# High Frequency Reproduction in Binaural Ambisonic Rendering

Thomas Thewlis McKenzie

PhD in Music Technology

Department of Electronic Engineering
University of York

December 2019

# Abstract

Humans can localise sounds in all directions using three main auditory cues: the differences in time and level between signals arriving at the left and right eardrums (interaural time difference and interaural level difference, respectively), and the spectral characteristics of the signals due to reflections and diffractions off the body and ears. These auditory cues can be recorded for a position in space using the head-related transfer function (HRTF), and binaural synthesis at this position can then be achieved through convolution of a sound signal with the measured HRTF. However, reproducing soundfields with multiple sources, or at multiple locations, requires a highly dense set of HRTFs. Ambisonics is a spatial audio technology that decomposes a soundfield into a weighted set of directional functions, which can be utilised binaurally in order to spatialise audio at any direction using far fewer HRTFs. A limitation of low-order Ambisonic rendering is poor high frequency reproduction, which reduces the accuracy of the resulting binaural synthesis.

This thesis presents novel HRTF pre-processing techniques, such that when using the augmented HRTFs in the binaural Ambisonic rendering stage, the high frequency reproduction is a closer approximation of direct HRTF rendering. These techniques include Ambisonic Diffuse-Field Equalisation, to improve spectral reproduction over all directions; Ambisonic Directional Bias Equalisation, to further improve spectral reproduction toward a specific direction; and Ambisonic Interaural Level Difference Optimisation, to improve lateralisation and interaural level difference reproduction. Evaluation of the presented techniques compares binaural Ambisonic rendering to direct HRTF rendering numerically, using perceptually motivated spectral difference calculations, auditory cue estimations and localisation prediction models, and perceptually, using listening tests assessing similarity and plausibility. Results conclude that the individual pre-processing techniques produce modest improvements to the high frequency reproduction of binaural Ambisonic rendering, and that using multiple pre-processing techniques can produce cumulative, and statistically significant, improvements.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

My funders for supporting this PhD: a Google Faculty Research Award and a University of York Department of Electronic Engineering Studentship.

My first supervisor Dr. Gavin Kearney for your infectious drive and guidance throughout this journey, and for always leaving your door open for me. It feels like only yesterday I was sat in your TFTV office listening to Freddie Mercury's isolated voice sing out of an Ambisonic loudspeaker array in virtual reality.

My second supervisor Prof. Damian Murphy for your wisdom and intuition. For interviewing me at undergraduate level and, after I chose a different University, humbly accepting me back years later as a PhD supervisee. Let us also not forget the AudioLab day outs.

Dr. Helena Daffern for being an excellent thesis advisor and for sending me up all those hills, and Tony Tew for valiantly stepping in on maternity cover thesis advisor duties and making me think.

Everyone at the AudioLab past and present for welcoming me and bringing out my true weird self, and of course for the lunchtime crosswords. Specifically but by no means exclusively, Frank for this glorious thesis template; Cal for sitting next to me the whole way through; Lewis, Christoph of Trouser and Tomasz for sharing an office with me.

Finally, my dear friends, family and cat, for pretending to listen while I rant.

# Declaration of Authorship

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references. I also declare that parts of this research have been presented in previous conference and journal publications, which are listed as follows:

- **"Diffuse-field equalisation of binaural Ambisonic rendering,"** *Applied Sciences*, vol. 8, no. 10, 2018, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"Interaural level difference optimisation of binaural Ambisonic rendering,"** *Applied Sciences*, vol. 9, no. 6, 2019, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"Diffuse-field equalisation of first order Ambisonics,"** in *20th International Conference on Digital Audio Effects*, (pp. 389–396), Edinburgh, UK, 2017, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"Directional bias equalisation of first order binaural Ambisonic rendering,"** in *AES Conference on Audio for Virtual and Augmented Reality*, Redmond, USA, 2018, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"Interaural level difference optimisation of first order binaural Ambisonic rendering,"** in *AES Conference on Immersive and Interactive Audio*, York, UK, 2019, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"An evaluation of pre-processing techniques for virtual loudspeaker binaural Ambisonic rendering,"** in *EAA Spatial Audio Signal Processing Symposium*, (pp. 149–154), Paris, France, 2019, T. McKenzie, D. T. Murphy, and G. Kearney.
- **"Towards a perceptually optimal bias factor for directional bias equalisation of binaural Ambisonic rendering,"** in *EAA Spatial Audio Signal Processing Symposium*, (pp. 97–102), Paris, France, 2019, T. McKenzie, D. T. Murphy, and G. Kearney.

- **“A perceptual spectral difference model for binaural signals,”** in *AES 145th Convention*, New York, USA, 2018, C. Armstrong, T. McKenzie, D. T. Murphy, and G. Kearney.

# Chapter 1

# Introduction

Sound is all around us, even when we close our eyes. While humans are often touted as visual animals, we do not have eyes in the back of our heads, yet we can hear in all directions. In many situations, we use our hearing first, and our other senses second. When we hear a bird chirp in a nearby tree, we turn and face it. When we hear a car driving past from behind, we stop before crossing the road. Our ability to localise sounds comes primarily from three auditory cues: the differences in time and level between signals arriving at the left and right eardrums (interaural time difference and interaural level difference, respectively) (Rayleigh, 1907), and the spectral characteristics of the signals due to reflections and diffractions off the torso, head and pinnae.

By gaining an understanding of the mechanisms we use to decipher the location and direction of sounds, we can then look to recreate them by rendering spatial audio. This has many potential applications and is not just confined to the entertainment industry (Rumsey, 2001); its uses range from health, such as exposure therapy (Johnston, Egermann and Kearney, 2019), wellbeing (Daffern et al., 2019) and accessibility (Cooper and Taylor, 1998), to historical purposes such as recording acoustics for posterity (Murphy, 2013; Postma and Katz, 2015; Postma and Katz, 2016; Postma et al., 2016) and architecture (Blesser and Salter, 2009), to improving safety in cars with directional collision warnings (House et al., 2017).

A common and relatively simple way of capturing the human auditory localisation cues is using the head-related transfer function (HRTF), which is typically measured by placing miniature microphones in the ear canals and recording a known signal from a specified position in space relative to the head (Begault, 1994). Spatial audio can then be synthesised at this position through convolution of a sound signal with the measured HRTF, which when played back through headphones can give the impression of that sound originating from the location of the measured HRTF with impressive realism. This is known as binaural synthesis. A drawback of binaural synthesis is that reproduction of soundfields with multiple sources, with varying widths, distances and locations, requires a highly dense set of HRTF measurements.

Ambisonics is a spatial audio technology that decomposes a soundfield into a weighted set of directional functions (Gerzon, 1973). This technology can be utilised binaurally (McKeag and McGrath, 1996; Noisternig et al., 2003b) in order to allow the spatialisation of audio at any position using far fewer HRTF measurements than direct binaural synthesis using HRTF convolution. However, low-order Ambisonic reproduction is poor at high frequencies, which reduces the accuracy of the resulting binaural synthesis (Daniel, Rault and Polack, 1998). Higher-order Ambisonics raises the frequency limit of accurate reproduction (Malham, 1999), but this comes with a requirement of more microphone capsules in recording, increased file size in storage and transmission, and a greater number of HRTF measurements in the binaural rendering stage. Therefore it is highly desirable to investigate methods of maximising the reproduction quality of low-order Ambisonic rendering.

The aim of the work presented in this thesis is to improve the high frequency reproduction of low-order binaural Ambisonic rendering. The purpose is to produce the most realistic spatial audio as possible within the same Ambisonic order, without altering the real-time rendering process, as improved realism produces greater immersion (Møller et al., 1996; Rumsey, 2002). Certain psychoacoustic characteristics can be used to guide research motivations, such as prioritising accuracy in timbre over localisation, which has been shown as a more important feature for the feeling of realism (Bregman, 1990). By striving to improve low-order binaural Ambisonic rendering, the need to use more complex higher-order Ambisonics can be reduced. This is beneficial for scenarios where the available computational power may be limited, such as in mobile phones and virtual reality headsets.

## 1.1 Statement of Hypothesis

The hypothesis that forms the motivation for the work presented in this thesis is as follows:

> *The use of head-related transfer function pre-processing techniques can improve the high frequency reproduction of binaural Ambisonic rendering.*

The key terms of this hypothesis, and how they relate to this thesis, are explained as follows:

- **Head-related transfer function:** A filter which describes the change in a sound signal between its source and the eardrums due to diffraction and reflections off the head, torso and ears.
- **Binaural Ambisonic rendering:** An alternative way of synthesising binaural audio using Ambisonic technology. This allows for soundfield reproduction using far fewer head-related transfer functions than direct binaural rendering.
- **Pre-processing techniques:** Algorithms that augment the spectral and temporal characteristics of head-related transfer functions used in binaural Ambisonic rendering.
- **High frequency reproduction:** Binaural Ambisonic rendering is inherently inaccurate at high frequencies. The effect of pre-processing techniques is measured through both numerical evaluation metrics, by comparing binaural Ambisonic rendering to reference head-related transfer functions, and perceptual evaluations in the form of listening tests with human participants.

## 1.2 Novel Contributions

The research presented in this thesis has produced the following novel contributions to the field:

- **Ambisonic Diffuse-Field Equalisation:** The development and evaluation of a head-related transfer function pre-processing technique for binaural Ambisonic rendering. This samples the Ambisonic reproduction of a specified binaural Ambisonic decoder evenly over all locations on the sphere and obtains an average response, before equalising it. This equalisation improves the overall spectral reproduction of the binaural Ambisonic rendering, when compared to direct head-related transfer function rendering.

- **Ambisonic Directional Bias Equalisation:** The development and evaluation of a head-related transfer function pre-processing technique for binaural Ambisonic rendering. This is an adaptation of the Ambisonic Diffuse-Field Equalisation technique, that instead samples the Ambisonic reproduction of a specified binaural Ambisonic decoder with the distribution of locations on the sphere skewed to a specific location. This is equalised as before, and with an additional equalisation, this technique further improves the spectral reproduction of binaural Ambisonic rendering at a specific location, when compared to direct head-related transfer function rendering.

- **Ambisonic Interaural Level Difference Optimisation:** The development and evaluation of a head-related transfer function pre-processing technique for binaural Ambisonic rendering. This augments the levels of the left and right signals of the head-related transfer functions used in the binaural Ambisonic rendering stage, such that the resulting rendering reproduces interaural level differences more accurately, when compared to direct head-related transfer function rendering.

- **Combinations of Ambisonic Pre-Processing Techniques:** Finally, the viability of combining multiple Ambisonic head-related transfer function pre-processing techniques is explored, both the presented ones and existing techniques, for whether they can produce cumulative improvements to the accuracy of reproduction of binaural Ambisonic rendering, when compared to direct head-related transfer function rendering.

## 1.3 Statement of Ethical Approval

The protocols for perceptual tests using human participants presented in this thesis, and the management of corresponding data, were approved by the University of York Physical Sciences Ethics Committee with reference McKenzie280217.

## 1.4 Thesis Structure

This thesis is structured as follows. Chapter 2 comprises a review of literature to give the reader a base of knowledge for the material covered in the rest of this thesis. It begins with an explanation of the relevant properties of sound and environmental factors that affect the way sound travels through space, and the changes that occur to sound signals from the source to a listener's eardrums. This is followed by an overview of the human auditory system: specifically, how humans decipher localisation cues to determine the position of a sound source. This chapter includes a review of current research in binaural audio technology, including both ways of rendering spatial audio binaurally using the head-related transfer function (HRTF), as well as methods for evaluating the quality of audio rendering.

Chapter 3 introduces Ambisonic technology, a way of separating the recording and encoding processes from the rendering process. A background review and the early forms of Ambisonics is first presented, followed by developments up to the current state-of-the-art techniques such as higher-order Ambisonics and psychoacoustic decoder weights. Binaural rendering of Ambisonic signals is then detailed, which offers a way of rendering binaural audio with far fewer HRTF measurements and potentially far fewer convolutions than direct HRTF rendering. The limitations of low-order binaural Ambisonic rendering are then demonstrated and explained, which form the justification for much of the work presented in this thesis.

Chapters 4, 5 and 6 present the motivation, methodology and evaluation of three novel HRTF pre-processing techniques for binaural Ambisonic rendering. The first two focus on improvements to spectral reproduction over the whole sphere and concentrated around a single position, respectively, and the third focuses on improvements to interaural level difference reproduction over the whole sphere, without reducing spectral quality. Chapter 7 then proposes the possibility of combining multiple HRTF pre-processing techniques for cumulative overall improvements. These include the techniques proposed in Chapters 4 and 6, as well as an existing technique (Evans, Angus and Tew, 1998; Richter et al., 2014; Zaunschirm, Schörkhuber and Höldrich, 2018). Evaluation of binaural Ambisonic rendering is performed both numerically and perceptually. Numerical evaluation is achieved through spectral difference calculations, interaural cue estimations and binaural localisation prediction models. Perceptual evaluation is achieved through listening tests with human participants.

This thesis is concluded in Chapter 8 with a summary of the key findings of the work presented, along with a restatement of the hypothesis. A comment is made on the objectives of the thesis and whether they have been achieved. Finally, areas of future work that have been identified throughout this thesis are considered in finer detail, and the scope of this thesis, along with its implications in the greater research context, are considered.

# Chapter 2

# A Review of Binaural Audio

In order to reproduce spatial sound as realistically as possible, it is first necessary to gain an understanding of the fundamental properties of sound. There are many factors that may affect a sound signal between its source and the eardrums, from environmental aspects such as the reverberation and temperature of the listening space to the morphology and age of the listener. This chapter begins with the basic principles of sound and wave propagation before describing the changes in time, level and frequency that occur between a sound source and its arrival at the eardrums, and the way in which the human auditory system deciphers these changes to deduce the position of the sound source. This chapter then focuses on binaural technology and techniques for simulating spatial audio over headphones, from recording and measurement methods to synthesis and playback systems. Finally, metrics for evaluating the quality of binaural audio are then investigated, from perceptual audio evaluation using listening tests to numerical calculations using binaural cue estimations and localisation prediction models.

## 2.1 Coordinate System

In this thesis, unless otherwise stated, angles of sound incidence are referred to in spherical coordinates of *azimuth* (denoted by $\theta$) for angles on the horizontal plane

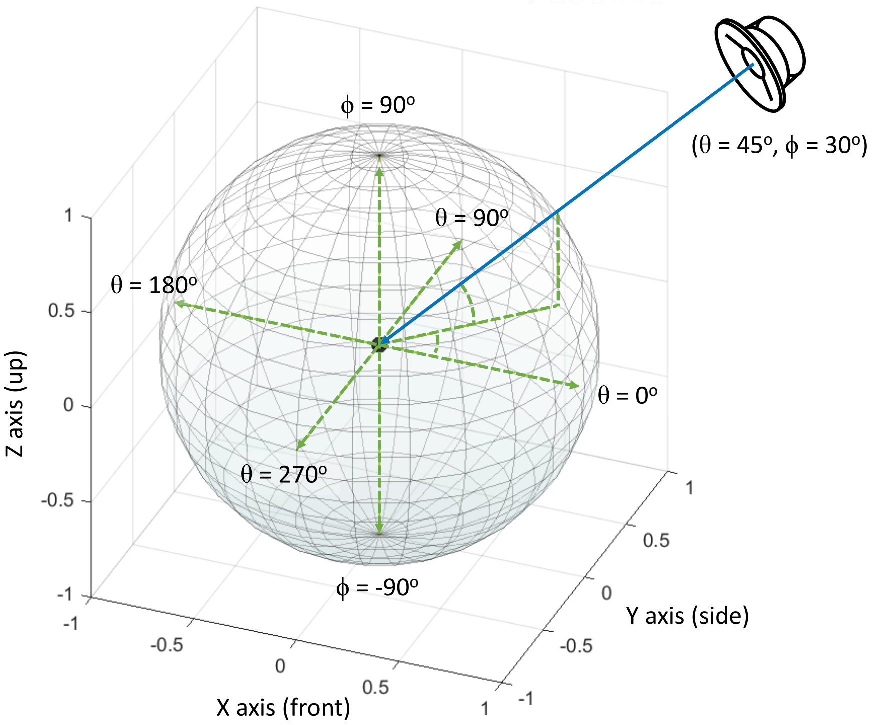


FIGURE 2.1: Illustration of the spherical coordinate system, with azimuth and elevation denoted by $\theta$ and $\phi$, respectively.

(in the region $-180° < \theta \leq 180°$) and *elevation* (denoted by $\phi$) for angles on the vertical plane (in the region $-90° \leq \phi \leq 90°$). An incidence of $(\theta = 0°, \phi = 0°)$ represents a direction straight in front of the listener at the height of the ears. Positive changes in angles move anticlockwise in azimuth and upwards in elevation. Another system used in this thesis is Cartesian coordinates, where positions around the origin are described as a combination of $x, y$ and $z$ values. The spherical and Cartesian coordinate systems are illustrated in Figure 2.1.

Conversion from the Cartesian coordinate system $(x, y, z)$ to the spherical coordinate system is as follows:

$$
\begin{aligned}
\theta &= \frac{180}{\pi}\tan^{-1}\left(\frac{y}{x}\right) \\
\phi &= \frac{180}{\pi}\tan^{-1}\left(\frac{z}{\sqrt{(x^2+y^2)}}\right)
\end{aligned}
\tag{2.1}
$$

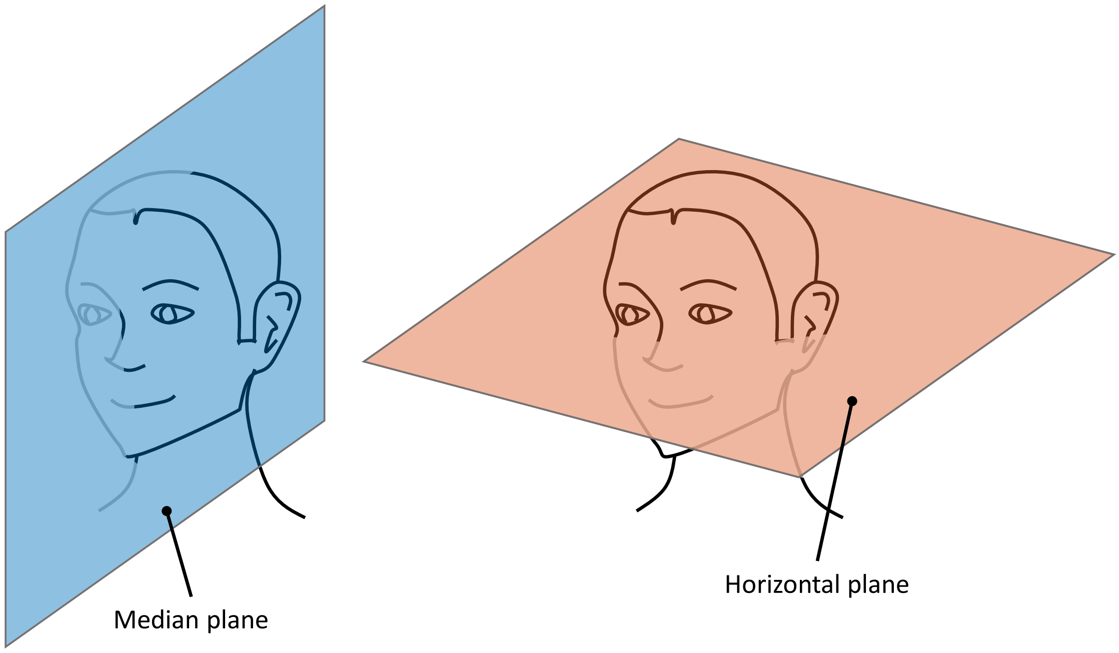


FIGURE 2.2: Illustration of the plane system used in this thesis.

where $\tan^{-1}$ denotes a four quadrant inverse tangent and the $\frac{180}{\pi}$ term converts the spherical coordinates from radians to degrees.

## 2.2 Plane System

Figure 2.2 illustrates plane system used in this thesis. The median plane bisects the head down the centre into left and right. Sagittal planes are parallel to the median plane. The horizontal plane bisects the head into above and below, with the vertical point of bisection aligning with the centre of the entrance to the ear canal.

## 2.3 Fundamentals of Sound

Sound is a longitudinal wave caused by the vibration of particles along the direction of propagation. Sound waves travel via changes in pressure at neighbouring particles, as illustrated in Figure 2.3, where compressions and rarefactions are points of increased

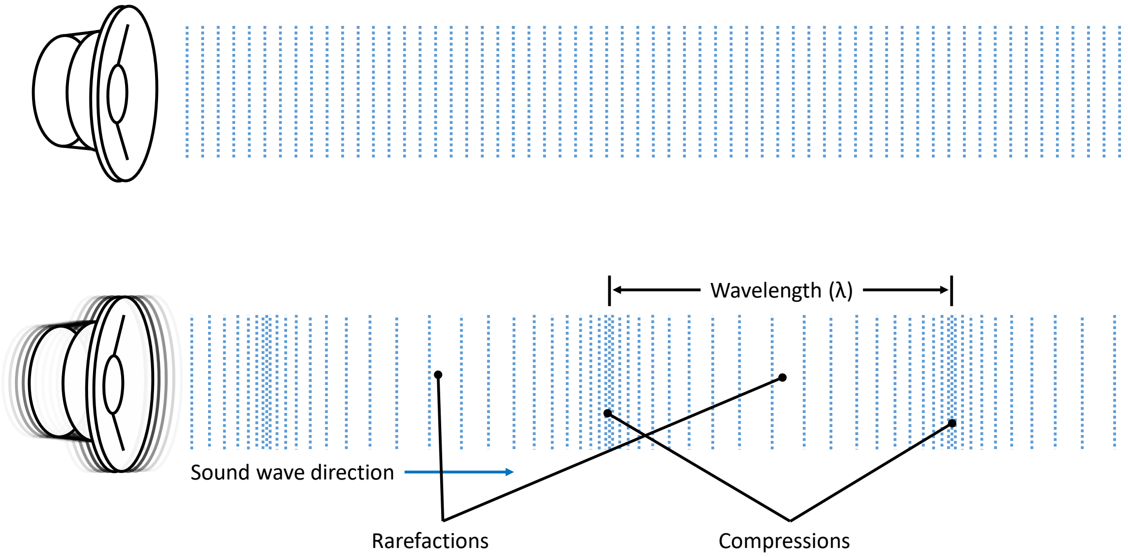


FIGURE 2.3: A longitudinal sound wave from a vibrating loudspeaker.

and decreased pressure, respectively. The speed of sound $c$ varies with pressure, temperature and the medium in which it travels. Sound therefore travels faster in solids and liquids than gases, and faster at higher temperatures and humidities. At 20°C in dry air, $c \approx 343$ m/s. The time $t$ a sound takes from source to destination is calculated as: $t = r/c$, where $r$ is the distance of the straight line direct path.

The relationship between frequency and wavelength of a sound is reciprocal:

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

where $f$ is frequency (measured in Hz) and $\lambda$ is wavelength (measured in m). The amplitude of a sound is typically measured using the decibel (dB), a logarithmic unit that expresses the ratio of two values of sound pressure. Sound pressure level (SPL) is measured using a reference pressure, which is usually the threshold of human hearing; the quietest sound audible to the human ear: 20 $\mu$Pa. The SPL of a sound is therefore calculated as:

$$\mathrm{SPL} = 20\log_{10}\left(\frac{\varphi}{\varphi_{\mathrm{ref}}}\right) \tag{2.3}$$

Table 2.1: Approximate sound pressure levels of common sounds at specified distances, reproduced from Sangpeilaudio (2019).

| Sound sources | Sound pressure level (dB SPL) |
|---|---|
| Human hearing threshold | 0 |
| Background in TV studio | 20 |
| Quiet library | 40 |
| Conversational speech at 1 m | 60 |
| Busy road at 5 m | 80 |
| Nightclub, 1 m from loudspeaker | 100 |
| Threshold of discomfort | 120 |
| Jet aircraft at 50 m | 140 |

where $\varphi$ is the measured sound pressure and $\varphi_{\text{ref}}$ is the reference sound pressure. The relationship between change in SPL and distance is calculated as

$$\text{SPL}_2 = \text{SPL}_1 - 20\log_{10}\left(\frac{r_1}{r_2}\right) \tag{2.4}$$

where $\text{SPL}_1$ is the SPL at distance $r_1$, and $\text{SPL}_2$ is the sound pressure level at the new distance $r_2$. SPL therefore drops by around 6 dB per doubling of distance. Some approximate SPL values for common sounds are presented in Table 2.1 (Sangpeilaudio, 2019).

Sound waves interact with the environment they are in. As shown in Figure 2.4, *direct sound* travels a straight line path from the sound source to the destination, and arrives first with the greatest amplitude. Next to arrive are the *early reflections*, which are the echoes of the direct sound reflected off a small number of surfaces (stated as $< 6$ in Martin, Van Maercke and Vian (1993)) and which arrive at the receiver having travelled the next shortest paths. In measured room impulse responses, early reflections are still visible as distinct impulses (a simplified example of which is shown in Figure 2.5). The time difference between the direct sound and the early reflections can in some cases be used to determine the path length difference and therefore infer where the reflection came from, and early reflections can be used in room acoustic analysis to give information on the size and geometry of a room (Khaykin and Rafaely, 2012; Lovedee-Turner and Murphy, 2018). *Diffuse*

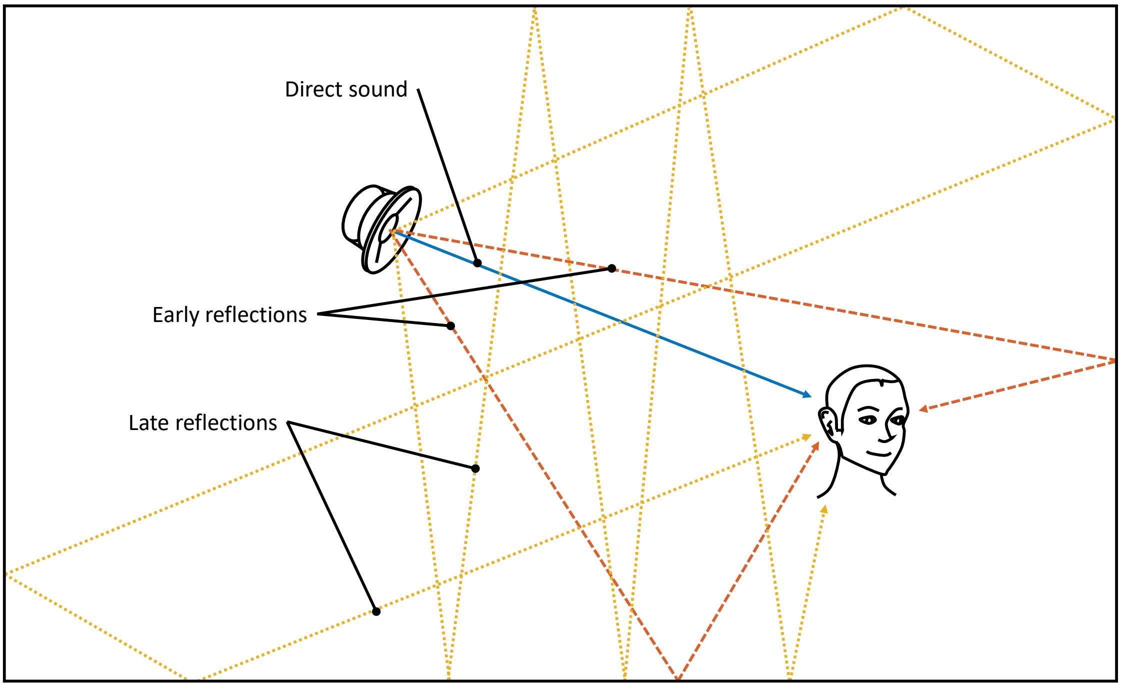


FIGURE 2.4: Simplified illustration of sound wave interaction in a room demonstrating the direct sound, early reflections and late reflections, which form diffuse reverberation.

*reverberation* arrives after the early reflections and typically contains no directional information (Karagiozov, 2014), due to the combination of multiple late reflections arriving at comparable times. Diffuse reverberation differs from early reflections in that there are no distinct impulses (reflections) observable. The transition from early reflections to diffuse reverberation occurs around the 4th to 6th order reflections (Martin, Van Maercke and Vian, 1993). The length of an environment's diffuse reverberation is generally measured by the time it takes for the SPL to reach 60 dB lower than the direct sound, once the direct sound has stopped. This is referred to as the *reverberation time*, or $\mathrm{RT}_{60}$. Referring to Figure 2.5, the reverberation time shown is $\mathrm{RT}_{60} \approx 4$ s.

A number of other factors change the characteristics of a sound from its source to receiver, and this interaction changes depending on the size of the environment, the objects inside the environment and the materials that make up the environment. The shape of a room also affects sound propagation: for example, parallel walls can produce room modes such as standing waves, and large rooms generally have longer

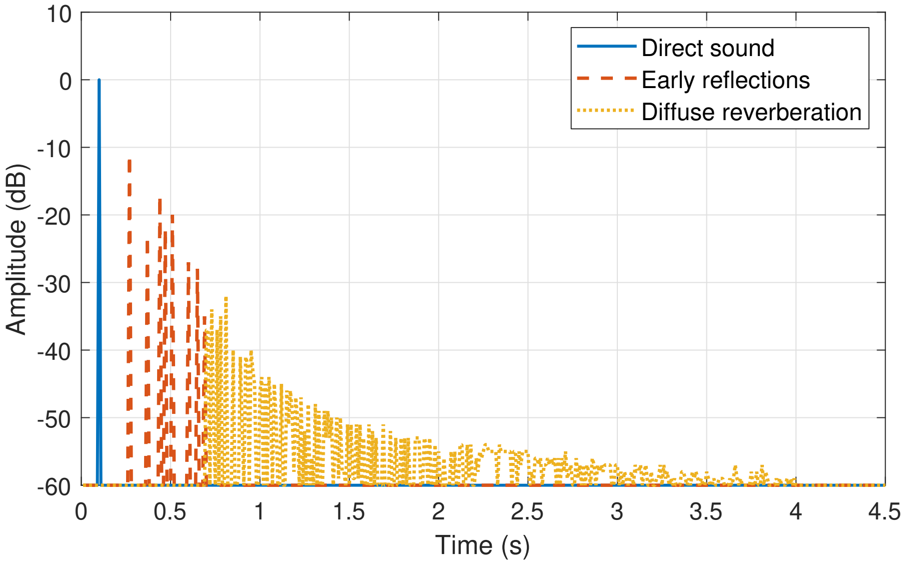


FIGURE 2.5: Simplified room impulse response formed from direct sound, early reflections and diffuse reverberation.

reverberation times. Different materials have particular acoustical properties: flat and shiny surfaces reflect sound more than soft and rough surfaces, which tend to absorb sound (Everest, 2014). Furthermore, sounds behave differently when incident upon an object depending on their frequency and the size of the diffracting object: low frequency waves have longer wavelengths and diffract around objects more than high frequencies, which have shorter wavelengths and are more directional.

An acoustic *free-field* refers to a theoretical situation where only the direct path from the source to the receiver exists, and the sound follows the Inverse Square Law whereby an increased distance $r$ reduces the amplitude by a rate of $r^2$. In practice a free-field is impossible to achieve, though efforts to get as close to this as possible are found in anechoic chambers which are special acoustically treated rooms with extremely short reverberation times and highly absorbent and non-reflective walls (Beranek and Sleeper, 1946). A *diffuse-field* refers to an environment where sound comes from no discernible direction - the sound pressure is equal at all directions and positions in the environment. Practical attempts at implementations of diffuse-field environments include reverberation chambers (Rettinger, 1957).

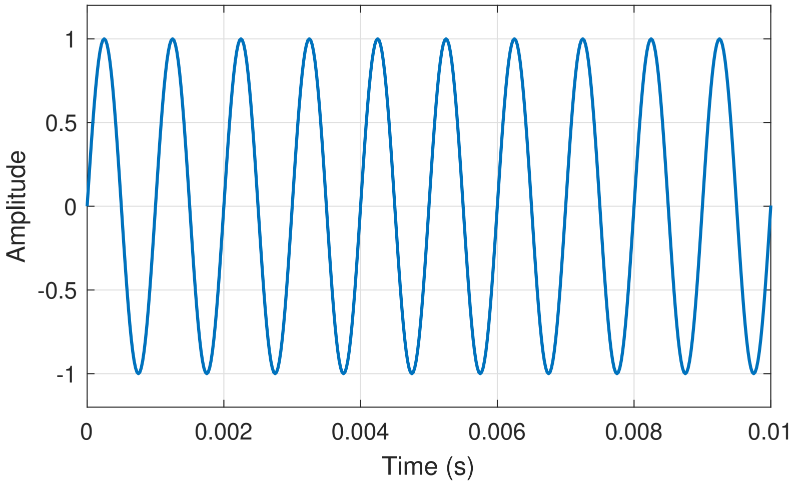


FIGURE 2.6: A 1 kHz sine wave.

### 2.3.1 Audio Sampling

Considering a sinusoidal wave of frequency 1 kHz, as illustrated in Figure 2.6: to represent this sound digitally, the amplitude must be sampled at regular points in time. Digital sampling of audio requires analogue-to-digital conversion, which uses pulse-code modulation (Reeves, 1942). The amount of samples per second is given by the sampling frequency, $f_s$, where $f_s/2$ is the *Nyquist frequency* which is the highest frequency that can be sampled and recovered without error. The value of $f_s$ must therefore fulfil $f_s > 2f_{max}$, where $f_{max}$ is the highest frequency to be recorded (usually around 20 kHz which is the limit of perceivable frequencies, as will be explained later in this chapter).

If the 1 kHz signal is sampled at $f_s = 48$ kHz, it is possible to obtain a digital approximation of the waveform, as depicted in Figure 2.7. To view the frequency spectrum of the signal, the *fast Fourier transform* (FFT) can be applied to the recorded data (Cooley and Tukey, 1965). The FFT breaks down a digitally sampled time-domain signal into a set of amplitudes for a number of frequency bins. Figure 2.8 shows an approximation of the frequency spectrum of a 1 kHz sine wave signal, obtained from an FFT of the time domain representation of the waveform. The curved nature of the peak at its base is due to the signal windowing process of the discrete Fourier transform algorithm.

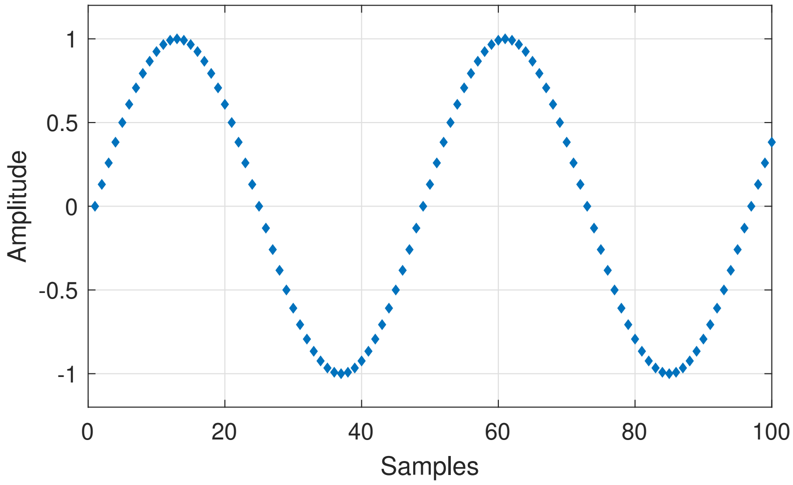


FIGURE 2.7: A discrete time 1 kHz sine wave sampled at $f_s = 48$ kHz.

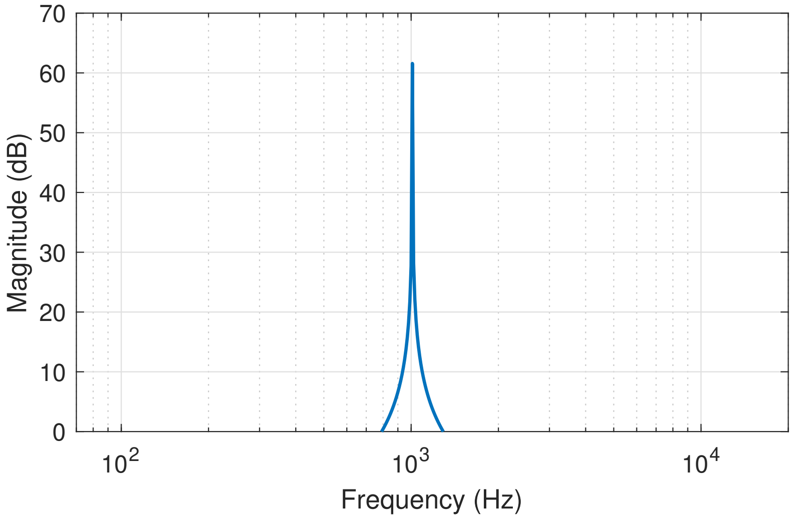


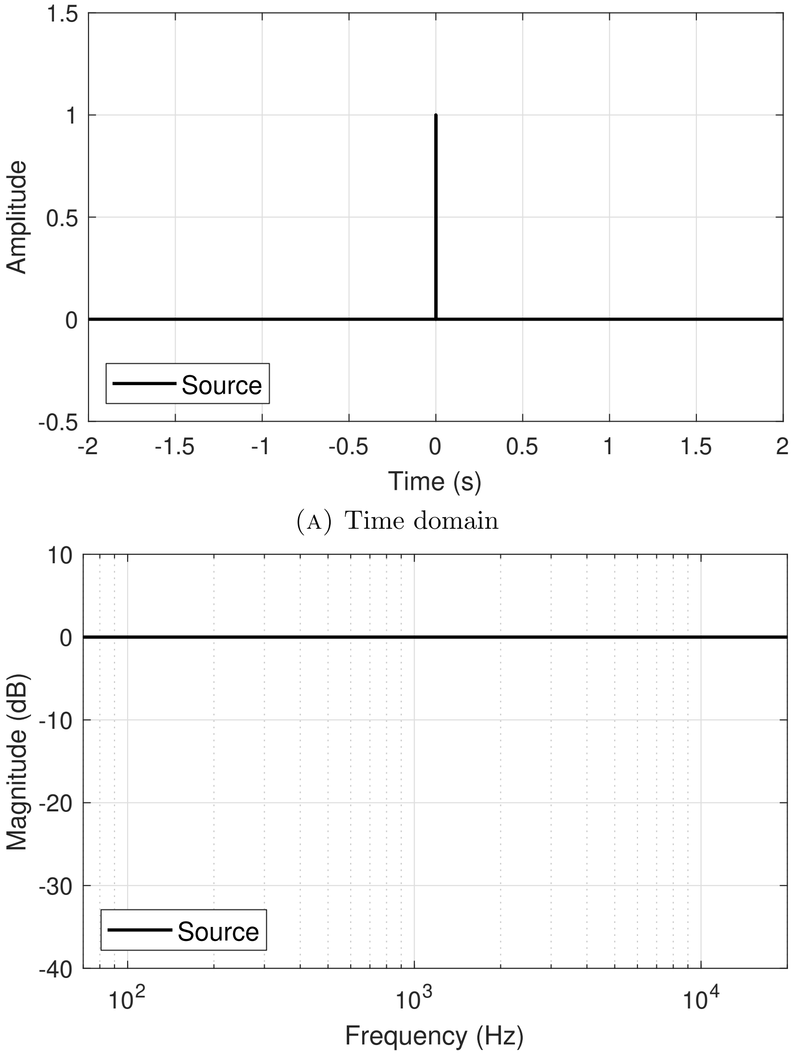
FIGURE 2.8: Frequency spectrum of a 1 kHz sine wave sampled at $f_s = 48$ kHz.

### 2.3.2 The Impulse Response

A theoretical impulse $\delta(t)$ has energy at all frequencies when time $t = 0$, and no energy at all other times:

$$\delta(t) = \begin{cases} 1, \text{for } t = 0 \\ 0, \text{for } t \neq 0 \end{cases} \tag{2.5}$$

A time-domain representation of $\delta(t)$ is shown in Figure 2.9a, and the frequency response of the impulse, obtained from the FFT of $\delta(t)$, is shown in Figure 2.9b as



(A) Time domain

(B) Frequency domain

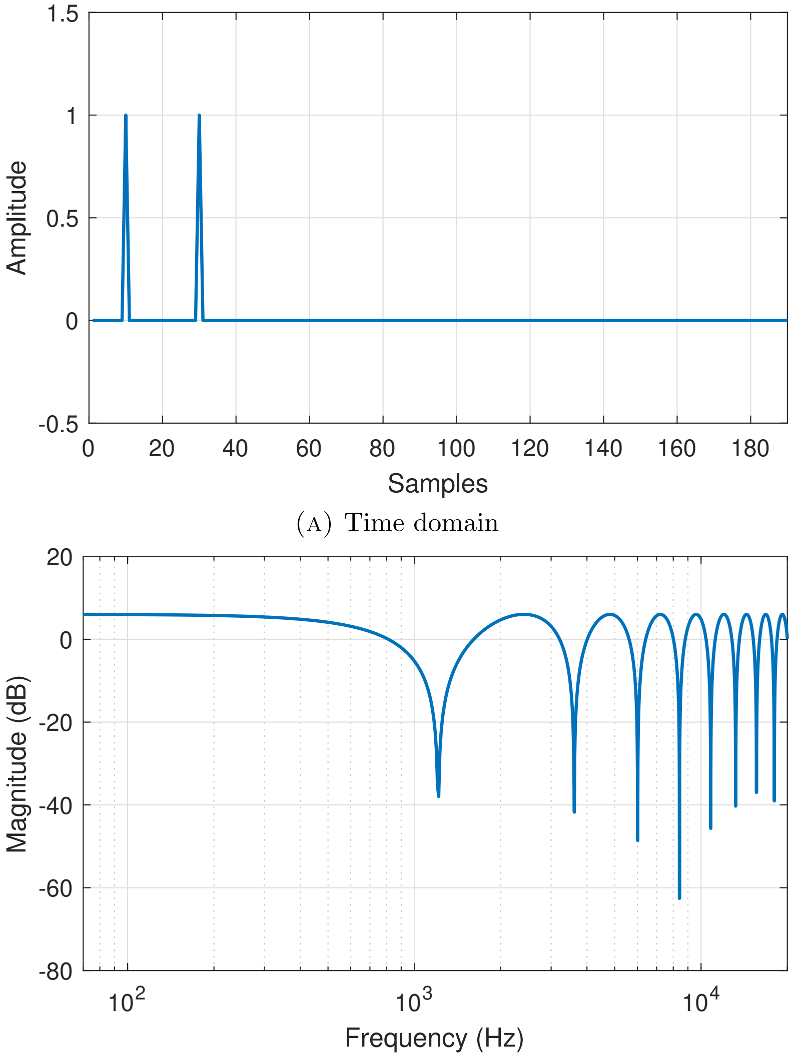
FIGURE 2.9: Time and frequency domain representations of an impulse.

including equal energy at all frequencies. By emitting an impulse from one position in space and recording from another, it is possible to measure the impulse response (IR): the changes that are subject to the impulse from the source to receiver. For room impulse responses for example, as depicted in Figure 2.5, this measured impulse response would include the direct sound, early reflections and diffuse reverberation of the room.

In practice, however, it is impossible to reproduce a perfect impulse due to the limitations of particle physics. Approximations of impulses can be made using sharp transient sounds, such as a starter pistol (Stevens and Murphy, 2014) or a balloon popping, both of which are used in IR measurement scenarios where other methods may be unavailable or infeasible. If the sharp transient sound is $< 25\ \mu$s in duration, it should contain relatively even energy levels in the frequency range of 16 Hz - 16 kHz (Blauert, 1997). However, recording IRs with sharp transient sounds can result in a low signal-to-noise ratio. Aiming for a signal-to-noise ratio greater than 80 dB is recommended for high quality auralisation (Stan, Embrechts and Archambeau, 2002).

To produce impulse responses with higher signal-to-noise ratios, more scientific methods have been developed that produce the excitation stimulus by the playback of a known signal through a loudspeaker. The *maximum length sequence* (Schroeder, 1979) and *inverse repeated sequence* (Dunn and Hawksford, 1993) use pseudo-random noise as the excitation stimulus followed by a circular cross-correlation for impulse response retrieval, but issues with these measurement techniques include peaks of distortion evenly spread in time throughout the impulse response, caused by the imperfect loudspeaker transducer reproduction. Reduction of these artefacts is possible by lowering the playback level, which in turn reduces the signal-to-noise ratio.

A more recent alternative excitation signal is suggested in the *exponential swept sine* technique (Farina, 2000). This is now the most widely used way of measuring impulse responses, and is achieved by playing an exponential sine sweep through a loudspeaker and then deconvolving the recorded sweep with an inverse of the original sweep. The sine sweep can be any desired duration, and doubling the duration will increase the signal-to-noise ratio by approximately 3 dB. Other benefits of sine sweep IR measurement include the ability to remove the harmonic distortion that can occur from over-driving the loudspeaker in measurement, as when the sine sweep is deconvolved into an impulse, the harmonic distortion effects will appear as smaller impulses that occur before the main (greatest amplitude) impulse. These can be removed through simple truncation. The exponential swept sine technique has been shown to perform better than other measurement methods for quiet environments (Stan, Embrechts and Archambeau, 2002).



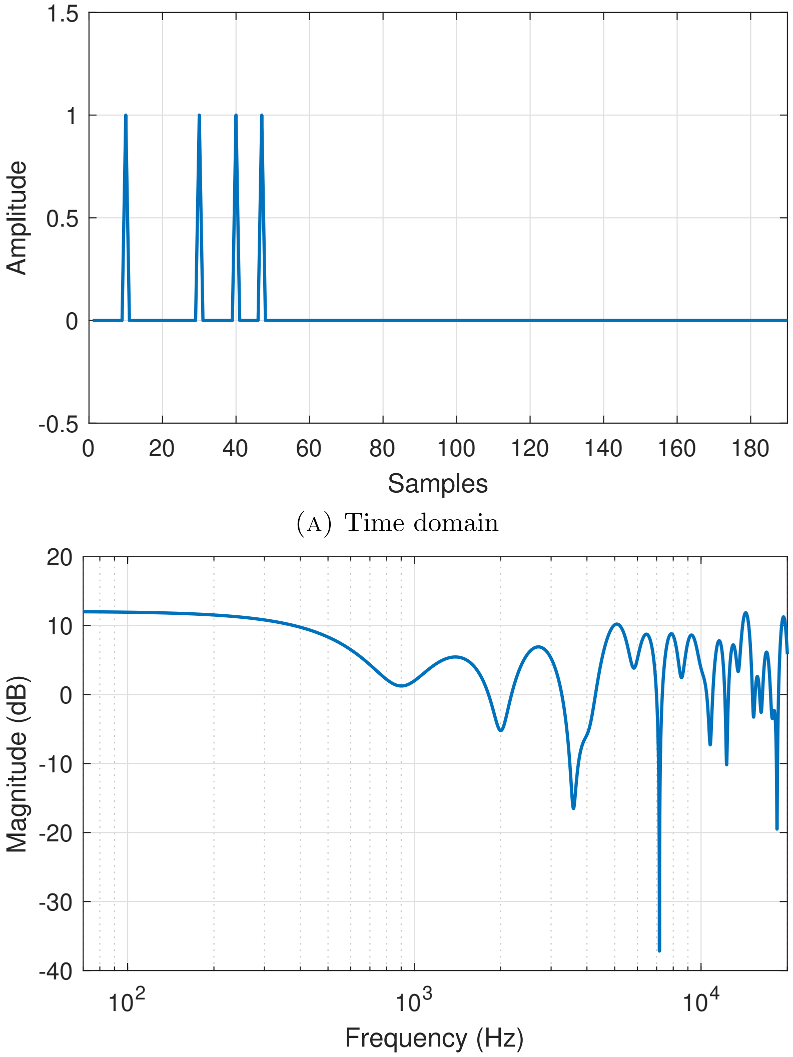
(A) Time domain

(B) Frequency domain

FIGURE 2.10: Time and frequency domain representations of a repeated impulse causing comb filtering.

### 2.3.3 Comb Filtering and Interference

When multiple versions of the same sound signal arrive at the ear with different time delays, such as in multiple-loudspeaker situations or with room reverberation, spectral colouration of the signal can occur. Consider if the $\delta(t)$ at $t = 0$ from 2.9a is recorded at sample 10 and then repeated 20 samples later, as depicted in Figure 2.10a. An acoustic phenomenon called *comb filtering* occurs, whereby several notches appear in the frequency response that resemble the shape of a comb (see Figure 2.10b). Also note how the overall magnitude is increased due to the extra impulse.



(A) Time domain

(B) Frequency domain

Figure 2.11: Time and frequency domain representations of multiple repeated impulses causing general destructive interference.

It is possible to calculate the frequency intervals $f_{\text{notch}}$ between the comb filtering notches from the time delay $t_{\text{delay}}$ between the two impulses, and vice versa, using the following equation:

$$f_{\text{notch}} = \frac{1}{t_{\text{delay}}} \tag{2.6}$$

For a delay of 20 samples as in 2.10 and using $f_s = 48$ kHz, the time delay can be calculated as approximately $t_{\text{delay}} = 0.417$ ms and therefore the frequency interval between notches $f_{\text{notch}} = 2.4$ kHz. The first notch occurs at $f_{\text{notch}}/2$.

To take this example one step further, consider several repeated impulses at varying delay times (see Figure 2.11a with duplicate impulses at 10, 30, 40 and 47 samples). The result is more general destructive interference, as shown in Figure 2.11, and the frequency response changes are less distinct. Comb filtering effects occur naturally when reflected signals reach the same destination with a small time delay. This will be relevant in the coming sections.

### 2.3.4 Convolution

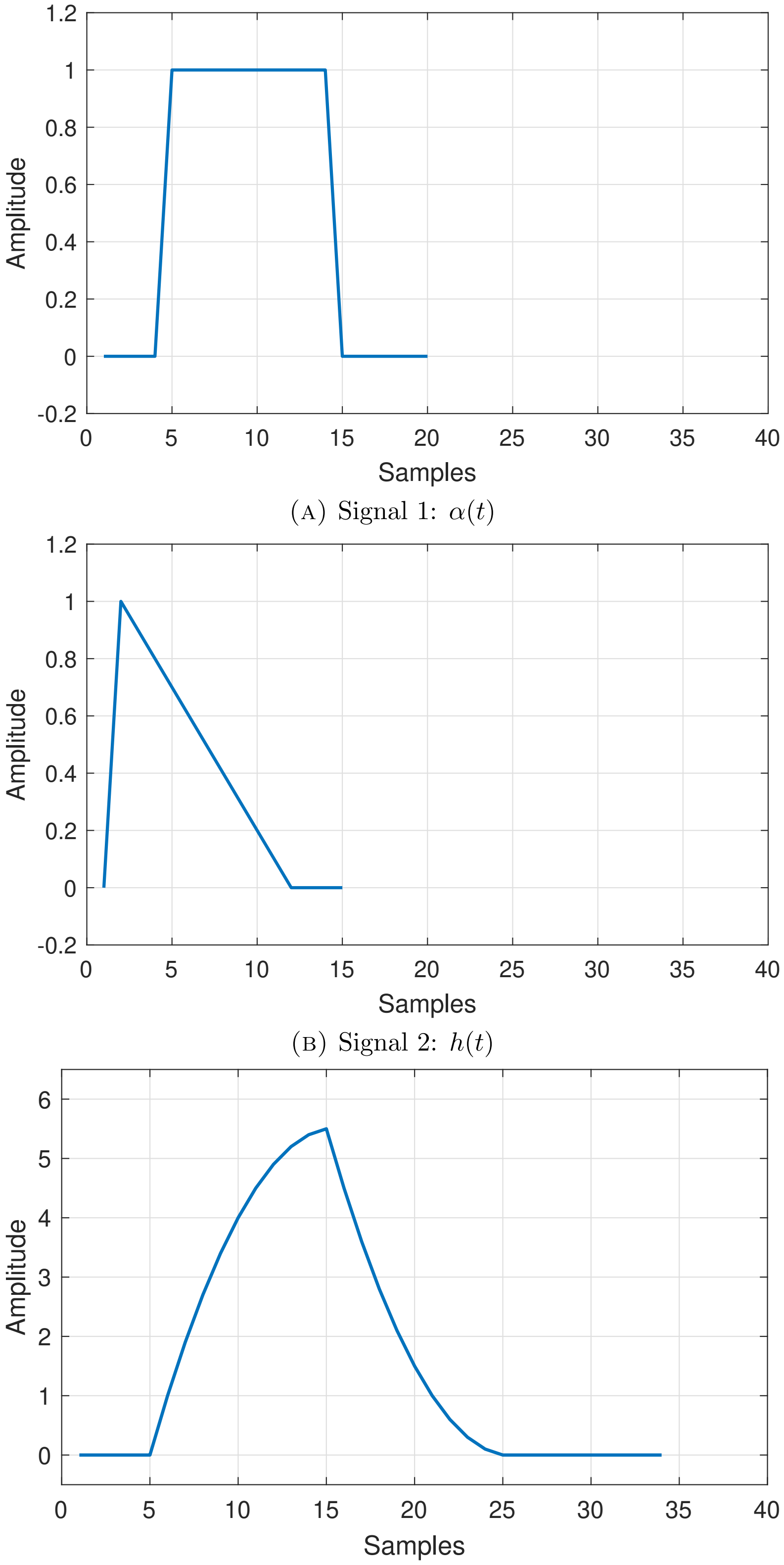
In digital signal processing, two signals $\alpha(t)$ and $h(t)$ can be combined to create a third signal (see Figure 2.12) using an operation called *convolution*, such that

$$\nu(t) = \alpha(t) * h(t) \tag{2.7}$$

where $*$ denotes convolution (Smith, 1997). Each sample of the first input signal is multiplied by every sample of the second input signal and the result is summed. The length of the output signal will therefore be one less than a summation of the two input signal lengths. Observing Figure 2.12, it is notable how the convolution result $\nu(t)$ has a much greater peak amplitude than the two input signals. Convolution is used widely in acoustics for processes such as imparting a measured impulse response to an anechoic signal and equalisation, and is used extensively in the work presented in this thesis.

## 2.4 The Human Auditory System

The ear is made up of three main parts: the outer ear, middle ear and inner ear (see Figure 2.13). The outer ear includes the *pinna* and *ear canal*. Sound waves travel down the ear canal and vibrate the eardrum, also known as the *tympanic membrane*, where the sound changes medium from air in the ear canal to liquid in the middle-ear.



(A) Signal 1: $\alpha(t)$

(B) Signal 2: $h(t)$

(C) Result of convolution: $\nu(t)$

FIGURE 2.12: Convolution of two signals, $\alpha(t)$ and $h(t)$, to give $\nu(t)$.

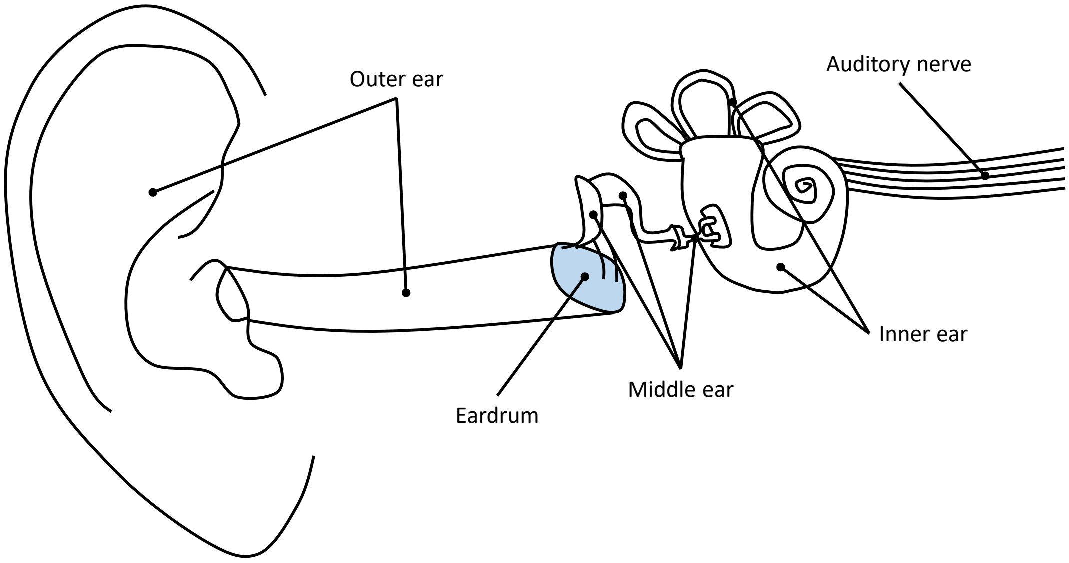


FIGURE 2.13: The anatomy of the ear, adapted from Howard and Angus (2017).

The three bones in the middle ear, known collectively as the *ossicles*, amplify the vibrations from the eardrum. The vibrations from the ossicles are then converted into nerve impulses in the inner ear and transmitted to the brain via the *auditory nerve*. The inner ear is made up of the *cochlea*, which is a coiled organ in which the *basilar membrane* sits, and the *semicircular canals*. On the basilar membrane are specific hair cells, called *cilia*, that respond to different frequencies to change vibrations into nerve impulses. High frequencies are picked up by cilia closest to the middle ear, and low frequencies are registered by the cilia towards the apex of the cochlea and closest to the auditory nerve.

As stated in Section 2.3, the quietest sound audible to the human ear is 20 $\mu$Pa. The frequency range of human hearing is often defined as from 20 Hz to 20 kHz, though sensitivity to frequencies approaching both extremes of this range is reduced. For example, sounds in the region of 15 kHz to 20 kHz are barely perceptible (Spagnol, Hiipakka and Pulkki, 2011), and high frequency hearing decreases with ageing (Dobreva, Neill and Paige, 2011), damage or disease. Within the audible frequency range, the sensitivity of the human ear to pressure levels is more variable. Sensitivity diminishes as frequency decreases below 1 kHz and increases above 15 kHz, and humans are most sensitive to frequencies around 1 kHz - 5 kHz, where speech

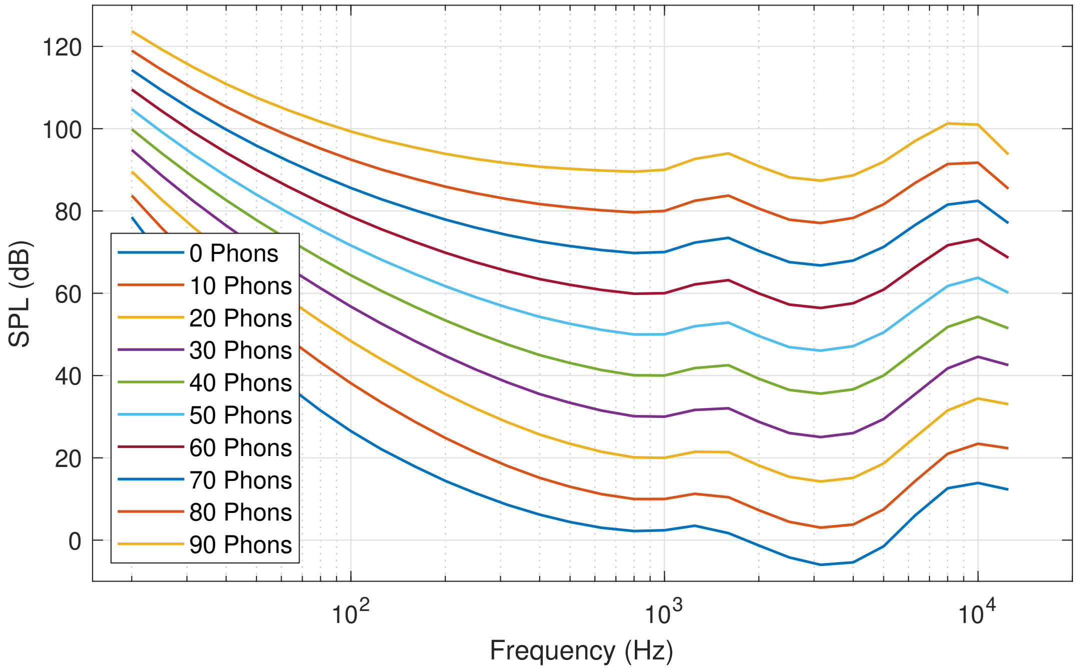


Figure 2.14: ISO 226 equal loudness curves illustrating the frequency and loudness variation of the human ear sensitivity.

articulation features such as sibilance and stops occur (Lourens, 1991; Stevens, 2000). The human auditory system frequency sensitivity has been measured by many researchers and the resulting frequency sensitivity plots are referred to as equal loudness curves (Fletcher and Munson, 1933; Fletcher and Munson, 1937; Bauer and Torick, 1966), which have been revised and are now part of the ISO 226 standard (International Organization for Standardization, 2003).

Figure 2.14 illustrates the ISO 226 equal loudness curves for 0 - 90 Phons. The *Phon* scale is an audio loudness unit that accounts for the variable human loudness sensitivity due to frequency. Two sinusoidal tones at different frequencies with the same Phon value will have the same perceived loudness, despite possibly having different amplitudes. The Phon value for any given frequency corresponds to the dB SPL level required to produce the same perceived loudness at 1 kHz. Looking at Figure 2.14, for example, a 1 kHz tone at a loudness of 10 Phons will have an amplitude of 10 dB SPL. The equivalent loudness of 10 Phons at 100 Hz will have an amplitude of 38.1 dB SPL, and a 10 Phon sound at 10 kHz will require an amplitude of 23.4 dB SPL.

The *sone* scale is based on human perception of loudness. The sone scale is calibrated such that 1 sone equals 40 Phons. Sones are calculated from Phons using the following calculation:

$$\text{sone} = 2\left(\frac{\text{Phon} - 40}{10}\right) \tag{2.8}$$

In the sone scale, therefore, a doubling in perceived loudness corresponds to twice the number of sones, where in the Phon scale, it would approximate to an increase in 10 Phons (Stevens, 1955; Bauer and Torick, 1966).

Anatomy varies greatly between humans and thus the cues for determining sound source location created by one person's body are highly specific to that individual. The size of one's head, spacing between the ears, size and shape of their pinnae, length of neck, shoulders, torso, posture and even the clothes and hair (Treeby, Pan and Paurobally, 2007) all have an effect on the sound arriving at the eardrums. The average human head radius is 8.75 cm (Kuhn, 1977), with ears slightly below and behind the centre point (Avendano, Algazi and Duda, 1999; Algazi, Avendano and Duda, 2001a).

The ear canal is on average 25 mm long and 8 mm wide (Chan and Geisler, 1990) and has the shape of a slightly curved cylinder with an oval shaped cross section (Blauert, 1997). The eardrum meets the canal at an angle between 40° - 50° (Blauert, 1997). It has been shown that, for most of the audible frequency spectrum, sound transmission from any point inside the ear canal is almost entirely independent of direction (Hammershøi and Møller, 1996; Hiipakka, Tikander and Karjalainen, 2010; Hiipakka, Kinnari and Pulkki, 2011; Hiipakka et al., 2012). Although it can be seen as not contributing to direction-dependent localisation cues, the ear canal does have an effect on incoming sound, with behaviour similar to that of a quarter-wave resonator (Hiipakka, Tikander and Karjalainen, 2010). It resonates at several frequencies determined by the length and shape of the ear canal, with the first resonance at roughly 3 kHz for the average adult human (Hiipakka, 2012). At the point of the eardrum, these resonances can be as high as 20 dB greater than at the entry to the ear canal in the frequency range of 1 kHz to 6 kHz (Griesinger, 2017). The exact frequencies at which the resonances occur vary significantly from person to person. Furthermore, any changes to the impedance at any point along the ear canal (for example wearing headphones) will alter the magnitude of the ear canal resonances or, in some cases, remove them completely (Griesinger, 2016).

A secondary way in which sound can travel to the inner ear is via the bones in the head and skull, through a process called *bone conduction*. However, bone-conducted sounds can be 40 dB lower in amplitude than sounds arriving at the eardrum, and therefore the localisation cues of bone-conducted sound are generally considered perceptually irrelevant (Griesinger, 1990; Blauert, 1997; Moore, 2009).

## 2.5 Binaural Sound Localisation

This section explains the cues for human binaural sound localisation. The accuracy of human sound localisation in the azimuthal plane ranges between 1° and 10° (Blauert, 1997) depending on the angle of incidence and the spectral content of the auditory stimuli. Head movements also improve the accuracy of horizontal localisation (Iwaya, Suzuki and Kimura, 2003), as do early reflections (Rakerd and Hartmann, 1985), whereas moving sound sources increase localisation blur (Gorzel et al., 2011). Vertical localisation accuracy ranges between 4° and 22° (Blauert, 1997) and is therefore less accurate than horizontal localisation, though again head movements also improve height localisation (Perrett and Noble, 1997b), even when high frequencies are absent from the stimuli (Dan and Xie, 2005).

It has been also shown that humans are more accurate at localising familiar sounds (Blauert, 1997), and can improve their localisation ability through participation in extended listening tests (Asano, Suzuki and Sone, 1990; Blauert, 1997; Steadman et al., 2017; Stitt, Picinali and Katz, 2019). Sounds that are short in duration are harder to localise (Hartmann et al., 2010), which is likely due to the limited opportunity to utilise head movements. Additionally, the direction in which people's

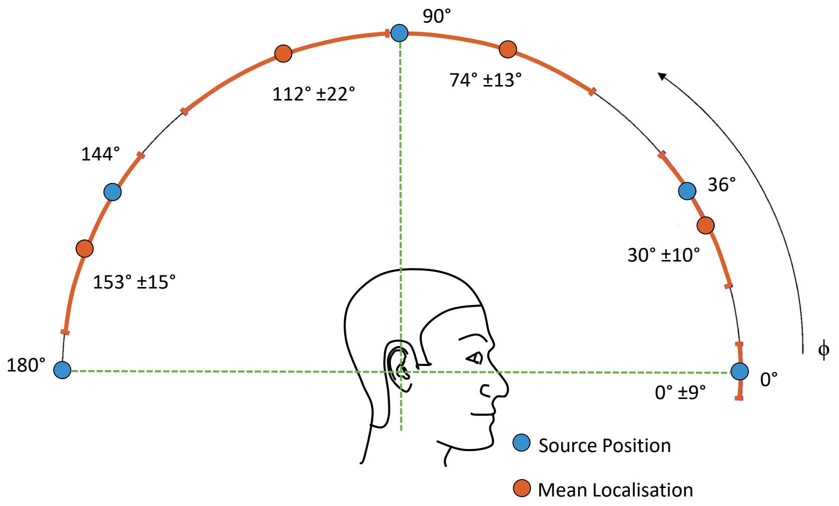


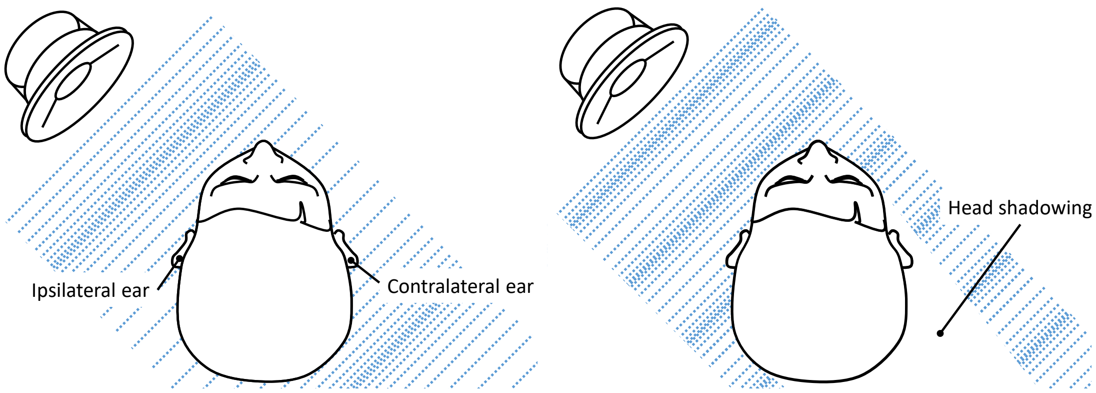
Figure 2.15: Elevation localisation blur, adapted from Damaske and Wagener (1969).

eyes are looking has an effect on perceived sound source location, as vision overrides hearing (Lewald, 1997). In some cases elevation localisation can be successfully relearned when pinnae are physically modified (Hofman, Van Riswick and Van Opstal, 1998; Shinn-Cunningham, Durlach and Held, 1998; Van Wanrooij and Van Opstal, 2005; Klein and Werner, 2015; Watson et al., 2017).

The human auditory system can localise frontal sounds more accurately than lateral, rear and elevated directions (Blauert, 1997). Damaske and Wagener studied localisation blur of continuous speech in the upper hemisphere. They found that for frontal sources, sagittal plane localisation blur is less than half the value at elevated rear sources, at ±9° for frontal and ±22° for elevated rear sources (Damaske and Wagener, 1969), as illustrated in Figure 2.15. They also showed that in the horizontal plane, localisation blur of frontal sources is again less than half that at lateral sources, at ±3.6° for frontal and $>$ ±9° for lateral sounds.

### 2.5.1 Horizontal Localisation

Because the ears are horizontally placed on the side of the head, sound localisation on the horizontal plane is largely determined by time and level discrepancies between the signals arriving at each eardrum (Rayleigh, 1907), known as *interaural differences*.



(A) Low frequency (B) High frequency

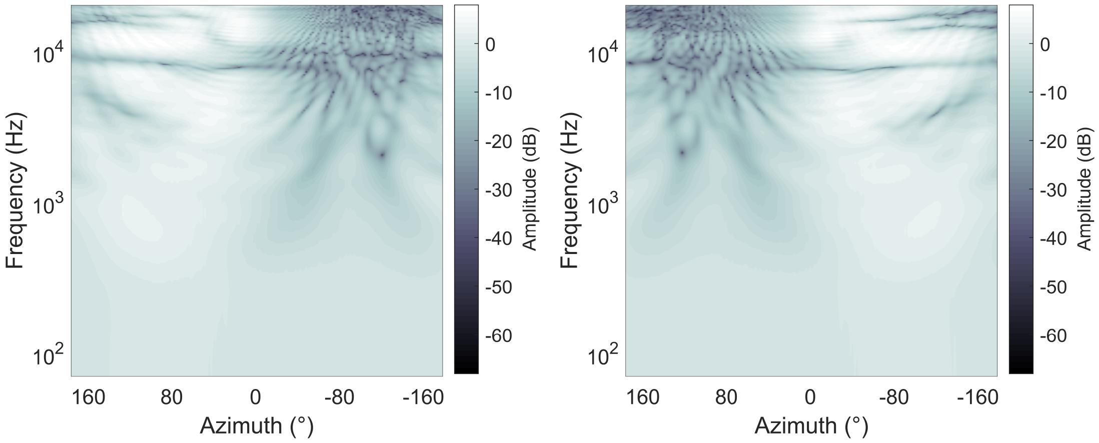
FIGURE 2.16: Illustration of both low and high frequency sounds arriving at an incidence of $\theta = 45°$.

Due to the frequency-dependent wave properties of sound, these interaural cues contribute to horizontal localisation at different frequency regions.

Consider a sound arriving at an azimuth of $\theta = 45°$ (as illustrated in Figure 2.16). The path between the source and the *ipsilateral*[1] ear is smaller than to the *contralateral*[2] ear, which causes the sound to arrive at the left ear earlier in time than at the right. At low frequencies, the sound will diffract around the head to the contralateral ear. This is the *interaural time difference* (ITD), sometimes referred to as interaural phase difference. ITD contributes to localisation in frequencies up to roughly 1.5 kHz (Kuhn, 1977; Blauert, 1997; Cheng and Wakefield, 1999), where the wavelength is approximately 23 cm, and humans can differentiate ITDs as small as 10 $\mu$s (Moore, 2012). Above this frequency, the wavelength of sounds become comparable or smaller than the size of the human head, which on average has a diameter of 17.5 cm (Kuhn, 1977). This causes phase ambiguity and makes ITD less detectable. However, at directions approaching the median plane, ITDs have been shown to be perceivable up to 3 kHz (Smith and Price, 2014a).

Looking at Figure 2.16 again, consider how the path to the contralateral ear is partially occluded by the head. At high frequencies, the head acts as a baffle and produces an acoustic shadow. This, along with the greater distance to the right ear,

[1] Referring to the side of the head closest to the incoming sound.
[2] Referring to the side of the head furthest from the incoming sound.



(A) Left ear

(B) Right ear

FIGURE 2.17: Frequency responses of left and right ears over the horizontal plane, from the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

means that sound arrives at the contralateral ear at a lower amplitude than at the ipsilateral ear. This is the *interaural level difference* (ILD), sometimes referred to as interaural intensity difference. The ILD contributes to horizontal localisation at frequencies above roughly 3 kHz, though some ILD is observable as low as 400 Hz.

In the period of crossover, between 1.5 kHz and 3 kHz, horizontal localisation accuracy is lower due to frequencies being too high to provide clear temporal cues (ITD) and too low to provide sufficient level cues (ILD) (Middlebrooks and Green, 1991). If the stimulus is wideband and includes both low and high frequencies, ITD has been shown to be the dominant horizontal localisation cue (Wightman and Kistler, 1992; Macpherson and Middlebrooks, 2002). Figure 2.17 demonstrates the frequency responses of the left and right ears with varying azimuth angle at an elevation of $\phi = 0°$ (more detail on these frequency responses is provided in Section 2.6.1). The plots are near-symmetric, and the bright spots show the higher amplitude of the ipsilateral signals, illustrating the changes in ILD with azimuth.

There do exist areas around the human head where the interaural cues will share the same values. This is referred to as the *cone of confusion*, and can cause errors in judgement of front-back or up-down location of sounds (Wenzel et al., 1993; Wightman and Kistler, 1999). Figure 2.18 illustrates the cone of confusion: incoming sounds situated on the cones will feature the same ITD and ILD. The way in which

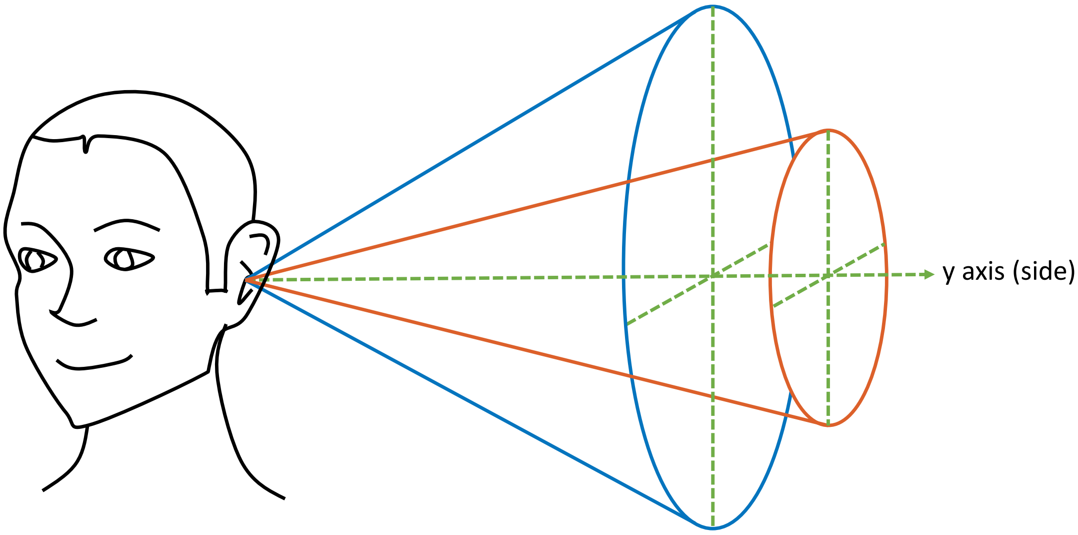


FIGURE 2.18: Illustration of the *cone of confusion*, where sounds incident on the cones have constant interaural time and level differences, adapted from Wenzel et al. (1993).

front-back and up-down confusions are usually resolved is through head movements (Thurlow and Runge, 1967; Perrett and Noble, 1997a; Noisternig et al., 2003b), as when rotating the head, the direction from which the sound is coming relative to the ears changes, and therefore the localisation cues change too. Asano *et al.* claim that frequencies between 500 Hz and 2 kHz must be accurately reproduced for front-back judgement (Asano, Suzuki and Sone, 1990).

### 2.5.2 Vertical Localisation

Vertical localisation cues are mainly made up of changes in the frequency spectrum of incoming sounds caused by constructive and destructive acoustic interference due to reflections and diffractions from various parts of the body such as the outer ears, head, shoulders and torso. The brain compares the frequency content of the incoming sound to memory. It is considered to be primarily a monaural process: changes due to elevation angle are predominantly present for both ears (Hebrank and Wright, 1974a). However, it is not entirely monaural (Morimoto, 2001), as changes in the elevation of a sound source do produce a subtle change in ITD (Avendano, Algazi and Duda, 1999; Hartmann et al., 2010), due to the asymmetric position of the ears on the head (they are slightly below and behind the centre point) (Algazi, Avendano and Duda, 2001a). Frequency content is not the only thing that affects elevation perception, as louder audio stimuli levels have been found to produce higher errors in elevation localisation (Hartmann and Rakerd, 1993; Vliegen, John and Opstal, 2004). The presence of early reflections has also been shown to produce a higher perceived elevation than the same sound without reflections (Begault, 1992a; Begault and Wenzel, 1993; Begault, Wenzel and Anderson, 2001).

Elevation cues exist between 700 Hz and 18 kHz, but are strongest above 5 kHz (Roffler and Butler, 1967; Asano, Suzuki and Sone, 1990). In order to determine the elevation angle of a sound, it must be broadband and include energy at frequencies above 700 Hz (Hartmann et al., 2010). Though the sound must include energy at a wide range of frequencies, the frequency spectra need not be *flat* (Vliegen, John and Opstal, 2004), where the term flat refers to a uniform response at all frequencies. Elevation cues are caused mainly by reflections from the *pinnae* (Hebrank and Wright, 1974b; Raykar, Duraiswami and Yegnanarayana, 2005), which cause comb filtering (as explained in Section 2.3.3) and produce notches in the frequency spectrum. The pinnae contribute significantly to sound source localisation in the median plane with effects on certain frequencies up to ±20 dB (Brinkmann et al., 2014b). Due to the physical size of pinnae, it can be deduced that the pinna contribute localisation cues in frequencies above roughly 3 kHz, though the most prominent effects of the pinnae appear in the range of roughly 6 kHz to 18 kHz. The main pinna cue is from the *concha* (Hebrank and Wright, 1974b; Raykar, Duraiswami and Yegnanarayana, 2005), which is the cavity closest to the eardrum (see Figure 2.19). The frequency of the notches is determined by the delay between the direct sound and the concha reflection, which changes depending on the elevation of the incoming sound due to the shape of the concha.

Figure 2.20 demonstrates the frequency responses of the left and right ears with varying elevation angle at an azimuth of $\theta = 0°$ (more detail on these frequency responses is provided in Section 2.6.1). The plots are near-identical which corroborates

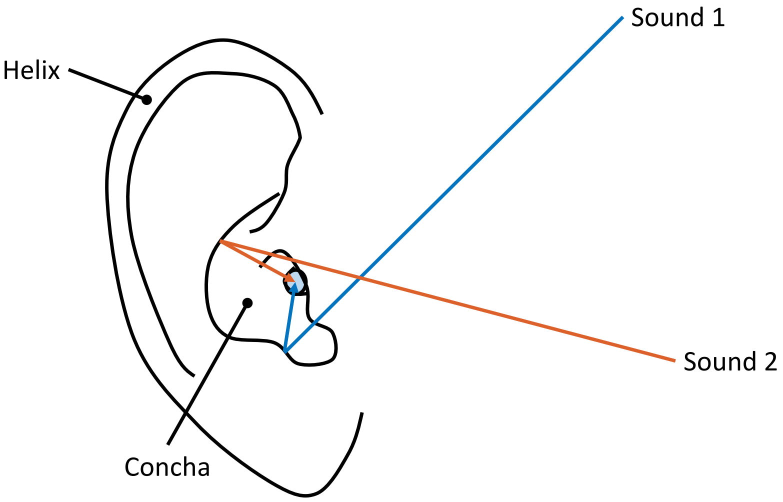
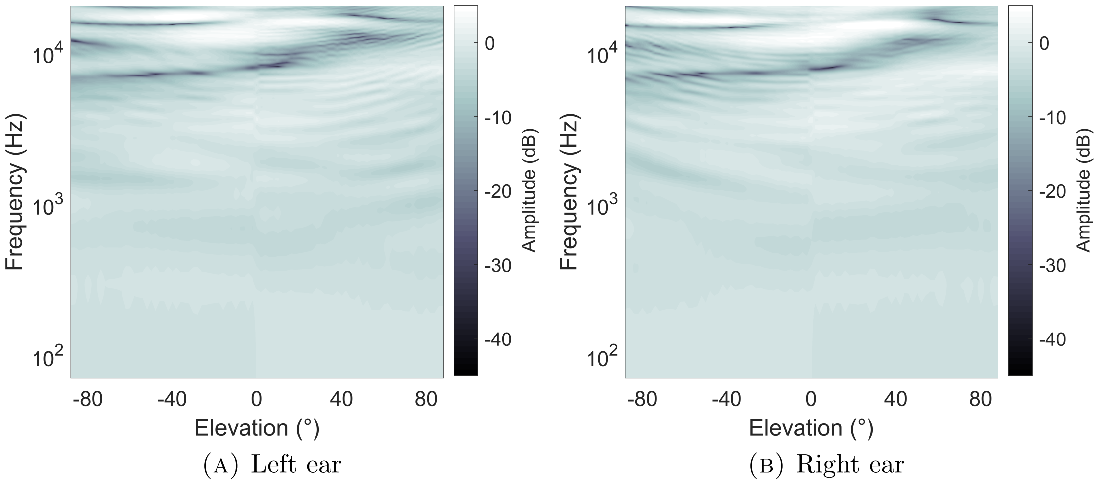


FIGURE 2.19: Illustration of the concha reflection for frontal sounds and how it differs with elevation, adapted from Hebrank and Wright (1974b).



(A) Left ear

(B) Right ear

FIGURE 2.20: Frequency responses of left and right ears over the median plane, from the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

the approximately monaural nature of elevation cues (Gardner, 1973; Wightman and Kistler, 1997), and the general increase in frequency of the deepest notch, from approximately 7 kHz - 12 kHz with rising elevation, illustrates the concha reflections.

The exact frequencies of elevation notches are dependent on the size and shape of the pinnae, which vary greatly between individuals (as illustrated in Figure 2.21). In general, as the angle of elevation increases, the frequency of the deepest notches tends to increase (Algazi et al., 2001). Other secondary features of the pinnae are elevation resonances (Raykar, Duraiswami and Yegnanarayana, 2005), which are frequencies

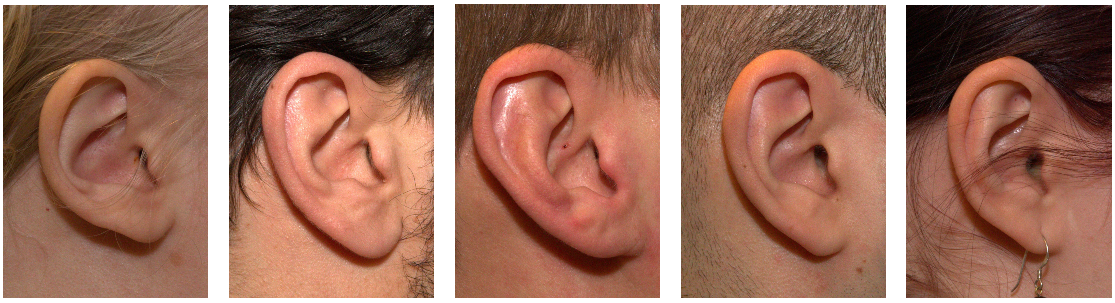


FIGURE 2.21: Variation in pinna shapes and sizes: subjects 1 - 5 from the Spagnol pinna database, reproduced from Spagnol, Hiipakka and Pulkki (2011).

where amplitude is increased, and rear incident sounds have more damped high frequencies due to pinnae shadowing.

Though the main cues for vertical localisation are caused by the pinnae, secondary cues do exist at frequencies as low as 700 Hz (Gardner, 1973; Algazi, Avendano and Duda, 2001a). These are from the torso (Brown and Duda, 1998; Avendano, Algazi and Duda, 1999). Unlike the pinna notches, these low frequency elevation cues decrease in frequency as elevation increases. Analysis of measurements in the time domain shows reflections that change in time depending on elevation angle, with the time of the reflection increasing as elevation approaches $\phi = 90°$ (Avendano, Algazi and Duda, 1999). The timing of these reflections correspond to the path from an incident sound to the shoulders and torso and then to the eardrum (Brown and Duda, 1998; Avendano, Algazi and Duda, 1999; Algazi, Avendano and Duda, 2001a). Some researchers have studied how the torso effects change when the head rotates on a stationary torso (Lewald, Dörrscheidt and Ehrenstein, 2000; Guldenschuh et al., 2008; Brinkmann et al., 2015a; Brinkmann et al., 2017a). Torso effects are more important when the sound incidence is of a low elevation (Algazi et al., 2002; Kirkeby et al., 2007). Additional low frequency elevation cues, at frequencies as low as 400 Hz, are from knee reflections (Algazi et al., 2001; Raykar et al., 2003; Raykar, Duraiswami and Yegnanarayana, 2005), though these are generally considered perceptually irrelevant due to their low relative amplitude.

Shoulder and torso cues are said to be of secondary importance to the pinnae (Gardner, 1973; Searle et al., 1975) as their impact on the frequency spectrum is of a smaller magnitude (up to ±5 dB) (Brinkmann et al., 2014b). However, they exist at frequencies where many real world sounds such as speech have more energy and should therefore still be treated as significant. Additionally, front-back confusion is reduced when the correct low frequency spectral content is present (Asano, Suzuki and Sone, 1990).

### 2.5.3 Distance Localisation

Humans can also determine the distance of sounds. This judgement is primarily made using the amplitude of the sound (Middlebrooks and Green, 1991) and the amount of reverberation present (Begault, 1992b), though distance is also judged to increase with low and high frequency roll off (Middlebrooks and Green, 1991). At distances greater than roughly 1 m, sound is said to be in the *far-field*, which means the sound waves can be modelled as plane waves and binaural localisation cues do not change (Brungart and Rabinowitz, 1999). At distances less than 1 m however, referred to as the *near-field*, the spherical nature of sound waves becomes relevant, so the distance between sound source and head affects the spectra of sounds (and therefore the binaural cues) at the eardrums. ILD increases with lateral sources as the distance gets smaller than 1 m, even at low frequencies (Brungart and Rabinowitz, 1999; Shinn-Cunningham, Santarelli and Kopco, 2000). ITD and elevation dependent cues do not appear to change as much in the near-field.

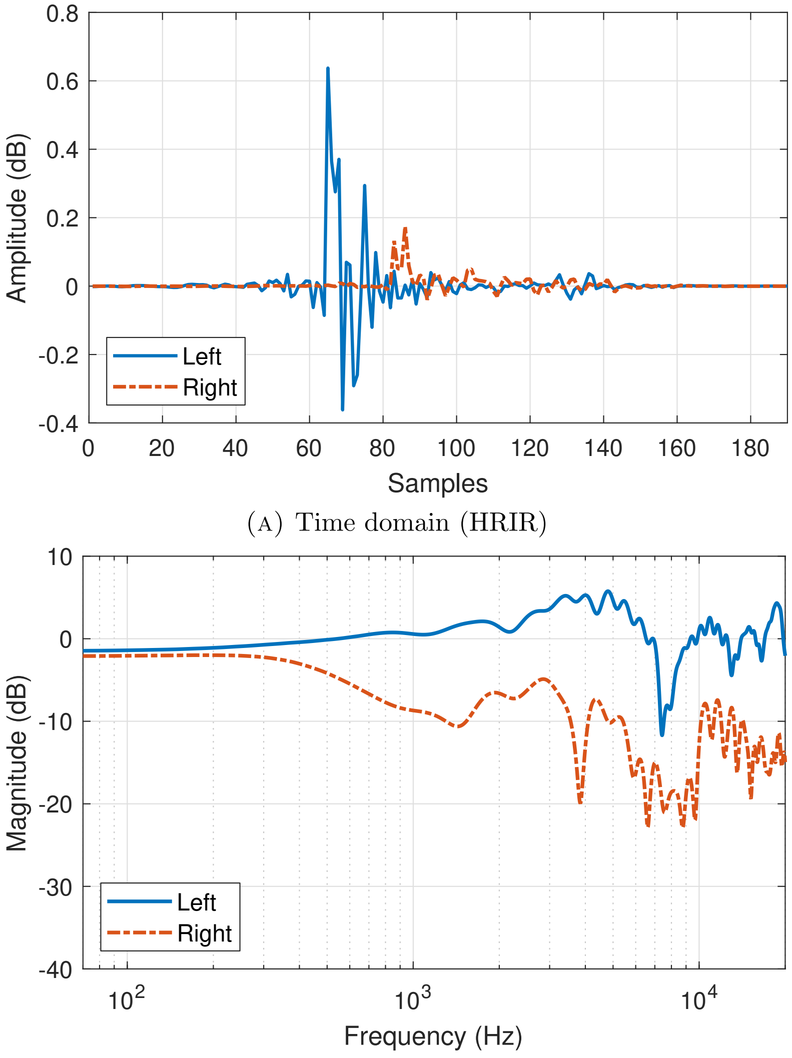
Another factor that can be placed into the category of distance localisation is the feeling of externalisation: that the sound source is outside of the head (Møller et al., 1996; Tan and Gan, 2000; Kim and Choi, 2005). The feeling of externalisation is mainly caused by the presence of early reflections (Völk, 2009) and reverberation (Schroeder, 1970; Durlach and Colburn, 1978; Begault, 1992b; Begault, Wenzel and Anderson, 2001; Catic, Santurette and Dau, 2015), though accurate timbre (Hartmann and Wittenberg, 1996) and head movements (Wenzel, 1995; Wightman and Kistler, 1997) also contribute. It is said to be a subjective feeling and therefore not one that can be categorically measured (Durlach et al., 1992).

## 2.6 Binaural Technology

There are many current systems that use the manner in which the human auditory system localises sound to produce a simulation of spatial audio. The simplest and most common systems are stereophonic and 5.1 surround sound, though these are both limited and do not provide height. As this thesis investigates spatial audio for virtual reality applications, this section will focus mainly on audio systems that can deliver three-dimensional sound over headphones. When the implementation is accurate, binaural audio can be virtually indistinguishable from loudspeaker audio (Langendijk and Bronkhorst, 2000; Martin, McAnally and Senova, 2001), and can be deemed perceptually authentic (Brinkmann, Lindau and Weinzierl, 2017).

### 2.6.1 The Head-Related Transfer Function

Consider again the scenario depicted in Figure 2.16 with a loudspeaker facing a human head at an incidence of $(\theta = 45°, \phi = 0°)$. If an *impulse* $\delta(t)$ (illustrated in Figure 2.9) is played out of the loudspeaker in free-field (anechoic) conditions, and microphones are placed at the point of the eardrums, it is possible to measure the change in sound between the source and eardrums due to the head, torso and ears (Xie, 2013). Therefore the effects of the human body shape, that allow us to determine the location of a sound, can be recorded. This is referred to as the *head-related transfer function* (HRTF) (Cooper, 1982). The time domain representation of the HRTF is the *head-related impulse response* (HRIR). In this thesis, the transfer function will be generally referred to as the HRTF when not specifically referring to the time-domain or frequency-domain representation. The HRIRs and HRTFs of a sound incident at $(\theta = 45°, \phi = 0°)$ are presented in Figure 2.22, obtained from the Bernschütz Neumann KU 100 database (Bernschütz, 2013). The plots demonstrate the ITD and ILD of a sound arriving from the left as described in Section 2.5.



(A) Time domain (HRIR)

(B) Frequency domain (HRTF)

FIGURE 2.22: Time domain and frequency domain representations of a head-related impulse response recorded at $(\theta = 45°, \phi = 0°)$ from the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

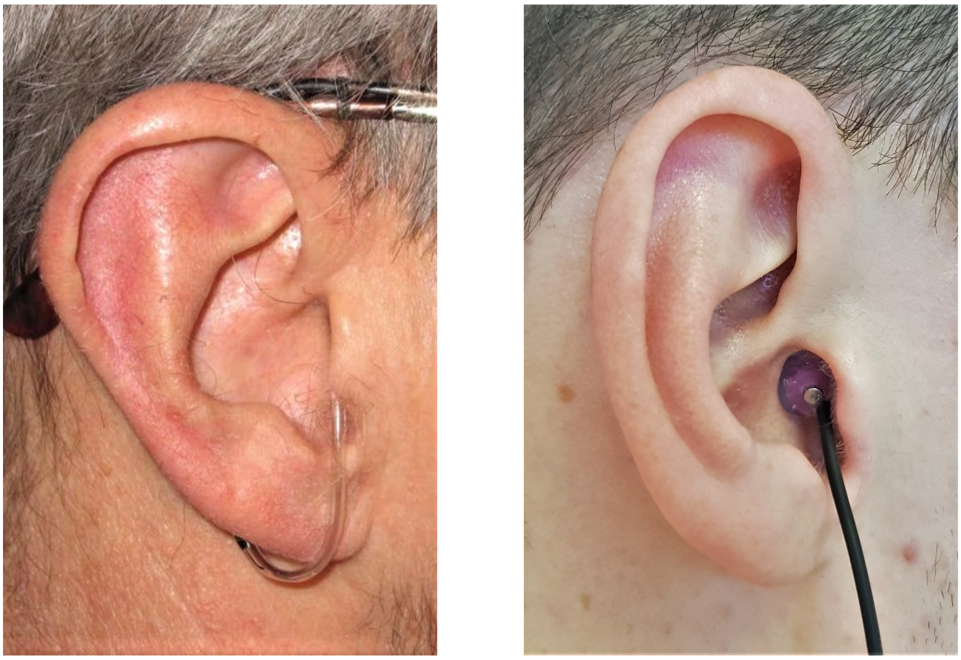
Convolution of a monophonic sound with an HRTF will generate a sound signal that could appear to originate from the position the HRTF was measured from, when played back at the eardrums.

### 2.6.2 HRTF Measurement

The HRTF is most commonly obtained by placing miniature microphones in the ear canals and recording the impulse response from a specific point in space (Algazi et al., 2001; Warusfel, 2003; Kearney and Doyle, 2015a; Armstrong et al., 2018a) using the IR measurement methods detailed in Section 2.3.2. As the HRTF aims to measure the change between a sound source and the eardrums, one could assume that the microphones for recording HRTFs should be placed at, or as close as possible to, the position of the eardrums (as in Figure 2.23a). However, physical measurements made at the position of the eardrum (Bronkhorst, 1995; Chen, Van Veen and Hecox, 1995; Langendijk and Bronkhorst, 2000; Hiipakka et al., 2012) require highly specific probe microphones and are potentially more dangerous than those made at the ear canal entrance, due to the delicate nature of the tympanic membrane which can be easily damaged or perforated. In some cases probe microphones have been fitted by medical doctors (Hammershøi and Møller, 1996). Furthermore, due to the small physical size of probe microphone designs, the frequency response of probe microphones are considered unreliable above 4 kHz (Hiipakka, 2012).

As the ear canal has little direction-dependent impact (Hammershøi and Møller, 1996) it is possible to instead measure HRTFs at the entrance to the ear canal. However, it is important that the playback system does not interfere with the impedance of the ear canal (Møller, 1992), as it has been shown that measurements made at the ear canal entrance can impart a greater timbral colouration than measurements made at the eardrum (Takanen, Hiipakka and Pulkki, 2012). Therefore when measuring HRTFs at the ear canal entrance, appropriate compensation is necessary. For measurements at the entrance of the ear canal, the ear canal is usually blocked, as shown in Figure 2.23b (Algazi et al., 2001; Takane et al., 2002; Kearney and Doyle, 2015a; Armstrong et al., 2018a), to reduce ear canal resonances (Hiipakka, Tikander and Karjalainen, 2010). This also reduces the magnitude of variations in measurements between individuals (Møller et al., 1995).

More than one HRTF measurement can be made at a time by using an *overlapped swept sine* technique (Majdak, Balazs and Laback, 2007), which can reduce the time required for acquisition of a large dataset of measurements. There are also other ways to measure HRTFs. Using the principle of reciprocity, the loudspeaker and microphone positions can be switched by placing a miniature loudspeaker in the ear



(A) Probe microphone at the eardrum with an open ear canal, reproduced from Griesinger (2013)

(B) Microphone at the entrance to a blocked ear canal, reproduced from Armstrong et al. (2018a)

FIGURE 2.23: Different microphone placement techniques for binaural audio recording.

canals and recording the output at a specific position using microphones. This can in theory allow rapid recording of multiple HRTFs using only one sweep, by placing numerous microphones around the head (Zotkin et al., 2006).

Alternatively, it is possible to simulate HRTF measurements using methods such as the *boundary element method* (BEM) and *finite element method* (FEM), by using a three-dimensional surface mesh of the head and torso (Brinkmann et al., 2017b). Simulations have been shown to offer close performance to measurements (Algazi et al., 2002; Brinkmann et al., 2015b), though accurate BEM suitable meshes are difficult to obtain for human subjects, and the HRTF simulation process is computationally expensive.

Unless the measurements are made in an anechoic environment, unwanted reflections such as those from the room or other loudspeakers can be present in HRTF measurements. These are usually unwanted, unless the measurements are *binaural room impulse responses* (BRIRs). Unwanted reflections can usually be removed by calculating the path difference between the direct sound and early reflections and truncating the measured HRTFs at the corresponding time (Algazi et al., 2002; Kearney and Doyle, 2015a). However, this will change the accuracy of low frequency reproduction, which as a consequence is frequently modelled (Algazi et al., 2002; Kan, Jin and Schaik, 2009; Bernschütz, 2013; Kearney and Doyle, 2015a). Modelling the low frequencies also helps to compensate for the often limited low frequency reproduction capabilities of the loudspeakers used in the HRTF measurement process.

Other unwanted attributes of HRTFs are the influence of the transducers in the microphones and loudspeakers used in measurement process, as transducers will not have an entirely flat frequency response. These can be removed in a process known as *free-field equalisation*, through deconvolution of the recorded HRTF measurement with a reference free-field measurement, such as the same transducers without the head present (Brinkmann, Lindau and Weinzierl, 2013). Alternatively, many HRTF databases employ *diffuse-field equalisation* to remove the direction-independent aspects of a measurement set (Merimaa, 2009), which will be explained in greater detail in the coming sections.

In order to synthesise sound from all directions, considering the spatial resolution of human hearing is as low as 1° in azimuth and 4° in elevation, a large amount of measurements are necessary which can be a time consuming task. As mentioned, one way to speed up the HRTF measurement process is through using multiple loudspeakers at once, and overlapping the playing of sweeps, offsetting them by the reverberation time of the environment. However, this must be done with care to avoid introducing other issues such as harmonic distortion and reflections from the other loudspeakers (Majdak, Balazs and Laback, 2007; Armstrong et al., 2018b).

### 2.6.3 Individualisation

As every individual's body is a different shape and size, individualised binaural recordings and HRTFs produce the most natural and believable binaural experience

(Møller et al., 1996), offering more accurate localisation and timbre than non-individualised HRTFs (Wenzel et al., 1993; Bronkhorst, 1995; Tan and Gan, 2000), and greater externalisation (Kim and Choi, 2005). Azimuth cues (both ITD and ILD) can produce relatively robust horizontal localisation for listeners using non-individualised HRTFs (Wenzel et al., 1993), due to the inter-individual variation in head size and ear spacing being relatively small. For spectral cues however, inter-individual variation in pinna size and shape is much larger (as demonstrated in Figure 2.21), which leads to front-back and up-down confusions, inaccurate perception of sound source distance, and lack of externalisation (Møller et al., 1996; Tan and Gan, 2000) when using non-individualised HRTFs. One feature of individualised HRTFs is the inherent asymmetry between the left and right ears, which has been shown to improve externalisation (Brookes and Treble, 2005).

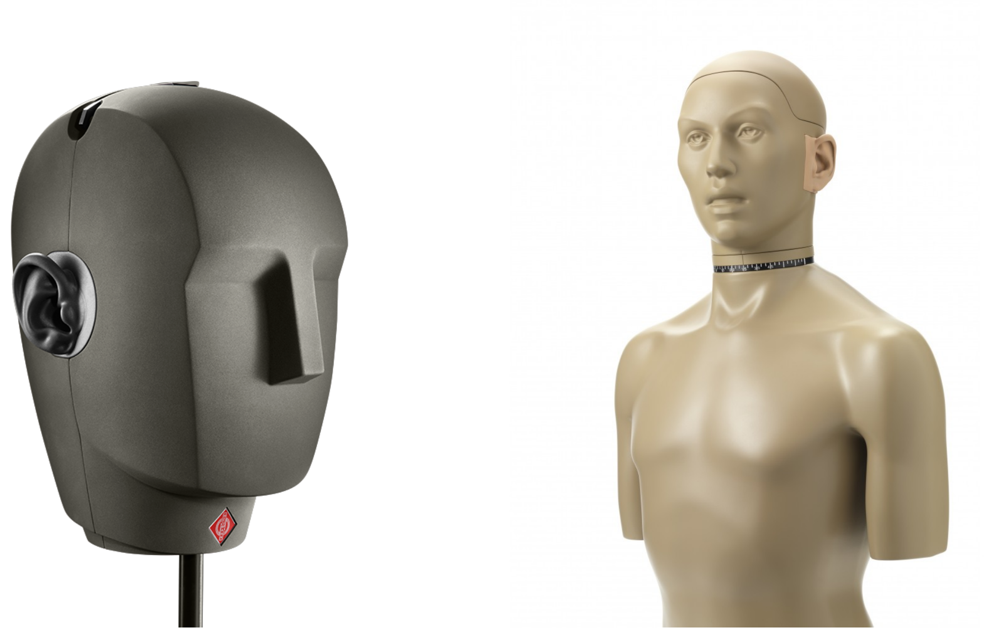
However, it is not always practical to measure every listener's HRTFs individually, as it is a time-consuming and laborious process that requires specific equipment, highly precise set-up and ideally an anechoic environment. The typical measurement process also requires a participant to stay motionless for a long period of time which can be fatiguing. For wide use individualised HRTFs are therefore not practical, and generic HRTFs produced from dummy heads (see Figure 2.24) are often utilised, such as the Neumann KU 100[3] or G.R.A.S. KEMAR[4]. Some efforts to personalise non-individualised HRTFs to improve localisation accuracy have been made by analysing anthropomorphic features (Dinakaran et al., 2006; Geronazzo et al., 2014), and some individualised HRTF sets are preferred to others (Katz and Parseihian, 2012).

### 2.6.4 Headphone Choice and Equalisation

With the exception of transaural systems that play back binaural sound over two or more loudspeakers using crosstalk cancellation filters (Cooper and Bauck, 1989;

[3] `https://en-de.neumann.com/ku-100`

[4] `https://www.gras.dk/products/head-torso-simulators-kemar/kemar-non-configured/product/749-45bc`



(A) Neumann KU 100, reproduced from Neumann (2013).

(B) G.R.A.S. KEMAR, reproduced from G.R.A.S. Sound & Vibration (2013).

FIGURE 2.24: Dummy heads for binaural recording.

Gardner, 1998), binaural audio is always reproduced over headphones. However, even with individualised binaural audio recordings and HRTFs, the headphone playback system can alter the acoustic impedance of the ear canals, which in practice can produce colouration and reduce the plausibility of the auditory experience. The type of headphone (in-ear, on-ear or over-ear) can have a dramatic effect on the sound (Satongar et al., 2015); for example, in-ear and on-ear headphones can suppress ear canal resonances completely (Griesinger, 2016; Schärer and Lindau, 2009). Typically, open-back over-ear (sometimes referred to as circumaural) headphones, that produce as close to free-air equivalent coupling (Møller et al., 1995) conditions as possible, are preferred for binaural audio as they have a smaller influence on ear canal resonances (Møller, 1992; Lindau and Brinkmann, 2012; Bolanos and Pulkki, 2015). Some researchers have even developed specific acoustically transparent headphones for practical in-situ comparisons to loudspeaker rendering (Schultz et al., 2010; Erbes et al., 2012; Brinkmann, Lindau and Weinzierl, 2017), though most researchers use commercially available open-back headphones such as the Sennheiser HD 650[5] or

[5] https://en-uk.sennheiser.com/high-quality-headphones-around-ear-audio-surround-hd-650

STAX headphones[6].

Even with a suitable choice of headphones, the transducers will not have a completely flat frequency response and therefore require compensation. Headphone equalisation has been shown to improve plausibility of binaural simulations when correctly implemented (Schärer and Lindau, 2009). Headphone equalisation is typically achieved by inverting the measured headphone transfer function (HpTF) between the headphones and the ear canals. The generated inverse filters are then applied to the binaural sounds. HpTF measurements are usually obtained from the average of multiple measurements taken with removal and replacement of the headphones on the ears, as even small displacements of the headphone on the ear can produce large changes in the HpTF (Kulkarni and Colburn, 2000; Masiero and Fels, 2011a; Masiero and Fels, 2011b), and equalisation based on just one measurement can produce poorer results than no equalisation at all (Kulkarni and Colburn, 2000). An additional benefit of using multiple measurements in the HpTF acquisition is that the deep notches in the response are smoothed out, producing a reduction in the sharp peaks in the inverse filter, which are more noticeable than troughs (Bücklein, 1981; Masiero and Fels, 2011b). When calculating inverse filters, the regularisation method by Kirkeby and Nelson (Kirkeby et al., 1998) with complex smoothing (Hatziantoniou and Mourjopoulos, 2000) is preferred perceptually to other inversion methods (Schärer and Lindau, 2009).

The HpTF is highly individual (Møller et al., 1995; Griesinger, 2016). Individualised headphone equalisation produces smaller deviations in reproduced sound than non-individual or generic compensation (Pralong and Carlile, 1996; Brinkmann and Lindau, 2010). However, as is the case with HRTFs and binaural recordings, individualised headphone equalisation is also not always feasible. Where individualised recordings are possible, individualised headphone equalisation should be used (Lindau and Brinkmann, 2010). When individualisation is not possible and in the case that the recordings are generic, such as from a dummy head, headphone equalisation filters generated using the same dummy head as for the binaural recordings have been

[6] `https://staxaudio.com/`

found to produce even greater naturalness than individual headphone compensation (Lindau and Brinkmann, 2010; Lindau and Brinkmann, 2012).

One alternative to free-air equivalent headphones is to equalise the transfer function between the headphone and the eardrum. Griesinger claims that in order to produce accurate frontal localisation in playback of binaural recordings, the ear canal resonances must be intact (Griesinger, 2017). Though the traditional way in which headphone equalisation with correct ear canal compensation is achieved through performing the equalisation at the position of the eardrum using a probe microphone, Griesinger developed another method whereby no microphones are necessary at all (Griesinger, 2016). The listener would adjust a multi-band equaliser whilst listening to noise from a frontal loudspeaker until it was perceived as having a flat frequency spectrum. The listener would then put on headphones and listen to the same sound, convolved with individual HRTF measurements of the same setup, repeat the equalisation process such that the headphone simulation of the noise appeared to have a flat frequency spectrum. The differences between the two equalisation curves would therefore be the desired headphone equalisation. With this, binaural sound would then appear to be externalised and frontal. However, no formal listening tests have been reported to evaluate this approach yet.

### 2.6.5 Dynamic Binaural Systems and Interpolation

As stated in Section 2.5, head movements are a key part of sound localisation and allow humans to resolve directional ambiguities such as the cone of confusion by subconsciously comparing the change in spectral and interaural cues to the change in the head's orientation (Blauert, 1997; Begault, Wenzel and Anderson, 2001). However in binaural spatial audio systems, headphones will remain at the same position relative to the ears when head movements are made. This will cause the soundfield to follow the head orientation changes unrealistically. Dynamic binaural systems use head orientation data from a head-tracking system, to update the headphone signals accordingly in order to maintain a stable orientation of the virtual sound scene (Mackensen et al., 2000).

When accurate and low latency, dynamic binaural systems can have little negative impact on the ability to localise sound (Hess, 2004). A latency of $< 64$ ms has been found to be necessary for detection rates lower than 4% of the time, however if the head movements are fast, the latency of the tracking and dynamic binaural synthesis system must be lower (Lindau, 2009). Furthermore, diffuse sounds are less critical to be updated spatially than highly directional sounds (Algazi and Duda, 2011). High accuracy and low latency tracking solutions include products such as OptiTrack[7], which uses multiple infra-red cameras to measure the position and orientation of a rigid body of light reflectors in three dimensions. Some commercially available headphones now even offer head orientation tracking, such as the Bose Noise Cancelling Headphones 700[8]. Virtual reality (VR) headsets track head orientation also, such as the Oculus Rift[9] and HTC Vive[10]. However, the influence of the head-tracking device on HRTF measurements should not be overlooked, as changes to ITD and ILD have been observed when wearing head-mounted displays (Porschmann, Arend and Gillioz, 2019).

A dynamic binaural system will typically utilise a large dataset of HRTF measurements taken from different positions around the subject, distributed in a spherical arrangement, in order to allow the rendering of a sound from whichever direction is desired. A dense distribution is required for a perceptually seamless transition between measurements, with a necessary resolution as fine as 2° in the azimuth and 2° in the elevation plane reported (Lindau, Maempel and Weinzierl, 2008). However, in some cases horizontal localisation resolution is reported to be as low as 1° (Blauert, 1997). Such fine resolution HRTF datasets are time consuming and impractical to measure for human subjects.

[7] `https://optitrack.com/`

[8] `https://www.bose.co.uk/en_gb/products/headphones/noise_cancelling_headphones/noise-cancelling-headphones-700.html`

[9] `https://www.oculus.com/rift/`

[10] `https://www.vive.com/uk/`

To obtain an HRTF at a position on the sphere where a measured HRTF is unavailable, interpolation can be used. One method of interpolation is to average between the nearest measurements, either in the time domain or the frequency domain (Hartung, Braasch and Sterbing, 1999). However, interpolating in this way between HRTFs at significantly different angles can produce blurred localisation (Duraiswami, Zotkin and Gumerov, 2004). Additionally, multiple sources of different distance, source width or incorporating movement requires complex interpolation and can therefore become problematic (Noisternig et al., 2003b).

Another approach for interpolating between HRTFs is to use the HRTFs as virtual loudspeakers, and to obtain any desired direction using a loudspeaker based spatial audio rendering method such as Vector Base Amplitude Panning (Pulkki, 1997), Wavefield Synthesis (Berkhout, Vries and Vogel, 1993) or Ambisonics (Gerzon, 1973). This will be discussed in greater detail in the following chapters.

## 2.7 Perceptual Audio Evaluation

The main goal of a binaural audio system is to produce an experience that is indistinguishable from reality. To evaluate how realistic the rendered spatial audio is requires comparison to a reference, whether that be a corresponding sound in reality or an inner expectation based on memory. Though it is possible to obtain some insight on the quality of a binaural audio system through numerical analytical methods by comparing measured data from two or more systems, results are ultimately estimations of human perception, and rigorous audio evaluation should always include some measurement of human perception. However, perceptual evaluations require statistical analysis in order to draw any conclusion on whether the results of one group of individuals are likely to apply to a larger population.

To perceptually evaluate a binaural audio system, listening tests with human participants are undertaken. In the design of a listening test, all aspects should be considered from the test methodology and choice of stimuli to the statistical analysis of the results. The methods of a listening test should be thoroughly documented such that the test can be repeated. This section covers the justifications and methodologies for the perceptual evaluation carried out in this thesis.

When perceptually evaluating binaural audio systems, the metric of evaluation is the first thing to be decided. *Basic audio quality* is defined in ITU-R BS.1116-3 as, 'a single, global attribute used to judge any and all detected differences between the reference and the object' (International Telecommunication Union, 2015a), however this does not account for the scale of perceived differences and is therefore appropriate only for specific testing scenarios. Another all-encompassing metric that lends itself more to scaled judgements is *realism*, in which the two main methods are *authenticity* and *plausibility*.

Measuring the authenticity of a spatial audio system requires presentation with a comparable real auditory event - 'if the two soundfields (simulation and real) cannot be distinguished, the simulation can be deemed perceptually authentic' (Blauert, 1997; Raake and Blauert, 2013; Brinkmann, Lindau and Weinzierl, 2017). However, authenticity is not always an appropriate question to ask when assessing quality of experience as it is often unachievable in a real listening test. For example, a test involving live human speech as the reference stimuli would not be repeatable, unless by use of a speech simulator (McKenzie, Murphy and Kearney, 2017). Therefore, the second realism metric, plausibility, has been considered as a more suitable aspect to measure in assessing the quality of experience in some previously published listening tests (Lindau and Brinkmann, 2012; Pike, Melchior and Tew, 2014). Plausibility is a variation on the concept of naturalness (Nicol et al., 2014; Raake and Blauert, 2013) defined as, 'a simulation in agreement with the listener's expectation towards a corresponding real event' (Lindau and Weinzierl, 2012). Therefore the reference in a question of plausibility uses the listener's memory. However, this means that an auditory experience could therefore be plausible without necessarily being authentic as well (Raake and Blauert, 2013). Furthermore, preference is a factor in what is considered natural, which undoubtedly differs between individuals (Brinkmann et al., 2014a).

An auditory assessment using a global attribute such as authenticity or plausibility requires evaluation of all aspects of the spatial audio system at the same time. This can be desirable in some cases, but problematic in a situation where the listener may judge contradictory changes between the stimuli and reference. For example, perceiving an increase in localisation accuracy but a decrease in timbral fidelity. For this reason, ways of evaluating spatial audio systems for single specific parameters separately have also been investigated previously (Nicol et al., 2014). Existing methods for evaluating separate perceptual features have mainly focused on *timbre* and *localisation* (Sontacchi et al., 2002; Le Bagousse et al., 2011), though some tests have been conducted that included *lateralisation* (Lewald and Ehrenstein, 1998; Lewald and Karnath, 2001), *externalisation* (Hartmann and Wittenberg, 1996; Moore, 2009; Catic et al., 2013; Armstrong et al., 2018a) and *distance perception* (Kearney et al., 2015).

