SIMULATION AND ANALYSIS OF SPATIAL AUDIO REPRODUCTION AND LISTENING AREA EFFECTS

Darius Satongar

School of Computing, Science and Engineering

University of Salford, UK

Submitted in Partial Fulfilment of the Requirements of the Degree of

Doctor of Philosophy

November 5, 2016

Abstract

Loudspeaker-based spatial audio systems are often designed with the aim to create an auditory event or scene to a listener positioned in the optimal listening position. However, in real-world domestic listening environments, listeners can be distributed across the listening area. Any translational change from the central listening position will introduce artefacts which can be challenging to evaluate perceptually. Simulation of a loudspeaker system using non-individualised dynamic binaural synthesis is one solution to this problem. However, the validity in using such systems is not well proven.

This thesis measures the limitations of using a non-individualised, dynamic binaural synthesis system to simulate the perception of loudspeaker-based panning methods across the listening area. The binaural simulation system was designed and verified in collaboration with BBC Research and Development. The equivalence of localisation errors caused by loudspeaker-based panning methods between in situ and binaural simulation was measured where it was found that localisation errors were equivalent to a +/-7° boundary in 75% of the spatial audio reproduction systems tested. Results were then compared to a computation localisation model which was adapted to utilise head-rotations. The equivalence of human acuity to sound colouration between in situ and when using non-individualised binaural simulation was measured using colouration detection thresholds from five directions. It was shown that thresholds were

equivalent within a +/-4 dB equivalence boundary, supporting the use for simulating sound colourations caused by loudspeaker-based panning methods. The binaural system was finally applied to measure the perception of multi-loudspeaker induced colouration artefacts across the listening area. It was found that the central listening position had the lowest perceived colouration. It is also shown that the variation in perceived colouration across the listening area is larger for reverberant reproduction conditions.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

This chapter provides an introduction to the research presented in this thesis. The research topic is defined and the general motivation behind the work is also introduced. Aims and objectives are explicitly presented followed by an itemised list of the contributions and publications resulting from this work.

1.1 Introduction

The subject of research presented in this thesis is the simulation and perception of domestic, loudspeaker-based spatial audio systems across the listening area. Spatial and timbral attributes of auditory perception have been shown to be important factors when considering overall perceived quality (Rumsey et al., 2005; Bech and Zacharov, 2006). Unfortunately, these two attributes are dramatically affected by changes in the listener's positioning within the listening area (Peters, 2010). These artefacts are caused predominantly by changes in the time-of-flight between loudspeakers, but other physical features also have an impact on the perception of loudspeaker panning methods. The physical causes and effects can be measured objectively but the perceptual relevance of these listening area effects is still not well understood (Ahrens, 2012, p. 14).

In this thesis, a state-of-the-art dynamic binaural synthesis system has been developed, validated and implemented to allow for the subjective evaluation of audio reproduction systems when a listener is virtually placed at positions across the listening area. The system has been designed to provide plausible auditory events that induce the main perceptual cues at off-centre listening positions. To enable this, a spatially-sampled binaural room impulse response dataset has been established. However, although binaural simulation methods have become a common tool in recent years, the perceptual validity in virtualising listening tests using non-individualised dynamic binaural simulation has yet to be defined. Validity in using a such a system for the assessment of spatial and timbral artefacts introduced by domestic loudspeaker systems is firstly addressed directly in this research. This data provides important information for researchers aiming to virtualise listening tests using binaural simulation. Results from the

simulation of spatial artefacts are also compared against a dynamic computational localisation model, which has novel developments to simulate a closed-loop localisation task.

For timbral fidelity, it can be predicted that comb-filtering will be introduced at non-central listening positions using loudspeaker panning methods (Solvang, 2008), one reason for this is the summation of delayed, coherent sound sources. However, even at the central listening position it can be shown that comb-filtering artefacts are present due to the spatial offset of the human ears (Toole, 2008, p. 151). Due to the questions raised by this problem, the binaural simulation system is finally applied to determine the subjective response to colouration artefacts across the listening area for two loudspeaker-based panning methods in both echoic and anechoic conditions. A graphical representation of the main technical chapters in the thesis is presented in Figure. 1.1.

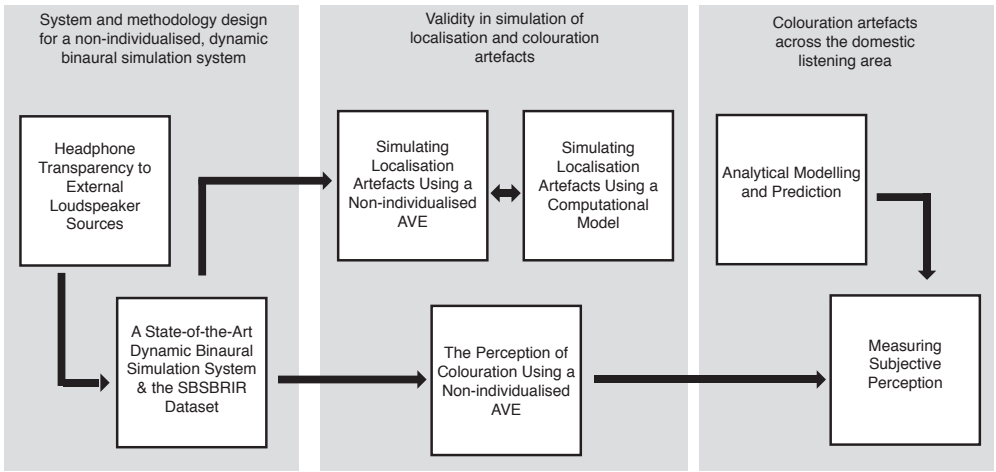


Figure 1.1: A graphical representation of the main technical chapters. The acronym AVE refers to an auditory virtual environment created using the binaural reproduction system defined in Chapter 4.

The purpose of this chapter is to define the motivation behind the research, set

the aims and objectives and finally state the scientific contributions generated by this work.

1.2 Motivation

Following from the stereophonic systems designed in the early 20th century (Blumlein, 1931; Steinberg and Snow, 1934), loudspeaker-based spatial audio systems have become a popular method of creating auditory scenes with spatial attributes. Many signal processing schemes have been developed to create auditory events at spatial locations where no physical sound sources are present (see Chapter. 2 for a summary). However, due to the limited number of loudspeakers that can be used for the reproduction, artefacts are introduced.

The term 'sweet spot' is applied in many different contexts, often referring to a location or condition where an optimal output of a measurable quantity can be observed. For loudspeaker-based spatial audio reproduction, the sweet spot can be defined as the spatial position or region within the listening area that induces the optimal intended physical quantities or subjective perception. The sweet spot in the context of loudspeaker reproduction is commonly known as the position of equidistance from each of the contributing loudspeakers (Toole, 2008).

For any given loudspeaker layout, many signal processing algorithms exist to control the amplitude and phase of loudspeaker signals to achieve auditory spatial perception by a listener; wave field synthesis (WFS) (Berkhout, 1988), Ambisonics (Gerzon, 1972) and Vector Base Amplitude Panning (VBAP) (Pulkki, 1997) are popular examples. Human perception at the central listening position of such systems has been widely tested for a range of perceptible

auditory attributes (see Section 3.3 for details of perceptual attributes). For these analyses, both objective and subjective metrics have been applied. However, the change in these attributes when a listener is placed at different positions across the domestic listening area remains largely untested. It is widely accepted that timbral fidelity represents one of the main auditory attributes contributing to overall audio quality (Bech and Zacharov, 2006). For this reason, the perception of colouration artefacts across the domestic listening area is addressed directly in this thesis using two popular panning methods.

Although an important area of research, the subjective evaluation of spatial audio systems across the listening area introduces a new set of practical problems. The subjective evaluation methodologies for off-centre listening can be separated into different types:

- In situ
- Artefact simulation
- Simulation of an auditory virtual environment (AVE)

In situ testing requires listeners to physically experience each of the tested conditions (listening positions, reverb, panning method). This is the most direct approach but is expensive and does not allow for direct-comparisons due to the physical time taken to adjust the geometry of the system (either by moving the listener or the loudspeakers). Another method would be to isolate the individual artefacts induced for off-centre listening and reintroduce them in a parametric way. This method allows for specific artefacts such as loudspeaker off-axis response or loudspeaker time-of-arrival to be compared against each other (Peters and McAdams, 2012).

Simulation systems such as binaural rendering over headphones provide listeners with an auditory virtual environment which will closely simulate the in situ environment to an almost indiscernible level of reality (plausibility) (Lindau and Weinzierl, 2012; Pike et al., 2014). In this scenario, the in situ environment is the loudspeaker-based spatial audio system perceived at a specific listening position reproduced in the same physical room as being simulated. Novo (2005) defines three types of AVE philosophies which are paraphrased below:

- *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 of plausibility constraints are imposed

To enable subjective evaluations, an authentic AVE is created allowing for the ability to perform direct, blind comparisons and have control over many independent variables.

Even though plausible auditory events can be created using dynamic binaural simulations (Lindau and Weinzierl, 2012; Pike et al., 2014), the perfect in situ reconstruction of acoustic pressure at the eardrum of a listener is not yet possible. Many practical factors contribute to the limitations of binaural simulation. Personalisation of the head-related transfer function and interaural cues, system latency, dynamic cues related to human movement, accuracy of room reflections and headphone design/coupling can be identified as some of the main limiting factors. The effect of these important factors also has a perceptual

impact which is not easily predicted. Therefore, in order to use non-individualised dynamic binaural simulation for the virtualisation of loudspeaker systems, the validity of the simulations presents a new set of research questions. More specifically, it should be shown that dominant perceptual artefacts are equivalently perceived when listening to loudspeakers in situ, or simulated using the AVE.

Studies presented in the early part of this thesis show the validity of using a non-individualised, dynamic binaural system for the simulation of off-centre listening position artefacts. The motivation for these studies is to allow for the use of such systems in the evaluation of spatial audio reproduction systems in more realistic listening conditions and further understand the perception of timbral and spatial artefacts. The validation of binaural simulation systems is a complex task in isolation, see studies by Møller et al. (1996), Begault et al. (2001), Völk (2012a) or Pike et al. (2014) for examples spanning the past two decades. For certain validation experiments, there are often situations when real loudspeakers must be evaluated whilst listening passively 'through' a headphone set used for binaural simulation[1]. Work conducted by the BBC in parallel to this project required that the binaural system under development be validated as 'plausible' in comparison to real loudspeakers. Therefore, the passive influence of headphones on the transmission path between loudspeakers and a human listener was considered specifically in the first contributory chapter of this thesis, Chapter. 5. The results from this experiment informed the method for plausibility studies conducted by the BBC (Pike et al., 2014) and also helped to define the experimental design of validation experiments in this research.

[1]consider a binaural plausibility study where real and virtual loudspeakers are auditioned and the listener must rate which is real/virtual

1.3 Aims

The following aims for the project are defined below.

1. To develop a perceptually valid non-individualised dynamic binaural synthesis system to simulate the perception of colouration and localisation artefacts induced at non-central listening positions in domestic, loudspeaker-based spatial audio systems.

2. To make objective predictions and determine the perceived magnitude of colouration artefacts induced at central *and* non-central listening positions in domestic, loudspeaker-based spatial audio systems.

1.4 Objectives

To achieve the aims defined above the following objectives have been set out for the research project.

1. Simulation of loudspeaker-based spatial audio reproduction at multiple listening positions
   (a) Establish and distribute the first freely available, spatially-sampled binaural room impulse response dataset
   (b) Determine the passive effect of headphones on external sounds for the application of binaural validation experiments
   (c) Design and verify a non-individualised, dynamic binaural synthesis system used to simulate loudspeakers in the BS.1116-1 compliant listening room at the University of Salford.
   (d) Determine the perceptual validity in using a non-individualised dynamic binaural simulation system to induce localisation artefacts found at non-central listening positions.
   (e) Determine the perceptual equivalence of acuity to sound colouration between in situ sound sources and simulations using a non-individualised dynamic binaural simulation system.
2. Prediction and analysis of colouration and localisation artefacts.
   (a) Use analytical models of sound propagation to predict listening area effects in loudspeaker-based spatial audio systems
   (b) Develop and validate an auditory localisation model to predict a closed-loop localisation task in anechoic and reverberant environments. This model will build upon the model implemented by Sheaffer (2013) using

the general approach of Faller and Merimaa (2004).

3. Perception of colouration across the listening area.

   (a) Determine the subjective perception of colouration artefacts across the listening area using a direct attribute scaling test.

   (b) Determine the subjective perception of colouration artefacts across the listening area using a indirect attribute scaling test. Specifically this test will compare against colouration predictions and compare the perception of colouration found at the central listening position.

   (c) Compare the differences in colouration perception across the listening area for VBAP and Ambisonic panning methods.

   (d) Compare the differences in colouration perception across the listening area for anechoic and reverberant reproduction conditions.

1.5 Publications and Original Contributions

Original contributions to the field of research have been created by work presented in this thesis. These contributions are explicitly defined below.

- Definition and distribution of the first freely available, spatially sampled binaural room impulse response dataset (SBSBRIR).
- Quantification of the passive filtering effect of headphones on external sound sources, applicable to situations of through-headphone listening tests and commercial headphone transparency applications.
- Using a selection of amplitude panning systems over two listening positions, the validity of using a non-individualised, dynamic binaural simulation system to simulate localisation artefacts caused by

loudspeaker-based spatial audio systems was defined.

- The change in colouration acuity has been determined for a non-individualised dynamic binaural simulation system. This will impact the possible use of similar systems for research into timbral fidelity.
- A computational localisation model has been adapted to use dynamic head-movements and validated for the prediction of a closed-loop localisation task in complex listening scenarios.
- The perceptual evaluation of coloration artefacts across the listening area using two panning methods was conducted. Results from two subjective evaluations indicated perceptual preference for colouration at non-central listening positions, however, colouration at the central listening position in amplitude panning systems is non-trivial.

The following publications have come as a direct result of work performed during this research project. The publications amount to 5 conference papers, 1 poster, 1 conference presentation, 2 internally reviewed white papers and 1 journal paper.

[1] D. Satongar, Y. W. Lam, F. F. Li, and C. Dunn, 'An Objective Investigation into the Auditory Localisation Cues Synthesised by Spatial Audio Systems' presented at the 9th International Symposium on Modern Acoustics (2012)

[2] D. Satongar, C. Pike, and Y. W. Lam, 'Psychoacoustic Evaluation of Spatial Audio Reproduction Systems'. Proceedings of the Institute of Acoustics . vol 34, no. 4 (2012)

[3] D. Satongar, C. Dunn, Y. W. Lam, and F. F. Li, 'Localisation Performance of Higher-Order Ambisonics for Off-Centre Listening' BBC White Paper WHP 254 (2013)

[4] D. Satongar, C. Pike, Y. W. Lam, and A. I. Tew, 'On the Influence of Headphones on Localisation of Loudspeaker Sources' presented at the 135th Audio Engineering Society Convention (2013)

[5] D. Satongar, C. Pike, Y. W. Lam, and A. I. Tew, 'On the Influence of Headphones on Localisation of Loudspeaker Sources' BBC White Paper WHP 276 (2013)

[6] D. Satongar, C. Pike, Y. W. Lam and F. F. Li, 'Measurement and Analysis of a Spatially Sampled Binaural Room Impulse Response Dataset' presented at the 21st International Congress on Sound and Vibration (2014)

[7] F. Melchior, D. Marston, C. Pike, D. Satongar, and Y. W. Lam, 'A Library of Binaural Room Impulse Responses and Sound Scenes for Evaluation of Spatial Audio Systems' poster presented at the 40th Annual German Congress on Acoustics (2014)

[8] D. Satongar, C. Pike, and Y. W. Lam, 'The Acuity of Colouration Perception Using Non-individualised Dynamic Binaural Synthesis' presented at the 22nd International Congress on Sound and Vibration (2015)

[9] D. Satongar, C. Pike, Y. W. Lam, and A. I. Tew, 'The Influence of Headphones on the Localisation of External Loudspeaker Sources' Journal of the Audio Engineering Society, vol. 63, no. 10 (2015 Oct.)

1.6 Thesis Structure

Chapter. 1 - Introduction

This chapter provides an introduction to the research presented in this thesis. The research topic is defined and the general motivation behind the work is also introduced. Aims and objectives are explicitly presented followed by an itemised list of the contributions and publications resulting from this work.

Chapter. 2 - General Concepts and Fundamental Theory

This chapter introduces general concepts, conventions and theory needed as a basis for the thesis. Firstly, naming conventions and mathematical nomenclature are introduced. The fundamental theory of spatial audio processing and perception are presented and discussed which will serve as a basis for the work presented in the main chapters of the thesis.

Chapter. 3 - Literature Review

This chapter presents a review of the previously documented literature that is relevant to the experiments presented in this thesis. The chapter reviews literature from both commercial and academic developments in spatial audio reproduction.

Chapter. 4 - A Non-individualised, Dynamic Binaural Simulation System

This chapter presents the specific design details on the non-individualised,

dynamic binaural simulation system used to create auditory virtual environments throughout this research project. System verification tests and specific details are presented to allow for experiments to be repeatable and applicable to future research.

Chapter. 5 - Headphone Transparency to External Loudspeaker Sources

This chapter presents experiments conducted into the passive effect of headphones on the transmission of sound from an external loudspeaker to a listener. Both physical measurements and a behaviour study was conducted to further understand the implications for binaural validation tests.

Chapter. 6 - Simulating Localisation Artefacts Across the Listening Area Using Non-individualised Dynamic Binaural Synthesis

This chapter presents experiments on the ability to use a non-individualised, dynamic binaural simulation system to simulate localisation artefacts in loudspeaker-based panning systems at central and non-central listening positions.

Chapter. 7 - Simulating Localisation Artefacts Across the Listening Area Using a Computational Model

This chapter presents the development to a computational localisation model proposed by Sheaffer (2013) built upon previous work by Faller and Merimaa (2004). The model is developed to include head/torso movements which resolve front-back confusions and in-turn, simulate a closed-loop localisation task in anechoic and reverberant environments. The current standing model is firstly introduced. Developments are then described before the model being validated against subjective data from Chapter. 6.

Chapter. 8 - The Perception of Colouration Using a Non-individualised Dynamic Binaural Simulation System

This chapter covers two experiments and accompanying analysis on the validity of using a non-individualised, dynamic binaural synthesis system to simulate colouration artefacts commonly found across the listening area. The colouration detection threshold (CDT) is a psychophysical metric commonly used to define a listener's acuity to changes in sound colour. Here, CDTs are measured for both in situ loudspeakers and auditory events created using the non-individualised dynamic binaural simulation system, using two assessment methods. CDTs are used to define the difference in colouration acuity between in situ and the simulation system.

Chapter. 9 - The Perception of Colouration Artefacts Across the Domestic Listening Area Using Loudspeaker-based Panning Methods

This chapter covers the results of two experiments implementing non-individualised dynamic binaural synthesis to measure the magnitude of perceived colouration found in spatial audio systems across the domestic listening area. In the second experiment, the comparison of colouration perception at central and non-central listening positions is considered specifically with analytical models to aid in analysis.

Chapter. 10 - Conclusions and Future Work

This chapter presents the general conclusions of the research activities presented in this thesis. Section 10.2 also presents a proposal for future research efforts following on from this research project.

1.7 Nomenclature

DT - Detection threshold

CDT - Colouration detection threshold

RDT - Reflection detection threshold

JND - Just-noticable difference

TD - Threshold of detection

AVE - Auditory virtual environment

in situ - Term used to describe the in situ real, or target system where auditory events are created by real sound events

GUI - Graphical user interface

CLP - Central listening position

VBAP - Vector-based amplitude panning

HOA - Higher-order Ambisonics

2.5-D - 2.5-Dimensional loudspeaker reproduction where loudspeakers are arranged in the horizontal plane, but inherently emit sound in 3-dimensions (spherical wave propagation)

NFC-HOA - Near-field compensated higher-order Ambisonics

LR2 / LR-2 - 2nd order Linkwitz-Riley filter

BRIR - Binaural room impulse response (time-domain)

BRTF - Binaural room transfer function (frequency-domain)

HRIR - Head-related impulse response (time-domain)

HRTF - Head-related transfer function (frequency-domain)

SBSBRIR - Salford/BBC spatially-sampled binaural room impulse response dataset

$max\, r_E$ - Energy maximised Ambisonic decoder

$max\, r_V$ - Velocity maximised Ambisonic decoder (basic, mode-matching)

θ_{LS} - Loudspeaker direction for CDT testing

θ_{direct} - Loudspeaker direction of the direct part of the feed-forward comb-filter

$\theta_{reflection}$ - Loudspeaker direction of the delayed part of the feed-forward comb-filter

BAQ - Basic audio quality

1.8 Conclusions

This chapter has introduced the main motivation, aims and objectives defined for the research project and thereby set the scope for the work presented within. It has been identified that the subjective evaluation of loudspeaker-based spatial audio reproduction systems across the listening is a non-trivial task. The use of dynamic binaural simulation of loudspeaker-based systems could provide researchers a powerful tool for the evaluation of primary artefacts such as localisation and colouration but until now, the validity in such systems has not been proven. It has also been identified that once validated, binaural simulations can be used to assess the perception of colouration artefacts across the listening area with a specific focus on the comb-filtering caused by the summation of coherent, delayed sound sources. The original contributions and research publications that have come from this research to-date have also been defined, supporting the importance and impact of the work to the research community.

CHAPTER 2

General Concepts and Fundamental Theory

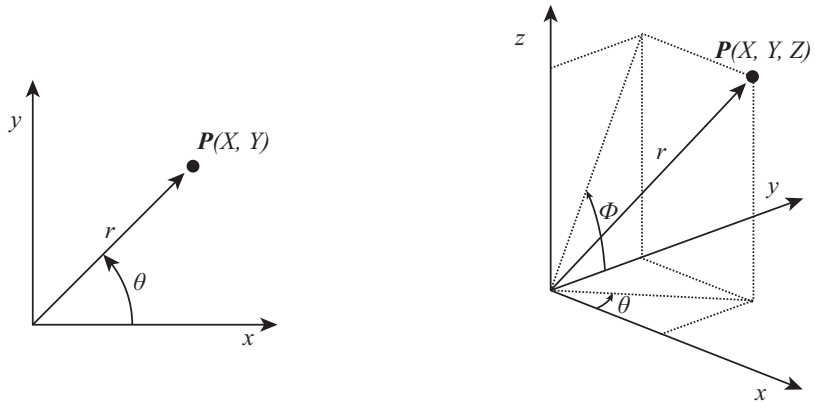
This chapter introduces general concepts, conventions and theory needed as a basis for the thesis. Firstly, naming conventions and mathematical nomenclature are introduced. The fundamental theory of spatial audio processing and perception are presented and discussed which will serve as a basis for the work presented in the main chapters of the thesis.

2.1 Introduction

Prior to the main contributory chapters, some basic concepts should be clearly defined. Firstly, due to the importance of spatial positioning and source/listener orientations in many elements of the thesis, the 2- and 3-dimensional coordinate systems are mathematically defined. The core features of the human auditory system are then introduced. This is vital to the understanding of human auditory perception for psychophysical tests, subjective evaluations and auditory modelling used in later chapters. The main loudspeaker-based processing techniques are then defined mathematically, followed by an introduction to dynamic binaural simulation from first principles. This basis of theory will serve as a technical reference for later chapters.

2.2 General Concepts

Figure 2.1 shows the coordinate system used to define cartesian, polar and spherical coordinates of a vector.



(a) 2-Dimensional. (b) 3-Dimensional

Figure 2.1: Coordinate system used throughout the thesis. θ and ϕ are enumerated in degrees using anti-clockwise as the positive direction from 0° through 180° to 360°

2.3 The Human Auditory System

The human auditory system is fundamental to the entirety of the work presented in this thesis. Therefore, it is important to define some key features of the system. Figure. 2.2 shows a simplified diagram of the anatomy of the human auditory system up to the cochlea.

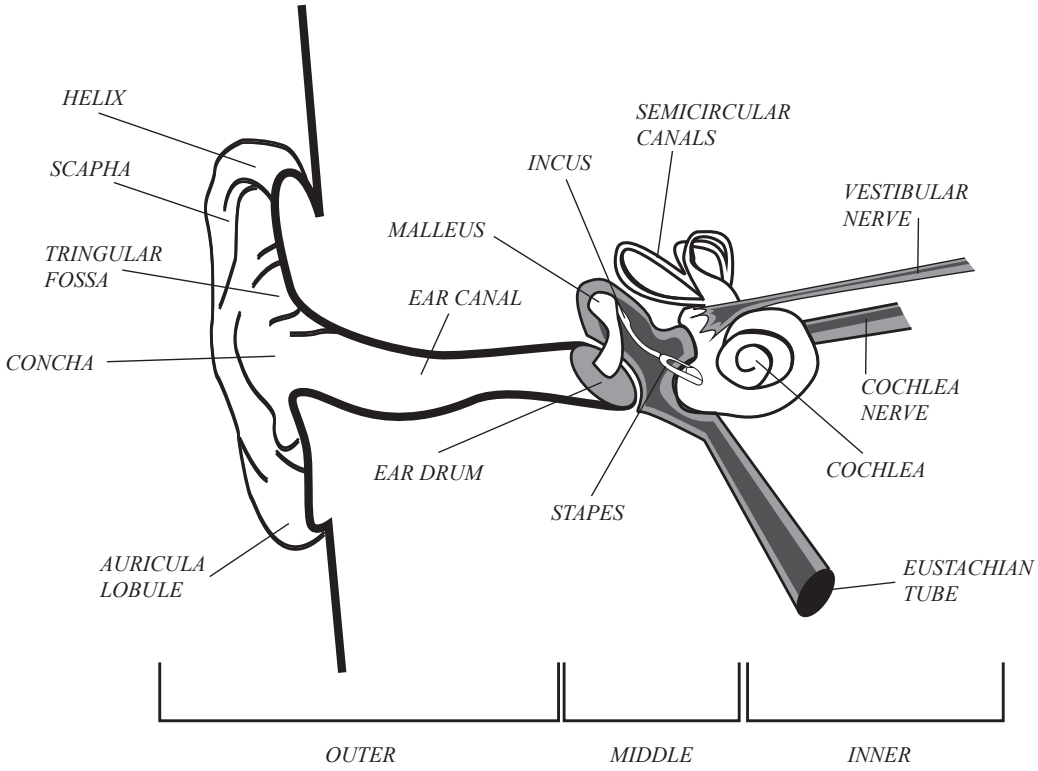


Figure 2.2: A simplified diagram of the human auditory system. Outer, middle and inner ear sections are indicated.

Sound waves propagate from a sound source and enter the auditory system through the ear canal. The sound is reflected and diffracted by the head, torso and folds of the pinna before entering the ear canal. A pressure division also occurs at the entrance to the ear canal due to impedance changes. The ear canal causes a peak in the 4 kHz frequency region which has been well characterised (Hammershøi and Møller, 1996). Sound waves created by an external acoustic stimulus will firstly interact with the physical shape of the listener, travel down the ear canal and create oscillations on the ear drum (tympanic membrane).

The middle ear begins with the 3 small bones connecting the ear-drum and the entrance to the cochlea. The bones serve to transmit sound from air to liquid

and therefore perform impedance matching between the air and the liquid (see work by Killion and Dallos (1979) for a summary on the topic).

The cochlea marks the start of the inner ear where sound waves are converted to nerve impulses transmitted to higher-level processing in the brain. An important element of the cochlea is the basilar membrane, which is fundamental in converting sound signals to electrical signals. The basilar membrane, due to its structure, is also responsible for the frequency selectivity of the human auditory system. Sound waves travelling into the basilar membrane create standing waves, spatially related to the driving wave's frequency. The detection of these excitations are defined by inner hair cells, spatially distributed along the basilar membrane. The Greenwood function maps the position of hair cells to the sound frequencies that stimulate the corresponding auditory nerves (Greenwood, 1961). The frequency selectivity of the basilar membrane structure has often been modelled using a series of overlapping band-pass filters. These filter-banks are often described as 'auditory filters' and can be implemented to give a more realistic understanding of the human auditory system's response to certain stimuli. Glasberg and Moore (1990) defined the equivalent rectangular bandwidth (ERB) and therefore showed the relationship between auditory filters and bandwidth across the frequency spectrum. Third-octave filtering also represents a simple approximation of the human auditory system's frequency selectivity. The magnitude response of a 1 ERB-spaced gammatone filter bank is shown in Figure. 2.3. This filter bank was implemented using the Auditory Modelling Toolbox (Søndergaard et al., 2011).

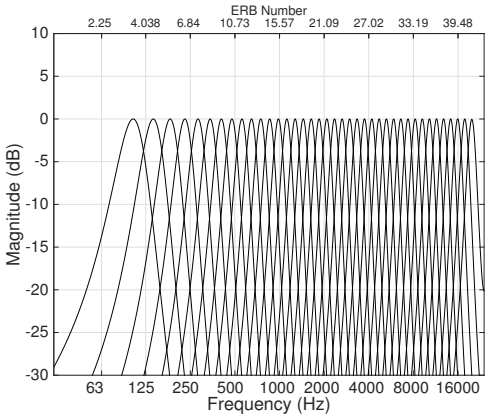


Figure 2.3: Magnitude response of a gammatone filter bank used to model the frequency selectivity of the basilar membrane. The filter bank uses 39 filters between 100 Hz and 20 kHz. Each filter is 1-ERB wide and normalised to 0 dB at the centre frequency. ERB scale is also shown in the upper x-axis.

The nomenclature regarding the description of physical properties and human perception of acoustics used in this paper follows those defined by Blauert (2001). The term *auditory event* describes the internal human perception when exposed to sounds. The term *sound event* describes the physical stimulus causing acoustical wave propagation. As an example of these labels particularly relevant to this thesis, a listener may be presented with a dynamic binaural simulation of a loudspeaker in a room using headphone reproduction. Assuming perfect performance of the simulation system, the listener will perceive the *auditory event* as being the simulated loudspeaker within the room. The *sound events* are the headphone transducers creating the pressure at the listener's ear drums, as shown in Figure. 2.4.

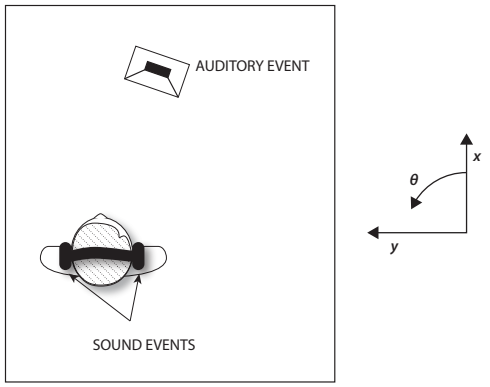


Figure 2.4: Description of auditory events and sound events for a binaural simulation.

2.4 Loudspeaker-based Spatial Audio Reproduction Systems

This thesis focuses on the simulation and performance of loudspeaker-based systems in a domestic listening environment. For this application, two systems have had recent popularity in research institutions: Ambisonics and Vector Base Amplitude Panning (VBAP). Following from the introduction of previous literature in Section. 3.2, this section introduces the fundamental theory of the two reproduction methods.

2.4.1 Vector Base Amplitude Panning

VBAP (Pulkki, 1997) is a vector-based reformulation of standard amplitude panning systems previously defined for stereophonic reproduction (specifically the tangent panning law). The method allows for the creation of auditory events using an arbitrary loudspeaker layout (sound events). For 2-dimensional loudspeaker layouts, the system uses amplitude panning between pairs of

loudspeakers to achieve the virtual sound sources. Figure. 2.5 shows the panning function for a target virtual sound source created using amplitude panning between a ±30° loudspeaker layout. The system is also often applied to ITU 5.0 layouts and represents a logical benchmark in domestic panning methods. The original mathematical formulation is shown below from Pulkki (1997).

In the vector system, the base is defined by unit-length vectors pointing towards the two loudspeakers $\boldsymbol{l}_1 = [l_{11}\ l_{12}]^T$ and $\boldsymbol{l}_2 = [l_{21}\ l_{22}]^T$ and the unit-length vector pointing to the virtual source direction is $\boldsymbol{p} = [p_1\ p_2]^T$.

Using gain coefficients g_1 and g_2, $\boldsymbol{p}$ can be represented as a linear combination of the loudspeaker vectors and written in matrix form:

$$\boldsymbol{p}^T = \boldsymbol{g}\boldsymbol{L}_{12} \tag{2.1}$$

where $\boldsymbol{g} = [g_1\ g_2]$ and $\boldsymbol{L}_{12} = [\boldsymbol{l}_1\ \boldsymbol{l}_2]^T$

If $\boldsymbol{L}_{12}^{-1}$ exists then the equation can be solved for $\boldsymbol{g}$.

$$g = p^T L_{12}^{-1} = [p_1\ p_2] \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}^{-1} \tag{2.2}$$

For 3-dimensional loudspeaker layouts, a triad of loudspeakers can be used to create the auditory event at a position within the loudspeakers. Figure. 2.6 shows a triad of loudspeakers positioned on the surface of a sphere. If a signal is equally weighted on each loudspeaker, the target virtual sound source will be positioned at a position between each of the speakers.

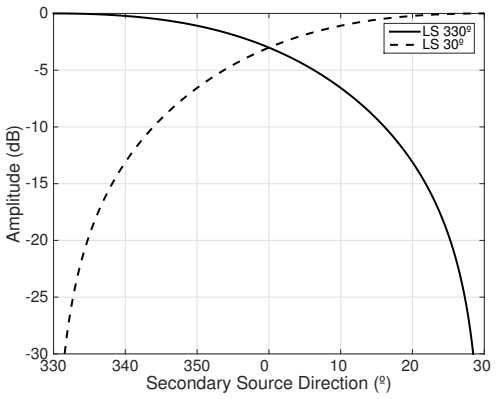


Figure 2.5: Stereophonic VBAP panning function.

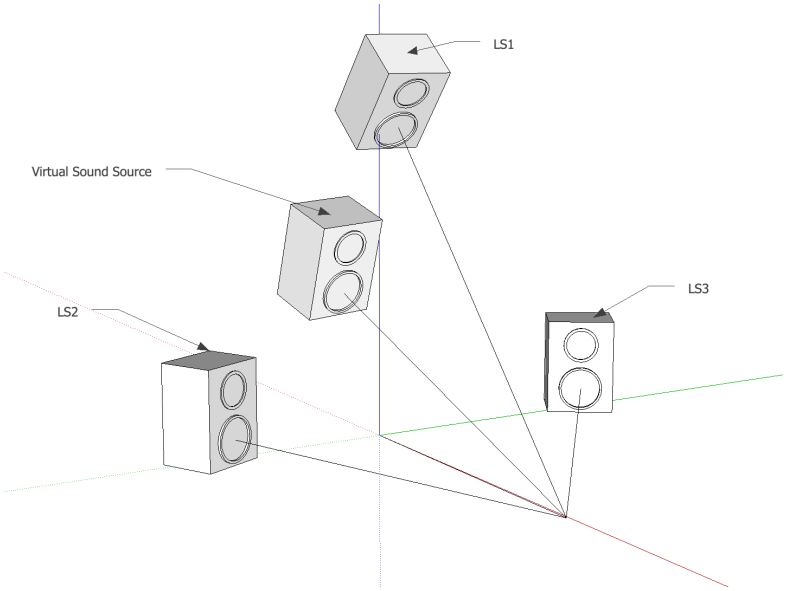


Figure 2.6: Geometrical layout of a 3D VBAP system.

2.4.2 Ambisonics

The name *Ambisonics* has become a general term to describe any audio processing in which a sound field is processed using an intermediate spherical harmonic representation. The first practical use of Ambisonic systems were created by Gerzon (1972) in which 0th and 1st order spherical harmonics were used to decompose and reconstruct a sound field. In the current state, Ambisonics can be used to describe anything from a microphone processing

theory through to virtual sound source panning functions and at the most complex end of the spectrum, sound field synthesis systems using near-field compensated Ambisonics with significantly higher numbers of spherical harmonic coefficients.

In this section, the mathematical formulation will firstly be derived from the wave equation. It will then be shown that using some basic assumptions, the Ambisonic derivation can be reduced to simple amplitude weightings to be used in a practical loudspeaker-based application. Recent developments of these weightings to improve perception will also be introduced.

Derivation

The sound field inside a source-free sphere can be recreated exactly by the control of a continuous source distribution across the surface of a sphere. Firstly, the time-domain homogeneous wave equation can be shown for a linear lossless medium in Equation. 2.3.

$$(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2})p = 0 \tag{2.3}$$

where ∇^2 is the Laplacian operator, p is acoustic pressure and c the speed of sound propagation.

From this, the homogeneous Helmholtz equation (Williams, 1999) can be found by applying a Fourier transform to Equation. 2.3.

$$(\nabla^2 + k^2)p = 0. \tag{2.4}$$

where $k = 2\pi f/c$ is the wave number.

The pressure at any point inside a source-free sphere can be found using a Fourier-Bessel decomposition as shown in Equation. 2.5.

$$p(kr, \theta, \phi) = \sum_{m=0}^{+\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi) \tag{2.5}$$

θ and ϕ here represent spherical coordinate angles with r being the distance. $j_m(kr)$ is the spherical Bessel functions responsible for the radial term between the origin and the measurement point. The 'outgoing' field can also be described by an additional divergent spherical Hankel function term, not shown in Equation. 2.5. Due to the assumptions of Ambisonic reproduction being free from sources inside the reproduced loudspeaker region, weighting coefficients for Hankel functions are made to equal zero and therefore this term is often removed (Daniel et al., 2003). The real-valued spherical harmonic functions, $Y_{mn}^{\sigma}(\theta, \phi)$ are described by Equation. 2.6.

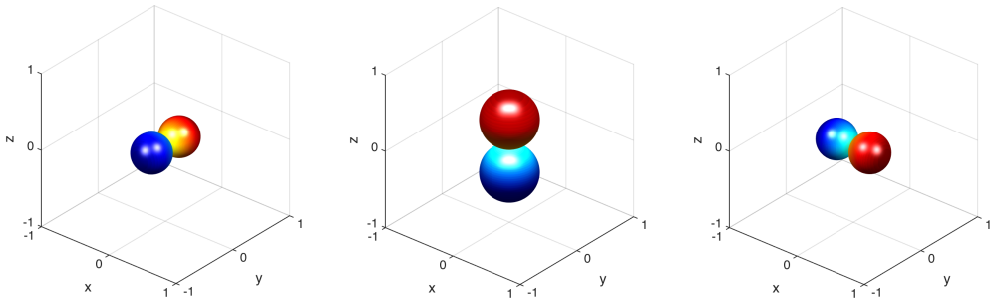
$$Y_{mn}^{\sigma}(\theta, \phi) = \sqrt{(2m+1)\epsilon_n \frac{(m-n)!}{(m+n)!}} P_{mn}(\sin\phi) \times \begin{cases} \cos(n\theta) \ \textit{if} \ \sigma = +1 \\ \sin(n\theta) \ \textit{if} \ \sigma = -1 \end{cases} \tag{2.6}$$

$P_{mn}(\sin\phi)$ are the associated Legendre functions. $\epsilon_n = 1$ when $n = 0$ and $\epsilon_n = 2$ when $n > 0$ (Nicol, 2010).

Due to the nature of spherical harmonics, it is possible to separate the spatial encoding and decoding processes. Encoding represents the conversion of a physical or theoretical sound field into the spherical harmonic domain for storage or transmission. Decoding represents the conversion back to the spatial domain where the spherical harmonic representation is put back into a format for acoustic reproduction. When encoding and decoding stages are separated, the

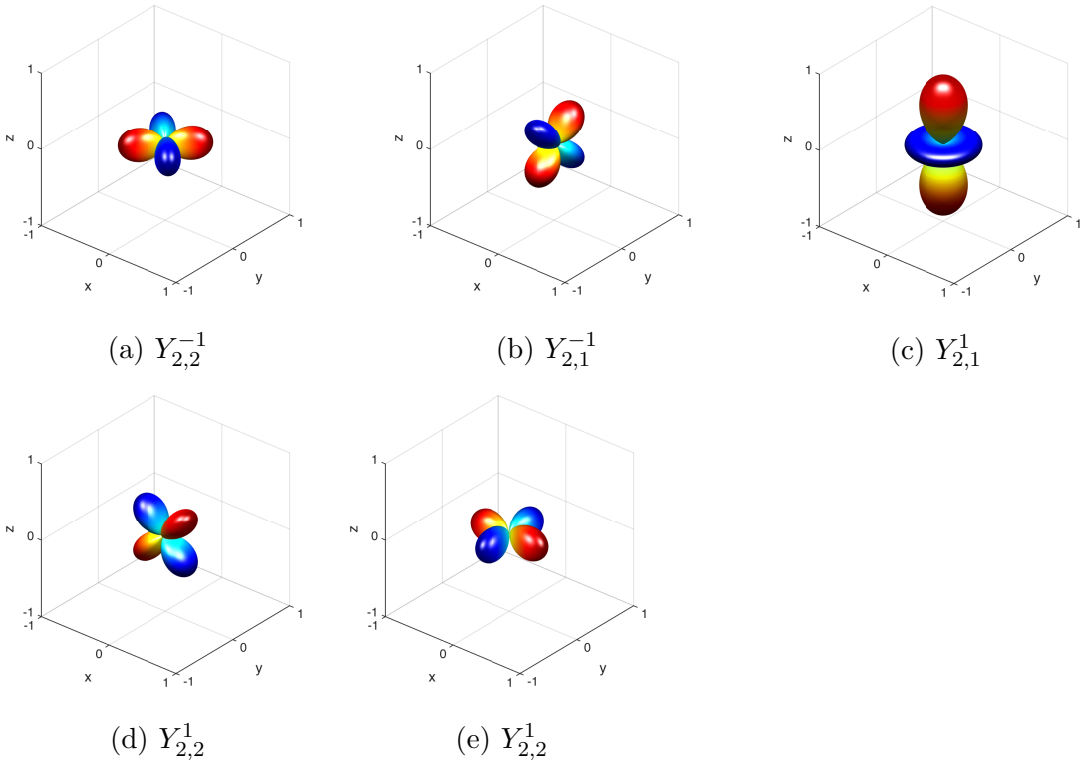
normalisation of $Y_{mn}^{\sigma}(\theta, \phi)$ must be carefully maintained (Daniel, 2001, p.156). For the practical use of higher-order Ambisonics, the Furse-Malham (Malham, 2005) normalisation scheme has become a popular standard. However, alternatives exist that can be extended to arbitrary orders (SN3D, N3D). m is the Ambisonic order where a spherical harmonic representation is truncated at M. For each order m, there are $(2m + 1)$ different spherical harmonic functions.

Real-valued spherical harmonic functions for 1st and 2nd orders are shown in Figure. 2.7 and Figure. 2.8 for $m = 1$ and $m = 2$ respectively.



(a) $Y_{1,1}^{-1}$ (b) $Y_{1,0}^{1}$ (c) $Y_{1,1}^{1}$

Figure 2.7: Real-valued spherical harmonic functions for first order (m=1)



(a) $Y_{2,2}^{-1}$ (b) $Y_{2,1}^{-1}$ (c) $Y_{2,1}^{1}$

(d) $Y_{2,2}^{1}$ (e) $Y_{2,2}^{1}$

Figure 2.8: Real-valued spherical harmonic functions for second order (m=2)

Encoding and Decoding

Following the derivation of Ambisonic theory it is important to consider the practical implementation of the technology, a plane-wave source can be encoded in to a spherical harmonic representation B_{mn}^{σ} defined by,

$$B_{mn}^{\sigma} = S \cdot Y_{mn}^{\sigma}(\theta, \phi) \tag{2.7}$$

This inherently describes how an amplitude S can be represented as a plane-wave from the direction (θ, ϕ). Equation. 2.5 defines that the summation extends to $+\infty$. Therefore, for a perfect reconstruction of the plane-wave, B_{mn}^{σ} also extends to ∞, which is practically unrealisable. The spherical harmonic representation of

the sound field is often truncated at an order M, motivated by the spatial resolution needed for the original sound field or other constraints such as number of values for B_{mn}^{σ} which are required.

Once the spherical harmonic representation has been achieved, loudspeaker driving signals must be derived. This is the decoding stage and the outcome is dependent on the geometry of the loudspeaker array. Perceptually motivated modifications to achieve loudspeaker signals have also been created (Gerzon, 1992a; Malham, 1992).

The decoding method can be defined using a similar premise to the encoding method, each loudspeaker is considered as a plane wave and the process is to define the plane-wave amplitudes needed to achieve the encoded spherical harmonic signal (therefore the original soundfield) (Hollerweger, 2006).

$$B_{mn}^{\sigma} = \sum_{j=1}^{L} Y_{mn}^{\sigma}(\theta_j, \phi_j) \cdot p_j \tag{2.8}$$

Where the summation occurs for each j loudspeaker at direction θ_j, ϕ_j, using a total of L loudspeakers. p_j is the loudspeaker signal (plane wave amplitude).

For an algorithmic implementation the loudspeaker layout must be 're-encoded',

defined by the re-encoding matrix, C (Daniel, 2001) where,

$$C = \begin{pmatrix} Y_{00}^{1}(\theta_1, \phi_1) & Y_{00}^{1}(\theta_2, \phi_2) & \cdots & Y_{00}^{1}(\theta_j, \phi_j) & \cdots & Y_{00}^{1}(\theta_L, \phi_L) \\ Y_{11}^{1}(\theta_1, \phi_1) & Y_{11}^{1}(\theta_2, \phi_2) & \cdots & Y_{11}^{1}(\theta_j, \phi_j) & \cdots & Y_{11}^{1}(\theta_L, \phi_L) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ Y_{M0}^{1}(\theta_1, \phi_1) & Y_{M0}^{1}(\theta_2, \phi_2) & \cdots & Y_{M0}^{1}(\theta_j, \phi_j) & \cdots & Y_{M0}^{1}(\theta_L, \phi_L) \end{pmatrix} \tag{2.9}$$

The decoding matrix D is then calculated by finding the inverse of C.

$$p_j = D \cdot B_{mn}^{\sigma} \tag{2.10}$$

If C is not square or does not have full rank i.e. the number of loudspeakers and Ambisonic channels is not equal, the inversion of C is not possible. In this situation a pseudo-inversion algorithm can be used (such as the commonly used, Moore-Penrose pseudo-inverse function). Modifications of the decoding matrix have been defined using a weighted diagonal matrix multiplication where the decoding matrix becomes,

$$D = D \cdot \Gamma \tag{2.11}$$

Daniel (2001) shows *basic*, *max* r_E and *in-phase* decoder types, each driving loudspeakers with different priorities. While basic decoders optimise pressure velocity, *max* r_E decoders optimise reconstruction energy in the direction of the encoded plane wave. In-phase decoding controls loudspeaker gains to avoid phase differences between loudspeakers of opposing directions. Each decoding 'flavour'

is chosen depending on the priorities of the reproduction system; Daniel (2001, p. 160) discusses these in detail. *In-phase* decoding processes were originally proposed by Malham (1992) Ambisonic systems.

$$\Gamma_{basic}(m) = (1) \tag{2.12}$$

$$\Gamma_{max\ r_E}(m) = \cos\frac{m\pi}{2M+2} \tag{2.13}$$

$$\Gamma_{in-phase}(m) = \frac{M!^2}{(M+m)!(M-n!)!} \tag{2.14}$$

Although tools were developed for this research project to enable the encoding, decoding and processing of Ambisonic panning methods, literature highlighted a general lack of documentation regarding Ambisonic tools used in subjective experiments, ultimately leading to difficult direct comparisons. Therefore, an open-source toolbox[1] was used to create Ambisonic decoders for all applications of Ambisonic panning methods used in this thesis. The authors of this toolbox include decoder settings specifically for the SBSBRIR dataset where subsets of the loudspeakers can be chosen. Although beneficial for 3-dimensional loudspeaker layouts, only loudspeaker arrays along the horizontal plane are considered in this thesis. This is often categorised as 2.5D, where loudspeakers lie in the 2-dimensional plane, but inherently reproduce sound in 3-dimensions.

[1] https://bitbucket.org/ambidecodertoolbox/adt.git

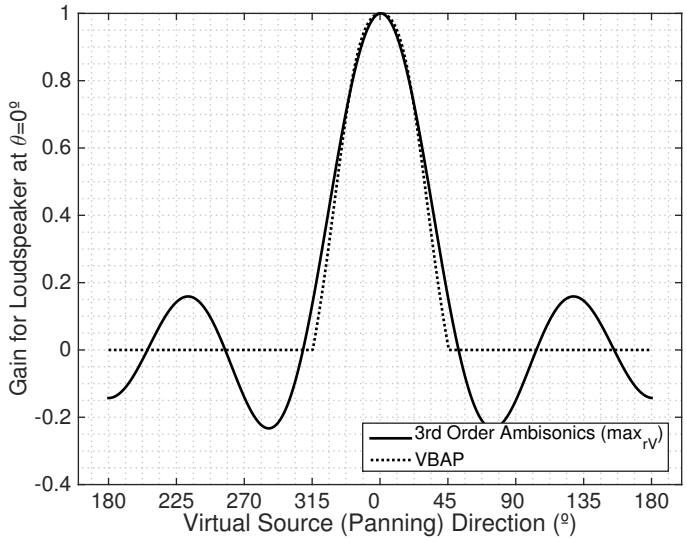


Figure 2.9: Panning functions for VBAP and Ambisonic reproduction methods. Lines show the gain coefficient for the loudspeaker at $\theta = 0°$ as the virtual/phantom sound source is panned across the full azimuth range. The loudspeaker layout is an octagon (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°). Ambisonic panning is, 3rd-order with max_{rV} decoding philosophy.

To illustrate the practical differences between VBAP and Ambisonic reproduction methods, Figure. 2.9 shows the panning function for a single loudspeaker in an Octagonal loudspeaker layout. It can be noted that whilst VBAP optimises energy in the direction of the phantom source, the Ambisonic panning function often has non-trivial gain coefficients, with opposite phase in opposing directions to the phantom source. Although beneficial to the reproduction system at low-frequencies, at high-frequencies this feature can become problematic.

2.5 Binaural Simulation

Preceding the literature review of binaural synthesis presented in Chapter. 3, a technical derivation of some fundamental concepts will be presented here. The concept of reproducing binaural recordings is firstly presented. This concept is

then parameterised to achieve the derivation of the dynamic binaural simulation system used in this thesis.

2.5.1 Binaural Recordings

Consider a listener positioned in a free-field environment in close proximity to sound events as shown in Figure. 2.10. Recording the sound pressure at the left and right ears of the listener P_L and P_R will inherently include the filtering effect on sound transmission from external sound events caused by reflection and diffraction around the physical geometry of the listener. This method of recording a sound scene (a composition of sound events) inherently encodes signals with the monaural and interaural localisation cues of the listener. Binaural recordings can be made with in-ear microphones on a human or using an artificial head and torso simulator. Replaying the signals over headphones alongside the inclusion of some headphone compensation filters will give a realistic reproduction of the sound scene, albeit with the limitation that ego-centric localisation cues are anchored to the movements made by the human or artificial head during the recording. A fundamental limitation of this method is that the sound scene is essentially fixed by what was recorded at $P_{L/R}$.

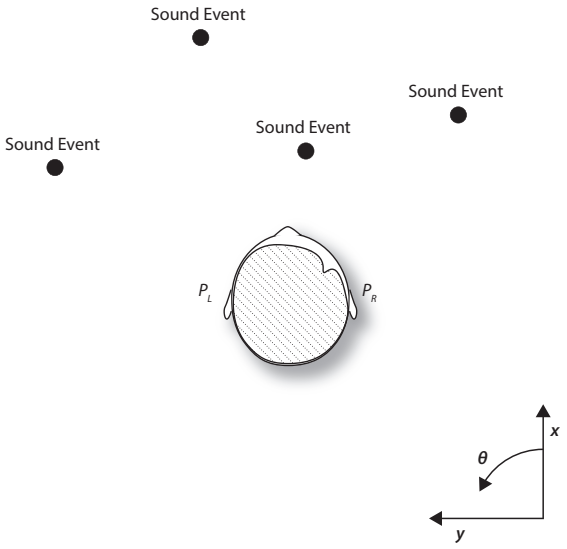


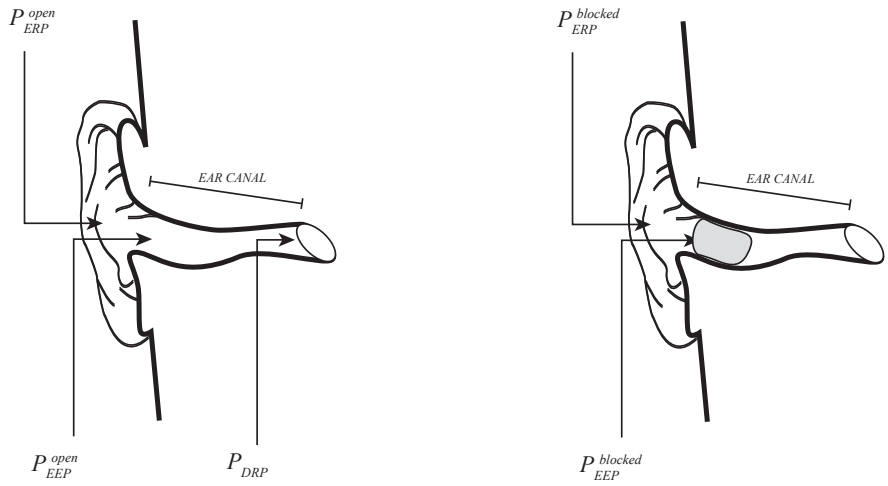
Figure 2.10: Listening scenario for a binaural recording where sound pressure recordings are made at the entrance of the listeners ears.

By treating the acoustic transmission from a sound event to a listener's ears as a linear and time-invariant system, it is possible to measure the head-related transfer function and separate the acoustic input to the system (the sound event) from the recording.

2.5.2 The Head-related Transfer Function (HRTF)

The head-related transfer function is defined as the free-field transfer function from a sound source to each of a listener's ears (Xie, 2013) and can be thought of as a linear, time-invariant (LTI) process. These frequency domain functions, one for each ear, contain *most* of the important localisation cues needed by a human listener. However, dynamic cues caused by in situ head-movements are not included. Møller (1992) states that transfer functions measured at any of the microphone positions shown in Figure. 2.12 constitute a HRTF due to the ear canal being regarded as a one-dimensional transmission line (Hammershøi and Møller, 1996). The HRTF is a function of sound source distance, azimuth,

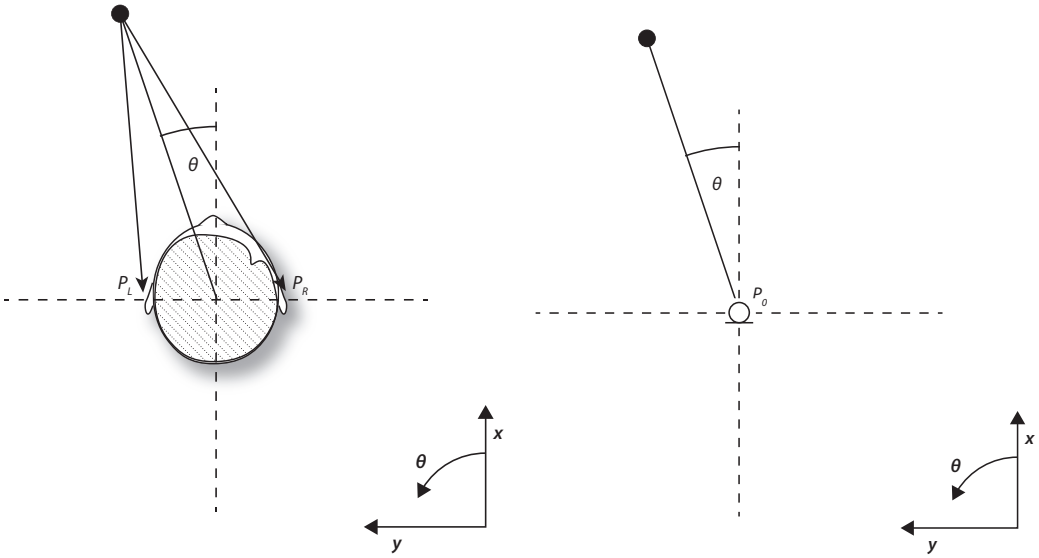
elevation and frequency but also varies between individuals due to anatomical differences. Figure. 2.12 shows a simple diagram of the human outer-ear anatomy and commonly-used measurement positions.



(a) Open ear canal. (b) Blocked ear canal.

Figure 2.11: Human outer-ear anatomy with pinna, ear-canal and ear-drum. ERP = ear reference point, EEP = ear entrance point and DRP = drum reference point. Ear-canal blocking is usually achieved using expanding foam ear-plugs.

HRTF measurements can be made using a far-field loudspeaker with a broad and flat frequency response. To remove the effect of the loudspeaker, microphone and propagation delay in an efficient way, a reference measurement can be made. This is the transfer function between loudspeaker input terminals and a microphone positioned at the centre of the head (without the head present) referred to as P_0. This method can be called the *measurement-equalised HRTF* and can be measured as shown in Figure. 2.12(b).

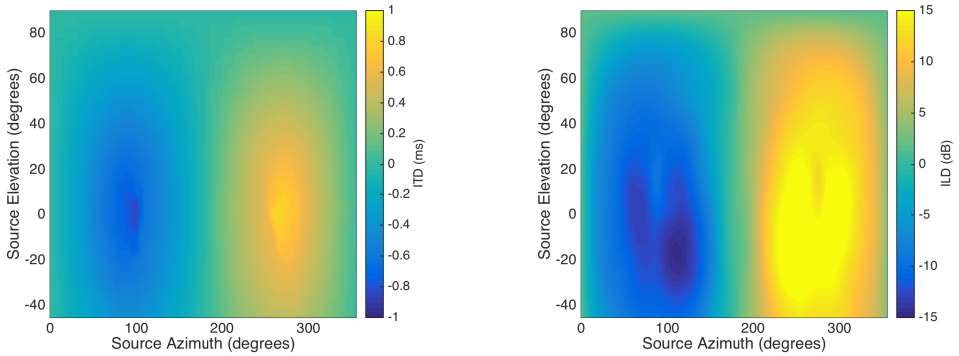


(a) P_L and P_R measurement. (b) P_0 measurement.

Figure 2.12: Setup for ear and reference measurements needed to derive HRTFs.

The HRTF is a frequency domain representation of the filtering effect where the inverse Fourier transform gives the head-related impulse response (HRIR).

When HRTFs are considered binaurally, differences between left and right ears can be used to highlight the two fundamental localisation cues: interaural time and level differences. Early investigations by Strutt (1907) on pure-tone localisation introduced the *Duplex Theory* which states that the human auditory system uses inter-aural time differences for the lateralisation of sounds in the low-frequency spectral region (<500 Hz) where shadowing by the head is negligible and lateralisation of high-frequency sound events is dominated by inter-aural level differences, where head shadowing is more dominant due to the relative size of the head compared to the wavelength. Experiments which revisit the topic still support the duplex theory today (Macpherson and Middlebrooks, 2002). Representative ITD and ILD values are shown in Figure. 2.13.



(a) ITD (b) ILD

Figure 2.13: Broadband interaural cues across azimuth and elevation calculated using a freely available HRTF dataset (Andreopoulou et al., 2015). ITD is calculated using the maximum IACC method.

The HRTF can be represented by three fundamental transfer function components: (1) Minimum-phase component, (2) all-pass component and (3) linear-phase component. The minimum-phase element has a magnitude response equal to that of the original HRTF but the smallest phase angle. Due to the natural log of the magnitude response being related to the phase angle of the minimum-phase transfer function's phase response, a Hilbert transform can be implemented to approximate this easily. The all-pass component is a unity magnitude filter with any excess phase response. The linear-phase filter is a pure delay. Minnaar et al. (1999) has shown that the omission of the all-pass component in binaural simulation is not perceptible by humans.

The pure-delay (linear-phase) parts of the HRTF, which are independent at each ear, represent the ITD. The delay between the ears is simple in concept. However, in practice it is difficult to estimate and many methods exist. Early procedures calculate the differences in time-of-arrival by finding the time at which the magnitude exceeds a certain threshold such as 5% of the maximum (Sandvad and Hammershøi, 1994), for each ear. Although efficient, this method

suffers from bias due to low interaural coherence on the contralateral ear (Nam et al., 2008). Kistler and Wightman (1992) implemented a method by calculating the argument of the maximum value in the interaural cross-correlation function. A similar method was used by Nam et al. (2008) where the cross-correlation was performed between HRIRs and their minimum-phase versions yielding the time of arrival for each ear separately. A linear-phase fitting method in frequency domain was also proposed by Jot et al. (1995). As shown in Figure. 2.13, the ITD values range from $0\mu s$ at a source azimuth of 0° to approximately $\pm 700\mu s$ at a source azimuth of ±90°.

Although a number of methods exist, all have been shown to produce valid results and the choice of implementation method may be defined simply by practicality. ITD approximations for the computational model presented in Chapter 7 use the method of Kistler and Wightman (1992) as this allows for ITD to be calculated at the same time as the running interaural coherence function. For ITD metrics presented in Chapter 5, the method of Nam et al. (2008) was used to improve stability of the approximation when interaural coherence is low.

The ILD can range between 0 and ±20 dB and is highly frequency dependent due to the inherent frequency dependance of the HRTF.

2.5.3 Dynamic Binaural Simulation

Due to the human anatomy, HRTFs are extremely directionally dependent. Whether movements of a sound source around a listener, or movements of a listener relative to a sound source, sound pressures at the ear drums of the listener will change. The human auditory system expects an ego-centric

movement to induce a specific change in the pressures at the ear drum, and the fusion of this motor-acoustic information helps to externalise sounds (Brimijoin et al., 2013). A *dynamic* binaural synthesis system uses real-time head rotation data to dynamically update signals at the ears of a listener using headphones. This allows for auditory events to be created at fixed locations within a reference frame (such as a room). The auditory system uses a dynamic feedback process whereby head-movements are 'expected' to induce changes in localisation cues for a plausible, stationary auditory event. If cues are not as expected by the movements of the listener, then the feedback process will begin to break down and auditory events are likely perceived as less plausible and less externalised.

A simple dynamic binaural simulation system will use a dataset of HRIRs measured at discrete head azimuth (and possibly elevation) directions. Realtime head-tracking data is then used to update the filters to each ear. By performing this process for numerous auditory events and summing at the headphone inputs it is possible to create a rich auditory scene.

When a listener is sitting in a reverberant environment, sounds arrive at the entrance to the auditory system through a direct path from the sound source, but also via indirect paths reflected off and diffracted around objects. An environment's natural reverberation can be included in the simulation by replacing the HRIR dataset with a binaural room impulse response (BRIR) dataset. These impulse responses are generally much longer than HRIRs due to the decay of energy in a reverberant environment. A representative BRIR for a left ear only is shown in Figure. 2.14.

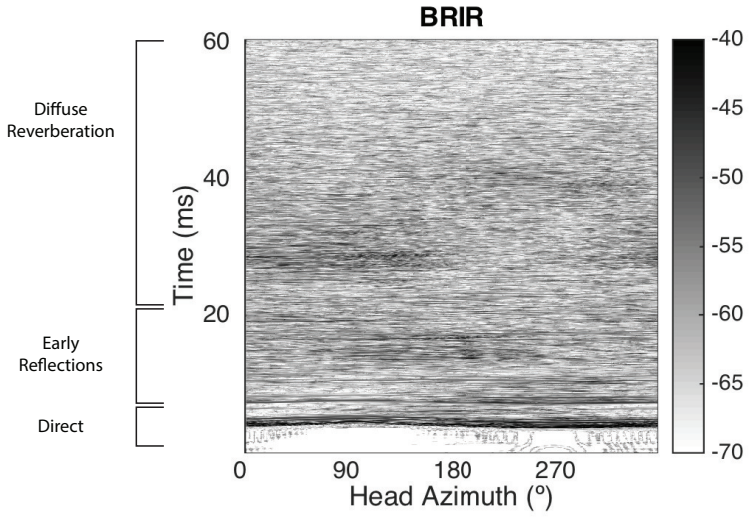


Figure 2.14: Magnitude of the left ear BRIR plotted on a dB scale across head-azimuth.

The colour axis is chosen to highlight different stages of the BRIR over time. The direct response can be seen clearly, arriving at around 5 ms where ITD can be seen in the curvature of the impulse. Early reflections are noticeable from the floor, ceiling and then later from side and rear walls in the room. After around 20ms the response becomes more diffuse due to reflections being scattered multiple times before arriving at the ear of the artificial head. It is important to note that the use of BRIRs, as opposed to HRIRs for dynamic binaural simulation have been used to improve the perception of distance (Begault, 1992) and therefore help to remove the perception of in-head localisation.

This section has provided a background on the fundamental theory needed to support the introduction of binaural simulations used in this thesis. Chapter. 4 continues from here to present a detailed technical introduction of the binaural simulation system created and validated for work conducted in this thesis.

2.6 Conclusions

This section of the thesis has introduced the most important fundamental concepts needed for later technical chapters. By considering the physiology of the human auditory system, it has been highlighted that the complex processing of acoustic stimuli (e.g. the frequency resolution of the basilar membrane) should be considered when analysing or modelling the human perception of loudspeaker-based reproduction systems. Two loudspeaker-based reproduction techniques (VBAP and Ambisonics) were then introduced. For VBAP, amplitude weightings are applied to stimulus signals over pairs of spatially distributed loudspeakers which could create audible artefacts. Ambisonic systems are shown to be constrained by the spherical harmonic order and number of loudspeakers available for reproduction. Following an introduction to binaural simulation, it has been shown that monaural, binaural and dynamic cues can be used to create an auditory virtual environment to a listener using headphones. This concept will be applied to define the simulation system in Chapter 4.

CHAPTER 3

Literature Review

This chapter presents a review of the previously documented literature that is relevant to the experiments presented in this thesis. The chapter reviews literature from both commercial and academic developments in spatial audio reproduction.

3.1 Introduction

This chapter presents a review of relevant literature up to the state-of-the-art. Firstly, a history of loudspeaker-based sound reproduction development is presented. This introduces the important reproduction systems and key experiments that have been performed to understand the reason for many of the standard practices and techniques found in today's reproduction systems. Secondly, it is important to understand what contributions have been made to the subjective perception of loudspeaker-based spatial audio systems. This is presented over two topics, localisation and colouration, to provide motivation for subjective experiments presented in this thesis. Complimentary to subjective evaluation, the modelling of human perception is often used to assess reproduction systems. The third main section covers previous literature relating to the computational modelling of human perception of localisation and colouration attributes. This sets out the scope for perceptual modelling presented in later chapters. Previously published literature on the assessment of loudspeaker-based spatial audio reproduction systems across the listening area is then reviewed in Section 3.5. This covers the relatively small number of publications relating to perception across the listening area and provides motivation for using binaural simulation of listening area effects. Although binaural simulation of listening area effects is a promising prospect, the final section of literature review addresses previous publications regarding the validity of using such simulations systems and provides motivations for the research presented in this thesis.

