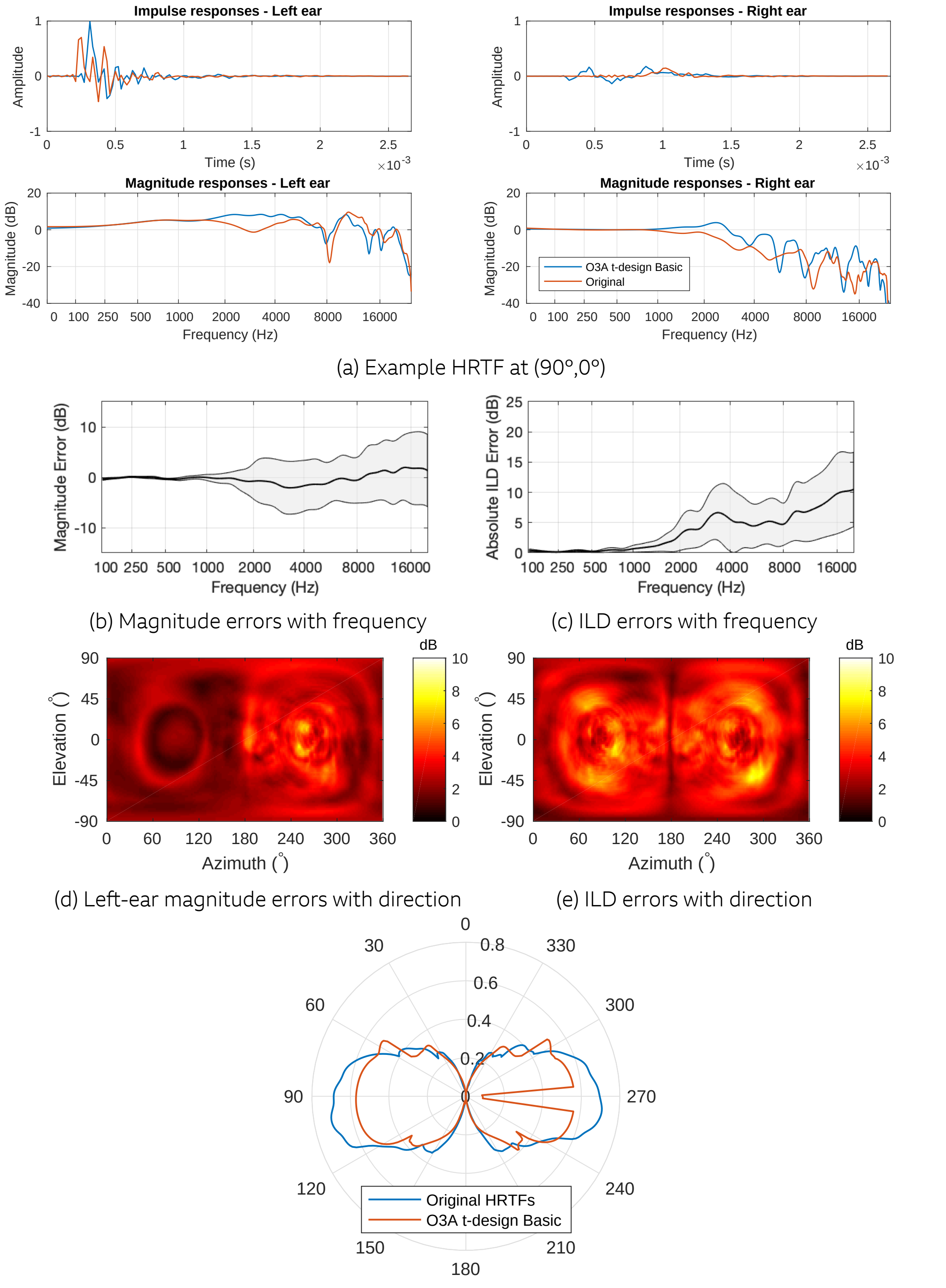
(c) ILD errors with frequency

(d) Left-ear magnitude errors with direction

(e) ILD errors with direction

(f) Horizontal-plane ITDs

Figure B.12: Auditory cue analysis of loudspeaker virtualisation using VBAP with a 4+5+4 layout.



(a) Example HRTF at (90°,0°)

(b) Magnitude errors with frequency

(c) ILD errors with frequency

(d) Left-ear magnitude errors with direction

(e) ILD errors with direction

(f) Horizontal-plane ITDs

Figure B.13: Auditory cue analysis of loudspeaker virtualisation using 3rd-order ambisonics with a 24-loudspeaker spherical *t*-design and a modal source strength matching decoder.

response angle for each target angle, a discretised resolution of 2° ϕ_{cc} was used. Figure B.14 shows the results of this analysis.

For VBAP rendering (figures B.14b and B.14d), a broad notch is introduced in the frontal region by comb-filtering between the loudspeakers of the horizontal and upper/lower layers. These notches vary in frequency from 2 kHz to 4 kHz with source ϕ_{cc}. There is also a broad region of large high-frequency errors (5 kHz–15 kHz) to the rear where loudspeaker coverage is sparse, with particularly large errors where notches exist in the target HRTF magnitude responses. The localisation model indicates that localisation blur in the polar angle is likely, including front-back reversals within the cone-of-confusion.

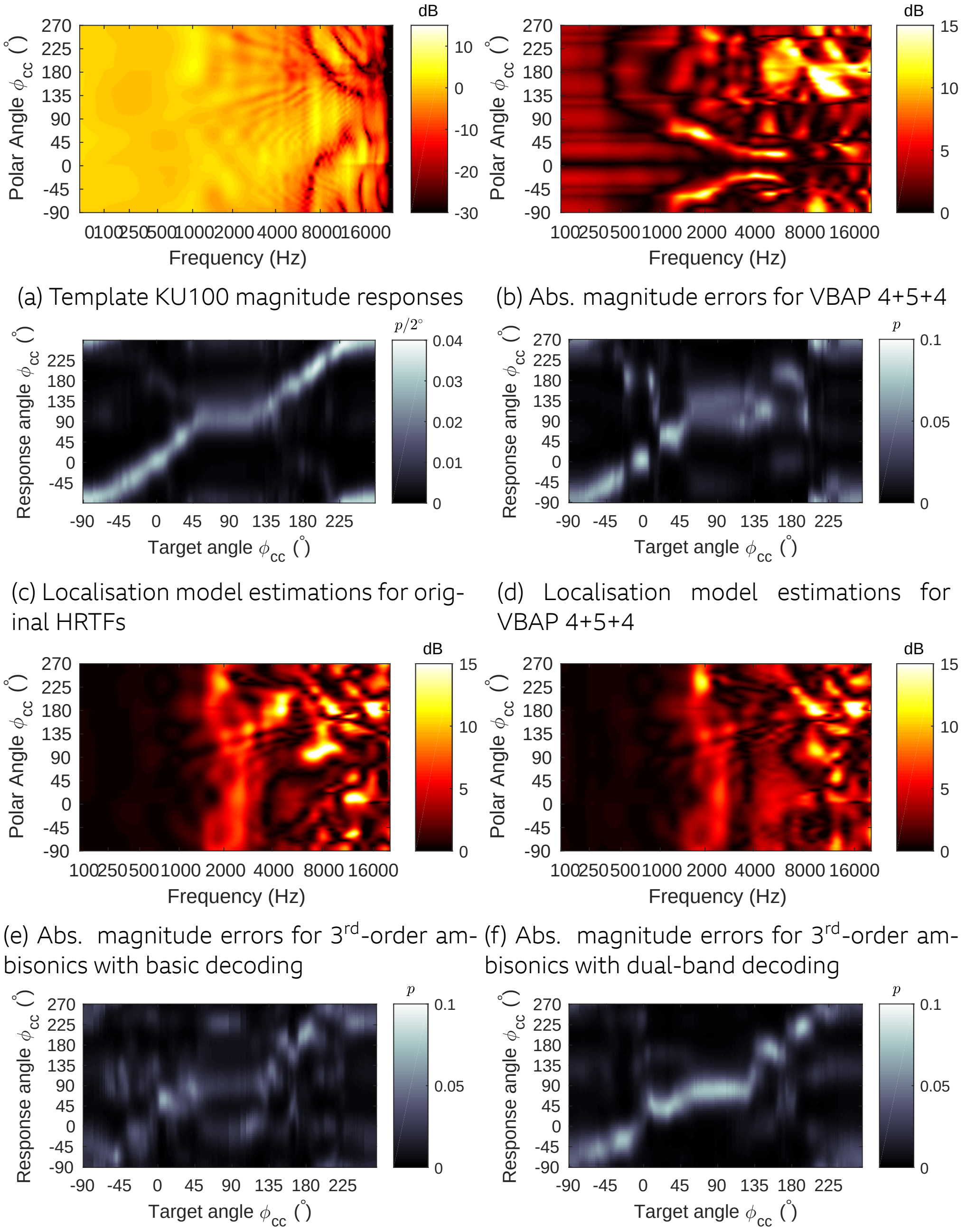
With 3$^{\text{rd}}$-order ambisonics decoding using the basic modal source strength matching decoder (figures B.14e and B.14g), there are large spectral errors above 2 kHz. There is a largely direction-independent notch in the ambisonics-to-binaural HRTFs at 2 kHz, and above 4 kHz there are spectral notch patterns that do not correspond to those of the target HRTFs and do not vary smoothly with source angle. The localisation model shows highly blurred results, with erratic polar angle response probability distribution, which is not limited to front-back reversals. It is worth noting that the spectral errors at low frequencies are lower for the ambisonics renderer than for the VBAP system. When a dual-band decoding approach is used, with max $\mathbf{r}_E$ modal weighting above 2.8 kHz, the errors below 7 kHz are reduced (figure B.14e). This leads to much improved performance from the localisation model (figure B.14g).

B.5.4 Summary

Techniques have been presented for the analysis of loudspeaker rendering algorithms. These provide tools for improving the design of such algorithms, including their use in virtualisation applications for headphone rendering. These techniques can also be used to gain insight into the causes of perceptual effects that might be observed in listening experiments. Some example analyses have been presented using both VBAP and ambisonics-based panning.

The Gerzon vector analysis provides simple measures for analysing panning algorithms, giving an estimate of low and high frequency localisation, as well as loudness and perceived source extent. Sound field analysis techniques allow reproduction across the listening area to be evaluated in both the time and frequency domains. This can be useful for visual analysis of effects like spatial aliasing and pre-echoes, which may cause perceptual artefacts. However, these techniques use idealised models of the reproduction process.

Simulation of the binaural signals generated at the listening position allows evaluation



(a) Template KU100 magnitude responses

(b) Abs. magnitude errors for VBAP 4+5+4

(c) Localisation model estimations for original HRTFs

(d) Localisation model estimations for VBAP 4+5+4

(e) Abs. magnitude errors for 3rd-order ambisonics with basic decoding

(f) Abs. magnitude errors for 3rd-order ambisonics with dual-band decoding

(g) Localisation model estimations for 3rd-order ambisonics with basic decoding

(h) Localisation model estimations for 3rd-order ambisonics with dual-band decoding

Figure B.14: Analysis of median-plane localisation afforded by virtual loudspeaker rendering techniques, using Baumgartner and Majdak (2015) model.

in terms of the reproduction of auditory cues, from which the perceptual effects can be more directly interpreted. This approach is closely related to loudspeaker virtualisation in binaural rendering systems, so it is particularly useful when designing and evaluating such applications. However, auditory analysis must be based on individualised binaural signals to capture individual variations. When considering non-individualised virtual loudspeaker systems, the target HRTF in the auditory model should be the individual's, whilst the rendering system simulation should use the non-individual HRTF.

An extension of such analysis would be to drive auditory models that predict perceptual results from the binaural input signals, often calibrated with data from perceptual experiments. Auditory modelling has made significant progress, yet there are still open questions and competing hypotheses under debate, which can lead to conflicting results (Dietz, Lestang, et al., 2018). Auditory models are available for relatively low-level percepts such as colouration, localisation and externalisation, but higher-level experiential quality features cannot be modelled yet. At this stage of cognitive processing, many influencing factors are involved and such models would be highly complex. Even with these tools, perceptual evaluation is vital for understanding the effects of virtual loudspeaker techniques, and the effects of binaural rendering techniques more generally.

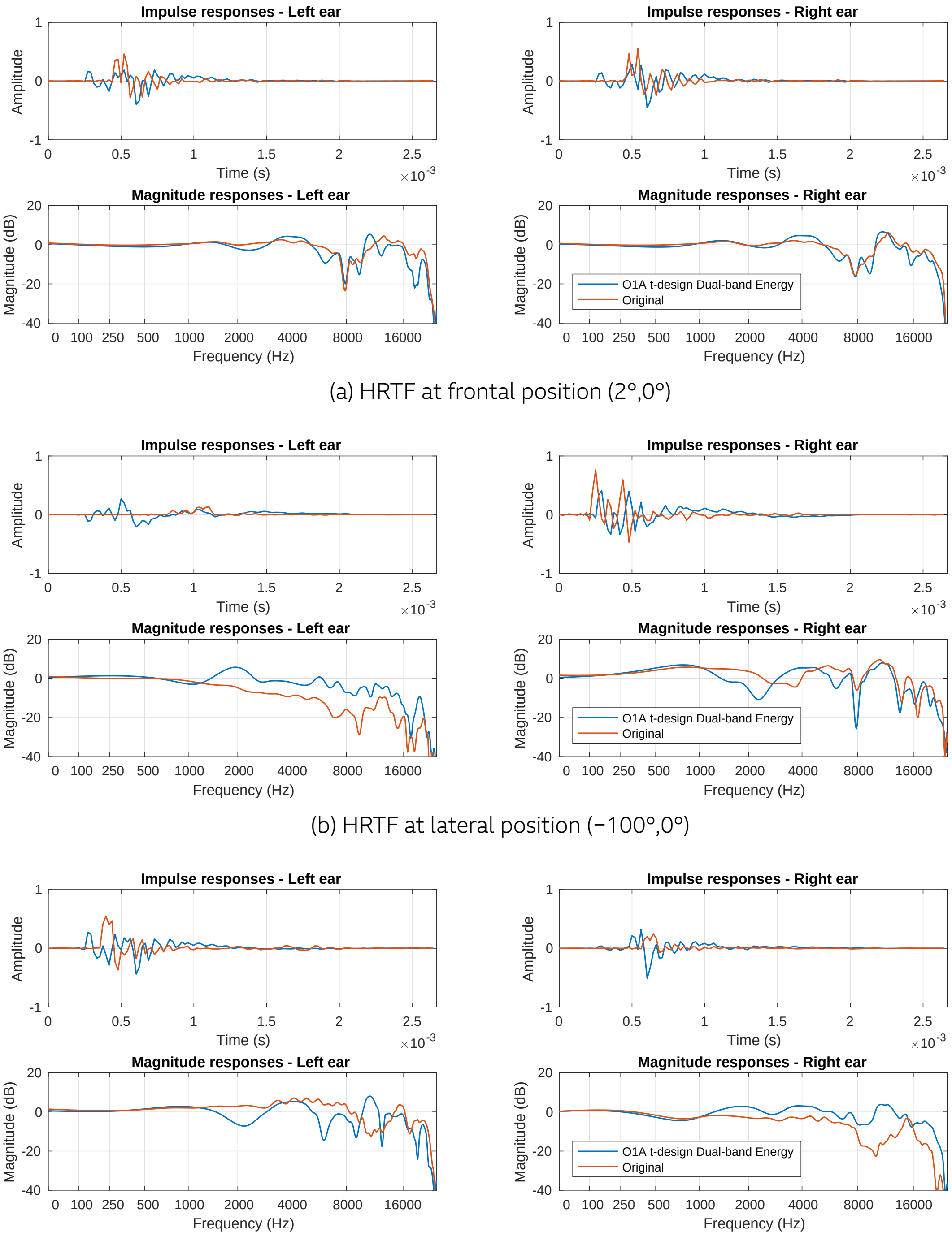
B.6 Analysis of Virtualisation Systems Used in CATA Experiment

This section presents figures resulting from auditory modelling analysis of the loudspeaker virtualisation rendering systems used in the experiment of chapter 8. For each system, the approximation of the measured HRTFs is shown at each of the three target directions used for the single source stimuli in the experiment. As in appendix B.5.3, analysis of the rendering was carried out for all directions in the 16 020-point Gauss-Legendre quadrature and used to visualise the approximation of spectral cues, ILDs and ITDs in the original HRTFs by each virtualisation approach.

These systems for binaural rendering via an intermediate loudspeaker virtualisation process are described in section 8.3.3.

- Figures B.15 and B.16 show analysis of the *A1* system, which uses 1$^{\text{st}}$-order ambisonics with an 8-loudspeaker spherical t-design and a dual-band decoder with crossover frequency $f_x = 748$ Hz.
- Figures B.17 and B.18 show analysis of the *A3* system, which uses 3$^{\text{rd}}$-order ambisonics with an 24-loudspeaker spherical t-design and a dual-band decoder with $f_x = 2805$ Hz.

- Figures B.19 and B.20 show analysis of the *A5* system, which uses 5$^{\text{th}}$-order ambisonics with an 70-loudspeaker spherical t-design and a dual-band decoder with $f_x = 5701$ Hz.
- Figures B.21 and B.22 show analysis of the *V1* system, which uses VBAP with a 0+8+0 loudspeaker layout and re-distributed dummy loudspeakers.
- Figures B.23 and B.24 show analysis of the *V3* system, which uses VBAP with a 4+8+4 loudspeaker layout and re-distributed dummy loudspeakers.
- Figures B.25 and B.26 show analysis of the *V5* system, which uses VBAP with a 9+12+9 loudspeaker layout.

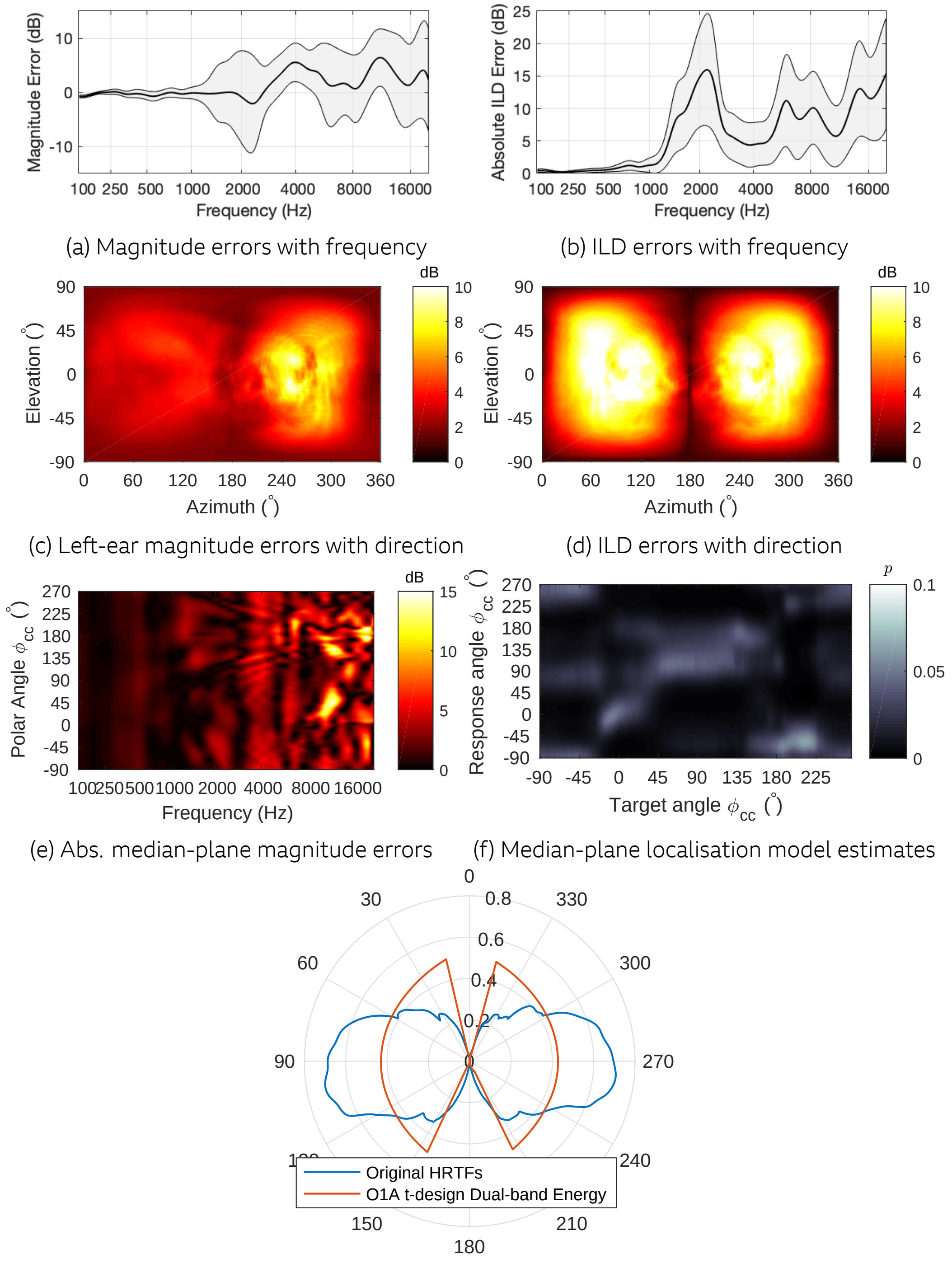


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.15: Approximation of HRTFs by loudspeaker virtualisation using *A1* system from chapter 8 experiment – 1$^{\text{st}}$-order ambisonics with an 8-loudspeaker spherical t-design and a dual-band decoder where $f_x = 748\,\text{Hz}$.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

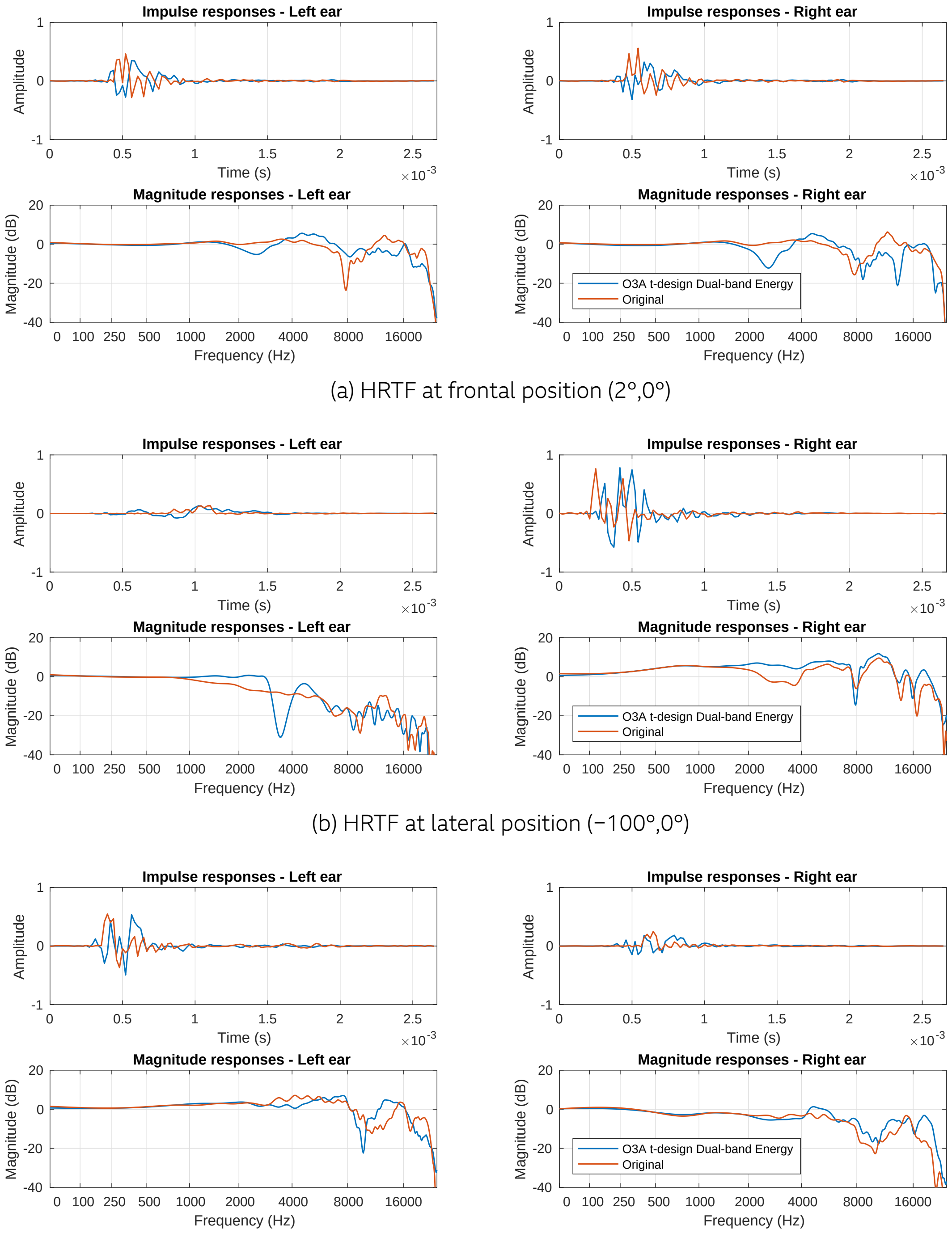
(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.16: Auditory cue analysis of loudspeaker virtualisation using *A1* system from chapter 8 experiment – 1st-order ambisonics with an 8-loudspeaker spherical t-design and a dual-band decoder where $f_x = 748$ Hz.

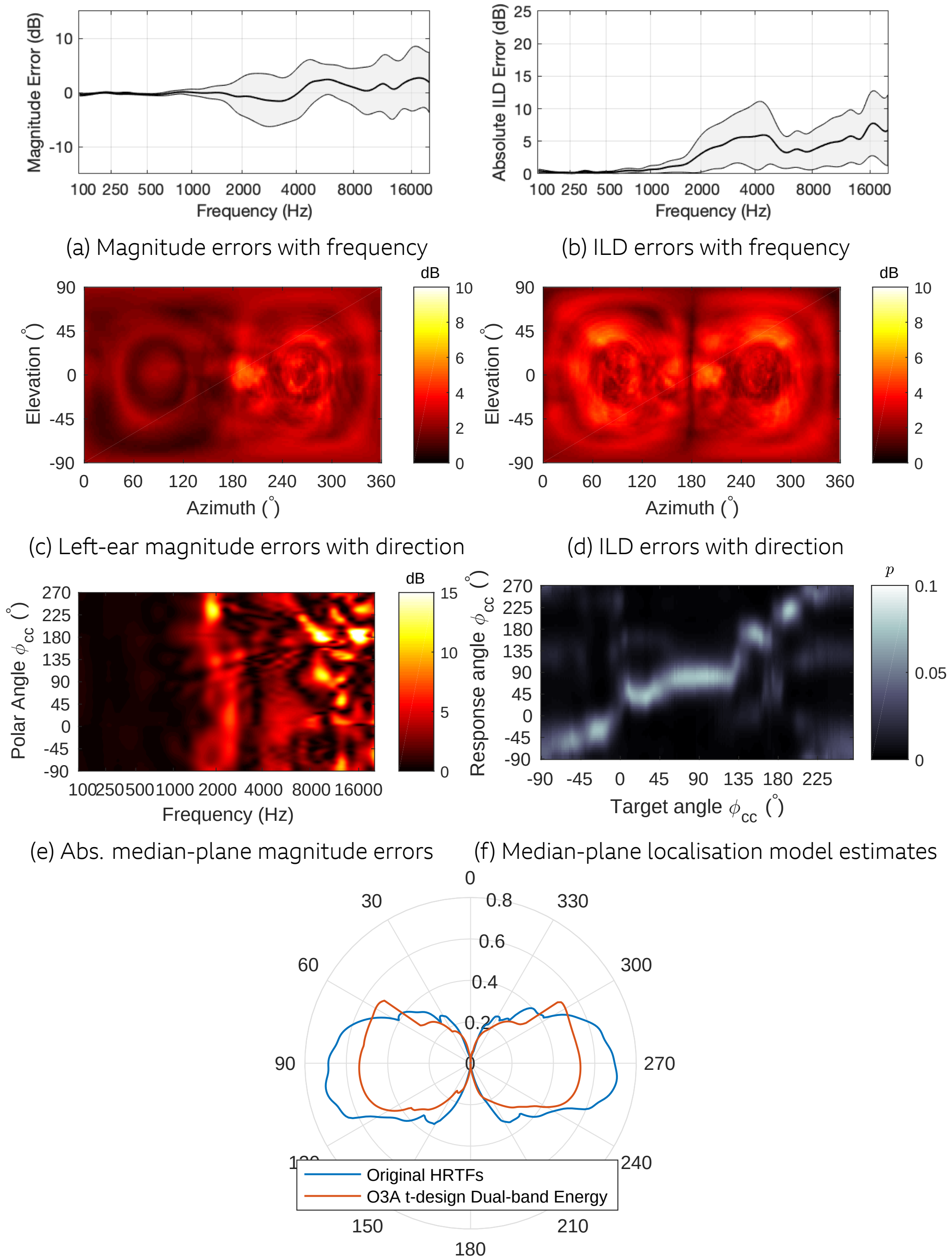


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.17: Approximation of HRTFs by loudspeaker virtualisation using *A3* system from chapter 8 experiment – 3$^{\text{rd}}$-order ambisonics with an 24-loudspeaker spherical t-design and a dual-band decoder where $f_x = 2805$ Hz.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

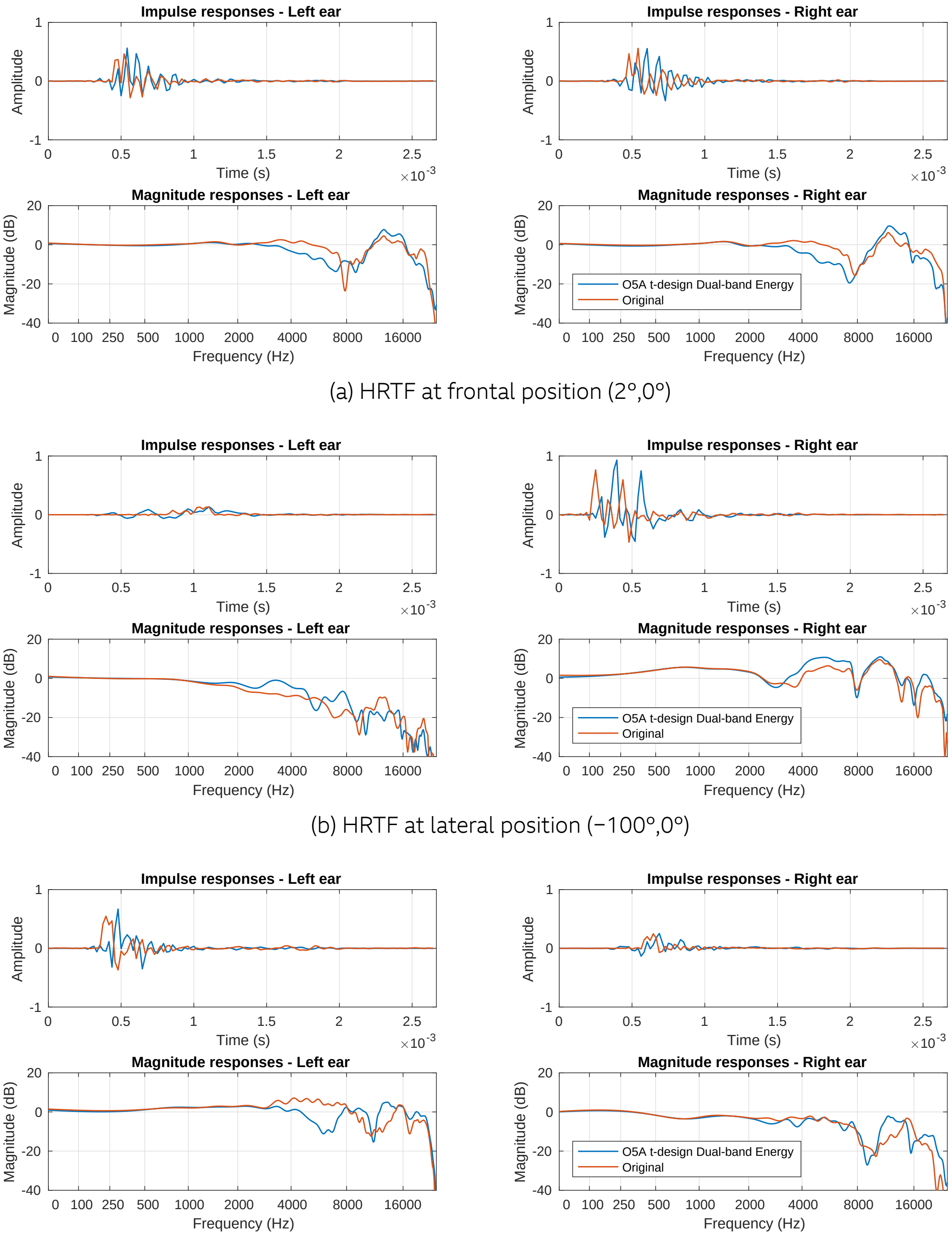
(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.18: Auditory cue analysis of loudspeaker virtualisation using *A3* system from chapter 8 experiment – 3$^{\text{rd}}$-order ambisonics with an 24-loudspeaker spherical t-design and a dual-band decoder where $f_x = 2805$ Hz.