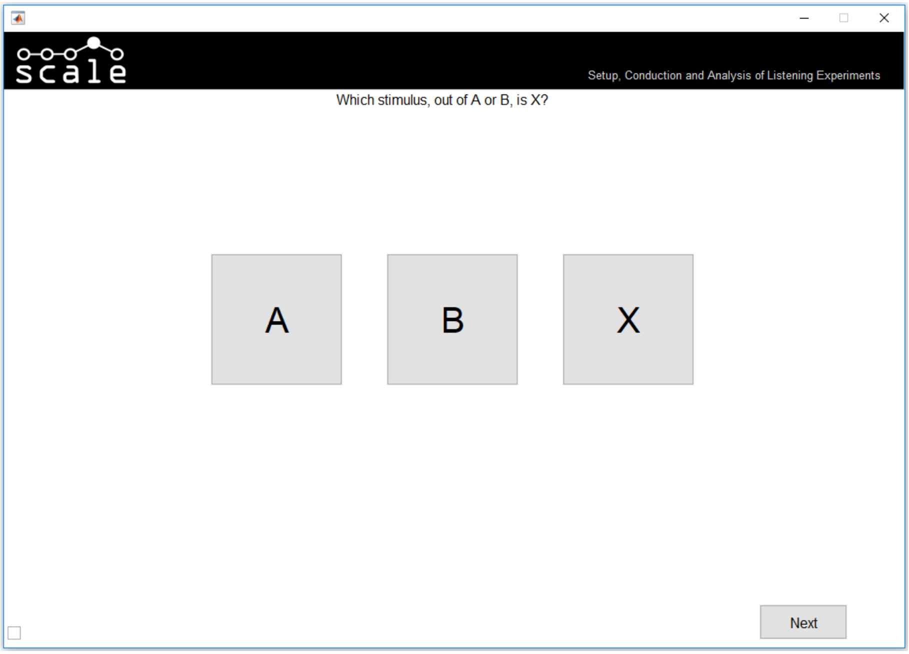
Timbre is defined in ANSI S1.1-1994 as, 'that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar' (American National Standards Institute, 1994). Localisation is defined as, 'the mathematical function relating the points of the physical (sound-source) space and those of the auditory space' (Blauert, 1997). Accurate timbre has been consistently rated as significantly more important than accurate localisation (Bregman, 1990; Rumsey et al., 2005a; Schärer and Lindau, 2009; Schoeffler and Herre, 2016). However, the quality of HRTFs has previously often only been measured in terms of localisation accuracy (Nicol et al., 2014), something that has been addressed more recently through HRTF evaluation tests including terms such as *brightness*, *richness* and *preference* (Armstrong et al., 2018a). Timbre and localisation cues are not entirely separate though, as spectral features can be partly interpreted as localisation cues in some circumstances (Nicol et al., 2014). As the aim of this thesis is to produce a plausible spatial audio experience, timbre is defined as the main concern for perceptual evaluation. Research on spatial audio specific perceptual evaluation has been conducted by Lindau *et al.*, which yielded the Spatial Audio Quality Inventory (SAQI) (Lindau et al., 2014): a collection of descriptive terms for spatial audio evaluation. This includes global attributes such as *difference* to more detailed timbral and localisation specific attributes, among others.

### 2.7.1 Listening Test Methodologies

There are many previously developed listening test methodologies for binaural audio evaluation. Different approaches can be more or less appropriate depending on the auditory conditions under test, so the choice of listening test methodology has been carefully considered in this thesis. Other factors that can have an effect on results include the experience, demographic and number of participants, as well as the environment in which the test is conducted.

Alternative forced choice (AFC) testing such as ABX (Munson and Gardner, 1950) (as detailed in ITU-R BS.1116-3 (International Telecommunication Union, 2015a)) is recommended when the test conditions under comparison have small differences. These systems typically evaluate two stimuli at a time. The ABX approach involves listening to three stimuli in a row, and choosing which, out of the first or second, is the third. This determines whether the stimuli under test can be differentiated from the reference. A typical ABX interface is illustrated in Figure 2.25 (Giner, 2013). Variations on the ABX test include the *oddball* paradigm (Langendijk and Bronkhorst, 2000; Moore, Tew and Nicol, 2010), whereby the two stimuli A and B are played back consecutively in four possible orders: AABA, ABAA, BABB or BBAB, and the participant must identify whether the odd one out occurred second or third.

Other test methodologies such as the multiple stimulus test with hidden reference and anchor (MUSHRA) paradigm, as found in ITU-R BS.1534-3 (International Telecommunication Union, 2015c), are recommended when there are medium to large differences between test conditions (Soulodre and Lavoie, 1999). A typical MUSHRA interface will compare a reference signal to several test signals simultaneously, rating each test signal on a scale of 1 - 100 in terms of overall perceived similarity to the reference signal. Included in the test signals are the reference signal again and



Which stimulus, out of A or B, is X?

FIGURE 2.25: Illustration of an ABX interface from the MATLAB based 'Scale' listening test toolbox, reproduced from Giner (2013).

one or two anchor signals. The mid anchor is usually the reference signal low-pass filtered with a cut off frequency $f_c = 7$ kHz and the low anchor is the reference low-pass filtered at $f_c = 3.5$ kHz (International Telecommunication Union, 2015c). An example of a MUSHRA interface is presented in Figure 2.26 (Schoeffler et al., 2018).

When conducting listening tests, care must be taken to ensure the participants are clear on the question they are being asked prior to the test. A training exercise should always be included to familiarise the participants with the test system. Tests are conducted double blind, meaning both the assessor and the participant did not know which stimuli is which, to avoid possible influence from the assessor. Additionally, the order of test conditions is randomised to avoid bias from inevitable factors such as learning as the test progresses. The environment in which the tests are conducted is chosen as a quiet listening room (background noise level of 41.8 dBA) with minimal distractions. In certain tests it may be necessary to perform the tests in an anechoic chamber, but for headphone based tests, a quiet listening test room is deemed sufficient in most cases (Brüggen, 2001; Faller and Baumgarte, 2003; Par et al., 2005; Katz and Parseihian, 2012; Brinkmann et al., 2014b; Ahrens and Andersson, 2019).

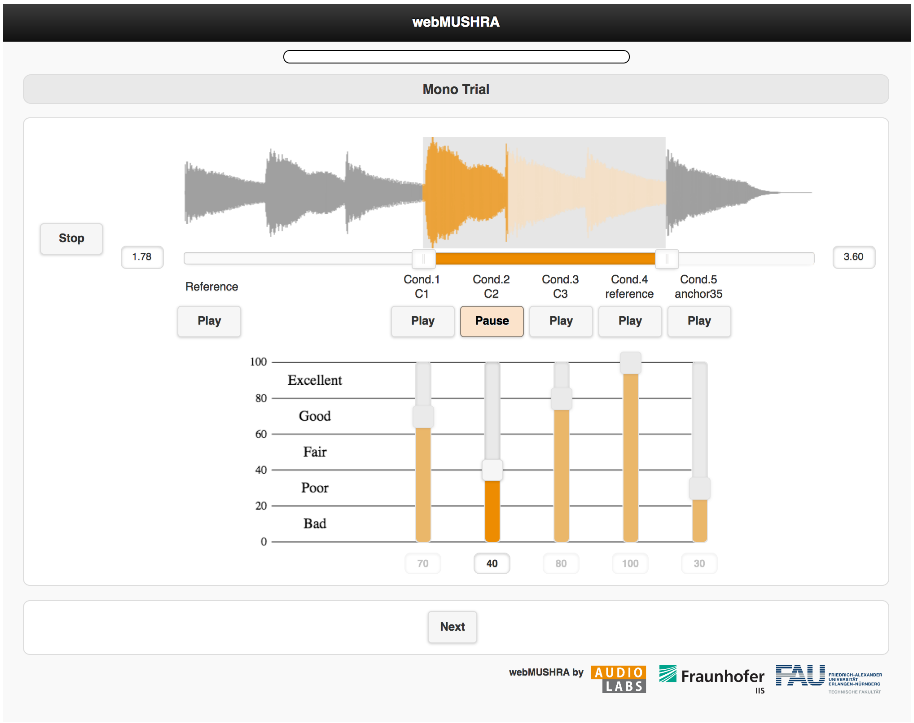


FIGURE 2.26: Illustration of a MUSHRA interface from the web based 'webMUSHRA' interface, reproduced from Schoeffler et al. (2018).

In evaluations where audio is accompanied by visuals, video quality has been found to have a notable effect on perceived quality of audio (Beerends and De Caluwe, 1999). Therefore, even in audio only tests, the accompanying visual components are still considered.

The type of listener is another consideration to make. If the binaural system under test is envisaged to be used by the wider population, the participants chosen for an evaluative listening test should ideally aim to cover a diverse range of genders, ages, races and listening ability such that results will better convey the likely opinions of a wider audience. However in audio research, inexperienced listeners can make more errors in localisation (Asano, Suzuki and Sone, 1990) and give their judgements more based on preference (Rumsey et al., 2005b). Therefore, it is often desirable to use experienced listeners in testing to ensure the questions of the listening test will be understood and answered correctly, and to give an accurate critical analysis of the test material (International Telecommunication Union, 2015c; Olive, 2003). Experienced listeners are defined in ITU-R BS.1534-3 as having, 'experience in listening to sound in a critical way' (International Telecommunication Union, 2015c). Therefore education or employment in music technology and audio engineering related fields is deemed sufficient in this thesis.

If a listening test is conducted with a small number of participants or trials, there is a greater probability of the results finding that no audible difference between stimuli exists when in fact there is. This is referred to as a Type 2 (beta) Error (Leventhal, 1986). Conversely, if a large number of participants and trials are used, this can increase the probability of finding an audible difference when in fact there is none. This is a Type 1 (alpha) Error. The probability of Type 1 and Type 2 Errors can be minimised by choosing an appropriate number of participants and trial repeats. In this thesis, participants chosen for a perceptual evaluation test do not include the primary investigator except for preliminary tests, in accordance with Blauert (Blauert, 1997). Around 20 participants is seen as sufficient for ABX and MUSHRA style tests in the ITU recommendation (International Telecommunication Union, 2015a; International Telecommunication Union, 2015c), and it is common to repeat each test condition at least once. Participants should be screened for hearing damage prior to starting the test, whether that be self reported or via an audiometry test. ISO 389 is one guideline to follow (International Organization for Standardization, 2016).

### 2.7.2 Test Stimuli

Once the chosen methodology has been established, it is necessary to determine appropriate test stimuli. First, the type of sound scene is decided. In this thesis, simple acoustic scenes are used to refer to a sound scene with a single source playing at any given moment, at one location on the sphere and with one type of sound, and a constant source width and distance. These can be appropriate for assessing specific locations under highly controlled conditions. Complex acoustic scenes are used to refer to sound scenes with multiple sounds occurring simultaneously, which can be of different stimuli type, source width, distance, position and direction. Complex acoustic scenes are beneficial for gauging a spatial audio system's character over multiple or all directions, and for simulating situations closer to real-life listening.

The type of stimuli used is another important consideration. A common choice is to use a broadband stimulus, which has energy at all frequencies, and is therefore appropriate for allowing the participant to effectively assess timbral and spectral cues (Hartmann et al., 2010). The most widely used examples of broadband stimuli include white Gaussian noise (equal amplitude at all frequencies) and pink noise (white Gaussian noise with a 6 dB attenuation per increase in octave). Other noise variants include grey noise (white noise with equal loudness weighting) and thermal noise (the noise generated by the thermal agitation of electrons in an electronic conductor) (Johnson, 1928; Thurlow, Mangels and Runge, 1967; Hartmann and Rakerd, 1993). It is sometimes desirable to assess transient responses with broadband stimuli, in which case pulsed noise bursts (Volk, Musialik and Fastl, 2009) or click trains (Moore, Tew and Nicol, 2010; Goupell, Majdak and Laback, 2010) can be used. An alternative real-world sound used in listening tests for assessing transient responses is percussion (Lindau, Hohn and Weinzierl, 2007). Another type of stimulus used in listening tests is human speech (Begault and Wenzel, 1993; Best et al., 2005; Brinkmann et al., 2014a; Catic, Santurette and Dau, 2015), as it is highly familiar to the human auditory system - the sensitivity of our ears is even tuned to the frequency response of speech (Blauert, 1997), as shown in equal loudness curves (Bauer and Torick, 1966).

The duration and level of stimuli also require careful consideration. Short bursts of noise are harder to localise than longer duration stimuli. However, when a system does not dynamically update the sound to the head orientation, longer duration stimuli can cause issues with front-back confusion (Hartmann et al., 2010). Stimuli should be played back at a consistent specified volume, and a general listening level is usually chosen between 55 - 75 dB SPL (Katz and Parseihian, 2012; Hartmann and Rakerd, 1993; Hammershøi and Møller, 1996), which corresponds approximately to the range of conversational human speech (Byrne et al., 1994). It is worth noting that louder test stimuli have been found to reduce perception of elevation in sound sources

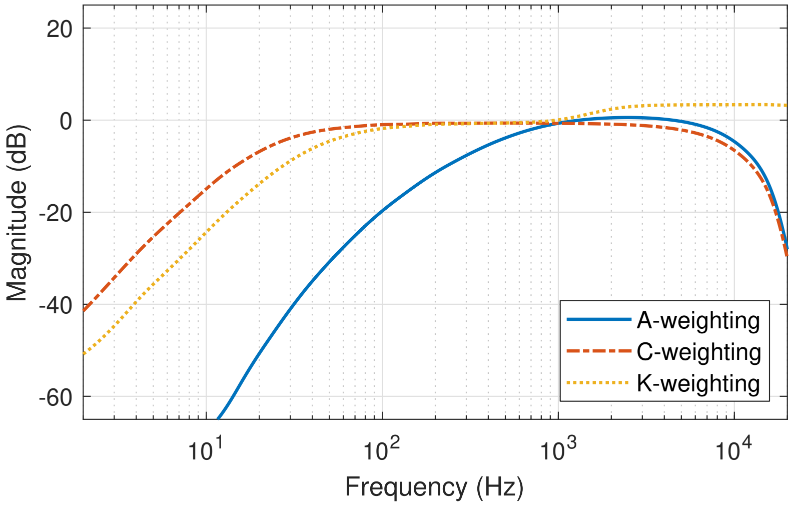


FIGURE 2.27: Frequency response curves of different SPL weightings.

(Hartmann and Rakerd, 1993). Furthermore, an onset and offset window on the test stimuli is always implemented to avoid unwanted clicks at the start and end of the stimuli (Schonstein, Ferré and Katz, 2008). If the stimuli have different spectral characteristics, the normalisation method used can affect the perceived loudness of the stimuli. In loudness measurement, various filtering options are available which can be matched to the type of stimuli. The frequency responses of three common SPL weighting filters are shown in Figure 2.27. A- and C-weighting filters are defined in ANSI S1.42-2001 (American National Standards Institute, 2001), where C-weighting is usually used for measuring louder levels and A-weighting is more often used. The K-weighting filter is defined in ITU-R BS.1770-4 (International Telecommunication Union, 2015d), and is used primarily for loudness measurement in the broadcast industry.

### 2.7.3 Statistical Analysis

The data collected from a listening test must be analysed accordingly. The first step in analysis of listening test results is to look for internal consistency, and to exclude the data of certain participants from analysis if they have been inconsistent. For example, if the same test conditions are repeated, analysis of the difference in answers between the two instances can be used to check the internal consistency of participants. For MUSHRA tests, the data of an individual participant would be excluded if the hidden reference was rated below 90 out of 100 for more than 15% of answers, or if the mid anchor was rated higher than 90 out of 100 for more than 15% of answers (International Telecommunication Union, 2015c).

To determine which type of analysis is appropriate, data is first checked for normality using a test such as the Kolmogorov-Smirnov test, or Levene's test for equal variance (Bech and Zacharov, 2007). If data is found to follow a normal distribution, then parametric statistical analysis is appropriate; otherwise non-parametric analysis should be used. Parametric statistical analysis methods include analysis of variance (ANOVA) and T-tests, and non-parametric variants of the ANOVA include the Friedman's ANOVA test. ANOVA tests assess whether two or more conditions have significantly different averages, where parametric analysis usually assesses the mean average and non-parametric analysis assesses the median average.

Typically, the confidence interval used in statistical analysis above which statistical significance can be claimed is 95%, though this value can be 99% for reporting high statistical significance. Significance is most often reported using $p$, where statistical significance at a confidence level of 95% can be claimed when $p < 0.05$. In AFC testing, a result approaching 50% for each stimuli can follow the assumption that the participants were not able to discern any differences and were therefore guessing. The exact value for results to be not statistically significantly different from chance (i.e. perceptually indistinguishable) is calculated from the cumulative binomial distribution, which is different depending on the number of trials (McKenzie, Murphy and Kearney, 2017).

## 2.8 Numerical Audio Evaluation

A convenient preliminary approach to audio evaluation, in order to avoid running perceptual listening tests which have limitations due to the time they take to perform and the physical setup needed, is to use numerical evaluation methods. This can be achieved by comparing test HRTFs to reference HRTFs (Wiggins, Paterson-Stephens and Schillebeeckx, 2001). Benefits of numerical evaluation include the ability to test a greater number of data points using several methods, as well as evaluating more than one binaural audio feature at a time, such as interaural cues and spectral difference. By using numerical estimation methods, it is possible to compare two or more datasets such that, if one is a reference dataset of ideal measurements, then comparing the difference between a test and reference datasets gives reasonable grounds for similarity - smaller differences can be seen as an improvement in accuracy. This section covers the methodologies for numerical evaluation used in this thesis.

### 2.8.1 Reference Datasets and Equalisation

The Bernschütz Neumann KU 100 HRTF database (Bernschütz, 2013) is used throughout this thesis as a reference dataset. Measurements are anechoic, taken at a distance of 3.25 m, which is sufficient distance to assume far-field conditions (Brungart and Rabinowitz, 1999), using a single fixed position Genelec 8260A loudspeaker[11]. One available configuration offers 16,020 measurements in a Gauss-Legendre arrangement, with 2° resolution in both azimuth and elevation (89 values of elevation for 180 values of azimuth, as illustrated in Figure 2.28).

The arrangement of points on a sphere presented in Figure 2.28 results in a clustering of measurements at the poles. This is an inherent characteristic of Gauss-Legendre quadrature. Figure 2.29 plots the vertices of an 8° Gauss-Legendre quadrature with 23 elevations at 45 different azimuth values in 8° increments totalling 1035 points (a lower resolution than the 2° Gauss-Legendre quadrature used in the rest of this thesis to aid visibility). Shading in the figure is based on the *solid angle*, denoted in this thesis using $\Omega$, of each point. The solid angle refers to the proportional amount of the area of the sphere in which a single point subtends (Oosterom and Strackee, 1983), such that

[11] `https://www.genelec.com/studio-monitors/sam-coaxial-studio-monitors/8260a-sam-studio-monitor`

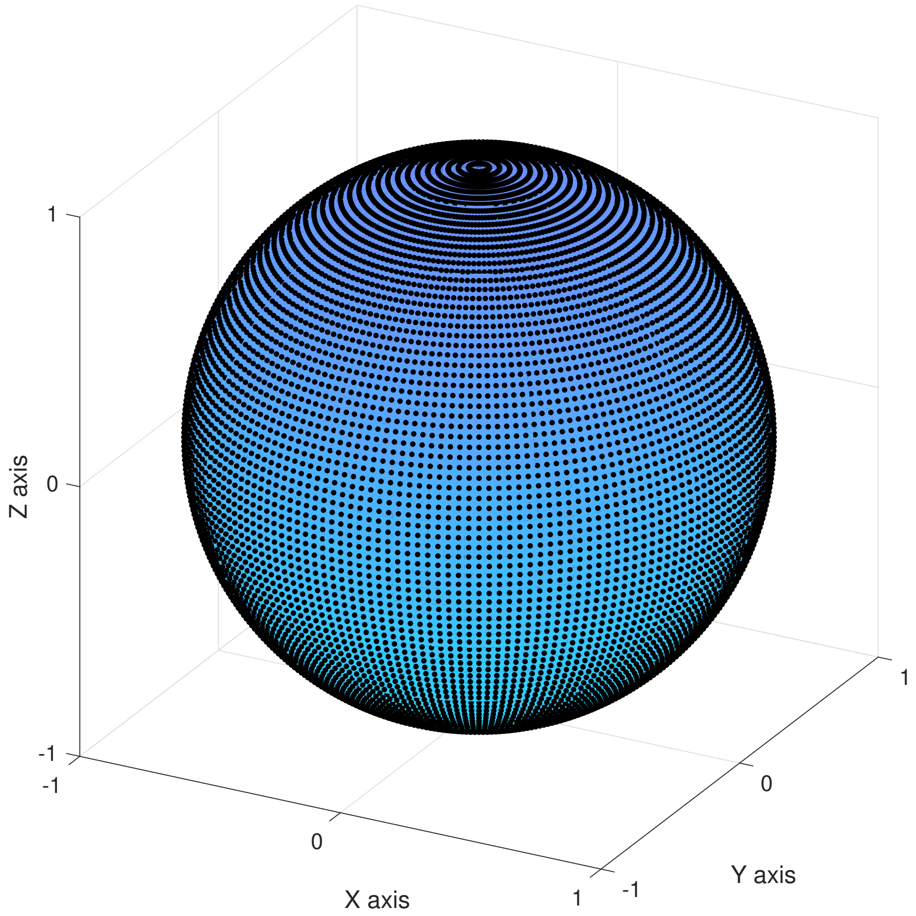


FIGURE 2.28: Distribution of points in the 16,020 point 2° Gauss-Legendre quadrature, as featured in the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

$$\sum_{q=1}^{Q} \Omega_q = 1 \tag{2.9}$$

where $q$ refers to the measurement number, of which $Q$ is the total number of measurements. The clustering of points at the poles in Figure 2.29 produces large variations in solid angle. For average calculations over the sphere, therefore, a mean average would produce a bias towards the poles. Solid angle weighting can be used to address this, by multiplying the value at each point on the sphere by its solid angle weight before taking the sum of all values.

The dense distribution of measurements in the 16,020 point 2° Gaussian configuration quadrature dataset from the Bernschütz Neumann KU 100 database (Bernschütz,

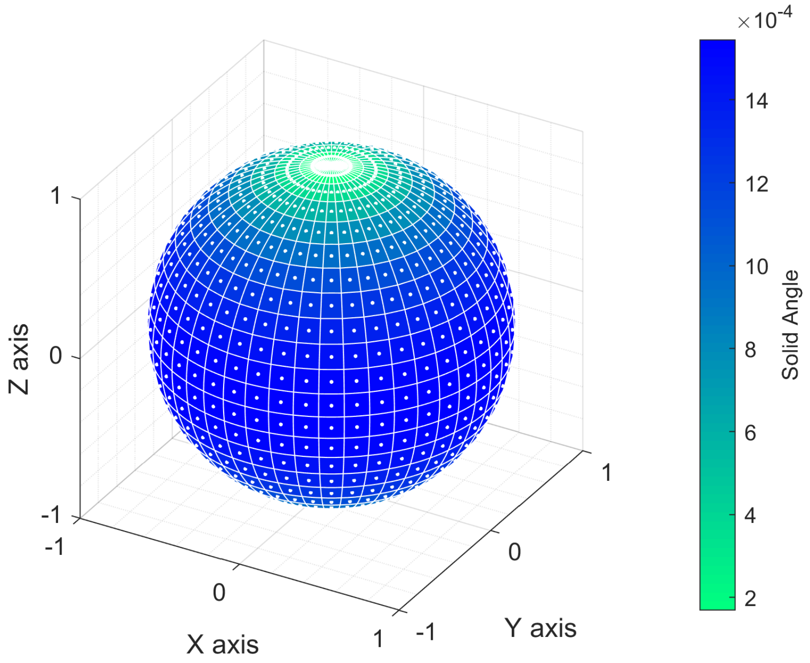


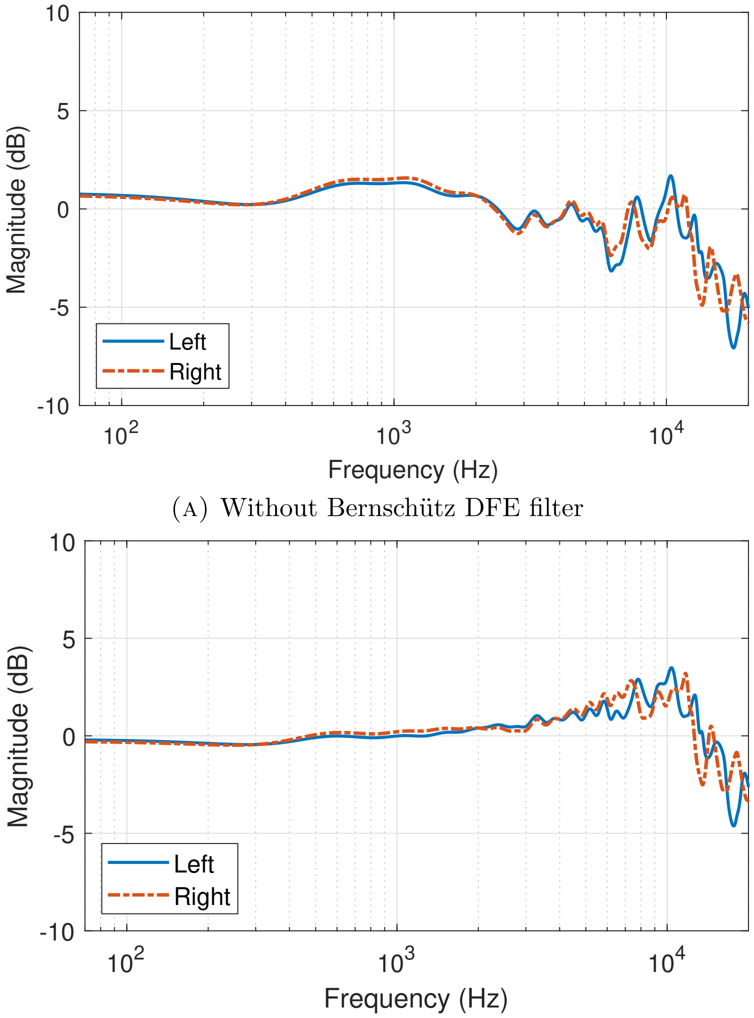
Figure 2.29: Voronoi sphere plot of an 8° Gauss-Legendre configuration.

2013) makes it highly suitable for numerical analysis of HRTFs at all positions on the sphere. However, the Bernschütz datasets do not feature a flat diffuse-field response, even with the Bernschütz diffuse-field equalisation (DFE) filters. Therefore prior to use in numerical evaluations in this thesis, the dataset was diffuse-field equalised. The diffuse-field response of the dataset was obtained from the root-mean-square (RMS) average of all the HRTFs. The diffuse-field HRTF $H_{\text{diff}}$ was calculated separately for both left and right ears as

$$H_{\text{diff}} = \sqrt{\sum_{q=1}^{Q} \Omega_q H_q(\theta, \phi)^2} \tag{2.10}$$

where $H(\theta, \phi)$ is a single HRTF, and $\Omega$ is the solid angle weight of the measurement $q$. The calculation was performed in the frequency domain. The left and right calculated diffuse-field responses of the 16,020 point dataset are presented in Figure 2.30, both without (2.30a) and with (2.30b) the Bernschütz DFE filters.

Linear-phase inverse filters were then calculated from the diffuse-field responses using least-mean-square regularisation (Kirkeby and Nelson, 1999).



(A) Without Bernschütz DFE filter

(B) With Bernschütz DFE filter

FIGURE 2.30: Diffuse-field responses of the 16,020 point dataset in the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

The process of inverse filtering using this method is as follows. First, the absolute values of an FFT of the signal to be inverted are obtained. The frequency domain signal is then smoothed using the fractional octave complex smoothing approach of Hatziantoniou and Mourjopoulos (2000), such that the FFT frequency sampling range is divided up into a number of frequency regions determined by the value of fractional octave smoothing (whereby a smaller fraction results in a larger number of frequency bins), the mean values of the FFT bins of each frequency region are calculated, and the signal is smoothed through one-dimensional interpolation of the FFT signal and the mean frequency region values using a spline interpolation method. The smoothed signal is then inverted by

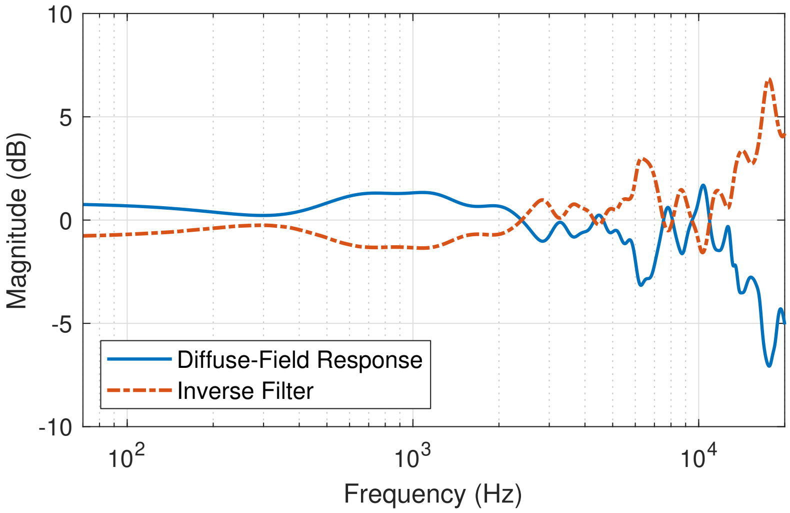


FIGURE 2.31: Diffuse-field response and equalisation inverse filters of the 16,020 point dataset in the Bernschütz Neumann KU 100 database (Bernschütz, 2013) (left ear).

$$H_{\text{inv}} = \frac{H^*}{(H^*H) + (B^*B)} \tag{2.11}$$

where superscript * denotes the complex conjugate, $H$ denotes the signal, and $B$ denotes the regularisation octave edges, which are calculated for each frequency bin of the FFT (of which the total number is determined by the inverse filter length), first converted from dB to linear values by

$$B_{\text{lin}} = 10^{\frac{-B}{20}} \tag{2.12}$$

before the minimum phase component is calculated in the time-domain, using a symmetric inverse FFT, before being windowed and returned to the frequency domain using an FFT.

For the Bernschütz Neumann KU 100 database equalisation, 1/4 octave smoothing was implemented, and the range of inversion was 2 Hz - 16.5 kHz, with in-band and out-band regularisation of 25 dB and 11 dB, respectively. The inverse filter for the left ear signal of the original dataset (i.e. without the Bernschütz DFE) is presented in Figure 2.31. The HRTF dataset was diffuse-field equalised through convolution of the original HRTFs with the calculated inverse filters.

In order to be rigorous, numerical evaluation should be performed using more than one dataset. The SADIE II HRTF database (Armstrong et al., 2018a) features datasets of two dummy heads and 18 human subjects, at dense distributions ranging from 2,114 - 8802 measurements. This is used as a second database for numerical tests in this thesis.

### 2.8.2 Interaural Level Difference Estimation

The ILD of an HRTF is estimated in this thesis as in Watanabe et al. (2007), whereby the HRTF is passed through a linear-phase high-pass filter of order 128 at a cut off frequency $f_c = 1.2$ kHz and a $-60$ dB stop band frequency $f_{\text{stop}} = 500$ Hz, followed by an FFT with a number of frequency bins of double the number of samples in the impulse response. The frequency bands of the FFT calculation are amplitude weighted using 30 equivalent rectangular bandwidth (ERB) frequency bands (Moore and Glasberg, 1983) between 20 Hz - 20 kHz (equating to approximately 1/3 octave intervals). The bandwidths $\text{BW}_{\text{ERB}}$ are calculated for each frequency in the frequency vector of the FFT calculation $f_c$ as

$$\text{BW}_{\text{ERB}} = 24.7(0.00437 f_c + 1) \tag{2.13}$$

and the amplitude value of each frequency bin is then weighted by $\text{BW}_{\text{ERB}}^{-1}$, such that the relative weight of high frequencies is reduced (Moore and Glasberg, 1983).

A single value of ILD in dB for an HRTF measured at an angle of incidence $(\theta, \phi)$ is then estimated as the mean magnitude difference between the absolute values of the left and right frequency bins such that

$$\text{ILD} = 20 \log_{10} \frac{|H_{\text{left}}(\theta, \phi)|}{|H_{\text{right}}(\theta, \phi)|} \tag{2.14}$$

where $H_{\text{left}}$ and $H_{\text{right}}$ are the left and right signals of the HRTF, respectively.

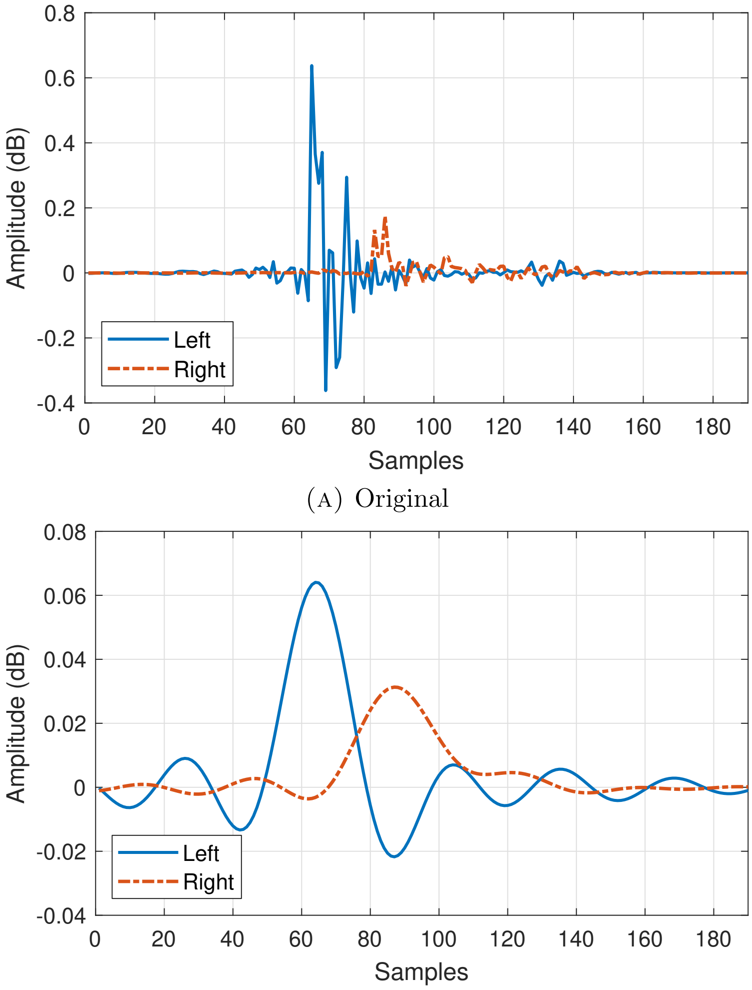
### 2.8.3 Interaural Time Difference Estimation

There are numerous different methods for estimating the ITD of an HRTF (Daniel, 2000, p. 59). The first is a simple calculation of the difference in time between detected signal onsets, which are determined when the amplitude of the left and right signals surpass a specified threshold level, as used in Kuhn (1977) and Algazi, Avendano and Duda (2001b). The second method calculates the time difference from the maximum value of cross-correlation between the left and right signals, as used in Kistler and Wightman (1992), MacCabe and Furlong (1994), Middlebrooks (1999), Macpherson and Middlebrooks (2002), and Langendijk and Bronkhorst (2002). Though differences exist between the two methods which may be perceivable (Katz and Noisternig, 2014), both are considered reasonable approximations (Andreopoulou and Katz, 2017).

The method of estimating ITD in this thesis uses the maximum value of the interaural cross-correlation as in Katz and Noisternig (2014), as humans are said to perceive ITD based on phase irregularities between the signals at the left and right ears rather than differences in onset time (Smith and Price, 2014b). Prior to the interaural cross-correlation (IACC) calculation, the HRIR is filtered by a low-pass linear phase filter at 1.5 kHz, due to ITD being perceptually irrelevant at high frequencies (Kuhn, 1977), with a filter order of 512 taps. Figure 2.32 illustrates a $(\theta = 90°, \phi = 0°)$ HRIR with and without 1.5 kHz low-pass filtering. The time delay is then estimated from the maximum value of IACC between the left and right signals of an HRIR measured at an angle of incidence $(\theta, \phi)$ such that

$$\text{ITD} = \text{arg max IACC}(\theta, \phi, \tau) \tag{2.15}$$

where IACC is calculated as



(A) Original

(B) Low-pass filtered

Figure 2.32: Original and 1.5 kHz low-pass filtered HRIRs recorded at ($\theta = 90°, \phi = 0°$) from the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

$$\text{IACC}(\theta, \phi, \tau) = \frac{\int_{t_1}^{t_2} H_{\text{left}}(\theta, \phi, t) H_{\text{right}}(\theta, \phi, t+\tau) dt}{\sqrt{\int_{t_1}^{t_2} H_{\text{left}}^2(\theta, \phi, t) dt \int_{t_1}^{t_2} H_{\text{right}}^2(\theta, \phi, t) dt}} \tag{2.16}$$

where the integration limits are defined as $t_1 = 0$ and $t_2 =$ the number of samples of the HRIR, and $\tau$ denotes the time delay in seconds (Katz and Noisternig, 2014).

### 2.8.4 Localisation Estimation

There are ways of estimating human sound localisation in order to avoid running localisation listening tests, for instance, the use of auditory models. Two commonly

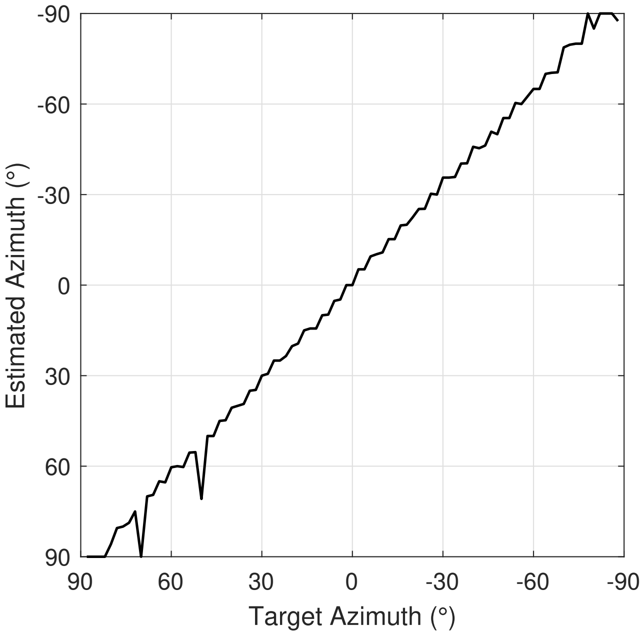


FIGURE 2.33: Estimated horizontal localisation (May, Van De Par and Kohlrausch, 2011) of the Bernschütz Neumann KU 100 database (Bernschütz, 2013). $\overline{E_\theta} = 3.52°$

used binaural localisation models are the horizontal model by May, Van De Par and Kohlrausch (2011) and the vertical model by Baumgartner, Majdak and Laback (2014). The May horizontal model is a probabilistic binaural localisation model which uses ITD and ILD estimations in conjunction with azimuth-dependent Gaussian mixture models to estimate the horizontal location of the input signal. Therefore, a reference signal is not required for comparison.

The estimated horizontal localisation of the Bernschütz Neumann KU 100 database (Bernschütz, 2013) at azimuth values at 2° intervals between $-90° < \theta < +90°$ at $\phi = 0°$ is presented in Figure 2.33. The model predicts the horizontal localisation highly accurate with $\theta_{\text{est}}$ within ±2° of the target azimuth for most positions, however some positions are inaccurate, such as $\theta = 50°$ and $\theta = 70°$. This shows the limitations of this model, which was not originally evaluated at azimuths greater than $\theta = 50°$ (May, Van De Par and Kohlrausch, 2011).

A single value of overall mean azimuth error $\overline{E_\theta}$ can be calculated as the mean absolute difference in degrees between the predicted azimuth $\theta_{\text{est}}$ and the target azimuth $\theta$ for all tested locations $Q$ as

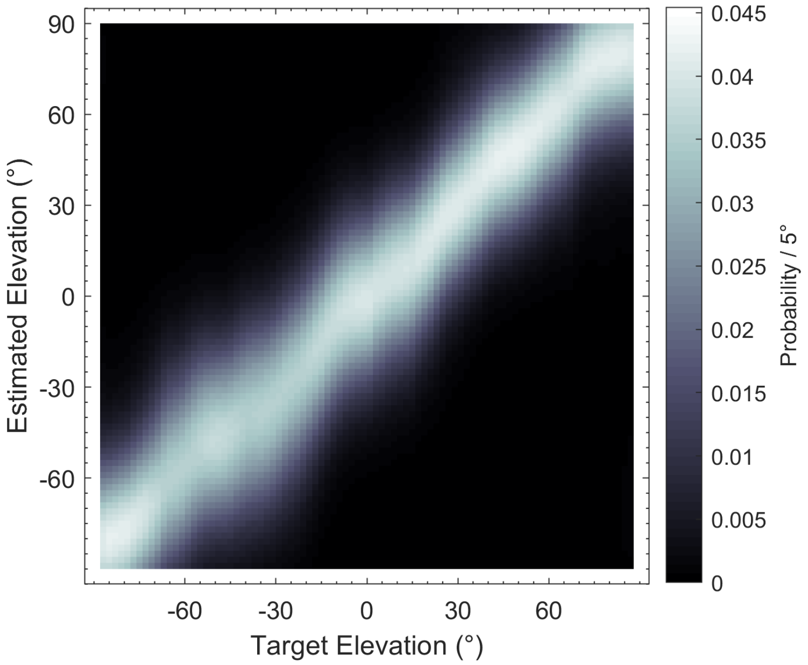


FIGURE 2.34: Estimated vertical localisation (Baumgartner, Majdak and Laback, 2014) of the Bernschütz Neumann KU 100 database (Bernschütz, 2013). QE = 0.6%, PE = 21.5°.

$$\overline{E_\theta} = \sum_{q=1}^{Q} \frac{|\theta_{\text{est}}(H_q) - \theta_q|}{Q} \tag{2.17}$$

where $H_q$ denotes a specific measurement at location $q$ in the dataset. For the Bernschütz dataset, $\overline{E_\theta} = 3.52°$.

For estimating elevation localisation, the Baumgartner sagittal plane localisation model uses a probabilistic functional model which removes direction-independent aspects, employs equivalent rectangular bandwidth filtering to approximate the cochlea's effect, and compares the extracted spectral gradients to those of a provided reference signal. It produces two psychoacoustic performance metrics: quadrant error (QE), a prediction of localisation confusion (presented as a percentage), and local polar RMS error (PE), a prediction of precision and accuracy in degrees.

The estimated vertical localisation of the Bernschütz Neumann KU 100 database (Bernschütz, 2013) at elevation values at 2° intervals between $\phi = -88°$ and $\phi = +88°$ at $\theta = 0°$ is presented in Figure 2.34. As this example uses the same reference dataset as the test set, the gradient is an even diagonal, as expected. For the Bernschütz dataset, QE = 0.6% and PE = 21.5°.

### 2.8.5 Timbre

For evaluating the timbre of a binaural system, methods such as spectral differences between magnitude responses are often used (Wiggins, Paterson-Stephens and Schillebeeckx, 2001; Moore, Tew and Nicol, 2010). However, a basic spectral difference calculation from the magnitude responses of two audio signals obtained using an FFT will not necessarily accurately represent the perceptual differences between the signals. The human auditory system's complex response to relative amplitude and temporal differences, many of which are exploited in audio compression techniques such as MPEG-1 Audio Layer III (also known as MP3) (International Organization for Standardization, 1993), require consideration in the spectral difference calculation. Methods such as the Composite Loudness Level (CLL) use ERB weightings and a Phon calculation (Pulkki et al., 1999; Ono, Pulkki and Karjalainen, 2001; Ono, Pulkki and Karjalainen, 2002).

## 2.9 Summary

This chapter has introduced the fundamental principles of sound, describing the mechanisms by which sound travels through and interacts with an environment to change the signals at the point of the eardrums. The functions of the human auditory system and psychoacoustics of sound localisation have been discussed: horizontal localisation is mainly achieved through interaural differences between the signals arriving at each eardrum and vertical localisation is mainly achieved through spectral features of the signals arriving at the eardrums as a result of interaction with the pinnae, head and torso.

Binaural technology has been introduced with a description of the HRTF and how it can be used to produce spatial audio over headphones. An emphasis has been placed on the importance of attention to detail in all aspects of binaural recording and reproduction for the production of a realistic auditory experience, from the impulse response measurement method when obtaining HRTFs to the choice and equalisation of headphones in reproduction. Accurate spectral reproduction, which has been shown to be more important than localisation accuracy, is identified as a focus for the work undertaken in this thesis. To aid in reproducing the auditory experience necessary for authentic reproduction, the binaural synthesis must be dynamic and react to rotations of the head. Some of the potential issues of dynamic binaural synthesis have also been introduced, such as the interpolation between HRTF measurements.

Finally, this chapter has discussed techniques for evaluating the quality of binaural audio. The motivation and listening test methodologies for perceptual audio evaluation have been introduced, and the appropriate methodologies have been identified, which depend on the stimuli: MUSHRA is more appropriate for medium to large differences between stimuli, and ABX is appropriate for small differences between stimuli. Numerical metrics, which offer a more convenient preliminary way of evaluating binaural spatial audio systems than running listening tests, have also been discussed. These include methods for interaural cue estimation, and a summary of two perceptually motivated models for azimuthal and elevation localisation prediction. These form the background for the binaural audio evaluation methods used throughout this thesis. The next chapter will introduce Ambisonics, a technology that allows binaural reproduction at any direction with as few as four HRTF convolutions.

# Chapter 3

# A Review of Ambisonics

Ambisonics is a spatial audio technology that first emerged in the 1970s. It was largely ignored by consumer culture until the resurgence of virtual reality (VR) technologies from the binaural reproduction of Ambisonic signals, such as in Google Resonance[1] (Gorzel et al., 2019). Ambisonics can be used to render binaural audio from all directions with as few as four convolutions, making it highly computationally efficient. Furthermore, Ambisonic soundfields can be easily rotated to account for head movements, which is crucial for VR applications. Recent developments have seen further improvements in the accuracy of reproduction, with low computational cost.

This chapter introduces Ambisonic technology and related workflows, from encoding and recording to decoding and binaural rendering, including state of the art methods for improving reproduction. It then discusses the limitations of binaural Ambisonic rendering, which will form the motivations for much of the work presented later in this thesis.

[1] `https://resonance-audio.github.io/resonance-audio/`

## 3.1 Fundamentals and First-Order Ambisonics

Ambisonics is a 3D spatial audio approach that allows the recording and encoding processes to be independent of reproduction. It was first introduced by Gerzon in the 1970s (Gerzon, 1973; Gerzon, 1977a; Gerzon, 1977b), though it should be noted that similar concepts involving separating the encode and decode processes of multichannel loudspeaker audio did arise around the same time (Cooper and Shiga, 1972). Digital implementations of the technology came from Malham and Myatt in the 1990s (Malham and Myatt, 1995; Farina, 1998). Ambisonics is based on spatial sampling and reconstruction of a soundfield using spherical harmonics (Lecomte et al., 2015), which began as an alternative to Quadraphonics (Bauer, Gravereaux and Gust, 1971), offering more stable panning between the front and rear pairs of loudspeakers (Fellgett, 1974; Furness, 1990). Though Ambisonics also exists in 2D, which uses cylindrical harmonics (Poletti, 2000; Benjamin, Lee and Heller, 2006; Solvang, 2008), the focus of this thesis is on 3D spatial audio, and therefore this chapter will concentrate on 3D Ambisonic audio.

Ambisonics has many advantages over other surround sound approaches: whereas for most surround sound systems each channel of the recording is the specific signal sent to an individual loudspeaker, such as Auro-3D (Theile and Wittek, 2011), the number and layout of loudspeakers for reproduction of Ambisonic format sound does not need to be considered in the encoding or recording process. By encoding into Ambisonic format, the soundfield is decomposed into orthogonal functions, whereby weighted combinations of the channels can produce a sound at any direction. Furthermore, the soundfield can be easily rotated, transformed and zoomed once in Ambisonic format (Gerzon and Barton, 1992; Wiggins, 2004; Pomberger and Zotter, 2011; Kronlachner and Zotter, 2014). A regular arrangement of loudspeakers in a sphere can produce an accurate representation (at low frequencies) of the original soundfield at the centre of the sphere, known as the *sweet spot* (Noisternig et al., 2003b).

The first form of Ambisonics was what is now known as first-order Ambisonics (FOA)

Table 3.1: Description and labelling of the tetrahedral capsules in the A-format Soundfield Ambisonic microphone (Gerzon, 1975; Craven and Gerzon, 1977).

| Capsule | Direction |
|---|---|
| *A* | Front left up |
| *B* | Front right down |
| *C* | Back left down |
| *D* | Back right up |

or B-format, which has 4 channels. B-format arranges these four channels as one with an omnidirectional polar pattern (W channel) and three with figure-of-eight polar patterns facing in the $x$, $y$ and $z$ Cartesian directions (X, Y and Z channels). A monophonic sound signal can be encoded into B-format FOA at a desired source direction of azimuth and elevation $\theta$ and $\phi$, respectively, through multiplication with the B-format gains for each channel, which are calculated as follows:

$$\begin{aligned}
\mathrm{W} &= \frac{1}{\sqrt{2}} \\
\mathrm{X} &= \cos\theta\cos\phi \\
\mathrm{Y} &= \sin\theta\cos\phi \\
\mathrm{Z} &= \sin\phi
\end{aligned} \tag{3.1}$$

The first microphone for recording Ambisonic signals was developed by Gerzon and Craven in 1975 (Gerzon, 1975; Craven and Gerzon, 1977) and launched commercially by Calrec Audio Ltd. in 1978, consisting of four cardioid capsules arranged in a tetrahedral array and capable of recording FOA. The arrangement of the capsules is described in Table 3.1.

The outputs of the tetrahedral capsules is known as A-format. The conversion from A-format to B-format (Craven and Gerzon, 1977) is achieved by

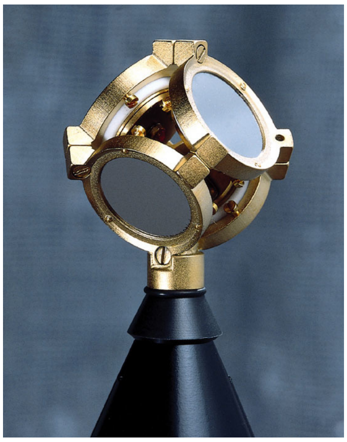


FIGURE 3.1: SoundField first-order Ambisonic microphone, reproduced from SoundField (2019).

$$
\begin{aligned}
\mathrm{W} &= A + B + C + D \\
\mathrm{X} &= A + B - C - D \\
\mathrm{Y} &= A - B + C - D \\
\mathrm{Z} &= A - B - C + D
\end{aligned}
\tag{3.2}
$$

The B-format signals then require equalisation and phase compensation (Gerzon, 1975). Figure 3.1 presents the tetrahedral microphone capsule layout.

Decoding B-Format Ambisonic audio requires the generation of a decoding matrix, which is dependent on the loudspeaker positions (Gerzon, 1992a; Jot, Larcher and Pernaux, 1999). A simple sampling decoder (Wiggins, 2004, p. 55) for an arbitrary number of loudspeakers gives the gain for each Ambisonic channel based on the loudspeaker's position, denoted by $\theta_l$ and $\phi_l$, as

$$
\begin{aligned}
g_{\mathrm{W}} &= \frac{1}{\sqrt{2}} \\
g_{\mathrm{X}} &= \cos\theta_l \cos\phi_l \\
g_{\mathrm{Y}} &= \sin\theta_l \cos\phi_l \\
g_{\mathrm{Z}} &= \sin\phi_l
\end{aligned}
\tag{3.3}
$$

and the resulting loudspeaker signal $s_l$ is given (Farina et al., 2001) as

$$s_l = \frac{(2-d)g_{\mathrm{W}}\mathrm{W} + d(g_{\mathrm{X}}\mathrm{X} + g_{\mathrm{Y}}\mathrm{Y} + g_{\mathrm{Z}}\mathrm{Z})}{2} \tag{3.4}$$

where $d$ is the directivity factor of the virtual microphone response, in the range $0 \leq d \leq 2$, such that $d = 0$ results in an omnidirectional virtual polar pattern, $d = 1$ results in a cardioid polar pattern, and $d = 2$ results in a figure-of-eight polar pattern (Wiggins, 2004). At low frequencies, this can reproduce the original soundfield accurately at the centre of the array. However, this simple decoding strategy fails when the loudspeakers are not situated in a regular (equally spaced) array, and the flexibility of directivity does not necessarily produce accurate reconstruction of the original soundfield (Gerzon and Barton, 1992).

## 3.2 Higher-Order Ambisonics

B-format and FOA, along with simple decoding strategies, offered a promising start to this spatial audio technology. First-order Ambisonics became well known for its ability to reproduce sound at any direction over the sphere with as few as four loudspeakers, reproducing distance cues well (Kearney et al., 2012), due to the relative simplicity of the cues (increased distance has less low frequency content and more reverberation). However low-order Ambisonics is only accurate up to a finite frequency, and localisation is inaccurate, with point sources perceived as blurry and large in width.

Different decoding methods soon began to be developed with psychoacoustic motivations (Gerzon, 1977b; Gerzon, 1992b; Gerzon and Barton, 1992) as well as for irregular loudspeaker arrays (Farina, 1998; Heller, Benjamin and Lee, 2010). Gerzon defined two models of sound localisation in his metatheory of auditory localisation (Gerzon, 1992a), the velocity vector, $\mathbf{r_V}$, and the energy vector, $\mathbf{r_E}$. For frequencies up to 700 Hz, the velocity vector is stated to contribute to auditory localisation with

ITDs, and for frequencies greater than 700 Hz, the energy vector is stated as the most important contributor to ILDs (Gerzon, 1977a) and height localisation (Wendt, Frank and Zotter, 2014).

When $\mathbf{r_V} = \mathbf{r_E}$, localisation should be accurate (Daniel, 2000, p. 159). However, if $\mathbf{r_V} < 1$ at low frequencies, or $\mathbf{r_E} < 1$ at high frequencies, localisation becomes more blurry (Daniel, Rault and Polack, 1998; Blauert, 1997), and values of $\mathbf{r_E} < 0.5$ are said to greatly reduce the soundfield image stability (Gerzon, 1980). Refinements of Ambisonic decoding techniques worked to optimise the reproduction based on this theory, as will be explored in Section 3.3.

A development in Ambisonic technology that allowed more accurate reproduction is the addition of higher-order spherical harmonics, referred to as higher-order Ambisonics. B-format Ambisonics was first practically extended beyond first-order to second-order horizontal (but still with first-order height) by Bamford and Vanderkooy through the introduction of two more channels to the B-format equations: U and V (Bamford and Vanderkooy, 1995), in the 1990s. This was extended to full 3D second-order by Malham with R, S and T (Malham, 1999), though the theory supporting up to third-order systems was defined by Gerzon as early as 1973 (Gerzon, 1973).

Higher-order Ambisonics uses a greater number of channels and requires more loudspeakers for playback, but offers the possibility of greater accuracy in the reproduced soundfield (Bamford and Vanderkooy, 1995; Daniel, Rault and Polack, 1998; Malham, 1999; Daniel, 2000; Kearney, 2010) leading to more accurate localisation (Bertet et al., 2007; Braun and Frank, 2011; Bertet et al., 2013; Kearney and Doyle, 2015b) as well as greater applicability to multiple listeners situated outside the exact centre of the loudspeaker array (Moore and Wakefield, 2010; Frank and Zotter, 2017). For near-perfect reconstruction up to 20 kHz in the centre of a loudspeaker array for a sweet spot the size of a human head, an Ambisonic order of 30 or greater is necessary (Palacino and Nicol, 2012; Wiggins, 2017). Table 3.2 presents the resulting source widths of point sources at varying Ambisonic order, calculated from the energy vector

Table 3.2: Estimated source width of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, calculated from the energy vector, reproduced from Bertet et al. (2013).

| $M$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Estimated source width (°) | 45.0 | 30.0 | 22.5 | 18.0 |

(Bertet et al., 2013). This illustrates the narrower source directivity of higher-order Ambisonics.

### 3.2.1 Encoding

The method for defining higher-order Ambisonics in this thesis is as follows. A monophonic audio signal $s$ can be encoded into Ambisonic format $\beta$ with Ambisonic order $M$ for a given location on the sphere of azimuth $\theta$ and elevation $\phi$ by

$$\beta = sY^{\sigma}_{mn} \tag{3.5}$$

where $Y^{\sigma}_{mn}$ are the three-dimensional full normalised (N3D) spherical harmonic (SH) functions of order $m$[2] and degree $n$, defined as

$$Y^{\sigma}_{mn}(\theta, \phi) = N_{mn}P_{mn}(\sin\phi) \times \begin{cases} \cos(n\theta), \text{if } \sigma = +1 \\ \sin(n\theta), \text{if } \sigma = -1 \end{cases} \tag{3.6}$$

where $\sigma = \pm 1$, $P_{mn}(\sin\phi)$ are the associated Legendre functions (Abramowitz and Stegun, 1972), and $N_{mn}$ denotes the normalisation strategy for the amplitudes of different SH orders. The most widely used normalisation strategies are three-dimensional normalised (N3D) and Schmidt semi-normalised (SN3D) (Daniel, 2000):

$$N^{\text{N3D}}_{mn} = \sqrt{(2 - \delta_{n,0})(2m + 1)\frac{(m - n)!}{4\pi(m + n)!}} \tag{3.7}$$

[2] In this thesis, the Ambisonic order, denoted by $M$, is separate to the spherical harmonic order, denoted by $m$, however $M$ is equal to the maximum value of $m$.

$$N_{mn}^{\mathrm{SN3D}} = \sqrt{(2 - \delta_{n,0}) \frac{(m-n)!}{4\pi(m+n)!}} \tag{3.8}$$

where $\delta_{n,0}$ is the Kronecker delta function,

$$\delta_{n,0} \equiv \begin{cases} 1, \text{for } n = 0 \\ 0, \text{for } n \neq 0 \end{cases} \tag{3.9}$$

The practical differences between N3D and SN3D are the variations in cumulative amplitudes as Ambisonic order changes: for the diffuse field, N3D retains a constant root-mean-square (RMS) level while that of SN3D declines as Ambisonic order increases, and for a single point source SN3D retains an approximately even peak level while that of N3D increases as Ambisonic order increases (Chapman et al., 2009). The conversion from SN3D to N3D is simple:

$$\beta_{mn}^{N3D} = \sqrt{2m+1}\beta_{mn}^{SN3D} \tag{3.10}$$

where $\beta_{mn}$ are the Ambisonic format signals. Other normalisation strategies include Max Normalisation (MaxN) (Hollerweger, 2006; Daniel, 2000, p. 156) and FuMa (Malham, 2003; Malham, 2019), which were only developed for $M \leq 3$, so will not be discussed further. In this thesis, N3D is used throughout.

The alphabetic labelling of channels used in B-format, though extended up to $M = 3$ by Malham (Malham, 2003), cannot be used for Ambisonic orders higher than $M = 4$, so a channel labelling that doesn't rely on letters is necessary. A numerical scheme called Ambisonic channel numbering (ACN) has since become the standard method for labelling SH channels (Chapman et al., 2009), which is much more applicable to greater Ambisonic orders. ACN is calculated as

$$\mathrm{ACN} = m^2 + m + n\sigma \tag{3.11}$$

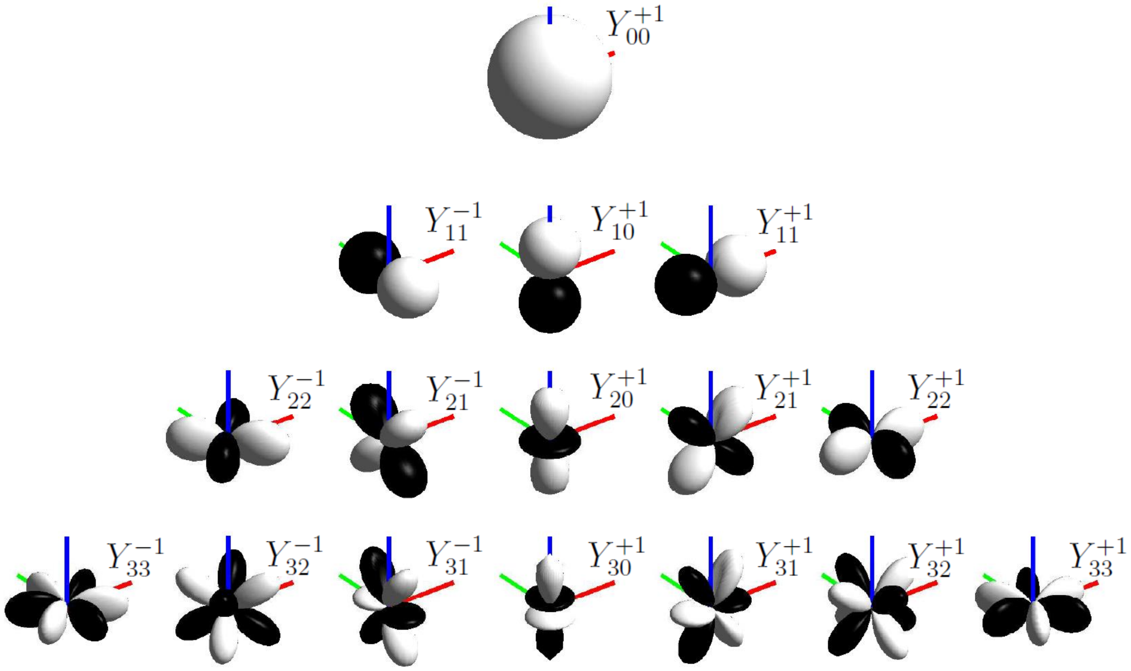
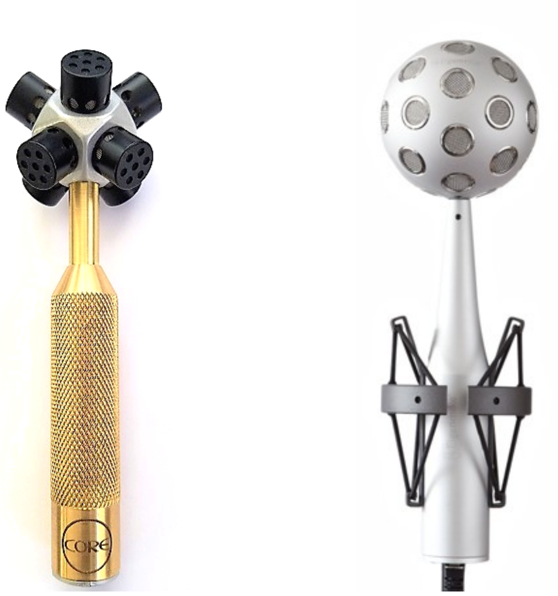


FIGURE 3.2: Illustration of the first 16 spherical harmonic polar patterns, following $Y_{mn}^{\sigma}$. Reproduced from Frank, Zotter and Sontacchi (2015) (edited to add labels).

such that ACN ordering of first-order B-format Ambisonics 0, 1, 2, 3 follows W, Y, Z, X. In this thesis, however, Ambisonic channels are referred to using $k$, where the total number of channels $K$ is determined by the order of Ambisonics such that $K = (M+1)^2$ and the number of channels in a single SH order is given by $k_m = (m-1)^2$. This channel numbering convention follows ACN, only without zero indexing, such that $k = \text{ACN} + 1$.

Figure 3.2 presents the spherical harmonics for channels $\{k = 1, k = 2, ..., k = 16\}$, representing Ambisonic orders up to $M = 3$. The components for $M = 1$ correspond to the omnidirectional $m = 0$ channel and the three figure-of-eight $m = 1$ channels along the y, z and x axes. As the Ambisonic order increases and the number of channels increases, the polar patterns of the spherical harmonics become more complex.

The development of spherical microphone arrays capable of recording higher-order Ambisonics became popular in the 2000s (Abhayapala and Ward, 2002; Rafaely, 2005; Rafaely, Weiss and Bachmat, 2007; Balmages and Rafaely, 2007; Li and Duraiswami,



(A) Core Sound OctoMic, reproduced from Core Sound (2018).

(B) MH Acoustics Eigenmike, reproduced from Acoustics (2013).

FIGURE 3.3: Some commercially available higher-order Ambisonic microphones.

2007). Some modern commercial microphones for recording higher-order Ambisonic signals are presented in Figure 3.3, including the mixed-order (second-order horizontal and first-order height) Core Sound OctoMic[3] and the fourth-order MH Acoustics Eigenmike[4], which uses beamforming to produce the Ambisonic signals from the microphone capsule outputs (Meyer and Elko, 2002; Meyer and Agnello, 2003) and has been shown to produce the most accurate localisation when compared to other commercially available Ambisonic microphones (Bates et al., 2017). The signals of Ambisonic microphone capsules require equalisation and processing (Zotter, Frank and Haar, 2015), and evaluating the audio quality of Ambisonic microphones involves objective models (Moreau, Daniel and Bertet, 2006) and perceptual listening tests (Bertet et al., 2009; Ahrens and Andersson, 2019).

[3] `http://www.core-sound.com/`

[4] `https://mhacoustics.com/`

### 3.2.2 Rotation and Transformation

One of the great benefits of Ambisonics is that, due to the orthogonal nature of spherical harmonics, Ambisonic format signals can be rotated through multiplication of simple matrices. This is used extensively in dynamic binaural Ambisonic rendering for rotating the Ambisonic soundfield to counter head movements, which will be discussed further in Section 3.4. Ambisonic format signals are rotated by multiplication with a rotation matrix $\mathbf{R}_{mK}$ such that

$$\hat{\beta}_K = \beta_K \mathbf{R}_{mK} \tag{3.12}$$

where $\beta_K$ is the original Ambisonic signal. A rotation around the z axis (yaw) requires an azimuth value, a rotation around the y axis (pitch) uses an elevation value, and rotation around the x axis (roll) uses gamma, $\gamma$. The $m = 0$ component ($k = 1$) is left untouched as it contains no directional information.

The order in which the matrix multiplications are applied can change the result (Hollerweger, 2006): the conventional ordering is z, y then x. For the SH components in $m = 1$, the rotation matrices for yaw, pitch and roll, corresponding to rotations around the z, y and x axes, respectively (Daniel, 2000, p. 165), are

$$\mathbf{R}_{\overrightarrow{Z},m=1} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{3.13}$$

$$\mathbf{R}_{\overrightarrow{Y},m=1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix} \tag{3.14}$$

$$\mathbf{R}_{\overrightarrow{X},m=1} = \begin{bmatrix} \cos\gamma & -\sin\gamma & 0 \\ \sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.15}$$

where the columns and rows (from left to right, and top to bottom, respectively) correspond to the $m = 1$ components $\{k = 2, k = 3, k = 4\}$, respectively. A rotation of $\theta = +45°$ around the z axis would therefore yield

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 \\ \hat{\beta}_2 &= 0.707\beta_2 - 0.707\beta_4 \\ \hat{\beta}_3 &= \beta_3 \\ \hat{\beta}_4 &= 0.707\beta_4 + 0.707\beta_2 \end{aligned} \tag{3.16}$$

Higher-order SH rotation matrices are found in Ivanic and Ruedenberg (1996). As well as rotating the soundfield, other manipulations of Ambisonic signals have been developed. The zoom function (Wiggins, 2004, p. 65), also referred to as dominance (Gerzon and Barton, 1992) has been developed for $M = 1$, which alters the polar patterns of the spherical harmonic channels to reproduce a greater amount of one direction. It works by including more of the SH channel for the desired direction into the $m = 0$ component ($k = 1$) as well as some of the $m = 0$ channel into the desired SH channel. Assuming the $\beta_1$ channel is attenuated by 3 dB, as in (3.1), zooming the soundfield in the x axis (Wiggins, 2004, p. 65) is achieved by

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + \frac{\zeta\beta_4}{\sqrt{2}} \\ \hat{\beta}_2 &= \sqrt{1-\zeta^2}\beta_2 \\ \hat{\beta}_3 &= \sqrt{1-\zeta^2}\beta_3 \\ \hat{\beta}_4 &= \beta_4 + \sqrt{2}\zeta\beta_1 \end{aligned} \tag{3.17}$$

where $\beta_K$ is the original Ambisonic signal and $\zeta$ is the amount of zoom in the range $-1 \leq \zeta \leq 1$ (where 0 produces no change, +1 a positive zoom in the x axis and $-1$

a negative zoom in the x axis). Manipulations of higher-order Ambisonic signals can be found in Pomberger and Zotter (2011) and Kronlachner and Zotter (2014) such as soundfield warping about the equator and toward a position such as the north pole. As these are not used later in this thesis they will not be discussed further.

### 3.2.3 Distance Coding

As mentioned in Section 2.5.3, far-field sound sources can be modelled as plane waves (Brungart and Rabinowitz, 1999) and near-field sources can be modelled as spherical waves. The addition of distance coding of sound sources in the Ambisonic encode and decoding processes has been addressed by Daniel *et al.* (Daniel, Nicol and Moreau, 2003; Daniel, 2003; Daniel and Moreau, 2004), which yielded SH order dependent filters whereby low frequencies of higher-order channels are boosted at near distances, and cut at far distances. This allows for approximately spherical waves at near-field distances.

The encoding and decoding equations presented so far in this thesis encode sources at an infinite distance, assuming plane wave properties. However, due to the finite distance of the loudspeakers and lack of correction, this reproduces sources at the distance of the loudspeakers, which is sufficient for the work presented in this thesis as loudspeakers are assumed to be in the far-field, at distances greater than 1 m from the centre of the array.

## 3.3 Decoding Higher-Order Ambisonics

The Ambisonic decode process involves producing a matrix whereby for each loudspeaker in a configuration, a gain for each Ambisonic channel is determined by the loudspeaker's spherical position. For example, a loudspeaker on the horizontal plane (no elevation) will not receive any of the encoded Z channel (as the Z channel does not have any horizontal sound - only vertical).

For an Ambisonic loudspeaker configuration with a total number of loudspeakers $L$, a re-encoding matrix $\mathbf{C}$ with $K$ rows and $L$ columns is calculated by encoding the position of each loudspeaker into SH coefficients using (3.6)[5], such that

$$\mathbf{C} = \begin{bmatrix} Y_1(\theta_1,\phi_1), & Y_1(\theta_l,\phi_l), & ..., & Y_1(\theta_L,\phi_L) \\ Y_k(\theta_1,\phi_1), & Y_k(\theta_l,\phi_l), & ..., & Y_k(\theta_L,\phi_L) \\ ..., & ..., & ..., & ... \\ Y_K(\theta_1,\phi_1), & Y_K(\theta_l,\phi_l), & ..., & Y_K(\theta_L,\phi_L) \end{bmatrix} \tag{3.18}$$

where $Y_k = Y_{mn}^{\sigma}$. A mode-matching (Poletti, 2000; Poletti, 2005) decoding matrix $\mathbf{D}$ is then calculated from the *pseudoinverse* of $\mathbf{C}$ (Daniel, Rault and Polack, 1998) such that

$$\mathbf{D} = \mathrm{pinv}(\mathbf{C}) = \mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1} \tag{3.19}$$

where transposition is notated by a superscript $^T$. Decode matrices therefore follow $L$ rows and $K$ columns. However, in the case that $L = K$, the pseudoinverse will simplify to

$$\mathbf{D} = \mathbf{C}^{-1} \tag{3.20}$$

which is a simple inverse (Daniel, 2000), also known as *projection* decoding. Additionally, if the normalisation scheme is the same, this can be equivalent to the sampling decoder as in (3.4) when using $d = \sqrt{2}$.

Finally, the signal of each loudspeaker $s_l$ for $\{l = 1, l = 2, ..., l = L\}$ is calculated by

$$s_l = \sum_{k=1}^{K} \beta_k * \mathbf{D}_{kl} \tag{3.21}$$

[5] The normalisation scheme implemented in the encode process must be the same as that used in the decode process.

Table 3.3: Approximate spatial aliasing frequency of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using (3.22), for a central listening area of $r = 0.09$ m.

| $M$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $f_{\text{alias}}$ (Hz) | 670 | 1270 | 1870 | 2470 | 3070 |

Table 3.4: Approximate spatial aliasing frequency of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ by an integrated D error of 20%, reproduced from Daniel, Rault and Polack (1998).

| $M$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $f_{\text{alias}}$ (Hz) | 743 | 1346 | 1960 | 2595 | 3230 |

It is possible to decode higher-order Ambisonic signals to lower orders - the unused higher-order channels can simply be discarded.

Theoretically, using a regular arrangement of loudspeakers, Ambisonics can reproduce the soundfield perfectly in the centre of a loudspeaker array at frequencies up to what is commonly referred to as the *spatial aliasing frequency*, $f_{\text{alias}}$ (Poletti, 1996), which can be approximated (Moreau, Daniel and Bertet, 2006; Bertet et al., 2013) as

$$f_{\text{alias}} = \frac{Mc}{4r(M+1)\sin\frac{\pi}{2M+2}} \tag{3.22}$$

where $r$ is the radius of the listening environment (such as the radius of the human head in the case of one listener situated in the centre of the loudspeaker array). Table 3.3 presents the calculated $f_{\text{alias}}$ for $\{M = 1, M = 2, ..., M = 5\}$, with $r = 0.09$ m, as per the radius of the Neumann KU 100 binaural dummy head microphone (Neumann, 2013).

An alternative method of calculating $f_{\text{alias}}$ is to measure the error between an ideal plane wave and the plane wave reconstructed using Ambisonics via the integrated wavefront (Bamford and Vanderkooy, 1995; Poletti, 1996; Daniel, Rault and Polack, 1998; Daniel, 2000). The resulting frequencies, as presented in Table 3.4, are similar to those in Table 3.3, albeit corresponding to approximations for a smaller listening area. In this thesis, the method in (3.22) is used.

At frequencies above $f_{\text{alias}}$, rendering can be inaccurate due to the limited spatial accuracy of recording and reproducing a physical soundfield with a finite number of transducers. The reasons for this will be explained in Section 3.5.

### 3.3.1 Loudspeaker Configurations

To accurately decode three-dimensional Ambisonic signals, a spherical array of loudspeakers distributed with at least semi-regularity is necessary, with a number of loudspeakers $L \geq K$ (Gerzon, 1985) that are diametrically opposed (Gerzon, 1980). However, though the least errors in spatial reproduction occur in the case of $L = K$, audible timbral shift artefacts can be heard when a sound is panned to the exact location of a loudspeaker (Poletti, 2005; Daniel, 2000), and the speaker detent effect is produced at all other panning locations, whereby the sound is pulled toward the closest loudspeaker (Gerzon, 1977a). Therefore, $L > K$ is used in this thesis.

Ideally, the loudspeaker configuration should have a regular distribution over the sphere (Daniel, 2000). The five platonic solids are the only known three-dimensional shapes to offer an entirely regular distribution; these include the tetrahedron, octahedron, cube, icosahedron and dodecahedron, with number of vertices $\{L = 4, L = 6, L = 8, L = 12, L = 20\}$, respectively.

To test the regularity of a loudspeaker configuration for SH sampling, orthonormality error $E_O$ is calculated as

$$E_O = \mathbf{I}_K - \frac{1}{L}\mathbf{C}^T\mathbf{C}, \tag{3.23}$$

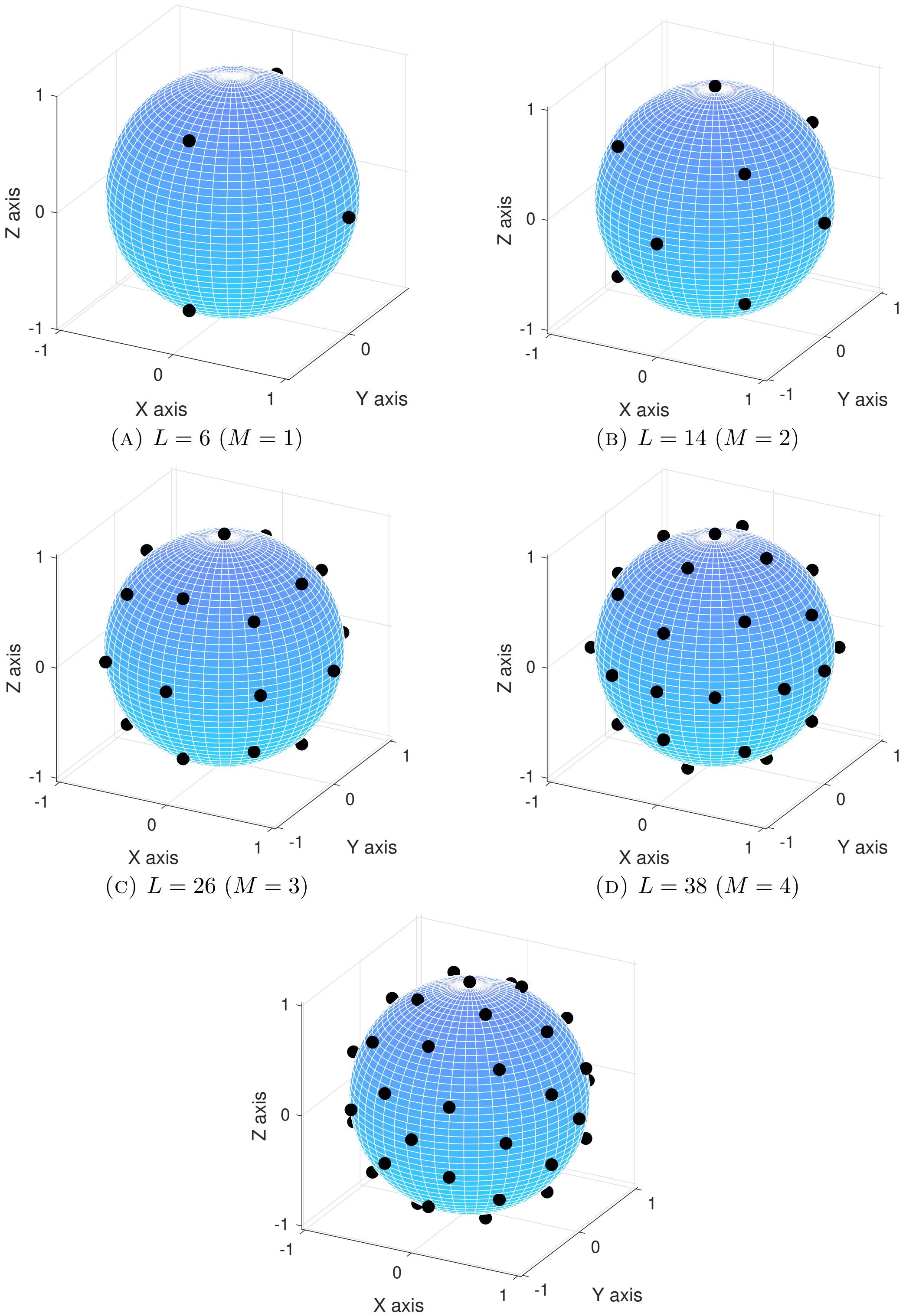
where $\mathbf{I}_K$ denotes the $K \times K$ identity matrix (Daniel, 2000). The orthonormality of the 5 platonic solids is calculated in Moreau, Daniel and Bertet (2006), which shows that the tetrahedron, octahedron and cube offer exact orthonormality up to $M = 1$. The icosahedron and dodecahedron offer exact orthonormality up to $M = 2$, despite the dodecahedron having a greater number of vertices ($L = 20$) than the number of channels at $M = 3$ ($K = 16$) (Moreau, Daniel and Bertet, 2006).

Therefore, for reproducing orders of Ambisonics greater than $M = 2$, alternative configurations of loudspeakers are necessary. Minimising the number of loudspeakers is desirable, as when listening outside the sweet spot, a greater number of loudspeakers produces a reduction in the accuracy of Ambisonic reconstruction (Solvang, 2008) due to increased destructive interference. This will be explained in more detail in Section 3.5. Additionally, fewer loudspeakers are more practical and cost effective.

Lebedev configurations (Lebedev, 1976) are particularly suited to practical reproduction of higher-order Ambisonic signals due to their near-exact orthonormal properties with relatively low number of loudspeakers (Lecomte et al., 2016), and diametrically opposed vertices. For Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, Lebedev configurations corresponding to $\{L = 6$[6]$, L = 14, L = 26, L = 38, L = 50\}$ can be used, respectively. The loudspeaker positions of the five used Lebedev configurations are illustrated in Figure 3.4; the exact vertices of which are obtained from Burkardt (2013b). An additional practical convenience of these Lebedev configurations is that the $L = 50$ configuration nests the lower order Lebedev configurations (apart from the $L = 38$ configuration), making comparisons of different Ambisonic orders practically viable over loudspeakers (Thresh, Armstrong and Kearney, 2017).

To assess the orthonormality of the five Lebedev configurations, $E_O$ has been calculated using (3.23). Figure 3.5 presents the orthonormality error matrices of the five Lebedev configurations for Ambisonic orders up to $M + 1$. In these plots, the orthonormality matrix of each SH order is displayed by the shade, whereby no error produces white and increased error is shown by a darker shade. The ideal plot (indicating exact orthonormality) would consist of a white grid up to the maximum Ambisonic reproduction order of the loudspeaker configuration. The exact orthonormality of the $L = 6$ configuration is evident from the lack of error along the diagonal of the plot. The small errors observed in the higher-order configurations show the near-orthonormality and therefore suitability for the chosen Ambisonic orders.

[6] Equivalent to an octahedron.



(A) $L = 6$ $(M = 1)$

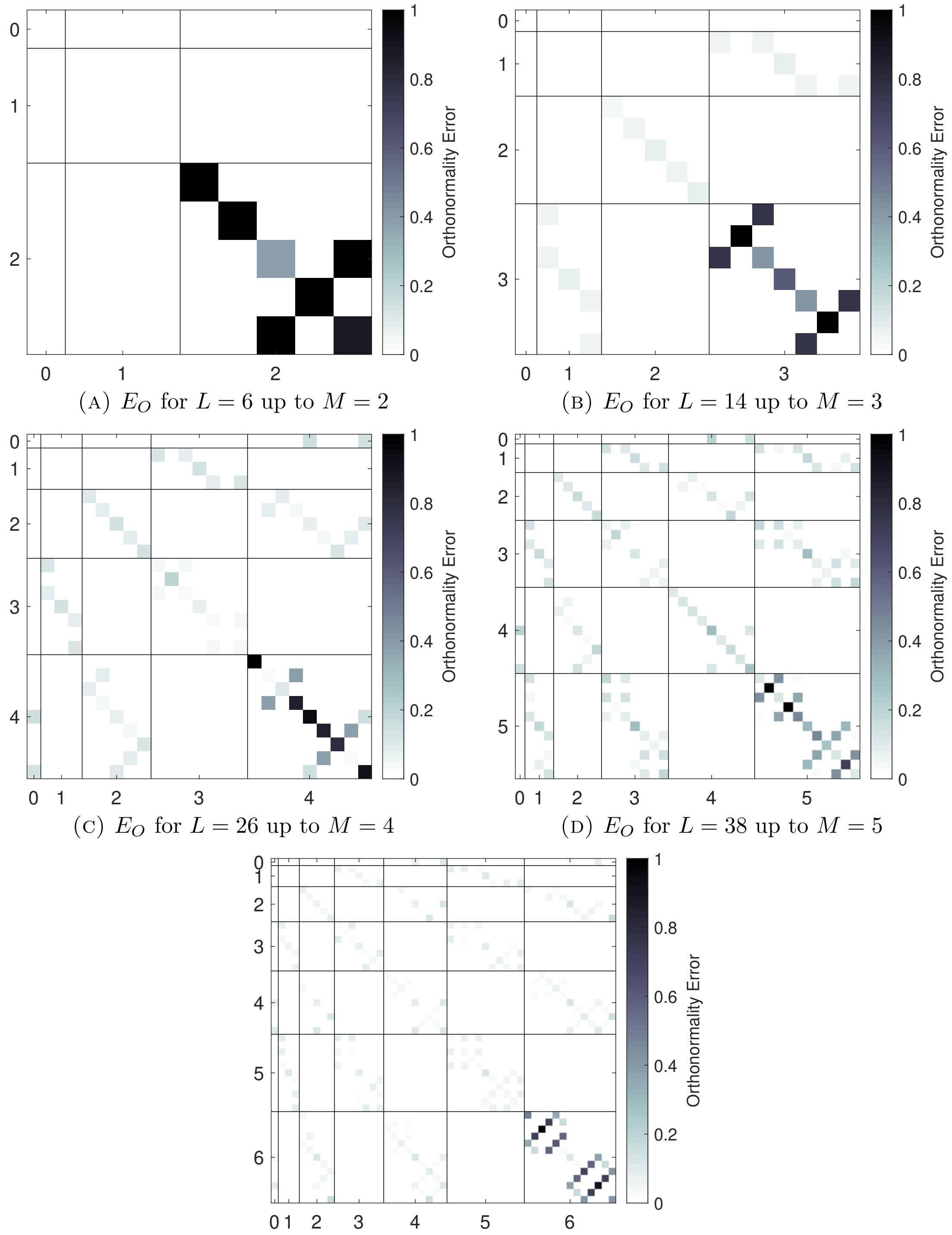
(B) $L = 14$ $(M = 2)$

(C) $L = 26$ $(M = 3)$

(D) $L = 38$ $(M = 4)$

(E) $L = 50$ $(M = 5)$

FIGURE 3.4: Loudspeaker layouts of the Lebedev configurations used in this thesis with corresponding order of Ambisonics.



(A) $E_O$ for $L = 6$ up to $M = 2$

(B) $E_O$ for $L = 14$ up to $M = 3$

(C) $E_O$ for $L = 26$ up to $M = 4$

(D) $E_O$ for $L = 38$ up to $M = 5$

(E) $E_O$ for $L = 50$ up to $M = 6$

FIGURE 3.5: Orthonormality error matrices for the Lebedev loudspeaker configurations used in this thesis corresponding to Ambisonic orders. In orthonormality error matrices plots, spherical harmonics of different orders are separated to aid visual clarity.

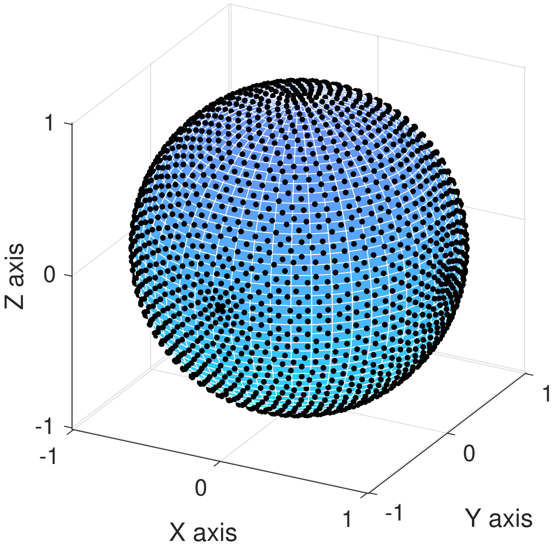
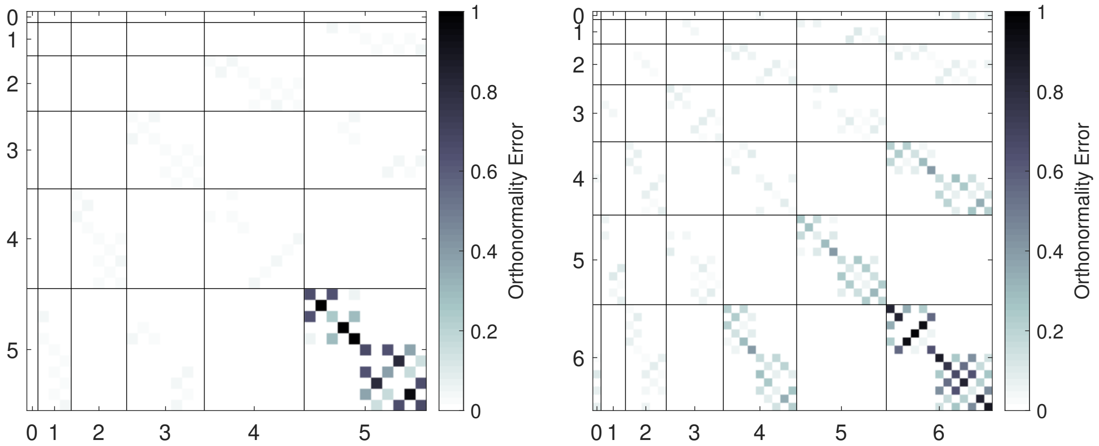


FIGURE 3.6: Loudspeaker layout of the $L = 2702$ Lebedev configuration used in this thesis for $M = 36$.

For comparisons to very high order ($M > 30$) Ambisonics in this thesis, the $L = 2702$ Lebedev configuration is used, which has low orthonormality error up to $M = 44$. Figure 3.6 presents the loudspeaker positions of the $L = 2702$ Lebedev configuration.

Other spherical loudspeaker configurations are suitable for higher-order Ambisonics, such as the Pentakis dodecahedron ($L = 32$) (Moreau, Daniel and Bertet, 2006; Lecomte et al., 2015) and the Pentakis icosidodecahedron ($L = 42$). These have been used in decoding (Kearney and Doyle, 2015b; Gorzel et al., 2019) and spherical microphone design (Moreau, Daniel and Bertet, 2006). Figure 3.7 presents the orthonormality error matrices of the two configurations; the exact vertices of which were obtained from Kearney and Doyle (2015a) and (Armstrong et al., 2018a) for the Pentakis dodecahedron and Pentakis icosidodecahedron, respectively. The Pentakis dodecahedron is capable of relatively low orthonormality error at $M = 4$. For the Pentakis icosidodecahedron, a low error is also observed up to $M = 4$, but for $M = 5$ a markedly higher error than the $L = 50$ Lebedev configuration is observed.

Another form of quadrature appropriate for higher-order Ambisonic reproduction is spherical T-designs (Hardin and Sloane, 1996). T-designs offer exact orthonormality for SH sampling if they fulfil the requirement $T \geq 2M + 1$ (Zotter, Frank and Sontacchi, 2010). For Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, T-designs



(A) $E_O$ for $L = 32$ up to $M = 5$

(B) $E_O$ for $L = 42$ up to $M = 6$

Figure 3.7: Orthonormality error matrices for the Pentakis configurations corresponding to Ambisonic orders. In orthonormality error matrices plots, spherical harmonics of different orders are separated to aid visual clarity.

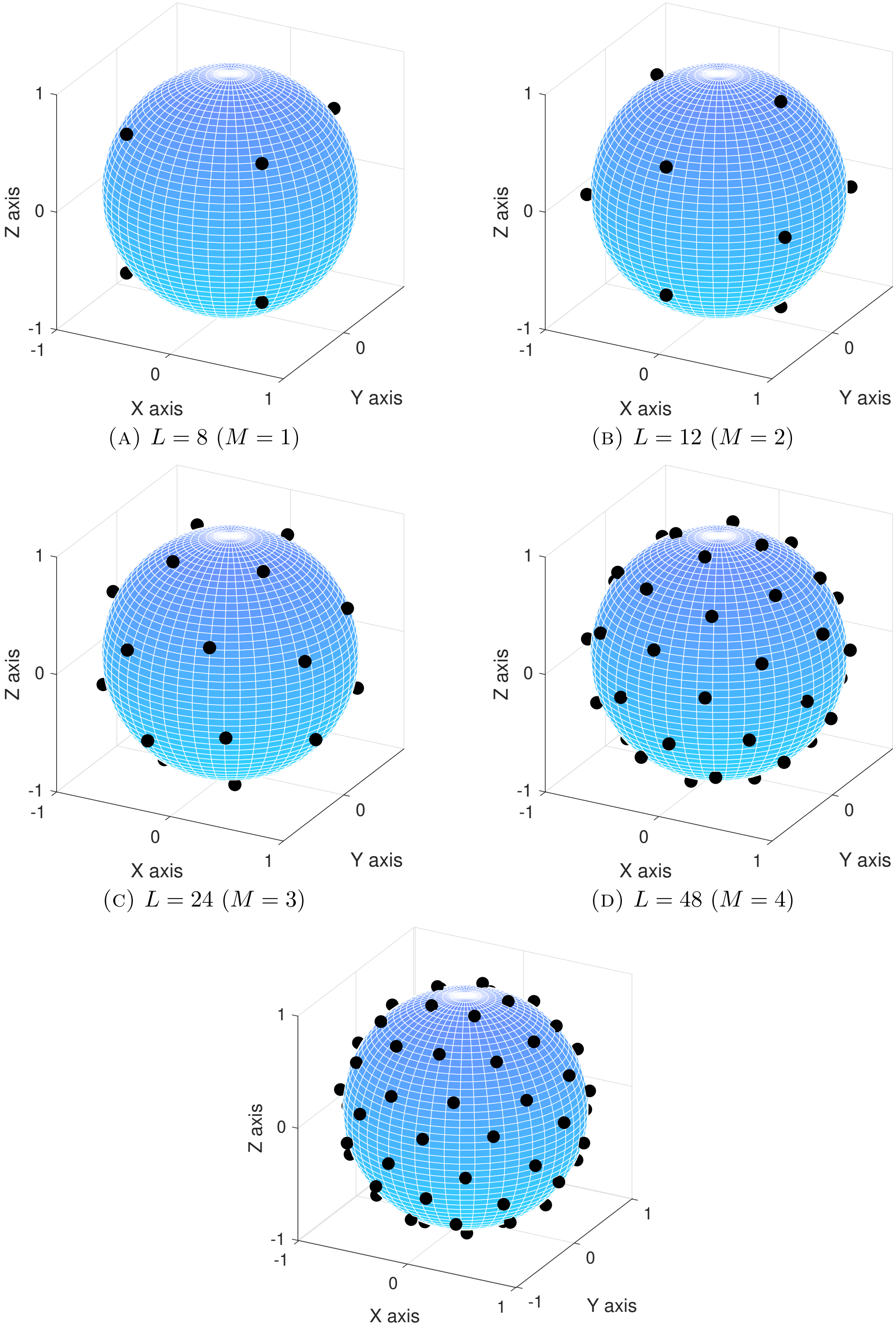
corresponding to $\{L = 8^{7}, L = 12^{8}, L = 24, L = 48, L = 70\}$ can be used, respectively. The loudspeaker positions of these configurations are shown in Figure 3.8; the exact vertices of which are obtained from Burkardt (2013b). The orthonormality error matrices are presented in Figure 3.9, which illustrates the exact orthonormal properties of the sampling scheme due to the lack of errors.

Though T-designs offer the best regularity for Ambisonic reproduction, the higher-order configurations feature a significantly greater number of loudspeakers than Lebedev configurations. Additionally, the lack of shared vertices between different configurations makes them less practical in real scenarios and HRTF measurements (see Section 3.4). The spherical coordinates of all loudspeaker configurations used in this thesis, unless stated otherwise, are therefore Lebedev configurations.

It is important that the decoder selected for a loudspeaker setup is correct. If the loudspeaker arrangement is irregular or incomplete (such as hemispherical), using pseudoinverse mode matching decoding reconstructs the soundfield poorly (Wiggins, 2007; Pomberger and Zotter, 2009; Trevino et al., 2010; Heller, Benjamin and Lee, 2010), which leads to poor localisation (Zotter, Pomberger and Noisternig, 2010) and inconsistent amplitudes with varying virtual source position. Therefore, decoding

[7] Equivalent to a cube.

[8] Equivalent to an icosahedron.



(A) $L = 8$ $(M = 1)$

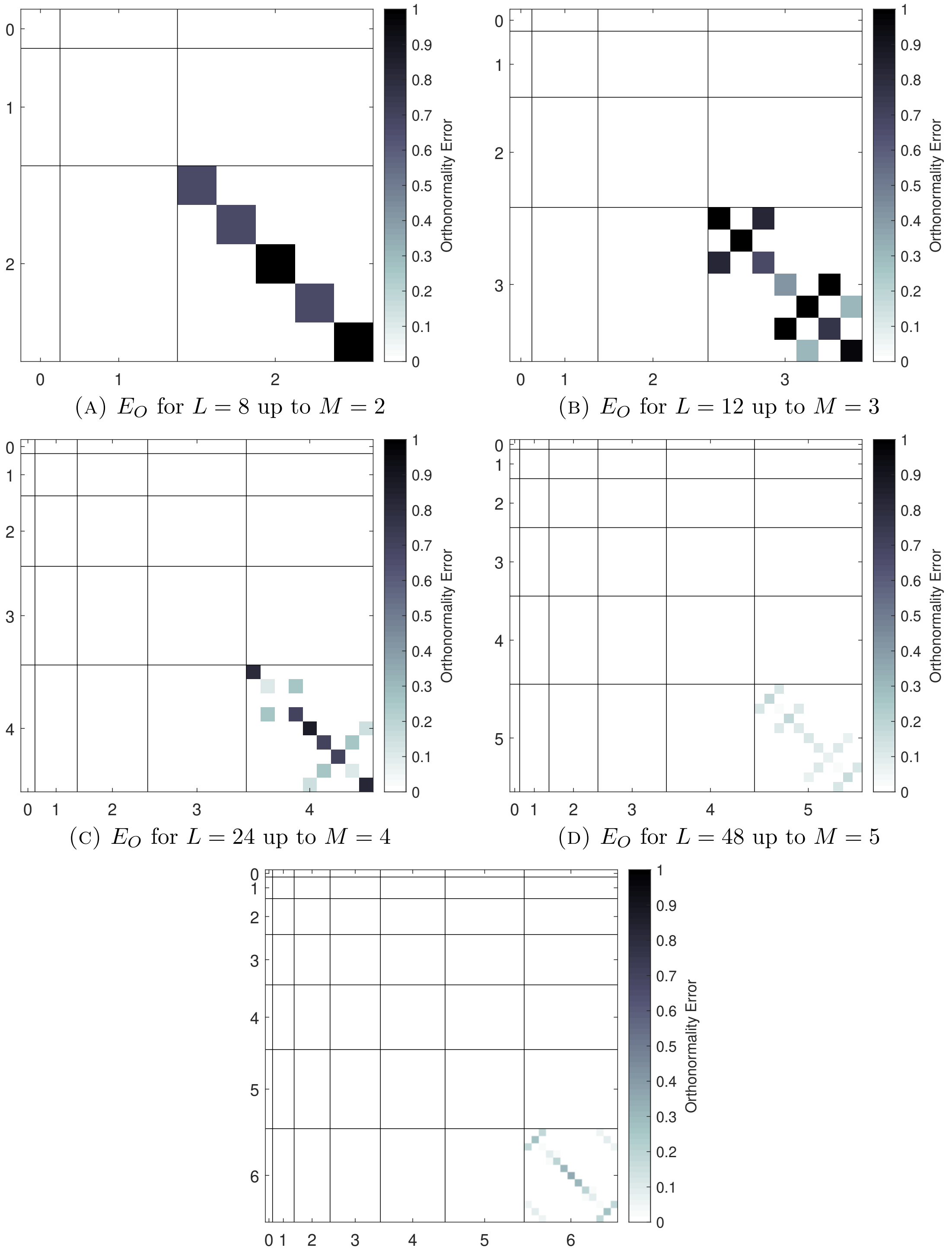
(B) $L = 12$ $(M = 2)$

(C) $L = 24$ $(M = 3)$

(D) $L = 48$ $(M = 4)$

(E) $L = 70$ $(M = 5)$

FIGURE 3.8: Loudspeaker layouts of the T-design configurations used in this thesis with corresponding order of Ambisonics.



(A) $E_O$ for $L = 8$ up to $M = 2$

(B) $E_O$ for $L = 12$ up to $M = 3$

(C) $E_O$ for $L = 24$ up to $M = 4$

(D) $E_O$ for $L = 48$ up to $M = 5$

(E) $E_O$ for $L = 70$ up to $M = 6$

FIGURE 3.9: Orthonormality error matrices for the five T-design configurations used in this thesis. In orthonormality error matrices plots, spherical harmonics of different orders are separated to aid visual clarity.

strategies for irregular arrays have been developed, such as energy-preserving (Zotter, Pomberger and Noisternig, 2012), which attempts to reproduce the energy evenly at any angle to reduce the source direction-dependent variation in loudness that comes with irregular arrays. Alternatively, All Round Ambisonic Decoding (AllRAD) (Zotter and Frank, 2012) decodes the soundfield to a virtual T-design loudspeaker arrangement before using VBAP to pan the loudspeaker feeds between the closest (up to 3 for three-dimensional reproduction) loudspeaker(s), and can be improved through combination with sampling decoding (Zotter, Frank and Pomberger, 2013; Zotter and Frank, 2018). However, these decoding methods can produce a change in apparent source width, which is addressed in constant angular spread decoding (Epain, Jin and Zotter, 2014), which finds the area of the array that produces the greatest source spread and attempts to recreate that spread for all directions to improve consistency.

### 3.3.2 SH Channel Weightings

Above $f_{\text{alias}}$ and outside the sweet spot, Ambisonic reproduction is inaccurate. Therefore, alternative weightings for the SH channels have been developed for this frequency region to maximise the accuracy of the reproduced psychoacoustic cues, as opposed to attempting to recreate the original soundfield as accurately as possible.

A decoding matrix can be altered with SH channel weightings by

$$\hat{\mathbf{D}}_m = \mathbf{D}_m g_m \tag{3.24}$$

where $\hat{\mathbf{D}}_m$ is the new decoding matrix, and $g_m$ denotes the SH order dependent gains. The standard mode-matching pseudoinverse decoding, as in (3.19), does not include any additional SH channel weightings. It is therefore described as *basic* weighted from here on, where $g_m = 1$ for all values of $m$.

Above $f_{\text{alias}}$, basic SH weighting produces an energy vector $\mathbf{r_E} = 1$ only when the virtual source direction coincides with the position of a loudspeaker and $L = K$, which causes only a single loudspeaker to output. When $L > K$, as used throughout this thesis, $\mathbf{r_E} < 1$ for all virtual source directions, as the sound is never emitted from only a single loudspeaker.. Max $\mathbf{r_E}$ SH channel weighting aims to reproduce $\mathbf{r_E} = 1$ for all directions (Gerzon and Barton, 1992; Gerzon, 1992a; Daniel, Rault and Polack, 1998; Daniel, 2000), and has been shown to improve the spectral response and auditory impression of height at low orders of Ambisonics (Gorzel, Kearney and Boland, 2014), as well as the reproduction of ILD cues (Daniel, Rault and Polack, 1998).

The gains $g_m$ to be applied in (3.24) to maximise $\mathbf{r_E}$ are found from differentiation of $\mathbf{r_E}$ with respect to $g_m$ (Daniel, 2000, p. 312), such that

$$\begin{aligned} &\frac{\delta \mathbf{r_E}}{\delta g_m} = 0 \\ &\qquad \Rightarrow \mathbf{r_E}(2m+1)g_m = (m+1)g_{m+1} + m g_{m-1} \end{aligned} \tag{3.25}$$

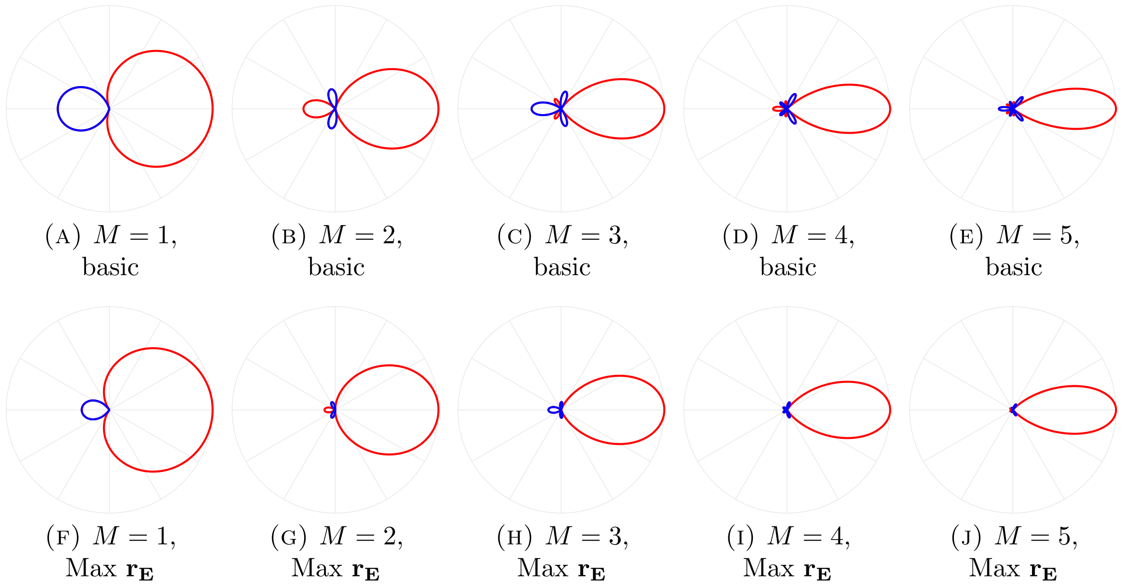
This recurrence equation can then be rewritten using Legendre polynomials to match Bonnets' Recursion Formula (Morse and Ingard, 1968). If the rules $g_{-1} = 0$ and $g_{M+1} = 0$ and therefore $g_0 = 1$ and $g_1 = r_E$ are followed, such that $\eta = \mathbf{r_E}$ and $g_m = P_m(\mathbf{r_E})$, for SH orders $\{m = 0, m = 1, ..., m = M\}$:

$$\eta(2m+1)P_m(\eta) = (m+1)P_{m+1}(\eta) + mP_{m-1}(\eta) \tag{3.26}$$

where $\mathbf{r_E}$ is the largest root of $P_{M+1}$:

$$P_{M+1} = g_{M+1} = 0 \tag{3.27}$$

In practice, Max $\mathbf{r_E}$ weighting reduces the amplitude of higher-order components, which changes the width of the reproduced source.



(A) $M = 1$, basic

(B) $M = 2$, basic

(C) $M = 3$, basic

(D) $M = 4$, basic

(E) $M = 5$, basic

(F) $M = 1$, Max $\mathbf{r_E}$

(G) $M = 2$, Max $\mathbf{r_E}$

(H) $M = 3$, Max $\mathbf{r_E}$

(I) $M = 4$, Max $\mathbf{r_E}$

(J) $M = 5$, Max $\mathbf{r_E}$

FIGURE 3.10: Horizontal virtual microphone pickup patterns of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, using both basic weighting and Max $\mathbf{r_E}$ (Lebedev loudspeaker configurations). Red and blue colours denote positive and negative phase, respectively.

One way of illustrating the Ambisonic source width is through the virtual microphone pickup pattern. These are generated in this thesis by encoding 360 point sources (1 sample impulses) at locations $\{\theta = 1°, \theta = 2°, ..., \theta = 360°\}$ and $\phi = 0°$, before decoding to Lebedev loudspeaker configurations corresponding to each order of Ambisonics. The Ambisonic encode and decode functions used in this thesis are from the Politis MATLAB library (Politis, 2016). The virtual microphone magnitude for each source location is then calculated as in (3.21) for the loudspeaker at $(\theta = 0°, \phi = 0°)$. Figure 3.10 presents the horizontal virtual microphone pickup patterns of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using both basic and Max $\mathbf{r_E}$ SH channel weightings. The plots display the absolute values, with positive amplitudes shown in red and negative amplitudes in blue. It is observable that, for all orders of Ambisonics, Max $\mathbf{r_E}$ increases the width of the frontal lobe due to the greater amplitude of lower SH order channels, and reduces the side and rear lobes.

For loudspeaker Ambisonic reproduction over a large listening area, in-phase weighting can be used. In-phase weighting was first introduced by Malham in the 1990s (Malham, 1992) and extended to higher orders in the early 2000s (Monro, 2000).

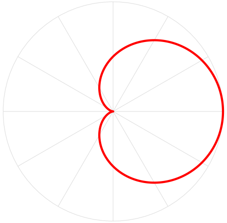


FIGURE 3.11: Virtual microphone pickup pattern of $M = 1$ using in-phase SH channel weighting.

TABLE 3.5: SH channel gains using different weighting schemes for $M = 1$.

| SH weighting | Basic | Max $\mathbf{r_E}$ | In-phase |
|---|---|---|---|
| $g_0$ | 1 | 1 | 1 |
| $g_1$ | 1 | 0.577 | 0.333 |

Like Max $\mathbf{r_E}$, in-phase increases the amplitude of the lower SH order channels. However, in-phase puts a greater weight on lower SH order channels in order to ensure that no out-of-phase signals are played back through opposing loudspeakers. This improves reproduction over a large listening area, though it reduces the accuracy of the reproduced soundfield as a compromise, causing a more blurred spatial image. For 3D reproduction, in-phase weightings are calculated (Neukom, 2007; Daniel, 2000, p. 184)[9] as

$$g_m = \frac{M!(M+1)!}{(M+m+1)!(M-m)!} \tag{3.28}$$

The virtual microphone pickup pattern of $M = 1$ using in-phase weighting is presented in Figure 3.11.

Table 3.5 presents the values of $g_m$ for $M = 1$ using the two presented SH channel weightings (basic, i.e. no weighting, is included for reference). It is evident that the reduced amplitudes of the higher SH order components will produce a reduction in overall amplitude.

To assess the amplitude reduction of different Ambisonic orders with psychoacoustic SH channel weightings, the root-mean-square (RMS) values of $g_m$ have been calculated

[9] The equation presented in (Daniel, 2000, p. 184) features $(M - n)!$ in the denominator. This is an error, and should be $(M - m)!$, as written in Neukom (2007).

Table 3.6: RMS SH channel weightings for varying Ambisonic orders.

| $M$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| RMS $g_m$ (Max $\mathbf{r_E}$) | 0.707 | 0.633 | 0.600 | 0.581 | 0.569 |
| RMS $g_m$ (In-phase) | 0.577 | 0.447 | 0.378 | 0.333 | 0.302 |

for Max $\mathbf{r_E}$ and in-phase for $\{k = 1, k = 2, ..., k = K\}$ (such that $g_m$ at higher SH orders is accounted for multiple times in the calculation). Table 3.6 presents the RMS $g_m$ values for $\{M = 1, M = 2, ..., M = 5\}$. It shows that the trend of reduced overall amplitude becomes more pronounced as the Ambisonic order increases, and is also more pronounced for in-phase weighting than Max $\mathbf{r_E}$.

An observation has been made that the reduction in amplitude due to SH channel weightings is not uniform across all frequencies. To illustrate this, Figure 3.12 presents three $M = 5$ Ambisonic renders at $(\theta = 0°, \phi = 0°)$ which have been generated and decoded to the $L = 50$ Lebedev loudspeaker configuration, before being rendered binaurally (using the method detailed further in Section 3.4), with different SH channel weightings. It is clear that there is little reduction at low frequencies; the reduction is focussed at frequencies $> 500$ Hz. This is likely due to the greater width of the frontal lobe with the SH channel weightings (as illustrated in Figure 3.10 and Figure 3.11), which produces a wider image spread and therefore increases spatial aliasing, which will be explained in further detail in Section 3.5.

The reduction in overall amplitude can be compensated (Gerzon and Barton, 1992; Daniel, Rault and Polack, 1998; Jot, Larcher and Pernaux, 1999) by rewriting (3.24) as

$$\hat{\mathbf{D}}_m = \mathbf{D}_m \frac{g_m}{g_{\text{norm}}} \tag{3.29}$$

where $g_{\text{norm}} = \text{RMS } g_m$. Figure 3.13 presents the same renders as shown in Figure 3.12, but with this amplitude compensation. Frequencies above $f_{\text{alias}}$ are now much closer in overall magnitude, and it is now the low frequencies that appear boosted with SH channel weights. This can be negated through dual-band decoding.

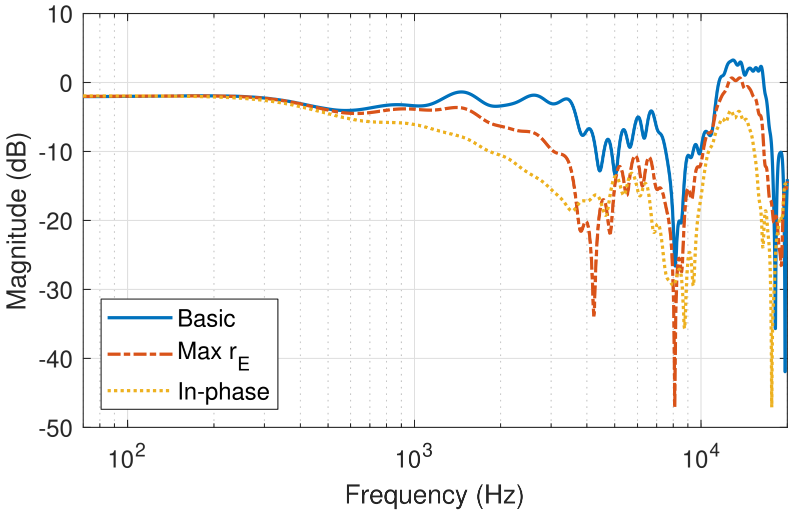


FIGURE 3.12: Example of the high frequency differences of decoding with different SH channel weightings, $M = 5$ binaural Ambisonic rendering at $(\theta = 0°, \phi = 0°)$ (left ear).