3.2 Loudspeaker Reproduction in Domestic Listening Environments

When a listener is situated in a sound scene, individual auditory events will have identifiable spatial characteristics such as direction, distance and size. Some auditory perceptions will be more diffuse, where the sound has no discrete direction. Simulating the spatial perception of sound using loudspeakers can be categorised by the term spatial audio reproduction and encompasses a broad range of concepts. The work of Blumlein (1931) was some of the first relating to the synthesis of spatial impression by controlling multiple loudspeakers to contribute to a reproduced sound field. Work by Steinberg and Snow (1934) went on to highlight the underlying principals where three spaced microphones signals (left, centre and right) were routed directly to three loudspeakers (left, centre and right) for a simple localisation task in an auditorium. The main psychoacoustic cues used for localisation are also presented as 'loudness', 'phase' and 'quality' differences between left and right ears. Stereophony is defined by the use of spaced loudspeakers which have independent driving signals used to simulate the impression of sound scenes with spatial characteristics. Individual sounds can be encoded using amplitude or time variations between channels to achieve the desired localisation cues. Almost a century after its initial development, this channel-based concept still remains the most popular format for the recording, production and reproduction of audio content.

The cinema industry has become the catalyst for developments in loudspeaker-based spatial audio reproduction. Technical innovations are often implemented in the cinema first and then transitioned to the home, car or personal entertainment products. Following the first implementation of multichannel reproduction for the film *Fantasia* (Sharpsteen et al., 1941), many

different reproduction systems and transmission methods have been used in cinemas. The last three decades has been dominated by the use of the 5.1 loudspeaker layout with accompanying transmission formats such as Dolby Digital, Sony SDDS and DTS. The Dolby Digital format was introduced commercially in 1992 where lossy AC3 perceptual encoding of discrete loudspeakers signals was used to reproduce 5.1 channels of audio (Dolby Laboratories, 2000). The 5-channel loudspeaker layout is also recommended in ITU-R 775-3 (ITU-R, 2012b) and can be considered the 'benchmark' for the current state of domestic[1] loudspeaker layouts.

In the home audio consumer space, two-channel radio broadcasts were first tested by the BBC using FM transmission in the early 1960s. Widespread broadcasting across multiple stations was not achieved until the mid 1970s (Denyer et al., 1979). The initial use of stereophonic broadcasts for television were closely related to radio where viewers could get stereophonic soundtracks for their television broadcasts by tuning to the corresponding radio station (simulcasting). Following this, more advanced methods were use to transmit multichannel audio content with television signals (NICAM, MTS).

In recent years, the popularisation of the internet has allowed for an alternative delivery format for both live and pre-recorded audio content in a multitude of formats. Music subscription services have recently reduced the demand for permanent downloads and physical media across many nations. 2014 was the first year where the share of industry revenues was equal between downloaded and physical media (46% each) (IFPI, 2015). However, other than special release versions, music and television broadcast still predominantly use 2-channel stereo

[1] the term 'domestic' in this thesis refers to a listening environment representative of broadcast/media content consumers.

and sound is reproduced in the home using two independent channels, either from a television/radio, separate loudspeakers or an integrated speaker device (dock).

In April 2012 Dolby Laboratories introduced their Dolby Atmos surround sound technology. This technology uses the object-based philosophy where individual audio elements have associated meta-data containing dynamic, 3-dimensional sound source positioning. Alongside static elements such as dialogue and atmospheric sounds, this allows the decoding system to create loudspeaker driving signals without prior knowledge of the loudspeaker layout. The system even allows for elevated loudspeakers. This type of system represents the future of spatial audio reproduction both in the cinema and home environments, allowing for a separation between creation (encoding) and reproduction (decoding) of audio content. Similar standardisation efforts are also being implemented for 3D audio by the Motion Picture Experts Group (MPEG) in the latest MPEG-H 3D Audio standards (Herre et al., 2015).

Although the cinema industry has pioneered the transmission and reproduction of spatial audio, it is important to consider how spatial audio mixes are created. The term 'panning' refers to the method of creating individual audio channels (signals) for storage/transmission based on the known loudspeaker layout. Sound signals are then replayed over specifically positioned loudspeakers to achieve an induced auditory event with the possibility of it being localised to a position where no real loudspeaker exists. These auditory events are often called virtual or phantom sound sources. For example, a snare drum track on a mixer is panned to the centre of a stereo loudspeaker layout. Using an amplitude panning technique, this causes equal amplitude to the loudspeaker pair and therefore the

auditory event is ideally perceived between the two loudspeakers. To achieve consistent loudness when amplitude-panning a sound for a stereo mix, different panning laws have been implemented and are often chosen based on the design of control rooms. -6, -4.5 and -3 dB panning laws are commonly used where the dB value determines the attenuation at each loudspeaker when the auditory event is positioned directly between. The choice of the three originates from the fact that coherent signals sum to +6dB whereas incoherent signals sum to +3dB. The technical details of a panning system is encompassed in how the loudspeaker signals are derived from the original sound signal, loudspeaker positions and the intended auditory event direction. Distance and width cues can also be incorporated into the algorithm and time-delay panning is another popular technique. Fundamentally, amplitude panning techniques similar to Blumlein (1931) are often implemented into broadcast or post-production mixing desks to create loudspeaker signals where a pan-pot gives a sound engineer simple control over the sound stage.

The work by Pulkki (1997) has since popularised the use of amplitude panning in research institutions by creating encoding equations for arbitrary, 2- and 3-dimensional loudspeaker layouts using Vector Base Amplitude Panning (VBAP). Later work (Pulkki et al., 1999; Pulkki and Karjalainen, 2001; Pulkki, 2001) was focused on the perception of amplitude- and time-based panning algorithms. Panning techniques have also been developed using Ambisonic principles (Gerzon, 1972; Craven, 2003; Wiggins, 2007; Neukom, 2007; Zotter and Frank, 2012) where the spherical harmonic decomposition and reconstruction of sounds fields are reduced to simple amplitude weighting functions which are easily implemented. However, these panning laws have not achieved commercial application. Quadraphonic systems were also developed in the 1950s where four

loudspeakers in the configuration: front-left, front-right, back-left, back-right were used to auralise spatial audio mixes. The format was also developed for FM transmission (Gaskell and Ratliff, 1977). The technology become obsolete due to inherent problems with, among others, wide loudspeaker spacing.

Ambisonics has now become a popular research area. Scalability and the ability to 'blind-encode' the content (have no knowledge of the reproduction loudspeaker layout) makes it an attractive technology in many disciplines. The early inspirations for Ambisonic theory can be found in the paper by Cooper and Shiga (1972) where 'harmonic analysis' was combined with loudspeaker reproduction of a sound field. Seminal work on Ambisonics by Gerzon (1972) developed upon this idea to introduce the spherical harmonic decomposition and reconstruction of a sound field using 0^{th} and 1^{st} order spherical harmonics. Gerzon (1972) introduced periphonic reproduction of a sound field and compatibility with legacy horizontal systems and also showed the tetrahedral microphone and loudspeaker layouts for 1^{st} order Ambisonic recording and reproduction. Gerzon importantly highlighted that any function on the sphere (such as a sound field) can be approximated as a summation of m spherical harmonics, whereby the degree of accuracy is defined by the order M. This type of spatial audio reproduction can be referred to as *transform-domain based* (Spors et al., 2013), where encoded signals have no relation to loudspeaker layouts.

Following theoretical derivation, psychoacoustic perception of Ambisonic theory was then considered in later work (Gerzon, 1974, 1980, 1985). Gerzon (1992a) proposed a 'meta-theory' for sound localisation which consisted of a number of models. Two of these models are based around the construction of vectors for energy and velocity which are applied to approximate human localisation ability

for low- (velocity vector, ITD) and high- (energy vector, ILD) frequency regions with low computational expense. These proved useful as design metrics for optimised decoders such as developed by Gerzon (1992b); Wiggins (2004) and also work by Daniel (2001).Frank (2013) has also compared the use of velocity and energy vector localisation models with binaural models and subjective responses where results showed only practically small differences in directional predictions.

The term 'Ambisonics' has now come to represent a broad range of technologies, all with fundamental roots in the spherical harmonic decomposition and/or reconstruction of a sound field and it is important to identify some of the key areas. As noted by Ahrens (2012), near-field compensated higher-order Ambisonics can be used to perfectly reconstruct a sound field across the whole listening area and audible frequency range under theoretical (and practically unrealistic) conditions. However, when the reconstruction region is reduced to a single point in space at the central listening position and the loudspeaker layout is a regular sampling of the loudspeaker boundary surface (either 2-D or 3-D), Ambisonics can be reduced to simple amplitude panning functions. Although not a necessity, this process can be split into encoding and decoding stages to capitalise on the fact that an encoded Ambisonic signal has no theoretical dependance on the loudspeaker layout or decoding method. The encoded sound field is a spatial representation where its resolution is defined by the number of spherical harmonic coefficients. This scalability makes it an attractive technology in audio broadcast situations; 'one sound mix for all'. However, in practice, the experience of the listener's will be, as with most domestic reproduction systems, different depending on their reproduction hardware (large number of speakers, portable device speakers or headphones for example). Although considered a

viable future tool for the transmission of spatial audio content (Herre et al., 2015), variability in the reproduction of transform-based methods has caused limited practical application. A mathematical derivation of Ambisonic theory is presented in Chapter. 2. This simplified Ambisonic approach means that a desired sound field can be encoded into amplitude weightings of spherical harmonics and then decoded to any specific reproduction system (decoded signals). Following from the work of Gerzon, the simplified Ambisonic formulations have been researched extensively. Perceptually motivated decoder designs were firstly introduced by Gerzon (1992b). Malham (1992) also provided decoders for larger listening areas by controlling the anti-phase signals produced by basic Ambisonic decoding. Further application of Ambisonics to the ITU 5.0 layout has also been considered extensively (Wiggins, 2004; Moore and Wakefield, 2007; Benjamin et al., 2010). Although Gerzon's focus was 1st order Ambisonics (0th and 1st order spherical harmonics), the natural development to higher-orders was later introduced (Daniel, 2001). Daniel (2001) also documents the theory and instruction for various Ambisonic decoding methods for listeners at both (a) a single listener at the central listening position or (b) multiple listeners distributed across the listening area. Earlier simplifications imposed on to Ambisonics also include the assumption that encoded sound sources and loudspeakers in the reproduction environment are in the far field (plane waves) but later studies have introduced the extension to point sources (Zotter et al., 2009).

It has been shown that artefacts are introduced when the spherical harmonic series is truncated to a practically usable number. Spectral unbalance was reported by (Daniel, 2001). Solvang (2008) also investigated spectral impairments in higher-order Ambisonics and defined that for higher-order

Ambisonics, the number of loudspeakers is a trade-off between spatial reproduction error and spectral impairments. Spatial artefacts have also been reported by Frank et al. (2008). Zotter et al. (2010) reported on Ambisonic decoding with and without mode-matching and results highlighted that for Ambisonics order, $N < 20$ the theoretical area (from objective simulations with ideal loudspeakers) of accurate sound field reconstruction is less than the size of the human head in the audible frequency range.

The same principles for Ambisonic reproduction can also be used to record a sound field by using specially designed microphones arrays and digital signal processing to create audio signals in the spherical harmonic domain. These inherently include spatial characteristics of the sound field and first order tetrahedral designs were proposed by Gerzon (1972). The SoundField microphone was the first to achieve this in a commercial product[2]. The SoundField microphone and accompanying hardware is capable of producing 1st order Ambisonic signals and is often used as general dynamic directivity microphone in non-Ambisonic applications. Studies have also been focused on the development of high-order Ambisonics microphones (Elko et al., 2005; Bertet et al., 2009) and the Eigenmike[3] is the most popular commercial product of this type, giving researchers a standardised format for the spherical microphone array. Although high-order spherical harmonic microphones can provide new possibilities in the spatial domain, spectral artefacts can often be unavoidable and the high number of physical microphones can cause substantial analogue noise.

With the advent of HD, 4K and stereoscopic 3D video reproduction, focus has

[2]The SoundField microphone was originally sold by Calrec Audio Limited in 1978 but now sold by TSL: http://www.tslproducts.com/soundfield-type/soundfield-microphones.

[3]http://www.mhacoustics.com/products

been placed on improved audio reproduction to a similar degree. Whereas the panning systems discussed so far attempt to induce some spatial cues, the more recent introduction of 'sound field synthesis' techniques represents a paradigm shift whereby a mathematical representation of a desired sound field (whether recorded or synthesised) is recreated over a specified listening area using carefully positioned loudspeakers. Higher-order Ambisonics with near-field compensation can be considered a sound field synthesis method. The fundamentals of sound field synthesis theory was first proposed by Jessel (1973). Wave field synthesis (WFS) and Ambisonics represent two of the most popularly used sound field synthesis techniques. Near-field compensated higher-order Ambisonics (NFC-HOA) (Daniel, 2001) and WFS (Berkhout, 1988) have been shown to be mathematically equivalent under high-order approximations (Ahrens, 2012). Wave field synthesis is derived from the Kirchoff-Helmholtz integral, which states that the wave field inside any arbitrarily defined region of space caused by primary sound sources outside that region is fully described by the wave field on the boundary of that region (Berkhout et al., 1993). Hence a large number of loudspeakers around a listening area can control the sound field within the listening area equivalently to sound sources outside the loudspeaker array boundary. Following definition of the fundamental theory of wave field synthesis many studies have since been conducted regarding aspects from theoretical development (Berkhout et al., 1993; Theile et al., 2003; Spors et al., 2008) to human perception (Bruijn, 2004; Melchior et al., 2011; Wierstorf et al., 2013). Although WFS has been a popular tool at academic and research institutions in recent years, the need for a large amount of loudspeakers and therefore current lack of applications to the domestic environment mean it is not implemented for testing in this thesis and will not be discussed further.

Object-based audio is another recently popularised concept for the creation, storage and transmission of spatial audio which will be discussed briefly here.

Dolby Digital using the ITU layout is the most commercially successful domestic surround sound system to date. Some research (Hamasaki et al., 2004a,b) has suggested a logical progression of this channel-based format by increasing the number of channels and therefore reproduction speakers whilst maintaining the underlying principals. However, recent innovations such as Dolby Atmos and MPEG technologies have used object-based audio production and delivery. Object-based content is created in an open-ended format where individual audio elements are produced and broadcast with meta-data describing dynamic temporal (amplitude) and spatial characteristics (such as 3D position in space). This allows for the optimisation of reproduced signals at the reproduction end of the broadcast chain depending on the desired loudspeaker layout. The real power of this shift in content creation and delivery is the ability to create a single piece of content for many different listening situations.

In practice, the object-based audio concept still requires reproduction technologies to control the loudspeakers and the standardisation of this process has yet to be achieved. Many different commercial and non-commercial establishments have their own methods and many of the underlying technologies are proprietary. Channel-based and transform-based reproduction technology will likely still form the underlying basis of object-based systems, but object-based creation, transmission and reproduction technologies will allow for scaleability and improved reproduction over different kinds of reproduction devices.

Although stereophonic recording, panning and reproduction techniques are

almost 100 years old, they remain today's dominant format. The techniques are easily understood by musicians, engineers and consumers and the huge collection of stereophonic-content that currently exists, from music to film, makes it difficult for the audio industry to progress, in spite of the possibility for improved quality of experience. This emphasises the need for a detailed understanding of the perception of loudspeaker-based spatial audio reproduction systems.

3.3 Perception of Audio Reproduction Systems

The perceptual assessment of spatial audio systems has been an area of investigation since the seminal designs of Steinberg and Snow (1934). The complexity of the human auditory system means that the most accurate method of evaluating system performance is using subjective evaluation. However, the underlying principles leading to the perception of good or bad sound quality is a complex and multidimensional concept (Bech and Zacharov, 2006).

Perceptual assessment of audio quality can be approached in two ways. One method is to ask participants to make ratings using more general impressions such as liking or pleasantness. However, it is also possible to split the multidimensional percept of sound into individual attributes such as loudness, pitch or sound location with stimuli and training chosen to evoke specific differences. Bech and Zacharov (2006) refer to these two different concepts as the integrative (for more global scales) or the analytical (for specific attributes) mindsets. In this report, testing is focused on the analytical mindset where ratings are based on specific underlying attributes.

Many auditory attribute lists have been defined and used (Gabrielsson, 1979;

Bunning and Wilkens, 1979; Gabrielsson and Sjögren, 1979), often found using methods such as interviews or statistical analysis like multidimensional scaling (MDS). Due to the growing number of available rating attributes for audio quality, Rumsey et al. (2005) investigated the relative importance of two main attributes for audio assessment: spatial and timbral fidelity. This study considered the prediction of Basic Audio Quality (BAQ) based on timbre and two spatial components. Results revealed that the changes in BAQ depend around two times more on changes in timbral fidelity than the two spatial fidelity attributes. The authors indicate the possible application of these results for acoustical design engineers when presented with a trade-off between spatial or timbral fidelities. It is therefore also applicable to use this weighting when considering the emphasis we place on subjective assessment of domestic reproduction systems. For the assessment of sound reproduction systems, Spors et al. (2013) provide a comprehensive summary of sub-attributes for spatial and timbral domains; these two lists are recreated in Table. 3.1 and Table. 3.2.

Table 3.1: Example spatial attributes from Spors et al. (2013)

Attribute	**Description**
Spatial fidelity	degree to which spatial attributes agree with reference
Spaciousness	perceived size of environment
Width	individual or apparent source width
Ensemble width	width of the set of sources present in the scene
Envelopment	degree to which the auditory scene is enveloping the listener
Depth	sense of perspective in the auditory scene as a whole
Distance	distance between listener and auditory event
Externalization	degree to which the auditory event is localized in- or outside of the head
Localization	measure of how well a spatial location can be attributed to an auditory event
Robustness	degree to which the position of an auditory event changes with listener movements
Stability	degree to which the location of an auditory event changes over time

Table 3.2: Example timbral attributes from Spors et al. (2013)

Attribute	**Description**
Timbral fidelity	degree to which timbral attributes agree with reference
Coloration	timbre-change considered as degradation of auditory event
Timbre, Color of tone	timbre of the auditory event(s)
Volume, Richness	perceived “thickness”
Brightness	perceived brightness or darkness (dullness)
Clarity	absense of distortion, clean sound
distortion, artifacts	noise or other disturbances in auditory event

Although many auditory attributes can be defined, a review of literature suggested that spatial and timbral attributes are most predominantly tested. For this reason, analysis of the literature on subjective perception will be focused on these two areas

specifically.

3.3.1 Localisation

Although results from Rumsey et al. (2005) show a greater importance for timbral attributes when predicting BAQ, the authors emphasise that spatial fidelity accounted for approximately 30% of the rating and is therefore an important concept when rating spatial audio systems. However, the term spatial fidelity encompasses many underlying attributes including direction, distance, source-size or spatial coherence. It should be highlighted that only sound source direction is primarily investigated in this thesis.

Many behavioural studies into the resolution of human localisation (Mills, 1958; Makous and Middlebrooks, 1990a; Perrott and Saberi, 1990; Carlile et al., 1997; Grantham et al., 2003) have indicated that in the frontal region, the minimum audible angle (MAA) in the horizontal plane, that is, the smallest change in angle of a sound event that is perceivable, is around 1°. An accurate spatial audio reproduction system should therefore aim to induce localisation cues with similar resolution.

Localisation was investigated for virtual sound sources in the earliest of systems (Steinberg and Snow, 1934) and since then, many localisation studies using loudspeaker-based spatial audio reproduction systems have been undertaken. Phantom sound source localisation using equally weighting signals on two loudspeakers was measured by Sandel et al. (1955) where results supported the dominance of ITDs in the low-frequency region and ILD in the higher-frequencies.

When considering the use of VBAP, Ambisonics and alternative amplitude-panning techniques, a number of experiments have been conducted to try to understand the performance of each.

Bamford (1995) considered first- and second-order Ambisonic panning methods when compared to stereo and *Dolby Surround* techniques using the objective metrics integrated wavefront error and wavefront mismatch error. General conclusions from the study indicated that an Ambisonic system was capable of achieving improved sound 'imaging'. After the introduction of VBAP, Pulkki and Karjalainen (2001) presents results on the localisation of amplitude-panned virtual sources where subjective results were compared against a computational model of human localisation. The results concluded that localisation ability was highly dependent on stimulus frequency due to ITD/ILD errors in the mid- and upper-frequency regions. Pulkki and Hirvonen (2005) also later implemented an auditory model to measure the spatial fidelity for Ambisonic, pair-wise panning and a spaced microphone technique (using 5 transducers). The model was compared against subjective test results and they found that large loudspeaker spacing in the ITU 5.0 layout caused problems for creating well-localisable auditory events with all techniques. Limitations for first-order Ambisonic reproduction were also found when the stimuli had higher frequency content, this is likely due to the upper-frequency artefacts caused by truncating the number of spherical harmonic coefficients and thereby introducing spatial aliasing. The localization performance of 2.5D[4] Ambisonic systems was investigated directly by Benjamin et al. (2006) where different sound stimuli and decoder types were tested. Decoders were optimised by maximising velocity (max_{rV}) and energy

[4]2.5D is the term used to described reproduction systems with loudspeakers positioned in the horizontal, 2-dimensional plane where loudspeakers inherently emit sound in all three dimensions.

(max_{rE}) vectors (Gerzon, 1992a) resulting in three types of decoder: max_{rV}, max_{rE} and dual-band max_{rV}/max_{rE} using a shelf filter. Shelf filtering was used to split the audio stimulus for separated processing bands with a 400Hz cross-over frequency. Localisation results were found to be different than predicted by the velocity and energy vectors under certain circumstances, specifically for the square array using four loudspeakers where panned sources at the front were different from those positioned at the front-diagonal. This highlights the limitations of using low number of loudspeakers or large loudspeaker spacing for Ambisonics reproduction. Localisation testing using a fixed 12-speaker array by Frank et al. (2008) found that increasing the Ambisonic order reduced localisation error and that off-centre listening positions suffered from reduced localisation accuracy. The authors also experimented with a delay compensation strategy by delaying loudspeaker input signals to achieve equal time-of-arrival at each listening position. This was tested for both central and non-central listening positions in an attempt to mitigate any artefacts caused specifically by time-of-arrival differences across the listening area and by loudspeaker positioning. Localisation results indicated an increase in front-back confusions when delay compensation was used but authors also indicated that sound colouration caused by the differing times of arrival in the system without delay compensation may have caused the lower reported mean-opinion score (MOS) values for 1st order Ambisonics, due to temporal artefacts. In-phase Ambisonic decoding was also tested by the authors following the original definitions by Malham (1992), whereby Ambisonic gain coefficients are mathematically constrained to only have positive gain values, thereby reducing energy contributions from the opposing side of the intended phantom-source position. However, it was found that in-phase decoding performed worst of all decoder combinations. More recent tests by Frank (2013) investigated phantom sound source localisation with a regularly spaced 8 loudspeaker array using 4

panning methods: VBAP, Multiple Direction Amplitude Panning (MDAP), and Ambisonics in max_{rV} and max_{rE} decoding variants. MDAP is an extension in VBAP where multiple loudspeakers are used at any time to increase the perceived width of an auditory event. Experiments were conducted at central and non-central listening positions. It was found that VBAP had the largest absolute deviation of localisation on average. At off-centre positions, max_{rV} decoding was found to create multiple auditory events and for all panning methods at off-centre listening, localisation to the nearest loudspeaker was reported which can be explained by the precedence effect.

On a more general level, Majdak et al. (2010) has presented work on the variability in conducting sound localisation experiments considering aspects such as the pointing method (head or manual pointing) and whether a visual environment was presented in the test. The effect of the pointing method was not found to be significant but the use of a virtual visual environment improved localisation precision. In a second experiment it was found that a training session improved the localisation accuracy of the listeners significantly. The work of Letowski and Letowski (2011) also provides thorough documentations of procedures and analysis methods for conducting localisation tests.

Some general conclusions should be made from the literature of localisation in loudspeaker-based reproduction systems.

- Spatial audio systems generally reduce the spatial fidelity of auditory events when compared to a single-speaker reference.
- The benefits of different amplitude panning methods such as VBAP,

pair-wise or Ambisonics are not clearly defined, often depending on source material, listening position and number of loudspeakers.

- Localisation studies have highlighted significant perceptual problems of using wide loudspeaker spacing for spatial audio reproduction

3.3.2 Colouration

The American Standard of Acoustical Terminology (American Standards Association and Acoustical Society of America, 1960) define timbre as

> "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."

Salomons (1995) later proposed a definition of sound colour (which incorporates timbre, rhythm and pitch) as

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

Where

> "The colouration of a signal is the audible distortion which alters the

natural colour of a sound"

Importantly, colouration is judged best when compared to an explicit reference signal (therefore avoiding the use of an implicit reference by a listener). It has also been shown that due to the effect of binaural decolouration in the human auditory system, colouration is perceived differently depending on whether the sound is perceived monaurally or binaurally (Zurek, 1979; Salomons, 1995; Brüggen, 2001; Buchholz, 2007).

Rather than describing what sound colouration *is*, the definition of colouration describes sound colouration as a difference when specific attributes are matched between two stimuli. This means that colouration can be characterised by many features and can be caused by processes such as room reflections (delayed, coherent and often filter versions of the direct sound sources) or electroacoustic transducer characteristics (changes to the signal's magnitude and phase). Although only one type, comb-filters have been used extensively when testing and characterising the perception of sound colouration (Atal and Schroeder, 1962; Bilsen, 1968; Bilsen and Ritsma, 1970; Zurek, 1979; Kates, 1985; Buchholz, 2007, 2011).

Colouration is the perceptual attribute caused by spectral alterations made to a sound stimulus. The colouration of a stimulus signal can occur when delayed but correlated signals are summed, resulting in comb-filtering (Toole, 2008) and sometimes a perceived pitch (Fastl and Zwicker, 2007, p. 126). Colouration has also been well investigated in the discipline of room acoustics (Beranek, 1996; Halmrast, 2001) where early reflections can cause complex spectral alterations to a listener. For listeners of a loudspeaker-based spatial audio reproduction system

in a non-anechoic listening environment (as used in this thesis), spectral artefacts will be induced caused by issues such as the effects of direct signal paths from loudspeakers, reflected signal paths from room reflections, room reflection characteristics, loudspeaker performance (directivity) and low-frequency room modes. For the perceptual assessment of colouration in spatial audio reproduction systems the magnitude of colouration perception in Ambisonic and VBAP reproduction systems is not well understood, especially when variations across the listening area are considered.

Some studies in the literature have considered the subjective effect of colouration in loudspeaker-based reproduction systems. Frank (2013) performed a subjective evaluation of colouration in Ambisonic and amplitude panning reproduction methods, specifically focused on the colouration induced when a virtual sound source is slowly panned around a listener. Colouration was found to be highest for VBAP with a smaller spacing between loudspeakers. Results indicated that reducing loudspeaker spacing had the effect of increasing perceived colouration. This indicates that comb-filtering caused by smaller delays (due to less differences in loudspeaker-to-ear path lengths) may induce higher perception of colouration. This is an interesting outcome that may be counter-intuitive as the decrease in loudspeaker spacing is often shown to improve localisation fidelity.

A report by Augspurger (1990) describes the effect of phantom centre speaker colouration in mix-down rooms whereby the author describes the use of third-octave pink noise signals to reveal a distinct null at 2kHz. The feature is reported to be so pronounced that it can be used to check mismatched phase in left/right drivers. Continuing on this topic, Shirley et al. (2007) measured speech intelligibility for real and virtual centre sound sources created using a

stereophonic loudspeaker layout. A 4.1% increase in correctly identified keywords was found for the real speaker. This result is likely caused by the first comb introduced using the stereo system, often at around 2.5kHz and directly related to the interaural time difference of the listener. This demonstrates the measurable impact of virtual sound source colouration from amplitude panning methods where coherent, delayed signals are summed at the listeners ears. This feature of amplitude panning was further described by Choisel and Wickelmaier (2007, Figure. 4) where the authors reported a reduction in brightness when a mono centre loudspeaker was replaced by a stereo phantom image. Two- and three-loudspeaker amplitude panning configurations were also used by Pulkki (2001) to measure the colouration effect in virtual sound sources. An interesting outcome of *this* study was that the inclusion of room reflections reduced the perception of colouration.

From this literature it is clear that colouration is audible when virtual sound sources are created using amplitude panning methods with two or more speakers. However, currently there is no standardised method for conducting a subjective evaluation of colouration in spatial audio reproduction systems. The 'timbre' perception of different wave-field synthesis systems was considered by Bruijn (2004) where a Thurstonian methodology using paired comparisons was undertaken. For that study, headphones were used to auralise the signals recorded by microphones placed at the positions of a listeners ears within the WFS listening area to approximate auditory perception. Due to colouration needing an explicit reference, two pairs were presented to the participant in each trial where the most dissimilar (in terms of colour) was selected. Comparisons of loudspeaker spacing were made between pairs where each pair consisted of the central and 1 non-central listening position. This means that colouration was

always rated relative to the CLP and thereby assuming this was the highest reference. Wittek et al. (2007) has also used a MUSHRA style subjective test to understand the intra-system colouration differences using WFS and stereophonic reproduction.

The most recent study into reproduction system colouration was conducted by Wierstorf (2014) where WFS systems with different loudspeaker spacings were compared in terms of the attribute 'timbre' using a MUSHRA style direct-scaling method. In a smaller additional study, different listening positions for one WFS system is also considered. Results here indicated that smaller spacing reduced the perceived colouration which differed from results of Frank (2013) for amplitude panning methods. All off-centre positions were considered relatively equally coloured apart from the central listening position which showed reduced colouration ratings. All testing was performed using (anechoic) HRTF simulation and head-tracking was not used. Indirect assessment (using paired comparisons) of colouration was also considered in another study by Merimaa (2006) where participants judged colouration alongside localisation, envelopment and other artefacts following an ITU-R BS.1284-1 standard. Three reproduction methods were analysed: Spatial Impulse Response Rendering (SIRR), Phase-randomisation and first-order Ambisonics all reproduced using a 3-D 16 channel loudspeaker layout in anechoic conditions. Although not possible to derive colouration effects independently from this study, on average SIRR had the highest graded difference whereas Ambisonics and phase-randomisation had significantly lower values.

This section has provided a summary of literature on the subjective perception of audio reproduction systems. Although localisation and direction-of-arrival

attributes are popular metrics in this field, colouration and timbral artefacts have been shown to be perceptually important. Colouration in loudspeaker-based reproduction systems has been considered in a number of experiments, but no standardised method for conducting experiments has yet been defined. Although not addressed directly in this thesis, the standardisation of experimental procedures on the perception of sound colouration would allow for higher quality comparisons between experiments. The lack of standardisation must be taken into consideration when designing future experiments to understand the perception of colouration. A number of studies have also revealed the perception of colouration artefacts at central and non-central listening positions caused by changes in the time-of-flight between coherent sound sources to each ear. This raises the question of whether colouration induced by smaller delays at the central listening position is perceptually better or worse than larger the time-of-flights found at non-central listening positions, and how this factor is affected by alternative panning methods and room reverberations.

3.4 Modelling Human Perception of Audio Reproduction Systems

Predictive models play an important role in estimating the human perception of audio quality. When proven adequate, these models can replace subjective evaluation of different sound stimuli. Often, models are structured in two parts; first a model of the human auditory system, then by some form of cognitive scoring model (Bech and Zacharov, 2006) dependent on the attribute of interest. Models exist for the evaluation of specific auditory attributes such as localisation (often called direction-of-arrival or DOA estimation), colouration or loudness and for more global percepts such as perceived quality. Recent contributions such as the Auditory Modelling Toolbox (Søndergaard et al., 2011) allow simplified

implementation of some of the fundamental building blocks for modelling the auditory system.

A simple example of a useful predictive model is the multichannel loudness algorithm described in Recommendation ITU-R BS.1770-3 (ITU-R, 2012a). The main element of this recommendation presents a method to determine subjective programme loudness from loudspeaker driving signals. The algorithm is split in to four stages: (1) physiologically-inspired frequency weighting, (2) mean square calculation for each channel, (3) channel weighted summation and (4) two-stage blocked gating. This provides, with some variability, an objective measure of perceived loudness for programme material.

When considering the design of loudspeaker-based spatial audio reproduction systems, auditory modelling provides a valuable tool for the evaluation of improved quality. Simple models for localisation have been defined since the birth of Ambisonics (Gerzon, 1992b) but with the development in computational power and more advanced models, it is now possible to predict human perception with increased accuracy. This section of the literature review will now discuss the important literature for the modelling of human judgements on sound localisation (direction of arrival or DoA) and sound colouration relevant to the specific application of quality assessment in loudspeaker-based reproduction systems.

3.4.1 Localisation

The function of the outer, middle and inner ear are well understood but the function of binaural and psychological stages of localisation has yet to be fully defined. However, many computational models have been applied to the prediction of spatial fidelity in loudspeaker-based spatial audio reproduction systems. One of the earliest models for predicting direction-of-arrival of auditory events in Ambisonic systems was defined by Gerzon (1992a). This work defined a number of auditory models for the perception of Ambisonic systems, two of which were the velocity and energy vectors. Total vectors, one for velocity and one for energy could be calculated from loudspeaker directions and gains. Although no physiological modelling was included, this high-level model could be implemented to predict the direction and 'confidence' of an auditory event. The low computational cost of these metrics also allows for the implementation into iterative optimisation algorithms such as those used by Wiggins (2007).

Pulkki et al. (1999) implemented a computation model of auditory perception to predict colouration and localisation of virtual sound sources using different reproduction methods. Auditory nerve signals were achieved by firstly modelling the outer ear using a HRTF. Middle-ear processing was omitted but the inner-ear's frequency selectivity was modelled using a 42 ERB-band gammatone filter bank and half-wave rectification. ITD, ILD and IACC cues were calculated per ERB band. Comparisons on interaural cues were then made between real and virtual sound sources as an estimate of spatial fidelity. Amplitude and time-delay panning methods were tested. The model was able to predict some key features in loudspeaker listening such as the increase in localisation error at high-frequencies and difficulty in positioning virtual sound sources at the sides of

a listener.

A development of the model was later reported by Pulkki and Hirvonen (2005) where peripheral processing stages were the same. Higher level perceptual stages were modelled using a database of stored ITD and ILD values for sounds at known angles. ITD and ILD 'angles' (ITDA/ILDA) are then computed for the virtual sound sources created with loudspeaker panning methods. Comparing the model results with a listening test it was found that the model performed generally well although in some situations deviations were found. The authors noted that the model suffered most for sound sources farther from the median plane.

One of the most difficult tasks when modelling the human auditory system is resilience to multiple sound sources and room reflections. Faller and Merimaa (2004) introduced a modelling mechanism so that direct sound sources can be localised whilst reflections and concurrently arriving sounds can be ignored. The general concept is that when analysing binaural nerve firing densities at the basilar membrane, only use ITD and ILD values at time instances when the interaural-coherence is above a certain threshold (therefore considered 'valid'). This theory is supported by the human ability to detect very small reductions in inter-aural coherence (Goupell and Hartmann, 2006). The mechanism was applied in a framework of binaural modelling for auditory periphery. The model was validated against a number of psychoacoustic phenomena reported in previous literature.

This mechanism was later applied in the PhD thesis of Sheaffer (2013) where a 'decision-maker' was included to achieve a single localisation direction and

confidence metric. Once binaural cues were achieved by the modelling mechanism proposed by Faller and Merimaa (2004), ITD and ILD cues in each critical band were compared using a 2-D cross-correlation with a referenced dataset to predict angular localisation judgements in each band. Statistical analysis was then applied to achieve across-frequency integration. The localisation model was validated under single-source localisation scenarios but importantly, also applied to stereophonic listening using level- and time-panning methods. The model was able to replicate subjective results for level difference and time-delay panning, summing localisation, localisation dominance and echo phenomenon could all be modelled. However, due to similarity of interaural cues around the cone of confusion, the model was not tested at azimuth angles greater than $\pm 90°$. Two models for the localisation and lateralisation of sounds were also implemented by Park (2007). The characteristic-curve (CC) model was implemented by comparing characteristic curves for free-field listening whereas a second model was based on pattern-matching (similar in nature to the model proposed by Sheaffer (2013)) on excitation-inhibition cell activity patterns. A characteristic curve represents the ITD and ILD values for discrete sound sources plotted against each other (ITD on x-axis, ILD on y-axis). Unknown ITD/ILD values can then be tested against the curve using nearest-neighbour processing. In Park's thesis, the models are compared to subjective data in the context of pairwise amplitude-panning systems, which will be discussed in more detail in Chapter. 6.

A computational model for direction-of-arrival estimation of concurrent speakers has also been recently presented by Dietz et al. (2009) and later with the addition of ILD processing stages (Dietz et al., 2011). Instead of interaural-coherence used by Faller and Merimaa (2004) as a binary reliability metric, an alternative metric based on interaural-phase difference fluctuations

was used (IVS_G). The authors reported that both IC and IVS_G gave identical results. This model was validated for five concurrent speakers in free-field and three concurrent speakers in the presence of noise where the direction of arrival prediction errors were always less than 5°. For the assessment of spatial audio reproduction systems, Wierstorf (2014) applied this model to the assessment of WFS and NFC-HOA systems where results were compared against subjective localisation results. Similar data using a computational localisation model was also presented by Takanen et al. (2014).

More recently, the use of head-movements to improve localisation models have been considered. Ashby et al. (2014) investigated the importance of head-movements in the localisation of sounds to understand the need for head-movements in modelling localisation in the vertical plane. Results indicated that the dynamic cues used whilst head-movements were made improved localisation. The implementation of head-movements to a localisation model was also performed by Braasch et al. (2013), where a small range of head-movements were programmed into the interaural cross-correlation modelling method for a small temporal window of head-movements.

The work of Conetta (2011) introduced the *QESTRAL* model which was designed to objectively predict spatial attributes of loudspeaker-based spatial audio systems. As opposed to considering localisation error specifically, this model attempts to predict the more global attribute of spatial quality, encompassing many different underlying elements. The model was found to be capable of predicting spatial quality with 11.06% error. The model used 14 signal analysis metrics to calibrate the prediction such as mean interaural cross-correlation, weighted frontal-localisation error and spectral roll-off.

A literature survey has been conducted of the currently available models for human sound localisation applicable to the assessment of loudspeaker-based spatial audio systems. It can be seen that many of the models have been tested in anechoic environments and often the ability to resolve front-back confusions is not been defined. The technical comparison of models found in the literature with the model applied in this thesis is further presented in Chapter. 7.

3.4.2 Colouration

The perception of colouration artefacts in spatial audio systems is also considered specifically in this thesis, although no attempt is made to model the perception of colouration. Compared with localisation and direction-of-arrival estimation, fewer researchers have attempted computational models for colouration perception. For colouration artefacts in amplitude panning systems, Pulkki (2001) considered the change in timbre when compared with a real sound source. The author first considered the comb-filtering introduced in stereophonic listening. 20 different individualised HRTFs were then used to calculate a composite loudness level spectrum which predicted, for stereophonic listening, a large null at around 2kHz. The model was then applied to reverberant listening scenarios which revealed that colouration would likely be reduced under reverberant conditions.

Models for colouration prediction were considered by Wittek et al. (2007) for the analysis of colouration in wave field synthesis. Wittek presents an overview of the objective predictors for colouration perception. The A_0 criterion (Atal and Schroeder, 1962) is the earliest metric of sound colouration whereby it is predicted that sound colouration is perceivable if the level difference between

minimum and maximum values in the short-time power spectrum exceeds the threshold A_0. The B_0 (Bilsen, 1968) criterion utilises a similar concept but this time, colouration is perceivable if the ratio of the maximum value of the short-time autocorrelation function for any non-zero delay to the value at zero delay exceeds the threshold B_0. However, these models are not applicable to delays shorter than 10 ms and cannot predict colourations outside the short-time window. Models were however later developed by Salomons (1995) to include the characteristics of auditory filtering. Berkley (1980) also defined the criteria of spectral deviation, whereby the standard deviation in the log-magnitude frequency response is used to predict colouration. Both A_0 and spectral deviation criteria were implemented by Wittek et al. (2007) where it was found using multiple regression that the A_0 criterion and spectral deviation metrics have $R^2 = 0.76$. However, Wittek finally concluded that there was no available model which was a reliable for the prediction of colouration in spatial audio reproduction systems.

Composite Loudness Levels (CLLs) were also used by Frank (2013, p. 90) to model colouration in amplitude panning systems. Composite loudness levels approximate the perceived loudness across frequency by modelling the peripheral processing of the human auditory system (middle ear, cochlea, auditory nerve). CLL was calculated by measuring incremental differences in the sum of third-octave band loudness levels across panning direction. Although the CLL does not have any modelling stage for the binaural decolouration phenomenon, using incremental differences helps to normalise the large differences in CLL that would be experienced for example between the phantom source at 0° and 60°. This model was then applied to the perception of colouration in 1, 2 and 3 loudspeaker systems as well as VBAP and Ambisonic systems. The validity of

the simplified CLL metric was not assessed directly. However, contrary to the work of Thiele (1980), it was found that colouration was perceivable using amplitude panning systems and VBAP at the central listening position caused the highest colouration compared with max_{rv} and max_{rE} Ambisonic systems. Berkley (1980) found that subjective responses to colouration were well correlated with the variance in the stimulus frequency response. This led to the criteria of 'spectral deviation' to predict stimulus colouration. However, this model did not differentiate between periodic and random frequency response fluctuations.

Binaural decolouration is the phenomenon by which the human auditory system's binaural processing is utilised to achieve a reduced perception of the colouration than, for example, if a listener were to hear with only one ear (Salomons, 1995). The delaying of interaural signals such as that caused by the difference in time-of-arrival to the two ears has been shown to reduce colouration acuity (Zurek, 1979; Salomons, 1995). One of the main limitations with using the currently available computational models (such as CLL or spectral deviation) for sound colouration is the inability to model binaural decolouration processes by the human auditory system. This would likely result in an overestimation in the magnitude of colouration predictions. For the models often applied to measure colouration detection thresholds (CDT) such as A_0 and B_0 criteria (Atal and Schroeder, 1962; Bilsen, 1968; Salomons, 1995; Buchholz, 2011), the application to more complex types of colouration has not yet been defined. Toole (2008, pp. 151) provides a selection of systems (both physical and perceptual) which help alleviate the perception of colouration in listening rooms; one of these being the ability of a human listener to spectrally 'adapt' to constant types of colouration. This could be, for example, the constant colouration provided by a single

listening room (and caused by reflections). This type of adaptation is also not currently represented by commonly used computational models for colouration. Similar to the conclusion by Wittek et al. (2007), the literature survey concludes that no model is currently able to predict the multi-dimensional percept of sound colouration. Therefore, the modelling of colouration induced by loudspeaker-based spatial audio systems across the listening area for quantitative analysis is not attempted any further in this thesis.

3.5 The Listening Area

Until only recently, the assessment of spatial audio systems has focused on the central listening position. This is referred to as the 'best' listening position (Toole, 2008) and is reported to be geometrically equidistant from the loudspeakers. Due to the reality of spatial audio reproduction in the domestic listening environment it has become increasingly important to consider the sound quality at non-central listening positions.

Landone and Sandler (2000) proposed a method for the simulation of non-central listening positions using binaural simulations. In the same year, Johnston and Lam (2000) presented results on the subjective judgement of the 'sweet spot' by asking listeners to place flags on the floor to indicate an area of acceptable listening. Results indicate that the area of acceptable listening was larger for a 5-channel system in comparison to a 2-channel system. Johnston also reported the limitation of the test not being blind and thereby directly suggesting the need for such simulation methods used in this thesis.

3.5.1 Localisation

Localisation and spatial impression was investigated by Kamekawa (2006) at the central and a number of non-central listening positions. For attributes of localisation and spatial impression, results indicated a general reduction in average scores as the listener moved away from the central listening positions. As discussed in Section. 3.3, Frank et al. (2008) performed localisation and mean opinion score evaluation using a 12-channel Ambisonic system at the central and one non-central listening position. Conclusions were presented that localisation accuracy deteriorates at the non-central listening position. The authors also experimented with delay calibration of loudspeakers signals for each listening positions which, surprisingly, gave worsened results. Frank (2013) considered the listening position factor when assessing localisation and colouration artefacts in a number of amplitude panning systems. Results indicated that for the $max\,r_v$ Ambisonic system where all loudspeaker contributed to the sound field, off-centre listening suffered from splitting of the auditory event and often localisation towards the nearest contributing loudspeaker. Tests for localisation accuracy were also conducted by Bates et al. (2007a) using in situ elicitation of sound source direction at 9 listening positions for loudspeaker arrays in a small concert hall. As a benchmark, monophonic reproduction was tested which gave mean reported localisation results at each listening position as within 5° of the actual direction. The test was then repeated using amplitude panning methods (Spat 1st order Ambisonics, 2nd order Ambisonics, VBAP and Delta Stereophony) over an 8 loudspeaker array. Localisation results for 2nd order Ambisonics showed that the seating position furthest from the CLP had comparable or sometimes better localisation accuracy than the CLP. 1st order Ambisonics showed significant localisation problems at non-central listening positions and also shared the problems of localisation towards the nearest contributing loudspeaker the

VBAP and Delta Stereophony systems. The results also indicated that higher-order Ambisonics performed better at non-central listening positions than 1st order Ambisonics which is inline with theoretical derivations of reconstruction accuracy (Daniel, 2001).

Investigations using the more integrative quality of evaluation have also been conducted for different listening positions across the listening area. Assessment using different multi-channel microphone techniques were reported by Peters et al. (2007) where precedence effect, comb filtering and proximity effect were defined as some of the fundamental artefacts causing degradation at non-central listening positions. Binaural recordings were made at each of the listening positions in two large listening spaces for each of the reproduction systems and then compared by listeners in a blind-comparison using headphones. Results were found to be highly affected by the stimuli choice. A paired comparison test was implemented where participants made ratings of perceived degradation. Results indicated that degradation was increased at non-central listening positions, but no analysis on the underlying attributes was made. Peters further developed the work on off-centre sound quality degradation in later years working towards his PhD thesis. In this thesis (Peters, 2010) qualitative and quantitative studies were conducted into overall perceived 'quality' at 5 listening positions using spatial audio reproduction. The first qualitative study found that four attributes could be used to describe the overall quality degradation where 'timbre' was the most dominant (similar to Rumsey et al. (2005)). In a second quantitative experiment which asked participants to judge timbre, loudness, position and reverberation, the effect of the three geometrical features of off-centre listening (direction, delay and level differences) were analysed. It was found that differences in levels of loudspeakers contributed most strongly to

sound degradation and although time-of-arrival difference is the dominant factor causing comb filtering, it was found to be the geometrical factor that had a lesser effect. The use of loudness as dependent variable in the test may have placed more emphasis on the perception of this geometric feature.

Stitt et al. (2014) presents the mostly recently documented reports on localisation abilities across the listening area. Using in situ presentation, auditory event directions are reported via a reverse acoustic target pointer for a number of listening positions and two Ambisonic reproduction systems. The reverse acoustic pointer methods allows for a user to have physical control over the direction of a perceived 'reference' auditory event (created using a high quality panning method or individual speakers) which is then matched in direction to the perceived 'test' auditory event. Once matched, the direction of the acoustic pointer is recorded as the reported direction. In line with previous reports, higher-order Ambisonic reproduction had reduced overall localisation error compared with the tested 1st order systems. It was reported that localisation error was increased at off-centre listening positions and the outcomes could not be explained by a the theory of precedence effect or law of the first wavefront. The report also commented on timbral artefacts at non-central listening positions with some participants reporting two simultaneous auditory events; indicating a splitting of the virtual image. However, no formal testing was presented on these two aspects.

3.5.2 Colouration

For the application of linear-array wave field synthesis systems, non-individualised dynamic binaural synthesis was implemented by Wierstorf et al. (2012a) to test localisation across multiple listening positions. Results indicated that again, for systems with a low number of loudspeakers, localisation accuracy deteriorated for non-central listening positions. An array with 19 cm loudspeaker spacing performed well. Work following this (Wierstorf et al., 2014) went on to consider colouration variations across the listening area for different wave field synthesis systems. A MUSHRA-style (ITU-R, 2003) test was implemented for direct-scaling of colouration and results showed that larger loudspeaker spacing introduced increased colouration at the central listening position, but when comparing across the listening area, colouration differences were small (maximum range of 3 points on a 10 point scale) with a mean standard deviation across the listening area of 1.2 points.

Very few studies have measured the subjective elicitation of colouration in domestic spatial audio systems.

3.5.3 Colouration at the Central Listening Position

It is important to carefully consider the reference for subjective assessment of colouration. It would be feasible to consider that the central listening position is a good reference for the 'best' the system can perform, but many studies have shown that at the central listening position, even in stereophonic reproduction, large dips at around 2 kHz are present due to both ears having delays from each loudspeaker (Augspurger, 1990; Pulkki, 2001; Choisel and Wickelmaier, 2007; Toole, 2008; Shirley et al., 2007; Vickers, 2009).

In Ambisonic reproduction, the underlying design of the system using spherical harmonics can be mathematically reduced to simple amplitude weightings depending on the Ambisonic order, number/position of the loudspeakers and the assumption of simulating sound sources at the same, far-field distances as the reproducing loudspeakers. Sometimes these weightings are split in to two frequency bands. Therefore the inherent problems of colouration in phantom imaging are also apparent for Ambisonic reproduction. Yao et al. (2015) discusses the issue briefly. This means that even for 3rd order HOA, the dominant effect on colouration is likely to be audible in the sensitive 2 kHz region. Although the delay that induces the comb-filtering is smallest at the central listening position compared with larger delays when moving off-centre, it is not well known how these different comb-filter structures are perceived or which is considered 'worse'. When modelling colouration, Atal and Schroeder (1962) and later Bilsen (1968) implemented exponential windows to model human perception of comb-filters, meaning that colouration induced by larger delays had less perceptual significance (due to the physiological auditory windowing) than delays arriving early, which implies that perceptually, colouration induced at, or near the central listening position could be more perceptually detrimental than colouration at off-centre listening positions. The comb-filtering causing the 2 kHz problem is, however, a function of both the listener's head rotation and the position of contributing loudspeakers and will therefore be correlated to a listener's movement.

Conetta (2011) also tested the spatial perception of various audio reproduction systems resulting in an objective model for spatial fidelity. Results from a subjective listening test indicated that spatial quality was influenced by listening position, inline with other literature.

Previously reported literature presented above shows that there has been a recent increase in interest for the perceptual assessment of spatial audio systems at non-central listening positions. Localisation has been the most commonly used audio attribute but colouration has also been reported upon and tested. Many results have supported the objective theory of reduced spatial fidelity at off-centre listening positions with a specific emphasis on the splitting and collapse of auditory events towards the nearest contributing loudspeakers. More global (integrative mindset) assessments have also been conducted. However, evidence shows that there is a lack of understanding in the field of colouration artefacts across the domestic listening area.

The subjective evaluation of spatial audio systems across the listening area has been undertaken in previous research. The most common methods for testing are:

- in situ - where the listener is physically positioned at the specific listening position under test
- artefact simulation - where specific artefacts of off-centre listening are parameterised and auralised to a listener for subjective evaluation
- simulation of an AVE - where the auditory environment perceived at a specific listening position is captured and reproduced to a listener

Although in situ tests have been undertaken, the inability to perform direct-blind comparisons is a problem. Unlike localisation tests which use a relatively objective attribute (albeit with some listener dependent variation), auditory attributes such as colouration require indirect scaling procedures where comparisons need to be made. The plausible or authentic simulation of domestic loudspeaker-based spatial audio reproduction systems using dynamic binaural synthesis allow for such test

designs. This method also means that the auditory perception of different listening positions can be compared directly whilst the auditory experience is still plausible. Auditory perception is a multimodal percept that is largely affected by our visual system, expectations and memories among other things. The simulation method using an AVE also provides the ability to run blind testing whereby listeners are not aware of their physical localisation within a room or loudspeaker array.

3.6 Binaural Simulation of Loudspeakers

Binaural simulation systems are categorised by using features of the effect on sound transmission from a sound event to the ear drums of a listener to induce auditory events. This effect on the transmission path is represented by the Head-Related Transfer Function measured from a point in space around a listener to a measurement point inside, at or near to the entrance to the ear canal. In binaural simulation systems, the HRTF allows cues to be reproduced to a listener using headphones and induce externalised auditory events markedly different from headphone reproduction of stereophonic signals common in personal entertainment devices. The roots of binaural simulation can be dated back to the Théâtrophone presented in the Electrical Exhibition at the Paris Grand Opera in 1881 (designed by Clement Ader), albeit with a more practical philosophy. Recent improvements in computing power/memory and an increased understanding of psychoacoustic perception has led to the popularity of such systems for entertainment (Staff Technical Writer, 2006) and simulation purposes (Rychtarikova et al., 2009; Olive and Welti, 2008; Lindau, 2014). A full technical derivation of the binaural simulation method is shown in Chapter 2.

The HRTF is a function of source azimuth and elevation (Blauert, 2001). Distance

also becomes a factor for point sources if the sound event is less than 0.5 m from the head (Xie, 2013). Human HRTFs are highly individualised (Middlebrooks, 1999a) due to physiological differences between humans. Artificial heads are often used to measure or simulate an 'average' HRTF. Bell Laboratories (Snow and Hammer, 1932) introduced one of the first artificial head models but there are currently many options to choose from. The G.R.A.S. Head and Torso simulator (KEMAR) and the Brüel and Kjær Head and Torso Simulator (HATS) are two popular choices with many variations and accessories available and for binaural recordings, the Neumann KU 100 is popular and can be found in many recording and foley studios around the world. A large number of freely available HRTF datasets measured from both artificial heads and humans are also available as shown in Table. 3.3.

Table 3.3: A selection of freely available HRTF datasets for artificial and human heads.

Head	**Affiliation**	**Reference**
Artificial	Massachusetts Institute of Technology	Gardner and Martin (1994)
Artificial	TU-Berlin	Wierstorf et al. (2011)
Artificial	Fachhochschule Köln	Bernschütz (2013)
Artificial	South China University of Technology	Xie (2013)
Artificial	Club Fritz Project	Andreopoulou et al. (2015)
Human	U. C. Davis CIPIC Interface Laboratory	Algazi et al. (2001)
Human	Listen Project (AKG, Ircam)	AKG and Ircam (2002)
Human & Artificial	Acoustics Research Institute	Majdak et al. (2010)

Many studies have considered the importance of features of the HRTF. Following measurements and localisation studies, Searle et al. (1975) reported that left-right pinna disparities created localisation cues for the median sagittal plane and that the frequency response inside the ear canal changed as a function of

elevation angle. Sagittal plane localisation cues were also studied by Butler and Belendiuk (1977) where results also highlighted the importance of pinna cues for localisation in this plane. Weinrich (1982) later considered the problem of front-back confusions concluding that individualised characteristics of the HRTF are used by listeners to distinguish between front and back. More recent consideration of front-back confusions using a simulation system were undertaken by Zhang and Hartmann (2010) who noted the large individual differences on the use of pinna cues to resolve front-back confusions. It was also found that dips were more important than peaks for accurate front-back localisation.

Localisation studies comparing individualised and artificial head HRTFs have also been undertaken. Møller et al. (1996) found that localisation was comparable to real sound events when using individualised simulations but when listening to non-individualised recordings, front-back confusions were more frequent and median plane errors were increased. Møller et al. (1999) measured localisation ability of binaural recordings made on 8 and 10 artificial heads in two experiments. Again, localisation using the artificial head recordings were worsened but the position of measurement within the ear was not found to be significant. Minnaar et al. (2001) later reported in localisation tests with binaural recordings that localisation was much poorer with non-individual recordings and artificial heads generally performed worse that human heads.

Binaural simulations can achieve improved externalisation and plausibility when designed to include the response of a reverberant listening environment as well as free-field HRTF (Begault, 1992). Longer FIR filters can be defined which not only simulate the free-field HRTF, but also the response of the head, torso and pinnae to room-specific reflections and diffuse field response. This transfer function is called the binaural room impulse response (BRIR). Due to increase in

computer specifications and therefore realtime audio processing, the use of longer BRIRs in binaural simulation have become increasingly popular for binaural simulations (Rychtarikova et al., 2009; Olive and Welti, 2008; Lindau, 2014).

Head-movements are a fundamental factor for localisation (Wallach, 1940) and egocentric movements have shown to improve localisation ability (Thurlow et al., 1967; Perrett and Noble, 1997; Wightman and Kistler, 1999). Movements have also been shown to be important to localisation of elevated sound sources (Ashby et al., 2014). Due to the important features of the HRTF such as spectral peaks/notches, ITD and ILD being a function of head azimuth (and elevation) relative to the sound event it is acceptable to say that head-movements that do not change these cues in an accurate way (non-dynamic binaural simulation) may break down the dynamic localisation process. Work by Wenzel et al. (1990); Bronkhorst (1995); Sandvad (1996); Begault et al. (2001); Algazi et al. (2004a); Brimijoin et al. (2013) have shown the importance of dynamic cues.
Although localisation has been the predominant attribute of evaluating HRTF personalisation, timbral distortions have also been noted (Lindau, 2014; Völk, 2013; Rumsey, 2011). Silzle (2002) reported that expert-tuning of HRTFs and headphone compensation filters to simulate a 5-channel loudspeaker system reduced colouration and improved localisation. Another method to reduce timbral artefacts in HRTFs was investigated by Merimaa (2009, 2010) by attempting to reduce the RMS spectral sum of HRTF pairs whilst maintaining inter-aural cues.

Developments in realtime signal processing and low-latency head-tracking has meant dynamic AVEs have become a popular simulation method at many research institutions. This is achieved by low-latency head-tracking and real-time

filtering with an indexed HRIR or BRIR database. The earliest design of these types of systems can be found in papers by Wenzel et al. (1990); Bronkhorst (1995); Sandvad (1996). It is even plausible that the inclusion of dynamic cues may lessen the perceptual importance of HRTF personalisation (Fisher and Freedman, 1968; Begault et al., 2001; Algazi et al., 2004b). Many research institutions now have active research projects using dynamic binaural systems (both individualised and non-individualised) (Estrella, 2011; Wierstorf et al., 2012a; Lindau and Weinzierl, 2012; Völk, 2013; Pike et al., 2014).

Following a review of the relevant literature it is clearly acceptable to consider the use non-individualised dynamic binaural simulation to create an auditory virtual environment suitable for simulation loudspeaker-based reproduction in domestic listening environments. An *authentic approach* is therefore applied, defined by Novo (2005) as:

> 'the *authentic approach*, aims at achieving an authentic reproduction of existing real environments. The objective here is to evoke in the listener the same percepts that would have been evoked in the corresponding auditory real environments.'

Although the real environment exists, this method gives the ability to achieve fast, blind comparisons of multiple listening positions or different loudspeaker systems. The specific percepts (attributes) of localisation characteristics induced by loudspeaker-based spatial audio systems and acuity to differences in sound colour will be validated in this thesis.

Literature described above has shown that auditory events simulating using an

AVE have worsened localisation when using non-individualised HRTFs when compared to individualised. However, for the assessment of spatial audio reproduction systems across the listening area, the personalisation and acoustic stimuli auralised using headphones is impractical and expensive. The direct measurement of personalised BRIRs for the SBSBRIR dataset would be unachievable. Even the most recent methods of personalised HRTF measurement are prone to variability/errors from factors such as microphone positioning, human movement or reflections from objects within the measurement environment. Non-individualised, dynamic systems using BRIRs with accurate room reflections have shown to have good plausibility and externalisation of auditory events and such systems have become popular for the simulation of loudspeaker-based reproduction systems at the central listening position (Wittek et al., 2007; Lindau et al., 2012; Wierstorf et al., 2013). To assess artefacts of loudspeaker-based panning methods across the listening area, it must be shown that the localisation artefacts caused by the change in proximity to, and direction of loudspeakers are maintained. This will therefore validate the use for, and limitations of, such systems for listening tests.

Toole (2008) provides a viewpoint on colouration in AVEs which is applicable here:

> *The task of a sound reproduction system is to accurately portray the panorama of resonances and other sounds in the original sources, not to "editorialize" by adding its own.*

Literature has shown however that the a non-individualised dynamic binaural simulation is likely to introduce absolute colouration artefacts. However, it has

not yet been addressed whether the colouration between two systems under test is perceived equivalently between in situ auralisation of the systems, or binaural simulation of the systems. Due to the complex nature of binaural decolouration processes, colouration acuity could be increased or decreased when imperfect binaural simulations are used. Olive and Schuck (1995) indirectly considered colouration by measuring the perception of loudspeakers using both in-situ and non-individualised binaural simulation without head-tracking. The results for binaural simulation showed good agreement with in situ results however, when performing statistical analysis it was found that the binaural simulation gave a higher number of statistically significant interactions of test factors. This could mean that the binaural simulation increased the listener's acuity to colouration. Similar results were also found in Olive and Welti (2009). Hiekkanen et al. (2009) also looked at loudspeaker testing using in situ and binaural methods where all binaural simulations used some form of individualisation of the HRTFs. Wittek et al. (2007) considered the same problem of simulating loudspeakers for the perception of colouration in WFS systems. The test used a direct scaling method to compare the perception of colouration in a loudspeaker-based system for both in-situ and binaural simulation (referred to as the BRS system). The BRS system used non-individualised BRIRs recorded in the same room where the test was conducted. Results showed good similarity, however, only one participant was used. Völk (2013) also considered multi-band loudness perception which indicated the perception of colouration artefacts between auditory events created with real speakers and those created by binaural simulations. The most recent and, to this authors knowledge, only objective measurement of colouration perception using non-individualised binaural simulation is by Wierstorf (2014). This test looked at the magnitude spectral difference between two different dummy heads for a virtual sound source generated by a number of WFS systems and at four source directions. The difference was considerable (up to 15dB for

high frequencies) but mostly consistent for the different WFS systems and directions from which the author concluded that the effect was linear. This AVE was later implemented for colouration testing which provides some grounds on the applicability of such systems to measuring colouration artefacts. This objective methodology is a logical approach to try and understand whether the spectral artefacts of non-individualised binaural simulations will introduce errors when different listeners make perceptual ratings on sound colouration for simulated loudspeaker systems. However, the method does not attempt to model the complex perception of auditory events and how the higher-level stages of the human auditory system account for limitations in the binaural simulation system (for example limited head-tracking, non-individualised HRTFs, effects caused by headphone transducers). To allow for future researchers to easily simulate loudspeaker-based spatial audio systems and measure the perception of colouration, it is clear that a more in-depth evaluation is necessary.

As described above, the assessment of individual sound events using binaural simulation has been studied extensively for many attributes and metrics. However, the interaction of multiple loudspeakers used in domestic reproduction systems introduces a new set of problems. Not only is it important for the individual auditory events created by real speakers to be perceived equivalently, the artefacts induced by sub-optimal systems or listening environment should also be accurately induced. This provides the motivation for the AVE validation work presented in this thesis, to provide researchers with a reference for the simulation of localisation and colouration artefacts created by loudspeaker-based reproduction systems using non-individualised, dynamic binaural simulation and also document the limitations of such methods.

3.7 Conclusions

This chapter of the thesis has presented the current standing literature related to the five main topics presented in this thesis. Firstly the fundamental literature surrounding the use of loudspeaker-based spatial audio reproduction systems in domestic listening environments was presented. It was shown that spatial audio processing philosophies can be categorised into different types and that there is a dominance of stereophonic reproduction in domestic applications. Following this, literature on the the subjective perception (and modelling thereof) spatial audio reproduction systems was presented where localisation and colouration were highlighted as two of the most important attributes. A number of studies which have considered localisation and colouration artefacts across the listening area were then presented and discussed where it was identified that the perception of colouration artefacts at central and non-central listening positions is not well understood; specifically the perception of comb-filtering caused by coherent signals with different delays. Finally, the literature on binaural simulation of loudspeaker signals was introduced with a specific focus on the ability to use a non-individualised binaural simulation system to induce listening area localisation and colouration effects. It is concluded that although studies have compared the binaural simulation and in situ reproduction of individual loudspeakers, the ability of such systems to induce localisation artefacts presented by complex listening situations has not been validated. It is also shown that the equivalence of human acuity to sound colour differences using non-individualised binaural simulation and in situ has not be addressed.

CHAPTER 4

A Non-individualised, Dynamic Binaural Simulation System

This chapter presents the specific design details on the non-individualised, dynamic binaural simulation system used to create auditory virtual environments throughout this research project. System verification tests and specific details are presented to allow for experiments to be repeatable and applicable to future research.

4.1 Introduction

In an ideal scenario, a binaural simulation system would be able to reconstruct the pressure at the ear drum of a listener exactly matching that of the intended environment and created an equivalent auditory virtual environment. This simulation would even hold under dynamic movements of the listener in real time. In reality however, a perfect system would require a grand scale of resources and for some features is computationally unobtainable. Therefore, in this project a system is designed that aims to reconstruct an *approximation* of the pressure at the ear drum that is perceptually equivalent to that of the real environment for certain auditory attributes. The two auditory attributes chosen were the localisation and colouration artefacts perceived using loudspeaker-based spatial audio systems across the domestic listening area.

Head-related transfer functions and room reflections are contained within the binaural room impulse responses which were measured using a generic dummy head. Dynamic interaction is achieved by tracking the listener's head-movements and changing headphone input signals accordingly. For this project a Vicon[1] optical motion tracking system was used to track the listener's head position with 6-degrees of freedom (3 rotations, 3 translations). However, only head-azimuth rotations (yaw) were utilised by the binaural rendering system, meaning pitch, roll and translational motions had no effect on dynamic filtering. The Vicon system consisted of 4x Bonita cameras, passive tracking markers mounted to the headphones/headband of a listener and Vicon Tracker software. The system is termed *non-individualised* due to the filtering effect of anthropometric features being that of a single dummy head and not changing for each listener.

[1] https://www.vicon.com/

This chapter focuses on the technical details of the systems implementation including factors such as binaural room impulse response measurement, dummy head and measurement-point choice, headphone-equalisation and tracking latency.

The purpose of the AVE is to simulate relevant physical effects which induce artefacts of domestic loudspeaker-based spatial audio reproduction systems at a selection of listening positions across the listening area. The system is then implemented to consider the equivalence to in situ loudspeaker reproduction for the perception of localisation and colouration separately. The system is then also applied to further understand the perception of colouration artefacts across the listening area for two loudspeaker-based panning methods.