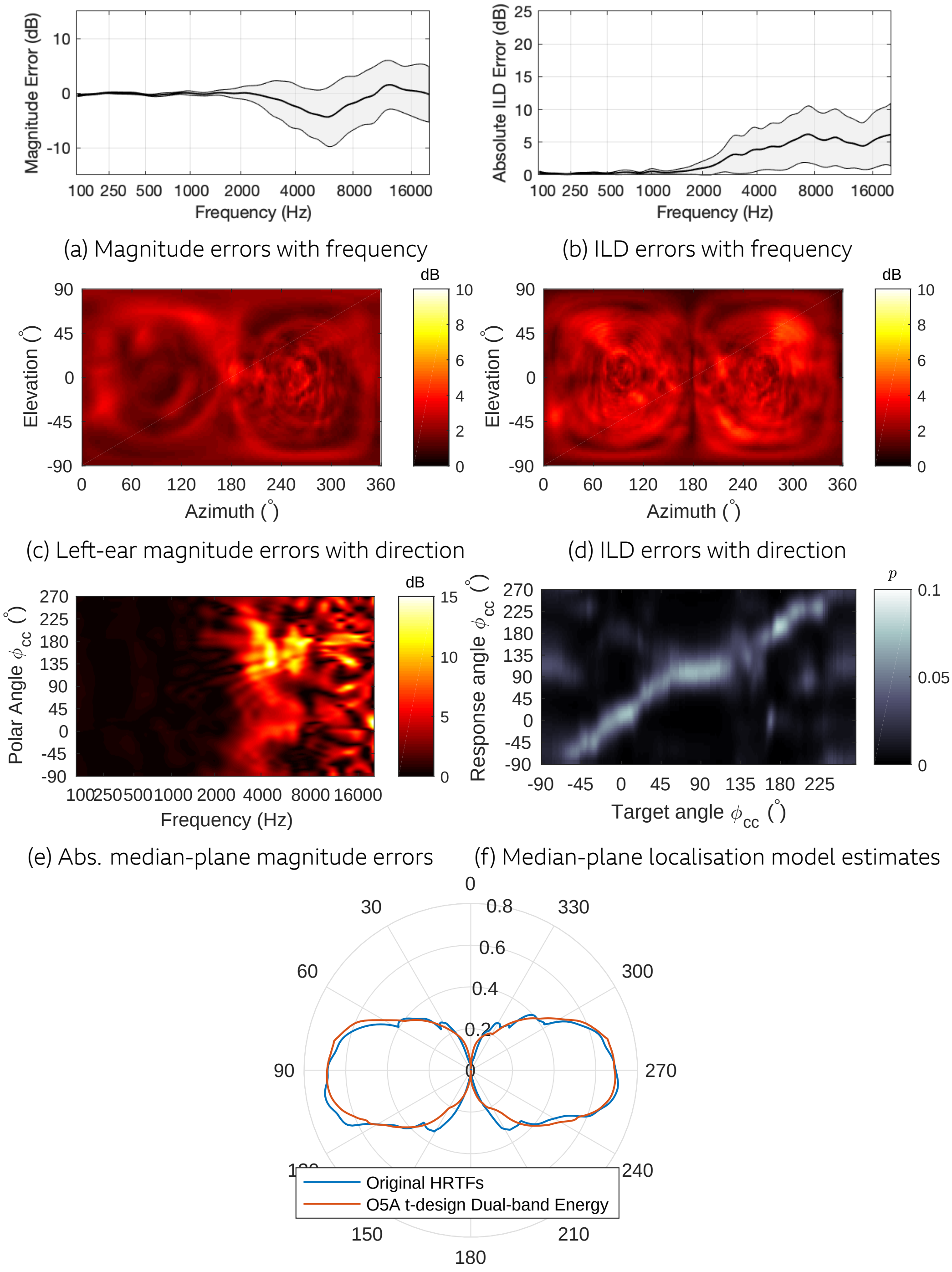


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.19: Approximation of HRTFs by loudspeaker virtualisation using *A5* system from chapter 8 experiment – 5th-order ambisonics with an 70-loudspeaker spherical t-design and a dual-band decoder where $f_x = 5701$ Hz.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

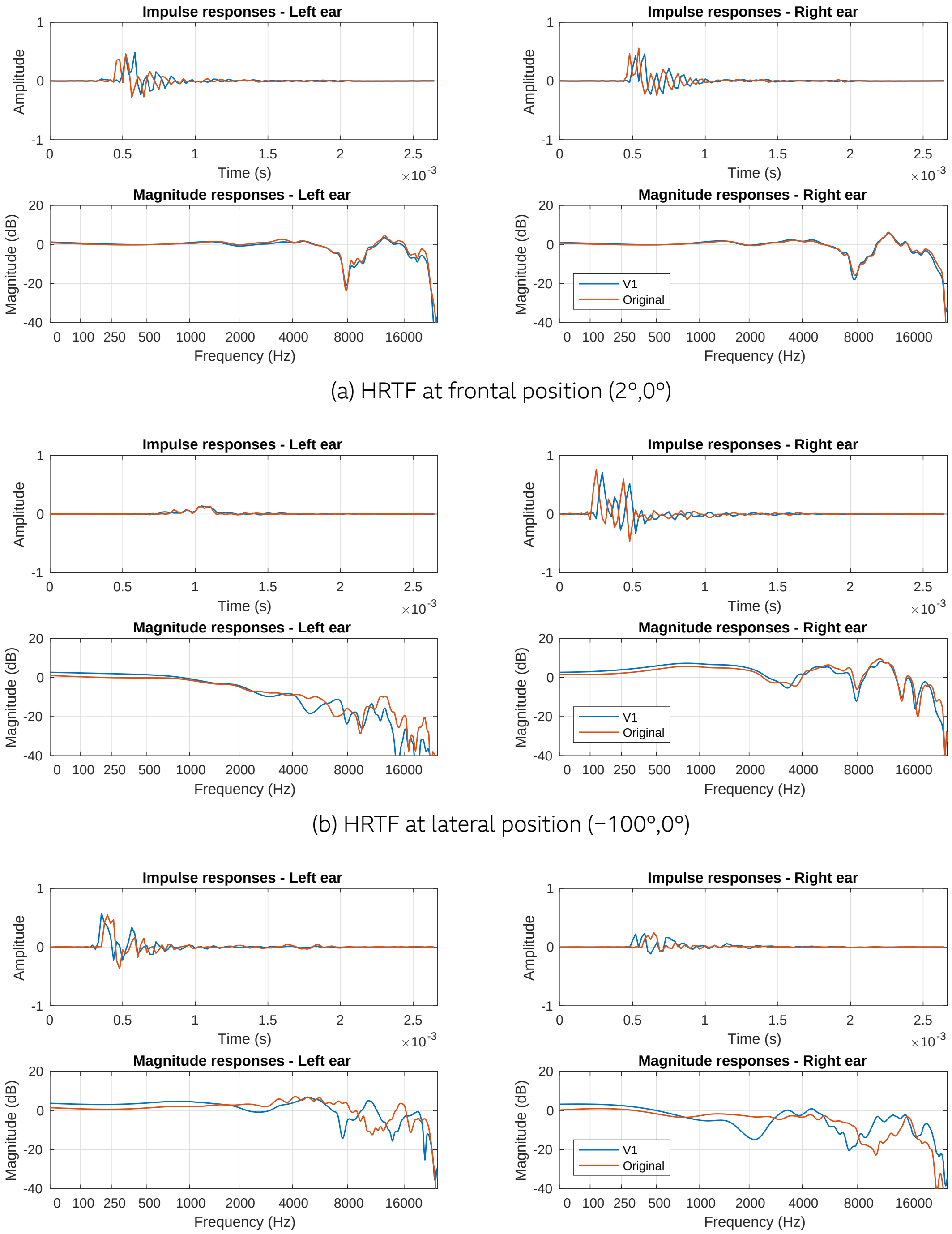
(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.20: Auditory cue analysis of loudspeaker virtualisation using *A5* system from chapter 8 experiment – 5th-order ambisonics with an 70-loudspeaker spherical t-design and a dual-band decoder where $f_x = 5701$ Hz.

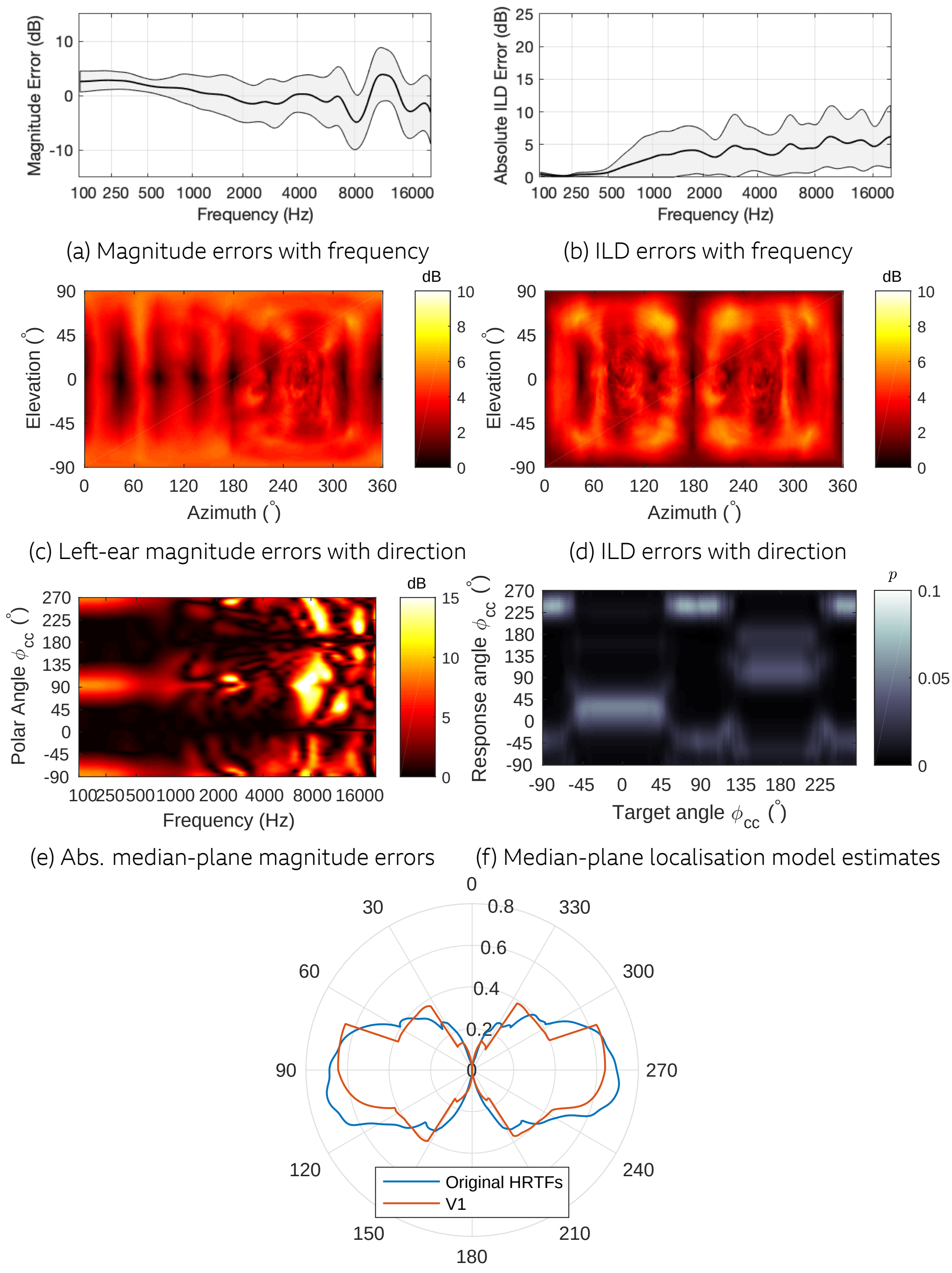


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.21: Approximation of HRTFs by loudspeaker virtualisation using *V1* system from chapter 8 experiment – based on VBAP with a 0+8+0 loudspeaker layout and re-distributed dummy loudspeakers.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

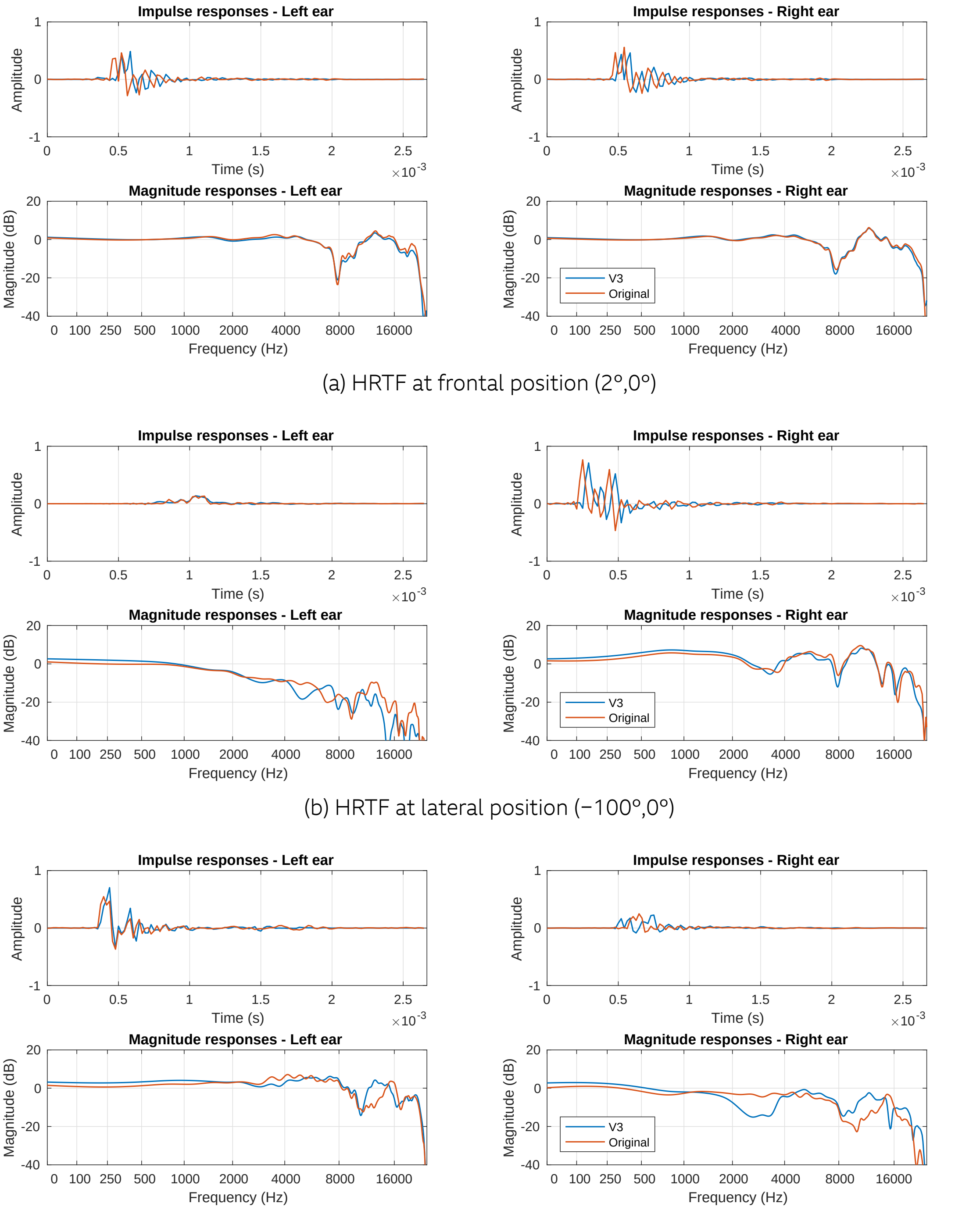
(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.22: Auditory cue analysis of loudspeaker virtualisation using *V1* system from chapter 8 experiment – based on VBAP with a 0+8+0 loudspeaker layout and re-distributed dummy loudspeakers.

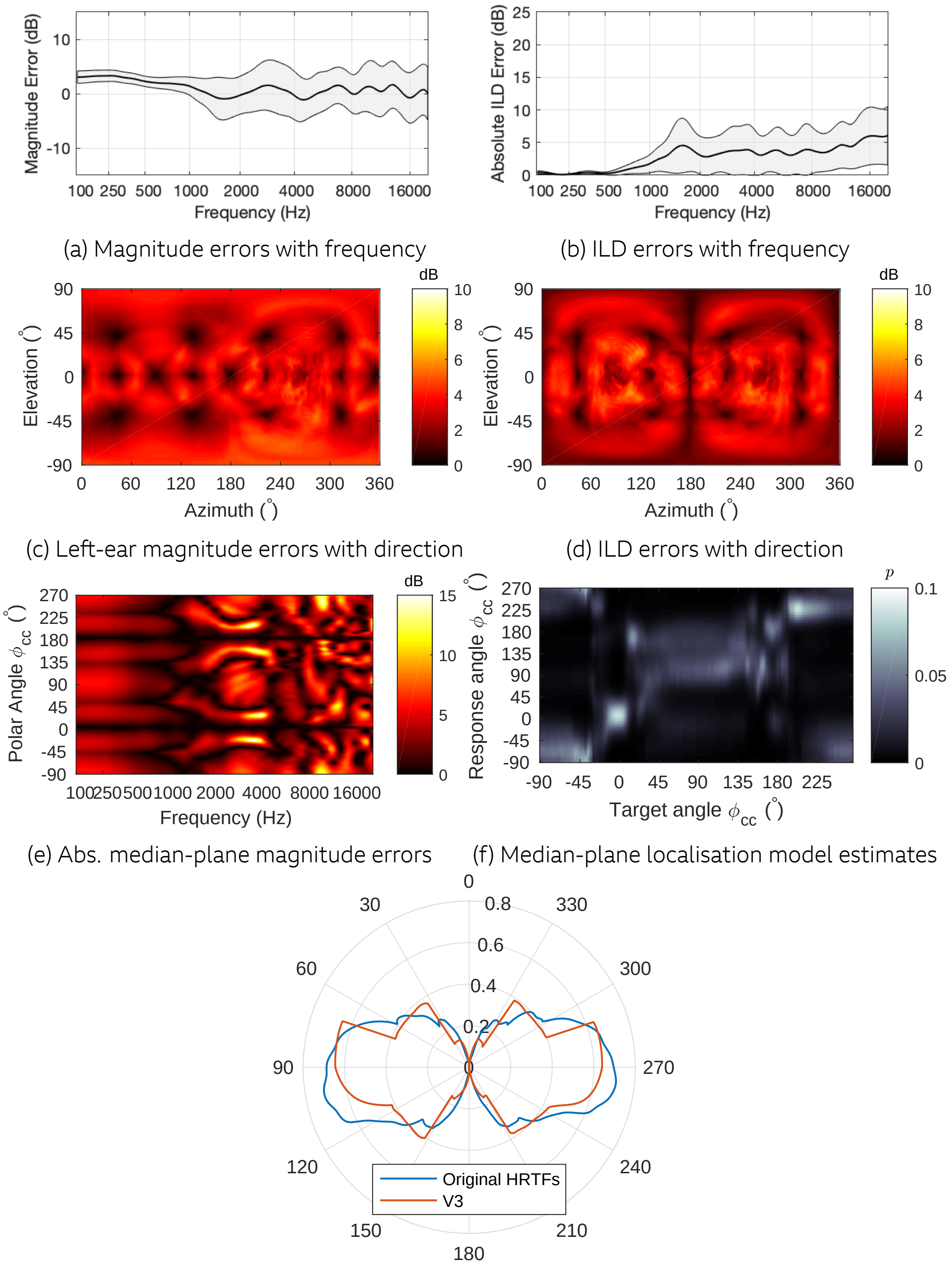


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.23: Approximation of HRTFs by loudspeaker virtualisation using *V3* system from chapter 8 experiment – based on VBAP with a 4+8+4 loudspeaker layout and re-distributed dummy loudspeakers.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

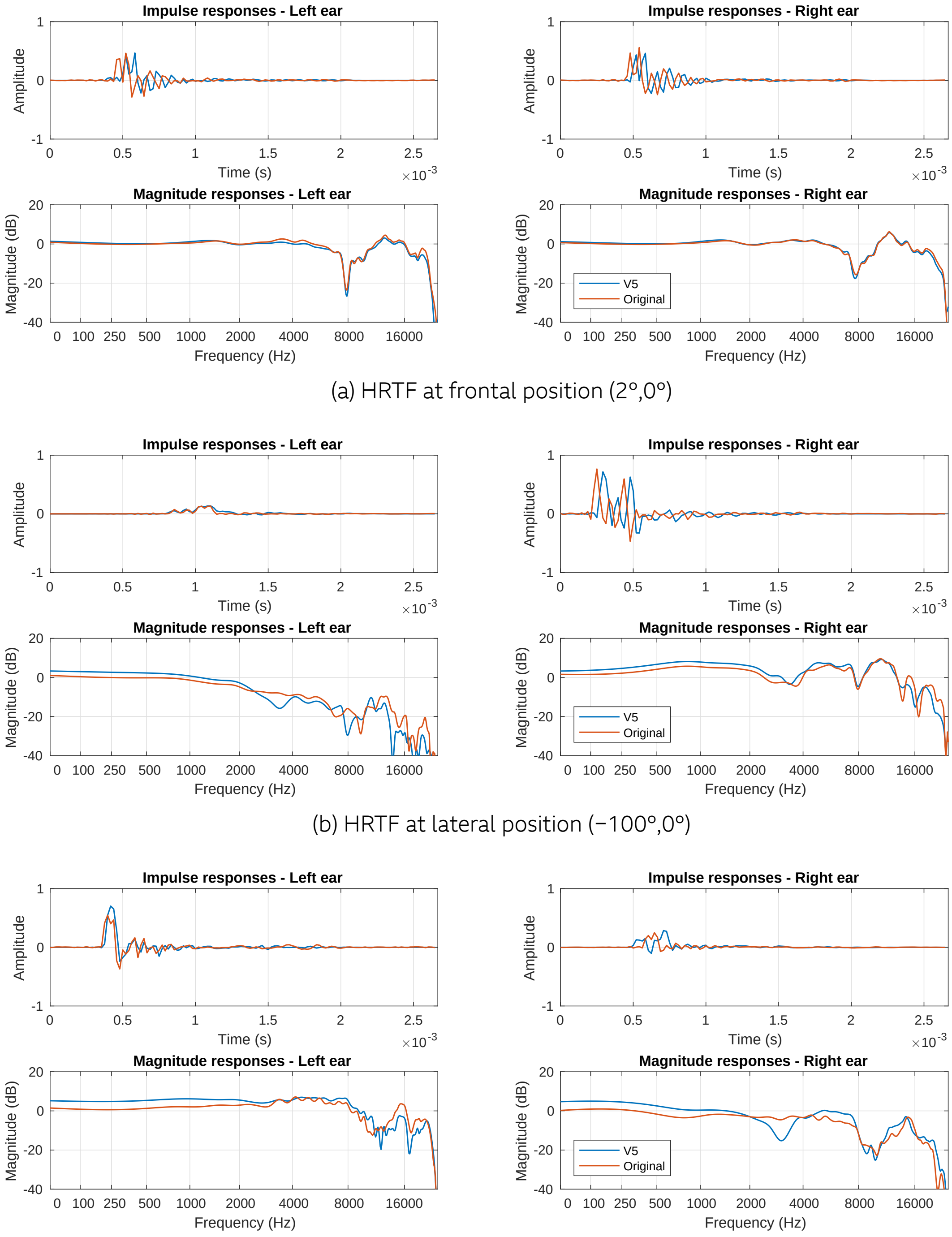
(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.24: Auditory cue analysis of loudspeaker virtualisation using *V3* system from chapter 8 experiment – based on VBAP with a 4+8+4 loudspeaker layout and re-distributed dummy loudspeakers.

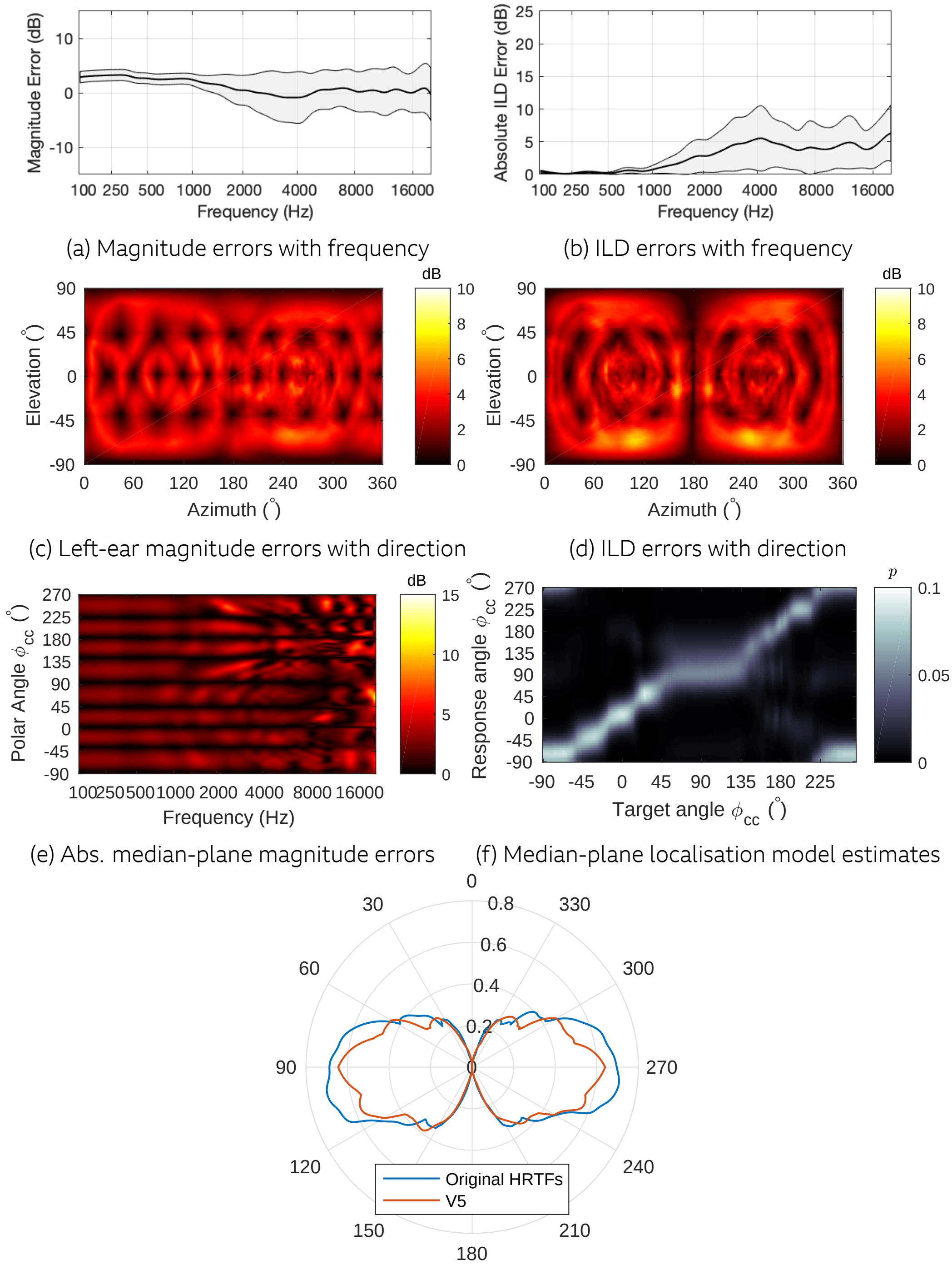


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.25: Approximation of HRTFs by loudspeaker virtualisation using *V5* system from chapter 8 experiment – based on VBAP with a 9+12+9 loudspeaker layout.



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.26: Auditory cue analysis of loudspeaker virtualisation using *V5* system from chapter 8 experiment – based on VBAP with a 9+12+9 loudspeaker layout.

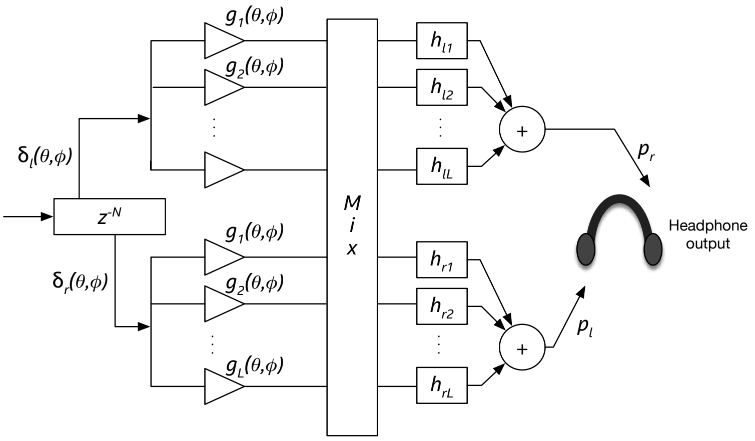


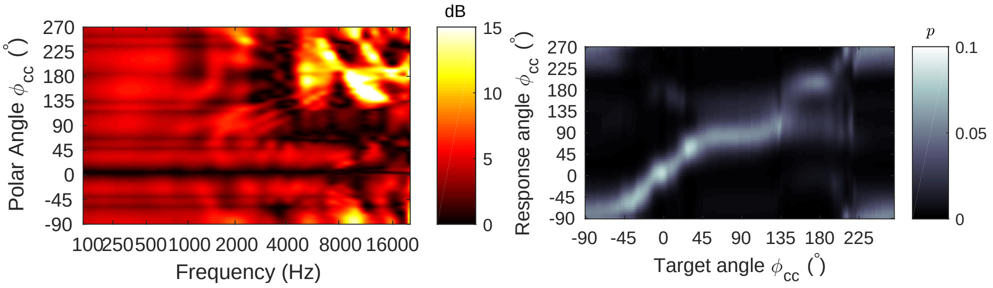
Figure B.27: Separate ITD processing during virtualisation, after Jot, Walsh, et al. (2006).

B.7 Advanced Virtualisation Techniques

Whilst the basic loudspeaker virtualisation technique presented in section 6.4 is widely used, along with VBAP and ambisonics rendering, more advanced techniques are available in the literature. For completeness they are reviewed here. Future work might compare these to the more basic techniques to establish whether they deliver benefits in terms of perceived quality.

The problem of inadequate ITD synthesis through virtualised loudspeaker rendering has long been acknowledged. This is not an issue for sources panned at loudspeaker positions with sparse panning techniques such as VBAP, since the HRTFs are precisely reconstructed in this case. Similarly, for channel-based input signals the inter-channel time differences (ICTDs), e.g. from spaced microphone recordings, are preserved. However, when amplitude panning or low-order ambisonics are used, the ITDs for virtual sources are incorrect. So, for example, if a channel-based input signal is based on a loudspeaker layout that does not correspond to the virtual loudspeaker positions in the renderer, amplitude panning might be used to place each channel at the correct location and then the ICTDs will not be correctly reproduced. With ambisonics-based virtualisation, this issue occurs irrespective of the alignment with virtual loudspeakers, due to the non-sparsity of the panning technique.