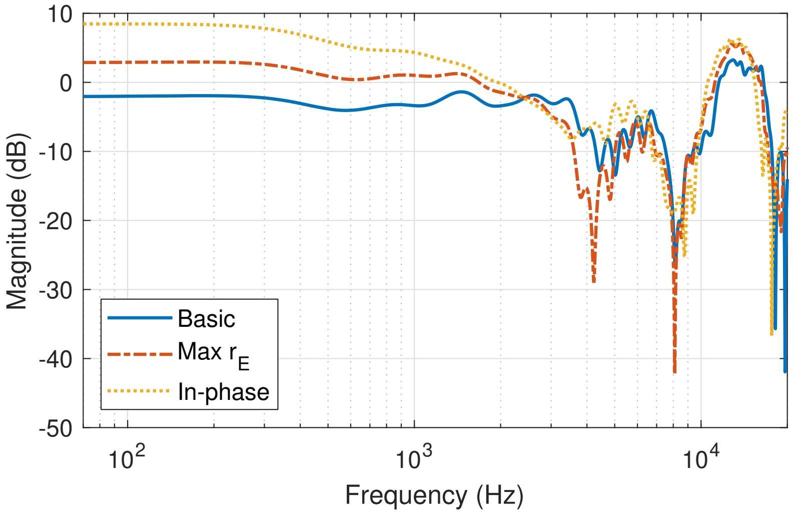


FIGURE 3.13: Example of amplitude normalisation when using different SH channel weightings, $M = 5$ binaural Ambisonic rendering at $(\theta = 0°, \phi = 0°)$ (left ear).

### 3.3.3 Dual-Band Decoding

It is possible to use more than one decode method simultaneously, by calculating separate decode matrices for low and high frequencies and implementing a crossover network between the two. This is referred to as a dual-band decode method (Heller, Lee and Benjamin, 2008). By decoding the Ambisonic format sound separately for low and high frequencies, the resulting soundfield can produce more accurate localisation (Benjamin, Lee and Heller, 2006).

In the case of a semi-regular loudspeaker configuration with a single listener in the centre of the array, psychoacoustic motivation for dual-band decoding leads to the use of mode-matching pseudoinverse decoding with basic SH channel weighting, as defined in (3.19), at frequencies up to $f_{\text{alias}}$, as calculated using (3.22). This produces the closest approximation of the original soundfield for near-regular loudspeaker arrangements with a non-square re-encoding matrix (Gerzon and Barton, 1992), and is optimised for the velocity vector (Gerzon, 1977a; Gerzon, 1992a; Daniel, Rault and Polack, 1999) so will therefore reproduce ITD with greater accuracy.

At frequencies above $f_{\text{alias}}$, mode-matching pseudoinverse decoding with Max $\mathbf{r_E}$ channel weighting should be used, which is optimised for the energy vector (Gerzon and Barton, 1992) and will therefore reproduce ILD more accurately than basic weighting. However, if a dual-band decode method is not possible, Max $\mathbf{r_E}$ is recommended for full-band reproduction in Benjamin, Lee and Heller (2006).

In the crossover network between the two decoders, the crossover should be gradual (Farina, 1998) and the filters must be phase matched to avoid unwanted destructive interference around the crossover frequency (Gerzon and Barton, 1992; Heller, Lee and Benjamin, 2008). The frequency of crossover $f_c$ should be informed by the size of the listening area: personal listening can therefore afford a higher value of $f_c$. In this thesis, with a listening area the size of the human head (i.e. for personal listening), the crossover frequency $f_c = f_{\text{alias}}$, therefore values of $f_c$ are the same as those presented in Table 3.3, for $\{M = 1, M = 2, ..., M = 5\}$.

By implementing dual-band decoding, the low frequency boost of amplitude normalised Max $\mathbf{r_E}$ decoding can be negated. Using the same $M = 5$ binaural example as before, a dual-band render has been calculated by using basic weights at low frequencies and Max $\mathbf{r_E}$ at high frequencies. Figure 3.14 presents the dual-band render, with the single band renders included for reference. The crossover is a linear phase finite impulse response (FIR) filter with Chebyshev windowing (Harris, 2004)

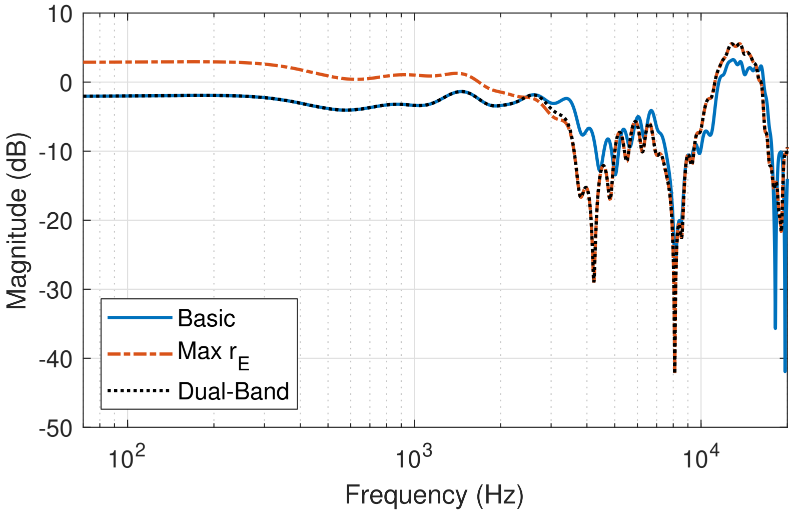


FIGURE 3.14: Example of a dual-band render using basic SH channel weighting at low frequencies and Max $\mathbf{r_E}$ above $f_{\text{alias}}$ (normalised), $M = 5$ binaural Ambisonic rendering at $(\theta = 0°, \phi = 0°)$ (left ear).

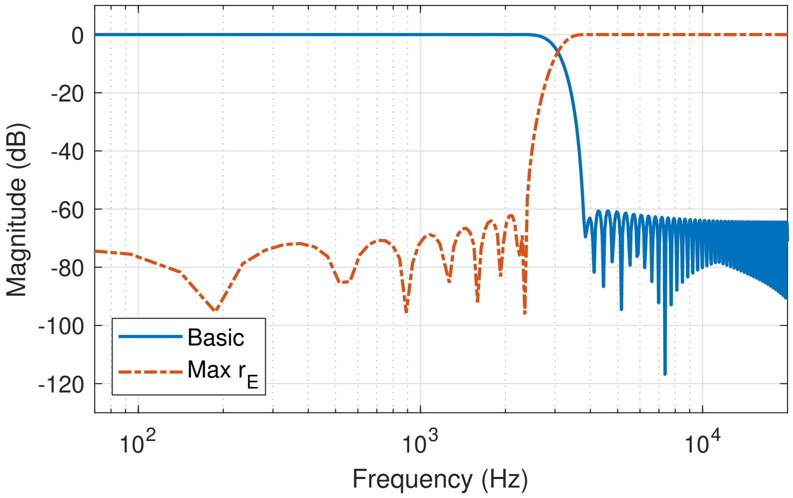


FIGURE 3.15: Magnitude response of the crossover for dual-band decoding, $M = 5$.

of 128 taps and 50 dB ripple, which produces a lower stop-band amplitude than alternative windowing methods. The frequency response of the crossover filter is presented in Figure 3.15.

## 3.4 Binaural Ambisonic Rendering

The binaural Ambisonic approach to spatial audio is popular in virtual reality applications due to the rotational capabilities of spherical harmonics. Additionally, binaural Ambisonic rendering removes the need for computationally expensive HRTF interpolation (Wenzel and Foster, 1993) of a highly dense grid of HRTFs. Furthermore, with standard binaural synthesis through HRTF convolution, each separate source requires its own convolution pair for the left and right ears. In Ambisonic binaural rendering, the total number of convolutions is only based on the amount of virtual loudspeakers used in the decode process. Binaural Ambisonic reproduction therefore allows spatial audio rendering at any direction with as few as four convolutions per ear (in the case $M = 1$).

Ambisonic signals can be rendered binaurally by performing a real-time convolution of the decoded Ambisonic loudspeaker signals with HRTFs at the position of each loudspeaker. This was first introduced by McKeag and McGrath (1996), developed further by (Noisternig et al., 2003b; Noisternig et al., 2003a), and labelled the *virtual loudspeaker* approach by Jot *et al.* (Jot, Wardle and Larcher, 1998; Jot, Larcher and Pernaux, 1999). When combined with a head-tracking system, and by using the head orientation data to inform the rotation matrices introduced in Section 3.2.2 to counter-rotate the Ambisonic soundfield prior to the decode process, binaural Ambisonic rendering can be updated dynamically and give the impression of a stable virtual soundfield. This can then help to deliver the dynamic binaural cues as detailed in Section 2.6.5.

Ambisonic signals can be rendered binaurally using the virtual loudspeaker approach (repeated for the left and right ears) by

$$B = \sum_{l=1}^{L} H_l * s_l \tag{3.30}$$

where $B$ denotes the binaural signals and $s_l$ denotes the loudspeaker signals, as calculated in (3.21). The amount of convolutions per ear is therefore given by $L$. However, if using dual-band decoding, the amount of convolutions must be doubled, and a real-time crossover network is necessary between the low and high frequency decoded binaural signals.

Alternatively, it is possible to encode the HRTFs into the spherical harmonic domain in the decode process by multiplication of the decoding matrix $\mathbf{D}$ gain coefficients with the HRTFs for each loudspeaker, followed by summation of the resulting SH channels for each loudspeaker:

$$\mathbf{D}^{\mathrm{SH}} = \sum_{l=1}^{L} H_l \mathbf{D}_l \tag{3.31}$$

to produce virtual loudspeaker binaural decoders (repeated for the left and right ears). In a dual-band decoding scenario, this can be repeated for both basic and Max $\mathbf{r_E}$ decoding matrices, whereby the two binaural decoders can then be combined through an offline crossover network to produce the dual-band binaural Ambisonic decoder.

Binaural Ambisonic rendering $B$ is then achieved through a summation of each SH channel of the encoded signal $\beta_K$ convolved with each SH channel of the decoder $\mathbf{D}_K^{SH}$ (repeated for the left and right ears) by

$$B = \sum_{k=1}^{K} \beta_k * \mathbf{D}_k^{\mathrm{SH}} \tag{3.32}$$

where the amount of convolutions per ear is given by $K$.

The approaches in (3.30) and (3.32) give equivalent results. However, decoding Ambisonic signals using the SH encoded HRTF binaural decoders has two advantages: in the case that $L > K$, the SH encoded binaural decoder approach requires fewer convolutions; dual-band decoding can be implemented offline in the decoder generation stage, thus removing the increased real-time computation from double the number of convolutions and crossover filtering in a dual-band decoding scenario. Hence, this method used is throughout this thesis for binaural Ambisonic rendering.

Some recent methods for binaural Ambisonic rendering have moved away from the virtual loudspeaker approach and instead focused on order truncation of an approximately spatially continuous spherical harmonic represented HRTF dataset (Avni et al., 2013; Bernschütz et al., 2014), whereby every order of Ambisonics uses the same dense loudspeaker configuration and decoder, as in (3.32), designed for a very high order of Ambisonics ($M > 30$). When using a lower order of Ambisonics in this technique, the higher-order data is simply discarded. For this reason, it is called order truncation. In this case, order-dependent decoding optimisation strategies such as Max $\mathbf{r_E}$ channel weights are not used.

Though the benefits of using a single dataset of HRTFs for all orders can be appreciated, such as a single decoding matrix and single dataset of measurements, this approach requires a highly dense dataset of HRTFs measured at points on the sphere distributed by a regular (or at least semi-regular) quadrature. For individualisation therefore, this is considered infeasible and impractical at present, despite techniques such as reciprocity (Zotkin et al., 2006) and multiple swept sines (Majdak, Balazs and Laback, 2007) offering faster measurement times, and BEM simulation techniques becoming more widely used. Recent research on the upsampling of sparse datasets to dense datasets has been conducted (Alon et al., 2018; Porschmann, Arend and Brinkmann, 2019), though these are approximations and not exact, and therefore this specific technique is not utilised in this thesis.

Additional issues caused by using a single HRTF dataset for all Ambisonic orders are that severe high frequency roll-off is observed at low truncation orders, as demonstrated for $M = 1$ in Figure 3.16. The reasons for this are explained in Section 3.5.

The binaural Ambisonic rendering in this thesis focuses on virtual loudspeaker binaural rendering of Ambisonic signals with sparse loudspeaker configurations as opposed to a single dense loudspeaker configuration for all Ambisonic orders. This allows the methods that will be presented to be directly applicable to individualised binaural Ambisonic rendering with the current physical measurement capabilities.

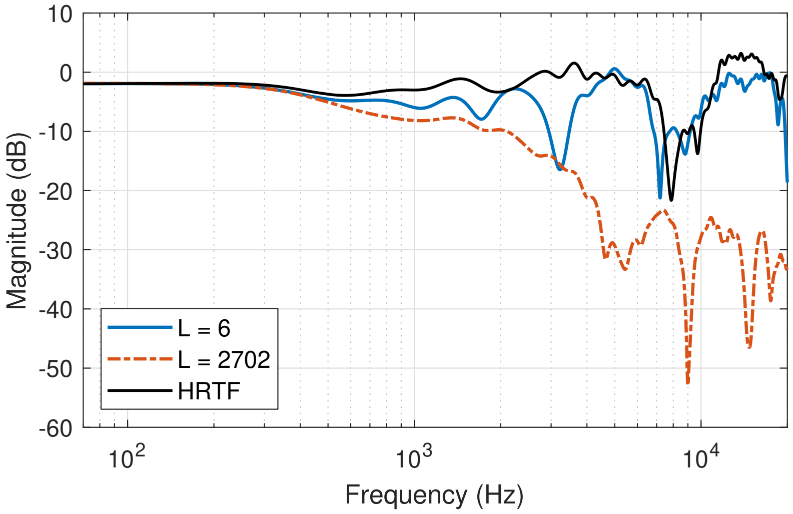


Figure 3.16: Comparison of virtual loudspeaker ($L = 6$) and order truncation ($L = 2702$) approaches at $M = 1$, with HRTF as reference, ($\theta = 0°, \phi = 0°$) (left ear).

## 3.5 Spatial Aliasing

For ideal Ambisonic soundfield rendering, the reproduction should be accurate up to 20 kHz. In binaural rendering, this means that an Ambisonic rendered HRTF at any direction should be equivalent to an unprocessed HRTF for the same direction. Figure 3.17 presents a comparison of binaural Ambisonic renders with an HRTF at ($\theta = 0°, \phi = 0°$) for $M = 1$, $M = 5$ and $M = 36$[10]. It is possible to observe the greater accuracy of higher Ambisonic orders at higher frequencies: $M = 1$ is accurate up to around 500 Hz, $M = 5$ to around 2.5 kHz, whilst the $M = 36$ remains a very close approximation of the HRTF up to 20 kHz. Small differences do exist however, which are likely due to the lack of exact orthonormality of the $L = 2702$ Lebedev loudspeaker configuration.

As discussed in Chapter 2, low frequencies are less directional than high frequencies. The source width of low frequencies can therefore be wider than at high frequencies. Referring back to Figure 3.10, it is evident that lower orders of Ambisonics feature wider virtual microphone pickup patterns - i.e. they are less precise than higher

[10] While the $M = 1$ and $M = 5$ renders are dual-band, the $M = 36$ is basic weighted for the entire frequency spectrum.

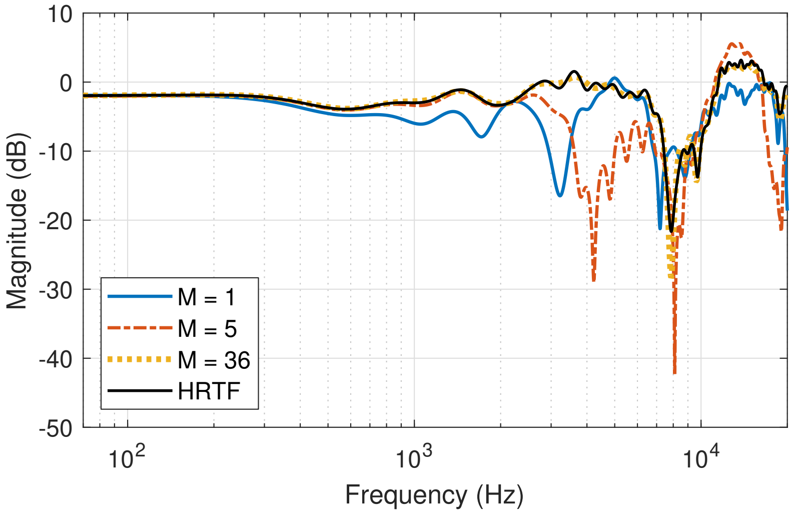


FIGURE 3.17: Binaural Ambisonic renders at varying Ambisonic order, with HRTF as reference, $(\theta = 0°, \phi = 0°)$ (left ear).

Ambisonic orders. However, this is not an issue at low frequencies, due to the lack of directionality at such low frequencies, as evidenced by the accurate reproduction of such frequencies at low orders of Ambisonics. It is only at high frequencies, being more directional and with narrower source widths, that low orders of Ambisonics deteriorate. With higher orders of Ambisonics and therefore narrower virtual microphone pickup patterns, combined with a denser distribution of loudspeakers around the sphere, high frequencies (with smaller wavelengths) can be reproduced more accurately in the centre of the array over an area the size of the head. The sweet spot is the area in the centre of the array where the reproduction is accurate. The size of the sweet spot depends on the spatial aliasing frequency, and vice versa. Depending on the order of Ambisonics, higher frequencies and positions further outside the sweet spot are more poorly reproduced (Wierstorf, Raake and Spors, 2013; Xie and Liu, 2014).

Above $f_{\text{alias}}$ or outside the sweet spot, spatial aliasing occurs due to the undersampling of the soundfield. One significant consequence of this is spectral interference and comb filtering (Jot, Larcher and Pernaux, 1999), which leads to timbral colouration (Yang and Bosun, 2015). At frequencies above the spatial aliasing frequency, loudspeakers that are close in position play out coherent signals that are summed at the ears. Consider the illustration in Figure 3.18. Though the human head is positioned in the centre of the array, the ears are not. Therefore, the path lengths

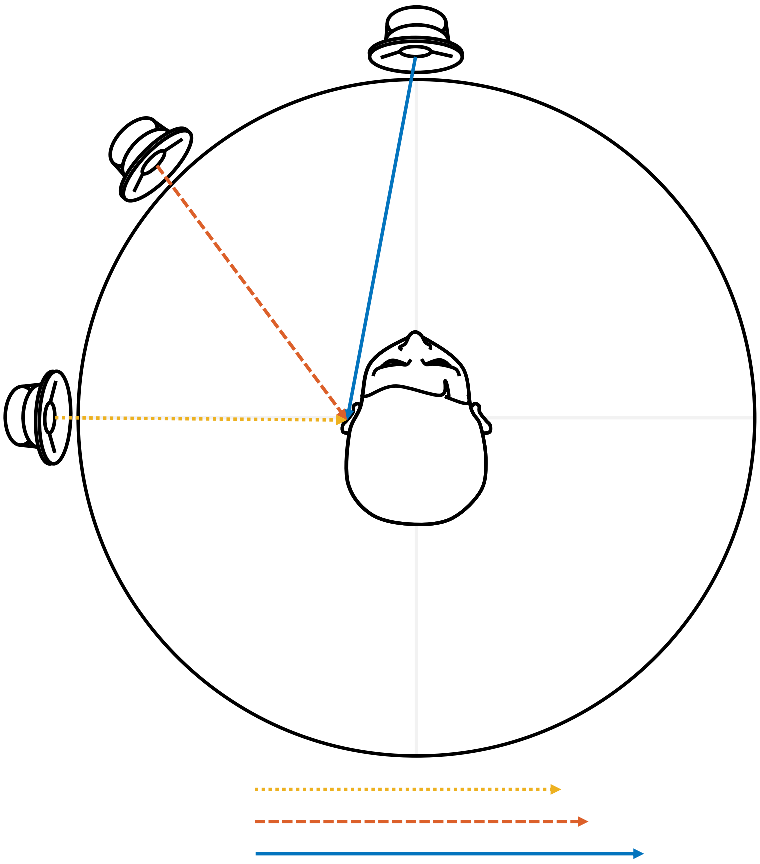


FIGURE 3.18: Illustration of the off-centre positions of the ears in head-centred Ambisonic rendering.

from the loudspeakers to the ears are not equal, and signals arrive at the ears at slightly different times. Additionally, the pinna filtering effects vary due to the differing angles of incidence. This explains why the effects are magnified when using a dense (virtual) loudspeaker array (Ben-hur et al., 2019), as demonstrated in Figure 3.16 with the significant reduction in amplitude at high frequencies.

Other issues that arise from spatial aliasing are the high frequency specific localisation cues, such as those used for determining source height. These are reproduced poorly at low-orders (Gorzel, Kearney and Boland, 2014; Millns, Mironovs and Lee, 2019), and improve at higher-orders (Kearney and Doyle, 2015b). Timbre between different loudspeaker layouts also varies substantially, even within the same Ambisonic order, which poses significant issues for content creators who desire a consistent timbre between different playback scenarios.

Spatial aliasing does not just produce comb filtering spectral artefacts. Other issues with Ambisonic reproduction at frequencies when the sweet spot is smaller than the human head include poor reproduction of ILD and ITD cues leading to reduced perception of lateralisation (Daniel, Rault and Polack, 1998; Daniel, 2000, p. 219). Localisation is also poorer at low Ambisonic orders (Bertet et al., 2007; Braun and Frank, 2011; Thresh, Armstrong and Kearney, 2017), and point sources appear more blurred (Bertet et al., 2013).

### 3.5.1 Improving High Frequency Reproduction

Many researchers have investigated ways of improving high frequency reproduction of Ambisonic rendering. This section focuses on existing techniques for specific improvements to binaural Ambisonic rendering, to which the work presented in this thesis aims to build on. One difference between loudspeaker reproduction and binaural reproduction of Ambisonic signals is that with binaural reproduction the left and right ear signals can be treated separately. This will be exploited in this thesis.

One attempt at improving lateralisation looked at introducing additional virtual loudspeakers at lateral directions (Collins, 2013). Though this is successful in improving lateralisation, it increases spectral artefacts (Yao, Collins and Jančovič, 2015) due to the higher number of loudspeakers.

Recording the HRTFs separately for the the left and right ear and centring the ear, as opposed to the centre of the head, has been investigated (Richter et al., 2014; Armstrong, Murphy and Kearney, 2018). This reduces some of the timbral issues as the sweet spot can be smaller, allowing for improved reproduction at high frequencies. However, a dual-band approach is necessary with a third measurement with the head in the centre of the array, in order to retain ITD information. Therefore, with specialised measurements necessary and triple the amount needed, this will not be investigated further in this thesis.

Methods for improving the reproduction when using order truncated binaural Ambisonic rendering have been developed, such as techniques for pre-processing the HRTFs (Brinkmann and Weinzierl, 2018). These include equalisation (Ben-Hur et al., 2017), which uses high frequency shelf boosts to negate the roll off that occurs from using a large amount of HRTFs at low Ambisonic orders. However, the required shelf filter is different for each order of Ambisonic reproduction, which negates some of the benefits of using a single decoder for all Ambisonic orders (as would be the case if implementing Max $\mathbf{r_E}$ SH weightings). Time-alignment (Evans, Angus and Tew, 1998; Richter et al., 2014; Zaunschirm, Schörkhuber and Höldrich, 2018) is the removal of ITDs between HRTFs above a certain frequency, which reduces comb filtering. However the cut-off frequency must be made sufficiently high to avoid removal of ITDs that are within the frequency range of being perceptually noticeable. A development of the time alignment technique exists in magnitude least squares (Schörkhuber, Zaunschirm and Höldrich, 2018), which attempts to remove not just time differences but all phase information above a certain frequency, further reducing spectral artefacts. Domain tapering uses Hanning amplitude windows on the higher SH order components to reduce the effects of order truncation (Hold et al., 2019), coupled with a high frequency shelf boosting filter similar to that in Ben-Hur et al. (2017). This follows the idea that, as order truncation sharply cuts off higher SH order components, a smoother transition from the SH orders used to those removed should produce improved results. This technique is not entirely dissimilar to the principles of Max $\mathbf{r_E}$ SH channel weighting, in that higher SH order channels are reduced in amplitude.

Other recent developments have included direction-dependent, or parametric, Ambisonic decoding strategies based on methods such as Directional Audio Coding (DirAC) (Pulkki, 2006; Pulkki, 2007), which analyses input signals in each frequency bin to separate out the diffuse and non-diffuse (directional) parts (Berge and Barrett, 2010a; Berge and Barrett, 2010b; Wabnitz, Epain and Jin, 2012; Politis, Vilkamo and Pulkki, 2015; Politis, McCormack and Pulkki, 2017; Politis, Tervo and Pulkki, 2018; Lecomte et al., 2018; McCormack and Politis, 2019; Schörkhuber and Höldrich, 2019; Giller and Schorkhuber, 2019). However, using a parametric decoding method requires additional computational cycles in the direction estimation. This thesis focuses on linear Ambisonic decoding methods.

## 3.6 Summary

This chapter has introduced Ambisonics, a technology for 3D full sphere spatial audio encoding and reproduction using spherical harmonics. The history and motivation for the development of Ambisonics has been discussed, including early encoding and decoding strategies. Higher-order Ambisonics has then been introduced, which offers more accurate reproduction over a greater area in the centre of the array to a higher frequency, but requires more microphone capsules in recording and more loudspeakers in reproduction. Psychoacoustic optimisations of decoding strategies have then been presented, such as dual-band decoding for improved rendering above the spatial aliasing frequency.

The application of Ambisonic technologies to binaural rendering has then been introduced, which allows for low channel 3D spatial audio reproduction over headphones, capable of soundfield rotation to counter head orientation changes, in order to facilitate real-time dynamic binaural synthesis. The issues of low-order binaural Ambisonic rendering have been discussed, including spatial aliasing due to the undersampling of a physical soundfield, resulting in high frequency spectral artefacts, poor localisation and reduced interaural cues. Finally, previous attempts to reduce the artefacts arising due to spatial aliasing are presented, which form the basis of the work to be presented in the rest of this thesis.

# Chapter 4

# Ambisonic Diffuse-Field Equalisation

As shown in Section 3.5, timbral inconsistencies exist in Ambisonic rendering above $f_{\text{alias}}$ due to spatial aliasing, which produces comb filtering from the summation of coherent loudspeaker signals with multiple delay paths to the ears, which are not situated at the exact centre of the loudspeaker array. By increasing the order of Ambisonics, which requires more microphones and encoded channels in production and storage and more loudspeakers in reproduction, $f_{\text{alias}}$ rises, improving both localisation and timbre, though for all practical Ambisonic rendering systems at present $f_{\text{alias}}$ is still much within the human hearing range. As spectral changes are the biggest differentiating factor between simulation and reality (Lindau, Hohn and Weinzierl, 2007), timbre is a vital consideration for binaural reproduction.

In this chapter, Diffuse-Field Equalisation (DFE) is applied to binaural Ambisonic rendering. An approximate diffuse-field response is typically calculated from the root-mean-square (RMS) of the magnitude responses of a large number of free-field measurements (Heller and Benjamin, 2012), and DFE is the removal of the direction-independent aspect of a set of frequency responses measured at many evenly distributed positions on a sphere, and is a technique often employed in HRTF databases. It has been discussed in Section 2.8.1 and implemented for the reference dataset of HRTFs from the Bernschütz Neumann KU 100 database (Bernschütz, 2013).

DFE has been implemented for SH order-truncated Ambisonic signals (Sheaffer, Villeval and Rafaely, 2014; Sheaffer and Rafaely, 2014; Ben-Hur et al., 2017), for both spherical head models and HRTF based filters, including the generation of simple shelving boosts to address the high frequency roll off of SH order-truncation. However, the approach taken in this thesis is to use an average of a large number of binaural Ambisonic renders made for all directions over the sphere. In particular, this chapter applies DFE to virtual loudspeaker binaural Ambisonic rendering, where the high frequency inconsistencies of Ambisonic rendering, though comparably lower in magnitude than SH order-truncation, are more complicated and cannot be characterised as simply a high frequency roll off. The method of sphere sampling, as presented in this chapter, also presents opportunities for further development such as directional biasing of the diffuse-field response to allow greater equalisation for a specific direction, as will be explored in Chapter 5.

The novel method of Ambisonic DFE used in this thesis is explained in detail in this chapter, with attention paid to the diffuse-field response calculation including the number of points used in sphere sampling and the method of distribution of the points. Ambisonic DFE can be applied to the SH binaural Ambisonic decoder in an offline process, thus producing no increase in real-time computational cost. Ambisonic DFE is then rigorously evaluated both numerically and perceptually. For the timbral evaluation of binaural signals, a new perceptually motivated model is presented which is designed to more accurately reflect human assessment of timbre than a basic spectral difference calculation. This is referred to as the perceptual spectral difference model. Numerical evaluation of Ambisonic DFE compares binaural Ambisonic rendering to a reference set of HRTFs in terms of perceptual spectral difference, estimated interaural cue similarity and predicted localisation accuracy. The applicability of DFE to other loudspeaker configurations and individualised HRTFs is also explored. Perceptual evaluation is carried out in the form of two listening tests. The first compares binaural Ambisonic rendering to HRTF convolution, both with and without DFE in terms of timbral similarity, and the second assesses the timbral consistency of different Ambisonic orders with and without DFE, to evaluate whether DFE improves the timbral consistency between different Ambisonic orders. The chapter ends with a summary of the findings and a recommendation on whether or not Ambisonic DFE should be implemented in binaural Ambisonic rendering.

## 4.1 Method

This section describes the method used for obtaining and equalising the diffuse-field response of a binaural Ambisonic decoder. In this thesis, diffuse-field equalisation filters are generated independently for left and right ears, due to the inherent asymmetry that exists between pinnae (especially in individualised HRTFs). This is also necessary for a second development of the technique that is explored further in Chapter 5, whereby the filters must be separate for left and right ears. A block diagram of the method is presented in Figure 4.1, and a brief summary of the method is as follows: binaural Ambisonic rendered HRTFs are generated for directions all over the sphere, an average of which is then obtained which gives a binaural impulse response that contains the direction-independent aspects of the binaural Ambisonic decoder. Equalising this using inverse filtering techniques, and convolving each channel of the binaural Ambisonic decoder with the calculated inverse filters, produces a diffuse-field equalised binaural Ambisonic decoder. All computation was carried out offline in Matlab version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4.

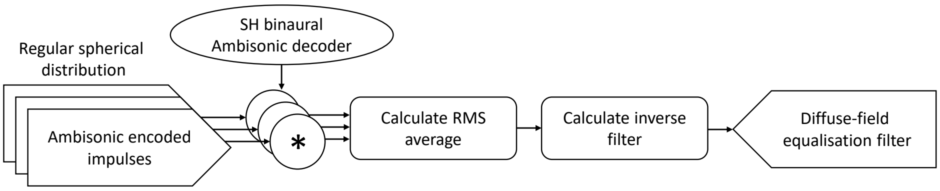


FIGURE 4.1: Block diagram of the Ambisonic DFE method.

### 4.1.1 Ambisonic Diffuse-Field Response

The Ambisonic diffuse-field response, $H_{\text{diff}}$, of the SH binaural Ambisonic decoder $\mathbf{D}_k^{\text{SH}}$ can be calculated from a sum of the RMS of the spherical harmonic channels, a process referred to in this thesis as numerical integration, as described in Sheaffer, Villeval and Rafaely (2014), Sheaffer and Rafaely (2014), and Ben-Hur et al. (2017), such that

$$H_{\text{diff}} = \sum_{k=1}^{K} |\mathbf{D}_k^{\text{SH}}|^2 \tag{4.1}$$

However, in this thesis an alternative approach is taken through spatial sampling of the sphere, so that further developments could be explored, such as directional bias in the diffuse-field response for localised spectral improvements, as will be explored in Chapter 5.

For the spatial sampling approach, the Ambisonic diffuse-field response $H_{\text{diff}}$ is calculated as follows. Ambisonic HRTFs are rendered as in (3.32), using $\delta(t)$ from (2.5) as the monophonic input signal, for virtually panned source locations on the sphere, denoted using $\iota$, of which the total is given by $\varrho$. The diffuse-field response of the binaural Ambisonic decoder is then obtained from the root-mean-square (RMS) average of the Ambisonic HRTFs, separately for the left and right ears, as

$$H_{\text{diff}} = \sqrt{\sum_{\iota=1}^{\varrho} \Omega_\iota H_\iota(\theta,\phi)^2} \tag{4.2}$$

where $H(\theta, \phi)$ is a single HRTF, and $\Omega$ is the solid angle weight of the measurement[1]. The calculation is performed in the frequency domain.

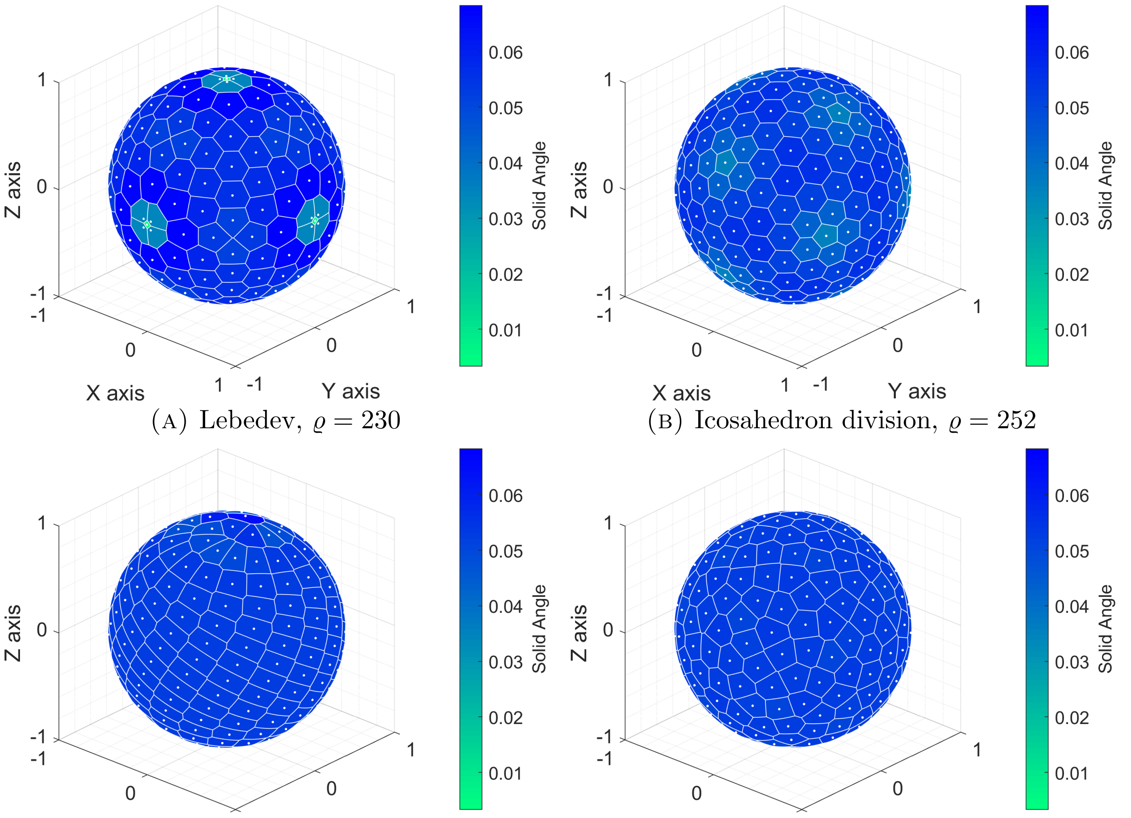
It is necessary to determine the optimal quadrature method and number of points necessary to produce an adequate approximate diffuse-field response. This is to ensure the calculated response is as accurate a representation of the diffuse-field response of the binaural Ambisonic decoder as possible. To do this, four quadrature methods are investigated with varying $\varrho$ by rendering approximate diffuse-field responses.

The four quadrature methods investigated for the distribution of points on a sphere are the Lebedev configuration (Lebedev, 1976), Icosahedron division (Burkardt, 2013a), Fibonacci spiral (Saff and Kuijlaars, 1997) and spherical T-design (Hardin and Sloane, 1996). Voronoi sphere plots of the quadrature methods with a similar value of $\varrho$ are shown in Figure 4.2 to compare the regularity of the quadrature methods. As the maximum possible value of $\varrho$ for T-design quadrature is $\varrho = 240$, the other quadrature methods were employed with similar[2] values of $\varrho$. The plots show T-design quadrature produces the highest regularity of the four methods.

Simulated diffuse-field responses at $M = 5$ using a $L = 50$ Lebedev loudspeaker configuration for the four quadrature methods are shown in Figure 4.3. Due to the HRTF diffuse-field equalisation of the reference HRTF dataset (as detailed in Section 2.8.1), the calculated Ambisonic diffuse-field responses are highly flat up to 3 kHz, approaching the $f_{\text{alias}}$ of $M = 5$. This illustrates the accuracy of low frequency reproduction in Ambisonic rendering as well as emphasises the need for Ambisonic diffuse-field equalisation above $f_{\text{alias}}$. The calculated Ambisonic diffuse-field responses differ by up to ± 0.5 dB at some frequencies without solid angle weighting, however implementation of solid angle weighting brings the variation to below ± 0.1 dB at all frequencies. Therefore, providing solid angle weighting is implemented, the quadrature method need not be highly regular.

[1] This equation uses different variable names for virtually panned source locations to those used in the HRTF diffuse-field response calculation in (2.10) to distinguish between Ambisonic DFE and standard HRTF DFE.

[2] Depending on the quadrature method, the exact same values of $\varrho$ are not obtainable.



(A) Lebedev, $\varrho = 230$

(B) Icosahedron division, $\varrho = 252$

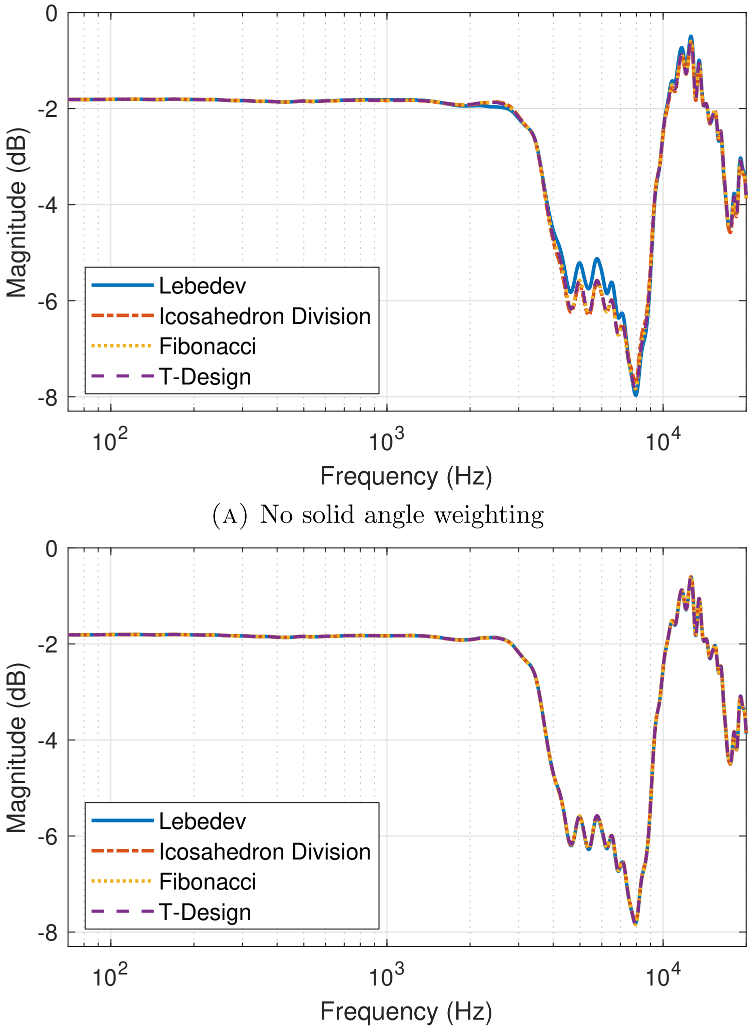
(C) Fibonacci, $\varrho = 240$

(D) T-design, $\varrho = 240$

FIGURE 4.2: Voronoi sphere plots demonstrating the regularity in spherical distribution of points for four quadrature methods.

The minimum number of measurements necessary to calculate a sufficient approximation of the diffuse-field is investigated by rendering diffuse-field responses with a varying number of measurements. The number of measurements ranged from $\varrho = L$ to $\varrho = 11L$ in intervals of 5 using Fibonacci quadrature, for $M = 5$. The calculated diffuse-field responses, with spectral difference between the response of each value of $\varrho$, are presented in Figure 4.4. With solid angle weighting implementation, variation in calculated diffuse-field response is less than ± 0.01 dB at all frequencies when $\varrho > 4L$.

To summarise, as T-designs are the most regular tested quadrature, the $\varrho = 240$ T-design configuration is used as the spherical distribution of points for calculating Ambisonic diffuse-field responses for the remainder of this thesis, unless otherwise stated, to ensure minimal error between the numerical integration method. However,

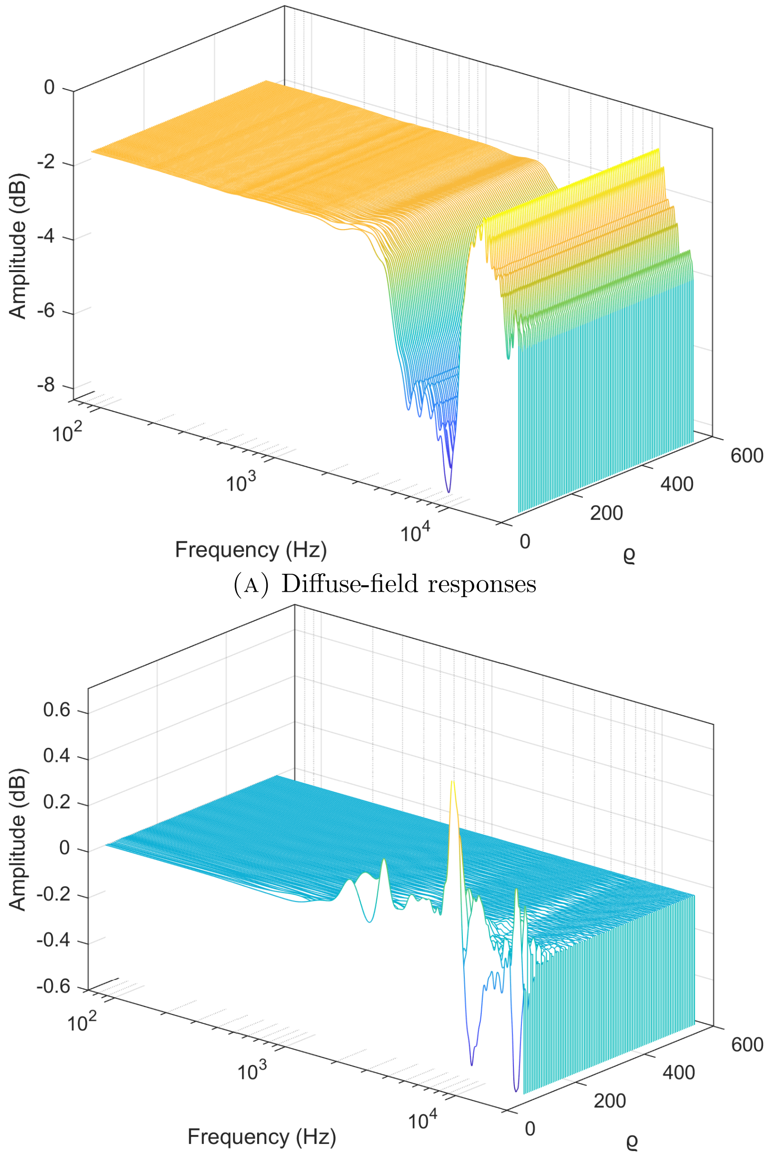


(A) No solid angle weighting

(B) With solid angle weighting

Figure 4.3: Diffuse-field responses of $M = 5$ using different quadrature methods, with and without solid angle weighting (left ear).

in the case $4L > 240$ (the largest possible T-design), it is recommended to use Fibonacci quadrature with $\varrho > 4L$ due to this being the second most regular quadrature method tested, capable of any value of $\varrho$. Solid angle weighting should be implemented in the diffuse-field response calculation which further improves the accuracy of the average calculation over the sphere.



(A) Diffuse-field responses

(B) Spectral difference

FIGURE 4.4: Diffuse-field responses of $M = 5$ calculated using Fibonacci quadrature with varying values of $\varrho$.

### 4.1.2 Inverse Filter Calculation

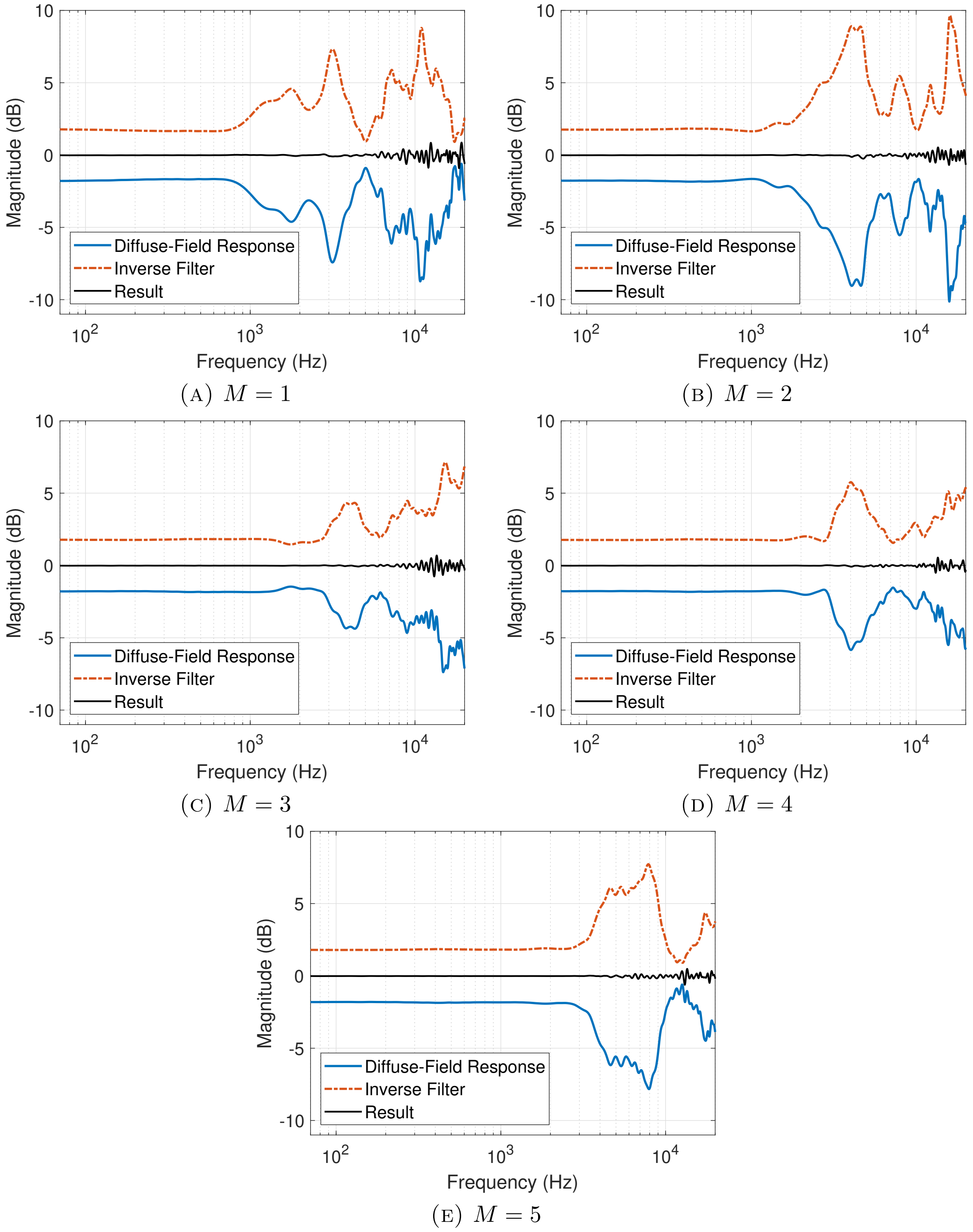
To equalise the calculated diffuse-field response of the binaural Ambisonic decoder, linear-phase inverse filters are produced using Kirkeby and Nelson's least-mean-square regularisation method (Kirkeby and Nelson, 1999), which produces perceptually preferred inversions to other currently available methods (Schärer and Lindau, 2009). 1/4 octave smoothing is implemented using the complex smoothing approach of (Hatziantoniou and Mourjopoulos, 2000), and the range of inversion is 2 Hz - 20 kHz, with in-band and out-band regularisation of 25 dB and 5 dB, respectively. For greater detail on the inverse filtering methods used, refer back to Section 2.8.1.

The diffuse-field responses, inverse filters and resulting equalised frequency responses (calculated by convolving the diffuse-field response with the inverse filter) of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using Lebedev loudspeaker configurations (the vertices of which are presented in Figure 3.4) are presented in Figure 4.5. The plots show how the diffuse-field responses of binaural Ambisonic loudspeaker configurations are highly even (within ±0.3 dB) up to $f_{\text{alias}}$, above which the frequency responses vary significantly for all orders, with deviations as large as 10 dB at some frequencies.

The implementation of the DFE filters is achieved offline, through convolution of each channel of the SH binaural Ambisonic decoder with the calculated inverse filter, for both left and right ears. With truncation and subsequent 50 sample half-Hanning windowing of the processed HRTFs, the resulting SH binaural Ambisonic decoders are the same sample size as before. Alternatively, the HRTFs can be convolved with the DFE filters separately, if desired.

To assess the effect of applying Ambisonic DFE filters to a SH binaural Ambisonic decoder in greater detail, Ambisonic DFE was calculated and applied to the $L = 6$ configuration for $M = 1$. Figure 4.6 shows the time-domain response of the left and right $k = 1$ channel, both without any pre-processing and with Ambisonic DFE. The X axis is zoomed into the middle 200 samples. It is clear that, after truncation, the Ambisonic DFE filtering does not produce an effect on the temporal nature of the resulting signals.

To assess whether the need for Ambisonic DFE applies to binaural Ambisonic rendering made using other HRTF datasets, the diffuse-field response calculations were repeated for $M = 5$ using the 18 human datasets of individualised HRTFs from the SADIE II database (Armstrong et al., 2018a). The calculated diffuse-field responses are presented in Figure 4.7. Note that the SADIE HRTF dataset has not been diffuse-field equalised using the technique detailed in Section 2.8.1 (which is the technique used for the Bernschütz HRTF dataset), which leads to low frequency



(A) $M = 1$

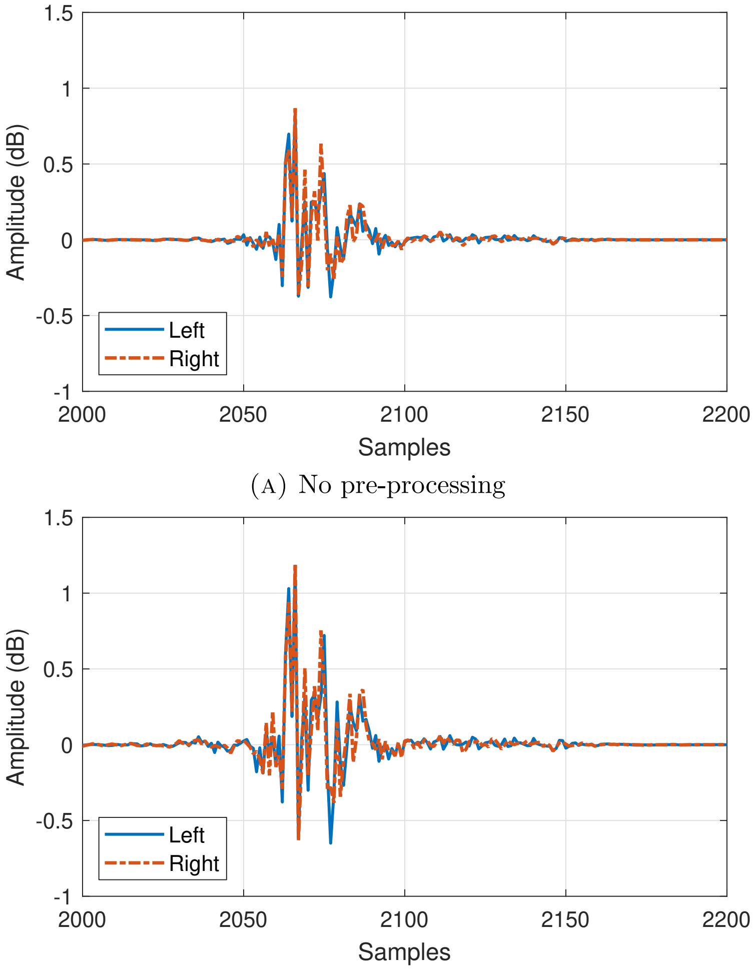
(B) $M = 2$

(C) $M = 3$

(D) $M = 4$

(E) $M = 5$

FIGURE 4.5: Diffuse-field response, inverse filters and resulting responses of the Lebedev loudspeaker configurations for $\{M = 1, M = 2, ..., M = 5\}$ (left ear).



(A) No pre-processing

(B) With Ambisonic DFE

FIGURE 4.6: Time-domain response of a SH binaural Ambisonic decoder without and with Ambisonic DFE filtering, for the $L = 6$ configuration at $M = 1$, $k = 1$ channel. X axis zoomed to between 2000 and 2200 samples.

deviations in the Ambisonic diffuse-field responses, that do not occur when using the Bernschütz HRTF dataset. As wide-band spectral variations up to ±8 dB occur, it is clear that Ambisonic DFE is still necessary when using individualised HRTFs. An observation is that there is a definite trend in the diffuse-field responses of the Lebedev loudspeaker configuration at $M = 5$ between different individuals, as highlighted by the average line. This trend is also loosely present in Figure 4.5e as a wideband notch around the region of 4 kHz, which suggests the arrangement of loudspeakers and Ambisonic order influence the Ambisonic diffuse-field response. This offers potential for future work in creating a generalised DFE filter for a specified loudspeaker arrangement and Ambisonic order, irregardless of whether reproduced

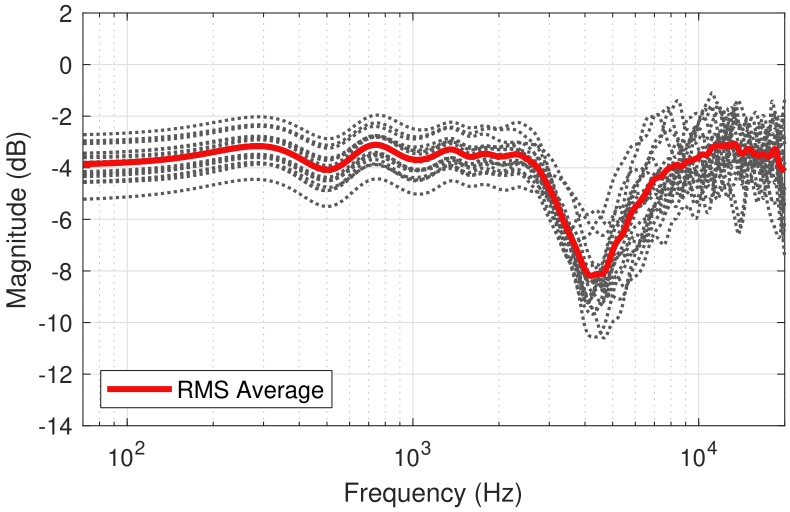


FIGURE 4.7: Diffuse-field responses of the $M = 5$ Lebedev loudspeaker configurations for the 18 human subjects of the SADIE II database, with average of all responses (left ear).

binaurally or over loudspeakers.

## 4.2 A Perceptually Motivated Binaural Spectral Difference Model

To evaluate the timbre of binaural Ambisonic rendering using HRTF pre-processing techniques more effectively than a basic spectral difference (BSD) calculation, a perceptually motivated model for judging the spectral difference between two datasets of audio signals has been developed in MATLAB. This is herein called the Perceptual Spectral Difference (PSD) model. This model is designed to be appropriate for assessing the perceptual difference between binaural signals by accounting for binaural cues such as ILDs by weighting louder signals with greater relevance. It takes inspiration from features present in ITU-T recommendation P.862 (Rix et al., 2001), as well as PEAQ (Thiede et al., 2000) and PSQM (Beerends and Stemerdink, 1994). Other previous spectral difference models exist (Wang, Sekey and Gersho, 1992; Moore, Glasberg and Baer, 1997; Pulkki et al., 1999), though this model has a number of different features that will be explained below. A block diagram of the

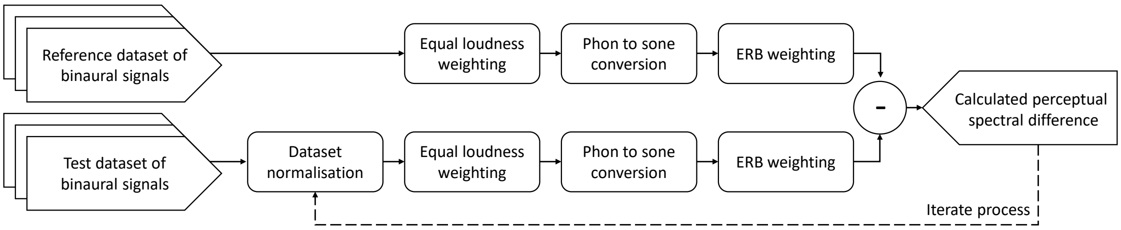


Figure 4.8: Block diagram of the PSD method.

PSD method is presented in Figure 4.8. The three main features that differentiate the PSD model from a BSD calculation are the frequency-varying amplitude weighting, the relative loudness amplitude weighting and accounting for the frequency spacing of the FFT operation. The model also uses solid angle weightings when normalising the two input datasets to the mean amplitude of the two datasets. Differences between the PSD and the Composite Loudness Level (CLL) are that, while both use ERB weightings and Phon calculations (Pulkki et al., 1999; Ono, Pulkki and Karjalainen, 2001; Ono, Pulkki and Karjalainen, 2002), the PSD also includes ISO 226 equal loudness weighting and sone amplitude weighting.

The perceptual spectral difference between two datasets of binaural signals is calculated as follows. Firstly, the two datasets must have the same number of signals and dimensions. An FFT of the time-domain audio signals is taken with a number of frequency bins of the input signal length. The amplitude data of the FFT calculations are converted into relative dB using (2.3), with no reference pressure used in this case. The amplitude values of each frequency bin are then weighted according to inverse equal loudness contours using the ISO 226 standard (International Organization for Standardization, 2003). This converts the amplitude data from dB to the Phon scale, and accounts for the frequency-varying sensitivity of human hearing in which the most sensitive frequency range lies between approximately 1 kHz and 5 kHz, with sensitivity decreasing outside this range. Therefore, frequencies where the human auditory system is less sensitive are weighted lower, and vice-versa. This approach differs from previous models, which use a single equal loudness contour filter based on the threshold of hearing (Moore and Glasberg, 1996; Moore, Glasberg and Baer, 1997), by utilising 90 magnitude dependent equal loudness contours in

1 dB increments from 0 to 90 dB SPL whereby the closest equal loudness contour is selected and used according to the magnitude of each frequency bin of the input signal. The playback level of the reference dataset is assumed to be 75 dB SPL, in line with a commonly reported typical listening level (Katz and Parseihian, 2012; Hartmann and Rakerd, 1993; Hammershøi and Møller, 1996), though this value is adjustable if desired.

The magnitude value of each frequency bin is then converted from Phons to sones using (2.8). As the sone scale is based on human perception of loudness using the approximate ratio of +10 phons per doubling of perceived loudness (Stevens, 1955; Bauer and Torick, 1966), this therefore accounts for human auditory system features such as spectral peaks being more perceptually significant than notches (Bücklein, 1981), and louder sounds carrying greater relative importance. This means that when calculating the spectral difference of binaural signals, the louder signal of the two ears (usually from the ipsilateral side) is therefore weighted with higher relevance, which is in accordance with Morimoto (Morimoto, 2001).

The FFT calculation samples a time-domain signal at linearly spaced frequency intervals. This is not a fair representation of the approximately logarithmic sensitivity of the cochlea, so the magnitude value of each frequency bin is weighted according to its equivalent rectangular bandwidth (ERB) (Moore and Glasberg, 1983) using (2.13). A single value of PSD between each signal in the test and reference datasets is calculated as the mean difference between the weighted amplitude values of each frequency bin.

In real-world listening comparisons, the loudness of the two systems can be adjusted to give a relative equal loudness. Therefore, in the PSD model, a normalisation stage has been included so the test dataset can be normalised relative to the reference dataset through an iterative process which finds the gain necessary to produce the lowest overall PSD value across all measurements. This is achieved by repeating the difference calculation between the two datasets with varying normalisation of the test dataset in order to find the optimal normalisation of the test dataset to produce

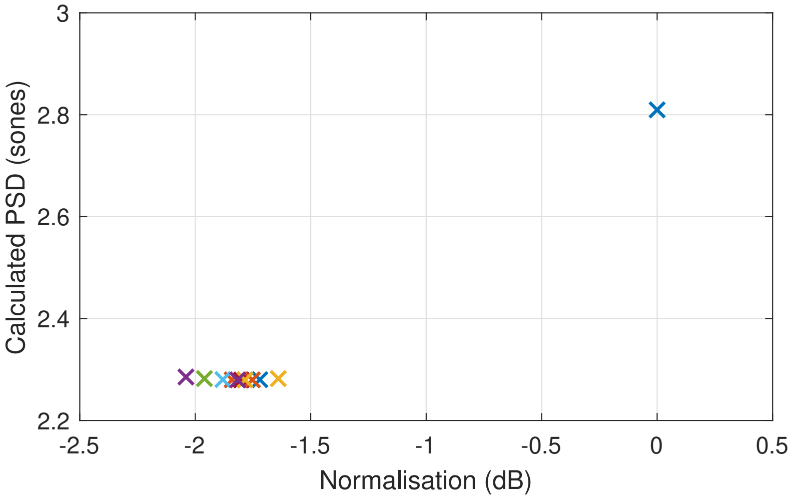


FIGURE 4.9: Iterative normalisation of PSD model test dataset. The first normalisation value is 0 dB.

the lowest overall PSD result. Figure 4.9 demonstrates the normalisation. The first iteration has no normalisation (a gain value of 0), and the subsequent calculations focus in on the normalisation gain that produces the lowest overall PSD result. The normalisation process uses solid angle weightings in calculating the overall PSD result between the two datasets. The resolution of normalisation is adjustable, or it can be turned off. The iterative process stops when a change in normalisation produces a change in calculated PSD lower than a specified value, which in this thesis is 0.01 dB as it a variation in calculated PSD within 0.01 sones (see again Figure 4.9).

### 4.2.1 Validation

The PSD model has been validated in two ways: firstly by comparison to a standard basic spectral difference calculation using specifically designed test signals produced from filtered impulses to demonstrate the various features of the model, and secondly by using the stimuli and results of a previously conducted perceptual listening test to demonstrate how the model correlates with real listening test scenarios.

Four test scenarios were created to demonstrate separate features of the model, using test signals created by passing an impulse, calculated using $\delta(t)$ from (2.5), through

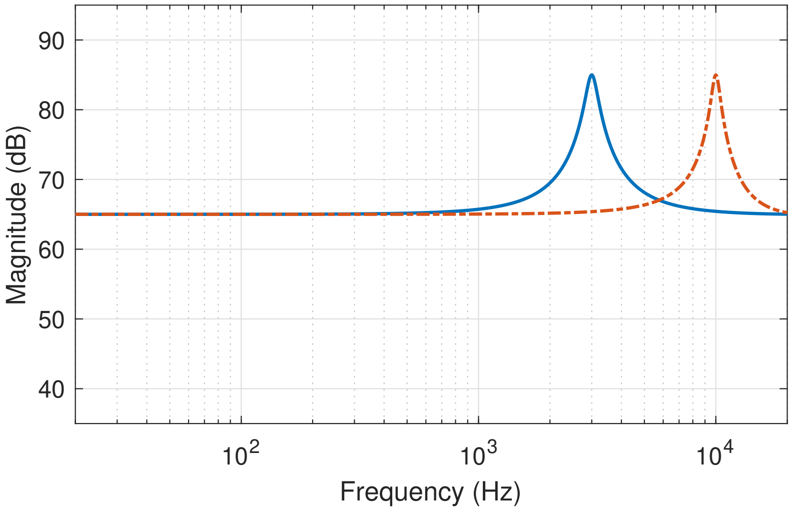


Figure 4.10: Frequency response of the 3 kHz and 10 kHz +20 dB peak filters with equal ERB filter bandwidths.

Table 4.1: Results of comparing the 3 kHz and 10 kHz +20 dB peak filtered signals with equal filter bandwidth at 65 dB SPL to flat response reference signals at the same level.

| Feature | 3 kHz | 10 kHz |
|---|---|---|
| BSD (dB) | 1.87 | 3.90 |
| PSD (sones) | 1.53 | 0.57 |

a filter. In each scenario, the filtered signals were compared to a flat frequency response signal of the same dB SPL level to compute the spectral difference. This was done using both a BSD calculation (the mean difference between each frequency bin of two FFT calculations) and the PSD model.

### Scenario 1: Equal Loudness

To demonstrate the use of ISO 226 equal loudness curves, two signals with +20 dB peaks at 3 kHz and 10 kHz at 65 dB SPL were compared to flat reference signals of the same level (see Figure 4.10 and Table 4.1).

Referring back to the equal loudness plots in Figure 2.14, 10 kHz is, in perceptual terms, a less sensitive frequency than 3 kHz. Therefore, whereas the PSD model

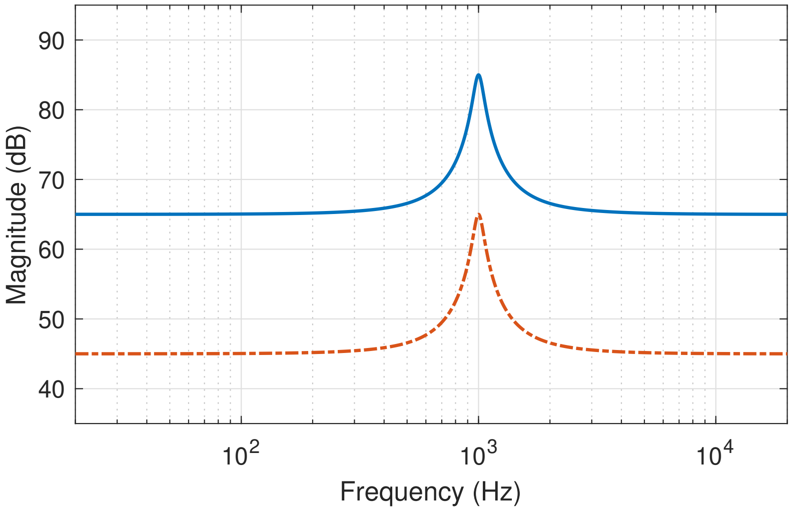


Figure 4.11: Frequency response of the 1 kHz +20 dB peak filters at 65 dB SPL and 45 dB SPL.

Table 4.2: Results of comparing the 1 kHz +20 dB peak filtered signals at 65 dB SPL and 45 dB SPL to flat response reference signals of the same respective levels.

| Feature | 0 dB | −20 dB |
|---|---|---|
| BSD (dB) | 0.59 | 0.59 |
| PSD (sones) | 1.06 | 0.26 |

produces a lower value of difference for the 10 kHz peak, the BSD calculation produces a higher value of difference. The PSD calculation is considered to be a closer approximation of human hearing.

### Scenario 2: Overall Loudness Difference

Secondly, to demonstrate the conversion from the Phon to sone scale, two comparisons were made. The first tests how a change in loudness at a lower amplitude is less perceptually noticeable than one at a higher amplitude by comparing two signals with 1 kHz +20 dB peaks at 65 dB SPL and 45 dB SPL to flat reference signals at the same respective levels (see Figure 4.11 and Table 4.2).

The second comparison assesses how a peak should be more noticeable than a notch (Bücklein, 1981) by comparing signals with a 1 kHz +20 dB peak and −20 dB notch

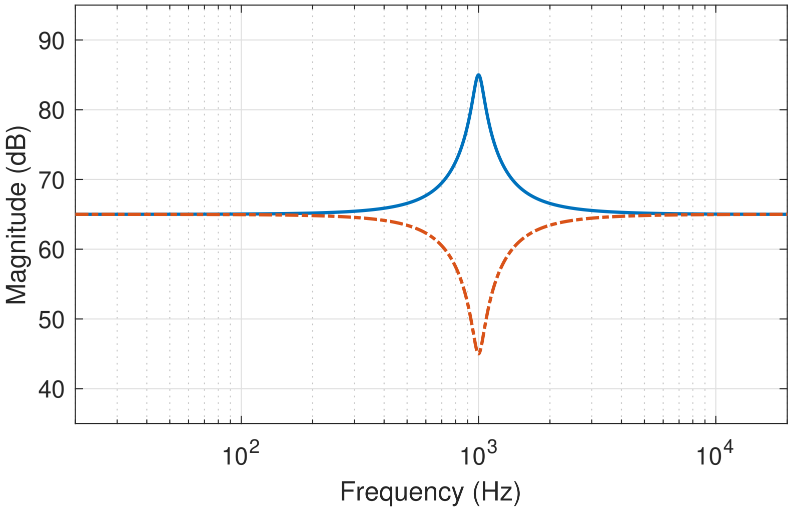


FIGURE 4.12: Frequency response of the 1 kHz +20 dB peak and −20 dB notch filters.

TABLE 4.3: Results of comparing the 1 kHz +20 dB peak and −20 dB notch filtered signals at 65 dB SPL to flat response reference signals of the same level.

| Feature | +20 dB | −20 dB |
|---|---|---|
| BSD (dB) | 0.59 | 0.59 |
| PSD (sones) | 1.06 | 0.53 |

at 65 dB SPL to flat reference signals at the same level (see Figure 4.12 and Table 4.3).

In both cases the BSD calculation produces the same value of spectral difference, whereas the PSD model produced results in line with what is expected from the human auditory system; the higher amplitude peak is rated as more perceptually different than the lower amplitude peak with the same characteristics, and the peak produces a greater PSD value than the notch.

### Scenario 3: Non-Linear Frequency Scaling

The third test scenario demonstrates the use of ERB weighting, which compensates for the linear frequency interval sampling of an FFT. To test this, two signals with +20 dB peaks at 1 kHz and 5.5 kHz, both with fixed 100 Hz −3 dB filter bandwidth,

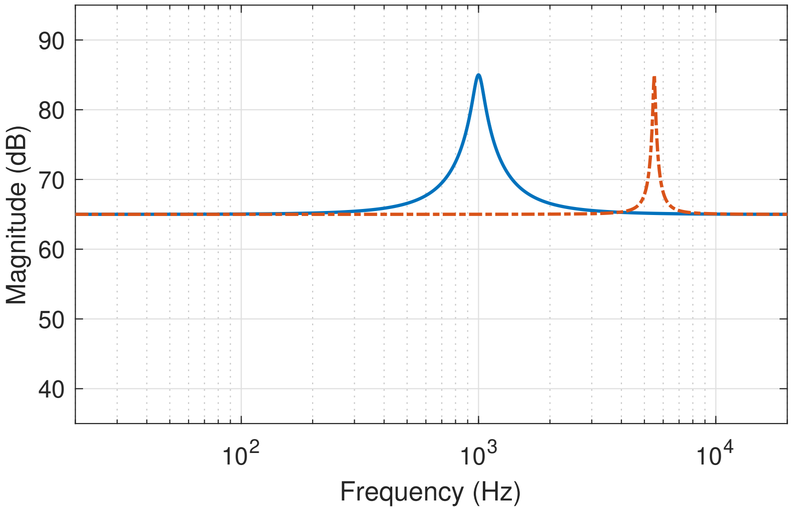


FIGURE 4.13: Frequency response of the 1 kHz and 5.5 kHz +20 dB peak filters of 100 Hz −3 dB filter bandwidths.

TABLE 4.4: Results of comparing the 1 kHz and 5.5 kHz +20 dB peak filtered signals with 100 Hz −3 dB filter bandwidths to flat response reference signals of the same level.

| Feature | 1 kHz | 5.5 kHz |
|---|---|---|
| BSD (dB) | 0.59 | 0.60 |
| PSD (sones) | 1.06 | 0.23 |

at 65 dB SPL level, were compared to flat reference signals at the same level (see Figure 4.13 and Table 4.4). The frequencies of 1 kHz and 5.5 kHz were chosen as these are frequencies at which the ear has approximately the same sensitivity (refer again to the equal loudness contours in Figure 2.14).

The perceptual relevance of the peak at 5.5 kHz should be less than one at 1 kHz as the peak is spread over fewer critical bands, but the BSD value shows little difference between the two signals due to the linear frequency interval sampling, whereas the PSD predicts a much greater difference for the wider perceptual bandwidth.

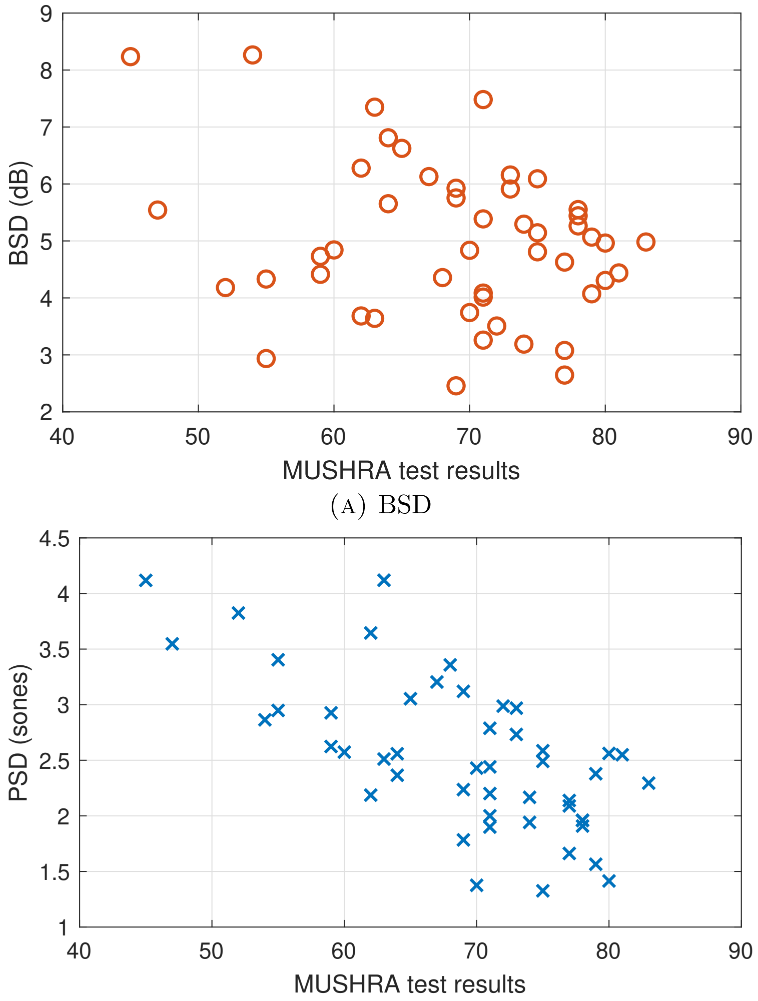
### Scenario 4: Perceptual Test Comparison

To further validate the PSD model, the perceptual results of a listening test on perceived timbral differences of binaural signals were compared to the PSD and BSD results of the spectral differences between the test stimuli and the reference stimuli. For specific details on the test paradigm, see Section 4.4.

The original listening test followed the MUSHRA paradigm and asked listeners to rate 6 different Ambisonic renders in timbral similarity to an HRTF reference over 8 different sound source directions, resulting in 48 separate conditions. Listener ratings were in the range of 0 - 100 with 100 being entirely the same as the reference signal and 0 being not at all similar. The test used 20 participants and every condition was repeated once resulting in 40 trials of each condition. A single value of perceived timbral similarity for each condition was taken as the mean average of the 40 trials.

The spectral difference was calculated between the stimuli, which were binaural stereo wav files of 1 second long, and the reference. This was done for both a BSD calculation and the PSD model. For each of the 48 conditions, the mean perceived timbral similarity between the 40 repetitions from the listening test is plotted against the calculated spectral difference value in Figure 4.14, for both BSD and PSD. A negative correlation between the spectral difference results and the perceptual results is visible in the PSD plot, with a trend that lower PSD values (indicating less difference) correspond with higher perceived similarity, whereas the BSD results appear somewhat less correlated and more spread out. The correlations between the listening test data and spectral difference calculations were tested using Pearson's correlation coefficient. Table 4.5 presents the correlation results, where $R$ denotes correlation and $p$ denotes statistical significance. The PSD correlation is highly statistically significant, whereas the BSD correlation is not statistically significant at a confidence of 95%.

The validation has shown how the PSD model presented produces a spectral difference for binaural signals that offers a closer representation of the human auditory system than a basic spectral difference calculation. The PSD calculation is therefore used throughout this thesis when assessing the spectral difference between datasets of binaural signals.



(A) BSD

(B) PSD

Figure 4.14: Comparing the MUSHRA test results from Section 4.4 to BSD and PSD calculations between the test stimuli and the references.

Table 4.5: Pearson correlation coefficient results comparing the MUSHRA test results from Section 4.4 to BSD and PSD calculations between the test stimuli and the references.

| Model | $R$ | $p$ |
|---|---|---|
| BSD | $-0.26$ | 0.07 |
| PSD | $-0.67$ | $< 0.01$ |

## 4.3 Numerical Evaluation

The effect of DFE was evaluated numerically by comparing binaural Ambisonic renders, with and without DFE, to a reference dataset of HRTFs. The metrics of evaluation include perceptual spectral difference, interaural cues and estimated horizontal and vertical localisation. This section also evaluates the effect of DFE when implemented in binaural Ambisonic rendering using different loudspeaker configurations and alternative human HRTFs.

In (Ben-Hur et al., 2017), the presented equalisation method is evaluated through comparison to unprocessed order-truncated binaural Ambisonic rendering with a dense HRTF configuration, which inherently produces much poorer spectral reproduction than unprocessed binaural Ambisonic rendering with sparse virtual loudspeaker configurations (Ben-hur et al., 2019). In this chapter, Ambisonic DFE is therefore evaluated using sparse virtual loudspeaker configurations, using the corresponding Lebedev and T-Design configurations stated in Figures 3.4 and 3.8, respectively.

For each measurement location in the reference dataset of $Q = 16{,}020$, as illustrated in Figure 2.28, binaural Ambisonic renders were generated for $\{M = 1, M = 2, ..., M = 5\}$, both with and without DFE. All HRTFs were truncated to 1024 taps with 50 sample in / out half-Hanning windows applied[3].

### 4.3.1 Perceptual Spectral Difference

The PSD between reference HRTFs and Ambisonic generated HRTFs, with and without DFE, was calculated for all measurement locations for each tested order, using the method detailed in Section 4.2. The solid angle weighted average value of PSD, $\overline{\text{PSD}}$, between reference HRTFs and Ambisonic HRTFs across all measurement locations on the sphere (a single value is calculated from the mean of the PSD calculations for the left and right ears) is calculated as

$$\overline{\text{PSD}} = \sum_{q=1}^{Q} \Omega_q \text{PSD}_q \tag{4.3}$$

[3] Note: exact values in the numerical evaluation in this chapter differ slightly from the published figures in McKenzie, Murphy and Kearney (2018) due to the use of $f_{\text{alias}}$ values from Table 3.3 throughout this thesis, as opposed to the values in Table 3.4 as used in McKenzie, Murphy and Kearney (2018).

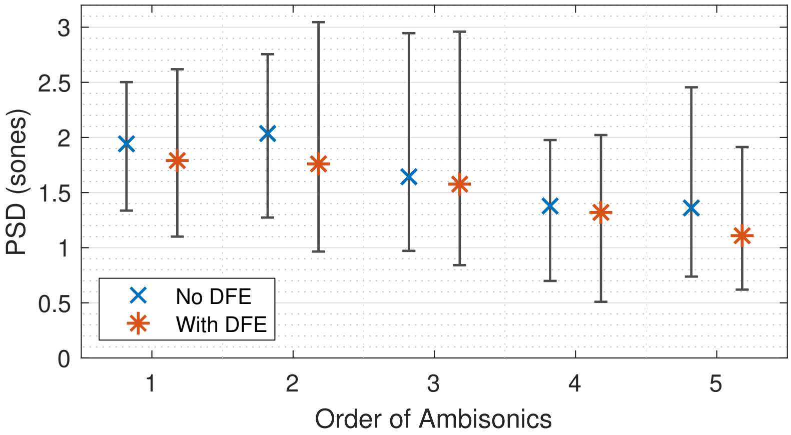
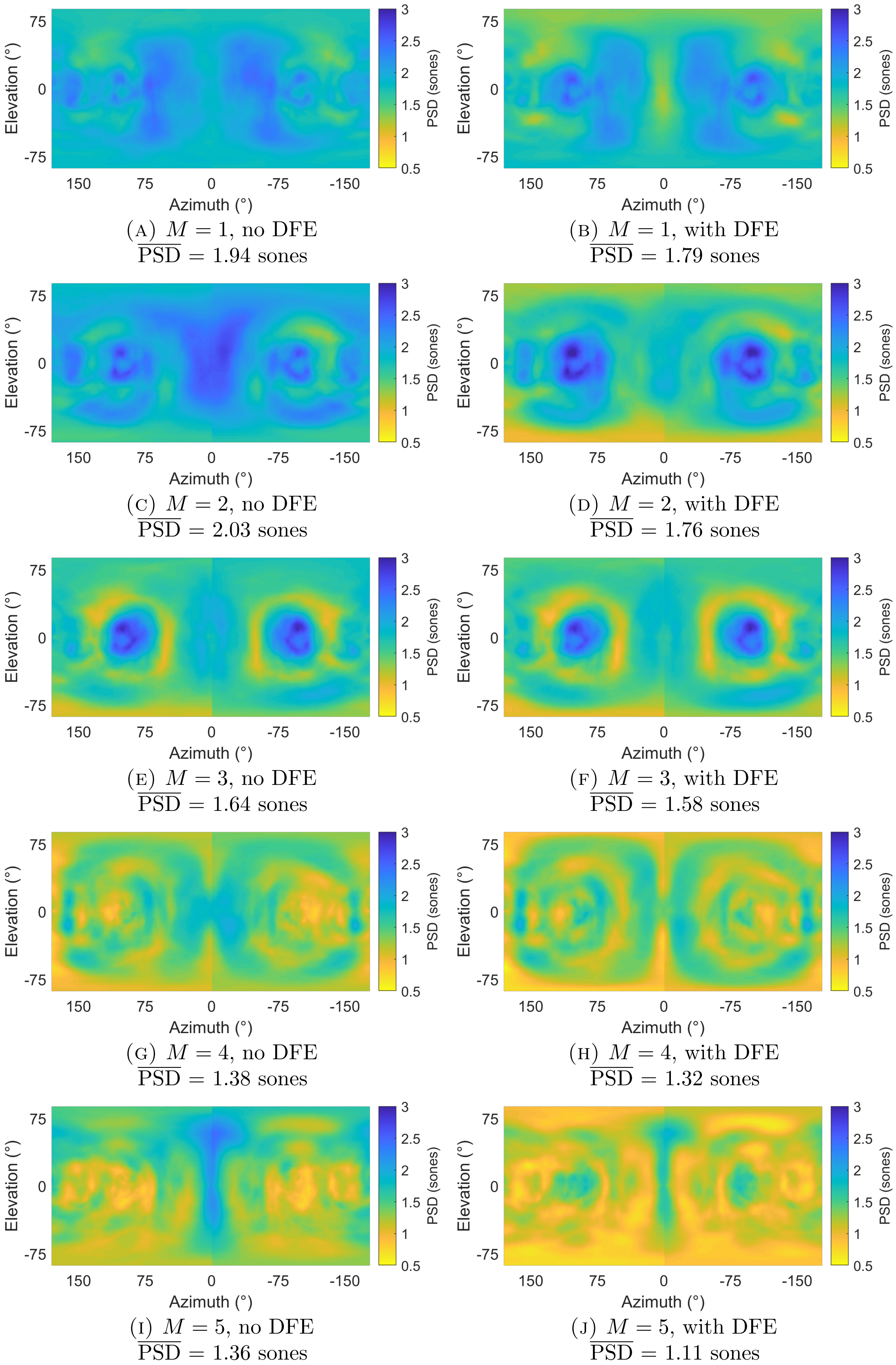


FIGURE 4.15: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the minimum and maximum absolute PSD values.

where $\Omega_q$ denotes the solid angle weight of the measurement location $q$. The $\overline{\text{PSD}}$ for $\{M = 1, M = 2, ..., M = 5\}$, with and without DFE, across all measurement locations, are presented in Figure 4.15, along with the minimum and maximum absolute PSD values.

DFE is observed to reduce $\overline{\text{PSD}}$ between Ambisonic HRTFs and reference HRTFs for all tested Ambisonic orders, as well as reduce the minimum value of PSD. This suggests that DFE improves the overall timbral reproduction in binaural Ambisonic rendering. However, it is also apparent that the maximum value of PSD increases for all tested Ambisonic orders except $M = 5$, suggesting that, though DFE improves the overall spectral reproduction, there may be locations at which spectral accuracy is reduced.

To investigate the directional effect of DFE on PSD, Figure 4.16 displays the absolute values of PSD between Ambisonic HRTFs and reference HRTFs with and without DFE for every measurement location on the sphere (mean of left and right ear PSD calculations). The plots show DFE implementation improves spectral reproduction for a large amount of the sphere, particularly for $M = 1, M = 2$ and $M = 5$, and in general, the greatest improvements appear closer to the median plane while the lateral directions are generally where PSD is higher.



(A) $M = 1$, no DFE
$\overline{\text{PSD}} = 1.94$ sones

(B) $M = 1$, with DFE
$\overline{\text{PSD}} = 1.79$ sones

(C) $M = 2$, no DFE
$\overline{\text{PSD}} = 2.03$ sones

(D) $M = 2$, with DFE
$\overline{\text{PSD}} = 1.76$ sones

(E) $M = 3$, no DFE
$\overline{\text{PSD}} = 1.64$ sones

(F) $M = 3$, with DFE
$\overline{\text{PSD}} = 1.58$ sones

(G) $M = 4$, no DFE
$\overline{\text{PSD}} = 1.38$ sones

(H) $M = 4$, with DFE
$\overline{\text{PSD}} = 1.32$ sones

(I) $M = 5$, no DFE
$\overline{\text{PSD}} = 1.36$ sones

(J) $M = 5$, with DFE
$\overline{\text{PSD}} = 1.11$ sones

FIGURE 4.16: PSD between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere (mean of left and right PSD values).

### 4.3.2 Interaural Cues

The DFE filters are generated independently for left and right ears, due to the asymmetry that exists between pinnae. However, as the DFE technique results in a single linear phase filter per ear, it should produce little directional effect on interaural cues, which are crucial for horizontal localisation. To assess this effect, both ITD and ILD were estimated using the methods detailed in Sections 2.8.3 and 2.8.2, respectively, for all measurement locations and Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$. The low-pass filter used in the ITD calculation was $f_c = 1.5$ kHz. The change in ITD between the reference HRTFs and the Ambisonic generated HRTFs was then calculated for each measurement location $q$ as

$$\Delta\text{ITD}_q = |\text{ITD}(H_q) - \text{ITD}(\hat{H}_q)| \tag{4.4}$$

where $H$ is the reference HRTF and $\hat{H}$ is the Ambisonic generated HRTF. Next, a single value of $\Delta$ITD for all locations on the sphere, $\overline{\Delta\text{ITD}}$, was calculated from the solid-angle weighted sum of all $\Delta$ITD values as

$$\overline{\Delta\text{ITD}} = \sum_{q=1}^{Q} \Omega_q \Delta\text{ITD}_q \tag{4.5}$$

Figure 4.17 displays the solid angle weighted $\overline{\Delta\text{ITD}}$ values between HRTFs and binaural Ambisonic rendering with and without DFE for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute $\Delta$ITD value. Detailed plots of $\Delta$ITD for every measurement location on the sphere are presented in Appendix A.1, which show the close similarity between renders with and without DFE, and an overall improvement in the reproduced ITD values with increased Ambisonic order.

The change in ILD between the reference HRTFs and the Ambisonic generated HRTFs, $\Delta$ILD, was calculated for each measurement location $q$ as

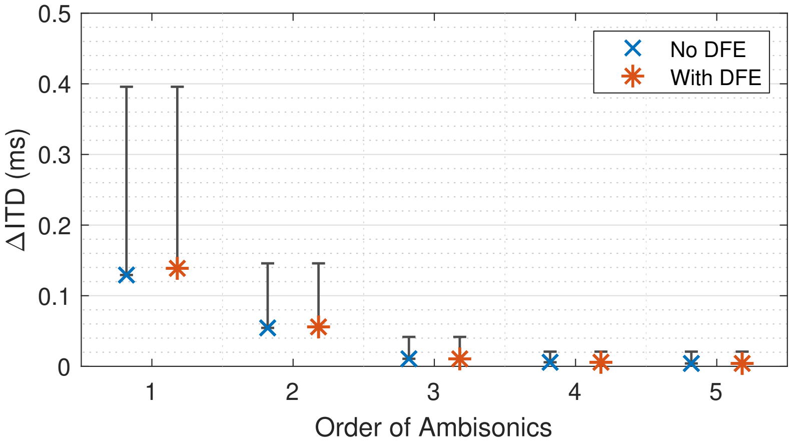


FIGURE 4.17: $\overline{\Delta \text{ITD}}$ between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔITD value.

$$\Delta \text{ILD}_q = |\text{ILD}(H_q) - \text{ILD}(\hat{H}_q)| \tag{4.6}$$

and the solid angle weighted $\overline{\Delta \text{ILD}}$ values are calculated as

$$\overline{\Delta \text{ILD}} = \sum_{q=1}^{Q} \Omega_q \Delta \text{ILD}_q \tag{4.7}$$

Figure 4.18 displays the solid angle weighted $\overline{\Delta \text{ILD}}$ values between HRTFs and binaural Ambisonic rendering with and without DFE for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute ΔILD value. Detailed plots of ΔILD for every measurement location on the sphere are presented in Appendix A.2, which show the close similarity between renders with and without DFE, and an overall improvement in the reproduced ILD values with increased Ambisonic order.

Observation of these figures confirms that DFE has a minimal effect on the Ambisonic rendering of interaural cues.

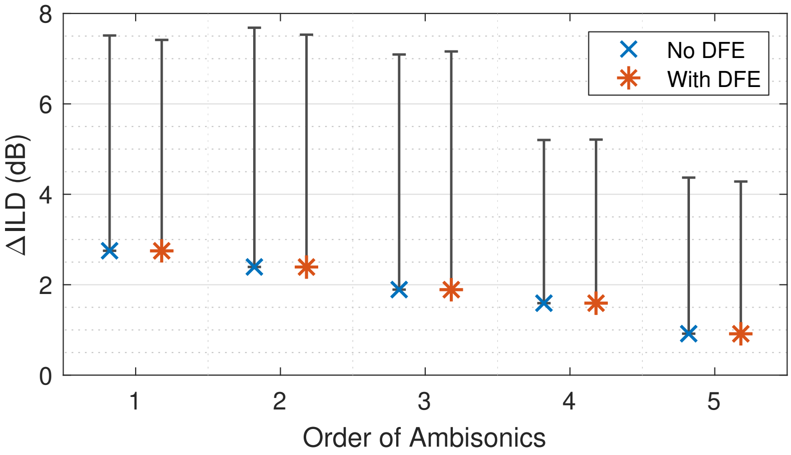


FIGURE 4.18: $\overline{\Delta\text{ILD}}$ between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔILD value.

### 4.3.3 Estimated Horizontal Localisation

The effect of DFE on horizontal localisation was estimated using the method detailed in Section 2.8.4, which utilises a horizontal model (May, Van De Par and Kohlrausch, 2011), producing a value of $\overline{E_\theta}$ for overall estimated localisation between $-90° < \theta < +90°$ at $\phi = 0°$ using (2.17). Figure 4.19 displays the overall estimated horizontal localisation of binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$. Appendix A.3 presents detailed individual plots for the estimated horizontal localisation of each simulated azimuth angle, which show similarity between renders with and without DFE, and an overall improvement in lateralisation with increased Ambisonic order.

Observation of these figures suggests that DFE has a small effect on the accuracy of estimated horizontal localisation. For $M = 1$, $M = 3$ and $M = 5$ the overall accuracy is improved, but for $M = 2$ and $M = 4$ it is poorer. As the model uses ITD and ILD calculations in the localisation estimation process, which have been shown in Section 4.3.2 to be largely unaffected by DFE, an explanation for the different localisation accuracy could be that the different spectral responses caused by DFE may mask or unmask certain frequencies in the localisation estimation process. This

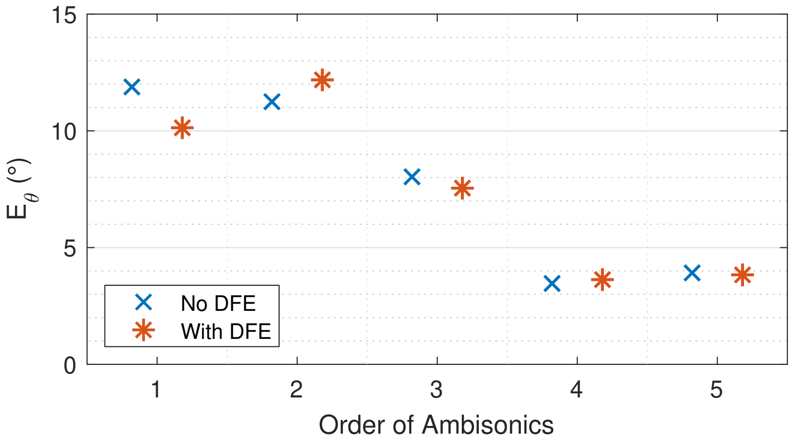


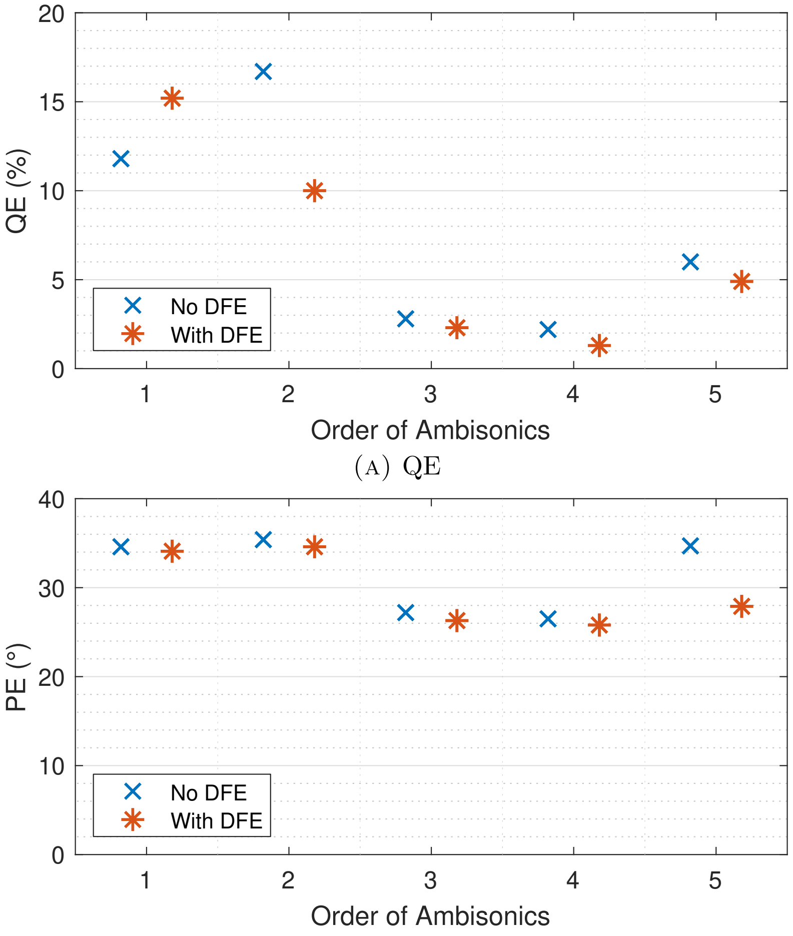
FIGURE 4.19: Estimated $\overline{E_\theta}$ of binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (May, Van De Par and Kohlrausch, 2011).

requires further testing to determine, such as real life localisation tests. As the focus of this thesis is timbral accuracy, this will not be discussed further.

### 4.3.4 Estimated Sagittal Plane Localisation

The effect of DFE on estimated elevation localisation in the sagittal plane was evaluated between $-90° < \phi < +90°$ at $\theta = 0°$ using the method detailed in Section 2.8.4 which utilises a localisation model (Baumgartner, Majdak and Laback, 2014) to produce two psychoacoustic performance metrics: quadrant error (QE), a prediction of localisation confusion (presented as a percentage), and polar RMS error (PE), a prediction of precision and accuracy in degrees. As the HRTFs used are of a Neumann KU 100, which has no torso, there will be no elevation cues present below 1.5 kHz (Algazi, Avendano and Duda, 2001a). Therefore, the frequency range of the model's filter bank was set to 1.5 kHz - 18 kHz, with the upper limit of the frequency range chosen as the highest frequency of perceivable elevation cues (Roffler and Butler, 1967; Asano, Suzuki and Sone, 1990).

Figure 4.20 shows the predicted QE and PE values of binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$. Detailed individual plots of predicted sagittal plane localisation are presented in Appendix A.4, which show



(A) QE

(B) PE

FIGURE 4.20: Estimated sagittal plane localisation of binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (Baumgartner, Majdak and Laback, 2014).

small improvements in localisation with increased Ambisonic order, as well as within Ambisonic order with DFE implementation, characterised by a bolder and straighter diagonal white line.

Results indicate a small improvement in sagittal plane localisation performance with the implementation of DFE for most Ambisonic orders. This is most likely due to the improved spectral reproduction. However, at $M = 1$, the QE value is higher with DFE implementation, despite the PE being marginally lower. Looking at Appendix A.4b, it appears that DFE increases up-down confusion, with a 'shadow' region appearing at $\phi < -30°$ and a predicted elevation $\phi > 60°$. This is likely due to the overall increase in high frequencies, which are associated with higher elevations.

### 4.3.5 Generalisability

To demonstrate the robust applicability of DFE, additional numerical evaluation of binaural Ambisonic rendering was performed using different loudspeaker configurations and an alternative HRTF dataset. In both cases, the effect of DFE was assessed as in Section 4.3.1, with PSD calculations comparing binaural Ambisonic renders to a reference dataset of HRTFs for all available measurement locations.

Firstly, different loudspeaker configurations were investigated[4]. Here, spherical T-designs (Hardin and Sloane, 1996), another commonly used loudspeaker configuration type for higher-order Ambisonic reproduction, are employed with corresponding loudspeaker vertices as illustrated in Figure 3.8. However, HRTFs measured at the exact spherical coordinates of the T-design vertices, as shown in Figure 3.8, are not all present in the Bernschütz Neumann KU 100 database (Bernschütz, 2013). Therefore, to approximate the T-design loudspeaker configurations in this and subsequent chapters, the closest available HRTFs are selected. Some HRTFs are, at worst, within ±1° of accuracy. This does cause small errors in orthonormality, meaning the practical orthonormality of the T-design loudspeaker configurations is not as accurate as depicted in Figure 3.9.

The diffuse-field responses, inverse filters and resulting equalised frequency responses of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations are presented in Appendix A.5, which show similar magnitude irregularities at frequencies above $f_{\text{alias}}$ to those shown in Figure 4.5.

The solid angle weighted $\overline{\text{PSD}}$ results, calculated from all locations on the sphere, are shown in Figure 4.21, along with the maximum and minimum absolute PSD values. This illustrates how DFE produces an overall improvement in PSD, regardless of the type of loudspeaker configuration. Detailed plots of PSD for every measurement location on the sphere are presented in Appendix A.6, which also follow similar trends to those shown in the Lebedev loudspeaker configurations in Figure 4.16 with overall

[4] For all other areas of this chapter, Lebedev loudspeaker configurations are used.

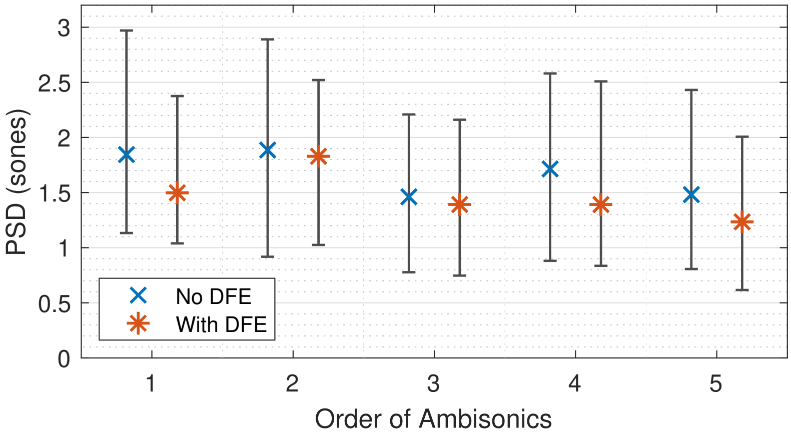


FIGURE 4.21: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the minimum and maximum absolute PSD values.

improved reproduction at higher Ambisonic orders, and improved reproduction approaching the median plane with the implementation of DFE.

Secondly, to assess how DFE works when using an alternative HRTF dataset, binaural Ambisonic renders were made for $\{M = 1, M = 2, M = 3, M = 5\}$[5] using the corresponding Lebedev loudspeaker configurations as used throughout this thesis and individualised HRTFs from the SADIE II database, human subject H20 (Armstrong et al., 2018a). The diffuse-field responses, inverse filters and resulting equalised frequency responses of Ambisonic orders $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs are presented in Appendix A.7, which show similar magnitude irregularities at frequencies above $f_{\text{alias}}$ to those shown in Figure 4.5 and Appendix A.5.

The solid angle weighted $\overline{\text{PSD}}$ results, calculated from 2,114 locations on the sphere, are shown in Figure 4.22 along with the minimum and maximum absolute PSD values. DFE produces an incremental improvement in $\overline{\text{PSD}}$ for all tested orders of Ambisonics. Detailed plots of PSD for every measurement location on the sphere are presented in Appendix A.8, which also follow similar trends to Figure 4.16 and Appendix A.6, with overall improved reproduction at higher Ambisonic orders, and

[5] The omission of $M = 4$ was due to a lack of necessary measurements.

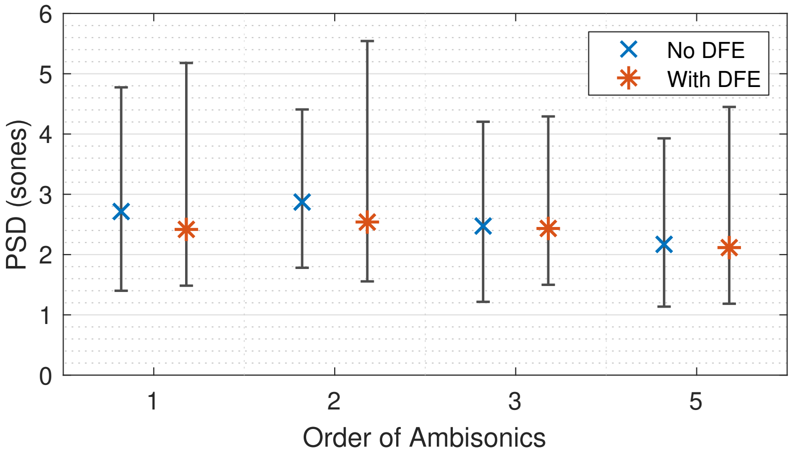


FIGURE 4.22: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the minimum and maximum absolute PSD values.

improved reproduction approaching the median plane with the implementation of DFE.

The tests on generalisability therefore show how DFE is applicable to binaural Ambisonic rendering using different virtual loudspeaker configurations and alternative HRTF datasets, through incremental improvements to spectral reproduction.

## 4.4 Perceptual Evaluation

It can be seen from the results obtained through numerical evaluation that the main effects of Ambisonic DFE are improvements to high frequency reproduction, hence perceptual evaluation focussed on timbre in binaural Ambisonic rendering. Two listening tests were conducted on 20 participants aged between 20 to 38 (17 male, 2 non-binary, 1 female) with self reported normal hearing according to ISO Standard 389 (International Organization for Standardization, 2016) and prior critical listening experience (such as education or employment in audio or music engineering). Tests were conducted using Ambisonic orders $M = 1$, $M = 3$ and $M = 5$, with Ambisonic orders $M = 2$ and $M = 4$ omitted to reduce the duration of the listening tests, and Lebedev loudspeaker configurations.

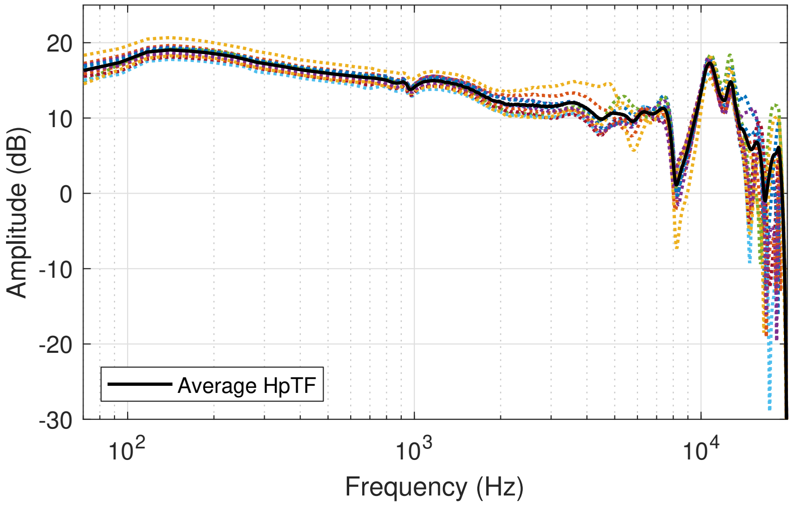


FIGURE 4.23: Frequency responses of 11 measured HpTFs for Sennheiser HD 650 headphones mounted on a Neumann KU 100 dummy head, with RMS average response (left ear).

Tests were conducted in a quiet listening room (background noise level of 41.8 dBA) using an Apple Macbook Pro with a Fireface UCX audio interface, which has software controlled input and output levels. The headphones used were a single set of Sennheiser HD 650 circumaural headphones, which were equalised using 11 measurements obtained from a Neumann KU 100 dummy head using the exponential swept sine impulse response technique (Farina, 2000) and re-fitting of the headphones between each measurement. The frequency responses of the 11 measurements are presented in Figure 4.23, which illustrates the variation in high frequency response caused by simply removing and refitting the headphones. Equalisation filters were calculated from the RMS average of the 11 deconvolved headphone transfer functions (HpTFs) using Kirkeby and Nelson's least-mean-square regularisation method (Kirkeby and Nelson, 1999). One octave smoothing was implemented using the complex smoothing approach of (Hatziantoniou and Mourjopoulos, 2000), and the range of inversion was 5 Hz - 4 kHz. In-band and out-band regularization of 25 dB and -2 dB respectively was used. These values were chosen empirically to reduce sharp peaks in the inverse filtering. Again, for greater detail on the inverse filtering methods used, refer back to Section 2.8.1. The RMS HpTF and inverse filter of the left HD 650 headphone, along with a resulting equalised response (calculated by convolving the RMS response with the inverse filter), are shown in Figure 4.24.

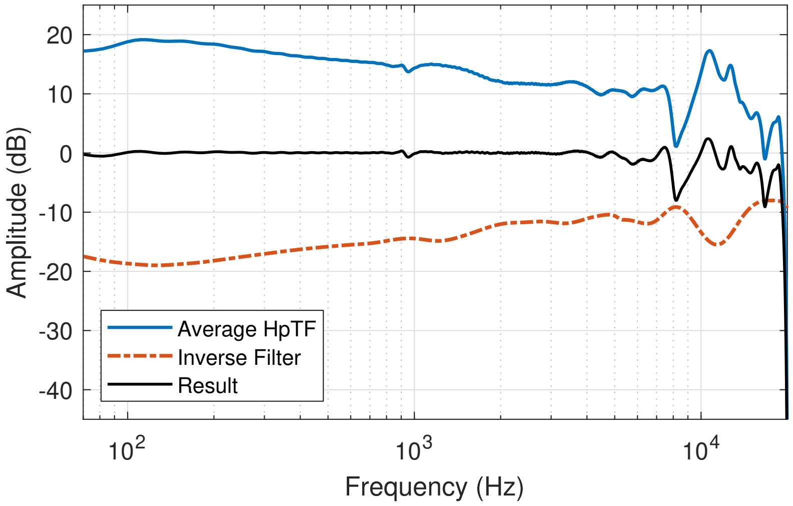


FIGURE 4.24: RMS of 11 measured HpTFs for Sennheiser HD 650 headphones mounted on a Neumann KU 100 dummy head, with inverse filter and resulting equalised response (left ear).

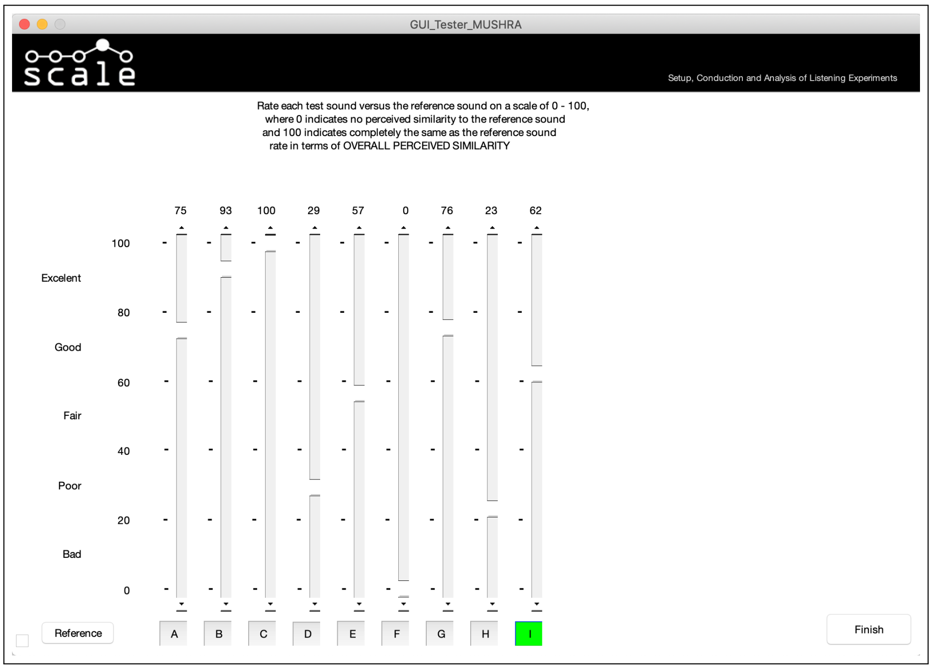
TABLE 4.6: Spherical coordinates of test sound locations.

| $\psi$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $\theta$ (°) | 180 | 50 | 118 | 0 | 180 | 62 | 130 | 0 |
| $\phi$ (°) | 64 | 46 | 16 | 0 | 0 | $-16$ | $-46$ | $-64$ |

The base stimulus was one second of monophonic pink noise at a sample rate of 48 kHz, windowed by onset and offset half-Hanning ramps of 5 ms. Each test sound was generated by convolving the pink noise with an HRTF; either Ambisonic or not. The test sound locations ($\psi$) corresponded to the central points of the faces of a dodecahedron. To reduce the total number of trials, symmetry was assumed and thus only locations in the left hemisphere were used, amounting to 8 locations (see Table 4.6). Test sounds were normalised to a consistent A-weighted RMS amplitude and participants were able to adjust the playback level. All binaural renders were static (fixed head orientation) to ensure consistency in the experience between participants.

### 4.4.1 Test Paradigms

In the perceptual evaluation used in this thesis, *trial* refers to each MUSHRA page (different trials will have different test sound locations or sound scene excerpts), and *condition* refers to the specific stimuli inside a trial (one trial will have multiple



Rate each test sound versus the reference sound on a scale of 0 - 100,
where 0 indicates no perceived similarity to the reference sound
and 100 indicates completely the same as the reference sound
rate in terms of OVERALL PERCEIVED SIMILARITY

FIGURE 4.25: Screenshot of the MUSHRA interface used in the Ambisonic DFE listening test using ‘Scale’ (Giner, 2013).

conditions, one of which will be the reference, the others of which will be the stimuli under test: i.e. the Ambisonic renders).

The first listening test followed the multiple stimulus test with hidden reference and anchor (MUSHRA) paradigm, ITU-R BS.1534-3 (International Telecommunication Union, 2015b). A screenshot of the MUSHRA interface, using the MATLAB based ‘Scale’ (Giner, 2013), is presented in Figure 4.25. The reference was a direct HRTF convolution, and medium and low anchors were low-pass filtered versions of the reference stimulus with an $f_c$ of 7 kHz and 3.5 kHz, respectively. The other 6 stimuli were the binaural Ambisonic renders for three Ambisonic orders, with and without DFE. For each trial, the listener was asked to rate the 9 stimuli in terms of timbral similarity to the reference. The 8 test sound locations were repeated giving a total of 16 trials. The presentation of stimuli and trials was randomised and double blind.

The second listening test was an AB comparison. Participants were presented with two sets of three consecutive stimuli (corresponding to Ambisonic renders at $\{M = 1, M = 3, M = 5\}$), one set of which was diffuse-field equalised. They were asked to rate each set in terms of timbral consistency. The 8 test sound locations were repeated with a different arrangement of the Ambisonic orders (the first was $\{M = 1, M = 3, M = 5\}$ and the second was $\{M = 1, M = 5, M = 3\}$), giving a total of 16 trials. The presentation of trials was again randomised and double blind.

Prior to the tests, participants were given the ANSI S1.1-1994 definition of timbre: 'that attribute of sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar' (American National Standards Institute, 1994) and taken through a training exercise to familiarise themselves with the test interface and task.

### 4.4.2 Results

Overall, the tests lasted between around 20 to 45 minutes to complete. No participant's results were excluded, based on the criteria of rating the hidden reference less than 90% for more than 15% of trials or rating the mid-anchor higher than 90% for more than 15% of trials. The results from both listening tests were tested for normality using the Kolmogorov-Smirnov test, which showed all data as non-normally distributed. As a result, all statistical analysis was conducted using non-parametric methods.