4.2 The SBSBRIR Dataset

The following section details the measurement procedure for the SBSBRIR dataset. The dataset was first presented as a poster (Melchior et al., 2014). Table 4.1 highlights the key details of the measurements. The SBSBRIR dataset is the first freely available dataset targeted for multiple listening positions. Files and extended information can be downloaded in WAV, SOFA and MIRO format from: http://www.bbc.co.uk/rd/publications/sbsbrir. The dataset webpage has received over 5500 unique visitors between April 2014 and March 2016 with more than 20 average dataset downloads per month.

Table 4.1: Key details of the SBSBRIR measurements.

Detail	Value
Listening Position Sampling Resolution	0.5m
Number of Listening Positions	15
Number of Loudspeakers	12
Loudspeaker Radius	2.1m
Loudspeaker Centre/Ear Height	1.06m
Measured Head-azimuth Resolution	2°
System Sampling Rate	48kHz

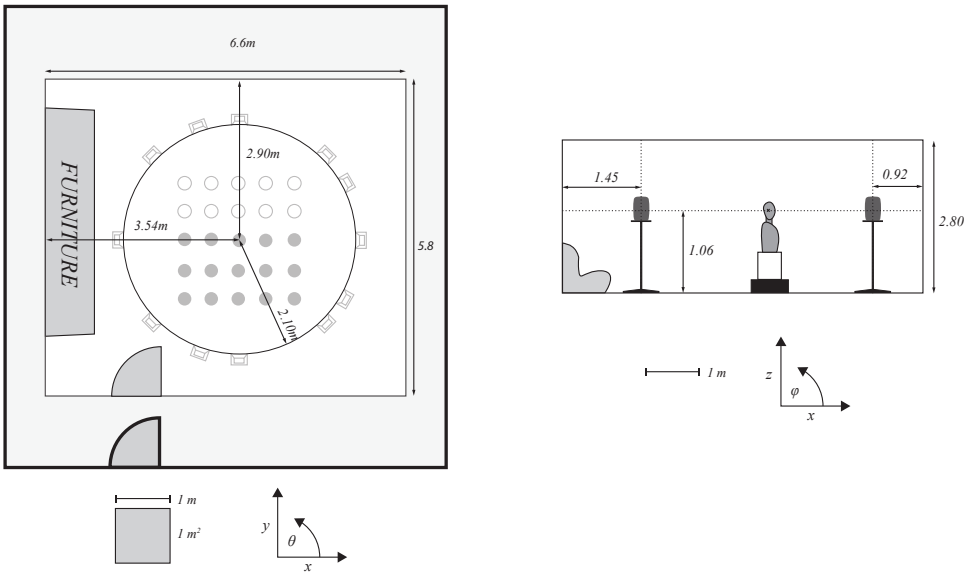
Three common measurement points are defined when measuring sound pressure at the entrance to the human auditory system (ITU-T, 2009): Ear Reference Point (ERP), Drum Reference Point (DRP) and Ear-canal Entrance Point (EEP). A B&K HATS Type 4100 was used to measure BRIRs and automated rotation was implemented using a B&K Type 9640 turntable. Silicon pinnae were fitted to the HATS which are anthropometrically feasible up to the entrance of the ear canal. B&K type 4190 free-field pressure microphone capsules were positioned to measure EEP pressures. No ear canal simulation was used therefore measurements simulated the pressure at the entrance to a blocked ear canal $P_{EEP}^{blocked}$. Genelec 8030a loudspeakers were used due to their popularity in research institutions.

Left and right ear microphones were calibrated to ensure no interaural level imbalances. Background noise measurements were made and analysis performed to check across-frequency signal to noise was adequate. BRIR measurements were made using a custom-written impulse response measurement software tool in MATLAB[2] which utilised the log-swept sinusoid method (Farina, 2007). Sweeps were overlapped to reduce the measurement time by beginning a neighbouring speaker's sweep before the current sweep had finished. The delay

[2]Chris Pike, BBC 2013

between sweeps starting was long enough to capture the full reverberant tail and resulted in increased non-linearities in the measured impulse responses after the deconvolution. An RME UFX USB audio interface was used to provide 6 analogue outputs, digital outputs were also routed to another digital-to-analogue converter (DAC) which provided the second 6 loudspeaker signals. Loopback signals were used to calibrate for level differences and delays between the different DACs.

Measurements were made in the BS.1116-1 (ITU, 1997) compliant listening room at the University of Salford over a two-week period. The measurements were set up as shown in Figures 4.1 and 4.2.



(a) Aerial view. (b) Side view.

Figure 4.1: A scaled diagram with aerial- and side-view of the measurement setup for the SBSBRIR dataset. Furniture, loudspeakers and mounting equipment are highlighted.

Figure 4.2: Photo of the SBSBRIR measurement positions. Markings on the floor indicate the 0.5m resolution measurement grid and loudspeaker positions relative to the size of the listening room. The photo graph is taken from position X = 3m, Y = 0m.

It should also be noted that all subjective evaluations using the non-individualised, dynamic binaural simulation system in this thesis were conducted in the same physical listening room that the BRIR measurements were made. Matching visual cues and auditory expectations to the auditory stimulus presented to the listeners is likely to have a significant positive effect on the plausibility of the auditory events. Users of the SBSBRIR dataset should consider this fact when using the measurements for binaural simulations.

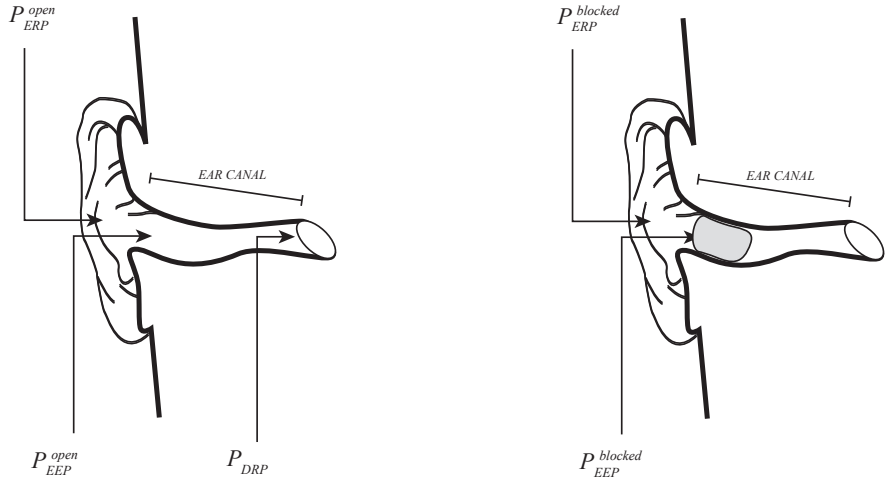
4.3 Ear Measurement Position

Section 4.2 defined that the blocked EEP position was used to make BRIR measurements on the HATS.

In a meta-analysis, Hammershøi and Møller (1996) used their own results and those from Wiener and Ross (1946); Middlebrooks et al. (1989) to show the

influences of changing sound source direction on the different measurement positions near to and inside the ear canal. Results for transmission from $P_{EEP}^{blocked}$ to P_{EEP}^{open} and P_{EEP}^{open} to P_{DRP} showed directional independence for up to around 12kHz and emphasised the fact that all three measurement positions shown in Figure 4.3 contain the cues for directional changes . This fact is also supported by results of Møller (1992).

Figure 4.3 shows the different reference measurement positions in a human auditory system.



(a) Open ear canal (b) Blocked ear canal

Figure 4.3: Ear measurement positions for both open and blocked ear canals scenarios.

Hammershøi and Møller (1996) further describe that the transmission from $P_{EEP}^{blocked}$ to the target P_{DRP} pressure can be achieved when two factors are accounted for:

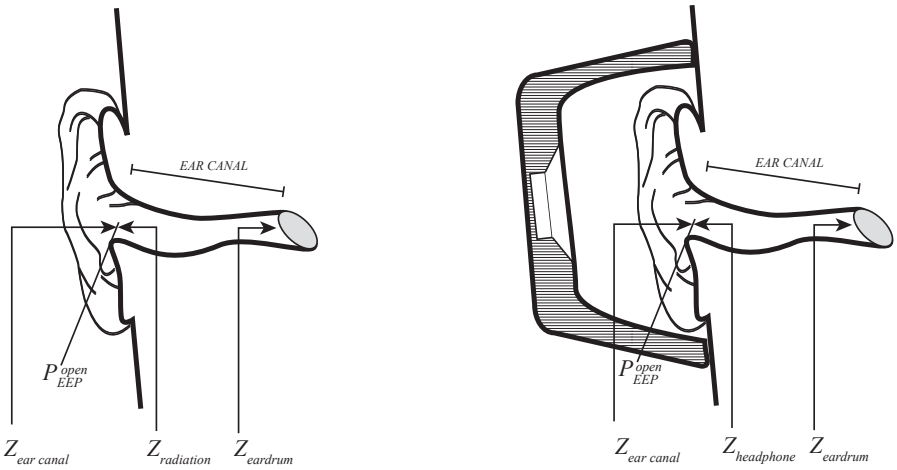
1. pressure division from $P_{EEP}^{blocked}$ to P_{EEP}^{open} when headphones are coupled
2. the transmission along the ear canal between P_{EEP}^{open} to P_{DRP}

Regarding point 1, Hammershøi and Møller (1996) describes a pressure division occurring at the entrance to the ear canal between the radiation impedance (the impedance looking outwards from the ear canal) $Z_{radiation}$ and the ear-canal input impedance $Z_{ear\ canal}$. This effect is described mathematically using Equation 4.1. Regarding point 2, if the AVE can simulate the P_{EEP}^{open} then transmission along the ear-canal is represented in-situ and therefore accounts for individualised differences in this transfer function caused by physiology of the ear canal. Møller (1992) confirms theoretically that transmission from P_{EEP}^{open} to P_{DRP} is identical for free-field or headphone listening.

If the pressure division is equivalent between free-air and headphone-coupled scenarios, then transmission from $P_{EEP}^{blocked}$ to P_{DRP} is also equivalent. However, this factor is dependent on the headphone choice.

$$\frac{P_{EEP}^{open}}{P_{EEP}^{blocked}} = \frac{Z_{ear\ canal}}{Z_{radiation} + Z_{earcanal}} \tag{4.1}$$

Figure 4.4 demonstrates the situation where $Z_{radiation}$ is affected by headphones being coupled. $Z_{headphone}$ is dependent on the headphones used for transmission.



(a) Free-Air Impedances. (b) Headphone-coupled Impedances.

Figure 4.4: Impedances in the outer ear. Pressure division at the entrance to the ear canal when headphones are mounted are different to the free-air scenario due to a change from $Z_{radiation}$ to $Z_{headphone}$.

Møller et al. (1995b) used the term 'Free-air Equivalent Coupling' to define a set of headphones where $Z_{radiation} \approx Z_{headphone}$. A headphone that meets the FEC criteria is objectively quantified by measuring the Pressure Division Ratio (PDR) to be unity. However a tolerance and upper-frequency limit is often imposed. PDR measurement data are seldom reported but can be found for a small selection of headphones (Møller et al., 1995b; Masiero and Fels, 2011; Paquier and Koehl, 2015). The main problem with PDR measurements is the requirement of resources due to needing four different measurements. The PDR was not measured in this experiment, but due to the design of the electrostatic STAX SR-307 headphones used in this project, it is likely they would be at the most favourable end of the 'open' headphones that could have been used for testing.

4.4 Headphone Equalisation

Following the measurements and description above, an approximation of P_{EEP}^{open} has been described which will change dynamically based on the listener's head movements. A method to recreate $P_{EEP}^{blocked}$ is now needed by way of headphones. To achieve this, the transfer function from headphone input terminals $E_{headphones}$ to $P_{EEP}^{blocked}$ must be accounted for so that effects such as transducer response and pinna reflections are not included (generic pinna reflections are already included in the BRIR measurements). The transfer function from $E_{headphones}$ to $P_{EEP}^{blocked}$ is named the headphone transfer function $HpTF$. This function must be measured and approximately inverted to mitigate the effect where the inverse is named the headphone equalisation $HpEQ$.

To define $HpEQ$ filters, measurements of HpTF were made at the University of Salford anechoic chamber on the same B&K HATS used in SBSBRIR measurements. Type 4190 microphones were positioned to simulate $P_{EEP}^{blocked}$. 10 measurements of $HpTF$ were made for each ear individually and headphones were removed and repositioned visually between measurements. Photographs of each positioning was made to allow for post-measurement analysis. Figure 4.5 shows the mean HpTF measurements and calculated HPEQ for each ear.

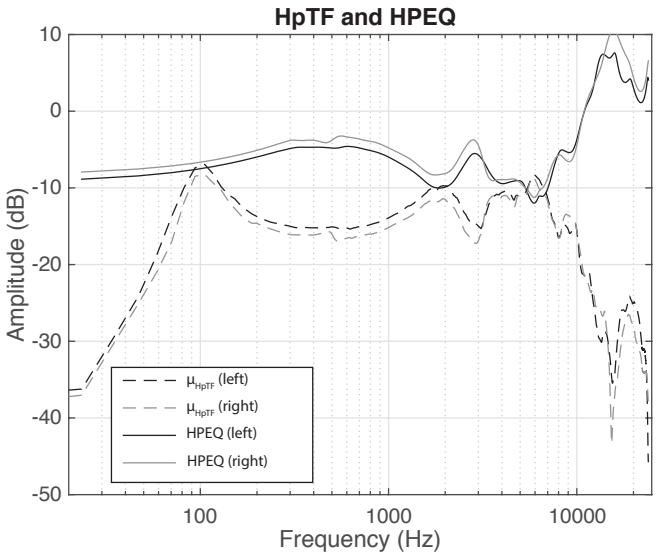


Figure 4.5: μ_{HpTF} and calculated HPEQ for each ear. μ_{HpTF} is the log-scale mean of 10x measurements and used to calculated $HpEQ$ filters independently for left and right ears. Smoothing was applied at upper and lower frequency band-extremes to avoid large gains in the $HpEQ$.

4.5 Rendering System and Signal Processing

A modified version of the the SoundScape Renderer (Geier et al., 2008) was implemented and verified in cooperation with BBC R&D. This system builds upon the low-latency and highly configurable SSR platform which performs block-wise FFT convolution for dynamic binaural synthesis. BBC R&D developments were made to also include variable early-to-late binaural mixing times which help to improve the efficiency of the renderer.

4.5.1 BRIR Perceptual Mixing Time

Because the SBSBRIR dataset contains long, independent BRIRs for each of the measured head-azimuths, realtime rendering of many loudspeakers and listening

positions can become expensive. Lindau et al. (2012) has shown that after a specified time following the BRIR onset, a cross-fade to a single BRIR tail can be perceptually equivalent and has large benefits for computation and computer memory usage. An example BRIR for the left ear, head-azimuth 90°, loudspeaker 0° and listening position (X = 0.5, Y = 0.5m) is shown in Figure 4.6 with the half cosine windows illustrated and the early and late regions plotted with different line styles. The amplitude shows $20log_{10}|BRIR|$. When the listener moves their head, only the initial region of the BRIRs change dynamically whereas the late part is loaded into memory once and reused. The late region is still maintained as a binaural signal and therefore has natural decorrelation between left and right ears. It can be seen that the direct region of the BRIR and early reflections are maintained dynamically but as the reverberation becomes more diffuse, less accuracy is needed.

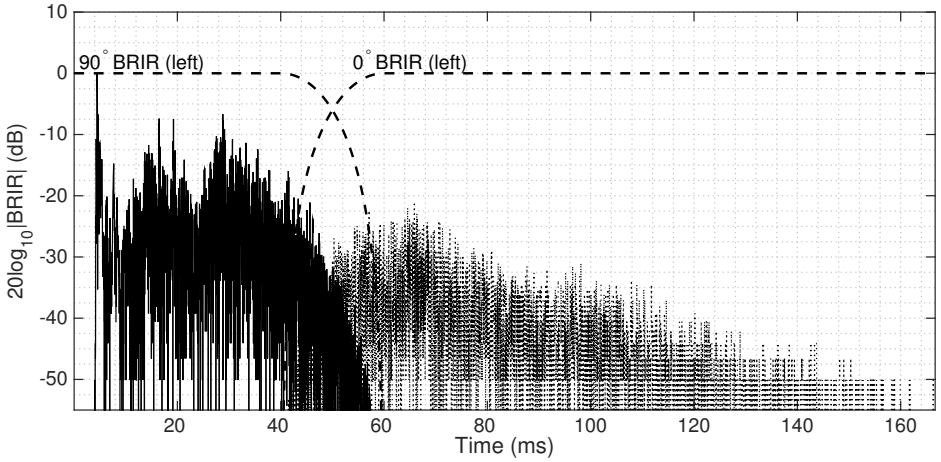
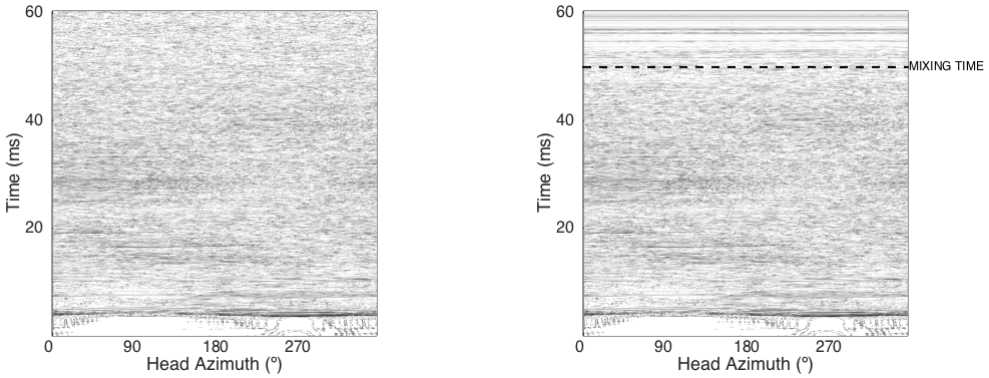


Figure 4.6: Example of BRIR split into static and dynamic regions. This is an example BRIR for the left ear where the initial region is dynamic and shows the BRIR where $\theta = 90°$. The static region is shown in the dashed line and is taken from the $\theta = 0°$ head azimuth. The y-axis is magnitude on a log scale to emphasise the early reflections in the dynamic region. Mixing time is 50 ms after the BRIR start.

The effect is more easily described by plotting resultant summated BRIR regions across both time and head-azimuth changes. Figure 4.7a and b show the functions

without and with mixing respectively.



(a) No mixing (b) 50 ms mixing time

Figure 4.7: $20log_{10}|BRIR|$ plotted against time and head-azimuth with and without dynamic mixing. (a) shows the raw BRIRs without any dynamic mixing and (b) shows the same BRIRs but with static tails from 0° BRIR used after 50 ms cross-over window.

A perceptual mixing time of 50 ms is supported by results of mixing times for small rooms and relevant model-based predictors in work by Lindau et al. (2012). This is due to smaller rooms having less distance between discreet reflections and therefore impulse responses take less time to achieve a 'diffuse' state.

4.5.2 Approximating Anechoic Simulations

For certain experiments within the project it was necessary to have anechoic versions of the SBSBRIR dataset i.e the same artificial head, loudspeakers and listening positions located in an anechoic environment. This would allow for the investigation of the importance of room reverberation on perception across the listening area and also serve as a learning dataset for localisation modelling presented in Chapter 7. A method to achieve anechoic versions of the SBSBRIR dataset was implemented which still maintained artefacts caused by changing

listening position. Two-band windowing was performed using onset detection to isolate only the direct part of the BRIR. This region includes the head, pinna and torso reflections from the full BRIR as well as maintaining loudspeaker effects. Truncating early regions of BRIRs have been applied for situations of comparing real and synthesised reverberant tails (Menzer and Faller, 2009). Frequency-dependent windowing has been applied to impulse responses of acoustics systems (Karjalainen and Pautero, 2001). A second order Linkwitz-Riley filter (LR2, LR-2), as shown in Appendix. A, was implemented to separate high- and low-frequency regions on an input BRIR. Each frequency band was then windowed independently, using a longer window for low-frequency components to avoid truncating the loudspeaker low-frequency response. The filter had a cross-over frequency of 400 Hz. Windowing was performed using onset detection to ensure that interaural and inter channel delays were not affected by the windowing. Following the detected direct-path onset time, windowing started after 45 samples[3] for the high-frequency region and 230 samples for the low-frequency region. The onset of the first room reflections varied slightly depending on the loudspeaker and listening position but visual inspection of the impulse response indicated they started around 150 samples after the onset of the direct path. The window lengths were optimised by iteratively lengthening the times for each frequency band until the optimal trade-off was found between maximising the length of the direct path and attenuating the first reflection. Figure 4.8 shows the magnitude response of an original BRIR against the anechoic version.

[3] sample-rate=48000 kHz.

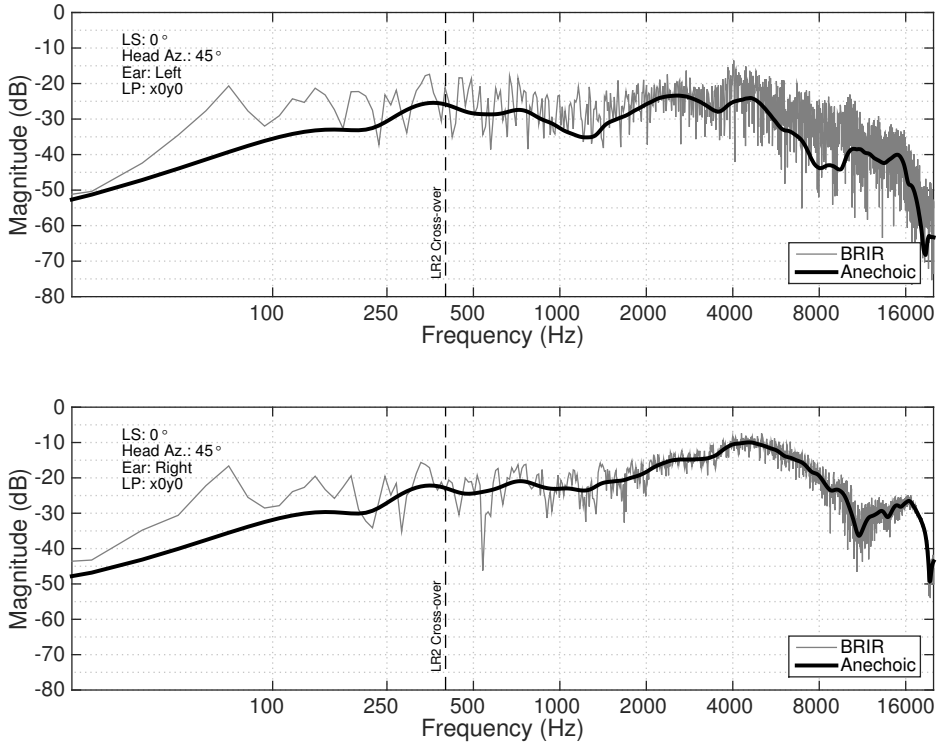


Figure 4.8: Power spectrum analysis of the reverberant and approximated anechoic BRIRs using dual-band windowing. Results are for L and R ears for the BRIR head azimuth of 45°, loudspeaker 0° at the central listening position.

4.6 Verifying Total System Latency

Due to the dynamic nature of the AVE it must be verified and quantified that the latency between a listener's head rotation and relative changes in the headphone signals are below the perceptual threshold. Previous work by Lindau (2009) has considered the latency requirements of dynamic binaural synthesis systems finding that out of 22 subjective responses (using 2 stimuli and 2 test environments) a minimum value of 53 ms was detected by one individual whilst the mean detectable latency was 107.63 ms. Only 3 values (3.4%) of all subjective latency threshold responses were below 64 ms. Yairi et al. (2006) has also undertaken research into detection thresholds. Their discussions highlight

that there are large subjective differences in the detection threshold and limen values and therefore the system latency. Therefore, averaging their results of detection threshold and detection limen with 2 previous studies of detection threshold values, they found the detection threshold for virtual auditory displays to be 75 ms.

From this it is justified to say that a measured total system latency (TSL) below 53ms would certainly be below perceptual threshold and values between 53 ms and 75 ms would also be acceptable.

To measure TSL the following method was undertaken:

1. Setup AVE renderer with HRIR database replacing 0° azimuth filter with zeros
2. Attach an accelerometer to the headphone set being tracked
3. Pass a continuous noise signal into the input of the renderer (when head azimuth is at 0°, headphones will output silence, movement of headphones to any other angle will trigger noise output)
4. Mount a microphone in audible range of headphones
5. Record outputs of accelerometer and headphone microphone simultaneously into a digital audio workstation
6. Align headphones to 0° and induce an impulse to the headphones which will give accelerometer impulse to indicate movement and trigger noise output due to rendering angle change
7. Repeat

Following this method, the TSL can be calculated by measuring the time between accelerometer onset and noise signal onset at the microphone output.

This value method was repeated 13 times to get a sample of TSL readings.

Table 4.2: Measurement statistics for total system latency (TSL)

Statistic	Value
Number of measurements	13
Mean TSL (ms)	49
Std. Dev. TSL (ms)	9

It is important to note that variability in TSL will be affected by changes in the audio rendering buffer size, BRIR onset length, audio hardware latency and tracker update rate and latency. Due to tracking data being sent over TCP/IP, network latency is also a factor but the effect is likely to be small (<1 ms).
These measurements verified that the system was able to dynamically change headphone signals with a latency lower than previously defined perceptual thresholds.

4.7 Conclusions

This chapter has presented a detailed description of the non-individualised, dynamic binaural simulation system used throughout the project. For the sake of consistency and repeatability it is important to define the many variables available for such a system and discuss the causes and implications of limitations of the system.

Firstly, the SBSBRIR dataset was presented which stands as the first publicly available spatially-sampled BRIR dataset designed for the evaluation of loudspeaker-based spatial audio systems at multiple listening positions. It has

been shown that the SBSBRIR dataset covers 15 positions across the listening area and simulates the main physical artefacts of off-centre listening. Acoustical considerations for BRIR microphone positioning and headphone reproduction were then defined where it was shown that when including headphone compensation filters to BRIR measurements made at the entrance to the blocked ear-canal, the acoustical transmission path from a sound source, to the ear-drum of a listener can be approximated. Perceptual mixing time values for dynamic binaural simulation were introduced from previous research. A perceptual mixing time for the dynamic binaural simulation system was chosen, allowing for a reduction in the amount of computational memory required. The use of dual-band windowing was then introduced and shown that anechoic simulations of loudspeaker layouts can be achieved from the SBSBRIR dataset whilst still maintaining inter-aural and inter-channels delays. Specific windowing values were also defined. It was finally shown using total system latency measurements that the latency between head-movement and changes in audio filtering were below previously reported perceptual thresholds.

CHAPTER 5

Headphone Transparency to External Loudspeaker Sources

This chapter presents experiments conducted into the passive effect of headphones on the transmission of sound from an external loudspeaker to a listener. Both physical measurements and a behaviour study was conducted to further understand the implications for binaural validation tests.

5.1 Introduction

An earlier version of this work was presented at 135th Audio Engineering Society, New York 2013 (Satongar et al., 2013) and further developed for publish in the Journal of the Audio Engineering Society (Satongar et al., 2015). Work in this chapter represents a 70:30 contribution between Darius Satongar and the BBC. The calculation of psychoacoustic metrics (ITD, ILD and ERB-smoothed transfer functions) was contributed by Chris Pike at BBC Research and Development.

The use of binaural rendering is popular in a number of audio applications; from hearing research (Minnaar, 2010; Zhang and Hartmann, 2010; Ericson and McLinley, 2001) to entertainment (Staff Technical Writer, 2006; Linkwitz, 2003). In each application, the specific requirements for the performance of a binaural system will be slightly different although generally, the aim is to induce the perception of intended auditory events as accurately as possible. Designing an assessment methodology that validates a binaural system within its intended application is often a difficult task. A common metric for a binaural system is the ability to produce a virtual sound source that is indistinguishable from a real sound source. Indirect comparisons have been investigated for example by Minnaar et al. (2001) and Møller et al. (1996, 1999) in which non-dynamic binaural simulation and real loudspeaker localisation tasks were considered in separated experiments. However, for direct comparisons where real and virtual loudspeakers are presented simultaneously, the validation of headphone-based binaural systems against a real loudspeaker reference can be problematic. The listener must wear the headphones throughout the experiment, which will affect the sound transmission from the external loudspeakers. A number of discrimination studies have involved direct comparison of real sources with

headphone-delivered virtual sources (Zahorik et al., 1995; Hartmann and Wittenberg, 1996; Langendijk and Bronkhorst, 2000; Lindau and Weinzierl, 2012; Fels et al., 2013) as well as some recent localisation tests (Wierstorf et al., 2012a,b) and loudness equalisation studies (Völk, 2013; Volk and Fastl, 2013). The passive use of headphones may have a significant effect on the perception of the external loudspeaker and therefore cause an unknown and possibly directionally dependent bias. Hartmann and Wittenberg (1996) noted that wearing headphones appeared to affect the listeners' ability to distinguish between front and back, although they also state that they were not aware of its effect on experiments in the azimuthal plane. To highlight the importance of the problem, Erbes et al. (2012) presented work on the development of an advanced headphone system specifically for the field of binaural reproduction.

This chapter investigates whether headphones mounted on a listener are likely to have an effect on the perception of external sound sources in the horizontal plane. The perceptual effect of the distortion in sound transmission from external loudspeakers, passively caused by headphones, is studied in two ways: (1) consideration of the physical differences in HRTFs measured with and without headphones and the implications on interaural cues and (2) a localisation test quantifying the passive effect of STAX SR-202 headphones on the localisation of external loudspeakers. Blauert (2001) states that the localisation of a sound event incorporates both direction and distance attributes. The term 'localisation' used in this paper refers only to the direction-of-arrival aspect.

There are a number of possible approaches to compensate for the effect of headphones on the perception of external sound sources. Moore et al. (2007) investigated the compensation of headphone transmission effects using the

headphones directly, where compensation filters were derived from HRTF measurements with and without headphones coupled. Their results highlighted attenuation at frequencies above 1kHz. The authors highlighted that at frequencies above 1kHz, headphones produced signals that were of the same order of magnitude as the loudspeaker source. Another possibility is to fit earphones with outward facing microphones to create a pseudoacoustic approximation of the external sounds as demonstrated by Harma et al. (2004). By filtering the signal received by the microphones to compensate for the earphone response and minimising leakage through the headset design and listening level, the system is a realistic possibility. Virtual sources are then synthesised using transfer functions also measured at the microphones on the binaural headset. Here both the 'real' and 'virtual' signals are approximations of the real loudspeaker sound at the ear canal entrance, since they are measured at a point outside the ear canal where some source direction dependence still exists (Møller, 1992). The pseudoacoustic loudspeaker sources also contain other errors, such as leakage of the external signal through the earphones, which varies individually due to earphone fitting, a delay introduced by filtering in comparison to the leaked signal and alteration of the pressure division at the entrance to the ear canal.

Making HRTF measurements with headphones worn would mean the transmission from both real and simulated loudspeakers is affected by the passive filtering effect of the headphones but would allow for direct comparison between the two systems. This approach was implemented by Völk (2013); Völk and Fastl (2011) and later studies (Wierstorf et al., 2012b; Fels et al., 2013) for both a dummy head and real listeners.

If the headphones do not have a perceptually significant effect on transmission from external sound sources to the ear then no additional processing is required

to compensate for the presence of the headphones. This is dependent upon the physical headphone construction. Previous studies have used this approach; Zahorik et al. (1995) state that the supra-aural headphones used in their study were chosen for 'minimal acoustic obstruction', while Lindau and Weinzierl (2012) state that their chosen circum-aural electrostatic headphones were 'relatively acoustically transparent'. However no verification of these statements is provided in those studies. Langendijk and Bronkhorst (2000) did provide physical measurements of the headphone effect and analysis in terms of interaural level and phase differences and time of arrival, showing minimal effects but in this test earphones were only suspended close to the pinnae and not directly coupled.

Regardless of whether the effect of headphones is perceptible, it is valuable to measure the effect that they have, so an informed decision can be made about methodologies for direct comparison of real and virtual sound sources.

5.2 Physical Measurements

To explore the perceptual significance of headphones on the distortion of transmission from external speakers to the ear, measurements were made on a number of available headphone sets. The measurements were taken to give an indication of the filtering effect the headphones had on the transmission from external sound sources. Similar perceptually motivated transfer function analysis has also been undertaken for head-related impulse response measurements (Völk et al., 2009). A range of headphones were chosen which are commonly used in binaural experiments as well as attempting to show a range of different models. The Sony MDR-V500 model was chosen as the only closed-back headphone to

give a 'worst-case scenario'. Table. 5.1 lists the headphone sets measured. The terminology 'open/closed' in Table. 5.1 refers to the manufacturer's design specification usually meaning that sounds from the outside can be heard when wearing the headphones as opposed to any measured objective criteria (Møller, 1992).

Table 5.1: Description of the headphones under test for physical measurements

Headphone Model	**Ear Coupling**	**Transducer**	**Open/Closed**
Sony MDR-V500	Supra-aural	Dynamic	Closed
Sennheiser HD650	Circum-aural	Dynamic	Open
AKG K601	Circum-aural	Dynamic	Open
Sennheiser HD800	Circum-aural	Dynamic	Open
STAX SR-202	Circum-aural	Electrostatic	Open

5.2.1 Method

Measurements were made in the semi-anechoic chamber in the University of Salford Acoustic Research Centre. This has a hard floor surface and acoustically absorbent walls and ceiling. The chamber has a working volume of 4.2 x 3.3 x 3.0 m and background noise level of 3.8 dBA[1]. Transfer function measurements were made using the exponential swept-sinusoid method. The B&K Head and Torso Simulator (HATS) Type 4100 was fitted with calibrated measurement microphones positioned at the entrance to the ear canal position therefore simulating measurement at the entrance to a blocked ear canal. The HATS was mounted on a hand-operated rotating turntable. A Genelec 8030A loudspeaker was used, mounted at ear height to the dummy head at a distance of 1.4m. It is assumed that a rotation of the HATS is equivalent to a rotation of the external source around the head in this environment. Measurements were made at both

[1]http://www.acoustics.salford.ac.uk/facilities/?content=semianechoic

ears at 15° increments in azimuth rotation from 0° to 180° for each headphone set and for a reference measurement without headphones. All measurements were made for a single headphone set before changing headphones and each set was positioned symmetrically by eye. The HATS has left/right head and torso symmetry so head rotations between 180° and 360° were not measured. Where data is presented for a single ear in this paper it is shown for the left ear and the contralateral data is actually measured on the right ear. In this paper an azimuth of 0° corresponds to directly in front of the head and positive rotation of the head is clockwise.

For each rotation angle and dummy-head ear, the transmission between the loudspeaker input and microphones at the blocked ear canal entrance point was measured for the two scenarios of (1) free-air and (2) headphones coupled. Both measurements contain electroacoustic transmission effects. Measurements were firstly converted to the complex frequency domain using a Fourier transform. The transfer function between measurements with and without headphones coupled will therefore show the effect of headphones on the blocked ear canal pressure as shown in equation (1).

$$H_{effect}(\omega, \theta) = \frac{P^{hp}_{blocked}/E_{loudspeaker}}{P_{blocked}/E_{loudspeaker}} \tag{5.1}$$

$H_{effect}(\omega, \theta)$ is the transfer function between pressures at the blocked ear canal with and without headphones and highlights the filtering effect of the headphones on the dummy head. $P^{hp}_{blocked}$ is the pressure at the entrance to the blocked ear canal with headphones mounted, $P_{blocked}$ is the pressure at the entrance to the blocked ear canal without headphones mounted and $E_{loudspeaker}$ is the input voltage at the loudspeaker terminals. Figure. 5.1 shows the measurement setup for all configurations.

Figure 5.1: Headphone sets mounted on B&K HATS for physical measurement stage of the investigation.

Perceptually motivated magnitude smoothing was applied by considering the auditory filter shapes and spacing (Glasberg and Moore, 1990). This was implemented using the Auditory Modelling Toolbox (Søndergaard et al., 2011) with a filter spacing of 0.1 ERBs. Each filterbank was applied to the inverse Fourier transforms of $P^{hp}_{blocked}/E_{loudspeaker}$ and $P_{blocked}/E_{loudspeaker}$ independently and for each ear. Taking the time-domain RMS value for each output meant the perceptually smoothed effect of the headphones, $|H^{ERB}_{effect}(k,\theta)|$, could be calculated by taking the difference in log power spectrum between the two cases of with and without headphones mounted. Note the change in notation from ω to k, where k represents the auditory filter centre frequency.

5.2.2 Results

Figure. 5.2 shows the spectral error across azimuth for each headphone.

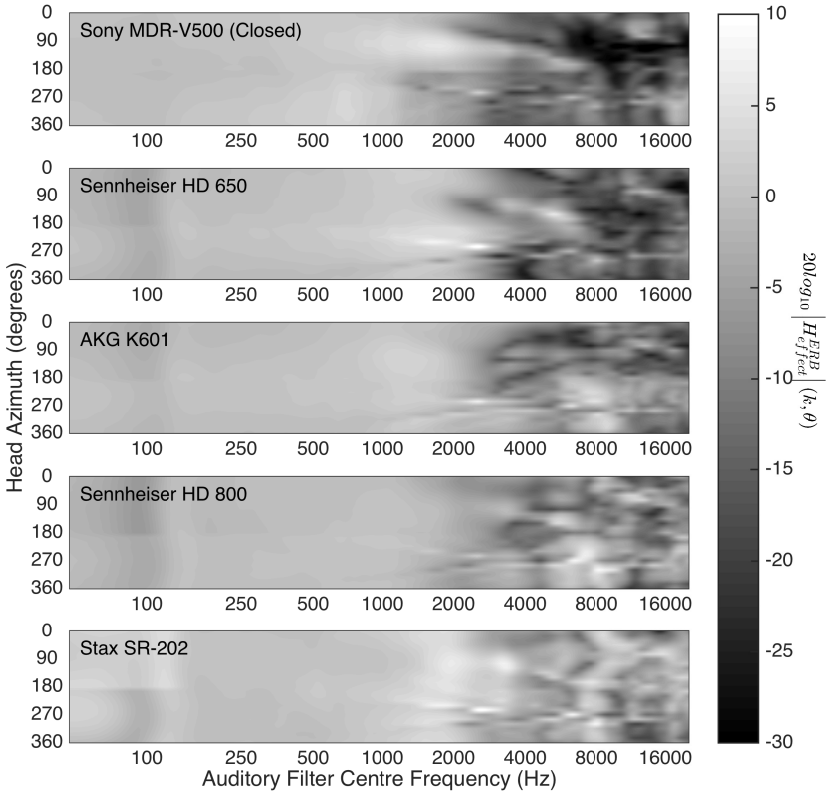


Figure 5.2: $20log_{10}|H^{ERB}_{effect}(k,\theta)|$ for each headphone set, at all angles. This ratio of transfer functions after perceptually motivated frequency smoothing demonstrates the filtering effect to external sound sources when listening with headphones coupled to the ears of a B&K HATS.

To achieve more insight into how headphones might affect localisation acuity of external sound sources, particularly in the horizontal plane, the interaural time and level differences (ITD and ILD) were approximated. The energy ratios for corresponding left and right auditory filter outputs were used to calculate the ILD in each frequency analysis band and source azimuth.

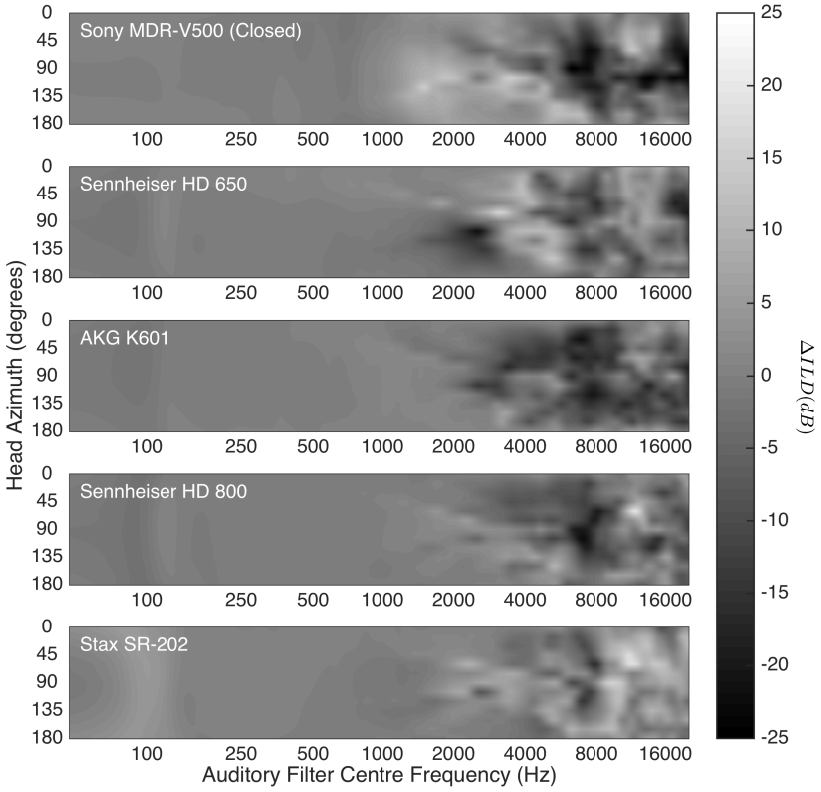


Figure 5.3: ILD error for all measured headphones for all measured head-azimuths. High frequency spectral differences show the changes in ILDs of external sound sources when headphones are mounted for a B&K HATS.

The difference from the case with no headphones was taken for each headphone to obtain the ΔILD error plots shown in Figure. 5.3.

Broadband ITD was calculated from the impulse responses using the minimum-phase cross-correlation method (Nam et al., 2008) and is plotted for each headphone in Figure. 5.4 alongside that of the reference measurements. This method, like others, generates some outliers at around 100° to 120° where the measured transfer function is not minimum-phase. Broadband ITD was used because it has been shown that we are not sensitive to frequency-dependence of interaural delay (Hartmann and Wittenberg, 1996).

Table 5.2: Root mean square, Standard Deviation and Maximum absolute values for ILD error (ΔILD) and broadband ITD error across all measured directions.

Headphone Model	**ITD Error (ms)**			**ILD Error (dB)**		
	RMS	**SD**	**MAX**	**RMS**	**SD**	**MAX**
Sony MDR-V500	0.081	0.084	0.23	6.83	2.66	26.52
Sennheiser HD650	0.033	0.024	0.08	5.04	1.64	21.40
AKG K601	0.045	0.036	0.10	6.10	1.62	21.86
Sennheiser HD800	0.044	0.045	0.15	4.57	1.41	22.13
STAX SR-202	0.059	0.040	0.15	3.87	0.97	18.50

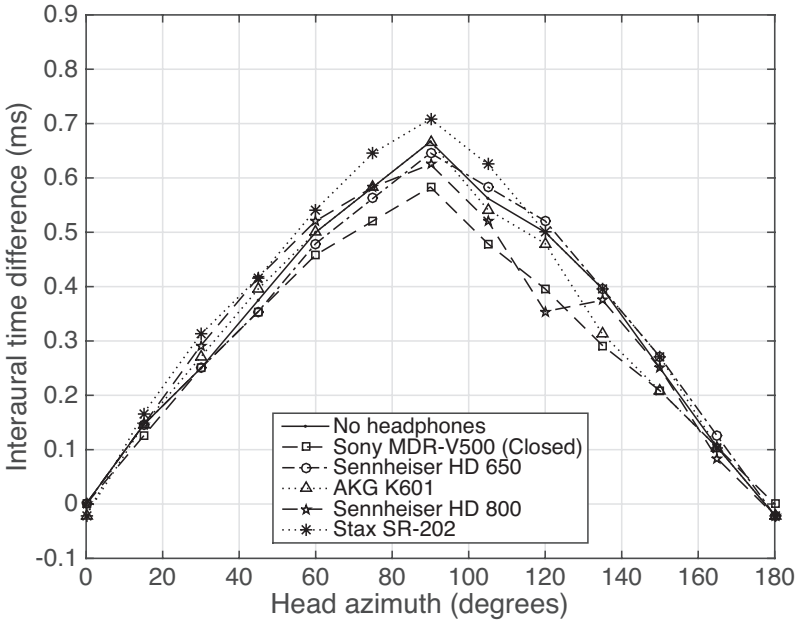


Figure 5.4: Broadband ITD with and without the measured headphones. The minimum-phase cross-correlation method was implemented on broadband impulse response measurements.

5.2.3 Effect of Repositioning

For the physical measurements presented above, no repositioning was performed. However as a post hoc study, the effect of repositioning was measured for the STAX SR-202 headphone set at 2 different angles 0° and 90°. The experimental setup was equivalent although post hoc measurements were made in the full anechoic chamber at the University of Salford Acoustic Research Centre. Statistical analysis was

performed to understand the significance of the different headphone-ear coupling in relation to the magnitude spectrum differences between the headphone sets measured. For each angle, the STAX SR-202 headphones were placed on the HATS and then completely removed and repositioned again before the next measurement. To consider the variance in $|H^{ERB}_{effect}(k,\theta)|$, the mean and standard deviation on the dB-scale magnitude responses was calculated for the output of each auditory filter band. Results are shown in Figure. 5.5.

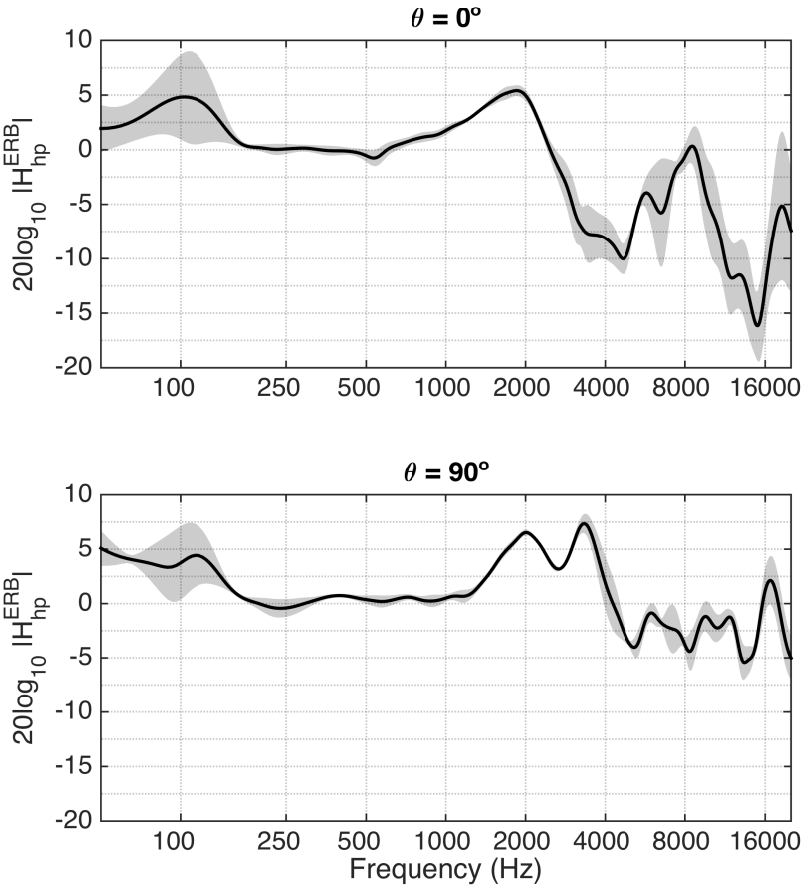


Figure 5.5: Variability in $|H^{ERB}_{effect}(k,\theta)|$ considered by 3 repeated measurements and 1 original measurement from initial experiment setup (4 total). Solid line represents mean and shaded regions represent 1 standard deviation both measured on a dB scale.

5.3 Behavioural Study - Localisation

The behavioural effect of the distortion to sound transmission from external loudspeaker sources, passively caused by headphones, was then investigated in a localisation test using the STAX SR-202. These headphones were chosen because they showed low errors in the physical measurements and have been used in previous comparison studies (Lindau and Weinzierl, 2012; Wierstorf et al., 2012b). The localisation test was performed both with and without the headphones, to see whether their presence had a significant effect on localisation acuity of an external loudspeaker source. Listeners were all recruited from the University of Salford.

5.3.1 Method

There have been a number of proposed methods for reporting perceived direction of a sound source in a localisation test, a summary can be found in Letowski and Letowski (2012). In this experiment, the egocentric method of head pointing was used by tracking the participants' head rotation in 6 degrees of freedom (DoF). This method is also comparable to the gun-pointing method used in Sandvad (1996) the difference being in the accuracy of the head opposed to hand for pointing. One disadvantage of this method is the possible disruption of natural listener behaviour due to the head being used to point. A Vicon optical tracking system (4x Bonita cameras, Tracker software) was used to track head motion, with passive markers that can be mounted unobtrusively. A number of trackers were piloted before the test and this system was found to be most accurate and reliable. Manufacturer reported tracking precision is 0.5° in rotation and 0.5mm in translation.

Two possible approaches when considering the localisation task are: (1) participant auditions a sound source of finite length, then subsequently points to the perceived direction, or (2) participant turns to face the direction of a continuous or repeating sound source. The first method is most common in localisation tests, assessing localisation using static cues at the tested directions. The latter method allows 'honing-in' on the source using dynamic localisation cue changes but the final judgement only highlights localisation error in the frontal region. The latter method was chosen to allow analysis of dynamic localisation processes and to minimise inaccuracies due to the reporting method, since minimum audible angles are smallest in the frontal region. Throughout this paper, a 'judgement period' refers to the period of time between the start of a sound event and the participant's decision on localisation direction.

The test was conducted in the BS.1116 compliant listening room at the University of Salford (ITU, 1997). Twelve loudspeakers were placed at randomly distributed angles around the listening area (59°, 105°, 118°, 126°, 158°, 188°, 211°, 245°, 273°, 294°, 312°, and 355°), at a distance of 2.1m from the centre and at ear height. The test was split into two sessions with an optional break: (1) localisation whilst wearing headphones (not connected to any sound source) and (2) localisation without headphones. The order of sessions was randomised in an attempt to normalise experimental bias. In each session the loudspeakers were selected in random order with 5 repeats, giving a total of 120 trials per session. A thin polyester curtain was positioned in front of the loudspeakers with a ≈2m radius to avoid visual biasing by the ability to see the loudspeaker. The participants were seated on a rotating chair, which could have an impact on the nature of movements but was not investigated in this study directly. Ten voluntary participants (3 inexperienced and 7 experienced in listening tests) were

used in the test. All participants reported normal hearing in a pre-test questionnaire but no audiometry tests were made.

Participants were asked to point their head towards the acoustic source and press a button to record their look direction. The next source then automatically started playing. A laser-pointing pen was mounted on the head to give a motor-visual indication as to the direction they were pointing. Participants were presented with repeating 500ms pink noise bursts with a rectangular window and 500ms silence between. The method focuses on frontal localisation acuity but the large number of source directions helped to reduce experimental bias due to e.g. room effects and increased the number of possible judgement patterns.

Participants performed a short initial training session to familiarise themselves with the method, in which they were asked to perform the localisation task for each of the twelve loudspeakers. No feedback on accuracy was given at any stage during the test. Figure. 5.6 shows an example participant conducting the test.

Figure 5.6: A picture demonstrating how listeners participated in the headphone transparency behavioural study. The response device can be seen on the participants lap. Reflective markers are visible mounted to the top of the headphones and one of the four tracking cameras can be seen in the background.

A calibration measurement preceded each session. The tracking system gave head position and orientation with 6 DoF relative to the room coordinate system with its origin at the centre of the loudspeaker array. Headphone and headband tracking was calibrated within the tracking system and aligned to the room coordinate system. Prior to each session the participant was firstly asked to ensure the laser pen output matched their gaze by adjusting the headset on their head. They were then asked to point the laser pen to a black marker located on the speaker circumference at 0° and at speaker height. The tracked position and head rotation values were then recorded and used to determine the listener's head position from the tracker data throughout that session. Real-time tracking data was recorded throughout the experiment.

When the listener's head position moves from the origin the source angle with respect to the listener will change. Therefore before calculating localisation error the real loudspeaker angle was geometrically corrected for the listener's head

position at the time of reporting the perceived angle. The standard deviation in head translation from the origin across all listeners and trials was 8.97cm. This meant that when processing the data, localisation error could be more accurately represented. It also meant that participants were given freedom of movement throughout the test.

5.3.2 Results

The most obvious is to analyse the absolute localisation error results but we also focus on the data captured during the decision making process. Since the chosen pointing method focuses on frontal localisation error, the movement profile during the decision making process is analysed in order to gain further insight.

Localisation Error

Localisation error was calculated by taking the angular difference between the translation-corrected real source directions and the calibrated reported source directions. However, results highlighted that when looking at the signed error distributions for each session, the arithmetic means or constant errors (CE or accuracy) (Letowski and Letowski, 2012) were not equal to zero. Figure. 5.7 shows the mean signed localisation error for each session with 95% confidence intervals.

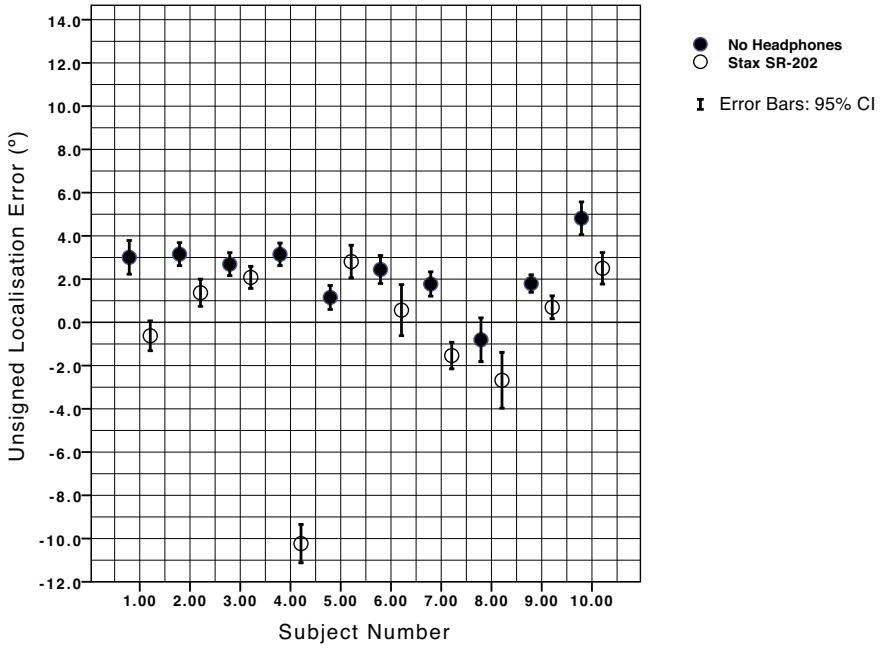


Figure 5.7: Signed localisation error without mean-correction for each subject. Filled markers represent the case of no headphones, hollow markers represented the case of listening with STAX coupled to ears. Error bars represent 95% confidence intervals.

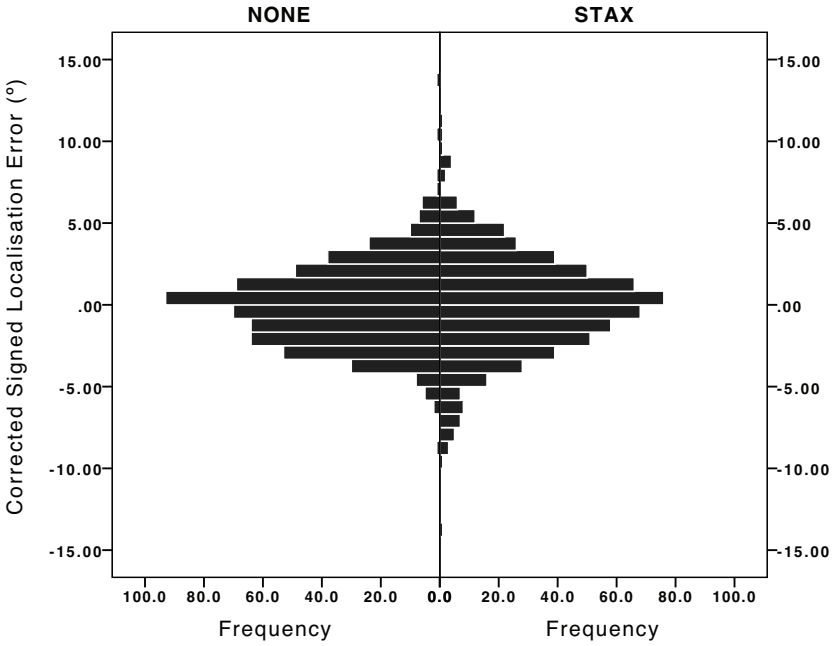


Figure 5.8: Mean-corrected signed localisation error histogram for (left) no headphones and (right) with headphones.

Letowski and Letowski (2012) explain that a non-zero mean signed localisation error could be due to a number of possible factors such as lack of symmetry in listener hearing or listening conditions (which could have been emphasised by the use of a reverberant room). In an attempt to separate any external factors influencing the relevant results, Letowski and Letowski (2012) also highlight that overall localisation error (LE) can be split into two separately identifiable statistics: accuracy (constant error, systematic error, validity, bias), and precision (random error, repeatability, reliability, reproducibility, blur). Due to uncontrollable parameters, which may affect the mean signed localisation error, it seems more experimentally justified to focus statistical analysis of localisation on precision to ensure separation from any external effects on CE. The method of 'mean correction' is also discussed by Letowski and Letowski. Signed error distribution means for each subject and session (STAX or NONE) can be seen in Figure. 5.7, these mean values were subtracted from the signed error samples for

each subject. The mean signed error before correction is also presented in Table. 5.3. Precision or random error (RE) is commonly identified by looking at the difference in distribution between the two cases (with or without headphones) with standard deviation and variance being popular metrics. Figure. 5.8 shows the mean-corrected distributions of all listeners for the two possible scenarios. It has been shown (Letowski and Letowski, 2012) that a reliable way of highlighting RE of localisation for normal distributions is to consider the signed standard deviation (SD) and mean unsigned error (MUE). The MUE (corrected) value is a compound statistic, which will highlight both RE and CE but due to the CE-correction applied here, values only show differences in RE. MUE (no correction) highlight changes in both RE and CE. Although standard deviation can be susceptible to the outliers usually recorded in real behavioural data, it gives a good overview of the comparison of distributions for the two cases. Results are shown in Table. 5.3.

Table 5.3: Localisation error and judgement statistics. SD is standard deviation and ToJ is the Time of Judgement.

Statistic	**NONE**	**STAX**
Uncorrected Mean Signed Error (°)	2.3	-0.5
Corrected SD Signed Error (°)	2.5	3.1
Mean ToJ (seconds)	3.2	3.4
SD ToJ (seconds)	1.3	1.4
Mean Turns (n)	1.3	1.4
SD Turns (n)	0.6	0.8

Time of Judgement

Due to the localisation task, any distortions introduced by the headphones at source angles other than close to 0° may not be directly apparent in localisation error, since the listener will arrive at a rotation with their head facing the source. However the effect of the headphones may change the process of forming the

judgement. Table. 5.3 shows the mean and standard deviation of the time-of-judgement (ToJ) values for the two cases.

Number of Head Movements

Another method of investigating the effect of the headphones is to consider the 'judgement profile'. Analysis of the participants' head-movements during their judgement period is made. This highlights the reliance on using dynamic cues when the participants were wearing headphones. Wallach (1940) describes the complex interaction between head movements and interaural cues. The number of times a participant changes their direction of head movement in each judgement can give another indication of the difficulty of localisation. If a participant is making lots of head turns, we can assume that they are using the interaction of movement and aural cues to improve localisation ability.

The number of head turns for each judgement was calculated using a Schmitt trigger on the angular head velocity with a threshold of 20°/s. Figure. 5.9 shows an example of a judgement profile with the relevant features highlighted. Similar analysis has been used for comparison of virtual/real sources in localisation tests by Wierstorf et al. (2012b). Table. 5.3 shows the mean and standard deviation for each headphone case.

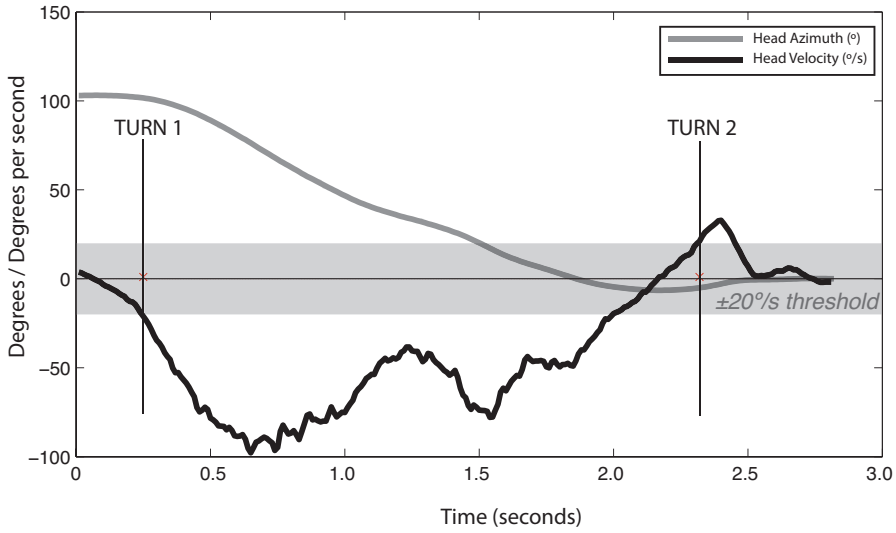


Figure 5.9: Calculation method for number of head turns per judgement. Grey area defines the threshold region of 20 degrees/sec. Vertical lines indicate a turn.

5.4 Discussion

Physical measurements showed that the headphones have a directionally dependent effect on the transmission from external sound sources above 1-2kHz, depending upon the model. Above 3kHz, errors are in the order of 10-20 dB, which is of the same order as variations across headphone-free measurements of 15° in azimuth separation. Most of the headphones cause a general attenuation at high frequencies, although sharp notches and peaks are present. The Sony headphones cause the greatest attenuation, due to their closed-back design. The STAX SR-202 cause the least attenuation overall. Several headphones show a prominent error peak at approximately 8kHz on the contralateral side, where there is a spectral notch in the $P_{blocked}/E_{loudspeaker}$ measurement.

The STAX headphones exhibit a consistent error peak at approximately 100Hz, which was approximately 5dB higher on the ipsilateral side. This could be caused by mechanical resonance of the drivers, which due to the electrostatic

design will be less damped than with other headphones. The other open-backed headphones show a small notch at roughly the same frequency so this could also highlight a specific interaction between the headphones and HATS coupling, which is exaggerated in the STAX measurements. The STAX headphones also showed another smaller peak at just below 2kHz.

For measurements of $|H^{ERB}_{effect}(k,\theta)|$ shown in Figure. 5.2, the observed errors are likely to have significant perceptual effects although further perceptual investigations are needed. The large abrupt changes to spectral level above 2kHz could cause audible colouration, as well as a possible dulling of the sound due to general high-frequency loss. It was found that localisation cues could be affected with similarly large sharp ILD errors above 2kHz. There is a large variation between headphone models in amount of error introduced. The spectral and ILD effects are less substantial for the Sennheiser HD800 and particularly the STAX SR-202, which is unsurprising due to their more open design. ITD is not as affected as ILD for the Sennheiser and AKG headphones, but the closed-back Sony headphones and the STAX cause a significant decrease and increase in ITDs respectively at lateral source positions. Inspection of the impulse responses showed that this increase in ITD for the STAX is mostly due to a delay of the contralateral time-of-arrival. For less open headphones the ipsilateral time-of-arrival is delayed at lateral source positions, causing a decrease in ITD. The STAX headphones show the lowest ILD error in terms of mean and maximum values, as shown in Table. 5.2. They also tend to increase ILD in contrast to the other tested headphones.

Figure. 5.5 highlights that there is a measurable effect of replacement on $|H^{ERB}_{effect}(k,\theta)|$, which is source direction dependent. This effect is larger for

regions of spectral peaks and notches highlighting that the repositioning of the headphones change the complex system of resonances caused by the headphone set and the pinna and ear-canal, this region is above 2-3 kHz which also corresponds to results of headphone transfer function variability with repositioning measured on human head (Völk, 2014). The Stax headphones chosen for the repositioning analysis have a large circum-aural design, which avoids deformation of the pinna which could improve robustness to repositioning. The effect of repositioning is small for the 200Hz-2kHz region with changes in the region of 1 dB. The 100Hz resonance found in the earlier physical measurements highlight increased variance, indicating the headphone-ear coupling as a variant factor of this parameter. Although not as dominant, similar increases at around 100 Hz can also been seen in results of headphone transfer function measurements with repositioning by Völk (2014) and also for Stax SR Lambda measurements specifically in measurements by Fels et al. (2013). Comparing against the magnitude headphone effect responses for different headphones, it seems that for the 90° angle (measurement ipsilateral to speaker), the variance was smaller than the difference between headphone models. At 0° measurement position, the variation in repositioning may cause the ranking of headphone models to overlap making the preference of headphones less defined.

Using a model of free-field sound transmission to the human external ear developed in Møller (1992), Møller et al. (1995a) presents results showing the influence of changes in radiation impedance when headphones are coupled to the ears of listeners. The term free-air-equivalent coupling (FEC) is presented (Møller et al., 1995b) to define a type of headphone set that does not disrupt the radiation impedance of ear canal looking outwards and therefore the ratio of pressure divisions between blocked and open ear canal pressures measured with

and without headphones coupled comes close to unity. A further developed selection criterion was later introduced by Völk (Völk, 2013, 2011, 2012b) which improves robustness of the criteria at high-frequencies. Although FEC is a separate consideration from the physical capsule design of the headphone, changes in the radiation impedance could additionally contribute to the effect of headphones on the perception of external sound sources. However, this effect will not depend on the direction of the sound source relative to the head.

For behavioural testing, it can be seen that the use of headphones did increase the RE of localisation error. However the increase was small: standard deviation by 0.6°. This magnitude of increase could be considered experimentally trivial when compared to the unimpaired human localisation ability. It should also be noted that the standard deviation measure will be affected a small amount by the tracking system error (manufacturer-reported as 0.5°). However, this error is likely to be balanced between the in situ and Stax measurements and therefore smaller than the 0.6° change in RE. On average the number of head turns made by participants when wearing the Stax was 0.1 more than when not wearing headphones (8% increase) and also the length of time taken to reach a judgement was 0.2 seconds longer (6% increase). This shows that normal localisation cues were disrupted and participants may have found it more difficult to arrive at the judgement direction. These dynamic cues, in addition to the small localisation precision error increase and large spectral changes highlight that care must be taken when implementing through-headphone listening test scenarios.