Jot, Larcher, et al. (1995) presented a technique for separate rendering of the ITDs during virtualisation. The virtualisation is performed with minimum-phase HRTFs and sources are panned into two separate virtual busses (channel-based or scene-based), for left and right ears, with the appropriate delays applied. This process is visualised in figure B.27. With



(a) Absolute magnitude errors

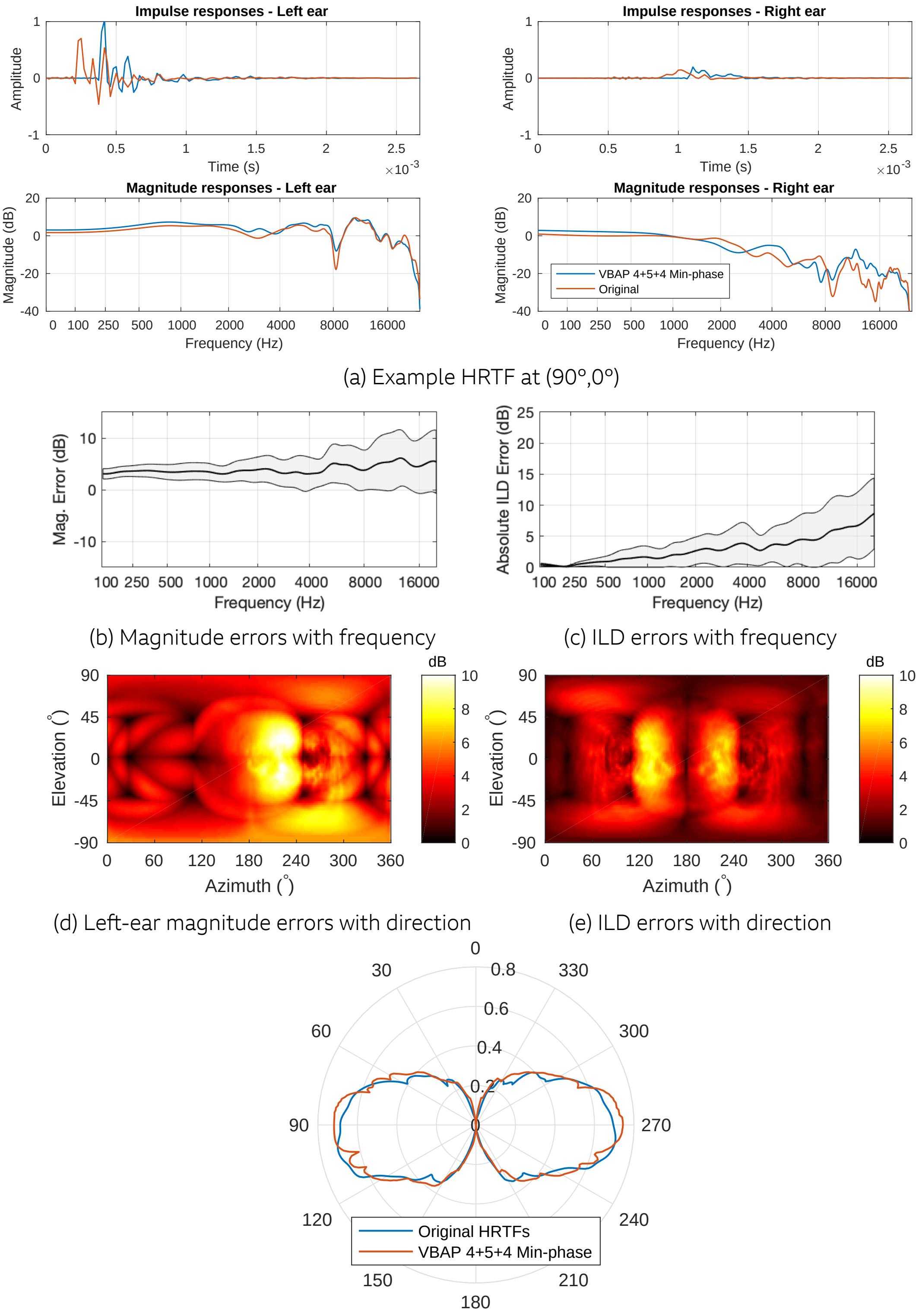
(b) Baumgartner and Majdak localisation model estimations

Figure B.28: Analysis of median-plane localisation afforded by VBAP virtual loudspeaker rendering using minimum-phase HRTFs and separate ITD insertion.

this method ITDs can be reconstructed more accurately. Comb-filtering effects during amplitude panning will be reduced also, but whether this leads to more accurate reconstruction of spectral magnitude cues and ILDs requires investigation.

Figure B.28 shows the results of the localisation model when applying this approach to VBAP rendering on the 4+5+4 layout. The notches in the range 2 kHz–4 kHz are avoided and this leads to improved elevation localisation in auditory modelling. Analysis of auditory cues across the full sphere is shown in figure B.29, for comparison to figure B.12. The intermediate format has twice as many channels, however, because a separate representation is needed for each ear. Jot, Larcher, et al. defined the “binaural B-format” where this approach was used with first-order ambisonics. It included recording with two SoundField microphones spaced with approximately inter-aural distance. This approach also prevents distribution of a pre-rendered intermediate format if head tracking is to be used at the receiver, since the ITDs are dependent on head orientation. It can be used when rendering an object-based input on a receiver device, though, since the ITD synthesis may be updated in real-time according to head orientation in this case. This approach is taken in (Jot, Walsh, et al., 2006).

Zaunschirm, Schörkhuber, et al. (2018) used a frequency-dependent time-alignment approach for designing ambisonics-to-binaural filters, taking into account that interaural phase relationships are not important above approximately 1.8 kHz. An optimisation-based approach is then used to derive ambisonics-to-binaural filters directly from the full-sphere HRTF set. This avoids restrictions introduced by selecting a specific sub-set of HRTF as virtual loudspeakers and using a single- or dual-band decoding approach. Instead the method directly minimises the errors in the diffuse-field HRTFs. Filters for rendering by this method are made available in the IEM Plug-in Suite (Rudrich, 2018) and were analysed using the



(a) Example HRTF at (90°,0°)

(b) Magnitude errors with frequency

(c) ILD errors with frequency

(d) Left-ear magnitude errors with direction

(e) ILD errors with direction

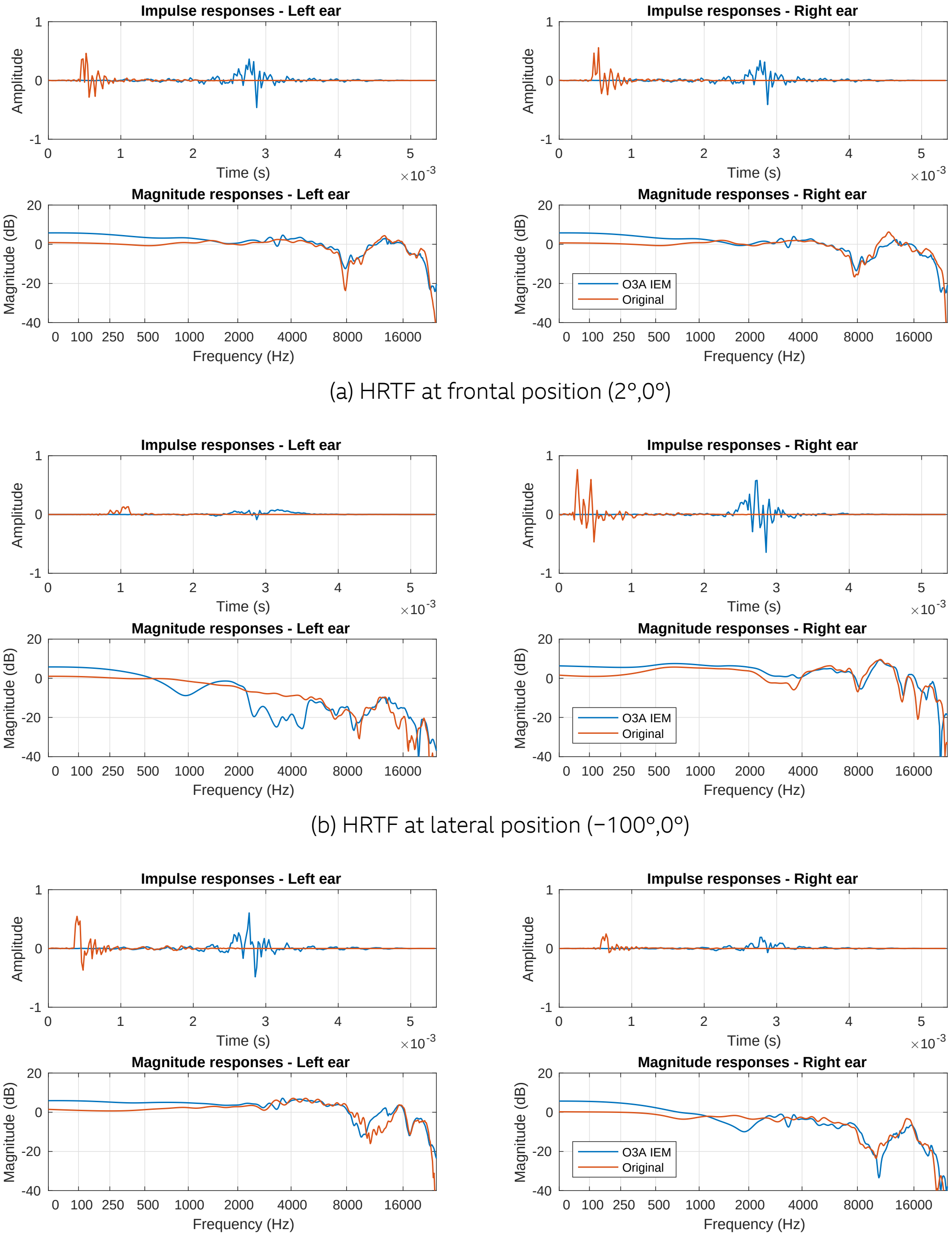
(f) Horizontal-plane ITDs

Figure B.29: Auditory cue analysis of loudspeaker virtualisation using VBAP virtual loudspeaker rendering using minimum-phase HRTFs and separate ITD insertion.

same techniques for other methods. Figures B.30 and B.31 show the analysis for 3rd order ambisonics rendering and can be compared with the approach used in chapter 8 shown in figures B.17 and B.18. This method appears to achieve superior performance at high frequencies, particularly with regard to spectral magnitude errors, but the low-frequency response is poor and leads to inaccurate reconstruction of ITDs. It appears that there may be a case for a dual-band approach, applying this time-alignment and diffuse-field response optimisation approach at high frequencies but the conventional modal source strength matching decoding approach at low frequencies.

Phantom sources can be better rendered by inter-channel decorrelation according to Jot, Walsh, et al. (2006) and Jot and Noh (2017). Decorrelation filters are used in loudspeaker rendering systems to achieve extended and diffuse sound sources, e.g. (EBU Tech 3388, 2018), though these filters can also introduce colouration issues and smearing of the temporal response (Franck, Fazi, et al., 2015). Structures can be designed to make virtualisation more efficient in applications (Jot and Noh, 2017). Assuming a left-right symmetrical HRTF representation is a common way to increase efficiency in non-individual rendering systems.

Time-frequency parametric analysis-synthesis approaches have also been presented. For channel-based signals Faller and Baumgarte (2003), Baumgarte and Faller (2003), and Goodwin and Jot (2006, 2007) and for ambisonics signals Vilkamo, Lokki, et al. (2009), Laitinen and Pulkki (2009), Berge and Barrett (2010), Politis, McCormack, et al. (2017), and Politis, Tervo, et al. (2018). Perceptual evaluations of these parametric approaches show significant improvements over conventional virtual loudspeaker techniques. For example, Politis, Tervo, et al. (2018) recently showed that binaural rendering of first-order ambisonics by parametric analysis-synthesis could achieve equivalent quality to third-order ambisonics rendering using normal virtual loudspeaker techniques.

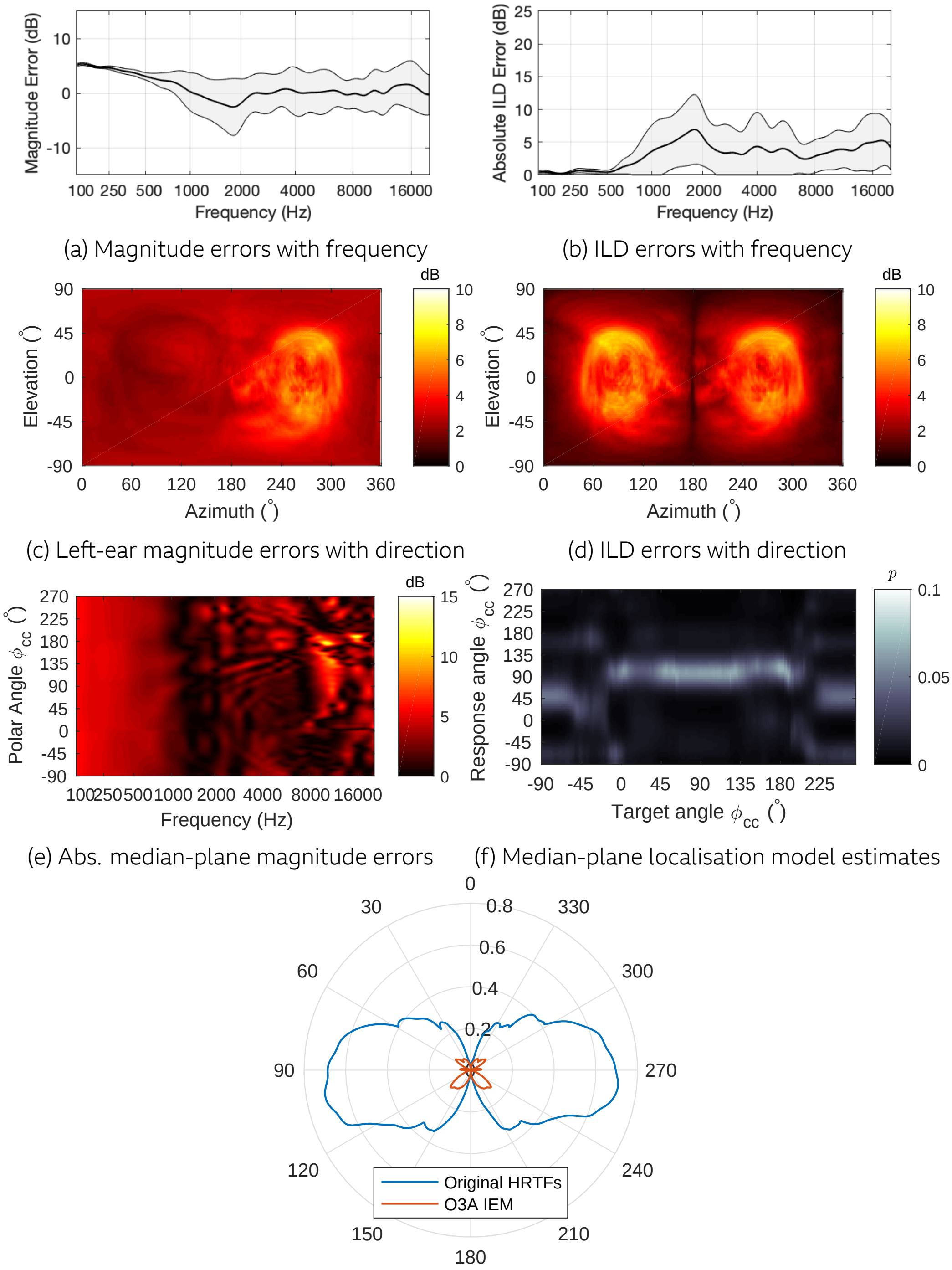


(a) HRTF at frontal position (2°,0°)

(b) HRTF at lateral position (−100°,0°)

(c) HRTF at lateral position (30°,30°)

Figure B.30: Approximation of HRTFs by loudspeaker virtualisation using 3rd order ambisonics-to-binaural rendering according to Zaunschirm, Schörkhuber, et al. (2018).



(a) Magnitude errors with frequency

(b) ILD errors with frequency

(c) Left-ear magnitude errors with direction

(d) ILD errors with direction

(e) Abs. median-plane magnitude errors

(f) Median-plane localisation model estimates

(g) Horizontal-plane ITDs

Figure B.31: Auditory cue analysis of loudspeaker virtualisation using 3rd order ambisonics-to-binaural rendering according to Zaunschirm, Schörkhuber, et al. (2018).

Appendix C

Attribute Definitions

The full list of attributes used in the experiment of chapter 8 are given in table C.1.

Table C.1: The 48 CATA attributes used in the listening experiment of chapter 8.

Attribute	Scale	Definition	Category
Dark	Tone color bright-dark	Timbral impression determined by the ratio of high to low frequency components. Dark means more low frequency (bass) and less high frequency (treble).	Timbral
Bright	Tone color bright-dark	Timbral impression determined by the ratio of high to low frequency components. Bright means more high frequency (treble) and less low frequency (bass).	Timbral
Treble - High	High-frequency tone colour	High energy in the high-frequency range (treble).	Timbral
Treble - Low	High-frequency tone colour	Low energy in the high-frequency range (treble).	Timbral
Mid-Frequency - High	Mid-frequency tone colour	High energy in the mid-frequency range.	Timbral
Mid-Frequency - Low	Mid-frequency tone colour	Low energy in the mid-frequency range.	Timbral
Bass - High	Low-frequency tone colour	High energy in the low-frequency range (bass).	Timbral
Bass - Low	Low-frequency tone colour	Low energy in the low-frequency range (bass).	Timbral

Continued on next page

Table C.1 – *Continued from previous page*

Attribute	Scale	Definition	Category
Sharp	Sharpness	Timbral impression which e.g., is indicative of the force with which a sound source is excited. Example: Hard beating of percussion instruments, hard plucking of string instruments (classical guitar or harp).	Timbral
Comb Filter Colouration	Comb Filter Colouration	'Hollow' sound. Often perceived as tonal colouration. Example: speaking through a tube.	Timbral
Phasey	Phasiness	Impression of time-varying phase relationships that result in modulated colouration.	Timbral
Metallic	Metallic tone colour	Colouration with pronounced narrow-band resonances, often as a result of low density of natural frequencies. Often audible when exciting metallic objects such as gongs, bells, rattling tin cans. Applicable to room simulations, plate reverb, spring reverb, too.	Timbral
Position Shifted Anti-Clockwise	Horizontal direction	Direction in the horizontal plane is shifted anti-clockwise.	Spatial
Position Shifted Clockwise	Horizontal direction	Direction in the horizontal plane is shifted clockwise.	Spatial
Position Shifted Up	Vertical direction	Direction in the vertical plane shifted up.	Spatial
Position Shifted Down	Vertical direction	Direction in the vertical plane shifted down.	Spatial
Close	Distance	Perceived position is close to the listener (short distance).	Spatial
Far	Distance	Perceived position is far from the listener (long distance).	Spatial
Front-back Reversal	Front-back reversal	A change of sound source position(s) from in front to behind, or vice versa.	Spatial
Wide	Width	Large perceived extent in horizontal direction.	Spatial
Narrow	Width	Small perceived extent in horizontal direction.	Spatial
Deep	Depth	Large perceived extent in radial direction.	Spatial
Shallow	Depth	Small perceived extent in radial direction.	Spatial
Tall	Height	Large perceived extent in vertical direction.	Spatial
Short	Height	Small perceived extent in vertical direction.	Spatial

Continued on next page

Table C.1 – *Continued from previous page*

Attribute	**Scale**	**Definition**	**Category**
Internal (inside the head)	Externalization	Perceived distinctly within the head, regardless of distance.	Spatial
External (outside the head)	Externalization	Perceived distinctly outside of the head, regardless of distance.	Spatial
Good Localizability	Localizability	Spatial extent and location are easy to determine.	Spatial
Poor Localizability	Localizability	Spatial extent and location are difficult to estimate, or appear diffuse.	Spatial
Envelopment - High	Envelopment	Sensation of being spatially surrounded by the sound source, scene, or ensemble. Typically, envelopment is associated with a scene. Being surrounded by reverberation would be considered highly enveloping. Being surrounded by a large number of dry sources may also be highly enveloping. This may be heard when standing and listening to the rain hitting the pavement. Envelopment may occur with reverberation or other aspects of the scene such as applause in a concert hall, atmosphere or air conditioning (room tone). Holes (an absence of sound from a certain directions) in the reproduction would normally reduce envelopment.	Spatial
Envelopment - Low	Envelopment	Lack of sensation of being spatially surrounded by the sound source, scene, or ensemble.	Spatial
Reverberation - High	Reverberation	Perception of strong reverberant energy and/or long duration of reverberant decay.	Room
Reverberation - Low	Reverberation	Perception of weak reverberant energy and/or short duration of reverberant decay.	Room
Imprecise Transients	Crispness	Perception of the reproduction of transients. Transients are soft/smoothed/imprecise.	Time behaviour
Unresponsive	Responsiveness	Characteristic that is affected by latencies in the reproduction system. Delayed response of a reproduction system with respect to user interaction.	Time behaviour
Poor Stability	Stability	Characteristics are not consistent over time or during source or listener movement.	Time behaviour
Loud	Loudness	High perceived loudness.	Dynamics

Continued on next page

Table C.1 – *Continued from previous page*

Attribute	**Scale**	**Definition**	**Category**
Quiet	Loudness	Low perceived loudness.	Dynamics
High Dynamic Range	Dynamic range	A large range in loudness.	Dynamics
Low Dynamic Range	Dynamic range	A small range in loudness.	Dynamics
Clear	Clarity	Clarity with respect to any characteristic of elements of a sound scene. Impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected.	General
Unclear	Clarity	Lack of clarity with respect to any characteristic of elements of a sound scene. Impression of how clearly different elements in a scene can be distinguished from each other, how well various properties of individual scene elements can be detected.	General
Natural	Naturalness	Impression that the sound source or scene is in accordance with the expectation/former experience of an equivalent sound source or scene.	General
Unnatural	Naturalness	Impression that the sound source or scene is not in accordance with the expectation/former experience of an equivalent sound source or scene.	General
Sense of Presence	Presence	Perception of “being-in-the-scene”, or “spatial presence”. Impression of being inside a presented scene or to be spatially integrated into the scene.	General
Poor Speech Intelligibility	Speech Intelligibility	The words of a speaker cannot be well understood.	General
Like	Degree-of-liking	Sense of pleasantness or enjoyment.	General
Dislike	Degree-of-liking	Sense of displeasure or lack of enjoyment.	General

Acronyms

2AFC two-alternative forced choice.

2D two-dimensional.

3AFC three-alternative forced choice.

3D three dimensions.

3D three-dimensional.

3GPP 3$^{\mathrm{rd}}$ Generation Partnership Project.

ADC analogue-to-digital converter.

ADM Audio Definition Model.

AES Audio Engineering Society.

ANOVA analysis of variance.

AR augmented reality.

AVE auditory virtual environment.

BAQ basic audio quality.

BBC British Broadcasting Corporation.

BEM boundary element method.

BILD binaural intelligibility level difference.

BIR binaural impulse response.

BRIR binaural room impulse response.

BWF Broadcast Wave Format.

CA correspondence analysis.

CATA check-all-that-apply.

CDT colouration detection threshold.

CT computed tomography.

CV coefficient of variation.

DA descriptive analysis.

DOA direction of arrival.

DoF degrees of freedom.

DRR direct-to-reverberant ratio.

DSFT discrete spatial Fourier transform.

DSP digital signal processing.

DTF directional transfer function.

EBU European Broadcasting Union.

EEG electroencephalography.

ERB equivalent rectangular bandwidth.

ERP event-related potential.

FDN feedback delay network.

FDTD finite difference time domain.

FEC free-air-equivalent coupling.

FFT fast Fourier transform.

FIR finite impulse response.

FM-BEM fast multipole accelerated boundary element method.

GPU graphics processing unit.

GUI graphical user interface.

HATS head and torso simulator.

HCA hierarchical cluster analysis.

HMD head-mounted display.

HOA higher-order ambisonics.

HpCF headphone-to-ear correction filter.

HpTF headphone-to-ear transfer function.

HRIR head-related impulse response.

HRTF head-related transfer function.

HSSP headphone surround sound processing.

IACC interaural cross-correlation.

IAVE interactive auditory virtual environment.

IC interaural coherence.

ICA independent component analysis.

ICTD inter-channel time difference.

IF influence factor.

IID interaural intensity difference.

IIR infinite impulse response.

ILD interaural level difference.

IPD interaural phase difference.

IR impulse response.

ISFT inverse spatial Fourier transform.

ITD interaural time difference.

ITU International Telecommunication Union.

ITU-R International Telecommunication Union Radiocommunication Sector.

IVAE interactive virtual acoustic environment.

JACK JACK audio connection kit.

JND just noticeable difference.

KEMAR Knowles Electronic Manikin for Acoustic Research.

LED light emitting diode.

LTI linear time-invariant.

MAA minimum audible angle.

MAMA minimum audible movement angle.

MFA multiple factor analysis.

MOS mean opinion score.

MPEG Moving Picture Experts Group.

MRI magnetic resonance imaging.

MUSHRA multiple stimulus presentation with hidden reference and anchor.

NGA next-generation audio.

OLE overall listening experience.

OSC Open Sound Control.

PCA principal component analysis.

PCM pulse code modulation.

PDR pressure division ratio.

PLS-R partial least squares regression.

QF quality feature.

QMF quadrature mirror filter.

QoE quality of experience.

QoS quality of service.

R&D Research & Development.

RATA rate-all-that-apply.

RGB red, green, blue.

RIR room impulse response.

RM-ANOVA repeated-measures analysis of variance.

RMS root mean square.

SAQI spatial audio quality inventory.

SDK software development kit.

SDT signal detection theory.

SFT spatial Fourier transform.

SIMD single instruction multiple data.

SNR signal-to-noise ratio.

SOFA spatially-oriented format for acoustics.

SSR SoundScape Renderer.

SVD singular value decomposition.

TCP transmission control protocol.

TOA time of arrival.

TSL total system latency.

TV television.

UDP user datagram protocol.

UK United Kingdom.

USB Universal Serial Bus.

VAE virtual acoustic environment.

VBAP vector base amplitude panning.

VFDL variable fractional delay line.

VR virtual reality.

WFS wave-field synthesis.

Symbols

This section presents a list mathematical symbols used in this thesis.

Physical Constants

c	Speed of sound	343.2 m/s in air at 20 °C and 1 bar

Number Sets

$\mathbb{R}$	Real Numbers
$\mathbb{C}$	Complex Numbers
$\mathbb{R}^{\mathrm{MxN}}$	Matrix of real numbers with M rows and N columns.
$\mathbb{S}^2$	Positions on the 2-sphere i.e. all 3D positions with unit range r.

Other Symbols

j	Imaginary number	$\sqrt{-1}$
k	Wavenumber	$2\pi/c$
Q	Quality factor	

Statistics

α	Significance level	$\alpha = 0.05$
μ	Mean	
ρ	Spearman's rank correlation coefficient	
σ	Standard deviation	
σ_μ	Standard error of the mean	

p	Probability value
r	Spearman’s correlation coefficient
W	Shapiro-Wilk test statistic
Z	Wilcoxon sign rank test statistic

Geometry

γ	Head yaw angle
ϕ	Elevation angle
ψ	Head tilt angle
θ	Azimuth angle
ζ	Head roll angle
r	Range/distance

Binaural Signals

$H_{l,r}$	Head-related transfer function for left and right ears
$Hp_{l,r}$	Headphone-to-ear correction filter for left and right ears
s	Sound source pressure
$p_{l,r}$	Sound pressure at the left and right ears

Bibliography

3D Sound Labs (2018). *3D Sound One*. URL: `http://www.3dsoundlabs.com/produit/3d-sound-one-headphones/` (visited on 11/21/2018).

3GPP TS 26.118 (2018). *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; 3GPP Virtual reality profiles for streaming applications (Release 15)*.

Aaron, P., M. Zeller, and M. Wojciakowski (2017). *Inside-out tracking*. URL: `https://docs.microsoft.com/en-us/windows/mixed-reality/enthusiast-guide/tracking-system` (visited on 10/21/2018).

Abdi, H., L. J. Williams, and D. Valentin (2013). "Multiple factor analysis: principal component analysis for multitable and multiblock data sets". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.2, pp. 149–179. ISSN: 19395108. DOI: `10.1002/wics.1246`.

Abel, J. S. and P. Huang (2006). "A Simple, Robust Measure of Reverberation Echo Density". In: *AES 121st Convention*. Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13819`.

Adams, J. et al. (2007). "Advantages and Uses of Check-All-That-Apply Response Compared to Traditional Scaling of Attributes for Salty Snacks". In: *7th Pangborn Sensory Science Symposium*.

AES69:2015 (2015). *AES standard for file exchange - Spatial acoustic data file format*.

Ahrens, J., M. Geier, and S. Spors (2008). "The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods". English. In: *124th AES Convention*. Amsterdam: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14460`.

Ahrens, J., H. Helmholz, and C. Andersson (2018). "Authentic Auralization of Acoustic Spaces based on Spherical Microphone Array Recordings". In: *Proceedings of the Institute of Acoustics*. Vol. 40. 3.

Ahrens, J., M. Thomas, and I. Tashev (2012). "HRTF magnitude modeling using a non-regularized least-squares fit of spherical harmonics coefficients on incomplete data". In: *ASIPA Annual Summit and Conference*. ISBN: 9780615700502. URL: `http://research.microsoft.com/pubs/179220/Ahrens_Thomas_Tashev_HRTFModeling_APSIPA_2012.pdf`.

Ahrens, J., H. Wierstorf, and S. Spors (2010). "Comparison of Higher Order Ambisonics And Wave Field Synthesis With Respect to Spatial Discretization Artifacts in Time Domain". In: *AES 40th International Conference: Spatial Sound*. Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=15563`.

Alais, D. (2004). "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration". In: *Current Biology* 14.3, pp. 257–262. ISSN: 09609822. DOI: `10.1016/S0960-9822(04)00043-0`.

Albrecht, R. and T. Lokki (2013). "Adjusting the Perceived Distance of Virtual Speech Sources by Modifying Binaural Room Impulse Responses". In: *International Conference on Auditory Display*. Lodz, Poland, pp. 233–241.

Algazi, V. R., C. Avendano, and R. O. Duda (2001a). "Elevation localization and head-related transfer function analysis at low frequencies". In: *The Journal of the Acoustical Society of America* 109.3, pp. 1110–1122. ISSN: 0001-4966. DOI: `10.1121/1.1349185`.

— (2001b). “Estimation of a Spherical-Head Model from Anthropometry”. In: *Journal of the Audio Engineering Society* 49.6, pp. 472–479. URL: http://www.aes.org/e-lib/browse.cfm?elib=10188.

Algazi, V. R., R. J. Dalton, et al. (2005). “Motion-Tracked Binaural Sound for Personal Music Players”. In: *119th AES Convention*. New York, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=13312.

Algazi, V. R., R. O. Duda, and D. M. Thompson (2004). “Motion-Tracked Binaural Sound”. In: *116th AES Convention*. Berlin, Germany. URL: http://www.aes.org/e-lib/browse.cfm?elib=12644.

Algazi, V., R. O. Duda, D. M. Thompson, and C. Avendano (2001). “The CIPIC HRTF database”. In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*. October. New Paltz, New York: IEEE, pp. 99–102. ISBN: 0-7803-7126-7. DOI: 10.1109/ASPAA.2001.969552.

Allen, J. B. and D. A. Berkley (1979). “Image method for efficiently simulating small room acoustics”. In: *The Journal of the Acoustical Society of America* 65.4, pp. 943–950. ISSN: 0001-4966. DOI: 10.1121/1.382599.

American Standards Association (1951). *American Standard – Acoustical Terminology*. URL: https://archive.org/details/ameri00amer/page/24.

Anderson, P. W. and P. Zahorik (2014). “Auditory/visual distance estimation: accuracy and variability”. In: *Frontiers in Psychology* 5.October. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01097.

Andreopoulou, A., D. R. Begault, and B. F. G. Katz (2015). “Inter-Laboratory Round Robin HRTF Measurement Comparison”. In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 895–906. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2015.2400417.

Andreopoulou, A. and B. F. G. Katz (2016). “Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees”. In: *Journal on Multimodal User Interfaces* 10.3, pp. 259–271. ISSN: 1783-7677. DOI: 10.1007/s12193-016-0214-y.

— (2017). “Identification of perceptually relevant methods of inter-aural time difference estimation”. In: *The Journal of the Acoustical Society of America* 142.2, pp. 588–598. ISSN: 0001-4966. DOI: 10.1121/1.4996457.

— (2018). “Comparing the effect of HRTF processing techniques on perceptual quality ratings”. In: *144th AES Convention*. Milan, Italy: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=19437.

Andreopoulou, A., A. Roginska, and J. P. Bello (2013). “Reduced Representations of HRTF datasets: A Discriminant Analysis Approach”. In: *Proceedings of 135th AES Convention*. New York, NY: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=16999.

Arend, J. M., A. Neidhardt, and C. Pörschmann (2016). “Measurement and Perceptual Evaluation of a Spherical Near-Field HRTF Set”. In: *29th Tonmeistertagung - VDT International Convention*. November, pp. 356–363.

Ares, G., F. Bruzzone, et al. (2014). “Evaluation of a rating-based variant of check-all-that-apply questions: Rate-all-that-apply (RATA)”. In: *Food Quality and Preference* 36, pp. 87–95. ISSN: 09503293. DOI: 10.1016/j.foodqual.2014.03.006.

Ares, G., R. Deliza, et al. (2010). “Comparison of two sensory profiling techniques based on consumer perception”. In: *Food Quality and Preference* 21.4, pp. 417–426. ISSN: 09503293. DOI: 10.1016/j.foodqual.2009.10.006.

Ares, G., E. Etchemendy, et al. (2014). “Visual attention by consumers to check-all-that-apply questions: Insights to support methodological development”. In: *Food Quality and Preference* 32, pp. 210–220. ISSN: 09503293. DOI: 10.1016/j.foodqual.2013.10.006.

Ares, G. and S. R. Jaeger (2013). “Check-all-that-apply questions: Influence of attribute order on sensory product characterization”. In: *Food Quality and Preference* 28.1, pp. 141–153. ISSN: 09503293. DOI: 10.1016/j.foodqual.2012.08.016.