The median results of the MUSHRA test, conducted to determine whether DFE reduces the differences in timbre between binaural Ambisonic rendering and HRTF convolution, are shown in Figure 4.26 for each condition across all test sound locations, with non-parametric 95% confidence intervals ($\mathrm{CI}_{95}$) (Mcgill, Tukey and Larsen, 1978), calculated from the median, 25th and 75th percentiles, denoted as $\rho_{50}$, $\rho_{25}$ and $\rho_{75}$, as

$$\begin{aligned} \mathrm{CI}_{95}^{-} &= \rho_{50} - \frac{1.57(\rho_{75} - \rho_{25})}{\sqrt{Q}} \\ \mathrm{CI}_{95}^{+} &= \rho_{50} + \frac{1.57(\rho_{75} - \rho_{25})}{\sqrt{Q}} \end{aligned} \tag{4.8}$$

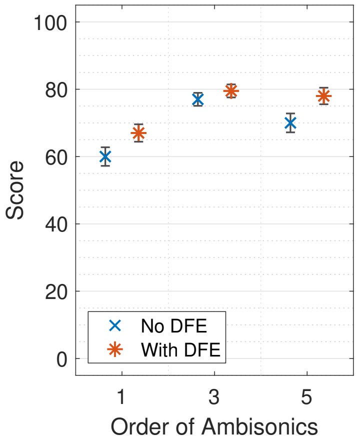


FIGURE 4.26: Median MUSHRA results with non-parametric $CI_{95}$ across all test sound locations, reference and anchor scores omitted. Score indicates perceived timbral similarity between test stimulus and HRTF reference.

TABLE 4.7: Significance results of the MUSHRA test over all test sound locations using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether DFE produced a statistically significant improvement to binaural Ambisonic rendering.

| $M$ | 1 | 3 | 5 |
|---|---|---|---|
| $h$ | 1* | 0 | 1* |

where $Q$ denotes the number of measurements. Friedman's Analysis of Variance (ANOVA) tests show a statistically significant difference ($\chi^2(5) = 247.6, p < 0.05$) between standard and DFE binaural Ambisonic rendering for all tested orders and sound locations. $M = 1$ shows the most improvement, followed by $M = 5$ and $M = 3$. To test whether these improvements are statistically significant, post-hoc Wilcoxon signed-rank tests were conducted for each Ambisonic order, and Table 4.7 presents the significance results. $M = 1$ and $M = 5$ are both highly statistically significant, and though a small improvement is shown at $M = 3$, it is not statistically significant at a confidence of 95% ($p = 0.198$).

The perceptual effect of DFE was found to vary with test sound location, with a Friedman's ANOVA showing this variation to be statistically significant ($\chi^2(7) = 127.8, p < 0.05$). Figure 4.27 shows the median results with non-parametric $CI_{95}$ for each test sound location $\psi$. Post-hoc Wilcoxon signed-rank tests determined which test conditions with DFE produced a significant improvement in timbre; the results of which are shown in Table 4.8. For both $M = 1$ and $M = 5$, DFE was shown to

Table 4.8: Significance results of the MUSHRA test using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether DFE produced a statistically significant improvement to binaural Ambisonic rendering.

| $\psi$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $h$ ($M = 1$) | 1 | 1* | 0 | 1* | 1* | 0 | 1* | 0 |
| $h$ ($M = 3$) | 0 | 0 | 0 | 0 | 0 | 1* | 0 | 0 |
| $h$ ($M = 5$) | 1* | 1* | 0 | 0 | 1 | 1* | 1* | 0 |

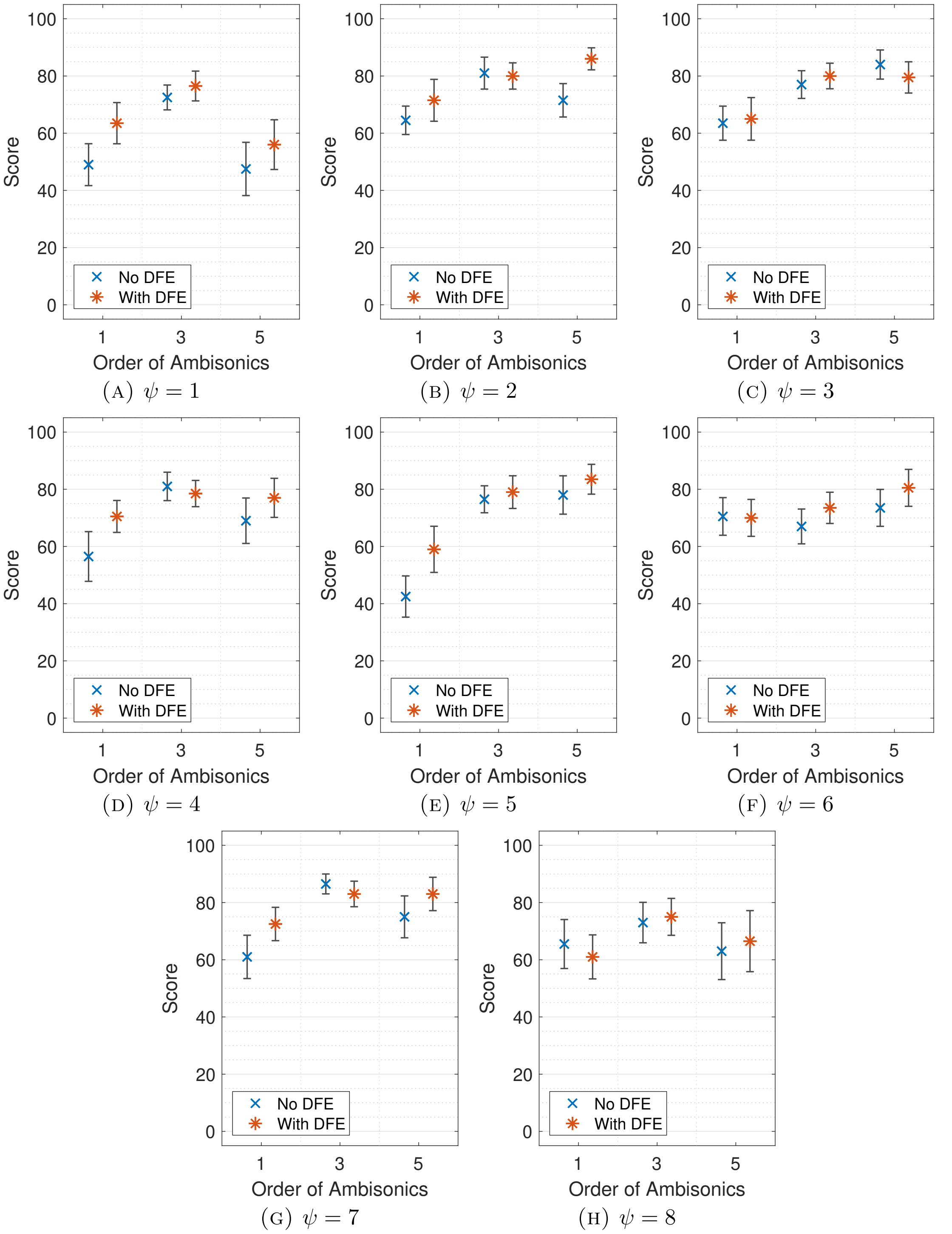
bring the timbre of binaural Ambisonic rendering closer to HRTF convolution with statistical significance for 5 of the 8 test sound locations. Results for $M = 3$ order were much less clear and only significant for one test sound location ($\psi = 6$). Results for $\psi = 8$ were not statistically significant for any tested order of Ambisonics.

The median results of the second listening test, the AB comparison conducted to determine whether DFE improved the consistency of timbre between Ambisonic orders, are shown in Figure 4.28 for both conditions across all test sound locations, with non-parametric $\text{CI}_{95}$.

Overall across all test sound locations, DFE produced higher timbral consistency between different Ambisonic orders, and a Friedman's ANOVA test showed that this was statistically significant ($\chi^2(1) = 8.45, p < 0.05$). To assess how perceived timbral consistency varied with test sound location, a second Friedman's ANOVA was conducted and showed significance ($\chi^2(7) = 37.5, p < 0.05$). Figure 4.29 shows the median AB results with non-parametric $\text{CI}_{95}$ for each test sound location $\psi$. Post-hoc Wilcoxon signed-rank tests to determine which test sound locations produced statistically significant results were conducted, the results of which are displayed in Table 4.9. Only $\psi = 2$ and $\psi = 4$ produced statistically significant results.

## 4.5 Discussion

The numerical evaluation of Ambisonic DFE has shown that, by calculating and applying DFE filters to the HRTFs used in binaural Ambisonic rendering, an improvement in high frequency reproduction is observed when compared to direct HRTF



(A) $\psi = 1$

(B) $\psi = 2$

(C) $\psi = 3$

(D) $\psi = 4$

(E) $\psi = 5$

(F) $\psi = 6$

(G) $\psi = 7$

(H) $\psi = 8$

FIGURE 4.27: Median MUSHRA results with non-parametric $CI_{95}$ for each test sound location ($\psi$), reference and anchor scores omitted. Score indicates perceived timbral similarity between test stimulus and HRTF reference.

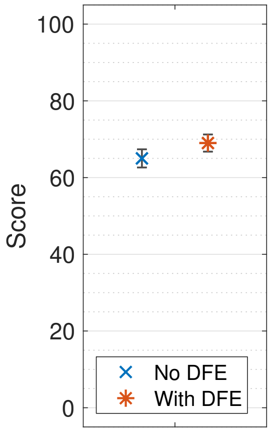


FIGURE 4.28: Median AB results with non-parametric CI$_{95}$ across all test sound locations. Score indicates perceived timbral consistency between the three tested orders of Ambisonics.

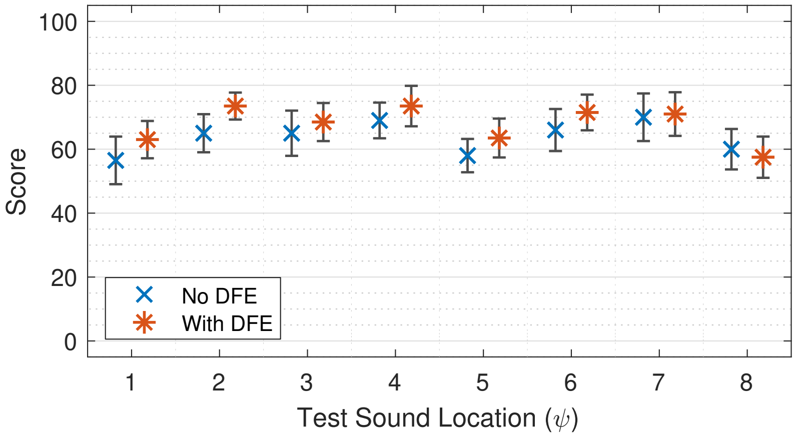


FIGURE 4.29: Median AB results with non-parametric CI$_{95}$ for each test sound location ($\psi$). Reference and anchor scores omitted. Score indicates perceived timbral consistency between the three tested orders of Ambisonics.

TABLE 4.9: Significance results of the AB test using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether DFE produced a statistically significant improvement to the timbral consistency of different Ambisonic orders.

| $\psi$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $h$ | 0 | 1* | 0 | 1 | 0 | 0 | 0 | 0 |

rendering. This has been shown through an overall reduction in perceptual spectral difference (see Figure 4.15), including when using different loudspeaker configurations or individualised HRTFs (Figures 4.21 and 4.22, respectively), as well as a reduction in predicted sagittal plane localisation error (Figure 4.20). The estimated rendering of interaural cues such as ITD and ILD, as well as predicted horizontal localisation, are largely unaffected by the implementation of DFE. This is most likely due to the linear phase nature of the equalisation filters and directional independence of the HRTF changes: every direction has the same filter applied. Anecdotally, the implementation of Ambisonic DFE tends to produce an increase in high frequency content.

A more general observation was made that the higher Ambisonic order did not always perform the best. This occurs in perceptual test results (see again Figure 4.26) where $M = 5$ is rated lower than $M = 3$, as well as in spectral difference calculations (see again Figure 4.15), where $M = 5$ performs very similarly to $M = 4$ (without DFE). This could be due to the HRTF measurement, as different Ambisonic order performance is seen when using T-design loudspeaker configurations or alternative HRTFs (see again Figures 4.21 and 4.22).

The perceptual evaluation of Ambisonic DFE has also demonstrated that binaural Ambisonic rendering with DFE produces closer high frequency reproduction to direct HRTF rendering than standard binaural Ambisonic rendering (see Figure 4.26). Timbral consistency is also improved across different orders of Ambisonics (Figure 4.28). However, results do not demonstrate high levels of statistical significance, indicating that even with DFE, binaural Ambisonic reproduction still varies considerably in timbre with Ambisonic order and with direct HRTF rendering. Listening test results for $M = 1$ and $M = 5$ produced greater statistical significance than $M = 3$, which is likely due to the more substantial variation in Ambisonic diffuse-field responses for $M = 1$ and $M = 5$, as illustrated in Figure 4.5, and at lower frequencies with more perceptual importance (see again the equal loudness curves presented in Figure 2.14).

Results were shown to vary with sound source location in both the numerical evaluation of perceptual spectral difference and the listening tests (see again Figures 4.16 and 4.27, respectively). In general, the greatest improvements appear closer to the median plane, and the lateral directions are made poorer. This is a similar effect to that found in DFE of SH order-truncated Ambisonic signals (Ben-Hur et al., 2017), though the effect is not as pronounced here due to the lack of high frequency roll off. A likely explanation of this is illustrated in (Bernschütz et al., 2014, Figure 3), which demonstrates the greater difference in path lengths for loudspeakers close to the median plane than at the lateral positions.

## 4.6 Summary

The inaccuracies of high frequency reproduction in Ambisonics, caused by comb filtering from the summation of multiple similar signals at the ears, have been addressed in this chapter through the introduction of the Ambisonic Diffuse-Field Equalisation (DFE) technique. By implementing DFE in binaural Ambisonic rendering as an offline low-computation virtual loudspeaker HRTF pre-processing technique, the diffuse-field response of the binaural Ambisonic loudspeaker configuration is flattened out, which changes the frequency response of renders at individual sound source locations. The perceptual spectral difference model, a numerical method for evaluating timbral difference between binaural signals based on traits of human auditory perception, has been presented. A validation has shown that it produces a closer portrayal of human timbral assessment than a basic spectral difference calculation. It is therefore used in this thesis for spectral difference calculations.

Numerical evaluation of Ambisonic DFE shows that the timbre and predicted height localisation of binaural Ambisonic rendering can be improved, with little effect on the estimated interaural cues and horizontal localisation accuracy. A perceptual evaluation in the form of two listening tests on timbre has corroborated the numerical results with some statistical significance, although not across all test conditions. Listening tests also showed that Ambisonic DFE produces a small, but not statistically significant, improvement in timbral consistency between different orders of Ambisonics.

This chapter has shown that a low-computation equalisation HRTF pre-processing stage can produce an improvement in the high frequency reproduction of binaural

Ambisonic rendering. It is therefore recommended for implementation in binaural Ambisonic rendering. However, there still exists a significant difference between binaural Ambisonic rendering and direct HRTF rendering. Therefore, Ambisonic DFE alone is not enough to minimise the timbral issues posed by Ambisonic rendering, and the coming chapters will look at alternative methods.

# Chapter 5

# Ambisonic Directional Bias Equalisation

Chapter 4 adapted the technique of diffuse-field equalisation (DFE) to virtual loudspeaker binaural Ambisonic rendering, which was shown to improve the overall spectral reproduction over the sphere of all tested orders of Ambisonics at frequencies above $f_{\text{alias}}$. Other effects of DFE implementation include higher accuracy in predicted median plane elevation localisation. However, there still exists a definite and perceivable difference in timbre between binaural Ambisonic rendering with DFE and direct HRTF convolution, even at $M = 5$.

This chapter introduces a novel method to further improve the spectral reproduction of binaural Ambisonic rendering for a specific direction, by adapting the method of Ambisonic DFE. This HRTF pre-processing technique is therefore not diffuse-field equalisation and is referred to in this chapter as Ambisonic Directional Bias Equalisation (DBE). Instead of removing the direction-independent aspects of binaural Ambisonic reproduction, the presented method aims to focus the improvement in spectral reproduction to a specific direction via directional biasing of the quadrature in the diffuse-field response calculation and an additional re-equalisation to the frequency response of a non-Ambisonic HRTF in the direction of bias.

The method of Ambisonic DBE used in this thesis is explained in detail in this chapter, with attention paid to the adaptation of the diffuse-field response calculation and the directional bias HRTF re-equalisation. As with Ambisonic DFE, Ambisonic DBE can be applied to the SH binaural Ambisonic decoder in an offline process, thus producing no increase in real-time computational cost. Ambisonic DBE is evaluated both numerically and perceptually, with numerical evaluation comparing binaural Ambisonic rendering to a reference set of HRTFs in terms of perceptual spectral difference, estimated interaural cue similarity and predicted localisation accuracy for $\{M = 1, M = 2, ..., M = 5\}$, with varying levels of directional bias. The applicability of DBE to other loudspeaker configurations and individualised HRTFs is also explored. Perceptual evaluation is carried out in the form of two listening tests, comparing binaural Ambisonic rendering with varying levels of DBE to HRTF convolution in terms of timbral similarity. The first listening test uses single sound sources in a simple acoustic scene, and the second uses multiple sound sources in a synthesised complex acoustic scene. The chapter ends with a summary of the findings and suggestions for future applications of Ambisonic DBE.

## 5.1 Method

This section details the approach taken in this thesis for improving the accuracy of spectral reproduction in binaural Ambisonic rendering for a specific direction by adaptation of the diffuse-field equalisation method. A block diagram of the method is presented in Figure 5.1, and a brief summary of the Ambisonic DBE method is as follows: the diffuse-field response calculation method as explained in Chapter 4 is adapted through a directional biasing in the distribution of points used for obtaining an average frequency response of the binaural Ambisonic decoder over the sphere. The subsequent generation of binaural Ambisonic rendered HRTFs remains the same, as does the RMS average calculation, however, the resulting RMS binaural impulse response more closely resembles the Ambisonic reproduction in the direction of bias. Next, a re-equalisation stage brings the frequency reproduction of binaural Ambisonic

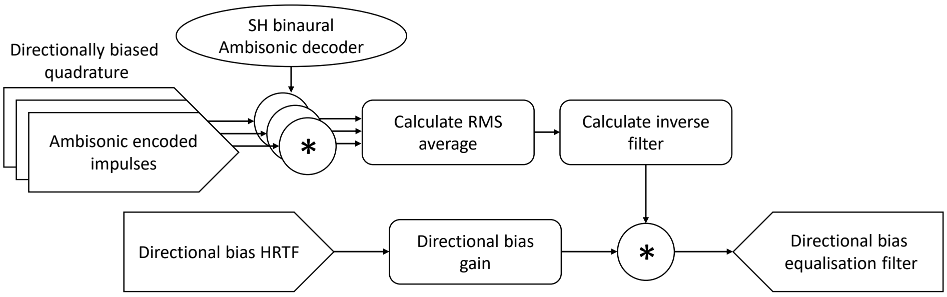

SH binaural Ambisonic decoder
Directionally biased quadrature
Ambisonic encoded impulses
*
Calculate RMS average
Calculate inverse filter
Directional bias HRTF
Directional bias gain
*
Directional bias equalisation filter

FIGURE 5.1: Block diagram of the Ambisonic DBE method.

rendering closer to HRTF rendering for the specified direction of bias. In this thesis, DBE filters are generated separately for left and right ears. All computation was carried out offline in MATLAB version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4.

### 5.1.1 Directional Biased Quadrature Response

An approximate diffuse-field response is typically calculated from the root-mean-square (RMS) of the magnitude responses of a large number of free-field measurements (Heller and Benjamin, 2012). The diffuse-field response calculation would usually account for an uneven spherical distribution of measurement positions by weighting each individual magnitude response relative to the amount of surface area on the sphere that each point contributes, known as the solid angle. In this chapter, however, it is necessary to deliberately skew the distribution of points on the sphere in order to bias the average response calculation in a specific direction. This means that solid-angle weighting of a near-regular spherical distribution of measurements is not appropriate. Therefore, the regularity of the spherical distribution is an important consideration to make.

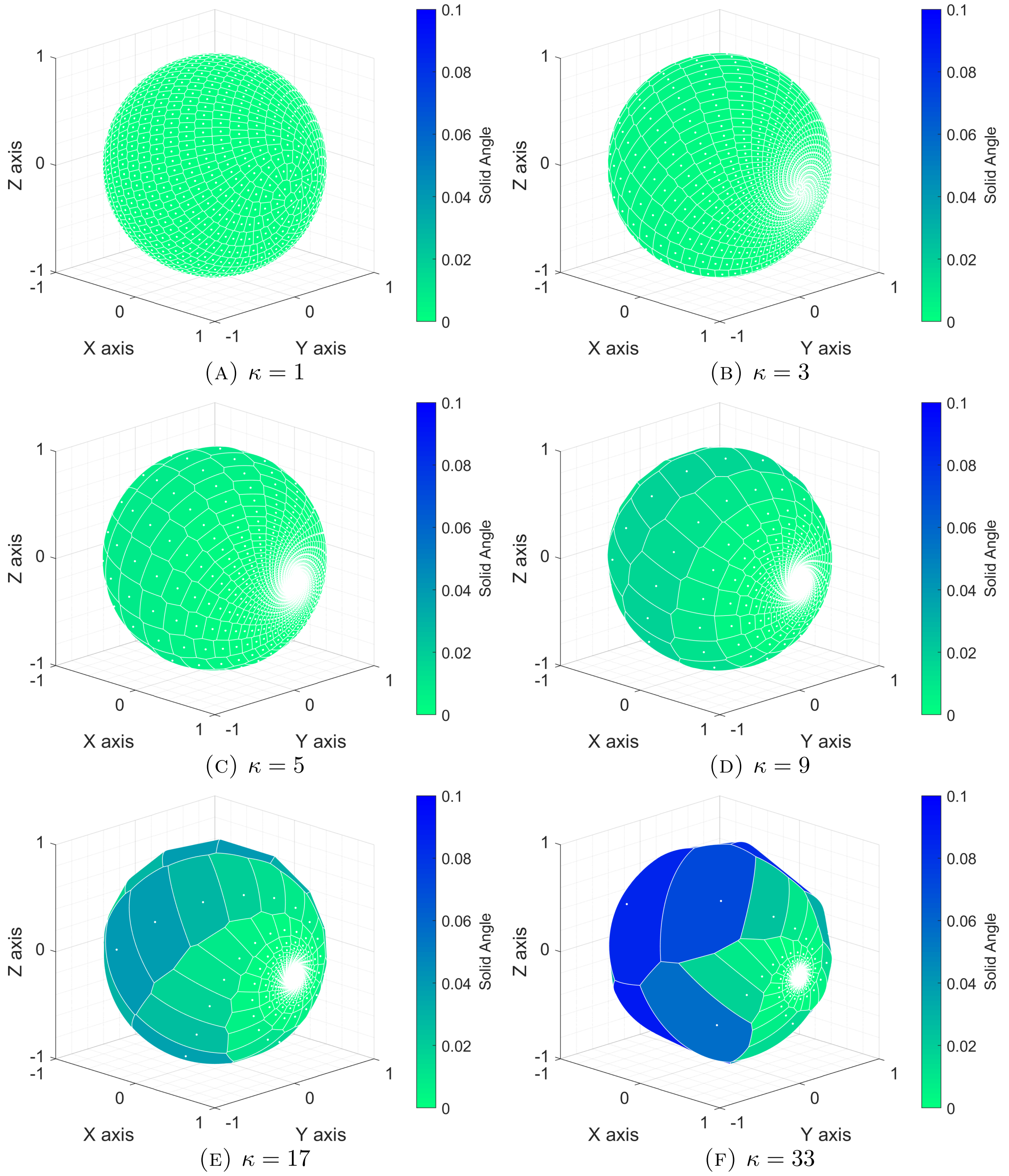
Figure 4.2 illustrated that T-design quadrature is the most regular of the four investigated spherical distribution methods. However, the highest resolution T-design available is $\varrho = 240$. For this reason, the quadrature used in DBE is the Fibonacci spiral, $\varrho = 1{,}000$, as it offers a greater resolution and is sufficiently regular for the majority of the sphere, producing only minor irregularities at the poles.

The distribution of points is then skewed in one direction, a process referred to in this thesis as directional biasing of quadrature (DBQ). This is achieved as follows: the spherical coordinates of each point in the quadrature are converted to Cartesian coordinates with values between ±1, and biased in the z axis due to this being the least regular region of Fibonacci quadrature; the directionally biased $z$ values $\hat{z}$ are calculated from the original values $z$ with respect to a bias factor $\kappa$ as

$$\hat{z} = \kappa(z + 1) - 1 \tag{5.1}$$

where $\kappa > 1$ produces a positive directional bias and $\kappa = 1$ produces no bias. The $x$ and $y$ values are unchanged. Cartesian coordinates are then converted back to spherical coordinates and rotated to the direction of bias. In this chapter, six values of directional bias, denoted in this thesis as $\kappa$, are investigated: $\kappa = 1, 3, 5, 9, 17$ and 33. These values have been empirically selected due to the gradual increase in bias that they produce, from no bias to extreme bias. Though the direction of bias can be chosen as any location on the sphere, this chapter will focus on frontal bias at a direction of $(\theta = 0°, \phi = 0°)$ unless stated otherwise. This direction is chosen as it is the position where human auditory localisation is the most accurate, as explained in Section 2.5, and it is a direction where the eyes face - humans are front oriented. The Voronoi sphere plots of the Fibonacci quadrature, $\varrho = 1{,}000$, with directionally biased quadrature using 6 values of $\kappa$ used in this chapter are shown in Figure 5.2.

As is the case in Chapter 4, the average response of the sphere is then calculated by generating binaural Ambisonic renders at the locations of every point $\iota$ in $\varrho$ (only in this case the points are directionally biased) and calculating an average response of the sphere from the RMS of each render in the frequency domain. The number



(A) $\kappa = 1$

(B) $\kappa = 3$

(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

FIGURE 5.2: Voronoi spheres demonstrating the directionally biased quadratures used in the first stage of DBE for the six values of $\kappa$ used in this thesis.

of frequency bins in the FFT calculation is the number of samples in the HRIR. However, in the case of DBE, no solid angle weighting is implemented, thus (4.2) is rewritten as

$$H_{\mathrm{RMS}} = \sqrt{\frac{1}{\varrho}\sum_{\iota=1}^{\varrho} H_{\iota}(\theta,\phi)^2} \tag{5.2}$$

with $1/\varrho$ now required. This average response is referred to here as the DBQ RMS response. Linear-phase inverse filters are then generated from the DBQ RMS responses between 2 Hz - 20 kHz using Kirkeby and Nelson's least-mean-square (LMS) regularisation method (Kirkeby and Nelson, 1999) with 1/8 octave smoothing and in / out-band inversion parameters of 30 dB and 20 dB, respectively. For greater detail on the inverse filtering methods used, refer back to Section 2.8.1.

Therefore the differences between the DBQ method and the diffuse-field response method presented in Chapter 4 are the directional skewing of points on the sphere used in the RMS calculation and removal of solid angle weighting, the different quadrature method, and the small variation in inverse filter parameters.

### 5.1.2 Equalisation Filter Calculation

As diffuse-field equalisation aims to flatten out the frequency response over all directions, equalising the binaural Ambisonic decoder with the inverse DBQ RMS filters is not sufficient: as $\kappa$ increases, the equalised frequency response in the direction of bias would become more uniform. As the aim of DBE is to produce Ambisonic rendering in the direction of bias closer to that of direct HRTF rendering, an additional re-equalisation stage is required. This is referred to in this thesis as the directional bias HRTF, which unless stated otherwise is a frontal bias at ($\theta = 0°, \phi = 0°$), as shown in Figure 5.3. The HRTF used for the directional bias HRTF stage is taken from the same HRTF data set as is used for the binaural Ambisonic rendering stage.

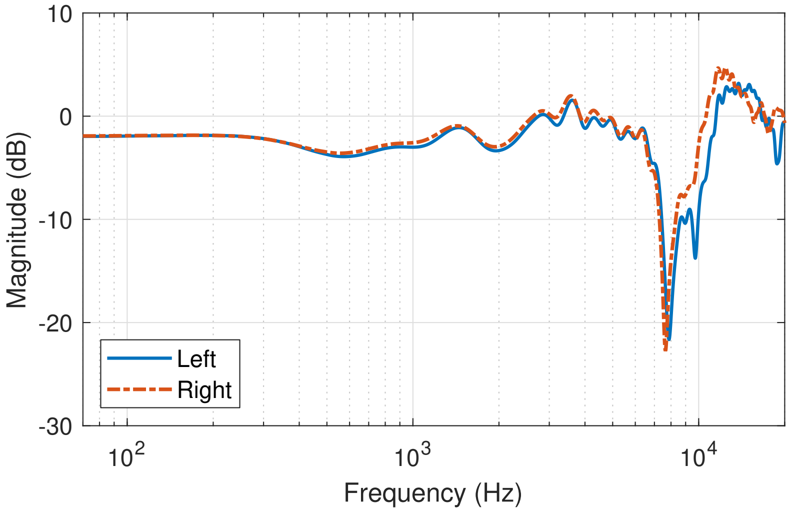


FIGURE 5.3: Frequency response of directional bias HRTF at $(\theta = 0°, \phi = 0°)$.

The gain for the directional bias HRTF $g_{\text{bias}}$ is calculated from the bias factor such that

$$g_{\text{bias}} = 1 - e^{-\frac{\kappa - 1}{10}} \tag{5.3}$$

The directional bias HRTF, $H_{\text{bias}}$ is then scaled by $g_{\text{bias}}$ in the frequency domain such that

$$\hat{H}_{\text{bias}} = H_{\text{bias}} g_{\text{bias}} + (1 - g_{\text{bias}}) \overline{H_{\text{bias}}} \tag{5.4}$$

where $\hat{H}_{\text{bias}}$ denotes the scaled directional bias HRTF and $\overline{H_{\text{bias}}}$ denotes the mean amplitude of all frequency bins of the directional bias HRTF. Therefore, a directional bias of $\kappa = 1$ (no bias) produces a filter with a completely flat frequency response at the mean amplitude of the directional bias HRTF, and $\kappa = \infty$ produces a filter equivalent to the frontal bias HRTF.

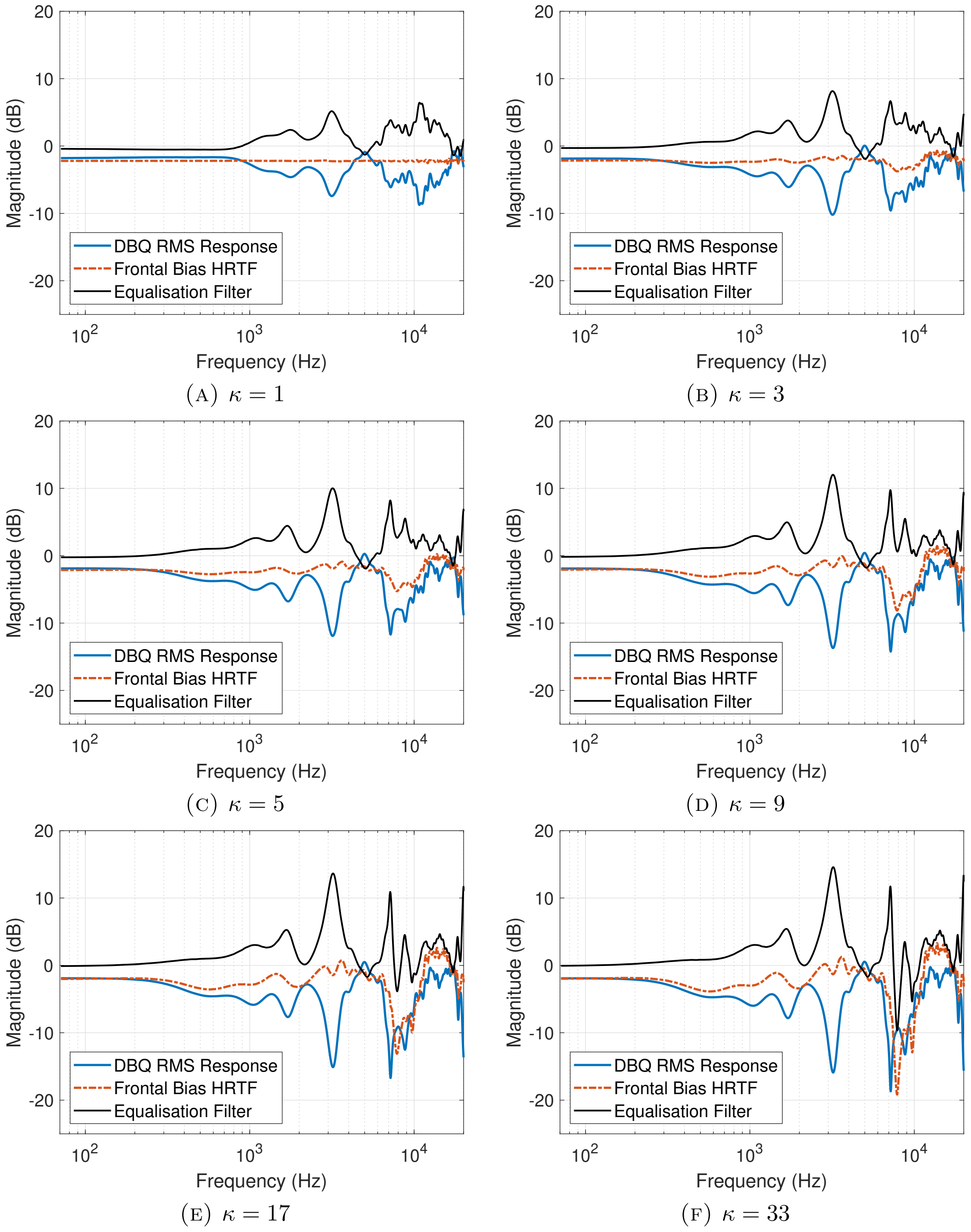
The frontal bias HRTF filter is then convolved with the inverse filter of the DBQ RMS resulting in the final DBE filters. Figure 5.4 presents the DBQ RMS responses, frontal bias HRTFs, and resulting final equalisation filters for $M = 1$, with varying $\kappa$. Corresponding Figures for Ambisonic orders $M = 2, M = 3, M = 4$ and $M = 5$

are presented in Appendices B.1, B.2, B.3 and B.4, respectively. With no directional bias at $\kappa = 1$, points are evenly distributed over the sphere, which, when equalised, is equivalent to the method of Ambisonic Diffuse-Field Equalisation. As $\kappa$ increases, the DBQ RMS response becomes more distorted as it more closely resembles an Ambisonic render in the direction of bias, and the directional bias HRTF filter begins to more closely resemble the magnitude of the directional bias HRTF (see again Figure 5.3). Therefore, the first equalisation stage removes the frontal Ambisonic response, and the second equalisation stage re-introduces the target HRTF response. DBE is implemented through offline convolution of the DBE filters with the channels of the SH binaural Ambisonic decoder.

## 5.2 Numerical Evaluation

In this section, the effect of DBE is evaluated numerically by comparing binaural Ambisonic renders, with DBE at a varying value of $\kappa$ and frontal direction of bias, to a reference dataset of HRTFs. The metrics of evaluation include perceptual spectral difference, interaural cues and estimated horizontal and vertical localisation. The applicability of DBE to other loudspeaker configurations and alternative HRTFs is also explored, as is the effect of altering the direction of bias to a lateral position.

For each measurement location in the reference HRTF dataset of $Q = 16{,}020$, as illustrated in Figure 2.28, binaural Ambisonic renders were generated for $\{M = 1, M = 2, ..., M = 5\}$, with DBE at a bias of $\kappa = 1, 3, 5, 9, 17$ and 33. The maximum value of bias is empirically chosen as $\kappa = 33$ due to observation of the frequency responses shown in Figure 5.4: when $\kappa = 33$, the directional-bias HRTF follows a similar shape to Figure 5.3. Therefore, this is considered sufficient for the highest value of $\kappa$ in this chapter. All HRTFs were truncated to 1024 taps with 50 sample in / out half-Hanning windows applied.



(A) $\kappa = 1$

(B) $\kappa = 3$

(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

FIGURE 5.4: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, $M = 1$ (left ear).

### 5.2.1 Perceptual Spectral Difference

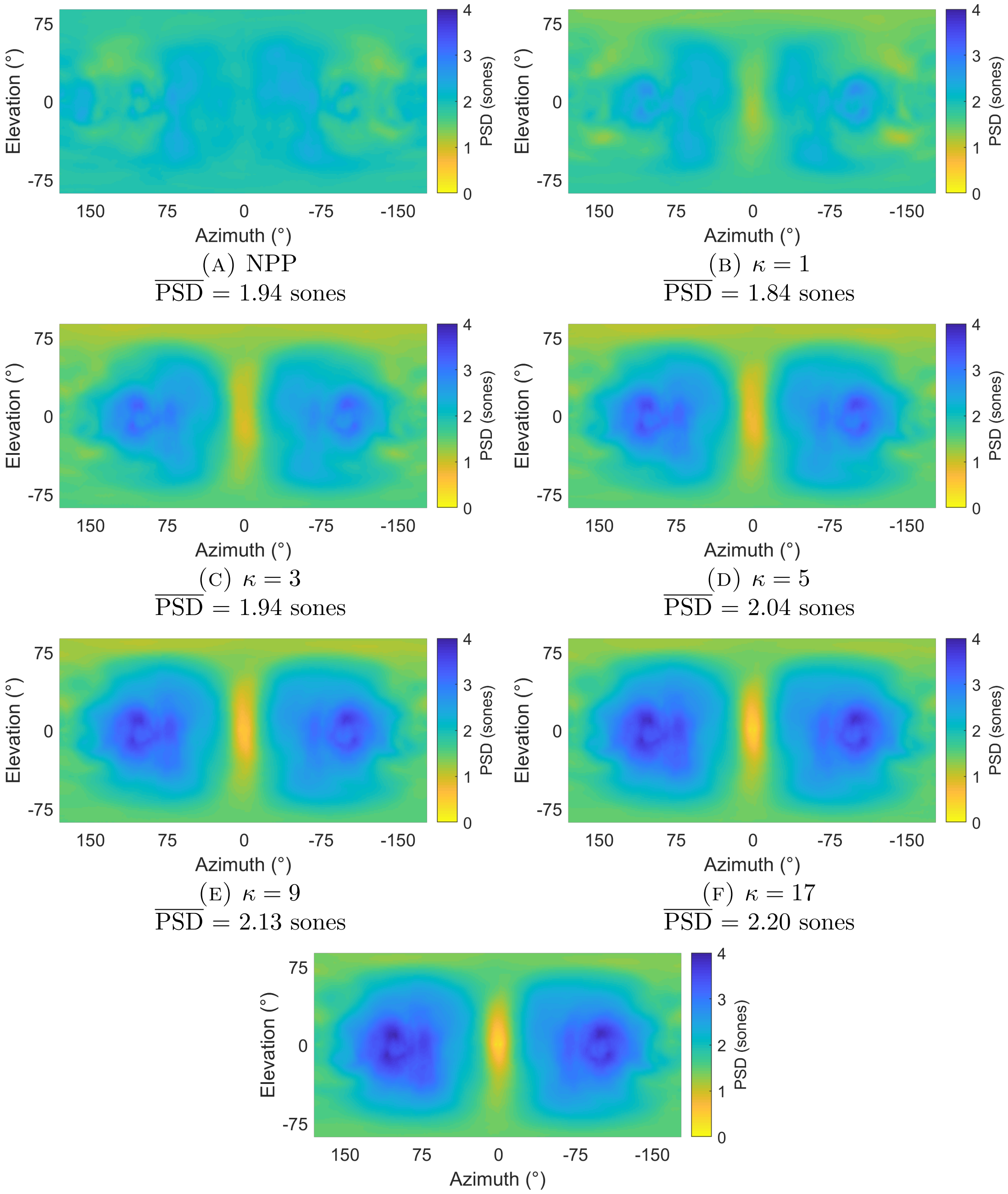
PSD between binaural Ambisonic renders and HRTFs was calculated for $\{M = 1, M = 2, ..., M = 5\}$ over all measurement locations on the sphere using the method detailed in Section 4.2. No normalisation was implemented in the PSD calculation in this chapter, in order to retain the true minimum values. This is because the PSD normalisation stage aims to find the closest overall perceptual loudness between the two datasets over all measurements.

To investigate the effect of DBE on PSD over the sphere, Figure 5.5 displays the absolute values of PSD between binaural Ambisonic rendering and reference HRTFs with DBE at varying values of $\kappa$ for every measurement location on the sphere (mean of left and right ear PSD calculations) at $M = 1$. The no pre-processing (NPP) case is also included as standard binaural Ambisonic rendering without DBE. Corresponding plots for Ambisonic orders $M = 2, M = 3, M = 4$ and $M = 5$ are presented in Appendices B.5, B.6, B.7 and B.8, respectively. These show how increasing the value of $\kappa$ produces an improvement in spectral accuracy for the frontal direction (direction of bias) with a reduction in spectral accuracy at the lateral locations.

Figure 5.6 presents the solid angle weighted $\overline{\text{PSD}}$ value for each tested value of $\kappa$ for $\{M = 1, M = 2, ..., M = 5\}$ calculated using (4.3), along with the maximum and minimum absolute PSD values. It is clear that increasing $\kappa$ reduces the minimum PSD value and increases the maximum PSD value and hence $\overline{\text{PSD}}$, suggesting that the overall timbral accuracy is reduced. At $\kappa = 33$, the spectral reproduction of binaural Ambisonic rendering is close to the reference HRTF at the direction of bias, even at $M = 1$.

### 5.2.2 Interaural Cues

The DBE filters are generated independently for left and right ears. To assess the effect of DBE on binaural Ambisonic reproduction of interaural cues, both ITD and ILD were estimated using the methods detailed in Sections 2.8.3 and 2.8.2, respectively,



(A) NPP
$\overline{\text{PSD}} = 1.94$ sones

(B) $\kappa = 1$
$\overline{\text{PSD}} = 1.84$ sones

(C) $\kappa = 3$
$\overline{\text{PSD}} = 1.94$ sones

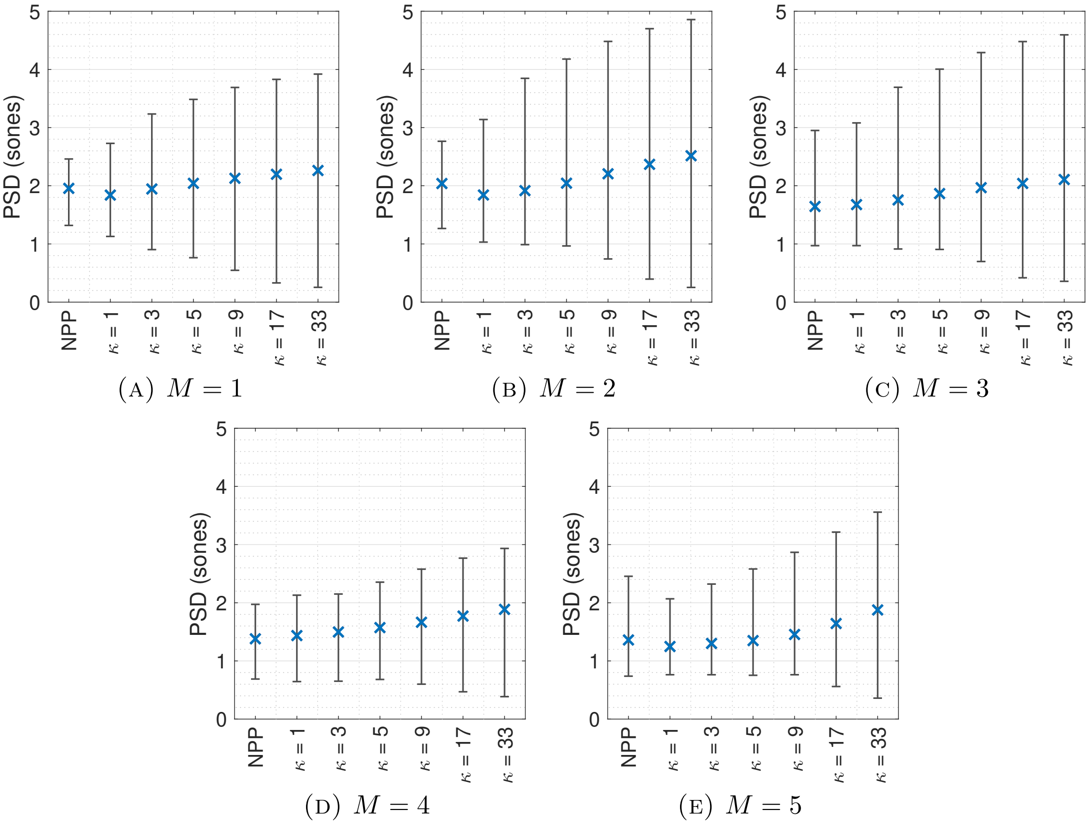
(D) $\kappa = 5$
$\overline{\text{PSD}} = 2.04$ sones

(E) $\kappa = 9$
$\overline{\text{PSD}} = 2.13$ sones

(F) $\kappa = 17$
$\overline{\text{PSD}} = 2.20$ sones

(G) $\kappa = 33$
$\overline{\text{PSD}} = 2.26$ sones

FIGURE 5.5: PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, $M = 1$ (mean of left and right PSD values). The no pre-processing (NPP) case is also included.



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$

(D) $M = 4$ (E) $M = 5$

FIGURE 5.6: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the minimum and maximum absolute PSD values.

for all measurement locations and Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, with DBE at a single bias factor of $\kappa = 33$ and frontal bias direction. The low-pass filter used in the ITD calculation is $f_c = 1.5$ kHz.

The change in ITD between the reference HRTFs and the Ambisonic generated HRTFs was then calculated for each measurement location $q$ using (4.4), and $\overline{\Delta\text{ITD}}$ was calculated using (4.5). Figure 5.7 displays the solid angle weighted $\overline{\Delta\text{ITD}}$ values between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute $\Delta$ITD value. This shows a small reduction in accuracy of ITD rendering when DBE is implemented for $M = 1$ and $M = 2$. However, this is at the highest tested amount of bias and is only marginally less accurate than the ITD reported with diffuse-field equalisation in Figure 4.17. Detailed plots of $\Delta$ITD for every measurement location on the sphere are presented in Appendix B.9 for binaural

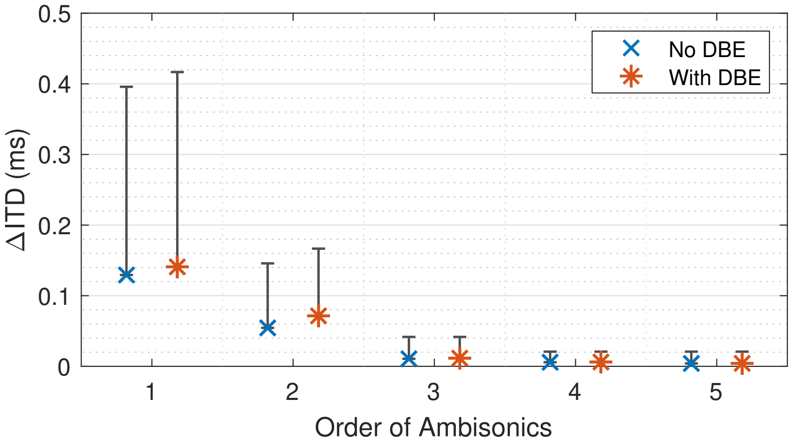


FIGURE 5.7: Estimated $\overline{\Delta\text{ITD}}$ between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔITD value.

Ambisonic rendering with DBE at $\kappa = 33$, which show the close similarity between renders with DFE as shown in Appendix A.1.

The change in ILD between the reference HRTFs and the Ambisonic generated HRTFs, ΔILD, was calculated for each measurement location $q$ using (4.6), and the solid angle weighted $\overline{\Delta\text{ILD}}$ values using (4.7). Figure 5.8 displays the solid angle weighted $\overline{\Delta\text{ILD}}$ values between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute ΔILD value. Though the maximum value of ΔILD is increased with DBE for $M = 2$ and $M = 3$, there is still not a substantial effect. Detailed plots of ΔILD for every measurement location on the sphere are presented in Appendix B.10 for binaural Ambisonic rendering with DBE at $\kappa = 33$, which show the close similarity between renders with DFE as shown in Appendix A.2.

Observation of these figures confirms that, even at high directional bias, DBE produces a similar minimal effect to DFE on the Ambisonic rendering of interaural cues when the direction of bias is in front.

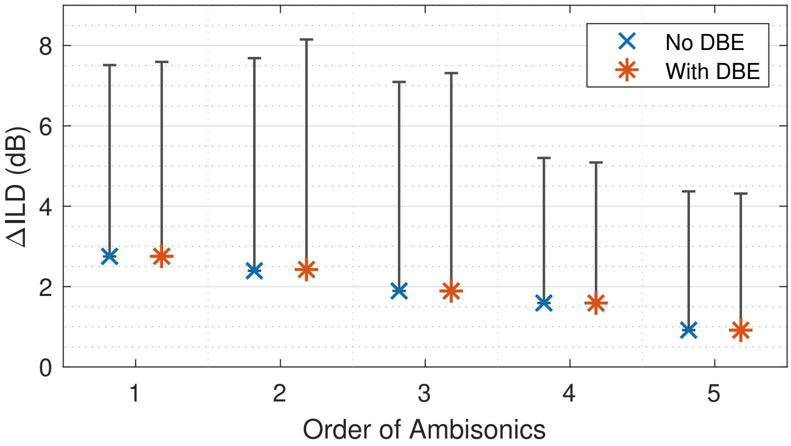


FIGURE 5.8: $\overline{\Delta\text{ILD}}$ between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔILD value.

### 5.2.3 Estimated Localisation

The effect of DBE with frontal bias on horizontal localisation of binaural Ambisonic rendering was estimated using the method detailed in Section 2.8.4, utilising a horizontal model (May, Van De Par and Kohlrausch, 2011), producing a value of $\overline{E_\theta}$ for overall estimated localisation between $-90° < \theta < +90°$ at $\phi = 0°$ using (2.17). Figure 5.9 displays the overall estimated horizontal localisation of binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$. Appendix B.11 presents detailed individual plots for the estimated horizontal localisation of each azimuth angle, which show accurate horizontal localisation at the direction of bias.

Observation of these figures suggests that DBE has a small effect on the accuracy of estimated horizontal localisation, with DBE causing a lower accuracy for $M = 1$ and $M = 2$. However, increases in accuracy are shown at $M = 3$, $M = 4$ and $M = 5$. Overall these figures are not dissimilar to the effect of DFE (see again Figure 4.19). These results agree with the ITD estimations, which are poorer for $M = 1$ and $M = 2$ and largely unchanged for $M = 3$, $M = 4$ and $M = 5$. It is important to note that these are at a very high directional bias ($\kappa = 33$). The explanation for relatively

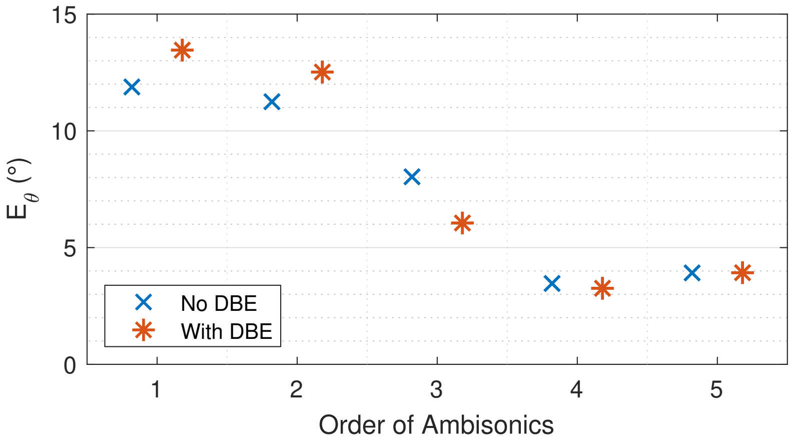
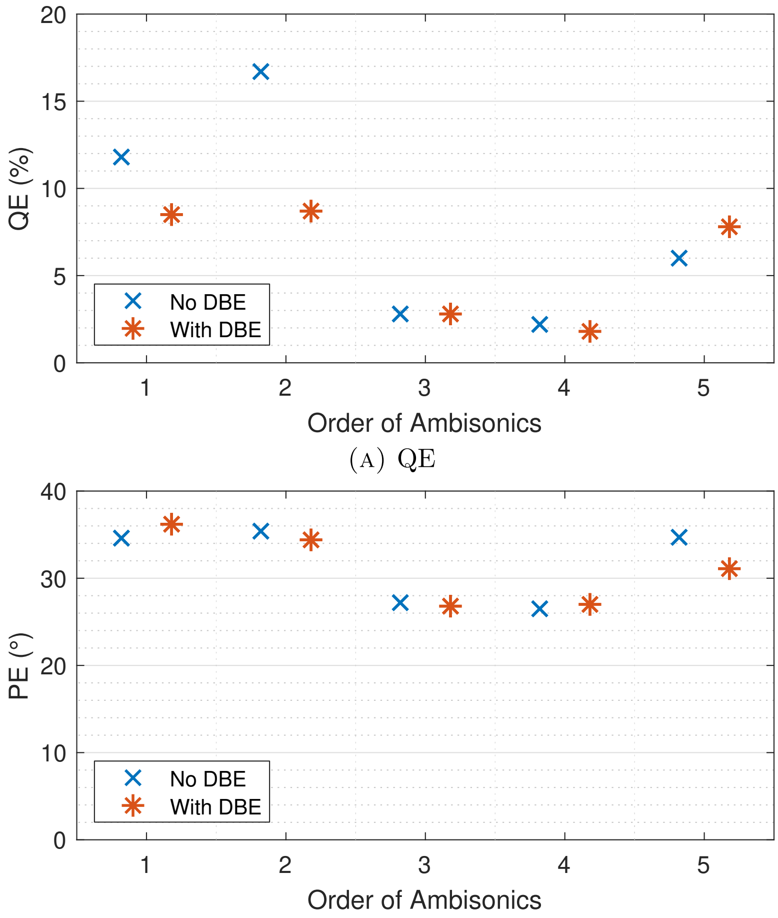


Figure 5.9: Estimated $\overline{E_\theta}$ of binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (May, Van De Par and Kohlrausch, 2011).

small effect on predicted horizontal localisation with DBE at a high directional bias is likely due to the largely unchanged rendering of interaural cues.

The effect of DBE on estimated elevation localisation in the sagittal plane was evaluated between $-90° < \phi < +90°$ at $\theta = 0°$ using the method detailed in Section 2.8.4 which utilises a localisation model (Baumgartner, Majdak and Laback, 2014) to produce two psychoacoustic performance metrics: quadrant error (QE), a prediction of localisation confusion (presented as a percentage), and polar RMS error (PE), a prediction of precision and accuracy in degrees. As the HRTFs used are of a Neumann KU 100, which has no torso, there will be no elevation cues present below 1.5 kHz (Algazi, Avendano and Duda, 2001a). Therefore, the frequency range of the model's filter bank is set to 1.5 kHz - 18 kHz, with the upper limit of the frequency range chosen as the highest frequency of perceivable elevation cues (Roffler and Butler, 1967; Asano, Suzuki and Sone, 1990).

Figure 5.10 shows the predicted QE and PE values of binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$. Detailed individual plots of predicted sagittal plane localisation are presented in Appendix B.12, which in general show an improvement in predicted elevation localisation accuracy at the direction of bias but reduced accuracy elsewhere. A curious observation is the effect of DBE on $M = 5$, which produces an increase in QE but a decrease in PE. Studying



(A) QE

(B) PE

Figure 5.10: Estimated sagittal plane localisation of binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (Baumgartner, Majdak and Laback, 2014).

the elevation plot of Ambisonic rendering with DBE at $M = 5$ in Appendix B.12e, there appears to be more up-down confusion at $\phi > 50°$ but a more even diagonal at other elevations with DBE.

### 5.2.4 Generalisability

To demonstrate the generalisability of DBE, additional numerical evaluation of binaural Ambisonic rendering was performed using different loudspeaker configurations and an alternative HRTF dataset. In both cases, the effect of DBE was assessed in terms of PSD using the method detailed in Section 4.2, with PSD calculations

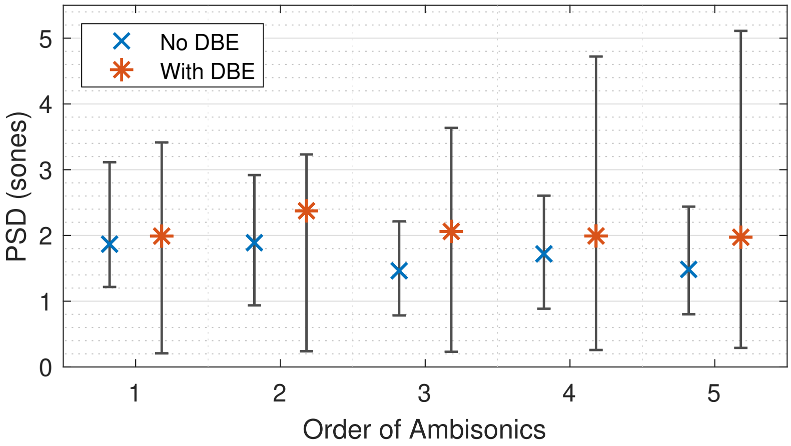


FIGURE 5.11: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the minimum and maximum absolute PSD values.

comparing binaural Ambisonic renders to a reference dataset of HRTFs for all available measurement locations at a single bias factor of $\kappa = 33$ and direction of bias $(\theta = 0°, \phi = 0°)$.

Firstly, different loudspeaker configurations were investigated[1]. Here, spherical T-designs (Hardin and Sloane, 1996) were employed with corresponding loudspeaker vertices as illustrated in Figure 3.8. The DBQ RMS responses, frontal bias HRTFs, and resulting final equalisation filters for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations are presented in Appendix B.13.

The solid angle weighted $\overline{\text{PSD}}$ results, calculated using (4.3) from 16,020 locations on the sphere, are shown in Figure 5.11, along with the maximum and minimum absolute PSD values. The trend is similar to that observed with Lebedev loudspeaker configurations, illustrating how DBE produces a reduction in the minimum PSD value, regardless of the type of loudspeaker configuration, and a higher average and maximum PSD value. Detailed plots of PSD for every measurement location on the sphere are presented in Appendix B.14, which also follow similar trends to those shown in the Lebedev loudspeaker configurations with improved spectral reproduction at the direction of bias, and reduced accuracy at lateral locations.

[1] For all other areas of this chapter, Lebedev loudspeaker configurations are used.

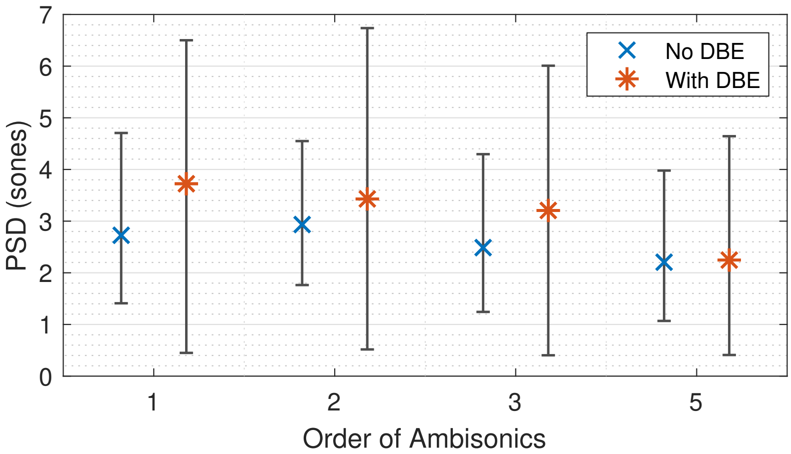


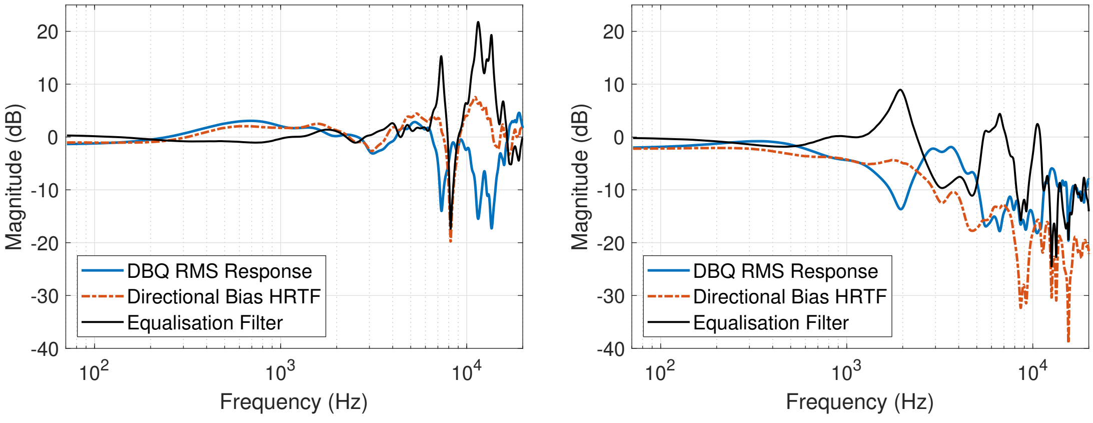
FIGURE 5.12: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the minimum and maximum absolute PSD values.

Secondly, to assess the effect of DBE when using an alternative HRTF dataset, binaural Ambisonic renders were made for $\{M = 1, M = 2, M = 3, M = 5\}$[2] using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, human subject H20 (Armstrong et al., 2018a). The DBQ RMS responses, frontal bias HRTFs, and resulting final equalisation filters of Ambisonic orders $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs are presented in Appendix B.15.

The solid angle weighted $\overline{\text{PSD}}$ results, calculated from 2,114 locations on the sphere, are shown in Figure 5.12 along with the minimum and maximum absolute PSD values, with detailed plots of PSD for every measurement location on the sphere presented in Appendix B.16. These show again how DBE greatly improves spectral reproduction at the direction of bias, and reduces accuracy at other directions, for all tested orders of Ambisonics.

The tests on generalisability therefore show how DBE is applicable to binaural Ambisonic rendering using different virtual loudspeaker configurations and alternative HRTF datasets. Increased directional bias improves the spectral reproduction of

[2] The omission of $M = 4$ is due to a lack of necessary measurements.



(A) Left ear

(B) Right ear

FIGURE 5.13: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering at a bias direction of $(\theta = 90°, \phi = 0°)$, $\kappa = 33$, $M = 1$.

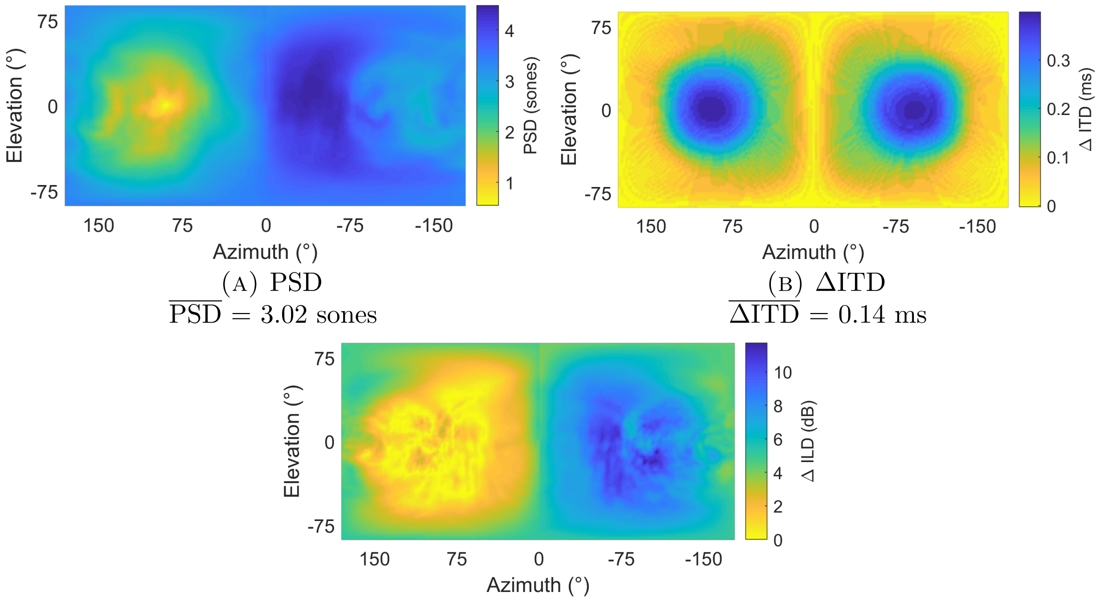
binaural Ambisonic rendering at the direction of bias, though this comes at the expense of reduced spectral accuracy at other directions.

### 5.2.5 Varying the Direction of Bias

To evaluate the effect of DBE with other directions of bias, DBE filters are calculated for $M = 1$ using a direction of bias $(\theta = 90°, \phi = 0°)$, at $\kappa = 33$. The DBQ RMS responses, frontal bias HRTFs, and resulting equalisation filters are presented in Figure 5.13. This illustrates the asymmetrical nature of DBE filters at a bias direction outside the median plane.

The effect of a bias direction at $(\theta = 90°, \phi = 0°)$ is evaluated numerically using PSD, ΔITD and ΔILD calculations between HRTFs and binaural Ambisonic rendering for 16,020 measurement locations on the sphere, which are presented in Figure 5.14.

Observations of the plots shows that a bias direction of $(\theta = 90°, \phi = 0°)$ still improves PSD at the direction of bias. ILD is improved for the hemisphere in the direction of bias, while greater ILD error arises in the contralateral hemisphere. ILD is also reproduced poorly on the median plane, which is likely due to the asymmetrical nature of the DBE filters (see again Figure 5.13). ITD remains largely unchanged, which is likely due to the linear phase nature of the filters.



(A) PSD
$\overline{\text{PSD}} = 3.02$ sones

(B) ΔITD
$\overline{\Delta\text{ITD}} = 0.14$ ms

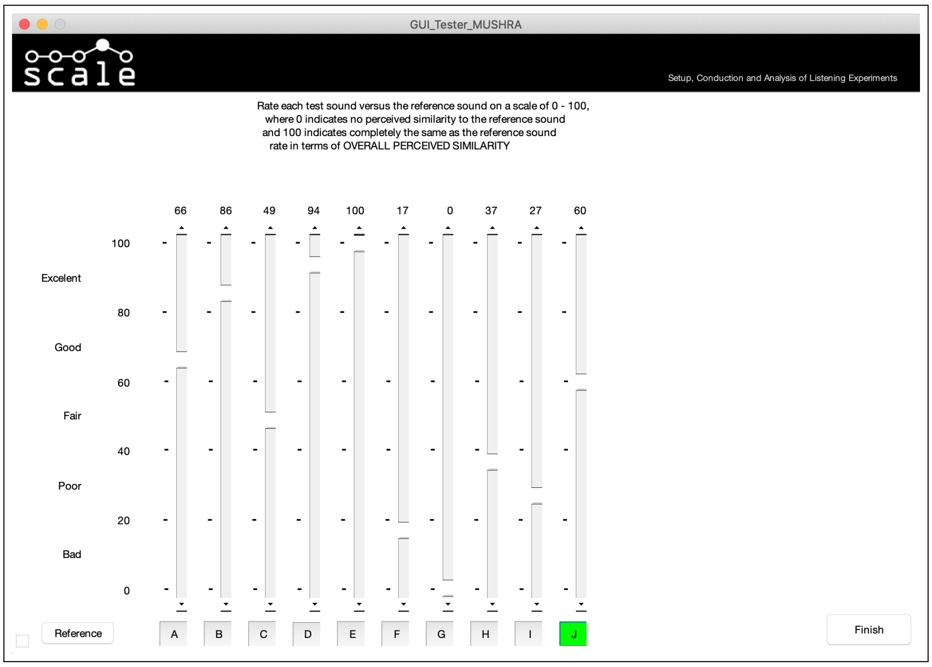
(C) ΔILD
$\overline{\Delta\text{ILD}} = 4.67$ dB

FIGURE 5.14: PSD (mean of left and right values), ΔITD and ΔILD between HRTFs and binaural Ambisonic rendering with DBE at a bias direction of $(\theta = 90°, \phi = 0°)$, $\kappa = 33$, for every measurement location on the sphere, $M = 1$.

## 5.3 Perceptual Evaluation

To assess the perceptual effect of DBE, listening tests were conducted using both simple and complex acoustic scenes for binaural Ambisonic rendering at $M = 1$, $M = 3$ and $M = 5$, with $M = 2$ and $M = 4$ omitted to reduce the overall duration of the listening tests, and Lebedev loudspeaker configurations. The direction of bias is $(\theta = 0°, \phi = 0°)$ and the amounts of bias investigated are $\kappa = 1, 3, 5, 9, 17$ and 33.

Tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 (International Telecommunication Union, 2015b), and were conducted in a quiet listening room (background noise level of 41.8 dBA) using a single set of Sennheiser HD 650 circumaural headphones and an Apple Macbook Pro with a Fireface 400 audio interface, which has software controlled input and output levels. A screenshot of the MUSHRA interface, using the MATLAB based 'Scale' (Giner, 2013), is presented in Figure 5.15. The headphones were equalised using a Neumann KU 100 as in Section 4.4. The RMS HpTF and inverse filter of



Rate each test sound versus the reference sound on a scale of 0 - 100,
where 0 indicates no perceived similarity to the reference sound
and 100 indicates completely the same as the reference sound
rate in terms of OVERALL PERCEIVED SIMILARITY

FIGURE 5.15: Screenshot of the MUSHRA interface used in the Ambisonic DBE listening test using 'Scale' (Giner, 2013).

the left HD 650 headphone, along with a resulting convolved response, are shown in Figure 4.24. 20 experienced listeners participated, aged between 22 and 41 (17 male, 3 female), with no reported knowledge of any hearing impairments according to ISO Standard 389 (International Organization for Standardization, 2016). All reported prior critical listening experience, which was deemed sufficient if the participant had education or employment in audio or music engineering.

The base stimulus was one second of monophonic pink noise at a sample rate of 48 kHz, windowed by onset and offset half-Hanning ramps of 5 ms. Each test sound was generated by convolving the pink noise with an HRTF, either Ambisonic or not. Test sounds were normalised to a consistent A-weighted RMS amplitude and participants were able to adjust the playback level. All binaural renders were static (fixed head orientation) to ensure consistency in the experience between participants.

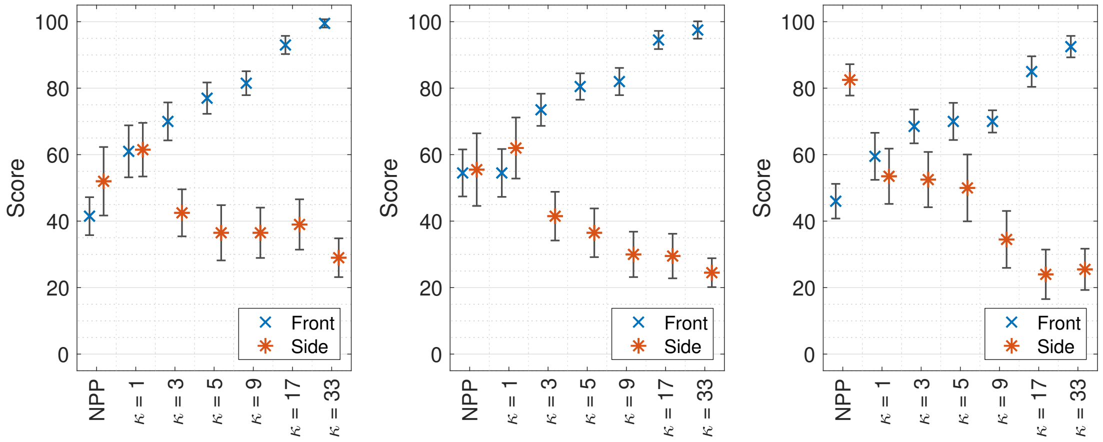
### 5.3.1 Test Paradigms

The simple scenes test comprised of a single pink noise source. Two sound source locations were used: directly in front of the listener at $(\theta = 0°, \phi = 0°)$, and directly to the left of the listener at $(\theta = 90°, \phi = 0°)$. The reference was a direct HRTF convolution, and low and mid anchors were the reference low-passed at 3.5 kHz and 7 kHz, respectively, giving a total of 10 conditions for each scenario. Participants were asked to rate each condition in overall perceived similarity to the reference with a score between 0 and 100. Stimuli and trial ordering was randomised and presented double blind.

The complex scene was simulated by mixing a pink noise burst with a diffuse soundscape. The noise burst consisted of half a second of pink noise followed by half a second of silence panned directly in front of the listener. The diffuse soundscape was synthesised from 24 excerpts of a monophonic sound scene recording of a train station (Green and Murphy, 2017), each 5 seconds long. The sound scene excerpts were panned to the vertices of an $L = 24$ spherical T-design quadrature (as illustrated in Figure 3.8c), to ensure minimal overlap between the positions of virtual loudspeakers in the binaural decoders and the sound sources in the complex scene. The frontal noise was set 3 dB RMS louder than the diffuse soundscape to approximate a centre of attention. The reference comprised of a sum of direct HRTF convolutions and the anchor an $M = 0$ Ambisonic render, giving a total of 9 conditions per trial. All test trials were repeated once. Participants were asked to rate each condition in overall perceived similarity to the reference with a score between 0 and 100. Stimuli and trial ordering was again randomised and presented double blind.

### 5.3.2 Results

Overall, the tests lasted between around 20 to 45 minutes to complete. No participant's results were excluded, based on the criteria of rating the hidden reference less than 90% for more than 15% of trials or rating the mid-anchor higher than



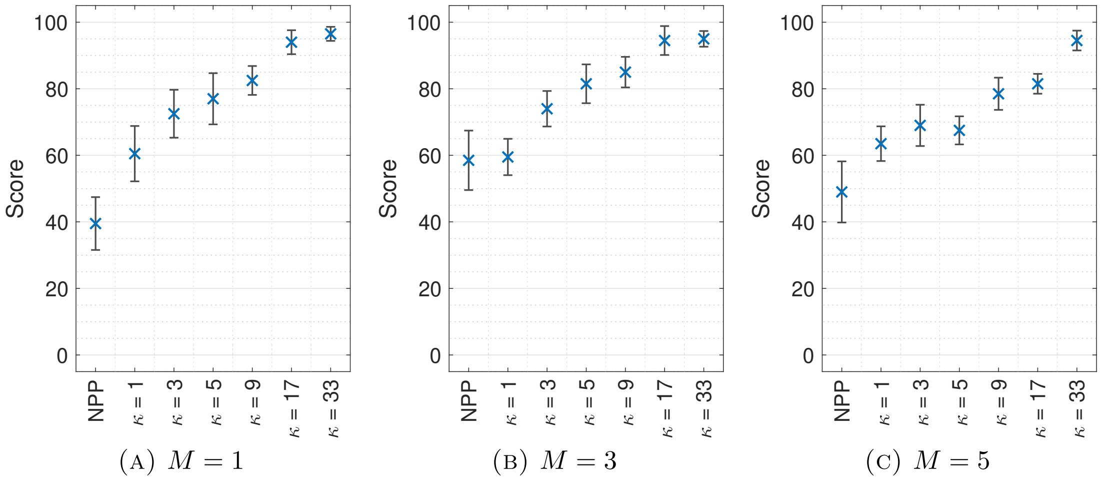
(A) $M = 1$ (B) $M = 3$ (C) $M = 5$

FIGURE 5.16: Median scores of the simple scene tests with non-parametric CI$_{95}$, reference and anchor scores omitted. Score indicates perceived similarity to the HRTF reference. NPP denotes no pre-processing.

90% for more than 15% of trials. Results data was tested for normality using the one-sample Kolmogorov-Smirnov test, which showed all data as non-normal. Results were therefore analysed using non-parametric statistics.

Figure 5.16 presents the simple scene median scores for Ambisonic orders $M = 1$, $M = 3$ and $M = 5$ with non-parametric 95% confidence intervals (CI$_{95}$) (Mcgill, Tukey and Larsen, 1978), calculated using (4.8). A Friedman's ANOVA, conducted on simple scene data from all orders and sound source locations, confirmed that changing the value of $\kappa$ had a highly statistically significant effect on the perceived similarity to the HRTF reference ($\chi^2(6) = 27.22, p < 0.01$). The results support the theory that increasing $\kappa$ improves the perceived similarity to the HRTF reference for the frontal stimuli for all three tested orders of Ambisonics, and reduces the similarity for the lateral stimuli. This shows that DBE performs as expected with simple scenes.

Figure 5.17 presents the complex scene median scores for for Ambisonic orders $M = 1$, $M = 3$ and $M = 5$ with non-parametric CI$_{95}$. A Friedman's ANOVA, conducted on complex scene data from all orders, again confirmed that changing the value of $\kappa$ had a highly statistically significant effect on the perceived similarity to the HRTF reference ($\chi^2(6) = 383.47, p < 0.01$). The results suggest that the diffuse sound is



(A) $M = 1$ (B) $M = 3$ (C) $M = 5$

Figure 5.17: Median scores of the complex scene tests with non-parametric $\text{CI}_{95}$, reference and anchor scores omitted. Score indicates perceived similarity to the HRTF reference. NPP denotes no pre-processing.

essentially ignored, as results for the complex scene are similar to the frontal stimuli in the simple scene, with increasing $\kappa$ producing a higher perceived similarity to the HRTF references for all three tested orders of Ambisonics.

## 5.4 Discussion

The numerical evaluation shows that Ambisonic DBE is capable of improving the accuracy of spectral reproduction of binaural Ambisonic rendering at a specified direction of bias, when compared to direct HRTF rendering. This comes at the expense of spectral reproduction at other locations, as observed in Figures 5.5 and 5.14. When the direction of bias is $(\theta = 0°, \phi = 0°)$, DBE has a small effect on the rendering of ITD and ILD, though this is not dissimilar to that of Ambisonic DFE. Anecdotally, the implementation of Ambisonic DBE produces a much more accurate timbral reproduction in the direction of bias. At other directions, audible ringing artefacts are present, the frequencies of which occur around the sharp peaks observed in the DBE equalisation filter plots in Figure 5.4.

The perceptual evaluation of Ambisonic DBE supports the numerical evaluation, with results suggesting increased directional bias improves the perceived timbral similarity between binaural Ambisonic rendering and HRTF convolution for the direction of bias while reducing timbral similarity for other directions (see again Figure 5.16). However when the acoustic scene is complex, with a frontal main source and additional diffuse sources, increased frontal bias still improves overall similarity (Figure 5.17). This suggests that, if the main sound source is at the front, one can afford to increase $\kappa$ without greatly reducing the perceived quality in lateral directions. As human auditory localisation is more accurate in front (see Figure 2.15), it is possible that sensitivity to timbral changes is also more finely tuned to the front, though this theory requires more investigation.

Without knowing the directional content of the input signal, it is not possible to recommend a single direction or amount of bias. However, if the dominant direction of the signal could be estimated, for example using a method such as Directional Audio Coding (Pulkki, 2007), the equalisation could be performed specifically for the direction of arrival, such as other signal dependent Ambisonic decoding methods (Politis, McCormack and Pulkki, 2017; Schörkhuber and Höldrich, 2019). This is a possible future application of Ambisonic DBE.

In head-tracked reproduction scenarios, such as in virtual reality, a possible use of Ambisonic DBE would be to improve the reproduction accuracy of the frontal region, as this can take advantage of the fixed nature of the virtual loudspeaker array in binaural Ambisonic rendering, whereby the direction of bias would be fixed in front of the listener, regardless of the listener's head orientation. This would improve the accuracy of scene rendering in the direction in which the user's head faced in a 'perceptual auditory spotlight'. Further investigation would be necessary to consider the viability of this.

## 5.5 Summary

Ambisonic Diffuse-Field Equalisation was shown in Chapter 4 to improve the spectral reproduction of low-order binaural Ambisonic rendering. However, there still exist notable differences between diffuse-field equalised binaural Ambisonic rendering and HRTF convolution. This chapter has built on Ambisonic DFE by adapting the technique with the introduction of a directional bias, such that the resulting equalisation produces a greater improvement in the spectral reproduction of binaural Ambisonic rendering at a specified direction. At high directional bias, this equalisation can produce binaural Ambisonic spectral reproduction highly similar to direct HRTF rendering in the direction of bias, though this comes at the detriment of spectral reproduction elsewhere. This technique is called Ambisonic Directional Bias Equalisation (DBE).

The numerical evaluation of DBE has shown that increasing the amount of bias improves spectral reproduction and predicted height localisation at the direction of bias, while producing little effect on the estimated interaural cues and horizontal localisation accuracy. The accuracy of spectral reproduction at other directions is reduced as a trade off. Perceptual evaluation in the form of listening tests on timbre has corroborated the numerical results, producing statistically significant results. When the direction of bias is in front and a dominant sound source is in the same direction, increased directional bias has been shown to produce an improvement in overall perceived similarity to HRTF convolution renders, even with diffuse background sounds at other directions. Input signal source direction estimation techniques could inform a future implementation of the DBE technique such that the direction of bias follows the estimated source direction.

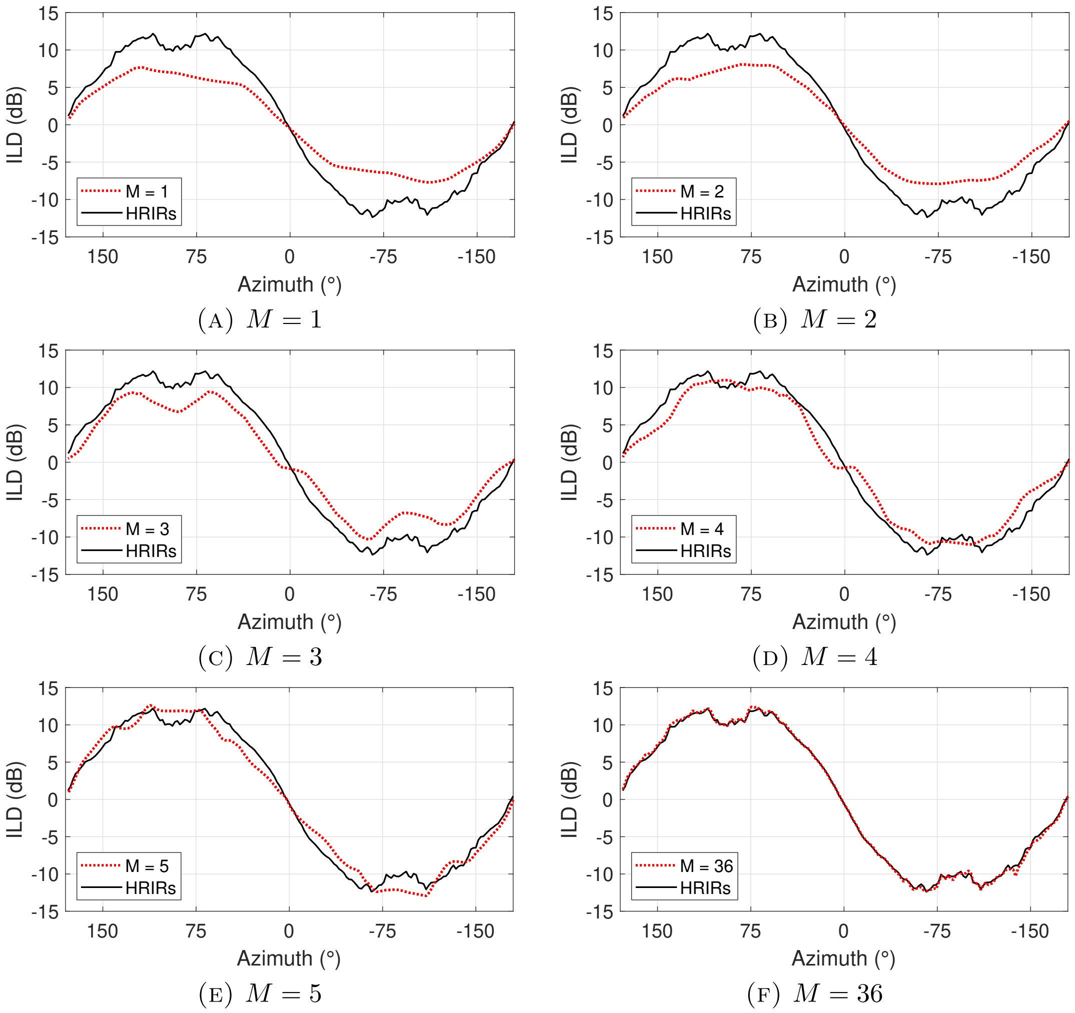
This chapter has shown that a directional equalisation HRTF pre-processing stage can produce an improvement in the high frequency reproduction of binaural Ambisonic rendering for a specific direction, bringing the spectral reproduction closer to HRTF convolution even at $M = 1$, though this comes at the expense of spectral accuracy at other directions. The coming chapters will look at other possible ways of improving the high frequency reproduction of binaural Ambisonic rendering, addressing the poor ILD reproduction of low Ambisonic orders.

# Chapter 6

# Ambisonic Interaural Level Difference Optimisation

Chapters 4 and 5 have presented HRTF pre-processing equalisation techniques for improving the spectral reproduction of virtual loudspeaker binaural Ambisonic rendering at frequencies above $f_{\text{alias}}$. However, there are other ways in which low-order binaural Ambisonic rendering is inaccurate, when compared to HRTFs, even at $M = 5$.

Ambisonic rendering of interaural level difference (ILD) has been shown to be poor at $M = 1$ and $M = 2$ (Daniel, Rault and Polack, 1998; Wiggins, Paterson-Stephens and Schillebeeckx, 2001; Kearney, 2010). To illustrate this for other Ambisonic orders, the ILD of binaural Ambisonic rendering has been estimated using (2.14) for measurement locations on the horizontal plane. Figure 6.1 presents the estimated ILD of HRTFs and binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5\}$, with $M = 36$ included for reference, calculated using (2.14). The reduced ILD reproduction of low-order Ambisonics is evident, and becomes less pronounced at higher orders of Ambisonics. At $M = 36$, ILD reproduction almost perfectly follows that of HRTFs. However, an interesting observation is that low-order Ambisonics does not just reproduce ILD at lower levels than HRTFs: there are regions where the



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$ (F) $M = 36$

Figure 6.1: Estimated horizontal plane ILD of HRTFs and binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5$ and $M = 36\}$.

Ambisonic reproduction of ILD is too high, such as $\theta = \pm 90°$ for $M = 5$. This may be due to the slightly irregular distribution of loudspeakers in Lebedev configurations.

One previous attempt to improve ILD reproduction of binaural Ambisonic rendering was by Collins, who introduced additional virtual loudspeakers at the lateral positions in the loudspeaker configuration (Collins, 2013). However, this caused localisation issues and poorer spectral reproduction due to increased comb filtering from the higher number of virtual loudspeakers (Yao, Collins and Jančovič, 2015).

This chapter presents a novel method for addressing the inadequate ILD reproduction of low-order binaural Ambisonic rendering through a pre-processing stage of the HRTFs used in the binaural rendering of Ambisonic signals. This technique is referred to as Ambisonic ILD Optimisation (AIO). AIO is evaluated both numerically and perceptually, with the numerical evaluation comparing binaural Ambisonic rendering with and without AIO to a reference set of HRTFs in terms of perceptual spectral difference, estimated ILD and ITD cues and predicted localisation accuracy for $\{M = 1, M = 2, ..., M = 5\}$. The ILD estimation evaluation investigates ILD reproduction over both sound source location and different frequency bands. The perceptual evaluation is presented through listening tests using both simple and complex acoustic scenes. Finally, results are discussed and further developments are discussed, along with a recommendation on whether or not AIO should be implemented in binaural Ambisonic rendering.

## 6.1 Method

This section describes the method for optimising the ILD reproduction of low-order binaural Ambisonic rendering. A block diagram of the method is presented in Figure 6.2, and a brief summary of the method is as follows: for each virtual loudspeaker in the configuration, binaural Ambisonic renders are created, and the ILD is estimated and compared to that of the original virtual loudspeaker HRTF. The amplitude of the ipsilateral and contralateral virtual loudspeaker HRTF signals is then augmented at high frequencies such that when the augmented virtual loudspeaker HRTFs are used to render Ambisonic signals, the ILD reproduction is closer to the original HRTFs. AIO is an HRTF pre-processing technique that must be implemented separately for each order of Ambisonics and each loudspeaker configuration. All computation was carried out offline in Matlab version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4.

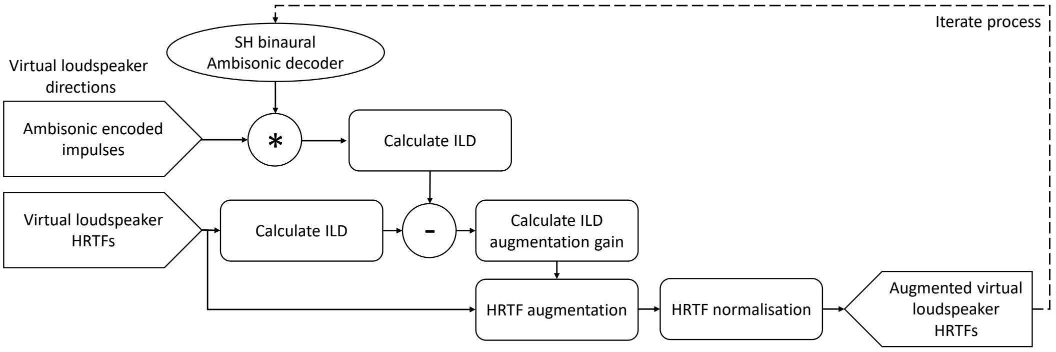

Iterate process
SH binaural Ambisonic decoder
Virtual loudspeaker directions
Ambisonic encoded impulses
*
Calculate ILD
Virtual loudspeaker HRTFs
Calculate ILD
-
Calculate ILD augmentation gain
HRTF augmentation
HRTF normalisation
Augmented virtual loudspeaker HRTFs

FIGURE 6.2: Block diagram of the AIO method.

### 6.1.1 Ambisonic ILD Comparison

For each loudspeaker in the configuration, an Ambisonic rendered HRTF is generated as in (3.32), by using $\delta(t)$ from (2.5) as the monophonic input signal. Figures 6.3a and 6.3b presents the virtual loudspeaker HRTF and Ambisonic rendered HRTF at $(\theta = 90°, \phi = 45°)$ corresponding to $l = 5$, $M = 1$. The ILD is estimated for both the Ambisonic HRTF and the original virtual loudspeaker HRTF using (2.14), and the difference in ILD between the two estimations is calculated as

$$\Delta\mathrm{ILD} = |\mathrm{ILD}(H)| - |\mathrm{ILD}(\hat{H})| \tag{6.1}$$

where $H$ refers to the original virtual loudspeaker HRTF (and thus $\mathrm{ILD}(H)$ is the target ILD) and $\hat{H}$ refers to the Ambisonic rendered HRTF. Note that (6.1) differs from (4.4) in that the change in ILD is here taken from the absolute values of both calculations, as opposed to the absolute value of the difference, as this allows for negative $\Delta$ILD values, which occur in the case that $\mathrm{ILD}(H) < \mathrm{ILD}(\hat{H})$. This means that if the Ambisonic rendered ILD is too large, it can be reduced.

### 6.1.2 Virtual Loudspeaker HRTF Augmentation

The next stage in AIO is to augment the virtual loudspeaker HRTF signals used in the binaural Ambisonic decoder above $f_{\mathrm{alias}}$. At loudspeakers where $\Delta\mathrm{ILD} > 0$, the ILD of the virtual loudspeaker HRTFs is increased, and at loudspeakers where $\Delta$ILD $< 0$, the ILD of the virtual loudspeaker HRTFs is decreased.

As ILD is calculated in dB, $\Delta$ILD from (6.1) is then converted to a gain value by the inverse of the dB SPL calculation, such that

$$g^{\Delta} = 10^{\frac{\Delta\mathrm{ILD}}{20}} \tag{6.2}$$

where ILD augmentation is conditional on the loudspeaker being situated away from the median plane, thus $g^{\Delta} = 1$ if $\theta_l = 0°$ or $\theta_l = 180°$.

This process is repeated for all loudspeakers in the configuration, and an array of $g^{\Delta}$ values is produced with $L$ length as $\mathbf{G}^{\mathbf{\Delta}} = \{g_1^{\Delta}, g_2^{\Delta}, ..., g_L^{\Delta}\}$.

The virtual loudspeaker HRTFs are augmented by applying the gains $g^{\Delta}$ to the contralateral signal of the HRTF for each loudspeaker, where values of $g^{\Delta} > 1$ produce an increase in ILD and values of $g^{\Delta} < 1$ produce a reduction in ILD of the unprocessed HRTFs as follows:

$$\begin{aligned} H_{\mathrm{left}}^{\mathrm{AIO}} &= \frac{H_{\mathrm{left}}}{g^{\Delta}}, \quad \text{if } \mathrm{ILD}(H) > 0 \\ H_{\mathrm{right}}^{\mathrm{AIO}} &= \frac{H_{\mathrm{right}}}{g^{\Delta}}, \quad \text{if } \mathrm{ILD}(H) < 0 \end{aligned} \tag{6.3}$$

where $H^{\mathrm{AIO}}$ is an augmented virtual loudspeaker HRTF. The ipsilateral signal of each HRTF remains unchanged ($H_{\mathrm{ips}}^{\mathrm{AIO}} = H_{\mathrm{ips}}$), as is the case for both signals of the HRTF if $g^{\Delta} = 1$.

### 6.1.3 Virtual Loudspeaker HRTF Normalisation

Each augmented virtual loudspeaker HRTF is then normalised to the same root-mean square (RMS) amplitude as the original virtual loudspeaker HRTF by

$$H^{\mathrm{AIO}} \times \frac{\mathrm{RMS}(H)}{\mathrm{RMS}(H^{\mathrm{AIO}})} \tag{6.4}$$

where RMS($H$) denotes the RMS amplitude of an HRTF, calculated as the arithmetic mean of the RMS of the left and right signals of the HRTF in the time domain.

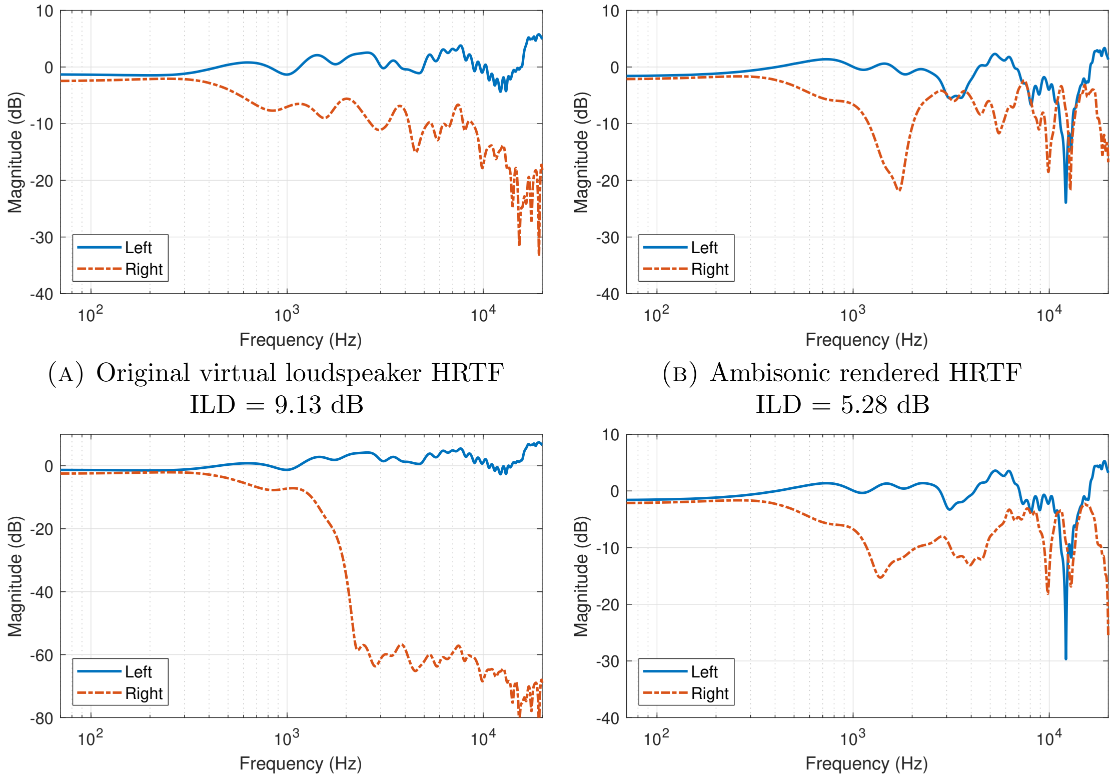
The augmented virtual loudspeaker HRTFs with AIO are then combined with the original virtual loudspeaker HRTFs using the same linear-phase crossover network as used in the dual-band decode design in Section 3.3.3, such that the resulting HRTFs are the same number of samples and RMS amplitude as the original HRTFs, identical at low frequencies, but with ILD augmentation at high frequencies. The cut off frequency of the filter $f_c$ is based on $f_{\mathrm{alias}}$ for the condition $f_{\mathrm{alias}} > 1.5$ kHz. This is due to ILDs being largely perceptually irrelevant below 1.5 kHz (Middlebrooks and Green, 1991). Therefore at $M = 1$ and $M = 2$, $f_c = 1.5$ kHz, and at $M \geq 3$, $f_c = f_{\mathrm{alias}}$.

### 6.1.4 Iteration

The pre-processed HRTFs are then switched into (3.31), and the process from Section 6.1.1 to 6.1.3 is repeated iteratively whereby the array of $g^{\Delta}$ values is taken as the product of the $g^{\Delta}$ values from each iteration $i$:

$$\mathbf{G}^{\boldsymbol{\Delta}} = \prod_{i=1}^{I} \mathbf{G}^{\boldsymbol{\Delta}}(i) \tag{6.5}$$

where $i = i + 1$ for each iteration. This ensures that the final AIO pre-processed HRTF dataset will be subject to the crossover filter only once, regardless of the number of iterations. The iteration runs until $\overline{\prod \mathbf{G}^{\boldsymbol{\Delta}}(i)} \approx \overline{\prod \mathbf{G}^{\boldsymbol{\Delta}}(i-1)}$ is satisfied to an accuracy of 5 significant figures, where the overline denotes arithmetic mean. Implementing AIO pre-processing as an iterative process also allows the consideration that changes in ILD to one virtual loudspeaker may influence the resulting ILD of other loudspeakers in the configuration. Figures 6.3c and 6.3d present the



(A) Original virtual loudspeaker HRTF
ILD = 9.13 dB

(B) Ambisonic rendered HRTF
ILD = 5.28 dB

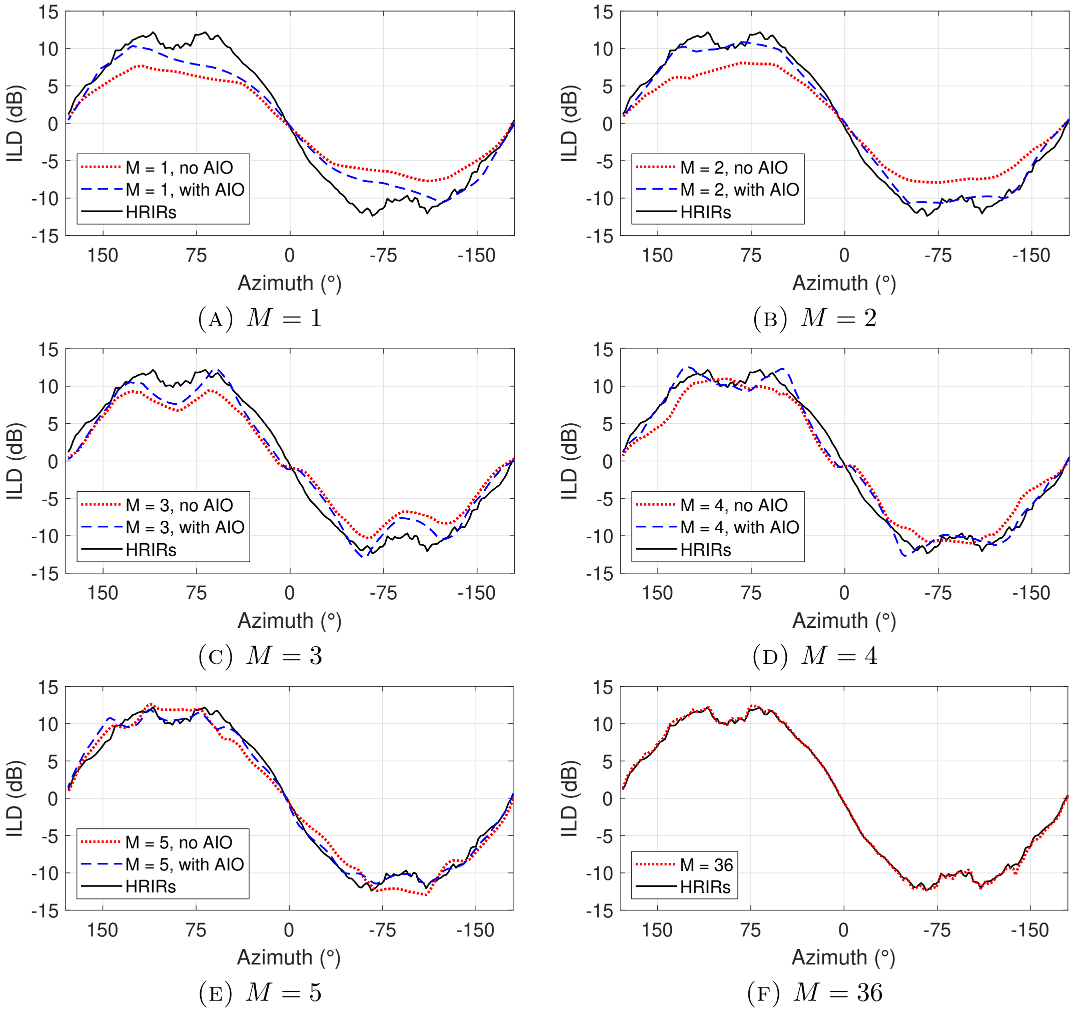
(C) AIO augmented virtual loudspeaker HRTF
ILD = 36.3 dB

(D) Ambisonic rendered HRTF with AIO
ILD = 6.34 dB

FIGURE 6.3: Comparing virtual loudspeaker HRTFs with Ambisonic rendered HRTF at ($\theta = 90°, \phi = 45°$) corresponding to $l = 5$, $M = 1$, with and without AIO.

augmented virtual loudspeaker HRTF and corresponding Ambisonic rendered HRTF ($\theta = 90°, \phi = 45°$) of $l = 5$ at $M = 1$, respectively. For this configuration, 18 iterations have been undertaken. Note how the Ambisonic reproduced ILD is improved with AIO, particularly in the region around 3 kHz.

Figure 6.4 presents the estimated ILD of binaural Ambisonic rendering and HRTFs on the horizontal plane, for $\{M = 1, M = 2, ..., M = 5\}$, both with and without AIO. The $M = 36$ (without AIO) is included again for reference. The figure shows how horizontal ILD reproduction is greatly improved with the implementation of AIO, producing values of ILD closer to those of HRTFs for the majority of locations on the horizontal plane. Though for the most part AIO produces an increase in reproduced ILD of binaural Ambisonic rendering (especially at $M = 1$ and $M = 2$), the renders at $M = 4$ and $M = 5$ illustrate how AIO can also produce a reduction in



(A) $M = 1$

(B) $M = 2$

(C) $M = 3$

(D) $M = 4$

(E) $M = 5$

(F) $M = 36$

FIGURE 6.4: Estimated horizontal plane ILDs of HRTFs and binaural Ambisonic rendering, with and without AIO for $\{M = 1, M = 2, ..., M = 5 \text{ and } M = 36\}$.

ILD for some locations when necessary - see azimuth values between $|75° < \theta < 105°|$ in Figures 6.4d and 6.4e, respectively.

## 6.2 Numerical Evaluation

In this section, the effect of AIO is evaluated numerically by comparing binaural Ambisonic renders, with and without AIO, to a reference dataset of HRTFs. The metrics of evaluation include perceptual spectral difference, interaural level difference, both over all directions on the sphere and over frequency, interaural time difference

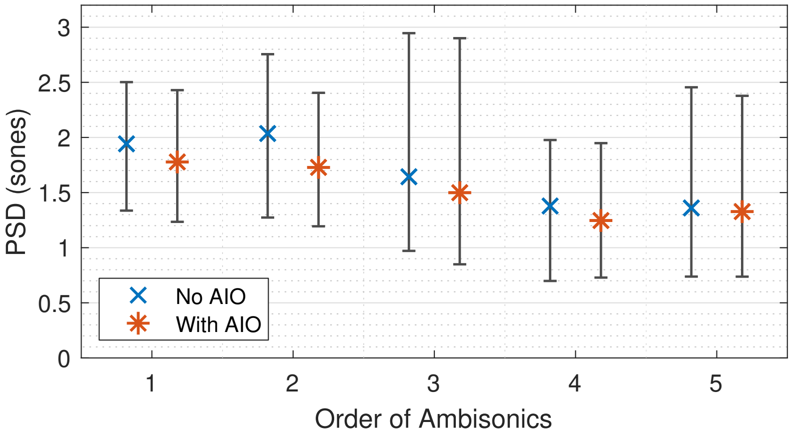


FIGURE 6.5: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum PSD value.

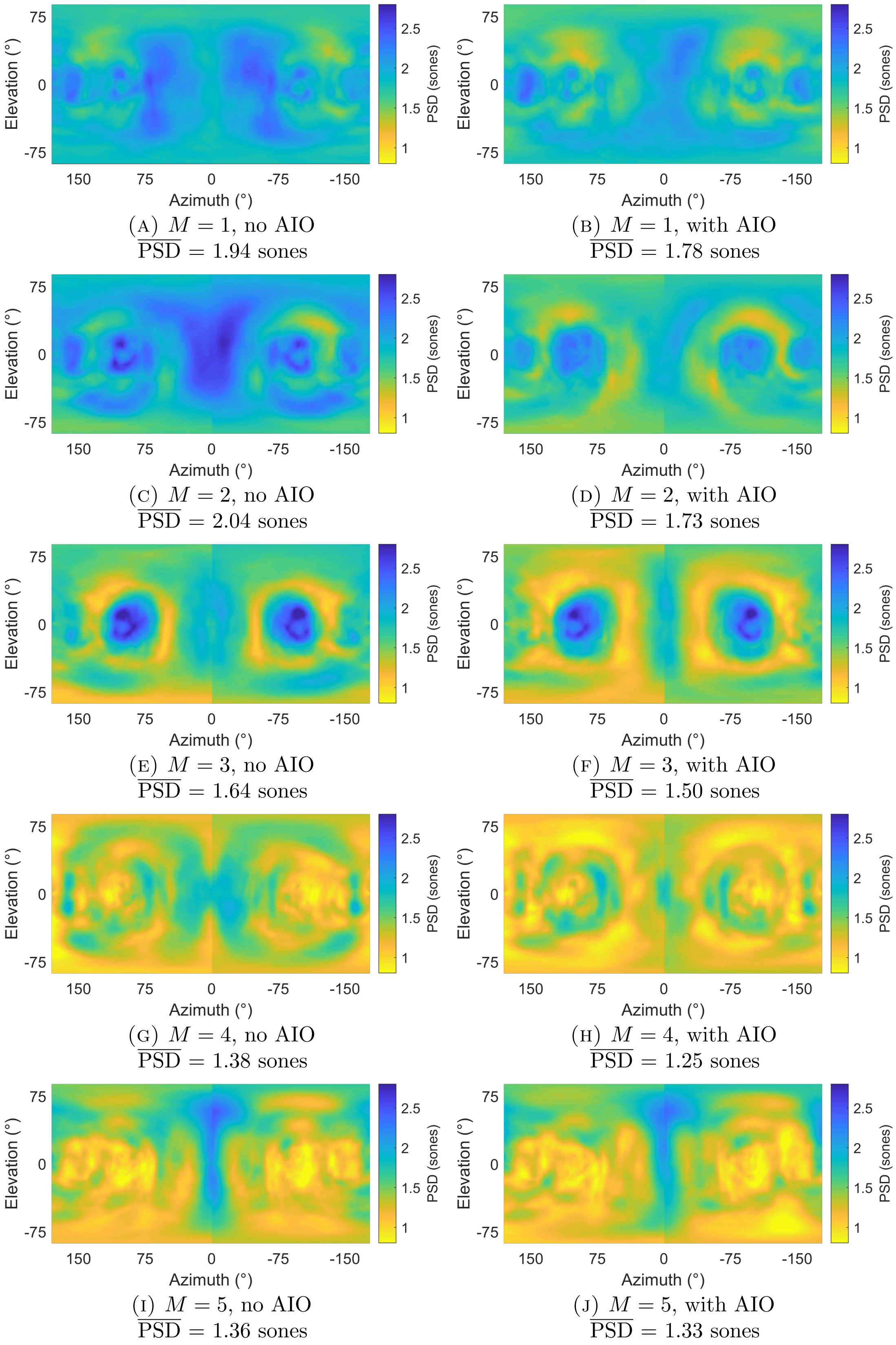
and estimated horizontal and vertical localisation. The applicability of AIO to other loudspeaker configurations and alternative HRTFs is also explored.

For each measurement location in the reference HRTF dataset of $Q$ = 16,020, as illustrated in Figure 2.28, binaural Ambisonic renders were generated for $\{M = 1, M = 2, ..., M = 5\}$, with and without AIO. All HRTFs were truncated to 1024 taps with 50 sample in / out half-Hanning windows applied.

### 6.2.1 Perceptual Spectral Difference

PSD between binaural Ambisonic renders and HRTFs was calculated for $\{M = 1, M = 2, ..., M = 5\}$ over all measurement locations on the sphere using the method detailed in Section 4.2, both with and without AIO. The solid angle weighted $\overline{\text{PSD}}$, calculated using (4.3), for $\{M = 1, M = 2, ..., M = 5\}$ with and without AIO, across all measurement locations, are presented in Figure 6.5, along with the minimum and maximum absolute PSD values. AIO is shown to produce an overall improvement in PSD for all tested orders of Ambisonics, as well as a lower maximum PSD value. A lower minimum PSD value is observed for all but $M = 4$.

To assess the directional effect of AIO on PSD between HRTFs and binaural Ambisonic rendering, Figure 6.6 presents the PSD for $\{M = 1, M = 2, ..., M = 5\}$ across



(A) $M = 1$, no AIO
$\overline{\text{PSD}} = 1.94$ sones

(B) $M = 1$, with AIO
$\overline{\text{PSD}} = 1.78$ sones

(C) $M = 2$, no AIO
$\overline{\text{PSD}} = 2.04$ sones

(D) $M = 2$, with AIO
$\overline{\text{PSD}} = 1.73$ sones

(E) $M = 3$, no AIO
$\overline{\text{PSD}} = 1.64$ sones

(F) $M = 3$, with AIO
$\overline{\text{PSD}} = 1.50$ sones

(G) $M = 4$, no AIO
$\overline{\text{PSD}} = 1.38$ sones

(H) $M = 4$, with AIO
$\overline{\text{PSD}} = 1.25$ sones

(I) $M = 5$, no AIO
$\overline{\text{PSD}} = 1.36$ sones

(J) $M = 5$, with AIO
$\overline{\text{PSD}} = 1.33$ sones

FIGURE 6.6: PSD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.