When localising sound events, anecdotal experience of the authors showed that head movements were often required to resolve front-back confusions and help to more accurately localise sound sources when wearing the Stax headphones.

Informal listening through the headphones also highlighted the spectral effects but showed that the Stax headphones had least noticeable distortion in line with physical measurements.

5.5 Conclusions

An assessment of the passive effect of headphones on the perception of external acoustic sources has been presented. Further analysis of physical measurements highlighted that headphones cause a measurable spectral error in HRTFs, with maximum spectral ILD distortion of 26.52dB for the close-back headphones (equivalent to a change in ILD corresponding to a large change in sound source direction). There was a difference between headphone sets with the closed-back headphones introducing the largest distortions overall and the STAX SR-202 electrostatic headphones introducing the smallest spectral distortions, although lateral ITDs were enlarged.

A behavioural test showed that wearing STAX SR-202 headphones reduced the precision of external loudspeaker source localisation, indicated by a 0.6° difference in the corrected standard deviation of signed localisation error. Further analysis of head movement to obtain judgement profiles showed that the participants on average took 0.2s longer to reach their final judgements (6% increase) and used 0.1 more head-turns (8% increase), which could imply an increase in complexity of the localisation process due to corrupted localisation cues.

In light of the findings in this study, it is recommended that care must be taken when choosing headphones for a scenario in which a listener is presented with

external acoustic sources. Results for different headphone designs highlight that the use of electrostatic transducers could help maintain natural acoustical perception. However, the effect on perception is still measurable and therefore headphone transparency should not be assumed. For an alternative solution it is recommended that headphones be worn during HRTF measurements to allow like-for-like comparison between the real and virtual sources, where in-situ HRTF measurement is possible (Völk, 2013; Völk and Fastl, 2011). As a result of the findings of this chapter, listening test evaluations between in situ and binaural simulation undertaken in this thesis are always conducted in separate trials and loudspeakers are never auditioned whilst wearing headphones.

CHAPTER 6

Simulating Localisation Artefacts Across the Listening Area Using Non-individualised Dynamic Binaural Synthesis

This chapter presents experiments on the ability to use a non-individualised, dynamic binaural simulation system to simulate localisation artefacts in loudspeaker-based panning systems at central and non-central listening positions.

6.1 Introduction

This chapter evaluates the use of an auditory virtual environment (AVE) created by a dynamic binaural synthesis system for the perceptual assessment of localisation artefacts commonly found in domestic loudspeaker-based panning systems. The experiment is designed to evaluate the degree to which localisation artefacts presented during off-centre listening (as caused by time-of-arrival changes from loudspeakers, loudspeaker directivity, relative loudspeaker positioning and room effects) are perceptually equivalent between in situ (using real loudspeakers) and a binaural simulation.

6.2 Method

A localisation test was undertaken in which participants were asked to indicate the direction-of-arrival of auditory events using two different presentation methods:

- In situ, where loudspeakers are used to create the auditory event
- AVE, where a non-individual dynamic binaural room impulse response convolution simulates loudspeakers in the listening environment using headphones as the sound events i.e. an AVE simulation of the in situ presentation method

Loudspeaker-based spatial audio reproduction methods were used to create virtual sound sources over a chosen loudspeaker layout using monophonic audio stimuli.

A selection of system parameter combinations, each defined by a combination number, were chosen to assess the performance of the AVE under a range of scenarios. These combinations were chosen to be representative of the most commonly used spatial audio reproduction systems. For each combination the panning method, panning direction, loudspeaker layout and stimuli were chosen as shown in Table. 6.2 and tested at both the central (X = 0m, Y = 0m) and one non-central (X = -0.5m, Y = -0.5m) listening position. This specific non-central listening position was chosen to represent a realistic, yet sub-optimal listening position in domestic listening conditions. This listening position also corresponds closely with one of the 'worst case listening positions' from Rec. ITU-R BS.1116-1 (ITU, 1997). Figure. 6.1 shows the geometry of the reproduction setup with all loudspeakers in place.

Alongside stimuli reproduced from a single loudspeaker (mono), Vector Base Amplitude Panning (VBAP) (Pulkki, 1997) and Ambisonics (Gerzon, 1972) were implemented using five different 2-D loudspeaker layouts. Ambisonic panning coefficients were calculated using a velocity decoder method (Gerzon, 1992a) as introduced in Chatper. 2. The Ambisonic decoding was implemented by taking Moore-Penrose pseudo-inverse of the re-encoding matrix C (as defined in equation. 2.9. 1st, 2nd and 3rd order Ambisonic systems were used. The loudspeakers indicated in Figure. 6.1 are at angles 0°, 30°, 45°, 90°, 110°, 135°, 180°, 225°, 250°, 270°, 315°, 330° with a 2.1 m radius. Subsets of these loudspeakers were chosen for each combination.

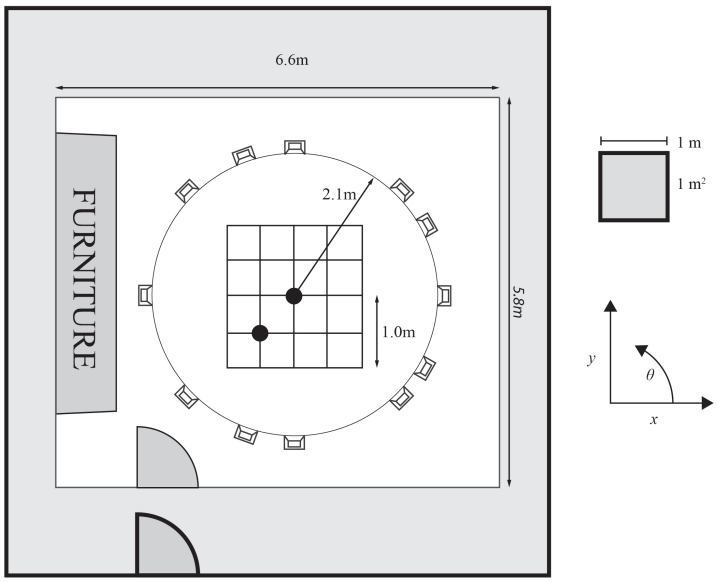


Figure 6.1: The geometry of the listening room (with interior and exterior walls), loudspeakers and listening positions used in the experiment. Localisation tests used the two listening positions highlighted with black markers. All loudspeakers are displayed where subsets make up the different loudspeaker layouts as shown in Table. 6.2. The front of the setup is shown to the right ($+x$) which corresponds to $\theta = 0°$ and positive rotation is counter clockwise.

The three types of audio sample used in the test are shown in Table. 6.1. The three types were chosen to represent the most commonly applied audio stimuli in perceptual evaluations. The music and speech excerpts were extracted from the EBU Sound Quality Assessment Material collection (EBU, 2008). All samples were repeated with a 500ms silence between repeats. This information is most critical for the noise stimulus as this induced a short, looping noise-burst stimulus.

Table 6.1: Description of the audio files used in the test.

Name	Description
Noise	500ms pink noise bursts with rectangular window
Music	Piano scale extract, 8s long
Speech	Female spoken voice, 28s long

The egocentric, head- or nose-pointing technique (Carlile et al., 1997; Letowski and Letowski, 2012) was chosen, where, upon hearing a continuously repeating

auditory event, participants turn to face in the perceived direction of the event whilst the stimulus continues to loop, pointing with a laser pointer attached to their head - a trigger button was then pressed by the participant to record their judgement and begin the next stimulus presentation. Makous and Middlebrooks (1990b) identify this method as a *closed-loop* localisation task where movements of the listener's head are inherently matched with changes in localisation cues. This can be compared to an open-loop task where audition is under a fixed listener position and the stimuli is stopped before reporting of the perceived direction. Practical implementations of the closed-loop localisation task can be found in (Bronkhorst, 1995; Carlile et al., 1997). The laser pointer was attached to the head-set and could be centrally aligned to their gaze to reduce over- or under-shooting the perceived direction (Lewald et al., 2000). This was done by mounting the laser on an adjustable ball-joint to reduce error caused by varied headphone positioning (Wierstorf et al., 2012a).

A potential disadvantage of the head-pointing method is that it predominantly measures changes in frontal localisation acuity. However the method allows for more accurate reporting of direction and the possibility to record head-movement data over time (Satongar et al., 2013). The benefits of this method are further highlighted by Carlile et al. (1997), stating it is a natural action and head-tracking can be performed feasibly. It has also been described that the importance of HRTF personalisation is greater in the median sagittal plane, influenced by the lack of interaural cues and therefore emphasis on spectral deviations caused by the pinna (Searle et al., 1975; Butler and Belendiuk, 1977; Wenzel et al., 1993). Therefore, this method is likely to further highlight realistic problems with non-individualised binaural simulation.

An optical motion tracking system (4x Vicon Bonita cameras and Tracker software) was used to track the participants' head position both for analysis and as input to the AVE rendering software. The tracking system is capable of 0.1° rotational and 1 mm translational accuracy and tracks with 6 degrees-of-freedom.

Participants were volunteers from the University of Salford Acoustics Research Centre (ARC). All participants reported their primary job as related to acoustics or audio and described themselves to be 'audio experts' when asked in a pre-test questionnaire. There were 15 participants in the test. No pre-screening audiometry tests were performed but all participants reported normal hearing and normal (corrected or uncorrected) sight.

Participants were given an instruction guide on the test procedure. They were then guided into the listening room in which loudspeakers were hidden behind an acoustically transparent curtain (curtain radius $\approx 2m$). Participants were given a MIDI controller with a button for submitting localisation decisions and a dial for audio volume control. They were allowed to adjust the volume at any point in the test. For each participant, a total of 120 judgements were given (3 repeats of 2 auralisation methods, 2 listening positions, 10 combinations of reproduction system parameters shown in Table. 6.2). The order of combinations with repeats was randomised for each participant but separate sessions were carried out for in situ and AVE and for each listening position. The test was preceded by a training session which consisted of a short trial test with a random selection of stimuli until participants reported they felt comfortable with the method. No feedback on localisation performance was given. Following from the conclusions in Chapter. 5, AVE and in situ testing was conducted in separate trials to avoid the audition of loudspeakers whilst a listener is wearing headphones.

Table 6.2: Detailed parameters of the combinations used in the localisation user study. No virtual source position exists for mono reproduction systems as the sound sources inherently comes from the specified loudspeaker - these directions are indicated with a *.

Combination #	Panning Method	Loudspeaker Layout (°)	Panning Direction (°)
1	VBAP	30, 330	15
2	Ambisonic 1st Order	0, 90, 180, 270	0
3	Mono	110	110*
4	Ambisonic 3rd Order	0, 45, 90, 135, 180, 225, 270, 315	45
5	VBAP	0, 45, 90, 135, 180, 225, 270, 315	100
6	Ambisonic 2nd Order	0, 90, 180, 270	115
7	Mono	315	315*
8	VBAP	0, 30, 110, 250, 330	290
9	VBAP	0, 30, 110, 250, 330	190
10	Ambisonic 1st Order	45, 135, 225, 315	0

A calibration stage was performed to align the coordinate systems of the motion-tracking sytem, the AVE renderer, and the physical room geometry. This was done by asking the participant to point their head-mounted laser at a marker placed at 0° (1.06 m from the floor) at which point the tracking data was recorded for post-hoc calibration of localisation judgements. In situ and AVE auralisation methods were specifically chosen to be tested separately to avoid any passive influence of the headphones on in situ loudspeaker reproduction if headphones were coupled to the listener during in situ reproduction (Satongar et al., 2013).

A note on the statistical analysis of circular data

When dealing with wrapped circular data, the normal arithmetic statical tools such as mean and standard deviation are often not valid. Consider a sample of localisation judgements

$$\alpha_{1,2,3,4} = [-189, -175, +178, +179] \tag{6.1}$$

All reported values in this hypothetical dataset point towards the rear of the coordinate system ($\alpha_n = \pm 180$. However, the arithmetic mean can be calculated using

$$\bar{\alpha} = \frac{1}{N} \sum_n \alpha_n \tag{6.2}$$

where $\bar{\alpha} = -1.75°$, a resultant angle in the opposite direction to any of the α_i values. This is only a problem for samples that span the wrapping part of the coordinate system ($\pm 180°$ in this example) but for samples with large variance, this is a crucial identification. To resolve this problem, the *mean angular direction* can be calculated (Fisher, 1995). Firstly angles are converted to cartesian form using complex notation

$$z_n = e^{i\alpha_n} \tag{6.3}$$

The angular mean is then calculated

$$\bar{z} = \frac{1}{N} \sum_{n=1}^{N} z_n \tag{6.4}$$

where the angle of the complex number $\bar{z}$ is the mean angular direction and the magnitude represents the spread of the circular sample. Implementations of

circular statistics for analysis in this chapter are from CircStat: A MATLAB Toolbox for Circular Statistics (Berens, 2009) who provides detailed descriptions of circular statistical analysis and their implementation. The *mean angular direction* for the example sample above now becomes $\mu = 178.25°$.

6.3 Methods of Analysis

The analysis of localisation error can be approached in a number of ways. Letowski and Letowski (2012) provide a detailed summary on the most commonly used statistics to analyse localisation judgements and localisation error. However, quantifying perceptual equivalence using statistical analyses is a non-trivial task. For this reason, the statistical analysis of the localisation judgments is separated into three sections. Firstly, signed localisation error is presented for each combination at each listening position independently. This provides a baseline for analysis using a simple metric. Following this, and to consider the data on-aggregate, a repeated-measures analysis of variance (ANOVA) is calculated on the mean unsigned error. Mean unsigned error provides an analysis on the localisation error where precision and accuracy are combined into a single, compound metric (Letowski and Letowski, 2012).

Statistical analysis of empirical data often aims to prove that the effect of two or more factors, when measuring a dependent variable, is statistically significant. Statistical methods such as t-tests and ANOVA help to inform the significance of the difference in the dependent variable. However, when attempting to show that two factors produce equivalent results, a failure to reject a null-hypothesis does *not* prove null hypothesis. In these situations, equivalence tests (Wellek, 2010)

are often implemented to understand the perceptual or practical equivalence of a factor. In the final stage of analysis, equivalence is measured using a two-one sided test methodology (TOST) to understand the perceptual equivalence of the binaural auralisation system to in situ auralisation. The motivation and calculation procedure for each method of analysis will now be presented separately.

6.3.1 Signed Localisation Error

To demonstrate the experimental data with a widely used statistic, firstly, the mean signed localisation error for each reported directional judgement is calculated. This is the angular difference between the intended panning direction, θ_{PAN} and the reported direction using the closed-loop task, θ_{REP}. For the in situ system, translational movements were accounted-for by a relative change in the panning direction due to the listeners' physical translation, measured using the tracking system. For each combination at each listening position, the angular mean value is calculated across participant and repeat factors. The standard deviation of this sample is also useful to understand the precision in localisation judgements.

6.3.2 Unsigned Localisation Error

To measure the global influence of the auralisation method on localisation error, a repeated measures ANOVA was conducted on the unsigned localisation error. Unsigned error was chosen in this analysis due to the ability to use linear algebraic tools which would not be valid for circular wrapped data such as signed

localisation error. Using a repeated measures analysis also accounts for inter-subject variability in localisation acuity and provides results of within-subject factor effects. The within-subject factors were (1) listening position, (2) auralisation method, (3) combination and (4) repeat. Table 6.3 presents the main results for the statistical analysis.

6.3.3 Equivalence Testing

To assess whether localisation error measured using the AVE can be considered perceptually equivalent to in situ listening, an equivalence test was performed. Although seldom used in the field of acoustics, the most common equivalence test is the two one-sided test (TOST) (Schuirmann, 1987) where a one-sided t-test is performed at each end of a difference sample. The difference sample, Δ Localisation Error (ΔLE) is firstly constructed in the following way:

1. Calculate localisation error as the angular difference between the intended panning direction, θ_{PAN} and the reported direction, θ_{REP}.
2. For each participant, combination and listening position independently, calculate all permutations of angular differences in localisation error between in situ and binaural simulation. Due to each participant making 3 judgements (repeats) there are 3^2 possible ΔLE values.
3. Aggregate ΔLE for each participant into a new sample for each combination at each listening position.

ΔLE represents a sample of differences in localisation error between in situ and the AVE at each combination and listening position. This calculation only considers within-subject differences in LE and therefore any inter-subject

variation in localisation acuity does not affect the data. ΔLE for each participant is then combined into a single composite sample and bootstrapping is used to provide a non-parametric estimate of the confidence intervals. The central tendency of this sample represents any systematic errors induced by the AVE.

Using a two one-sided t-test, equivalence is found at the α significance level if the $(1-2\alpha)\cdot 100\%$ confidence intervals for the difference sample, ΔLE in this case, are fully contained within the pre-defined equivalence boundaries (θ_1, θ_2). Note that 90% confidence intervals must be used to test equivalence at the $\alpha = 0.05$ significance level. Due to the localisation error also accounting for translational changes when listening in situ, ΔLE cannot be calculated using the difference between judgement directions alone. Using a bootstrapping method (N=1000) with replacement (Fisher, 1995), the non-parametric 90% confidence intervals of the variation in ΔLE circular mean are presented in Fig. 6.5 and Fig. 6.6. Results for the ΔLE samples for four combinations are also presented in Fig. 6.4 to demonstrate some examples of shape and parameters of the error samples using circular statistical analysis (Berens, 2009).

To quantify the perceptual equivalence, it is clear that (θ_1, θ_2) must be carefully defined. Due to the closed-loop localisation task chosen for this study, it would be logical to consider the minimum audible angle (MAA) in the frontal region, of which many results have been published (Mills, 1958; Makous and Middlebrooks, 1990a; Perrott and Saberi, 1990; Grantham et al., 2003) to show MAA in the range between 0.5° and 2.0°. The choice of equivalence boundary will depend on the accuracy needed by the AVE. For panning methods with accurate localisation ability, equivalence boundaries set at ±1° may be applicable. However, for panning methods which reduce localisation acuity and therefore

increase the variance in ΔLE, more realistic equivalence boundaries should be defined. MAA has been shown to increase as the sound event moves away from 0° in the horizontal plane, up to a value of $MAA \approx 6.5°$ when the sound event azimuth reaches 50° (Makous and Middlebrooks, 1990a). Therefore, considering the panning methods used in this test, an equivalence boundary of ±7° is an informed and realistic choice for evaluating the AVE. Regions of equivalence are highlighted on the graphs at $\pm 7°(\theta_1 = -7°, \theta_2 = +7°)$. Mean values for ΔLE are also presented on the graph.

6.4 Results

For each combination at each listening position, the angular directional mean of the signed localisation error sample is shown in Fig. 6.2. Data for in situ reproduction and simulation using the AVE is displayed on the same plot for analysis and error bars represent the standard deviation of the sample.

6.4.1 Signed Localisation Error

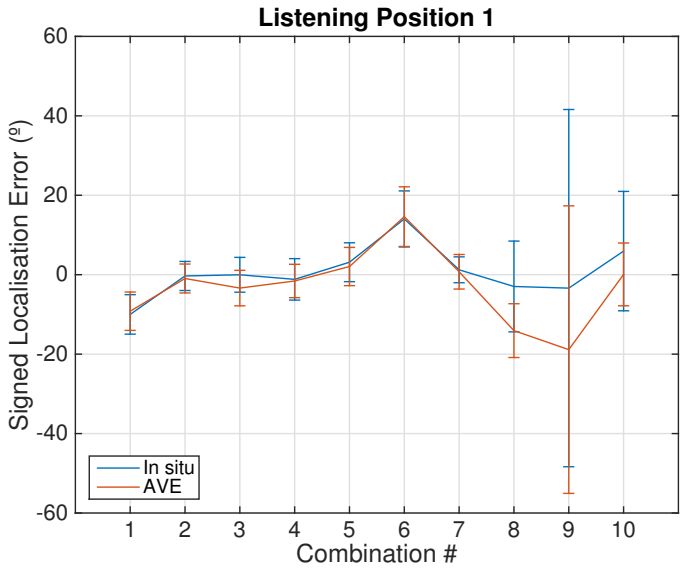


Figure 6.2: In situ and AVE simulated signed localisation error at the central listening position. Lines represent the mean signed localisation error and length of the error bars show one standard deviation of sample.

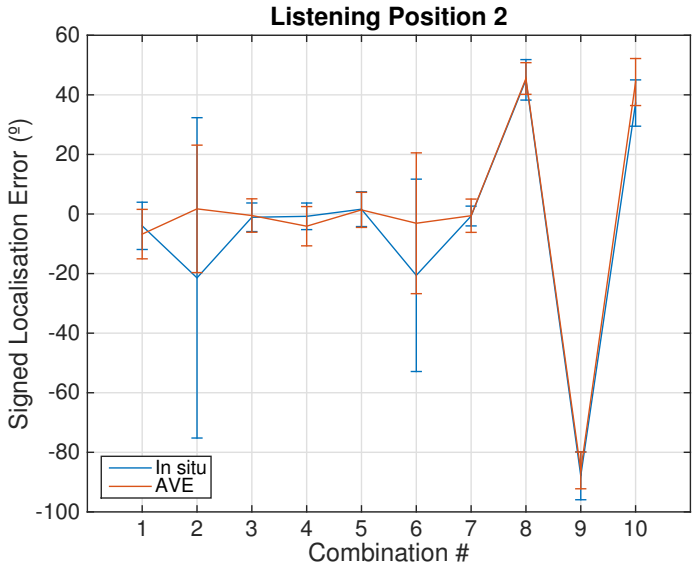


Figure 6.3: In situ and AVE simulated signed localisation error at the non-central listening position. Lines represent the mean signed localisation error and length of the error bars show one standard deviation of sample.

6.4.2 Unsigned Localisation Error

Table 6.3: Results from repeated-measures ANOVA for unsigned localisation error.

Effect	Type III sum of squares	DF	Mean Square	F	sig.	η_p^2	η^2
Listening Position	101890.008	1	101890.008	145.692	0.000	0.912	0.0928
Combination	537221.695	9	59691.299	126.321	0.000	0.900	0.4893
Auralisation Method	1642.791	1	1642.791	3.355	0.088	0.193	0.0015
Repeat	287.146	2	143.573	2.444	0.105	0.149	0.0003

6.4.3 Equivalence Testing

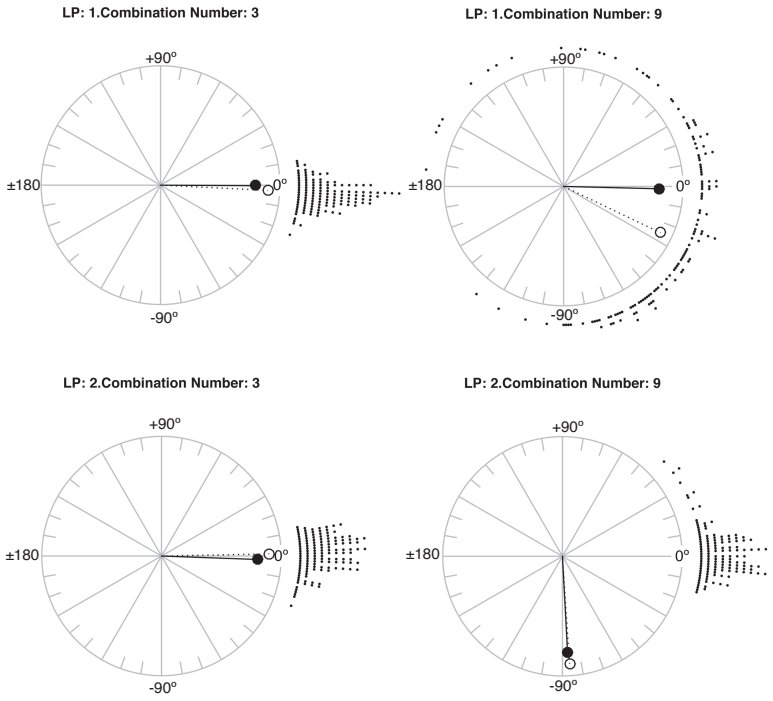


Figure 6.4: Examples of circular ΔLE histograms for combinations 3 (left) and 9 (right) for both listening position X=0 m, Y=0 m (top) and X=-0.5 m, Y=-0.5 m (bottom). The exterior markers indicate the counts of ΔLE (between in situ and AVE) with 1° resolution. Interior circular markers represent mean signed error where filled markers represent in situ listening and unfilled markers represent AVE listening. 0° is to the right of the circle and positive direction is counter clockwise. LP:2, Combination Number: 9 is specifically interesting as the mean signed localisation is high (-90°) but the error between in situ and AVE is centered around 0°.

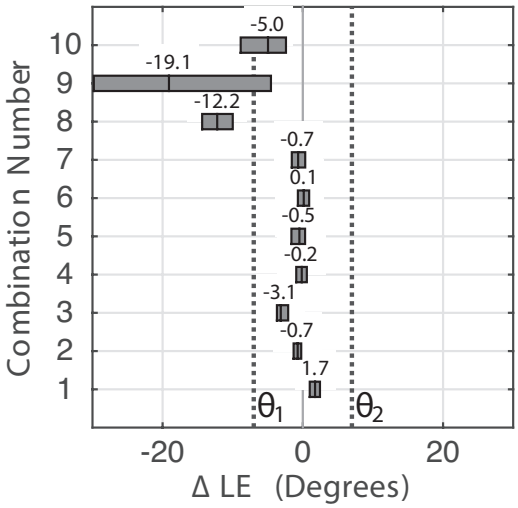


Figure 6.5: ΔLE results for each combination used to perform two one-sided test (TOST) equivalence tests. Perceptual equivalence boundaries, θ_1, θ_2 are shown at ±7°. Listening position 1, X = 0m, Y = 0m.

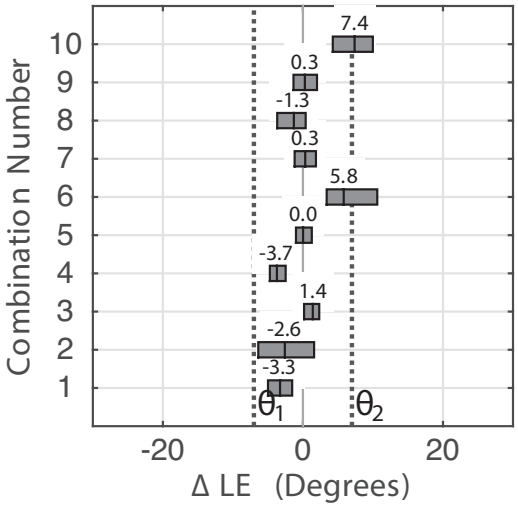


Figure 6.6: ΔLE results for each combination used to perform two one-sided test (TOST) equivalence tests. Perceptual equivalence boundaries, θ_1, θ_2 are shown at ±7°. Listening position 2, X = -0.5m, Y = -0.5m.

6.5 Discussion

Mean signed localisation error is a commonly used metric for measuring human sound localisation in a closed-loop localisation task. To assess a baseline for the results found in this experiment, signed localisation data was presented for listening

positions 1 and 2 in Fig. 6.2 and Fig. 6.3 respectively. For both listening positions the mean signed localisation errors for AVE follow the in situ results well, even when the underlying localisation error is large such as combinations 8, 9 and 10 at listening position 2. To consider the data on aggregate, the average absolute deviation in mean signed localisation error across all combinations was found to be 4.0° for listening position 1 and 5.7° for listening position 2. It is clear from the results that some cases were more problematic such as combinations 8, 9 and 10 at listening position 1 and combinations 2 and 6 at listening position 2. Referring to Table 6.2 it can be noted that these combinations had 2 or 4 loudspeakers active in the reproduction and had large loudspeaker spacings. It can also be seen that although the difference in mean signed error values are visibly larger than other combinations, the error bars are also larger, indicating a larger apparent source width for the auditory event. For this reason it is difficult to understand the perceptual significance of these deviations without defining clear perceptual limens.

Results from the rANOVA showed that the *Combination* and *Listening position* within-subject factors had the largest effect sizes of the main four factors and both were statistically significant at the $p < 0.05$ level. The interaction *Listening position * Combination* was also found to be significant.

To consider the influence of using the AVE to simulate localisation artefacts it is possible to look at the *auralisation method* factor alone. This was found to fail to reject the null-hypothesis ($F(1) = 3.355$, $sig. = .088$, $\eta_p^2 = 0.193$) at the 0.05 significance level. However, it is important to clearly define what the p-value represents here: the p-value tells us that if the mean unsigned localisation error of the two samples were *equal* (i.e. the null-hypothesis is true), the probability of achieving the values measured in this study is 0.088. Although the p-value was found to be greater than 0.05, this value does not provide information about how

the significance of the auralisation method changes across combination or listening position. The p-value also suffers from the issue that only statistical, not practical significance is considered. The scientific use of p-values has become so confused that the Journal of the American Statistician recently released a statement on the topic (Wasserstein and Lazar, 2016). To understand the practical impact, the effect size estimate η^2 was calculated to understand the contribution of variance from each of the factors and their interactions. Both η_p^2 and η^2 are presented in Table 6.3 but η^2 is favoured for analysis as it gives the ability to compare within-subject factors against each other (summing η^2 for each factor, interaction of factors and errors accounts for 100% of the measured variance). η^2 values highlight the large differences in effect size, with *Combination* accounting for 0.4893 of the variance and *Repeat* accounting for only 0.0003.

Although some inference about the equality of localisation artefacts could be drawn from the p-values result, unfortunately, the null-hypothesis is inherently flawed from the outset. For this reason, an alternative analysis was performed where a two one-sided test (TOST) method is used to analyse the perceptual equivalence of the AVE and in situ localisation error results.

Results from the two one-sided tests for each combination at listening positions 1 and 2 separately can be seen in Fig. 6.5 and Fig. 6.6. It is firstly worth considering combinations 3 and 7 separately during evaluation of the AVE. These were mono sound sources and therefore can be used as a benchmark in this test. For these combinations the AVE performed well with mean values close to zero and small confidence boundaries comparable to previous studies (Rychtarikova et al., 2009; Hiekkanen et al., 2009) and literature results for the minimum

audible angle.

At listening position 1 (X = 0, Y = 0), it can be seen from Fig. 6.5 that Combinations 1 - 7 have means close to zero with small variance. Combinations 8 - 10 highlight an increase in both variance and central tendency. Combination 9 specifically highlighted a substantial increase in the confidence boundaries and performed worst out of all combination*listening position interactions. It is important to note that the variance in ΔLE is not only influenced by the error introduced by the auralisation method (in situ or AVE) but also, the listeners localisation precision, tracking system measurement error and reporting method error. Results for ΔLE of combination 9 may indicate that the variance of ΔLE is dominated by the localisation precision, which would show an increase in the difficulty of the localisation task. A systematic change in ΔLE of -19.1° also highlights that the AVE failed to accurately simulate in situ reproduction for this combination. Combinations 8 and 10 did not show an increase in confidence boundaries to the extent of combination 9 but the variance in ΔLE was slightly larger than for 1-7, possibly due to the AVE failing to create the exact localisation cues. Combinations 8, 9 and 10 had loudspeaker spacing of 140°, 80° and 90° respectively, which would likely stimulate confusing localisation cues. Reports (Stitt et al., 2014; Bates et al., 2007b) for similar systems have highlighted the perceptually challenging auditory artefacts found in loudspeaker systems with low spatial resolution. For in situ, listeners may have been forced to seek more complex localisation cues such as translational changes or monaural spectral characteristics, which may not have been well represented by the AVE.

At listening position 2 (X = -0.5m, Y = -0.5m), combinations 1, 3, 4, 5, 7, 8 and 9 all fall within the equivalence regions and can be said to perform well with

narrow confidence intervals. Combinations 6 and 10 have confidence regions outside the equivalence boundaries and therefore would be perceptually problematic when using AVE for simulation. Both combinations had loudspeaker spacing of 90° and again may have suffered from the same limitations of combinations 9-10 at listening position 1. Combination 9 is particularly interesting where ΔLE has a mean close to zero and narrow confidence intervals but the underlying localisation error is large ($\approx -90°$) as seen by the interior markers in Fig. 6.4. At the off-centre position the virtual image is collapsed to the nearest (surround right) loudspeaker and therefore although the error between intended and perceived direction is considerable, localisation precision is improved and the ΔLE is reduced. This means the AVE correctly stimulated the adverse artefacts of the reproduction system most likely related to the precedence effect.

10 combinations were tested at the central and one non-central listening position. Aside from the specific differences discussed above, 15 of the 20 combination-listening position interactions were found to be equivalent within the pre-defined equivalence boundaries of ±7° . The 5 combination-listening position interactions that failed the equivalence test were from combinations 6, 8, 9 and 10. These combinations all had substantial loudspeaker spacings (greater than 80°) using a square or ITU loudspeaker layout. This may have induced challenging localisation cues for the AVE to simulate. This shows that although the localisation judgements were not the same for all combinations tested, for well localisable sound sources, artefacts of reproduction systems and listening positions are well maintained.

Having a single-judgement pointing method for this test could have been a

limitation. If participants were presented with a sound source that is un-localisable or had a split or broad image they were still required to make a directional judgement that was inherently treated with the same weighting as a well-localised auditory event. Anecdotal reports by subjects in a recent localisation study by Stitt et al. (2014) noted the perception of multiple auditory events which could have also been possible in this study. The introduction of a confidence or image width reporting method or the ability to report multiple auditory event perceptions would have aided this problem and would have avoided ΔLE samples to be dominated by the random error in directional judgements.

As a general conclusion, the applicability of the AVE system tested for simulating reproduction system artefacts depends on the type of system being simulated and the general accuracy required by an experimenter. For well-localisable auditory events, such as those created by mono sound sources, amplitude panning with a stereophonic layout or high-order Ambisonic systems with small loudspeaker spacing, non-individualised dynamic binaural simulations such as the one used in this paper are equivalent ($\pm 7°$) to in situ auralisation. However, for systems which may induce confusing localisation cues such as large loudspeaker spacing or low-order Ambisonics, results for binaural simulations may be non-equivalent and a more accurate binaural simulation should be used. Possible developments of the AVE used in this thesis could include the use of personalised HRTFs/BRIRs, head-elevation tracking or translation tracking.

6.6 Conclusions

The aim of this chapter was to evaluate the degree to which localisation artefacts presented during off-centre listening are perceptually equivalent between in situ (using real loudspeakers) and a binaural simulation. A closed-loop localisation test was performed using a selection of representative panning methods, loudspeaker layouts and stimuli at both the central and one non-central listening position. The test was performed in situ and repeated using a non individualised, dynamic binaural simulation system that incorporates all important perceptual cues.

The localisation judgements were used to compute localisation error across each of the independent variables of the test. The baseline metric of mean unsigned localisation error for each of the combinations indicated that the binaural system was capable of maintaining large localisation errors. The aggregated absolute deviation between in situ and the binaural system was found to be 4.0° for the central listening position and 5.7° for the non-central listening position.

A repeated measures ANOVA showed that factors *combination* (selections of panning method, loudspeaker layout, stimuli and panning direction) and *listening position* were statistically significant and also accounted for the largest effect sizes of the main factors. The auralisation method factor (in situ or AVE) failed to reject the null hypothesis.

To understand the practical equivalence of the in situ results, to those recorded using binaural simulation, equivalence tests were performed on each system combination using a two one-side test (TOST) framework. It was found that 15 out of 20 system combination/listening position interactions were equivalent

within pre-defined equivalence boundaries of ±7°. However, certain system-combinations with poor spatial resolution and large loudspeaker spacings created larger differences between in situ and AVE results. It has been concluded that for loudspeaker-based spatial audio reproduction systems with well localisable auditory events the AVE simulation was equivalent (±7°) to in situ auralisation but care should be taken when simulating loudspeaker systems with poor localisation fidelity.

CHAPTER 7

Simulating Localisation Artefacts Across the Listening Area Using a Computational Model

---

This chapter presents the development to a computational localisation model proposed by Sheaffer (2013) built upon previous work by Faller and Merimaa (2004). The model is developed to include head/torso movements which resolve front-back confusions and in turn, simulate a closed-loop localisation task in anechoic and reverberant environments. The current standing model is firstly introduced. Developments are then described before the model being validated against subjective data from Chapter. 6.

7.1 Introduction

The computational model in this thesis is a development of the model introduced and used by Sheaffer (2013). This model was chosen due to its resilience to complex listening situations (Faller and Merimaa, 2004) and ability to model temporal localisation cues as highlighted by Sheaffer (2013). Other localisation models have been applied to the prediction of localisation performance in loudspeaker-based systems as described in Chapter 3. A recent adaptive binaural model has also been implemented by Braasch et al. (2013) where a short region of head-motion (0°-30°) is simulated to resolve multiple peaks in the interaural cross-correlation and EI (excitation-inhibition) cell patterns. This simulation of cognitive integration over head-rotations is similar in philosophy to the model applied here and has been shown to work in resolving front-back confusions caused by the similarities in interaural cues around the cone-of-confusion. Interaural level differences in the model were handled in a different manner than was chosen here, whereby an excitation-inhibition process was used to compare ILD patterns. The work of Braasch et al. (2013), however, highlights the importance of utilising head-rotations in the computational modelling of human localisation.

The model used here utlises the principle of interaural cue selection based on interaural coherence (IC) (Faller and Merimaa, 2004) which has been shown to simulate the human auditory system's ability to localise sounds in complex listening situations. Sheaffer (2013) showed that a number of psychoacoustic phenomena could be explained using the model when localising sounds in anechoic and reverberant conditions with BRIRs simulated from a finite-difference time-domain model. One limitation of the computational

localisation model is the inability to resolve front-back confusions due to the similarity of interaural cues about the coronal plane. After introducing the fundamental concepts of the computational model proposed by Sheaffer (2013), a development is described that can be used to resolve front-back and back-front confusions. This method can also be used to simulate a closed-loop localisation task, where a listener is asked to judge the direction of a continuous sound source by pointing with their head or nose (as used in Chapter 6). The model is then applied to binaural stimuli from the SBSBRIR dataset and results are compared to subjective data from Chapter 6. The model proposed by Sheaffer (2013) is firstly introduced in section 7.2, followed by the definition of an additional processing stage used to model the dynamic localisation process of a human listener. The model is then applied to predict the direction of a sound source within a reverberant listening environment, utilising data from the SBSBRIR dataset. The model is finally compared to the subjective localisation data for a number of reproduction system combinations to validate the ability to use such models for the simulation of localisation artefacts found in complex listening situations.

7.2 The Existing Computational Model

In this section, the current localisation model presented by Sheaffer (2013), built upon fundamental concepts of Faller and Merimaa (2004) is described. The aim of the model is to allow the prediction of sound source localisation in complex listening scenarios using only the binaural sound stimulus at the entrance to the auditory system.

7.2.1 The Peripheral Auditory System

The first stage of the computational model simulates the physical effects of the outer, middle and inner ear. The structure and description of this part of the auditory system has been described in Chapter. 2.

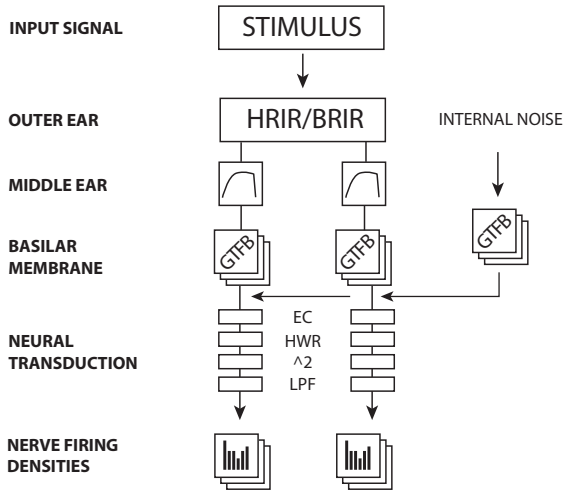


Figure 7.1: Processing stages of the peripheral auditory system model. The neural transduction modelling is separated into four parts: envelope compression [EC], half-wave rectification [HWR], squaring [ˆ2] and low-pass filtering [LPF].

For anechoic listening conditions, the pressure at the entrance to the auditory system can be modelled as the convolution of a stimulus signal with a pair of HRIRs for a specific sound position. The middle ear response is modelled using a Butterworth band-pass filter (-3 dB/octave) with the passband between 1kHz and 4kHz. A gammatone filter bank (GTFB) is then applied to the left and right signals to approximate the frequency selectivity of the basilar membrane. 44 overlapping gammatone filters, each with a width of 1 ERB were used for the filter bank (Søndergaard et al., 2011). The dynamic range of the human auditory system is accounted for in the model by also passing scaled Gaussian noise through the same GTFB. Noise-scaling values according to ISO 389 (ISO, 1975)

as recommended by Faller and Merimaa (2004) were used and interpolated for each filter centre-frequency. A model of neural transduction (Bernstein et al., 1999) is then used to simulate the process of energy transduction from sound to electrical signals. The neural transduction stage can be split into four parts, computed per auditory filter band and for left and right ears separately: (1) envelope compression by raising the signal envelope to the power of 0.23, (2) half-wave rectification, (3) squaring and (4) low-pass filtering using the filter definition by Bernstein and Trahiotis (1996). The left-ear nerve firing density for an impulse input is shown in Figure. 7.2. Raising the signal *envelope* to the power of 0.23 (van de Par and Kohlrausch, 1998), has been shown to simulate the compressive effect of the basilar membrane. This compressive effect can be demonstrated by the fact that an increase in the level of a stimulus by 1 dB, causes an increase in basilar membrane response by approximately 0.2 dB (Bernstein et al., 1999). The application of a low-pass filter to the half-wave rectified and squared signal simulates the physiological inability of the human auditory system to synchronise neural responses at high-frequencies. At low frequencies, both fine structure and signal envelopes are transduced by the auditory system. However, in the high-frequency region only signal envelopes are transduced (Bernstein and Trahiotis, 1996).

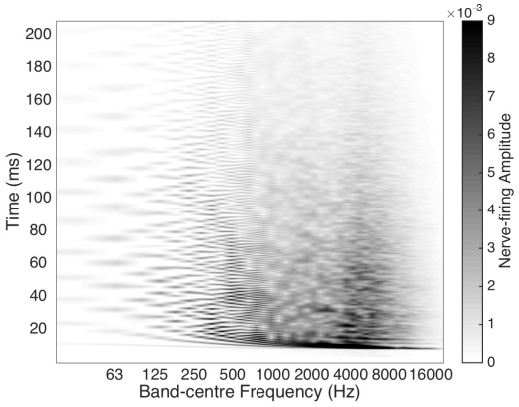


Figure 7.2: Nerve firing density data for the left ear. Data is shown for the response of the auditory model to a Dirac delta input to a loudspeaker at 0°, head-azimuth 0° for the SBSBRIR dataset.

The left and right ear nerve firing densities can be represented as $x_L(k,t)$ and $x_R(k,t)$ respectively where k is the auditory filter index and t is the instantaneous time (sample) index.

7.2.2 Binaural and Central Processing

The binaural processing stage uses the selection criteria implemented by Faller and Merimaa (2004). Firstly, the time-domain nerve firing densities in each auditory filter band are split into time windows so that interaural cues can be processed. All processing in this chapter uses the 10 ms time window proposed by Faller and Merimaa (2004). By calculating the running inter-aural cross-correlation function, IC and ITD values are estimated as the maximum of the interaural cross-correlation and the argument of the maximum of the interaural cross-correlation respectively. ILD is calculated as the energy difference between left and right nerve-firing densities and all values are calculated as a function of time window index, n and auditory filter index k. These three values are represented by:

$$\begin{matrix} ITD(k,n) \\ ILD(k,n) \\ IC(k,n) \end{matrix} \tag{7.1}$$

ITD and ILD values are then selected as valid according to a frequency dependent threshold C_0. At each time constant n, $ITD(k,n)$ and $ILD(k,n)$ are only used if the corresponding $IC(k,n)$ value is greater than C_0. Faller and Merimaa (2004) proposed that the physiological representation of C_0 is adaptive, depending on the room the listener is in or other external factors. Sheaffer (2013) used empirical data to define Equation. 7.2, the primary feature being an increase in the threshold at high-frequencies.

$$C_0(k) = (1 - e^{-\mu k}) \tag{7.2}$$

where k is the auditory filter band index $(1, 2, ..K-1, K)$ where $K = 44$, the total number of auditory filters. μ is the control parameter for changing the slope of C_0 with respect to frequency. It is important to note that this selection function is not scaled according to frequency, but the number of filters in the auditory filterbank and therefore, the tuning parameter μ should be selected carefully when a different number of auditory filters is used. For calculations shown here, $\mu = 0.15$ (Sheaffer, 2013).

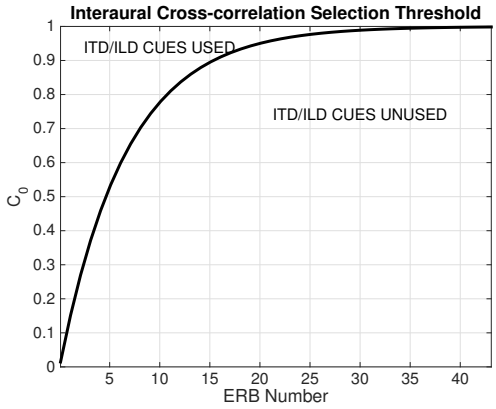


Figure 7.3: C_0 function used to select valid ITD and ILD values. $\mu = 0.15$

Probability density functions, $PDF_{ITD}(k, \tau)$ and $PDF_{ILD}(k, \alpha)$ are created by counting the occurrences of ITD and ILD values across n after the C_0 selection criteria. τ and α values are constrained to the minimum and maximum possible range across which ITD and ILD are calculated. Figure. 7.4 shows a time-domain representation of the calculation process.

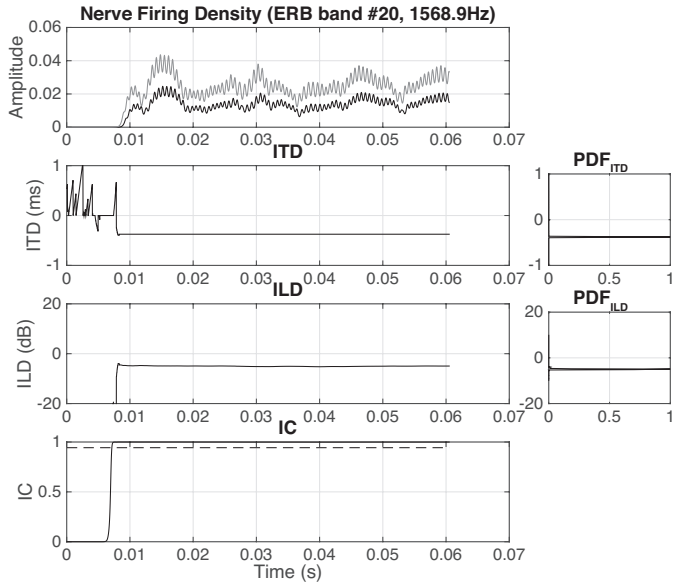
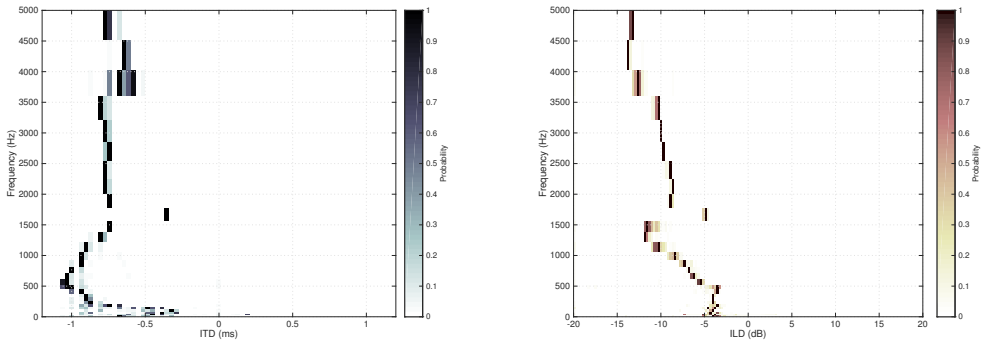


Figure 7.4: The process for calculating valid ITD and ILD values and probability density functions. Data is shown for a sound source at $\theta = 315°$ and ERB band 20 with a centre-frequency of 1568.9Hz. Left hand side, from the top (a) input left and right nerve firing densities, (b) calculated ITD value, (c) calculated ILD value (post peripheral processing), (d) IC function with corresponding C_0 threshold value.

Histogram plots on the right side show the probability density values for the chosen ERB filter band. The PDFs are created for both ITD and ILD values and for each auditory filter band, represented by $PDF_{ITD}(k, \tau)$ and $PDF_{ILD}(k, \alpha)$ respectively. Figure. 7.5 shows the PDF data across all frequency bands. PDF data is normalised for each auditory filter k.



(a) $PDF_{ITD}(k, \tau)$ (b) $PDF_{ILD}(k, \alpha)$

Figure 7.5: Cue probability functions for a sound source at $\theta = 45°$.

Sheaffer (2013) combines $PDF_{ITD}(k, \tau)$ and $PDF_{ILD}(k, \alpha)$ into compact matrix form named the *cue probability pattern* $CPP(k, \theta)$. For each sound source azimuth θ and auditory filter k, a 2-dimensional matrix can be calculated to define the localisation characteristics where,

$$CPP(k, \theta) = \begin{bmatrix} PDF_{ITD}(k, \tau) \\ PDF_{ILD}(k, \alpha) \end{bmatrix} \tag{7.3}$$

The assumption of this stage of the computational model is that for each of the possible directions of sound stimuli presented to the model, there is an almost unique CPP matrix. A reference dataset is created by calculating CPP values for anechoic data at known sound source directions. Each CPP is labelled according to the sound source direction. An unknown binaural *test* stimulus can then be analysed by firstly calculating CPP data. The correlation between the *test* CPP and the *reference* CPP data for each angle will provide a prediction of the most likely stimulus directions.

Correlation analysis is performed by taking a 2D cross-correlation of the CPP data per frequency band. For computations of the localisation model used in this

chapter, the reference dataset was calculated by taking an anechoic approximation of the SBSBRIR dataset as described in Chapter. 4. This ensured that the same electroacoustic equipment was used for the reference and test stimuli.

The resulting *cue correlation function*, $CC(k, \theta)$ gives a visual representation of the localisation angle of the test stimulus for each frequency band, k. For modelling localisation cues across frequency, Stern et al. (1988) provides a best-fit third order polynomial model based on original data from Raatgever (1980) as,

$$\omega(f) = 10^{-(b_1 f + b_2 f^2 + b_3 f^3)/10} \tag{7.4}$$

Where $b_1 = -9.383 \times 10^{-2}$, $b_2 = 1.126 \times 10^{-4}$, $b_3 = -3.992 \times 10^{-8}$. At frequencies above 1200 Hz, the function is set to the constant value of $\omega(1200)$. This integration process provides the summarised localisation prediction, $S(\theta)$.

7.3 Modelling Dynamic Cues

Now that the model implemented by Sheaffer (2013) has been introduced, this section describes the novel developmentss. It has been shown that head-movements made by human listeners help to resolve front-back confusions in sound localisation (Blauert, 2001; Algazi et al., 2004a). Consider a sound event placed at $\theta = 45°, \phi = 0°$ relative to a listener. This sound event will have very similar ITD and ILD cues to a sound placed at $\theta = 135°, \phi = 0°$. However, when the listener rotates their head, ITD and ILD cues will change accordingly and oppositely for a sound source placed in front, or behind the listener. This

dynamic effect is included in this computational localisation model to account for front-back confusions.

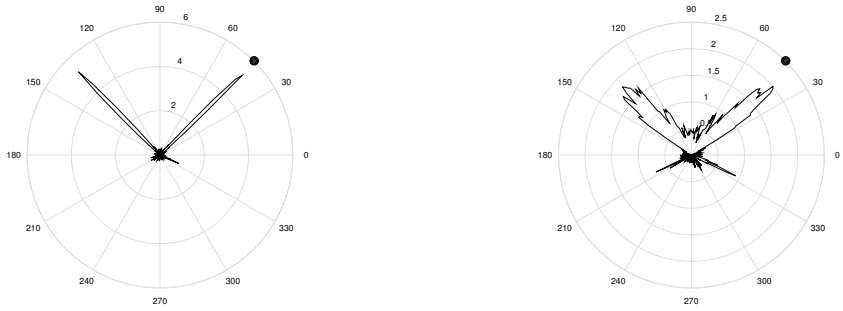
Using the dataset of CPPs from free-field measurements, cue correlation data indicates the egocentric direction of a sound relative to the listener's head/torso. However, when localising a sound source in practice, head-movements (movements of the listener's body frame) can be made by the listener to achieve a judgement of the sound source's position within an external reference frame, which could be defined by the coordinate system of the listening room.

A closed-loop localisation task as used in Chapter. 4 and Chapter. 6 can be simulated by introducing an additional stage to the central processing. The model output is calculated for a number of head positions within the environment and a memory and integration stage (introduced in the following pages) is used to resolve front/back confusions and improve the model's precision.

The reference set of CPPs were firstly calculated using a free-field approximation of the SBSBRIR dataset as demonstrated in Chapter. 4. As an example of the dynamic processing, consider a loudspeaker placed a 45° in a listening room. This can be simulated by taking BRIR measurements from the SBSBRIR dataset at the central listening position.

With the head-azimuth $\theta_{head} = 0°$, binaural stimuli can be processed using the localisation model. The resultant summary correlation data $S(\theta)$ shows the azimuth direction of the sound source relative to the head. The data indicates high correlation of interaural cues with angles 45° and 135° as indicated by Figure. 7.6. The similarity of interaural cues around the cone-of-confusions

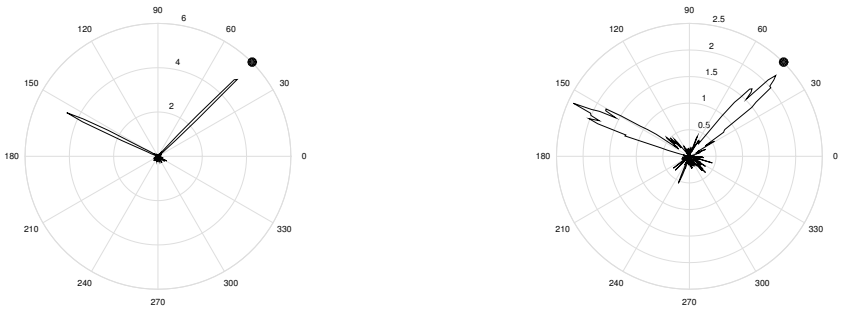
means that the static model cannot easily resolve front/back confusions.



(a) Anechoic (b) Reverberant

Figure 7.6: $S(\theta)$ localisation model predictions without head-rotations. Binaural stimuli is for a loudspeaker at 45°. The head is facing forward in this prediction therefore θ_{head}=0°.

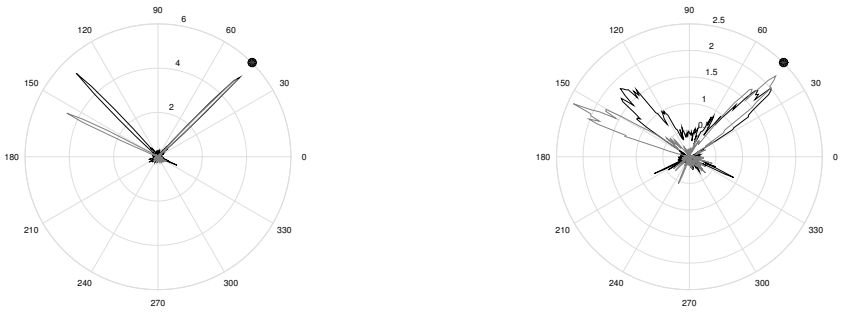
Using the SBSBRIR dataset it is possible to virtually rotate the head and torso of the listener in the listening room and re-calculate the summary correlation data, $S(\theta)$. Setting $\theta_{head} = 10°$, the ego-centric data now predicts high correlation at 35° and 145°. The resultant directional prediction, $(S(\theta))$, is then rotated by the $-\theta_{head}$ to align the coordinate systems of the room and the head direction. Thereby correct directional judgement at 45° is achieved. However, due to the change in interaural axis, the erroneous correlation around the cone-of-confusion is now found at 155°, compared with 135° at $\theta_{head} = 0°$ as shown in Figure. 7.7. $S(\theta)$ for $\theta_{head} = 0°$ and $\theta_{head} = 10°$ is overlaid in Figure. 7.8 to show the deviation in the directional prediction of the front-back confusion.



(a) Anechoic (b) Reverberant

Figure 7.7: $S(\theta)$ localisation model predictions with head-rotation. Binaural stimuli is for a loudspeaker at 45°. The head and torso simulator is rotated by 10° in this prediction therefore θ_{head}=10°.$S(\theta)$ is corrected by rotating the resultant data by $-\theta_{head}$ to give a prediction within the global coordinate system.

The resultant localisation judgement data for $\theta_{head} = 0°$ and $\theta_{head} = 10°$ is shown together in Figure. 7.8, where the data is rotated to align head rotation with the environment geometry. Localisation judgements in the direction of the sound source are in agreement for both θ_{head} values but front-back confusion errors are separated.



(a) Anechoic (b) Reverberant

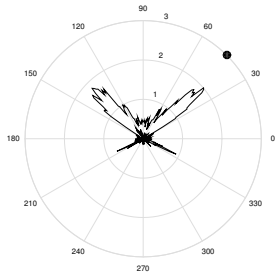
Figure 7.8: $S(\theta)$ predictions for $\theta_{head} = 0°$ and $\theta_{head} = 10°$ overlaid to highlight the effect of resolving front/back confusions.

$\hat{S}(\theta_{GC})$, the directional judgement of the sound source within the listening environment global coordinate system can be calculated by iterating the

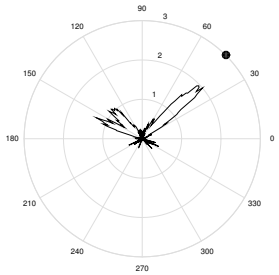
head-rotation θ_{head} across a selection of discrete angles and averaging the resultant localisation judgements, $S(\theta)$.

$$\hat{S}(\theta_{GC}) = \frac{1}{N} \sum_{\theta_{head}=0}^{N} S(\theta) \tag{7.5}$$

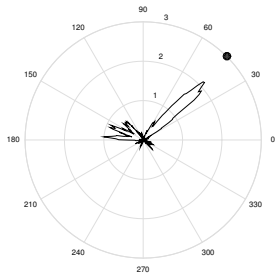
Where θ_{GC} is the angle within the listening environment global coordinate system and N is the number of discrete head rotations used in the dynamic localisation process. Before the mean is calculated, the coordinate system is aligned by accounting for the head-rotation, θ_{head}. As shown in Figure. 7.9, as the model iterates through θ_{head} values, the erroneous front-back confusion is resolved into noise via the averaging and realignment of $S(\theta)$ back to the global coordinate system.



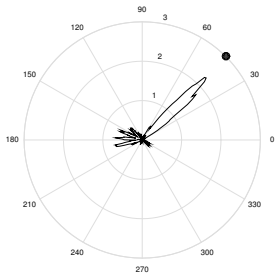
(a) $\theta_{head} = 0°$



(b) $\theta_{head} = 0°, 10°$



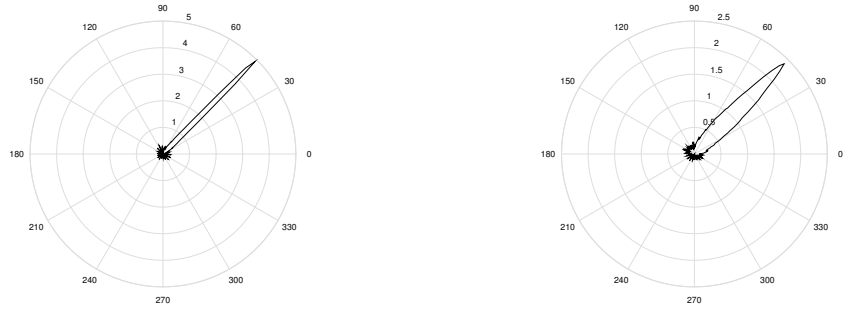
(c) $\theta_{head} = 0°, 10°, 20°$



(d) $\theta_{head} = 0°, 10°, 20°, 30°$

Figure 7.9: $\hat{S}(\theta_{GC})$ for 1, 2, 3 and 4 θ_{head} iterations. It can be seen that as the number of head positions increases, the erroneous front-back confusion prediction is reduced into noise.

The use of dynamic cues improves the directional judgement and simulates the dynamic localisation process of humans to resolve front-back confusions around the cone-of-confusion. Although the computational model implemented here goes through a fixed set of head-rotations to achieve the final localisation task, a closed-loop localisation task will likely be more variable in the type of head-movements made, depending on the starting and final head directions and time allowed to make the localisation judgement.



(a) Anechoic (b) Reverberant

Figure 7.10: $\hat{S}(\theta_{GC})$ predictions for $\theta_{head} = 0° : 10° : 350°$ for binaural stimuli from a loudspeaker at 45°.

The resultant circular cue correlation data, averaged over the iterated head-directions provides a probability density function for the localisation judgement direction with a dynamic feedback process to resolve front-back confusions as shown in Figure. 7.10. To achieve a single direction judgement from this data, the angular argument of the maximum value in the distribution, $argmax(\hat{S}(\theta_{GC}))$, provides the most probable localisation direction. However, the variance and shape of the distribution can be used to inform the confidence of the localisation judgement (as used by Sheaffer (2013) on the model output without head-rotations). The variance of the model could also be used to predict the stability of the image under head-rotations or translation if the model were adapted to account for lateral movements.

7.4 Results

In this section of the chapter, a qualitative analysis of the artefacts introduced at non-central listening positions using loudspeaker-based spatial audio reproduction is performed. To validate the localisation model, localisation predictions are firstly made on the localisation of a single loudspeaker by utilising data from the SBSBRIR dataset. Following this, localisation predictions from the model are compared against subjective results from the closed-loop localisation test found from a number of loudspeaker-based panning methods found in Chapter. 6.

7.4.1 Single Loudspeaker in a Reverberant Environment

As a benchmark, it is important to understand how well the computational model can predict the localisation accuracy of a single loudspeaker in a reverberant listening room at multiple listening positions. To analyse this qualitatively, BRIRs from the SBSBRIR dataset are utilised and the dynamic localisation model described above is applied to achieve a prediction of a closed-loop localisation task. The stimulus signal for this analysis is a rectangular windowed Gaussian noise burst which lasts 63 ms (3000 samples at 48kHz). 36 θ_{head} positions at 10° intervals were used to resolve front-back confusions and data was used on the full-length BRIRs to ensure all the reflections were included in the model.

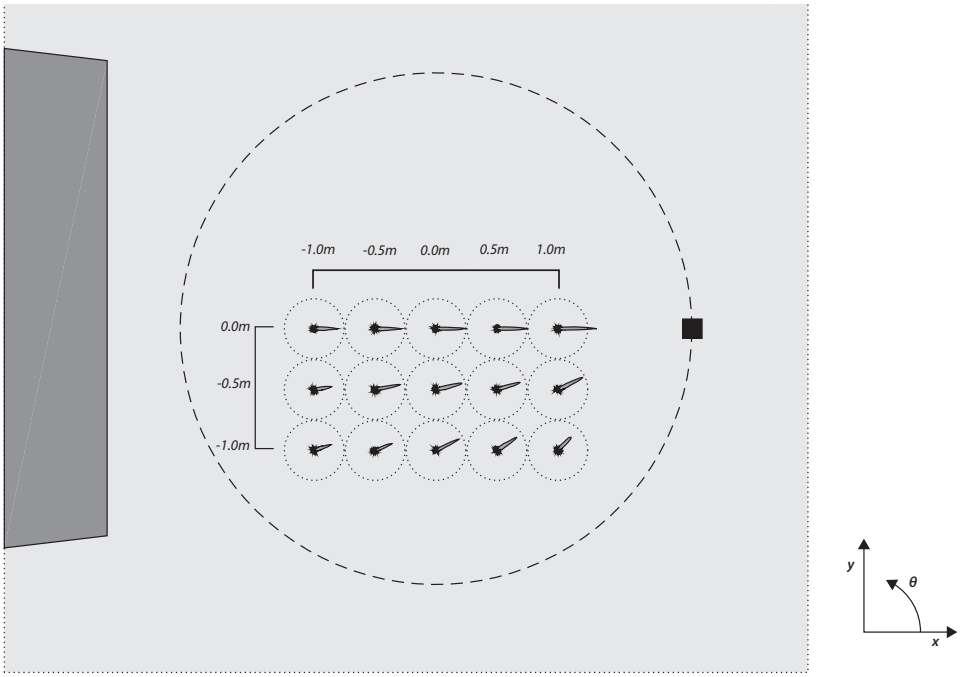


Figure 7.11: $\hat{S}(\theta_{GC})$ functions at each listening position for a single loudspeaker at 0°. BRIRs were taken from the SBSBRIR dataset where no truncation of the reverberation was applied. The black square indicates the loudspeaker position and interior walls of the listening room are shown.

By using $argmax(\hat{S}(\theta_{GC}))$, it is possible to achieve a single angular value that represents the model's directional judgement and therefore an approximation of a closed-loop localisation task. Comparing this data to the physical angle between a forward facing listener and the loudspeaker at each listening position it is possible to describe how accurate the model is at predicting localisation. The mean and standard deviation for signal localisation error is shown in Table. 7.1.

Table 7.1: Mean and standard deviation in localisation error between actual speaker directions and model predcitions for a single speaker at 0°. Compare data with model predictions shown in Figure. 7.11.

	Signal Localisation Error
Mean	0.81°
Std. Dev.	0.84°

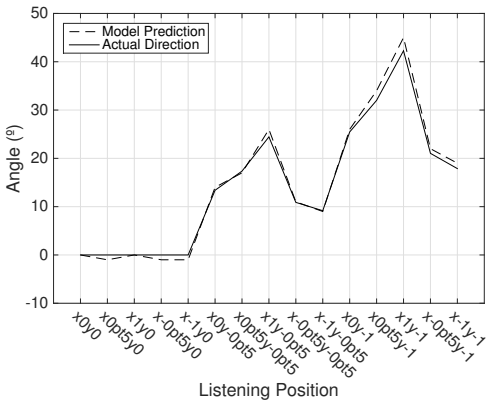


Figure 7.12: Comparison of model directional prediction and actual loudspeaker direction. Data for the model shows $argmax(\hat{S}(\theta_{GC}))$ for each listening position for a loudspeaker at 0°, 2.1m (model data shown in Figure. 7.11. Actual direction data is the physical angle between the forward facing listener and the loudspeaker.