Ares, G., A. Picallo, et al. (2018). “A comparison of RATA questions with descriptive analysis: Insights from three studies with complex/similar products”. In: *Journal of Sensory Studies*, e12458. ISSN: 08878250. DOI: 10.1111/joss.12458.

Ares, G., A. Tárrega, et al. (2014). “Investigation of the number of consumers necessary to obtain stable sample and descriptor configurations from check-all-that-apply (CATA) questions”. In: *Food Quality and Preference* 31.1, pp. 135–141. ISSN: 09503293. DOI: 10.1016/j.foodqual.2013.08.012.

Armstrong, M. et al. (2014). *White Paper 285 – Object-based broadcasting-curation, responsiveness and user experience*. Tech. rep. BBC Research & Development. URL: http://downloads.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP285.pdf.

ART (2013). *SmartTrack*. URL: http://www.ar-tracking.com/products/tracking-systems/smarttrack/ (visited on 11/09/2013).

Ascension Technology Corporation (2013). *Flock of Birds*. URL: http://www.ascension-tech.com/realtime/rtflockofbirds.php (visited on 11/09/2013).

Ashby, T., R. Mason, and T. Brookes (2013). “Head Movements in Three-Dimensional Localization”. English. In: *134th AES Convention*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=16782.

Ashmead, D. H., D. L. Davis, and A. Northington (1995). “Contribution of listeners’ approaching motion to auditory distance perception.” In: *Journal of Experimental Psychology: Human Perception and Performance* 21.2, pp. 239–256. ISSN: 1939-1277. DOI: 10.1037/0096-1523.21.2.239.

Azizi, S.-A. and T. Munch (2008). *Patent US8121319B2 – Tracking System Using Audio Signals Below Threshold*.

Backman, J. et al. (2017). “A Self-Calibrating Earphone”. In: *AES 142nd Convention*. Berlin, Germany. URL: http://www.aes.org/e-lib/browse.cfm?elib=18722.

Bailer, W. et al. (2015). “Multi-sensor concert recording dataset including professional and user-generated content”. In: *Proceedings of the 6th ACM Multimedia Systems Conference*. New York, New York, USA: ACM Press, pp. 201–206. ISBN: 9781450333511. DOI: 10.1145/2713168.2713191.

Bailey, W. and B. M. Fazenda (2018). “The effect of visual cues and binaural rendering method on plausibility in virtual environments”. In: *144th AES Convention*. URL: http://www.aes.org/e-lib/browse.cfm?elib=19438.

Barber, C. B. et al. (1996). “The Quickhull Algorithm for Convex Hulls”. In: *ACM Transactions on Mathematical Software* 22.4, pp. 469–483. ISSN: 00983500. DOI: 10.1145/235815.235821.

Barco Audio Technologies (2017). *Spatial Audio Workstation User Manual*. URL: http://www.iosono-sound.com/professional-audio/.

Barreiro, C. et al. (2010). “Application of a check-all-that-apply question to the development of chocolate milk desserts”. In: *Journal of Sensory Studies* 25.s1, pp. 67–86. ISSN: 08878250. DOI: 10.1111/j.1745-459X.2010.00290.x.

Bartz, P. (2012). *Razor AHRS Head-tracker, 9-DOF Razor IMU*. URL: https://github.com/ptrbrtz/razor-9dof-ahrs.

Bauer, B. B. (1961a). “Phasor Analysis of Some Stereophonic Phenomena”. In: *The Journal of the Acoustical Society of America* 33.11, pp. 1536–1539. ISSN: 0001-4966. DOI: 10.1121/1.1908492.

— (1961b). “Stereophonic Earphones and Binaural Loudspeakers”. In: *Journal of the Audio Engineering Society* 9.2, pp. 148–151.

Baumgarte, F. and C. Faller (2003). “Binaural cue coding-part I: psychoacoustic fundamentals and design principles”. In: *IEEE Transactions on Speech and Audio Processing* 11.6, pp. 509–519. ISSN: 1063-6676. DOI: 10.1109/TSA.2003.818109.

Baumgartner, R. and P. Majdak (2015). “Modeling Localization of Amplitude-Panned Virtual Sources in Sagittal Planes”. In: *J. Audio Eng. Soc* 63.7/8, pp. 562–569. URL: `http://www.aes.org/e-lib/browse.cfm?elib=17842`.

Baumgartner, R., P. Majdak, and B. Laback (2014). “Modeling sound-source localization in sagittal planes for human listeners”. In: *The Journal of the Acoustical Society of America* 136.2, pp. 791–802. ISSN: 0001-4966. DOI: `10.1121/1.4887447`.

BBC (2006). *BBC to show World Cup and Wimbledon in HDTV*. URL: `http://www.bbc.co.uk/pressoffice/pressreleases/stories/2006/03_march/23/hdtv.shtml` (visited on 12/08/2018).

— (2007). *BBC iPlayer to launch on 27 July*. URL: `http://www.bbc.co.uk/pressoffice/pressreleases/stories/2007/06_june/27/iplayer.shtml` (visited on 12/08/2018).

— (2011a). *BBC iPlayer apps, coming soon to Android and iPad*. URL: `http://www.bbc.co.uk/blogs/bbcinternet/2011/02/bbc_iplayer_apps_coming_soon_t.html` (visited on 12/08/2018).

— (2011b). *BBC iPlayer: iPhone app and 3G streaming across all mobile networks*. URL: `http://www.bbc.co.uk/blogs/bbcinternet/2011/12/iplayer_bbciplayer_iphone_android.html` (visited on 12/08/2018).

— (2018a). *Annual Report and Accounts 2017/18*. URL: `http://downloads.bbc.co.uk/aboutthebbc/insidethebbc/reports/pdf/bbc_annualreport_201718.pdf`.

— (2018b). *History of the BBC*. URL: `https://www.bbc.com/timelines/zxqc4wx` (visited on 12/07/2018).

BBC Charter (2016). *Royal Charter for the continuance of the British Broadcasting Corporation*. URL: `http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/about/how_we_govern/2016/charter.pdf`.

BBC Radio 4 - Drama (2016). *States of Mind: The Sky Is Wider - In Studio with The Sky Is Wider*. URL: `https://www.bbc.co.uk/programmes/p04038vr/p04034rg` (visited on 12/15/2018).

BBC R&D (2018a). *BBC R&D – Our Vision of the Future*. URL: `https://www.bbc.co.uk/rd/about/vision` (visited on 12/07/2018).

— (2018b). *History of BBC R&D*. URL: `https://www.bbc.co.uk/rd/about/history` (visited on 12/07/2018).

Bech, S. (1996). “Timbral aspects of reproduced sound in small rooms. II”. In: *The Journal of the Acoustical Society of America* 99.6, pp. 3539–3549. ISSN: 0001-4966. DOI: `10.1121/1.414952`.

Bech, S. (1995). “Timbral aspects of reproduced sound in small rooms. I”. In: *The Journal of the Acoustical Society of America* 97.3, pp. 1717–1726. ISSN: 0001-4966. DOI: `10.1121/1.413047`.

Bech, S. and N. Zacharov (2006). *Perceptual Audio Evaluation: Theory, Method and Application*. Wiley. ISBN: 978-0-470-86923-9.

Begault, D. R. (1991). “Challenges to the Succesful Implementation of 3-D Sound”. In: *Journal of the Audio Engineering Society* 39.11, pp. 864–870. URL: `http://www.aes.org/e-lib/browse.cfm?elib=10281`.

— (1992). “Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems”. In: *Journal of the Audio Engineering Society* 40.11, pp. 895–904. URL: `http://www.aes.org/e-lib/browse.cfm?elib=7027`.

Begault, D. R. and E. M. Wenzel (1993). “Headphone Localization of Speech”. In: *Human Factors* 35.2, pp. 361–376. DOI: `10.1177/001872089303500210`.

Begault, D. R., E. M. Wenzel, and M. R. Anderson (2001). “Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualised 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. URL: `http://www.aes.org/e-lib/browse.cfm?elib=10175`.

Begault, D. R., E. M. Wenzel, M. Godfroy, et al. (2010). “Applying Spatial Audio to Human Interfaces: 25 Years of NASA Experience”. In: *AES 40th International Conference: Spatial Sound*. Tokyo, Japan: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=15546`.

Bell, R. (2018). *BBC iPlayer monthly report: October 2018*. URL: http://downloads.bbc.co.uk/mediacentre/iplayer/iplayer-performance-oct18.pdf.

Ben Hagai, I. et al. (2011). "Acoustic centering of sources measured by surrounding spherical microphone arrays". In: *The Journal of the Acoustical Society of America* 130.4, pp. 2003–2015. ISSN: 0001-4966. DOI: 10.1121/1.3624825.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

Bennett, J. C., K. Barker, and F. O. Edeko (1985). "A New Approach to the Assessment of Stereophonic Sound System Performance". In: *Journal of the Audio Engineering Society* 33.5, pp. 314–321. URL: http://www.aes.org/e-lib/browse.cfm?elib=4449.

Bentley, J. L. (1975). "Multidimensional binary search trees used for associative searching". In: *Communications of the ACM* 18.9, pp. 509–517. ISSN: 00010782. DOI: 10.1145/361002.361007.

Beresford, K. et al. (2006). "Contextual effects on sound quality judgements: Part II – multi- stimulus vs. single stimulus method". In: *121st AES Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=13747.

Berg, J. and F. Rumsey (1999). "Spatial Attribute Identification and Scaling by Repertory Grid Technique and Other Methods". English. In: *16th International Conference of the Audio Engineering Society*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8049.

— (2000a). "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction". In: *AES 109th Convention*. Los Angeles, CA, USA. URL: http://www.aes.org/e-lib/browse.cfm?elib=9132.

— (2000b). "In Search of The Spatial Dimensions of Reproduced Sound: Verbal Protocol Analysis and Cluster Analysis of Scaled Verbal Descriptors". In: *108th AES Convention*. Paris, France. URL: http://www.aes.org/e-lib/browse.cfm?elib=9199.

— (2001). "Verification and correlation of attributes used for describing the spatial quality of reproduced sound". In: *AES 19th International Conference*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=10057.

— (2003). "Systematic Evaluation of Peceived Spatial Quality". In: *AES 24th International Conference on Multichannel Audio*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=12272.

— (2006). "Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique". In: *Journal of the Audio Engineering Society* 54.5, pp. 365–379. URL: http://www.aes.org/e-lib/browse.cfm?elib=13681.

Berge, S. and N. Barrett (2010). "A new method for B-format to binaural transcoding". In: *AES 40th International Conference: Spatial Sound*. Tokyo, Japan. URL: http://www.aes.org/e-lib/browse.cfm?elib=15527.

Berger, C. C. et al. (2018). "Generic HRTFs May be Good Enough in Virtual Reality. Improving Source Localization through Cross-Modal Plasticity". In: *Frontiers in Neuroscience* 12.February. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00021.

Bergstrom, I. et al. (2017). "The Plausibility of a String Quartet Performance in Virtual Reality". In: *IEEE Transactions on Visualization and Computer Graphics* 23.4, pp. 1352–1359. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2657138.

Bernfeld, B. (1973). "Attempts for Better Understanding of the Directional Sterephonic Listening Mechanism". In: *44th AES Convention*. Rotterdam, The Netherlands: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=1743.

Bernschütz, B. (2013). "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100". In: *Proceedings the AIA-DAGA*. Merano, Italy, pp. 592–595.

— (2016). "Microphone Arrays and Sound Field Decomposition for Dynamic Binaural Recording". Doctoral Thesis. Technische Universität Berlin. DOI: `10.14279/depositonce-5082`.

Bernschütz, B. et al. (2014). "Binaural Reproduction of Plane Waves With Reduced Modal Order". In: *Acta Acustica united with Acustica* 100.5, pp. 972–983. ISSN: 16101928. DOI: `10.3813/AAA.918777`.

Beyerdynamic (2018). *Headzone Pro XT V2.1.1*. URL: `https://europe.beyerdynamic.com/catalog/product/view/_ignore_category/1/id/40/s/headzone-pro-xt/` (visited on 11/20/2018).

Blake, I. (2018). *Daily Mail Online | BBC Blue Planet II inspired shoppers to go plastic free*. URL: `https://www.dailymail.co.uk/femail/food/article-5946339/BBC-Blue-Planet-II-inspired-Lakeland-shoppers-buy-reusable-eco-friendly-plastic-free-products.html` (visited on 12/09/2018).

Blanco, J. L. and P. K. Rai (2014). *nanoflann: a C++ header-only fork of FLANN, a library for Nearest Neighbor (NN) with KD-trees*. URL: `https://github.com/jlblancoc/nanoflann`.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. Revised Ed. MIT Press. ISBN: 0262024136.

— (2014). *The Technology of Binaural Listening*. Ed. by J. Blauert. 1st Editio. Springer. ISBN: 9783642377617.

Blauert, J. and U. Jekosch (1997). "Sound-Quality Evaluation - A Multi-Layered Problem". In: *Acta Acustica united with Acustica* 83, pp. 747–753.

— (2003). "Concepts Behind Sound Quality: Some Basic Considerations". In: *32nd International Congress and Exposition on Noise Control Engineering*. Seogwipo, Korea.

— (2012). "A Layer Model of Sound Quality". In: *Journal of the Audio Engineering Society* 60.1/2, pp. 4–12. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16160`.

Blauert, J., D. Kolossa, et al. (2013). "Further Challenges and the Road Ahead". In: *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 477–501. DOI: `10.1007/978-3-642-37762-4_18`.

Blauert, J. and W. Lindemann (1986). "Auditory spaciousness: Some further psychoacoustic analyses". In: *The Journal of the Acoustical Society of America* 80.2, pp. 533–542. ISSN: 0001-4966. DOI: `10.1121/1.394048`.

Blumlein, A. (1931). *Patent GB394325 – Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems*.

Boerger, G., P. Laws, and J. Blauert (1977). "Stereophone Kopfhörerwiedergabe mit Steuerung bestimmter Übertragungsfaktoren durch Kopfdrehbewegungen [Stereophonic reproduction through headphones with control of special transfer functions by head movements]". In: *Acustica* 39, pp. 22–26.

Bomhardt, R., M. de la Fuente Klein, and J. Fels (2016). "A high-resolution head-related transfer function and three-dimensional ear model database". In: *Proceedings of Meetings on Acoustics 29*. DOI: `10.1121/2.0000467`.

Borg, I. and P. J. F. Groenen (2005). *Modern Multidimensional Scaling: Theory and Application*. Springer. ISBN: 9780387289816.

Borish, J. (1984). "Extension of the image model to arbitrary polyhedra". In: *The Journal of the Acoustical Society of America* 75.6, pp. 1827–1836. ISSN: 0001-4966. DOI: `10.1121/1.390983`.

Bork, I. (2000). "A Comparison of Room Simulation Software - The 2nd Round Robin on Room Acoustical Computer Simulation". In: *Acta Acustica united with Acustica,* 86.6, pp. 943–956. URL: `http://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000006/art00008`.

Borß, C. et al. (2016). *Patent US9729995B2 – Apparatus and method for generating a plurality of audio channels*. URL: `https://patents.google.com/patent/US9729995`.

Brandenburg, K. et al. (2018). “Plausible Augmentation of Auditory Scenes Using Dynamic Binaural Synthesis for Personalized Auditory Realities”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=19691.

Breebaart, J. (2007). “Analysis and Synthesis of Binaural Parameters for Efficient 3D Audio Rendering in MPEG Surround”. In: *IEEE International Conference on Multimedia and Expo*. Beijing, China: IEEE, pp. 1878–1881. ISBN: 1-4244-1016-9. DOI: 10.1109/ICME.2007.4285041.

— (2017). “No correlation between headphone frequency response and retail price”. In: *The Journal of the Acoustical Society of America* 141.6, EL526–EL530. ISSN: 0001-4966. DOI: 10.1121/1.4984044.

Breebaart, J. et al. (2006). “Multi-Channel Goes Mobile: MPEG Surround Binaural Rendering”. In: *AES 29th International Conference*. Seoul, Korea: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=13864.

Bregman, A. S. (1999). *Auditory Scene Analysis*. 2nd Editio. MIT Press. ISBN: 9780262521956.

Brimijoin, W. O. and M. A. Akeroyd (2014). “The moving minimum audible angle is smaller during self motion than during source motion”. In: *Frontiers in Neuroscience* 8. ISSN: 1662-453X. DOI: 10.3389/fnins.2014.00273.

Brimijoin, W. O., A. W. Boyd, and M. A. Akeroyd (2013). “The Contribution of Head Movement to the Externalization and Internalization of Sounds”. In: *PLoS ONE* 8.12. Ed. by C. Alain, e83068. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0083068.

Brinkmann, F. (2011). “Individual headphone compensation for binaural synthesis”. Master’s thesis. Technische Universität Berlin.

Brinkmann, F., A. Lindau, and S. Weinzierl (2014). “Assessing the Authenticity of Individual Dynamic Binaural Synthesis”. In: *EAA Joint Smyposium on Auralization and Ambisonics*. April. Berlin, pp. 62–68. DOI: 10.14279/depositonce-11.

Brinkmann, F., A. Lindau, S. Weinzierl, G. Geissler, and S. V. D. Par (2013). “A high resolution head-related transfer function database including different orientations of head above the torso”. In: *Proceedings of AIA-DAGA 2013*. Merano, Italy.

Brinkmann, F., A. Lindau, S. Weinzierl, G. Geissler, S. van de Par, et al. (2017). *The FABIAN head-related transfer function data base*. DOI: 10.14279/depositonce-5718.

Brinkmann, F., R. Roden, et al. (2014). “Audibility of head-above-torso orientation in head-related transfer functions”. In: *Forum Acusticum*. Kraków, Poland.

— (2015). “Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology”. In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 931–942. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2015.2414905.

Bronkhorst, A. W. (1995). “Localization of real and virtual sound sources”. In: *The Journal of the Acoustical Society of America* 98.5, pp. 2542–2553. ISSN: 0001-4966. DOI: 10.1121/1.413219.

— (2000). “The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions”. In: *Acustica* 86, pp. 117–128. URL: http://www.ingentaconnect.com/content/dav/aaua/2000/00000086/00000001/art00016.

Brown, C. and R. O. Duda (1998). “A structural model for binaural sound synthesis”. In: *IEEE Transactions on Speech and Audio Processing* 6.5, pp. 476–488. ISSN: 10636676. DOI: 10.1109/89.709673.

Brüel & Kjær (2018). *Type 9640 Turntable System*. URL: https://www.bksv.com/en/Products/transducers/ear-simulators/electroacoustic-accessories/turntable-type-9640.aspx (visited on 10/31/2018).

Brüggen, M. (2001). “Coloration and Binaural Decoloration in Natural Environments”. In: *Acta Acustica united with Acustica* 87.3, pp. 400–406. URL: `https://www.ingentaconnect.com/content/dav/aaua/2001/00000087/00000003/art00012`.

Brungart, D. S. (1999a). “Auditory localization of nearby sources. III. Stimulus effects”. In: *The Journal of the Acoustical Society of America* 106.6, pp. 3589–3602. ISSN: 0001-4966. DOI: `10.1121/1.428212`.

— (1999b). “Auditory parallax effects in the HRTF for nearby sources”. In: *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 171–174. ISBN: 0-7803-5612-8. DOI: `10.1109/ASPAA.1999.810877`.

Brungart, D. S. and W. M. Rabinowitz (1999). “Auditory localization of nearby sources. Head-related transfer functions”. en. In: *The Journal of the Acoustical Society of America* 106.3, p. 1465. ISSN: 00014966. DOI: `10.1121/1.427180`.

Brungart, D. S. and G. D. Romigh (2009). “Spectral HRTF enhancement for improved vertical-polar auditory localization”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 305–308. ISSN: 1931-1168. DOI: `10.1109/ASPAA.2009.5346479`.

Brungart, D. S. and K. R. Scott (2001). “The effects of production and presentation level on the auditory distance perception of speech”. In: *The Journal of the Acoustical Society of America* 110.1, pp. 425–440. ISSN: 0001-4966. DOI: `10.1121/1.1379730`.

Brungart, D. S., B. D. Simpson, and A. J. Kordik (2005). “The detectability of headtracker latency in virtual audio displays”. In: *Proceedings of the 11th International Conference on Auditory Display*. Limerick, Ireland, pp. 37–42.

Bücklein, R. (1981). “The Audibility of Frequency Response Irregularities”. In: *Journal of the Audio Engineering Society* 29.3, pp. 126–131. URL: `http://www.aes.org/e-lib/browse.cfm?elib=10291`.

Burkhard, M. D. (1978). “Non hearing-aid uses of the KEMAR manikin”. In: *Manikin Measurements*. Ed. by M. D. Burkhard, pp. 63–65.

Burkhard, M. D. and R. M. Sachs (1975). “Anthropometric manikin for acoustic research”. In: *The Journal of the Acoustical Society of America* 58.1, pp. 214–222. ISSN: 0001-4966. DOI: `10.1121/1.380648`.

Business Wire (2017). *South Korea Launches UHD TV with MPEG-H Audio | Business Wire*. URL: `https://www.businesswire.com/news/home/20170601006518/en/South-Korea-Launches-UHD-TV-MPEG-H-Audio` (visited on 11/18/2018).

Busson, S., R. Nicol, and B. F. G. Katz (2005). “Subjective investigations of the interaural time difference in the horizontal plane”. In: *AES 118th Convention*. Barcelona, Spain: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13040`.

Cairncross, S. E. and L. B. Sjostrom (1950). “Flavor Profiles: A New Approach to Flavor Problems”. In: *Descriptive Sensory Analysis in Practice*. Vol. 4. Trumbull, Connecticut, USA: Food & Nutrition Press, Inc., pp. 15–22. DOI: `10.1002/9780470385036.ch1b`.

Campo, E., J. Ballester, et al. (2010). “Comparison of conventional descriptive analysis and a citation frequency-based descriptive method for odor profiling: An application to Burgundy Pinot noir wines”. In: *Food Quality and Preference* 21.1, pp. 44–55. ISSN: 09503293. DOI: `10.1016/j.foodqual.2009.08.001`.

Campo, E., B. V. Do, et al. (2008). “Aroma properties of young Spanish monovarietal white wines: a study using sorting task, list of terms and frequency”. In: *Australian Journal of Grape and Wine Research* 14, pp. 104–115. DOI: `10.1111/j.1755-0238.2008.00010.x`.

Carlile, S., C. T. Jin, and V. V. Raad (2000). “Continuous Virtual Auditory Space Using HRTF Interpolation: Acoustic & Psychophysical Errors”. In: *International Symposium on Multimedia Information Processing*. IEEE. URL: `http://www.ee.usyd.edu.au/carlab/CARlabPublicationsData/PDF/S123-2623056392/S123.pdf`.

Carpentier, T., H. Bahu, et al. (2014). "Measurement of a head-related transfer function database with high spatial resolution". In: *7th Forum Acusticum (EAA)*. Kraków, Poland.

Carpentier, T., M. Noisternig, and O. Warusfel (2015). "Twenty years of Ircam Spat: looking back, looking forward". In: *International Computer Music Conference Proceedings*, pp. 270–277. ISSN: 0270-6474.

Carpentier, T., T. Szpruch, et al. (2013). "Parametric control of convolution based room simulators". In: *International Symposium on Room Acoustics*. Toronto, Canada, pp. 1–11.

Catic, J. et al. (2013). "The effect of interaural-level-difference fluctuations on the externalization of sound". In: *The Journal of the Acoustical Society of America* 134.2, pp. 1232–1241. ISSN: 0001-4966. DOI: 10.1121/1.4812264.

Chandler, D. W. and D. W. Grantham (1992). "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity". In: *The Journal of the Acoustical Society of America* 91.3, pp. 1624–1636. ISSN: 0001-4966. DOI: 10.1121/1.402443.

Chechik, G. and I. Nelken (2012). "Auditory abstraction from spectro-temporal features to coding auditory entities". In: *Proceedings of the National Academy of Sciences* 109.46, pp. 18968–18973. ISSN: 0027-8424. DOI: 10.1073/pnas.1111242109.

Chittka, L. and A. Brockmann (2005). "Perception Space—The Final Frontier". In: *PLoS Biology* 3.4, e137. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0030137.

Choisel, S. and F. Wickelmaier (2007). "Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference." In: *The Journal of the Acoustical Society of America* 121.1, pp. 388–400. ISSN: 00014966. DOI: 10.1121/1.2385043.

Christensen, A. T., W. Hess, et al. (2013). "Magnitude and Phase Response Measurement of Headphones at the Eardrum". In: *51st AES International Conference: Loudspeakers and Headphones*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=16877.

Christensen, F., C. B. Jensen, and H. Møller (2000). "The Design of VALDEMAR-An Artificial Head for Binaural Recording Purposes". In: *109th AES Convention*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=9085.

Chun, C. J. et al. (2017). "Deep Neural Network Based HRTF Personalization Using Anthropometric Measurements". In: *AES 143rd Convention*. New York, NY, USA, pp. 1–5. URL: http://www.aes.org/e-lib/browse.cfm?elib=19257.

Cieciura, C. et al. (2018). "Survey of Media Device Ownership, Media Service Usage, and Group Media Consumption in UK Households". In: *AES 145th Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=19721.

Civille, G. V. and I. H. Liska (1975). "Modifications and Applications to Foods of the General Foods Sensory Texture Profiles Technique". In: *Journal of Texture Studies* 6.1, pp. 19–31. DOI: 10.1111/j.1745-4603.1975.tb01115.x.

Cobos, M. et al. (2015). "Subjective quality assessment of multichannel audio accompanied with video in representative broadcasting genres". In: *Multimedia Systems* 21.4, pp. 363–379. ISSN: 0942-4962. DOI: 10.1007/s00530-013-0340-2.

Cochran, W. G. (1950). "The Comparison of Percentages in Matched Samples". In: *Biometrika* 37.3-4, pp. 256–266. ISSN: 0006-3444. DOI: 10.1093/biomet/37.3-4.256.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Second edi. Lawrence Erlbaum Associates. ISBN: 0805802835.

Coleman, P., A. Franck, J. Francombe, et al. (2018). “An Audio-Visual System for Object-Based Audio: From Recording to Listening”. In: *IEEE Transactions on Multimedia* 20.8, pp. 1919–1931. ISSN: 1520-9210. DOI: `10.1109/TMM.2018.2794780`.

Coleman, P., A. Franck, P. Jackson, et al. (2017). “Object-Based Reverberation for Spatial Audio”. In: *Journal of the Audio Engineering Society* 65.1/2, pp. 66–77. ISSN: 15494950. DOI: `10.17743/jaes.2016.0059`.

Cote, N., V. Koehl, and M. Paquier (2012). “Ventriloquism effect on distance auditory cues”. In: *Proceedings of Acoustics 2012*. April. Nantes, France, pp. 1063–1067.

Courtois, G. et al. (2018). “Effects of Binaural Spatialization in Wireless Microphone Systems for Hearing Aids on Normal-Hearing and Hearing-Impaired Listeners”. In: *Trends in Hearing* 22, pp. 1–17. ISSN: 2331-2165. DOI: `10.1177/2331216517753548`.

Cowan, N. (1984). “On short and long auditory stores.” In: *Psychological Bulletin* 96.2, pp. 341–370. ISSN: 1939-1455. DOI: `10.1037/0033-2909.96.2.341`.

Craven, P. and M. A. Gerzon (1975). *Patent US4042779A – Coincident microphone simulation covering three dimensional space and yielding various directional outputs*. URL: `https://patents.google.com/patent/US4042779A/en`.

Damaske, P. and B. Wagener (1969). “Directional Hearing Tests by the Aid of a Dummy Head”. In: *Acta Acustica united with Acustica* 21.1, 30–35(6).

Daniel, J. (2000). “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD thesis. Université Pierre et Marie Curie (Paris VI): Paris. URL: `http://pcfarina.eng.unipr.it/Public/phd-thesis/jd-these-original-version.pdf`.

— (2003). “Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format”. In: *AES 23rd International Conference*. Copenhagen, Denmark: Audio Engineering Society.

Daniel, J., R. Nicol, and S. Moreau (2003). “Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging”. In: *114th AES Convention*. Amsterdam, The Netherlands: Audio Engineering Society.

Daniel, J., J.-B. Rault, and J.-D. Polack (1998). “Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions”. In: *AES 105th Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=8385`.

Dejardin, H. (2018). *Personal correspondence regarding spatial audio production methods at Radio France*. URL: `http://hyperradio.radiofrance.fr/son-3d/`.

Delarue, J. and J.-M. Sieffermann (2004). “Sensory mapping using Flash profile. Comparison with a conventional descriptive method for the evaluation of the flavour of fruit dairy products”. In: *Food Quality and Preference* 15.4, pp. 383–392. ISSN: 09503293. DOI: `10.1016/S0950-3293(03)00085-5`.

Dick, S., N. Schinkel-Bielefeld, and S. Disch (2017). “Generation and Evaluation of Isolated Audio Coding Artifacts”. In: *AES 143rd Convention*. New York, NY, USA: Audio Engineering Society.

Dietrich, P., B. Masiero, and M. Vorländer (2013). “On the Optimization of the Multiple Exponential Sweep Method”. English. In: *Journal of the Audio Engineering Society* 61.3, pp. 113–124. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16672`.

Dietz, M., S. D. Ewert, and V. Hohmann (2011). “Auditory model based direction estimation of concurrent speakers from binaural signals”. In: *Speech Communication* 53.5, pp. 592–605. ISSN: 01676393. DOI: `10.1016/j.specom.2010.05.006`.

Dietz, M., J.-H. Lestang, et al. (2018). “A framework for testing and comparing binaural models”. In: *Hearing Research* 360, pp. 92–106. ISSN: 03785955. DOI: 10.1016/j.heares.2017.11.010.

DIN EN (2002). *ISO 11904-1 – Determination of sound immission from sound sources placed close to the ear - part 1: Technique using microphone in real ear (MIRE technique)*.

Dolby (2015). *Dolby AC-4: Audio Delivery for Next-Generation Entertainment Services*. URL: https://www.dolby.com/us/en/technologies/ac-4/Next-Generation-Entertainment-Services.pdf.

Dooley, L., Y. seung Lee, and J. F. Meullenet (2010). “The application of check-all-that-apply (CATA) consumer profiling to preference mapping of vanilla ice cream and its comparison to classical external preference mapping”. In: *Food Quality and Preference* 21.4, pp. 394–401. ISSN: 09503293. DOI: 10.1016/j.foodqual.2009.10.002.

Dravnieks, A. (1982). “Odor Quality: Semantically Generated Multidimensional Profiles Are Stable”. In: *Science* 218.4574, pp. 799–802.

DTG (2018). *Next Generation Audio Study Group Terms of Reference*. URL: https://dtg.org.uk/wp-content/uploads/2018/07/NGA_001-Next-Generation-Audio-Group-Draft-Terms-of-reference-v2.docx (visited on 11/18/2018).

Du Moncel, T. (1881). “The International Exhibition and Congress of Electricity at Paris 1”. In: *Nature* 24.625, pp. 585–589. ISSN: 0028-0836. DOI: 10.1038/024585b0.

Duda, R. O., V. R. Algazi, and D. M. Thompson (2002). “The Use of Head-and-Torso Models for Improved Spatial Sound Synthesis”. In: *113th AES Convention*. Los Angeles, CA, USA: Audio Engineering Society.

Duda, R. O., C. Avendano, and V. Algazi (1999). “An adaptable ellipsoidal head model for the interaural time difference”. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. IEEE, 965–968 vol.2. ISBN: 0-7803-5041-3. DOI: 10.1109/ICASSP.1999.759855.

Duraiswaini, R., D. N. Zotkin, and N. Gumerov (2004). “Interpolation and range extrapolation of HRTFs [head related transfer functions]”. In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 4. 5. Montreal, Canada: IEEE, pp. 45–48. ISBN: 0-7803-8484-9. DOI: 10.1109/ICASSP.2004.1326759.

Duraiswami, R. et al. (2005). “High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues”. In: *119th AES Convention*. New York, NY, USA: Audio Engineering Society.

Durlach, N. I., A. Rigopulos, et al. (1992). “On the Externalization of Auditory Images”. In: *Presence: Teleoperators and Virtual Environments* 1.2, pp. 251–257. ISSN: 1054-7460. DOI: 10.1162/pres.1992.1.2.251.

Durlach, N. and H. S. Colburn (1978). “Binaural Phenomena”. In: *Hearing*. Ed. by E. C. Carterette and M. P. Friedman. Academic Press, pp. 365–466. ISBN: 9780121619046. DOI: 10.1016/B978-0-12-161904-6.50017-8.

EBU R128 (2014). *Loudness Normalisation and Permitted Maximum Level of Audio Signals*. URL: https://tech.ebu.ch/docs/r/r128.pdf.

EBU TECH 3253 (2008). *Sound Quality Assessment Material recordings for subjective tests*.

EBU TECH 3285 (2011). *Specification of the Broadcast Wave Format (BWF)*.