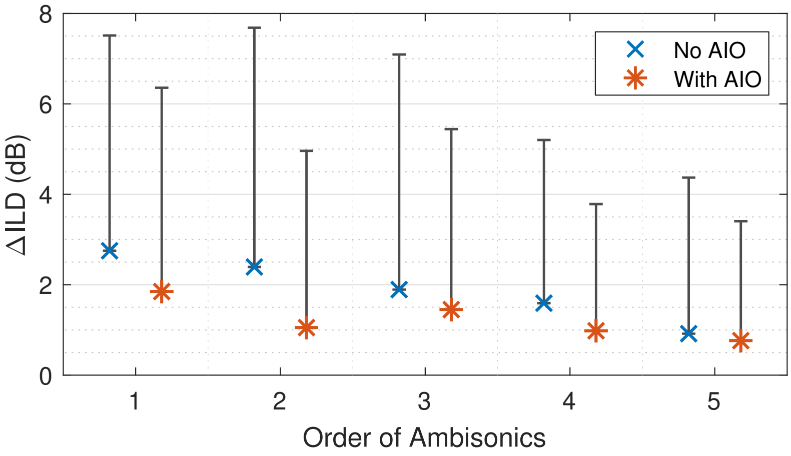
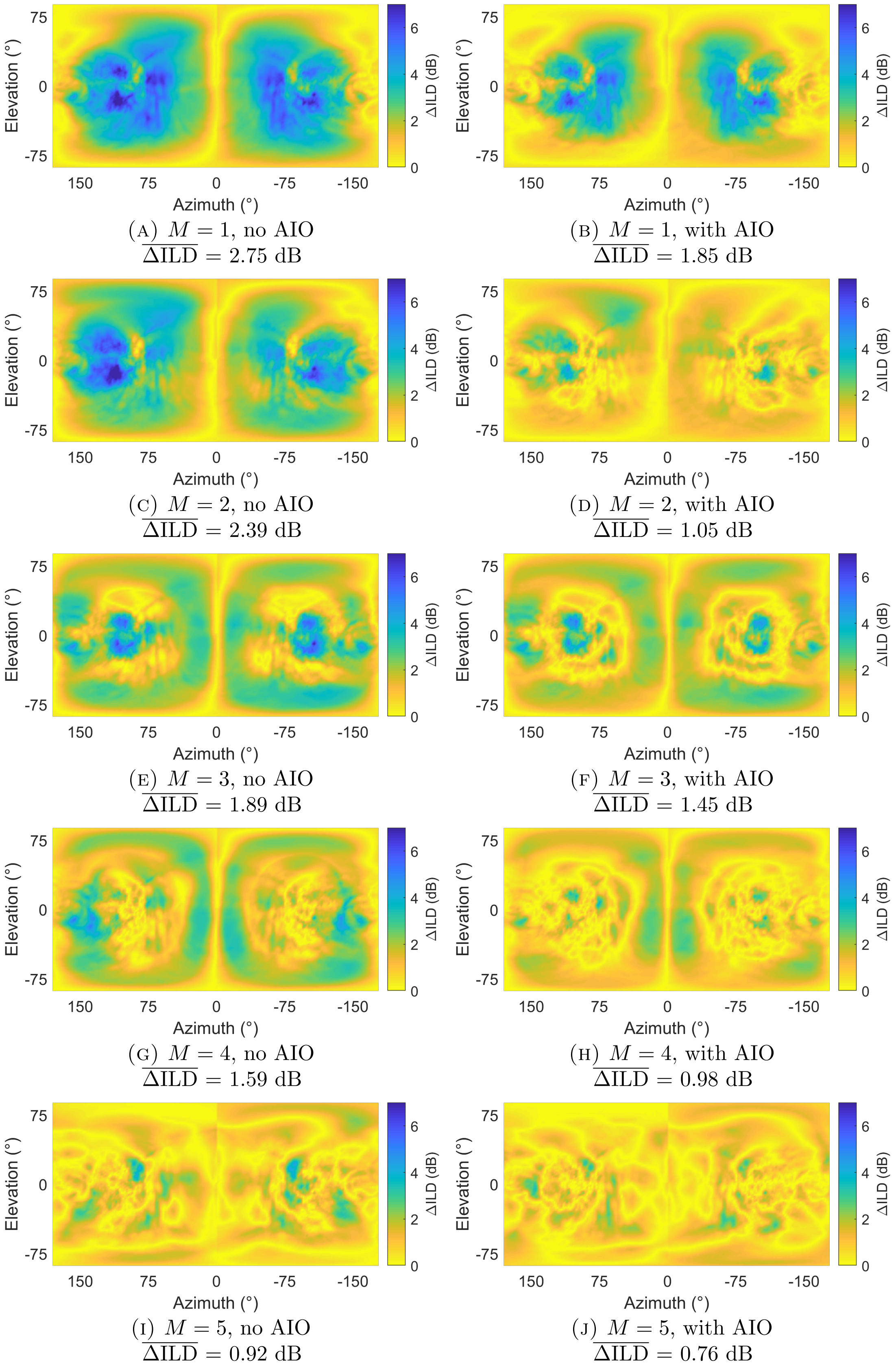


FIGURE 6.7: $\overline{\Delta \text{ILD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum $\Delta$ILD value.

every measurement location on the sphere, with and without AIO. It is clear that, for the majority of Ambisonic orders, the greatest improvements in PSD lie at the lateral regions. This is to be expected, as AIO does not effect virtual loudspeakers on the median plane. However, for $M = 2$, PSD is also improved along the median plane. The numerical tests on spectral reproduction show positive results: not only does AIO appear to improve Ambisonic ILD reproduction, it also improves the spectral reproduction.

### 6.2.2 Interaural Level Difference

To assess the effect of AIO on binaural Ambisonic ILD reproduction in detail, $\Delta$ILD between the reference HRTFs and the Ambisonic generated HRTFs was calculated for each measurement location $q$ using (4.6), and the solid angle weighted $\overline{\Delta \text{ILD}}$ values using (4.7). Figure 6.7 displays the solid angle weighted $\overline{\Delta \text{ILD}}$ values between HRTFs and binaural Ambisonic rendering with and without AIO for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute $\Delta$ILD value. This shows that with AIO, ILD is reproduced with greater accuracy for all tested orders of Ambisonics; indeed, a greater accuracy than $M + 1$ without AIO for all tested orders apart from the $M = 4$ instance. The improvement is greatest at



(A) $M = 1$, no AIO
$\overline{\Delta\text{ILD}} = 2.75$ dB

(B) $M = 1$, with AIO
$\overline{\Delta\text{ILD}} = 1.85$ dB

(C) $M = 2$, no AIO
$\overline{\Delta\text{ILD}} = 2.39$ dB

(D) $M = 2$, with AIO
$\overline{\Delta\text{ILD}} = 1.05$ dB

(E) $M = 3$, no AIO
$\overline{\Delta\text{ILD}} = 1.89$ dB

(F) $M = 3$, with AIO
$\overline{\Delta\text{ILD}} = 1.45$ dB

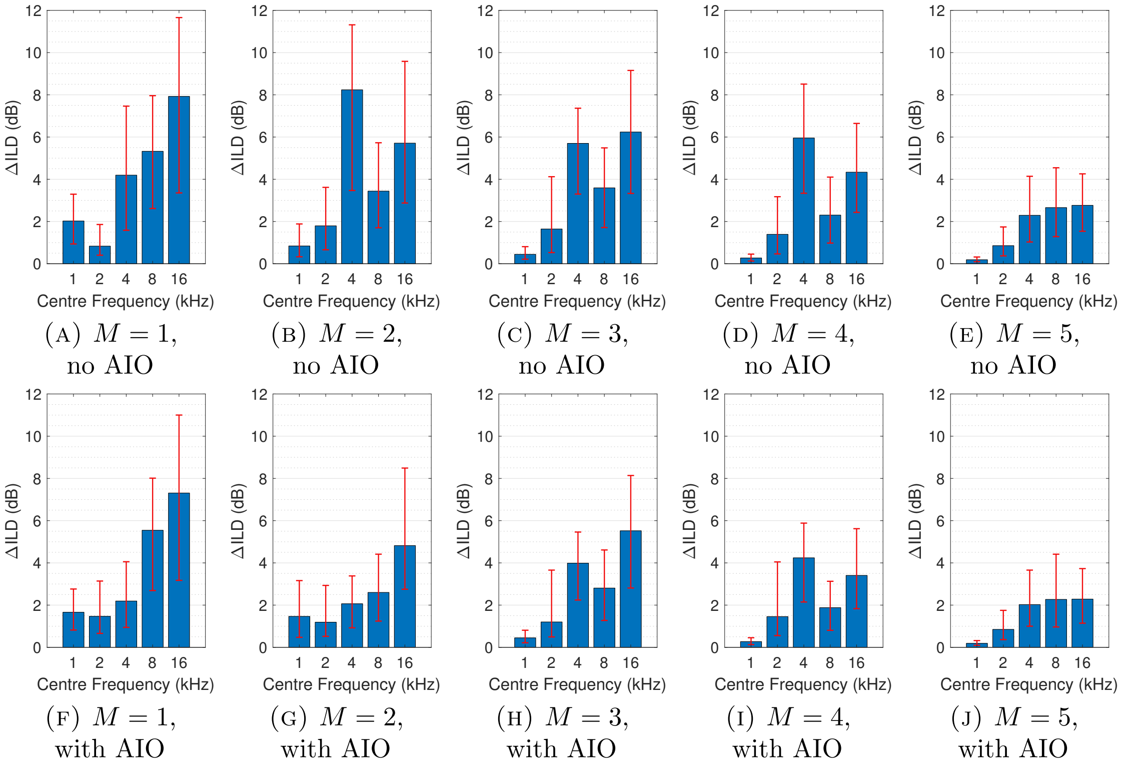
(G) $M = 4$, no AIO
$\overline{\Delta\text{ILD}} = 1.59$ dB

(H) $M = 4$, with AIO
$\overline{\Delta\text{ILD}} = 0.98$ dB

(I) $M = 5$, no AIO
$\overline{\Delta\text{ILD}} = 0.92$ dB

(J) $M = 5$, with AIO
$\overline{\Delta\text{ILD}} = 0.76$ dB

FIGURE 6.8: ΔILD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$, no AIO

(B) $M = 2$, no AIO

(C) $M = 3$, no AIO

(D) $M = 4$, no AIO

(E) $M = 5$, no AIO

(F) $M = 1$, with AIO

(G) $M = 2$, with AIO

(H) $M = 3$, with AIO

(I) $M = 4$, with AIO

(J) $M = 5$, with AIO

FIGURE 6.9: Median values of ΔILD between HRTFs and binaural Ambisonic rendering (Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$) for five frequency bands over all directions on the sphere, with 25% and 75% percentile bars.

orders $M = 1$ and $M = 2$ though, where Ambisonic ILD reproduction is inherently the least accurate.

Figure 6.8 presents detailed plots of ΔILD for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without AIO. Smaller values of ΔILD indicate ILD rendering closer to the HRTF. It is clear that ILD is improved for lateral locations for all tested orders of Ambisonics, though the effect is most pronounced at $M = 1$ and $M = 2$.

To look closer at the effect of AIO on the ILD reproduction of binaural Ambisonic rendering, a second ILD calculation was made to observe how ΔILD changes with frequency. Instead of producing one value of ILD for all frequencies using 30 ERB bands as in (2.14), ILD was calculated separately for 5 frequency bands with centre frequencies of 1 kHz, 2 kHz, 4 kHz, 8 kHz and 16 kHz. Figure 6.9 illustrates the median value of ΔILD over all measurement locations between binaural Ambisonic rendering and HRTFs for $\{M = 1, M = 2, ..., M = 5\}$, both with and without AIO, across the 5 frequency bands. 25% and 75% percentile bars are included to demonstrate the divergence from the median. Observations of the graph show that in general, $\Delta$ILD between binaural Ambisonic rendering and HRTFs increases with frequency, however the 4 kHz band sees greater improvement than the other bands. For $M = 1, M = 3$ and $M = 4$, the median $\Delta$ILD for the 4 kHz band is more than 1 dB lower, and for $M = 2$ more than 6 dB lower with AIO. This informs a potential future development of the algorithm, whereby altering the algorithm to produce an ILD augmentation for different frequencies could produce greater results.

### 6.2.3 Interaural Time Difference

To assess the effect of AIO on the Ambisonic reproduction of ITD, $\Delta$ITD between the reference HRTFs and the Ambisonic generated HRTFs was calculated for each measurement location $q$ using (4.4), and $\overline{\Delta \text{ITD}}$ is calculated using (4.5). The low-pass filter used in the ITD calculation was $f_c = 1.5$ kHz. Figure 6.10 displays the solid angle weighted $\overline{\Delta \text{ITD}}$ values between HRTFs and binaural Ambisonic rendering with and without AIO for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute $\Delta$ITD value. This shows a small reduction in overall accuracy of ITD rendering when AIO is implemented for $M = 1$ but a marginal improvement for $M = 2$, while $M \geq 3$ remains largely unaffected. The likely reason for reduced accuracy at $M = 1$ is due to the great reduction in amplitude at high frequencies of the contralateral signal of the HRTFs at lateral positions, in the frequency region between the low-pass filter used in the AIO process and the high-pass filter used in the ITD estimation. At these positions the contralateral signal is likely to be delayed, and the reduction in amplitude reduces the significance of this delay in the resulting Ambisonic renders, therefore producing a marginal reduction in ITD accuracy. Using $f_c = 500$ Hz in the ITD calculation low-pass filter causes the AIO and non-AIO values to converge. Detailed plots of $\Delta$ITD for every measurement location on the sphere are presented in Appendix C.1, which show the minimal effect of AIO on ITD.

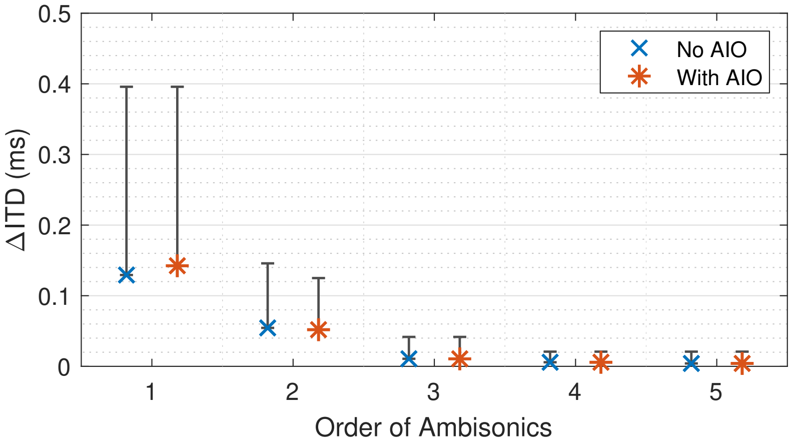


FIGURE 6.10: Estimated $\overline{\Delta\text{ITD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔITD value.

### 6.2.4 Estimated Horizontal Localisation

The effect of AIO on horizontal localisation of binaural Ambisonic rendering was estimated using the method detailed in Section 2.8.4, which utilises a horizontal model (May, Van De Par and Kohlrausch, 2011), producing a value of $\overline{E_\theta}$ for overall estimated localisation between $-90° < \theta < +90°$ at $\phi = 0°$ using (2.17). Figure 6.11 displays the overall estimated horizontal localisation of binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$. This shows a small improvement for $M = 1$ and $M = 2$, but greater error for $M \geq 3$. Appendix C.2 presents detailed individual plots for the estimated horizontal localisation of each tested azimuth angle with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$. For $M = 1$, the maximum absolute value of $\theta_{\text{est}}$ is greater with AIO implementation, which suggests an increase in lateralisation and explains the improvement in $\overline{E_\theta}$, whereas AIO reduces the overestimated $\theta_{\text{est}}$ in $M = 2$. It is interesting to see an increase in $\overline{E_\theta}$ at $M \geq 3$, despite improved ILD reproduction and unchanged ITD reproduction. Further testing, such as real life localisation tests, are required to explain this. As the focus of this thesis is timbral accuracy, this will not be discussed further.

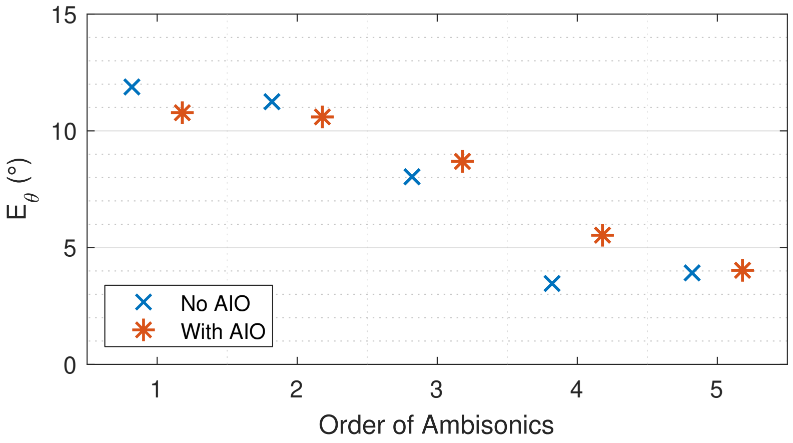
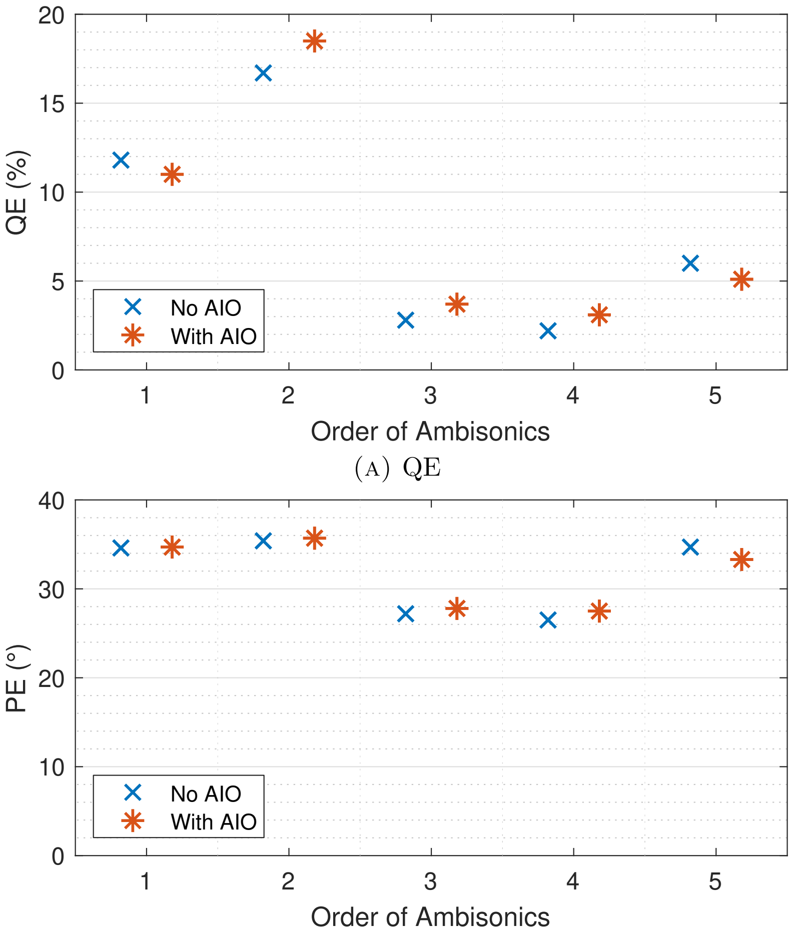


FIGURE 6.11: Estimated $\overline{E_\theta}$ of binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (May, Van De Par and Kohlrausch, 2011).

### 6.2.5 Estimated Sagittal Plane Localisation

The effect of AIO on estimated elevation localisation in the sagittal plane was evaluated between $-90° < \phi < +90°$ at $\theta = 0°$ using the method detailed in Section 2.8.4 which utilises a localisation model (Baumgartner, Majdak and Laback, 2014) to produce quadrant error (QE), a prediction of localisation confusion (in %), and polar RMS error (PE), a prediction of precision and accuracy (in °). As the HRTFs used are of a Neumann KU 100, which has no torso, there will be no elevation cues present below 1.5 kHz (Algazi, Avendano and Duda, 2001a). Therefore, the frequency range of the model's filter bank was set to 1.5 kHz - 18 kHz, with the upper limit of the frequency range chosen as the highest frequency of perceivable elevation cues (Roffler and Butler, 1967; Asano, Suzuki and Sone, 1990).

Figure 6.12 shows the predicted QE and PE values of binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, and detailed individual plots of predicted sagittal plane localisation are presented in Appendix C.3. Results indicate that AIO has a small effect on estimated sagittal plane localisation accuracy, though results vary for different Ambisonic orders. Comparing predicted elevation localisation with the PSD plots in Figure 6.6, the regions where AIO appears to



(A) QE

(B) PE

FIGURE 6.12: Estimated sagittal plane localisation of binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (Baumgartner, Majdak and Laback, 2014).

improve spectral reproduction on the mid-sagittal plane appear to correlate with the regions of improved sagittal plane localisation in Appendix C.3.

### 6.2.6 Generalisability

To demonstrate how general the applicability of AIO is, additional simulations were run using both different loudspeaker configurations and an alternative HRTF dataset. In both sets of simulations, the effect of AIO was assessed by comparing Ambisonic renders to the original HRTFs for all available measurement locations. Two numerical evaluation metrics were investigated: PSD and ILD, with PSD calculations made

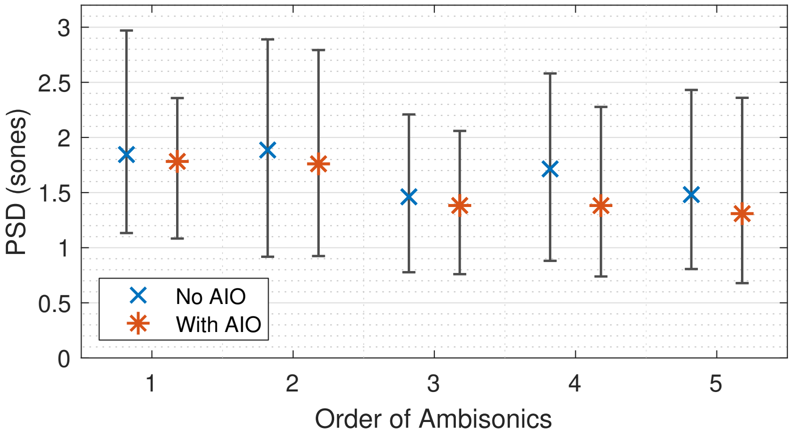


FIGURE 6.13: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the minimum and maximum absolute PSD values.

using the method detailed in Section 4.2 and single values of $\overline{\text{PSD}}$ then calculated using (4.3), and ILD calculations using (4.6), and $\overline{\Delta\text{ILD}}$ calculated using (4.7).

Firstly, different loudspeaker configurations were investigated[1]. Here, spherical T-designs (Hardin and Sloane, 1996) are employed with corresponding loudspeaker vertices as illustrated in Figure 3.8. The solid angle weighted $\overline{\text{PSD}}$ results, calculated from 16,020 locations on the sphere, are shown in Figure 6.13, along with the maximum and minimum absolute PSD values. This illustrates how AIO produces an overall improvement in PSD, regardless of the type of loudspeaker configuration. Additionally, the maximum absolute PSD value is reduced with AIO for all tested orders of Ambisonics, and the minimum absolute PSD value is reduced for all but $M = 2$. Detailed plots of PSD for every measurement location on the sphere are presented in Appendix C.4, which also follow similar trends to those shown in the Lebedev loudspeaker configurations in Figure 6.6, however with T-designs the improvements in spectral reproduction also occur closer to the median plane.

Values of $\overline{\Delta\text{ILD}}$ for orders of Ambisonics $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations are presented in Figure 6.14, along with the maximum $\Delta$ILD value. These show significant improvements at $M = 1$ and $M = 2$. Marginally lower $\overline{\Delta\text{ILD}}$ values are also observed for $M \geq 3$, however the maximum $\Delta$ILD value is

[1] For all other areas of this chapter, Lebedev loudspeaker configurations are used.

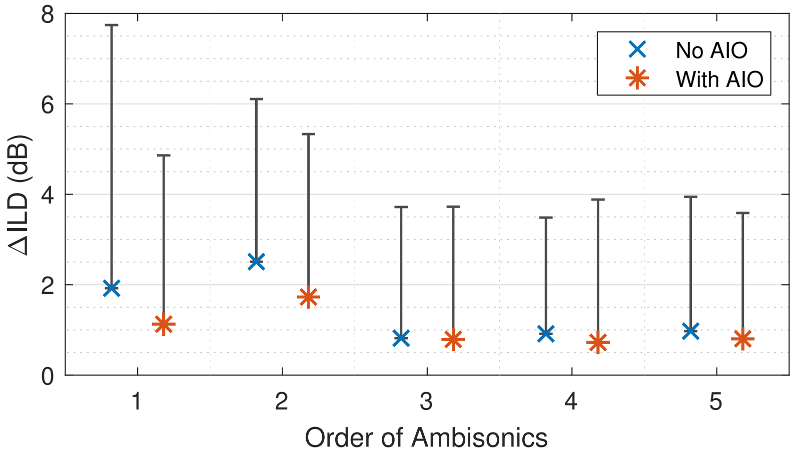


FIGURE 6.14: $\overline{\Delta \text{ILD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the maximum ΔILD value.

greater at $M = 4$, suggesting there are areas where ILD is reproduced less accurately when AIO is implemented for this configuration. Detailed plots of ΔILD for every measurement location on the sphere are presented in Appendix C.5, which follow similar trends to those observed in the Lebedev configuration plots in Figure 6.8, with AIO producing improved ILD reproduction at lateral regions.

Secondly, to assess the effect of AIO when using an alternative HRTF dataset, binaural Ambisonic renders were made for $\{M = 1, M = 2, M = 3, M = 5\}$[2] using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, human subject H20 (Armstrong et al., 2018a). The solid angle weighted $\overline{\text{PSD}}$ results, calculated from 2,114 locations on the sphere, are shown in Figure 6.15 along with the minimum and maximum absolute PSD values. This illustrates how, again, AIO produces a marginal improvement in $\overline{\text{PSD}}$ for all tested orders of Ambisonics, regardless of the HRTF database or subject used, though the effect is less prominent here. While AIO produces marginally lower minimum values of PSD for all tested Ambisonic orders but $M = 3$, the maximum values are increased with individualised HRTFs. Detailed plots of PSD for every measurement location on the sphere are presented in Appendix C.6, which also follow similar trends to Figure 6.6

[2] The omission of $M = 4$ was due to a lack of necessary measurements.

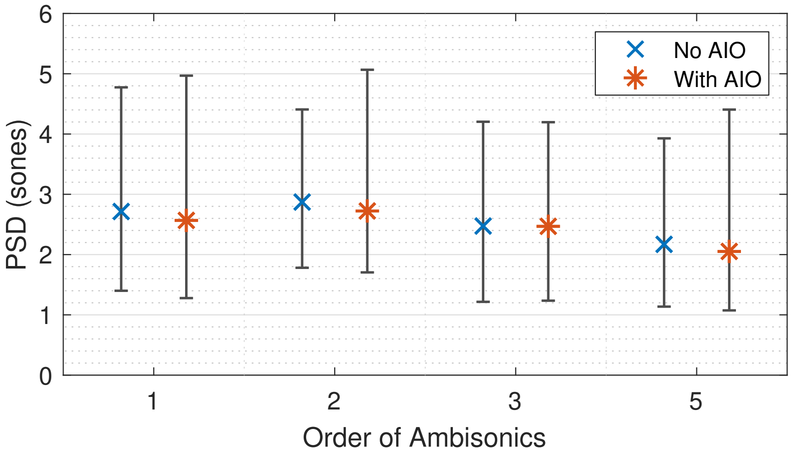


FIGURE 6.15: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the minimum and maximum absolute PSD values.

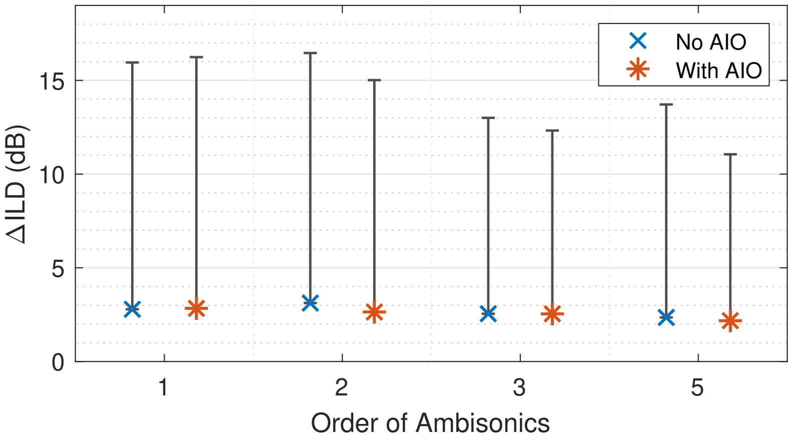


FIGURE 6.16: $\overline{\Delta\text{ILD}}$ between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the maximum ΔILD value.

and Appendix C.4, with improved spectral reproduction largely constrained to the lateral regions.

To evaluate the effect of AIO on Ambisonic reproduction of ILD when using an alternative HRTF dataset, values of $\overline{\Delta\text{ILD}}$ for orders of Ambisonics $\{M = 1, M = 2, M = 3, M = 5\}$ are presented in Figure 6.16, along with the maximum ΔILD value, and detailed plots of ΔILD for every measurement location on the sphere are presented in Appendix C.7. AIO produces marginally lower $\overline{\Delta\text{ILD}}$ values with a lower maximum $\Delta$ILD for $M > 1$. It is clear that AIO is less effective in improving ILD reproduction for these specific HRTFs.

The tests on generalisability therefore show how AIO improves the overall spectral reproduction and ILD reproduction of binaural Ambisonic rendering when using different virtual loudspeaker configurations and alternative HRTF datasets.

## 6.3 Perceptual Evaluation

To assess the perceptual effect of AIO in binaural Ambisonic rendering, two listening tests were conducted, corresponding to simple and complex acoustic scenes. As the objective evaluation showed AIO to produce the most notable effects for low order Ambisonics (in particular, $M = 1$ and $M = 2$), the perceptual evaluation focused on low order ($M < 5$) rendering, using Lebedev loudspeaker configurations.

Tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 (International Telecommunication Union, 2015b) using the Matlab based 'Scale' (Giner, 2013), as shown in Figure 4.25. Tests were conducted in a quiet listening room (background noise level of 41.8 dBA) using an Apple MacBook Pro with a Fireface 400 audio interface, which has software-controlled input and output levels. A single set of Sennheiser HD 650 circumaural headphones was used for all tests, which were equalised as in Section 4.4. The RMS HpTF and inverse filter of the left HD 650 headphone, along with a resulting convolved response, are shown in Figure 4.24. Listening tests were conducted on 18 participants aged between 23 to 71 (14 male, 3 female, 1 non-binary). All reported normal hearing as in accordance with ISO Standard 389 (International Organization for Standardization, 2016) and prior critical listening experience, which was deemed sufficient if the participant had education or employment in audio or music engineering.

Test sounds were normalised to a consistent A-weighted RMS amplitude and participants were able to adjust the playback level. All binaural renders were static (fixed head orientation) to ensure consistency in the experience between participants.

### 6.3.1 Test Paradigms

The first listening test assessed the perceptual effect of AIO in binaural Ambisonic rendering for simple scenes. The base stimulus was a one second burst of monophonic pink noise at a sample rate of 48 kHz, windowed by onset and offset half-Hanning ramps of 5 ms, with half a second of silence between each burst. Test sound locations $\psi$ were the same as used in Section 4.4 and presented in Table 4.6, chosen as the central points of the faces of a dodecahedron, to avoid test sound locations coinciding with loudspeaker locations. The reference was a direct HRTF convolution, the medium anchor was a low-pass filtered version of the reference with an $f_c = 7$ kHz, and the low anchor was the monophonic base stimulus low-pass filtered at an $f_c = 3.5$ kHz. The other 6 stimuli were binaural Ambisonic renders for three Ambisonic orders $\{M = 1, M = 2, M = 3\}$, with and without AIO, totaling 9 test stimuli per trial. For each trial, the listener was asked to rate the 9 stimuli with a score between 0 and 100 in terms of overall perceived similarity to the reference, in accordance with the Spatial Audio Quality Inventory (SAQI) (Lindau et al., 2014) whereby increased similarity would be rated higher. Each trial was repeated once, giving a total of 16 trials. Stimuli and trial ordering was randomised and presented double blind.

The second listening test used four complex scenes, which were 3 - 5 second excerpts of soundscape recordings from the open source EigenScape database of $M = 4$ Ambisonic recordings made using an MH Acoustics em32 Eigenmike[3] at various locations in northern England (Green and Murphy, 2017). The initial format of recordings follows Schmidt semi-normalised (SN3D) normalisation, which therefore was converted to N3D normalisation using (3.10). The soundscapes used in the complex scenes listening test, along with a description of the specific excerpt used are as follows:

The composition of scenes featured mainly horizontally located sounds, though elevated sources were present such as the birdsong in scene 2 and travel announcement in scene 4, as well as the room reverberation in scene 3 and 4 due to the recordings

[3] https://mhacoustics.com/

1. **Beach:** Waves breaking against the shore.

2. **Quiet Street:** A single car drives past with birdsong.

3. **Pedestrian Zone:** Pedestrians walking around and talking.

4. **Train Station:** Travel announcement on the station platform.

having been made indoors. The complex-scenes listening test loosely followed the MUSHRA paradigm; however, due to the nature of the stimuli no ideal reference was available. Partly for this reason, the $M = 4$ renders were included in the complex-scenes test which are the highest available Ambisonic order of Eigenmike recordings. Lower order renders were obtained by simply discarding the higher-order channels. An $M = 0$ render was used as an anchor, and test stimuli were binaural Ambisonic renders for orders $\{M = 1, M = 2, ..., M = 4\}$, with and without AIO, totalling 9 test stimuli per trial. For each trial, participants were asked to rate each stimuli with a score between 0 and 100 on plausibility and spaciousness, whereby natural, wide, full and externalised stimuli would be rated higher, and boxed in, lacking lateralisation, internalised stimuli would be rated lower. Each trial was repeated once, giving a total of 8 trials. Stimuli and trial ordering was again randomised and presented double blind.

### 6.3.2 Results

Overall, the tests lasted between around 20 to 45 minutes to complete. Results were post-screened for unreliable participants based on the following criteria. For simple scenes: rating the hidden reference lower than 90% for >15% of trials or rating the mid-anchor higher than 90% for >15% of trials, and for complex scenes: rating the anchor higher than 90% for >15% of trials. Based on these criteria, one participant's results were excluded from analysis. The raw results from both listening tests were tested for normality using the Kolmogorov-Smirnov test, which showed all data as non-normally distributed. Therefore, all statistical analysis was conducted using non-parametric methods.

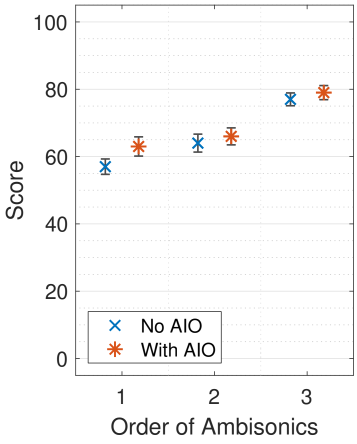


Figure 6.17: Median simple-scene scores with non-parametric $\mathrm{CI}_{95}$ across all participants and test sound locations ($\psi$), reference and anchor scores omitted. Score indicates overall perceived similarity between binaural Ambisonic rendering and HRTF convolution.

Table 6.1: Significance results of the simple-scene test over all test sound locations using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether AIO produced a statistically significant improvement to binaural Ambisonic rendering.

| $M$ | 1 | 2 | 3 |
|---|---|---|---|
| $h$ | 1* | 1 | 0 |

The median scores of the simple-scenes test, conducted to determine whether AIO improves the overall perceived similarity between binaural Ambisonic rendering and HRTF convolution, are shown in Figure 6.17 (reference and anchor scores are omitted) for each order of Ambisonics across all participants and test sound locations, with non-parametric 95% confidence intervals ($\mathrm{CI}_{95}$) (Mcgill, Tukey and Larsen, 1978) calculated using (4.8). The different conditions of the test were tested for statistical significance using a Friedman's analysis of variance (ANOVA) test, which showed high statistical significance ($\chi^2(5) = 203.71, p < 0.01$). AIO is shown to produce an increase in overall perceived similarity for all tested orders of Ambisonics. To test whether this improvement is statistically significant, post-hoc Wilcoxon signed-rank tests were conducted for each Ambisonic order, and Table 6.1 presents the significance results. For $M = 1$ and $M = 2$, AIO produced a statistically significant improvement in overall perceived similarity between binaural Ambisonic rendering and HRTF convolution. Though an improvement can be observed for $M = 3$, it was

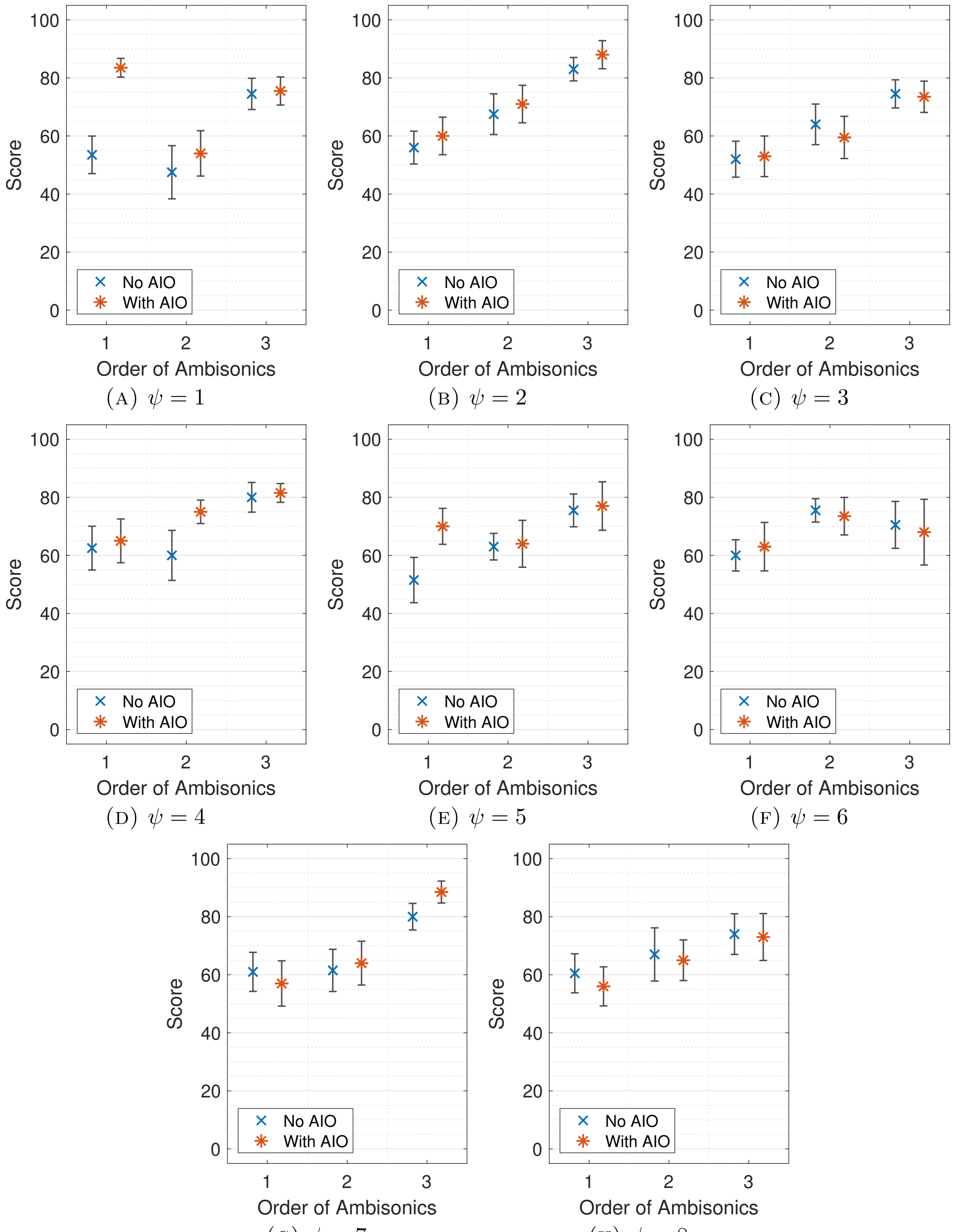
Table 6.2: Significance results of the simple-scene test for each test sound location using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether AIO produced a statistically significant improvement to binaural Ambisonic rendering.

| $\psi$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $h$ $(M = 1)$ | 1* | 0 | 0 | 0 | 1* | 1 | 0 | 0 |
| $h$ $(M = 2)$ | 1 | 0 | 0 | 1* | 0 | 0 | 0 | 0 |
| $h$ $(M = 3)$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

not statistically significant at a confidence of 95% ($p = 0.743$).

To assess whether the perceptual effect of AIO varied with test sound location, a Friedman's ANOVA was conducted, which showed high statistical significance ($\chi^2(7) = 39.61, p < 0.01$). Figure 6.18 illustrates the median scores with non-parametric CI$_{95}$ for each individual test sound location $\psi$ across all participants. Post-hoc Wilcoxon signed-rank tests were conducted to determine which test sound locations produced a significant improvement in overall perceived similarity for AIO, the results of which are shown in Table 6.2. It is clear that results varied for test sound location differently for each tested Ambisonic order. Additionally, some participants noted minor listening fatigue in the simple scenes due to repeated pink noise bursts, so future tests should look at addressing this.

The median scores of the complex-scenes test, conducted to determine whether AIO improves plausibility and spaciousness of binaural Ambisonic rendering, are shown in Figure 6.19 for each condition across all participants and test sound locations, with non-parametric CI$_{95}$. A Friedman's ANOVA confirmed that the test conditions produced highly statistically significantly different results ($\chi^2(7) = 264.4, p < 0.01$). An observation of Figure 6.19 indicates that ratings increase with Ambisonic order, tapering off as order increases, and AIO improves the ratings for all tested orders, though the improvement is greatest at $M = 1$ and $M = 2$. To test whether this improvement for each order is statistically significant, post-hoc Wilcoxon signed-rank tests were conducted. Table 6.3 presents the significance results. For $M = 1$ and $M = 2$, AIO produces a highly statistically significant improvement. Though



(A) $\psi = 1$

(B) $\psi = 2$

(C) $\psi = 3$

(D) $\psi = 4$

(E) $\psi = 5$

(F) $\psi = 6$

(G) $\psi = 7$

(H) $\psi = 8$

FIGURE 6.18: Median simple-scene scores with non-parametric $\mathrm{CI}_{95}$ across all participants for each test sound location ($\psi$), reference and anchor scores omitted. Score indicates overall perceived similarity between binaural Ambisonic rendering and HRTF convolution.

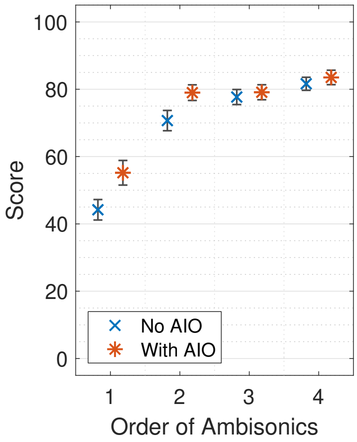


FIGURE 6.19: Median complex-scene scores with non-parametric $\mathrm{CI}_{95}$ across all participants and soundscapes, $M = 0$ scores omitted. Score indicates perceived plausibility and spaciousness.

TABLE 6.3: Significance results of the complex-scene test over all soundscapes using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether AIO produced a statistically significant improvement to the plausibility of binaural Ambisonic rendering.

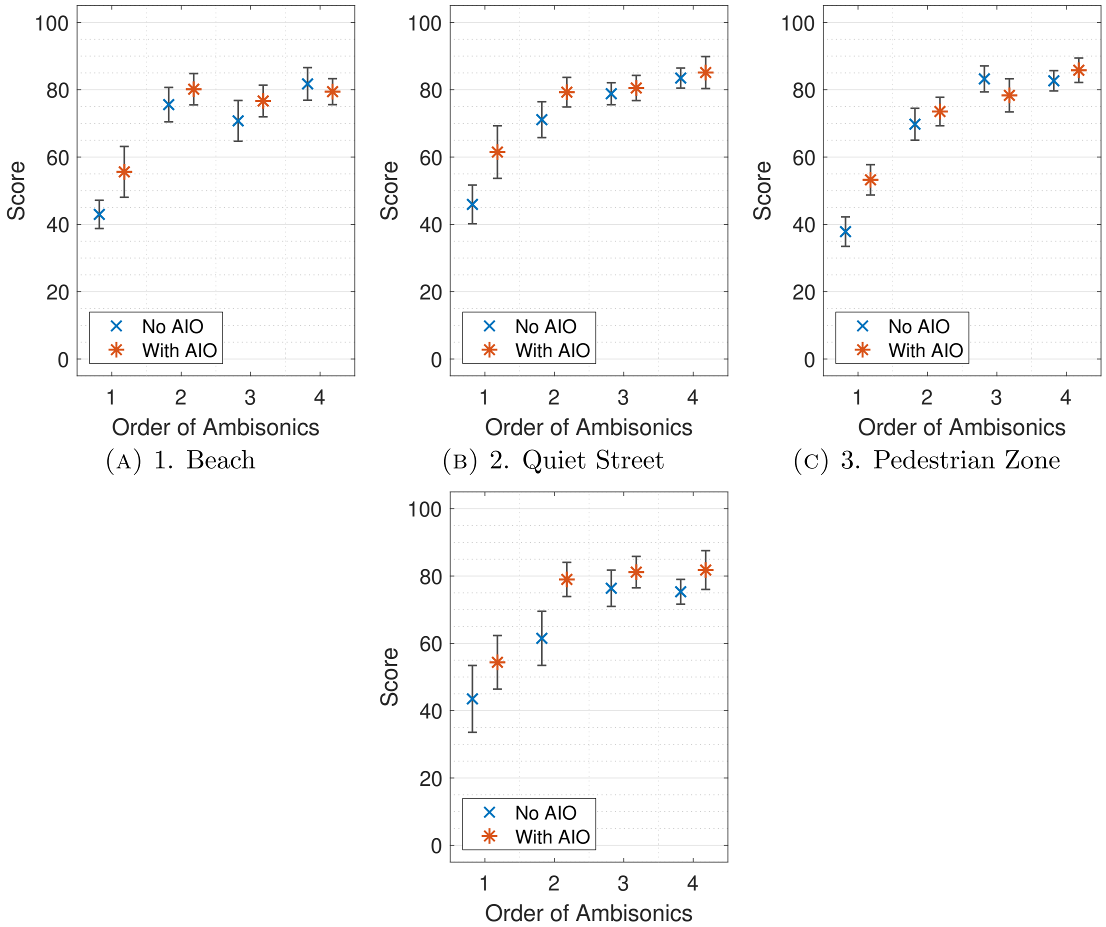
| $M$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $h$ | 1* | 1* | 0 | 0 |

improvements are still observed for $M = 3$ and $M = 4$, they are not statistically significant at 95% confidence ($p = 0.1$ and $p = 0.07$, respectively).

Figure 6.20 shows the median complex-scene scores with non-parametric $\mathrm{CI}_{95}$ across all participants for each individual soundscape. AIO produced a higher median score than without AIO for all soundscapes and tested orders, apart from the conditions of $M = 4$, soundscape 1 and $M = 3$, soundscape 3. To observe whether soundscape type had a statistically significant effect on results, a Friedman's ANOVA was conducted, which showed no significance ($\chi^2(3) = 1.9, p = 0.59$). Therefore, no post-hoc tests were conducted.

## 6.4 Discussion

The numerical evaluation of Ambisonic ILD optimisation has shown that, by manipulating left and right signals of the HRTFs used in binaural Ambisonic rendering,



(A) 1. Beach

(B) 2. Quiet Street

(C) 3. Pedestrian Zone

(D) 4. Train Station

FIGURE 6.20: Median complex-scene scores with non-parametric $\mathrm{CI}_{95}$ across all participants for each soundscape, $M = 0$ scores omitted. Score indicates perceived plausibility and spaciousness.

an improvement in the ILD reproduction of binaural Ambisonic rendering can be achieved, when compared to direct HRTF rendering (see again Figure 6.8). In most cases this comes in the form of an increase in values of ILD (especially at $M = 1$ and $M = 2$), but not all - at some locations on the sphere AIO reduces ILD of the Ambisonic rendering (as shown in Figure 6.4e). The evaluation of Ambisonic ILD reproduction for all directions over the sphere shows that AIO improves ILD reproduction for all tested Ambisonic orders, including when using different loudspeaker configurations or individualised HRTFs. The greatest benefits are observed where ILD is inherently reproduced the worst: at $M = 1$ and $M = 2$, though AIO has been shown to improve ILD reproduction even at $M = 5$. In general, $\Delta$ILD between HRTFs and binaural Ambisonic rendering has been shown to increase with frequency (see Figure 6.9), which is likely caused by Ambisonic spatial aliasing, which increases with frequency once above $f_{\text{alias}}$. The AIO algorithm produces a single augmentation gain value for all frequencies. A future development could investigate implementing frequency specific ILD optimisation, which may yield further improvements. The regions of the sphere where AIO affects ILD reproduction the most also produce the greatest improvements in PSD and listening tests. $M = 2$ produces the largest improvement in ILD reproduction over the sphere, and this is followed by the biggest improvement in PSD. With AIO, the value of $\overline{\Delta \text{ILD}}$ is lower than the next Ambisonic order (without AIO), for all but $M = 4$ (see again Figure 6.7). Anecdotally, the implementation of AIO tends to produce an increase in lateralisation at low Ambisonic orders, pushing the sound stage away from the head.

Estimated horizontal plane localisation is also improved at $M = 1$ and $M = 2$ through an observation of greater lateralisation (see Figures C.2b and C.2d, respectively). Estimated sagittal plane localisation tends to correlate with the mid-sagittal plane PSD results, suggesting the regions where AIO improves PSD also produce improved elevation localisation. ITD is largely unchanged with AIO.

Concerning the listening test results, AIO produced notable improvements for $M = 1$ and $M = 2$, and small (but generally not statistically significant) improvements for $M = 3$ and $M = 4$. However, in the simple-scenes test, sound source location was found to be a significant influence on results. In the complex-scenes listening test, the type of soundscape did not affect results with statistical significance. A general observation is the considerable differences between simple and complex-scene results. There is a much greater difference in scores between Ambisonic orders in complex scenes (see Figure 6.19), and AIO produced more significant improvements here. A likely explanation for this is that recorded soundscapes of complex acoustic scenes have more of a focus on lateralization and spaciousness due to the numerous simultaneous sources, whereas pink noise, used in the simple-scenes tests, causes the listener to focus more on timbre. Further investigation is warranted to conclude the reason for the variation in results between the two tests.

Some additional observations have been made during this study. Despite the iterative pre-processing stage, the ILD augmentation gains for $M = 1$ plateau, meaning the Ambisonic reproduced ILDs do not quite reach those of the HRTF targets (as illustrated in Figure 6.6a). This is due to the normalisation of HRTFs post ILD augmentation using (6.4), which normalises the processed virtual loudspeaker HRTFs to the same RMS as the unprocessed virtual loudspeaker HRTFs. With this normalisation, the contralateral signals of the HRTFs with AIO processing for $M = 1$ have a very low amplitude and are essentially muted (see Figure 6.3c). Therefore, a further increase in ILD does not produce a change in results. Some preliminary experimentation found that if the normalisation is changed such that ILD augmented virtual loudspeaker HRTFs are normalised with respect to the RMS amplitude of the Ambisonic reproduced HRTF, AIO HRTFs can then become much louder than unprocessed HRTFs at high frequencies, which can produce Ambisonic ILDs much greater. However, this comes at the expense of spectral quality on the median plane. As accurate timbre is the most important spatial audio quality metric (Bregman, 1990; Rumsey et al., 2005a), the initial normalisation method was retained.

## 6.5 Summary

ILD reproduction of binaural Ambisonic rendering has been shown as inaccurate at low orders of Ambisonics. This chapter has presented a method for Ambisonic Interaural Level Difference Optimisation (AIO), aiming to improve the ILD reproduction of binaural Ambisonic rendering. This has been achieved through an iterative pre-processing stage whereby the ILD of the HRTFs for binaural rendering are measured and then augmented accordingly at frequencies above $f_{\text{alias}}$ by applying a gain to the contralateral signal of the HRTF such that when used for binaural Ambisonic rendering, the resulting rendered ILD matches that of the original HRTF more closely. The effect of AIO has been evaluated both numerically and perceptually. When compared to direct HRTF rendering, ILD and spectral reproduction is improved over the sphere, with little effect on the estimated vertical localisation accuracy. AIO is most effective at $M = 1$ and $M = 2$ where Ambisonic ILD reproduction is inherently the least accurate, and implementing AIO produces an improvement in lateralisation, which helps to reduce the perceptual differences between Ambisonic orders.

This chapter has shown that an iterative HRTF pre-processing stage that augments the levels of the ipsilateral and contralateral virtual loudspeaker HRTFs can produce an improvement in the ILD reproduction of binaural Ambisonic rendering, while also producing a small improvement in spectral reproduction. Therefore, a general statement can be suggested that for binaural Ambisonic rendering, AIO offers a clear improvement at $M = 1$ and $M = 2$, and an incremental improvement at $M \geq 3$. As AIO pre-processing of HRTFs can be implemented offline, it is hence recommended for improving lateralisation and spaciousness for all orders of Ambisonics, without producing a reduction in timbral quality. Future developments could look at adapting the AIO algorithm to implement frequency-dependent gains for each virtual loudspeaker, instead of a single gain as is the current case. The next chapter will investigate the combination of AIO with other binaural Ambisonic HRTF pre-processing techniques such as presented in Chapter 4, to assess whether cumulative improvements can be obtained.

# Chapter 7

# Combinations of Ambisonic Pre-Processing Techniques

The previous three chapters have presented novel HRTF pre-processing techniques for improving the high frequency reproduction of binaural Ambisonic rendering. These have included Ambisonic Diffuse-Field Equalisation (DFE), which removes the direction-independent frequency response characteristics of the binaural Ambisonic decoder; Ambisonic Directional Bias Equalisation (DBE), which improves the spectral reproduction of the binaural Ambisonic decoder at a specified direction; and Ambisonic Interaural Level Difference Optimisation (AIO), which improves the accuracy of ILD reproduction by level augmentation of the left and right signals of the HRTFs used in the binaural Ambisonic decoder.

This chapter investigates the possibility of using these HRTF pre-processing techniques together, along with other state-of-the-art techniques, in order to achieve even greater cumulative improvements to the high frequency reproduction of binaural Ambisonic rendering. Different pre-processing technique combinations are evaluated both numerically and perceptually, with the numerical evaluation comparing binaural Ambisonic rendering to a reference set of HRTFs in terms of perceptual spectral difference, estimated ILD and ITD cues and predicted localisation accuracy for $\{M = 1, M = 2, ..., M = 5\}$. The perceptual evaluation is presented through listening tests using three different types of acoustic scene for Ambisonic orders $M = 1, M = 2$ and $M = 3$. Finally, results are discussed and the chapter is concluded, along with proposed further work.

## 7.1 Method

In this chapter, different HRTF pre-processing techniques are combined. The aim is that, by running one after the other, the resulting binaural Ambisonic decoder will produce greater results than just one of the pre-processing techniques alone. The pre-processing techniques investigated in this chapter are Ambisonic Time Alignment (TA) (Evans, Angus and Tew, 1998; Richter et al., 2014; Zaunschirm, Schörkhuber and Höldrich, 2018), Ambisonic Interaural Level Difference Optimisation (AIO), as presented in Chapter 6, and Ambisonic Diffuse-Field Equalisation (DFE), as presented in Chapter 4. Ambisonic DBE is not utilised in this chapter as the evaluation covers directions over the whole sphere, whereas DBE is aimed at focussing improvements at a single location. All computation was carried out offline in MATLAB version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4.

### 7.1.1 Ambisonic Time Alignment

TA is the complete removal of interaural time differences (ITDs) of the virtual loudspeaker HRTFs at high frequencies (Evans, Angus and Tew, 1998; Richter et al., 2014; Zaunschirm, Schörkhuber and Höldrich, 2018), which reduces the comb filtering caused by the off-centre position of the ears in the virtual loudspeaker array. TA has previously only been implemented for dense sets of HRTFs and is here applied to sparse virtual loudspeaker sets. The previous implementations of TA in SH binaural reproduction have used basic SH channel weighting for the whole frequency spectrum. This is due to order truncation principles which utilise the same HRTF dataset regardless of reproduction order. In this chapter, the effects of implementing TA for Ambisonic reproduction are investigated with both basic SH weighting over the whole frequency spectrum and for dual-band decoding.

Time alignment of virtual loudspeaker HRTFs is achieved in this chapter as follows. The virtual loudspeakers are filtered using a low-pass filter at $f_c = 500$ Hz, using a filter order of 8. The time difference between both left and right filtered signals for all HRIRs $\{l = 1, l = 2, ..., l = L\}$ and the left signal of the first filtered HRIR in the loudspeaker configuration ($l = 1$) is then calculated in samples using the maximum of the cross-correlation method detailed in Section 2.8.3. For each signal of the virtual loudspeaker HRTFs (without the low-pass filtering), the time delay is then removed by shifting the signals forward or backward by the necessary amount of samples.

As the time delays are calculated with respect to the left signal of the first HRIR in the loudspeaker configuration, a second step is necessary to remove any shared time delay between all HRIRs, to instead align with the centre of the head. This is achieved by calculating the arithmetic mean of all HRIRs in the loudspeaker configuration, which gives a single average HRIR, a process repeated for both the original virtual loudspeaker HRIRs and those with time delay removal. The shared time delay between the two is then calculated in samples using the cross-correlation method in Section 2.8.3 and removed as above from every individual HRIR.

As TA is only implemented at high frequencies, the time delay removed HRIRs are then combined with the original virtual loudspeaker HRIRs using a crossover filter such that high frequencies have time-aligned HRIRs and low frequencies preserve the original ITD. The crossover used is the same as in Section 3.3.3. The crossover frequency $f_c$ is dependent on the Ambisonic order. However, according to listening test results in Schörkhuber, Zaunschirm and Höldrich (2018) which showed that removal of high frequency ITD is perceivable at frequencies as high as 2.5 kHz,

TABLE 7.1: Combinations of the pre-processing techniques used in this chapter, including whether or not dual-band decoding is used.

| Condition | Dual-band | TA | AIO | DFE |
|---|---|---|---|---|
| NPP | ✓ | | | |
| PP 1 | ✓ | | ✓ | ✓ |
| PP 2 | | ✓ | | ✓ |
| PP 3 | | ✓ | ✓ | ✓ |
| PP 4 | ✓ | ✓ | ✓ | ✓ |

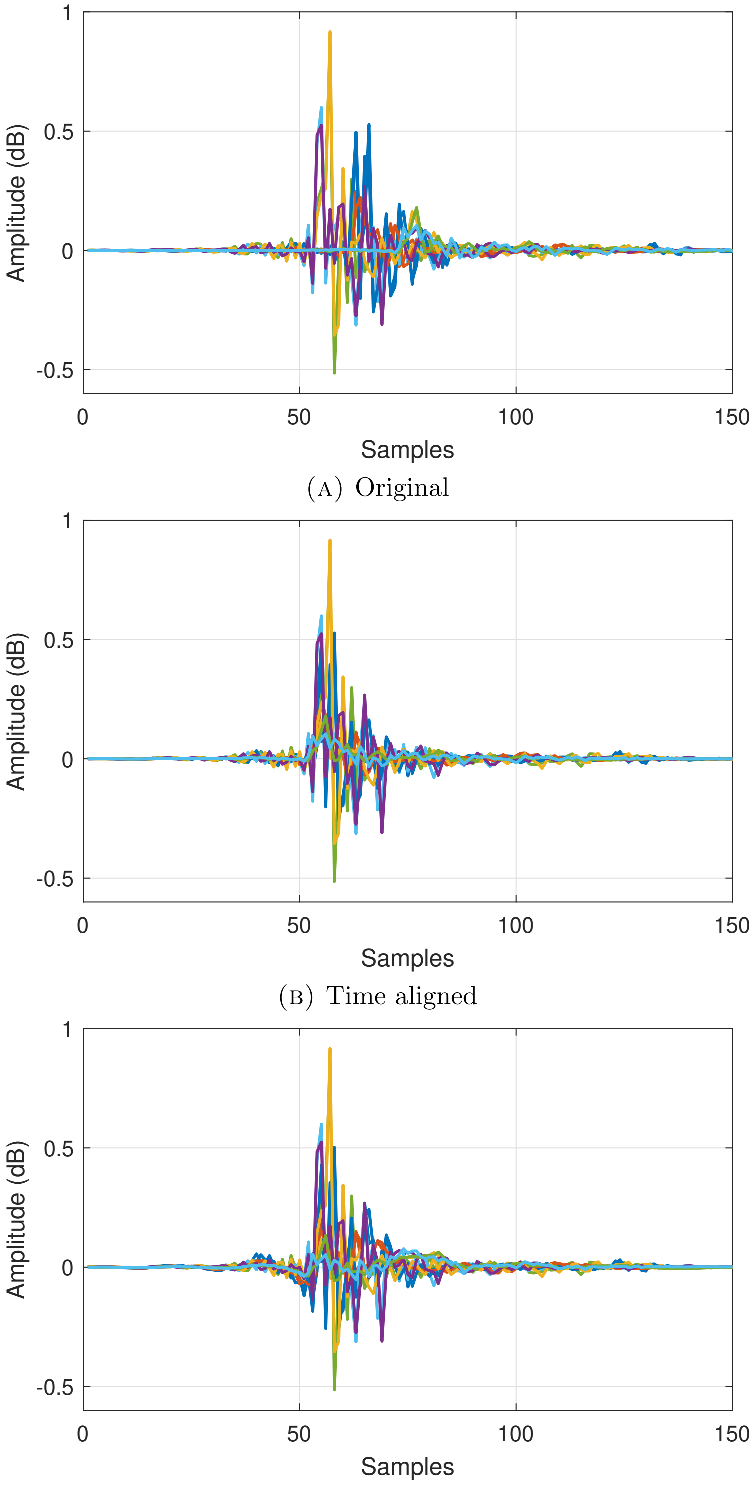
the value of $f_c$ utilised in this chapter was chosen as $f_c = 2.5$ kHz if $M < 4$, and $f_c = f_{\text{alias}}$ otherwise.

To illustrate the temporal changes of time alignment, Figure 7.1 presents the three stages of time alignment for $M = 1$ using the $L = 6$ Lebedev configuration. The original HRIRs show significant time delays between the different HRIR signals. The full-band time aligned HRIRs show how the time delays have been reduced for all HRIR signals, which is especially evident by the shared peak at approximately sample 55. The final dual-band time aligned HRIRs show a greater amount of low frequency ripple than the full-band time aligned, illustrating the original timing of low frequency content, and new timing of high frequency content.

### 7.1.2 Pre-Processing Technique Ordering

The ordering of pre-processing techniques is as follows: TA is implemented first as it affects the rendering of ILD and the diffuse-field response. AIO also affects the diffuse-field response, so follows TA. DFE is implemented last, as it addresses any changes in average frequency response and the other pre-processing techniques can affect the diffuse-field response. The five binaural Ambisonic decoders under test in this chapter (along with their abbreviations) are presented in Table 7.1, where PP denotes pre-processing and NPP denotes no pre-processing.

For all other chapters in this thesis, dual-band Ambisonic decoding is utilised, with basic channel weightings at frequencies below $f_{\text{alias}}$, and Max $\mathbf{r_E}$ weightings above (Gerzon and Barton, 1992; Daniel, Rault and Polack, 1998). When using TA however,



(A) Original

(B) Time aligned

(C) Time aligned at high frequencies, original at low frequencies

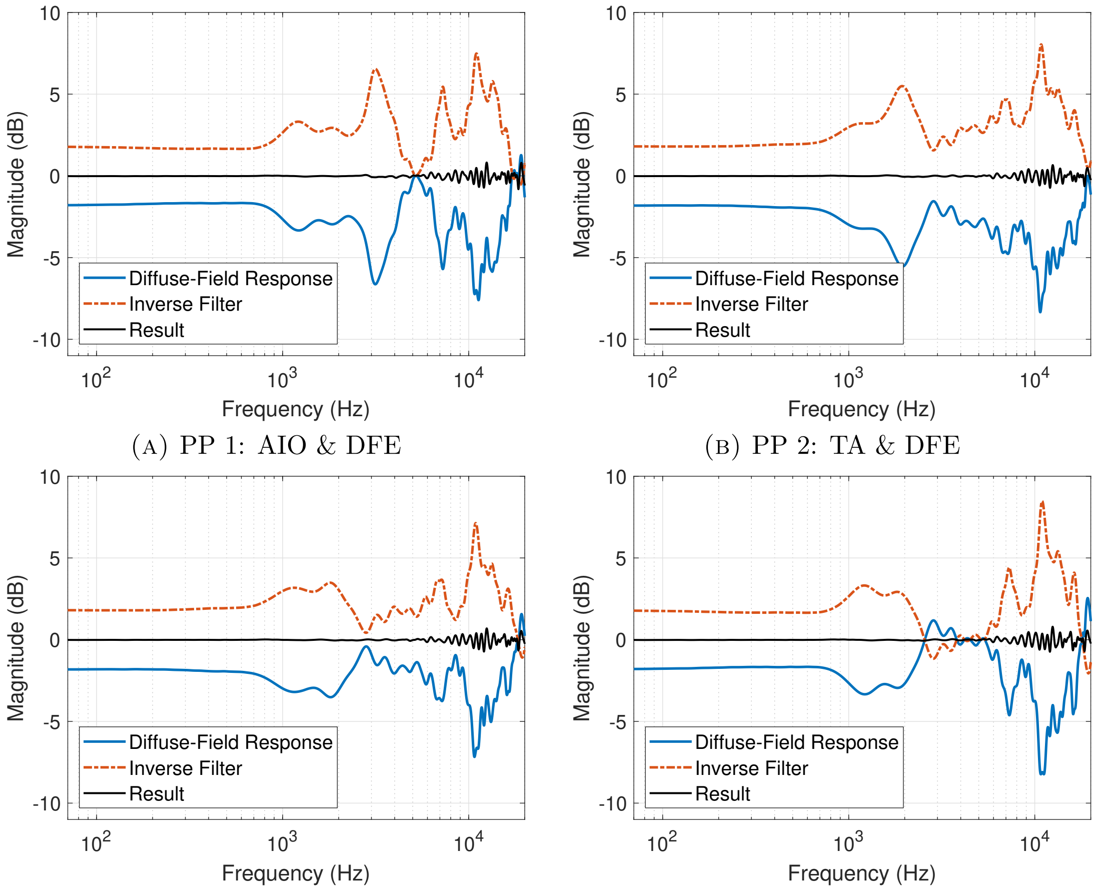
Figure 7.1: Virtual loudspeaker HRIRs of $L = 6$ for $M = 1$ with and without time alignment.

the method presented in Zaunschirm, Schörkhuber and Höldrich (2018) uses basic weighting for the full frequency spectrum. Therefore, PP 2 and PP 3 use single band basic SH channel weighting for the entire frequency spectrum. As Ambisonic ILD reproduction is improved with Max $\mathbf{r_E}$ SH channel weighting (Daniel, Rault and Polack, 1998), PP 1 and PP 4 use dual-band decoding. Therefore, in PP 3 and PP 4 with the combination of all three pre-processing techniques, both basic weighted and dual-band instances are included.

The diffuse-field responses, inverse filters and resulting convolved responses of PP 1 - 4 for $M = 1$ are presented in Figure 7.2, and corresponding plots for $M = 2, M = 3, M = 4$ and $M = 5$ are presented in Appendices D.1, D.2, D.3 and D.4, respectively. In general, these show how the implementation of TA with basic weighting (PP 2 and PP 3), produces a more uniform diffuse-field response at high frequencies: a trend that becomes more defined as Ambisonic order increases, and TA with Max $\mathbf{r_E}$ (PP 4) produces a boost in high frequencies. This is due to the reduced comb filtering from time-alignment, in conjunction with the Max $\mathbf{r_E}$ normalisation as explained in Section 3.3.2 and Figure 3.13. A narrow notch is observed in the diffuse-field responses with TA at $f_c$, which is due to the crossover between the time-aligned and non-time-aligned HRTFs which causes destructive interference. One way to potentially reduce this effect in future developments is to calculate the group delay and use all-pass filters with frequency dependent phase delay (Zaunschirm, Schörkhuber and Höldrich, 2018).

## 7.2 Numerical Evaluation

In this section, the effect of different pre-processing technique combinations is evaluated numerically by comparing binaural Ambisonic renders to a reference dataset of HRTFs. The metrics of evaluation include perceptual spectral difference (PSD), interaural cue accuracy and estimated horizontal and vertical localisation. Finally, the effect of different pre-processing technique combinations on binaural



(A) PP 1: AIO & DFE

(B) PP 2: TA & DFE

(C) PP 3: TA & AIO & DFE (basic)

(D) PP 4: TA & AIO & DFE (dual-band)

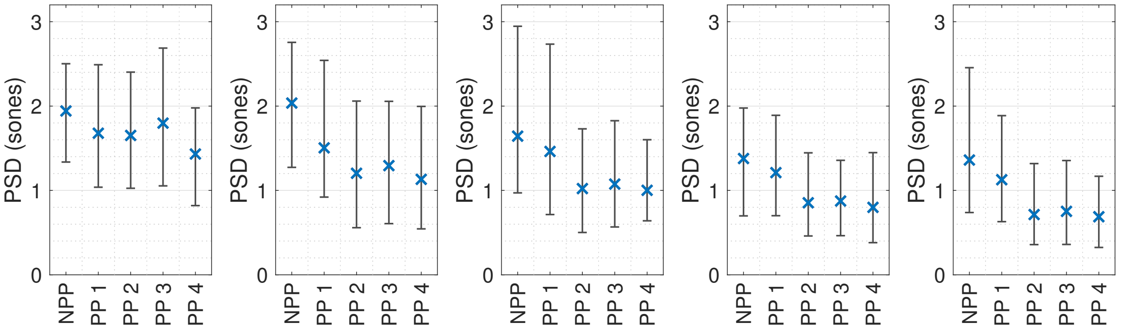
FIGURE 7.2: Diffuse-field response, inverse filters and resulting responses of different pre-processing technique combinations, $M = 1$ (left ear).

Ambisonic rendering using other loudspeaker configurations and alternative HRTFs is also explored.

For each measurement location in the reference HRTF dataset of $Q = 16{,}020$, as illustrated in Figure 2.28, binaural Ambisonic renders were generated for $\{M = 1, M = 2, ..., M = 5\}$, with the four pre-processing technique combinations as detailed in Table 7.1. All HRTFs were truncated to 1024 taps with 50 sample in / out half-Hanning windows applied.

### 7.2.1 Perceptual Spectral Difference

PSD between binaural Ambisonic renders and HRTFs for all measurement locations on the sphere with different combinations of pre-processing techniques using the

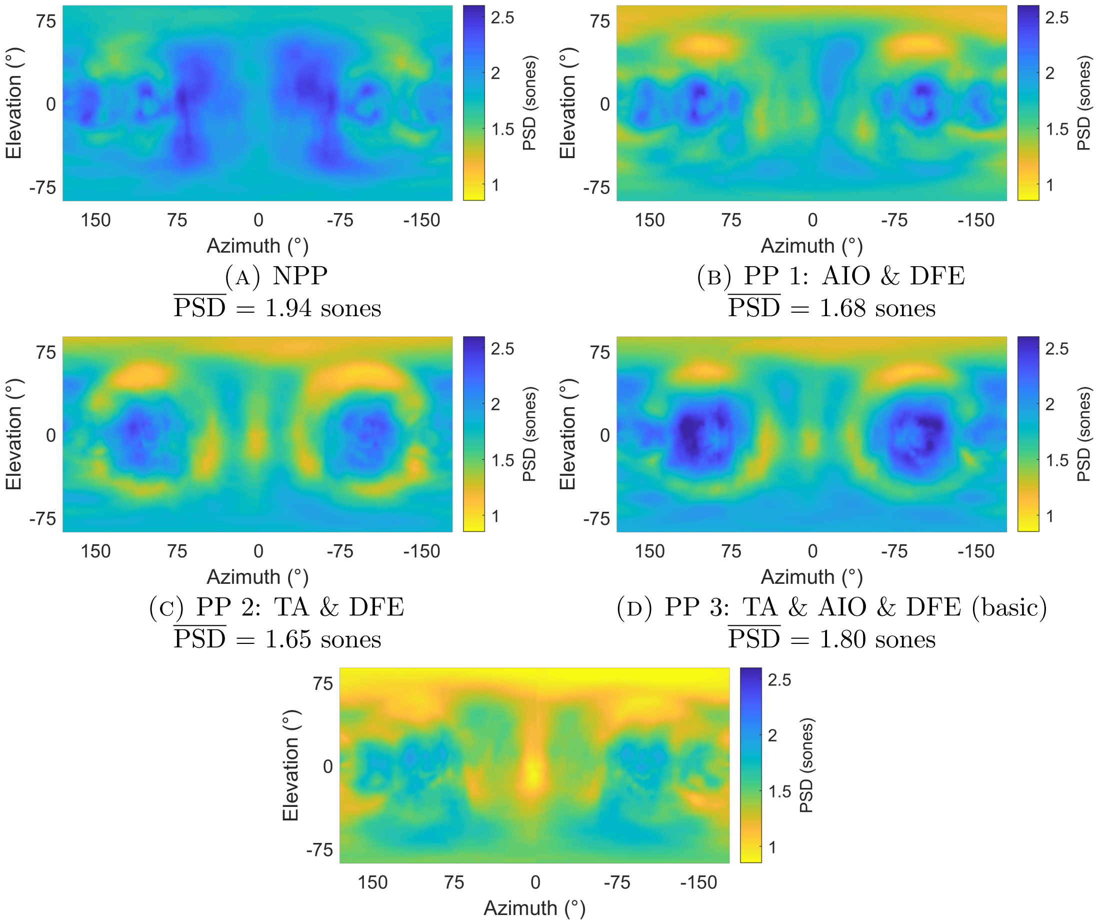


(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

FIGURE 7.3: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the minimum and maximum absolute PSD values.

method detailed in Section 4.2. Figure 7.3 shows the solid angle weighted $\overline{\text{PSD}}$ value for each pre-processing combination for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum and minimum PSD values. In all tested Ambisonic orders, every pre-processing technique combination improves the overall spectral accuracy over binaural Ambisonic decoding with NPP, but PP 4 (the dual-band combination of TA, AIO and DFE) produces the greatest improvements with the lowest $\overline{\text{PSD}}$ for all tested orders of Ambisonics. Additionally, PP 4 produces the lowest maximum PSD value for all but $M = 4$ and the lowest minimum value for all but $M = 3$. Other observable trends are that TA significantly improves spectral reproduction at $M > 1$, and there are not significant differences between the spectral reproduction of PP 2, PP 3 and PP 4 at $M > 1$.

To illustrate how PSD changes over direction, Figure 7.4 presents the absolute values of PSD between binaural Ambisonic rendering and reference HRTFs for each tested combination of pre-processing techniques for every measurement location on the sphere (mean of left and right ear PSD calculations) at $M = 1$. Corresponding plots for Ambisonic orders $M = 2, M = 3, M = 4$ and $M = 5$ are presented in Appendices D.5, D.6, D.7 and D.8, respectively. In general, they corroborate the similarities between PP 2, PP 3 and PP 4 at $M > 1$, though PP 4 does appear to produce a more even spectral reproduction over the sphere, especially for lateral locations.



(A) NPP
$\overline{\text{PSD}} = 1.94$ sones

(B) PP 1: AIO & DFE
$\overline{\text{PSD}} = 1.68$ sones

(C) PP 2: TA & DFE
$\overline{\text{PSD}} = 1.65$ sones

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\text{PSD}} = 1.80$ sones

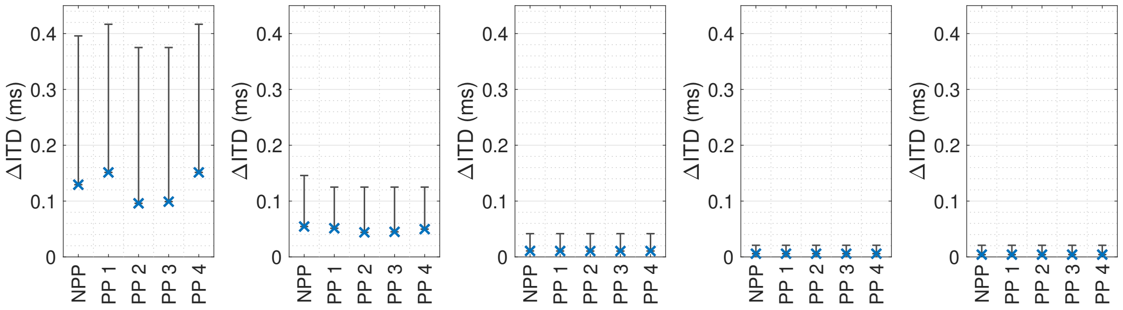
(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\text{PSD}} = 1.43$ sones

FIGURE 7.4: PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 1$ (mean of left and right PSD values).

### 7.2.2 Interaural Cues

To assess the effect of different pre-processing technique combinations on binaural Ambisonic reproduction of interaural cues, both ITD and ILD were estimated using the methods detailed in Sections 2.8.3 and 2.8.2, respectively, for all measurement locations and Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$. The low-pass filter used in the ITD calculation was $f_c = 1.5$ kHz.

The change in ITD between the reference HRTFs and the Ambisonic generated HRTFs was then calculated for each measurement location using (4.4), and $\overline{\Delta\text{ITD}}$ is calculated using (4.5). Figure 7.5 displays the solid angle weighted $\overline{\Delta\text{ITD}}$ values



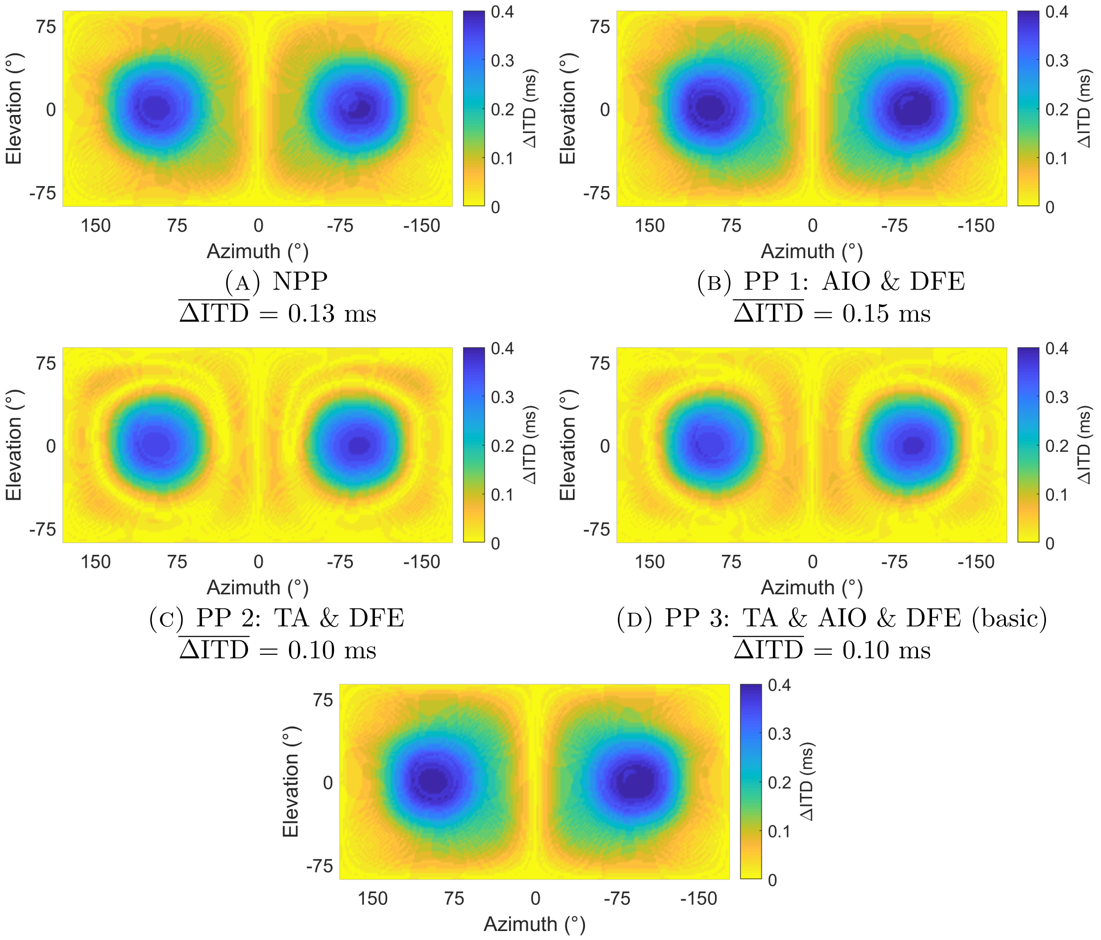
(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

FIGURE 7.5: Estimated $\overline{\Delta \text{ITD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔITD value.

between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute ΔITD value. This shows there is an insignificant difference in ITD reproduction between different pre-processing technique combinations for $M \geq 3$, and only very minor differences for $M = 2$. However, for $M = 1$, it is possible to see how PP 2 and PP 3 produce somewhat improved ITD reproduction. This is likely due to the use of basic channel weighting throughout the frequency spectrum, which maximises the reproduction of $\mathbf{r_V}$ and therefore reproduces temporal cues more accurately. This suggests that increasing the $f_c$ of the dual-band crossover network in low-order Ambisonics could improve ITD reproduction.

Detailed plots of ΔITD for every measurement location on the sphere are presented in Figure 7.6 and Appendix D.9 for $M = 1$ and $M = 2$, respectively. Plots for $M \geq 3$ are omitted because the change is insignificant, as shown in Figure 7.5. The plots for $M = 1$ show that, in general, the improvements in ITD reproduction for PP 2 and PP 3 occur in the regions approximately ±30° from the median plane, thus the ITD reproduction at the lateral extremes is still poorly reproduced.

The change in ILD between the reference HRTFs and the Ambisonic generated HRTFs, ΔILD, was calculated for all measurement locations using (4.6), and the solid angle weighted $\overline{\Delta \text{ILD}}$ values calculated using (4.7). Figure 7.7 presents the solid angle weighted $\overline{\Delta \text{ILD}}$ values between HRTFs and binaural Ambisonic rendering



(A) NPP
$\overline{\Delta \text{ITD}} = 0.13$ ms

(B) PP 1: AIO & DFE
$\overline{\Delta \text{ITD}} = 0.15$ ms

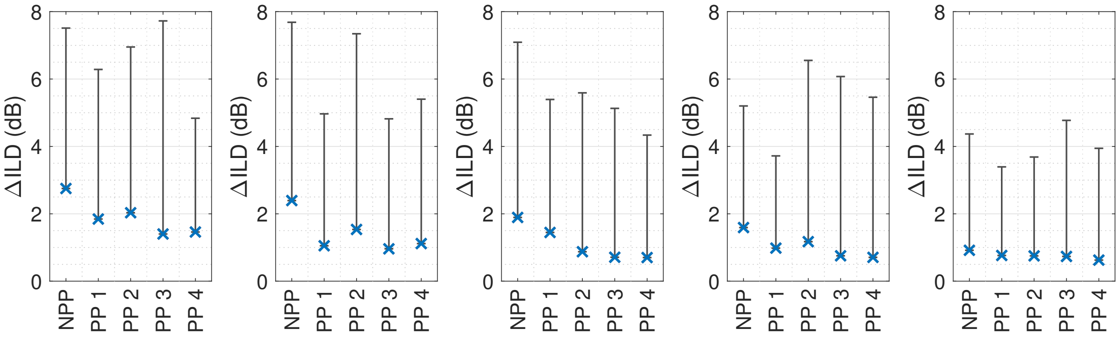
(C) PP 2: TA & DFE
$\overline{\Delta \text{ITD}} = 0.10$ ms

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta \text{ITD}} = 0.10$ ms

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta \text{ITD}} = 0.15$ ms

FIGURE 7.6: ΔITD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 1$. NPP denotes no pre-processing.

with different pre-processing technique combinations for $\{M = 1, M = 2, ..., M = 5\}$, across all measurement locations, along with the maximum absolute ΔILD value. As expected, higher orders of Ambisonics produce improved ILD rendering, and in all tested Ambisonic orders, every pre-processing technique combination improves the overall ILD reproduction over standard binaural Ambisonic rendering, as shown by the lower $\overline{\Delta \text{ILD}}$ values. When TA is implemented in conjunction with AIO (as in PP 3 and PP 4), the lowest $\overline{\Delta \text{ILD}}$ values are produced for all but $M = 2$, where PP 1 shows better performance than PP 4. However, results in general vary depending on Ambisonic order. There is not a single pre-processing technique combination that offers the best Ambisonic ILD reproduction.



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

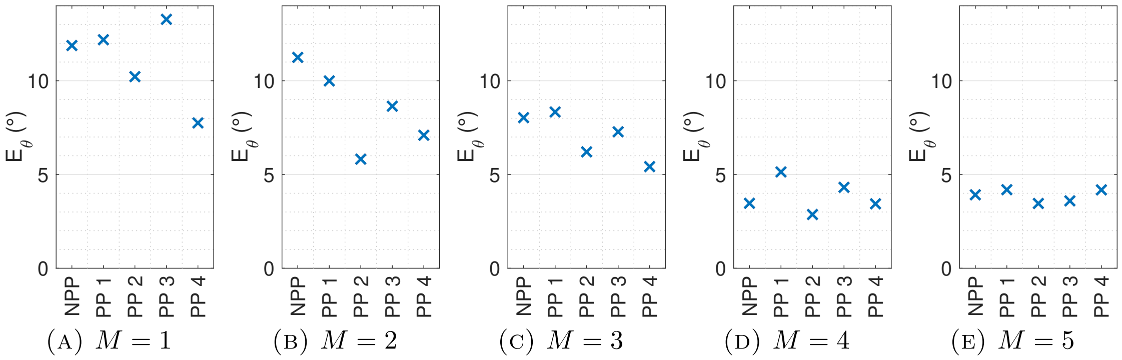
FIGURE 7.7: Estimated $\overline{\Delta \text{ILD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$, with whiskers to denote the maximum ΔILD value.

Detailed plots of ΔILD for every measurement location on the sphere for $\{M = 1, M = 2, ..., M = 5\}$ with different pre-processing technique combinations are presented in Appendices D.10, D.11, D.12, D.13 and D.14, respectively. The least accurate ILD reproduction occurs at lateral regions.

### 7.2.3 Estimated Localisation

The effect of different pre-processing technique combinations on estimated auditory localisation of binaural Ambisonic rendering was assessed. Horizontal localisation was estimated using the method detailed in Section 2.8.4, utilising a horizontal model (May, Van De Par and Kohlrausch, 2011), producing a value of $\overline{E_\theta}$ for overall estimated localisation between $-90° < \theta < +90°$ at $\phi = 0°$ using (2.17).

Figure 7.8 displays the overall estimated horizontal localisation of binaural Ambisonic rendering with different pre-processing technique combinations for $\{M = 1, M = 2, ..., M = 5\}$. For $M \leq 3$, the two pre-processing technique combinations that produce the most accurate estimated horizontal localisation are PP 2 and PP 4. For $M > 3$ standard binaural Ambisonic rendering is reasonably accurate at $\overline{E_\theta} < 5°$, and the pre-processing technique combinations make less of an improvement, though PP 2 does produce the lowest value of $\overline{E_\theta}$ here. Considering how the results from using the reference dataset in the model in Section 2.8.4 produce $\overline{E_\theta} = 3.52°$, this shows



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

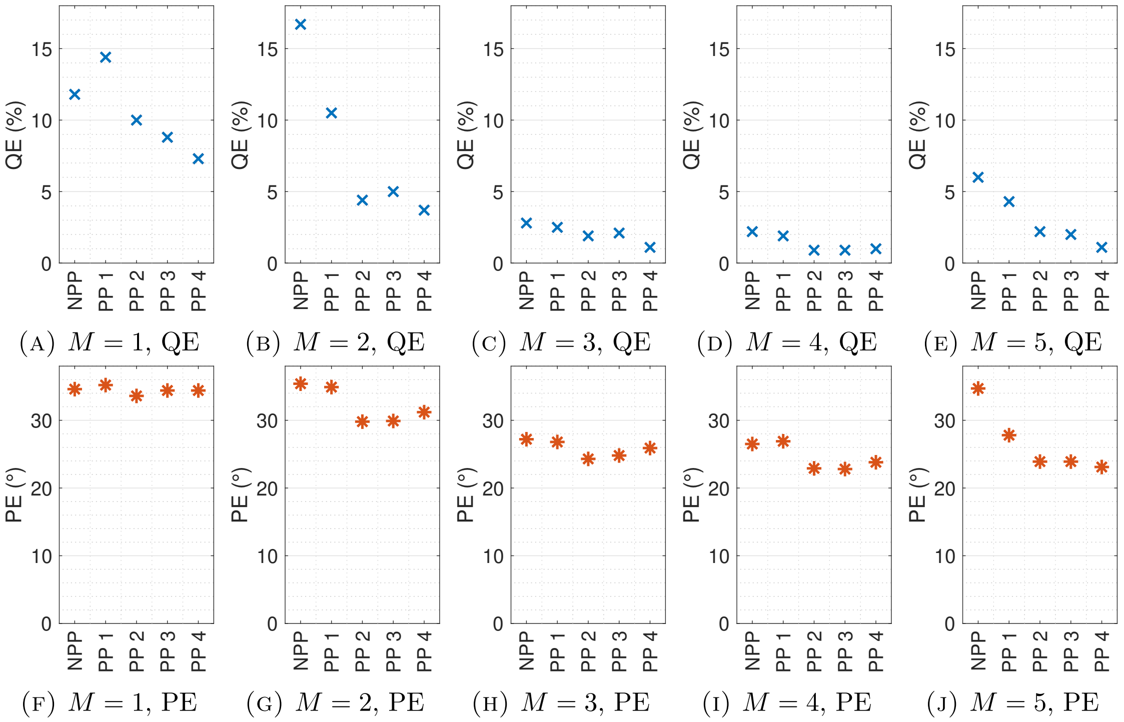
FIGURE 7.8: Estimated $\overline{E_\theta}$ of binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (May, Van De Par and Kohlrausch, 2011).

how binaural Ambisonic rendering is capable of achieving comparable horizontal localisation accuracy to HRTFs, for $M \geq 4$.

Detailed individual plots of the estimated horizontal localisation of each azimuth angle for $\{M = 1, M = 2, ..., M = 5\}$ are presented in Appendices D.15, D.16, D.17, D.18 and D.19, respectively, which show the predicted improvement in lateralisation for $M = 1$, which is greatest at PP 4. They also show that, for $M = 4$ and $M = 5$, there is little error between $-70° < \theta < +70°$.

The effect of different pre-processing technique combinations on estimated elevation localisation in the sagittal plane was evaluated between $-90° < \phi < +90°$ at $\theta = 0°$ using the method detailed in Section 2.8.4 which utilises a localisation model (Baumgartner, Majdak and Laback, 2014) producing two metrics: quadrant error (QE), a prediction of localisation confusion (in %), and polar RMS error (PE), a prediction of precision (in °). As the HRTFs used are of a Neumann KU 100, which has no torso, there will be no elevation cues present below 1.5 kHz (Algazi, Avendano and Duda, 2001a). Therefore, the frequency range of the model's filter bank was set to 1.5 kHz - 18 kHz, with the upper limit of the frequency range chosen as the highest frequency of perceivable elevation cues (Roffler and Butler, 1967; Asano, Suzuki and Sone, 1990).

Figure 7.9 shows the predicted QE and PE values of binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$,



(A) $M = 1$, QE (B) $M = 2$, QE (C) $M = 3$, QE (D) $M = 4$, QE (E) $M = 5$, QE

(F) $M = 1$, PE (G) $M = 2$, PE (H) $M = 3$, PE (I) $M = 4$, PE (J) $M = 5$, PE

FIGURE 7.9: Estimated sagittal plane localisation plots of binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$, calculated using a perceptual model (Baumgartner, Majdak and Laback, 2014).

and detailed individual plots of predicted sagittal plane localisation for $\{M = 1, M = 2, ..., M = 5\}$ are presented in Appendices D.20, D.21, D.22, D.23 and D.24, respectively. The pre-processing technique combinations with TA (PP 2, PP 3 and PP 4) produce the lowest QE and PE values for all tested Ambisonic orders, with the lowest QE produced by PP 4 for all but $M = 4$. The lowest PE values are produced by PP 2 for $M \leq 3$. These suggest that all tested pre-processing technique combinations improve sagittal plane localisation over standard binaural Ambisonic rendering. For $M = 5$ both QE and PE are lowest at PP 4, which is supported by the plots in Appendix D.24 which shows increased precision around $\phi = -60°$. When considering the results from using the reference dataset in the model in Section 2.8.4 produced QE = 0.6% and PE = 21.5°, this shows how binaural Ambisonic rendering is capable of producing predicted sagittal plane localisation accuracy approaching HRTF rendering at $M \geq 4$.

### 7.2.4 Generalisability

To demonstrate the generalisable effect of different pre-processing technique combinations on binaural Ambisonic rendering, additional numerical tests were performed using both different loudspeaker configurations and an alternative HRTF dataset. In both sets of simulations, the effect of AIO was assessed by comparing Ambisonic renders to the original HRTFs for all available measurement locations. Two numerical evaluation metrics were investigated: PSD and ILD, with PSD calculations made using the method detailed in Section 4.2 and single values of $\overline{\text{PSD}}$ then calculated using (4.3), and ILD calculations using (4.6), and $\overline{\Delta\text{ILD}}$ calculated using (4.7).

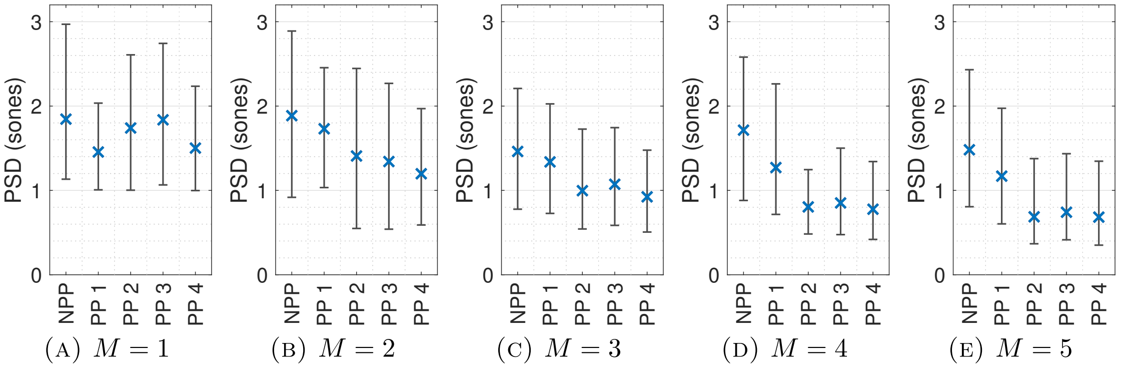
Firstly, different loudspeaker configurations were investigated[1]. Here, spherical T-designs (Hardin and Sloane, 1996) were employed with corresponding loudspeaker vertices as illustrated in Figure 3.8. The solid angle weighted $\overline{\text{PSD}}$ results, calculated from all locations on the sphere, are shown in Figure 7.10, along with the maximum and minimum absolute PSD values. These show similar results to the corresponding calculations using Lebedev configurations in Figure 7.3, with PP 4 producing the lowest $\overline{\text{PSD}}$ value for all but $M = 1$ and the lowest absolute PSD value for all but $M = 2$.

Values of $\overline{\Delta\text{ILD}}$ for orders of Ambisonics $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations are presented in Figure 7.11, along with the maximum $\Delta$ILD value. PP 4 produces the lowest $\overline{\Delta\text{ILD}}$ value for $M = 2, M = 3$ and $M = 5$. An interesting observation is that PP 2 produces high values of maximum $\Delta$ILD for $M \leq 3$.

Secondly, to assess the effect of different pre-processing technique combinations when using an alternative HRTF dataset, binaural Ambisonic renders were made for $\{M = 1, M = 2, M = 3, M = 5\}$[2] using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, human subject H20 (Armstrong et al., 2018a). The solid angle weighted $\overline{\text{PSD}}$ results, calculated from 2,114 locations

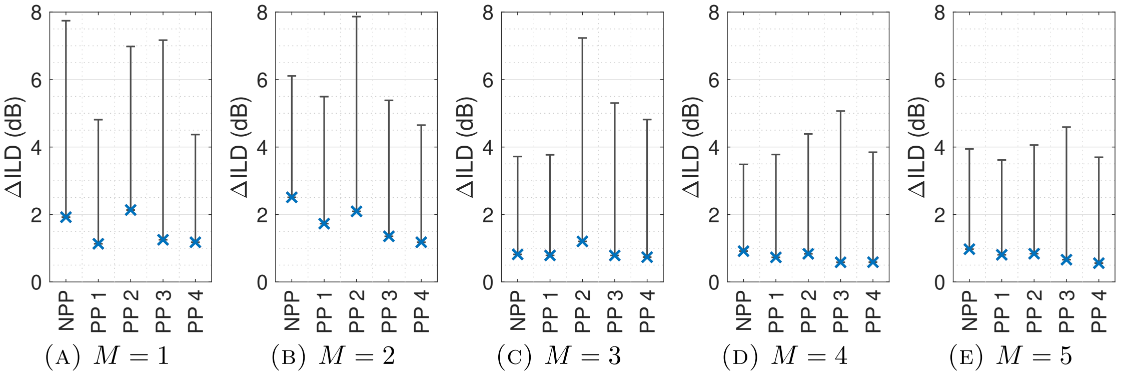
[1] For all other areas of this chapter, Lebedev loudspeaker configurations are used.

[2] The omission of $M = 4$ was due to a lack of necessary measurements.



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

FIGURE 7.10: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the minimum and maximum absolute PSD values.



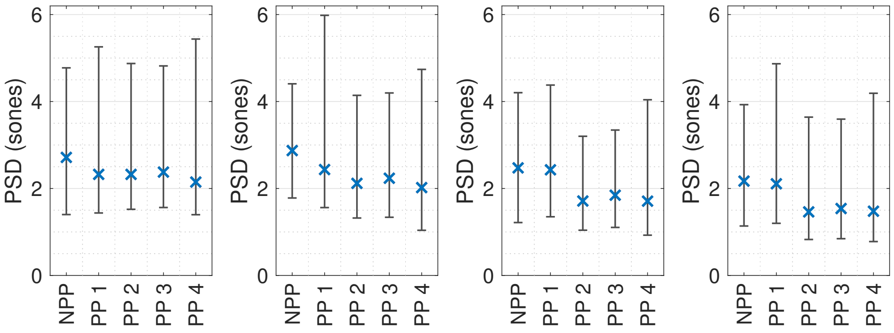
(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 4$ (E) $M = 5$

FIGURE 7.11: Estimated $\overline{\Delta\text{ILD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations, with whiskers to denote the maximum ΔILD value.

on the sphere, are shown in Figure 7.12 along with the minimum and maximum absolute PSD values. Here, PP 4 produces the lowest $\overline{\text{PSD}}$ value for all but $M = 5$, and the lowest minimum absolute PSD value for all tested orders of Ambisonics.

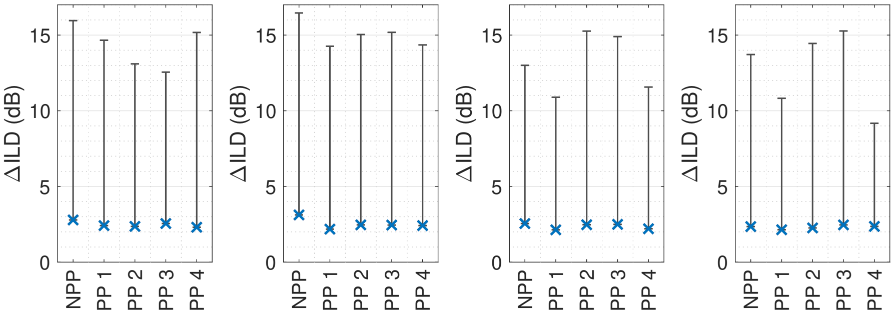
Values of $\overline{\Delta\text{ILD}}$ for orders of Ambisonics $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database are presented in Figure 7.13, along with the maximum ΔILD value. PP 4 produces the lowest $\overline{\Delta\text{ILD}}$ value for $M = 2, M = 3$ and $M = 5$.

The tests on generalisability therefore show how combinations of pre-processing techniques can be used in binaural Ambisonic rendering with different virtual loudspeaker configurations and alternative HRTF datasets, and that improvements are



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 5$

FIGURE 7.12: $\overline{\text{PSD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the minimum and maximum absolute PSD values.



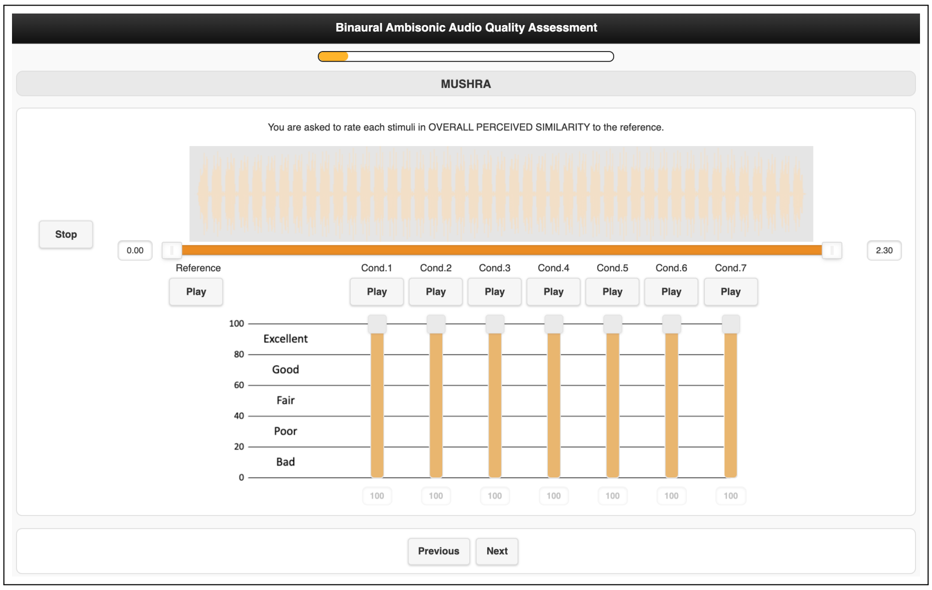
(A) $M = 1$ (B) $M = 2$ (C) $M = 3$ (D) $M = 5$

FIGURE 7.13: Estimated $\overline{\Delta\text{ILD}}$ between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, M = 3, M = 5\}$ using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a), with whiskers to denote the maximum ΔILD value.

still observed. It appears that PP 4 produces the best overall improvements.

## 7.3 Perceptual Evaluation

To assess the perceptual effect of different pre-processing technique combinations, listening tests were conducted using three different acoustic scenes. As the objective evaluation showed the biggest differences between pre-processing technique combinations occur at lower Ambisonic orders (particularly interaural cue reproduction and predicted localisation), the perceptual evaluation focused on $M \leq 3$ to reduce the overall duration of the tests, using Lebedev loudspeaker configurations.



Binaural Ambisonic Audio Quality Assessment

MUSHRA

You are asked to rate each stimuli in OVERALL PERCEIVED SIMILARITY to the reference.

Figure 7.14: Screenshot of the MUSHRA interface used in the pre-processing technique comparison listening test using ‘webMUSHRA’ (Schoeffler et al., 2018).

The tests followed the multiple stimulus with hidden reference and anchors (MUSHRA) paradigm, ITU-R BS.1534-3 (International Telecommunication Union, 2015b). A screenshot of the MUSHRA interface, using the web based ‘webMUSHRA’ (Schoeffler et al., 2018), is presented in Figure 7.14. Tests were conducted in a quiet listening room (background noise level of 41.8 dBA) using an Apple Macbook Pro with a Fireface 400 audio interface, which has software controlled input and output levels. A single set of Sennheiser HD 650 circumaural headphones were used, which were equalised using the Neumann KU 100 as in Section 4.4 (see Figure 4.24 for the RMS HpTF and inverse filter of the left headphone, along with a resulting convolved response). 20 experienced listeners took part, aged between 22 and 41 (16 male, 4 female), with no reported knowledge of any hearing impairments according to ISO Standard 389 (International Organization for Standardization, 2016). All reported prior critical listening experience, which was deemed sufficient if the participant had education or employment in audio or music engineering.

Test sounds were normalised to a consistent A-weighted RMS amplitude and participants were able to adjust the playback level. All binaural renders were static (fixed head orientation) to ensure consistency in the experience between participants.

### 7.3.1 Test Paradigms

Listeners compared binaural Ambisonic renders created using the pre-processing combinations as throughout this chapter. Three types of stimuli were used in the listening test. All trials were repeated once, and stimuli and trial ordering was randomised and presented double blind.

The first stimuli was a pseudo-moving pink noise sound. This was generated using 45 bursts of pink noise played consecutively and lasting 0.05 seconds long each, panned between $(\theta = 44°, \phi = 0°)$ and $(\theta = 132°, \phi = 0°)$ in 2° increments, which creates the impression of a moving noise source. Only one noise burst would be played at any given time, so is similar in temporal structure to a click train, as used in Goupell, Majdak and Laback (2010) and Moore, Tew and Nicol (2010), which helps assess any temporal changes to the binaural rendering. The pink noise stimuli was chosen due to it featuring energy at all frequencies. The panning trajectory was chosen as it passes through the lateral extreme, which is beneficial for assessing lateralisation of the binaural rendering. Each burst was windowed using a 50 sample Hanning window, resulting in a stimulus of duration 2.25 seconds. The reference was made from the summation of direct HRTF convolutions, and a monophonic version of the HRTF reference low-passed at 3.5 kHz was used as the low anchor, giving a total of 7 conditions per trial. Participants were asked to rate each condition in overall perceived similarity to the reference with a score between 0 and 100.

The second stimuli was a synthesised complex scene which comprised of 8 monophonic percussive sounds panned to 8 of the centre vertices of the faces of a dodecahedron. Table 7.2 presents the test sound locations, along with a description of each sound. The percussive nature of the sounds was chosen to allow the assessment of any temporal changes to the binaural rendering, and the sounds varied in frequency response and duration, such as the kick drum, which features high amplitude at low frequencies, the hi-hat, which features high amplitude at high frequencies, and the pitched staccato chord on a keyboard. This soundscene therefore offers a closer representation of a realistic musical soundfield than the moving noise stimuli, whilst

TABLE 7.2: Spherical coordinates of percussion sound stem locations.

| $\psi$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Sample | Tom | Hi-Hat | Keys | Block | Snare | Kick | Shaker | Tambourine |
| $\theta$ (°) | 50 | 310 | 118 | 242 | 0 | 180 | 62 | 230 |
| $\phi$ (°) | 46 | 46 | 16 | 16 | 0 | 64 | $-16$ | $-46$ |

still offering an HRTF reference. The reference was created by summing direct HRTF convolutions of the 8 original tracks, and again a monophonic version of the reference low-passed at 3.5 kHz was used as the low anchor, giving a total of 7 conditions per trial. Participants were again asked to rate each condition in overall perceived similarity to the reference with a score between 0 and 100.

The third stimuli was a 5 second excerpt of a beach soundscape recording (the same as used in Chapter 6) from the open source EigenScape database of $M = 4$ Ambisonic recordings made using an MH Acoustics em32 Eigenmike[3] (Green and Murphy, 2017). The initial format of recording follows SN3D normalisation, which therefore was converted to N3D normalisation using (3.10). This stimuli offers a real recording of a natural soundfield and thus allows the assessment of the plausibility of the binaural rendering, due to the varied distance, width, movement and frequency content of the multiple sound sources in the scene. As the Eigenmike recording test could not use a direct HRTF convolution render as a reference, listeners were in this case asked to rate the stimuli in terms of plausibility, which was defined as, 'a simulation in agreement with the listener's expectation towards a corresponding real event' (Lindau and Weinzierl, 2012). An anchor was included as a monophonic version of the Ambisonic render ($M = 0$) with no pre-processing, giving a total of 6 conditions per trial.

### 7.3.2 Results

Overall, the tests lasted between around 20 to 35 minutes to complete. No participant's results were excluded, based on the criteria of rating the hidden reference

[3] `https://mhacoustics.com/`

less than 90% for more than 15% of trials or rating the mid-anchor higher than 90% for more than 15% of trials. Listening test data was checked for normality using the one-sample Kolmogorov-Smirnov test, which showed all data as non-normal. Therefore, results were analysed using non-parametric statistics.

Figure 7.15 presents the median scores with non-parametric 95% confidence intervals ($\text{CI}_{95}$) (Mcgill, Tukey and Larsen, 1978), calculated using (4.8), of the moving noise stimuli for $M = 1, M = 2$ and $M = 3$. For all tested Ambisonic orders, NPP was rated as the worst condition. To assess the statistical significance of the differences between pre-processing combinations, Friedman's ANOVA tests were conducted on all test stimuli and orders. For the moving noise stimuli, statistical significance was only found at $M = 3$ ($\chi^2(4) = 3.4, p = 0.5$; $\chi^2(4) = 6.3, p = 0.18$; $\chi^2(4) = 15.7, p < 0.01$ for $M = 1, M = 2$ and $M = 3$, respectively). To test whether the different pre-processing technique combinations produce a statistically significant improvement over standard binaural Ambisonic rendering, Wilcoxon signed-rank tests were conducted; the results of which are presented in Table 7.3. PP 4 produces a statistically significant improvement over NPP for all tested Ambisonic orders, and PP 1 is not statistically significant for any tested Ambisonic orders. This is surprising considering the results of PP 1 and PP 4 at $M = 1$ (Figure 7.15a) which show highly similar median results between the two conditions, yet different statistical significance results using Wilcoxon signed-rank tests. This is likely due to the pairwise nature of the Wilcoxon data comparison method.

Figure 7.16 presents the median scores with non-parametric $\text{CI}_{95}$ of the percussion stimuli for $M = 1, M = 2$ and $M = 3$. Again, NPP was rated as the worst condition for all tested orders of Ambisonics. To test for statistical significance between different test conditions, Friedman's ANOVA tests were conducted on all test stimuli and orders, which showed that different pre-processing technique combinations have a highly statistically significant effect on the similarity of binaural Ambisonic rendering and HRTF rendering for all tested orders ($\chi^2(4) = 19.7, p < 0.01$; $\chi^2(4) = 17.4, p < 0.01$; $\chi^2(4) = 34.2, p < 0.01$ for $M = 1, M = 2$ and $M = 3$, respectively). To test whether the different pre-processing technique combinations

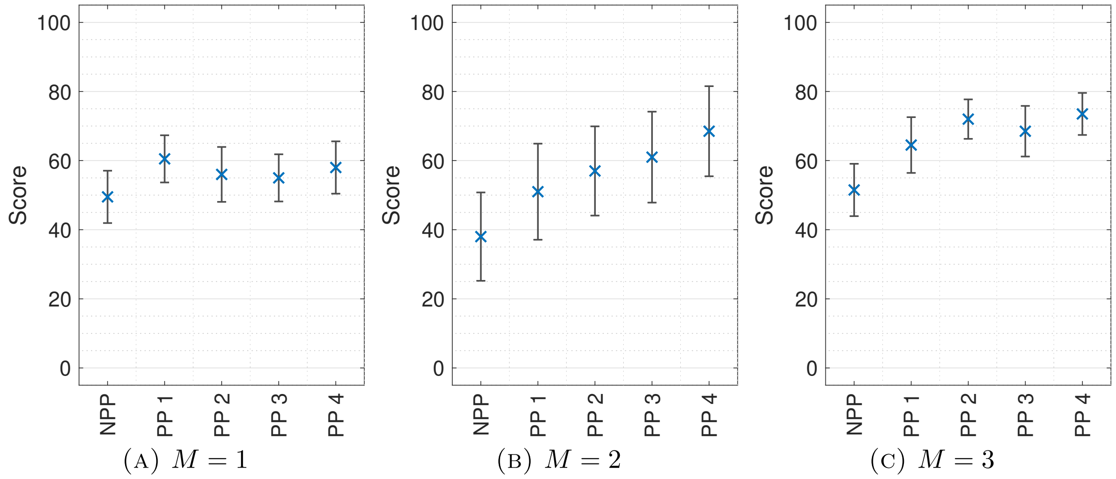


FIGURE 7.15: Median scores of the moving noise stimuli tests with non-parametric $\text{CI}_{95}$, reference and anchor scores omitted. Score indicates perceived similarity to the HRTF reference.

TABLE 7.3: Significance results of the moving noise stimuli tests using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether the pre-processing technique combination produced a statistically significant improvement to standard binaural Ambisonic rendering (NPP).

| Condition | PP 1 | PP 2 | PP 3 | PP 4 |
|---|---|---|---|---|
| $h$ ($M = 1$) | 0 | 0 | 0 | 1 |
| $h$ ($M = 2$) | 0 | 1* | 1* | 1* |
| $h$ ($M = 3$) | 0 | 1* | 1* | 1* |

produce a statistically significant improvement over standard binaural Ambisonic rendering, Wilcoxon signed-rank tests were conducted; the results of which are presented in Table 7.4. PP 2, PP 3 and PP 4 produce highly statistically significant improvements over NPP for all tested Ambisonic orders, and PP 1 is significant for $M = 2$ and $M = 3$.

Figure 7.17 presents the median scores with non-parametric $\text{CI}_{95}$ of the beach stimuli for $M = 1, M = 2$ and $M = 3$. As with the other two stimuli, NPP was rated as the worst condition for all tested orders of Ambisonics. To test for statistical significance between different test conditions, Friedman's ANOVA tests were conducted on all test stimuli and orders, which showed statistically significantly different results again only for $M = 3$ ($\chi^2(4) = 9.3, p = 0.05$; $\chi^2(4) = 7.3, p = 0.12$; $\chi^2(4) = 16.2, p < 0.01$ for $M = 1, M = 2$ and $M = 3$, respectively). To test whether the different pre-processing

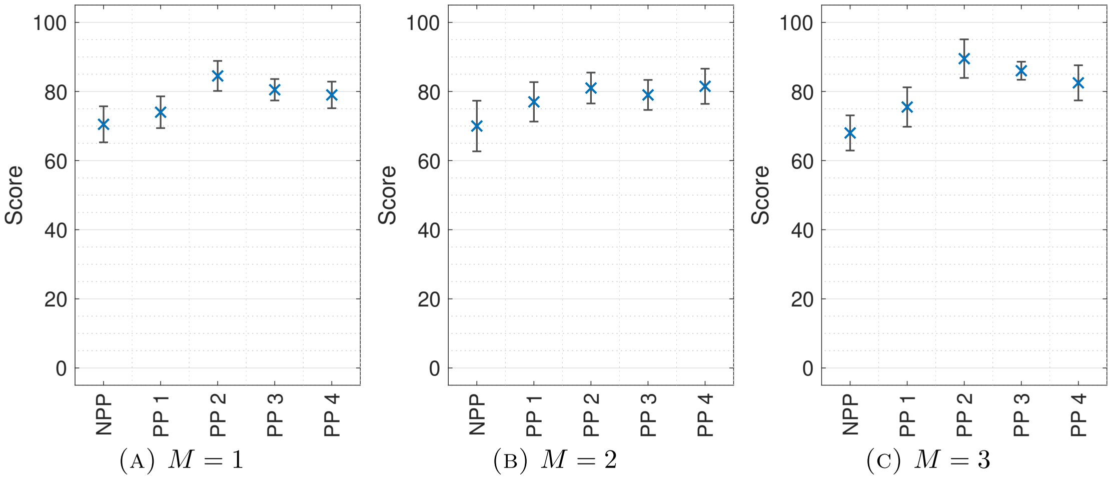


FIGURE 7.16: Median scores of the percussion stimuli tests with non-parametric $CI_{95}$, reference and anchor scores omitted. Score indicates perceived similarity to the HRTF reference.