7.4.2 Comparison with Subjective Results

In Chapter. 6, 10 reproduction system combinations were implemented in a localisation test using both in situ loudspeaker reproduction and simulation using the AVE. To understand how well the computational localisation model can represent the artefacts of off-centre listening in loudspeaker-based systems, model predictions will be made based on the systems used in the previous test and results will be compared.

Firstly, the binaural stimuli were presented to the computational model using data from the SBSBRIR dataset. Head rotations every 10 degrees were implemented to resolve F/B confusions and simulate the closed-loop localisation task. The model output data is shown in Figures. 7.13 and 7.14. The actual stimuli used in the localisation task were applied to the model prediction. Reference CPP data was calculated using noise bursts.

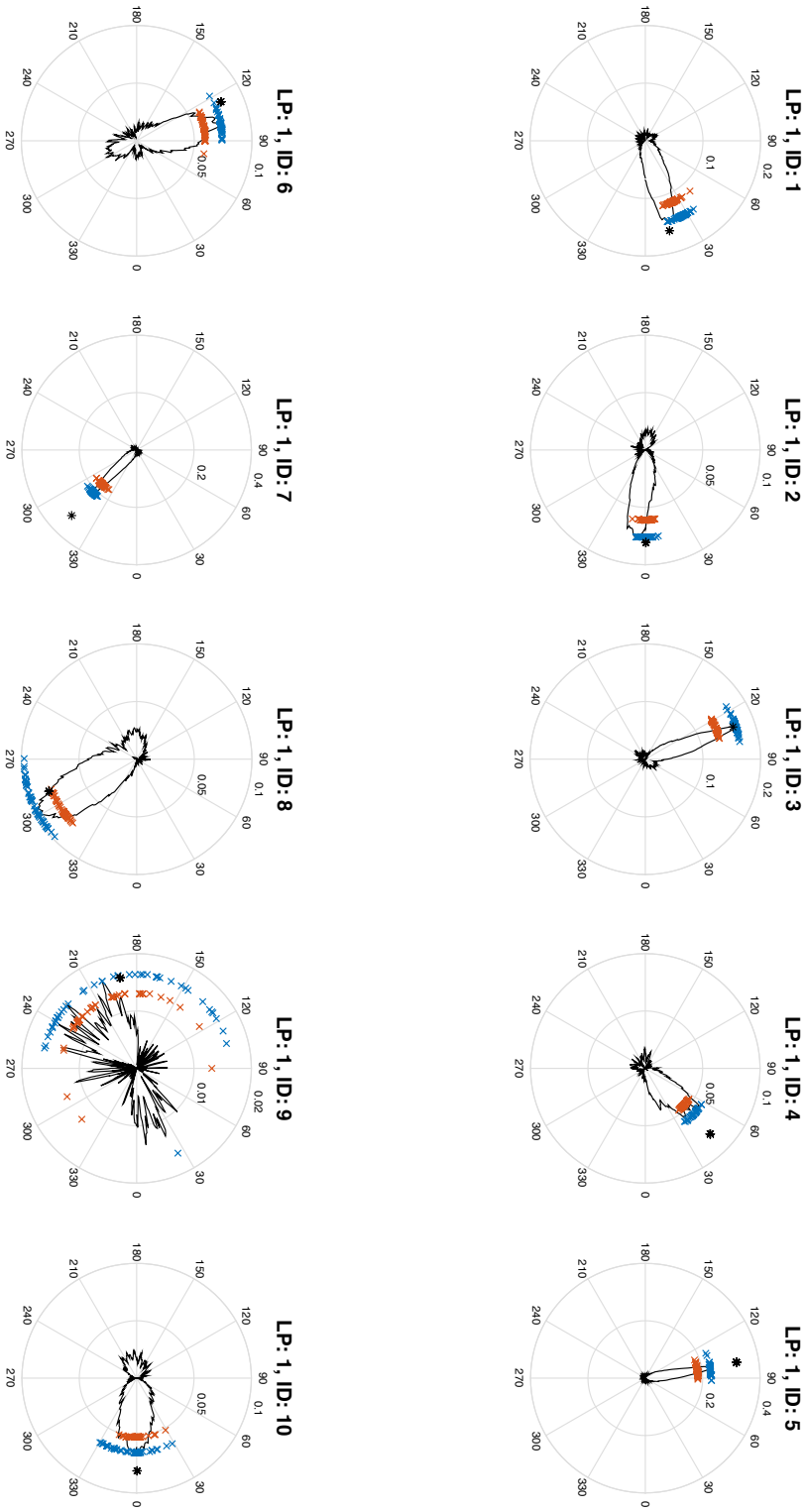


Figure 7.13: Model predictions compared against subjective data for the 10 combinations used in Chapter. 6, listening position x=0, y=0. $\hat{S}(\theta_{GC})$ data is plotted direction with a black line. Blue crosses represent the subjective localisation judgements of in situ reproduction and red crosses show the subjective localisation judgements with AVE simulation. Black * show the intended panning direction.

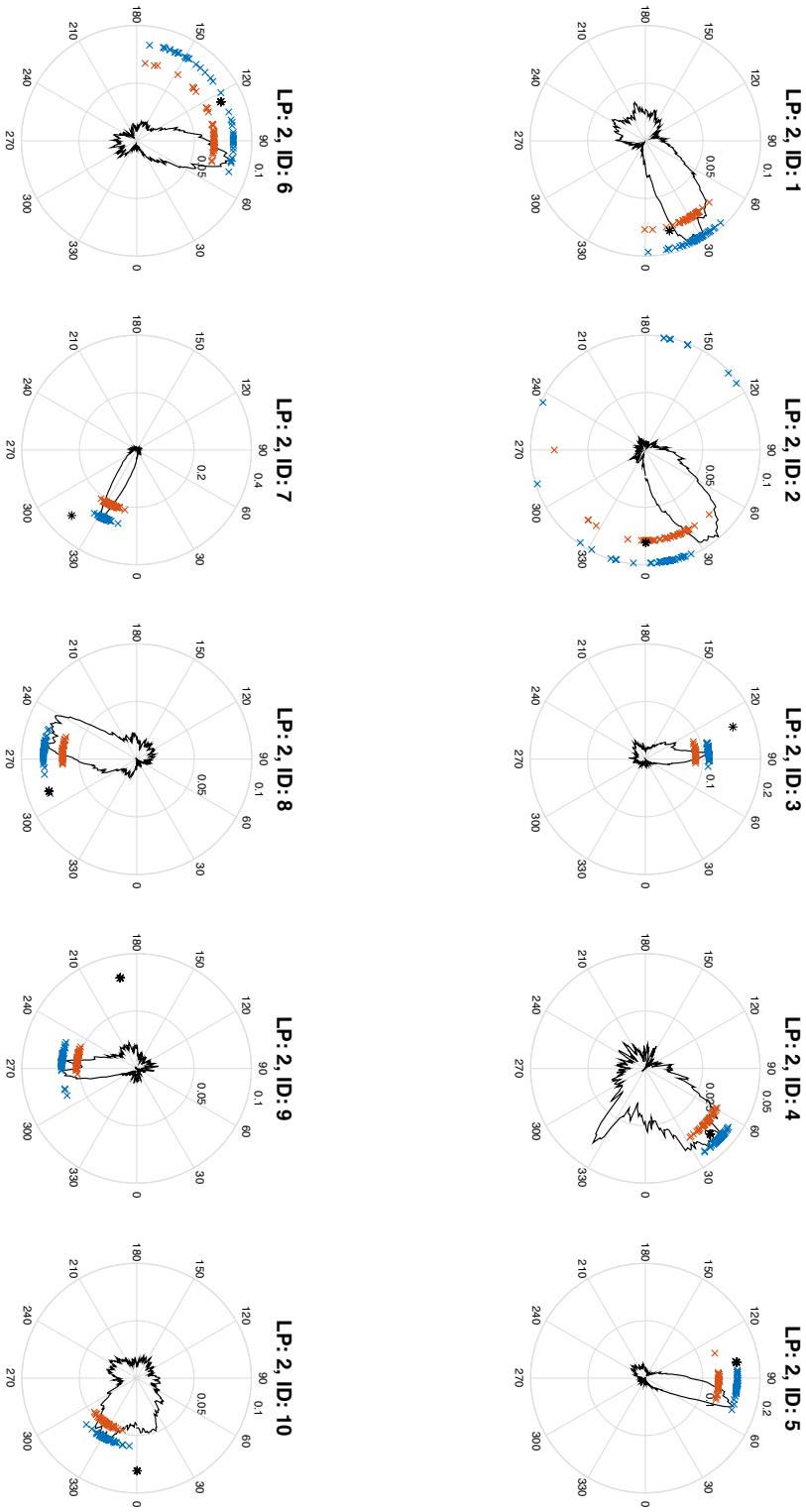


Figure 7.14: Model predictions compared against subjective data for the 10 combinations used in Chapter. 6, listening position x=-0.5, y=-0.5. $\hat{S}(\theta_{GC})$ data is plotted direction with a black line. Blue crosses represent the subjective localisation judgements of in situ reproduction and red crosses show the subjective localisation judgements with AVE simulation. Black * show the intended panning direction.

Figures. 7.15 and 7.16 show the signed localisation error, where subjective responses are compared against model prediction for each combination at both listening positions.

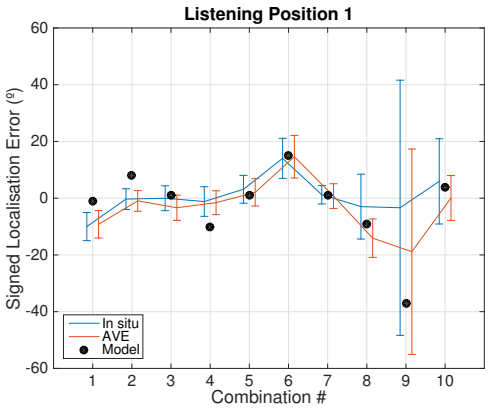


Figure 7.15: Localisation error for in situ, AVE and model for combinations 1-10 at listening positions x=0, y=0. The full height of the error bar represents 1 standard deviation of underlying sample.

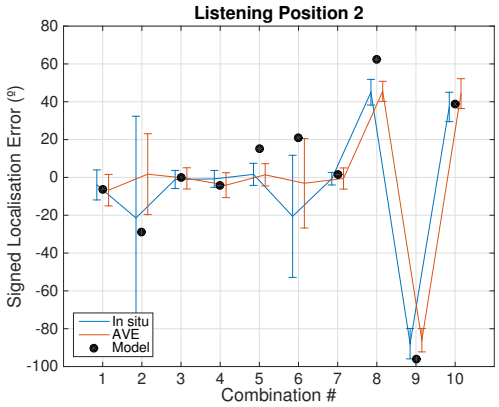


Figure 7.16: Localisation error for in situ, AVE and model for combinations 1-10 at listening positions x=-0.5, y=-0.5. The full height of the error bar represents 1 standard deviation of underlying sample.

7.5 Discussion

In this chapter a computational model initially proposed by Sheaffer (2013) and based on previous work on the selection of binaural cues has been defined and

modified for the specific application of simulating a closed-loop localisation task in a reverberant environment. The model was specifically implemented to understand the validity in using the model for the simulation of localisation error found at off-centre listening positions using loudspeaker-based panning methods.

The model was firstly applied to predict the direction of a single loudspeaker in a reverberant listening room at 15 different listening positions, using data from the SBSBRIR dataset. Head-rotations at 10° increments were applied to resolve front-back confusions and systematically simulate the dynamic localisation process performed by listeners in a closed-loop localisation task. The model output, $\hat{S}(\theta_{GC})$ at each listening position was presented in Figure. 7.11 where the azimuth scale, θ_{GC} represents the localisation direction of the sound source relative to the specified listening position within the global coordinate system (the listening room). The resultant probability density functions $\hat{S}(\theta_{GC})$ for each listening position indicated a clear, directional estimate of the loudspeaker at 0° from each of the 15 listening positions. The listening position closest to the loudspeaker (x=1, y=0) showed the largest peak value which is likely due to a reduction in the signal to noise ratio and therefore less influence from discrete reflections in the BRIRs. Listening positions at the extremities of the listening area (x=-1, y=-1) and (x=1, y=-1) however showed smaller peak values and increased noise in the peripheral azimuth directions.

For each listening position, the angular direction to the loudspeaker was calculated (relative to a forward facing listener). The argument of the maximum value of $\hat{S}(\theta_{GC})$ was then used to predict the directional judgement of the computational model. This data is shown in Figure. 7.12 where the actual sound source direction and localisation angle predicted by the model are shown to be

extremely close. The mean signed localisation error for the model prediction was 0.81° with a standard deviation in signed localisation error of 0.84°. This indicates that in a reverberant listening environment, the model can localise sound sources comparably to a human listener (Mills, 1958; Makous and Middlebrooks, 1990a; Perrott and Saberi, 1990; Grantham et al., 2003). This highlights the model's resilience to room reflections, similar in the way the human auditory system can localise a sound source within a room despite reflections.

However, the purpose of a computational localisation model is not to calculate the most accurate direction-of-arrival prediction as possible, but to simulate the natural localisation acuity of a human listener. Therefore, it must also be shown that the model can simulate the limitations of human sound localisation and, specifically, that the localisation artefacts introduced by loudspeaker panning methods at non-central listening positions are indicated by the model prediction.

In Chapter. 6, a closed-loop localisation task was performed by human listeners presented with a number of loudspeaker-based panning methods. Both in situ loudspeaker reproduction and a simulation using the AVE described in Chapter. 4 were used to highlight the limitations of using the AVE for simulating localisation cues at non-central listening positions. To understand how well the model can simulate the same localisation cues, the binaural data was used as input to the computational model and the results are compared. Figures. 7.13 and 7.14 show the model output, plotted against the localisation judgements from the subjective test. The test data is shown for both the central (x=0, y=0) and one non-central (x=-0.5, y=-0.5) listening position.

Figures. 7.15 and 7.16 also show angular mean signed localisation error for combinations 1-10 at the central and non-central listening positions respectively. Angular prediction data from the localisation model is also presented alongside the subjective samples for in situ and AVE simulation. Error bars represent 1 standard deviation of the underlying sample of judgements. Samples 3 and 7 at both listening positions were mono sounds sources and it can be seen from both the mean signed error and the corresponding $\hat{S}(\theta_{GC})$ data that the localisation model performed well, with confident predictions of the sound source direction, inline with in situ and AVE results. The mean error between the model's localisation judgement and mean signed error of the in situ data was found to be 7.7° for central listening position and 9.9° for the non-central listening position. It is important to note that these values represent the model's error in prediction of the localisation error. For the panning algorithm and speaker layout combination 9 at the non-central listening position, the mean signed localisation error was -87.9° with the model prediction being -95.9°. This means that, like the AVE simulation, the model detected the distinct localisation artefact where the auditory event collapsed to the nearest loudspeaker.

Looking at $\hat{S}(\theta_{GC})$ data shown in Figures. 7.13 and 7.14, it can be seen that the distribution of the data is more converged at some combinations. Although combination 9 at listening position 1 produced a model prediction within the error bars of the subjective data, the distribution is noisy. The prediction of the model in this scenario is likely very susceptible to localisation error. However, this result compares well with the subjective data where the variance of localisation judgements is much larger than many of the other combinations tested.

To contextualise the results presented in Section. 7.4 it is important to compare the difference in localisation prediction between the model and the subjective data against comparable models from the literature.

Similar computational models with the purpose of predicting the direction-of-arrival in loudspeaker-based reproduction systems have been presented in the literature (Pulkki and Hirvonen, 2005; Park, 2007; Takanen et al., 2014; Wierstorf, 2014; Härmä et al., 2014). Pulkki and Hirvonen (2005) presented an auditory model for the prediction of sound source localisation using amplitude panning methods at the central listening position. The output of the model was in the form of ITD and ILD angles (ITDa/ILDa) calculated across a range of frequencies. The authors reported that the model performed best at low-frequencies and when the intended sound source was near the median plane. For some of the reproduction systems tested, the authors reported that the model had ITD and ILD maximal directional deviations as large as 50° (for 1st order Ambisonic reproduction in the 800 Hz frequency region). For higher-order systems with more loudspeakers the model prediction improved. The authors concluded that the model and subjective results matched generally well. Two models were also tested by Park (2007), one based on characteristic-curve analysis where a nearest neighbour technique was applied to find match ITD and ILD cues to a reference dataset. The second model was based on pattern-matching excitation-inhibition cell activity patterns as an internal representation of localisation cues. These models were evaluated for horizontal sound source localisation in anechoic listening conditions without simulated head-movements. For the localisation of single loudspeakers, the model predictions gave an increased number of front-back confusions compared with subjectively reported localisation judgements. Park also noted that subjective judgements generally underestimated the actual sound source direction whereas the model made over-estimations. For the localisation of virtual sound sources

using an amplitude panning system, comparisons between subjective judgements and the model predictions were reported. Differences between model predictions and subjective judgements were found to be in the region of 5°-10° but the values differed depending on the loudspeaker spacing of the panning system. For amplitude panning using an ITU 5.0 layout, many of the key localisation characteristics were well represented by the auditory models (Park, 2007, p. 203). No aggregate localisation error was provided for the model predictions.

For the localisation of concurrent human speakers, Dietz et al. (2011) presented a computational model which is able to isolate multiple speakers and predict their location. This situation represents a complex localisation task comparable to the localisation of an auditory event at non-central listening positions using loudspeaker-based spatial audio reproduction. The model was found to have localisation error predictions with less than 5° error. This model was later applied and modified for use by Takanen et al. (2014) and Wierstorf (2014) for the evaluation of wave field synthesis and near-field compensated higher-order Ambisonic systems. Takanen et al. (2014) reported the model had a mean absolute deviation in localisation direction of 7° (relative to subjectively reported directions) and noted the relative difficulty of the model to predict low-order Ambisonic systems. Wierstorf (2014) reported an average of 8° deviation (again, to subjectively reported data) with a maximum of 40° using 14 loudspeakers and 7th order NFC-HOA reproduction at a non-central listening position.

Although not investigating direction-of-arrival predictions directly, Conetta (2011) introduces the *QESTRAL* model for the prediction of the *spatial quality* induced by loudspeaker-based spatial audio reproduction systems. The model was optimised using subjective test data and was able to use objective signal

processing metrics to achieve an root-mean-square error of 11.06% for listening at both the central and non-central listening positions.

Although it is difficult to aggregate data from different experiments and analysis methods, it is clear from the results of other model-based predictors of sound source localisation, that the predictions presented in this chapter are comparable and in many cases, are lower than previous models reported in the literature. Unlike many of the comparable models, it has been shown that the model with novel developments presented here can also resolve front-back confusions and localise sound sources within a reverberant environment. The model has also been shown to be capable of inducing the distinct spatial characteristics of loudspeaker-based spatial audio systems at non-central listening positions. The mean difference between localisation error and signed localisation error from subjective data was found to be 7.7° for central listening position and 9.9° for the non-central listening position. Comparing this to similar models reported in the literature, and in consideration of the large localisation errors induced by the combinations chosen in this analysis, the model can be seen as performing well and provides a valid tool for the assessment of localisation artefacts induced by loudspeaker-based panning methods across the domestic listening area. The dynamic movements used in the model resolve front-back confusions and provide a prediction for a closed-loop localisation task.

7.6 Conclusion

This chapter has introduced a development to an existing computational localisation model whereby a closed-loop localisation task can be simulated in a

reverberant environment. The current-standing model was firstly introduced followed by a novel approach to simulate the dynamic localisation process used by humans to resolve front-back and back-front confusions. A process is implemented that re-aligns and averages the head-centric localisation predictions between systematic head rotations giving a new, resolved directional prediction in the room coordinate system. Using BRIRs form the SBSBRIR dataset, the model is then applied to the localisation of a single loudspeaker in a reverberant listening room environment at multiple listening positions across the listening area. The mean signed localisation error for the model prediction was found to be 0.81° with a standard deviation in signed localisation error of 0.84°; comparable to the literature data for minimum audible angle for a frontal sound source. The model was then compared to subjective closed-loop localisation task data from Chapter. 6. The model was found to have a mean deviation in localisation error of 7.7° for central listening position and 9.9° for the non-central listening position. When comparing these values to comparable models for predicting localisation of loudspeaker-based reproduction systems, the model can be considered state-of-the-art and has also shown the ability to account for localisation artefacts caused by time-of-arrival problems when a listener moves across the listening area.

Some logical developments to the model used and developed in this chapter can be defined. Currently the model iterates through a fixed set of head-rotations. However, the use of movement data from humans in a closed-loop localisation task could be implemented to make this process more representative of real-world conditions. Also, translatory movements by the listener could be incorporated to give improved localisation accuracy and even the ability to judge distance. More developed statistical analysis of the resulting sound source localisation

distributions could also be used to achieve improved information about the auditory event such as image-width or stability under head-rotations.

CHAPTER 8

The Perception of Colouration Using a Non-individualised Dynamic Binaural Simulation System

This chapter covers two experiments and accompanying analysis on the validity of using a non-individualised, dynamic binaural synthesis system to simulate colouration artefacts commonly found across the listening area. The colouration detection threshold (CDT) is a psychophysical metric commonly used to define a listener's acuity to changes in sound colour. Here, CDTs are measured for both in situ loudspeakers and auditory events created using the non-individualised dynamic binaural simulation system, using two assessment methods. CDTs are used to define the difference in colouration acuity between in situ and the simulation system.

8.1 Introduction

Following from Chapter 4 and relevant literature on non-individualised binaural simulation presented in Section. 3.6, it is apparent that absolute colouration differences will be audible between the AVE and intended auditory environment. However, it is not well understood whether relative differences in colouration can be accurately judged using a non-individualised AVE. Tests in this chapter aim to answer the following research questions:

1. To what extent is the acuity of colouration perceptually equivalent when listening to the AVE to that of a real auditory environment?
2. Is colouration acuity increased or decreased when using an AVE?
3. How does colouration acuity change over different source directions?

This chapter presents results of two colouration detection threshold experiments using two different non-parametric adaptive psychometric methods. Possible methods for finding perceptual thresholds are firstly discussed in Section 8.1.4. Following this, the psychometric tests are presented in two separate sections. The results and implications are discussed and conclusions are drawn from the results.

Due to the differences in human physiology, the HRTF is highly individualised for each listener. The physiological differences in the pinna geometry, head and torso all change the characteristics of reflections and diffractions of sound when entering the auditory system. Many studies have presented empirical data on the acoustic effect of individual differences, see the work by Møller et al. (1995b) or Middlebrooks (1999b) for examples.

The use of non-individualised binaural simulation will therefore introduce spectral artefacts (colouration) of auditory events when compared to an absolute, in situ, reference. However, to the authors knowledge, no results have yet been published to define whether the introduction of stimulus colouration can be considered a fixed offset for the simulation system. For the evaluation of loudspeaker-based colouration at non-central listening positions, the ability to use a non-individualised binaural simulation presents numerous advantages, primarily not needing to measure separate HRTF or BRIR datasets for each individual listener. In this chapter, the human sensitivity to colouration artefacts (colouration acuity) is measured for both real and simulated loudspeakers using the non-individualised AVE. Colouration artefacts introduced in loudspeaker-based systems are caused by at-ear summation of coherent, delayed sounds from different directions. Therefore, localisation and colouration artefacts are introduced simultaneously. By using individual sound sources with colouration induced prior to the binaural simulation stage, colouration artefacts can be tested in isolation to any localisation artefacts. The use of non-individualised HRTFs in the binaural simulation will cause mismatched inter-aural cues. These interaural cues are used by the human auditory system as part of a binaural decolouration process, which has been shown to decrease colouration acuity (Salomons, 1995; Brüggen, 2001), therefore indicating a reduction in human's acuity to sound colouration. Any differences in colouration acuity due to non-individualised binaural simulation of auditory events will be highlighted by significant changes in colouration acuity for in situ versus the AVE. A Colouration Detection Threshold (CDT) measured on human listener is the result of a psychometric test which can be used to measure human acuity to sound colouration.

Psychometric analysis using colouration, reflection and image-shift thresholds will firstly be discussed to provide a foundation for the experiments presented in this chapter.

8.1.1 Colouration Detection Threshold

One specific type of psychophysical test considers the ability to perceive 'colouration' specifically. This means that the parameter being varied between high colouration and low colouration is the intensity of a comb-filter network and the response parameter is the probability of a listener being able to positively select the coloured signal when presented alongside an uncoloured signal.

The earliest work on colouration detection thresholds introduced computational models of colouration perception by assuming the human ear performs an autocorrelation analysis (Licklider, 1956). This theory takes an auto-correlation of the short-time power spectrum to derive a resonant delay time. A short-time power spectrum model was also later suggested by Atal and Schroeder (1962) who measured CDTs using 8 participants; model thresholds that are still widely used today were defined in both time (B_0) and frequency (A_0) domains. The short-time analysis was implemented by windowing the autocorrelation function and weighting functions were further developed by Bilsen (1968). A modelling method based around central spectrum analysis was reported by Kates (1985) which compared model results to measured CDTs by Atal and Schroeder (1962) in the comb-filter delay range from $T = 0\,ms$ to $T = 40\,ms$. The central spectrum model was also applied to pitch perception by Bilsen (1977). The most recent work in the specific area of CDT modelling is that by Buchholz (2011) in

which subjective testing was performed on 3 participants which also introduced the concept of band-limited CDT measurements. A monaural quantitative model is defined by passing the difference spectrum (internal spectrum) though a number of auditory modelling stages, then predicting the auto-correlation function from the power-spectral density. The use of internal noise defines whether the colouration is perceivable at the threshold. The psychometric function is predicted directly to simulate the subjective psychometric testing.

Recent scientific contributions have shown that diotic and dichotic presentation of coloured signals have different effects on colouration perception. Binaural perception of sound colouration was firstly reported by Koenig et al. (1975) using dichotic JND measurements and later Zurek (1979) also presented work on CDTs with a focus on the binaural suppression effect. Their tests simulated ITDs by delaying one ear signal by 500 μs. They found the ITD increased the CDT slightly (less acuity to colouration). The experiment used only 3 participants, one being the author. However, the most comprehensive study of binaural decolouration effects was reported by Salomons (1995) who conducted subjective tests and developed auditory models to further understand the effect of binaural listening to coloured signals. Salomons (1995) presents subjective results for CDTs measured under a number of stimulus scenarios and also proposed and tested auditory models for the binaural decolouration process. This work also provides an excellent reference for the methodology of implementing CDT tests.

Timbre has been shown as a fundamental parameter of the colouration attribute. Toole and Olive (1988) introduced the timbre detection threshold for a number of different test scenarios. An adjustment method was used to find the DTs of 2 experienced listeners. Similar experiments for timbre were also conducted by

Bech (1995, 1996) using just-noticable-difference (JND) and threshold of detection (TD) methods; 8 and 4 participants were used respectively. The results showed the influence of reflections in a non-anechoic environment on the perception of timbre. The in-situ reverberant environment was found to increase the DTs. The authors reported that participants were able to isolate timbre, level and localisation cues and that level differences were small enough to ensure JND and TD values were only related to timbre.

Results of from the literature show that CDT values are a commonly used tool to measure the human perception to sound colouration. Due to binaural decolouration effects, changes in inter-aural cues are likely to cause a change in the measured CDT values which must be considered when creating auditory events with non-individualised binaural synthesis.

8.1.2 Reflection Detection Thresholds

Reflection detection threshold (RDT) experiments represent another category of psychometric tests which look at the ability of listeners to perceive reflections. Unlike CDT tests, the artefacts of reflections may contain other parameters alongside pitch and timbre such as localisation changes in the auditory event since reflections do not generally arrive from the same direction as the direct sound. Reflection detection thresholds were not measured directly in any of the experiments presented in this thesis, however, the design and results of RDT experiments found in the literature is directly related to both image-shift threshold and colouration detection threshold measurements. Key results and features of previously conducted RDT experiments are presented here to help inform the design of CDT experiments in this chapter.

Lochner and Burger (1958) presented some of the first work on the ability of participants to perceive reflections. Testing was split into two parts: (1) echo (reflection) detection thresholds and (2) measurement of the auditory integration window. A total of 5 participants performed the tests and results highlighted the perception are artefacts when longer reflection delays (up to 100 ms) were used. Sepharim (1961) later developed the work by testing DTs for loudspeakers in an anechoic environment (meaning any binaural decolouration would have been present) and Burgtorf (1961) also conducted extensive experiments measuring the threshold of perception for various sound field conditions. For the application of concert hall perception, Barron (1971) conducted threshold tests for single lateral reflections in an anechoic chamber. Results showed that spatial impression was influenced by the reflections but delays were constrained to the region between 10 ms and 80 ms, larger than would be achieved from the direct sound delays in off-centre loudspeaker-based listening.

Koenig et al. (1975) researched to find the binaural perception of reverberant sound. Diotic and phase-inverted diotic (named dichotic in the publication) signals were presented to the listeners and JNDs in reflection levels were measured. Results showed that listeners were less sensitive (lower acuity) to spectral artefacts for phase-inverted diotic presentation. Work by Olive and Toole (1989) reported on tests for detection thresholds using both reflection and image-shift thresholds under a broad-range of conditions; using both anechoic and reverberant rooms. One of the major conclusions from this study was that listeners were less sensitive to reflections arriving from the same direction and the direct sound than reflections from any other directions.

Reflection detection thresholds were also experimented by Buchholz et al. (2001), where all the parameters affecting RDTs were clearly defined. The work also introduced the room-reflection masked model. Buchholz (2007) later presented subjective results for reflection-masked thresholds measured using a 3-interval, 3-alternative forced choice JND design. 3 participants each with at least 4 hours training per participant were used in the test.

8.1.3 Image-shift Thresholds

In this thesis, validity of the AVE is considered specifically in separate experiments for localisation and colouration. Although interrelated and often caused by the same time-of-arrival differences between coherent sound sources, isolating colouration artefacts by using a CDT methodology allowed for the assessment of colouration acuity. However, the next logical step in the evaluation of the AVE is to consider whether combined colouration and localisation cues are perceived equivalently to in situ reproduction. The image-shift threshold was implemented by Olive and Toole (1988) as the just-noticeable shift in an auditory event's direction or size when the output of a feed-forward filter has spatially separated direct and delayed signals. This measurement will test the AVE's ability to induce correct precedence effect cues at magnitudes close to the just-perceptible threshold. Olive reported that when compared to absolute thresholds, image-shift thresholds were much higher but standard deviations across participants were comparable. Toole (2008) further comments on the use of image-shift thresholds especially when compared to a room's early decay curve (EDC). Image-shift thresholds are measured for a selection of participants in Section. 8.4.

Following a survey of the relevant literature for colouration detection, reflection detection and image-shift thresholds, it can be seen that the use of psychometric tests to strictly define human perception to colouration is common. Due to the implementation of reflections (i.e. delayed coherent signals arriving from a different direction to the direct sound) inducing localisation artefacts alongside colouration, colouration detection thresholds will be implemented in this thesis primarily to limit the scope to colouration acuity. Also, it is noted that many of the studies for detection thresholds use a small number of participants (3-4 is common) despite results showing non-trivial inter-subject differences in thresholds.

8.1.4 CDT Test Methodologies

Many methods to determine colouration detection thresholds have been proposed and implemented. Each have their own strengths and weaknesses and a method is often chosen based on individual needs of the experiment. Three of the most popular methods are presented and discussed below.

Trajectory method - Although less common than other methods, this method shown by Salomons (1995) for CDT measurement has the benefit of reducing the predictability of the threshold test, meaning there is less chance of bias from listeners predicting the test format. The aim is to predict the psychometric function directly: that is, a sigmoid-like function that defines the probability of a listener correctly identifying the coloured stimulus, for all intervals of colouration value (g_{delay}). Predefined intervals of colouration are firstly selected by the experimenter both above and below a predicted CDT value. A listener is then presented with randomised, unlabelled pairs of signals ([uncoloured followed by

uncoloured] and [uncoloured followed by coloured]) at each of the predefined colouration intervals and they must select the pair which contains the coloured signal. Repetitions of this task for each colouration interval gives a course prediction of the psychometric function which can be interpolated to find the CDT value. However, care must be taken when choosing the levels to avoid listener fatigue.

Method of adjustment - The method of adjustment allows for listeners to attempt to find their own threshold. Both coloured and uncoloured signals are presented to the listener and based on their ability to judge the difference, they must select whether to increase or decrease the colouration until they find their threshold. The method is efficient, requiring fewer judgements to be made before achieving the CDT value. However, this method allows for participants to falsely report perceived colouration when in reality they are much lower than their actual CDT value. The experimenter must also choose step sizes carefully, too large and the CDT value may not have enough resolution to be accurately reported. If the step size is too small then it is difficult for a listener to oscillate above and below their threshold easily.

Two-alternative forced choice - Utilising an adaptive test whereby the colouration amount is altered step-by-step, a two-interval two-alternative forced choice test can be used. In this test the listener is presented with four stimuli in two intervals groups; [uncoloured followed by uncoloured] and [uncoloured followed by coloured] and 'tested' on their ability to find the coloured signal. The amount of colouration is changed depending on their answer. For a simple

one-up, one-down test, a correct answer reduces the colouration and an incorrect answer raises the amount of colouration. Starting at maximum colouration, the listener will eventually oscillate above and below their 50% correct threshold and this can be taken as their CDT value. Rules are enforced to determine when the test is ended and how many repeats are performed.

8.2 CDT Experiment A: Adjustment

In CDT experiment A, a method of adjustment was implemented to find the JND for coloured white-noise signals. The indirect-dependent variable of this experiment is the CDT which represents the value of delayed signal level in dB re. the direct signal where colouration is just-detectable. Measurements are made separately for a real loudspeaker and a binaurally simulated loudspeaker using the AVE described in Chapter 4, results are then compared to ensure that colouration acuity is acceptable using the AVE to test colouration perception at off-centre listening positions in domestic, loudspeaker-based spatial audio reproduction.

8.2.1 Method

The method of adjustment has been implemented in numerous detection threshold experiments (Sepharim, 1961; Toole and Olive, 1988; Olive and Toole, 1989; Salomons, 1995; Lindau, 2014). In this experiment, a top-down procedure was used where the first stimulus presented to the participants is of maximum colouration. Before each judgement both reference and coloured signals were replayed to the participant. The reference signal for the test was the original

(uncoloured) white noise signal with a uniformly distributed power density function. For the coloured signal, comb filtering was artificially introduced using a feed-forward comb filter structure. Figure 8.2 shows the block processing for the experiment and Figure. 8.1 shows the graphical user interface used by the participants.

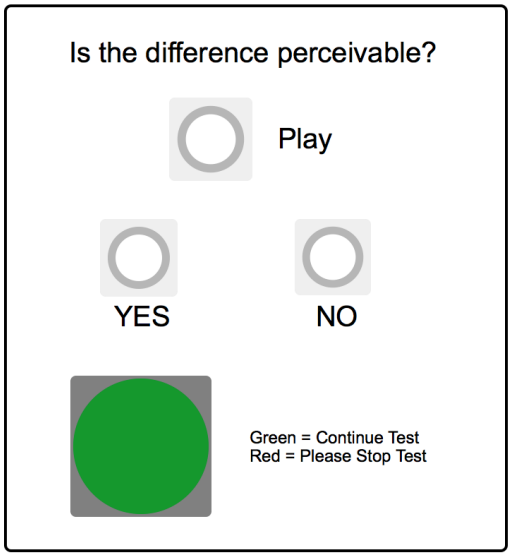


Figure 8.1: The graphical user interface used for Experiment: A. Participants could press play to audition the two audio samples. If the two samples sounded different then the 'YES' response was used. If the two samples were considered the same the 'NO' response was used. The green circle turned to red after the adaptive method had converged.

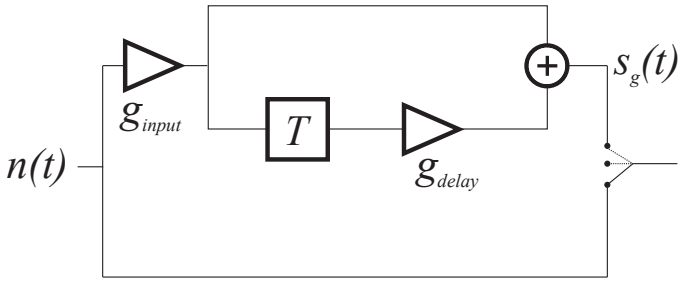


Figure 8.2: Feed-forward comb filter structure with gated output.g_{input} is scaled by g_{delay} to ensure that $s_g(t)$ has the same power as the uncoloured signal. T is repetition delay. The switch selects whether coloured or uncoloured signal is played and also truncates the delayed signal offset (see Figure 8.4)

g_{input} can be calculated using Equation 8.1 (Salomons, 1995).

$$g_{input} = \frac{1}{\sqrt{1 + g_{delay}^2}} \tag{8.1}$$

The shape of the spectral magnitude response of the comb filter and therefore the nature of the colouration is dependent on the delay-time of the delayed signal path, T. The purpose of this study is for the application of perceptual colouration tests in domestic spatial audio systems across the listening area and therefore, AVE colouration acuity must be equivalent for delays representative of this scenario. A simple analysis can be conducted to show the maximum delay between nearest and furthest loudspeakers in a domestic listening area to further understand the region of delays needed as shown in Figure. 8.3.

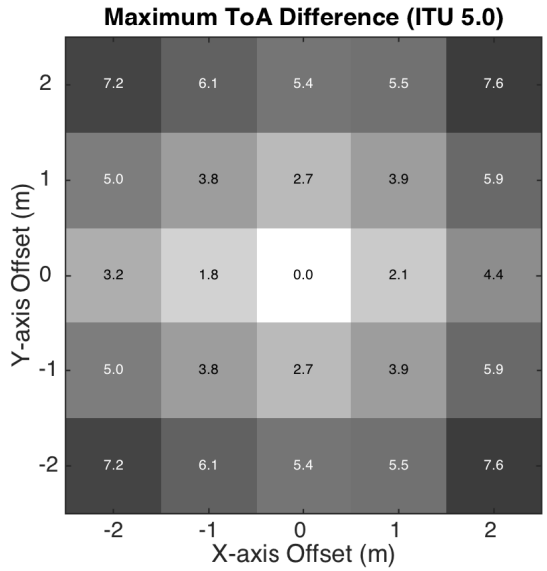


Figure 8.3: Maximum delays across the listening area for ITU 5.0 speaker layout. Numbers show time-of-arrival delay between nearest and furthest loudspeaker in *milliseconds* for each listening position. Loudspeaker radius is 2.1 m. Speed of sound $c = 343ms^{-1}$.

Considering the delays presented above, T was made constant at 2 ms for CDT experiment A. It has been shown in numerous studies that CDT values are dependant on the delay constant, T (see Salomons (1995) or Buchholz (2007) for data on CDT as a function of T). However, the practical difference for the purpose of comparing CDT values between AVE and in situ is trivial. Values of T between 0 ms and 10 ms are representative of the comb-filtering induced at off-centre listening positions.

As seen in Figure 8.4, the repetition 'offset' is the remaining region of the delayed signal. Some experimental procedures have chosen to keep this region in tact (Buchholz, 2007), others have truncated it to the end of the direct signal (Zurek, 1979). However, this subtlety has been shown to influence localisation in rooms (Litovsky et al., 1999) and if the offset provides additionally perceived artefacts it could lower the measured CDT values relative to CDTs measured with the

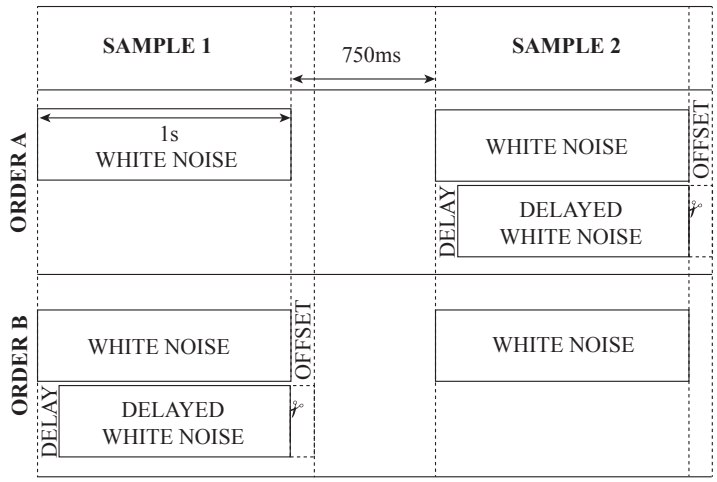


Figure 8.4: Order A or B was randomly chosen for each step. Time is not to scale. Repetition signal offset caused by delay was truncated.

offset silenced. For CDT experiment A, offsets were truncated by implementing a switch on the *output* as shown in Figure 8.2. Although this is not representative of room reflections or delays caused by loudspeakers, the practical impact of this choice is likely only to cause an equivalent offset in the measured CDT values, making the comparison of CDTs between in situ and the AVE still valid.

Looking at the effect of this feed-forward filter in the frequency domain highlights linearly spaced notches with equal magnitude across the full frequency range. Notch frequencies can be calculated using Equation. 8.2 where f_i is the frequency for notch integer i and T is the repetition delay.

$$f_i = \frac{2i - 1}{2T} \tag{8.2}$$

However, due to the specific spacing and bandwidth of auditory filters, narrow notches cannot be resolved in the upper-frequency bands. Passing the filtering effect of the harmonic cosine noise generator through a ERB spaced gamma-tone

filter bank using the Auditory Modelling Toolbox (Søndergaard et al., 2011), Figure 8.5 shows clear notches, which decrease in magnitude in the upper frequency regions due to the filter spacing and bandwidth. The plot also shows that as the repetition delay T is increased, the fundamental notch frequency is reduced. The magnitude of the notch depths is proportional to the repetition gain g_{delay}. T values less than around 50 ms have been shown to induce colouration artifacts (Bilsen and Ritsma, 1970).

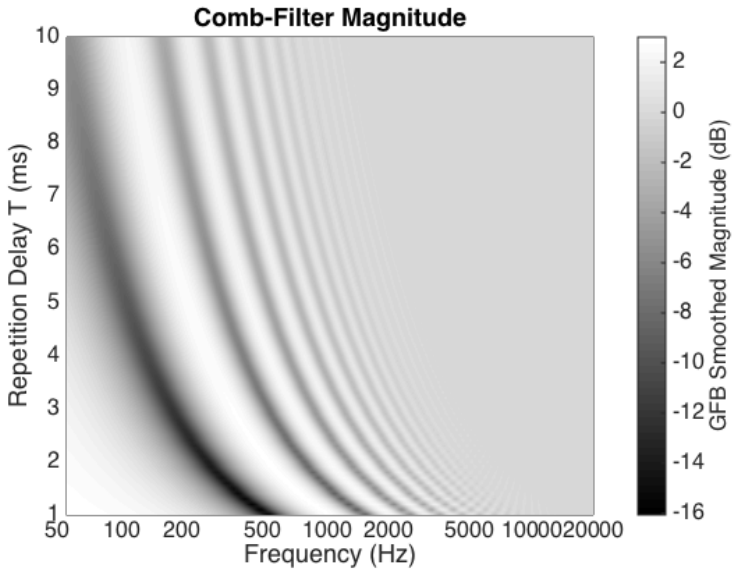


Figure 8.5: Magnitude spectrum of the comb-filtering effect caused by different repetition delays (T) across the frequency range. The spectrum is smoothed by a Gamma-tone filter bank, using 100 filters per Equivalent Rectangular Bandwidth (ERB). Outputs of each filter are scaled relative to the response of the filter to a Dirac; thereby removing the global frequency weighting and instead providing perceptually-motivated smoothing. $g_{delay} = 0\,dB$.

A test using the top-down, adjustment method (Salomons, 1995) was implemented to allow participants to find their own colouration detection thresholds. The test was carried out in the University of Salford BS.1116-1 conforming listening room. Digital signal processing and the adaptive test procedure was all performed in realtime using software.

Some important parameters of the adaptive testing procedure are defined in Table. 8.1.

Table 8.1: Description of parameters used in the adaptive testing procedures for colouration detection thresholds.

Parameter	Description	Symbol
Trial	each decision made by the participant	n
Run	a set of trials in a consistent direction either up or down	-
Reversal	a trial that changes the direction and therefore divides runs	R
Level change	a trial which causes a change in g_{delay}	-
Step size	the size of the change in g_{delay}	-
Convergent region	the sample of trials selected to calculate CDT and σCDT	

Upon the 'PLAY' button being pressed two noise signals were played; one uncoloured reference white noise signal and one coloured white noise signal with switching highlighted in Figure. 8.2. Noise signals had duration of 1 s with a 750 ms silence in-between. The order of reference/coloured was randomised for each play and the participants were informed of this. The ordering of stimuli can be seen in Figure. 8.4. If a difference in colouration was perceived (answer YES) the amount of colouration was decreased in the next trial if no colouration is perceived (answer NO) the amount of colouration was increased. The aim of the adaptive procedure is for the participants to converge on their just-noticeable threshold. Following the initial run and reversal, participants continued to move above and below their threshold until a green light on the GUI became red, this indicated 15 reversals had been made and the session was ended. The initial step size of decreased colouration was randomly chosen from between 4 dB to 8 dB with 1 dB resolution and this intial step size was not revealed to the participant to avoid predictability bias. The step size was halved after each reversal until the lowest step of 1 dB was achieved.

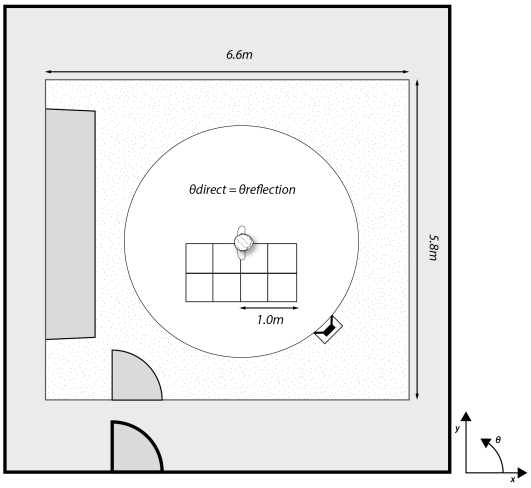


Figure 8.6: Layout of CDT Experiment A. A single loudspeaker was simulated using the AVE where $\theta_{direct} = \theta_{reflection} = 315°$ i.e. the same loudspeaker used for both direct and reflected signals.

All signals were replayed from a single loudspeaker at 315° (front and right of the listener) relative to the listener seated in the central listening position (X=0, Y=0) as shown in Figure. 8.6. Due to the summation of direct and delayed signals as shown in Figure. 8.2 the direction of these two signals were the same therefore $\theta_{direct} = \theta_{reflection} = 315°$ [1].

For the in situ scenario the loudspeaker was real, for the AVE scenario the loudspeaker was simulated. A loudspeaker direction of 0° was specifically avoided due to anecdotal perceptions of front-back confusions if no head-movements were made. Participants were allowed and instructed to move naturally during the test. Many participants were unaware the the AVE was actually a simulation and not the real loudspeaker.

[1] In this part of the experiment 'direct' and 'reflected' signal directions are equal yet they are defined separately to improve clarity for later tests where the directions are not equal.

Participants were given a training session before the test allowing them to audition the interface and hear the effect of their decisions. Experienced participants were a prerequisite for this test due to the reliance on finding their own threshold; this was assessed by a pre-test questionnaire. Each participant undertook two threshold tests for both auralisation methods giving a total of four colouration detection threshold values per participant. The order of auralisation method presented to the participant was randomised between participants in either AABB or BBAA sequence.

The colouration detection threshold is the g_{delay} magnitude that the adaptive test procedure converges upon. The first run of the adaptive test is a 'coarse' alignment of the colouration. After four reversals the function is nearing the convergence value and the participant is likely to be moving above and below their perceivable threshold for colouration. Therefore, CDT is calculated by taking the mean g_{delay} at each trial between the 4th reversal and the 15th reversals (the test was stopped after the 15th reversal) as shown in Equation. 8.3.

$$CDT = \frac{1}{N} \sum_{n(R=5)}^{n(R=15)} g_{delay} \tag{8.3}$$

Where g_{delay} is the magnitude of the delayed signal path specified in dB relative to the direct path. n is the trial number and R the number of repetitions therefore $n(R = 5)$ is the trial number at the 5th reversal. N is the number of trials between $n(R = 4) + 1$ and $n(R = 15)$.

Another important parameter in the adaptive testing procedure is the standard deviation of the assumed converged region of responses defined to calculate the mean value over. This provides information on how stable the convergence was;

large standard deviation means that runs were larger where as smaller standard deviation values mean small runs and therefore more confident responses from the participant. Similar to Equation. 8.3, the standard deviation can be calculated using Equation. 8.4.

$$\sigma CDT = \left(\frac{1}{N-1} \sum_{n(R=5)}^{n(R=15)} (g_{delay} - CDT)^2 \right)^{\frac{1}{2}} \tag{8.4}$$

8.2.2 Results

A total of 6 experienced listeners from the University of Salford undertook the experiment all of which had used the AVE in different tests at the university. For each detection threshold measured, the final data point was taken as the average over two repeats. The complete adaptive test results are shown in Figure. 8.7.

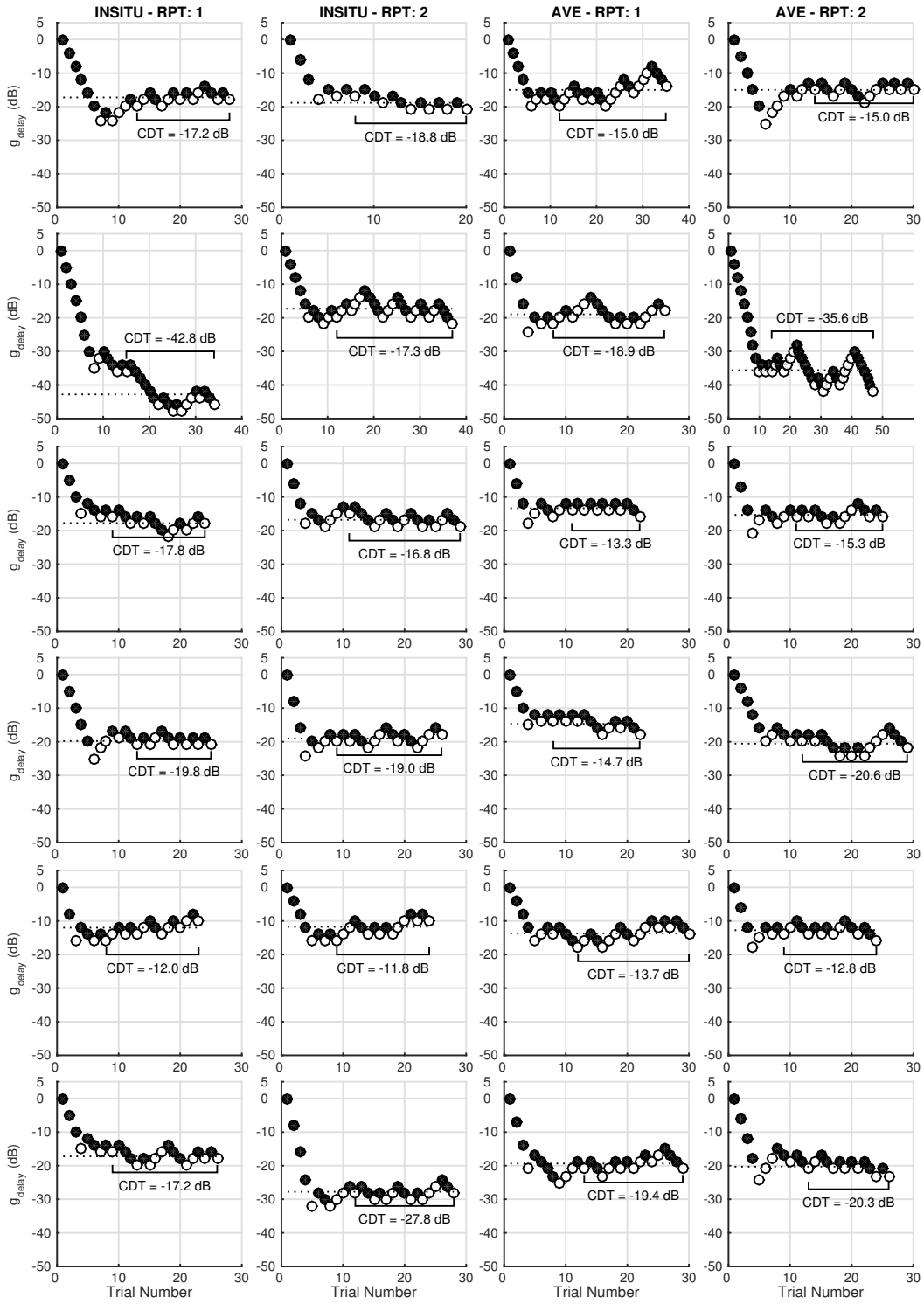


Figure 8.7: Adaptive colouration detection threshold test results for experiment: A. Each row is a participant in the test and each column shows the different conditions. Squared brackets show the region where the CDT value is calculated and the dotted line intercepting the y-axis indicates the CDT value. Filled markers = YES response, hollow makers = NO response.

From the results in Figure. 8.7 it is now possible to show the CDT values for each participant and auralisation method. CDT values are averaged over repeats for each participant.

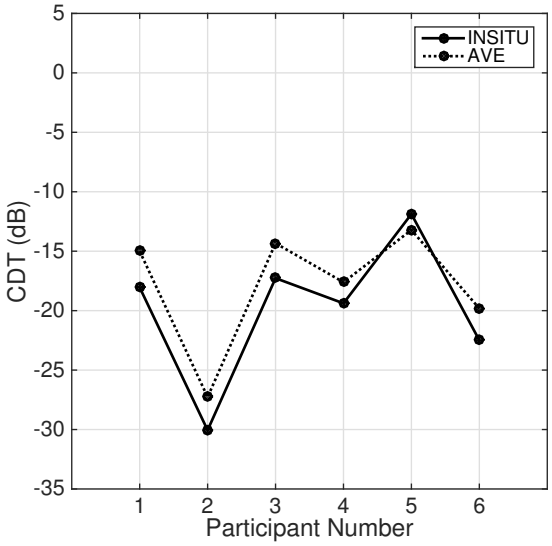


Figure 8.8: Colouration detection thresholds for in situ and AVE simulation. Results are shown for each participant independently.

Now the standard deviation of the convergence region is shown for each session by each participant in Figure. 8.9.

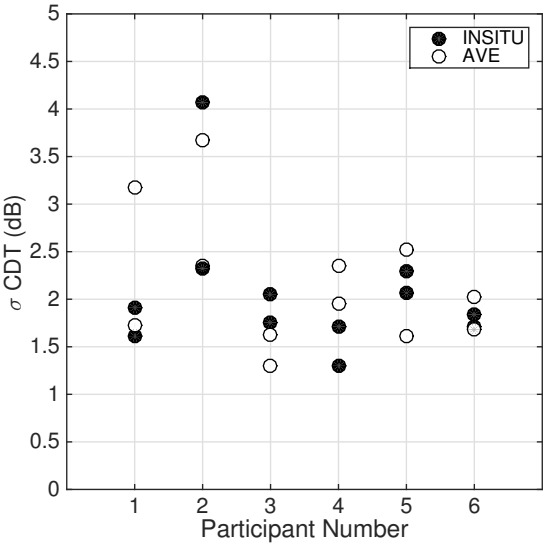


Figure 8.9: μCDT values for in situ and AVE simulation. This data shows the stability of the convergence of the CDT approximation method.

8.2.3 Discussion

The raw response data of the adaptive test procedure shown in Figure. 8.7 show that the method achieved good convergence towards each participant's colouration detection threshold. This is highlighted by a consistent increase towards the first reversal and then a consistent movement above and below their threshold. It can be seen in these results however that there is some inter-subject variation in the standard deviation of the sample used to calculated the CDT value highlighted by the movements above and below the threshold having varying size. This is particularly noticeable for participant number 2, in the in situ, repeat 1 and AVE, repeat 2 sessions.

Another noticeable feature of the response data is the size of the run following the first reversal. For some of the session this run was larger than would be expected by considering σCDT for the same session. One explanation for this is that when participants are approaching their CDT, they have an explicit

reference for the degradation and therefore know 'what to listen for'. When going below their threshold and then returning above it after their first reversal the explicit reference is no longer audible and the listener is relying either on an implicit reference or require a larger amount of colouration for them to achieve their second reversal. However, this feature should not affect the results CDT due to the mean value being calculated following 4 reversals.

The resultant CDT values shown in Figure 8.8 show very positive results for the performance of the AVE. Firstly, inter-subject variation in CDT appears to be large with a range of 18.2 dB when in situ and 15.1 dB for the AVE. This range of values is not surprising when compared with previous literature values such as Salomons (1995), where results show an inter-subject range of approximately 11 dB for $T = 4\,ms$, also implementing the method of adjustment. Results from Olive and Toole (1989) also indicate inter-subject threshold ranges of approximately 10 dB, again using the method of adjustment. However, the slightly larger range of measured CDT values than those found in the literature could have been a result of the participant training or general experience of the listeners.

For each participant, the AVE CDT matches the in situ CDT with low error especially when compared to large inter-subject variation. This indicates that differences between participants CDT values are well represented using the AVE. On average the AVE CDT is 2.0 dB higher than the in situ CDT indicating that the AVE induced lower colouration acuity.

Measuring the variation in g_{delay} for each trial over the region used to calculated CDT was used to consider whether participants found it harder to converge upon their CDT when using an AVE as opposed to in situ. Figure. 8.9 shows σCDT

for each session by each participant of the trial. The data shows that the AVE σCDT values were not significantly higher or lower than with in situ auralisation. The variance in values shown by participant 2 were also well represented by the AVE.

Although the data present in this section highlights that the AVE can simulate colouration differences well it was found that the method of adjustment relied heavily on the participants' abilities to find their own thresholds; this problem has been highlighted in previous work by Salomons (1995) who indicated that CDT values may be lower than thresholds measured with more statistically robust methods. Also, the sample size of participants in the test may not fully reflect the subjectivity of CDT responses.
Because of this, a second experiment was implemented to build upon and extend CDT measurements in this section to support the research question of measuring the extent of perceptual equivalence between the AVE and that of a real auditory environment.

8.3 CDT Experiment B: 2AFC

To improve on the reliability of convergence of the CDT and also include more loudspeaker directions, CDT experiment B was designed using a 2-interval, 2-alternative force choice design with a two-down, one-up adaptive procedure. The concept of this procedure will be explained in the following sections.

No literature is currently available for the change in CDT as the sound source is positioned at different angles across the horizontal plane around the listener. Due to the localisation errors introduced at off-centre listening positions in

loudspeaker-based spatial audio, it is important to understand the nature of colouration acuity at multiple sound source directions. CDTs were again measured for both in situ and AVE scenarios and results were compared using statistical testing to present the equivalence boundaries of colouration detection using the AVE.

In a small-scale additional study, colouration and localisation artefacts were combined to measure image-shift thresholds for three participants as a preliminary study which could be developed upon in future work.

8.3.1 Method

Whereas the method of adjustment approximates the amount of colouration that is 'just detectable' a 2AFC method can be implemented in specific ways to converge on an exact point on the psychometric function by using transformed up-down scaling (Levitt, 1971).

Levitt (1971) has shown that the desired ordinate of a perceptual threshold can be predefined by the design of the test. This value is represented by X_p where p is the percentage of positive responses.

For this experiment, $X_{70.7}$ was defined by using a (transformed) 2-down, 1-up design (two correct answers are needed to reduce the colouration amount, 1 incorrect answer will increase the colouration amount). Using a transformed method means that once converged, the amount of colouration is higher and therefore less tiring for the listener. A simple example of a 1-up, 1-down psychometric function is shown in Figure. 8.10. In this example the number of correct responses will increase as the stimulus level (amount of colouration in a CDT test) also increases.

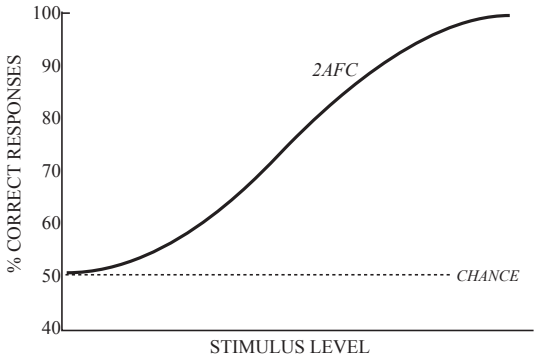


Figure 8.10: An example psychometric function for a 1-up, 1-down 2AFC test.

Similar to CDT experiment A, the resultant CDT value from each session was calculated by estimating the value upon which the function converges. A sample of g_{delay} values was taken following the 4^{th} reversal and the 15^{th} level change after the 4^{th} reversal; this value also signified the end of the session. The initial step size was set to 4 dB, which was reduced to 2 dB after the 1^{st}reversal. This change in step size helped to reduce the session duration by increasing the initial adjustment towards the final convergent CDT value. A failsafe was also implemented so that if 25 level changes were made before the 4^{th} reversal the session was stopped and flagged as a failed session. Two sessions for each loudspeaker direction, auralisation method and participant were measured.

Signal Processing

The delay was fixed at $T = 5$ ms. As shown in Figure. 8.3, maximum delays at a single point in space can range from 0 ms at the central listening position up to around 10 ms at the extremities for an ITU layout. CDT experiment A used $T = 2\,ms$ therefore a slightly increased delay was used in CDT experiment B, still within the desired range yet inducing slightly different comb-filter characteristics. CDT values have been shown to be dependent on delay time and

therefore absolute values are not comparable between CDT experiments A and B. However, if equivalence can be shown between in situ and AVE reproduction for two different delays this will support the use of AVE for testing colouration across the listening area.

The comb filter design is shown in Figure. 8.11. The stimulus signal $n(t)$ was white noise with a uniform amplitude distribution. It should also be noted in Figure. 8.11 that the position of the switch moved from the output (as in CDT experiment A) to the input. This meant that delay-signal offsets were maintained, which has been shown to lower measured CDT values (Buchholz, 2007). This factor is also highlighted by looking at the signal envelopes shown in Figure. 8.13.

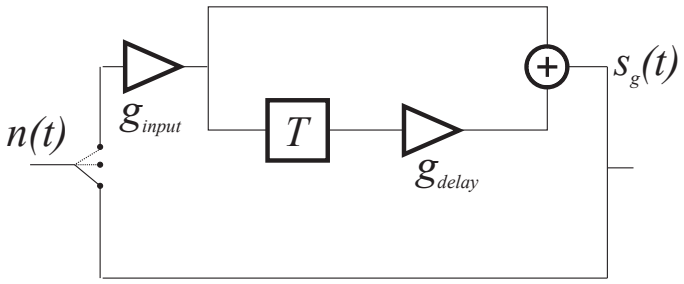


Figure 8.11: Feed-forward comb filter structure used in the 2AFC colouration experiment. Notation remains the same as Figure. 8.2.

An important addition to this test is the inclusion of θ_{LS} as a discrete independent variable, which defines the direction of the reproduced and simulated test signals. Due to the HRTF function being highly direction dependent, changes in colouration acuity across signal direction is an outcome that is likely to vary between individuals. The coloured signal can be separated into two parts: direct and reflection. Direct represents the un-delayed element of the comb-filter block processing and 'reflection' represents the delayed element of the signal processing. For this experiment the direction of the direct and

reflection element were always kept the same and their directions in the horizontal plane only are referred to as θ_{direct} and $\theta_{reflection}$ and the combined direction of the coloured signal is referred to as θ_{LS} where $\theta_{LS} \equiv \theta_{direct} \equiv \theta_{reflection}$. Table. 8.2 documents the directions used in the test.

Table 8.2: θ_{LS} directions used in CDT measurements for experiment B.

Loudspeaker Index	θ_{LS}
1	0°
2	45°
3	90°
4	135°
5	180°

Procedure

The listeners were centrally seated in the listening area and loudspeakers were hidden behind a white acoustically transparency curtain. Figure. 8.12 shows the experimental layout for the test. Loudspeakers were calibrated with the uncoloured stimulus signal to 62 dBA using an A-weighted SPL meter at the central listening position and headphone volume was fixed to give equivalent loudness to the loudspeakers.

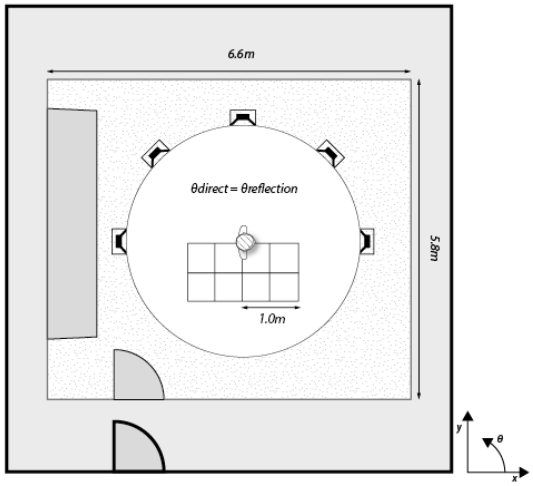


Figure 8.12: Experimental layout for CDT test with multiple θ_{LS} directions.

The experiment was split into two halves each lasting approximately 45 minutes. Each participant either performed AVE or in situ in the first half and the remaining system in the second half to avoid any presentation order bias and to ensure listeners were never required to listen to in situ loudspeakers whilst wearing headphones. Between halves the participants were allowed to take a 10 minute break if they wanted it.

One of two presentation orders was chosen randomly for each trial. These are shown schematically in Figure. 8.13 with the signal envelopes. As discussed for the previous CDT experiment, the effect of reflection onset and offset will likely caused a fixed change in measured CDT values for this experiment, due to the listener having more cues to infer that a sample is 'coloured'. However, the fixed change is not likely to affect the measurement of equivalence between in situ and AVE, but should be noted when comparing data to other studies from the literature.

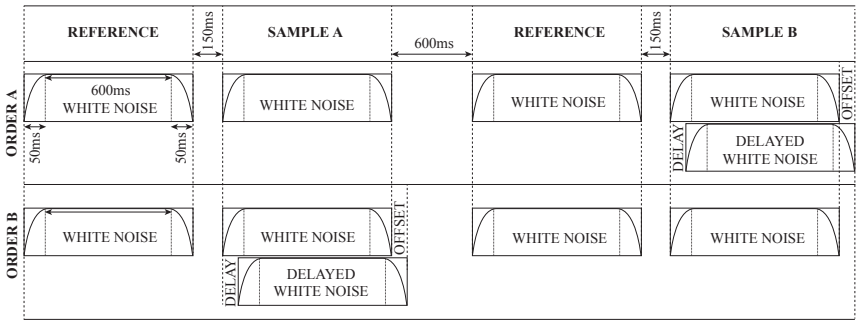


Figure 8.13: Presentation order for each trial. Order A or Order B was chosen randomly. Comb-filter offsets were left in tact; an important note which has been discussed as a reason for a lowered DT by Buchholz (2007). Note the x-axis is not to scale.