EBU TECH 3364 (2014). *Audio Definition Model (ADM)*.

EBU Tech 3388 (2018). *ADM Renderer for use in Next Generation Audio Broadcasting – Specification Version 1.0*. URL: https://tech.ebu.ch/publications/tech3388.

Egan, D. et al. (2016). “An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments”. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: 10.1109/QoMEX.2016.7498964.

Engelke, U. et al. (2017). “Psychophysiology-Based QoE Assessment: A Survey”. In: *IEEE Journal on Selected Topics in Signal Processing* 11.1, pp. 6–21. ISSN: 19324553. DOI: 10.1109/JSTSP.2016.2609843.

Erbes, V. et al. (2012). “An extraaural headphone system for optimized binaural reproduction”. In: *Proceedings of 38th DAGA*. Darmstadt, Germany, pp. 17–18.

ETSI TS 101 154 (2018). *Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications*.

ETSI TS 102 366 (2017). *Digital Audio Compression (AC-3, Enhanced AC-3) Standard – V1.4.1*.

ETSI TS 103 190 (2014). *Digital Audio Compression (AC-4) Standard*.

ETSI TS 103 491 (2017). *DTS-UHD Audio Format; Delivery of Channels, Objects and Ambisonic Sound Fields*.

Evans, M. J., J. A. S. Angus, and A. I. Tew (1998). “Analyzing head-related transfer function measurements using surface spherical harmonics”. In: *The Journal of the Acoustical Society of America* 104.4, pp. 2400–2411. ISSN: 0001-4966. DOI: 10.1121/1.423749.

Evans, M., T. Ferne, et al. (2017). “Creating Object-Based Experiences in the Real World”. In: *SMPTE Motion Imaging Journal* 126.6, pp. 1–7. ISSN: 1545-0279. DOI: 10.5594/JMI.2017.2709859.

Facebook (2018a). *Audio360 SDK*. URL: https://facebookincubator.github.io/facebook-360-spatial-workstation/Documentation/SDK/Audio360_SDK_GettingStarted.html.

— (2018b). *Facebook 360 Spatial Workstation Knowledge Base - Creating Videos with Spatial Audio for Facebook 360*. URL: https://facebookincubator.github.io/facebook-360-spatial-workstation/KB/CreatingVideosSpatialAudioFacebook360.html (visited on 06/25/2018).

Fallahi, M., F. Brinkmann, and S. Weinzierl (2015). “Simulation and analysis of measurement techniques for the fast acquisition of head-related transfer functions of head-related transfer functions”. In: *Proceedings of DAGA 2015*. Nürnberg, Germamny, pp. 1107–1110.

Faller, C. and F. Baumgarte (2003). “Binaural cue coding-part II: schemes and applications”. In: *IEEE Transactions on Speech and Audio Processing* 11.6, pp. 520–531. ISSN: 1063-6676. DOI: 10.1109/TSA.2003.818108.

Faller, C. and J. Breebaart (2011). “Binaural Reproduction of Stereo Signals Using Upmixing and Diffuse Rendering”. In: *131st AES Convention*. New York, NY, USA: Audio Engineering Society.

Faller, C. and J. Merimaa (2004). “Source localization in complex listening situations: Selection of binaural cues based on interaural coherence”. In: *The Journal of the Acoustical Society of America* 116.5, pp. 3075–3089. ISSN: 0001-4966. DOI: 10.1121/1.1791872.

Farina, A. (2007). “Advancements in impulse response measurements by sine sweeps”. In: *Proceedings of 122nd AES Convention*. Vienna, Austria: Audio Engineering Society.

Faul, F. et al. (2007). “G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences”. In: *Behavior Research Methods* 39.2, pp. 175–191. ISSN: 1554-351X. DOI: 10.3758/BF03193146.

Fayek, H. M. et al. (2017). “On Data-Driven Approaches to Head-Related Transfer Function Personalization”. In: *143rd AES Convention*. New York, NY, USA: Audio Engineering Society.

Fink, K. J. and L. Ray (2012). “Tuning principal component weights to individualize HRTFs”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2. IEEE, pp. 389–392. ISBN: 978-1-4673-0046-9. DOI: 10.1109/ICASSP.2012.6287898.

Fleischmann, F., J. Plogsties, and B. Neugebauer (2013). “Design of a Headphone Equalizer Control Based on Principal Component Analysis”. English. In: *134th AES Convention*. Rome, Italy: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=16770.

Fleischmann, F., A. Silzle, and J. Plogsties (2012). “Identification and Evaluation of Target Curves for Headphones”. In: *Proceedings of 133rd AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

Foster, S., E. M. Wenzel, and R. Tayior (1991). “Real Time Synthesis of Complex Acoustic Environments”. In: *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA: IEEE, pp. 47–48. DOI: `10.1109/ASPAA.1991.634098`.

Four Audio (2018). *ELF – Computer-Controlled 3D Loudspeaker Measurement System for Balloon Measurements*. URL: `http://fouraudio.com/en/products/elf.html` (visited on 10/28/2018).

Franck, A. (2014). “Efficient Frequency-Domain Filter Crossfading for Fast Convolution with Application to Binaural Synthesis”. In: *Proceedings of the 55th AES International Conference*. Helsinki, Finland: Audio Engineering Society.

Franck, A., F. M. Fazi, and F. Melchior (2015). “Optimization-based reproduction of diffuse audio objects”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–14. ISSN: 10715797. DOI: `10.1109/WASPAA.2015.7336892`.

Franck, A., W. Wang, and F. M. Fazi (2017). “Sparse l1-Optimal Multiloudspeaker Panning and Its Relation to Vector Base Amplitude Panning”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.5, pp. 996–1010. ISSN: 2329-9290. DOI: `10.1109/TASLP.2017.2674975`.

Francombe, J., T. Brookes, and R. Mason (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. ISSN: 15494950. DOI: `10.17743/jaes.2016.0070`.

Francombe, J., R. Mason, et al. (2014). “Elicitation of attributes for the evaluation of audio-on-audio interference”. In: *The Journal of the Acoustical Society of America* 136.5, pp. 2630–2641. ISSN: 0001-4966. DOI: `10.1121/1.4898053`.

Frank, M. (2013). “Source Width of Frontal Phantom Sources: Perception, Measurement, and Modeling”. In: *Archives of Acoustics* 38.3, pp. 311–319. ISSN: 0137-5075. DOI: `10.2478/aoa-2013-0038`.

Freeman, T. C. A. et al. (2017). “Auditory compensation for head rotation is incomplete.” In: *Journal of Experimental Psychology: Human Perception and Performance* 43.2, pp. 371–380. ISSN: 1939-1277. DOI: `10.1037/xhp0000321`.

FreeTrack (2013). *FreeTrack*. URL: `http://www.free-track.net/english/` (visited on 11/09/2013).

Frigo, M. and S. Johnson (2005). “The Design and Implementation of FFTW3”. In: *Proceedings of the IEEE* 93.2, pp. 216–231. ISSN: 0018-9219. DOI: `10.1109/JPROC.2004.840301`.

Fritz, C. O., P. E. Morris, and J. J. Richler (2012). “Effect size estimates: Current use, calculations, and interpretation.” In: *Journal of Experimental Psychology: General* 141.1, pp. 2–18. ISSN: 1939-2222. DOI: `10.1037/a0024338`.

Furse, R. (2010). *Patent GB 2467534 A – Sound System – Methods and systems for using transforms to modify the spatial characteristics of audio data*. URL: `http://doi.apa.org/getdoi.cfm?doi=10.1037/a0024338`.

Futuresource (2018). *Worldwide Headphone Market Outlook April 2018*.

Gabrielsson, A. (1979). “Dimension analyses of perceived sound quality of sound-reproducing systems”. In: *Scandinavian Journal of Psychology* 20.1, pp. 159–169. ISSN: 0036-5564. DOI: `10.1111/j.1467-9450.1979.tb00697.x`.

Gamper, H. (2013). “Head-related transfer function interpolation in azimuth, elevation, and distance”. In: *The Journal of the Acoustical Society of America* 134.6, EL547. ISSN: 00014966. DOI: `10.1121/1.4828983`.

Garcia-Gomez, V. and J. J. Lopez (2018). “Binaural room impulse responses interpolation for multimedia real-time applications”. In: *144th AES Convention*. Milan, Italy: Audio Engineering Society, pp. 1–13. URL: `http://www.aes.org/e-lib/browse.cfm?elib=19479`.

Gardner, M. B. (1973). “Some monaural and binaural facets of median plane localization”. In: *The Journal of the Acoustical Society of America* 54.6, pp. 1489–1495. ISSN: 0001-4966. DOI: `10.1121/1.1914447`.

Gardner, W. G. (1995). "Efficient Convolution without Input-Output Delay". In: *Journal of the Audio Engineering Society* 43.3, pp. 127–136. URL: `http://www.aes.org/e-lib/browse.cfm?elib=7957`.

Gartner (2012). *Gartner Says Worldwide Smartphone Sales Soared in Fourth Quarter of 2011 With 47 Percent Growth*. URL: `https://www.gartner.com/newsroom/id/1924314` (visited on 12/08/2018).

— (2018). *Gartner Says Worldwide Sales of Smartphones Returned to Growth in First Quarter of 2018*. URL: `https://www.gartner.com/newsroom/id/3876865` (visited on 12/08/2018).

Gedemer, L. A. and T. Welti (2013). "Validation of the Binaural Room Scanning Method for Cinema Audio Research". In: *135th AES Convention*. New York, NY, USA: Audio Engineering Society.

Geier, M., J. Ahrens, and S. Spors (2008). "ASDF: Ein XML Format zur Beschreibung von virtuellen 3D-Audioszenen". In: *34th German Annual Conference on Acoustics (DAGA)*. Dresden, Germany. URL: `http://www.deutsche-telekom-laboratories.de/~sporssas/publications/2008/Geier_et_al_ASDF_DAGA2008.pdf`.

Geier, M., T. Hohn, and S. Spors (2012). "An Open-Source C ++ Framework for Multithreaded Realtime Multichannel Audio Applications". In: *Linux Audio Conference*.

Genovese, A. F. (2014). "Individualisation and Reverberation Factors in the Subjective Assessment of Plausibility in a Binaural Auditory Display". Masters Thesis. University of York.

Geronazzo, M., S. Spagnol, and F. Avanzini (2018). "Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.7, pp. 1247–1260. ISSN: 2329-9290. DOI: `10.1109/TASLP.2018.2821846`.

Gerzon, M. A. (1973). "Periphony: With-Height Sound Reproduction". In: *Journal of the Audio Engineering Society* 21.1, pp. 2–10. URL: `http://www.aes.org/e-lib/browse.cfm?elib=2012`.

— (1975). "The Design of Precisely Coincident Microphone Arrays for Stereo and Surround Sound". In: *AES 50th Convention*. London, UK: Audio Engineering Society.

— (1992a). "General Metatheory of Auditory Localisation". In: *92nd AES Convention*. Vienna, Austria: Audio Engineering Society.

— (1992b). "Panpot Laws for Multispeaker Stereo". In: *AES 92nd Convention*. Vienna, Austria: Audio Engineering Society.

— (1992c). "Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound". In: *AES 93rd Convention*. San Francisco, CA, USA: Audio Engineering Society.

Gerzon, M. A. and G. J. Barton (1992). "Ambisonic Decoders for HDTV". In: *92nd AES Convention*. Vienna, Austria: Audio Engineering Society.

Giacalone, D., W. L. P. Bredie, and M. B. Frøst (2013). ""All-In-One Test" (AI1): A rapid and easily applicable approach to consumer product testing". In: *Food Quality and Preference* 27, pp. 108–119. DOI: `10.1016/j.foodqual.2012.09.011`.

Giacalone, D., M. Nitkiewicz, et al. (2017). "Sensory profiling of high-end loudspeakers using rapid methods - Part 2: Projective mapping with expert and naïve assessors". In: *142nd AES Convention*. Berlin, Germany: Audio Engineering Society.

Gibson, O. (2012). *Mo Farah powers to a sensational 10,000m Olympic gold for Britain | Sport | The Guardian*. URL: `https://www.theguardian.com/sport/2012/aug/04/mo-farah-team-gb-olympic-medals` (visited on 12/09/2018).

Goodwin, M. M. and J.-M. Jot (2006). "A frequency-domain framework for spatial audio coding based on universal spatial cues". In: *120th AES Convention*. Paris, France: Audio Engineering Society. ISBN: 9781604235975. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13555`.

Goodwin, M. M. and J.-M. Jot (2007). “Binaural 3-D audio rendering based on spatial audio scene coding”. In: *123rd AES Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=14335.

Google (2018a). *Daydream View*. URL: https://vr.google.com/daydream/smartphonevr/.

— (2018b). *Resonance Audio*. URL: https://developers.google.com/resonance-audio/ (visited on 07/28/2018).

— (2018c). *Spatial Audio Resources*. URL: https://github.com/google/spatial-media/tree/master/spatial-audio (visited on 10/28/2018).

Goupell, M. J. and W. M. Hartmann (2007). “Interaural fluctuations and the detection of interaural incoherence. III. Narrowband experiments and binaural models”. In: *The Journal of the Acoustical Society of America* 122.2, pp. 1029–1045. ISSN: 0001-4966. DOI: 10.1121/1.2734489.

Gower, J. C. (1975). “Generalized procrustes analysis”. In: *Psychometrika* 40.1, pp. 33–51. ISSN: 0033-3123. DOI: 10.1007/BF02291478.

Gray, H. (1918). *Anatomy of the Human Body*. Ed. by W. H. Lewis. 20th ed. Lea and Febiger. URL: https://www.bartleby.com/107/.

Green, D. M. and J. A. Swets (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley. ISBN: 0932146236.

Greenacre, M. (2007). *Correspondence Analysis in Practice*. Third edit. Chapman and Hall/CRC.

Grey, J. M. (1977). “Multidimensional perceptual scaling of musical timbres”. In: *The Journal of the Acoustical Society of America* 61.5, pp. 1270–1277. ISSN: 0001-4966. DOI: 10.1121/1.381428.

Guastavino, C. and B. F. G. Katz (2004). “Perceptual evaluation of multi-dimensional spatial audio reproduction”. In: *The Journal of the Acoustical Society of America* 116.2, p. 1105. ISSN: 00014966. DOI: 10.1121/1.1763973.

Guezenoc, C. and R. Séguier (2018). “HRTF Individualization: A Survey”. In: *AES 145th Convention*. New Paltz, NY, USA: Audio Engineering Society.

Guillon, P., R. Nicol, and L. S. R. Simon (2008). “Head-Related Transfer Functions Reconstruction from Sparse Measurements Considering a Priori Knowledge from Database Analysis: A Pattern Recognition Approach”. In: *125th AES Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=14761.

Guru, A., W. L. Martens, and D. Lee (2010). “Effects of individualised headphone correction on front/back discrimination of virtual sound sources displayed using individualised head related transfer functions”. In: *AES 40th International Conference: Spatial Sound*. Tokyo, Japan: Audio Engineering Society.

Hafter, E. R. and J. De Maio (1975). “Difference thresholds for interaural delay”. In: *The Journal of the Acoustical Society of America* 57.1, pp. 181–187. ISSN: 0001-4966. DOI: 10.1121/1.380412.

Hamasaki, K., K. Hiyama, and R. Okumura (2005). “The 22.2 Multichannel Sound System and Its Application”. In: *118th AES Convention*. Barcelona, Spain: Audio Engineering Society.

Hammer, K. and W. Snow (1932). “Binaural Transmission System at Academy of Music in Philadelphia”. In: *Memorandum MM-3950*.

Hammershøi, D. and H. Møller (1996). “Sound transmission to and within the human ear canal”. In: *The Journal of the Acoustical Society of America* 100.1, pp. 408–427. ISSN: 0001-4966. DOI: 10.1121/1.415856.

— (2005). “Binaural Technique — Basic Methods for Recording, Synthesis, and Reproduction”. In: *Communication Acoustics*. Ed. by J. Blauert. Berlin/Heidelberg: Springer-Verlag. Chap. 9, pp. 223–254. ISBN: 9783540221623. DOI: 10.1007/3-540-27437-5_9.

Hardin, R. H. and N. J. A. Sloane (1996). “McLaren’s improved snub cube and other new spherical designs in three dimensions”. In: *Discrete & Computational Geometry* 15.4, pp. 429–441. ISSN: 0179-5376. DOI: 10.1007/BF02711518.

Härmä, A., J. Jakka, et al. (2004). “Augmented Reality Audio for Mobile and Wearable Appliances”. In: *Journal of the Audio Engineering Society* 52.6, pp. 618–639. URL: http://www.aes.org/e-lib/browse.cfm?elib=13010.

Härmä, A., R. van Dinther, et al. (2012). “Personalization of headphone spatialization based on the relative localization error in an auditory gaming interface”. In: *AES 132nd Convention*. Budapest, Hungary: Audio Engineering Society.

Hartigan, J. A. and M. A. Wong (1979). “Algorithm AS 136: A K-Means Clustering Algorithm”. In: *Applied Statistics* 28.1, p. 100. ISSN: 00359254. DOI: 10.2307/2346830.

Hartmann, W. M. (1983). “Localization of sound in rooms”. In: *The Journal of the Acoustical Society of America* 74.5, pp. 1380–1391. ISSN: 0001-4966. DOI: 10.1121/1.390163.

Hartmann, W. M. and A. Wittenberg (1996). “On the externalization of sound images”. In: *The Journal of the Acoustical Society of America* 99.6, pp. 3678–3688. ISSN: 0001-4966. DOI: 10.1121/1.414965.

Haustein, B. G. and W. Schirmer (1970). “Messeinrichtung zur Untersuchung des Richtungslokalisationsvermögens [A measuring apparatus for the investigation of the faculty of directional localization].” In: *Hocfrequenztech. u. Elektroakustik* 79, pp. 79–101.

HbbTV Association (2018). *HbbTV 2.0.2 Specification*. URL: https://www.hbbtv.org/resource-library/.

He, J. et al. (2018). “Fast Continuous Measurement of HRTFs with Unconstrained Head Movements for 3D Audio”. In: *Journal of the Audio Engineering Society* 66.11, pp. 884–900. ISSN: 15494950. DOI: 10.17743/jaes.2018.0050.

Hebrank, J. and D. Wright (1974). “Spectral cues used in the localization of sound sources on the median plane”. en. In: *The Journal of the Acoustical Society of America* 56.6, pp. 1829–1834. ISSN: 0001-4966. DOI: 10.1121/1.1903520.

Heinz, R. (1993). “Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail”. In: *Applied Acoustics* 38.2-4, pp. 145–159. ISSN: 0003682X. DOI: 10.1016/0003-682X(93)90048-B.

Heller, A. J., R. Lee, and E. M. Benjamin (2008). “Is My Decoder Ambisonic ?” In: *125th AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

Hendrickx, E. et al. (2017). “Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis”. In: *The Journal of the Acoustical Society of America* 141.3, pp. 2011–2023. ISSN: 0001-4966. DOI: 10.1121/1.4978612.

Hendrix, C. M. and W. Barfield (1996). “The Sense of Presence within Auditory Virtual Environments”. In: *Presence: Teleoperators and Virtual Environments* 5.3, pp. 290–301. ISSN: 1054-7460. DOI: 10.1162/pres.1996.5.3.290.

Herre, J., J. Hilpert, et al. (2015). “MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio”. In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 770–779. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2015.2411578.

Herre, J., H. Purnhagen, et al. (2012). “MPEG Spatial Audio Object Coding — The ISO / MPEG Standard for Efficient Coding of Interactive Audio Scenes”. In: *Journal of the Audio Engineering Society* 60.9, pp. 655–673. URL: http://www.aes.org/e-lib/browse.cfm?elib=16371.

Hertz, B. F. (1981). “100 Years with Stereo: The Beginning”. In: *Journal of the Audio Engineering Society* 29.5, pp. 368–370.

Hess, W. (2012). “Head-Tracking Techniques for Virtual Acoustics Applications”. In: *Proceedings of 133rd AES Convention*. Vol. 133. San Francisco, CA, USA: Audio Engineering Society.

Hess, W. and T. Mayer (2012). *Patent EP2428813A1 – Head Tracking System with Improved Detection of Head Rotation*.

Hicks, L., S. Moulin, and S. Bech (2018). “Sensory Profiling of High-End Loudspeakers Using Rapid Methods—Part 3: Check-All-That-Apply with Naïve Assessors”. In: *Journal of the Audio Engineering Society* 66.5, pp. 329–342. ISSN: 15494950. DOI: 10.17743/jaes.2018.0015.

Hiipakka, M., T. Kinnari, and V. Pulkki (2012). “Estimating head-related transfer functions of human subjects from pressure–velocity measurements”. In: *The Journal of the Acoustical Society of America* 131.5, pp. 4051–4061. ISSN: 0001-4966. DOI: 10.1121/1.3699230.

Hinde, S. J. (2017). “Attention While Watching Movies”. PhD thesis. University of Bristol. DOI: 10.13140/RG.2.2.11800.80648.

Hirahara, T. et al. (2010). “Head movement during head-related transfer function measurements”. In: *Acoustical Science and Technology* 31.2, pp. 165–171. ISSN: 1347-5177. DOI: 10.1250/ast.31.165.

Hobden, L. J. and A. I. Tew (2015). “Investigating head-related transfer function smoothing using a sagittal-plane localization model”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY, USA: IEEE, pp. 1–5. ISBN: 978-1-4799-7450-4. DOI: 10.1109/WASPAA.2015.7336955.

Hoffman, J., W. Hess, and M. Mittmann (2009). *Patent US20090147993A1 – Head-tracking system*.

Hoffman, L. and M. J. Rovine (2007). “Multilevel models for the experimental psychologist: Foundations and illustrative examples”. In: *Behavior Research Methods* 39.1, pp. 101–117. ISSN: 1554-351X. DOI: 10.3758/BF03192848.

Hofman, P. M., J. G. Van Riswick, and A. J. Van Opstal (1998). “Relearning sound localization with new ears.” In: *Nature neuroscience* 1.5, pp. 417–21. ISSN: 1097-6256. DOI: 10.1038/1633.

Hom, R. C.-M., V. R. Algazi, and R. O. Duda (2006). “High-Frequency Interpolation for Motion-Tracked Binaural Sound”. In: *Proceedings of the 121st AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

Honeyborne, J. (2018). *Blue Planet II – Creating an underwater soundscape*. URL: http://www.bbc.co.uk/programmes/articles/34j4WPCGZnFJvj2Xq9NGK7M/creating-an-underwater-soundscape (visited on 12/09/2018).

Horbach, U. et al. (1999). “Design and applications of a data-based auralization system for surround sound”. In: *AES 106th Convention*. Munich, Germany: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8204.

Hu, H. et al. (2008). “HRTF personalization based on artificial neural network in individual virtual auditory space”. In: *Applied Acoustics* 69, pp. 163–172. DOI: 10.1016/j.apacoust.2007.05.007.

Hughes, R. J. et al. (2016). “The Room-in-Room Effect and its Influence on Perceived Room Size in Spatial Audio Reproduction”. In: *AES 141st Convention*. Los Angeles, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=18425.

Huopaniemi, J., N. Zacharov, and M. Karjalainen (1999). “Objective and Subjective Evaluation of Head Related Transfer Function Filter Design”. In: *Journal of the Audio Engineering Society* 47.4, pp. 218–239. URL: http://www.aes.org/e-lib/browse.cfm?elib=12109.

Husson, F., J. Josse, and J. Pagès (2010). “Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data?” In: *Technical Report of the Applied Mathematics Department (Agrocampus)* September, pp. 1–17. URL: http://factominer.free.fr/more/HCPC_husson_josse.pdf.

Husson, F., S. Lê, and M. Cadoret (2017). *{SensoMineR}: Sensory Data Analysis*. URL: `https://cran.r-project.org/package=SensoMineR`.

Husson, F., S. Lê, and J. Pagès (2010). *Exploratory Multivariate Analysis by Example Using R*. 1st ed. CRC Press. ISBN: 9780429190032.

Huttunen, T. and A. Vanne (2017). "End-to-end process for HRTF personalization". In: *AES 142nd Convention*. Berlin, Germany: Audio Engineering Society.

Huttunen, T., A. Vanne, et al. (2014). "Rapid generation of personalized HRTFs". In: *AES 55th International Conference: Spatial Audio*. Helsinki, Finland: Audio Engineering Society. ISBN: 9780937803981.

IEC 60268-7 (2010). *Sound system equipment - Part 7: Headphones and earphones*.

IEC TR 60959 (1990). *Provisional head and torso simulator for acoustic measurements on air condution hearing aids*.

IEC TS 60318 (2017). *Electroacoustics - Simulators of human head and ear - Part 7: Head and torso simulator for the measurement of air-conduction hearing aids*.

Iljazovic, A. et al. (2012). "The Influence of 2D and 3D Video Playback on the Perceived Quality of Spatial Audio Rendering for Headphones". In: *Proceedings of 133rd AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

InterSense (2013). *InertiaCube 4*. URL: `http://www.intersense.com/pages/18/234/` (visited on 11/09/2013).

ISO 1996-1:2016 (2016). *Acoustics – Description, measurement and assessment of environmental noise – Part 1: Basic quantities and assessment procedures*.

ISO 3382-1:2009 (2009). *Acoustics – Measurement of room acoustic parameters –- part 1: Performance spaces*. Geneva.

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

ISO 9000 (2005). *Quality management systems – Fundamentals and vocabulary*.

ISO/IEC 23008-3:2015 (2015). *Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio*.

ITU-R (1997). *Recommendation BS.1286 - Methods for the subjective assessment of audio systems with accompanying picture*.

— (2001). *Recommendation BS.1387-1 – Method for objective measurements of perceived audio quality*.

— (2002). *Recommendation BS.708 - Determination of the electro-acoustical properties of studio monitor headphones*.

— (2003a). *Recommendation BS.1284 - General methods for the subjective assessment of sound quality*.

— (2003b). *Recommendation BS.1534-1 – Method for the subjective assessment of intermediate quality level of coding systems (MUSHRA)*.

— (2011). *Recommendation BS.1770-3 - Algorithms to measure audio programme loudness and true-peak audio level*.

— (2012a). *Recommendation BS. 775-3 - Multichannel stereophonic sound system with and without accompanying picture*.

— (2012b). *Recommendation BT.500-13 - Methodology for the subjective assessment of the quality of television pictures*. Tech. rep.

— (2014a). *Recommendation BS.2051-0 - Advanced sound system for programme production*.

— (2014b). *Report ITU-R BS.2300-0 - Methods for Assessor Screening*.

— (2015a). *Question 139/6 – Methods for rendering of advanced audio formats*.

ITU-R (2015b). *Recommendation BS.1534-3 - Method for the subjective assessment of intermediate quality level of audio systems*.

— (2015c). *Recommendation BS.2076 - Audio Definition Model*.

— (2015d). *Recommendation BS.2088 - Long-form file format for the international exchange of audio programme materials with metadata*.

— (2015e). *Recommendation ITU-R BS.1116-3 - Methods for the subjective assessment of small impairments in audio systems*.

— (2017a). *Recommendation BS.2051-1 - Advanced sound system for programme production*.

— (2017b). *Recommendation BS.2076-1 - Audio Definition Model*.

— (2017c). *Report ITU-R BS.2399-0 - Methods for selecting and describing attributes and terms, in the preparation of subjective tests*.

— (2018). *Recommendation BS.2051-2 – Advanced sound system for programme production*.

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

— (1998). *P.911 - Subjective audiovisual quality assessment methods for multimedia applications*.

— (2001). *Recommendation P.862 – Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*.

Ivanic, J. (1996). “Rotation matrices for real spherical harmonics. Direct determination by recursion”. In: *The Journal of Physical Chemistry* 100.15, pp. 6342–6347. ISSN: 0022-3654. DOI: 10.1021/jp953350u.

Jack, C. E. and W. R. Thurlow (1973). “Effects of Degree of Visual Association and Angle of Displacement on the “Ventriloquism” Effect”. In: *Perceptual and Motor Skills* 37.3, pp. 967–979. DOI: 10.1177/003151257303700360.

Jack, F. R. and J. Piggott (1991). “Free choice profiling in consumer research”. In: *Food Quality and Preference* 3.3, pp. 129–134. ISSN: 09503293. DOI: 10.1016/0950-3293(91)90048-J.

Jaeger, S. R., M. K. Beresford, et al. (2015). “Check-all-that-apply (CATA) questions for sensory product characterization by consumers: Investigations into the number of terms used in CATA questions”. In: *Food Quality and Preference* 42, pp. 154–164. ISSN: 09503293. DOI: 10.1016/j.foodqual.2015.02.003.

Jaeger, S. R., S. L. Chheang, et al. (2013). “Check-all-that-apply (CATA) responses elicited by consumers: Within-assessor reproducibility and stability of sensory product characterizations”. In: *Food Quality and Preference* 30.1, pp. 56–67. ISSN: 09503293. DOI: 10.1016/j.foodqual.2013.04.009.

Jaeger, T. F. (2008). “Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models”. In: *Journal of Memory and Language* 59.4, pp. 434–446. ISSN: 0749596X. DOI: 10.1016/j.jml.2007.11.007.

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. URL: http://www.ingentaconnect.com/content/dav/aaua/2004/00000090/00000006/art00002.

— (2005). *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer. ISBN: 9783540288602.

Jin, C. T., P. Leong, et al. (2003). “Enabling individualized virtual auditory space using morphological measurements”. In: *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pp. 235–238. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.451.6779%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf.

Jin, C. T., A. I. Tew, et al. (2013). “Creating the Sydney York Morphological and Acoustic Recordings of Ears Database”. English. In: *IEEE Transactions on Multimedia* PP.99, pp. 1–1. ISSN: 1520-9210. DOI: 10.1109/TMM.2013.2282134.

Jot, J.-M. (1992). “An analysis/synthesis approach to real-time artificial reverberation”. In: *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 221–224 vol.2. ISBN: 0-7803-0532-9. DOI: 10.1109/ICASSP.1992.226080.

— (1998). “Approaches to binaural synthesis”. In: *105th AES Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8319.

— (1999). “Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces”. In: *Multimedia Systems* 7.1, pp. 55–69. ISSN: 0942-4962. DOI: 10.1007/s005300050111.

Jot, J.-M., L. Cerveau, and O. Warusfel (1997). “Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model”. In: *103rd AES Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/inst/browse.cfm?elib=7150%7B%5C&%7Drndx=570714.

Jot, J.-M. and A. Chaigne (1991). “Digital delay networks for designing artifical reverberators”. In: *Proceedings of the 90th AES Convention*. Paris, France: Audio Engineering Society.

Jot, J.-M., J.-P. Jullien, and O. Warusfel (1998). *Patent US5812674 – Method to simulate the acoustical quality of a room and associated audio-digital processor*.

Jot, J.-M., V. Larcher, and O. Warusfel (1995). “Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony”. In: *Proceedings of the 98th AES Convention*. Paris, France: Audio Engineering Society.

Jot, J.-M. and K.-S. Lee (2016). “Augmented Reality Headphone Environment Rendering”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Los Angeles, CA, USA: Audio Engineering Society.

Jot, J.-M. and D. Noh (2017). “Efficient Structures for Virtual Multi-Channel Immersive Audio Rendering”. In: *143rd AES Convention*. New York, NY, USA: Audio Engineering Society.

Jot, J.-M., M. Walsh, and A. Philp (2006). “Binaural Simulation of Complex Acoustic Scenes for Interactive Audio”. In: *AES 121st Convention*. San Francisco, CA, USA: Audio Engineering Society, pp. 1–20.

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., V. K. Malamal Vadakital, and M. M. Hannuksela (2008). “Acceptance Threshold: A Bidimensional Research Method for User-Oriented Quality Evaluation Studies”. In: *International Journal of Digital Multimedia Broadcasting* 2008, pp. 1–20. ISSN: 1687-7578. DOI: 10.1155/2008/712380.

Kan, A., C. Jin, and A. van Schaik (2009). “A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function”. In: *The Journal of the Acoustical Society of America* 125.4, pp. 2233–2242. ISSN: 0001-4966. DOI: 10.1121/1.3081395.

Kaplanis, N. et al. (2017). “Perceptual aspects of reproduced sound in car cabin acoustics”. In: *The Journal of the Acoustical Society of America* 141.3, pp. 1459–1469. ISSN: 0001-4966. DOI: 10.1121/1.4976816.

Katz, B. F. G. (2001a). “Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation”. In: *The Journal of the Acoustical Society of America* 110.5, pp. 2440–2448. ISSN: 0001-4966. DOI: 10.1121/1.1412440.

— (2001b). “Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements”. In: *The Journal of the Acoustical Society of America* 110.5, pp. 2449–2455. ISSN: 0001-4966. DOI: 10.1121/1.1412441.

Katz, B. F. G. and M. Noisternig (2014). “A comparative study of interaural time delay estimation methods”. In: *The Journal of the Acoustical Society of America* 135.6, pp. 3530–3540. ISSN: 0001-4966. DOI: 10.1121/1.4875714.