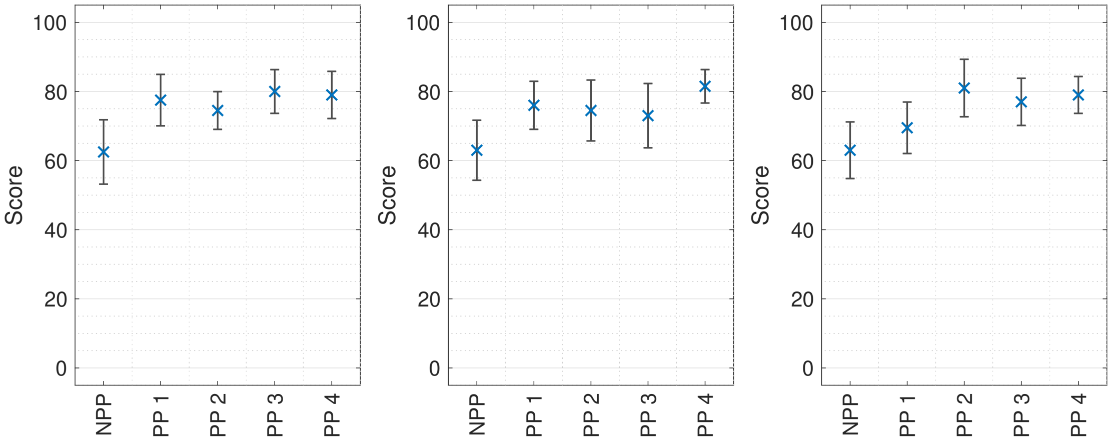
TABLE 7.4: Significance results of the percussion stimuli tests using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether the pre-processing technique combination produced a statistically significant improvement to standard binaural Ambisonic rendering (NPP).

| Condition | PP 1 | PP 2 | PP 3 | PP 4 |
|---|---|---|---|---|
| $h$ ($M = 1$) | 0 | 1* | 1* | 1* |
| $h$ ($M = 2$) | 1* | 1* | 1* | 1* |
| $h$ ($M = 3$) | 1 | 1* | 1* | 1* |

technique combinations produce a statistically significant improvement over standard binaural Ambisonic rendering, Wilcoxon signed-rank tests were conducted; the results of which are presented in Table 7.5. PP 2 and PP 4 produce statistically significant improvements over NPP for all tested Ambisonic orders.

## 7.4 Discussion

The evaluation has shown that a combination of HRTF pre-processing techniques can improve the reproduction of binaural Ambisonic rendering, when compared to HRTFs. These show greater improvements in PSD than those found when just using a single pre-processing technique.



(A) $M = 1$ (B) $M = 2$ (C) $M = 3$

FIGURE 7.17: Median scores of the beach stimuli tests with non-parametric $CI_{95}$, anchor scores omitted. Score indicates perceived plausibility.

TABLE 7.5: Significance results of the beach stimuli tests using Wilcoxon signed-rank analysis (1 indicates statistical significance at $p < 0.05$; * indicates $p < 0.01$). Values indicate whether the pre-processing technique combination produced a statistically significant improvement to the plausibility of standard binaural Ambisonic rendering (NPP).

| Condition | PP 1 | PP 2 | PP 3 | PP 4 |
|---|---|---|---|---|
| $h$ $(M = 1)$ | 0 | 1 | 1* | 1 |
| $h$ $(M = 2)$ | 1 | 1 | 0 | 1* |
| $h$ $(M = 3)$ | 1* | 1 | 1* | 1* |

This chapter has looked to determine the optimal pre-processing technique combination, which has produced the following observations. As Ambisonic order increases, the inclusion of time-alignment (TA) becomes more beneficial, with greater improvements observed in PSD and estimated sagittal plane localisation (see Figures 7.3 and 7.9, respectively). In the numerical evaluation, the pre-processing technique combinations that include TA (PP 2, PP 3 and PP 4) overall perform closer to direct HRTF rendering than those without (PP 1 and NPP). This is supported by the perceptual evaluation, with PP 1 producing the lowest number of statistically significantly different performance to standard binaural Ambisonic rendering over all tested conditions and Ambisonic orders. As found in Chapter 6, the influence of AIO is greatest at $M < 3$, and this remains true when AIO is used in conjunction with other pre-processing techniques.

Perceptual results differed with test stimuli. PP 2 performed better for the percussion stimuli type, whereas pre-processing combinations with AIO (PP 1, PP 3 and PP 4) performed better for the other two test stimuli. One possible explanation for this is that there was greater lateralisation present in the moving noise and soundscape stimuli.

Another observation is that the listening test results show comparable ratings between different Ambisonic orders. This contrasts with the results in Chapters 4 and 6, which produced results in line with the expectation that higher orders of Ambisonics are closer to HRTF rendering. A likely explanation for this is that in the listening tests in Chapters 4 and 6, each trial of the MUSHRA test required participants to directly compare different Ambisonic orders on the same MUSHRA trial. This was not feasible in this test due to the number of conditions per Ambisonic order, and required the separation of Ambisonic orders to separate trials.

Overall, it appears that PP 4 offers the greatest improvements in binaural Ambisonic rendering. This is supported by the lowest $\overline{\text{PSD}}$ values over all tested Ambisonic orders, and is the only pre-processing technique combination to show a statistically significant difference from standard binaural Ambisonic rendering for all tested conditions and Ambisonic orders.

The tests on generalisability show that, though individual results vary between Ambisonic orders, loudspeaker configurations and HRTF datasets, certain trends do emerge. AIO has a greater effect at the lowest tested Ambisonic orders, and TA has the greatest effect at higher Ambisonic orders, and again, PP 4 is the most consistent pre-processing technique combination for low PSD and low ΔILD between binaural Ambisonic rendering and HRTFs.

An interesting discovery is how an increased amplitude at high frequencies arises from implementing TA using a dual-band decoding, as illustrated in Figure 7.2d and the corresponding plot of PP 4 in Appendices D.4 to D.4. This is due to the decrease in high frequency destructive interference caused by the time-aligned HRTFs, and the normalisation stage of the Max $\mathbf{r_E}$ SH channel weights, when combined. Though in this chapter, the high frequency boost was mitigated by the diffuse-field equalisation stage, future work could look at removing the normalisation stage when implementing Max $\mathbf{r_E}$ SH channel weights and evaluating the effect, when combining pre-processing techniques.

## 7.5 Summary

Combinations of HRTF pre-processing techniques for virtual loudspeaker binaural Ambisonic rendering have been presented in this chapter. Ambisonic Time-Alignment and Ambisonic Diffuse-Field Equalisation are implemented to reduce high frequency comb filtering, and Ambisonic ILD optimisation is utilised to improve high frequency ILD reproduction. Four variations of these pre-processing technique combinations have been tested, including using basic SH channel weightings for the entire frequency spectrum as well as dual-band decoding.

Numerical evaluation has shown that combining pre-processing techniques improves the high frequency reproduction of binaural Ambisonic rendering compared to direct HRTF rendering with greater results than when using just one pre-processing technique. Though results have shown the tested combinations produce varied improvements to different aspects of reproduction accuracy, a combination of high frequency time-alignment, ILD optimisation and diffuse-field equalisation with dual-band decoding appears to produce the best overall results.

# Chapter 8

# Conclusions

A summary of the work presented in this thesis is as follows. Chapter 2 introduces the fundamental principles of sound, including the human auditory system and the psychoacoustics of sound localisation, such as the three main auditory cues: the interaural time difference and interaural level difference between the signals at the left and right eardrums, and the spectral characteristics of the signals due to interactions with the body and ears. Binaural technology, a way of synthesising spatial audio over headphones, is then introduced with the head-related transfer function (HRTF), which captures the auditory localisation cues and allows the rendering of spatial sound at a specific location. Chapter 2 ends by discussing techniques for binaural audio quality evaluation, including both numerical calculations and perceptual listening tests.

Ambisonics is introduced in Chapter 3 as a spatial audio technology that uses spherical harmonics to decompose a soundfield into a weighted set of directional functions. The binaural rendering of Ambisonic signals using virtual loudspeakers is then detailed, which allows binaural rendering with far fewer HRTF measurements than direct HRTF rendering. However, Ambisonics is only accurate up to a specific frequency. Higher-orders of Ambisonics raise the frequency limit of accurate reproduction, but require more microphone capsules in recording, increased file size in storage and transmission, and a greater number HRTF measurements in the binaural rendering stage. Chapter 3 ends by discussing the limitations of Ambisonics, including the inaccurate high frequency reproduction, which causes spatial blurring and spectral artefacts, and the current state-of-the-art approaches for improving rendering quality within the same Ambisonic order.

Chapter 4 introduces the first novel virtual loudspeaker HRTF pre-processing technique called Ambisonic Diffuse-Field Equalisation. By generating binaural Ambisonic rendered HRTFs at directions all over the sphere and then taking an average of them, an approximate diffuse-field response of the binaural Ambisonic decoder can be obtained. Equalising this using inverse filtering techniques and convolving the original virtual loudspeaker HRTFs with the calculated inverse filters produces a diffuse-field equalised binaural Ambisonic decoder. This produces an incremental improvement in the high frequency reproduction of binaural Ambisonic rendering when compared to direct HRTF rendering, however there still exist considerable differences between binaural Ambisonic rendering and direct HRTF rendering.

The method detailed in Chapter 4 is adapted in Chapter 5 to form a second novel HRTF pre-processing technique called Ambisonic Directional Bias Equalisation. Instead of producing a small improvement in spectral reproduction at all directions on the sphere, this technique produces a more significant improvement in spectral reproduction at a specified direction, to the detriment of other directions. This is achieved by introducing a directional bias in the distribution of points in the diffuse-field response calculation used to generate the equalisation filters, before introducing an additional re-equalisation stage to bring the frequency reproduction of binaural Ambisonic rendering closer to direct HRTF rendering for the specified direction of bias.

Chapter 6 introduces a third novel HRTF pre-processing technique called Ambisonic Interaural Level Difference Optimisation. This is achieved by measuring the Ambisonic interaural level difference at the position of each virtual loudspeaker of the binaural Ambisonic decoder, before augmenting left and right signals of the virtual loudspeaker HRTFs accordingly at high frequencies such that, when used for binaural

Ambisonic rendering, the resulting rendered interaural level differences match those of the direct HRTF rendering more closely. This also produces a small improvement in spectral reproduction.

Finally, Chapter 7 introduces novel combinations of multiple HRTF pre-processing techniques at once, for greater overall improvements to binaural Ambisonic rendering. The pre-processing techniques considered in this chapter are Ambisonic Time-Alignment and Ambisonic Diffuse-Field Equalisation, which improve high frequency spectral reproduction, and Ambisonic Interaural Level Difference Optimisation, which improves high frequency interaural level difference reproduction. Four variations of these pre-processing technique combinations are tested, including using basic SH channel weightings for the entire frequency spectrum as well as dual-band decoding. The combination of all three tested pre-processing techniques with dual-band decoding is shown to produce the best overall results, and the observed improvements are greater than when using a single pre-processing technique.

## 8.1 Restatement of Hypothesis

The hypothesis originally stated in Section 1.1, which has informed the work presented in this thesis, is now restated as follows:

> *The use of head-related transfer function pre-processing techniques can improve the high frequency reproduction of binaural Ambisonic rendering.*

The research presented in this thesis confirms this hypothesis. The HRTF pre-processing techniques developed and tested have shown how the high frequency reproduction of binaural Ambisonic rendering can be improved, with no alterations of the Ambisonic decoding process. This has been shown in different ways: from spectral reproduction improvements when using Ambisonic Diffuse-Field Equalisation, Ambisonic Directional Bias Equalisation or Ambisonic Time-Alignment, to the ILD reproduction improvements when using Ambisonic Interaural Level Difference

Optimisation. More than one technique can be used at once, which produces more significant improvements to the Ambisonic rendering accuracy. The thorough numerical and perceptual evaluations have shown that these techniques are robust and produce positive results, regardless of Ambisonic order, loudspeaker configuration or HRTF dataset.

## 8.2 Future Work

Throughout this thesis, from the work undertaken, a number of suggestions for future research have been identified. These are as follows:

### Ambisonic Diffuse-Field Equalisation for Loudspeaker Reproduction

In Chapter 4, the diffuse-field response of the $L = 50$ Lebedev loudspeaker configuration using $M = 5$ is calculated for the datasets of all 18 human subjects in the SADIE II database (Armstrong et al., 2018a). The plots are presented in Figure 4.7, which show a clear trend in the diffuse-field responses. This trend is also loosely observed in Figure 4.5e as a wideband notch around the region of 4 kHz, which suggests an Ambisonic diffuse-field response is dependent on the loudspeaker configuration and Ambisonic order. This poses a potential avenue for future work, whereby a generalised Ambisonic diffuse-field equalisation filter could be generated for a specified loudspeaker arrangement and Ambisonic order. This could produce an improvement in overall spectral reproduction for the majority of listeners, without the need to calculate the diffuse-field equalisation filters separately for each HRTF dataset. An additional application of generalised Ambisonic diffuse-field equalisation filters for a specified loudspeaker arrangement and Ambisonic order would be to apply it to loudspeaker reproduction (as opposed to binaural reproduction). However, an important consideration in this would be to measure the effect on the generalised diffuse-field response when the head is rotated.

### Signal-Dependent Ambisonic Directional Bias Equalisation

The Ambisonic Directional Bias Equalisation technique presented in Chapter 5 requires a specified direction and amount of bias. A future development of the method could employ a source direction estimation technique such as Directional Audio Coding (Pulkki, 2007) to produce an estimated source direction and confidence level (which could be calculated from the direct-to-reverberant ratio, for example). If the directional bias equalisation filters for a series of bias directions and bias amounts were pre-computed, the estimated source direction could inform the bias direction and the confidence level could inform the bias amount, such that diffuse sounds or multiple sources could have lower bias amounts and non-diffuse, focussed single sources could have a higher bias amount. This method would make the real-time rendering process more complex however, so methods such as simplifying of the filters would be necessary.

### Frequency-Dependent Ambisonic Interaural Level Difference Optimisation

In Chapter 6, the presented method of Ambisonic Interaural Level Difference Optimisation (AIO) augments the left and right signals of the virtual loudspeaker HRTFs using one value of gain for all frequencies above $f_{\text{alias}}$. However, the change in interaural level difference plots presented in Figure 6.9 show that, in general, AIO produces the greatest improvement in ILD reproduction in the region around 4 kHz. Therefore, a future development could investigate the possibility of generating frequency-dependent ILD optimisation gains, in order to maximise the improvement in ILD reproduction at different frequencies.

### Ambisonic Interaural Time Difference Optimisation

As with ILDs, Ambisonic reproduction of ITDs is also inaccurate at low orders of Ambisonics, as alluded to in the literature for $M = 1$ (Kearney, 2010, p. 87). This was

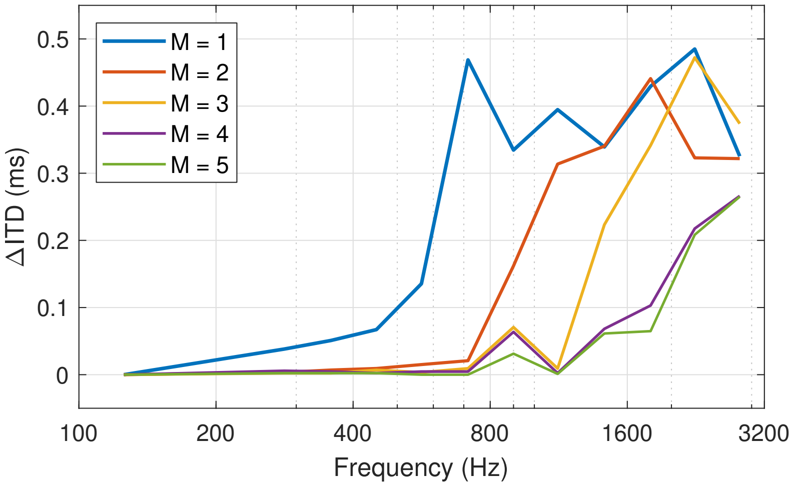


FIGURE 8.1: Mean values of ΔITD between HRTFs and binaural Ambisonic rendering at $\{M = 1, M = 2, ..., M = 5\}$ for twelve frequency bands on the horizontal plane.

demonstrated for $\{M = 1, M = 2, ..., M = 5\}$ in Figure A.1 for all locations on the sphere. To investigate the accuracy of Ambisonic ITD reproduction over frequency, binaural Ambisonic renders have been created for $\{M = 1, M = 2, ..., M = 5\}$ using the corresponding Lebedev loudspeaker configurations as presented in Figure 3.4 and Bernschütz Neumann KU 100 HRTFs (Bernschütz, 2013) for locations on the horizontal plane between $-180° < \theta < +180°$ at $\phi = 0°$ in 20° increments. This was using basic SH channel weightings for the entire frequency spectrum to avoid any possible temporal effects from the dual-band crossover. The ITD was then calculated for original HRTFs and binaural Ambisonic renders across twelve frequency bands, corresponding to third-octave bands between 250 Hz and 3150 Hz (except the first band which is between 1 Hz and 250 Hz), achieved by bandpass filtering the HRTFs prior to the ITD calculation, which used the cross-correlation method in (2.15). Figure 8.1 presents the mean value of ΔITD across all tested locations for each frequency band.

It is clear that the frequency at which ITD becomes inaccurate is dependent on Ambisonic order, as expected. For $M = 1$, $M = 2$ and $M = 3$, ITD reproduction becomes significantly inaccurate at approximately 700 Hz, 1100 Hz and 1400 Hz, respectively. As ITD is widely considered to be less perceptually relevant at frequencies above 1.5 kHz (Minnaar et al., 2000), Ambisonic ITD reproduction is therefore likely to be perceived as inaccurate for $M \leq 3$. A possible future HRTF pre-processing technique could look at using the principles of Ambisonic Interaural Level Difference Optimisation, as detailed in Chapter 6), and applying them to ITD instead. By time-shifting a frequency portion of the virtual loudspeaker HRTFs accordingly, it may be possible to improve the Ambisonic rendered ITD.

## 8.3 Final Remarks

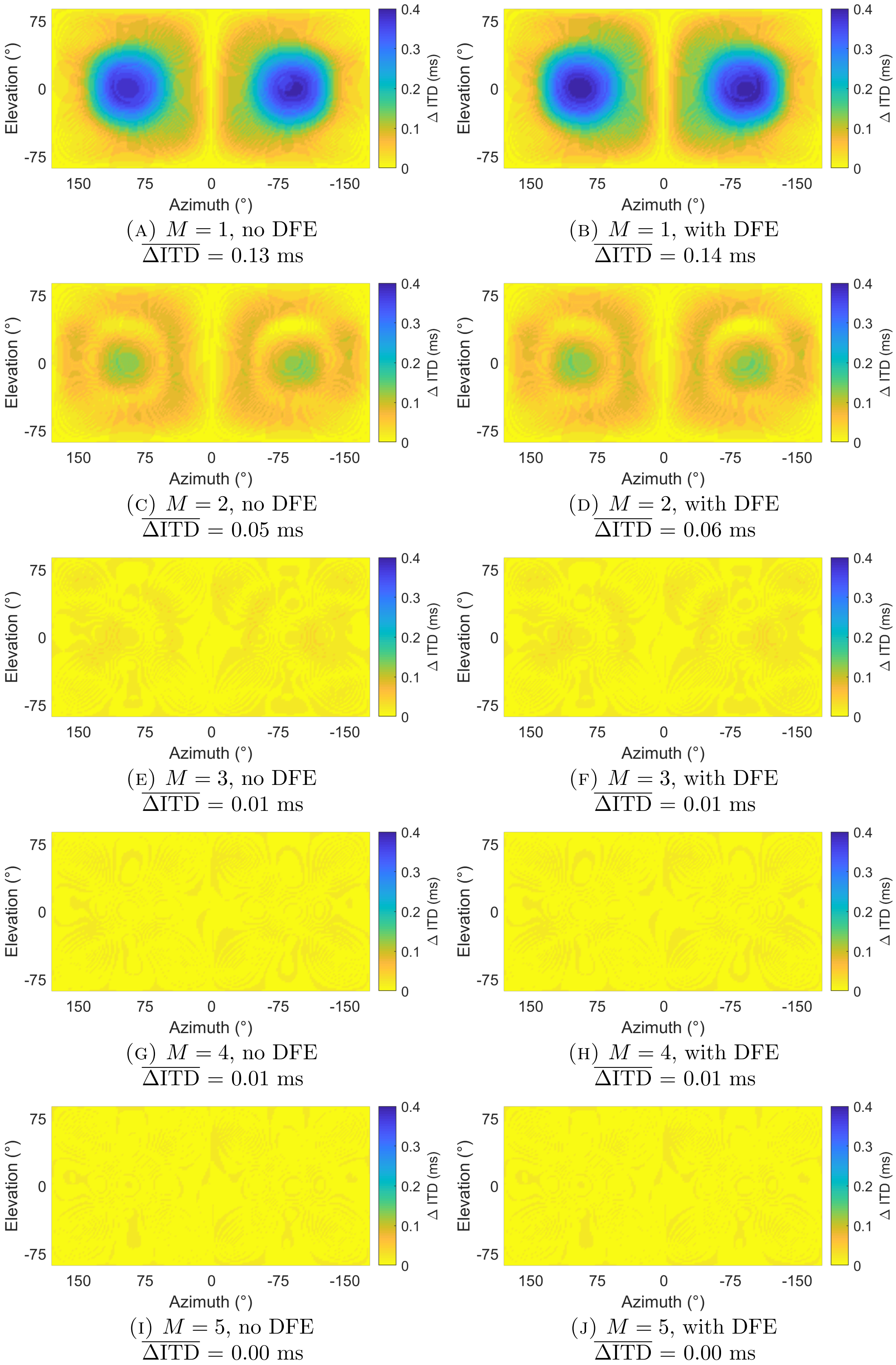
The research presented in this thesis has investigated low-order binaural Ambisonic rendering, a technology that facilitates the reproduction of spatial sound in any direction to a great degree of accuracy, using a minimal number of head-related transfer function measurements. The head-related transfer function pre-processing techniques developed and presented in this thesis have addressed some of the limitations of Ambisonic technology, such as the inaccurate spectral reproduction and interaural cue reproduction due to spatial aliasing. Though the techniques presented in this thesis do not solve the problem in its entirety, they do offer progress. The fact that these techniques can be implemented offline means they require no changes to the real-time rendering process and are therefore easy to implement.

It is hoped that the research presented in this thesis will offer a basis upon which future investigations can build, as with more realistic binaural audio, the resulting experiences can be more immersive. When used in entertainment, this can help to create an adventure the user gets lost in, and when used in medical and educational training, this can offer a more authentic simulation to better train professionals for the corresponding real world tasks.

# Appendix A

# Supplementary Plots for Ambisonic Diffuse-Field Equalisation

This appendix presents supplementary plots for Chapter 4. All computation was carried out offline in Matlab version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4. Figure A.1 presents ΔITD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without DFE. Figure A.2 presents ΔILD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without DFE. Figure A.3 presents the estimated $\theta$ values of binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \theta < +90°$ and $\phi = 0°$, both with and without DFE. Figure A.4 presents the estimated sagittal plane localisation plots of binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \phi < +90°$ and $\theta = 0°$, both with and without DFE. Figure A.5 presents the diffuse-field responses, inverse filters and resulting equalised frequency responses of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations. Figure A.6 presents PSD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without DFE, using T-design loudspeaker configurations. Figure A.7 presents the diffuse-field responses, inverse filters and resulting equalised frequency responses of Ambisonic orders $\{M = 1, M = 2, M = 3, M = 5\}$ using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a). Figure A.8 presents PSD values for $\{M = 1, M = 2, M = 3, M = 5\}$ for every measurement location on the sphere, both with and without DFE, using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).



(A) $M = 1$, no DFE
$\overline{\Delta\text{ITD}} = 0.13$ ms

(B) $M = 1$, with DFE
$\overline{\Delta\text{ITD}} = 0.14$ ms

(C) $M = 2$, no DFE
$\overline{\Delta\text{ITD}} = 0.05$ ms

(D) $M = 2$, with DFE
$\overline{\Delta\text{ITD}} = 0.06$ ms

(E) $M = 3$, no DFE
$\overline{\Delta\text{ITD}} = 0.01$ ms

(F) $M = 3$, with DFE
$\overline{\Delta\text{ITD}} = 0.01$ ms

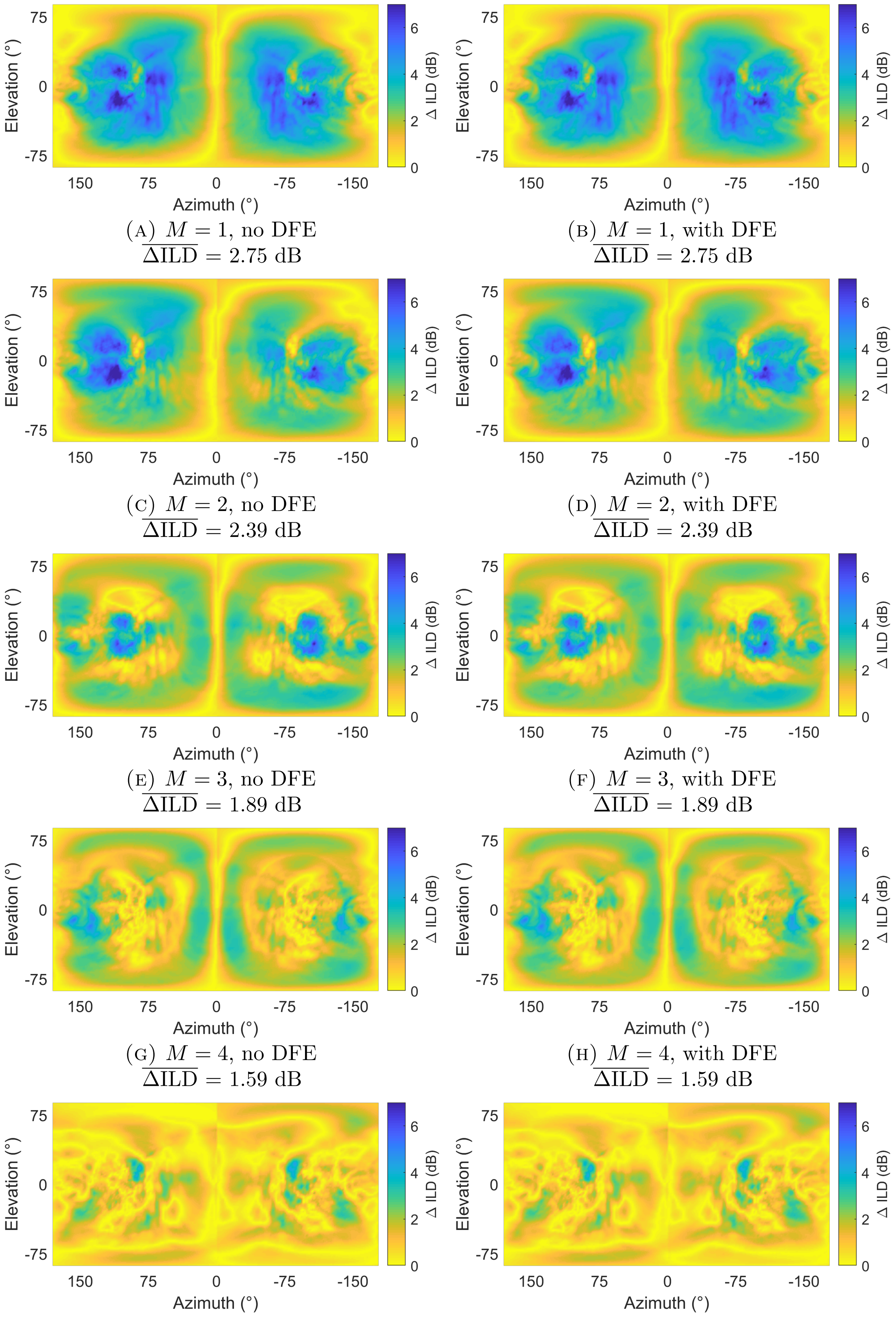
(G) $M = 4$, no DFE
$\overline{\Delta\text{ITD}} = 0.01$ ms

(H) $M = 4$, with DFE
$\overline{\Delta\text{ITD}} = 0.01$ ms

(I) $M = 5$, no DFE
$\overline{\Delta\text{ITD}} = 0.00$ ms

(J) $M = 5$, with DFE
$\overline{\Delta\text{ITD}} = 0.00$ ms

FIGURE A.1: ΔITD between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$, no DFE
$\overline{\Delta\text{ILD}} = 2.75$ dB

(B) $M = 1$, with DFE
$\overline{\Delta\text{ILD}} = 2.75$ dB

(C) $M = 2$, no DFE
$\overline{\Delta\text{ILD}} = 2.39$ dB

(D) $M = 2$, with DFE
$\overline{\Delta\text{ILD}} = 2.39$ dB

(E) $M = 3$, no DFE
$\overline{\Delta\text{ILD}} = 1.89$ dB

(F) $M = 3$, with DFE
$\overline{\Delta\text{ILD}} = 1.89$ dB

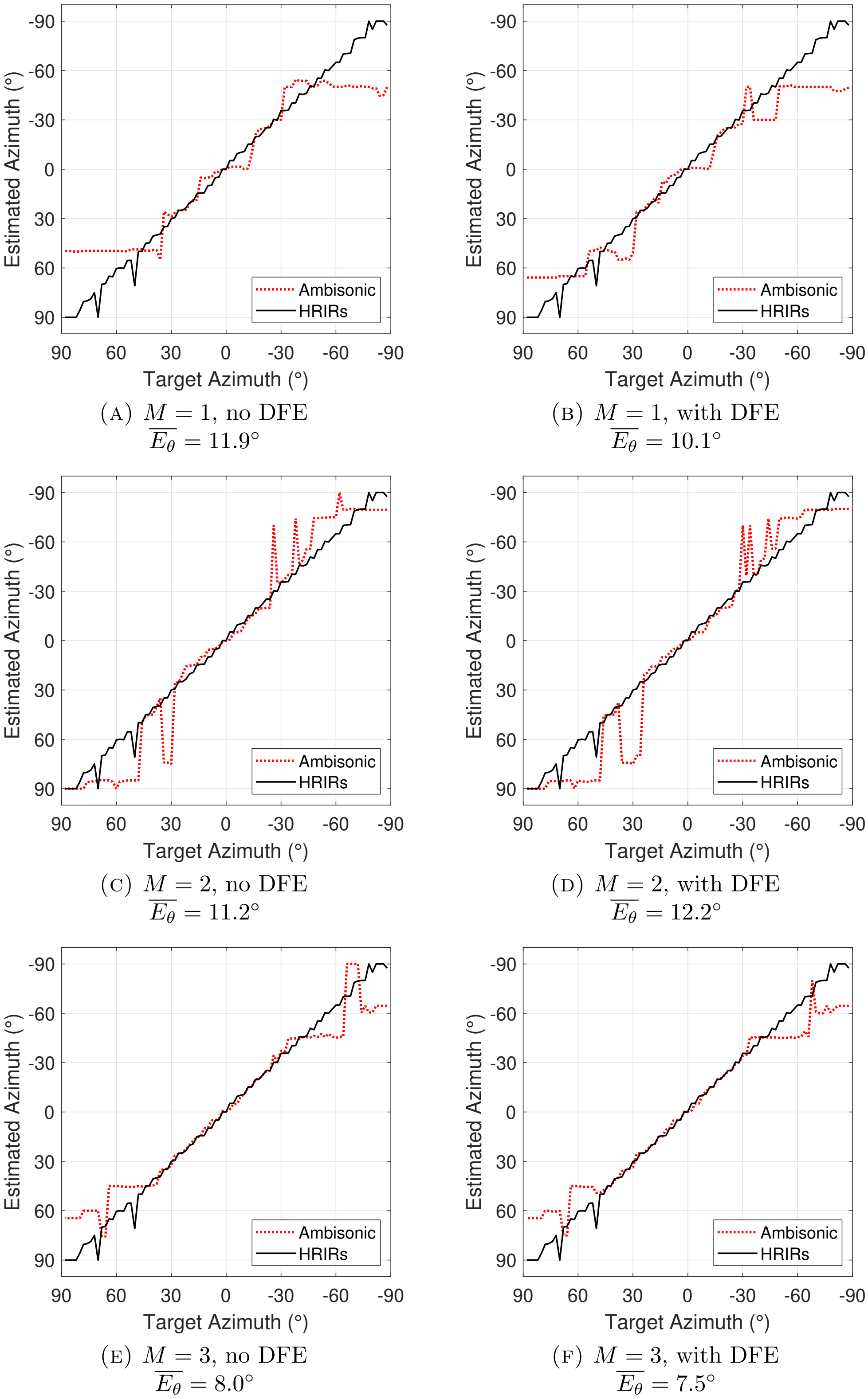
(G) $M = 4$, no DFE
$\overline{\Delta\text{ILD}} = 1.59$ dB

(H) $M = 4$, with DFE
$\overline{\Delta\text{ILD}} = 1.59$ dB

(I) $M = 5$, no DFE
$\overline{\Delta\text{ILD}} = 0.92$ dB

(J) $M = 5$, with DFE
$\overline{\Delta\text{ILD}} = 0.91$ dB

FIGURE A.2: ΔILD between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$, no DFE
$\overline{E_\theta} = 11.9°$

(B) $M = 1$, with DFE
$\overline{E_\theta} = 10.1°$

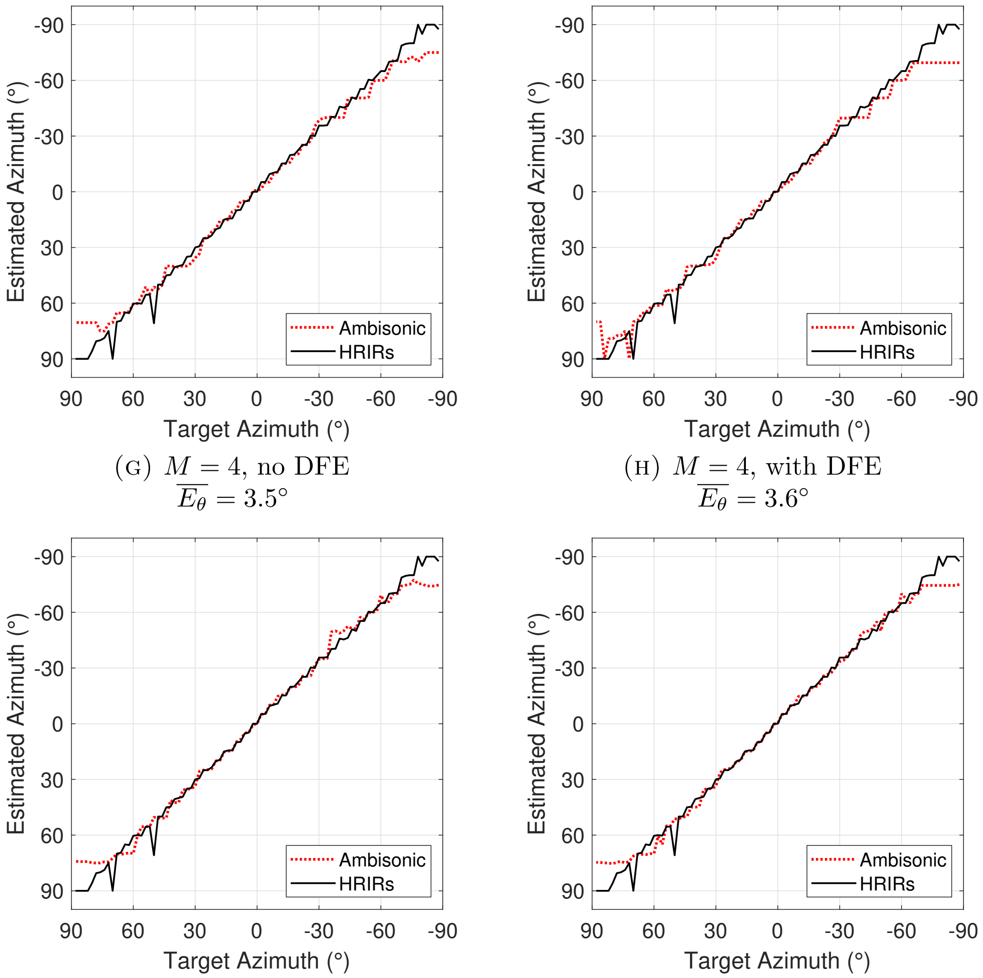
(C) $M = 2$, no DFE
$\overline{E_\theta} = 11.2°$

(D) $M = 2$, with DFE
$\overline{E_\theta} = 12.2°$

(E) $M = 3$, no DFE
$\overline{E_\theta} = 8.0°$

(F) $M = 3$, with DFE
$\overline{E_\theta} = 7.5°$

FIGURE A.3: *Cont.*



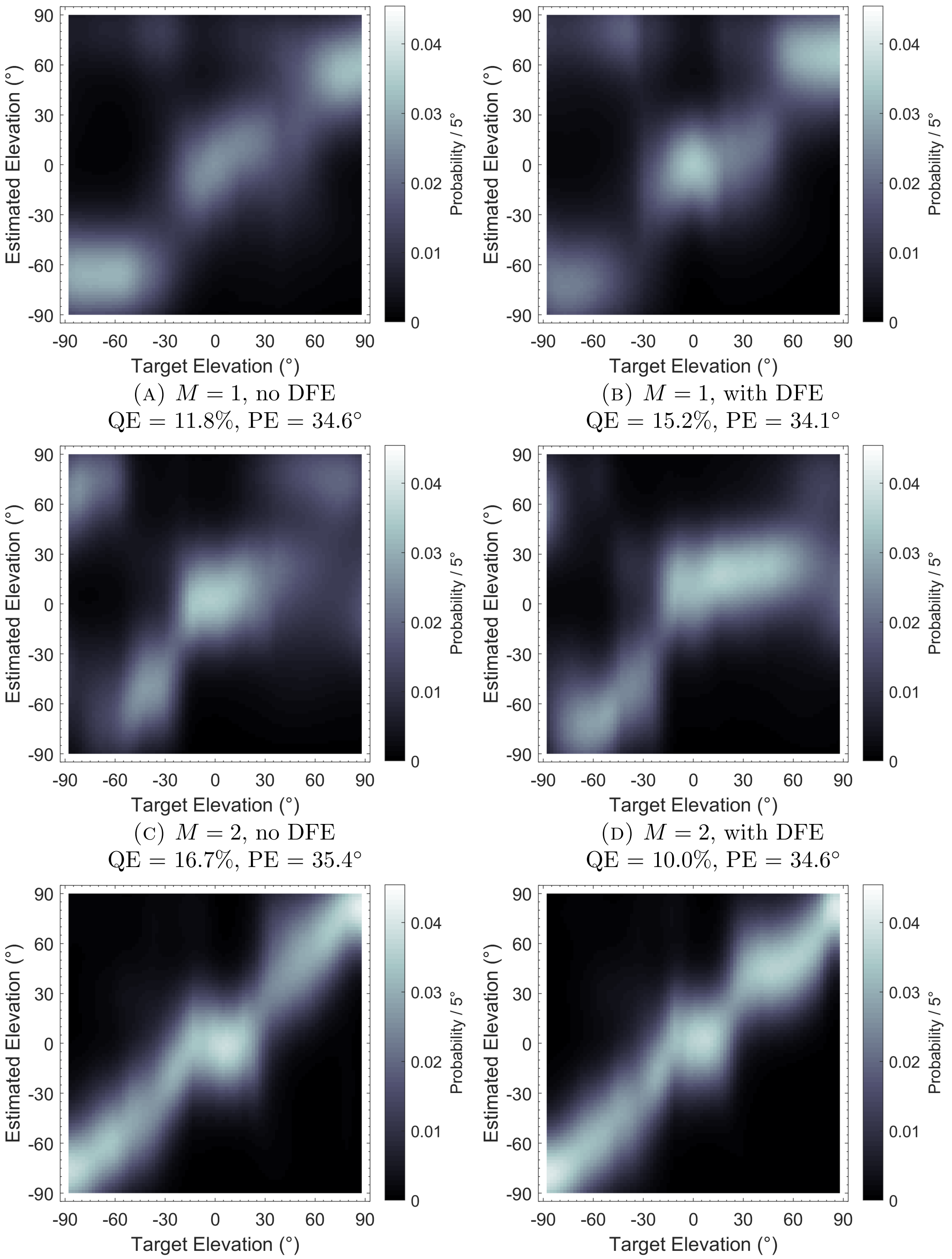
(G) $M = 4$, no DFE
$\overline{E_\theta} = 3.5°$

(H) $M = 4$, with DFE
$\overline{E_\theta} = 3.6°$

(I) $M = 5$, no DFE
$\overline{E_\theta} = 3.9°$

(J) $M = 5$, with DFE
$\overline{E_\theta} = 3.8°$

FIGURE A.3: Horizontal localisation model plots of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, with and without Ambisonic DFE.



(A) $M = 1$, no DFE
QE = 11.8%, PE = 34.6°

(B) $M = 1$, with DFE
QE = 15.2%, PE = 34.1°

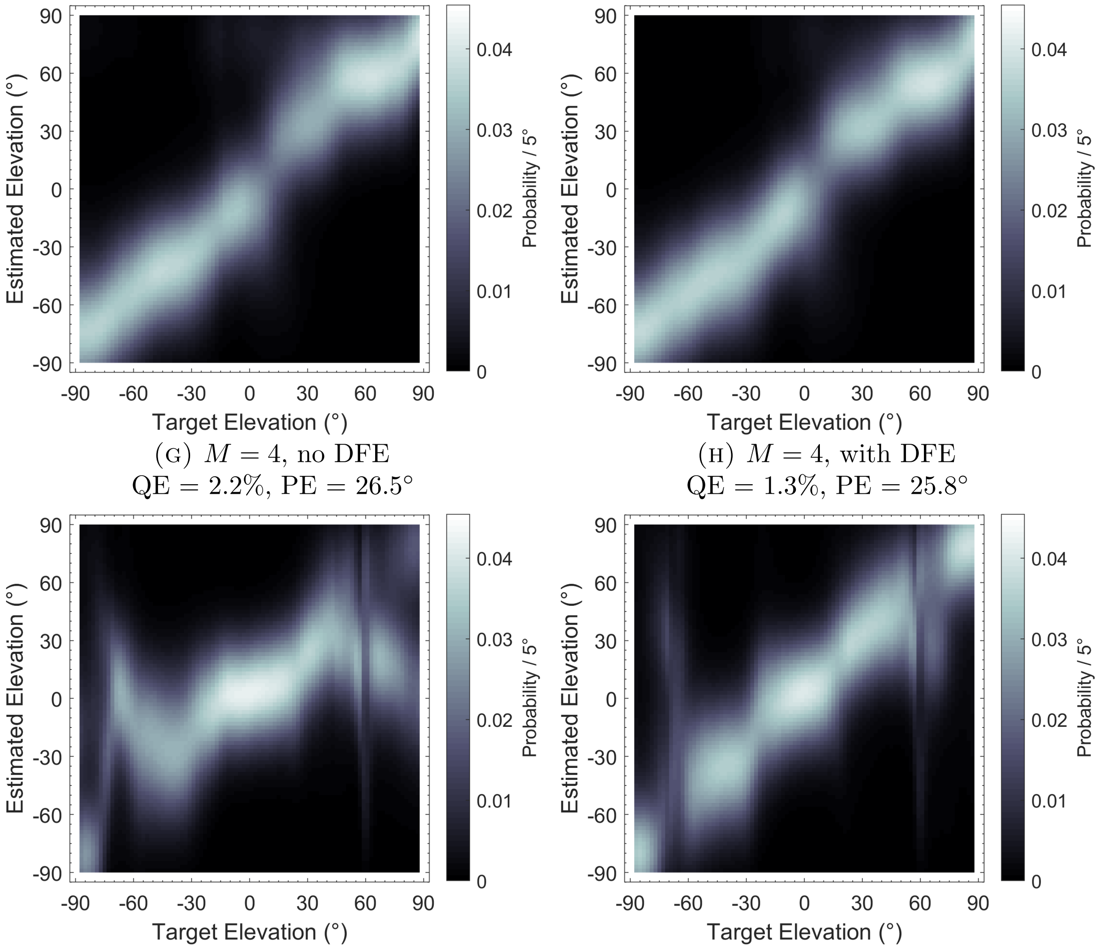
(C) $M = 2$, no DFE
QE = 16.7%, PE = 35.4°

(D) $M = 2$, with DFE
QE = 10.0%, PE = 34.6°

(E) $M = 3$, no DFE
QE = 2.8%, PE = 27.2°

(F) $M = 3$, with DFE
QE = 2.3%, PE = 26.3°

FIGURE A.4: *Cont.*



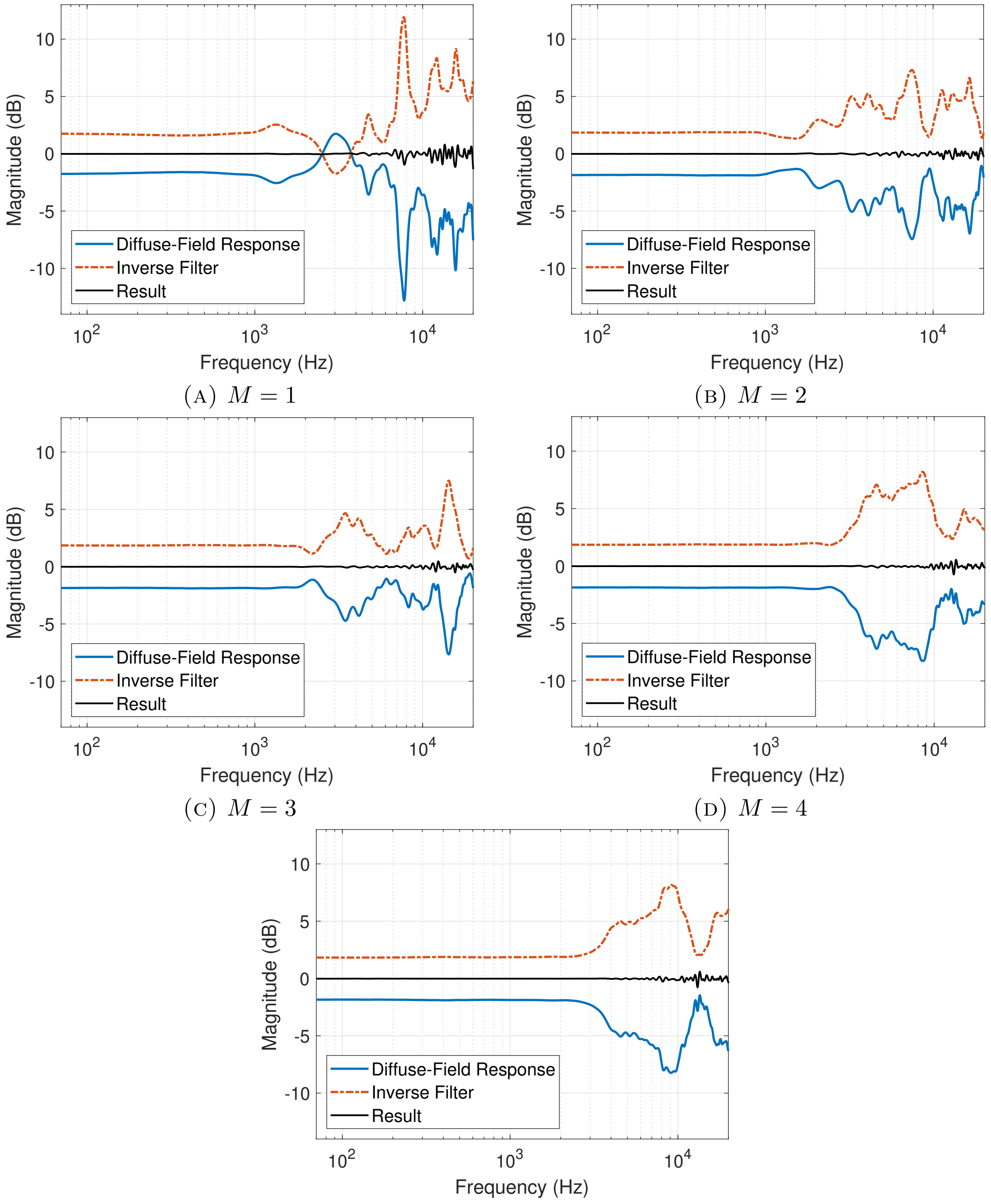
(G) $M = 4$, no DFE
QE = 2.2%, PE = 26.5°

(H) $M = 4$, with DFE
QE = 1.3%, PE = 25.8°

(I) $M = 5$, no DFE
QE = 6.0%, PE = 34.7°

(J) $M = 5$, with DFE
QE = 4.9%, PE = 27.9°

FIGURE A.4: Sagittal plane localisation model plots of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, with and without Ambisonic DFE.



(A) $M = 1$

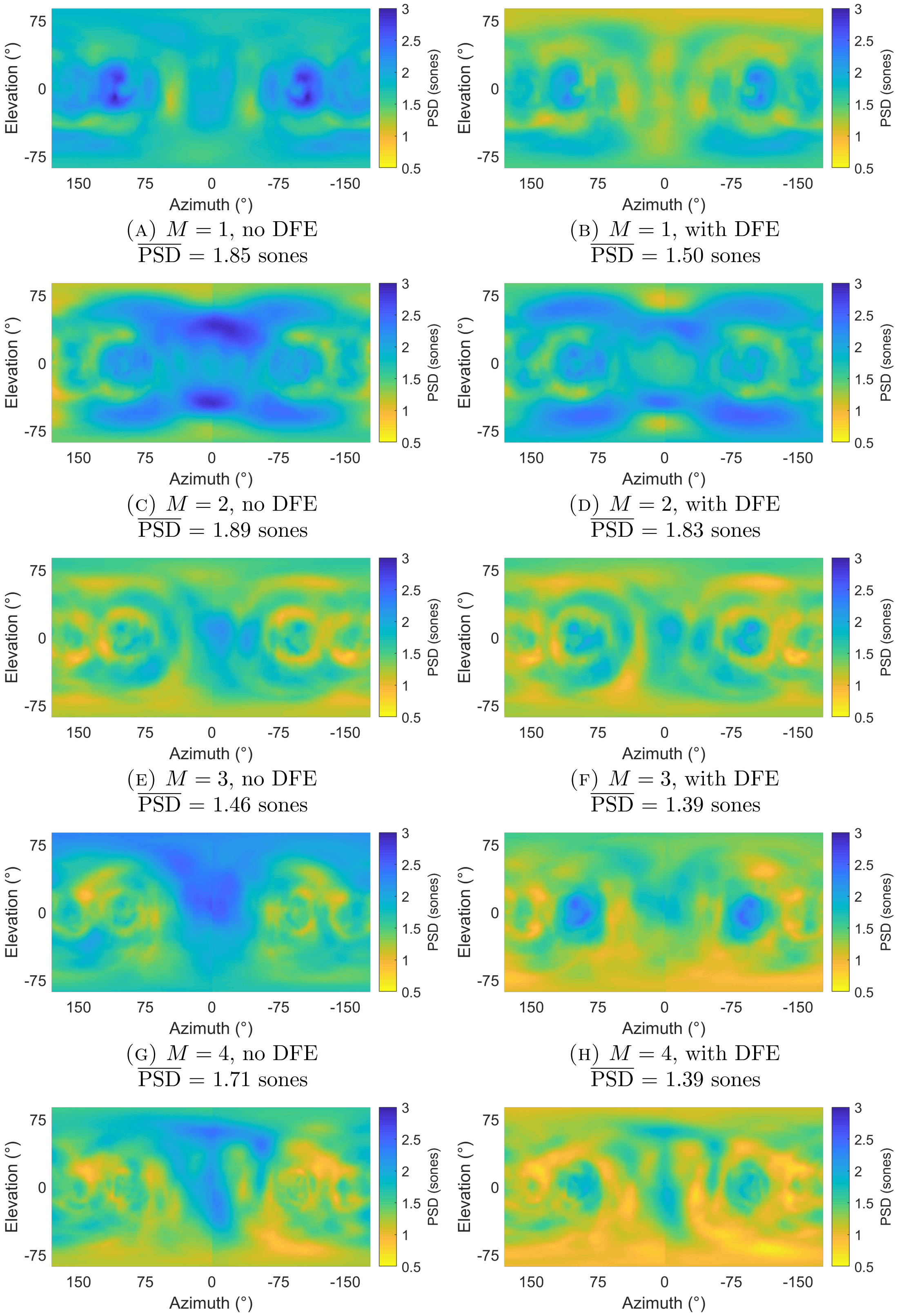
(B) $M = 2$

(C) $M = 3$

(D) $M = 4$

(E) $M = 5$

FIGURE A.5: Diffuse-field response, inverse filters and resulting responses of the T-design loudspeaker configurations for $\{M = 1, M = 2, ..., M = 5\}$ (left ear).



(A) $M = 1$, no DFE
$\overline{\text{PSD}} = 1.85$ sones

(B) $M = 1$, with DFE
$\overline{\text{PSD}} = 1.50$ sones

(C) $M = 2$, no DFE
$\overline{\text{PSD}} = 1.89$ sones

(D) $M = 2$, with DFE
$\overline{\text{PSD}} = 1.83$ sones

(E) $M = 3$, no DFE
$\overline{\text{PSD}} = 1.46$ sones

(F) $M = 3$, with DFE
$\overline{\text{PSD}} = 1.39$ sones

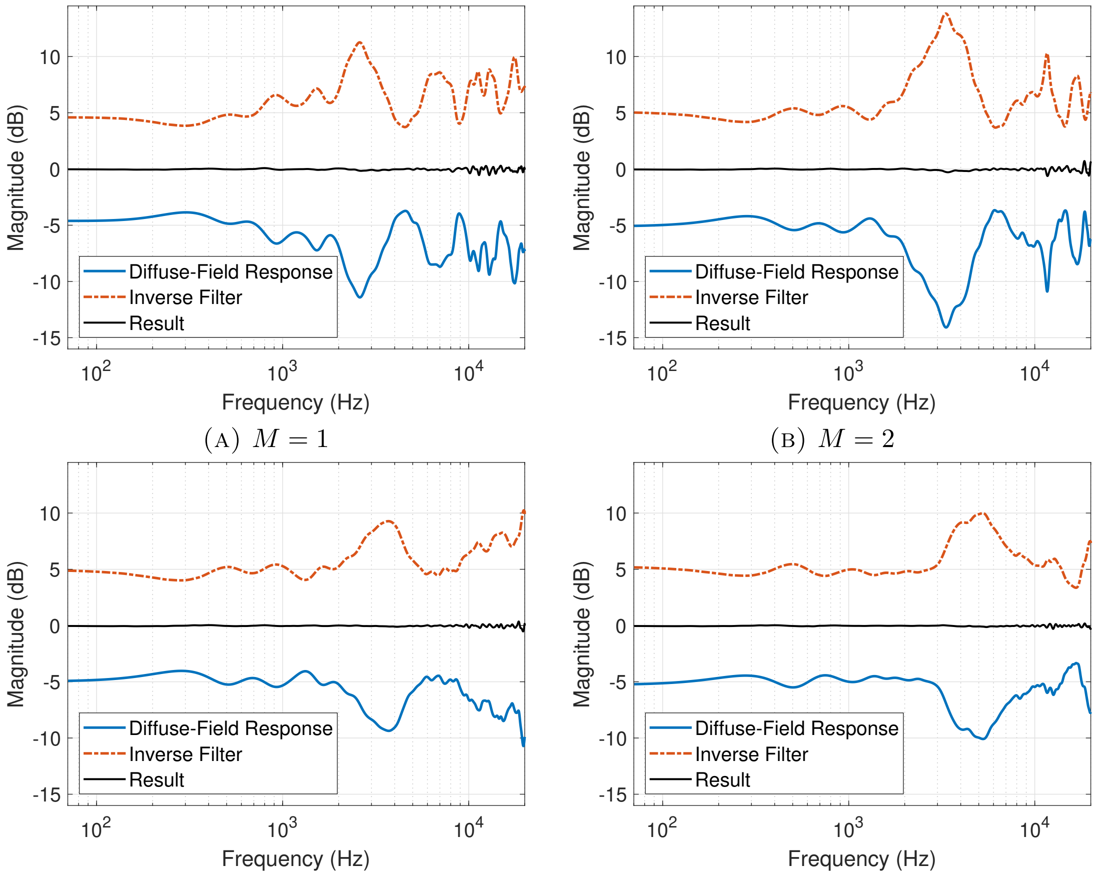
(G) $M = 4$, no DFE
$\overline{\text{PSD}} = 1.71$ sones

(H) $M = 4$, with DFE
$\overline{\text{PSD}} = 1.39$ sones

(I) $M = 5$, no DFE
$\overline{\text{PSD}} = 1.48$ sones

(J) $M = 5$, with DFE
$\overline{\text{PSD}} = 1.23$ sones

FIGURE A.6: PSD between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, ..., M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using T-design loudspeaker configurations.



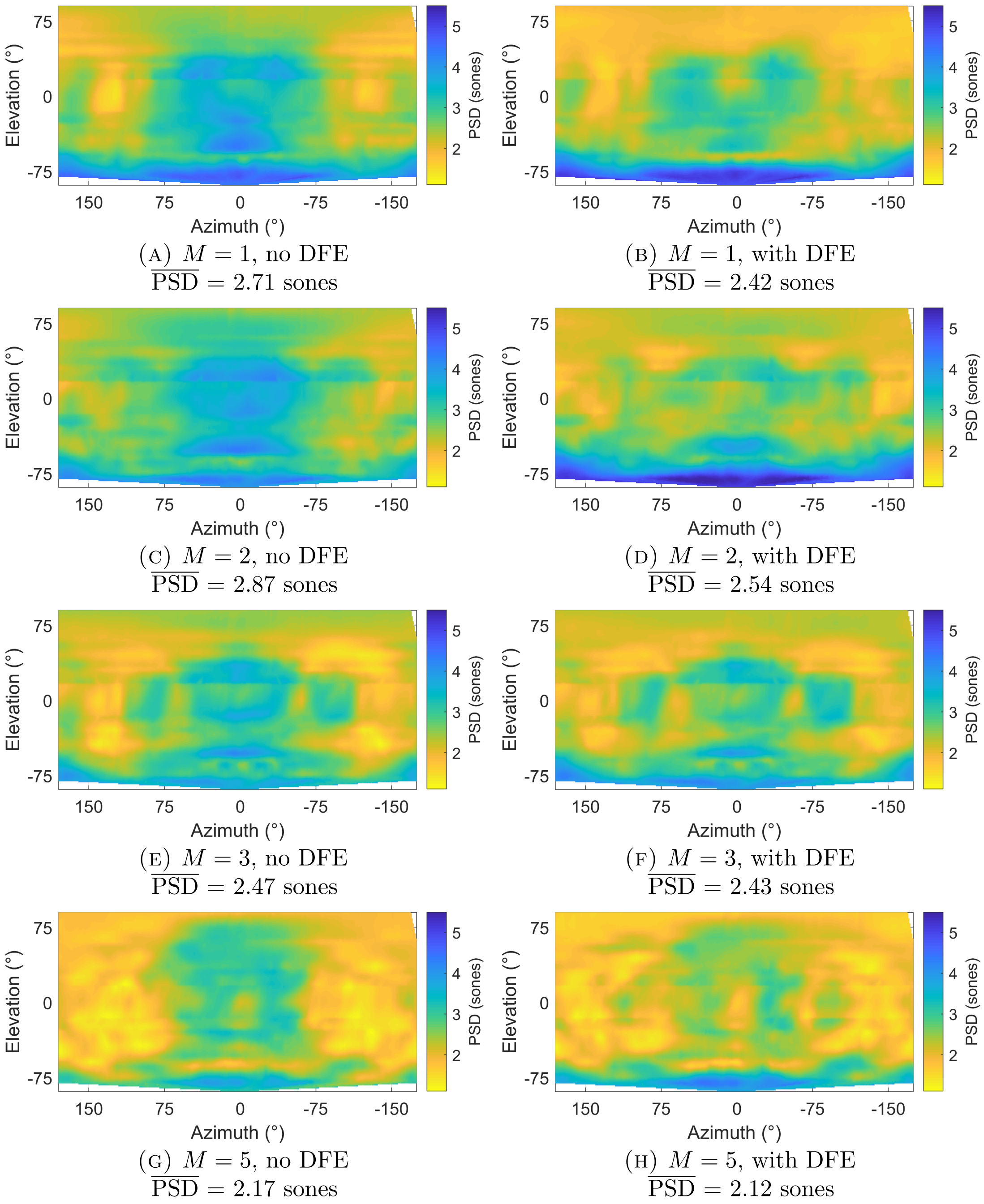
(A) $M = 1$

(B) $M = 2$

(C) $M = 3$

(D) $M = 5$

FIGURE A.7: Diffuse-field response, inverse filters and resulting responses of the Lebedev configurations for $\{M = 1, M = 2, M = 3, M = 5\}$, using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a) (left ear).



(A) $M = 1$, no DFE
$\overline{\text{PSD}} = 2.71$ sones

(B) $M = 1$, with DFE
$\overline{\text{PSD}} = 2.42$ sones

(C) $M = 2$, no DFE
$\overline{\text{PSD}} = 2.87$ sones

(D) $M = 2$, with DFE
$\overline{\text{PSD}} = 2.54$ sones

(E) $M = 3$, no DFE
$\overline{\text{PSD}} = 2.47$ sones

(F) $M = 3$, with DFE
$\overline{\text{PSD}} = 2.43$ sones

(G) $M = 5$, no DFE
$\overline{\text{PSD}} = 2.17$ sones

(H) $M = 5$, with DFE
$\overline{\text{PSD}} = 2.12$ sones

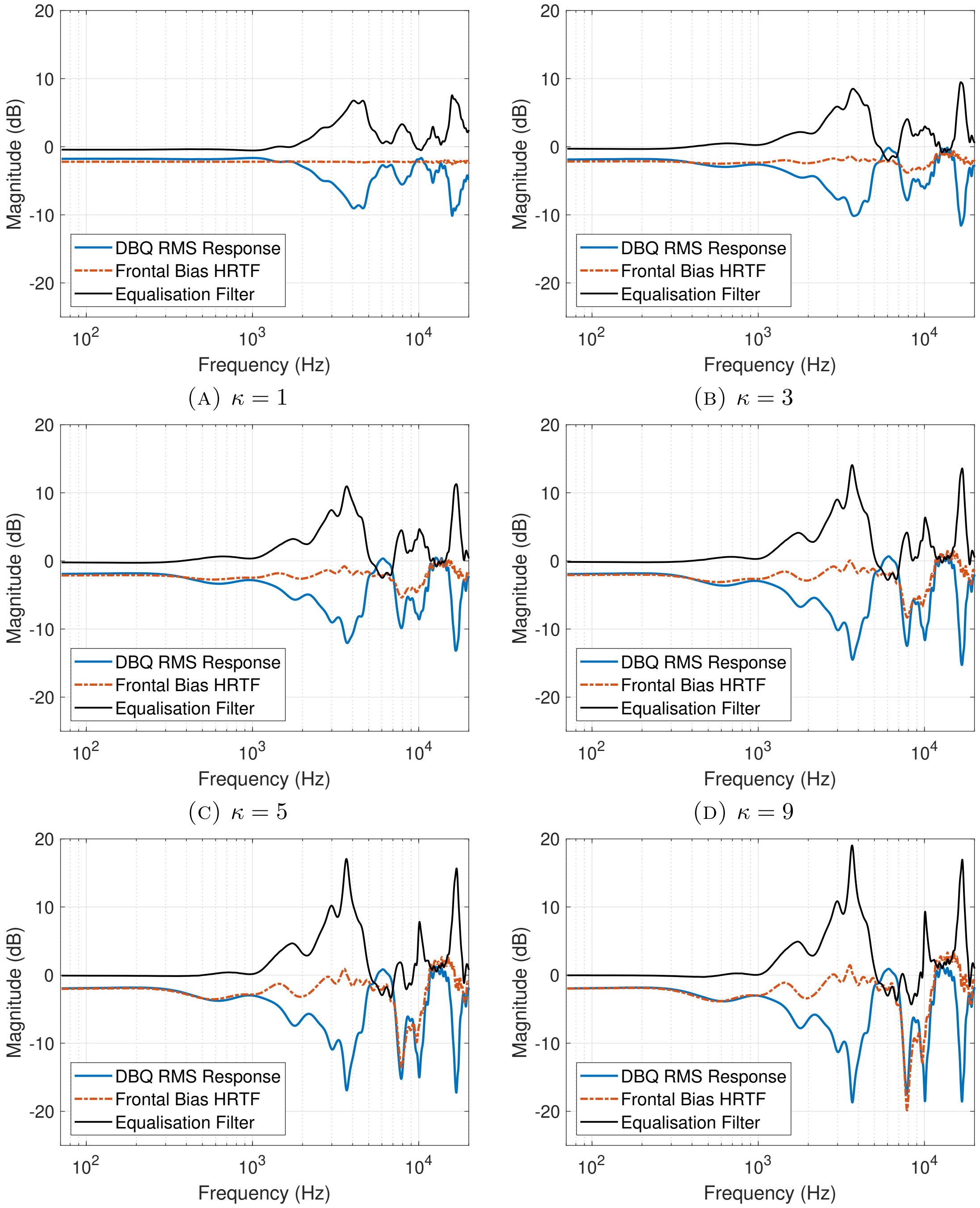
FIGURE A.8: PSD between HRTFs and binaural Ambisonic rendering with and without DFE, for $\{M = 1, M = 2, M = 3, M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).

# Appendix B

# Supplementary Plots for Ambisonic Directional Bias Equalisation

This appendix presents supplementary plots for Chapter 5. The direction of bias in all these plots is ($\theta = 0°, \phi = 0°$). All computation was carried out offline in Matlab version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4. Figures {B.1, B.2, B.3 and B.4} present the DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, for Ambisonic orders $\{M = 2, M = 3, ..., M = 5\}$, respectively. Figures {B.5, B.6, B.7 and B.8} present the PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, for Ambisonic orders $\{M = 2, M = 3, ..., M = 5\}$, respectively. Figure B.9 presents ΔITD values between HRTFs and binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere.

Figure B.10 presents ΔILD values between HRTFs and binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere. Figure B.11 presents the estimated $\theta$ values of binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ between $-90^\circ < \theta < +90^\circ$ and $\phi = 0^\circ$. Figure B.12 presents the estimated sagittal plane localisation plots of binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ between $-90^\circ < \phi < +90^\circ$ and $\theta = 0^\circ$. Figure B.13 presents the DBQ RMS response, directional HRTF response and resulting DBE filters at $\kappa = 33$ of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$ using T-design loudspeaker configurations. Figure B.14 presents PSD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without DBE at $\kappa = 33$, using T-design loudspeaker configurations. Figure B.15 presents the DBQ RMS response, directional HRTF response and resulting DBE filters at $\kappa = 33$ of Ambisonic orders $\{M = 1, M = 2, M = 3, M = 5\}$ using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a). Figure B.16 presents PSD values for $\{M = 1, M = 2, M = 3, M = 5\}$ for every measurement location on the sphere, both with and without DBE at $\kappa = 33$, using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).



(A) $\kappa = 1$

(B) $\kappa = 3$

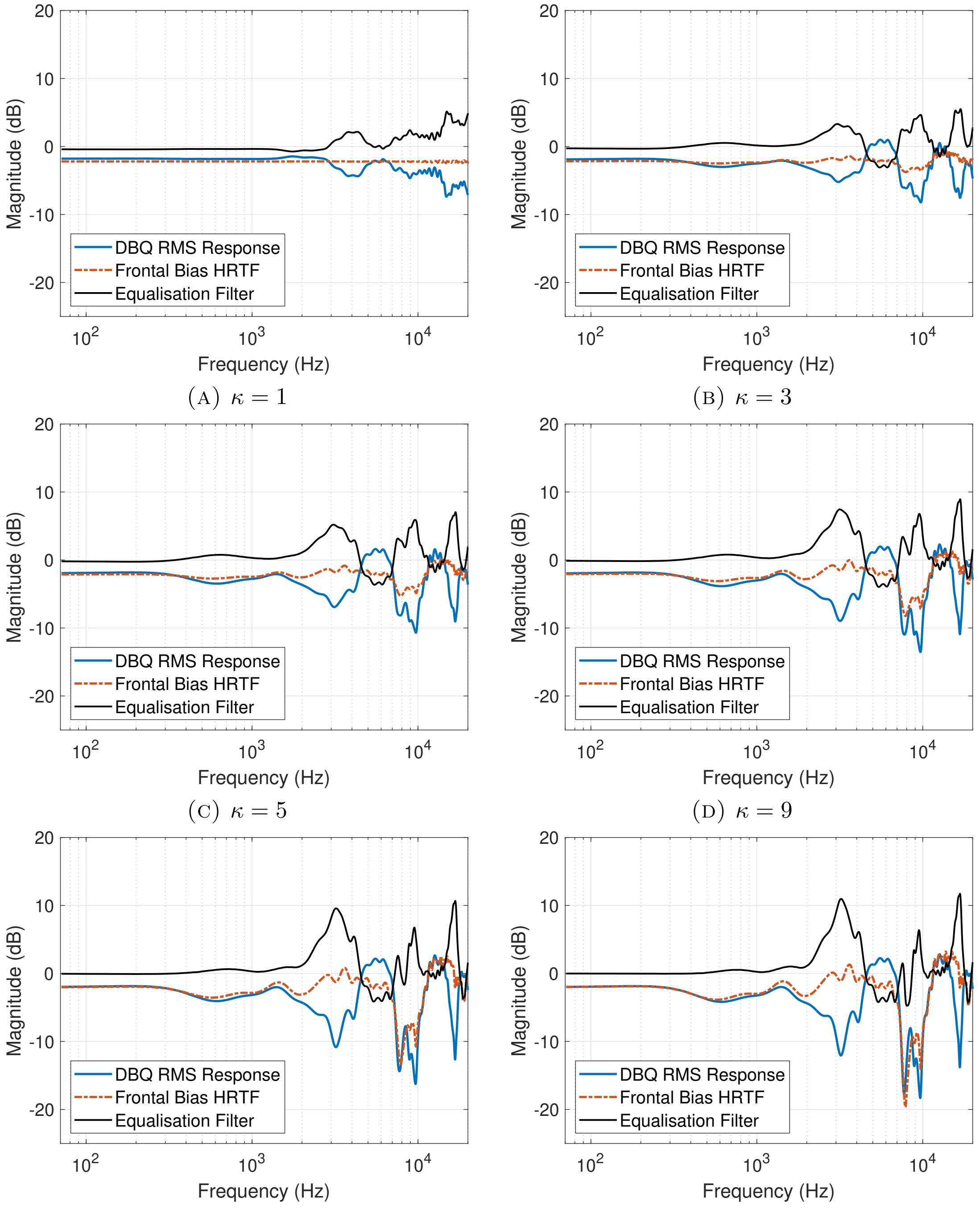
(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

FIGURE B.1: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, $M = 2$ (left ear).



(A) $\kappa = 1$

(B) $\kappa = 3$

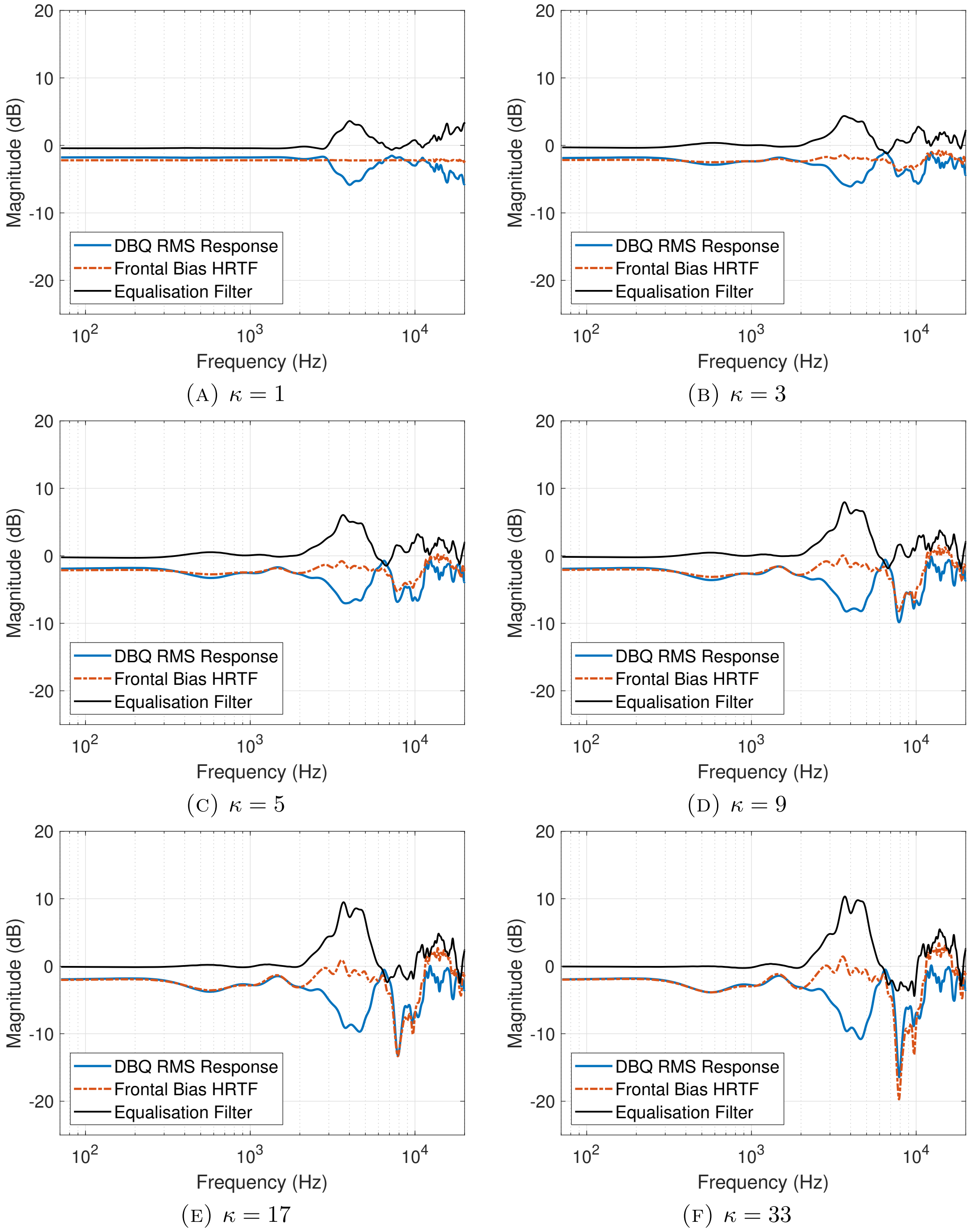
(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

Figure B.2: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, $M = 3$ (left ear).



(A) $\kappa = 1$

(B) $\kappa = 3$

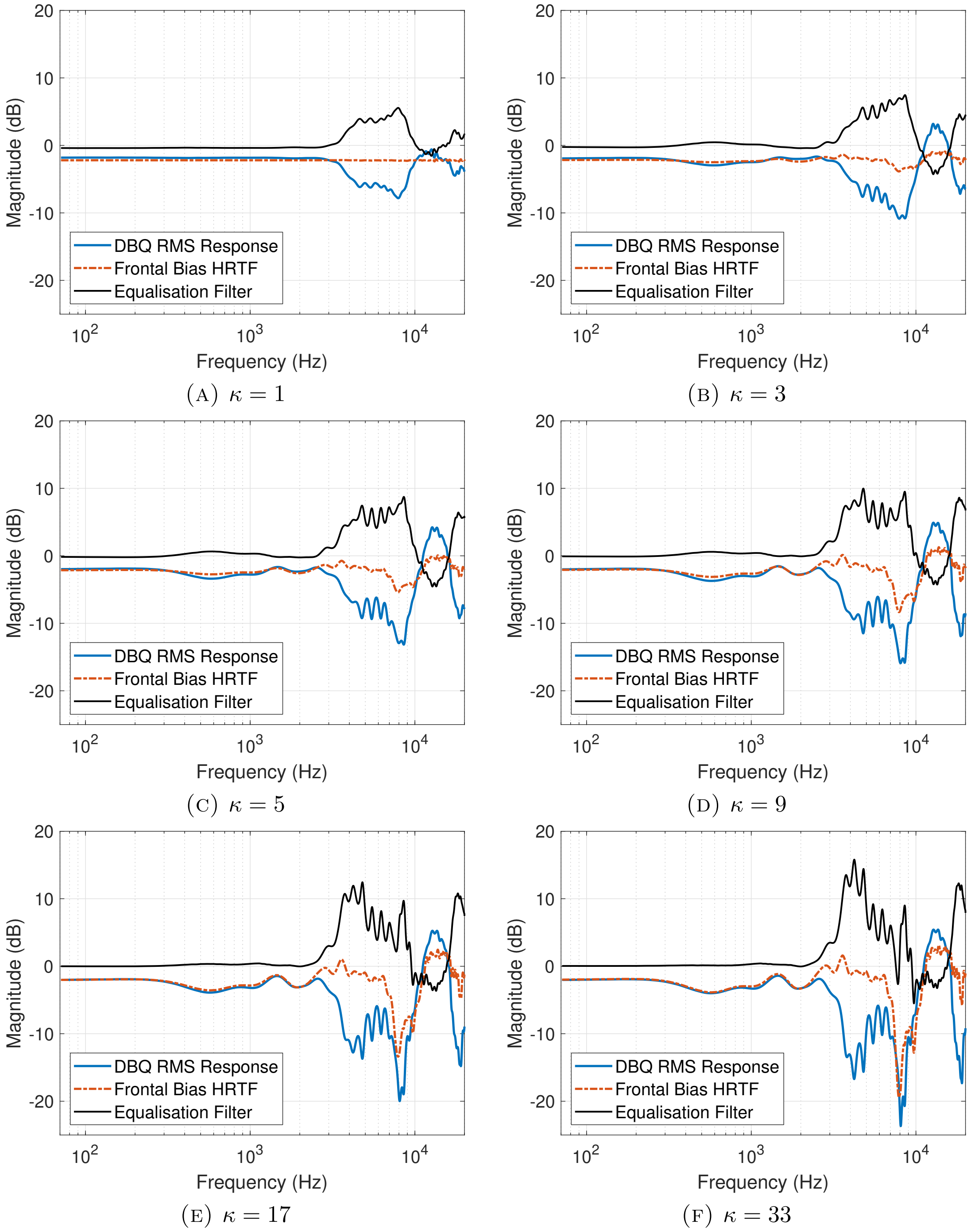
(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

FIGURE B.3: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, $M = 4$ (left ear).



(A) $\kappa = 1$

(B) $\kappa = 3$

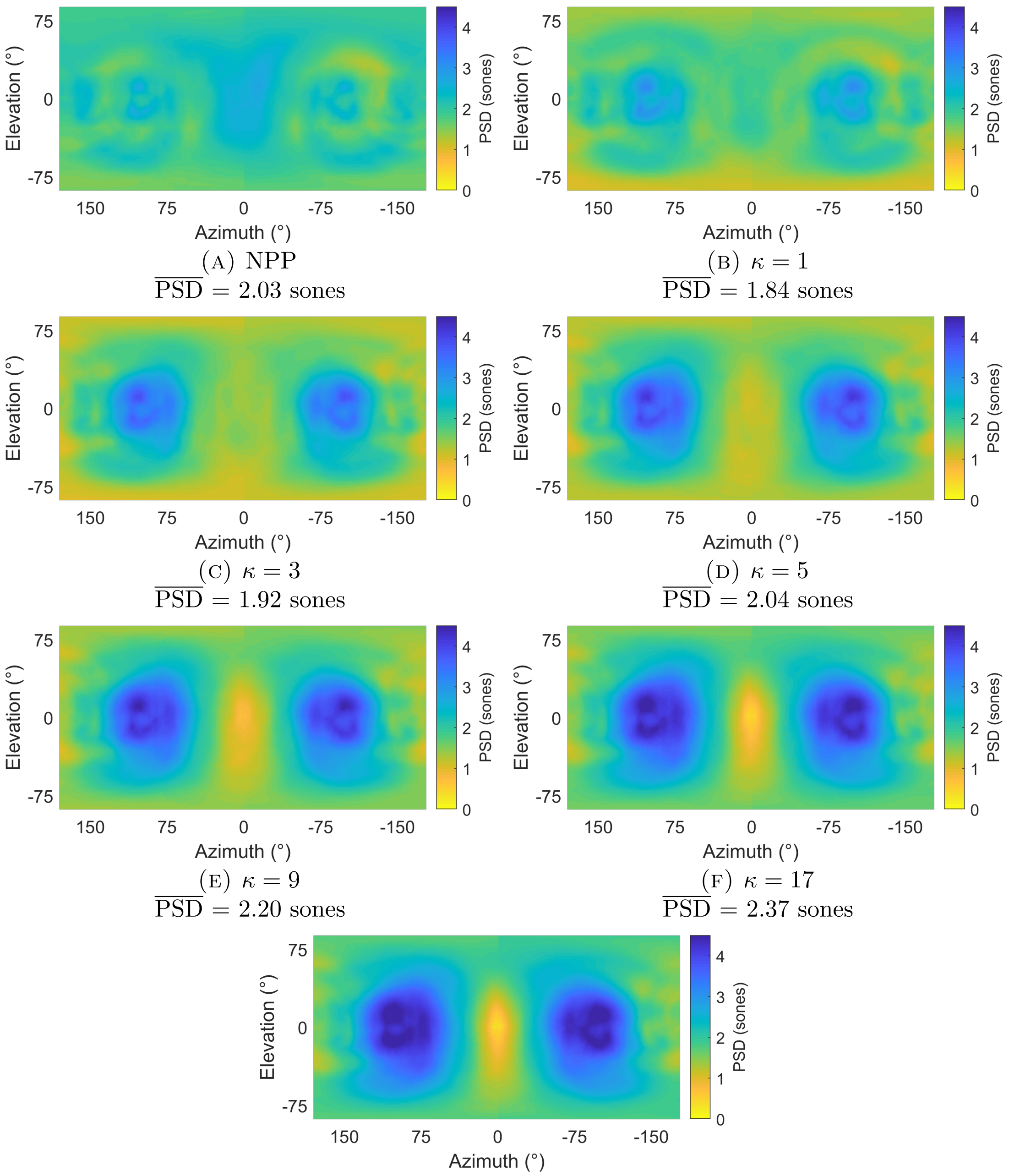
(C) $\kappa = 5$

(D) $\kappa = 9$

(E) $\kappa = 17$

(F) $\kappa = 33$

FIGURE B.4: DBQ RMS response, directional HRTF response and resulting DBE filters of binaural Ambisonic rendering with varying $\kappa$, $M = 5$ (left ear).



(A) NPP
$\overline{\text{PSD}}$ = 2.03 sones

(B) $\kappa = 1$
$\overline{\text{PSD}}$ = 1.84 sones

(C) $\kappa = 3$
$\overline{\text{PSD}}$ = 1.92 sones

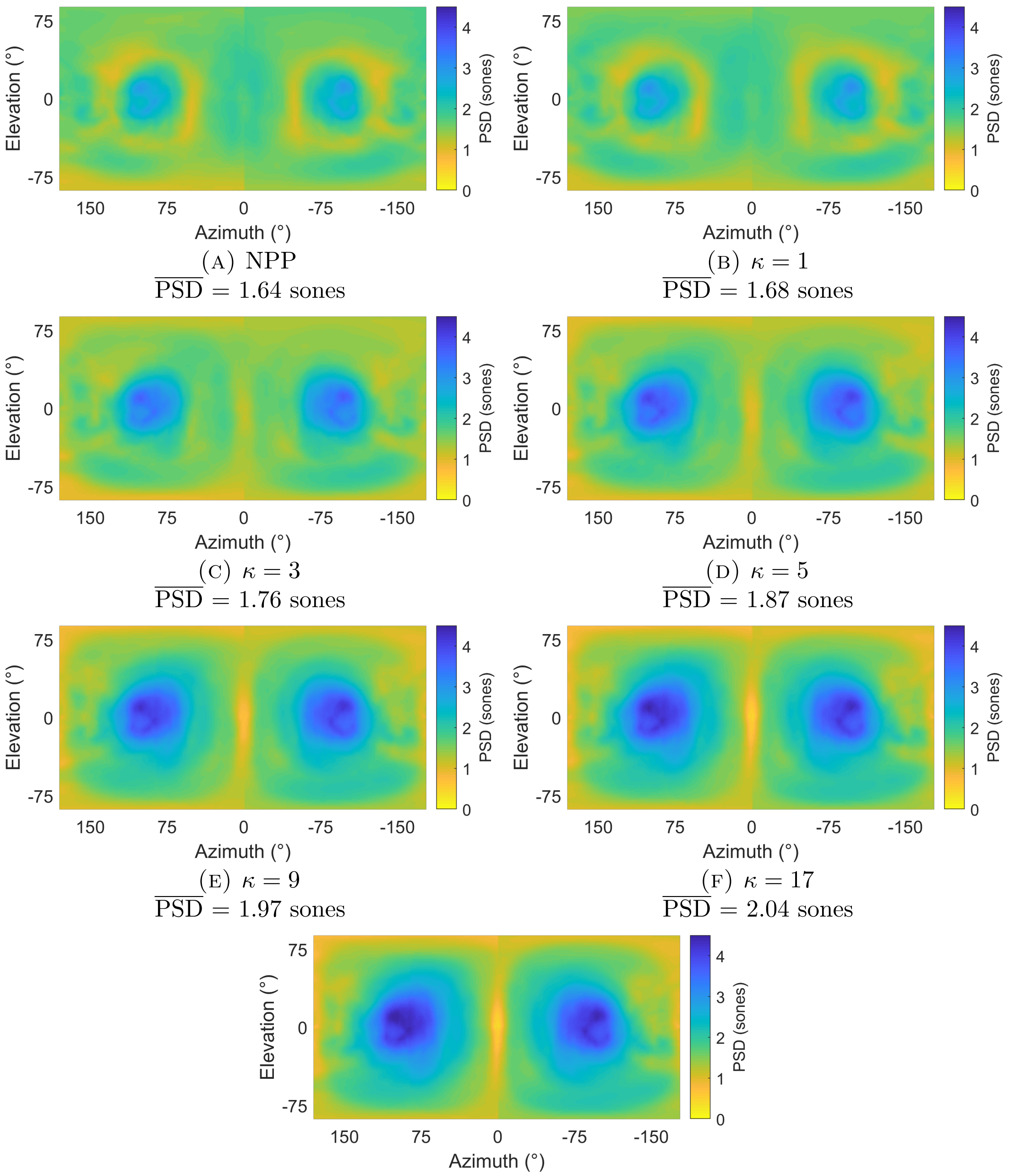
(D) $\kappa = 5$
$\overline{\text{PSD}}$ = 2.04 sones

(E) $\kappa = 9$
$\overline{\text{PSD}}$ = 2.20 sones

(F) $\kappa = 17$
$\overline{\text{PSD}}$ = 2.37 sones

(G) $\kappa = 33$
$\overline{\text{PSD}}$ = 2.52 sones

FIGURE B.5: PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, $M = 2$ (mean of left and right PSD values). NPP included, which denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.64$ sones

(B) $\kappa = 1$
$\overline{\text{PSD}} = 1.68$ sones

(C) $\kappa = 3$
$\overline{\text{PSD}} = 1.76$ sones

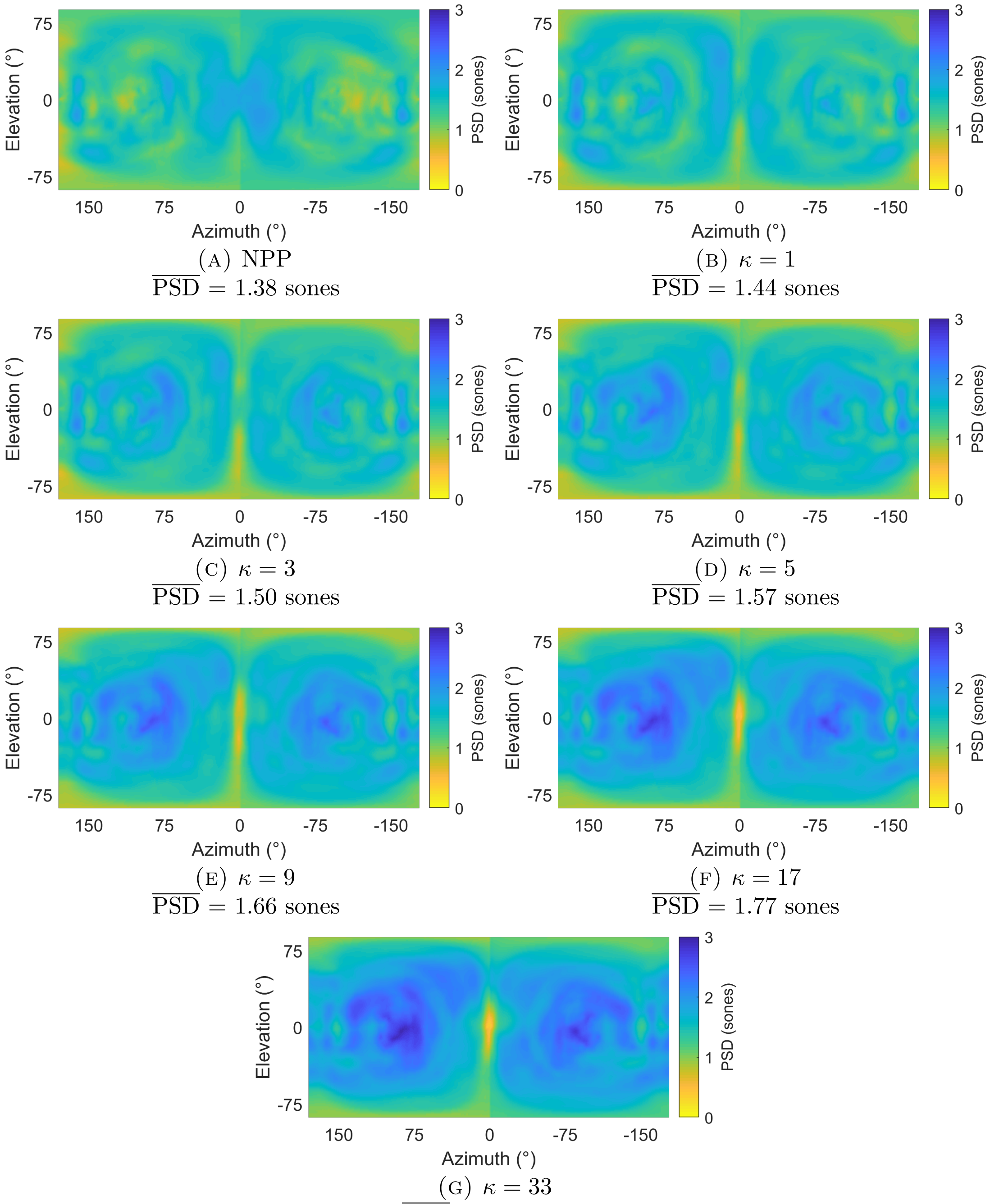
(D) $\kappa = 5$
$\overline{\text{PSD}} = 1.87$ sones

(E) $\kappa = 9$
$\overline{\text{PSD}} = 1.97$ sones

(F) $\kappa = 17$
$\overline{\text{PSD}} = 2.04$ sones

(G) $\kappa = 33$
$\overline{\text{PSD}} = 2.11$ sones

FIGURE B.6: PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, $M = 3$ (mean of left and right PSD values). NPP included, which denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.38$ sones

(B) $\kappa = 1$
$\overline{\text{PSD}} = 1.44$ sones

(C) $\kappa = 3$
$\overline{\text{PSD}} = 1.50$ sones

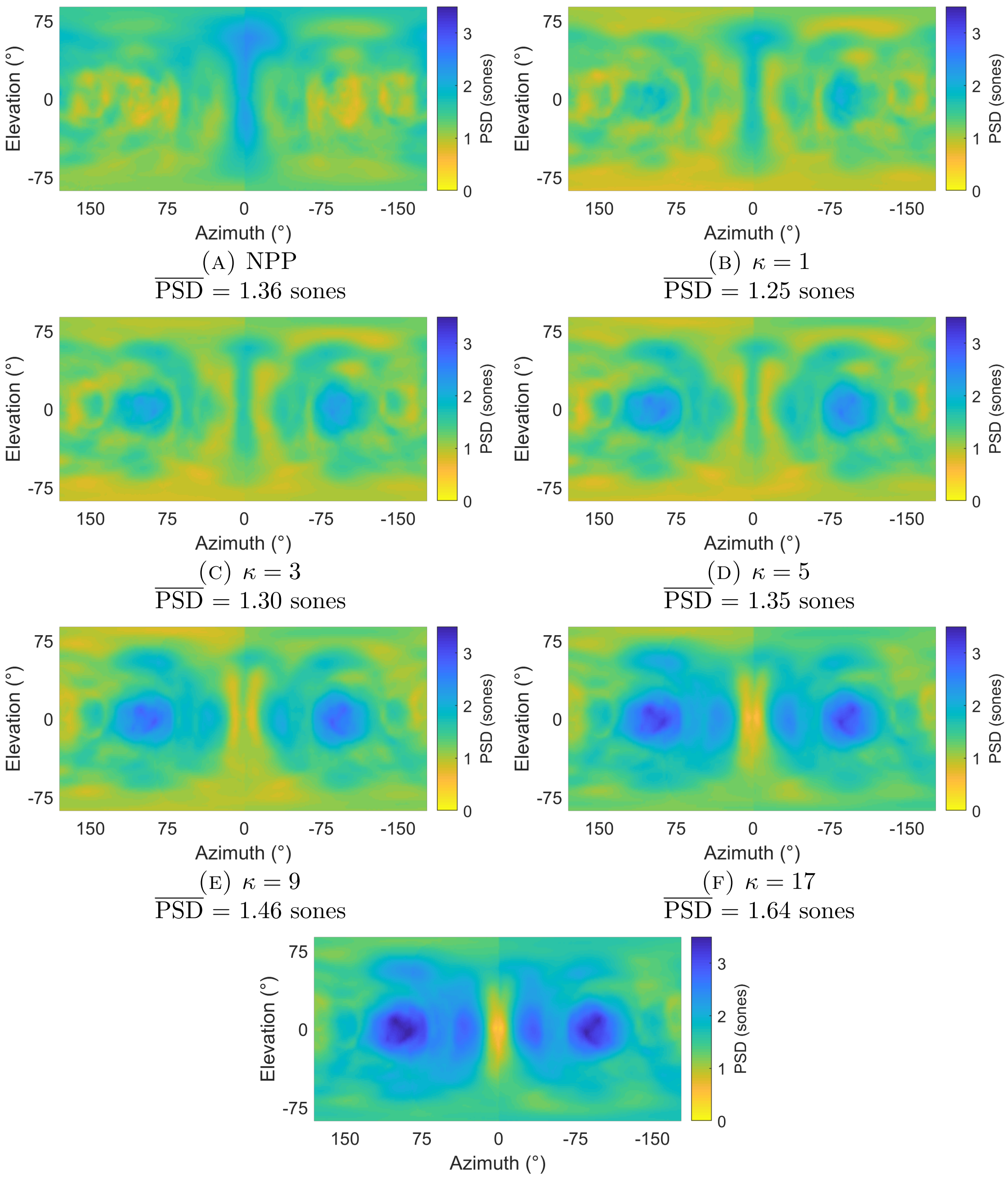
(D) $\kappa = 5$
$\overline{\text{PSD}} = 1.57$ sones

(E) $\kappa = 9$
$\overline{\text{PSD}} = 1.66$ sones

(F) $\kappa = 17$
$\overline{\text{PSD}} = 1.77$ sones

(G) $\kappa = 33$
$\overline{\text{PSD}} = 1.89$ sones

FIGURE B.7: PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, $M = 4$ (mean of left and right PSD values). NPP included, which denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.36$ sones

(B) $\kappa = 1$
$\overline{\text{PSD}} = 1.25$ sones

(C) $\kappa = 3$
$\overline{\text{PSD}} = 1.30$ sones

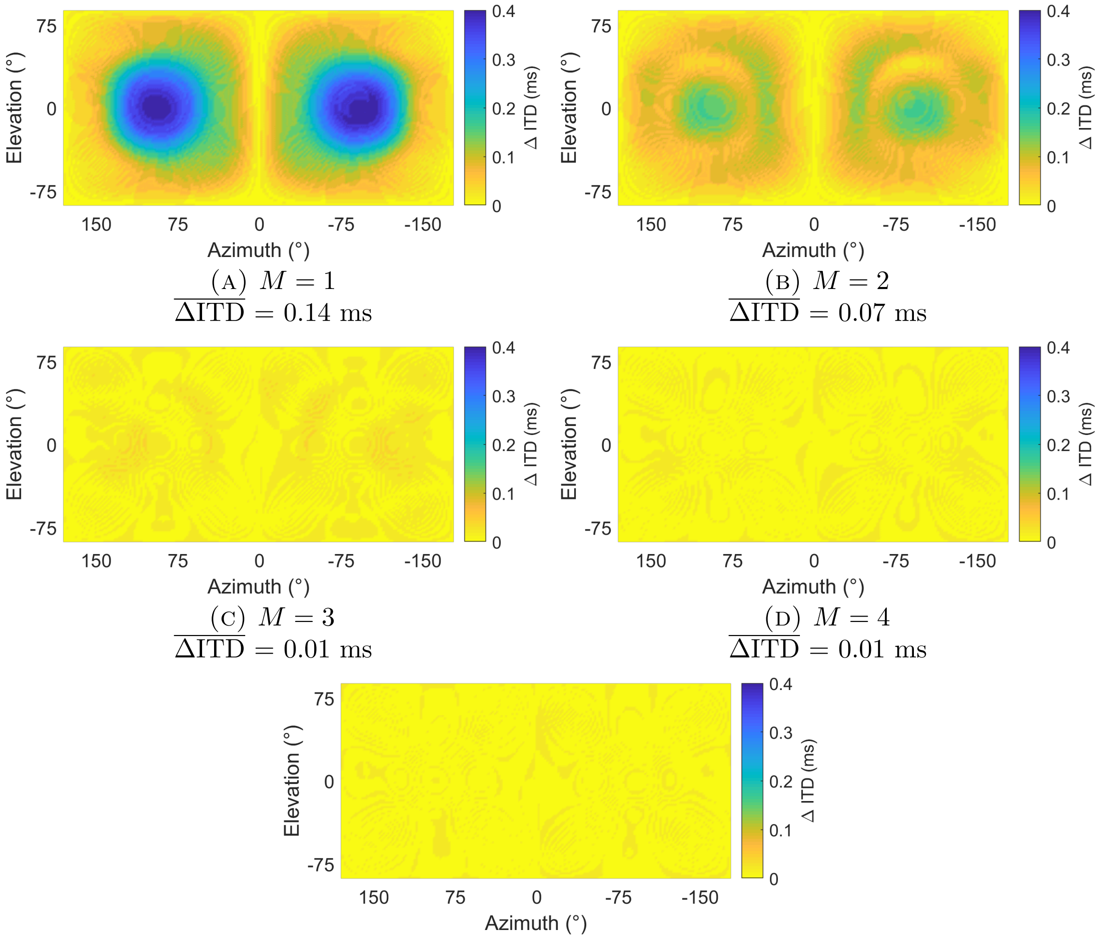
(D) $\kappa = 5$
$\overline{\text{PSD}} = 1.35$ sones

(E) $\kappa = 9$
$\overline{\text{PSD}} = 1.46$ sones

(F) $\kappa = 17$
$\overline{\text{PSD}} = 1.64$ sones

(G) $\kappa = 33$
$\overline{\text{PSD}} = 1.88$ sones

FIGURE B.8: PSD between HRTFs and binaural Ambisonic rendering with DBE and varying $\kappa$, for every measurement location on the sphere, $M = 5$ (mean of left and right PSD values). NPP included, which denotes no pre-processing.



(A) $M = 1$
$\overline{\Delta \text{ITD}} = 0.14$ ms

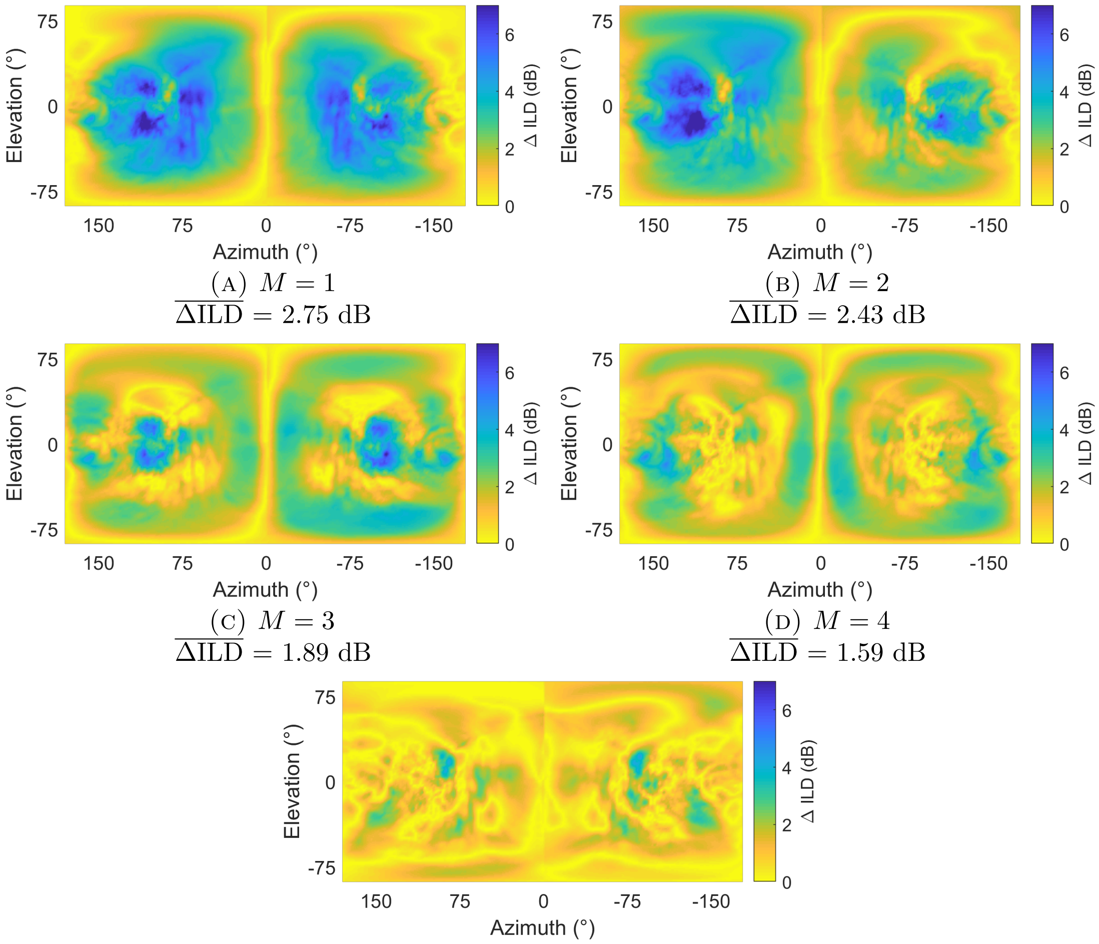
(B) $M = 2$
$\overline{\Delta \text{ITD}} = 0.07$ ms

(C) $M = 3$
$\overline{\Delta \text{ITD}} = 0.01$ ms

(D) $M = 4$
$\overline{\Delta \text{ITD}} = 0.01$ ms

(E) $M = 5$
$\overline{\Delta \text{ITD}} = 0.00$ ms

FIGURE B.9: ΔITD between HRTFs and binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$
$\overline{\Delta\text{ILD}} = 2.75$ dB

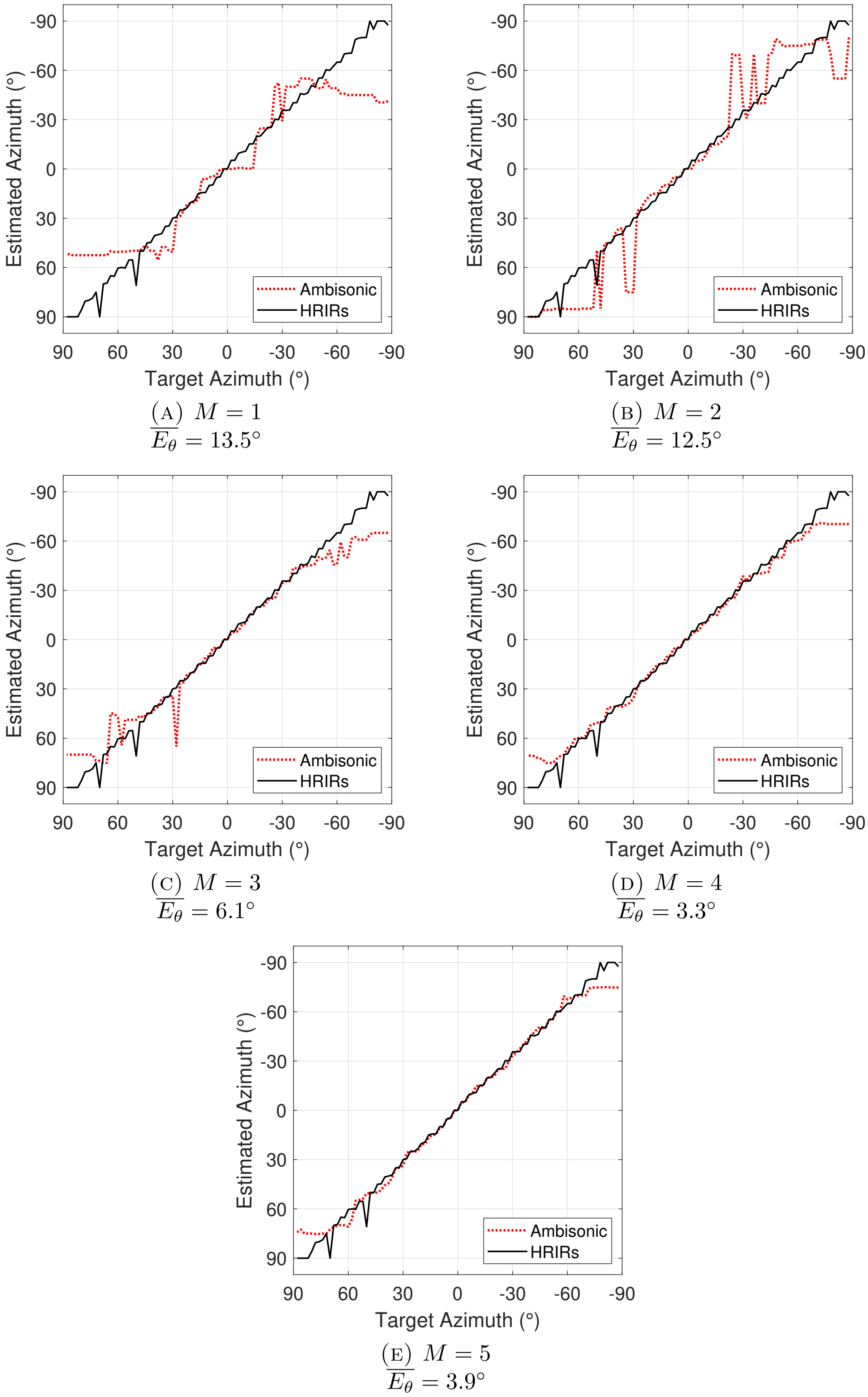
(B) $M = 2$
$\overline{\Delta\text{ILD}} = 2.43$ dB

(C) $M = 3$
$\overline{\Delta\text{ILD}} = 1.89$ dB

(D) $M = 4$
$\overline{\Delta\text{ILD}} = 1.59$ dB

(E) $M = 5$
$\overline{\Delta\text{ILD}} = 0.91$ dB

FIGURE B.10: ΔILD between HRTFs and binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$
$\overline{E_\theta} = 13.5°$

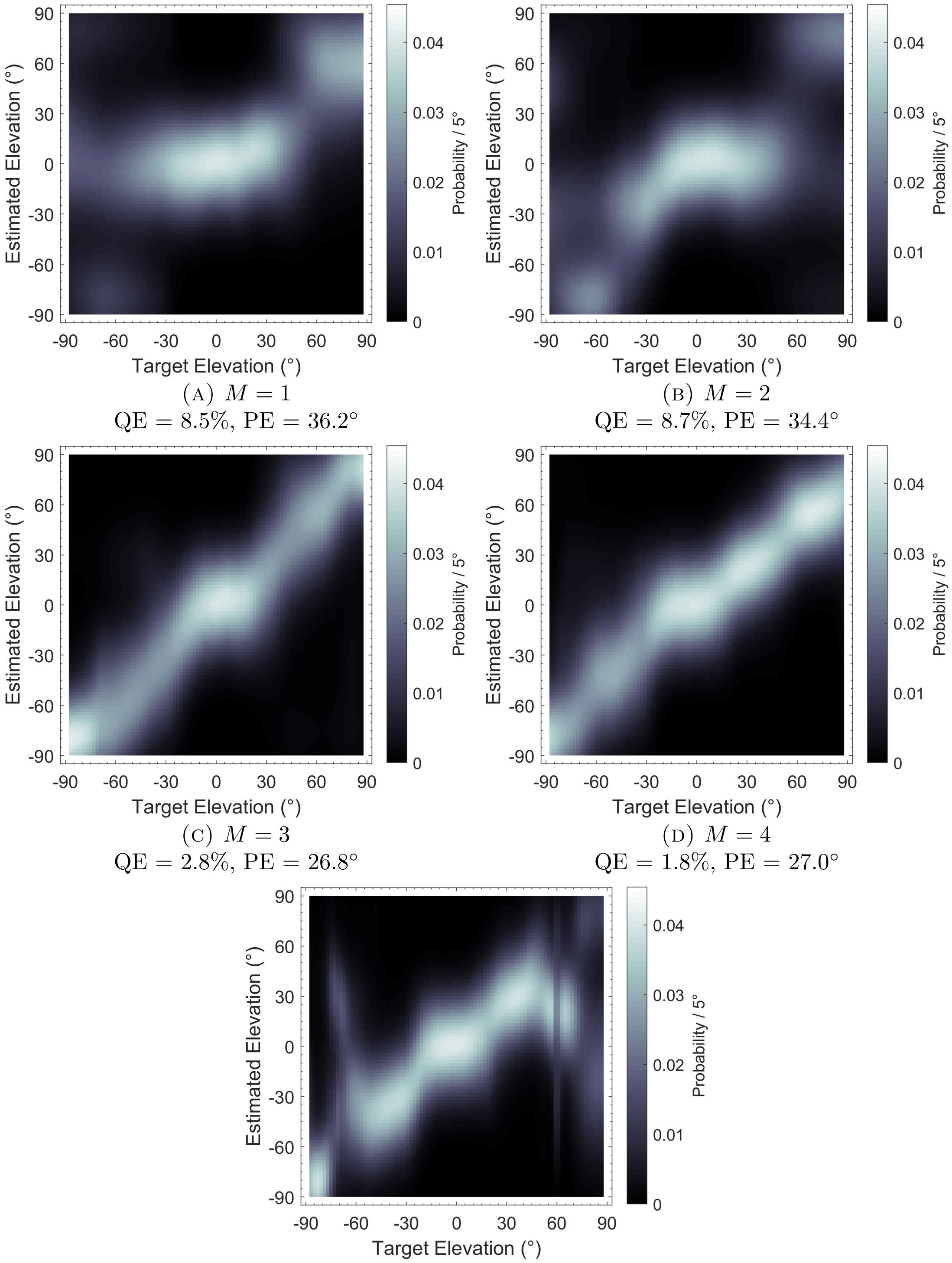
(B) $M = 2$
$\overline{E_\theta} = 12.5°$

(C) $M = 3$
$\overline{E_\theta} = 6.1°$

(D) $M = 4$
$\overline{E_\theta} = 3.3°$

(E) $M = 5$
$\overline{E_\theta} = 3.9°$

FIGURE B.11: Horizontal localisation model plots of binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$.