A graphical user interface was designed for the method which allowed the participants to run the experiment themselves. The experimenter explained the purpose of the experiment to the participant and gave an overview of the task. Before the test proper, the participant was allowed to begin one session in situ until they felt comfortable with procedure. Before each session began, a start page was shown to the participants. A button labelled 'START' on this test began the session and started the audio playback automatically, where the stimuli were presented in the order REF, A, REF, B and the coloured signal was randomly applied to either A or B. The direct task of the participant was to identify the coloured signal by clicking the button labelled A or B. Arrow keys on the keyboard could also be used to make the selection. Following selection the next trial began automatically and this process was repeated until the test was complete. The main trial GUI is shown in Figure. 8.14.

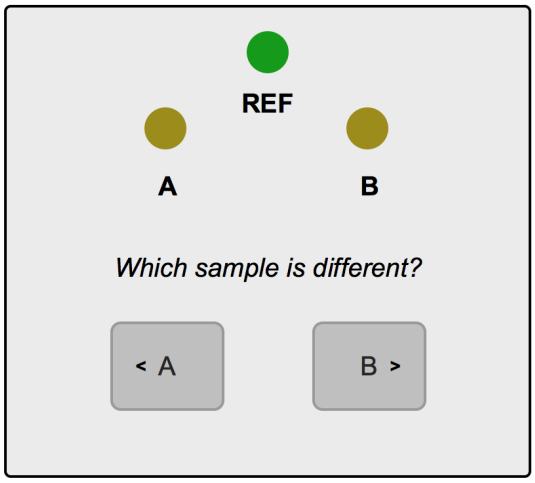


Figure 8.14: The graphical user interface used in Experiment: B. The stimuli played sequentially (REF, A, REF, B) and the LEDs indicated which sound was currently playing.

Participants

11 male participants from the University of Salford Acoustic Research Centre were used in the experiment. All reported normal hearing and normal (corrected or uncorrected) vision. All participants worked or studied in the field of audio/acoustics and reported experience in audio-related used studies and were remunerated for their time. During the initial training participants were given a definition of the term 'sound colour' from Salomons (1995).

The colour 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 therefore comprises timbre, rhythm and pitch.

In a pre-test questionnaire, participants were also asked the question 'Do you feel you understand the term 'colouration'?'. All participants answered 'yes' unnanimously.

Head-width measurements were also made on all participants as this has been

shown to correlate well with maximum ITD Middlebrooks (1999b). No ITD scaling was used in the non-individualised AVE and therefore large differences in participants' ITD compared with the dummy head's ITD could introduce additional distortion to the simulation, mainly for localisation cues but also for changes in static sound positions as the listener rotates their head. Head-width (J_w) measurements were taken as the distances between a reference point just in front of the tragus (defined by the condyle of the mandible) using a calliper. The mean measured distances was $\mu J_w = 137.6\,mm$ with a standard deviation across participants of $\sigma J_w = 5.8\,mm$. Comparing with the mean $\mu J_w = 134\,mm$ and standard deviation $\sigma J_w = 8\,mm$ from Middlebrooks (1999b) (from 33 subjects) shows that participants of this experiment may have had, on average, a slightly increased maximum ITD with a smaller inter-subject variation but the similar results for mean and standard deviation indicate a normal sampling of participants based on the variation in maximum ITDs.

8.3.2 Equivalence Testing

For the same justification as the use of a two one-side equivalence test in Chapter. 6, this statistical test is applied here to understand the practical equivalence of CDT values measured in situ or using the AVE. In CDT experiment B, equivalence tests were implemented.

To compute equivalence samples between AVE and in situ, CDT results for AVE were subtracted from CDT results for in situ for each participant and θ_{LS} independently. This gave a new sample for each θ_{LS}. From the difference samples, non-parametric confidence intervals were calculated by bootstrap resampling (N=1000). The mean and 90% confidence intervals for each sample

were then calculated. These results not only highlight any systematic changes in CDT using the AVE but also describe the magnitude of equivalence that can be expected by defining a region which contains all upper and lower confidence intervals. Any samples with 90% CIs that do not overlap the 0 dB region indicate that the difference is largely systematic in either direction and 90% confidence intervals that do not fall within an equivalence boundary are not equivalent to a 0.05 significance level[2]. For statistical tests of equivalence, boundaries are commonly defined prior to testing to have a target range within which the difference samples can lie (based on perceptual limens). For this study, the equivalence boundary is defined following the test results by defining the $\pm\Delta CDT$ values that encompass all confidence intervals of the ΔCDT for each θ_{LS}. The physical effect of the change in CDT caused by the AVE will then be considered with practical examples. The effect of the AVE on CDT should also be compared to inter- and intra-subjective variations in CDT measurements. Future researchers wishing to implement similar binaural simulation systems to colouration testing can also apply their own boundaries based on their needs for equivalent colouration acuity.

8.3.3 Results

Due to the large number of sessions for CDT experiment B (2x repeats, 5x θ_{LS}, 11x participants for both AVE and in situ = 220 sessions), individual participant session data for each scenario is shown in Appendix. B.

Non-individualised binaural simulation has been shown in the literature to induce absolute colouration differences which are also likely change depending on head-azimuth relative to the intended auditory event. To assess the influence of

[2]It is important to note that because the TOST performs to t-tests at either direction, 90% CIs equate to a significance test at the $\alpha = 0.05$ level.

colouration differences introduced by significant head movements, head-orientation data was recorded for all AVE trials for all sessions. Participants were asked to move naturally, but maintain their head direction facing forward which was also indirectly imposed by the use of the graphical user interface. Using the tracked headphones it was possible to construct a vector pointing in the direction of the listeners' forward-facing head with a base at the centre of the head, Figure. 8.15 shows the direction of the head-point vector for each trial of the AVE CDT tests.

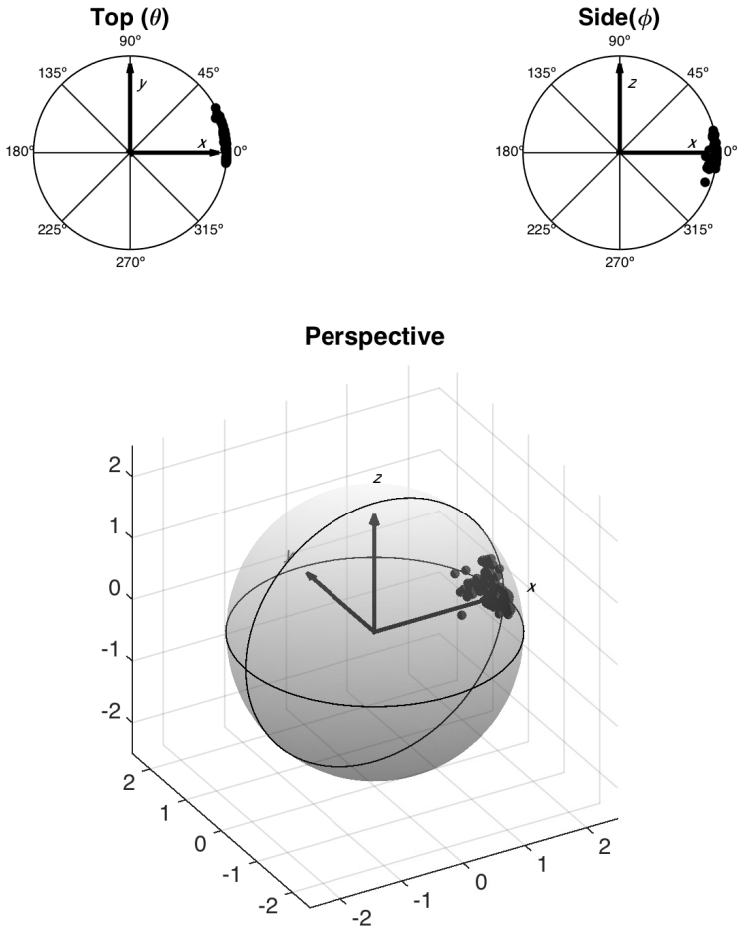


Figure 8.15: 3-dimensional head-point vector directions for every AVE trial of CDT experiment B. Three different views show top, side and a perspective view of the data which shows clustering of data around central head-point but a slight shift in + azimuth direction when looking at the 'top' view. Each marker represents one trial direction.

Next, calculated CDT values for each session are shown.

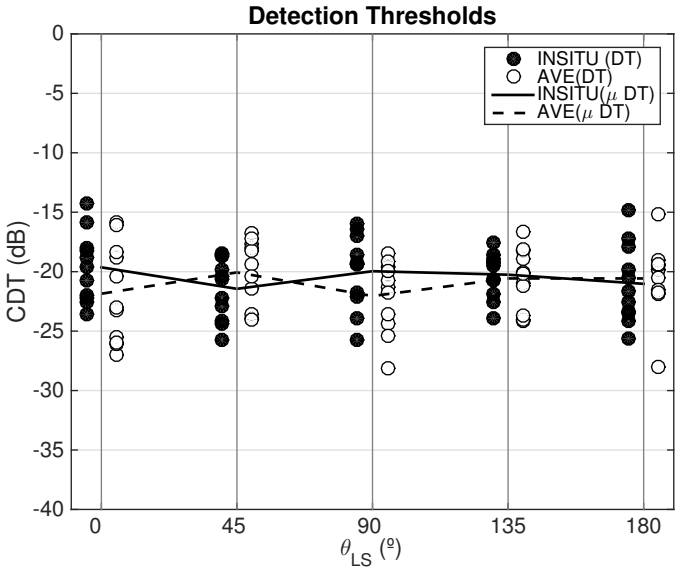


Figure 8.16: Measured CDT values for each session of CDT experiment B. Markers represent measured values for AVE and in situ independently. Lines show the mean values μDT for each θ_{LS}.

It is important to also consider the standard deviation of g_{delay} values over the convergent region of each session. This helps to understand how well the psychometric function had converged and if there was any difference between the AVE and in situ sessions.

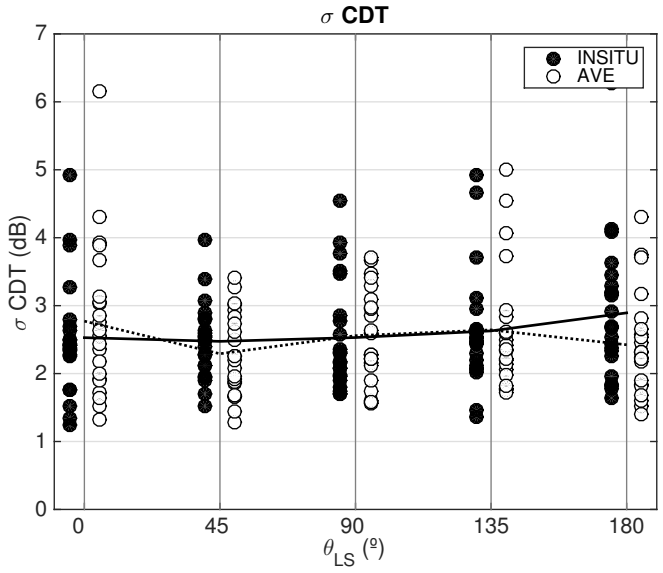


Figure 8.17: Standard deviation for CDT values for each session of CDT experiment B. Markers represent measured values for AVE and in situ independently. Lines show the mean values of σCDT for each θ_{LS}.

Equivalence Testing

Figure. 8.16 shows that results may be comparable but any intra-subject systematic increase or decrease in CTD values may be hidden due to between-subjective variation. Figure 8.18 shows data on the equivalence of CDT values measured in situ and using the AVE. It is important to note that the y-axis represents the change in g_{delay} at the perceptual threshold, which causes a much smaller difference in actual coloured sound stimuli. See Figure. 8.22 for a graph showing the effect of a difference of $\pm 4\,dB$ in g_{delay}.

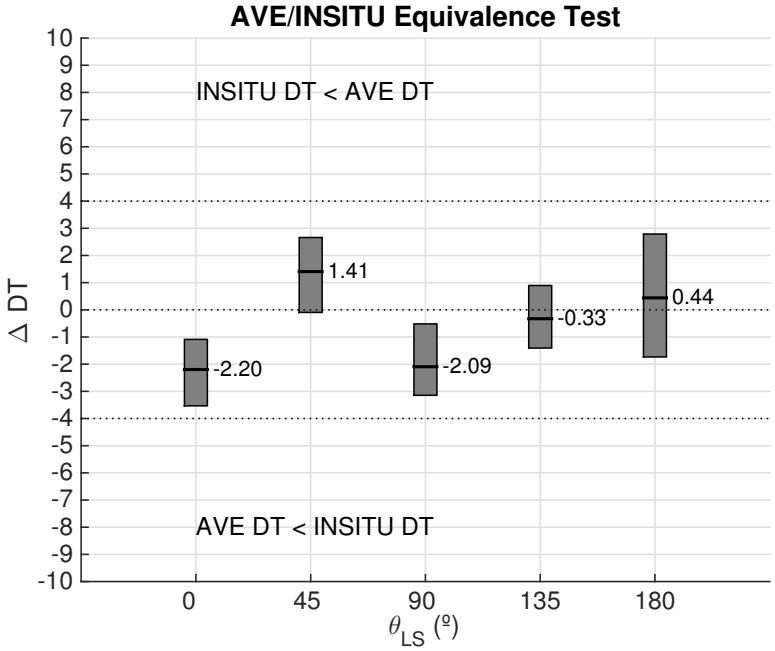


Figure 8.18: Results for RDT equivalence testing between AVE and in situ reproduction. Bootstrap resamples ($N = 1000$) were used to define the mean and 90% CIs highlighted by the boxes.

An additional parameter of the study is the standard deviation in g_{delay} calculated across the same trials that CDT is calculated. This provides information on the stability of the convergence. If participants found the AVE CDT more difficult to converge upon, variations between reversals are likely to be larger and therefore an increase in the average standard deviation would be recorded. Again, an equivalence sample was calculated by taking the difference in the standard deviation (SD) between participants. Results are shown in Figure. 8.19.

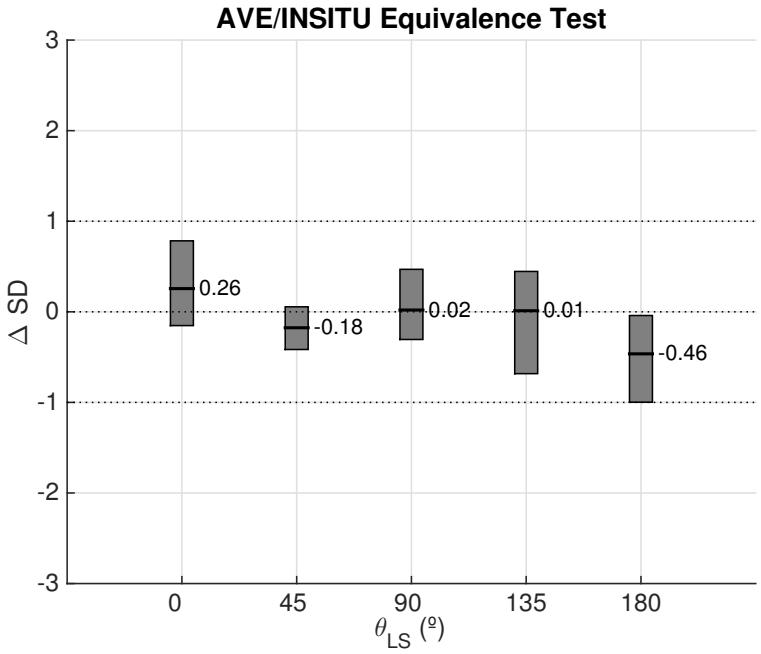


Figure 8.19: Results for SD equivalence which indicates the stability of convergence on the CDT value.

The time taken between playing the audio trial and the participant making a decision was also recorded. Figure. 8.20 shows a box-plot for all time-of-judgement values.

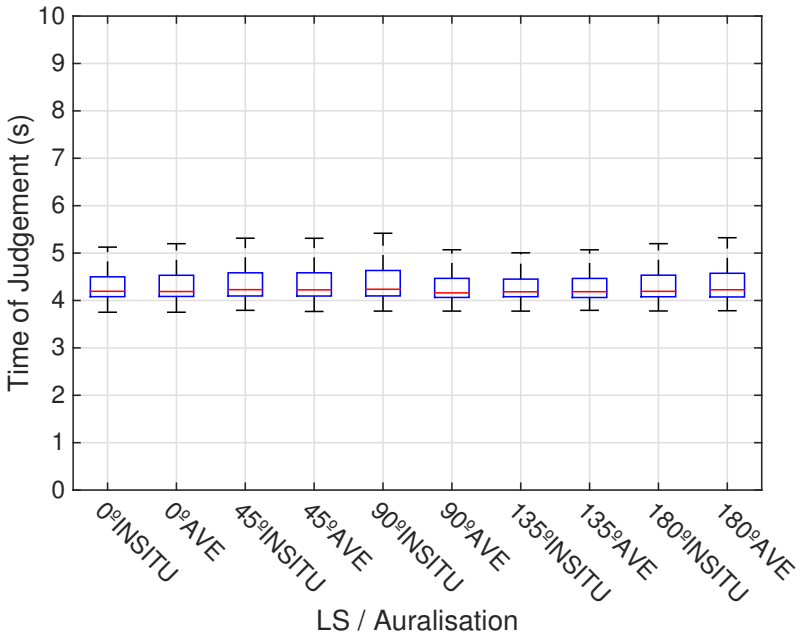


Figure 8.20: Time taken by participants between the start of the stimuli and reporting their judgement.

8.3.4 Discussion

In CDT experiment B, a top-down, 2-alternative forced choice method was implemented to measure the equivalence of CDTs between in situ loudspeakers and loudspeakers simulated using a non-individualised AVE. Five loudspeakers were used in the study to also include the variation in CDT equivalence across source azimuth. θ_{direct} and $\theta_{reflection}$ were kept equal so that only colouration differences were used by the listener to respond to the comparisons.

Due to the HRTF being a function of head-azimuth, it is likely that the use of non-individualised binaural simulation will induce colouration artefacts that change as the user rotates their head. Resonant notches in the HRTF, caused by pinna reflections have been shown to track across head-rotation (Xie, 2013). If the notches present in the binaural simulation are largely mismatched to the

listener then colouration will be perceived. Head-tracking implemented in the study allowed for natural dynamic interaction of the user and the virtual auditory environment but to avoid colouration changes affecting CDT values, the listener should predominently face forwards during audition. Although the graphical user interface loosely enforces listeners to face forward, head-point vectors were recorded for each judgement made by each listener of CDT experiment B as a post-hoc validation. Figure. 8.15 shows that head-point vectors were mainly grouped around the forward direction ($+x$, $\theta = 0°$, $\phi = 0°$).

Raw CDT values are firstly displayed in Figure. 8.16 which shows that inter-subject variation in measured CDTs were non-trivial. The mean CDT value, μCDT, is taken for each θ_{LS} for in situ and simulation using the AVE and therefore does not directly account for inter-subject variations in CDTs. Considering the data in this way, the difference in μCDT is small between AVE and in situ loudspeakers. The maximum differences occurs at $\theta_{LS} = 0°$ where μCDT is 2.2 dB higher than the AVE.

Results for the standard deviation of g_{delay} are shown in Figure. 8.17. Although the inter-subject variation in σCDT is larger than the measured CDT values, the average σCDT is, like CDT values, very similar between AVE and in situ. The data does highlight some outliers however, which may indicate sessions where participants moved too far above or below their CDT value and found it difficult to realign. These outliers occur in both AVE and in situ auralisation methods.

One interesting assumption imposed by Levitt (1971) for the use of up-down testing procedures states:

'Responses obtained from the observer are independent of each other and of the preceding stimuli'.

However, large outliers shown in σCDT could result from adaptive procedure whereby the amount of colouration is often increasing or decreasing. As a specific example, consider the raw test data from participant 7, using an in situ loudspeaker at $\theta_{LS} = 0°$ during the first repeat shown in Figure. 8.21.

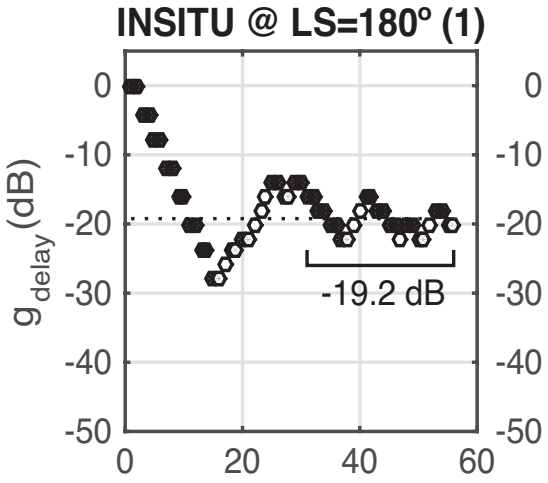


Figure 8.21: Raw CDT response data for participant 7, $\theta_{LS} = 0°$, repeat 2. This example is shown to highlight the possible influence of previous stimuli on currently trials.

The first run continues much further below the final measured CDT value. The second run also continues much higher than values shown in later runs. One hypothesis for this is that each trial *is* dependent on the preceding stimuli because when a listener perceives a coloured signal above their threshold they are 'tuned' in to what to listen for, then as the amount of colouration drops below their threshold they no longer have a reference for what the colouration sounds like and must step much higher than their CDT to be 'reminded'. This profile was found to occur in other raw response plots (Appendix. B) but no further experimentation into the cause was performed. If the hypothesis is correct then

this profile could have contributed to some of the larger measured σCDT values.

Following analysis of the raw CDT values, equivalence testing was performed to define the limits of equivalence for CDTs measured in situ and using the AVE. A new difference sample was firstly created by finding all possible intra-subject differences between in situ and AVE CDT values. This sample was then bootstrapped (N=1000) to provide a new sample where mean and 90% confidence intervals are shown in Figure. 8.18.

Firstly, these results show that a $\pm 4\,dB$ equivalence region can be defined across all five θ_{LS} directions tested. This region was defined post-hoc from the test results and provides a range which can be used to understand the physical effect of the change in CDT, to further understand the practical impact. It was found that $\theta_{LS} = 180°$ highlighted the largest confidence intervals meaning that at this angle, participants had the largest variability in difference between AVE and in situ CDT values. $\theta_{LS} = 0°$ and $\theta_{LS} = 90°$ both had CIs not overlapping 0° meaning that the reduction in CDT using the AVE (increased colouration acuity) was significant. Although CDT values were measured across 5 different θ_{LS} directions, results show that there does not seem to be a conclusive pattern for CDT equivalence over θ_{LS}. More θ_{LS} angles would need to be measured to define this.

However, it is important to quantify what this means for the practical use of such AVEs for the assessment of colouration artefacts. To help demonstrate this, Figure. 8.22 shows the perceptually smoothed magnitude frequency response of a coloured signal simulated using the response of the system shown in Figure. 8.11 to

a Dirac delta function. Inter-subject variation has shown to be non-trivial in CDT measurements and therefore a value of $g_{delay} = -25\,dB$ is taken as a representative CDT value for these examples. The 4 dB equivalence region ($g_{delay} = 21\,dB$ & $g_{delay} = 29\,dB$) defines the region where CDTs could be measured when simulating the signal using the AVE if the CDT for in situ reproduction was 25 dB.

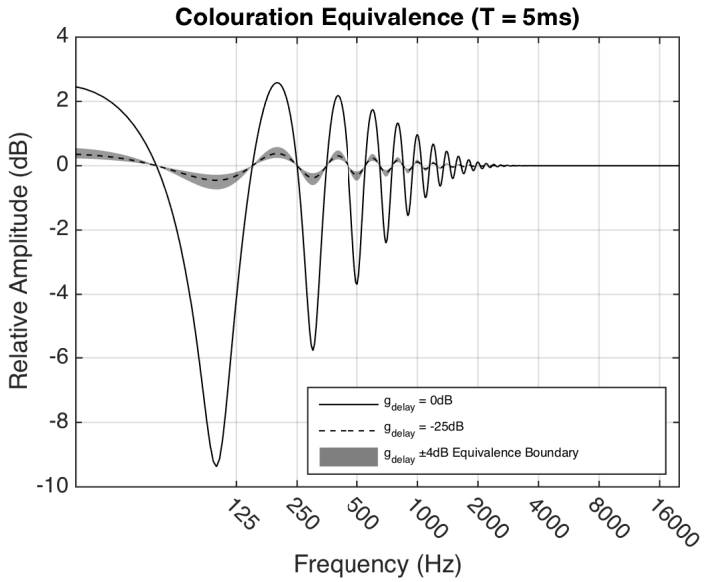


Figure 8.22: Magnitude frequency response of coloured signal corresponding to different CDT values. Perceptually motivated smoothing is applied using an ERB spaced gammatone filter-bank (ERB spacing = 0.1). This represents the internal magnitude spectrum of a white noise signal played into the loudspeaker input. T = 5 ms.

It is possible to see the effect of the AVE on CDTs by convolving the colouration processing's response to a Dirac delta (as shown in Figure. 8.22) with a HRIR pair. The magnitude frequency response of this is shown for the right ear only in Figure. 8.23. The HRIR was used from an anechoic approximation of the listening scenario for a loudspeaker at $\theta_{LS} = 315°$.

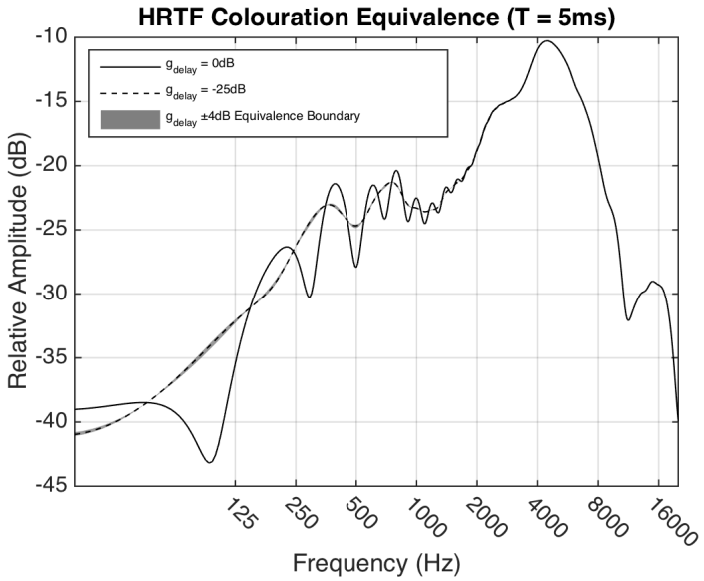


Figure 8.23: Magnitude frequency response of the anechoic approximated HRTF for coloured signal corresponding to different g_{delay} values. The response shown is for the right ear, central listening position, $\theta_{LS} = 315°$ and head-azimuth $\theta = 0°$. Perceptually motivated smoothing is applied using an ERB spaced gammatone filter-bank (ERB spacing = 0.1). T = 5 ms.

When g_{delay} is maximal at 0 dB, the comb filtering is clearly identifiable with large notches as shown in Figure. 8.22. However, notches are largely reduced when $g_{delay} = -25\,dB$ corresponding to realistic colouration at the CDT value. Values of g_{delay} at 21 dB and 29 dB correspond to the upper- and lower-limits of shift in CDT introduced by listening using the AVE. These values change the HRTF by only a small amount as shown in Figure. 8.23 and notches appear objectively less dramatic due to the larger range of magnitudes in the HRTF. The fact that the colouration is perceivable with such small modulations in the magnitude response also highlights the sensitivity of the auditory system. These plots highlight the fact that the AVE only introduces small distortions in the perception of weak sound colouration.

8.4 Image-shift Threshold

As an additional study on the ability of listeners to rate artefacts when colouration and localisation are combined, image-shift thresholds were measured. Three listeners from CDT experiment B participated in the study which was conducted on a different day. The simulation of a delayed signal from a different direction will help to verify that the AVE can, alongside simulating localisation and colouration cues independently, induce important image-shift cues which hinder localisation performance at off-centre listening positions in loudspeaker-based spatial audio systems.

The same procedure was followed as described above for CDT experiment B, but for image-shift thresholds where,

$\theta_{direct} \neq \theta_{reflection}$.

Figure. 8.24 shows the setup of the image-shift threshold test. For this test $\theta_{direct} = 0°$ and $\theta_{reflection} = 45°$. The delay time remained at $T = 5ms$. The geometrical setup for IST measurements are shown in Figure. 8.24.

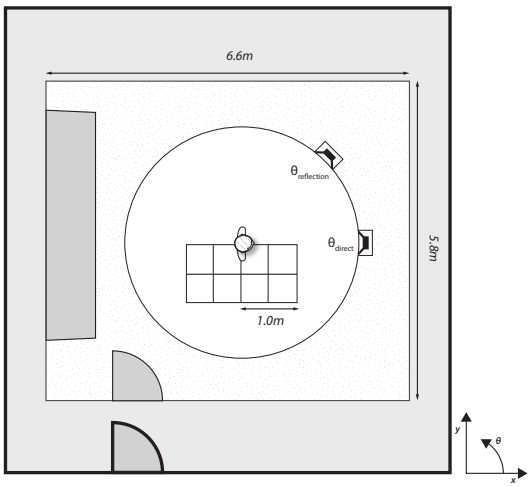


Figure 8.24: Experimental setup of the image-shift threshold measurements. $\theta_{direct} = 0°$ and $\theta_{reflection} = 45°$. $T = 5ms$.

Image-shift thresholds (IST) were calculated by taking the mean for each participant across 2 repeats. As with the CDT measurements, pariticipants were asked to identify the shifted image of their perceived auditory events when comparing REF to A and REF to B, selecting either A or B as the image-shift stimulus.

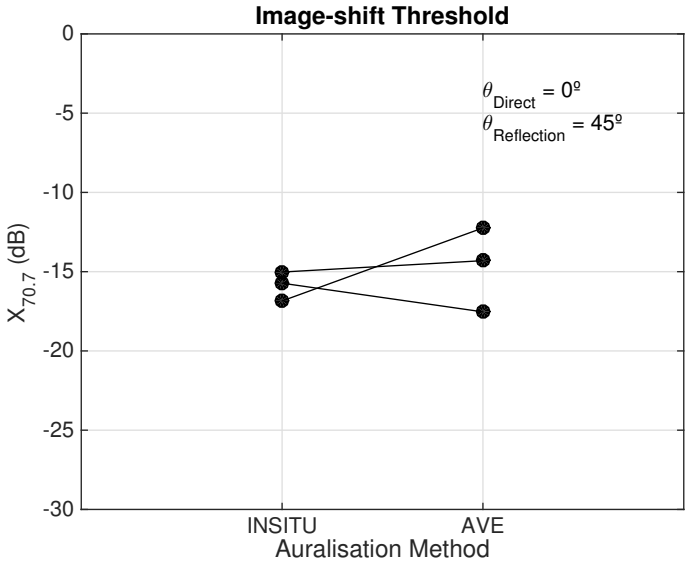


Figure 8.25: Mean image-shift thresholds for $\theta_{direct} = 0°$ and $\theta_{reflection} = 45°$. Results are averaged over 2 repeats of each trial and solid lines join results for each participant.

The equivalence of ISTs measured using in situ and AVE represents the logical progression from localisation and colouration acuity measured individually. Although only measured using a small sample of listeners, some features can be identified from the data. IST values show more variance for the AVE than in situ. For two of the three participants the AVE had a higher value meaning image-shift acuity was reduced using the AVE. However, a detailed study would need to be undertaken to achieve more conclusive results following from this study.

8.5 Conclusions

This chapter proposed a number of research questions regarding the use a non-individualised auditory virtual environment for the simulation of colouration artefacts. Specifically, colouration artefacts found at non-central listening positions in domestic spatial audio reproduction systems were considered. CDTs

were measured using both the adjustment method and a 2-alternative forced choice design. For the 2AFC method, CDT values were measured from five directions around the left side of the listener for both auralisation methods. However, in both methods, no significant difference in colouration acuity was found across the tested directions. The equivalence boundaries for CDTs measured using the AVE were found to be $\pm 4\,dB$ across 5 loudspeaker directions. Inter-subject variation in CDT values was found to be significantly larger than differences between in situ and AVE CDTs. Equivalence testing also indicated that variance in the difference between AVE and in situ CDTs (δDT) was largest when the coloured signal came from 180° (behind the listener). It was also found that CDTs, and therefore colouration acuity, were not consistently increased or decreased by the use of the AVE. The physical impact of the change in colouration acuity using the AVE was also demonstrated using perceptually motivated smoothing and head-related transfer functions. Image-shift thresholds were also measured for a small sample of listeners and results indicated that inter-subject variation was larger for ISTs measured using the AVE. However, further testing is required for this metric.

CHAPTER 9

The Perception of Colouration Artefacts Across the Domestic Listening Area Using Loudspeaker-based Panning Methods

This chapter covers the results of two experiments implementing non-individualised dynamic binaural synthesis to measure the magnitude of perceived colouration found in spatial audio systems across the domestic listening area. In the second experiment, the comparison of colouration perception at central and non-central listening positions is considered specifically with analytical models to aid in analysis.

9.1 Introduction

To understand the perception of colouration across the domestic listening area for two different amplitude panning systems, two subjective experiments have been undertaken. The first experiment was conducted with direct attribute scaling where the second experiment used an indirect attribute scaling approach. The aim of both experiments was to measure the magnitude of colouration at different listening positions relative to an explicit reference. In both experiments, the explicit reference was chosen to be an auditory event created using a single loudspeaker. This allows the evaluation of colouration artefacts at central and non-central listening positions to be compared.

In these experiments the aim is to understand the magnitude (scale) of perception of colouration when presented with different controlled stimuli. Scaling procedures can be split into two categories (Bech and Zacharov, 2006): (1) Direct and (2) Indirect. For direct scaling, a listener's task is to report the magnitude of an attribute (from small to large) based on their perception of an auditory event. Indirect scaling procedures developed around the initial work by Thurstone (1927) require listeners to select one of two (or more) presented stimuli based on which has the largest magnitude of percept under evaluation. Binomial distributions are then used to compute choice probabilities. In this regard the analysis is applied in a more stochastic way.

9.2 Physical Sound Field

The aim of a loudspeaker-based reproduction system is to simulate the perception of an intended auditory event using a finite number of loudspeakers.

For Ambisonic systems, a plane-wave can be decomposed into spherical harmonics and then reconstructed using a regular, 2-dimensional loudspeaker layout. Under certain assumptions already discussed in Chapter. 3, simulation can be achieved relatively simply by weighting a desired mono signal onto each loudspeaker using a set of gain coefficients.

It is possible to show the physical error in pressure-field reconstruction by truncating the spherical harmonics and number of loudspeakers. Spatio-temporal pressure fields can be mapped under the assumption that loudspeakers act as free-field point sources. Comparison with a free-field monopole at the virtual source position will show the difference between intended and actual sound fields. The complex pressure at any point in the listening area generated by a monopole can be found using Equation. 9.1.

$$p(r, \omega) = \frac{A}{r} e^{i(\omega t - kr)} \tag{9.1}$$

Where p is the complex valued, free-field pressure with radial source frequency $\omega = 2\pi f$. The resultant physical pressure is given by $\Re(p)$. r is the distance from the monopole and k the wavenumber. A is the point source strength (or volume flow). t is the instantaneous point in time. The resultant spatio-temporal pressure fields for a single monopole at 20° azimuth from the front-facing listener at different driving frequencies are shown in Figure. 9.1.

It is possible to plot the error in the pressure-field reconstruction by considering the difference between a monopole at the virtual sound source direction and the pressure field created by Ambisonic weightings applied to monopoles at the intended loudspeaker directions. To demonstrate the spatial limitations of truncating the number of loudspeakers and number of spherical harmonics, the

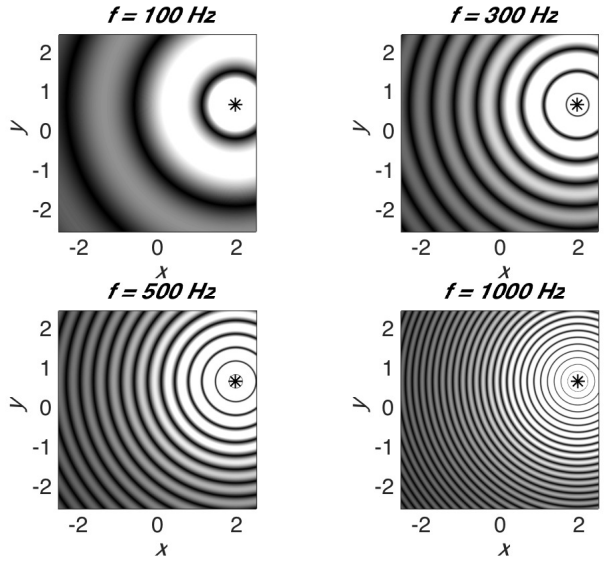


Figure 9.1: Pressure field for a monopole at 20°. The monopole is at a radius of 2.1m from the central listening position, indicated using the (∗) symbol. Calculations were made using p from Equation. 9.1 at 100 Hz, 300 Hz, 500 Hz and 1000 Hz.

log-squared pressure field error map was created for an octagonal loudspeaker layout with a loudspeaker radius of 2.1 m. A third-order max_{rV} decoder was implemented to achieve the loudspeaker weightings.

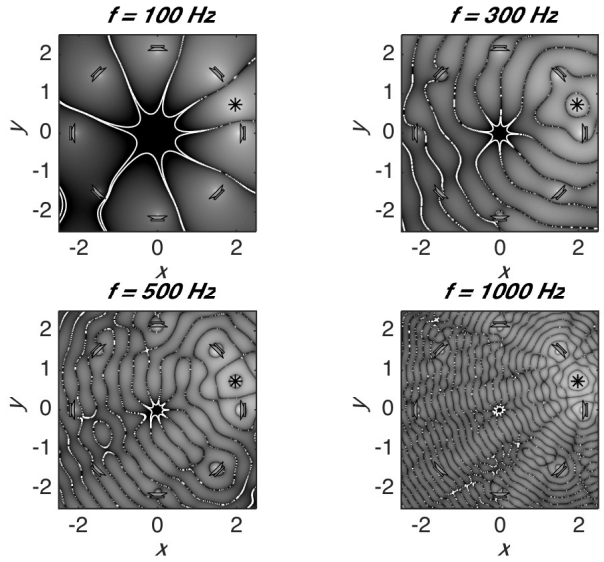


Figure 9.2: log-squared pressure-field error using Ambisonics for a virtual source at 20° (indicated by *). max rV 3rd order Ambisonic decoder with Octagonal loudspeaker layout. White contour lines indicate -70 dB error threshold.

As shown in Figure. 9.2, at $f = 100Hz$, the reconstruction of the monopole pressure field is largest but as the driving frequency increases, the region of accurate reconstruction is reduced. Contour lines indicating -70 dB were chosen to highlight the effect of changing driving frequency but may not necessarily reflect the threshold of human perception.

Although the delay that induces the comb-filtering is smallest at the central listening position compared with larger maximal delays when moving off-centre, it is not well known how these different comb-filter structures are perceived or which is considered worse in a perceptual sense.

It is possible to continue this analysis and choose one point in the listening area, with coordinates denoted X_{LP} and Y_{LP} and calculate the pressure magnitude response across frequency for this point as an estimate of the comb filtering introduced. For a listener seated at the central listening positions, their ear positions are likely to be at $\pm 0.08\,m$ along the y axis for a listener looking towards $\theta = 0°$ ($\pm 0.08\,m$ is chosen as a representative human head radius). As a fundamental example, consider the resultant frequency response caused by the summation of two coherent sound sources positioned at $\theta = \pm 30°$ (phantom centre) for (1) $X_{LP} = 0.0\,m$, $Y_{LP} = -0.08\,m$ corresponding to the right ear of a centrally seated listener and (2) $X_{LP} = 0.0\,m$, $Y_{LP} = -0.58\,m$ corresponding to the right ear of an offset seated listener shown in Figures. 9.3 and 9.4. This leads to the question: if comb-filtering is audible at the CLP, and also at non-CLP (larger delays), which is perceptually worse?

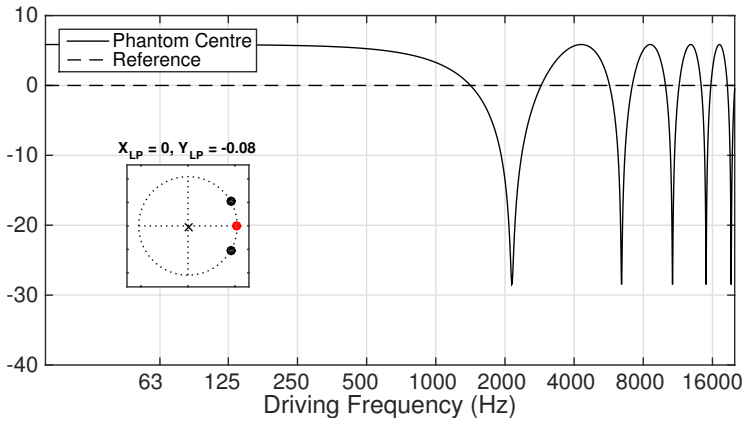


Figure 9.3: Simulated comb-filter magnitude response of stereo phantom centre at position $X_{LP} = 0.0$, $Y_{LP} = -0.08$. This simulates the magnitude response at the position of the right ear for a centrally seated listener for coherent signals on a stereophonic layout.

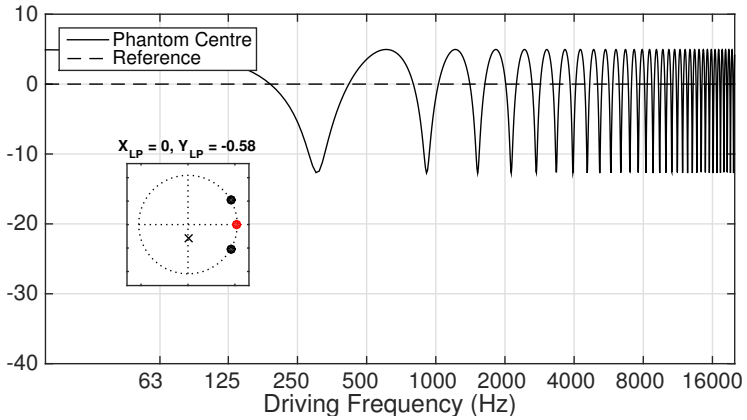


Figure 9.4: Simulated comb-filter magnitude response of stereo phantom centre at position $X_{LP} = 0.0$, $Y_{LP} = -0.58$. This simulates the magnitude response at the position of the right ear for a non-centrally seated listener for coherent signals on a stereophonic layout.

To compare the comb filtering at central and one non-central listening positions, two reference points are chosen shown in Figures. 9.5 and 9.6. The examples are shown for max_{rV} Ambisonic decoders by scaling A from equation (9.1) by the loudspeaker gain coefficients.

For the 1st order system, listening at the CLP indicates worsened spectral artefacts will be introduced at the ears of the listener (with no head present) due

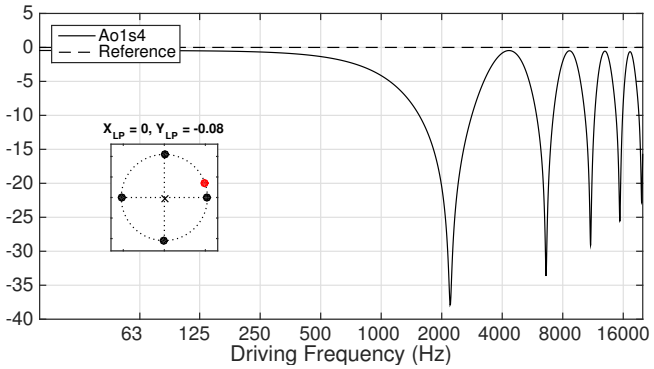


Figure 9.5: Simulated magnitude response at the position of the right ear of a centrally seated listener - 1st order Ambisonics with a cross loudspeaker layout.

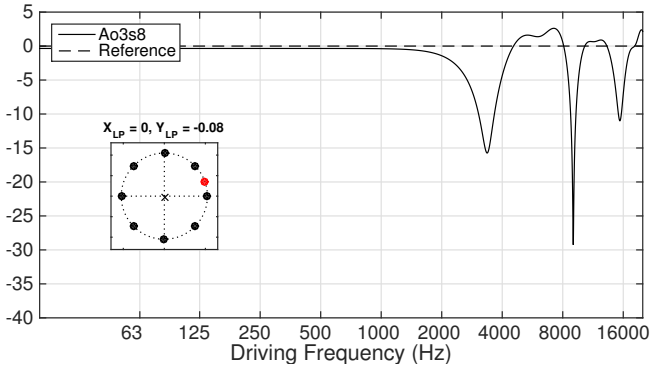


Figure 9.6: Simulated magnitude response at the position of the right ear of a centrally seated listener - 3rd order Ambisonics with an octagonal loudspeaker layout.

to the deeper notches which are wider in the higher frequency regions. Due to the specifics of the loudspeaker layout and virtual source position, the magnitude response is similar to a feed-forward comb-filter network. Increasing the Ambisonic order from 1st to 3rd order and moving to an octagonal loudspeaker layout produces similar results. The introduction of different transmission path delays seems to reduce the magnitude of the first notch and change the fundamental frequency. Generally, the comb-filtering effect seems to be slightly reduced for the higher order case, matching with the pressure-field error plots.

It is possible to now consider the comb-filtering at a non-CLP for the same

reproduction systems in Figures. 9.5 and 9.6. Due to the larger offset from the CLP, the magnitude response has a more complex nature due to an increase in the number of delayed transmission paths. This is shown in Figures. 9.7 and 9.8. However, it should be clearly noted that Figures. 9.3 - 9.6 are shown without the effect of head-scattering present and purposely highlight the physical cause of any perceptual effects. In real-life scenarios, the effect of the head scattering and diffusing and room reflections will perceptually lessen the effects highlighted above.

Considering these objectively it is non-trivial to predict which will induce the largest subjectively perceived artefacts. The inability of the human auditory system to resolve narrow, closely-spaced combs at high-frequencies may mean that comb-filtering for non-CLP scenarios may be less problematic than suggested by the pressure magnitude response plots. When modelling human response to coloured signals Atal and Schroeder (1962) and later Bilsen (1968) implemented exponential windows (10-20 ms in length) to model human perception of comb-filters, meaning that colouration induced by larger delays had less perceptual significance (due to the physiological auditory windowing) than delays arriving early, which implies that perceptually, colouration induced at, or near the central listening position could be more perceptually detrimental than colouration at off-centre listening positions. For colouration induced by larger delays it is known that timbral effects such as 'roughness' or 'rumble' are more dominant as the auditory windowing of the ear cannot resolve the spectral artefacts (Bilsen, 1977; Rubak and Johansen, 2003).

Although the reconstruction error can be defined objectively, the way that a listener integrates the errors across driving frequency is not well understood. Two studies are described in this chapter to measure the subjective judgement of

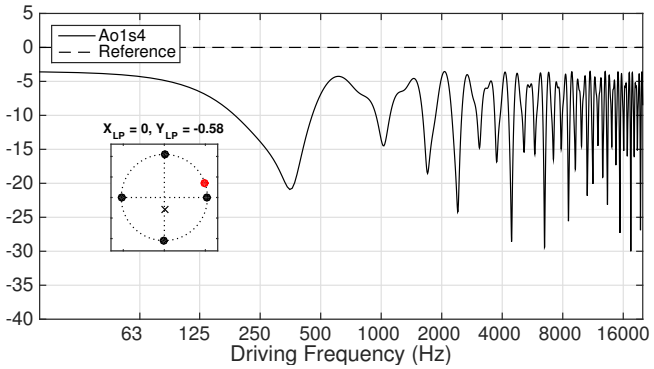


Figure 9.7: Simulated magnitude response at the position of the right ear of a non-centrally seated listener - 1st order Ambisonics with a cross loudspeaker layout.

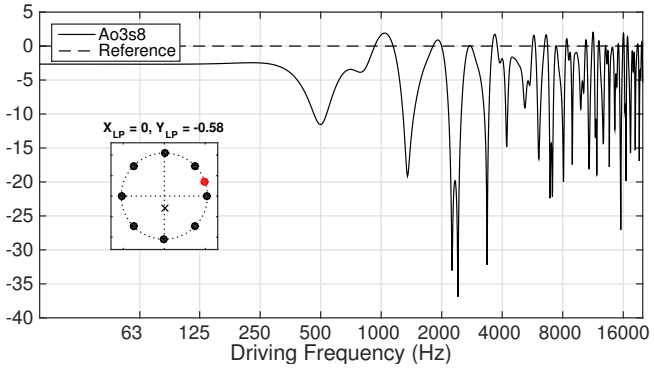


Figure 9.8: Simulated magnitude response at the position of the right ear of a non-centrally seated listener - 3rd order Ambisonics with an octagonal loudspeaker layout.

the magnitude of colouration artefacts perceived across the listening area. Due to the nature of panning methods to driving coherent sound signals, the small delays between signals arriving at the ears of a listener seated even at the central listening position have been shown to cause significant colouration effects. It is therefore important to understand how the differences in comb-filtering caused by off-centre listening relate to comb-filtering at the central listening position.

9.3 Experiment A: Direct Scaling

Research Questions:

1. How does perceived colouration change over the listening area of each system / stimulus?
2. Does Ambisonic reproduction increase the area of reduced colouration relative to benchmark amplitude panning (VBAP)?

9.3.1 Procedure

The auditory virtual environment simulating listening at multiple listening positions was created using the non-individualised binaural system described in Chapter. 4. The stimulus was repeating pink noise bursts, 800 ms in length including 50 ms fade-in/fade-out and 500 ms silence. Wierstorf et al. (2014) noted high similarity between results for music and noise; therefore, noise was isolated for test efficiency. A MUSHRA (BS.1534) inspired test design with an explicit reference, hidden reference and anchors was used to ensure both inter-system and inter-positional differences in colouration can be compared. The explicit reference for the test was a real loudspeaker simulated at the central listening position. The anchor was chosen to be the same binaural simulation as the reference but the loudspeaker input signal went through an additional process, performing a feed-forward comb-filter with 2 sample delay constant (at 48 kHz sample rate) and low-pass filtering with 7 kHz cut-off frequency. The delay was chosen to ensure colouration was perceptible.

Stimuli after binaural convolution for each listening position and system were loudness normalised using a ITU.1770-1 specification to avoid loudness

differences. Each listening position's reference (forward) direction was also rotated to maintain the virtual source direction and reduce direction cues influencing ratings. The loudness normalisation was achieved by running the loudness model on pre-processed versions of the binaural stimuli for a listener facing forward ($\theta = 0°$). Once the normalisation values were calculated, they were applied to each of the simulated systems in real-time.

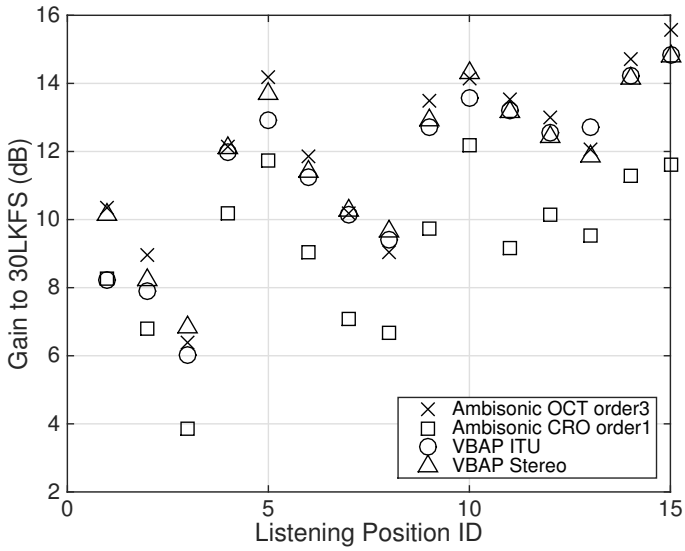


Figure 9.9: Loudness normalisation values for each listening position and panning method in direct-scaling experiment A.

Reference, hidden reference and hidden anchors ensure range equalising biases (Zielinski et al., 2008) were kept constant between the test repeats and centring bias is reduced. Rating was performed on a 100-point continuous scale from 0 (no difference) to 100 (very different). The main test was split into two parts where each part used either anechoic or reverberant binaural simulation. The order was chosen randomly. Listeners were asked to rate colouration across the listening area for 15 listening positions for each of the 4 panning methods separately. The test lasted approximately 30 minutes. 13 participants (all male) from the University of Salford Acoustics Research Centre participated in the experiment. All had

experience with audio or acoustics related user studies and were remunerated for their time.

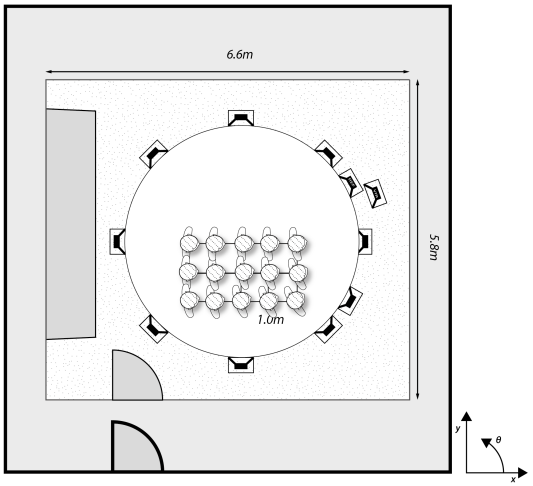


Figure 9.10: The experiment layout simulated by the AVE for the indirect-scaling colouration experiment. All virtual listeners are rotated to maintain the relative virtual source direction of 20°. All loudspeakers for the reproduction systems are shown as well as the virtual source direction / reference highlighted on the external perimeter at 20°. The exterior walls of the listening room are also shown.

9.3.2 Panning Methods

Two loudspeaker-based spatial audio systems were used in both the direct- and indirect-scaling experiments. Vector Base Amplitude Panning (VBAP) Pulkki (1997) was used to simulate basic amplitude panning between pairs of loudspeakers. This is a commonly-used system, which can be considered the current state of spatial audio reproduction used in stereophonic and surround sound for broadcast and cinema platforms. Calculation is performed by making the centrally-seated listening the base of a vector describing the direction of the intended virtual sound source. Loudspeaker directions are then used to create weighting coefficients to a mono sound source.

A second system was investigated using Higher-Order Ambisonic reproduction. This method 'encodes' a mono sound source using spherical harmonic weighting functions depending on the loudspeaker layout and intended source direction. This gives another set of weighting coefficients for the loudspeakers but will often use all of the loudspeakers in the array. Ambisonic reproduction has been optimised to have two stages of weighting coefficients. In the low-frequency region less than 380 Hz a $max\, r_V$ decoder is used to ensure that velocity is reconstructed accurately, in the upper-frequency region larger than 380 Hz energy is optimised ($max\, r_E$) using spherical harmonic weighting functions defined by Daniel (2001). When combined using a phase-match 2nd order Linkwitz-Riley filter (shown in Appendix. A) this gives a frequency dependent gain function as described by (Heller et al., 2008). In an attempt to standardise Ambisonic reproduction systems, the Ambisonic Decoder Toolbox[1] was used to create Ambisonic decoders. Table. 9.1 shows the panning methods and loudspeaker configurations used in this experiment. See chapter 2.4 for a fundamental derivation of VBAP and Ambisonic theory. Although the frequency limit of accurate sound field reconstruction goes up as Ambisonic order increases, the cross-over frequency of the band splitting filter was kept constant at 380 Hz for both 1st- and 3rd-order Ambisonic decoders. The gains and loudspeaker positions for the panning methods used in this experiment are shown in Appendix. D, Table. D.2.

Simulations using the binaural system were auralised using two reverberation modes. 'Full' in which the natural BRIRs of the listening room at the University of Salford is simulated (from the SBSBRIR dataset). 'Anechoic' mode used only the direct part of the BRIRs and therefore maintaining directional and

[1] https://bitbucket.org/ambidecodertoolbox/adt.git

Table 9.1: Panning methods and details used in direct scaling experiment A

Panning Method	Loudspeaker Spacing	Order / Decoding Method	Short ID
VBAP	30°	N/A	VbITU
VBAP	60°	N/A	VbST
Ambisonics	90° (square)	1st / $max\, r_V$ / $max\, r_E$	Ao1s4
Ambisonics	45° (octagon)	3rd / $max\, r_V$ / $max\, r_E$	Ao3s8

time-of-arrival cues alongside loudspeaker off-axis effects without any contributions from room reflections. This process is described in chapter 4.5.2.

9.3.3 Results

The results for the direct-scaling experiment will now be presented. Firstly, the distribution of colouration judgement values for each system is shown in Figure. 9.11.

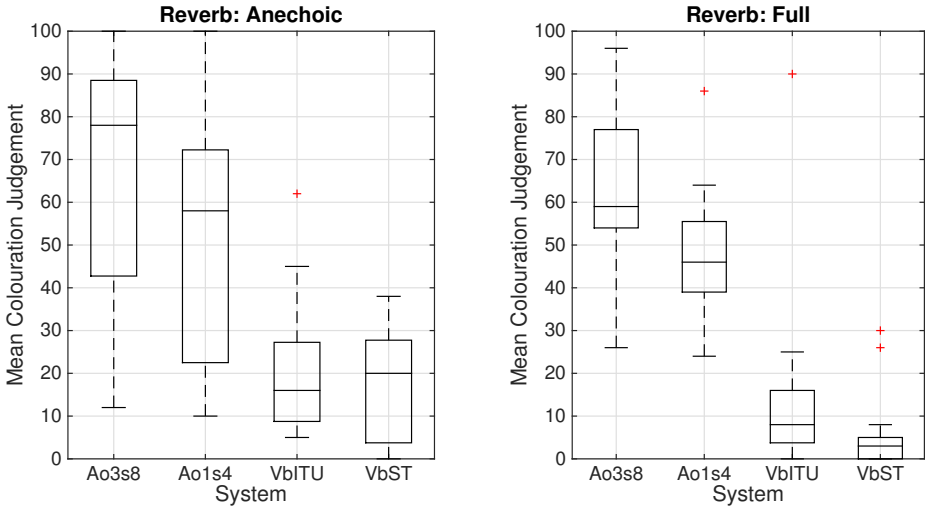
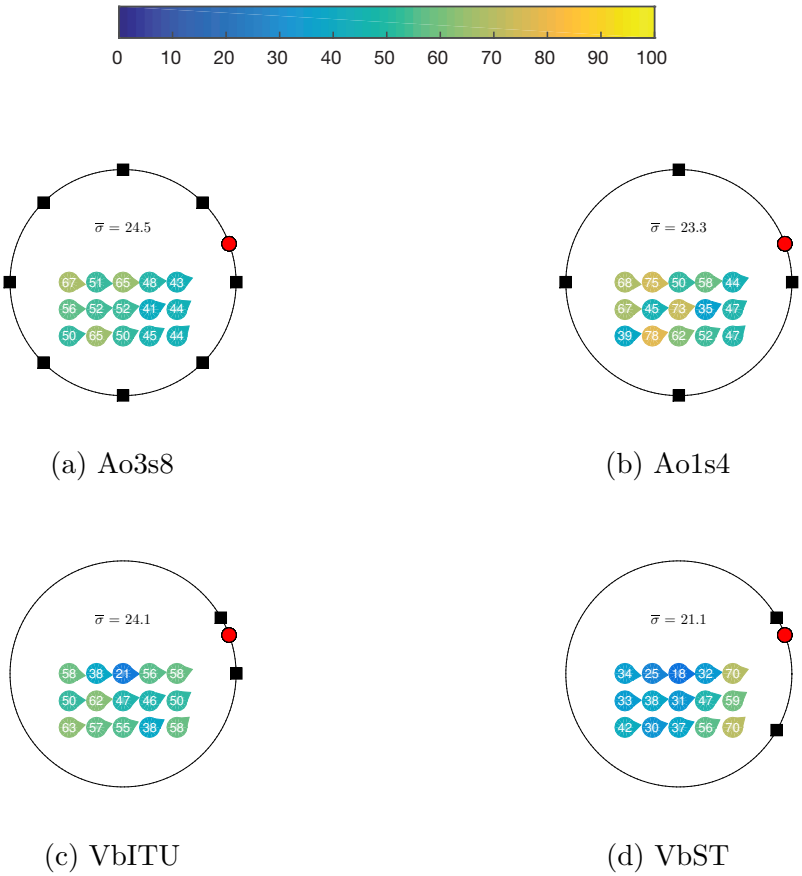


Figure 9.11: Results of direct scaling of colouration perception experiment at the central listening position only. The tops and bottoms of each box are the 25th and 75th percentiles of the samples, respectively and the line is the sample median. The length of the box is the inter-quartile range (IQR) and whiskers are the minimum and maximum observed values. Outliers show values more than 1.5 times the IQR from the top or bottom of the boxes.



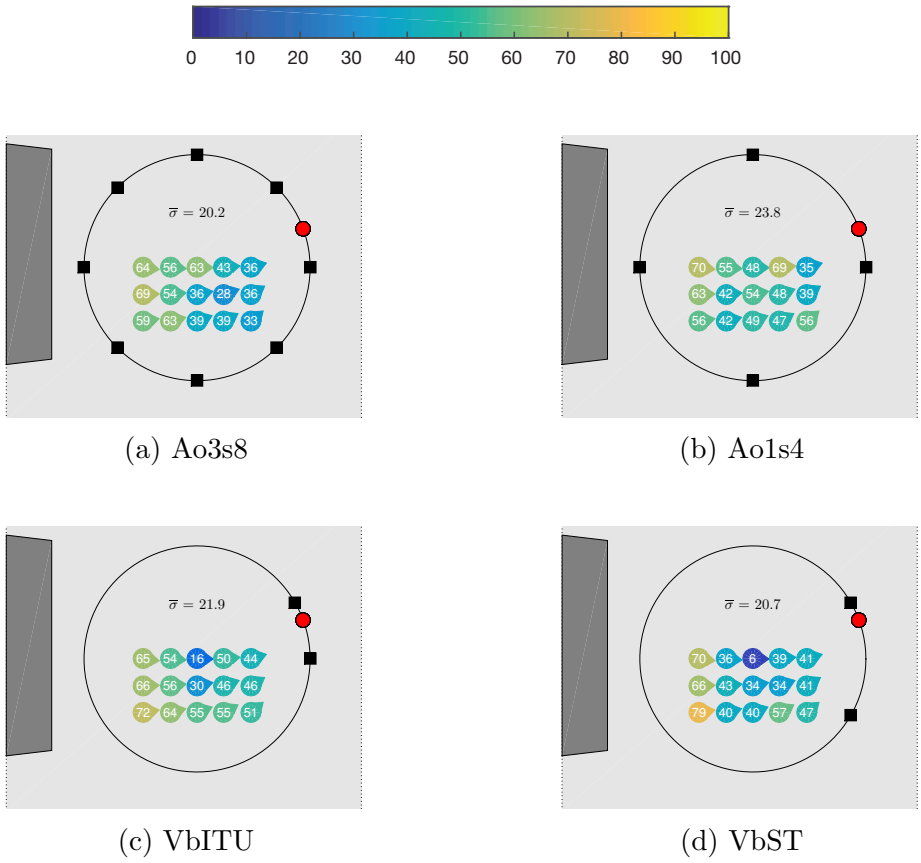
(a) Ao3s8

(b) Ao1s4

(c) VbITU

(d) VbST

Figure 9.12: Anechoic listening scenario mean colouration judgement results. Results are on a 0 - 100 point scale. Arrows indicate the virtual orientation of the listening in the auditory virtual environment. Colours change from cold to hot relate to the mean rated colouration at each listening position, indicated by the number. Black squares indicate loudspeaker positions and red circles show the virtual sound source position. Standard deviation values for reported colouration across listeners were averaged across all listening positions ($\bar{\sigma}$) and shown for each system.



(a) Ao3s8 (b) Ao1s4

(c) VbITU (d) VbST

Figure 9.13: Reverberant listening scenario mean colouration judgement results. Results are on a 0 - 100 point scale. Arrows indicate the virtual orientation of the listening in the auditory virtual environment. Colours change from cold to hot relate to the mean rated colouration at each listening position, indicated by the number. Black squares indicate loudspeaker positions and red circles show the virtual sound source position. The grey bounding box highlights the listening room walls. Standard deviation values for reported colouration across listeners were averaged across all listening positions ($\bar{\sigma}$) and shown for each system.

9.3.4 Discussion

In experiment A the central listening position was firstly considered independently where colouration was rated highest for Ambisonic systems in both anechoic and reverberant environments. There was also larger variation for Ambisonic systems which means participants had less agreement in their ratings. VBAP had lowest colouration at the CLP in the reverberant environment.

Figure. 9.12 shows that mean colouration ratings were generally increased at

non-central listening positions, but in contrast to objective metrics of sound field reconstruction error shown in the Figure. 9.2, the central listening position was not rated with the lowest colouration for the two Ambisonic systems. For VBAP systems, mean reported colouration was found to be lowest at the CLP.

The specific systems and virtual source positions were chosen to attempt to represent a fair comparison of different spatial audio panning methods. However, two specifically surprising outcomes from this experiment were found.

1. The CLP did not have the lowest mean rated colouration for all systems
2. VbST had lowest colouration ratings for both anechoic and reverberant environments at the CLP

The reason for CLP not having the lowest mean value can be partially explained by looking at the resultant BRIRs at each ear after the Ambisonic weightings. Summation of correlated signals with small time-of-arrival differences cause strong comb filtering effects. Due to the smaller loudspeaker spacing for the 3rd order Ambisonic system (45°), loudspeakers at 0° and 45° had almost equal energy. Small delays caused by the ears not being perfectly equidistant from each loudspeaker meant that comb filtering was a dominant factor. A possible reason that VbST and VbITU performed more favourably at the CLP than Ambisonics was again due to the energy levels at multiple loudspeakers. Because of the relationship between the virtual source direction and the loudspeakers,VbST had very little energy in the right loudspeaker (330°) and therefore comb-filtering at CLP was reduced. At listening positions x = 1, y = 0 and x = 1, y = -1 colouration was increased, possibly due to moving closer to the right loudspeaker and therefore comb-filtering becoming more apparent. Therefore, although an experimental choice was made to maintain the virtual source direction for each

system, this may have lead to the virtual source not being equidistantly placed between the two nearest loudspeakers and therefore, in this scenario, favouring VBAP systems.