Katz, B. F. G. and G. Parseihian (2012). “Perceptually based head-related transfer function database optimization”. In: *The Journal of the Acoustical Society of America* 131.2, EL99–EL105. ISSN: 0001-4966. DOI: 10.1121/1.3672641.

Katz, B. F. G., D. Poirier-Quinot, et al. (2018). “Objective and perceptive evaluations of high-resolution room acoustic simulations and auralizations”. In: *Proceedings of Euronoise*. Crete.

Kearney, G. and T. Doyle (2015a). “A HRTF Database for Virtual Loudspeaker Rendering”. In: *139th AES Convention*. New York, NY, USA: Audio Engineering Society.

— (2015b). “Height Perception in Ambisonic Based Binaural Decoding”. In: *AES 139th Convention*. New York, NY, USA: Audio Engineering Society, pp. 1–10.

Kearney, G., M. Gorzel, et al. (2012). “Distance Perception in Interactive Virtual Acoustic Environments using First and Higher Order Ambisonic Sound Fields”. In: *Acta Acustica united with Acustica* 98.1, pp. 61–71. DOI: 10.3813/AAA.918492.

Kearney, G., X. Liu, et al. (2015). “Auditory Distance Perception with Static and Dynamic Binaural Rendering”. In: *AES 57th International Conference*. Hollywood, CA, USA: Audio Engineering Society.

Kearney, G., C. Masterson, et al. (2009). “Approximation of binaural room impulse responses”. In: *IET Irish Signals and Systems Conference (ISSC 2009)*. Dublin, Ireland, pp. 36–36. ISBN: 978 1 84919 213 2. DOI: 10.1049/cp.2009.1713.

Kendall, G. S. (1995). “The Decorrelation of Audio Signals and Its Impact on Spatial Imagery”. In: *Computer Music Journal* 19.4, pp. 71–87. ISSN: 01489267. DOI: 10.2307/3680992.

Kendall, G. S. and W. L. Martens (1984). “Simulating the cues of spatial hearing in natural environments”. In: *Proceedings of the International Computer Music Conference*. September. Paris, France. URL: http://www.garykendall.net/papers/SimulatingTheCues1984.pdf.

Kim, C., R. Mason, and T. Brookes (2013). “Head Movements Made by Listeners in Experimental and Real-Life Listening Activities”. In: *Journal of the Audio Engineering Society* 61.6, pp. 425–438.

Kim, S.-M. and W. Choi (2005). “On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach”. In: *The Journal of the Acoustical Society of America* 117.6, pp. 3657–3665. ISSN: 0001-4966. DOI: 10.1121/1.1921548.

Kirby, D. G., N. A. F. Cutmore, and J. A. Fletcher (1996). “Program Origination of Five-Channel Surround Sound”. In: *Journal of the Audio Engineering Society* 46.4. URL: http://www.aes.org/e-lib/browse.cfm?elib=12153.

Kirkeby, O. and P. A. Nelson (1999). “Digital Filter Design for Inversion Problems in Sound Reproduction”. In: *Journal of the Audio Engineering Society* 47.7/8, pp. 583–595.

Kirkeby, O., P. A. Nelson, et al. (1998). “Fast deconvolution of multichannel systems using regularization”. In: *IEEE Transactions on Speech and Audio Processing* 6.2, pp. 189–194. ISSN: 10636676. DOI: 10.1109/89.661479.

Kistler, D. J. and F. L. Wightman (1992). “A model of head⌧related transfer functions based on principal components analysis and minimum⌧phase reconstruction”. In: *The Journal of the Acoustical Society of America* 91.3, pp. 1637–1647. ISSN: 0001-4966. DOI: 10.1121/1.402444.

Koenig, W. (1950). “Subjective Effects in Binaural Hearing”. In: *The Journal of the Acoustical Society of America* 22.1, pp. 61–62. ISSN: 0001-4966. DOI: 10.1121/1.1906578.

Kohlrausch, A. et al. (2014). “An Introduction to Binaural Processing”. In: *The Technology of Binaural Listening. Modern Acoustics and Signal Processing.* Ed. by J. Blauert. 1st. Springer, Berlin, Heidelberg. Chap. 1. ISBN: 978-3-642-37761-7. DOI: https://doi.org/10.1007/978-3-642-37762-4_1.

Koivuniemi, K. and N. Zacharov (2001). "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training". In: *AES 111th Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=9815.

Kolarik, A. J. et al. (2016). "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss". In: *Attention, Perception, & Psychophysics* 78.2. ISSN: 1943-3921. DOI: 10.3758/s13414-015-1015-1.

König, W. and W. Sussmann (1955). "Zum Richtungshören in der Median-sagittal-ebene [On directional hearing in the median-sagittal plane]". In: *Archiv für Ohren-, Nasen-, und Kehlkopfheilkunde* 167.2-6, pp. 303–307. DOI: 10.1007/BF02107754.

Kopčo, N., B. G. Shinn-cunningham, and N. Kopc (2011). "Effect of stimulus spectrum on distance perception for nearby sources". In: *The Journal of the Acoustical Society of America* 130.1530. DOI: 10.1121/1.3613705.

Kreuzer, W., P. Majdak, and Z. Chen (2009). "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range." In: *The Journal of the Acoustical Society of America* 126.3, pp. 1280–90. ISSN: 1520-8524. DOI: 10.1121/1.3177264.

Krosnick, J. A. (1999). "Survey Research". In: *Annual Review of Psychology* 50.1, pp. 537–567. ISSN: 0066-4308. DOI: 10.1146/annurev.psych.50.1.537.

Kuchinsky, S. E. et al. (2013). "Pupil size varies with word listening and response selection difficulty in older adults with hearing loss Stefanie". In: *Psychophysiology* 50.1, pp. 23–34. DOI: 10.1111/j.1469-8986.2012.01477.x.Pupil.

Kuhn, G. F. (1977). "Model for the interaural time differences in the azimuthal plane". In: *The Journal of the Acoustical Society of America* 62.1, pp. 157–167. ISSN: 00014966. DOI: 10.1121/1.381498.

Kulkarni, A. and H. S. Colburn (1998). "Role of spectral detail in sound-source localization." In: *Nature* 396.6713, pp. 747–9. ISSN: 0028-0836. DOI: 10.1038/25526.

— (2000). "Variability in the characterization of the headphone transfer-function". In: *The Journal of the Acoustical Society of America* 107.2, pp. 1071–1074. ISSN: 0001-4966. DOI: 10.1121/1.428571.

Kulkarni, A., S. K. Isabelle, and H. S. Colburn (1999). "Sensitivity of human subjects to head-related transfer-function phase spectra". In: *The Journal of the Acoustical Society of America* 105.5, pp. 2821–2840. ISSN: 0001-4966. DOI: 10.1121/1.426898.

Kürer, R., G. Plenge, and H. Wilkens (1969). "Correct Spatial Sound Perception Rendered by a Special 2-Channel Recording Method". In: *AES 37th Convention*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=1346.

Laakso, T. et al. (1996). "Splitting the unit delay [FIR/all pass filters design]". In: *IEEE Signal Processing Magazine* 13.1, pp. 30–60. ISSN: 10535888. DOI: 10.1109/79.482137.

Lacouture-Parodi, Y. and E. A. P. Habets (2012). "Crosstalk Cancellation System Using a Head Tracker Based on Interaural Time Differences". In: *IEEE International Workshop on Acoustic Signal Enhancement 2012*. September. URL: https://ieeexplore.ieee.org/document/6309392.

Laitinen, M.-V., T. Pihlajamäki, et al. (2012). "Influence of Resolution of Head Tracking in Synthesis of Binaural Audio". In: *AES 132nd Convention*. Budapest, Hungary: Audio Engineering Society.

Laitinen, M.-V. and V. Pulkki (2009). "Binaural reproduction for Directional Audio Coding". In: *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 337–340. ISBN: 978-1-4244-3678-1. DOI: 10.1109/ASPAA.2009.5346545.

Langendijk, E. H. A. and A. W. Bronkhorst (2000). "Fidelity of three-dimensional-sound reproduction using a virtual auditory display". In: *The Journal of the Acoustical Society of America* 107.1, pp. 528–537. ISSN: 0001-4966. DOI: 10.1121/1.428321.

Larcher, V., J.-M. Jot, J. Guyard, et al. (2000). "Study and Comparison of Efficient Methods for 3D Audio Spatialization Based on Linear Decomposition of HRTF Data". In: *Proceedings of the 108th Convention of the AES*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=9241.

Larcher, V., J.-M. Jot, and G. Vandernoot (1998). "Equalization Methods in Binaural Technology". In: *105th AES Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8322.

Larsen, E. et al. (2008). "On the minimum audible difference in direct-to-reverberant energy ratio". In: *The Journal of the Acoustical Society of America* 124.1, pp. 450–461. ISSN: 0001-4966. DOI: 10.1121/1.2936368.

LaScalza, S., J. Arico, and R. Hughes (2003). "Effect of metal and sampling rate on accuracy of Flock of Birds electromagnetic tracking system". In: *Journal of Biomechanics* 36.1, pp. 141–144. ISSN: 00219290. DOI: 10.1016/S0021-9290(02)00322-6.

Lawless, H. T. (1999). "Descriptive analysis of complex odors: reality, model or illusion?" In: *Food Quality and Preference* 10.4-5, pp. 325–332. ISSN: 09503293. DOI: 10.1016/S0950-3293(98)00052-4.

Lawless, H. T. and H. Heymann (2010). *Sensory Evaluation of Food*. 2nd. Springer. ISBN: 9781441964878. DOI: 10.1007/978-1-4419-6488-5.

Lawless, L. J. and G. V. Civille (2013). "Developing Lexicons: A review". In: *Journal of Sensory Studies* 28.4, pp. 270–281. ISSN: 08878250. DOI: 10.1111/joss.12050.

Le Bagousse, S., M. Paquier, and C. Colomes (2010). "Families of Sound Attributes for Assessment of Spatial Audio". In: *129th AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

— (2012). "Assessment of spatial audio quality based on sound attributes". In: *Acoustics 2012*. April. Nantes, pp. 867–871.

Le Bagousse, S., M. Paquier, C. Colomes, and S. Moulin (2011). "Sound Quality Evaluation based on Attributes - Application to Binaural Contents". In: *Proceedings of 131st AES Convention*. New York, NY, USA: Audio Engineering Society.

Le Callet, P., S. Möller, and A. Perkis, eds. (2012). *Qualinet White Paper on Definitions of Quality of Experience*.

Lê, S., J. Josse, and F. Husson (2008). "{FactoMineR}: A Package for Multivariate Analysis". In: *Journal of Statistical Software* 25.1, pp. 1–18. DOI: 10.18637/jss.v025.i01.

Lee, J. C. (2008). "Hacking the Nintendo Wii Remote". English. In: *IEEE Pervasive Computing* 7.3, pp. 39–45. ISSN: 1536-1268. DOI: 10.1109/MPRV.2008.53.

Lee, Y., C. Findlay, and J. F. Meullenet (2013). "Experimental consideration for the use of check-all-that-apply questions to describe the sensory properties of orange juices". In: *International Journal of Food Science and Technology* 48.1, pp. 215–219. ISSN: 09505423. DOI: 10.1111/j.1365-2621.2012.03165.x.

Lentz, T. et al. (2007). "Virtual Reality System with Integrated Sound Field Simulation and Reproduction". In: *EURASIP Journal on Advances in Signal Processing* 2007.1, p. 070540. ISSN: 1687-6180. DOI: 10.1155/2007/70540.

Letowski, T. (1989). "Sound Quality Assessment: Concepts and Criteria". In: *87th AES Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=5869.

Letowski, T. and S. Letowski (2011). "Localization Error: Accuracy and Precision of Auditory Localization". In: *Advances in Sound Localisation*. Ed. by P. Strumillo. Chap. 4. ISBN: 978-953-307-224-1. DOI: 10.5772/597.

Leventhal, L. (1986). "Type 1 and Type 2 Errors in the Statistical Analysis of Listening Tests". In: *Journal of the Audio Engineering Society* 34.6, pp. 437–453. URL: http://www.aes.org/e-lib/browse.cfm?elib=5265.

Li, L. and Q. Huang (2013). "HRTF personalization modeling based on RBF neural network". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3707–3710. ISBN: 9781479903566. DOI: 10.1109/ICASSP.2013.6638350.

Lindau, A. (2009). “The Perception of System Latency in Dynamic Binaural Synthesis”. In: *Proceedings of NAG/DAGA 2009*, pp. 1063–1066.

— (2015). *Spatial Audio Quality Inventory (SAQI). Test Manual.* DOI: 10.14279/depositonce-1.2.

Lindau, A. and F. Brinkmann (2012). “Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings”. In: *Journal of the Audio Engineering Society* 60.1/2, pp. 54–62. URL: http://www.aes.org/e-lib/browse.cfm?elib=16166.

Lindau, A., F. Brinkmann, and S. Weinzierl (2014). “Sensory Profiling of Individual and Non-individual Dynamic Binaural Synthesis Using the Spatial Audio Quality Inventory”. In: *Forum Acusticum*. Kraków.

Lindau, A., V. Erbes, et al. (2014). “A Spatial Audio Quality Inventory (SAQI)”. In: *Acta Acustica united with Acustica* 100.5, pp. 984–994. ISSN: 16101928. DOI: 10.3813/AAA.918778.

Lindau, A., J. Estrella, and S. Weinzierl (2010). “Individualisation of dynamic binaural synthesis by real time manipulation of the ITD”. In: *128th AES Convention*. London, UK: Audio Engineering Society.

Lindau, A., L. Kosanke, and S. Weinzierl (2012). “Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses”. In: *Journal of the Audio Engineering Society* 60.11, pp. 887–898. URL: http://www.aes.org/e-lib/browse.cfm?elib=16633.

Lindau, A. and S. Roos (2010). “Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings”. In: *VDT International Convention*. November, pp. 680–701. ISBN: 978-3-9812830-1-3.

Lindau, A. and S. Weinzierl (2006). “FABIAN - An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom”. In: *VDT International Convention*. November, pp. 1–5.

— (2009). “On The Spatial Resolution of Virtual Acoustic Environments for Head Movements in Horizontal, Vertical and Lateral Direction”. In: *EAA Symposium on Auralization*, pp. 13–18.

— (2012). “Assessing the Plausibility of Virtual Acoustic Environments”. In: *Acta Acustica united with Acustica* 98.5, pp. 804–810. ISSN: 16101928. DOI: 10.3813/AAA.918562.

Linkwitz, S. (1976). “Active Crossover Networks for Noncoincident Drivers”. In: *Journal of the Audio Engineering Society* 24.1, pp. 2–8. URL: http://www.aes.org/e-lib/browse.cfm?elib=2649..

Litovsky, R. Y. et al. (1999). “The precedence effect”. In: *The Journal of the Acoustical Society of America* 106.4, pp. 1633–1654. ISSN: 0001-4966. DOI: 10.1121/1.427914.

Logitech (2013). *3D Mouse & Head Tracker*. URL: http://www.vrdepot.com/manual-tracker.pdf (visited on 11/09/2013).

Lokki, T. (2014). “Tasting music like wine: Sensory evaluation of concert halls”. In: *Physics Today* 67.1, pp. 27–32. ISSN: 00319228. DOI: 10.1063/PT.3.2242.

Lokki, T., J. Patynen, et al. (2011). “Concert hall acoustics assessment with individually elicited attributes.” In: *The Journal of the Acoustical Society of America* 130.2, pp. 835–49. ISSN: 1520-8524. DOI: 10.1121/1.3607422.

Lokki, T., J. Pätynen, et al. (2012). “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles.” In: *The Journal of the Acoustical Society of America* 132.5, pp. 3148–61. ISSN: 1520-8524. DOI: 10.1121/1.4756826.

Loomis, J. M., C. Hebert, and J. G. Cicinelli (1990). “Active localization of virtual sounds”. In: *The Journal of the Acoustical Society of America* 88.4, pp. 1757–1764. ISSN: 0001-4966. DOI: 10.1121/1.400250.

Lopez-Poveda, E. A. and R. Meddis (1996). “A physical model of sound diffraction and reflections in the human concha”. In: *The Journal of the Acoustical Society of America* 100.5, pp. 3248–3259. ISSN: 0001-4966. DOI: 10.1121/1.417208.

Lord Rayleigh (1907). “On our perception of sound direction”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232. ISSN: 1941-5982. DOI: 10.1080/14786440709463595.

Lorho, G. (2005a). “Evaluation of Spatial Enhancement Systems for Stereo Headphone Reproduction by Preference and Attribute Rating”. In: *Proceedings of 118th AES Convention*. Barcelona, Spain: Audio Engineering Society.

— (2005b). “Individual Vocabulary Profiling of Spatial Enhancement Systems for Stereo Headphone Reproduction”. In: *119th AES Convention*. New York, NY, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13360`.

— (2009). “Subjective Evaluation of Headphone Target Frequency Responses”. In: *126th AES Convention*. Munich, Germany: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14966`.

— (2010). “Perceived quality evaluation: An application to sound reproduction over headphones”. PhD Thesis. Aalto University. ISBN: 9789526031958.

Lorho, G., D. Isherwood, et al. (2002). “Round Robin Subjective Evaluation of Stereo Enhancement Systems for Headphones”. In: *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, pp. 1–10.

Lorho, G., G. Le Ray, and N. Zacharov (2010). “eGauge—A Measure of Assessor Expertise in Audio Quality Evaluations”. In: *AES 38th International Conference: Sound Quality Evaluation*. Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=15471`.

Lorho, G. and N. Zacharov (2004). “Subjective Evaluation of Virtual Home Theatre Sound Systems for Loudspeakers and Headphones”. In: *Proceedings of 116th AES Convention*. Berlin, Germany: Audio Engineering Society.

Mackensen, P. (2004). “Auditive Localization. Head movements, an additional cue in Localization”. PhD. TU Berlin.

Mackensen, P., U. Felderhof, et al. (1999). “Binaural room scanning—A new tool for acoustic and psychoacoustic research”. In: *The Journal of the Acoustical Society of America* 105.2, p. 1343. ISSN: 00014966. DOI: `10.1121/1.426373`.

Mackensen, P., M. Fruhmann, et al. (2000). “Head Tracker-Based Auralization Systems: Additional Consideration of Vertical Head Movements”. In: *108th AES Convention*. Paris, France: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=9203`.

Macpherson, E. A. (2013). “Cue weighting and vestibular mediation of temporal dynamics in sound localization via head rotation”. In: *Proceedings of Meetings on Acoustics*. Vol. 19, pp. 050131–050131. DOI: `10.1121/1.4799913`.

Macpherson, E. A. and J. C. Middlebrooks (2000). “Localization of brief sounds: Effects of level and background noise”. In: *The Journal of the Acoustical Society of America* 108.4, pp. 1834–1849. ISSN: 0001-4966. DOI: `10.1121/1.1310196`.

— (2002). “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited”. In: *The Journal of the Acoustical Society of America* 111.5, pp. 2219–2236. ISSN: 00014966. DOI: `10.1121/1.1471898`.

— (2003). “Vertical-plane sound localization probed with ripple-spectrum noise”. In: *The Journal of the Acoustical Society of America* 114.430, pp. 430–445. ISSN: 0001-4966. DOI: `10.1121/1.1582174`.

Macpherson, E. A. and A. T. Sabin (2013). “Vertical-plane sound localization with distorted spectral cues.” In: *Hearing research* September, pp. 1–17. ISSN: 1878-5891. DOI: `10.1016/j.heares.2013.09.007`.

Maempel, H.-J. and A. Lindau (2012). “Opto-acoustic simulation of concert halls - a data-based approach”. In: *27th Tonmeistertagung - VDT International Convention*, pp. 293–309.

Mahony, R., T. Hamel, and J.-M. Pflimlin (2008). “Nonlinear Complementary Filters on the Special Orthogonal Group”. In: *IEEE Transactions on Automatic Control* 53.5, pp. 1203–1218. ISSN: 0018-9286. DOI: `10.1109/TAC.2008.923738`.

Majdak, P., P. Balazs, and B. Laback (2007). “Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions”. In: *Journal of the Audio Engineering Society* 55.7/8, pp. 623–637. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14190`.

Majdak, P., M. J. Goupell, and B. Laback (2010). “3-D localization of virtual sound sources : Effects of visual environment, pointing method, and training”. In: *Attention, Perception, & Psychophysics* 72.2, pp. 454–469. DOI: `10.3758/APP`.

Majdak, P., Y. Iwaya, et al. (2013). “Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions”. In: *Proceedings of 134th AES Convention*. Rome, Italy: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16781`.

Makous, J. C. and J. C. Middlebrooks (1990). “Two⍰dimensional sound localization by human listeners”. In: *The Journal of the Acoustical Society of America* 87.5, pp. 2188–2200. ISSN: 0001-4966. DOI: `10.1121/1.399186`.

Martens, H. and M. Martens (2001). *Multivariate analysis of quality: an introduction*. Wiley. ISBN: 978-0-471-97428-4. URL: `https://www.wiley.com/en-gb/Multivariate+Analysis+of+Quality:+An+Introduction-p-9780471974284`.

Martens, W. L. (1987). “Principal Components Analysis and Resynthesis of Spectral Cues to Perceived Direction”. In: *International Computer Music Conference*.

— (2003). “Individualized and Generalized Earphone Correction Filters for Spatial Audio Reproduction”. In: *International Conference on Auditory Display*, pp. 263–266.

Martin, R. L. and K. I. McAnally (2007). *Interpolation of Head-Related Transfer Functions*. Tech. rep. Victoria: Air Operations Division Defence Science and Technology Organisation.

Martin, R. L., K. I. McAnally, and M. A. Senova (2001). “Free-Field Equivalent Localization of Virtual Audio”. In: *Journal of the Audio Engineering Society* 49.1/2, pp. 14–22. URL: `http://www.aes.org/e-lib/browse.cfm?elib=10204`.

Masiero, B. (2012). “Individualized Binaural Technology”. Doctoral Thesis. Aachen University. ISBN: 9783832532741.

Masiero, B., P. Dietrich, et al. (2012). “Design of a Fast Individual HRTF Measurement System”. In: *Proceedings of DAGA*. Darmstadt, Germany.

Masiero, B. and J. Fels (2011). “Perceptually Robust Headphone Equalization for Binaural Reproduction”. In: *AES 130th Convention*. London, UK: Audio Engineering Society.

Mason, R. et al. (2001). “Verbal and Nonverbal Elicitation Techniques in the Subjective Assessment of Spatial Sound Reproduction”. In: *Journal of the Audio Engineering Society* 49.5, pp. 366–384. URL: `http://www.aes.org/e-lib/browse.cfm?elib=10190`.

Mattila, V.-V. (2001). “Descriptive Analysis of Speech Quality in Mobile Communications: Descriptive Language Development and External Preference Mapping”. In: *111th AES Convention*. New York, NY, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=9880`.

McAnally, K. I. and R. L. Martin (2014). “Sound Localization with Head Movement: Implications for 3-D Audio Displays”. In: *Frontiers in Neuroscience* 8.210, pp. 1–6. ISSN: 1662453X. DOI: `10.3389/fnins.2014.00210`.

McArthur, A. (2016). “Disparity in horizontal correspondence of sound and source positioning: The impact on spatial presence for cinematic VR”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Los Angeles, CA, USA: Audio Engineering Society.

McEwan, J. A. (1996). “Preference Mapping for Product Optimization”. In: *Data Handling in Science and Technology*. Vol. 16, pp. 71–102. DOI: `10.1016/S0922-3487(96)80027-X`.

McGill, R., J. W. Tukey, and W. A. Larsen (1978). “Variations of Box Plots”. In: *The American Statistician* 32.1, p. 12. ISSN: 00031305. DOI: `10.2307/2683468`.

McKeeg, A. and D. S. McGrath (1997). "Using Auralization Techniques to Render 5.1 Surround to Binaural and Transaural Playback". In: *AES 102nd Convention*. Munich, Germany: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=7321.

Meares, D. J. and E. W. Taylor (1982). *Technical Memorandum PH-1739 – An Assessment of the Neumann Artificial Head Type KU81*. Tech. rep. BBC Research Department.

Meddis, R. et al. (2013). *Basic Aspects of Hearing*. Ed. by B. C. J. Moore et al. Vol. 787. Advances in Experimental Medicine and Biology May. New York, NY: Springer New York. Chap. 2. ISBN: 978-1-4614-1589-3. DOI: 10.1007/978-1-4614-1590-9.

Mehra, R., A. Nicholls, et al. (2016). "Comparison of localization performance with individualized and non-individualized head-related transfer functions for dynamic listeners". In: *The Journal of the Acoustical Society of America* 140.4, pp. 2956–2957. ISSN: 0001-4966. DOI: 10.1121/1.4969129.

Mehra, R., A. Rungta, et al. (2015). "WAVE: Interactive Wave-based Sound Propagation for Virtual Environments". In: *IEEE Transactions on Visualization and Computer Graphics* 21.4, pp. 434–442. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2391858.

Mehrgardt, S. and V. Mellert (1977). "Transformation characteristics of the external human ear". In: *The Journal of the Acoustical Society of America* 61.6, pp. 1567–1576. ISSN: 0001-4966. DOI: 10.1121/1.381470.

Meier, M., M. Weitnauer, and J. Groh (2011). "Spherical Surface Microphone Interpolation". In: *International Conference on Spatial Audio*, pp. 117–124.

Melchior, F., C. Sladeczek, and D. de Vries (2008). "Spatial Sound Design: From Special Effects to Spatial Effects". In: *VDT International Convention*. November.

Melchior, F., C. Sladeczek, A. Partzsch, et al. (2010). "Design and Implementation of an Interactive Room Simulation for Wave Field Synthesis". In: *AES 40th International Conference: Spatial Sound*. Tokyo, Japan: Audio Engineering Society, pp. 1–8.

Melchior, F., O. Thiergart, et al. (2009). "Dual radius spherical cardioid microphone arrays for binaural auralization". In: *Proceedings of 127th AES Convention*. New York, NY, USA: Audio Engineering Society.

Mendonça, C., G. Campos, et al. (2013). "Learning Auditory Space: Generalization and Long-Term Effects". In: *PLoS ONE* 8.10. Ed. by M. S. Malmierca, e77900. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0077900.

Mendonça, C. and S. Delikaris-Manias (2018). "Statistical test with MUSHRA data". In: *Audio Engineering Society 144th Convention*. Milan, Italy: Audio Engineering Society.

Menzer, F. and C. Faller (2009). "Binaural reverberation using a modified Jot reverberator with frequency-dependent interaural coherence matching". In: *126th AES Convention*. Munich, Germany: Audio Engineering Society. URL: http://www.aes.org/e-lib/online/browse.cfm?elib=14961.

Menzer, F., C. Faller, and H. Lissek (2011). "Obtaining Binaural Room Impulse Responses From B-Format Impulse Responses Using Frequency-Dependent Coherence Matching". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.2, pp. 396–405. ISSN: 1558-7916. DOI: 10.1109/TASL.2010.2049410.

Menzies, D. (2010). "Nearfield Synthesis of Complex Sources with High-Order Ambisonics and Binaural Rendering". In: *Proceedings of the 13th International Conference on Auditory Display*.

Merimaa, J. (2009). "Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis". In: *Proceedings of 127th AES Convention*. New York, NY, USA: Audio Engineering Society.

— (2010). "Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis, Part 2: Individual HRTFs". In: *Proceedings of 129th AES Convention*, pp. 1–13.

Merimaa, J. and V. Pulkki (2006). "Spatial Impulse Response Rendering I: Analysis and Synthesis". In: *Journal of the Audio Engineering Society* 53.12, pp. 1115–1127.

Mershon, D. H., D. H. Desaulniers, et al. (1980). “Visual capture in auditory distance perception: Proximity image effect reconsidered.” In: *Journal of Auditory Research* 20.2, pp. 129–136.

Mershon, D. H. and L. E. King (1975). “Intensity and reverberation as factors in the auditory perception of egocentric distance”. In: *Perception & Psychophysics* 18.6, pp. 409–415. ISSN: 0031-5117. DOI: 10.3758/BF03204113.

Meyners, M. and J. C. Castura (2014). “Check-All-That-Apply Questions”. In: *Novel Techniques in Sensory Characterization and Consumer Profiling*. Ed. by P. Varela and G. Ares. CRC Press. Chap. 11, pp. 271–306. ISBN: 9781138034273.

— (2016). “Randomization of CATA attributes: Should attribute lists be allocated to assessors or to samples?” In: *Food Quality and Preference* 48, pp. 210–215. ISSN: 09503293. DOI: 10.1016/j.foodqual.2015.09.014.

Meyners, M., J. C. Castura, and B. T. Carr (2013). “Existing and new approaches for the analysis of CATA data”. In: *Food Quality and Preference* December, pp. 309–319. DOI: 10.1016/j.foodqual.2013.06.010.

Meyners, M., S. R. Jaeger, and G. Ares (2016). “On the analysis of Rate-All-That-Apply (RATA) data”. In: *Food Quality and Preference* 49, pp. 1–10. ISSN: 09503293. DOI: 10.1016/j.foodqual.2015.11.003.

Microsoft (2018a). *Kinect for Windows*. URL: https://developer.microsoft.com/en-us/windows/kinect (visited on 11/09/2018).

— (2018b). *Project Acoustics*. URL: https://docs.microsoft.com/en-us/azure/cognitive-services/acoustics/what-is-acoustics.

Middlebrooks, J. C. (1992). “Narrow⌧band sound localization related to external ear acoustics”. In: *The Journal of the Acoustical Society of America* 92.5, pp. 2607–2624. ISSN: 0001-4966. DOI: 10.1121/1.404400.

— (1999a). “Individual differences in external-ear transfer functions reduced by scaling in frequency”. In: *The Journal of the Acoustical Society of America* 106.3, pp. 1480–1492. ISSN: 0001-4966. DOI: 10.1121/1.427176.

— (1999b). “Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency”. In: *The Journal of the Acoustical Society of America* 106.3, pp. 1493–1510. ISSN: 0001-4966. DOI: 10.1121/1.427147.

Middlebrooks, J. C., E. A. Macpherson, and Z. A. Onsan (2000). “Psychophysical customization of directional transfer functions for virtual sound localization”. In: *The Journal of the Acoustical Society of America* 108.6, pp. 3088–3091. ISSN: 0001-4966. DOI: 10.1121/1.1322026.

Miller, J. D. et al. (2003). “Latency Measurement of a Real-Time Virtual Acoustic Environment Rendering System”. In: *Proceedings of the 2003 International Conference on Auditory Display*, pp. 1–21.

Millns, C. and H. Lee (2018). “An Investigation into Spatial Attributes of 360° Microphone Techniques for Virtual Reality”. In: *144th AES Convention*. Milan, Italy: Audio Engineering Society.

Mills, A. W. (1958). “On the Minimum Audible Angle”. In: *The Journal of the Acoustical Society of America* 30.4, pp. 237–246. ISSN: 0001-4966. DOI: 10.1121/1.1909553.

Minnaar, P., F. Christensen, et al. (1999). “Audibility of All-Pass Components in Binaural Synthesis”. In: *106th AES Convention*. Munich, Germany: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8269.

Minnaar, P., S. K. Olesen, et al. (2001). “Localization with Binaural Recordings from Artificial and Human Heads”. In: *Journal of the Audio Engineering Society* 49.5, pp. 323–336. URL: http://www.aes.org/e-lib/browse.cfm?elib=10193.

Minnaar, P., J. Plogsties, and F. Christensen (2005). “Directional resolution of head-related transfer functions required in binaural synthesis”. In: *Journal of the Audio Engineering Society* 53.10, pp. 919–929. URL: http://www.aes.org/e-lib/browse.cfm?elib=13392.

Mohan, A. et al. (2003). “Using computer vision to generate customized spatial audio”. In: *2003 International Conference on Multimedia and Expo. ICME ’03. Proceedings (Cat. No.03TH8698)*. IEEE, pp. III–57. ISBN: 0-7803-7965-9. DOI: 10.1109/ICME.2003.1221247.

Mokhtari, P. et al. (2010). “Acoustic sensitivity to micro-perturbations of KEMAR’s pinna surface geometry”. In: *AAS International Congress on Acoustics*. August, pp. 1–8.

— (2015). “Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry”. In: *The Journal of the Acoustical Society of America* 137.2, pp. 3–5. DOI: 10.1121/1.4906160.

— (2016). “Vertical normal modes of human ears : Individual variation and frequency estimation from pinna anthropometry”. In: *The Journal of the Acoustical Society of America* 140.2, pp. 3–5. DOI: 10.1121/1.4960481.

Møller, H. (1992). “Fundamentals of binaural technology”. In: *Applied Acoustics* 36.3-4, pp. 171–218. ISSN: 0003682X. DOI: 10.1016/0003-682X(92)90046-U.

Møller, H. and D. Hammershøi (1999). “Evaluation of Artificial Heads in Listening Tests”. In: *Journal of the Audio Engineering Society* 47.3, pp. 83–100.