(A) $M = 1$
QE = 8.5%, PE = 36.2°

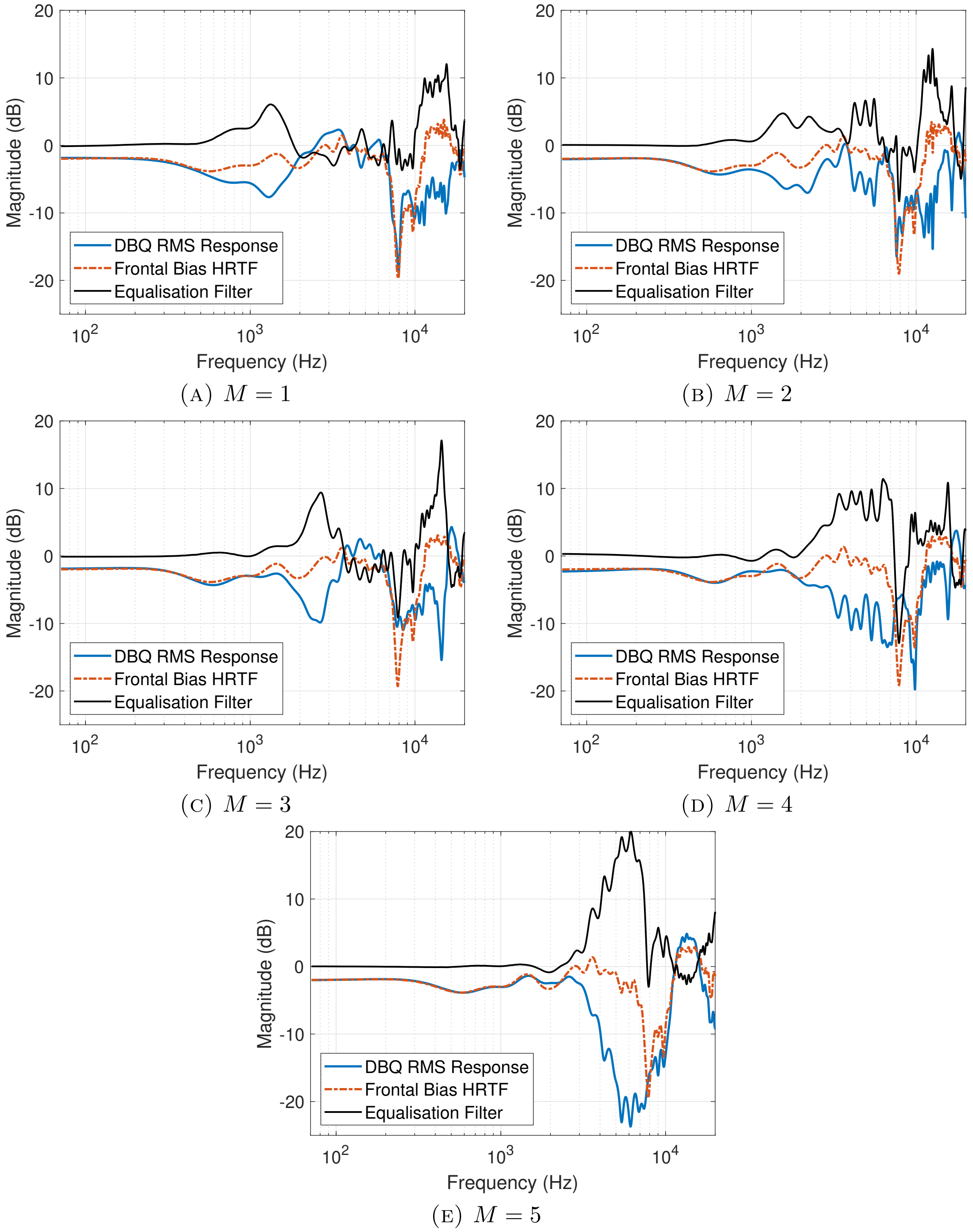
(B) $M = 2$
QE = 8.7%, PE = 34.4°

(C) $M = 3$
QE = 2.8%, PE = 26.8°

(D) $M = 4$
QE = 1.8%, PE = 27.0°

(E) $M = 5$
QE = 7.8%, PE = 31.1°

FIGURE B.12: Sagittal plane localisation model plots of binaural Ambisonic rendering with DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$.



(A) $M = 1$

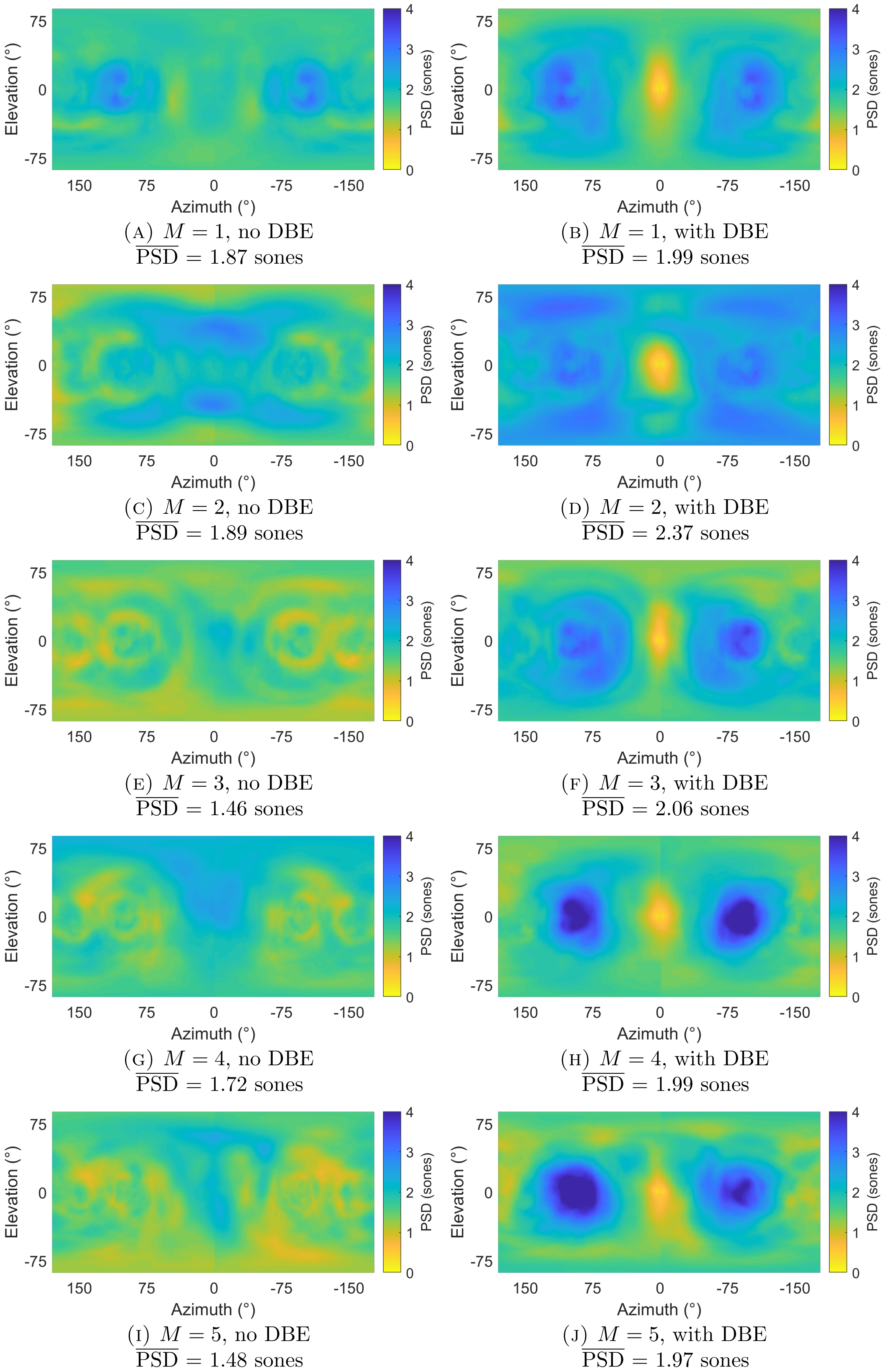
(B) $M = 2$

(C) $M = 3$

(D) $M = 4$

(E) $M = 5$

FIGURE B.13: DBQ RMS response, directional HRTF response and resulting DBE filters at $\kappa = 33$ of binaural Ambisonic rendering of the T-design loudspeaker configurations for $\{M = 1, M = 2, ..., M = 5\}$ (left ear).



(A) $M = 1$, no DBE
$\overline{\text{PSD}} = 1.87$ sones

(B) $M = 1$, with DBE
$\overline{\text{PSD}} = 1.99$ sones

(C) $M = 2$, no DBE
$\overline{\text{PSD}} = 1.89$ sones

(D) $M = 2$, with DBE
$\overline{\text{PSD}} = 2.37$ sones

(E) $M = 3$, no DBE
$\overline{\text{PSD}} = 1.46$ sones

(F) $M = 3$, with DBE
$\overline{\text{PSD}} = 2.06$ sones

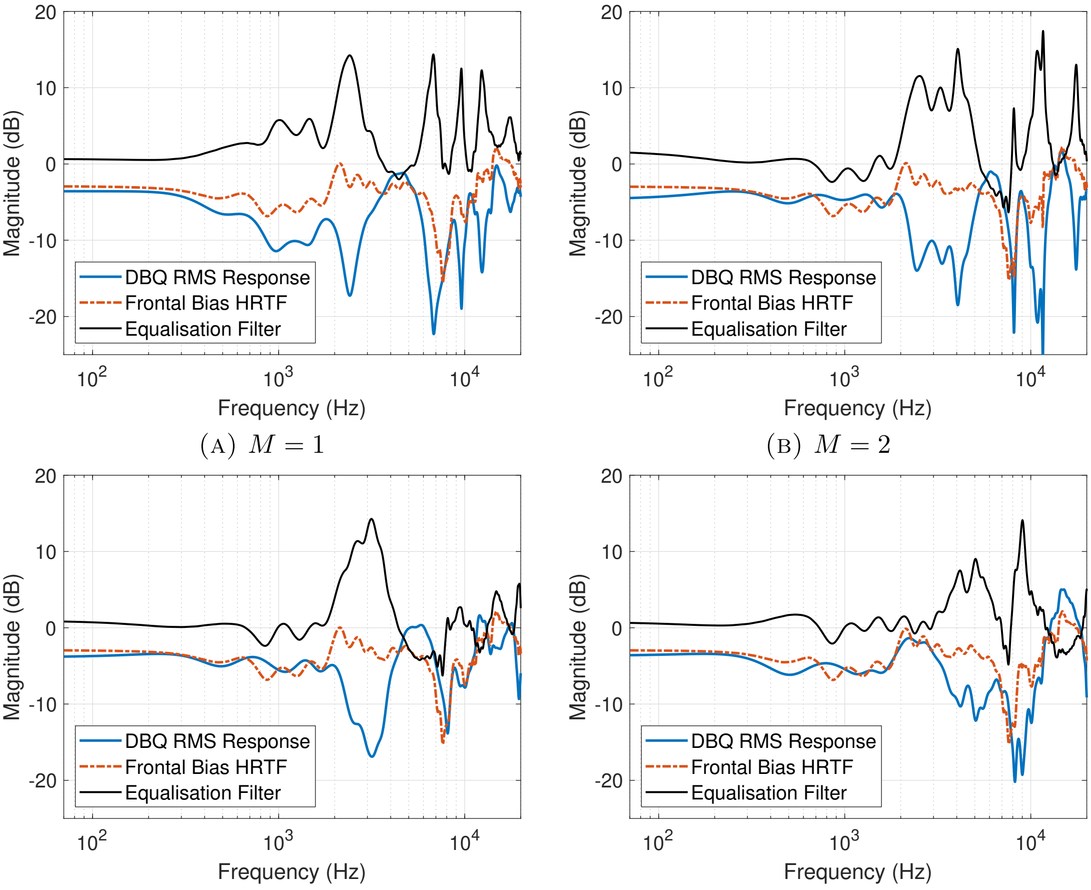
(G) $M = 4$, no DBE
$\overline{\text{PSD}} = 1.72$ sones

(H) $M = 4$, with DBE
$\overline{\text{PSD}} = 1.99$ sones

(I) $M = 5$, no DBE
$\overline{\text{PSD}} = 1.48$ sones

(J) $M = 5$, with DBE
$\overline{\text{PSD}} = 1.97$ sones

FIGURE B.14: PSD between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, ..., M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using T-design loudspeaker configurations.



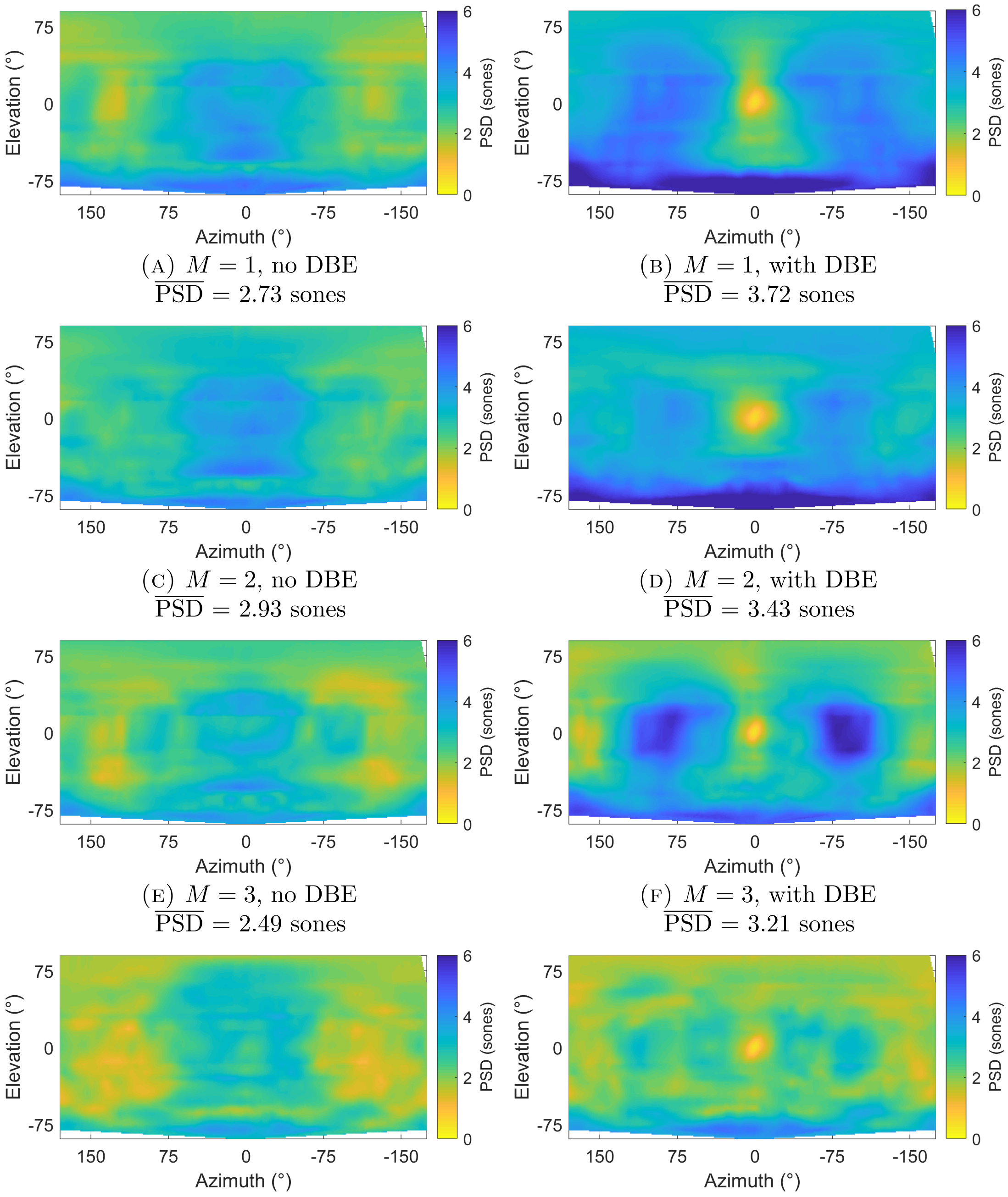
(A) $M = 1$

(B) $M = 2$

(C) $M = 3$

(D) $M = 5$

FIGURE B.15: DBQ RMS response, directional HRTF response and resulting DBE filters at $\kappa = 33$ of binaural Ambisonic rendering for $\{M = 1, M = 2, M = 3, M = 5\}$, using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a) (left ear).



(A) $M = 1$, no DBE
$\overline{\text{PSD}} = 2.73$ sones

(B) $M = 1$, with DBE
$\overline{\text{PSD}} = 3.72$ sones

(C) $M = 2$, no DBE
$\overline{\text{PSD}} = 2.93$ sones

(D) $M = 2$, with DBE
$\overline{\text{PSD}} = 3.43$ sones

(E) $M = 3$, no DBE
$\overline{\text{PSD}} = 2.49$ sones

(F) $M = 3$, with DBE
$\overline{\text{PSD}} = 3.21$ sones

(G) $M = 5$, no DBE
$\overline{\text{PSD}} = 2.21$ sones

(H) $M = 5$, with DBE
$\overline{\text{PSD}} = 2.25$ sones

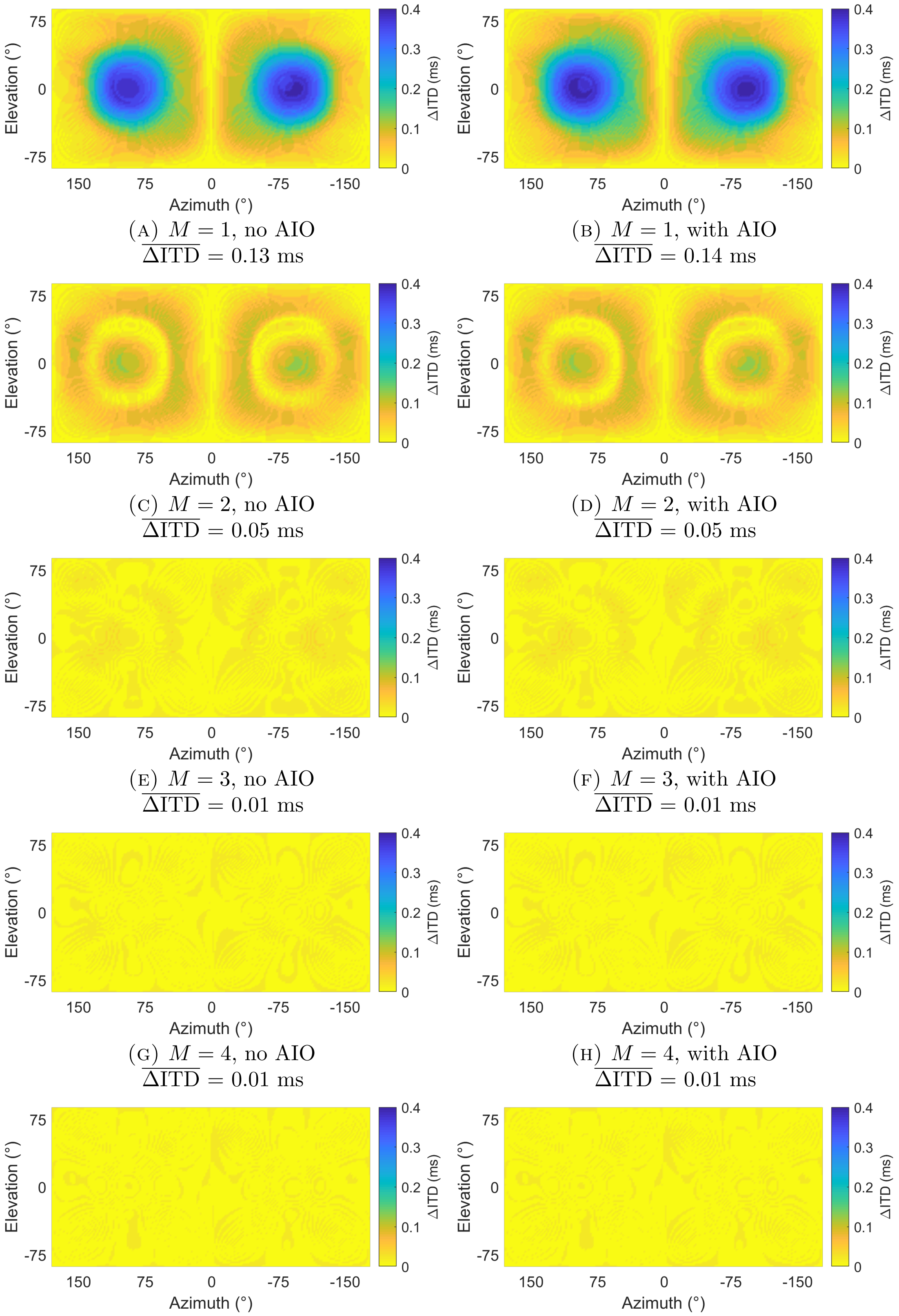
FIGURE B.16: PSD between HRTFs and binaural Ambisonic rendering with and without DBE at $\kappa = 33$, for $\{M = 1, M = 2, M = 3, M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).

# Appendix C

# Supplementary Plots for Ambisonic Interaural Level Difference Optimisation

This appendix presents supplementary plots for Chapter 6. All computation was carried out offline in MATLAB version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4. Figure C.1 presents ΔITD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without AIO. Figure C.2 presents the estimated $\theta$ values of binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \theta < +90°$ and $\phi = 0°$, both with and without AIO. Figure C.3 presents the estimated sagittal plane localisation plots of binaural Ambisonic rendering for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \phi < +90°$ and $\theta = 0°$, both with and without AIO. Figure C.4 presents PSD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without AIO, using T-design loudspeaker configurations. Figure

C.5 presents ΔILD values for $\{M = 1, M = 2, ..., M = 5\}$ for every measurement location on the sphere, both with and without AIO, using T-design loudspeaker configurations. Figure C.6 presents PSD values for $\{M = 1, M = 2, M = 3, M = 5\}$ for every measurement location on the sphere, both with and without AIO, using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a). Figure C.7 presents ΔILD values for $\{M = 1, M = 2, M = 3, M = 5\}$ for every measurement location on the sphere, both with and without AIO, using Lebedev loudspeaker configurations and individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).



(A) $M = 1$, no AIO
$\overline{\Delta\text{ITD}} = 0.13$ ms

(B) $M = 1$, with AIO
$\overline{\Delta\text{ITD}} = 0.14$ ms

(C) $M = 2$, no AIO
$\overline{\Delta\text{ITD}} = 0.05$ ms

(D) $M = 2$, with AIO
$\overline{\Delta\text{ITD}} = 0.05$ ms

(E) $M = 3$, no AIO
$\overline{\Delta\text{ITD}} = 0.01$ ms

(F) $M = 3$, with AIO
$\overline{\Delta\text{ITD}} = 0.01$ ms

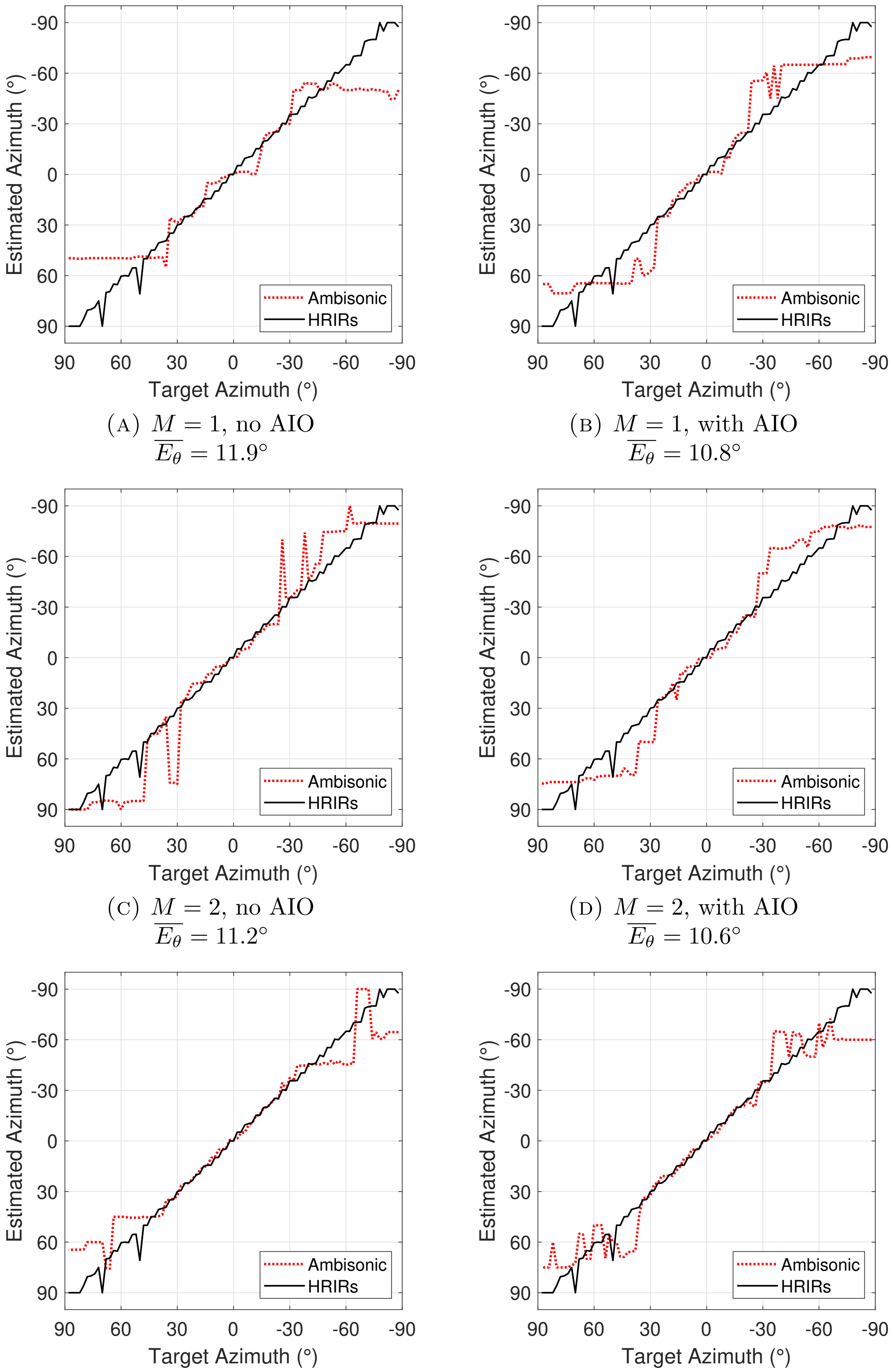
(G) $M = 4$, no AIO
$\overline{\Delta\text{ITD}} = 0.01$ ms

(H) $M = 4$, with AIO
$\overline{\Delta\text{ITD}} = 0.01$ ms

(I) $M = 5$, no AIO
$\overline{\Delta\text{ITD}} = 0.00$ ms

(J) $M = 5$, with AIO
$\overline{\Delta\text{ITD}} = 0.00$ ms

FIGURE C.1: ΔITD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere.



(A) $M = 1$, no AIO
$\overline{E_\theta} = 11.9°$

(B) $M = 1$, with AIO
$\overline{E_\theta} = 10.8°$

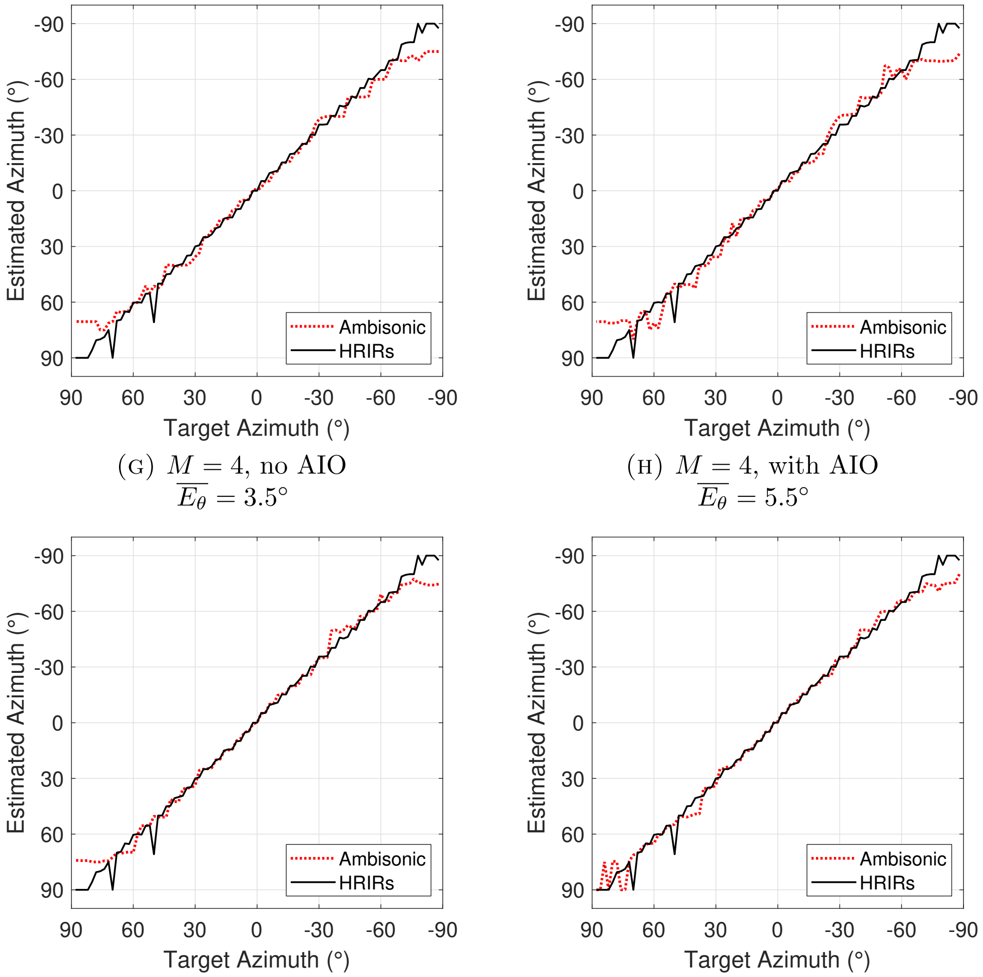
(C) $M = 2$, no AIO
$\overline{E_\theta} = 11.2°$

(D) $M = 2$, with AIO
$\overline{E_\theta} = 10.6°$

(E) $M = 3$, no AIO
$\overline{E_\theta} = 8.0°$

(F) $M = 3$, with AIO
$\overline{E_\theta} = 8.7°$

FIGURE C.2: *Cont.*



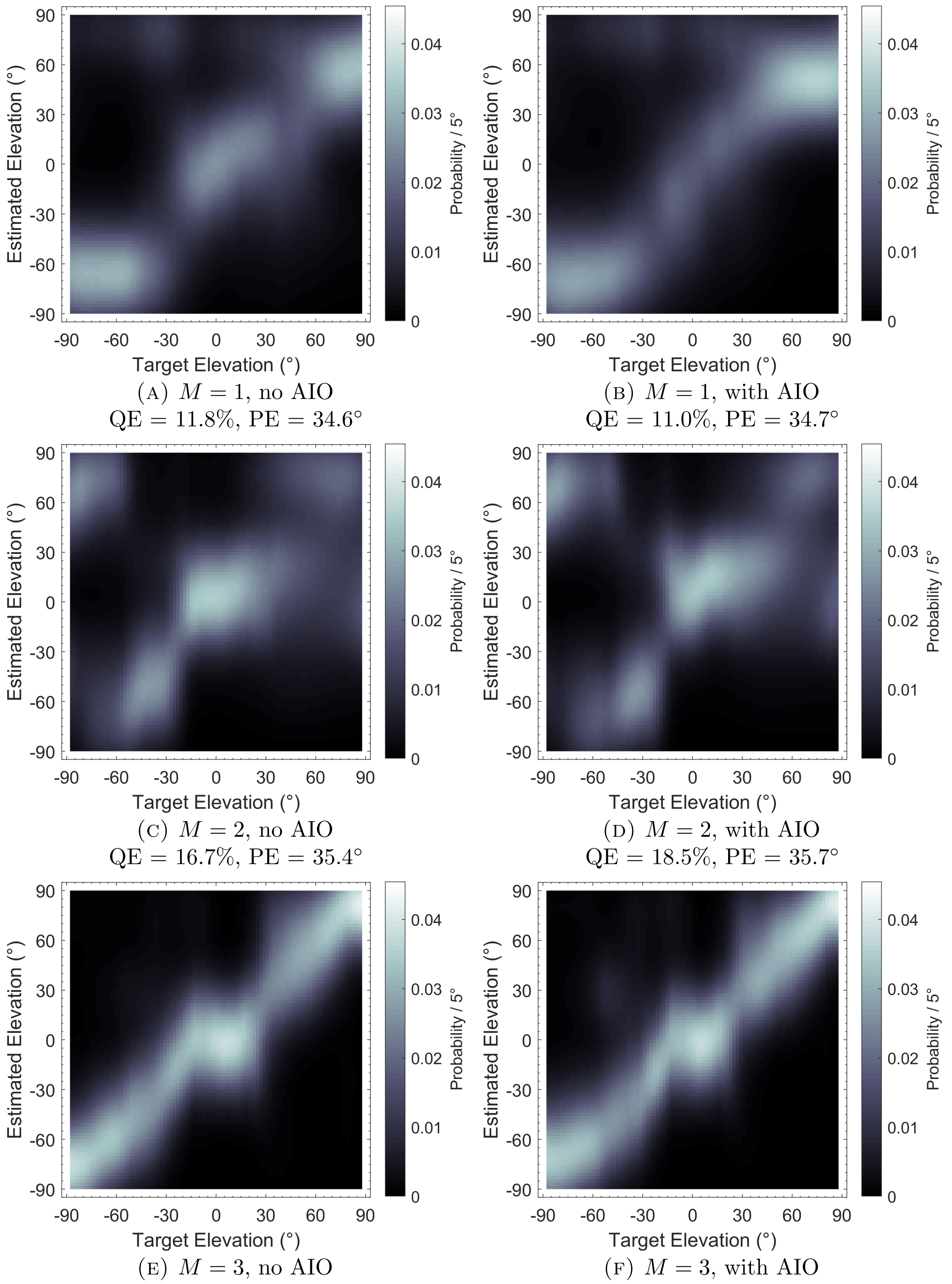
(G) $M = 4$, no AIO
$\overline{E_\theta} = 3.5°$

(H) $M = 4$, with AIO
$\overline{E_\theta} = 5.5°$

(I) $M = 5$, no AIO
$\overline{E_\theta} = 3.9°$

(J) $M = 5$, with AIO
$\overline{E_\theta} = 4.0°$

FIGURE C.2: Horizontal localisation model plots of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, with and without AIO.



(A) $M = 1$, no AIO
QE = 11.8%, PE = 34.6°

(B) $M = 1$, with AIO
QE = 11.0%, PE = 34.7°

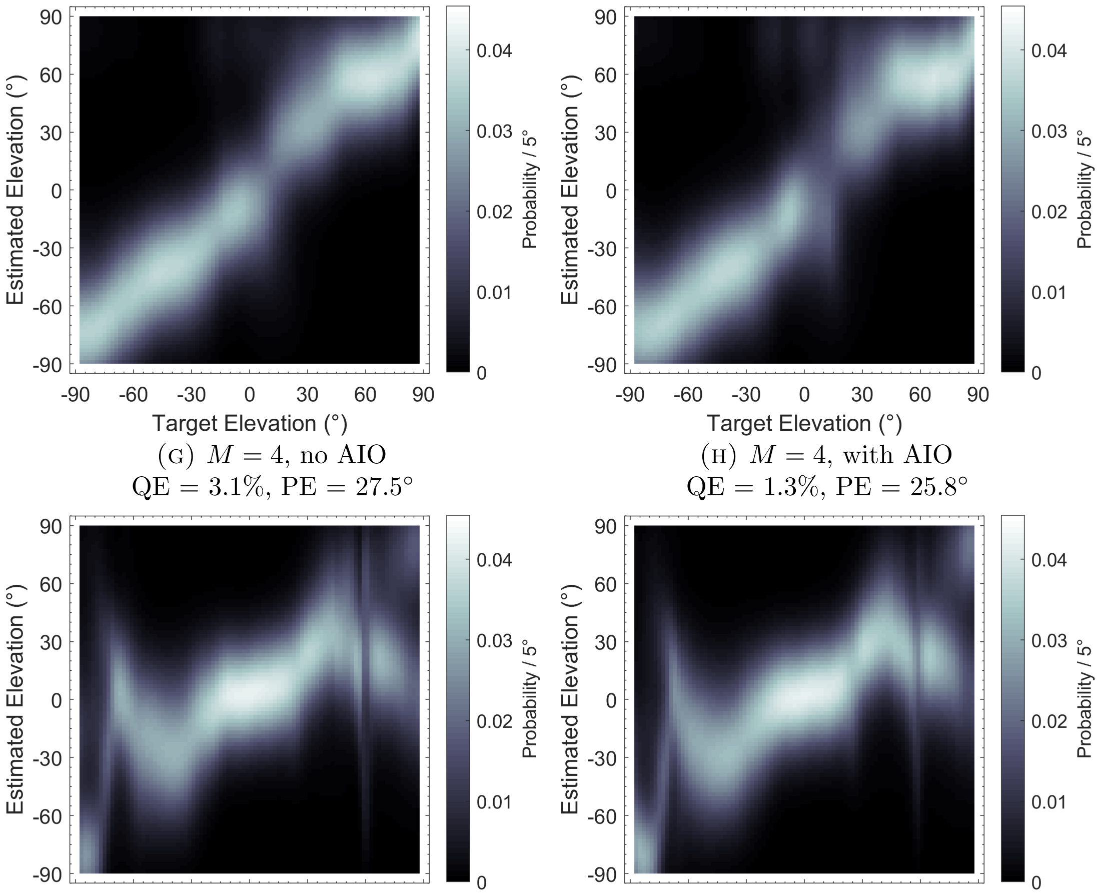
(C) $M = 2$, no AIO
QE = 16.7%, PE = 35.4°

(D) $M = 2$, with AIO
QE = 18.5%, PE = 35.7°

(E) $M = 3$, no AIO
QE = 2.8%, PE = 27.2°

(F) $M = 3$, with AIO
QE = 3.7%, PE = 27.8°

FIGURE C.3: *Cont.*



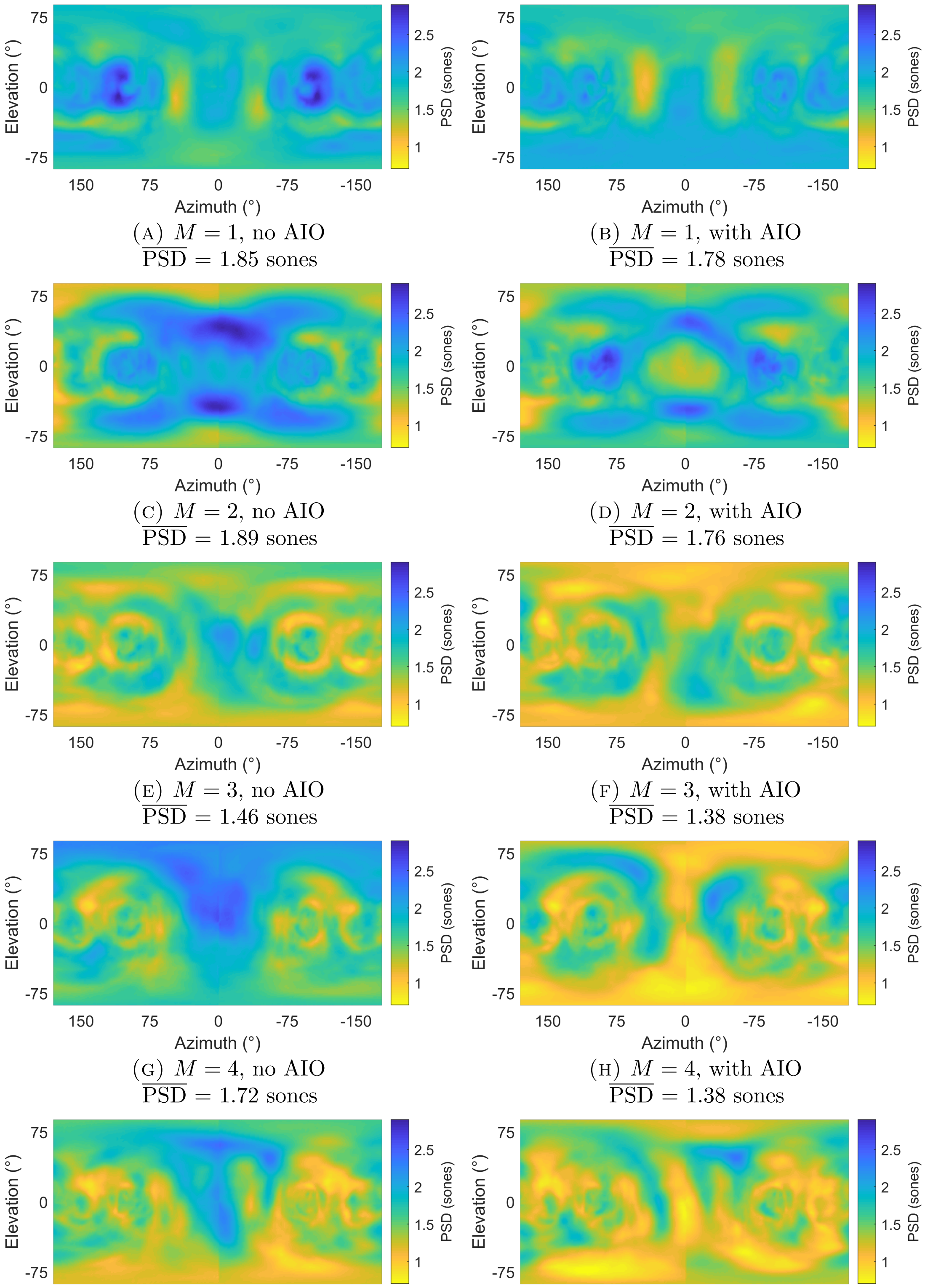
(G) $M = 4$, no AIO
QE = 3.1%, PE = 27.5°

(H) $M = 4$, with AIO
QE = 1.3%, PE = 25.8°

(I) $M = 5$, no AIO
QE = 6.0%, PE = 34.7°

(J) $M = 5$, with AIO
QE = 5.1%, PE = 33.3°

FIGURE C.3: Sagittal plane localisation model plots of Ambisonic orders $\{M = 1, M = 2, ..., M = 5\}$, with and without AIO.



(A) $M = 1$, no AIO
$\overline{\text{PSD}} = 1.85$ sones

(B) $M = 1$, with AIO
$\overline{\text{PSD}} = 1.78$ sones

(C) $M = 2$, no AIO
$\overline{\text{PSD}} = 1.89$ sones

(D) $M = 2$, with AIO
$\overline{\text{PSD}} = 1.76$ sones

(E) $M = 3$, no AIO
$\overline{\text{PSD}} = 1.46$ sones

(F) $M = 3$, with AIO
$\overline{\text{PSD}} = 1.38$ sones

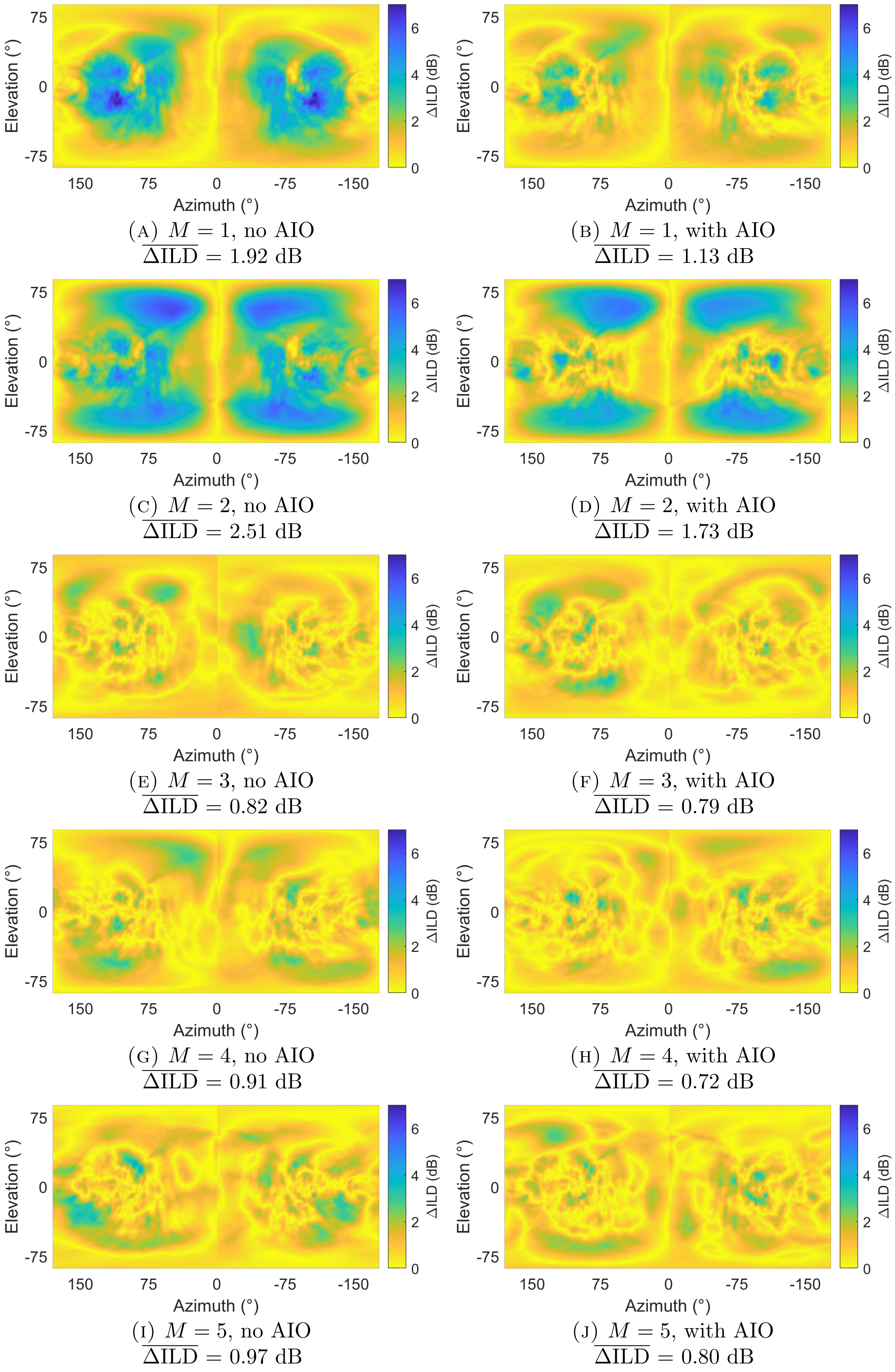
(G) $M = 4$, no AIO
$\overline{\text{PSD}} = 1.72$ sones

(H) $M = 4$, with AIO
$\overline{\text{PSD}} = 1.38$ sones

(I) $M = 5$, no AIO
$\overline{\text{PSD}} = 1.48$ sones

(J) $M = 5$, with AIO
$\overline{\text{PSD}} = 1.31$ sones

FIGURE C.4: PSD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using T-design loudspeaker configurations.



(A) $M = 1$, no AIO
$\overline{\Delta\text{ILD}} = 1.92$ dB

(B) $M = 1$, with AIO
$\overline{\Delta\text{ILD}} = 1.13$ dB

(C) $M = 2$, no AIO
$\overline{\Delta\text{ILD}} = 2.51$ dB

(D) $M = 2$, with AIO
$\overline{\Delta\text{ILD}} = 1.73$ dB

(E) $M = 3$, no AIO
$\overline{\Delta\text{ILD}} = 0.82$ dB

(F) $M = 3$, with AIO
$\overline{\Delta\text{ILD}} = 0.79$ dB

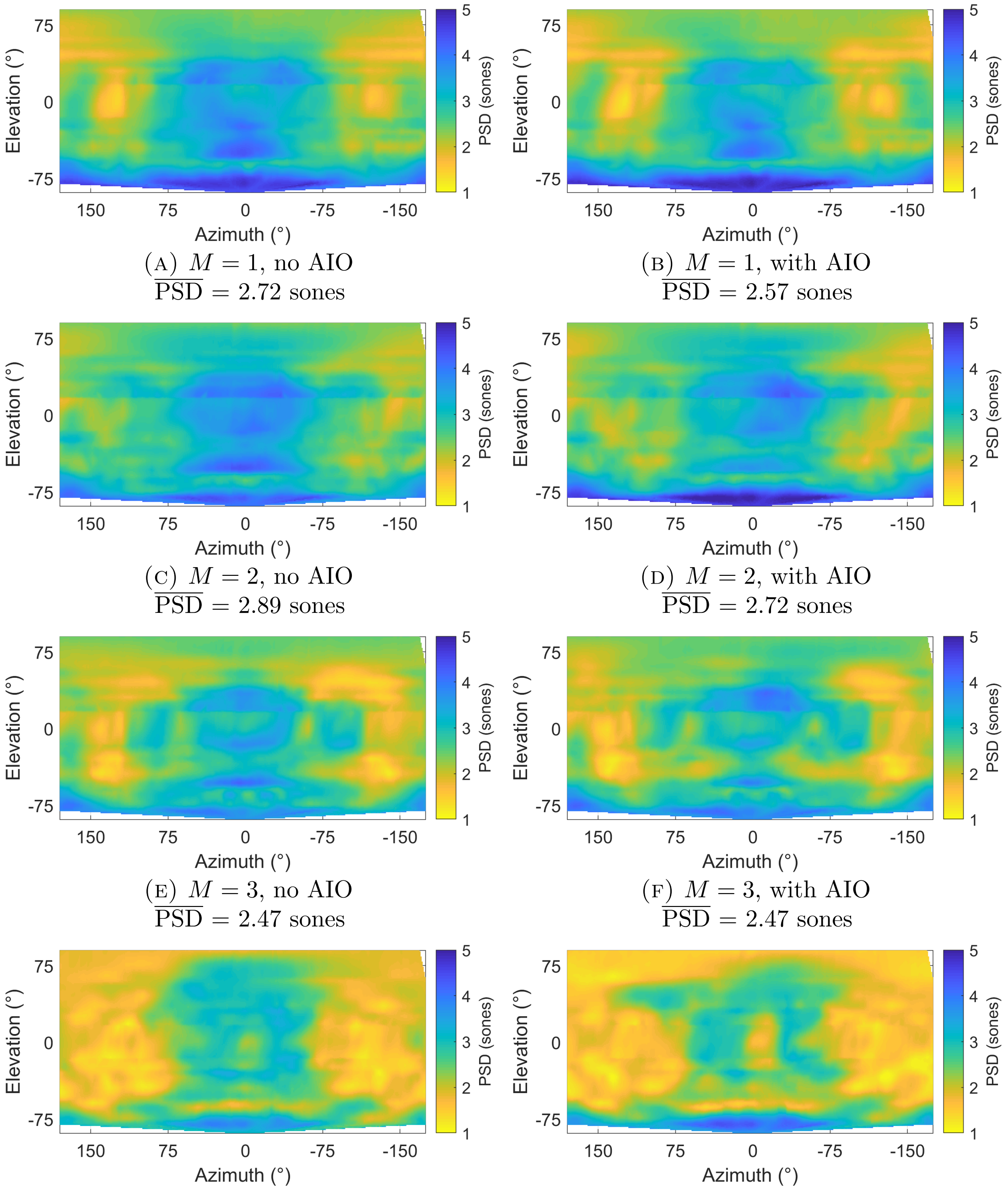
(G) $M = 4$, no AIO
$\overline{\Delta\text{ILD}} = 0.91$ dB

(H) $M = 4$, with AIO
$\overline{\Delta\text{ILD}} = 0.72$ dB

(I) $M = 5$, no AIO
$\overline{\Delta\text{ILD}} = 0.97$ dB

(J) $M = 5$, with AIO
$\overline{\Delta\text{ILD}} = 0.80$ dB

FIGURE C.5: ΔILD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere, using T-design loudspeaker configurations.



(A) $M = 1$, no AIO
$\overline{\text{PSD}} = 2.72$ sones

(B) $M = 1$, with AIO
$\overline{\text{PSD}} = 2.57$ sones

(C) $M = 2$, no AIO
$\overline{\text{PSD}} = 2.89$ sones

(D) $M = 2$, with AIO
$\overline{\text{PSD}} = 2.72$ sones

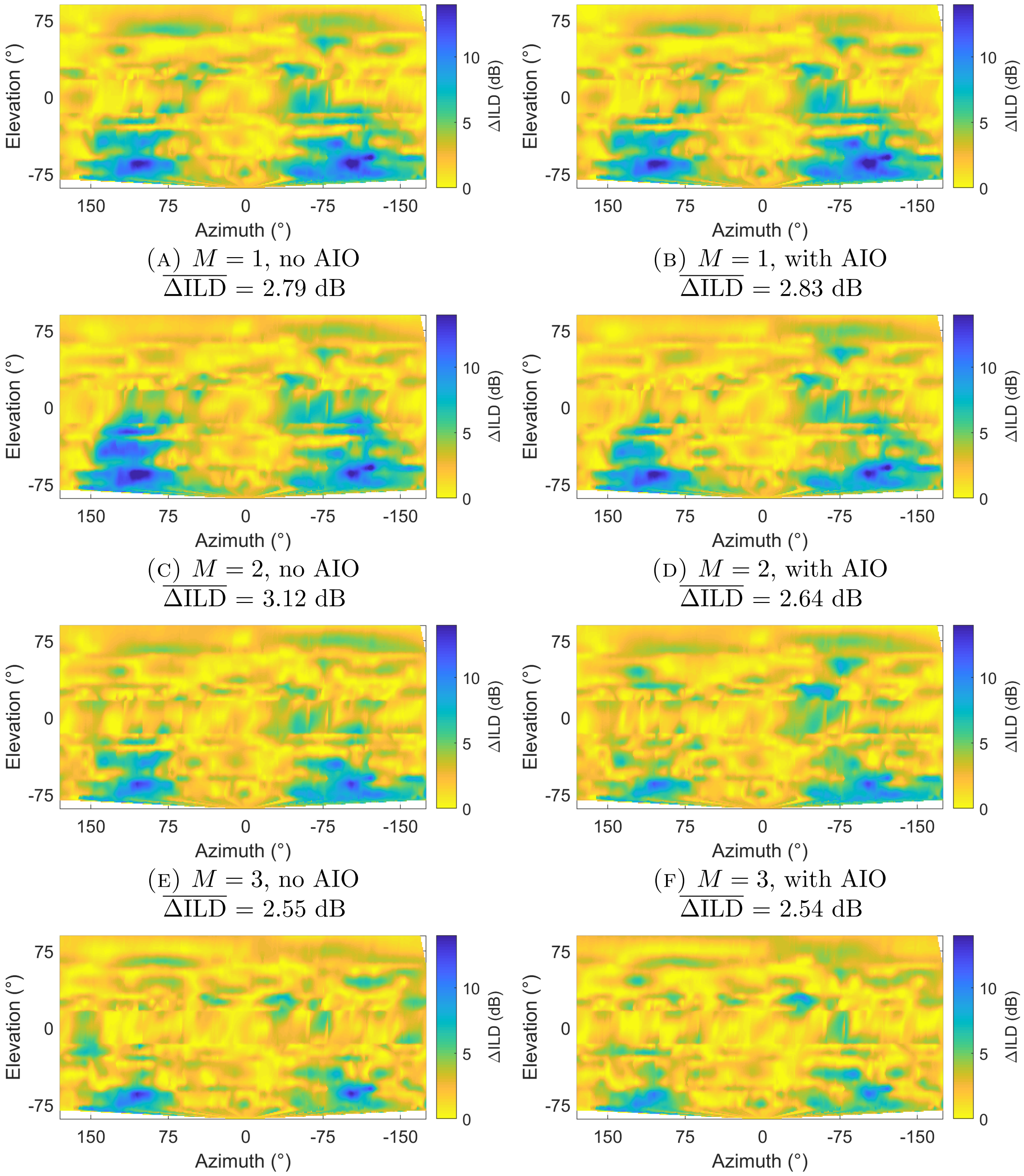
(E) $M = 3$, no AIO
$\overline{\text{PSD}} = 2.47$ sones

(F) $M = 3$, with AIO
$\overline{\text{PSD}} = 2.47$ sones

(G) $M = 5$, no AIO
$\overline{\text{PSD}} = 2.17$ sones

(H) $M = 5$, with AIO
$\overline{\text{PSD}} = 2.05$ sones

FIGURE C.6: PSD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ and every measurement location on the sphere (mean of left and right PSD values), using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).



(A) $M = 1$, no AIO
$\overline{\Delta\text{ILD}} = 2.79$ dB

(B) $M = 1$, with AIO
$\overline{\Delta\text{ILD}} = 2.83$ dB

(C) $M = 2$, no AIO
$\overline{\Delta\text{ILD}} = 3.12$ dB

(D) $M = 2$, with AIO
$\overline{\Delta\text{ILD}} = 2.64$ dB

(E) $M = 3$, no AIO
$\overline{\Delta\text{ILD}} = 2.55$ dB

(F) $M = 3$, with AIO
$\overline{\Delta\text{ILD}} = 2.54$ dB

(G) $M = 5$, no AIO
$\overline{\Delta\text{ILD}} = 2.35$ dB

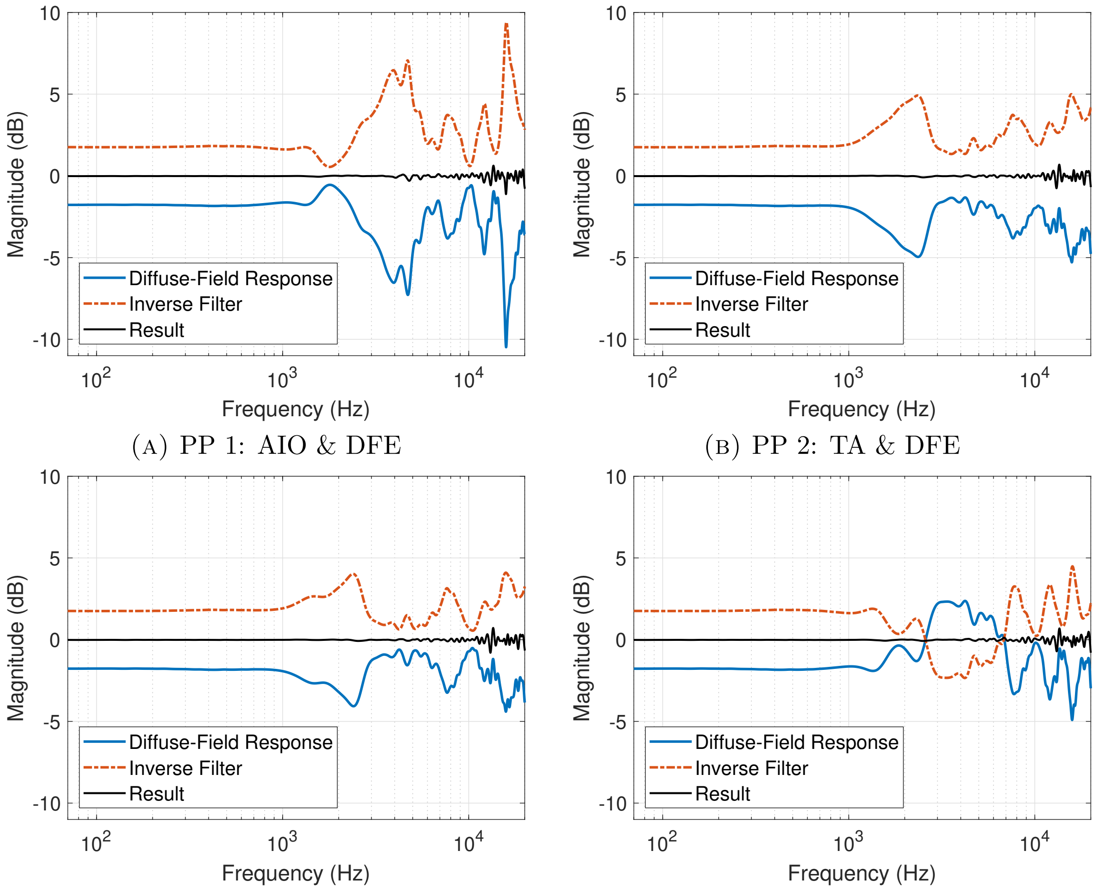
(H) $M = 5$, with AIO
$\overline{\Delta\text{ILD}} = 2.18$ dB

FIGURE C.7: ΔILD between HRTFs and binaural Ambisonic rendering with and without AIO, for $\{M = 1, M = 2, ..., M = 5\}$ across every measurement location on the sphere, using individualised HRTFs from the SADIE II database, subject H20 (Armstrong et al., 2018a).

# Appendix D

# Supplementary Plots for Combinations of Ambisonic Pre-Processing Techniques

This appendix presents supplementary plots for Chapter 7. All computation was carried out offline in MATLAB version 9.3.0 - R2017b and Ambisonic encoding and decoding utilised the Politis Ambisonic library (Politis, 2016). All HRTFs, unless otherwise stated, are from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), diffuse-field equalised as in Section 2.8.1. All corresponding loudspeaker configurations, unless otherwise stated, are Lebedev arrangements as displayed in Figure 3.4. Figures D.1, D.2, D.3 and D.4 present the diffuse-field responses, inverse filters and resulting equalised frequency responses of binaural Ambisonic rendering with different pre-processing technique combinations, for Ambisonic orders $M = 2, M = 3, M = 4$ and $M = 5$, respectively. Figures D.5, D.6, D.7 and D.8 present the PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, for Ambisonic orders $M = 2, M = 3, M = 4$ and $M = 5$, respectively. Figure D.9 presents $\Delta$ITD values between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for $M = 2$ for every measurement



(A) PP 1: AIO & DFE

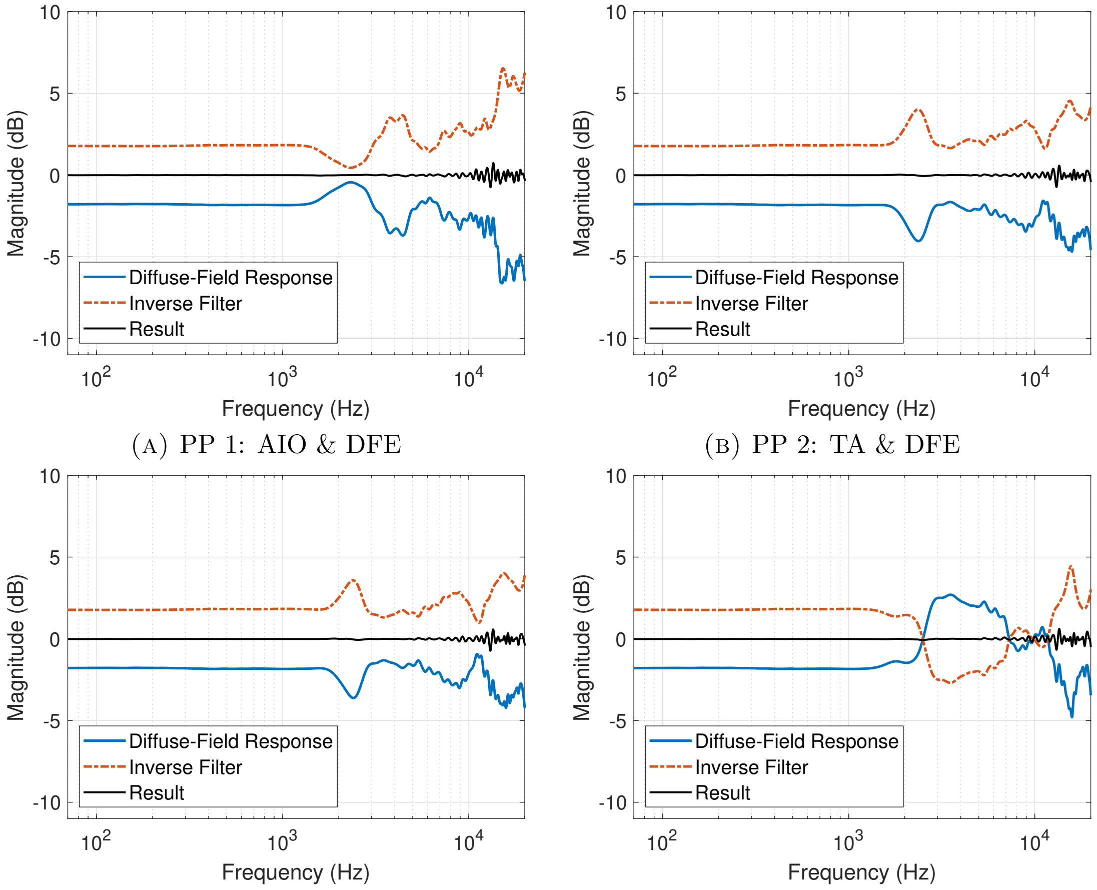
(B) PP 2: TA & DFE

(C) PP 3: TA & AIO & DFE (basic)

(D) PP 4: TA & AIO & DFE (dual-band)

FIGURE D.1: Diffuse-field response, inverse filters and resulting responses of different pre-processing technique combinations, $M = 2$ (left ear).

location on the sphere. Figures D.10, D.11, D.12, D.13 and D.14 present ΔILD values between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations for $\{M = 1, M = 2, ..., M = 5\}$, for every measurement location on the sphere. Figures D.15, D.16, D.17, D.18 and D.19 present the estimated $\theta$ values of binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \theta < +90°$ and $\phi = 0°$. Figures D.20, D.21, D.22, D.23 and D.24 present the estimated sagittal plane localisation plots of binaural Ambisonic rendering with different pre-processing technique combinations, for $\{M = 1, M = 2, ..., M = 5\}$ between $-90° < \phi < +90°$ and $\theta = 0°$.



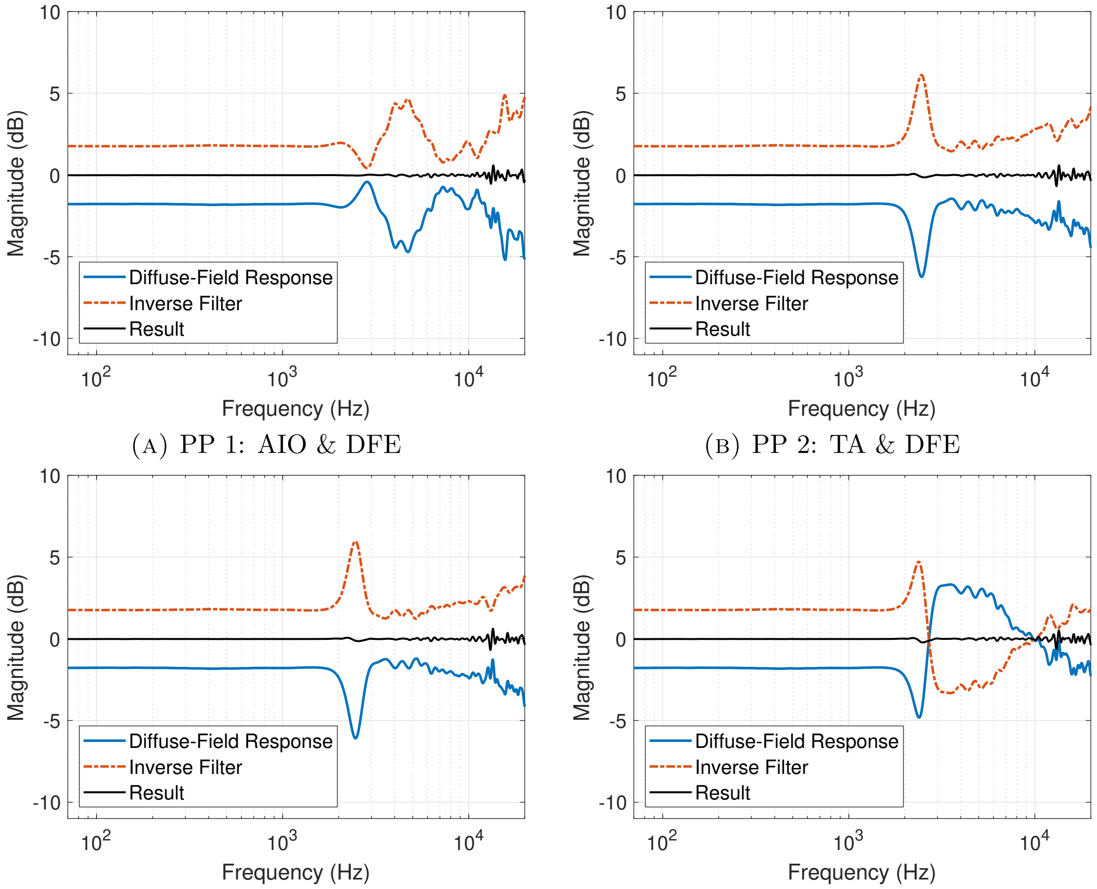
(A) PP 1: AIO & DFE

(B) PP 2: TA & DFE

(C) PP 3: TA & AIO & DFE (basic)

(D) PP 4: TA & AIO & DFE (dual-band)

FIGURE D.2: Diffuse-field response, inverse filters and resulting responses of different pre-processing technique combinations, $M = 3$ (left ear).



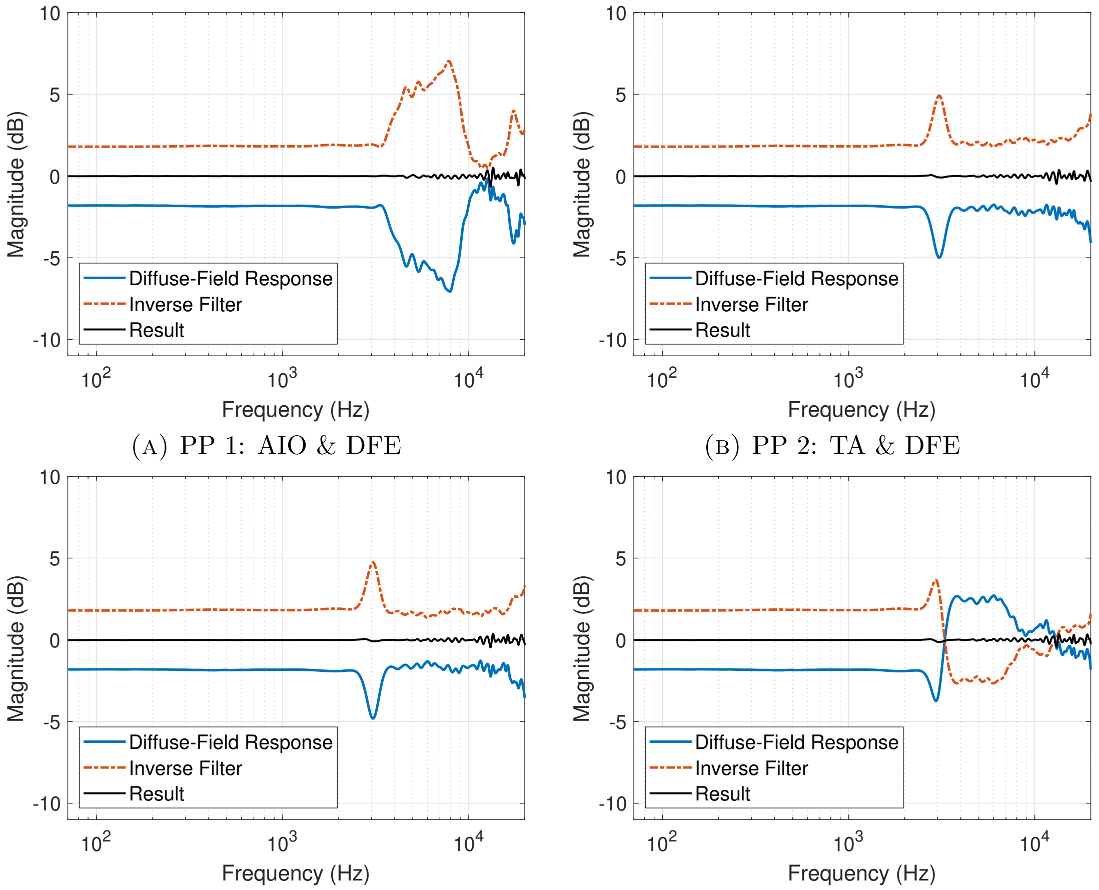
(A) PP 1: AIO & DFE

(B) PP 2: TA & DFE

(C) PP 3: TA & AIO & DFE (basic)

(D) PP 4: TA & AIO & DFE (dual-band)

FIGURE D.3: Diffuse-field response, inverse filters and resulting responses of different pre-processing technique combinations, $M = 4$ (left ear).



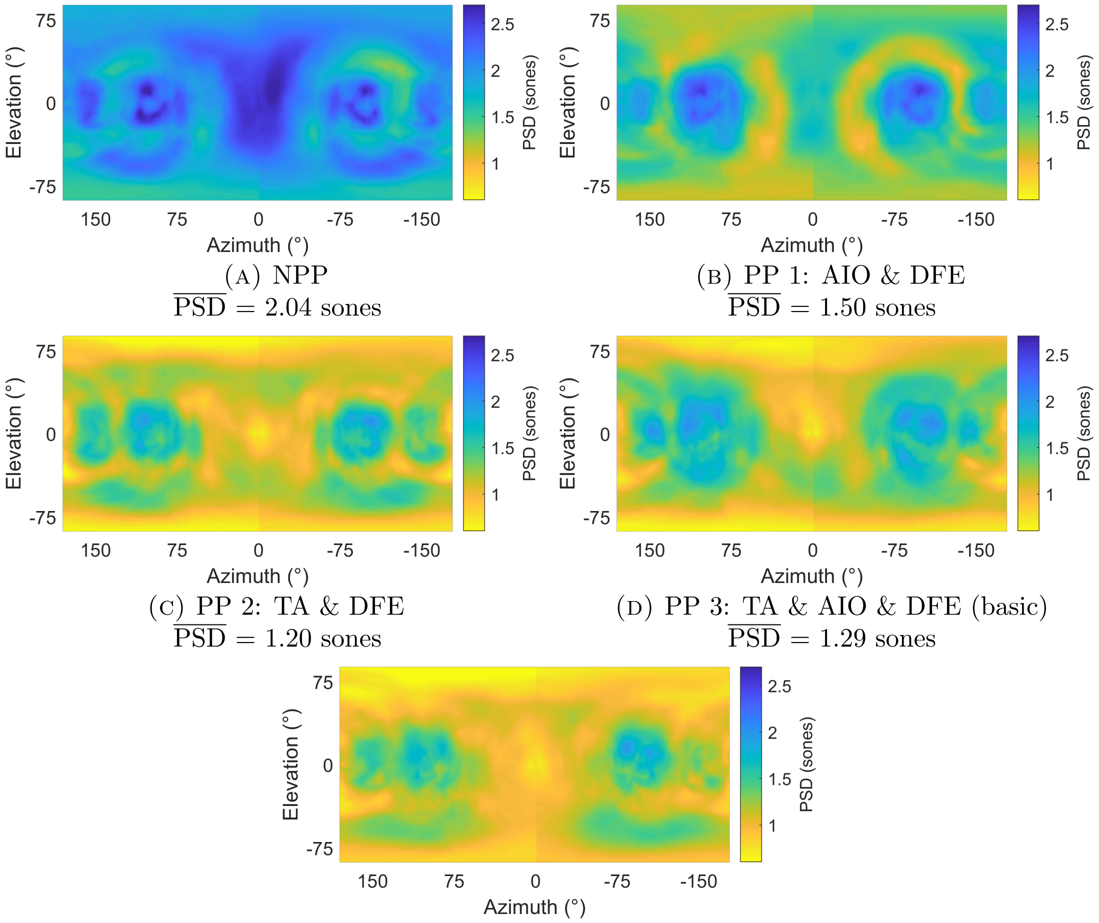
(A) PP 1: AIO & DFE

(B) PP 2: TA & DFE

(C) PP 3: TA & AIO & DFE (basic)

(D) PP 4: TA & AIO & DFE (dual-band)

FIGURE D.4: Diffuse-field response, inverse filters and resulting responses of different pre-processing technique combinations, $M = 5$ (left ear).



(A) NPP
$\overline{\text{PSD}} = 2.04$ sones

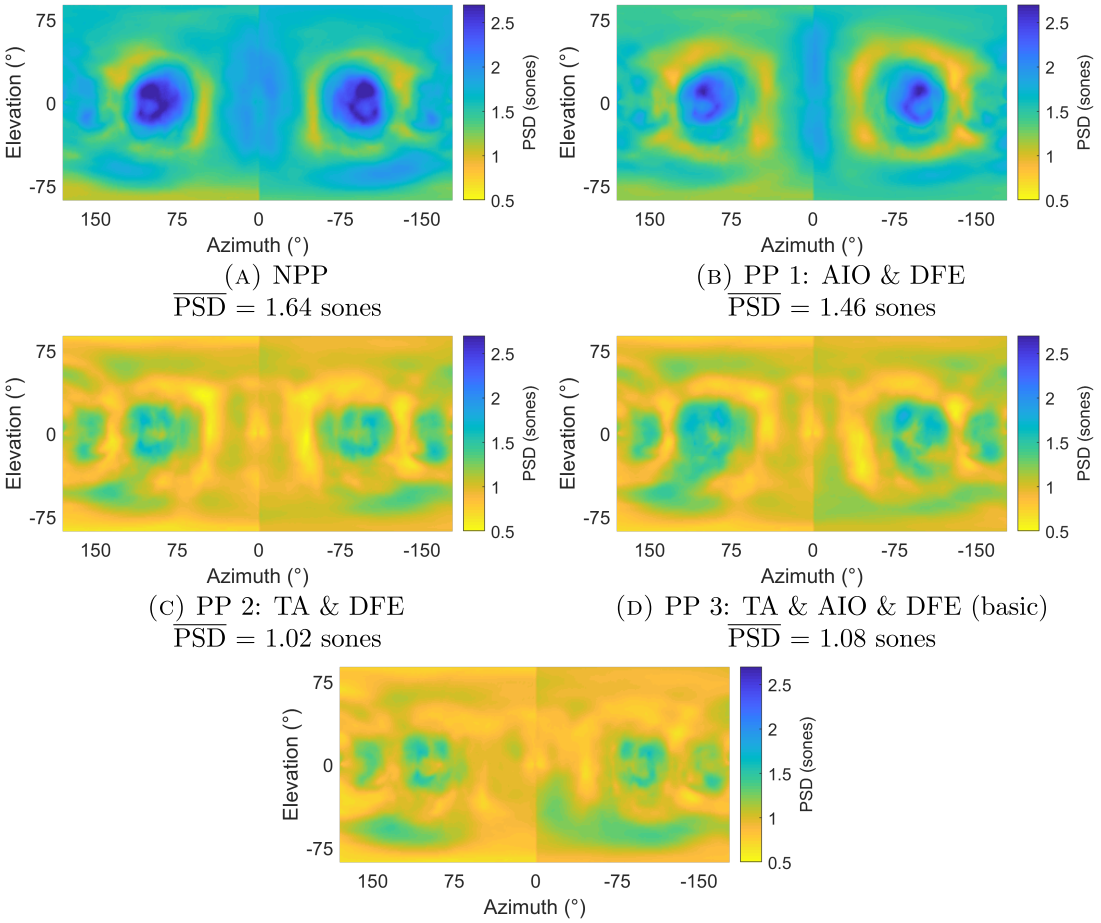
(B) PP 1: AIO & DFE
$\overline{\text{PSD}} = 1.50$ sones

(C) PP 2: TA & DFE
$\overline{\text{PSD}} = 1.20$ sones

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\text{PSD}} = 1.29$ sones

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\text{PSD}} = 1.13$ sones

FIGURE D.5: PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 2$ (mean of left and right PSD values). NPP denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.64$ sones

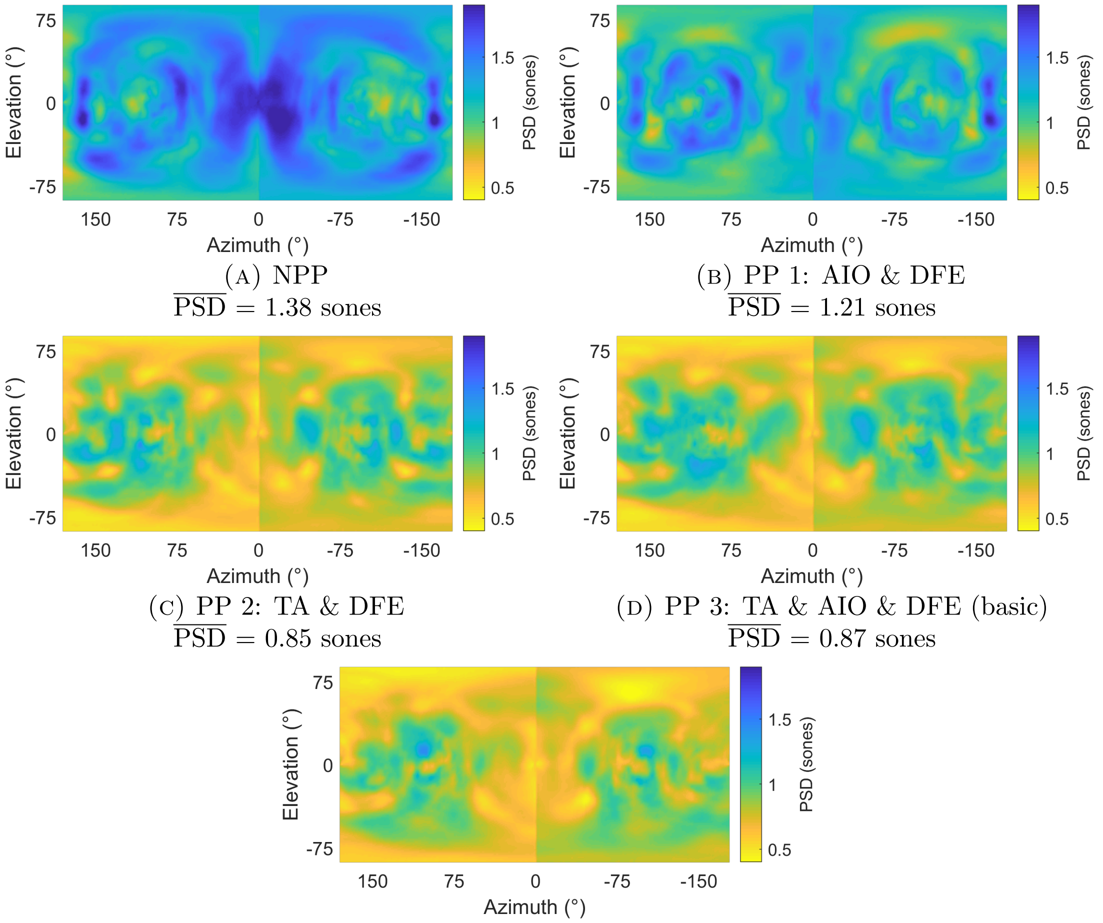
(B) PP 1: AIO & DFE
$\overline{\text{PSD}} = 1.46$ sones

(C) PP 2: TA & DFE
$\overline{\text{PSD}} = 1.02$ sones

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\text{PSD}} = 1.08$ sones

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\text{PSD}} = 1.00$ sones

FIGURE D.6: PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 3$ (mean of left and right PSD values). NPP denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.38$ sones

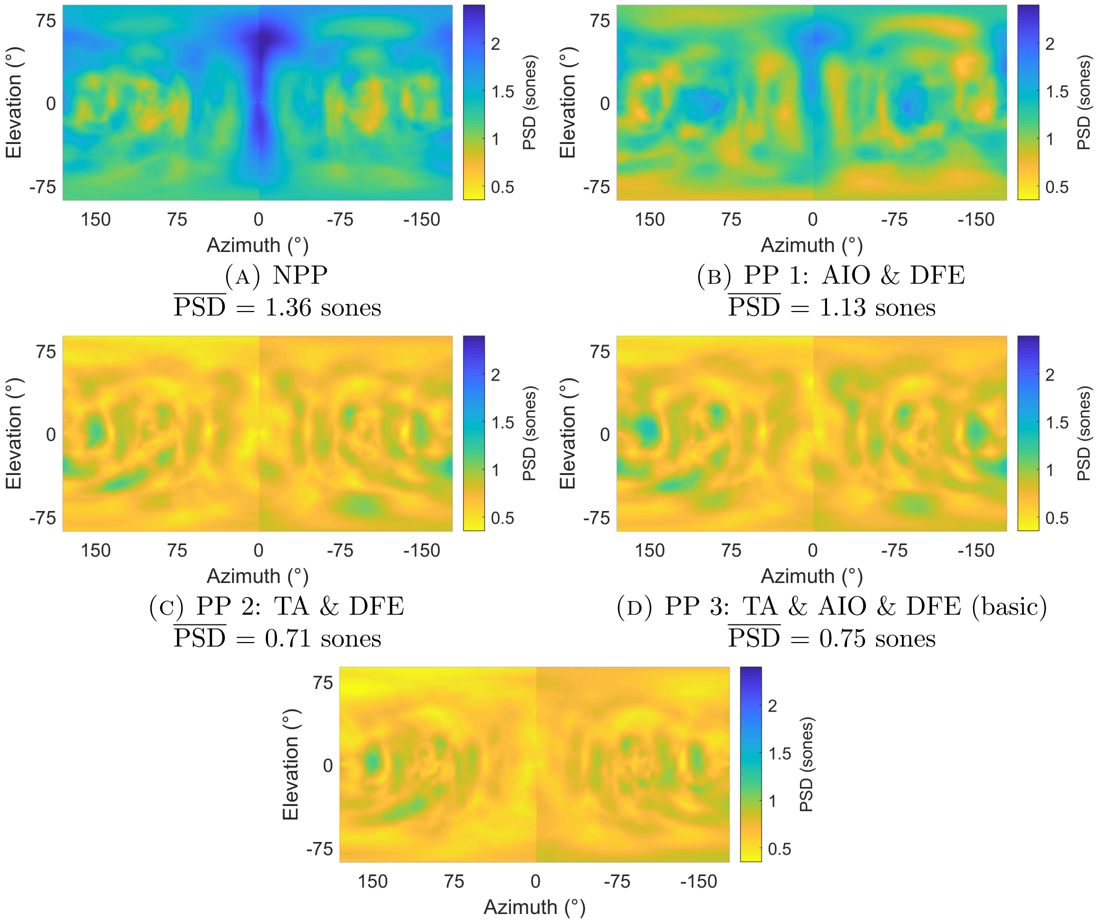
(B) PP 1: AIO & DFE
$\overline{\text{PSD}} = 1.21$ sones

(C) PP 2: TA & DFE
$\overline{\text{PSD}} = 0.85$ sones

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\text{PSD}} = 0.87$ sones

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\text{PSD}} = 0.80$ sones

FIGURE D.7: PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 4$ (mean of left and right PSD values). NPP denotes no pre-processing.



(A) NPP
$\overline{\text{PSD}} = 1.36$ sones

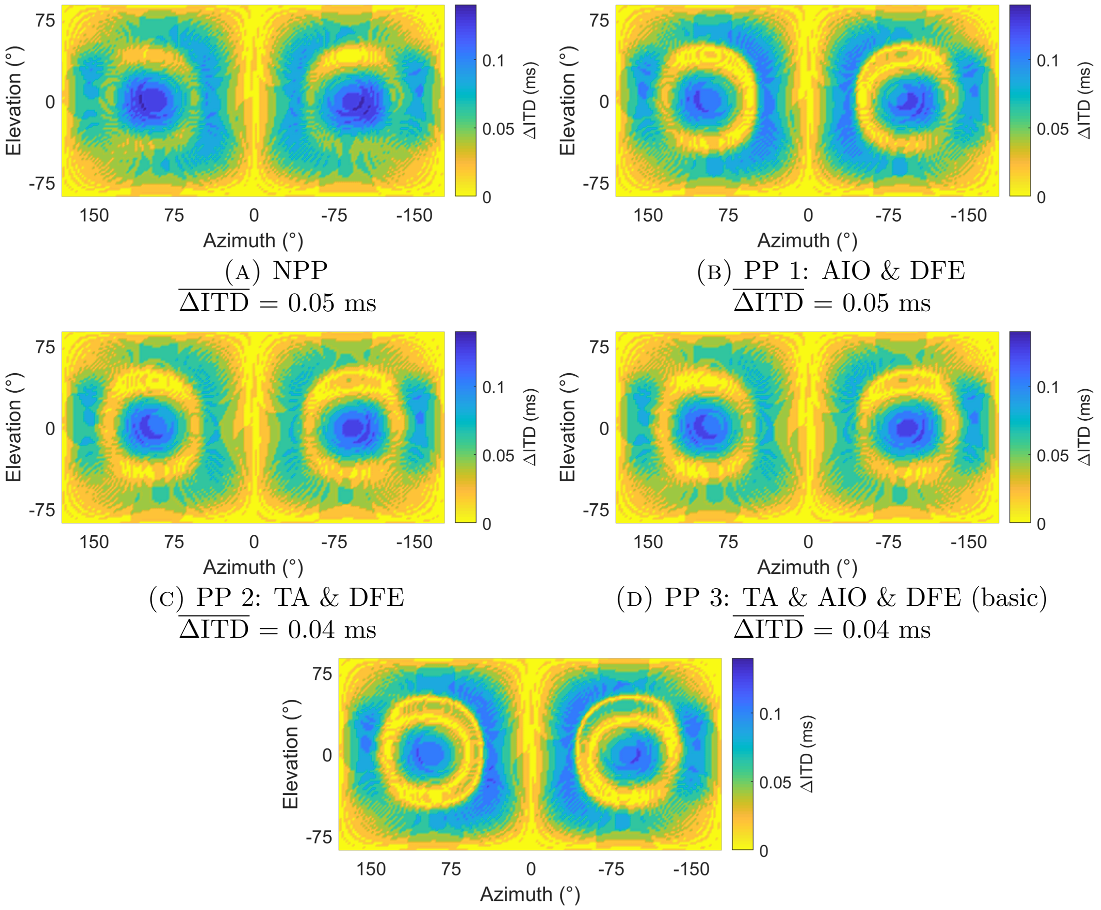
(B) PP 1: AIO & DFE
$\overline{\text{PSD}} = 1.13$ sones

(C) PP 2: TA & DFE
$\overline{\text{PSD}} = 0.71$ sones

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\text{PSD}} = 0.75$ sones

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\text{PSD}} = 0.69$ sones

FIGURE D.8: PSD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 5$ (mean of left and right PSD values). NPP denotes no pre-processing.



(A) NPP
$\overline{\Delta\text{ITD}} = 0.05$ ms

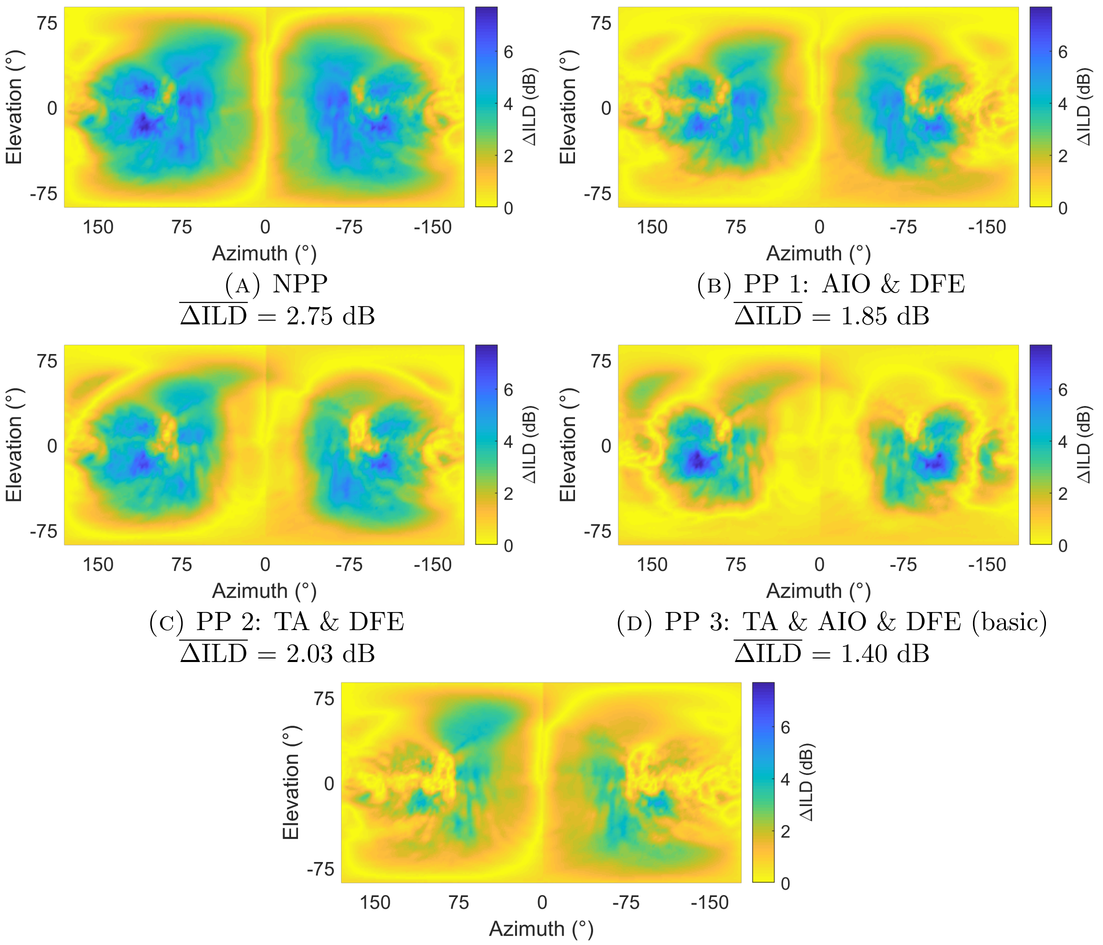
(B) PP 1: AIO & DFE
$\overline{\Delta\text{ITD}} = 0.05$ ms

(C) PP 2: TA & DFE
$\overline{\Delta\text{ITD}} = 0.04$ ms

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta\text{ITD}} = 0.04$ ms

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta\text{ITD}} = 0.05$ ms

FIGURE D.9: ΔITD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 2$. NPP denotes no pre-processing.



(A) NPP
$\overline{\Delta\text{ILD}} = 2.75$ dB

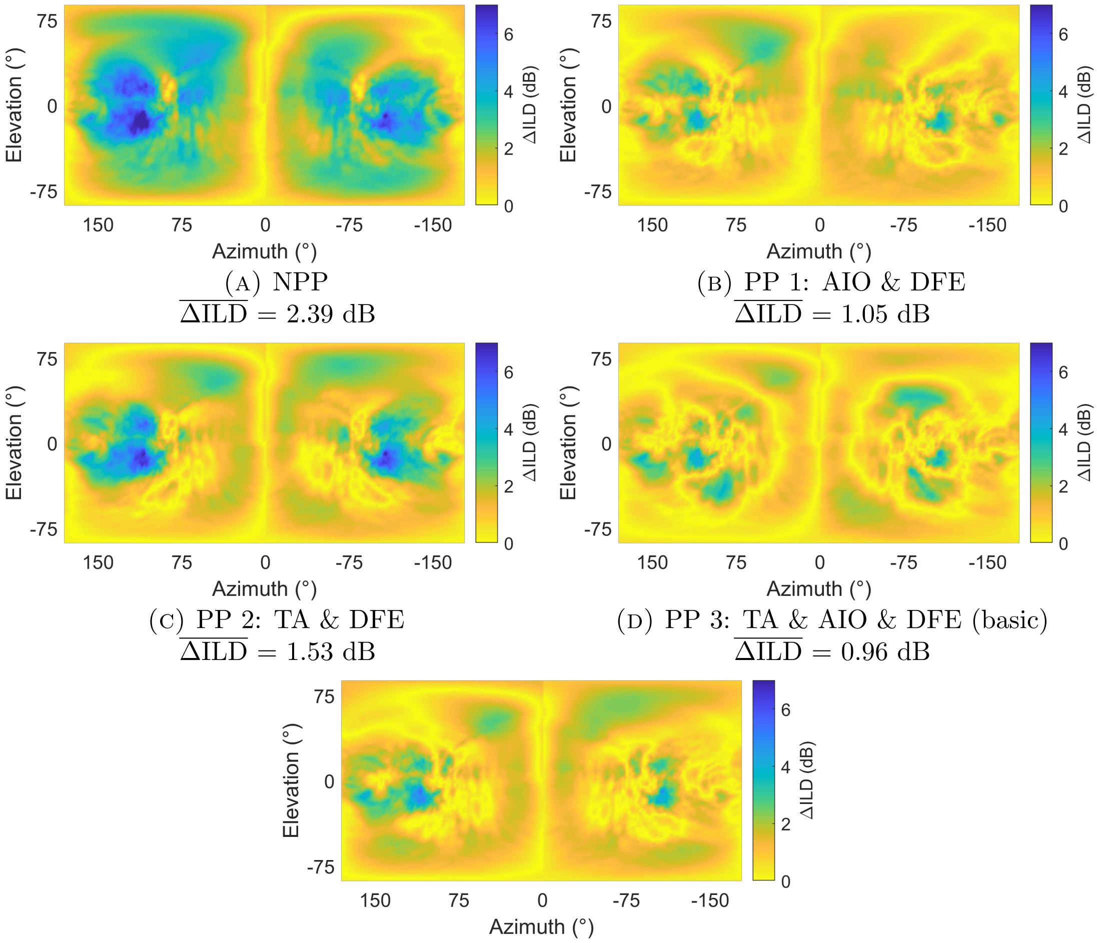
(B) PP 1: AIO & DFE
$\overline{\Delta\text{ILD}} = 1.85$ dB

(C) PP 2: TA & DFE
$\overline{\Delta\text{ILD}} = 2.03$ dB

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta\text{ILD}} = 1.40$ dB

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta\text{ILD}} = 1.46$ dB

FIGURE D.10: ΔILD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 1$.



(A) NPP
$\overline{\Delta\text{ILD}} = 2.39$ dB

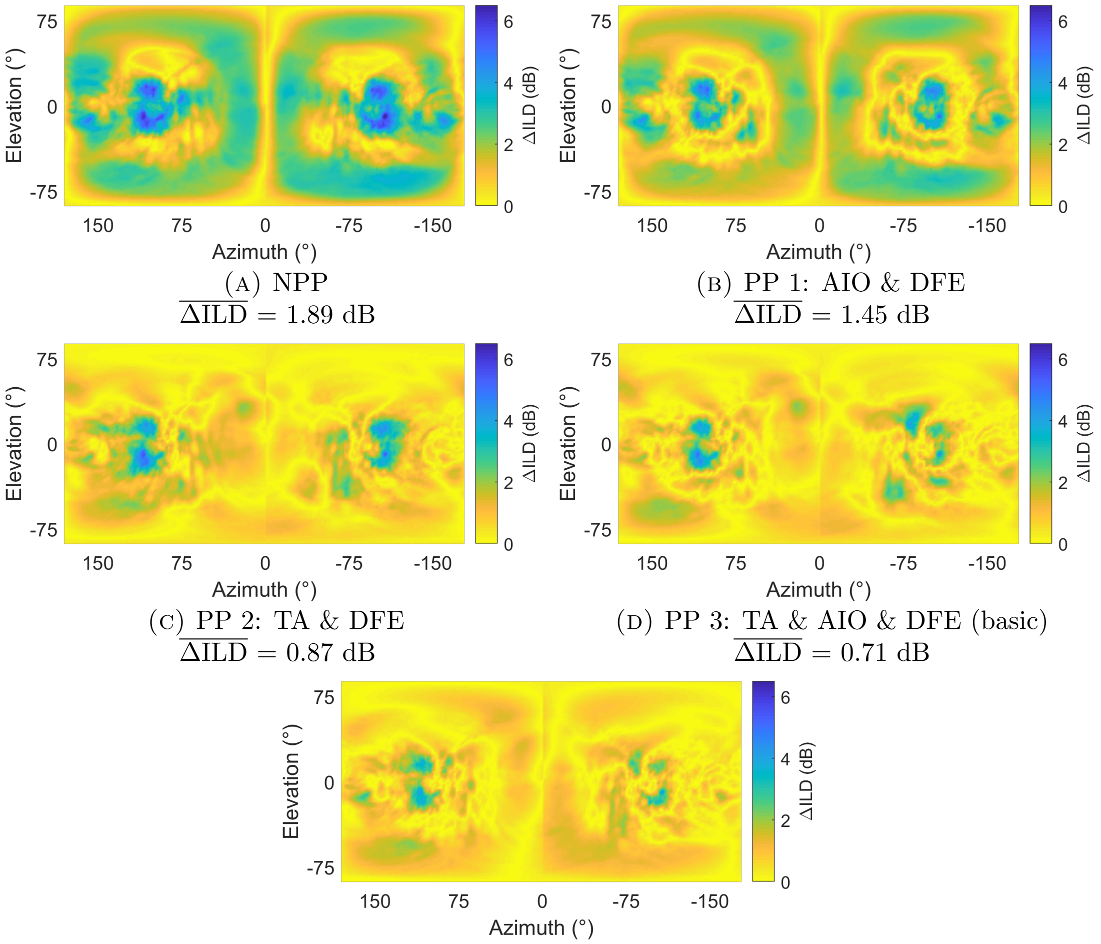
(B) PP 1: AIO & DFE
$\overline{\Delta\text{ILD}} = 1.05$ dB

(C) PP 2: TA & DFE
$\overline{\Delta\text{ILD}} = 1.53$ dB

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta\text{ILD}} = 0.96$ dB

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta\text{ILD}} = 1.12$ dB

FIGURE D.11: ΔILD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 2$.



(A) NPP
$\overline{\Delta\text{ILD}} = 1.89$ dB

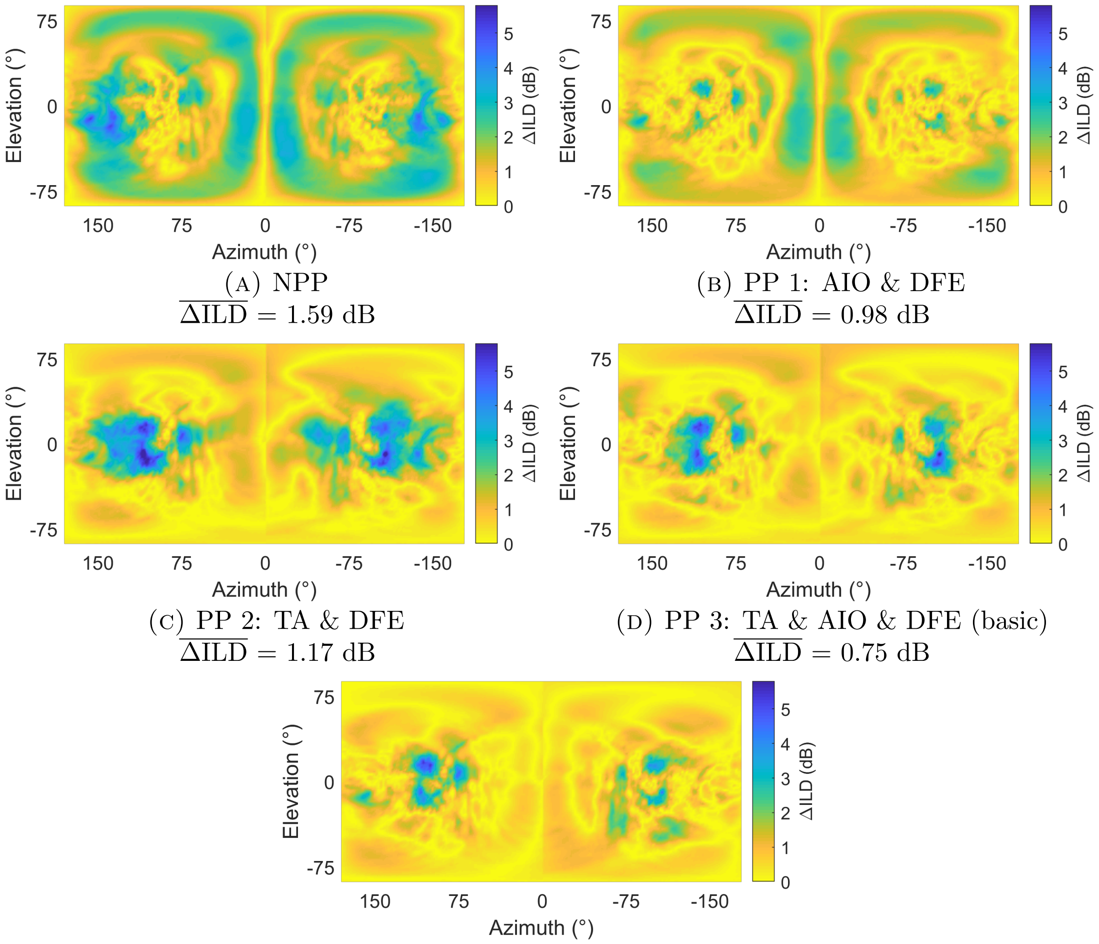
(B) PP 1: AIO & DFE
$\overline{\Delta\text{ILD}} = 1.45$ dB

(C) PP 2: TA & DFE
$\overline{\Delta\text{ILD}} = 0.87$ dB

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta\text{ILD}} = 0.71$ dB

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta\text{ILD}} = 0.70$ dB

FIGURE D.12: ΔILD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 3$.



(A) NPP
$\overline{\Delta\text{ILD}} = 1.59$ dB

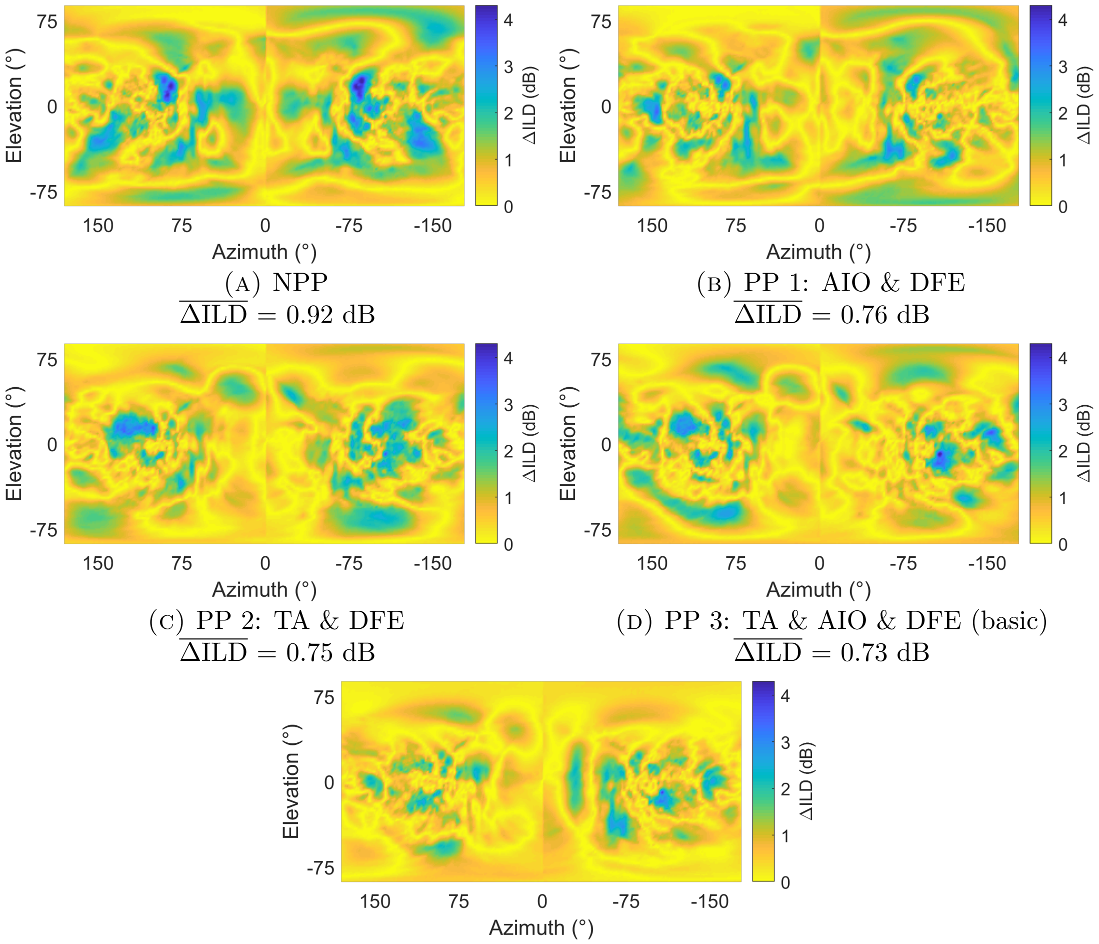
(B) PP 1: AIO & DFE
$\overline{\Delta\text{ILD}} = 0.98$ dB

(C) PP 2: TA & DFE
$\overline{\Delta\text{ILD}} = 1.17$ dB

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta\text{ILD}} = 0.75$ dB

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta\text{ILD}} = 0.71$ dB

FIGURE D.13: ΔILD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 4$.



(A) NPP
$\overline{\Delta \text{ILD}} = 0.92$ dB

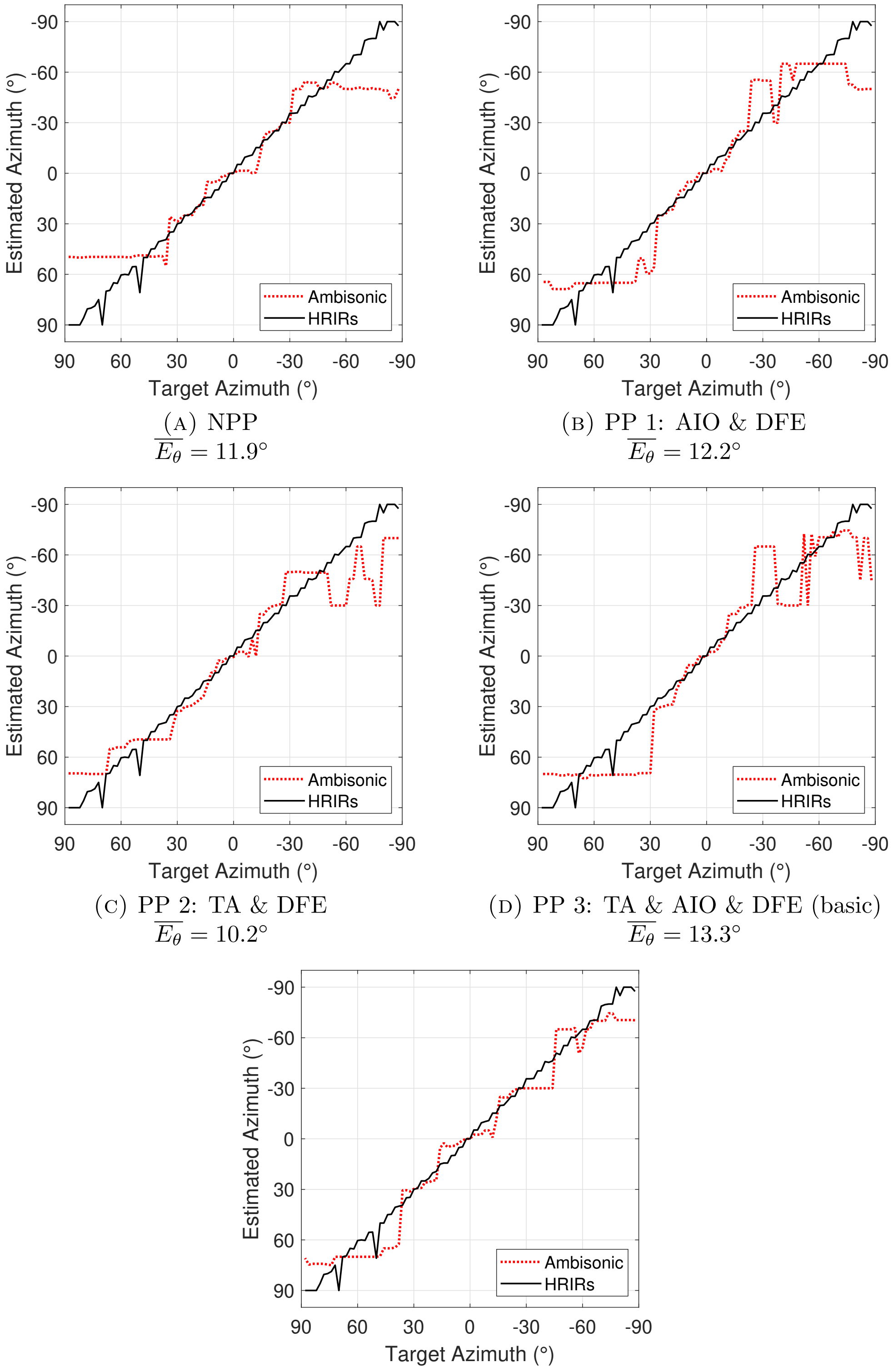
(B) PP 1: AIO & DFE
$\overline{\Delta \text{ILD}} = 0.76$ dB

(C) PP 2: TA & DFE
$\overline{\Delta \text{ILD}} = 0.75$ dB

(D) PP 3: TA & AIO & DFE (basic)
$\overline{\Delta \text{ILD}} = 0.73$ dB

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{\Delta \text{ILD}} = 0.63$ dB

FIGURE D.14: ΔILD between HRTFs and binaural Ambisonic rendering with different pre-processing technique combinations, for every measurement location on the sphere, $M = 5$.



(A) NPP
$\overline{E_\theta} = 11.9°$

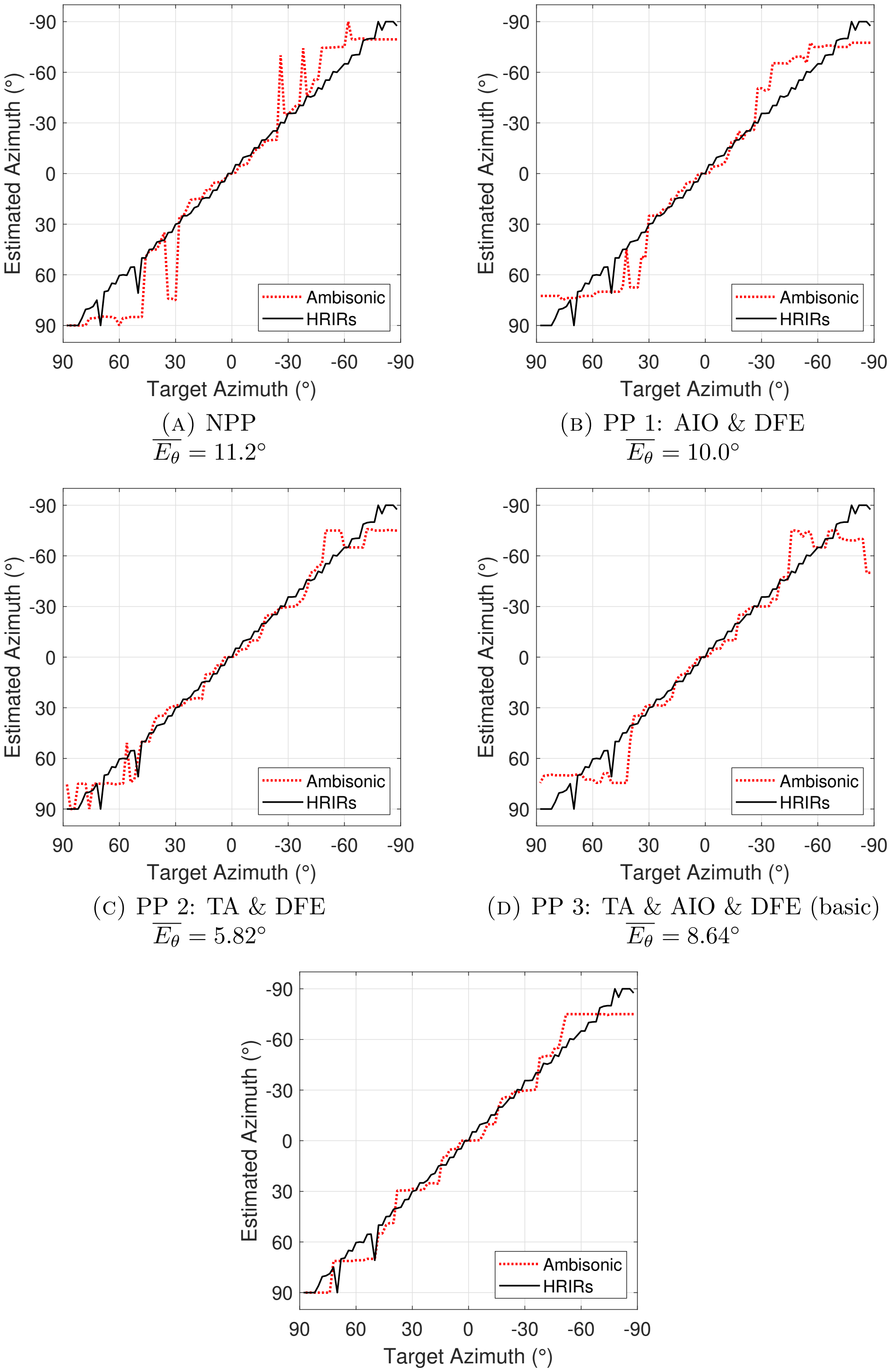
(B) PP 1: AIO & DFE
$\overline{E_\theta} = 12.2°$

(C) PP 2: TA & DFE
$\overline{E_\theta} = 10.2°$

(D) PP 3: TA & AIO & DFE (basic)
$\overline{E_\theta} = 13.3°$

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{E_\theta} = 7.75°$

FIGURE D.15: Horizontal localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 1$.