If colouration due to comb-filtering can be perceived at the central listening position as well as non-central listening positions it is not directly obvious which type of comb-filtering is perceptually worse. The outcome from this study indicated that further experimentation was need.

9.4 Experiment B: Indirect Scaling

Following the direct-scaling method, a more in-depth assessment was performed. It can be shown that spectral changes at the central listening position are non-trivial yet the perceptual relevance of the effect is not well understood. The following test takes 6 listening positions and uses a paired-comparison method to understand how listeners rate the magnitude of colouration under independent variable conditions or two panning methods, and two simulated reverberation types. A training and familiarisation session was also implemented to ensure that listeners were aware of the different types of colouration and how large the artefacts would be in the test.

Each sample contained two coloured stimuli and a single reference. The reference was always a single simulated loudspeaker at 110° and the reproduction systems were encoded to simulate a virtual sound source a 112.5° (equidistant between 90° and 135° loudspeakers in the array). Each coloured signal was always preceded by the reference signal which was indicated to the participants using GUI indicators. The task of the participants was to choose the most coloured

signal relative to the reference and then rate the magnitude of colouration between the two coloured samples. This data is used in the analysis stage to scale the preferences and achieve a better approximation of choice probabilities. Indirect-scaling using paired comparisons has been used to avoid the complex task of having to rate many stimuli at the same time.

The type of stimuli used in paired comparison data must be carefully selected. If the differences in colouration are extreme, meaning that one stimulus is always chosen to be more, or less coloured; Thurstonian models can fail. If the stimuli are too similar then resulting choice probabilities will converge to 50%.
The research questions were defined as:

1. Is the magnitude of colouration perception reduced at the central listening position?
2. Does the use of Ambisonic reproduction change the perception of colouration?
3. Does the natural room reverberation decrease the magnitude of colouration differences across the listening positions?

9.4.1 Method

The binaural simulation system was setup as shown in Figure. 9.14. The auditory virtual environment was simulated using the non-individualised binaural system described in Chapter. 4.

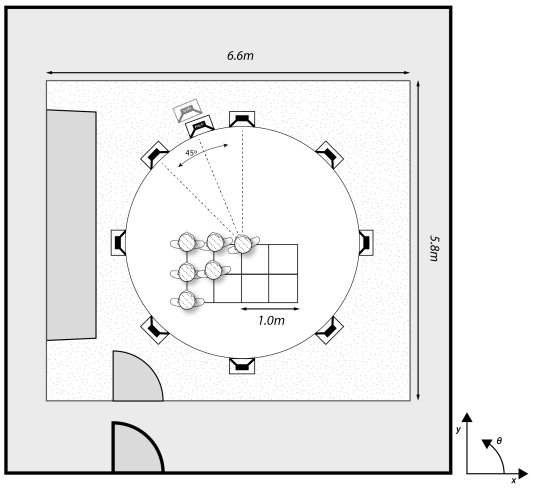


Figure 9.14: The physical setup simulated using the AVE for indirect-scaling colouration experiment. 6 listening positions are indicated with the loudspeakers used for VBAP and Ambisonic reproduction. Reference and virtual sources are also shown in the figure at 110° and 112.5° respectively.

The indirect-scaling experiment had three independent variables:

- Reproduction System (VBAP 45° or Ambisonic 3rd order with dual-band $max\ r_E$)
- Listening position (6 including CLP as shown in Figure. 9.14).
- Reverberation type (either 'Reverberant' simulating the room or 'Anechoic' with no room reverberation)

Loudspeaker gain coefficients for the panning methods are shown in Table. D.3 of Appendix. D.

The experiment was split into 4 trials, one for each combination of reproduction system and reverberation type. Within each trial pairs of 6 possible listening positions were compared giving 15 possible combinations using $\frac{n!}{(n-r)!(r!)}$ where $n = 6$ and $r = 2$.

9.4.2 Participants and Training

11 participants were chosen from the staff and students at the University of Salford Computing, Science and Engineering department. All were male and reported experience in previous audio or acoustics related user studies. 7 of the participants had previously done CDT experiment B. All participants reported that they understood the term 'colouration' in a pre-test questionnaire and all had primary employment in either audio or acoustics.

Participants were first verbally taught by the experimenter about the test and the term 'colouration' was defined. The test was split into 4 sections using a GUI: (1) Introduction, (2) Demonstration of the GUI, (3) Training and familiarisation of sound samples and (4) Main test. The four elements of the test GUI are shown in Appendix C. Parts 1 and 2 were guided through with the experimenter and parts 3 and 4 were completed by the participant alone. Part (2) allowed the participant to see the GUI and use it with a working example. Part (3) was implemented to allow the participants to identify the range of colouration magnitudes in the test and thereby aid judgements of the relative differences in colouration. The page was split into two halves as shown in Appendix C, Figure. C.3. The first half contained a training section on colouration where participants were able to choose from two different types of colouration (implemented using a feed-forward comb filter with two delay times), vary the amount of colouration induced and compare against a reference. The whole page was also determined by a 'Reverberation Type' switch which changed the global BRIR reverb from either 'Reverberant' or 'Anechoic'. The second half of the page contained all samples that would be used in the test (two reproduction systems at 6 listening positions). The order was randomised.

In the final page the participants undertook the main rating part of the test which took around 15 minutes. Pressing 'A' or 'B' firstly played the reference then followed with the either sample A or B (LED indicators showed which sample was playing). Following this the participants were asked to rate which sample had the highest magnitude of colouration relative to the reference and then how larger the difference was between the two samples. Only after answering both questions could the participant proceed to the next sample pair.

9.4.3 Paired Comparisons

An important assumption that is seldom discussed when using comparative judgements is unidimensionality of the attribute continuum (Bech and Zacharov, 2006). For the current experiment this means that listeners must be able to rate the amount of 'colouration' on a one dimensional scale. Level normalisation and adjusting the listener rotation to face the virtual source will help to reduce the effect of other perceptual attributes influencing the judgement of colouration. Training and familiarisation will also guide listeners to focus on the perception of colouration only. Although colouration is a multi-dimensional attribute (defined by underlying attributes such as timbre, pitch and roughness), subjective listening tests are often conducted where listeners are asked to integrate multiple dimensions into a single value (Bech and Zacharov, 2006; Merimaa, 2006; Peters, 2010). Some examples of compound auditory attribute (with multiple underlying attributes) can be seen in Chapter 3, Tables. 3.1 and 3.2.

The first thorough statistical analysis of paired-comparisons was proposed by Thurstone (1927) and has since been applied to fields such as taste testing,

personnel ratings and general preference David (1959). For audio testing it is possible to use paired comparison ratings to create a scale of 'colouration magnitude' for each listening position in each trial, all judged relative to the explicit reference.

Choice probabilities were firstly found from the subjective response data, each judgement was weighted using the answer of 'How large was the difference in colouration between A and B' by each participant. This method has been previously implemented by Salomons (1995) and represents a more accurate approximation of choice probabilities by using not only the binary decision but a scaling factor. In previous applications from the literature, the choice frequencies have been adjusted post-hoc using a priori knowledge, smoothing or other methods to improve the estimation, or avoid problematic 1 or 0 choice probabilities (Morrissey, 1955; Gulliksen, 1956) but explicit weighting of values by the participants is a more reliable approach.

The law of comparative judgement assumes that for two stimuli rated repeatedly by a listener, the parameter of interest can be approximated by two normal distributions with mean values μ_A and μ_B separated along the parameter continuum (in this case magnitude of colouration relative to the reference) as shown at the top of Figure. 9.15. If the two stimuli are presented simultaneously, the probability of a listener choosing one to have a higher value of colouration than another depends on the probability of the random quality difference being greater than 0 i.e. the highlighted region in the lower plot of Figure. 9.15.

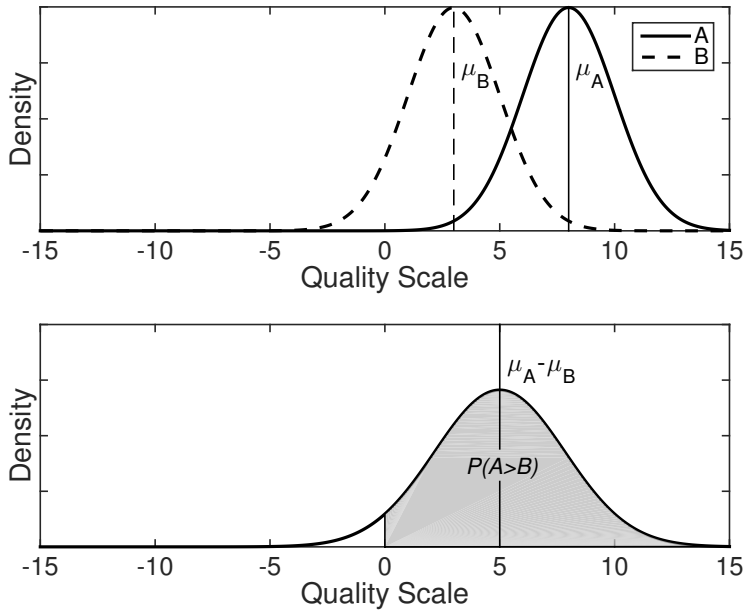


Figure 9.15: Theoretical distributions highlighting the underlying principles of Thurstonian law of comparative judgement

Once $P(A > B)$ has been calculated it is possible to use a standard normal cumulative distribution function to find the z-score of $P(A > B)$ and therefore the mean colouration magnitude difference μ_{AB}. The Case V simplification imposed by Thurstone also assumes that each stimulus distribution has equal variance and correlations. This simplifies the colouration magnitude difference to:

$$\mu_{AB} = \Phi^{-1}(C_{A,B}) \tag{9.2}$$

Where μ_{AB} is the colouration magnitude scale, Φ^{-1} is the standard normal cumulative distribution function estimated empirically and $C_{A,B}$ is the choice probabilities.

9.4.4 Results and Analysis

The z-scores, which represent the colouration scalings for each listening position are shown in Figure. 9.16.

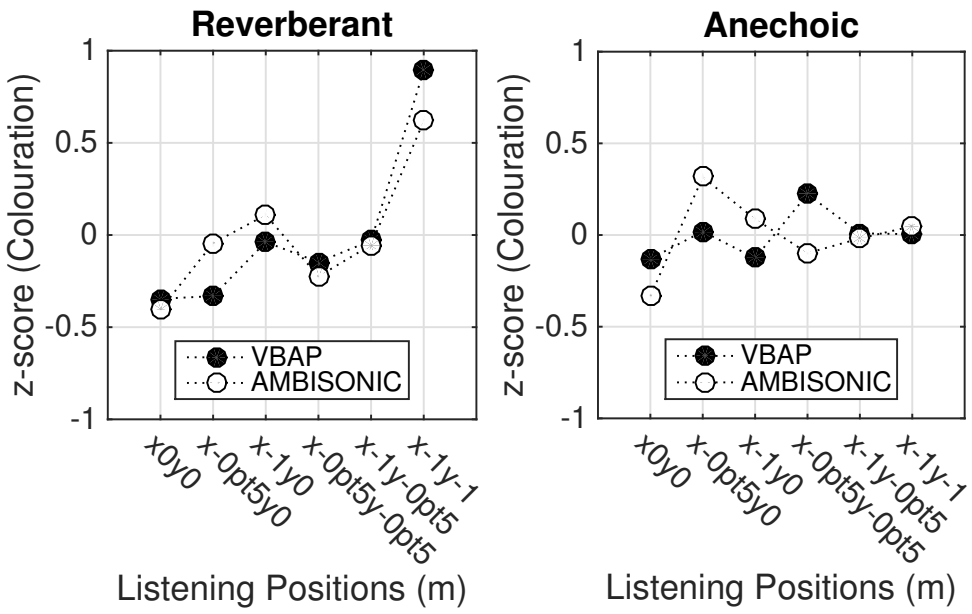


Figure 9.16: Scaling values for colouration using indirect-scaling method of paired comparisons. Analysis with Thurstone's law of comparative judgement under Case V assumptions.

The underlying data from these plots are presented in Table. 9.2.

Table 9.2: Table of z-scores corresponding to data shown in Figure. 9.16.

	Reverberant		Anechoic	
Listening Position	VBAP	Ambisonic	VBAP	Ambisonic
x=0,y=0	-0.3460	-0.4051	-0.1293	-0.3338
x=-0.5,y=0	-0.3314	-0.0441	0.0191	0.3159
x=-1,y=0	-0.0379	0.1092	-0.1260	0.0871
x=-0.5,y=-0.5	-0.1468	-0.2247	0.2225	-0.0987
x=-1,y=-0.5	-0.0304	-0.0539	0.0094	-0.0187
x=-1,y=-1	0.8925	0.6186	0.0044	0.0482

Transitivity can be used as an objective measure for the validation of paired comparison data. Taken for each participant individually it can reveal the ability of participants to make ratings consistently and on a unidimensional continuum. A stochastic representation by using choice probabilities for all participants is also useful as a global metric.

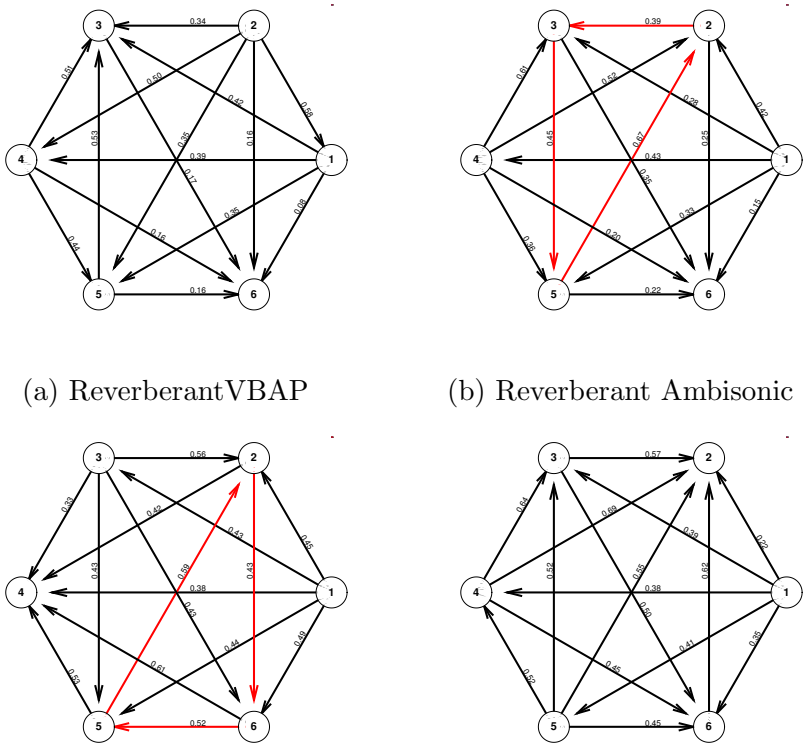
To define the fundamental concept of transitivity, consider a simple example of 3 'samples' A, B and C each rated for preference in pairs. The results can be considered transitive if the resultant scores indicated that $A < B$, $B < C$ *and* $A < C$. This would show a likelihood that all judgements could be made on a single dimension of the parameter. However, if $C < A$, either the participants failed to rate the stimuli consistently or the stimuli were rated on different dimensions for different comparisons.

If $P_{B>A} > 0.5$ and $P_{C>B} > 0.5$ then weak stochastic transitivity can be defined as (Block and Marschak, 1960; Tversky, 1969):

$$P_{C>A} \geq 0.5 \tag{9.3}$$

where $P_{A>C}$ defines the stochastic probability of a listener choosing A to be more coloured than C. For this experiment the number of listening positions was 6 per trial therefore weak stochastic transitivity (WST) violations can be calculated for

each triad of listening positions. Figure. 9.17 shows the WST digraph plots for each system*reverberation combination. Each node represents a listening position with the listening position ID shown in Table. 9.3 and red arrows highlight triads where transitivity (WST) violations were found.



(a) ReverberantVBAP (b) Reverberant Ambisonic

(c) AnechoicVBAP (d) Anechoic Ambisonic

Figure 9.17: Digraph for each trial of the indirect-scaling experiment. Triadic WST violations are highlighted in red. Numbers on lines indicated the stochastic choice probabilities of choosing the smaller node ID over the larger node ID, arrows indicate choice direction. Each node ID is a listener position (see Table. 9.3).

Numbers placed on the lines represent the probability of a participant choosing the lower ID number over the higher ID number [2] which is calculated from the paired comparison data. Arrows always point to the direction of highest probability. If participants were reliable (they could accurately discriminate between the test samples) and were able to perceive differences in colouration

[2]For example, the number shown above the arrow between ② and ③ represents $P_{2>3}$. The number shown above the arrow between ⑤ and ① represents $P_{1>5}$.

Table 9.3: Table showing listening position ID to co-ordinates. Please see Figure. 9.14 for visualisation of the listening positions.

ID	x, y (m)
1	0, 0
2	-0.5, 0
3	-1, 0
4	-0.5, -0.5
5	-1, -0.5
6	-1, -1

and rate them on a one-dimensional scale, there would never be a situation where arrows form a triadic loop. Such an outcome would indicate that for those three nodes, each one was rated higher than the other; a weak stochastic transitivity violation. WST violations can be identified by the red arrows indicating a circular triad. Only triadic transitivities were considered in this analysis. The graphs show that, when the data is considered stochastically i.e. the probability is calculated from all participant's judgements, only 2 trials had any WST violations and for these systems there was only 1 violation each. This is a positive outcome that suggests that participants were, in most cases, reliable and able to rank the coloured stimuli differences on a one-dimensional scale.

9.4.5 Transitivity Violations Correlated with CDTs

In Chapter. 8 CDT values were measured for each participant with independent variables of loudspeaker direction, θ_{LS} both in situ and simulated with an AVE. It was found that CDT values varied between participants and therefore intra-subject equivalence of CDT values were considered. It is possible to now calculate the number of weak stochastic transitivity violations that were reported by each participant. Looking at the number of transitivity violations found in the indirect scaling of colouration experiment gives some insight into the ability of participants

to rate colouration per trial. Therefore, if CDT values and the number of circular triads in the indirect-scaling comparison correlate well, CDT values could be used as a pre-screening test where colouration acuity is important.

7 participants from the CDT experiment also participated in the indirect-scaling colouration study. Correlation between the mean CDT values for each participant and mean number of circular triads (transitivity violations) was found. $R = 0.45$, $p = 0.31$ under the hypothesis of no correlation. This means that although data was correlated to some extent, the correlation failed to reject the hypothesis of no correlation.

9.4.6 Discussion

The aim of the test was to assess the magnitude of reported sound colouration across the domestic listening area when using two panning algorithms. This test was simulated using both the natural reverberation of the listening room and in anechoic conditions. Salomons (1995) defines sound colour as 'that attribute of cochlea sensation in terms of which a listener judge that two sounds similarly presented and having the same loudness are dissimilar'. Therefore colouration is 'the audible distortion which alters the natural colour of a sound'.

Figure. 9.16 shows the reported magnitude of colouration at each listening position and for both panning methods and reverberation modes. The z-scale represents a dimensionless, relative metric for the magnitude of perceived colouration and should be interpreted independently for each independent variable (no paired comparisons were made between different panning methods or reverberation modes). It is also important to note that although an explicit reference was given for each paired comparison, the magnitude of colouration

relative to the reference was never reported directly; only the difference in colouration *between* the two samples was reported. Figure. 9.16 and Table. 9.2 shows that the central listening position ($x = 0$, $y = 0$) was found to have the lowest colouration for both panning methods in both reverberant and anechoic conditions. The anechoic reproduction of the VBAP panning method also showed the lowest variation in colouration both across panning methods and reverberation modes used in the test. The range ($maximum - minimum$) of z-scores for VBAP/anechoic was 0.35 compared with 1.24 of the VBAP/reverberant condition.

Due to the nature of paired comparison tests, if the central listening position was rated consistently as more coloured than any other listening position then the z-score would have gone to ∞[3]. Although the central listening position was reported to have the lowest colouration in each trial, comparisons must have existed where the central listening position was rated to be more coloured than another listening position, indicating non-trivial colouration at the central listening position. This result is comparable to literature reports of central listening position colouration (Choisel and Wickelmaier, 2007; Pulkki, 2001; Shirley et al., 2007) and also the spectral artefact predictions shown in Figure. 9.5 and Figure. 9.6. However, results here indicated that the specific nature of the comb filtering at the central listening position, caused by small delays and characterised by the fundamental comb tooth between 2-3 kHz, is less perceptually detrimental in terms of colouration than the comb-filtering caused at non-central listening positions. Aside from the comb-filtering induced by the summation of delayed coherent sound sources, factors of colouration such as loudspeaker directivity may also have an influence on the results. Measurements

[3]This problem can be identified with knowledge that $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = \infty$

made for the SBSBRIR dataset did not pre-equalise any of the loudspeakers, therefore all reproduction utilised the relative flat frequency response from the Genelec 8030A loudspeakers.

The variation in colouration across the listening area did not vary substantially between the two panning methods which indicates that the Ambisonic panning method chosen did not reduce relative colouration at non-central listening positions. The variation in colouration across the listening area was largest for the reverberant listening mode where, for both panning methods, listening positions *x=-1, y=-1* was found to have the highest colouration. This was also the listening position furthest from the central listening position. The reverberant listening condition also indicated the lowest colouration of all the z-scores for both systems. Due to the nature of paired comparison test designed here, absolute colouration of the panning method may have been different for VBAP and Ambisonic systems which cannot be highlighted from these results directly due to comparisons between panning methods never being made by participants directly.

The large change in colouration found for the reverberant condition could have been influenced by the level normalisation applied to avoid level differences influencing the paired comparisons. Listening positions further from the active loudspeakers will result in a reduced signal to noise ratio which was likely an underlying attribute used to rate colouration. This result does indicate however that one of the predominant causes for colouration using these panning methods is caused by the reverberant field as opposed to the interaction of direct paths from the loudspeakers solely.

9.5 Conclusions

In this chapter, two experiments are presented intended to further understand the perception of colouration artefacts across the domestic listening area caused by loudspeaker-based panning methods. In experiment A, a direct-scaling method was applied using a MUSHRA-type evaluation design. Results showed that even at the central listening position, colouration was non-trivial, induced by the addition of coherent, delayed signals at the listener's ears from individual loudspeakers. Experiment A also suggested that the tested Ambisonic reproduction system had increased colouration perception than VBAP with Ambisonics having a smaller area of reduced colouration. However, the improved performance of VBAP was likely due to the virtual sound source being positioned close to the speaker position, meaning contributions from the second coherent source was small when panned using VBAP.

In experiment B, an indirect-scaling methodology was applied to further understand the differences in colouration caused by small delays (found at the central listening position), or larger delays at non-central listening positions. The test was conducted by simulating both reverberant and anechoic listening conditions with VBAP and Ambisonic panning methods. It was found that for both reverberation types and using both panning methods, the central listening position had the lowest colouration. However, the relatively small change in colouration values to non-central listening positions indicated that colouration at the central listening position is non-trivial. Variation in colouration across the six listening positions was found to be higher for reverberant simulation than anechoic simulation, indicating that artefacts caused by room reflections a significant in the perception of colouration artefacts. However, the nature of

paired comparison tests meant that the absolute panning-method-specific colouration cannot be estimated. Ambisonic and VBAP systems exhibited similar colouration profiles across the listening area for the reverberant listening condition. This indicates only a small difference in colouration perception between panning methods in comparison to changes in colouration perception across the listening area likely caused by room reflections. For the anechoic condition, smaller variations in colouration across the listening area were found which showed larger differences between Ambisonic and VBAP panning methods, where the Ambisonic system had a larger range of colouration magnitudes.

CHAPTER 10

Conclusions and Future Work

This chapter presents the general conclusions of the research activities presented in this thesis. Section 10.2 also presents a proposal for future research efforts following on from this research project.

10.1 Conclusions

The work presented in this thesis covers a number of experiments contributing to the knowledge of perception in loudspeaker-based spatial audio reproduction systems across the domestic listening area. Experiments were conducted with a focus on two important auditory attributes: localisation and colouration. For the subjective assessment at multiple listening positions, a non-individualised, dynamic binaural synthesis system was designed, verified and the validity of the system was assessed on the ability to induce the spatial and timbral artefacts found at non-central listening positions in situ. Following the aims and objectives set out in Chapter. 1, this chapter will present the general conclusions from the research.

Chapter 2 firstly introduced the important fundamental concepts needed as a basis for technical chapters presented in this thesis. VBAP and Ambisonic panning methods were described which highlighted that spatial limitations of the systems (summation of coherent, spatial distributed signals or truncation of spherical harmonic order) are likely to introduce spatial and timbral artefacts to a listener. Through a derivation of binaural simulation methods it was also shown that monaural, binaural and dynamic cues can be used to create an auditory virtual environment to a listener using headphone reproduction.

A literature review was then presented covering the main five topics relevant to this thesis. A history of loudspeaker-based spatial audio reproduction highlighted the dominance of stereophonic amplitude panning systems for domestic applications. It was also shown that timbral and spatial artefacts are two primary auditory attributes for the perception of overall audio quality in spatial

audio reproduction systems. However, these two attributes are likely to be primarily affected by off-centre listening. Computational models have been applied to predict localisation perception in loudspeaker systems. However, the modelling of colouration percepts is more complex and fewer models have been previously applied to predict loudspeaker-based spatial audio system performance. It was finally highlighted that although binaural systems are popular tools for the simulation of loudspeaker-based systems, the equivalence of perception in localisation and colouration artefacts to in situ loudspeaker systems has not yet been defined.

When designing a dynamic binaural synthesis system that includes the accurate simulation of room reflections, a binaural room impulse response (BRIR) dataset is required. Many open-source datasets are available but at the time of conducting this work, no datasets were available that included multiple listening positions across the domestic listening area. In collaboration with BBC Research and Development, the Salford-BBC spatially-sampled binaural room impulse response (SBSBRIR) dataset was measured as described in Chapter. 4. BRIR measurements were made using an artificial head and torso for 12 loudspeakers at 15 discreet listening positions in a BS.1116-1 compliant listening room at the University of Salford. Each listening position and loudspeaker was measured for every 2° in azimuth rotation of the artificial head and torso. This represents the first standardised, open-source dataset of its kind[1] allowing researchers to analyse and simulate loudspeaker systems at multiple listening positions. All measurements were also repeated using an omni-directional microphone for further analysis.

When considering the design of binaural validation experiments, the passive

[1]SBSBRIR data is available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 licence: http://creativecommons.org/licenses/by-nc-sa/4.0/.

effect of headphones on the transmission path from external loudspeakers must be considered for certain situations (due to the presentation of loudspeaker signals 'through' the headphones coupled to the listeners ear/head). In preparation for a plausibility study (Pike et al., 2014), the effect of headphones was measured using both physical measurements and a behavioural study investigating the effect of headphones on the localisation precision of external loudspeakers. Transfer function analysis using an auditory filter-bank showed a measurable spectral error in the HRTF, particularly in the upper-frequency regions. Perceptual modelling and binaural analysis also showed a maximum spectral ILD distortion of 26.52 dB. It was found that the electrostatic transducers (STAX SR-202) caused the least distortions for HRTF magnitude responses when compared to three other headphone sets commonly used in binaural reproduction systems (Sennheiser HD650, AKG K601, Sennheiser HD800) and one closed-back set selected as a low-reference (Sony MDR-V500). However, for studies where direct comparisons between virtual and real loudspeakers are required, it is recommended that headphones be coupled during the HRTF measurement to normalise the error for both scenarios, as later implemented by Pike et al. (2014). Given the findings of this work, the decision was made to validate the use of the dynamic binaural simulation system indirectly by comparing localisation errors and colouration acuity and therefore avoiding the passive effect of headphones which has been shown to affect the perception of external sound sources.

Localisation error and timbral colouration are known perceptual artefacts caused by off-centre listening in loudspeaker systems. However, subjective evaluation of systems across the listening area is practically problematic. The use of non-individualised dynamic binaural simulation systems allow for direct, blind

comparisons for multiple listening positions. Many studies have considered the localisation equivalence between individual loudspeakers presented in situ and using binaural simulation. However, it has yet to be defined whether such non-individualised binaural systems can induce localisation artefacts, caused by multi-loudspeaker systems equivalently to in situ (real) loudspeaker reproduction. To define the validity in measuring the artefacts of localisation, a subjective evaluation was conducted where a selection of representative reproduction system combinations were presented to participants in a localisation test. The term 'combination' used here is the specific selection of loudspeaker layout, audio stimulus, panning method and panning direction, where 10x 'combinations' were defined for testing intended to be representative of commonly used spatial audio reproduction systems. The test was performed using both in situ loudspeakers and loudspeakers simulated using non-individualised dynamic binaural simulation. Using a repeated measures ANOVA it was found that *combination* and *listening position* factors accounted for the largest effect sizes of the main factors. Following this, a two one-side test (TOST) concluded that 15 out of 20 reproduction system combinations, across two listening positions had perceptual equivalence between in situ and AVE simulation. Perceptual equivalence was defined by a $\pm 7^\circ$ equivalence boundary and tested at the $p = 0.05$ significance level. The 5 systems that did not achieve perceptual equivalence were found to have larger loudspeaker spacing and were likely to induce confusing localisation cues. For single loudspeakers (mono) the difference in localisation error (ΔLE) from two different directions was reported as -3.1° and -0.7° at the central listening position and 1.4° and 0.3° at the non-central listening position each with small confidence intervals, indicating the AVE simulated single loudspeaker auditory events well. For specific combinations with large absolute localisation error ($\approx 90^\circ$), the error was reproduced by the AVE well, highlighting the ability to induce important localisation artefacts and

the applicability of non-individualised dynamic binaural simulation to virtualising localisation artefacts found in loudspeaker-based reproduction systems. Importantly, results from this study validate the use of such non-individualised dynamic binaural simulation systems (within the limits specified above) for the virtualisation of loudspeaker arrays, even when complex localisation characteristics need to be well represented.

As an alternative method to simulate localisation artefacts induced by loudspeaker-based panning methods across the listening area, a computational model from the literature was implemented. A novel approach to resolve front-back confusions was added by simulating the dynamic movements made by an artificial listener using BRIRs from the SBSBRIR dataset. A process is implemented that re-aligns and averages the head-centric localisation predictions between systematic head rotations giving a new, resolved directional prediction in the room coordinate system. Model predictions for the localisation of a single loudspeaker in a reverberant environment were found to be comparable to minimum audible angle in the frontal region. When compared against subjective data for the localisation of 10 different combinations of panning algorithm/loudspeaker layout from Chapter. 6 it was found that the model had a mean signed deviation in localisation error of 7.7° for central listening position and 9.9° for the non-central listening position. When comparing to similar models for this specific application found in the literature, it can be seen the model performs comparably and often with lower error. The addition of the ability to use the model in complex, reverberant environments supports the novelty of the methodology.

As with localisation artefacts, the validity in the use of non-individualised

dynamic binaural systems for the measurement of colouration across the domestic listening area has not yet been defined. The equivalence of (single-loudspeaker) colouration acuity between in situ and AVE simulated loudspeakers was measured directly in two subjective experiments. This was implemented using tests of colouration detection thresholds for both in situ loudspeakers and simulated loudspeakers in a BS.1116-1 compliant listening room. The first experiment used a method of adjustment to compare CDTs for in situ and the AVE at one loudspeaker direction. The second experiment extended this study to consider multiple loudspeaker azimuth directions in a 2-interval alternative forced choice CDT test. Results from experiment A showed that inter-subject variation in CDT was much higher than the differences in CDT between in situ and AVE, defined by Figure. 8.8. In Experiment B equivalence boundaries were $\pm$ 4 dB across all five loudspeaker directions and as with experiment A, inter-subject variations in CDT values were non-trivial and larger than the differences between in situ and the AVE. Differences in CDTs were found to be largest when sound was placed behind the listener ($\theta_{LS} = 180°$). When differences in CDTs were applied to coloured Dirac delta responses and HRTF magnitudes, the resultant errors indicated that the differences in CDTs were small. This shows that when colouration artefacts are substantially larger than artefacts produced by feed-forward delay networks with parameters set to measured JND thresholds (g_{delay} at around -25 dB), the use of non-individualised dynamic binaural simulation (similar to the system defined in this thesis) is valid.

Although the binaural simulation system in this thesis was defined as a research tool, the specific knowledge attained from this thesis will also likely help support the application of binaural simulation systems to broadcast/media platforms

directly, through the collaborative work with BBC Research & Development[2].

The perception of (multi-loudspeaker) colouration across the listening area was finally addressed directly in two subjective evaluations. Both tests used vector base amplitude panning and Ambisonic reproduction simulated in both reverberant and anechoic environments using the AVE. In a direct-scaling experiment where a MUSHRA-style methodology was used, it was found that the central listening position induced colouration artefacts due to addition of coherent delayed signals at the listeners ears. This phenomenon has been reported in the literature for pair-wise stereophonic panning, but analytical models indicate that the effect is comparable to timbral artefacts in Ambisonic panning methods. A second experiment measured the magnitude of colouration at the central and 5 non-central listening positions using a paired comparison methodology. Results showed that the central listening positions had the lowest reported colouration on average, but absolute colouration was still perceivable at the central listening position when compared to a mono reference. Weak-stochastic transitivity was tested to assess the reliability of the results. Only 2 triadic WST violations were found which indicated participant reliability and the ability to rate colouration magnitude on a one-dimensional scale.

The difference in colouration magnitude variation across the listening area was small between the two panning methods but variation across the listening positions was larger for the reverberant environment. These results show that binaural timbral artefacts induced by very small delays (experienced at the central listening position) are likely to be perceived with less magnitude than artefacts with larger delays at non-central listening positions. However,

[2]http://www.bbc.co.uk/rd/projects/binaural-broadcasting

colouration at the central listening position is non-trivial in an absolute sense. The interaction of binaural decolouration in this process is still unknown. Analytical modelling showed that the phantom centre comb-filter effect is also related in nature to spectral artefacts caused by low-order Ambisonic systems.

Overall, this thesis has validated the use of non-individualised, dynamic binaural synthesis for assessing localisation and colouration artefacts of panning techniques in domestic listening environments. Some limitations of the system have been identified but in many cases, it shows to be a powerful tool. This is demonstrated in the application of the system to measuring colouration artefacts across the domestic listening area for two different panning methods. Subjective assessment has revealed trends in colouration magnitudes across different reproduction scenarios and techniques. This provides a mechanism for further probing of issues of colouration among loudspeaker-based reproduction methods and alternative scenarios that could be investigated by other researchers. Perceptual results indicate that the specific characteristics of colouration at central listening positions, despite being measurable, are less perceptually problematic than colourations caused off-centre listening positions. This has the potential to change the emphasis of system design for future spatial audio reproduction systems and also provides important information on the perception of sound colouration in general.

10.2 Future Work

The results presented in this thesis have addressed the original aims and objectives set out in Chapter. 1. However, alongside the *answers* from this research, it is important to consider the *questions* that have also been raised and

how these can lead to future research topics.

When considering binaural plausibility experiments (or similar types of tests) where headphones are coupled to a listeners ears whilst listening to external loudspeakers, the effect of the headphones on the upper-frequency bands in the HRTF were measurable. The effect of headphones also caused a change in the localisation precision. One solution to this problem is to measure HRTFs *with* the headphones coupled so that both binaural and real loudspeakers are equally influenced by the effect of the headphones. Since conducting the work, this issue has resurfaced numerous times when considering the validation of real and virtual sound sources. Signal processing algorithms have been investigated (Moore et al., 2007) but the concept is not without its own issues. The work in Chapter. 5 has clearly defined the problem and set fourth some of the possible solutions. However, there is currently a gap in the literature to serve as a clear solution to this problem.

For the validation of a non-individualised dynamic binaural simulation system, the ability of the system to induce localisation and colouration artefacts was considered separately. A 2AFC experiment was conducted to consider the equivalence of colouration acuity between the AVE and in situ reproduction with the loudspeaker direction (θ_{LS}) as an independent variable. 5 directions were tested but a logical development of this work would be to consider a higher resolution of loudspeaker directions to further understand the change in CDTs, and equivalence of CDTs for the AVE, across sound source direction. Although CDTs were primarily measured for single loudspeaker in this thesis, further experiments should also be undertaken to understand the acuity to colouration artefacts caused by multiple loudspeakers.

Although image-shift thresholds were measured in a small pilot study, this work should be developed further to increase the understanding of the interaction of localisation and colouration cues in binaural simulation. A development of this work would logically consider localisation and colouration attributes together for the validation of the binaural system, possibly using an alternative objective metric such as plausibility (Lindau and Weinzierl, 2012; Pike et al., 2014).

The AVE designed, verified and validated in this research served also as a foundation for part of the S3A[3] research project. This is a collaborative project between the University of Surrey, Salford and Southhampton and BBC Research and Development on future spatial audio for an immersive listener experience.

One of the limitations of the AVE found in this research is that the system uses fixed time-of-arrivals between loudspeakers to the measurement microphones. Analytically, it has been shown that these small ToA changes cause audible colouration when simulating coherent sound sources (such as a phantom centre speaker). For in situ listening, micro-translatory movements of the listener will cause a dynamic change in the comb-filtering at each listeners ear but for the AVE, the delays are fixed and therefore so are the comb-filters. The small changes in delays will also induce localisation artefacts. One hypothesis could be that because changes in the comb-filtering are dynamically related to a listeners movement, they are somehow decoloured by the auditory system and removing this dynamic change may hinder the decolouration process. Further work considering the effect of micro- and macro-translatory movements on dynamic binaural synthesis must be conducted to understand the importance of this

[3]http://www.s3a-spatialaudio.org/

parameter.

A perceptual analysis of colouration artefacts across the listening area using loudspeaker-based spatial audio reproduction systems has been presented in this thesis. However, further work must be undertaken to understand the application of these results to domestic listening scenarios. Noise-bursts were used in both direct- and indirect-scaling tests which allowed for the listeners to have good acuity to colouration artefacts. However, for multi-dimensional sound stages the results may differ. A logical progression would be to implement the indirect scaling procedures for the perception of colouration artefacts across the listening area using real broadcast content such as radio or television material and the extension of the testing to reproduction techniques, even using 3-dimensional loudspeaker layouts.

APPENDIX A

2nd order Linkwitz-Riley filter (LR2) used for dual-band processing.

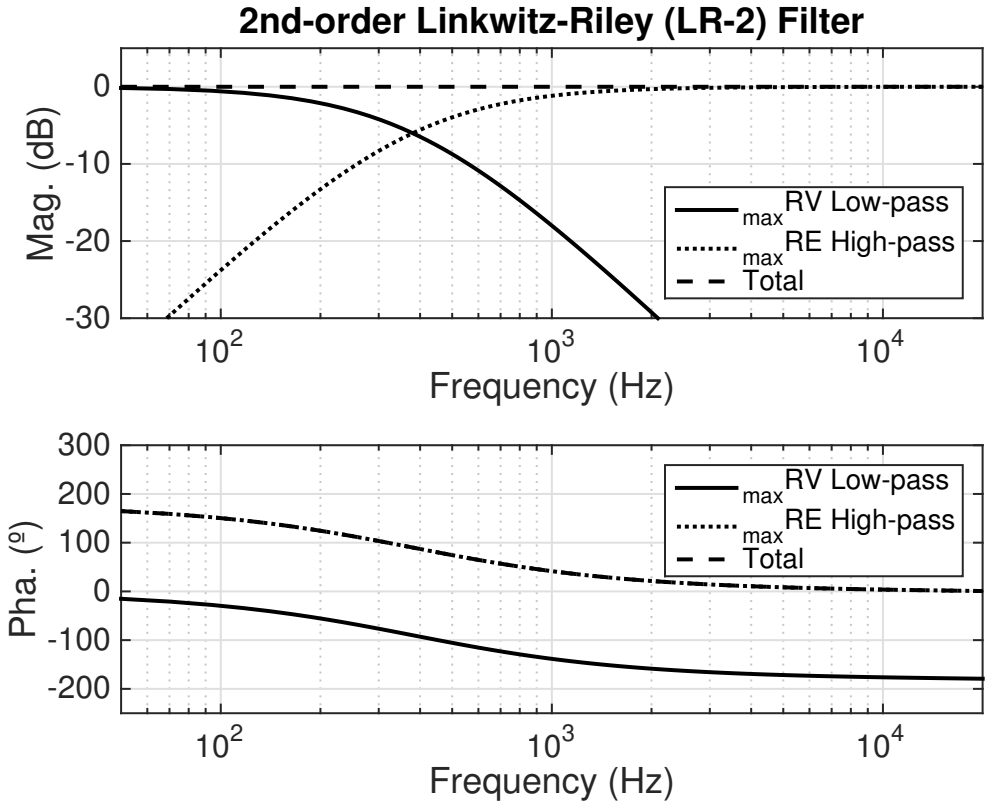


Figure A.1: Magnitude and phase response of phase-matched Linkwitz-Riley filters (Linkwitz, 1976) used in dual-band Ambisonic decoding .Direct-form I IIR filter coefficients from Heller et al. (2008) with a cross-over frequency on 380 Hz and a sample rate of 48000 Hz are $b_{lp} = [0.589 \times 10^{-3}, 1.178 \times 10^{-3}, 0.589 \times 10^{-3}]$, $b_{hp} = [0.952, -1.904, 0.952]$, $a = [1.00, -1.903, 0.905]$.

APPENDIX B

Raw CDT measurement results for CDT experiment B, from Chapter 8.3.

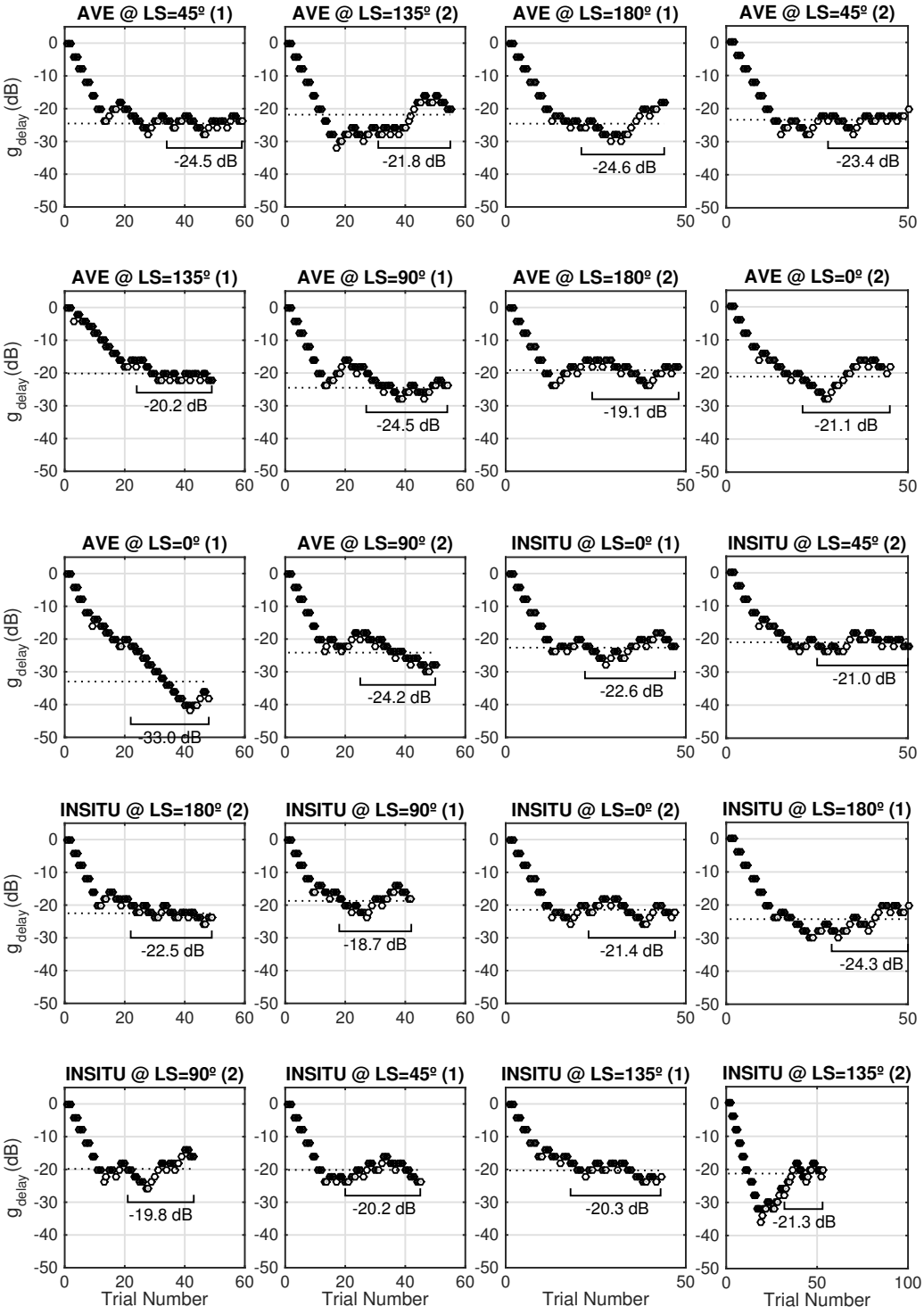


Figure B.1: CDT response data for colouration validation experiment B: Participant 1

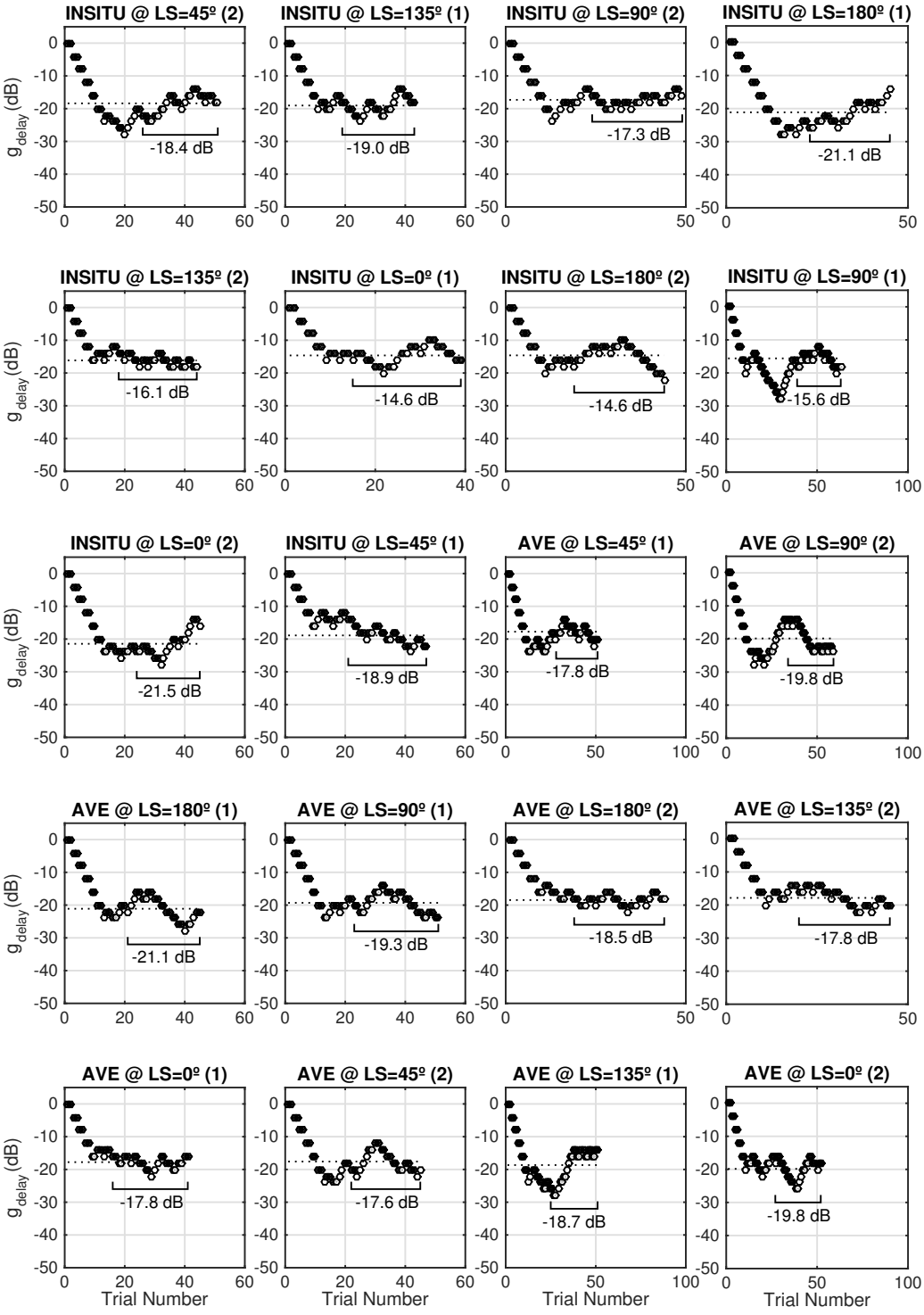


Figure B.2: CDT response data for colouration validation experiment B: Participant 2

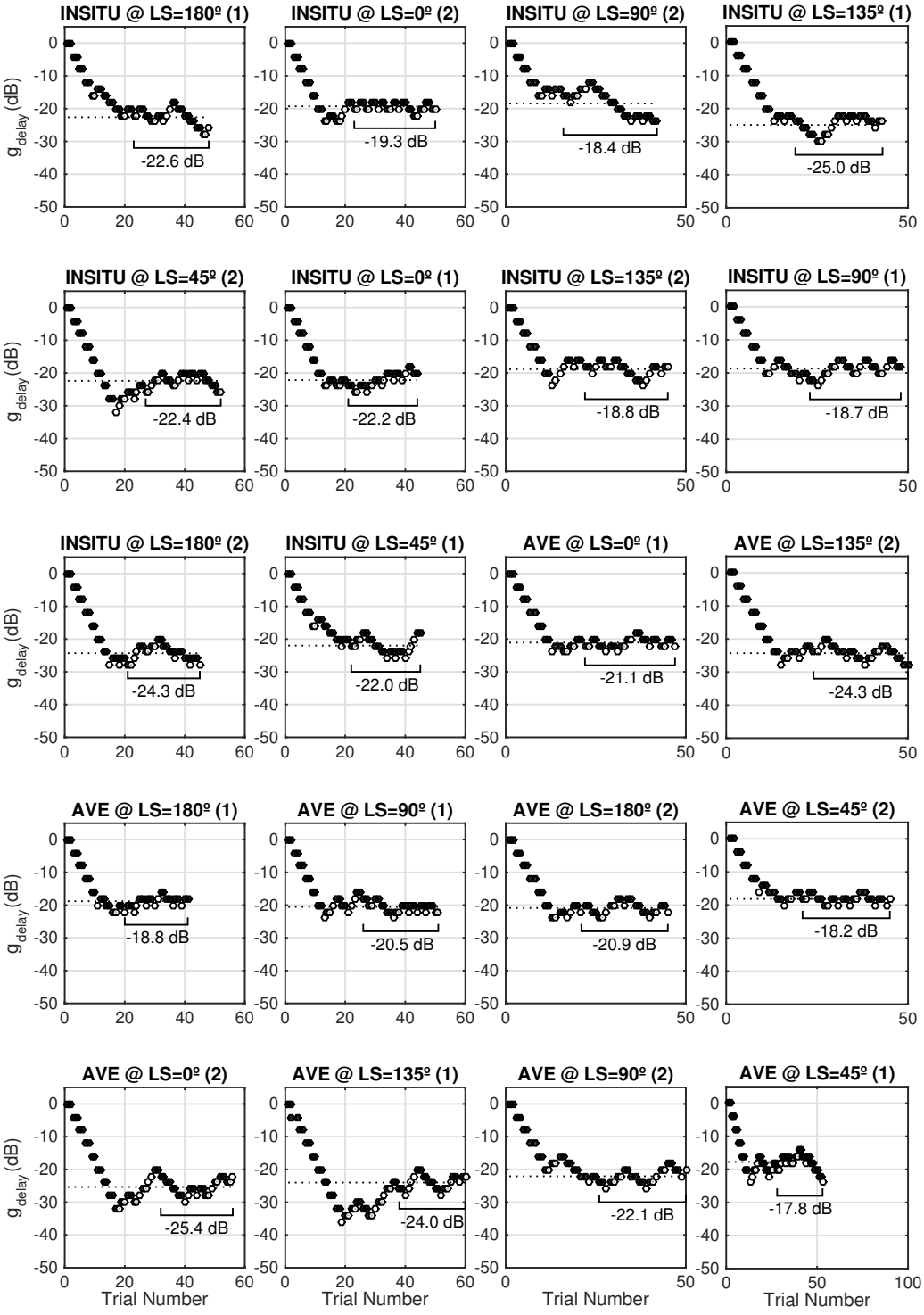


Figure B.3: CDT response data for colouration validation experiment B: Participant 3

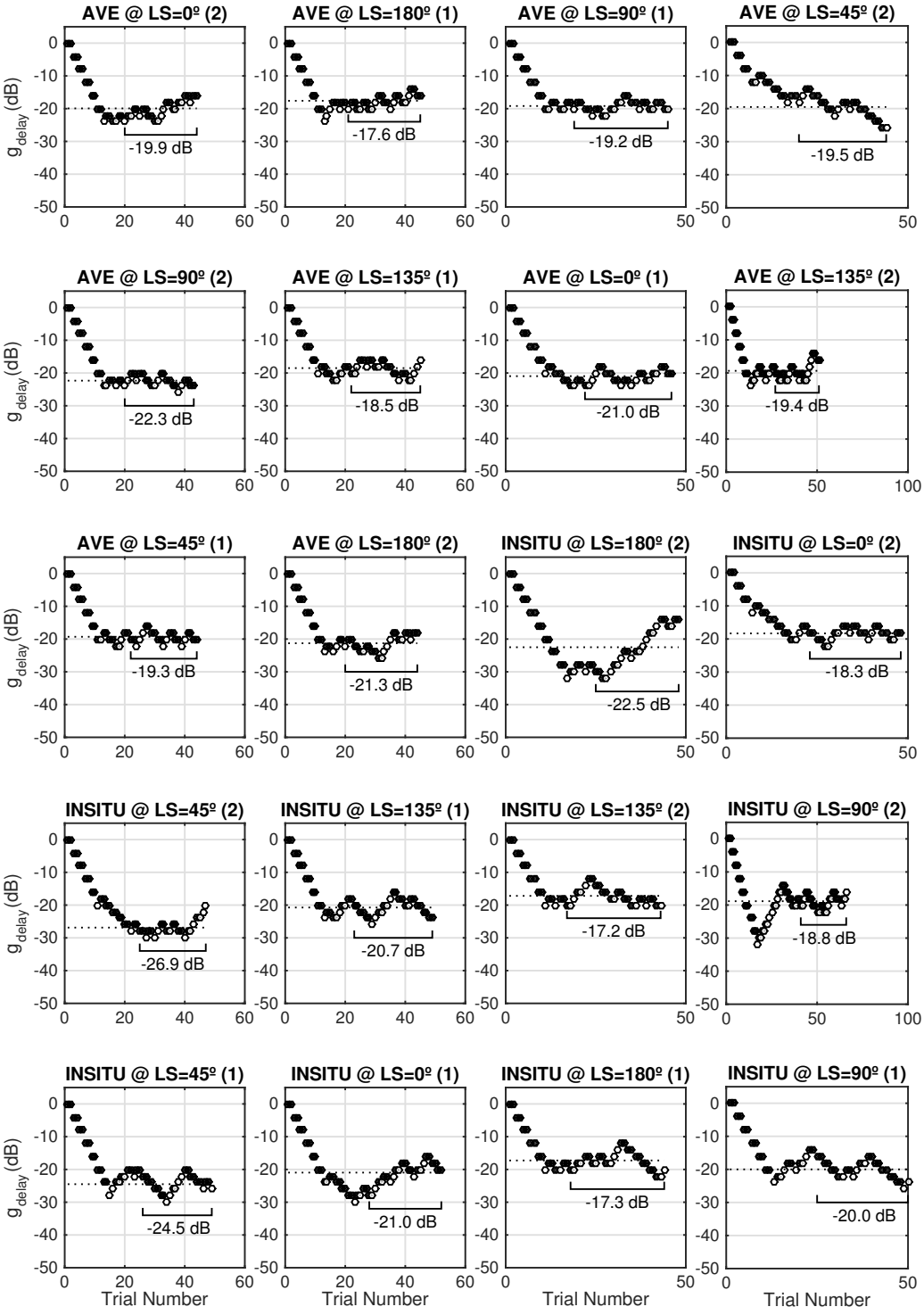


Figure B.4: CDT response data for colouration validation experiment B: Participant 4

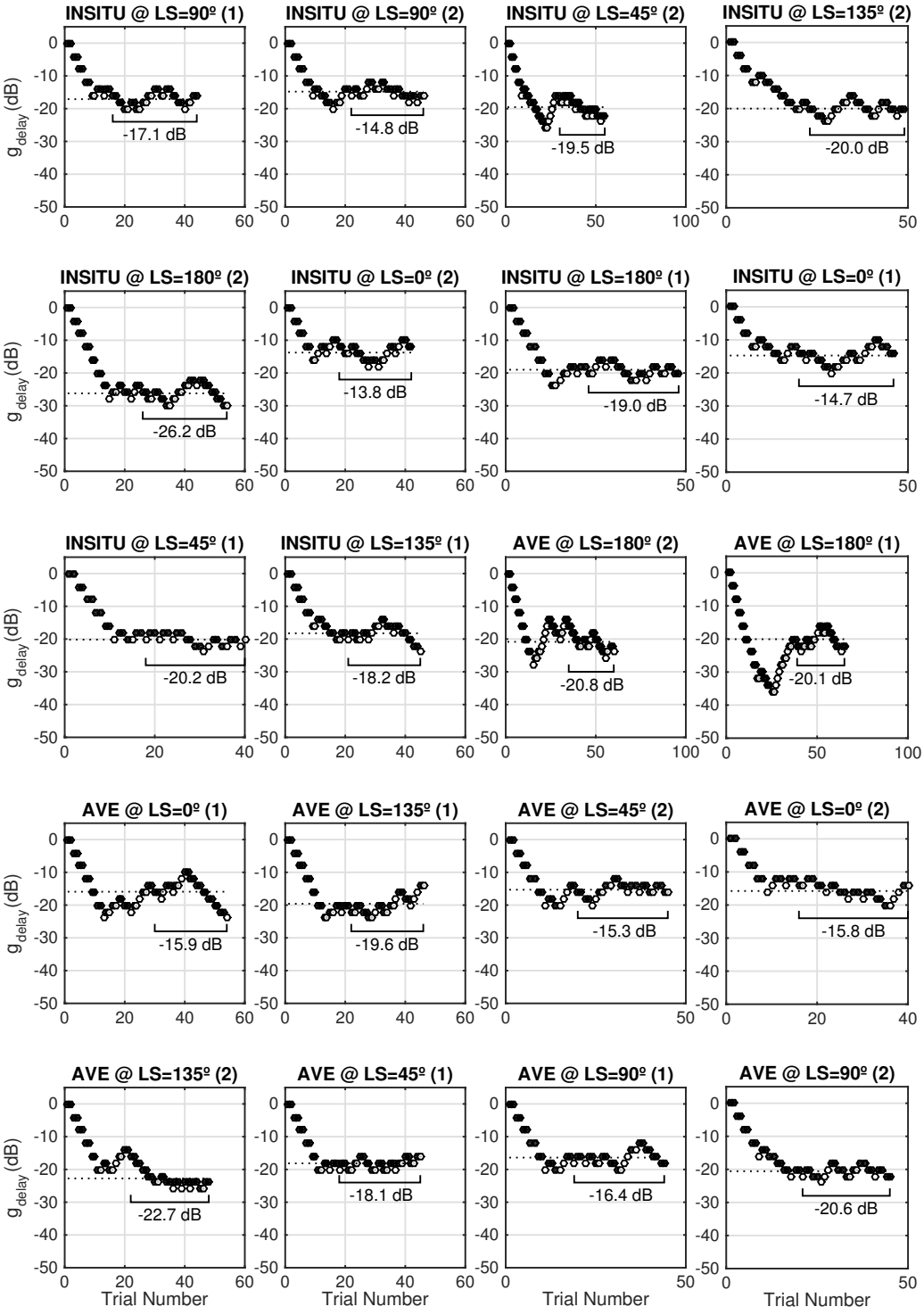


Figure B.5: CDT response data for colouration validation experiment B: Participant 5

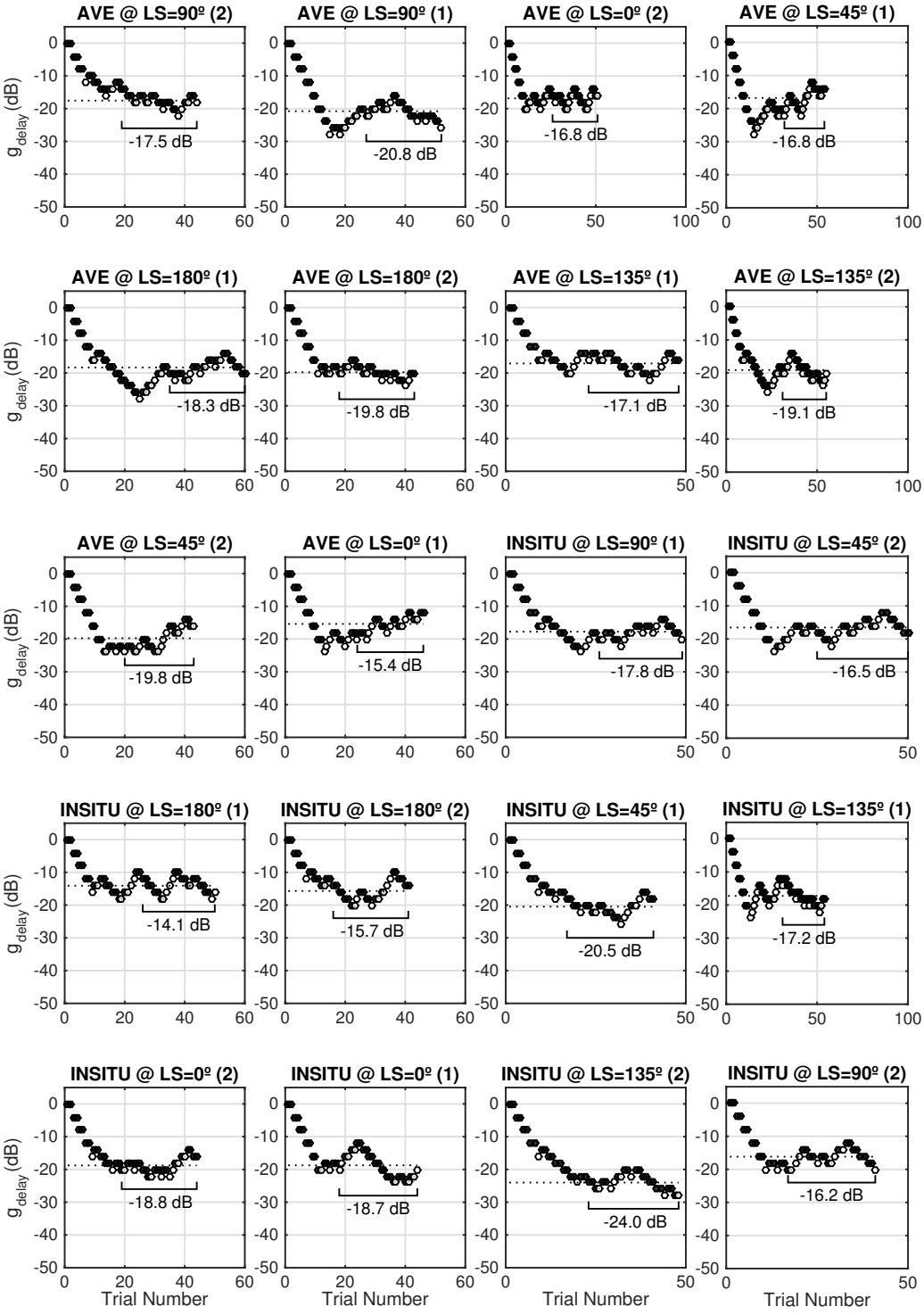


Figure B.6: CDT response data for colouration validation experiment B: Participant 6

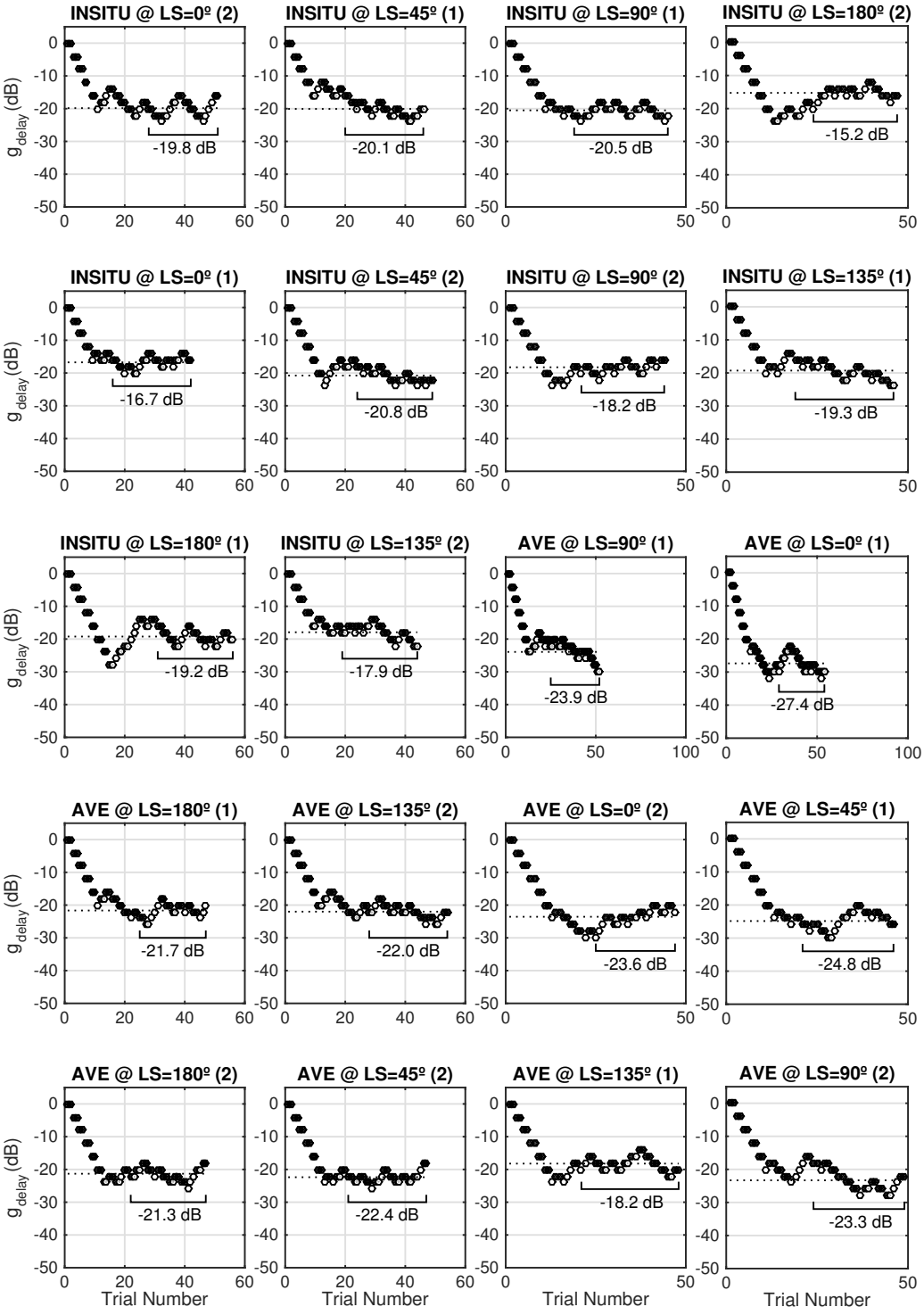


Figure B.7: CDT response data for colouration validation experiment B: Participant 7