Møller, H., D. Hammershøi, et al. (1995). “Transfer Characteristics of Headphones Measured on Human Ears”. English. In: *Journal of the Audio Engineering Society* 43.4, pp. 203–217. URL: http://www.aes.org/e-lib/browse.cfm?elib=7954.

Møller, H., C. B. Jensen, et al. (1995). “Design Criteria for Headphones”. English. In: *Journal of the Audio Engineering Society* 43.4, pp. 218–232. URL: http://www.aes.org/e-lib/browse.cfm?elib=10274.

— (1996). “Using a Typical Human Subject for Binaural Recording”. In: *100th AES Convention*. Copenhagen, Denmark: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=7614.

Møller, H., M. F. Sorensen, et al. (1995). “Head-Related Transfer Functions of Human Subjects”. In: *Journal of the Audio Engineering Society* 43.5, pp. 300–321. URL: http://www.aes.org/e-lib/browse.cfm?elib=7949.

Møller, H., M. F. Sørensen, et al. (1996). “Binaural Technique: Do We Need Individual Recordings?” In: *Journal of the Audio Engineering Society* 44.6, pp. 451–469. URL: http://www.aes.org/e-lib/browse.cfm?elib=7897.

Möller, S. and A. Raake (2014). *Quality of Experience*. Ed. by S. Möller and A. Raake. Springer. ISBN: 9783319026800.

Möller, S., M. Wältermann, and M.-N. Garcia (2014). “Features of Quality of Experience”. In: *Quality of Experience*. Ed. by S. Möller and A. Raake. Springer. Chap. 5, pp. 73–84. DOI: 10.1007/978-3-319-02681-7_5.

Moore, A. H., A. I. Tew, and R. Nicol (2010). “An Initial Validation of Individualized Crosstalk Cancellation Filters for Binaural Perceptual Experiments”. In: *Journal of the Audio Engineering Society* 58.1, pp. 36–45. URL: http://www.aes.org/e-lib/browse.cfm?elib=15240.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing*. 5th. Elsevier, Academic Press. ISBN: 0125056281.

Moreau, S., J. Daniel, and S. Bertet (2006). “3D Sound Field Recording with Higher Order Ambisonics – Objective Measurements and Validation of Spherical Microphone”. In: *120th AES Convention*. Paris, France: Audio Engineering Society.

Morse, P. M. and K. U. Ingard (1987). *Theoretical Acoustics*. Princeton University Press. ISBN: 9780691024011.

Moulin, S., S. Bech, and T. Stegenborg-Andersen (2016). “Sensory profiling of high-end loudspeakers using rapid methods - Part 1: Baseline experiment using headphone reproduction”. In: *AES International Conference on Headphone Technology*. Aalborg, Denmark: Audio Engineering Society.

Muja, M. and D. G. Lowe (2009). “Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration”. In: *International Conference on Computer Vision Theory and Applications (VISAPP ’09)*, pp. 1–10.

Müller, S. and P. Massarani (2001). “Transfer Function Measurement with Sweeps”. In: *Journal of the Audio Engineering Society* 49.6, pp. 443–471.

Munch, T., W. Hess, and S. Beyer (2009). *Patent US 2009/0058606 A1 – Tracking System Using Radio Frequency Identification Technology*.

Murgai, P., M. Rau, and J.-M. Jot (2017). “Blind Estimation of the Reverberation Fingerprint of Unknown Acoustic Environments”. In: *143rd AES Convention*. New York, NY, USA: Audio Engineering Society.

Murphy, D., S. Shelley, et al. (2017). “Acoustic Heritage and Audio Creativity: the Creative Application of Sound in the Representation, Understanding and Experience of Past Environments”. In: *Internet Archaeology* 44. ISSN: 13635387. DOI: 10.11141/ia.44.12.

Murphy, K. R. and B. Myors (1999). “Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model”. In: *Journal of Applied Psychology* 84.2, pp. 234–248. ISSN: 00219010. DOI: 10.1037/0021-9010.84.2.234.

Murphy-Chutorian, E. and M. M. Trivedi (2009). “Head Pose Estimation in Computer Vision: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.4, pp. 607–626. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2008.106.

Musil, T., M. Noisternig, and R. Höldrich (2005). “A Library for Real Time 3D Binaural Sound Reproduction in Pure Data (PD)”. In: *International Conference on Digital Audio Effects*. Madrid, Spain, pp. 1–5.

Nachbar, C. et al. (2011). “AmbiX - A Suggested Ambisonics Format”. In: *International Symposium on Ambisonics and Spherical Acoustics*. Lexington, KY, USA.

Nam, J., J. S. Abel, and J. O. Smith III (2008). “A Method for Estimating Interaural Time Difference for Binaural Synthesis”. In: *Proceedings of 125th AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

Narbutt, M. et al. (2017). “Streaming VR for Immersion : Quality aspects of Compressed Spatial Audio”. In: *23rd International Conference on Virtual System & Multimedia (VSMM)*. Dublin, Ireland: IEEE. ISBN: 9781538644942.

Natural Point (2018a). *OptiTrack systems*. URL: http://www.naturalpoint.com/optitrack/.

— (2018b). *Track IR 5*. URL: http://www.naturalpoint.com/trackir/products/trackir5/.

Neidhardt, A. and N. Knoop (2015). “Investigating the room divergence effect in binaural playback”. In: *International Conference on Spatial Audio*. Graz, Austria.

Netflix (2018). *Netflix Dolby Atmos Home Mix Deliverable Requirements v1.9 – Netflix | Partner Help Center*. URL: https://partnerhelp.netflixstudios.com/hc/en-us/articles/115001539991-Netflix-Dolby-Atmos-Home-Mix-Deliverable-Requirements-v1-9 (visited on 11/18/2018).

Nicol, R. et al. (2016). “How to make immersive audio available for mass-market listening”. In: *EBU Technical Review*. ISSN: 16091469.

Niehorster, D. C., L. Li, and M. Lappe (2017). “The accuracy and precision of position and orientation tracking in the HTC vive virtual reality system for scientific research”. In: *i-Perception* 8.3, pp. 1–23. ISSN: 20416695. DOI: 10.1177/2041669517708205.

Nixon, T., A. Bonney, and F. Melchior (2015). “A Reference Listening Room for 3D Audio Research”. In: *International Conference on Spatial Audio*. Graz, Austria.

Norcross, S. G., M. Bouchard, and G. A. Soulodre (2006). “Inverse Filtering Design Using a Minimal-Phase Target Function from Regularization”. In: *121st AES Convention*. San Francisco, CA, USA: Audio Engineering Society.

Novo, P. (2005). “Auditory Virtual Environments”. In: *Communication Acoustics*. Ed. by J. Blauert. Berlin/Heidelberg: Springer-Verlag. Chap. 11, pp. 277–297. DOI: 10.1007/3-540-27437-5_11.

Nunnally, J. C. and I. H. Bernstein (1994). *Psychometric Theory*. 3rd editio. New York: McGraw-Hill. ISBN: 007047849X.

NVIDIA (2017). *VRWorks Audio*. URL: `https://developer.nvidia.com/vrworks/vrworks-audio`.

Oberem, J., J. Fels, and B. Masiero (2013). "Experiments on authenticity and naturalness of binaural reproduction via headphones". In: *Applied Acoustics* 114, pp. 71–78. ISSN: 1939800X. DOI: `10.1121/1.4799533`.

Oberem, J., B. Masiero, and J. Fels (2016). "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods". In: *Applied Acoustics* 114. ISSN: 1872910X. DOI: `10.1016/j.apacoust.2016.07.009`.

Oculus (2018). *Oculus Audio SDK*. URL: `https://developer.oculus.com/documentation/audiosdk/latest/concepts/book-audiosdk/` (visited on 11/18/2018).

Olive, S. E. (2003). "Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study". In: *Journal of the Audio Engineering Society* 51.9, pp. 806–825.

Olive, S. E., O. Khonsaripour, and T. Welti (2018). "A Survey and Analysis of Consumer and Professional Headphones Based on Their Objective and Subjective Performances". In: *AES 145th Convention*. New York, NY, USA: Audio Engineering Society.

Olive, S. E. and T. Welti (2009). "Validation of a Binaural Car Scanning System for Subjective Evaluation of Automotive Audio Systems". In: *AES 36th International Conference: Automotive Audio*. Dearborn, Michigan, USA: Audio Engineering Society, pp. 1–7.

— (2015). "Factors that Influence Listeners' Preferred Bass and Treble Balance in Headphones". In: *AES 139th Convention*. New York, NY, USA: Audio Engineering Society, pp. 1–12.

Olive, S. E., T. Welti, and O. Khonsaripour (2016). "The Preferred Low Frequency Response of In-Ear Headphones". In: *AES International Conference on Headphone Technology*. Aalborg, Denmark: Audio Engineering Society. ISBN: 9781942220091. URL: `http://www.aes.org/e-lib/browse.cfm?elib=18369`.

— (2017a). "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 1 – Listening Test Results and Acoustic Measurements". In: *AES 143rd Convention*. New York, NY, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=19237`.

— (2017b). "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 2 – Development and Validation of the Model". In: *AES 143rd Convention*. New York, NY, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=19275`.

— (2018). "A Statistical Model that Predicts Listeners' Preference Ratings of Around-Ear and On-Ear Headphones". In: *AES 144th Convention*. Milan, Italy: Audio Engineering Society, pp. 1–15.

Olive, S. E., T. Welti, and E. McMullin (2013a). "A Virtual Headphone Listening Test Methodology". In: *Audio Engineering Society 51st International Conference*. Audio Engineering Society, pp. 1–10. ISBN: 9781629933283. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16874`.

— (2013b). "Listener Preferences for Different Headphone Target Response Curves". In: *134th AES Convention*. Rome, Italy: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16768`.

Olive, S. E., T. Welti, and E. Mcmullin (2014). "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences". In: *AES 137th Convention*. Los Angeles, CA, USA: Audio Engineering Society. ISBN: 9781634397483.

Oppenheim, A. V. and R. W. Schafer (1998). *Discrete-Time Signal Processing*. Second edi. Prentice Hall. ISBN: 0137549202.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum (1957). *The Measurement of Meaning*. University of Illinois Press. ISBN: 0252745396.

Paquier, M. and V. Koehl (2015). "Discriminability of the placement of supra-aural and circumaural headphones". In: *Applied Acoustics* 93, pp. 130–139. ISSN: 0003682X. DOI: 10.1016/j.apacoust.2015.01.023.

Pardoe, L. and C. Pike (2018). *Adding dynamic spatial audio to SMP360*. Tech. rep. BBC Research & Development.

Parente, M. E., A. V. Manzoni, and G. Ares (2011). "External Preference Mapping of Commercial Antiaging Creams Based on Consumers' Responses to a Check-All-That-Apply Question". In: *Journal of Sensory Studies* 26, pp. 158–166. DOI: 10.1111/j.1745-459X.2011.00332.x.

Parseihian, G. and B. F. G. Katz (2012). "Rapid head-related transfer function adaptation using a virtual auditory environment." In: *The Journal of the Acoustical Society of America* 131.4, pp. 2948–2957. ISSN: 1520-8524. DOI: 10.1121/1.3687448.

Paul, S. (2009). "Binaural Recording Technology: A Historical Review and Possible Future Developments". In: *Acta Acustica united with Acustica* 95, pp. 767–788. DOI: 10.3813/AAA.918208.

Pedersen, T. H. and N. Zacharov (2015). "The development of a Sound Wheel for Reproduced Sound". In: *138th AES Convention*. Warsaw, Poland: Audio Engineering Society. ISBN: 9781510806597.

Pellegrini, R. S. (1999). "Comparison of Data- and Model-Based Simulation Algorithms for Auditory Virtual Environments". In: *106th AES Convention*. Munich, Germany: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=8227.

— (2001). "A Virtual Reference Listening Room as an Application of Virtual Environments". PhD. Ruhr-Universität Bochum. ISBN: 3898254038.

— (2002). "Perception-Based Design of Virtual Rooms for Sound Reproduction". In: *AES 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=11137.

Pellegrini, R. S., C. Kuhn, and M. Gebhardt (2007). "Headphones Technology for Surround Sound Monitoring - A Virtual 5.1 Listening Room." In: *AES 122nd Convention*. Vienna, Austria: Audio Engineering Society.

Perrett, S. and W. Noble (1997). "The effect of head rotations on vertical plane sound localization". In: *The Journal of the Acoustical Society of America* 102.4, pp. 2325–2332. ISSN: 0001-4966. DOI: 10.1121/1.419642.

Perrin, L. et al. (2008). "Comparison of three sensory methods for use with the Napping® procedure: Case of ten wines from Loire valley". In: *Food Quality and Preference* 19.1, pp. 1–11. DOI: 10.1016/J.FOODQUAL.2007.06.005.

Perrott, D. R. (1984). "Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity". In: *The Journal of the Acoustical Society of America* 75.4, p. 1201. ISSN: 00014966. DOI: 10.1121/1.390771.

Perrott, D. R. and A. D. Musicant (1977). "Minimum auditory movement angle: Binaural localization of moving sound sources". In: *The Journal of the Acoustical Society of America* 62.6, pp. 1463–1466. ISSN: 0001-4966. DOI: 10.1121/1.381675.

Peus, S. (1985). *Natural Listening with a Dummy Head*. Tech. rep. Berlin: Georg Neumann GmbH.

Pihlajamäki, T., O. Santala, and V. Pulkki (2014). "Synthesis of Spatially Extended Virtual Source with Time-Frequency Decomposition of Mono Signals". In: *Journal of the Audio Engineering Society* 62.7/8, pp. 467–484. ISSN: 15494950. DOI: 10.17743/jaes.2014.0031.

Pike, C. and F. Melchior (2014). *Tommies in 3D - Binaural Headphone Mix - BBC R&D*. URL: https://www.bbc.co.uk/rd/blog/2014-10-tommies-in-3d (visited on 12/15/2018).

Pike, C. and T. Nixon (2014). *Lend Us Your Ears! Immersive Headphone Experiments - BBC R&D*. URL: https://www.bbc.co.uk/rd/blog/2014-10-binaural-experiment-surveys (visited on 12/15/2018).

Pike, C. and M. Romanov (2017a). "An impulse response dataset for dynamic data-based auralisation of advanced sound systems". In: *142nd AES Convention*. Berlin, Germany: Audio Engineering Society.

Pike, C. and M. Romanov (2017b). *BBC R&D BRIRs – An impulse response dataset for dynamic data-based auralisation of advanced sound systems*. URL: `https://github.com/bbc/bbcrd-brirs` (visited on 10/28/2018).

Pike, C. and H. Stenzel (2017). “Direct and indirect listening test methods-a discussion based on audio-visual spatial coherence experiments”. In: *AES 143th Convention*. New York, NY, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib.`.

Pinheiro, J. et al. (2018). *{nlme}: Linear and Nonlinear Mixed Effects Models*. URL: `https://cran.r-project.org/package=nlme`.

Plaehn, D. (2012). “CATA penalty/reward”. In: *Food Quality and Preference* 24.1, pp. 141–152. ISSN: 09503293. DOI: `10.1016/j.foodqual.2011.10.008`.

Plenge, G. (1974). “On the differences between localization and lateralization”. In: *The Journal of the Acoustical Society of America* 56.3, pp. 944–951. ISSN: 0001-4966. DOI: `10.1121/1.1903353`.

Plogsties, J. et al. (2000). “Audibility of All-Pass Components in Head-Related Transfer Functions”. In: *108th AES Convention*. Vol. 108. Paris, France: Audio Engineering Society.

Poirier-Quinot, D. and B. F. G. Katz (2018a). “Impact of HRTF individualization on player performance in a VR shooter game I”. In: *AES Conference on Spatial Reproduction*. Tokyo, Japan: Audio Engineering Society.

— (2018b). “Impact of HRTF individualization on player performance in a VR shooter game II”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society.

Poirier-Quinot, D., B. F. G. Katz, and M. Noisternig (2017). “EVERTims: Open Source Framework for Real-Time Auralization in Architectural Acoustics and Virtual Reality”. In: *Digital Audio Effects (DAFx)*, pp. 323–328.

Poletti, M. (1996). “The Design of Encoding Functions for Stereophonic and Polyphonic Sound Systems”. In: *Journal of the Audio Engineering Society* 44.11, pp. 948–963.

Polhemus (2018). *Fastrak*. URL: `https://polhemus.com/motion-tracking/head-trackers/` (visited on 11/09/2018).

Polich, J. (2007). “Updating P300: An integrative theory of P3a and P3b”. In: *Clinical Neurophysiology* 118.10, pp. 2128–2148. ISSN: 13882457. DOI: `10.1016/j.clinph.2007.04.019`.

Politis, A., L. McCormack, and V. Pulkki (2017). “Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing”. In: *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, pp. 379–383. ISBN: 978-1-5386-1632-1. DOI: `10.1109/WASPAA.2017.8170059`.

Politis, A. and D. Poirier-Quinot (2016). “JSAmbisonics : A Web Audio library for interactive spatial sound processing on the web”. In: *Interactive Audio Systems Symposium*. York, United Kingdom.

Politis, A., S. Tervo, and V. Pulkki (2018). “COMPASS: Coding and Multidirectional Parameterization of Ambisonic Sound Scenes”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* May, pp. 6802–6806. ISSN: 15206149. DOI: `10.1109/ICASSP.2018.8462608`.

Pollow, M. et al. (2012). “Calculation of Head-Related Transfer Functions for Arbitrary Field Points Using Spherical Harmonics Decomposition”. In: *Acta Acustica united with Acustica* 98.1, pp. 72–82. ISSN: 1610-1928. DOI: `10.3813/AAA.918493`.

Pörschmann, C., J. M. Arend, and P. Stade (2017). “Influence of head tracking on distance estimation of nearby sound sources”. In: *DAGA*. Kiel, Germany.

Pörschmann, C., P. Stade, and J. M. Arend (2017). “Binauralization of Omnidirectional Room Impulse Responses - Algorithm and Technical Evaluation”. In: *Digital Audio Effects (DAFx)*. Edinburgh, UK, pp. 345–352.

Pralong, D. and S. Carlile (1996). “The role of individualized headphone calibration for the generation of high fidelity virtual auditory space”. In: *The Journal of the Acoustical Society of America* 100.6, pp. 3785–3793. ISSN: 0001-4966. DOI: `10.1121/1.417337`.

Pulkki, V. (1997). “Virtual Sound Source Positioning Using Vector Base Amplitude Panning”. In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

— (2001a). “Coloration of Amplitude-Panned Virtual Sources”. In: *AES 110th Convention*. Amsterdam, The Netherlands: Audio Engineering Society.

— (2001b). “Localization of Amplitude-Panned Virtual Sources II: Two-and Three-dimensional Panning”. In: *Journal of the Audio Engineering Society* 49.9, pp. 753–767.

— (2001c). “Spatial sound generation and perception by amplitude panning techniques”. Doctoral Thesis. Helsinki University of Technology. ISBN: 9512255324. DOI: `ISBN951-22-5531-6ISSN1456-6303`.

Pulkki, V. and M. Karjalainen (2001). “Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning”. In: *Journal of the Audio Engineering Society* 49.9, pp. 739–752.

— (2015). “Spatial Hearing”. In: *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics*. Wiley. Chap. 12. ISBN: 978-1-118-86654-2.

QSound Labs (2007). *Virtual Barber Shop (Audio...use headphones, close ur eyes) - YouTube*. URL: `https://www.youtube.com/watch?v=IUDTlvagjJA` (visited on 12/09/2018).

Quen, H. and H. V. D. Bergh (2004). “On multi-level modeling of data from repeated measures designs : a tutorial”. In: *Speech Communication* 43, pp. 103–121. DOI: `10.1016/j.specom.2004.02.004`.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. URL: `https://www.r-project.org/`.

Raake, A. and S. Egger (2014). “Quality and Quality of Experience”. In: *Quality of Experience*. Ed. by S. Möller and A. Raake. Springer. Chap. 2, pp. 11–34. ISBN: 978-3-319-02680-0. DOI: `10.1007/978-3-319-02681-7`.

Rafaely, B. (2005). “Analysis and design of spherical microphone arrays”. In: *IEEE Transactions on Speech and Audio Processing* 13.1, pp. 135–143. ISSN: 1063-6676. DOI: `10.1109/TSA.2004.839244`.

— (2015). *Fundamentals of Spherical Array Processing*. Springer. ISBN: 9783662456637.

Rakerd, B. and W. M. Hartmann (2010). “Localization of sound in rooms, V. Binaural coherence and human sensitivity to interaural time differences in noise.” In: *The Journal of the Acoustical Society of America* 128.5, pp. 3052–63. ISSN: 1520-8524. DOI: `10.1121/1.3493447`.

Reardon, G. et al. (2018). “Evaluation of Binaural Renderers: Multidimensional Sound Quality Assessment”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society. ISBN: 0780309405.

Reinbach, H. C. et al. (2014). “Comparison of three sensory profiling methods based on consumer perception: CATA, CATA with intensity and Napping®”. In: *Food Quality and Preference* 32, pp. 160–166. ISSN: 09503293. DOI: `10.1016/j.foodqual.2013.02.004`.

Reiter, U. et al. (2014). “Factors Influencing Quality of Experience”. In: *Quality of Experience*. Ed. by S. Möller and A. Raake. Springer International Publishing. Chap. 4. DOI: `10.1007/978-3-319-02681-7_4`.

Richter, J.-G. et al. (2014). “Spherical Harmonics Based HRTF Datasets: Implementation and Evaluation for Real-Time Auralization”. In: *Acta Acustica united with Acustica* 100.4, pp. 667–675. ISSN: 16101928. DOI: `10.3813/AAA.918746`.

Risvik, E. et al. (1994). “Projective mapping: A tool for sensory analysis and consumer research”. In: *Food Quality and Preference* 5.4, pp. 263–269. DOI: `10.1016/0950-3293(94)90051-5`.

Roginska, A., G. Wakefield, and T. S. Santoro (2010). “Stimulus-Dependent HRTF Preference”. In: *Proceedings of 129th AES Convention*. San Francisco, CA, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=15690`.

Romblom, D. and B. Cook (2008). “Near-Field Compensation for HRTF Processing”. In: *125th AES Convention*. San Francisco, CA, USA: Audio Engineering Society, pp. 1–6. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14762`.

Romigh, G. D., D. S. Brungart, and B. D. Simpson (2015). “Free-Field Localization Performance with a Head-Tracked Virtual Auditory Display”. In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 943–954. ISSN: 19324553. DOI: `10.1109/JSTSP.2015.2421874`.

Romigh, G. D., D. S. Brungart, R. M. Stern, et al. (2015). “Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions”. In: *IEEE Journal of Selected Topics in Signal Processing* 9.5, pp. 921–930. ISSN: 1932-4553. DOI: `10.1109/JSTSP.2015.2421876`.

Romigh, G. D. and B. D. Simpson (2014). “Do you hear where I hear?: isolating the individualized sound localization cues”. In: *Frontiers in Neuroscience* 8.370. ISSN: 1662-453X. DOI: `10.3389/fnins.2014.00370`.

Rothbucher, M. et al. (2013). “Comparison of head-related impulse response measurement approaches”. In: *The Journal of the Acoustical Society of America* 134.2, EL223. ISSN: 00014966. DOI: `10.1121/1.4813592`.

Rubak, P. and L. G. Johansen (2003). “Coloration in Natural and Artificial Room Impulse Responses”. In: *AES 23rd International Conference*. Copenhagen, Denmark: Audio Engineering Society.

Rudrich, D. (2018). *IEM Plug-in Suite*. URL: `https://plugins.iem.at/` (visited on 12/23/2018).

Rummukainen, O., T. Robotham, et al. (2018). “Audio Quality Evaluation in Virtual Reality: Multiple Stimulus Ranking with Behavior Tracking”. In: *AES Conference on Audio for Virtual and Augmented Reality*. Redmond, WA, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib`.

Rummukainen, O., J. Wang, et al. (2018). “Influence of Visual Content on the Perceived Audio Quality in Virtual Reality”. In: *AES 145th Convention*. New York, USA: Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=19854`.

Rumsey, F. (1999). “Controlled Subjective Assessments of Two-to-Five-Channel Surround Sound Processing Algorithms”. In: *Journal of the Audio Engineering Society* 47.7/8, pp. 563–582. URL: `http://www.aes.org/e-lib/browse.cfm?elib=12099`.

— (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. URL: `http://www.aes.org/e-lib/browse.cfm?elib=11067`.

— (2011). “Whose head is it anyway? Optimizing Binaural Audio”. In: *Journal of the Audio Engineering Society* 59.9, pp. 672–675.

Rumsey, F., S. K. Zieliński, P. Jackson, et al. (2008). “QESTRAL (Part 1): Quality Evaluation of Spatial Transmission and Reproduction using an Artificial Listener”. In: *125th AES Convention*.

Rumsey, F., S. K. Zieliński, R. Kassier, et al. (2005a). “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality”. In: *The Journal of the Acoustical Society of America* 118.2, pp. 968–976. ISSN: 0001-4966. DOI: `10.1121/1.1945368`.

— (2005b). “Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences”. In: *The Journal of the Acoustical Society of America* 117.6, pp. 3832–3840. ISSN: 0001-4966. DOI: `10.1121/1.1904305`.

Sakamoto, N. et al. (1978). “Linear-Drive Headphones with Eardrum Response”. In: *60th AES Convention*. Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=3013`.

Salomons, A. M. (1995). “Coloration and binaural decoloration of sound due to reflections”. PhD thesis. TU Delft.

Samet, H. (1989). “Implementing ray tracing with octrees and neighbor finding”. In: *Computers & Graphics* 13.4, pp. 445–460. ISSN: 00978493. DOI: `10.1016/0097-8493(89)90006-X`.

Samsung (2018). *Gear VR*. URL: https://www.samsung.com/global/galaxy/gear-vr/ (visited on 10/24/2018).

Sandvad, J. (1996). “Dynamic Aspects of Auditory Virtual Environments”. In: *100th AES Convention*. Copenhagen, Denmark: Audio Engineering Society.

Santala, O. and V. Pulkki (2011). “Directional perception of distributed sound sources”. In: *The Journal of the Acoustical Society of America* 129.3, pp. 1522–1530. ISSN: 0001-4966. DOI: 10.1121/1.3533727.

Satongar, D. (2016). “Simulation and analysis of spatial audio reproduction and listening area effects”. PhD thesis. University of Salford.

Satongar, D., Y. Lam, and C. Pike (2014). “Measurement and analysis of a spatially sampled binaural room impulse response dataset”. In: *21st International Congress on Sound and Vibration 2014, ICSV 2014*. Vol. 2. ISBN: 9781634392389.

Satongar, D., C. Pike, et al. (2015). “The Influence of Headphones on the Localization of External Loudspeaker Sources”. In: *Journal of the Audio Engineering Society* 63.10, pp. 799–810.

Savioja, L. et al. (1999). “Creating Interactive Virtual Acoustic Environments”. In: *Journal of the Audio Engineering Society* 47.9, pp. 675–705. URL: http://www.aes.org/e-lib/browse.cfm?elib=12095.

Scaini, D. and D. Arteaga (2014). “Decoding of Higher Order Ambisonics to Irregular Periphonic Loudspeaker Arrays”. In: *AES 55th International Conference*. Helsinki, Finland: Audio Engineering Society, pp. 1–8.

Schärer, Z. and A. Lindau (2009). “Evaluation of Equalization Methods for Binaural Signals”. In: *126th AES Convention*. Munich, Germany: Audio Engineering Society.

Schatz, R. and P. Reichl (2011). “Quality of Experience – Just another Buzzword?” In: *Euroview 2011*. Würzburg, Germany.

Scheirer, E., R. Vaananen, and J. Huopaniemi (1999). “AudioBIFS: Describing audio scenes with the MPEG-4 multimedia standard”. In: *IEEE Transactions on Multimedia* 1.3, pp. 237–250. ISSN: 15209210. DOI: 10.1109/6046.784463.

Schinkel-Bielefeld, N., N. Lotze, and F. Nagel (2013). “Audio quality evaluation by experienced and inexperienced listeners”. In: *Proceedings of Meetings on Acoustics*. Vol. 19. DOI: 10.1121/1.4799190.

Schissler, C., C. Loftin, and D. Manocha (2018). “Acoustic Classification and Optimization for Multi-Modal Rendering of Real-World Scenes”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.3, pp. 1246–1259. ISSN: 1077-2626. DOI: 10.1109/TVCG.2017.2666150.

Schissler, C., R. Mehra, and D. Manocha (2014). “High-Order Diffraction and Diffuse Reflections for Interactive Sound Propagation in Large Environments”. In: *ACM SIGGRAPH*. ISBN: 0730-0301. DOI: 10.1145/2601097.2601216.

Schissler, C., A. Nicholls, and R. Mehra (2016). “Efficient HRTF-based Spatial Audio for Area and Volumetric Sources”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.4, pp. 1356–1366. ISSN: 10772626. DOI: 10.1109/TVCG.2016.2518134.

Schlich, P. (1995). “Preference mapping: relating consumer preferences to sensory or instrumental measurements”. In: *Colloques de l’INRA*.

Schmidt, S. (2009). “Finite Element simulation of external ear sound fields for the optimization of eardrum-related measurements”. PhD. Ruhr-Universitat Bochum. ISBN: 978-3-8325-2262-9.

Schoeffler, M. (2013). “About the Impact of Audio Quality on Overall Listening Experience”. In: *Sound and Music Computing Conference*. Stockholm, Sweden, pp. 48–53.

Schoeffler, M., S. Conrad, and J. Herre (2014). “The Influence of the Single / Multi-Channel-System on the Overall Listening Experience”. In: *AES 55th International Conference*. Helsinki, Finland: Audio Engineering Society, pp. 1–8.

Schoeffler, M., B. Edler, and J. Herre (2013). “How Much Does Audio Quality Influence Ratings of Overall Listening Experience ?” In: *10th International Symposium on Computer Music Multidisciplinary Research*. Marseille, France.

Schoeffler, M. and J. Herre (2014). “About the Different Types of Listeners for Rating the Overall Listening Experience”. In: *Sound and Music Conference*.

— (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. ISSN: 0001-4966. DOI: 10.1121/1.4963078.

Schoeffler, M., A. Silzle, and J. Herre (2017). “Evaluation of Spatial/3D Audio: Basic Audio Quality Versus Quality of Experience”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.1, pp. 75–88. ISSN: 1932-4553. DOI: 10.1109/JSTSP.2016.2639325.

Schönstein, D. and B. F. G. Katz (2012). “Variability in Perceptual Evaluation of HRTFs”. In: *Journal of the Audio Engineering Society* 60.10, pp. 783–793. URL: http://www.aes.org/e-lib/browse.cfm?elib=16552.

Schröder, D. and M. Vorländer (2011). “RAVEN : A Real-Time Framework for the Auralization of Interactive Virtual Environments”. In: *Forum Acusticum*. Aalborg, pp. 1541–1546. ISBN: 9788469415207.

Schroeder, M. R. (1970). “Digital Simulation of Sound Transmission in Reverberant Spaces”. In: *The Journal of the Acoustical Society of America* 47.2A, pp. 424–431. ISSN: 0001-4966. DOI: 10.1121/1.1911541.

Schultz, F., A. Lindau, and S. Weinzierl (2009). “Just Noticeable BRIR Grid Resolution for Lateral Head Movements”. In: *NAG-DAGA*. Rotterdam, The Netherlands, pp. 200–201. URL: http://asa.scitation.org/doi/10.1121/1.1911541.

Schultz, F. and S. Spors (2013). “Data-based Binaural Synthesis Including Rotational and Translatory Head-Movements”. In: *AES 52nd International Conference*. Guildford, UK: Audio Engineering Society.

Scott, F. S. and A. Roginska (2008). “Room-dependent preference of Virtual Surround Sound”. In: *AES 124th Convention*. Amsterdam, The Netherlands: Audio Engineering Society.

Searle, C. L. et al. (1975). “Binaural pinna disparity: another auditory localization cue.” In: *The Journal of the Acoustical Society of America* 57.2, pp. 448–55. ISSN: 0001-4966. URL: http://www.ncbi.nlm.nih.gov/pubmed/1117098.

Seeber, B. U. and H. Fastl (2003). “Subjective Selection of Non-Individual Head-Related Transfer Functions”. In: *International Conference on Auditory Display*. Boston, MA, USA, pp. 259–262.

Seeingmachines (2018). *Face API*. URL: https://sourceforge.net/p/facetracknoir/wiki/faceAPI/ (visited on 11/10/2018).

Shaw, E. A. G. (2007). “Acoustical Characteristics of the Outer Ear”. In: *Encyclopedia of Acoustics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 1325–1335. DOI: 10.1002/9780470172537.ch105.

Shinn-Cunningham, B. G. (2000). “Learning Reverberation : Considerations for Spatial Auditory Displays”. In: *International Conference on Auditory Display*.

Shinn-Cunningham, B. G., N. I. Durlach, and R. M. Held (1998a). “Adapting to supernormal auditory localization cues. I. Bias and resolution”. In: *The Journal of the Acoustical Society of America* 103.6, pp. 3656–3666. ISSN: 0001-4966. DOI: 10.1121/1.423088.