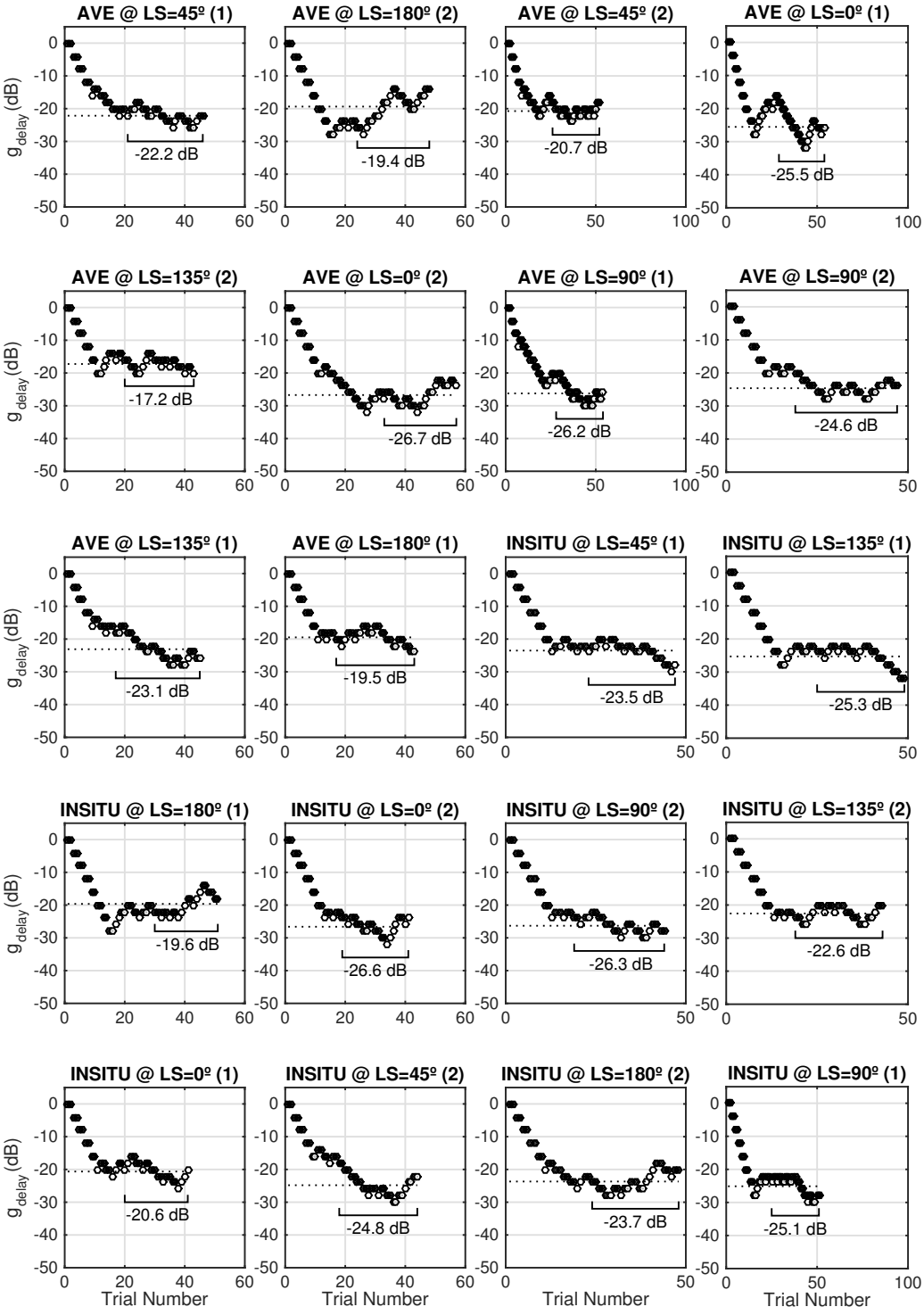


Figure B.8: CDT response data for colouration validation experiment B: Participant 8

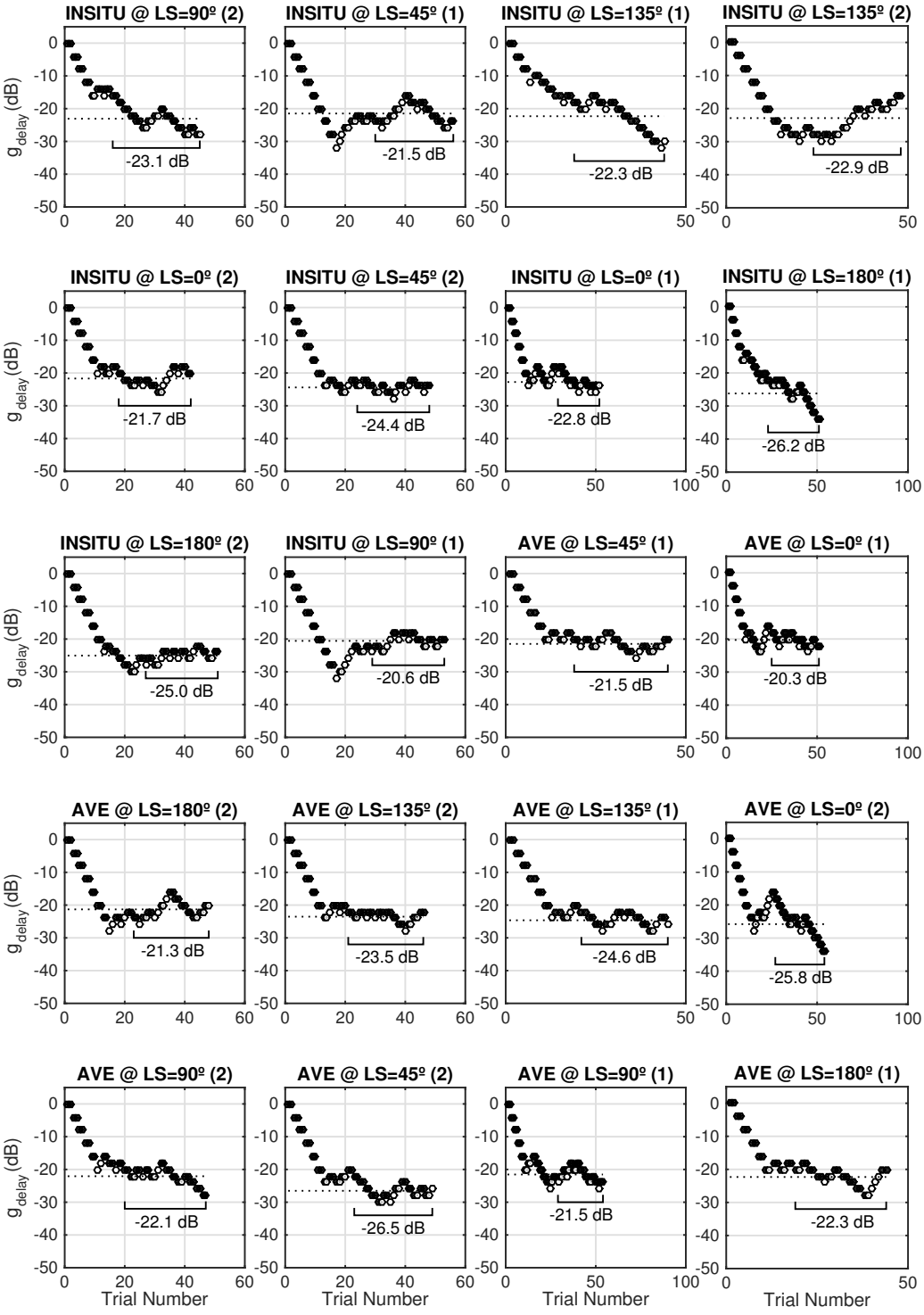


Figure B.9: CDT response data for colouration validation experiment B: Participant 9

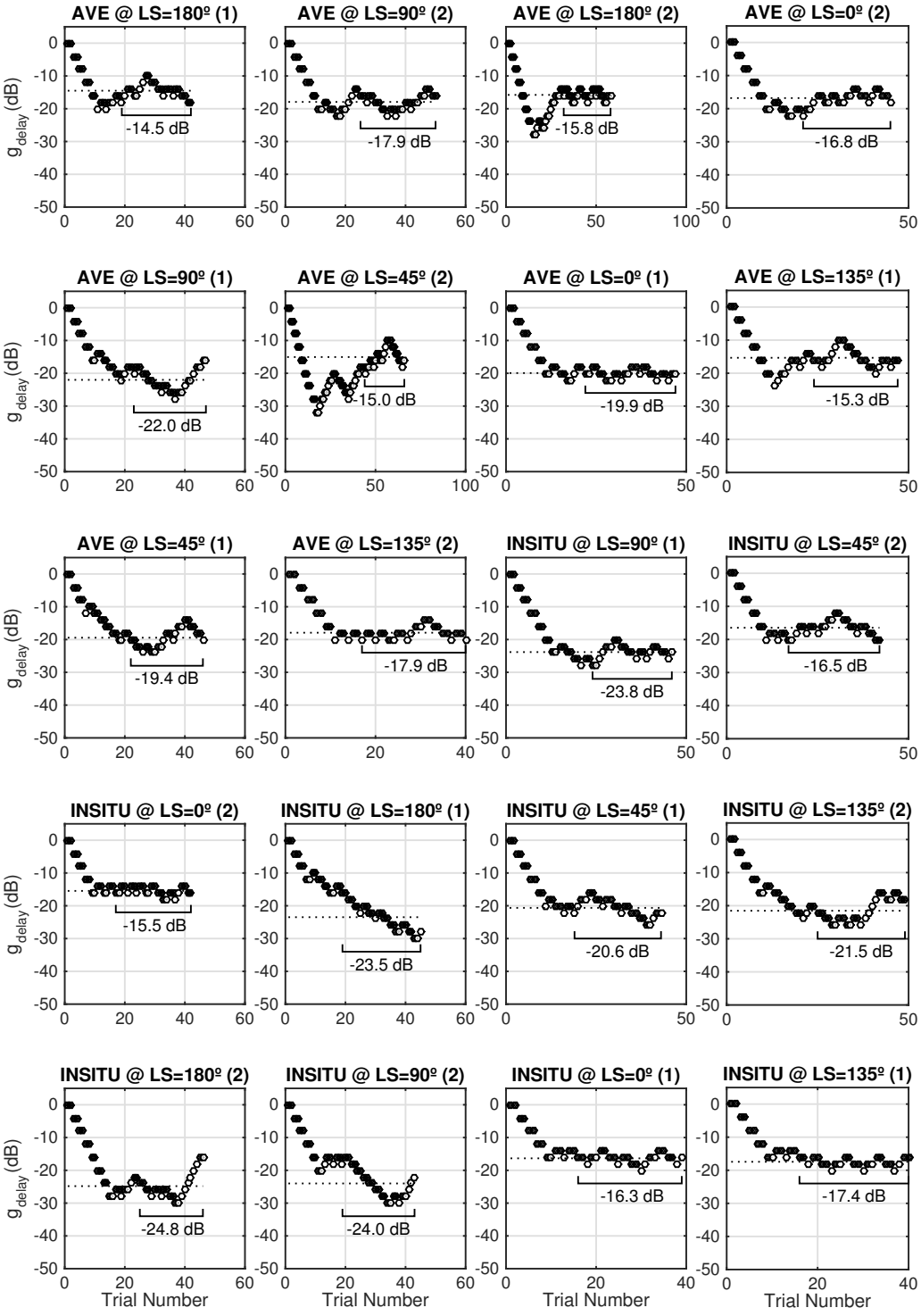


Figure B.10: CDT response data for colouration validation experiment B: Participant 10

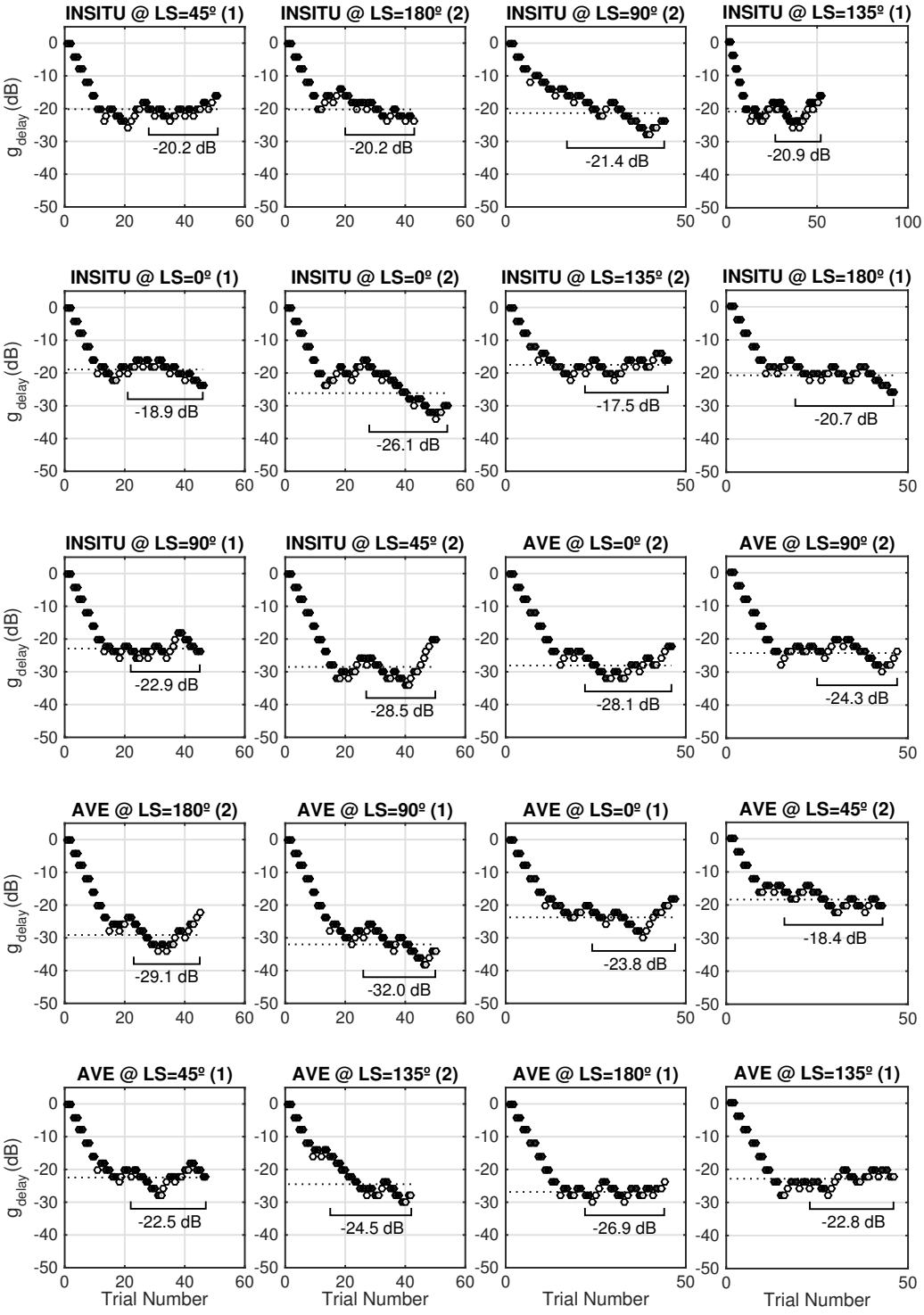


Figure B.11: CDT response data for colouration validation experiment B: Participant 11

APPENDIX C

Graphical user interface components used in the direct-scaling experiment from Chapter 9.4.

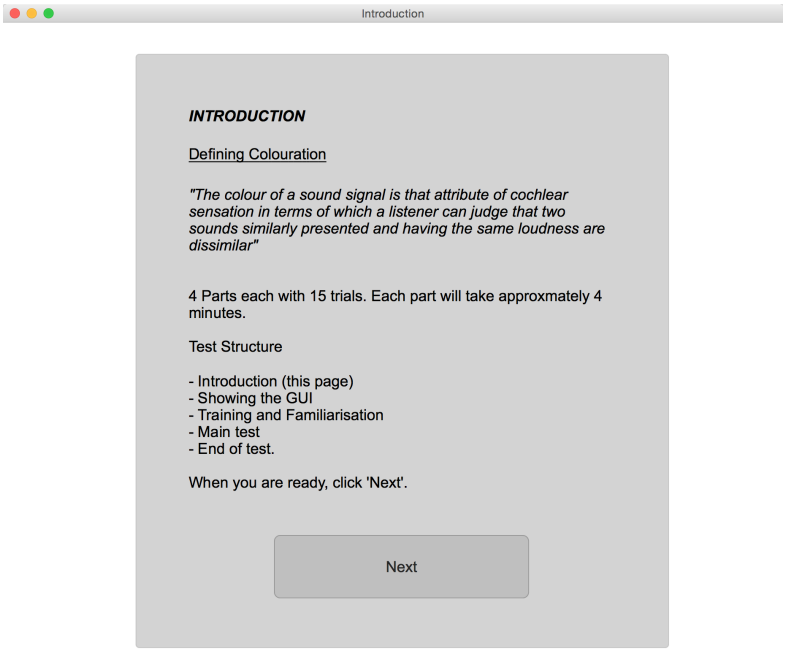


Figure C.1: Section 1: Introduction.

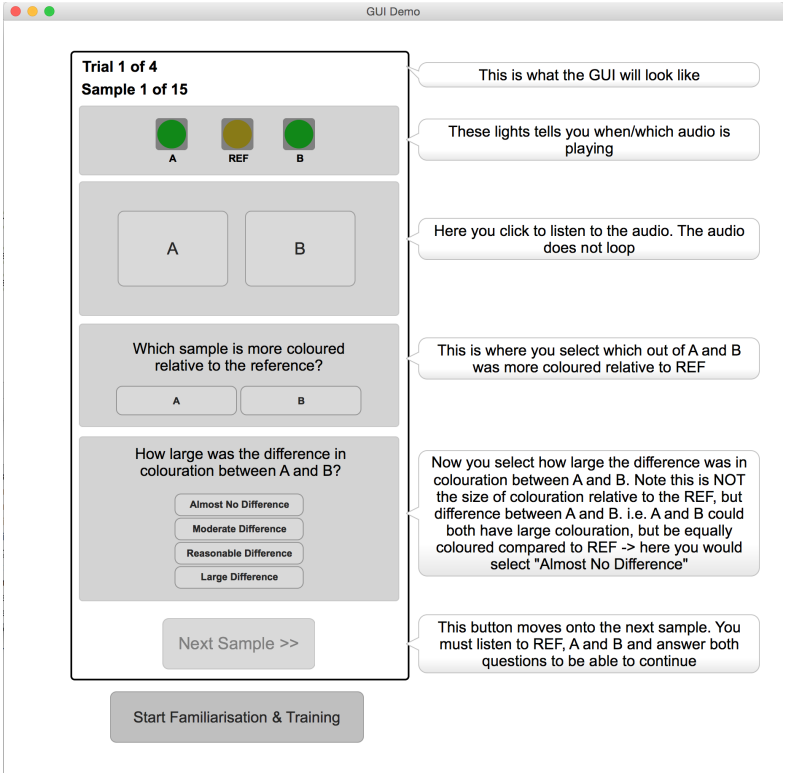


Figure C.2: Section 2: Demonstration of the GUI.

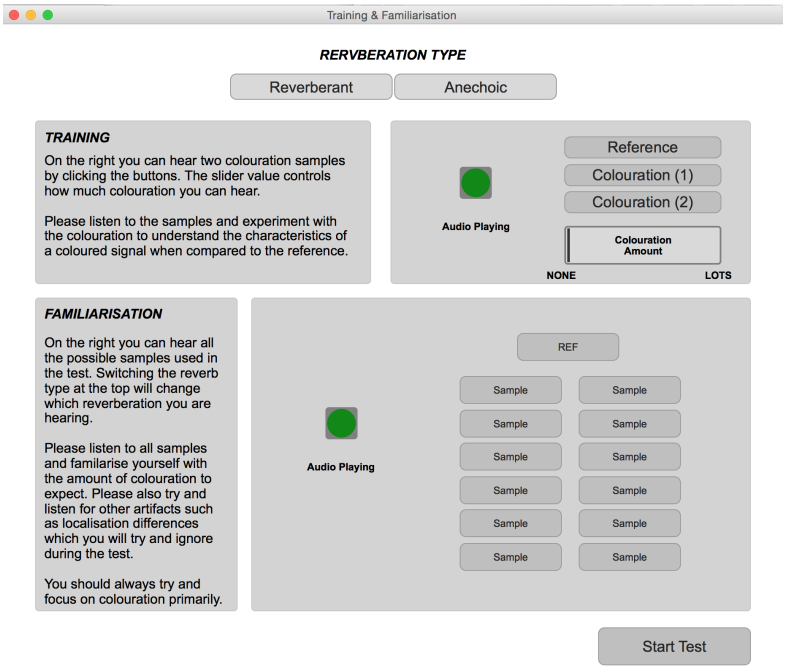


Figure C.3: Section 3: Training and familiarisation of sound samples.

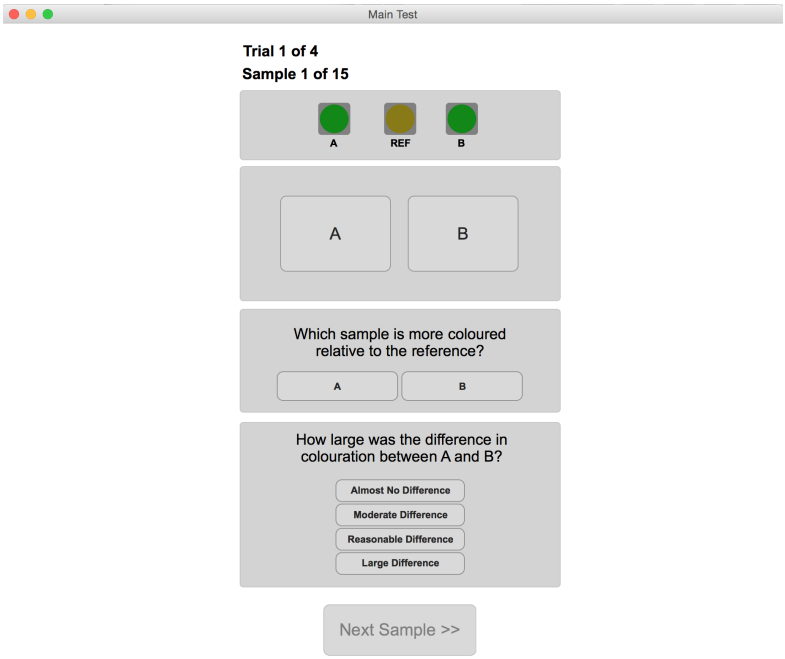


Figure C.4: Section 4: Main test.

APPENDIX D

Gain Coefficients for Panning Methods used in Sections 6, 9.3 and 9.4.

Table D.1: A table showing loudspeaker positions and gain coefficients for the panning methods used in Chapter. 6.

Combination Number	Loudspeaker Angles (°)	Gain Coefficients
1	30, 330	0.9391, 0.3437
2	0, 90, 180, 270	0.8536, 0.3536, -0.1464, 0.3536
3	110	1.0000
4	0, 45, 90, 135, 180, 225, 270, 315	0.2138, 0.8409, 0.2138, -0.0397, 0.1398, -0.0544, 0.1398, -0.0397
5	0, 45, 90, 135, 180, 225, 270, 315	0.0000, 0.0000, 0.9571, 0.0000, 0.2898, 0.0000, 0.0000, 0.0000
6	0, 90, 180, 270	0.0031, 0.9459, 0.4257, 0.0396
7	315	1.0000
8	0, 30, 110, 250, 330	0.0000, 0.0000, 0.0000, 0.7071, 0.7071
9	0, 30, 110, 250, 330	0.0000, 0.0000, 0.6604, 0.7509, 0.0000
10	45, 135, 225, 315	0.7071, 0.0000, 0.0000, 0.7071

Table D.2: A table showing loudspeaker positions and gain coefficients for the panning methods used in Section. 9.3.

Panning Method (short ID)	Loudspeaker Angles ()	Gain Coefficient	
Ao3s8	0, 45, 90, 135, 180, 225, 270, 315	max rV	0.6764, 0.5770, -0.1975, 0.1001, -0.0434, -0.0056, 0.0645,-0.1715
		max rE	0.7429, 0.6685, -0.0202, 0.0087, -0.0063, 0.0064, -0.0093, 0.0234
Ao1s4	0, 90, 180, 270	max rV	0.7198, 0.4210, -0.2198, 0.0790
		max rE	0.8234, 0.5246, -0.1163, 0.1825
VbITU	0, 30	0.4527, 0.8917	
VbST	30, 330	0.9753, 0.2211	

Table D.3: A table showing loudspeaker positions and gain coefficients for the panning methods used in Section. 9.4.

Loudspeaker Angles (°)	Ambisonics		VBAP
	max rV (LF)	max rE (HF)	
0	0.0835	0.0000	0.0000
45	-0.1871	0.0000	0.0000
90	0.6284	0.7071	0.7071
135	0.6284	0.7071	0.7071
180	-0.1871	0.0000	0.0000
225	0.0835	0.0000	0.0000
270	-0.0249	0.0000	0.0000
315	-0.0249	0.0000	0.0000

Bibliography

Ahrens, J. (2012). *Analytical Methods of Sound Field Synthesis*. Springer.

AKG and Ircam (2002). Listen HRTF Database. `http://recherche.ircam.fr/equipes/salles/listen/`. Accessed: 2016-10-24.

Algazi, V. R., Duda, R. O., and Thompson, D. M. (2004a). Motion-Tracked Binaural Sound. *Journal of the Audio Engineering Society*, 52(11).

Algazi, V. R., Duda, R. O., and Thompson, D. M. (2004b). Motion-Tracked Binaural Sound. In *116th Audio Engineering Society Convention*, Berlin, Germany.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). The CIPIC HRTF Database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, New York, USA.

American Standards Association and Acoustical Society of America (1960). American standard acoustical terminology. Technical report, American Standards Association.

Andreopoulou, A., Begault, D. R., and Katz, B. F. G. (2015). Inter-Laboratory Round Robin HRTF Measurement Comparison. *IEEE Journal of Selectic Topics in Signal Processing*, 9(5):895–906.

Ashby, T., Brookes, T., and Mason, R. (2014). Towards a Head-Movement-Aware Spatial Localisation Model : Elevation. In *The 21st International Congress on Sound and Vibration*, Beijing, China.

Atal, B. S. and Schroeder, M. R. (1962). Perception of Coloration in Filtered Gaussian Noise - Short-time Analysis by the Ear. In *Fourth International Cogress on Acoustics*, Copenhagen, Denmark.

Augspurger, G. L. (1990). Loudspeakers in Control Rooms and Living Rooms. In *AES 8th International Conference*, pages 171–178, Washington, D.C., USA.

Bamford, J. S. (1995). *An Analysis of Ambisonic Sound Systems of First and Second Order by*. Master of science, University of Waterloo.

Barron, M. (1971). The subjective effects of first reflections in concert halls - The need for lateral reflections. *Journal of Sound and Vibration*, 15:475–494.

Bates, E., Kearney, G., Boland, F., and Furlong, D. (2007a). Localization Accuracy of Advanced Spatialization Techniques in Small Concert Halls. *Journal of the Acoustical Society of America*, 121(5):3069.

Bates, E., Kearney, G., and Furlong, D. (2007b). Localization accuracy of advanced spatialisation techniques in small concert halls. *Journal of the Acoustical Society of America*, 121(5):3069–3070.

Bech, S. (1995). Timbral aspects of reproduced sound in small rooms. I. *Journal of the Acoustical Society of America*, 97(3):1717–1726.

Bech, S. (1996). Timbral aspects of reproduced sound in small rooms . II. *Journal of the Acoustical Society of America*, 99(6):3539–3549.

Bech, S. and Zacharov, N. (2006). *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley and Sons, Ltd.

Begault, D. R. (1992). Perceptual Effects of Synthetic Reverberationon Three-Dimensional Audio Systems. *Journal of the Audio Engineering Society*, 40(11):895–904.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatil Perception of a Virtual Speech Source. *Journal of the Audio Engineering Society*, 49(10).

Benjamin, E., Heller, A., and Lee, R. (2010). Design of Ambisonic Decoders for Irregular Arrays of Loudspeakers by Non-Linear Optimization. In *129th Audio Engineering Society Convention*, San Francisco, USA.

Benjamin, E., Lee, R., and Heller, A. J. (2006). Localization in Horizontal-Only Ambisonic Systems (revision). In *Audio Engineering Society Convention*, pages 1–13, San Francisco, USA.

Beranek, L. L. (1996). Acoustics and musical qualities. *The Journal of the Acoustical Society of America*, 99(5):2647.

Berens, P. (2009). CircStat: A MATLAB Toolbox for Circular Statistics. *Journal of Statistical Software*, 31(10).

Berkhout, A. (1988). A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36(1):977–995.

Berkhout, A. J., Vries, D. D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778.

Berkley, D. A. (1980). Normal Listeners in Typical Rooms - Reverberation Perception, Simulation and Reduction. In Studebaker, G. A. and Hochberg, editors, *Acoustical Factors Affecting Hearing Aid Performance*, pages 3–24. University Park Press, Baltimore, MD.

Bernschütz, B. (2013). A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100. In *Fortschritte der Akustik AIA-DAGA 2013*.

Bernstein, L. R. and Trahiotis, C. (1996). The normalized correlation: accounting

for binaural detection across center frequency. *The Journal of the Acoustical Society of America*, 100(6):3774–84.

Bernstein, L. R., van de Par, S., and Trahiotis, C. (1999). The normalized interaural correlation: accounting for NoS pi thresholds obtained with Gaussian and 'low-noise' masking noise. *The Journal of the Acoustical Society of America*, 106(2):870–6.

Bertet, S., Daniel, J., Parizet, E., and Warusfel, O. (2009). Influence of Microphpne and Loudspeaker Setup on Perceived Higher Order Ambisonics Reproduced Sound Field. In *Ambisonics Symposium*, Graz, Austria.

Bilsen, F. A. (1968). Thresholds of Perception of Repetition Pitch. Conclusions Concerning Coloration in Room Acoustics and Correlation in the Hearing Organ. *Acoustica*, 19(1).

Bilsen, F. a. (1977). Pitch of noise signals: evidence for a 'central spectrum'. *The Journal of the Acoustical Society of America*, 61(1):150–161.

Bilsen, F. A. and Ritsma, R. J. (1970). Some Parameters Influencing the Perceptibility of Pitch. *Journal of the Acoustical Society of America*, 47(2):469–475.

Blauert, J. (2001). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, London, UK, revised edition.

Block, H. D. and Marschak, J. (1960). Random Orderings and Stochastic Theories of Responses. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, chapter 2, pages 97–132. Stanford University Press.

Blumlein, A. D. (1931). Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems. British Patent 394325.

Braasch, J., Clapp, S., Parks, A., Pastore, T., and N., X. (2013). A Binaural

Model that Analyses Acoustic Spaces and Stereophonic Reproduction Systems by Utilizing Head Rotations. In Blauert, J., editor, *The Technology of Binaural Listening*, pages 201–223. Springer Berlin Heidelberg, Berlin, Heidelberg.

Brimijoin, W. O., Boyd, A. W., and Akeroyd, M. a. (2013). The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE*, 8(12):1–12.

Bronkhorst, A. W. (1995). Localization of real and virtual sound sources. *Journal of the Acoustical Society of America*, 98(5):2542–2553.

Brüggen, M. (2001). Coloration and Binaural Decoloration in Natural Environments. *Acta Acustica: Acustica*, 87:400–406.

Bruijn, W. D. (2004). *Application of Wave Field Synthesis in Videoconferencing*. PhD thesis, Delft University of Technology.

Buchholz, J. M. (2007). Characterizing the monaural and binaural processes underlying reflection masking. *Hearing Research*, 232:52–66.

Buchholz, J. M. (2011). A quantitative analysis of spectral mechanisms involved in auditory detection of coloration by a single wall reflection. *Hearing Research*, 277(1-2):192–203.

Buchholz, J. M., Mourjopoulos, J., and Blauert, J. (2001). Room masking: Understanding and Modelling the Masking of Room Reflections. In *110th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

Bunning, H. and Wilkens, H. (1979). *Mehrdimensionale Verknüpfung der Höreindrücke von Lautsprechern mit deren physikalischen Daten*. Abschlussbericht bo 421/7, Heinrich-Hertz-Institut für Nachrichtentechnik, Berlin, Germany.

Burgtorf, W. (1961). Untersuchungen zur Wahrnehmbarkeit verzögerter Schallsignale. *Acustica*, 11.

Butler, R. a. and Belendiuk, K. (1977). Spectral cues utilized in the localization of sound in the median sagittal plane. *The Journal of the Acoustical Society of America*, 61(5):1264–1269.

Carlile, S., Leong, P., and Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners. *Hearing research*, 114:179–196.

Choisel, S. and Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference. *The Journal of the Acoustical Society of America*, 121:388–400.

Conetta, R. (2011). *Towards the automatic assessment of spatial quality in the reproduced sound environment*. PhD thesis, University of Surry.

Cooper, D. H. and Shiga, T. (1972). Discrete-Matrix Multichannel Stereo. *Journal of the Audio Engineering Society*, 20(5).

Craven, P. G. (2003). Continuous Surround Panning for 5-Speaker Reproduction. In *24th Internation Audio Engineering Society Conference on Multichannel Audio*, pages 1–6, Banff, Canada.

Daniel, J. (2001). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. PhD thesis, University of Paris.

Daniel, J., Nicol, R., and Moreau, S. (2003). Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging. In *Audio Engineering Society Convention*, Amsterdam, The Netherlands. Audio Engineering Society; 1999.

David, H. A. (1959). The Method of Paired Comparisons. In *Proceedings of the*

Fifth Conference on the Design of Experiments in Army Research Developments and Testing, volume 60-2.

Denyer, M. C., Gilchrist, N. H. C., Kallaway, M. J., and Palmer, R. J. (1979). Digital radio links for stereophonic outside broadcasts. Technical report, Research Department, Engineering Division, BBC.

Dietz, M., Ewert, S. D., and Hohmann, V. (2009). Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities. *The Journal of the Acoustical Society of America*, 125(3):1622–35.

Dietz, M., Ewert, S. D., and Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605.

Dolby Laboratories (2000). Frequently Asked Questions about Dolby Digital. Technical report, Dolby Laboratories.

EBU (2008). Sound Quality Assessment Material recordings for subjective tests EBU- TECH 3253. Technical Report September, EBU UER, Geneva, Switzerland.

Elko, G., Kubli, R., and Meyer, J. (2005). Audio System Based on at Least Second-Order Eigenbeams.

Erbes, V., Schultz, F., Lindau, A., and Weinzierl, S. (2012). An extraaural headphone system for optimized binaural reproduction. In *DAGA: The 38TH German Annual Conference on Acoustics*, pages 17–18, Darmstadt, Germany.

Ericson, M. A. and McLinley, R. L. (2001). The intelligibility of multiple talkers separated spatially in noise. In Gilkey, R. and Anderson, T., editors, *Binaural and Spatial Hearing in Real and Virtual Environments*, chapter 32, pages 701–724. Lawrence Erlbaum, New Jersey, USA.

Estrella, J. (2011). Real Time Individualization of Interaural Time Differences for Dynamic Binaural Synthesis. *Ak.Tu-Berlin.De*.

Faller, C. and Merimaa, J. (2004). Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116(5):3075.

Farina, A. (2007). Advancements in impulse response measurements by sine sweeps. In *122nd Audio Engineering Society Convention*, Vienna, Austria.

Fastl, H. and Zwicker, E. (2007). *Psychoacoustics: Facts and Models*. Springer-Verlag, Berlin Heidelberg, 3 edition.

Fels, J., Oberem, J., and Masiero, B. (2013). Experiments on authenticity and naturalness of binaural reproduction via headphones. In *Proceedings of Meetings on Acoustics*, volume 19, Montreal, Canada.

Fisher, G. H. and Freedman, S. J. (1968). The Role of the Pinna in Auditory Localization. *Journal of Auditory Research*, 8(1):15–26.

Fisher, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge University Press.

Frank, M. (2013). *Phantom Sources using Multiple Loudspeakers in the Horizontal Plane*. PhD thesis, University of Music and Performing Arts. Graz, Austria.

Frank, M., Zotter, F., and Sontacchi, A. (2008). Localization Experiments Using Different 2D Ambisonics Decoders. In *VDT International Convention*.

Gabrielsson, A. (1979). Dimension analyses of perceived sound quality of sound-reproducing systems. *Scandinavian Journal of Psychology*, 20(1).

Gabrielsson, A. and Sjögren, H. (1979). Perceived sound quality of sound-

reproducing systems. *The Journal of the Acoustical Society of America*, 65(4):1019–33.

Gardner, B. and Martin, K. (1994). HRTF Measurements of a KEMAR Dummy-Head Microphone. Technical report, MIT Media Laboratory, Cambridge, MA, USA.

Gaskell, P. S. and Ratliff, P. A. (1977). QUADRAPHONY: developments in Matrix H decoding. Technical report, BBC.

Geier, M., Ahrens, J., and Spors, S. (2008). The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods. In *124th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

Gerzon, M. A. (1972). Periphony : With-Height Sound Reproduction. *Journal of the Audio Engineering Society*, 21(1):2–10.

Gerzon, M. A. (1974). Surround-sound psychoacoustics. *Wireless World*.

Gerzon, M. A. (1980). Practical Periphony: The Reproduction of Full-Sphere Sound. In *65th Audio Engineering Society Convention*, London, UK.

Gerzon, M. A. (1985). Ambisonics in Multichannel Broadcasting and Video. *Journal of the Audio Engineering Society*, 33(11):859–871.

Gerzon, M. A. (1992a). General Metatheory of Auditory Localisation. In *92nd Audio Engineering Society Convention*, Vienna, Austria.

Gerzon, M. A. (1992b). Panpot Laws for Multispeaker Stereo. In *92nd Audio Engineering Society Convention*, Vienna, Austria.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–38.

Goupell, M. J. and Hartmann, W. M. (2006). Interaural fluctuations and the

detection of interaural coherence: bandwidth effects. *Journal of the Acoustical Society of America*, 119(6):3971–3986.

Grantham, D. W., Hornsby, B. W. Y., and Erpenbeck, E. A. (2003). Auditory spatial resolution in horizontal, vertical, and diagonal planes. *The Journal of the Acoustical Society of America*, 114(2):1009–1022.

Greenwood, D. D. (1961). Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane. *Journal of the Acoustical Society of America*, 33:1344.

Gulliksen, H. (1956). A Least Squares Solution For Paired Comparisons with Incomplete Data. *Psychometrika*, 21(2):125–134.

Halmrast, T. (2001). Sound coloration from (very) early reflections. *Journal of the Acoustical Society of America*, 109(5):2303.

Hamasaki, K., Hiyama, K., Nishiguchi, T., and Ono, K. (2004a). Advanced Multichannel Audio Systems with Superior Impression of Presence and Reality. In *Audio Engineering Society Convention*, pages 1–12, Berlin, Germany.

Hamasaki, K., Komiyama, S., Okubo, H., Hiyama, K., and Hatano, W. (2004b). 5.1 and 22.2 Multichannel Sound Productions Using an Integrated Surround Sound Panning System. In *117th International Audio Engineering Society Convention*, San Francisco, USA.

Hammershøi, D. and Møller, H. (1996). Sound transmission to and within the human ear canal. *The Journal of the Acoustical Society of America*, 100(1):408–27.

Harma, A., Jakka, J., Tikander, M., Karjalainen, M., Lokki, T., Hiipakka, J., and Lorho, G. (2004). Augmented Reality Audio for Mobile and Wearable Appliances. *Journal of the Audio Engineering Society*, 52(6).

Härmä, A., Park, M., and Kohlrausch, A. (2014). Predicting the subjective

evaluation of spatial audio systems. In *International Conference on Spatial Audio*, Erlangen, Germany.

Hartmann, W. M. and Wittenberg, A. (1996). On the externalization of sound images. *Journal of the Acoustical Society of America*, 99(6):3678–3688.

Heller, A. J., Lee, R., and Benjamin, E. (2008). Is My Decoder Ambisonic? In *125th International Audio Engineering Society Convention*, San Francisco, USA.

Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J. (2015). MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5).

Hiekkanen, T., Mäkivirta, A., and Karjalainen, M. (2009). Virtualized Listening Tests for Loudspeakers. *Journal of the Audio Engineering Society*, 57(4):237–252.

Hollerweger, F. (2006). *Periphonic sound spatialization in multi-user virtual environments*. Msc, Institute of Electronic Music and Acoustics (IEM).

IFPI (2015). IFPI Digital Music Report 2015. Technical report, International Federation of the Phonographic Industry.

ISO (1975). ISO 389. Acoustics - standard reference zero for the calibration of pure-tone audiometers.

ITU (1997). Methods For The Subjective Assessment Of Small Impairments In Audio Systems Including Multichannel Sound Systems (Rec. ITU-R BS.1116-1). Technical report, ITU-R.

ITU-R (2003). Method For The Subjective assessment Of Intermediate Quality Level Of Coding (Rec. ITU-R BS.1534-1). Technical report, ITU-R.

ITU-R (2012a). Algorithms to measure audio programme loudness and true-peak audio level BS Series (Rec. ITU-R BS.1770-3). Technical report, ITU-R.

ITU-R (2012b). Multichannel stereophonic sound system with and without accompanying picture (Rec. ITU-R BS.775-3). Technical report, ITU-R.

ITU-T (2009). Use of head and torso simulator (HATS) for hands-free and handset terminal testing (Rec. ITU-T P.581). Technical report, ITU-T.

Jessel, M. (1973). *Acoustique Théorique - propagation et holophonie*. PhD thesis, Paris: Masson et Cie.

Johnston, J. D. and Lam, Y. H. V. (2000). Perceptual Soundfield Reconstruction. In *109th International Audio Engineering Society Convention*, Los Angeles, USA.

Jot, J. M., Larcher, V., and Warusfel, O. (1995). Digital signal processing issues in the context of binaural and transaural stereophony. In *98th Audio Engineering Society Convention*.

Kamekawa, T. (2006). The Effect on Spatial Impression of the Configuration and Directivity of Three Frontal Microphones Used in Multi-Channel Stereophonic Recording. In *28th International Audio Engineering Society Conference*, pages 1–8, Pitea, Sweden.

Karjalainen, M. and Pautero, T. (2001). Frequency-dependent signal windowing. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 35–38, New York, USA.

Kates, J. M. (1985). A central spectrum model for the perception of coloration in filtered Gaussian noise. *Journal of the Acoustical Society of America*, 77(4):1529–1534.

Killion, M. C. and Dallos, P. (1979). Impedance matching by the combined effects

of the outer and middle ear. *Journal of the Acoustical Society of America*, 66(2):599–602.

Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3):1637–47.

Koenig, A. H., Berkley, D. A., Curtis, T. H., and B, A. J. (1975). Magnitude of JNDs for diotic and dichotic perception of spectrally colored noise. *Journal of the Acoustical Society of America*, 58(S55).

Landone, C. and Sandler, M. (2000). Applications of binaural processing to surround sound reproduction in large spaces. In *IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland. IEEE.

Langendijk, E. H. A. and Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual. *Journal of the Acoustical Society of America*, 107(1):528–537.

Letowski, T. and Letowski, S. (2011). Localization Error : Accuracy and Precision of Auditory Localization. In Strumillo, P., editor, *Advances in Sound Localization*, chapter 4. InTech, 1st edition.

Letowski, T. R. and Letowski, S. T. (2012). Auditory Spatial Perception : Auditory Localization. Technical Report May, Army Research Laboratory.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America*, 49:467–477.

Lewald, J., Dörrscheidt, G. J., and Ehrenstein, W. H. (2000). Sound localization with eccentric head position. *Behavioural brain research*, 108(2):105–125.

Licklider, J. C. R. (1956). Auditory Frequency Analysis. *Information Theory*.

Lindau, A. (2009). The Perception of System Latency in Dynamic Binaural Synthesis. In *Proceedings of NAG/DAGA*, pages 1063–1066, Rotterdam.

Lindau, A. (2014). *Binaural Resynthesis of Acoustical Environments. Technology and Perceptual Evaluation.* PhD thesis, Technische Universität Berlin.

Lindau, A., Kosanke, L., and Weinzierl, S. (2012). Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses. *Journal of the Audio Engineering Society*, 60(11):887–898.

Lindau, A. and Weinzierl, S. (2012). Assessing the Plausibility of Virtual Acoustic Environments. *Acta Acustica united with Acustica*, 98(5):804–810.

Linkwitz, S. (2003). Binaural Audio in the Era of Virtual Reality: A digest of research papers presented at recent AES conventions. *Journal of the Audio Engineering Society*, 51(11):1066–1072.

Linkwitz, S. H. (1976). Active Crossover Networks for Noncoincident Drivers. *Journal of the Audio Engineering Society*, 24(1):2–8.

Litovsky, R. Y., Colburn, H. S., Yost, W. a., and Guzman, S. J. (1999). The precedence effect. *The Journal of the Acoustical Society of America*, 106(4 Pt 1):1633–54.

Lochner, J. and Burger, J. (1958). The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech. *Acta Acustica united with Acustica*, 8:1–10.

Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5):2219–2236.

Majdak, P., Goupell, M. J., and Laback, B. (2010). 3-D Localization of Virtual

Sound Sources: Effects of Visual Environment, Pointing Method, and Training. *Atten Percept Psychophys*, 72(2):454–469.

Makous, J. C. and Middlebrooks, J. C. (1990a). Two-dimensional sound localization by human listeners. *Journal of the Acoustical Society of America*, 87(5):2188–2200.

Makous, J. C. and Middlebrooks, J. C. (1990b). Two-dimensional sound localization by human listeners. *The Journal of the Acoustical Society of America*, 87(5):2188–2200.

Malham, D. G. (1992). Experience With Large Area 3D Ambisonic Sound Systems. In *Proceedings of the Institute of Acoustics 14*, pages 209–216, St. Albans, UK.

Malham, D. G. (2005). Second and third order ambisonics - the furse-malham set. `http://www.york.ac.uk/inst/mustech/3d_audio/secondor.html`. Accessed: 2016-10-24.

Masiero, B. and Fels, J. (2011). Perceptually Robust Headphone Equalization for Binaural Reproduction. In *130th Audio Engineering Society Convention*, pages 1–7, London, UK.

Melchior, F., Heusinger, U., and Liebetrau, J. (2011). Perceptual evaluation of a spatial audio algorithm based on wave field synthesis using a reduced number of loudspeakers. In *131st Audio Engineering Society Convention*, New York, USA.

Melchior, F., Marston, D., Pike, C., Satongar, D., and Lam, Y. W. (2014). A Library of Binaural Room Impulse Responses and Sound Scenes for Evaluation of Spatial Audio Systems. In *40th Annual German Congress on Acoustics*, Oldenburg.

Menzer, F. and Faller, C. (2009). Investigations on modeling BRIR tails with

filtered and coherence-matched noise. In *127th Audion Engineering Society Convention*, New York, USA.

Merimaa, J. (2006). *Analysis, synthesis and perception of spatial sound - binaural localization modeling and multichannel loudspeaker reproduction*. PhD thesis, Helsinki University of Technology.

Merimaa, J. (2009). Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis. In *127th Audion Engineering Society Convention*, New York, USA.

Merimaa, J. (2010). Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis, Part 2: Individual HRTFs. In *129th Audio Engineering Society Convention*, San Francisco, USA.

Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of the Acoustical Society of America*, 106(3):1480–1492.

Middlebrooks, J. C. (1999b). Individual differences in external-ear transfer functions reduced by scaling in frequency. *Journal of the Acoustical Society of America*, 106(3):1480–1492.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989). Directional sensitivity of sound-pressure levels in the human ear canal. *The Journal of the Acoustical Society of America*, 86(1):89–108.

Mills, A. W. (1958). On the Minimum Audible Angle. *Journal of the Acoustical Society of America*, 30(4):237–246.

Minnaar, P. (2010). Enhancing music with virtual sound sources. *The Hearing Journal*, 63(9):38–43.

Minnaar, P., Christensen, F., Moller, H., Olesen, S. K., and Plogsties, J. (1999).

Audibility of All-Pass Components in Binaural Synthesis. In *106th Audio Engineering Society Convention*, pages 113–, Munich, Germany.

Minnaar, P., Olesen, S. K., Christensen, F., and Møller, H. (2001). Localization with Binaural Recordings from Artificial and Human Heads. *Journal of the Audio Engineering Society*, 49(5):323—-336.

Møller, H. (1992). Fundamentals of Binaural Technology. *Applied Acoustics*, 36(December 1991):171–218.

Møller, H., Hammershøi, D., boje Jensen, C., and Sørensen, M. F. (1995a). Transfer Characteristics of Headphones Measured on Human Ears. *Journal of the Audio Engineering Society*, 43(4).

Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. (1999). Evaluation of Artificial Heads in Listening Tests. *Journal of the Audio Engineering Society*, 47(3):83–100.

Møller, H., Sørensen, M. F., boje Jensen, C., and Hammershøi, D. (1996). Binaural technique: Do We Need Individual Recordings? *Journal of the Audio Engineering Society*, 44(6):451–469.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995b). Head-Related Transfer-Functions of Human-Subjects. *Journal of the Audio Engineering Society*, 43(5):300–321.

Moore, A. H., Tew, A. I., and Nicol, R. (2007). Headphone transparification: A novel method for investigating the externalisation of binaural sounds. In *123rd Audio Engineering Society Convention*, New York, USA.

Moore, D. and Wakefield, J. (2007). The Design and Detailed Analysis of First Order Ambisonic Decoders for the ITU Layout loudspeakers being placed in

diametrically opposed. In *Audio Engineering Society Convention*, Vienna, Austria.

Morrissey, J. H. (1955). New method for the assignment of psychometric scale values from incomplete paired comparisons. *Journal of the Optical Society of America*, 45(5):373–378.

Nam, J., Abel, J. S., and Smith III, J. O. (2008). A Method for Estimating Interaural Time Difference for Binaural Synthesis. In *125th Audio Engineering Society Convention*, San Francisco, USA.

Neukom, M. (2007). Ambisonic Panning. In *123rd Audio Engineering Society Convention*, New York, USA.

Nicol, R. (2010). Sound Spatialization by Higher Order Ambisonics: Encoding and Decoding A Sound Scene in Practice from a Theoretical Point of View. In *Proc. of the 2nd International Symposium on Ambisonic and Spherical Acoustics*, Paris, France.

Novo, P. (2005). Communication Acoustics. In Blauert, J., editor, *Communication Acoustics*, chapter Auditory V. Springer-Verlag, Berlin, Germany.

Olive, S. and Toole, F. (1989). The Detection of Reflections in Typical Rooms. *Journal of the Audio Engineering Society*, 37(7/8):539–553.

Olive, S. E. and Schuck, P. L. (1995). The Variability of Loudspeaker Sound Quality Among Four Domestic-Sized Rooms. In *99th Audio Engineering Society Convention*, New York, USA.

Olive, S. E. and Toole, E. (1988). The Detection of Reflections in Typical Rooms. In *85th Audio Engineering Society Convention*.

Olive, S. E. and Welti, T. (2008). The Calibration and Validation of a Binaural

Room Scanning System Used for Subjective Evaluation of Automotive Audio Systems. *The Journal of the Acoustical Society of America*, 123(5):3246.

Olive, S. E. and Welti, T. (2009). Validation of a Binaural Car Scanning System for Subjective Evaluation of Automative Audio Systems. In *36th International Audio Engineering Society Conference*, pages 1–7, Dearborn, Michigan, USA.

Paquier, M. and Koehl, V. (2015). Discriminability of the placement of supra-aural and circumaural headphones. *Applied Acoustics*, 93:130–139.

Park, M. (2007). *Models of binaural hearing for sound lateralisation and localisation*. PhD thesis, University of Southampton, UK.

Perrett, S. and Noble, W. (1997). The contribution of head motion cues to localization of low-pass noise. *Perception & psychophysics*, 59(7):1018–1026.

Perrott, D. R. and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, 87(4):1728–1731.

Peters, N. (2010). *Developing Sound Spatialization Tools for Musical Applications with Emphasis on Sweet Spot and Off-Center Perception*. PhD thesis, McGill University, Montreal, Canada.

Peters, N. and McAdams, S. (2012). A perceptual analysis of off-center sound degradation in surround-sound reproduction based on geometrical properties. *The Journal of the Acoustical Society of America*, 131(4):3256.

Peters, N., Mcadams, S., and Braasch, J. (2007). Evaluating Off-Center Sound Degradation in Surround Loudspeaker Setups for Various Multichannel Microphone Techniques. In *123rd Audio Engineering Society Convention*, New York, USA.

Pike, C., Melchior, F., and Tew, T. (2014). Assessing the Plausibility of Non-

Individualised Dynamic Binaural Synthesis in a Small Room. In *55th AES International Conference*, pages 1–8, Helsinki, Finland.

Pulkki, V. (1997). Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *Journal of the Audio Engineering Society*, 45(6):456–466.

Pulkki, V. (2001). Coloration of Amplitude-Panned Virtual Sources. In *110th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

Pulkki, V. and Hirvonen, T. (2005). Localization of Virtual Sources in Multichannel Audio Reproduction. *IEEE Transactions on Speech and Audio Processing*, 13(1):105–119.

Pulkki, V. and Karjalainen, M. (2001). Localization of Amplitude-Panned Virtual Sources I : Stereophonic Panning. *Journal of the Audio Engineering Society*, 49(9):739–752.

Pulkki, V., Karjalainen, M., and Jyri, H. (1999). Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model. *Journal of the Audio Engineering Society*.

Raatgever, J. (1980). *On the Binaural Processing of Stimuli with Different Interaural Phase Relations*. PhD thesis, Technische Hogesehool Delft, The Netherlands.

Rubak, P. and Johansen, L. G. (2003). Coloration in Natural and Artificial Room Impulse Responses. In *23rd International AES Conference*, pages 1–19, Copenhagen, Denmark.

Rumsey, F. (2011). Whose head is it anyway? Optimizing binaural audio. *Journal of the Audio Engineering Society*, 59(9).

Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. (2005). On the relative importance of spatial and timbral fidelities in judgments of degraded

multichannel audio quality. *The Journal of the Acoustical Society of America*, 118(2):968–976.

Rychtarikova, M., Van Den Bogaert, T., Vermeir, G., and Wouters, J. (2009). Binaural Sound Source Localization in Real and Virtual Rooms. *Journal of the Audio Engineering Society*, 57(4):205–220.

Salomons, A. M. (1995). *Coloration and Binaural Decoloration of Sound Due to Reflections*. PhD thesis, Technische Universiteit Delft.

Sandel, T. T., Teas, D. C., Feddersen, W. E., and Jeffress, L. A. (1955). Localization of Sound from Single and Paired Sources. *Journal of the Acoustical Society of America*, 27:842.

Sandvad, J. (1996). Dynamic Aspects of Auditory Virtual Environments. In *100th Audio Engineering Society Convention*, Copenhagen, Denmark.

Sandvad, J. and Hammershøi, D. (1994). Binaural Auralization, Comparison of FIR and IIR Filter Representation of HIRs. In *96th Audion Engineering Society Convention*, Amsterdam, The Netherlands.

Satongar, D., Pike, C., Lam, Y. W., and Tew, A. I. (2013). On the Influence of Headphones on Localisation of Loudspeaker Sources. In *135th Audio Engineering Society Convention*, New York, USA.

Satongar, D., Pike, C., Lam, Y. W., and Tew, A. I. (2015). The Influence of Headphones on the Localization of External Loudspeaker Sources. *Journal of the Audio Engineering Society*, 63(10):799–810.

Schuirmann, D. J. (1987). A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Searle, C. L., Braida, L. D., Cuddy, D. R., and Davis, M. F. (1975). Binaural

pinna disparity: another auditory localization cue. *The Journal of the Acoustical Society of America*, 57(2):448–455.

Sepharim, H. P. (1961). Uber die wahrnehmbarkeit mehrerer ruckwurfe yon sprachschall. *Acoustica*, 11:80–91.

Sharpsteen, B., Roberts, B., Beebe Jr, F., Luske, H., and Algar, J. (1941). Fantasia [Motion Picture].

Sheaffer, J. (2013). *From Source to Brain: Modelling Sound Propagation and Localisation in Rooms*. PhD thesis, University of Salford, UK.

Shirley, B., Kendrick, P., and Churchill, C. (2007). The effect of stereo crosstalk on intelligibility: Comparison of a phantom stereo image and a central loudspeaker source. *AES: Journal of the Audio Engineering Society*, 55(10):852–863.

Silzle, A. (2002). Selection and Tuning of HRTFs. In *112th Audio Engineering Society Convention*, Munich, Germany.

Snow, W. B. and Hammer, K. (1932). Binaural Transmission System at Academy of Music in Philadelphia.

Solvang, A. (2008). Spectral impairment of two-dimensional higher order Ambisonics. *Journal of the Audio Engineering Society*, 56(4):267–279.

Søndergaard, P. L., Culling, J. F., Dau, T., Goff, N. L., Jepsen, M. L., Majdak, P., and Wierstorf, H. (2011). Towards a binaural modelling toolbox. *Forum Acousticum*.

Spors, B. S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F. (2013). Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State. *Proceedings of the IEEE*, 101(9):1–19.

Spors, S., Rabenstein, R., and Ahrens, J. (2008). The Theory of Wave

Field Synthesis Revisited. In *124th Audio Engineering Society Convention*, Amsterdam, The Netherlands.

Staff Technical Writer (2006). Binaural Technology for Mobile Applications. *Journal of the Audio Engineering Society*, 54(10):990–995.

Steinberg, J. C. and Snow, W. B. (1934). Auditory Perspective - Physical Factors. *Electrical Engineering*, pages 12–17.

Stern, R. M., Zeiberg, A. S., and Trahiotis, C. (1988). Lateralization of complex binaural stimuli: A weighted-image model. *Journal of the Acoustical Society of America*, 84(1):156–165.

Stitt, P., Bertet, S., and van Walstijn, M. (2014). Off-Centre Localisation Performance of Ambisonics and HOA For Large and Small Loudspeaker Array Radii. *Acta Acustica united with Acustica*, 100(5):937–944.

Strutt, J. V. (1907). On our perception of sound direction. *Philos. Mag.*, 13:214–232.

Takanen, M., Wierstorf, H., Pulkki, V., and Raake, A. (2014). Evaluation of sound field synthesis techniques with a binaural auditory model. In *55th AES International Conference*, pages 1–8.

Theile, G., Wittek, H., and Reisinger, M. (2003). Potential Wavefield Synthesis Applications in the Multichannel Stereophonic World. In *24th Internation Audio Engineering Society Conference on Multichannel Audio*.

Thiele, G. (1980). *On the localisation in the superimposed soundfield*. PhD thesis, Technische Universität Berlin.

Thurlow, W. R., Mangels, J. W., and Runge, P. S. (1967). Head Movements During Sound Localization. *Journal of the Acoustical Society of America*, 42(2):489–493.

Thurstone, L. L. (1927). A Law of Comparative Judgment. *Psychological Review*, 34:273–286.

Toole, F. (2008). *Sound Reproduction: Loudspeakers and Rooms*. Focal Press, Burlington, MA.

Toole, F. E. and Olive, S. E. (1988). The Modification of Timbre by Resonances: Perception and Measurement. *Journal of the Audio Engineering Society*, 36:122–142.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76(1):31.

van de Par, S. and Kohlrausch, A. (1998). Diotic and dichotic detection using multiplied-noise maskers. *Journal of the Acoustical Society of America*, 103(4):2100–2110.

Vickers, E. (2009). Fixing the Phantom Center: Diffusing Acoustical Crosstalk. *127th Convention New York*, pages 1–17.

Völk, F. (2011). System Theory of Binaural Synthesis. In *131st Audio Engineering Society Convention*.

Völk, F. (2012a). Headphone Selection for Binaural Synthesis with Blocked Auditory Canal Recording. In *132nd Audio Engineering Society Convention*.

Völk, F. (2012b). Headphone Selection for Binaural Synthesis with Blocked Auditory Canal Recording. In *132nd Audio Engineering Society Convention*, Budapest, Hungary.

Völk, F. (2013). *Interrelations of Virtual Acoustics and Hearing Research by the Example of Binaural Synthesis*. Doktor-ingenieurs, Technische Universität München.

Völk, F. (2014). Inter- and intra-individual variability in the blocked auditory

canal transfer functions of three circum-aural headphones. *Journal of the Audio Engineering Society*, 62(5):315–323.

Völk, F. and Fastl, H. (2011). Locating the Missing 6 dB by Loudness Calibration of Binaural Synthesis. In *131st Audio Engineering Society Convention*, volume 131, pages 1–12.

Völk, F., Straubinger, M., Roalter, L., and Fastl, H. (2009). Measurement of Head Related Impulse Responses for Psychoacoustic Research. In *Nag/Daga 2009*, pages 164—-167.

Volk, F. and Fastl, H. (2013). Physical Correlates of Loudness Transfer Functions in Binaural Synthesis. In *Proceedings of Meetings on Acoustics (ICA 2013)*, volume 19.

Wallach, H. (1940). The Role of Head Movements and Vestibular and Visual Cues in Sound Localization. *Journal of Experimental Psychology*, 27(4):339–368.

Wasserstein, R. L. and Lazar, N. a. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*.

Weinrich, S. (1982). The problem of front-back localization in binaural hearing. *Scand Audiol Suppl.*, 15:135–145.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. Chapman and Hall/CRC, 2nd edition.

Wenzel, E., Stone, P., Fisher, S., and Foster, S. (1990). A system for three-dimensional acoustic 'visualization' in a virtual environment workstation. *Proceedings of the First IEEE Conference on Visualization: Visualization '90*, pages 329–337.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993).

Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94(1):111–123.

Wiener, F. M. and Ross, D. A. (1946). The Pressure Distribution in the Auditory Canal in a Progresive Sound Field. *Journal of the Acoustical Society of America*, 18(2):401–408.

Wierstorf, H. (2014). *Perceptual Assessment of Sound Field Synthesis*. PhD thesis, Technische Universität Berlin.

Wierstorf, H., Geier, M., Raake, A., and Spors, S. (2011). A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. In *130th Audio Engineering Society Convention*, pages 3–6, London, UK. Audio Engineering Society.

Wierstorf, H., Hohnerlein, C., Spors, S., and Raake, A. (2014). Coloration in Wave Field Synthesis. In *55th AES International Conference*, pages 1–8, Helsinki, Finland.

Wierstorf, H., Raake, A., Geier, M., and Spors, S. (2013). Perception of Focused Sources in Wave Field Synthesis. *Journal of the Audio Engineering Society*, 61(1/2):5–16.

Wierstorf, H., Raake, A., and Spors, S. (2012a). Localization of a virtual point source within the listening area for Wave Field Synthesis. In *133rd Audio Engineering Society Convention*, pages 1–9, San Francisco, USA.

Wierstorf, H., Spors, S., and Raake, A. (2012b). 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. (2007). The Generation of Panning Laws for Irregular Speaker

Arrays Using Heuristic Methods. In *31st Internation Audio Engineering Society Conference*, London, UK.

Wightman, F. L. and Kistler, D. J. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America*, 105(5):2841–2853.

Williams, E. G. (1999). *Fourier Acoustics*. Academic Press, London, UK.

Wittek, H., Rumsey, F., and Theile, G. (2007). Perceptual Enhancement of Wavefield Synthesis by Stereophonic Means. *Journal of the Audio Engineering Society*, 55(9):723–751.

Xie, B. (2013). *Head-related Transfer Function and Virtual Auditory Display*. J Ross Publishing, 2nd edition.

Yairi, S., Iwaya, Y., and Suzuki, Y. (2006). Investigation of System Latency Detection Threshold of Virtual Auditory Display. In *Proceedings of the 12th International Conference on Auditory Display*, pages 217–222, London, UK.

Yao, S.-N., Collins, T., and Jancovic, P. (2015). Timbral and spatial fidelity improvement in ambisonics. *Applied Acoustics*, 93:1–8.

Zahorik, P., Wightman, F., and Kistler, D. (1995). On the discriminability of virtual and real sound sources. In *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, pages 76–79.

Zhang, P. X. and Hartmann, W. M. (2010). On the ability of human listeners to distinguish between front and back. *Hearing research*, 260(1-2):30–46.

Zielinski, S., Rumsey, F., and Bech, S. (2008). On Some Biases Encountered in Modern Audio Quality Listening Tests A Review. *Journal of the Audio Engineering Society*, 56(6):427–451.

Zotter, F. and Frank, M. (2012). All-Round Ambisonic Panning and Decoding. *Journal of the Audio Engineering Society*, 60(10):807–820.

Zotter, F., Pomberger, H., and Matthias, F. (2009). An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Spatial Aliasing. In *Audio Engineering Society Convention*, pages 1–12, Munich, Germany.

Zotter, F., Pomberger, H., and Noisternig, M. (2010). Ambisonics Decoding With And Without Mode-Matching: A Case Study Using The Hemisphere. In *International Symposium on Ambisonics and Spherical Acoustics*, Paris, France.

Zurek, P. M. (1979). Measurements of binaural echo suppression. *Journal of the Acoustical Society of America*, 66(6):1750–1757.