— (1998b). “Adapting to supernormal auditory localization cues. II. Constraints on adaptation of mean response”. In: *The Journal of the Acoustical Society of America* 103.6, pp. 3667–3676. ISSN: 0001-4966. DOI: 10.1121/1.423107.

Shinn-Cunningham, B. G., S. Santarelli, and N. Kopco (2000). “Tori of confusion: Binaural localization cues for sources within reach of a listener”. In: *The Journal of the Acoustical Society of America* 107.3, pp. 1627–1636. ISSN: 0001-4966. DOI: 10.1121/1.428447.

Shirley, B., P. Kendrick, and C. Churchill (2007). "The effect of stereo crosstalk on intelligibility: Comparison of a phantom stereo image and a central loudspeaker source". In: *Journal of the Audio Engineering Society* 55.10, pp. 852–863. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14174`.

Shotton, M., C. Pike, and F. Melchior (2014). "A Motorised Telescope Mount as a Computer-Controlled Rotational Platform for Dummy Head Measurements". In: *136th AES Convention*. Berlin, Germany: Audio Engineering Society.

Silzle, A. (2002). "Selection and Tuning of HRTFS". In: *112th AES Convention*. Munich, Germany: Audio Engineering Society.

— (2007). "Quality Taxonomies for Auditory Virtual Environments". In: *Proceedings of 122nd AES Convention*. Vienna, Austria: Audio Engineering Society.

Silzle, A., S. George, et al. (2011). "Investigation on the Quality of 3D Sound Reproduction". In: *Proceedings of ICSA 2011*. Detmold, Germany.

Silzle, A., B. Neugebauer, et al. (2009). "Binaural Processing Algorithms: Importance of Clustering Analysis for Preference Tests". In: *Proceedings of 126th AES Convention*. Vol. 126. Munich, Germany: Audio Engineering Society.

Silzle, A., P. Novo, and H. Strauss (2004). "IKA-SIM: A System to Generate Auditory Virtual Environments". In: *116th AES Convention*. Berlin, Germany: Audio Engineering Society.

Simon, G., A. Fitzgibbon, and A. Zisserman (2000). "Markerless tracking using planar structures in the scene". In: *Proceedings IEEE and ACM International Symposium on Augmented Reality*. IEEE, pp. 120–128. ISBN: 0-7695-0846-4. DOI: `10.1109/ISAR.2000.880935`.

Simon, L. S. R., N. Zacharov, and B. F. G. Katz (2016). "Perceptual attributes for the comparison of head-related transfer functions". In: *The Journal of the Acoustical Society of America* 140.5, pp. 3623–3632. ISSN: 0001-4966. DOI: `10.1121/1.4966115`.

Skottun, B. C. et al. (2001). "The ability of inferior colliculus neurons to signal differences in interaural delay." In: *Proceedings of the National Academy of Sciences of the United States of America* 98.24, pp. 14050–4. ISSN: 0027-8424. DOI: `10.1073/pnas.241513998`.

Slater, M. (2004). "How Colorful Was Your Day? Why Questionnaires Cannot Assess Presence in Virtual Environments". In: *Presence: Teleoperators and Virtual Environments* 13.4, pp. 484–493. ISSN: 1054-7460. DOI: `10.1162/1054746041944849`.

Smyth Research (2018a). *Realiser A16*. URL: `https://smyth-research.com/a16/`.

— (2018b). *Realiser A8*. URL: `https://smyth-research.com/a8/` (visited on 10/27/2018).

Smyth, S. (2005). *Patent US2006/0045294 A1 – Personalized Headphone Virtualization*.

Smyth, S., M. Smyth, and S. Cheung (2008). "Headphone Surround Monitoring for Studios". In: *23rd UK Conference of the Audio Engineering Society*. Audio Engineering Society, pp. 1–7.

SOFA (2018). *SOFA Database*. URL: `http://sofacoustics.org/data/database/` (visited on 10/28/2018).

Søndergaard, P. L. and P. Majdak (2013). "The Auditory Modeling Toolbox". In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 33–56. ISBN: 978-3-642-37762-4. DOI: `10.1007/978-3-642-37762-4_2`.

Spagnol, S., M. Geronazzo, and F. Avanzini (2013). "On the Relation Between Pinna Reflection Patterns and Head-Related Transfer Function Features". In: *IEEE Transactions on Audio, Speech, and Language Processing* 21.3, pp. 508–519. ISSN: 1558-7916. DOI: `10.1109/TASL.2012.2227730`.

Spagnol, S., D. Rocchesso, and F. Avanzini (2013). "Extraction of Pinna Features for Customized Binaural Audio Delivery on Mobile Devices". In: *International Conference on Advances in Mobile Computing & Multimedia*, pp. 2–4. ISBN: 9781450321068.

Spors, S. and J. Ahrens (2008). “A Comparison of Wave Field Synthesis and Higher-Order Ambisonics with Respect to Physical Properties and Spatial Sampling”. In: *125th AES Convention*. San Francisco, CA, USA: Audio Engineering Society, pp. 1–17.

Spors, S., H. Wierstorf, et al. (2013). “Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State”. English. In: *Proceedings of the IEEE* 101.9, pp. 1920–1938. ISSN: 0018-9219. DOI: `10.1109/JPROC.2013.2264784`.

Sproule, D. S. (2018). “An Evaluation of Ambisonic and VR Microphones”. MSc. Birmingham City Unviersity.

Sra, S. (2012). “A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$”. In: *Computational Statistics* 27, pp. 177–190. DOI: `10.1007/s00180-011-0232-x`.

Stanislaw, H. and N. Todorov (1999). “Calculation of signal detection theory measures”. In: *Behavior Research Methods, Instruments, & Computers* 31.1, pp. 137–149. ISSN: 0743-3808. DOI: `10.3758/BF03207704`.

Starr, G. E. (1997). “Interference effects in short-term memory for timbre”. In: *The Journal of the Acoustical Society of America* 102.1, p. 486. ISSN: 00014966. DOI: `10.1121/1.419722`.

Steam (2018). *Steam Audio*. URL: `https://valvesoftware.github.io/steam-audio/` (visited on 11/18/2018).

Stone, H. et al. (1974). “Sensory Evaluation by Quantitative Descriptive Analysis”. In: *Descriptive Sensory Analysis in Practice*. Vol. 28. 11. Trumbull, Connecticut, USA: Food & Nutrition Press, Inc., pp. 23–34. DOI: `10.1002/9780470385036.ch1c`.

Strohmeier, D., S. Jumisko-Pyykkö, and K. Kunze (2010). “Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception”. In: *Advances in Multimedia*, pp. 1–28. ISSN: 1687-5680. DOI: `10.1155/2010/658980`.

Strohmeier, D., S. Jumisko-Pyykkö, K. Kunze, and M. O. Bici (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. ISSN: 1687-5176. DOI: `10.1155/2011/538294`.

Su, H., A. Marui, and T. Kamekawa (2017). “Frequency bands distribution for virtual source widening in binaural synthesis”. In: *143rd AES Convention*. New York, NY, USA: Audio Engineering Society.

Sundareswara, R. and P. Schrater (2003). “Extensible point location algorithm”. In: *2003 International Conference on Geometric Modeling and Graphics, 2003. Proceedings*, pp. 84–89. DOI: `10.1109/GMAG.2003.1219670`.

Supper, B. (2010). “Processing and Improving a Head-Related Impulse Response Database for Auralization”. In: *Audio Engineering Society Convention*. San Francisco, CA, USA: Audio Engineering Society.

Takanen, M., M. Hiipakka, and V. Pulkki (2012). “Audibility of coloration artifacts in HRTF filter designs”. In: *Proceedings of 45th AES International Conference*. Helsinki, Finland: Audio Engineering Society. ISBN: 9781622760077.

Tate, M. W. and S. M. Brown (1970). “Note on the Cochran Q Test”. In: *Journal of the American Statistical Association* 65.329, pp. 155–160. ISSN: 01621459. DOI: `10.2307/2283582`.

Taubin, G. (2011). “3D Rotations”. In: *IEEE Computer Graphics and Applications* 31.6, pp. 84–89. ISSN: 0272-1716. DOI: `10.1109/MCG.2011.92`.

Temme, S. et al. (2014). “The Correlation Between Distortion Audibility and Listener Preference in Headphones”. In: *137th AES Convention*. Los Angeles, CA, USA: Audio Engineering Society. ISBN: 9781634397483. URL: `http://www.aes.org/e-lib/browse.cfm?elib=17441`.

Tew, A. I., C. T. Hetherington, and J. B. A. Thorpe (2012). “Morphoacoustic perturbation analysis : principles and validation”. In: *Proceedings of Acoustics*. Nantes, France.

Theile, G. (1986). “On the Standardization of the Frequency Response of High-Quality Studio Headphones”. English. In: *Journal of the Audio Engineering Society* 34.12, pp. 956–969. URL: `http://www.aes.org/e-lib/browse.cfm?elib=5233`.

Theile, G. and G. Plenge (1977). “Localization of Lateral Phantom Sources”. In: *Journal of the Audio Engineering Society* 25.4, pp. 196–200. URL: `http://www.aes.org/e-lib/browse.cfm?elib=3376`.

Theile, G. and H. Wittek (2011). “Principles in Surround Recordings with Height”. In: *International Conference on Spatial Audio*. Detmold, Germany, pp. 527–543.

Thomson, D. M. and J. A. McEwan (1988). “An application of the repertory grid method to investigate consumer perceptions of foods”. In: *Appetite* 10.3, pp. 181–193. ISSN: 01956663. DOI: `10.1016/0195-6663(88)90011-6`.

Thresh, L., C. Armstrong, and G. Kearney (2017). “A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering”. In: *143rd AES Convention*. New York, NY, USA: Audio Engineering Society, pp. 1–9.

Thurlow, W. R., J. W. Mangels, and P. S. Runge (1967). “Head Movements During Sound Localization”. In: *The Journal of the Acoustical Society of America* 42.2, pp. 489–493. ISSN: 0001-4966. DOI: `10.1121/1.1910605`.

Toole, F. E. (1985). “Subjective Measurements of Loudspeaker Sound Quality and Listener Performance”. English. In: *Journal of the Audio Engineering Society* 33.1/2, pp. 2–32. URL: `http://www.aes.org/e-lib/browse.cfm?elib=4465`.

— (2008). *Sound Reproduction*. Focal Press. ISBN: 9780240520094.

Torres-Gallegos, E. A., F. Orduña-Bustamante, and F. Arámbula-Cosío (2015). “Personalization of head-related transfer functions (HRTF) based on automatic photo-anthropometry and inference from a database”. In: *Applied Acoustics* 97, pp. 84–95. ISSN: 1872910X. DOI: `10.1016/j.apacoust.2015.04.009`.

Travis, C. (1996). “A Virtual Reality Perspective on Headphone Audio”. In: *AES UK Conference on Audio for New Media*. Audio Engineering Society. URL: `http://www.aes.org/e-lib/browse.cfm?elib=7425`.

Tsilfidis, A. et al. (2013). “Binaural Dereverberation”. In: *The Technology of Binaural Listening*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 359–396. DOI: `10.1007/978-3-642-37762-4_14`.

Twells, J. (2013). *The Sounds of Time: 50 Years of Doctor Who’s influential music and SFX | FACT Mag*. URL: `https://www.factmag.com/2013/11/23/the-sounds-of-time-50-years-of-doctor-whos-influential-music-and-sfx/` (visited on 12/10/2018).

Udesen, J., T. Piechowiak, and F. Gran (2014). “Vision Affects Sound Externalization”. In: *Audio Engineering Society 55th International Conference*. Helsinki, Finland: Audio Engineering Society, pp. 27–30.

Valentin, D. et al. (2012). “Quick and dirty but still pretty good: A review of new descriptive methods in food science”. In: *International Journal of Food Science and Technology* 47.8, pp. 1563–1578. ISSN: 09505423. DOI: `10.1111/j.1365-2621.2012.03022.x`.

Välimäki, V. and T. Laakso (2000). “Principles of fractional delay filters”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* 6.June, pp. 3870–3873. DOI: `10.1109/ICASSP.2000.860248`.

Välimäki, V., J. D. Parker, et al. (2012). “Fifty Years of Artificial Reverberation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.5, pp. 1421–1448. ISSN: 1558-7916. DOI: `10.1109/TASL.2012.2189567`.

Väljamäe, A. et al. (2004). “Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments”. In: *Proc. of Seventh Annual Workshop Presence 2004*, pp. 141–147.

Varela, P. and G. Ares, eds. (2016). *Novel techniques in sensory characterization and consumer profiling*. 1st. CRC Press. ISBN: 9781466566293.

VICON (2013). *Bonita motion capture systems*. URL: https://www.vicon.com/products/archived-products/bonita (visited on 11/15/2018).

Vidal, L. et al. (2018). “Comparison of rate-all-that-apply (RATA) and check-all-that-apply (CATA) questions across seven consumer studies”. In: *Food Quality and Preference* 67, pp. 49–58. ISSN: 09503293. DOI: 10.1016/j.foodqual.2016.12.013.

Vilkamo, J., T. Lokki, and V. Pulkki (2009). “Directional Audio Coding : Virtual Microphone-Based Synthesis and Subjective Evaluation *”. In: *Journal of the Audio Engineering Society* 57.9, pp. 709–724. ISSN: 15494950.

Vilkamo, J., B. Neugebauer, and J. Plogsties (2010). “Sparse Frequency-Domain Reverberator”. In: *AES 40th International Conference: Spatial Sound*. Tokyo, Japan: Audio Engineering Society, pp. 3–11.

Volk, C. P. et al. (2016). “Identifying the dominating perceptual differences in headphone reproduction”. In: *The Journal of the Acoustical Society of America* 140.5, pp. 3664–3674. ISSN: 0001-4966. DOI: 10.1121/1.4967225.

Völk, F. (2009). “Externalization in data-based Binaural Synthesis: Effects of Impulse Response Length”. In: *NAG/DAGA*. Rotterdam, The Netherlands.

Völk, F., F. Heinemann, and H. Fastl (2008). “Externalization in binaural synthesis: effects of recording environment and measurement procedure”. In: *Proceedings of Acoustics 2008*. Paris, France, pp. 6421–6426.

VRVCA (2018). *VR/AR Global Investment Report & Outlook 2018*. Tech. rep. Virtual Reality Venture Capital Alliance. URL: www.vrvca.com.

W3C (2018). *Web Audio API – W3C Candidate Recommendation*. URL: https://www.w3.org/TR/webaudio/.

Wade, N. J. and D. Detutsch (2008). “Binaural Hearing - Before and After the Stethophone”. In: *Acoustics Today* 4.3, pp. 16–27.

Wallach, H. (1939). “On Sound Localization”. In: *The Journal of the Acoustical Society of America* 10.4, pp. 270–274. ISSN: 0001-4966. DOI: 10.1121/1.1915985.

Wältermann, M., A. Raake, and S. Möller (2010). “Quality Dimensions of Narrowband and Wideband Speech Transmission”. In: *Acta Acustica united with Acustica* 96.6, pp. 1090–1103. DOI: 10.3813/AAA.918370.

— (2012). “Direct Quantification of Latent Speech Quality Dimensions”. In: *Journal of the Audio Engineering Society* 60.4, pp. 246–254. URL: http://www.aes.org/e-lib/browse.cfm?elib=16218.

Walton, T. (2017). “The Overall Listening Experience of Binaural Audio”. In: *International Conference on Spatial Audio*. Graz, Austria.

Walton, T. and M. Evans (2018). “The role of human influence factors on overall listening experience”. In: *Quality and User Experience* 3.1, pp. 1–16. ISSN: 2366-0139. DOI: 10.1007/s41233-017-0015-4.

Walton, T., M. J. Evans, D. Kirk, et al. (2016a). “A Subjective Comparison of Discrete Surround Sound and Soundbar Technology by Using Mixed Methods”. In: *AES 140th Convention*. Paris, France: Audio Engineering Society.

— (2016b). “Does environmental noise influence preference of background-foreground audio balance?” In: *141st AES Convention*. Los Angeles, CA, USA: Audio Engineering Society.

Ward, J. H. (1963). “Hierarchical Grouping to Optimize an Objective Function”. In: *Journal of the American Statistical Association* 58.301, pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

Watanabe, K. et al. (2014). “Dataset of head-related transfer functions measured with a circular loudspeaker array”. In: *Acoustical Science and Technology* 35.3, pp. 159–165. ISSN: 1346-3969. DOI: 10.1250/ast.35.159.

Wechsung, I. and K. De Moor (2014). “Quality of Experience Versus User Experience”. In: *Quality of Experience*. Ed. by S. Möller and A. Raake. Springer. Chap. 3, pp. 35–54. DOI: 10.1007/978-3-319-02681-7_3.

Wefers, F. (2014). “Partitioned convolution algorithms for real-time auralization”. PhD thesis. Aachen University. ISBN: 9783832539436.

Weiping, T. et al. (2010). “Measurement and analysis of just noticeable difference of interaural level difference cue”. In: *2010 International Conference on Multimedia Technology, ICMT 2010*, pp. 5–7. DOI: 10.1109/ICMULT.2010.5630980.

Welch, G. and E. Foxlin (2002). “Motion tracking: No silver bullet, but a respectable arsenal”. In: *IEEE Computer Graphics and Applications* 22.6, pp. 24–38. ISSN: 02721716. DOI: 10.1109/MCG.2002.1046626.

Welti, T., S. E. Olive, and O. Khonsaripour (2016). “Validation of a Virtual In-ear Headphone Listening Test Method”. In: *AES 141st Convention*. Los Angeles, CA, USA: Audio Engineering Society.

Wendt, T., S. van de Par, and S. D. Ewert (2014). “A Computationally-Efficient and Perceptually-Plausible Algorithm for Binaural Room Impulse Response Simulation”. In: *Journal of the Audio Engineering Society* 62.11, pp. 748–766. URL: http://www.aes.org/e-lib/browse.cfm?elib=17550.

Wenzel, E. M. (1997). “Analysis of the Role of Update Rate and System Latency in Interactive Virtual Acoustic Environments”. In: *103rd AES Convention*. New York, NY, USA: Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=7146.

— (2001). “Effect of Increasing System Latency on Localisation of Virtual Sounds with Short and Long Duration”. In: *International Conference on Auditory Display*. Espoo, Finland, pp. 185–190.

Wenzel, E. M., M. Arruda, et al. (1993a). “Localisation using non-individualised head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123.

— (1993b). “Localization using nonindividualized head‐related transfer functions”. In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123. ISSN: 0001-4966. DOI: 10.1121/1.407089.

Wenzel, E. M. and S. H. Foster (1993). “Perceptual consequences of interpolating head-related transfer functions during spatial synthesis”. In: *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 102–105. ISBN: 0-7803-2078-6. DOI: 10.1109/ASPAA.1993.379986.

Wenzel, E. M., F. L. Wightman, and S. H. Foster (1988). “A Virtual Display System for Conveying Three-Dimensional Acoustic Information”. In: *Proceedings of the Human Factors Society Annual Meeting* 32.2, pp. 86–90. ISSN: 0163-5182. DOI: 10.1177/154193128803200218.

Wenzel, E. (1995). “The relative contribution of interaural time and magnitude cues to dynamic sound localization”. In: *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE, pp. 80–83. ISBN: 0-7803-3064-1. DOI: 10.1109/ASPAA.1995.482963.

Werner, S., F. Klein, et al. (2016). “A summary on acoustic room divergence and its effect on externalization of auditory events”. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: 10.1109/QoMEX.2016.7498973.

Werner, S. and J. Liebetrau (2013). “Effects of shaping of binaural room impulse responses on localization”. English. In: *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 88–93. ISBN: 978-1-4799-0738-0. DOI: 10.1109/QoMEX.2013.6603216.

Wersenyi, G. (2009). “Effect of Emulated Head-Tracking for Reducing Localization Errors in Virtual Audio Simulation”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.2, pp. 247–252. ISSN: 1558-7916. DOI: 10.1109/TASL.2008.2006720.

Wickelmaier, F. et al. (2009). “Comparing three methods for sound quality evaluation with respect to speed and accuracy”. In: *126th AES Convention*. Munich, Germany: Audio Engineering Society.

Wierstorf, H. (2014). “Perceptual Assessment of Sound Field Synthesis”. PhD thesis. Technische Universität Berlin. DOI: 10.14279/depositonce-4310.

Wierstorf, H., C. Hohnerlein, et al. (2014). “Coloration in Wave Field Synthesis”. In: *AES 55th International Conference*. Helsinki, Finland: Audio Engineering Society.

Wierstorf, H., A. Raake, M. Geier, et al. (2013). "Perception of Focused Sources in Wave Field Synthesis". In: *Journal of the Audio Engineering Society* 61.1/2, pp. 5–16.

Wierstorf, H., A. Raake, and S. Spors (2013). "Binaural Assessment of Multichannel Reproduction". In: *The Technology of Binaural Listening*. Ed. by J. Blauert. Springer. Chap. 10. ISBN: 9783642377624. DOI: 10.1007/978-3-642-37762-4.

Wierstorf, H. and S. Spors (2012). "Sound Field Synthesis Toolbox". In: *132nd AES Convention*. Budapest, Hungary: Audio Engineering Society, pp. 2–5.

Wierstorf, H., S. Spors, and A. Raake (2012). "Perception and evaluation of sound fields". In: *59th Open Seminar on Acoustics*.

Wiggins, B. (2004). "An investigation into the real-time manipulation and control of three-dimensional sound fields". PhD thesis. University of Derby.

Wiggins, B., I. Paterson-Stephens, and V. Lowndes (2003). "The design and optimisation of surround sound decoders using heuristic methods". In: *Proceedings of UKSim*, pp. 2–9.

Wightman, F. L. and D. J. Kistler (1989a). "Headphone simulation of free-field listening. I: Stimulus synthesis". In: *The Journal of the Acoustical Society of America* 85.2, pp. 858–867. ISSN: 0001-4966. DOI: 10.1121/1.397557.

— (1989b). "Headphone simulation of free-field listening. II: Psychophysical validation". In: *The Journal of the Acoustical Society of America* 85.2, pp. 868–878. ISSN: 0001-4966. DOI: 10.1121/1.397558.

— (1992). "The dominant role of low-frequency interaural time differences in sound localization". In: *The Journal of the Acoustical Society of America* 91.3, pp. 1648–1661. ISSN: 0001-4966. DOI: 10.1121/1.402445.

— (1999). "Resolution of front–back ambiguity in spatial hearing by listener and source movement". In: *The Journal of the Acoustical Society of America* 105.5, pp. 2841–2853. ISSN: 0001-4966. DOI: 10.1121/1.426899.

Williams, A. A. and S. P. Langron (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. ISSN: 1097-0010. DOI: 10.1002/jsfa.2740350513.

Williams, M. (1991). "Microphone Arrays for Natural Multiphony". In: *Proceedings of 91st AES Convention*. Paris, France.

Winkler, I. and N. Cowan (2005). "From Sensory to Long-Term Memory". In: *Experimental Psychology* 52.1, pp. 3–20. ISSN: 1618-3169. DOI: 10.1027/1618-3169.52.1.3.

Wisniewski, M. G. et al. (2016). "Enhanced auditory spatial performance using individualized head-related transfer functions: An event-related potential study". In: *The Journal of the Acoustical Society of America* 140.6, EL539–EL544. ISSN: 0001-4966. DOI: 10.1121/1.4972301.

Woodcock, J., C. Pike, P. Coleman, et al. (2016). *S3A Object-based Audio Drama Dataset*. DOI: 10.17866/rd.salford.3043921.

Woodcock, J., C. Pike, F. Melchior, et al. (2016). "Presenting the S3A Object-Based Audio Drama dataset". In: *140th AES Convention*. Paris, France: Audio Engineering Society.

Worch, T. et al. (2013). "Ideal Profile Method (IPM): The ins and outs". In: *Food Quality and Preference* 28.1, pp. 45–59. ISSN: 09503293. DOI: 10.1016/j.foodqual.2012.08.001.

Wozny, D. R., U. R. Beierholm, and L. Shams (2008). "Human trimodal perception follows optimal statistical inference". In: *Journal of Vision* 8.3, p. 24. ISSN: 1534-7362. DOI: 10.1167/8.3.24.

Xiao, T. and Q. Huo Liu (2003). "Finite difference computation of head-related transfer function for human hearing". In: *The Journal of the Acoustical Society of America* 113.5, pp. 2434–2441. ISSN: 0001-4966. DOI: 10.1121/1.1561495.

Xie, B., X. Zhong, and N. He (2015). “Typical data and cluster analysis on head-related transfer functions from Chinese subjects”. In: *Applied Acoustics* 94, pp. 1–13. ISSN: 0003682X. DOI: `10.1016/j.apacoust.2015.01.022`.

Xu, S. and Z. Li (2007). “Individualization of head-related transfer function for three-dimensional virtual auditory display: A review”. In: *International Conference on Virtual Reality* 4563, pp. 397–407. URL: `http://www.springerlink.com/index/D024G2P2717053M4.pdf`.

Xu, S., Z. Li, and G. Salvendy (2008). “Improved method to individualize head-related transfer function using anthropometric measurements”. In: *Acoustical Science and Technology* 29.6, pp. 388–390. ISSN: 1347-5177. DOI: `10.1250/ast.29.388`.

Yairi, S., Y. Iwaya, and Y. Suzuki (2007). “Estimation of detection threshold of system latency of virtual auditory display”. In: *Applied Acoustics* 68.8, pp. 851–863. ISSN: 0003682X. DOI: `10.1016/j.apacoust.2006.12.005`.

YouGov (2017). *VR: A Deeper Perspective Study*. Tech. rep. URL: `https://yougov.co.uk/topics/consumer/articles-reports/2017/05/19/vr-headsets-more-popular-tablets-and-wearables-wer`.

YouTube (2018). *YouTube Help: Use spatial audio in 360-degree and VR videos*. URL: `https://support.google.com/youtube/answer/6395969` (visited on 06/25/2018).

Yu, G. et al. (2018). “Near-field head-related transfer-function measurement and database of human subjects”. In: *The Journal of the Acoustical Society of America* 143.3, EL194–EL198. ISSN: 0001-4966. DOI: `10.1121/1.5027019`.

Zacharov, N. and K. Koivuniemi (2001a). “Unravelling the perception of spatial sound reproduction: Analysis & external preference mapping”. In: *AES 111th Convention*. New York, NY, USA: Audio Engineering Society.

— (2001b). “Unravelling the perception of spatial sound reproduction: Techniques and experimental design”. In: *AES 19th International Conference*. Schloss Elmau, Germany: Audio Engineering Society.

Zacharov, N. and T. H. Pedersen (2015). “Spatial sound attributes - development of a common lexicon”. In: *139th AES Convention*. New York, NY, USA: Audio Engineering Society.

Zacharov, N., T. Pedersen, and C. Pike (2016). “A common lexicon for spatial sound quality assessment - latest developments”. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: `10.1109/QoMEX.2016.7498967`.

Zacharov, N., C. Pike, et al. (2016a). “Next generation audio system assessment using the multiple stimulus ideal profile method”. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: `10.1109/QoMEX.2016.7498966`.

— (2016b). “Next generation audio system assessment using the multiple stimulus ideal profile method”. In: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, pp. 1–6. ISBN: 978-1-5090-0354-9. DOI: `10.1109/QoMEX.2016.7498966`.

Zacharov, N., C. P. Volk, and T. Stegenborg-Andersen (2017). “Comparison of hedonic and quality rating scales for perceptual evaluation of high-and intermediate-quality stimuli”. In: *AES 143rd Convention*. New York, USA. URL: `http://www.aes.org/e-lib/browse.cfm?elib=19276`.

Zahorik, P. (2000). “Distance localization using nonindividualized head￾related transfer functions”. In: *The Journal of the Acoustical Society of America* 108.5, pp. 2597–2597. ISSN: 0001-4966. DOI: `10.1121/1.4743664`.

— (2002a). “Assessing auditory distance perception using virtual acoustics”. In: *The Journal of the Acoustical Society of America* 111.4, pp. 1832–1846. ISSN: 0001-4966. DOI: `10.1121/1.1458027`.

— (2002b). “Auditory display of sound source distance”. In: *Proceedings of the International Conference on Auditory Display*. Kyoto, Japan.

— (2009). “Perceptually relevant parameters for virtual listening simulation of small room acoustics”. In: *The Journal of the Acoustical Society of America* 126.2, pp. 776–791. ISSN: 0001-4966. DOI: `10.1121/1.3167842`.

Zahorik, P., P. Bangayan, et al. (2006). “Perceptual recalibration in human sound localization: Learning to remediate front-back reversals”. In: *The Journal of the Acoustical Society of America* 120.1, pp. 343–359. ISSN: 0001-4966. DOI: 10.1121/1.2208429.

Zahorik, P., D. S. Brungart, and A. W. Bronkhorst (2005). “Auditory Distance Perception in Humans: A Summary of Past and Present Research”. In: *Acta Acustica united with Acustica* 91.3, pp. 409–420.

Zahorik, P., D. J. Kistler, et al. (1994). “Sound localization in varying virtual acoustic environments”. In: *International Conference on Auditory Display*.

Zahorik, P., F. Wightman, and D. Kistler (1995). “On the discriminability of virtual and real sound sources”. In: *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE, pp. 76–79. ISBN: 0-7803-3064-1. DOI: 10.1109/ASPAA.1995.482951.

Zahorik, P. and F. L. Wightman (2001). “Loudness constancy with varying sound source distance”. In: *Nature Neuroscience* 4.1, pp. 78–83. ISSN: 1097-6256. DOI: 10.1038/82931.

Zaunschirm, M., M. Frank, and F. Zotter (2018). “BRIR synthesis using first-order microphone arrays”. In: *AES 144th Convention*. Milan, Italy: Audio Engineering Society.

Zaunschirm, M., C. Schörkhuber, and R. Höldrich (2018). “Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint”. In: *The Journal of the Acoustical Society of America* 143.6, pp. 3616–3627. ISSN: 0001-4966. DOI: 10.1121/1.5040489.

Zhang, W. et al. (2011). “On high resolution head-related transfer function measurements: An efficient sampling scheme”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.2, pp. 575–584. ISSN: 1558-7916. DOI: 10.1109/TASL.2011.2162404.

Ziegelwanger, H., W. Kreuzer, and P. Majdak (2015). “Mesh2HRTF: An open-source software package for the numerical calculation of head-related transfer functions”. In: *22nd International Congress on Sound and Vibration*. July, pp. 1–8. ISBN: 9788888942483. DOI: 10.13140/RG.2.1.1707.1128.

Ziegelwanger, H. and P. Majdak (2014). “Modeling the direction-continuous time-of-arrival in head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 135.3, pp. 1278–1293. ISSN: 0001-4966. DOI: 10.1121/1.4863196.

Ziegelwanger, H., P. Majdak, and W. Kreuzer (2015). “Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization”. In: *The Journal of the Acoustical Society of America* 138.1, pp. 208–222. ISSN: 0001-4966. DOI: 10.1121/1.4922518.

Zieliński, S. K., F. Rumsey, and S. Bech (2002). “Subjective audio quality trade-offs in consumer multichannel audio-visual delivery systems Part I: Effects of high frequency limitation.” In: *Proceedings of 112th AES Convention*, pp. 1–18.

Zieliński, S. K., F. Rumsey, R. Kassier, et al. (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. URL: http://www.aes.org/e-lib/browse.cfm?elib=13407.

Zotkin, D. N., R. Duraiswami, and N. A. Gumerov (2009). “Regularized HRTF fitting using spherical harmonics”. In: *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 257–260. ISBN: 978-1-4244-3678-1. DOI: 10.1109/ASPAA.2009.5346521.

Zotkin, D., J. Hwang, et al. (2003). “HRTF personalization using anthropometric measurements”. In: *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, pp. 157–160. ISBN: 0-7803-7850-4. DOI: 10.1109/ASPAA.2003.1285855.

Zotter, F. (2009). “Analysis and synthesis of sound-radiation with spherical arrays”. PhD thesis. University of Music and Performing Arts, Austria.

Zotter, F. and M. Frank (2012). “All-round ambisonic panning and decoding”. In: *Journal of the Audio Engineering Society* 60.10, pp. 807–820. ISSN: 15494950. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16554`.

— (2013). “Efficient phantom source widening”. In: *Archives of Acoustics* 38.1, pp. 27–37. ISSN: 01375075. DOI: `10.2478/aoa-2013-0004`.

Zotter, F., H. Pomberger, and M. Frank (2009). “An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Spatial Aliasing”. In: *126th AES Convention*. Munich, Germany: Audio Engineering Society.

Zotter, F., H. Pomberger, and M. Noisternig (2010). “Ambisonic Decoding With And Without Mode-Matching: A Case Study Using the Hemisphere”. In: *International Symposium on Ambisonics and Spherical Acoustics*.

— (2012). “Energy-Preserving Ambisonic Decoding”. In: *Acta Acustica united with Acustica* 98.1, pp. 37–47. ISSN: 1610-1928. DOI: `10.3813/AAA.918490`.