(A) NPP
$\overline{E_\theta} = 11.2°$

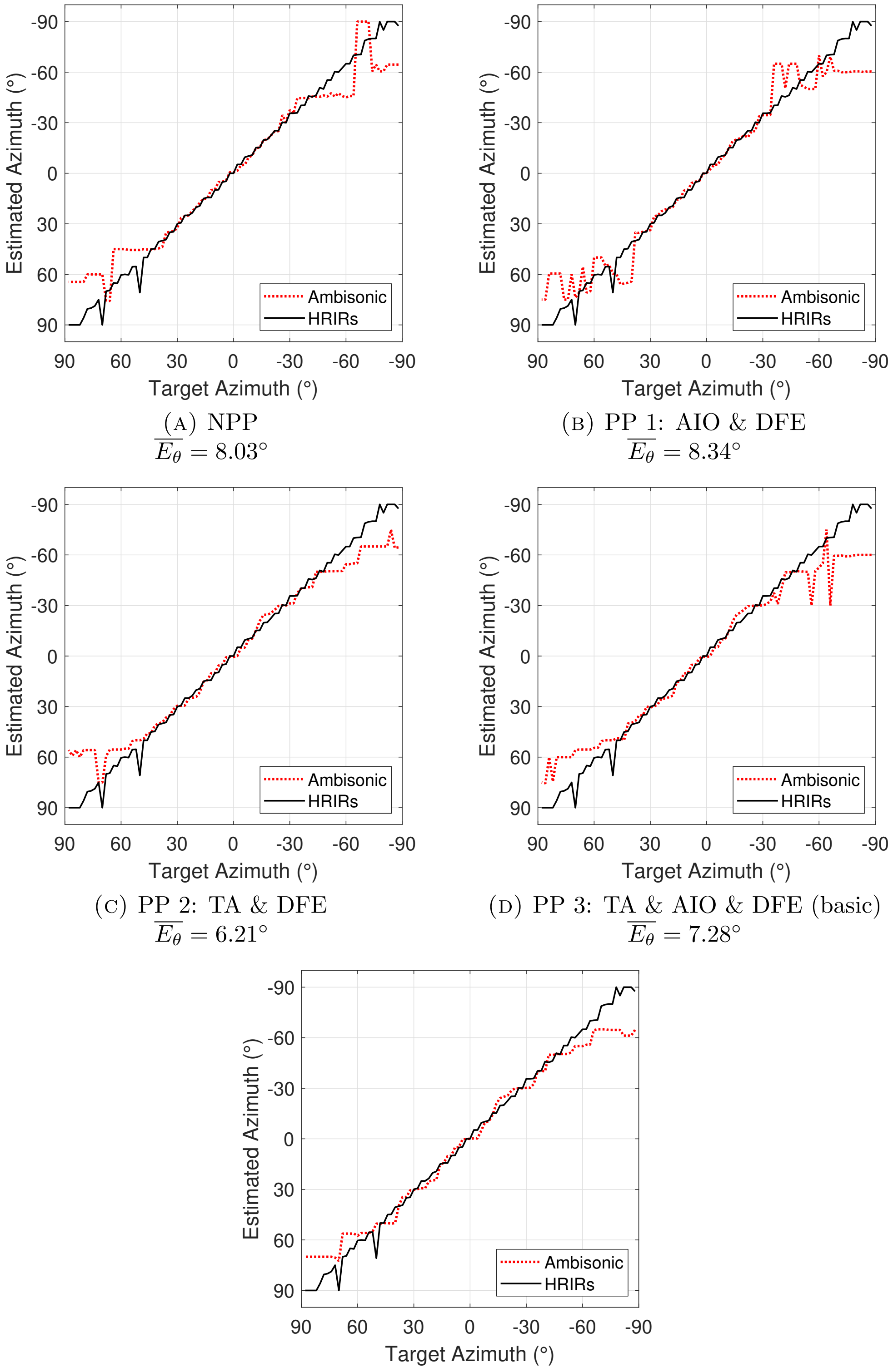
(B) PP 1: AIO & DFE
$\overline{E_\theta} = 10.0°$

(C) PP 2: TA & DFE
$\overline{E_\theta} = 5.82°$

(D) PP 3: TA & AIO & DFE (basic)
$\overline{E_\theta} = 8.64°$

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{E_\theta} = 7.10°$

FIGURE D.16: Horizontal localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 2$.



(A) NPP
$\overline{E_\theta} = 8.03°$

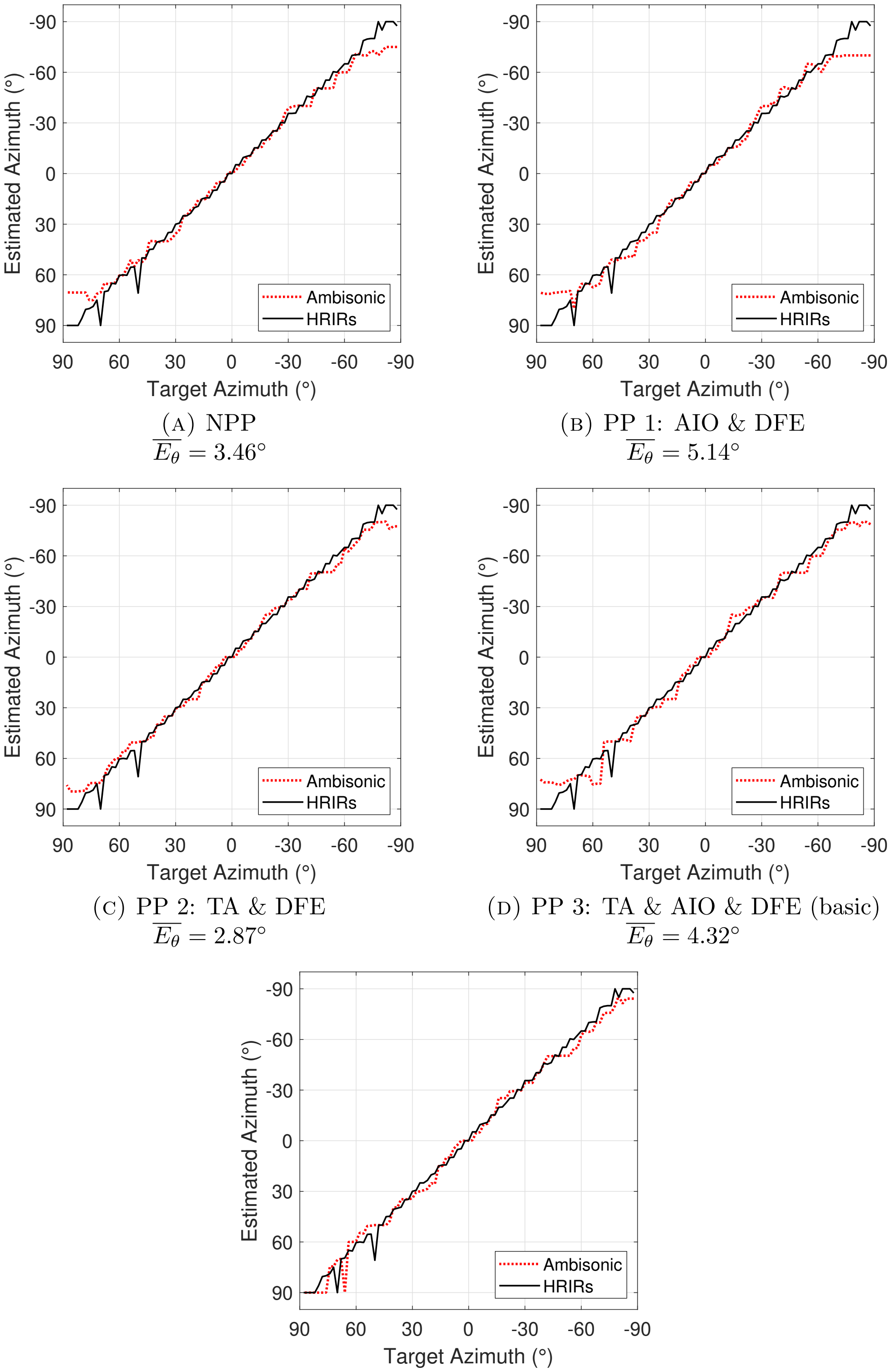
(B) PP 1: AIO & DFE
$\overline{E_\theta} = 8.34°$

(C) PP 2: TA & DFE
$\overline{E_\theta} = 6.21°$

(D) PP 3: TA & AIO & DFE (basic)
$\overline{E_\theta} = 7.28°$

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{E_\theta} = 5.42°$

FIGURE D.17: Horizontal localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 3$.



(A) NPP
$\overline{E_\theta} = 3.46°$

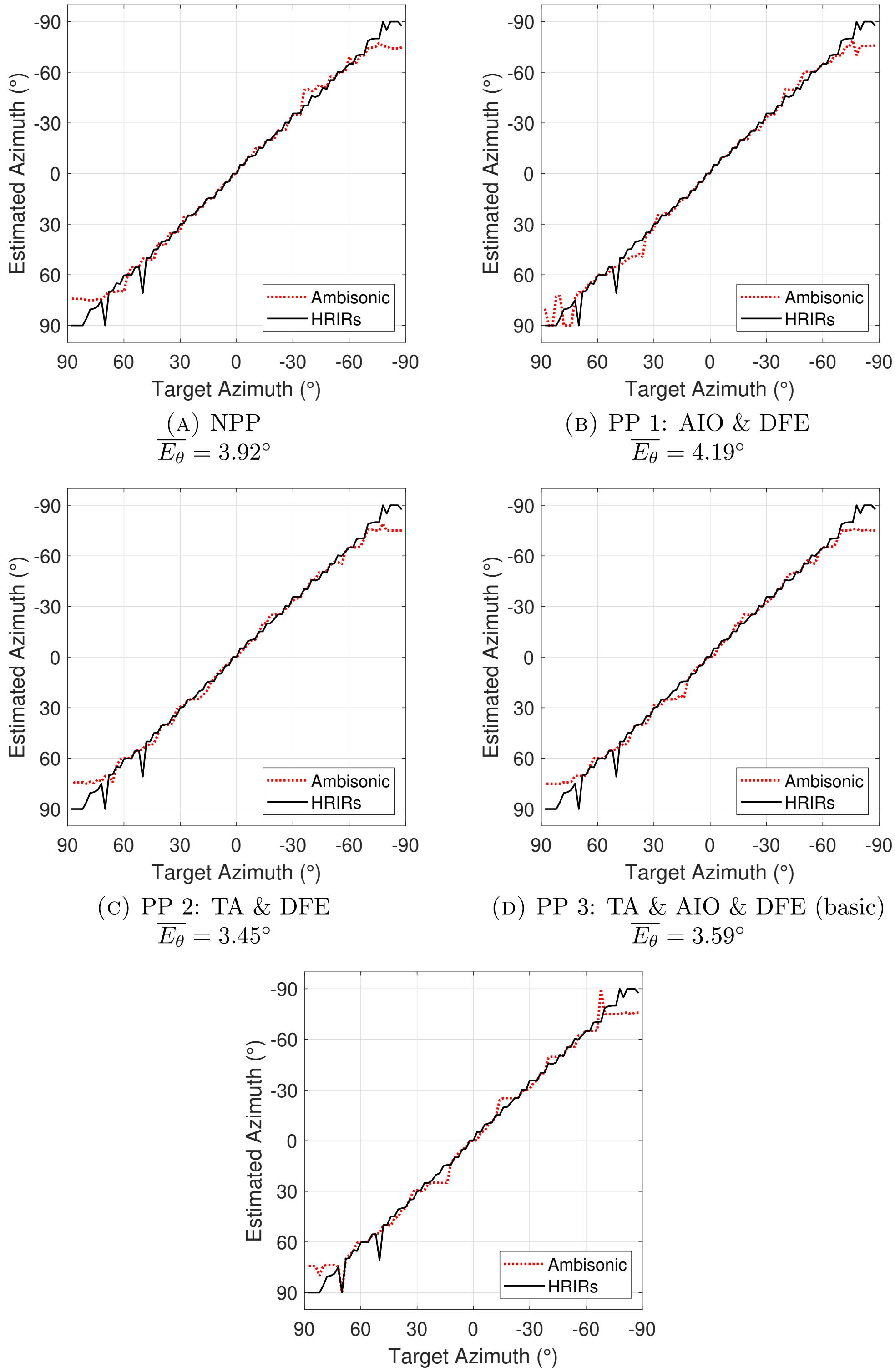
(B) PP 1: AIO & DFE
$\overline{E_\theta} = 5.14°$

(C) PP 2: TA & DFE
$\overline{E_\theta} = 2.87°$

(D) PP 3: TA & AIO & DFE (basic)
$\overline{E_\theta} = 4.32°$

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{E_\theta} = 3.44°$

FIGURE D.18: Horizontal localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 4$.



(A) NPP
$\overline{E_\theta} = 3.92°$

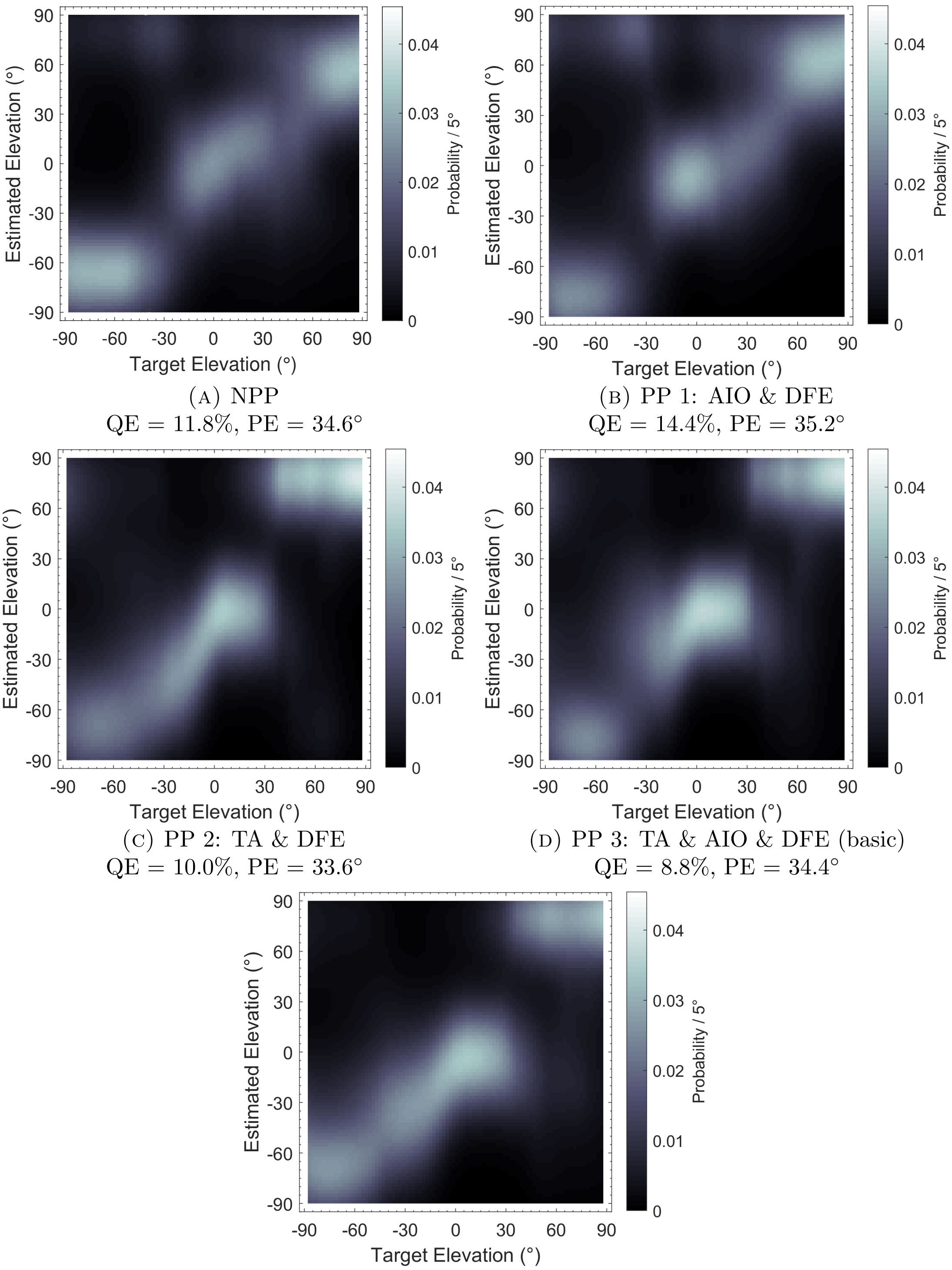
(B) PP 1: AIO & DFE
$\overline{E_\theta} = 4.19°$

(C) PP 2: TA & DFE
$\overline{E_\theta} = 3.45°$

(D) PP 3: TA & AIO & DFE (basic)
$\overline{E_\theta} = 3.59°$

(E) PP 4: TA & AIO & DFE (dual-band)
$\overline{E_\theta} = 4.18°$

FIGURE D.19: Horizontal localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 5$.



(A) NPP
QE = 11.8%, PE = 34.6°

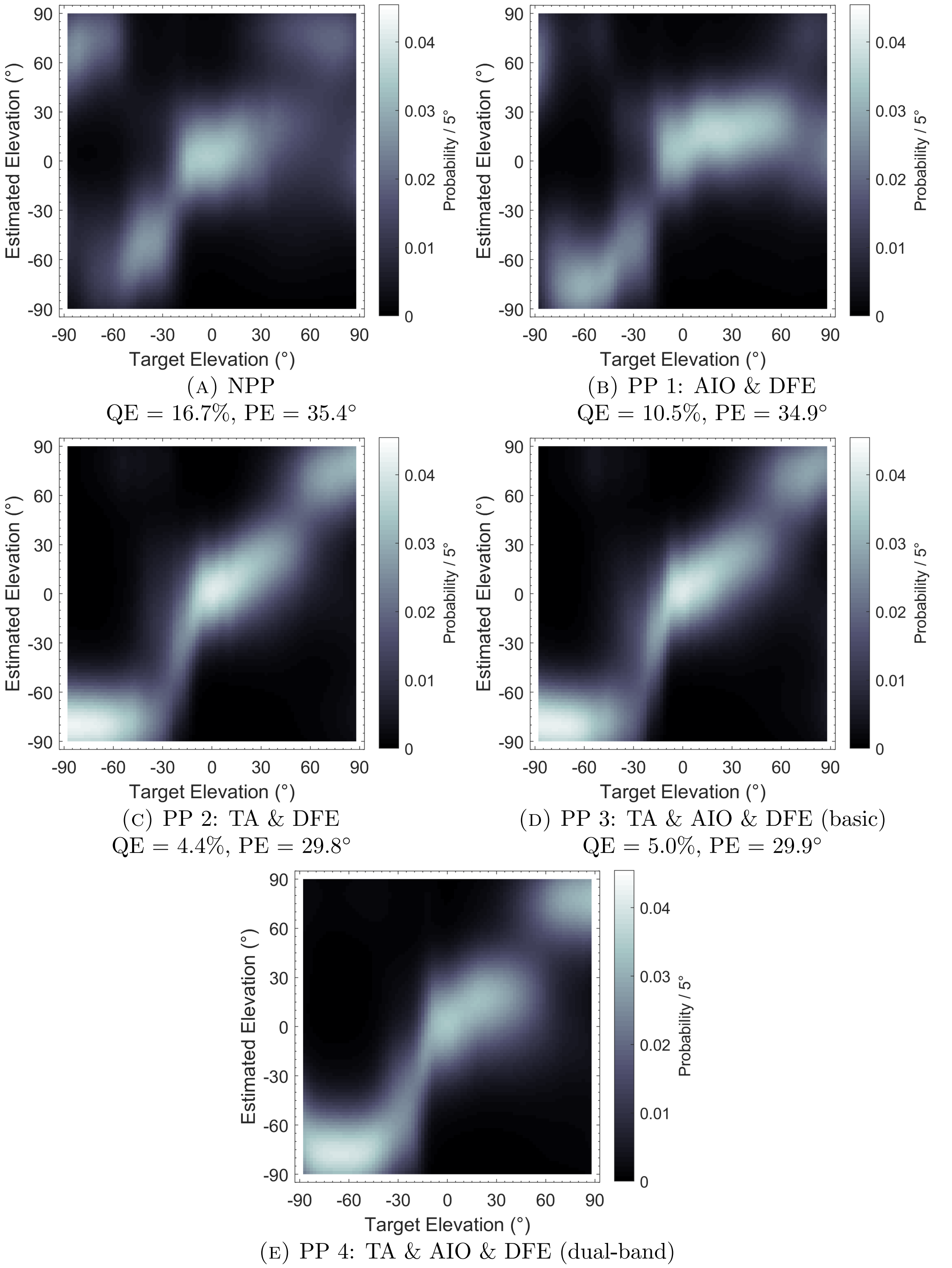
(B) PP 1: AIO & DFE
QE = 14.4%, PE = 35.2°

(C) PP 2: TA & DFE
QE = 10.0%, PE = 33.6°

(D) PP 3: TA & AIO & DFE (basic)
QE = 8.8%, PE = 34.4°

(E) PP 4: TA & AIO & DFE (dual-band)
QE = 7.3%, PE = 34.4°

FIGURE D.20: Sagittal plane localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 1$.



(A) NPP
QE = 16.7%, PE = 35.4°

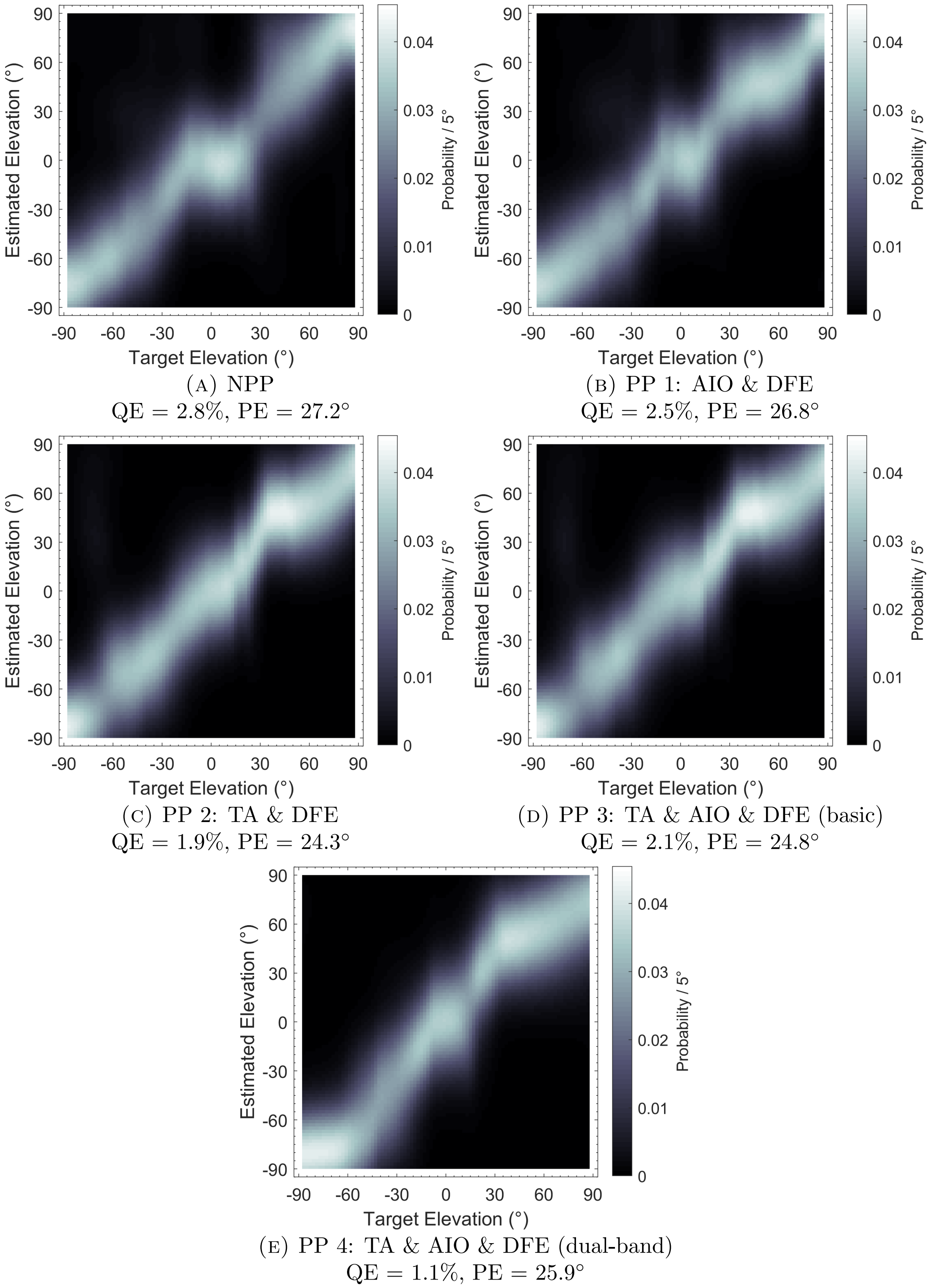
(B) PP 1: AIO & DFE
QE = 10.5%, PE = 34.9°

(C) PP 2: TA & DFE
QE = 4.4%, PE = 29.8°

(D) PP 3: TA & AIO & DFE (basic)
QE = 5.0%, PE = 29.9°

(E) PP 4: TA & AIO & DFE (dual-band)
QE = 3.7%, PE = 31.2°

FIGURE D.21: Sagittal plane localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 2$.



(A) NPP
QE = 2.8%, PE = 27.2°

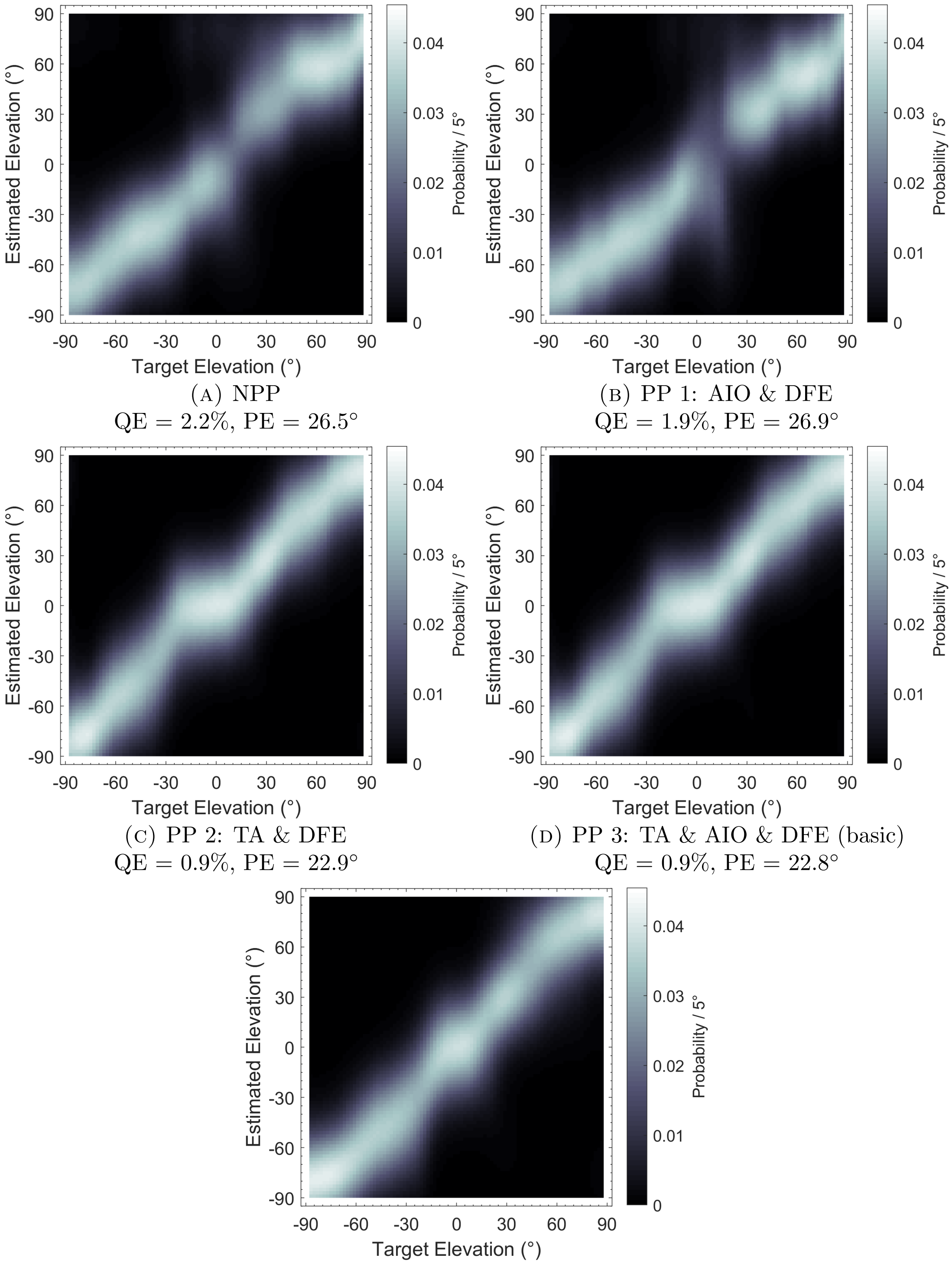
(B) PP 1: AIO & DFE
QE = 2.5%, PE = 26.8°

(C) PP 2: TA & DFE
QE = 1.9%, PE = 24.3°

(D) PP 3: TA & AIO & DFE (basic)
QE = 2.1%, PE = 24.8°

(E) PP 4: TA & AIO & DFE (dual-band)
QE = 1.1%, PE = 25.9°

FIGURE D.22: Sagittal plane localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 3$.



(A) NPP
QE = 2.2%, PE = 26.5°

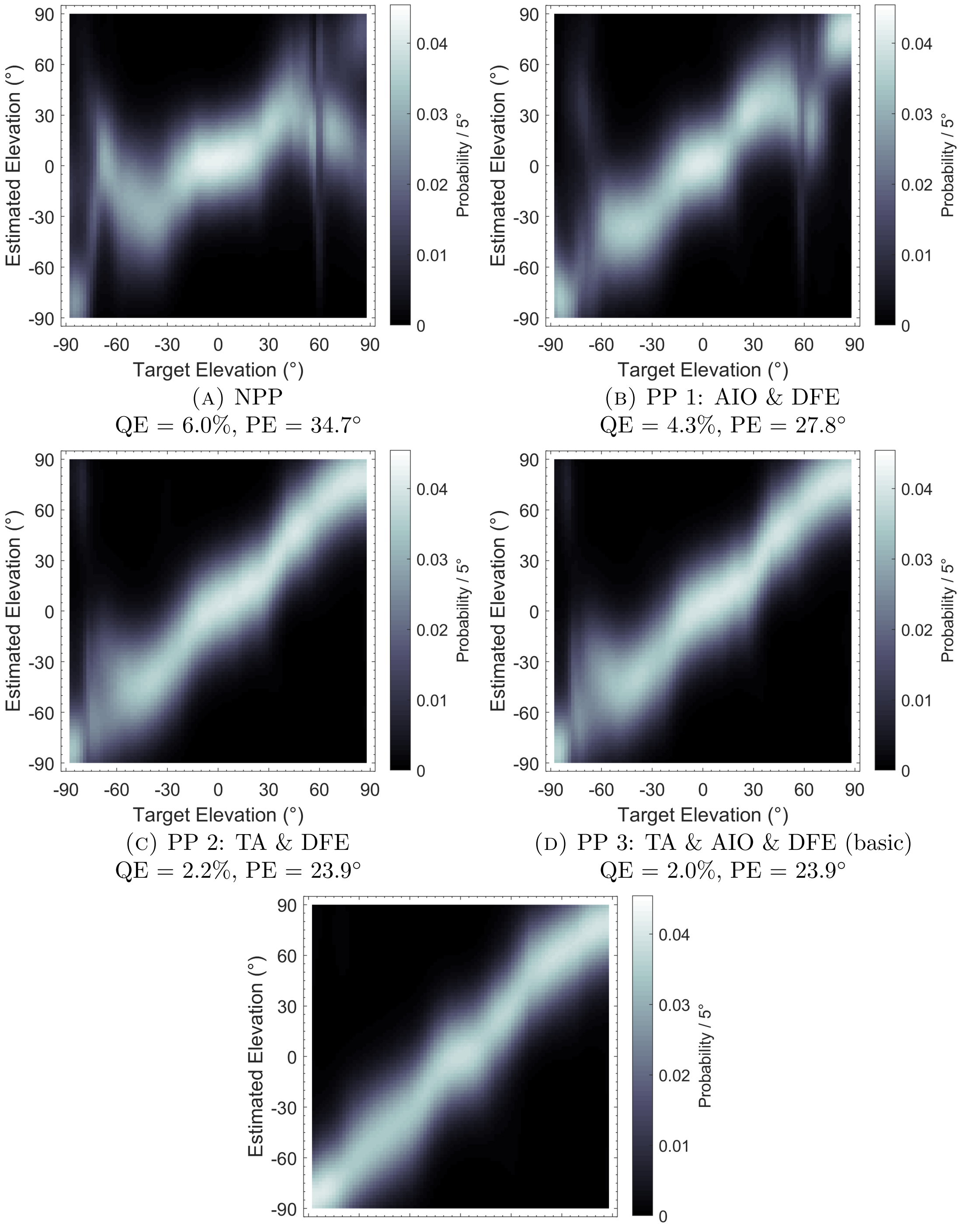
(B) PP 1: AIO & DFE
QE = 1.9%, PE = 26.9°

(C) PP 2: TA & DFE
QE = 0.9%, PE = 22.9°

(D) PP 3: TA & AIO & DFE (basic)
QE = 0.9%, PE = 22.8°

(E) PP 4: TA & AIO & DFE (dual-band)
QE = 1.0%, PE = 23.8°

Figure D.23: Sagittal plane localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 4$.



(A) NPP
QE = 6.0%, PE = 34.7°

(B) PP 1: AIO & DFE
QE = 4.3%, PE = 27.8°

(C) PP 2: TA & DFE
QE = 2.2%, PE = 23.9°

(D) PP 3: TA & AIO & DFE (basic)
QE = 2.0%, PE = 23.9°

(E) PP 4: TA & AIO & DFE (dual-band)
QE = 1.1%, PE = 23.1°

FIGURE D.24: Sagittal plane localisation model plots of binaural Ambisonic rendering with different pre-processing technique combinations, $M = 5$.

# Appendix E

# Index of Accompanying Materials

The accompanying materials folder is laid out as follows:

- Listening Test Documents
  - **Consent Form DFE:** Consent form for the listening test presented in Chapter 4.
  - **Consent Form DBE:** Consent form for the listening test presented in Chapter 5.
  - **Consent Form AIO:** Consent form for the listening test presented in Chapter 6.
  - **Consent Form Combinations:** Consent form for the listening test presented in Chapter 7.
  - **Information Sheet DFE:** Information sheet for the listening test presented in Chapter 4.
  - **Information Sheet DBE:** Information sheet for the listening test presented in Chapter 5.
  - **Information Sheet AIO:** Information sheet for the listening test presented in Chapter 6.

  - **Information Sheet Combinations:** Information sheet for the listening test presented in Chapter 7.

- Listening Test Stimuli
  - Diffuse-Field Equalisation
    - **MUSHRA:** Folder containing stimuli for the MUSHRA listening test in Chapter 4. Wav files are labelled as CONDITION _ AMBISONIC ORDER _ TEST SOUND AZIMUTH _ TEST SOUND ELEVATION _ TEST SOUND LOCATION ($\psi$), where condition refers to either standard Ambisonic (AMBI), Ambisonic with DFE (DFE), HRIR reference (HRIR), mid anchor (HRIR_MIDANCHOR) or low anchor (HRIR_LOWANCHOR).
    - **AB:** Folder containing stimuli for the AB test in Chapter 4. Wav files are labelled as CONDITION _ AMBISONIC ORDERS AND ARRANGEMENT _ TEST SOUND AZIMUTH _ TEST SOUND ELEVATION _ TEST SOUND LOCATION ($\psi$), where condition refers to either standard Ambisonic (AMBI) or Ambisonic with DFE (DFE).
  - Directional Bias Equalisation
    - **Simple:** Folder containing stimuli for the simple scene listening test in Chapter 5. Wav files are labelled as CONDITION _ AMBISONIC ORDER _ BIAS ($\kappa$) _ TEST SOUND AZIMUTH _ TEST SOUND ELEVATION, where condition refers to either standard Ambisonic (AMBI), Ambisonic with DBE (DBE), HRIR reference (HRIR), mid anchor (HRIR_MIDANCHOR) or low anchor (HRIR_LOWANCHOR).
    - **Complex:** Folder containing stimuli for the complex scene test in Chapter 5. Wav files are labelled as CONDITION _ AMBISONIC ORDER _ BIAS ($\kappa$), where condition refers to either standard Ambisonic (AMBI), Ambisonic with DBE (DBE), HRIR reference (HRIR), mid anchor (HRIR_MIDANCHOR) or low anchor (HRIR_LOWANCHOR).
  - ILD Optimisation

        - **Simple:** Folder containing stimuli for the simple scene listening test in Chapter 6. Wav files are labelled as Condition _ Ambisonic order _ test sound location ($\psi$) _ test sound azimuth _ test sound elevation, where condition refers to either standard Ambisonic (Ambi), Ambisonic with AIO (AIO), HRIR reference (HRIR), mid anchor (HRIR_MidAnchor) or low anchor (HRIR_LowAnchor).
        - **Complex:** Folder containing stimuli for the complex scene listening test in Chapter 6. Wav files are labelled as Condition _ Ambisonic order _ soundscape number _ soundscape description, where condition refers to either standard Ambisonic (Amb) or Ambisonic with AIO (AIO).
    - Combining Pre-Processing Techniques
        - **Moving Noise:** Folder containing stimuli for the moving noise listening test in Chapter 7. Wav files are labelled as Condition _ Ambisonic order _ pre-processing combinations, where condition refers to either standard Ambisonic (Amb), Ambisonic with pre-processing (AmP) or HRIR reference (HRIR).
        - **Percussion:** Folder containing stimuli for the percussion listening test in Chapter 7. Wav files are labelled as Condition _ Ambisonic order _ pre-processing combinations, where condition refers to either standard Ambisonic (Amb), Ambisonic with pre-processing (AmP) or HRIR reference (HRIR).
        - **Beach:** Folder containing stimuli for the beach listening test in Chapter 7. Wav files are labelled as Condition _ Ambisonic order _ pre-processing combinations, where condition refers to either standard Ambisonic (Amb) or Ambisonic with pre-processing (AmP).
- Matlab
    - **ambisonics:** Folder containing Matlab scripts for binaural Ambisonic rendering.

- **hrirs:** Folder containing 50 HRIRs from the Bernschütz Neumann KU 100 database (Bernschütz, 2013), corresponding to the measurements of the $L = 50$ Lebedev grid (as shown in Figure 3.4). These have been diffuse-field equalised as part of a larger dataset using the method detailed in Section 2.8.1.
- **pre_processing_techniques:** Folder containing functions for virtual loudspeaker HRTF pre-processing techniques presented in this thesis.
- **signal_processing:** Folder containing various functions used in the pre-processing techniques, such as inverse filtering, ILD and ITD estimation, and plotting techniques.
- **test_scripts:** Folder containing test scripts for producing binaural Ambisonic decoders with pre-processing, and subsequent scripts for testing the decoders. One tests the decoder in relation to HRTF rendering with numerical methods, and two produce sample binaural sounds.
- **voronoi_solid_angle:** Folder containing functions, scripts and data for generating points on the sphere, including spherical coordinate rotation and directional biasing.
- **readme.txt:** A readme file detailing the contents of the Matlab folder. This includes details and links to the materials included that have been developed by others.

# List of Acronyms

| Acronym | Description |
|---|---|
| ACN | Ambisonic channel numbering |
| AFC | alternative forced choice |
| AIO | Ambisonic ILD Optimisation |
| AllRAD | All Round Ambisonic Panning |
| ANOVA | Analysis of Variance |
| BEM | boundary element method |
| BRIR | binaural room impulse response |
| BSD | basic spectral difference |
| BW | bandwidth |
| CI | confidence interval |
| CLL | Composite Loudness Level |
| dB | decibel |
| DBE | Directional Bias Equalisation |
| DBQ | directionally biased quadrature |
| DFE | Diffuse-Field Equalisation |
| DirAC | Directional Audio Coding |
| ERB | equivalent rectangular bandwidth |
| FEM | finite element method |
| FFT | fast Fourier transform |
| FIR | finite impulse response |
| FOA | first-order Ambisonics |
| IACC | interaural cross-correlation |

| | |
|---|---|
| ILD | interaural level difference |
| IR | impulse response |
| ISO | International Standards Organisation |
| ITD | interaural time difference |
| HRIR | head-related impulse response |
| HpTF | headphone transfer function |
| HRTF | head-related transfer function |
| MUSHRA | multiple stimulus test with hidden reference and anchor |
| N3D | three-dimensional full normalised |
| NPP | no pre-processing |
| PE | polar error |
| PP | pre-processing |
| PSD | perceptual spectral difference |
| QE | quadrant error |
| RMS | root-mean-square |
| RT | reverberation time |
| SAQI | Spatial Audio Quality Inventory |
| SH | spherical harmonic |
| SN3D | Schmidt semi-normalised |
| SPL | sound pressure level |
| TA | Time Alignment |
| VBAP | Vector Base Amplitude Panning |
| VR | virtual reality |

# List of Symbols

| Symbol | Description | Unit |
|---|---|---|
| $\beta$ | Ambisonic format signal | |
| $\delta(t)$ | time domain impulse | |
| $\delta_{n,0}$ | Kronecker delta function | |
| $\theta$ | azimuth | $^\circ$ |
| $\theta_l$ | loudspeaker azimuth | $^\circ$ |
| $\theta_{\text{est}}$ | estimated azimuth | $^\circ$ |
| $\phi$ | elevation | $^\circ$ |
| $\rho$ | percentile | |
| $\iota$ | virtual source location | |
| $\varrho$ | total number of virtual source locations | |
| $\kappa$ | amount of bias | |
| $\lambda$ | wavelength | m |
| $\pi$ | Pi | |
| $\psi$ | test sound location | |
| $\Omega$ | solid angle | |
| $\varphi$ | pressure | Pa |
| $\tau$ | time delay | s |
| $\sigma$ | spin | |
| $\zeta$ | zoom | |
| | | |
| $B$ | binaural signal | |
| $c$ | speed of sound | m/s |

| | | |
|---|---|---|
| $\mathbf{C}$ | re-encoding matrix | |
| $d$ | directivity | |
| $\mathbf{D}$ | decoding matrix | |
| $E$ | error | |
| $E_\theta$ | azimuth error | |
| $f$ | frequency | Hz |
| $f_{\text{alias}}$ | spatial aliasing frequency | Hz |
| $f_c$ | cut off frequency | Hz |
| $f_s$ | sampling frequency | Hz |
| $g$ | gain | |
| $g_{\text{bias}}$ | directional bias gain | |
| $g_m$ | spherical harmonic order dependent gain | |
| $g^{\Delta}$ | ILD difference gain | |
| $\mathbf{G}$ | gain matrix | |
| $H$ | head-related transfer function | |
| $H_{\text{bias}}$ | directional bias head-related transfer function | |
| $H_{\text{diff}}$ | diffuse-field head-related transfer function | |
| $H_{\text{RMS}}$ | RMS average head-related transfer function | |
| $\mathbf{I}$ | identity matrix | |
| $k$ | Ambisonic channel | |
| $K$ | total number of Ambisonic channels | |
| $l$ | Ambisonic loudspeaker | |
| $L$ | total number of Ambisonic loudspeakers | |
| $m$ | SH order | |
| $M$ | Ambisonic order | |
| $n$ | SH degree | |
| $N$ | normalisation | |
| $p$ | significance | |
| $P$ | Legendre function | |
| $q$ | measurement number | |
| $Q$ | total measurements | |

| | | |
|---|---|---|
| $r$ | distance | m |
| $\mathbf{r}_{\mathbf{E}}$ | energy vector | |
| $\mathbf{r}_{\mathbf{V}}$ | velocity vector | |
| $R$ | correlation | |
| $\mathbf{R}$ | rotation matrix | |
| $s$ | signal | |
| $t$ | time | s |
| ${}^{T}$ | transposition | |
| $T$ | order of T-design | |
| $x$ | Cartesian coordinate | |
| $y$ | Cartesian coordinate | |
| $Y$ | spherical harmonic | |
| $z$ | Cartesian coordinate | |

# Bibliography

Abhayapala, Thushara D. and Darren B. Ward (2002). "Theory and design of high order sound field microphones using spherical microphone array". In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* Orlando, pp. 1949–1952. DOI: 10.1109/ICASSP.2002.5745011.

Abramowitz, Milton and Irene Stegun (1972). *Handbook of Mathematical Functions.* 10th ed. Washington, D.C.: Dover Publications.

Acoustics, mh (2013). *Eigenmike.* URL: https://mhacoustics.com/products#eigenmike1 (visited on 09/23/2019).

Ahrens, Jens and Carl Andersson (2019). "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre". In: *Journal of the Acoustical Society of America* 145.4, pp. 2783–2794. DOI: 10.1121/1.5096164.

Algazi, V. and Richard Duda (2011). "Headphone-based spatial sound". In: *IEEE Signal Processing Magazine* 28.1, pp. 33–42. DOI: 10.1109/MSP.2010.938756.

Algazi, V. Ralph, Carlos Avendano and Richard O. Duda (2001a). "Elevation localization and head-related transfer function analysis at low frequencies". In: *Journal of the Acoustical Society of America* 109.3, pp. 1110–1122. DOI: 10.1121/1.1349185.

– (2001b). "Estimation of a spherical-head model from anthropometry". In: *Journal of the Audio Engineering Society* 49.6, pp. 472–479.

Algazi, V Ralph et al. (2002). "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *Journal of the Acoustical*

*Society of America* 112.5, pt. 1, pp. 2053–2064. DOI: http://dx.doi.org/10.1121/1.1508780.

Algazi, V.R. et al. (2001). "The CIPIC HRTF database". In: *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics.* October. New Paltz, NY, pp. 99–102. DOI: 10.1109/ASPAA.2001.969552.

Alon, David Lou et al. (2018). "Sparse head-related transfer function representation with spatial aliasing cancellation". In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* Vol. 2018-April. April. IEEE, pp. 6792–6796. DOI: 10.1109/ICASSP.2018.8462101.

American National Standards Institute (1994). "ANSI S1.1-1994". In: *American National Standard Acoustical Terminology.*

– (2001). "ANSI S1.42-2001". In: *Design Response of Weighting Networks for Acoustical Measurements.*

Andreopoulou, Areti and Brian F. G. Katz (2017). "Identification of perceptually relevant methods of inter-aural time difference estimation". In: *Journal of the Acoustical Society of America* 142.2, pp. 588–598. DOI: 10.1121/1.4996457.

Armstrong, Cal, Damian Murphy and Gavin Kearney (2018). "A Bi-RADIAL approach to Ambisonics". In: *AES International Conference on Audio for Virtual and Augmented Reality.* Redmond.

Armstrong, Cal et al. (2018a). "A perceptual evaluation of individual and non-individual HRTFs: a case study of the SADIE II database". In: *Applied Sciences* 8.11. DOI: 10.3390/app8112029.

Armstrong, Cal et al. (2018b). "A perceptual spectral difference model for binaural signals". In: *145th Convention of the Audio Engineering Society.* New York, E–Brief 457.

Asano, F, Y Suzuki and T Sone (1990). "Role of spectral cues in median plane localization". In: *Journal of the Acoustical Society of America* 88.1, pp. 159–168. DOI: 10.1121/1.399963.

Avendano, Carlos, V. Ralph Algazi and Richard O. Duda (1999). "A head-and-torso model for low-frequency binaural elevation effects". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* New York, pp. 179–182.

Avni, Amir et al. (2013). “Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution”. In: *Journal of the Acoustical Society of America* 133.5, pp. 2711–2721. DOI: 10.1121/1.4795780.

Balmages, Ilya and Boaz Rafaely (2007). “Open-sphere designs for spherical microphone arrays”. In: *IEEE Transactions on Audio, Speech and Language Processing* 15.2, pp. 727–732. DOI: 10.1109/TASL.2006.881671.

Bamford, Jeffery S and John Vanderkooy (1995). “Ambisonic sound for us”. In: *99th Convention of the Audio Engineering Society*. New York, Preprint 4138.

Bates, Enda et al. (2017). “Comparing Ambisonic microphones – part 2”. In: *142nd Convention of the Audio Engineering Society*. Berlin, Paper 9730.

Bauer, B. B. and E. L. Torick (1966). “Researches in loudness measurement”. In: *IEEE Transactions on Audio and Electroacoustics* 14.3, pp. 141–151.

Bauer, Benjamin B., Daniel W. Gravereaux and Arthur J. Gust (1971). “A compatible stereo-quadraphonic (SQ) record system”. In: *Journal of the Audio Engineering Society* 19.8, pp. 638–646.

Baumgartner, Robert, Piotr Majdak and Bernhard Laback (2014). “Modeling sound-source localization in sagittal planes for human listeners.” In: *Journal of the Acoustical Society of America* 136.2, pp. 791–802. DOI: 10.1121/1.4887447.

Bech, Søren and Nick Zacharov (2007). *Perceptual audio evaluation - theory, method and application*. Chichester: John Wiley & Sons.

Beerends, John and Jan Stemerdink (1994). “A perceptual speech-quality measure based on a psychoacoustic sound representation”. In: *Journal of the Audio Engineering Society* 42.3, pp. 115–123.

Beerends, John G. and Frank E. De Caluwe (1999). “The influence of video quality on perceived audio quality and vice versa”. In: *Journal of the Audio Engineering Society* 47.5, pp. 355–362.

Begault, D. R. (1994). *3-D sound for virtual reality and multimedia*. Academic Press Inc.

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

Begault, D.R. (1992a). "Perceptual similarity of measured and synthetic HRTF filtered speech stimuli". In: *Journal of the Acoustical Society of America* 92.4, p. 2334. DOI: 10.1121/1.404988.

Begault, Durand R. (1992b). "Perceptual effects of synthetic reverberation on three-dimensional audio systems". In: *Journal of the Audio Engineering Society* 40.11, pp. 895–904.

Begault, Durand R. and Elizabeth Wenzel (1993). "Headphone localization of speech." In: *Human factors* 35.2, pp. 361–376. DOI: 10.1518/107118191786755797.

Ben-Hur, Zamir et al. (2017). "Spectral equalization in binaural signals represented by order-truncated spherical harmonics". In: *Journal of the Acoustical Society of America* 141.6, pp. 4087–4096. DOI: 10.1121/1.4983652.

Ben-hur, Zamir et al. (2019). "Loudness stability of binaural sound with spherical harmonic representation of sparse head-related transfer functions". In: *EURASIP Journal on Audio, Speech, and Music* 2019.5.

Benjamin, Eric, Richard Lee and Aaron Heller (2006). "Localization in horizontal-only Ambisonic systems". In: *121st Convention of the Audio Engineering Society.* San Francisco, Paper 6967.

Beranek, Leo L. and Harvey P. Sleeper (1946). "The design and construction of anechoic sound chambers". In: *Journal of the Acoustical Society of America* 18.1, pp. 140–150. DOI: 10.1121/1.1916351.

Berge, Svein and Natasha Barrett (2010a). "A new method for B-format to binaural transcoding". In: *40th AES International conference.* Tokyo. URL: http://www.aes.org/e-lib/browse.cfm?elib=15527.

– (2010b). "High angular resolution planewave expansion". In: *2nd International Symposium on Ambisonics and Spherical Acoustics.* Paris.

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

Bernschütz, B. et al. (2014). "Binaural reproduction of plane waves with reduced modal order". In: *Acta Acustica united with Acustica* 100.5, pp. 972–983. DOI: 10.3813/AAA.918777.

Bernschütz, Benjamin (2013). "A spherical far field HRIR/HRTF compilation of the Neumann KU 100". In: *Fortschritte der Akustik – AIA-DAGA 2013*. Merano, pp. 592–595. URL: http://www.audiogroup.web.fh-koeln.de/FILES/AIA-DAGA2013_HRIRs.pdf.

Bertet, Stéphanie et al. (2007). "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: a subjective evaluation involving virtual and real 3D microphones". In: *AES 30th International Conference*. URL: http://www.aes.org/e-lib/browse.cfm?elib=13925.

Bertet, Stéphanie et al. (2009). "Influence of microphone and loudspeaker setup on perceived higher order ambisonics reproduced sound field". In: *Ambisonics Symposium*. URL: http://hal.upmc.fr/hal-00418553.

– (2013). "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources". In: *Acta Acustica united with Acustica* 99.4, pp. 642–657. DOI: 10.3813/AAA.918643.

Best, Virginia et al. (2005). "The role of high frequencies in speech localization". In: *Journal of the Acoustical Society of America* 118.1, pp. 353–363. DOI: 10.1121/1.1926107.

Blauert, Jens (1997). *Spatial hearing: the psychophysics of human sound localization*. Cambridge, MA: MIT press.

Blesser, Barry and Linda-Ruth Salter (2009). *Spaces speak, are you listening?: experiencing aural architecture*. MIT press.

Bolanos, Javier Gomez and Ville Pulkki (2015). "Estimation of pressure at the eardrum in magnitude and phase for headphone equalization using pressure-velocity measurements at the ear canal entrance". In: *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*. DOI: 10.1109/WASPAA.2015.7336910.

Braun, Sebastian and Matthias Frank (2011). "Localization of 3D Ambisonic recordings and Ambisonic virtual sources". In: *International Conference on Spatial Audio.*

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: The MIT Press.

Brinkmann, Fabian and Alexander Lindau (2010). "On the effect of individual headphone compensation in binaural synthesis". In: *36th DAGA.* January, pp. 1055–1056.

Brinkmann, Fabian, Alexander Lindau and Stefan Weinzierl (2013). "A high resolution head-related transfer function database including different orientations of head above the torso". In: *AIA-DAGA 2013 Conference on Acoustics*, pp. 596–599. URL: `http://www.ak.tu-berlin.de/fileadmin/a0135/Publikationen/2013/Brinkmann_2013_A_high_resolution_head-related_transfer_function_database_including_different_orientations_of_head_above_the_torso.pdf`.

– (2017). "On the authenticity of individual dynamic binaural synthesis". In: *Journal of the Acoustical Society of America* 142.4, pp. 1784–1795. DOI: `10.1121/1.5005606`.

Brinkmann, Fabian and Stefan Weinzierl (2018). "Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition". In: *AES Conference on Audio for Virtual and Augmented Reality.*

Brinkmann, Fabian et al. (2014a). "Assessing the authenticity of individual dynamic binaural synthesis". In: *EAA Joint Symposium on Auralization and Ambisonics.* Vol. 71. 4, pp. 62–68. DOI: `10.1121/1.5005606`.

Brinkmann, Fabian et al. (2014b). "Audibility of headabove-torso orientation in head-related transfer functions". In: *Forum Acusticum.*

– (2015a). "Audibility and interpolation of head-above-Torso orientation in binaural technology". In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 931–942. DOI: `10.1109/JSTSP.2015.2414905`.

Brinkmann, Fabian et al. (2015b). "Cross-validation of measured and modeled head-related transfer functions". In: *DAGA*, pp. 1118–1121.

Brinkmann, Fabian et al. (2017a). "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations". In: *Journal of the Audio Engineering Society* 65.10, pp. 841–848. DOI: `10.17743/jaes.2017.0033`.

Brinkmann, Fabian et al. (2017b). *The FABIAN head-related transfer function data base*. DOI: `http://dx.doi.org/10.14279/depositonce-5718`.

Bronkhorst, Adelbert W. (1995). "Localization of real and virtual sound sources". In: *Journal of the Acoustical Society of America* 98.5, pp. 2542–2553. DOI: `10.1121/1.413219`. arXiv: `1109.6529`.

Brookes, Tim and Chris Treble (2005). "The effect of non-symmetrical left/right recording pinnae on the perceived externalisation of binaural recordings". In: *118th Convention of the Audio Engineering Society*. Barcelona, Paper 6439. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13155`.

Brown, C. Phillip and Richard O. Duda (1998). "A structural model for binaural sound synthesis". In: *IEEE Transactions on Speech and Audio Processing* 6.5, pp. 476–488. DOI: `10.1109/89.709673`.

Brüggen, Marc (2001). "Coloration and binaural decoloration in natural environments". In: *Acta Acustica* 87.2001, pp. 400–406.

Brungart, D S and W M Rabinowitz (1999). "Auditory localization of nearby sources. Head-related transfer functions". In: *Journal of the Acoustical Society of America* 106.3, pp. 1465–1479.

Bücklein, Roland (1981). "The audibility of frequency response irregularities". In: *Journal of the Audio Engineering Society* 29.3, pp. 126 –131.

Burkardt, John (2013a). "SPHERE_GRID - points, lines, faces on a sphere". accessed February 09, 2017. URL: `http://people.sc.fsu.edu/~jburkardt/datasets/sphere_grid/sphere_grid.html`.

– (2013b). "SPHERE_LEBEDEV_RULE - Quadrature Rules for the Unit Sphere". accessed September 11, 2018. URL: `https://people.sc.fsu.edu/~jburkardt/m_src/sphere_lebedev_rule/sphere_lebedev_rule.html`.

Byrne, Denis et al. (1994). "An international comparison of long-term average speech spectra". In: *Journal of the Acoustical Society of America* 96.4, pp. 2108–2120. DOI: `10.1121/1.410152`.

Catic, Jasmina, Sébastien Santurette and Torsten Dau (2015). "The role of reverberation-related binaural cues in the externalization of speech". In: *Journal of the Acoustical Society of America* 138.2, pp. 1154–1167. DOI: `10.1121/1.4928132`.

Catic, Jasmina et al. (2013). "The effect of interaural-level-difference fluctuations on the externalization of sound". In: *Journal of the Acoustical Society of America* 134.2, pp. 1232–1241. DOI: `10.1121/1.4812264`.

Chan, J C and C D Geisler (1990). "Estimation of eardrum acoustic pressure and of ear canal length from remote points in the canal". In: *Journal of the Acoustical Society of America* 87.3, pp. 1237–1247. DOI: `10.1121/1.398799`.

Chapman, Michael et al. (2009). "A standard for interchange of Ambisonic signal sets". In: *International Symposium on Ambisonics and Spherical Acoustics*. URL: `http://iem.kug.ac.at/fileadmin/media/iem/projects/2009/ambixchange09.pdf`.

Chen, Jiashu, Barry D. Van Veen and Kurt E. Hecox (1995). "A spatial feature extraction and regularization model for the head-related transfer function". In: *Journal of the Acoustical Society of America* 97.1, pp. 439–452. DOI: `10.1121/1.413110`. URL: `http://scitation.aip.org/content/asa/journal/jasa/97/1/10.1121/1.413110`.

Cheng, Corey and Gregory H Wakefield (1999). "Introduction to head-related transfer functions (HRTFs): representations of HRTFs in time, frequency, and space". In: *107th Convention of the Audio Engineering Society*. New York, Preprint 5026.

Collins, T (2013). "Binaural Ambisonic decoding with enhanced lateral localization". In: *134th Convention of the Audio Engineering Society*. Rome, Paper 8878. URL: `http://www.aes.org/e-lib/browse.cfm?elib=16779`.

Cooley, James W and John W Tukey (1965). "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19.90, pp. 297–301.

Cooper, Duane H. (1982). “Calculator program for head-related transfer function”. In: *Journal of the Audio Engineering Society* 30.1/2, pp. 34–38.

Cooper, Duane H and Jerald L Bauck (1989). “Prospects for Transaural recording”. In: *Journal of the Audio Engineering Society* 37.1.

Cooper, Duane H. and Takeo Shiga (1972). “Discrete-matrix multichannel stereo”. In: *Journal of the Audio Engineering Society* 20.5, pp. 346–360.

Cooper, M and M E Taylor (1998). “Ambisonic sound in virtual environments and applications for blind people”. In: *2nd European Conference on Disability, Virtual Reality & Associated Technology.* September 1998. Skovde, Sweden.

Core Sound (2018). *OctoMic.* URL: `http://www.core-sound.com/OctoMic/1.php` (visited on 09/23/2019).

Craven, Peter and Michael A. Gerzon (1977). *Coincident microphone simulation covering three dimensional space and yielding various directional outputs.*

Daffern, H. et al. (2019). “Exploring the potential of virtual reality technology to investigate the health and well being benefits of group singing”. In: *International Journal of Performance Arts and Digital Media* 15.1. DOI: `10.1080/14794713.2018.1558807`.

Damaske, P. and B. Wagener (1969). “Richtungshorversuche uber einen nachgebildeten Kopf [Investigations of directional hearing using a dummy head]”. In: *Acustica* 19, pp. 198–213.

Dan, Rao and Bosun Xie (2005). “Head rotation and sound image localization in the median plane”. In: *Chinese Science Bulletin* 50.5. DOI: `10.1360/04WW0030`.

Daniel, Jérôme (2000). “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD thesis. l’Université Paris. URL: `http://pcfarina.eng.unipr.it/Public/phd-thesis/jd-these-original-version.pdf`.

– (2003). “Spatial sound encoding including near field effect: introducing distance coding filters and a viable, new Ambisonic format”. In: *AES 23rd International Conference.* Copenhagen.

Daniel, Jérôme and Sébastien Moreau (2004). "Further study of sound field coding with higher order Ambisonics". In: *116th Convention of the Audio Engineering Society*. Berlin, Paper 6017.

Daniel, Jérôme, Rozenn Nicol and Sébastien Moreau (2003). "Further investigations of high order Ambisonics and Wavefield Synthesis for holophonic sound imaging". In: *114th Convention of the Audio Engineering Society*. Amsterdam, Paper 5788.

Daniel, Jérôme, Jean-Bernard Rault and Jean-Dominique Polack (1998). "Ambisonics encoding of other audio formats for multiple listening conditions". In: *105th Convention of the Audio Engineering Society*. San Francisco, Preprint 4795.

– (1999). "Acoustic properties and perceptive implications of stereophonic phenomena". In: *AES 16th International Conference: Spatial Sound Reproduction*. Rovaniemi, pp. 91–102.

Dinakaran, Manoj et al. (2006). "Extraction of anthropometric measures from 3D-meshes for the individualization of head-related transfer functions". In: *140th Convention of the Audio Engineering Society*. Paris, Paper 9579.

Dobreva, Marina S, William E O Neill and Gary D Paige (2011). "Influence of aging on human sound localization". In: *Journal of neurophysiology* 105, pp. 2471–2486. DOI: 10.1152/jn.00951.2010..

Dunn, Chris and Malcolm Omar Hawksford (1993). "Distortion immunity of MLS-derived impulse response measurement". In: *Journal of the Audio Engineering Society* 41.5, pp. 314–335.

Duraiswami, Ramani, Dmitry N. Zotkin and Nail A. Gumerov (2004). "Interpolation and range extraction of HRTFs". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 277–280.

Durlach, N. I. et al. (1992). "On the externalization of auditory images". In: *Presence: Teleoperators and Virtual Environments* 1.2, pp. 251–257. DOI: 10.1162/pres.1992.1.2.251.

Durlach, Nathaniel I. and H. Steven Colburn (1978). "Binaural phenomena". In: *Handbook of Perception*. New York: Academic Press. Chap. 10, pp. 405–466.

Epain, N., C. T. Jin and F. Zotter (2014). "Ambisonic decoding with constant angular spread". In: *Acta Acustica united with Acustica* 100.5, pp. 928–936. DOI: 10.3813/AAA.918772.

Erbes, Vera et al. (2012). "An extraaural headphone system for optimized binaural reproduction". In: *DAGA*. Darmstadt, pp. 313–314.

Evans, Michael J., James A. S. Angus and Anthony I. Tew (1998). "Analyzing head-related transfer function measurements using surface spherical harmonics". In: *Journal of the Acoustical Society of America* 104.4, pp. 2400–2411. DOI: 10.1121/1.423749.

Everest, F Alton (2014). *Master handbook of acoustics.* Ed. by McGraw-Hill Education. 6th ed. ASA.

Faller, Christof and Frank Baumgarte (2003). "Binaural cue coding - part II: schemes and applications". In: *IEEE Transactions on Speech and Audio Processing* 11.6, pp. 520–531. DOI: 10.1109/TSA.2003.818108.

Farina, Angelo (1998). "Software implementation of B-format encoding and decoding". In: *104th Convention of the Audio Engineering Society.* Amsterdam, Preprint 4691.

– (2000). "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *108th Convention of the Audio Engineering Society.* Paris, Preprint 5093.

Farina, Angelo et al. (2001). "Ambiophonic principles for the recording and reproduction of surround sound for music". In: *AES 19th International Conference.* Schloss Elmau.

Fellgett, P. B. (1974). "Ambisonic reproduction of directionality in surround-sound systems". In: *Nature* 252, pp. 534–538. DOI: 10.1038/252534a0.

Fletcher, Harvey and W. A. Munson (1933). "Loudness, its definition, measurement and calculation". In: *The Bell System Technical Journal* 12.4, pp. 337–430. DOI: 10.1016/S0016-0032(23)90506-5.

– (1937). "Relation between loudness and masking". In: *Journal of the Acoustical Society of America* 9.1. DOI: 10.1121/1.1915904.

Frank, Matthias and Franz Zotter (2017). "Exploring the perceptual sweet area in Ambisonics". In: *142nd Convention of the Audio Engineering Society.* Berlin, Paper 9727.

Frank, Matthias, Franz Zotter and Alois Sontacchi (2015). "Producing 3D audio in Ambisonics". In: *AES 57th International Conference.* Hollywood.

Furness, Roger K. (1990). "Ambisonics - an overview". In: *AES 8th International Conference: The Sound of Audio.* DOI: `10.1007/s11427-011-4247-x`.

Gardner, Mark B. (1973). "Some monaural and binaural facets of median plane localization". In: *Journal of the Acoustical Society of America* 54.6, pp. 1489–1495. DOI: `10.1121/1.1914447`.

Gardner, W.G. (1998). "3-D audio using loudspeakers". PhD thesis. DOI: `10.1109/ASPAA.1997.625598`.

Geronazzo, Michele et al. (2014). "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions". In: *IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 4463–4467.

Gerzon, Michael A. (1973). "Periphony: with-height sound reproduction". In: *Journal of the Audio Engineering Society* 21.1, pp. 2–10.

– (1975). "The design of precisely coincident microphone arrays for stereo and surround sound". In: *50th Convention of the Audio Engineering Society.* London. DOI: `10.1017/CBO9781107415324.004`. arXiv: `arXiv:1011.1669v3`.

– (1977a). "Criteria for evaluating surround-sound systems". In: *Journal of the Audio Engineering Society* 25.6, pp. 400–408.

– (1977b). "Design of Ambisonic decoders for multispeaker surround sound". In: *58th Convention of the Audio Engineering Society.*

– (1980). "Practical periphony: the reproduction of full-sphere sound". In: *65th Convention of the Audio Engineering Society.* London, Preprint 1571.

– (1985). "Ambisonics in multichannel broadcasting and video". In: *Journal of the Audio Engineering Society* 33.11, pp. 859–871.

– (1992a). "General metatheory of auditory localization". In: *92nd Convention of the Audio Engineering Society.* Vienna, Preprint 3306. DOI: `10.1111/j.1365-2141.1992.tb04620.x`.

Gerzon, Michael A. (1992b). “Psychoacoustic decoders for multispeaker stereo and surround sound”. In: *93rd Convention of the Audio Engineering Society.* San Francisco, Preprint 3406.

Gerzon, Michael A. and Geoffrey J. Barton (1992). “Ambisonic decoders for HDTV”. In: *92nd Convention of the Audio Engineering Society.* Vienna, Preprint 3345. URL: http://www.aes.org/e-lib/browse.cfm?elib=6788.

Giller, Peter Maximilian and Christian Schorkhuber (2019). “A super-resolution Ambisonics-to-binaural rendering plug-in”. In: *DAGA*. March. Rostock.

Giner, Arnau Vasquez (2013). “Scale - a software tool for listening experiments”. In: *AIA-DAGA*. Merano, pp. 1315–1319. URL: http://audiogroup.web.th-koeln.de/FILES/Vazquez_DEGA2013.pdf.

Gorzel, Marcin, Gavin Kearney and Frank Boland (2014). “Investigation of Ambisonic rendering of elevated sound sources”. In: *AES 55th International Conference.* Helsinki.

Gorzel, Marcin et al. (2011). “On the perception of dynamic sound sources in Ambisonic Binaural Renderings”. In: *AES 41st International Conference on Game Audio.* London.

Gorzel, Marcin et al. (2019). “Efficient encoding and decoding of binaural sound with Resonance Audio”. In: *AES Conference on Immersive and Interactive Audio.* York, UK, Paper 68.

Goupell, Matthew J, Piotr Majdak and Bernhard Laback (2010). “Median-plane sound localization as a function of the number of spectral channels using a channel vocoder”. In: *Journal of the Acoustical Society of America* 127.2, pp. 990–1001. DOI: 10.1121/1.3283014.

G.R.A.S. Sound & Vibration (2013). *Instruction manual: G.R.A.S. KEMAR 45BC.* URL: http://www.gras.dk/media/docs/files/items/m/a/man_45BB_45BC.pdf (visited on 07/04/2017).

Green, Marc and Damian Murphy (2017). “EigenScape: a database of spatial acoustic scene recordings”. In: *Applied Sciences* 7.12. DOI: 10.3390/app7111204. URL: http://www.mdpi.com/2076-3417/7/11/1204.

Griesinger, David (1990). “Binaural techniques for music reproduction”. In: *AES 8th International Conference*, pp. 197–207. DOI: 10.1016/j.nec.2015.03.011.

– (2013). *Binaural hearing, ear canals, and headphone equalization (presentation slides)*. URL: http://www.davidgriesinger.com/Binaural_hearing_and_headphones.ppt (visited on 07/18/2019).

– (2016). “Accurate timbre and frontal localization without head tracking through individual eardrum equalization of headphones”. In: *141st Convention of the Audio Engineering Society*. Los Angeles, Paper 9620.

– (2017). “Laboratory reproduction of binaural concert hall measurements through individual headphone equalization at the eardrum”. In: *142nd Convention of the Audio Engineering Society*. Berlin, Paper 9691.

Guldenschuh, Markus et al. (2008). “HRTF modeling in due consideration variable torso reflections”. In: *Acoustics*. Vol. 123. 5. Paris. DOI: 10.1121/1.2932888.

Hammershøi, Dorte and Henrik Møller (1996). “Sound transmission to and within the human ear canal”. In: *Journal of the Acoustical Society of America* 100.1, pp. 408–427.

Hardin, R. H. and N. J. A. Sloane (1996). “McLaren’s improved Snub Cube and other new spherical designs in three dimensions”. In: *Discrete & Computational Geometry* 15.4, pp. 429–441. DOI: 10.1007/BF02711518. arXiv: 0207211 [math].

Harris, Fredric J (2004). *Multirate signal processing for communication systems*. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Hartmann, W M and B Rakerd (1993). “Auditory spectral discrimination and the localization of clicks in the sagittal plane.” In: *Journal of the Acoustical Society of America* 94.4, pp. 2083–2092. DOI: 10.1121/1.407481.

Hartmann, William M. and Andrew Wittenberg (1996). “On the externalization of sound images”. In: *Journal of the Acoustical Society of America* 99.6, pp. 3678–3688. DOI: 10.1121/1.414965.

Hartmann, William M et al. (2010). “Phase effects on the perceived elevation of complex tones.” In: *Journal of the Acoustical Society of America* 127.5, pp. 3060–3072. DOI: 10.1121/1.3372753.

Hartung, Klaus, Jonas Braasch and Susanne J. Sterbing (1999). “Comparison of different methods for the interpolation of head-related transfer functions”. In: *AES 16th International Conference: Spatial Sound Reproduction*, pp. 319–329. URL: http://www.aes.org/e-lib/browse.cfm?elib=8026.

Hatziantoniou, Panagiotis D and John N Mourjopoulos (2000). “Generalized fractional-octave smoothing of audio and acoustic responses”. In: *Journal of the Audio Engineering Society* 48.4, pp. 259–280.

Hebrank, J H and D. Wright (1974a). “Are two ears necessary for localization of sound in the median plane?” In: *Journal of the Acoustical Society of America* 56.3, pp. 935–938. DOI: 10.1121/1.1903351.

Hebrank, Jack and D. Wright (1974b). “Spectral cues used in the localization of sound sources on the median plane”. In: *Journal of the Acoustical Society of America* 56.6, pp. 1829–1834.

Heller, Aaron J., Eric Benjamin and Richard Lee (2010). “Design of Ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization”. In: *129th Convention of the Audio Engineering Society*. San Francisco, Paper 8243.

Heller, Aaron J. and Eric M. Benjamin (2012). “Calibration of soundfield microphones using the diffuse-field response”. In: *133rd Convention of the Audio Engineering Society*. San Francisco, Paper 8711.

Heller, Aaron J., Richard Lee and Eric M. Benjamin (2008). “Is my decoder ambisonic?” In: *125th Convention of the Audio Engineering Society*. San Francisco, Paper 7553.

Hess, Wolfgang (2004). “Influence of head-tracking on spatial perception”. In: *117th Convention of the Audio Engineering Society*. San Francisco, Paper 6288.

Hiipakka, Marko (2012). “Estimating pressure at the eardrum for binaural reproduction”. PhD Thesis. Aalto University. URL: https://aaltodoc.aalto.fi/handle/123456789/6315.

Hiipakka, Marko, Teemu Kinnari and Ville Pulkki (2011). “HRTF measurements with pressure-velocity sensor”. In: *6th Forum Acusticum*. Aalborg.

Hiipakka, Marko, Miikka Tikander and Matti Karjalainen (2010). "Modeling of external ear acoustics for insert headphone usage". In: *Journal of the Audio Engineering Society* 58.4, pp. 269–281. DOI: `10.1121/1.397957`.

Hiipakka, Marko et al. (2012). "Localization in binaural reproduction with insert headphones". In: *132nd Convention of the Audio Engineering Society*. Budapest, Paper 8666. URL: `http://www.aes.org/e-lib/browse.cfm?conv=132&papernum=8666`.

Hofman, Paul M., Jos G.A. Van Riswick and A. John Van Opstal (1998). "Relearning sound localization with new ears". In: *Nature Neuroscience* 1.5. DOI: `10.1038/1633`.

Hold, Christoph et al. (2019). "Improving binaural Ambisonics decoding by spherical harmonics domain tapering and coloration compensation". In: *IEEE International Conference on Acoustic, Speech and Signal Processing*. Brighton, pp. 261–265. DOI: `10.1109/icassp.2019.8683751`.

Hollerweger, Florian (2006). "Periphonic sound spatialization in multi-user virtual environments". PhD thesis. University of California at Santa Barbara, p. 125.

House, C. et al. (2017). "Personal spatial audio in cars: development of a loudspeaker array for multi-listener transaural reproduction in a vehicle". In: *Institute of Acoustics*. Vol. 39. 2.

Howard, David M. and Jamie Angus (2017). *Acoustics and psychoacoustics*. 5th ed. New York: Routledge.

International Organization for Standardization (1993). *ISO 11172-3:1993, Coding of moving pictures and associated audio for digital storage media at up to about 1,5 mbit / s - part 3: audio*. Tech. rep.

– (2003). *ISO 226:2003, Normal equal-loudness-level contours*. Tech. rep.

– (2016). *ISO 389, Acoustics - reference zero for the calibration of audiometric equipment*. Tech. rep.

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

– (2015b). *ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems BS Series Broadcasting service*. Tech. rep.

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

– (2015d). *ITU-R BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level.* Tech. rep.

Ivanic, Joseph and Klaus Ruedenberg (1996). "Rotation matrices for real spherical harmonies. direct determination by recursion". In: *Journal of Physical Chemistry* 100.15, pp. 6342–6347. DOI: 10.1021/jp9833350.

Iwaya, Yukio, Yoiti Suzuki and Daisuke Kimura (2003). "Effects of head movement on front-back error in sound localization". In: *Acoustical Science And Technology* 24.5, pp. 322–324. DOI: 10.1250/ast.24.322.

Johnson, J. B. (1928). "Thermal agitation of electricity in conductors". In: *Physical Review* 32, pp. 97–109.

Johnston, Daniel, Hauke Egermann and Gavin Kearney (2019). "Measuring the behavioral response to spatial audio within a multi-modal virtual reality environment in children with autism spectrum disorder". In: *Applied Sciences* 9.15. DOI: 10.3390/app9153152.

Jot, Jean-Marc, Véronique Larcher and Jean-Marie Pernaux (1999). "A comparative study of 3-D audio encoding and rendering techniques". In: *AES 16th International Conference*, pp. 281–300.

Jot, Jean-Marc, Scott Wardle and Veronique Larcher (1998). "Approaches to binaural synthesis". In: *105th Convention of the Audio Engineering Society.* San Francisco, Preprint 4861. URL: http://www.aes.org/e-lib/browse.cfm?elib=8319.

Kan, Alan, Craig Jin and André van Schaik (2009). "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function". In: *Journal of the Acoustical Society of America* 125.4, pp. 2233–2242. DOI: 10.1121/1.3081395.

Karagiozov, Hristo (2014). *Reverberation Lecture.* URL: http://karagioza.com/?p=676 (visited on 07/25/2017).

Katz, Brian F. G. and Markus Noisternig (2014). "A comparative study of interaural time delay estimation methods". In: *Journal of the Acoustical Society of America* 135.6, pp. 3530–3540. DOI: 10.1121/1.4875714.

Katz, Brian F. G. and Gaëtan Parseihian (2012). “Perceptually based head-related transfer function database optimization.” In: *Journal of the Acoustical Society of America* 131.2, EL99–105. DOI: `10.1121/1.3672641`.

Kearney, Gavin (2010). “Auditory scene synthesis using virtual acoustic recording and reproduction”. PhD thesis.

Kearney, Gavin and Tony Doyle (2015a). “A HRTF database for virtual loudspeaker rendering”. In: *139th Convention of the Audio Engineering Society.* New York, Paper 9424.

– (2015b). “Height perception in Ambisonic based binaural decoding”. In: *139th Convention of the Audio Engineering Society.* New York, Paper 9423.

Kearney, Gavin et al. (2012). “Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields”. In: *Acta Acustica united with Acustica* 98.1, pp. 61–71. DOI: `10.3813/AAA.918492`.

Kearney, Gavin et al. (2015). “Auditory distance perception with static and dynamic binaural rendering”. In: *AES 57th International Conference.* Hollywood.

Khaykin, Dima and Boaz Rafaely (2012). “Acoustic analysis by spherical microphone array processing of room impulse responses.” In: *Journal of the Acoustical Society of America* 132.1, pp. 261–270. DOI: `10.1121/1.4726012`.

Kim, Sang-Myeong and Wonjae Choi (2005). “On the externalization of virtual sound images in headphone reproduction: a Wiener filter approach.” In: *Journal of the Acoustical Society of America* 117.6, pp. 3657–3665. DOI: `10.1121/1.1921548`.

Kirkeby, Ole and Philip A. Nelson (1999). “Digital filter design for inversion problems in sound reproduction”. In: *Journal of the Audio Engineering Society* 47.7/8, pp. 583–595. URL: `http://www.aes.org/e-lib/browse.cfm?elib=12098`.

Kirkeby, Ole et al. (1998). “Fast deconvolution of multichannel systems using regularization”. In: *IEEE Transactions on Speech and Audio Processing* 6.2, pp. 189–194.

Kirkeby, Ole et al. (2007). “Some effects of the torso on head-related transfer functions”. In: *122nd Convention of the Audio Engineering Society.* Vienna, Paper 7030.

Kistler, Doris J. and Frederic L. Wightman (1992). “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction.” In: *Journal of the Acoustical Society of America* 91.3, pp. 1637–47. DOI: 10.1121/1.402444.

Klein, Florian and Stephan Werner (2015). “Auditory Adaptation in Spatial Listening Tasks”. In: *138th Convention of the Audio Engineering Society.* Warsaw, Paper 9281.

Kronlachner, Matthias and Franz Zotter (2014). “Spatial transformations for the enhancement of Ambisonic recordings”. In: *2nd International Conference on Spatial Audio.*

Kuhn, George F. (1977). “Model for the interaural time differences in the azimuthal plane”. In: *Journal of the Acoustical Society of America* 62.1, pp. 157–167.

Kulkarni, Abhijit and H. Steven Colburn (2000). “Variability in the characterization of the headphone transfer-function”. In: *Journal of the Acoustical Society of America* 107.2, pp. 1071–1074. DOI: 10.1121/1.428571.

Langendijk, Erno H. and Adelbert W. Bronkhorst (2002). “The contribution of spectral cues to human sound localization”. In: *Journal of the Acoustical Society of America* 105.2, pp. 1036–1036. DOI: 10.1121/1.424945.

Langendijk, Erno H. A. and Adelbert W. Bronkhorst (2000). “Fidelity of three-dimensional-sound reproduction using a virtual auditory display”. In: *Journal of the Acoustical Society of America* 107.1, pp. 528–537. URL: http://asa.scitation.org/doi/10.1121/1.428321.

Le Bagousse, Sarah et al. (2011). “Sound quality evaluation based on attributes - application to binaural contents”. In: *131st Convention of the Audio Engineering Society.* New York, Paper 8542.

Lebedev, V. I. (1976). “Quadratures on a sphere”. In: *USSR Computational Mathematics and Mathematical Physics* 16.2, pp. 10–24. DOI: 10.1016/0041-5553(76)90100-2.

Lecomte, Pierre et al. (2015). “On the use of a Lebedev grid for Ambisonics”. In: *139th Convention of the Audio Engineering Society.* New York, Paper 9433.

Lecomte, Pierre et al. (2016). "A fifty-node Lebedev grid and its applications to Ambisonics". In: *Journal of the Audio Engineering Society* 64.11, pp. 868–881. DOI: 10.17743/jaes.2016.0036.

Lecomte, Pierre et al. (2018). "Directional filtering of Ambisonic sound scenes". In: *AES Conference on Spatial Reproduction.*

Leventhal, Les (1986). "Type 1 and Type 2 errors in the statistical analysis of listening tests". In: *Journal of the Audio Engineering Society* 34.6, pp. 437–453; 664.

Lewald, Jörg (1997). "Eye-position effects in directional hearing". In: *Behavioural Brain Research* 87.1, pp. 35–48. DOI: 10.1016/S0166-4328(96)02254-1.

Lewald, Jörg, Gerd J. Dörrscheidt and Walter H. Ehrenstein (2000). "Sound localization with eccentric head position". In: *Behavioural Brain Research* 108.2, pp. 105–125. DOI: 10.1016/S0166-4328(99)00141-2.

Lewald, Jörg and Walter H. Ehrenstein (1998). "Influence of head-to-trunk position on sound lateralization". In: *Experimental Brain Research* 121.3, pp. 230–238. DOI: 10.1007/s002210050456.

Lewald, Jörg and Hans Otto Karnath (2001). "Sound lateralization during passive whole-body rotation". In: *European Journal of Neuroscience* 13.12, pp. 2268–2272. DOI: 10.1046/j.0953-816X.2001.01608.x.

Li, Zhiyun and Ramani Duraiswami (2007). "Flexible and optimal design of spherical microphone arrays for beamforming". In: *IEEE Transactions on Audio, Speech and Language Processing* 15.2, pp. 702–714. DOI: 10.1109/TASL.2006.876764.

Lindau, Alexander (2009). "The perception of system latency in dynamic binaural synthesis". In: *NAG-DAGA 2009 International Conference on Acoustics.* 1. Rotterdam, pp. 1063–1066. URL: https://www.ak.tu-berlin.de/fileadmin/a0135/Publikationen/2009/Lindau_2009_The_Perception_of_System_Latency_in_Dynamic_Binaural_Synthesis.pdf.

Lindau, Alexander and Fabian Brinkmann (2010). "Perceptual evaluation of individual headphone compensation in binaural synthesis based on non-individual recordings". In: *ISCA / DEGA Tutorial and Research Workshop on Perceptual Quality of Systems.* Dresden, Germany, pp. 137–142.

Lindau, Alexander and Fabian Brinkmann (2012). "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings". In: *Journal of the Audio Engineering Society* 60.1-2, pp. 54–62.

Lindau, Alexander, Torben Hohn and Stefan Weinzierl (2007). "Binaural resynthesis for comparative studies of acoustical environments". In: *22nd Convention of the Audio Engineering Society*. Vienna, Paper 7032. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14017`.

Lindau, Alexander, Hans Joachim Maempel and Stefan Weinzierl (2008). "Minimum BRIR grid resolution for dynamic binaural synthesis". In: *Acoustics Conference*. Paris, pp. 3851–3856.

Lindau, Alexander and Stefan Weinzierl (2012). "Assessing the plausibility of virtual acoustic environments". In: *Acta Acustica united with Acustica* 98.5, pp. 804–810. DOI: `10.3813/AAA.918562`.

Lindau, Alexander et al. (2014). "A spatial audio quality inventory (SAQI)". In: *Acta Acustica united with Acustica* 100.5, pp. 984–994. DOI: `10.3813/AAA.918778`.

Lourens, J. G. (1991). "On the sibilance problem in FM sound transmission". In: *IEEE Transactions on Broadcasting* 37.3, pp. 115–120.

Lovedee-Turner, Michael and Damian Murphy (2018). "Application of machine learning for the spatial analysis of binaural room impulse responses". In: *Applied Sciences* 8.1. DOI: `10.3390/app8010105`. URL: `http://www.mdpi.com/2076-3417/8/1/105`.

MacCabe, C. J. and D. J. Furlong (1994). "Virtual imaging capabilities of surround sound systems". In: *Journal of the Audio Engineering Society* 42.1/2, pp. 38–49.

Mackensen, Philip et al. (2000). "Head-tracker based Auralization Systems: Additional Consideration of Vertical Head Movements". In: *108th Convention of the Audio Engineering Society*. Paris, Preprint 5135. URL: `http://www.aes.org/e-lib/browse.cfm?elib=9203`.

Macpherson, Ewan A. and John C. Middlebrooks (2002). "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited". In: *Journal of the Acoustical Society of America* 111.5, pp. 2219–2236. DOI: `10.1121/1.1471898`.

Majdak, Piotr, Peter Balazs and Bernhard Laback (2007). “Multiple exponential sweep method for fast measurement of head-related transfer functions”. In: *Journal of the Audio Engineering Society* 55.7-8, pp. 623–636.

Malham, David G. (1992). “Experience with a large area 3D Ambisonic sound systems”. In: *Institute of Acoustics Autumn Conference on Reproduced Sound*. Windermere, pp. 209–216.

– (1999). “Higher order Ambisonic systems for the spatialisation of sound”. In: *International Computer Music Conference*. Beijing, pp. 484–487.

– (2003). “Higher order Ambisonic systems”. Mphil. University of York, pp. 484–487. URL: http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf.

– (2019). “The early years of Ambisonics at York”. In: *AES Conference on Immersive and Interactive Audio*. York, Paper 59.

Malham, David G. and Anthony Myatt (1995). “3-D sound spatialization using Ambisonic techniques”. In: *Computer Music Journal* 19.4, pp. 58–70. DOI: 10.2307/3680991.

Martin, J., D. Van Maercke and J. P. Vian (1993). “Binaural simulation of concert halls: A new approach for the binaural reverberation process”. In: *Journal of the Acoustical Society of America* 94.6, pp. 3255–3264. DOI: 10.1121/1.407232.

Martin, Russell L., Ken I. McAnally and Melis A. Senova (2001). “Free-field equivalent localization of virtual audio”. In: *Journal of the Audio Engineering Society* 49.1/2, pp. 14–22. URL: http://www.aes.org/e-lib/browse.cfm?elib=10204.

Masiero, Bruno and Janina Fels (2011a). “Equalization for binaural synthesis with headphone”. In: *Fortschritte der Akusitk - DAGA*. Düsseldorf, pp. 675–676. URL: http://publications.rwth-aachen.de/record/128503.

– (2011b). “Perceptually robust headphone equalization for binaural reproduction”. In: *130th Convention of the Audio Engineering Society*. London, Paper 8388.

May, Tobias, Steven Van De Par and Armin Kohlrausch (2011). “A probabilistic model for robust localization based on a binaural auditory front-end”. In: *IEEE Transactions on Audio, Speech and Language Processing* 19.1. DOI: 10.1109/TASL.2010.2042128.

McCormack, Leo and Archontis Politis (2019). “SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods”. In: *AES Conference on Immersive and Interactive Audio*, e–brief 111.

Mcgill, Robert, John W Tukey and Wayne A Larsen (1978). “Variations of box plots”. In: *The American Statistician* 32.1, pp. 12–16.

McKeag, Adam and David McGrath (1996). “Sound field format to binaural decoder with head-tracking”. In: *AES 6th Australian Regional Convention.* Melbourne, Preprint 4302.

McKenzie, Thomas, Damian Murphy and Gavin Kearney (2017). “Assessing the authenticity of the KEMAR mouth simulator as a repeatable speech source”. In: *143rd Convention of the Audio Engineering Society.* New York, Paper 9820.

– (2018). “Diffuse-field equalisation of binaural Ambisonic rendering”. In: *Applied Sciences* 8.10. DOI: 10.3390/app8101956.

Merimaa, Juha (2009). “Modification of HRTF Filters to reduce timbral effects in binaural synthesis”. In: *127th Convention of the Audio Engineering Society.* New York, Paper 7912.

Meyer, Jens and Tony Agnello (2003). “Spherical microphone array for spatial sound recording”. In: *115th Convention of the Audio Engineering Society.* New York, Paper 5975.

Meyer, Jens and Gary Elko (2002). “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* Vol. 2. IEEE, pp. 1781–1784. DOI: 10.1109/ICASSP.2002.5744968.

Middlebrooks, John C. (1999). “Individual differences in external-ear transfer functions reduced by scaling in frequency”. In: *Journal of the Acoustical Society of America* 106.3, pp. 1480–1492. DOI: 10.1121/1.427176.

Middlebrooks, John C and David M Green (1991). “Sound localization by human listeners”. In: *Annual Review of Psychology* 42, pp. 135–159.

Millns, Connor, Maksims Mironovs and Hyunkook Lee (2019). "Vertical localisation accuracy of binauralised First Order Ambisonics across multiple horizontal positions". In: *146th Convention of the Audio Engineering Society*. Dublin, Paper 10167.

Minnaar, Pauli et al. (2000). "The interaural time difference in binaural synthesis". In: *108th Convention of the Audio Engineering Society*. Paris, Preprint 5133. DOI: `10.1109/ASPAA.1999.810884`.

Møller, Henrik (1992). "Fundamentals of binaural technology". In: *Applied Acoustics* 36.3-4, pp. 171–218. DOI: `10.1016/0003-682X(92)90046-U`.

Møller, Henrik et al. (1995). "Transfer characteristics of headphones measured on human ears". In: *Journal of the Audio Engineering Society* 43.4, pp. 203–217.

Møller, Henrik et al. (1996). "Binaural technique: do we need individual recordings?" In: *Journal of the Audio Engineering Society* 44.6, pp. 451–469. URL: `http://www.aes.org/e-lib/browse.cfm?elib=7897`.

Monro, Gordon (2000). "In-phase corrections for Ambisonics". In: *International Computer Music Conference*.

Moore, Alastair H. (2009). "Towards the perception of externalised auditory images using binaural technology". PhD thesis. University of York.

Moore, Alastair H., Anthony I. Tew and Rozenn Nicol (2010). "An initial validation of individualized crosstalk cancellation filters for binaural perceptual experiments". In: *Journal of the Audio Engineering Society* 58.1-2, pp. 36–45.

Moore, Brian C. J. (2012). *An introduction to the psychology of hearing*. 6th ed. Emerald Group Publishing Limited.

Moore, Brian C. J. and Brian R. Glasberg (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns". In: *Journal of the Acoustical Society of America* 74.3, pp. 750–753. DOI: `10.1121/1.389861`.

– (1996). "A revision of Zwicker's loudness model". In: *Acustica* 82, pp. 335–345.

Moore, Brian C. J., Brian R. Glasberg and Thomas Baer (1997). "A model for the prediction of thresholds, loudness, and partial loudness". In: *Journal of the Audio Engineering Society* 45.4, pp. 224–240.

Moore, David and Jonathan Wakefield (2010). "Optimisation of the localisation performance of irregular Ambisonic decoders for multiple off-centre listeners". In: *128th Convention of the Audio Engineering Society*. London, Paper 8061.

Moreau, S, Jérôme Daniel and S Bertet (2006). "3D sound field recording with higher order Ambisonics-objective measurements and validation of a 4th order spherical microphone". In: *120th Convention of the Audio Engineering Society*. Paris. URL: `http://www.aes.org/e-lib/online/browse.cfm?elib=13661`.

Morimoto, M (2001). "The contribution of two ears to the perception of vertical angle in sagittal planes." In: *Journal of the Acoustical Society of America* 109.4, pp. 1596–1603. DOI: `10.1121/1.1352084`.

Morse, Philip McCord and Uno Ingard (1968). *Theoretical Acoustics*. Princeton University Press.

Munson, W. A. and Mark B. Gardner (1950). "Standardizing auditory tests". In: *Journal of the Acoustical Society of America* 22.5, p. 675. DOI: `10.1121/1.1917190`.

Murphy, Damian T. (2013). "Archaeological acoustic space measurement for convolution reverberation and auralization applications". In: *9th International Conference on Digital Audio Effects*, pp. 221–226.

Neukom, Martin (2007). "Ambisonic panning". In: *123rd Convention of the Audio Engineering Society*, Paper 7297.

Neumann (2013). *KU 100*. URL: `https://www.neumann.com/?lang=en&id=current_microphones&cid=ku100_description`.

Nicol, Rozenn et al. (2014). "A roadmap for assessing the quality of experience of 3D audio binaural rendering". In: *EAA Joint Symposium on Auralization and Ambisonics*, pp. 100–106.

Noisternig, Markus et al. (2003a). "3D binaural sound reproduction using a virtual ambisonic approach". In: *IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*. Lugano, pp. 174–178. DOI: `10.1109/VECIMS.2003.1227050`.

Noisternig, Markus et al. (2003b). "A 3D ambisonic based binaural sound reproduction system". In: *AES 24th International Conference on Multichannel Audio*. Banff.

Olive, Sean E. (2003). "Differences in performance and preference of trained versus untrained listeners in loudspeaker tests: a case study". In: *Journal of the Audio Engineering Society* 51.9, pp. 806–825. URL: `http://www.aes.org/e-lib/browse.cfm?elib=12206`.

Ono, Kazuho, Ville Pulkki and Matti Karjalainen (2001). "Binaural modeling of multiple sound source perception: methodology and coloration experiments". In: *111th Convention of the Audio Engineering Society*, Paper 5446.

– (2002). "Binaural modeling of multiple sound source perception: coloration of wideband sound". In: *112th Convention of the Audio Engineering Society*. Munich, Paper 5550. URL: `http://www.aes.org/e-lib/browse.cfm?elib=11331`.

Oosterom, A. Van and J. Strackee (1983). "The solid angle of a plane triangle". In: *IEEE Transactions on Biomedical Engineering* BME-30.2, pp. 125–126.

Palacino, J and R Nicol (2012). "Perceptual assessment of binaural decoding of first-order Ambisonics". In: *Acoustics Conference*. April. Nantes. URL: `http://hal.archives-ouvertes.fr/hal-00810918/`.

Par, Steven van de et al. (2005). "A perceptual model for sinusoidal audio coding based on spectral integration". In: *EURASIP Journal on Applied Signal Processing* 2005.9, pp. 1292–1304.

Perrett, Stephen and William Noble (1997a). "The contribution of head motion cues to localization of low-pass noise". In: *Perception & Psychophysics* 59.7, pp. 1018–1026. DOI: `10.3758/BF03205517`.

– (1997b). "The effect of head rotations on vertical plane sound localization". In: *Journal of the Acoustical Society of America* 102.4, pp. 2325–2332. DOI: `10.1121/1.419642`.

Pike, Chris, Frank Melchior and Tony Tew (2014). "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room". In: *AES 55th International Conference: Spatial Audio*. Helsinki.

Poletti, M. A. (2005). "Three-dimensional surround sound systems based on spherical harmonics". In: *Journal of the Audio Engineering Society* 53.11, pp. 1004–1024.

Poletti, Mark (1996). “The design of encoding functions for stereophonic and polyphonic sound systems”. In: *Journal of the Audio Engineering Society* 44.11, pp. 948–963.

– (2000). “A unified theory of horizontal holographic sound systems”. In: *Journal of the Audio Engineering Society* 48.12, pp. 1155–1182. URL: `http://www.aes.org/e-lib/browse.cfm?elib=12033`.

Politis, Archontis (2016). “Microphone array processing for parametric spatial audio techniques”. PhD Thesis. Aalto University.

Politis, Archontis, Leo McCormack and Ville Pulkki (2017). “Enhancement of ambisonic binaural reproduction using directional audio coding with optimal adaptive mixing”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* 2017, pp. 379–383. DOI: `10.1109/WASPAA.2017.8170059`.

Politis, Archontis, Sakari Tervo and Ville Pulkki (2018). “COMPASS: Coding and multidirectional parameterization of ambisonic sound scenes”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* DOI: `10.1109/ICASSP.2018.8462608`.

Politis, Archontis, Juha Vilkamo and Ville Pulkki (2015). “Sector-based parametric sound field reproduction in the spherical harmonic domain”. In: *IEEE Journal on Selected Topics in Signal Processing* 9.5, pp. 852–866. DOI: `10.1109/JSTSP.2015.2415762`.

Pomberger, Hannes and Franz Zotter (2009). “An Ambisonics format for flexible playback layouts”. In: *Ambisonics Symposium.* Graz. URL: `http://ambisonics.iem.at/symposium2009/proceedings`.

– (2011). “Warping of 3D Ambisonic recordings”. In: *3rd International Symposium on Ambisonics and Spherical Acoustics*. Lexington.

Porschmann, Christoph, Johannes M. Arend and Fabian Brinkmann (2019). “Directional equalization of sparse head-related transfer function sets for spatial upsampling”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 27.6, pp. 1060–1071. DOI: `10.1109/TASLP.2019.2908057`.

Porschmann, Christoph, Johannes M. Arend and Raphael Gillioz (2019). “How wearing headgear affects measured head-related transfer functions”. In: *EAA Spatial Audio Signal Processing Symposium.* Paris.

Postma, Barteld N J and Brian F G Katz (2016). “Acoustics of Notre-Dame cathedral de Paris”. In: *22nd International Congress on Acoustics.* November. Buenos Aires.

Postma, Barteld N. J. et al. (2016). “Virtual reality performance auralization in a calibrated model of Notre-Dame cathedral”. In: *EuroRegio2016.* Porto.

Postma, Barteld N.J. and Brian F.G. Katz (2015). “Creation and calibration method of acoustical models for historic virtual reality auralizations”. In: *Virtual Reality* 19.3-4, pp. 161–180. DOI: `10.1007/s10055-015-0275-3`.

Pralong, Daniele and Simon Carlile (1996). “The role of individualized headphone calibration for the generation of high fidelity virtual auditory space”. In: *Journal of the Acoustical Society of America* 100.6, pp. 3785–3793. DOI: `10.1121/1.417337`.

Pulkki, Ville (1997). “Virtual sound source positioning using vector base amplitude panning”. In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

– (2006). “Directional audio coding in spatial sound reproduction and stereo upmixing”. In: *AES 28th International Conference.* Piteå. URL: `http://www.aes.org/e-lib/browse.cfm?elib=13847`.

– (2007). “Spatial sound reproduction with directional audio coding”. In: *Journal of the Audio Engineering Society* 55.6, pp. 503–516. URL: `http://www.aes.org/e-lib/browse.cfm?elib=14170`.

Pulkki, Ville et al. (1999). “Analyzing virtual sound source attributes using a binaural auditory model”. In: *Journal of the Audio Engineering Society* 47.4, pp. 203–217.

Raake, A. and J. Blauert (2013). “Comprehensive modeling of the formation process of sound-quality”. In: *5th International Workshop on Quality of Multimedia Experience*, pp. 76–81.

Rafaely, Boaz (2005). “Analysis and design of spherical microphone arrays”. In: *IEEE Transactions on Speech and Audio Processing* 13.1, pp. 135–143. DOI: `10.1109/TSA.2004.839244`.

Rafaely, Boaz, Barak Weiss and Eitan Bachmat (2007). “Spatial aliasing in spherical microphone arrays”. In: *IEEE Transactions on Signal Processing* 55.3, pp. 1003–1010. DOI: 10.1109/TSP.2006.888896.

Rakerd, Brad and W. M. Hartmann (1985). “Localization of sound in rooms, II: The effects of a single reflecting surface”. In: *Journal of the Acoustical Society of America* 78.2, pp. 524–533. DOI: 10.1121/1.392474.

Raykar, V C et al. (2003). “Extracting significant features from the HRTF”. In: *International Conference on Auditory Display*, pp. 115–118. URL: http://icad.org/Proceedings/2003/RaykarDuraiswami2003.pdf.

Raykar, Vikas C, Ramani Duraiswami and B Yegnanarayana (2005). “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses.” In: *Journal of the Acoustical Society of America* 118.1, pp. 364–374. DOI: 10.1121/1.1923368.

Rayleigh, Lord (1907). “On our perception of sound direction”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232. DOI: 10.1080/14786440709463595.

Reeves, Alec Harley (1942). *Electric signalling system.*

Rettinger, Michael (1957). “Reverberation chambers for broadcasting and recording studios”. In: *Journal of the Audio Engineering Society* 5.1, pp. 18–22.

Richter, Jan Gerrit et al. (2014). “Spherical harmonics based hrtf datasets: Implementation and evaluation for real-time auralization”. In: *Acta Acustica united with Acustica* 100.4, pp. 667–675. DOI: 10.3813/AAA.918746.

Rix, A.W. et al. (2001). “Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752. DOI: 10.1109/ICASSP.2001.941023. arXiv: arXiv:1011.1669v3.

Roffler, Suzanne K. and Robert A. Butler (1967). “Factors that influence the localization of sound in the vertical plane”. In: *Journal of the Acoustical Society of America* 43.6, pp. 1255–1259.

Rumsey, Francis (2001). *Spatial Audio.* Oxford: Focal Press, pp. 1689–1699. arXiv: arXiv:1011.1669v3.

Rumsey, Francis (2002). "Spatial quality evaluation for reproduced sound: terminology, meaning, and a scene-based paradigm". In: *Journal of the Audio Engineering Society* 50.9, pp. 651–666.

Rumsey, Francis et al. (2005a). "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality". In: *Journal of the Acoustical Society of America* 118.2, pp. 968–976. DOI: `10.1121/1.1945368`.

Rumsey, Francis et al. (2005b). "Relationships between experienced listener ratings of multichannel audio quality and naïve listener preferences". In: *Journal of the Acoustical Society of America* 117.6, pp. 3832–3840. DOI: `10.1121/1.1904305`.

Saff, E. B. and A. B. J. Kuijlaars (1997). "Distributing many points on a sphere". In: *The Mathematical Intelligencer* 19.1, pp. 5–11. DOI: `10.1007/BF03024331`.

Sangpeilaudio (2019). *Table of sound levels.* URL: `http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm` (visited on 08/14/2019).

Satongar, Darius et al. (2015). "The influence of headphones on the localization". In: *Journal of the Audio Engineering Society* 63.10, pp. 799–810.

Schärer, Zora and Alexander Lindau (2009). "Evaluation of equalization methods for binaural signals". In: *126th Convention of the Audio Engineering Society*. Munich.

Schoeffler, Michael and Jürgen Herre (2016). "The relationship between basic audio quality and overall listening experience". In: *Journal of the Acoustical Society of America* 140.3, pp. 2101–2112. DOI: `10.1121/1.4963078`.

Schoeffler, Michael et al. (2018). "webMUSHRA - a comprehensive framework for web-based listening tests". In: *Journal of Open Research Software* 6.8. DOI: `10.5334/jors.187`.

Schonstein, David, Laurent Ferré and Brian Katz (2008). "Comparison of headphones and equalization for virtual auditory source localization". In: *Acoustics Conference*. Vol. 123. 5. Paris, pp. 3724–3729. DOI: `10.1121/1.2935199`.

Schörkhuber, Christian and Robert Höldrich (2019). "Linearly and quadratically constrained least-squares decoder for signal-dependent binaural rendering of Ambisonic signals". In: *AES Conference on Immersive and Interactive Audio*. York, Paper 22.

Schörkhuber, Christian, Markus Zaunschirm and Robert Höldrich (2018). "Binaural rendering of Ambisonic signals via magnitude least squares". In: *DAGA*, pp. 339–342.

Schroeder, M. R. (1970). "Digital simulation of sound transmission in reverberant spaces". In: *Journal of the Acoustical Society of America* 47.2, pp. 424–431.

– (1979). "Integrated-impulse method measuring sound decay without using impulses". In: *Journal of the Acoustical Society of America* 66.2, pp. 497–500. DOI: 10.1121/1.383103.

Schultz, Frank et al. (2010). "An extraaural headphone for optimized binaural reproduction". In: *26th Tonmeister Convention*, pp. 702–714.

Searle, C L et al. (1975). "Binaural pinna disparity: another auditory localization cue." In: *Journal of the Acoustical Society of America* 57.2, pp. 448–455. DOI: 10.1121/1.380442.

Sheaffer, Jonathan and Boaz Rafaely (2014). "Equalization strategies for binaural room impulse response rendering using spherical arrays". In: *IEEE 28th Convention of Electrical and Electronics Engineers in Israel.* DOI: 10.1109/EEEI.2014.7005804.

Sheaffer, Jonathan, Shahar Villeval and Boaz Rafaely (2014). "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction". In: *EAA Joint Symposium on Auralization and Ambisonics.* Berlin, pp. 81–85.

Shinn-Cunningham, B G, N I Durlach and R M Held (1998). "Adapting to supernormal auditory localization cues. II. Constraints on adaptation of mean response." In: *Journal of the Acoustical Society of America* 103.6, pp. 3667–3676. DOI: 10.1121/1.423107.

Shinn-Cunningham, B G, Scott Santarelli and Norbert Kopco (2000). "Tori of confusion: binaural localization cues for sources within reach of a listener." In: *Journal of the Acoustical Society of America* 107.3, pp. 1627–36. DOI: 10.1121/1.428447.

Smith, Rosanna C.G. and Stephen R. Price (2014a). "Modelling of human low frequency sound localization acuity demonstrates dominance of spatial variation of interaural time difference and suggests uniform just-noticeable differences in interaural time difference". In: *PLoS ONE* 9.2. DOI: 10.1371/journal.pone.0089033.

Smith, Rosanna C.G. and Stephen R. Price (2014b). "Modelling of human low frequency sound localization acuity demonstrates dominance of spatial variation of interaural time difference and suggests uniform just-noticeable differences in interaural time difference". In: *PLoS ONE* 9.2. DOI: 10.1371/journal.pone.0089033.

Smith, Steven W. (1997). *The scientist and engineer's guide to digital signal processing*. San Diego, CA: California Technical Publishing.

Solvang, Audun (2008). "Spectral impairment for two-dimensional higher order ambisonics". In: *Journal of the Audio Engineering Society* 56.4, pp. 267–279.

Sontacchi, Alois et al. (2002). "Subjective validation of perception properties in binaural sound reproduction systems". In: *AES 21st International Conference*. St. Petersburg.

Soulodre, Gilbert A and Michel C Lavoie (1999). "Subjective evaluation of large and small impairments in audio codecs". In: *AES 17th International Conference*, pp. 329–336.

SoundField (2019). *ST450 MKII*. URL: https://www.soundfield.com/#/products/st450mk2 (visited on 09/23/2019).

Spagnol, Simone, Marko Hiipakka and Ville Pulkki (2011). "A single-azimuth pinna-related transfer function database". In: *14th International Conference on Digital Audio Effects*. Paris, pp. 209–212.

Stan, Guy-Bart, Jean-Jacques Embrechts and Dominique Archambeau (2002). "Comparison of different impulse response measurement techniques". In: *Journal of the Audio Engineering Society* 50.4, pp. 249–262.

Steadman, Mark A. et al. (2017). "Effects of gamification and active listening on short-term sound localization training in virtual reality". In: *bioRxiv*. DOI: 10.1101/207753.

Stevens, Francis and Damian Murphy (2014). "Spatial impulse response measurement in an urban environment". In: *Aes 55th International Conference*. Helsinki.

Stevens, K. N. (2000). *Acoustic phonetics.* MIT Press.

Stevens, S S (1955). "The measurement of loudness". In: *Journal of the Acoustical Society of America* 27.5, pp. 815–829. DOI: `10.1121/1.1908048`.

Stitt, Peter, Lorenzo Picinali and Brian F.G. Katz (2019). "Auditory accommodation to poorly matched non-individual spectral localization cues through active learning". In: *Scientific Reports* 9.1, pp. 1–14. DOI: `10.1038/s41598-018-37873-0`.

Takane, Shouichi et al. (2002). "A database of head-related transfer functions in whole directions on upper hemisphere". In: *Acoustical Science and Technology* 23.3, pp. 160–162. DOI: `10.1250/ast.23.160`.

Takanen, Marko, Marko Hiipakka and Ville Pulkki (2012). "Audibility of coloration artifacts in HRTF filter designs". In: *AES 45th International Conference.* Helsinki.

Tan, CJ and Woon Seng Gan (2000). "Direct concha excitation for the introduction of individualized hearing cues". In: *Journal of the Audio Engineering Society* 48.7, pp. 642–653. URL: `http://www.aes.org/e-lib/browse.cfm?elib=12055`.

Theile, Günther and Helmut Wittek (2011). "Principles in surround recordings with height". In: *130th Convention of the Audio Engineering Society.* London, Paper 8403.

Thiede, Thilo et al. (2000). "PEAQ– The ITU standard for objective measurement of perceived audio quality". In: *Journal of the Audio Engineering Society* 48.1/2, pp. 3–29.

Thresh, Lewis, Calum Armstrong and Gavin Kearney (2017). "A direct comparison of localisation performance when using first, third and fifth order Ambisonics for real loudspeaker and virtual loudspeaker rendering". In: *143rd Convention of the Audio Engineering Society.*

Thurlow, Willard R., John W. Mangels and Philip S. Runge (1967). "Head movements during sound localization". In: *Journal of the Acoustical Society of America* 42.2, pp. 489–493. DOI: `10.1121/1.1910605`.

Thurlow, Willard R and Philip S Runge (1967). "Effect of induced head movements on localization of direction of sounds". In: *Journal of the Acoustical Society of America* 42.2, pp. 480–488.

Treeby, BE, Jie Pan and RM Paurobally (2007). “The effect of hair on auditory localization cues”. In: *Journal of the Acoustical Society of America* 122.6, pp. 3586–3597. DOI: 10.1121/1.2793607.

Trevino, Jorge et al. (2010). “High order Ambisonic decoding method for irregular loudspeaker arrays”. In: *International Congress on Acoustics* August, pp. 1–8.

Van Wanrooij, M. M. and A John Van Opstal (2005). “Relearning sound localization with a new ear”. In: *Journal of Neuroscience* 25.22, pp. 5413–5424. DOI: 10.1523/JNEUROSCI.0850-05.2005.

Vliegen, Joyce, A John and Van Opstal (2004). “The influence of duration and level on human sound localization”. In: *Journal of the Acoustical Society of America* 115.4, pp. 1705–1713. DOI: 10.1121/1.1687423.

Völk, Florian (2009). “Externalization in data-based binaural synthesis: effects of impulse response length”. In: *NAG-DAGA 2009 International Conference on Acoustics*, pp. 1075–1078.

Volk, Florian, Thomas Musialik and Hugo Fastl (2009). “Crosstalk cancellation between phantom sources”. In: *126th Convention of the Audio Engineering Society.* Munich, Paper 7722.

Wabnitz, Andrew, Nicolas Epain and Craig T. Jin (2012). “A frequency-domain algorithm to upscale Ambisonic sound scenes”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, pp. 385–388. DOI: 10.1109/ICASSP.2012.6287897.

Wang, Shihua, Andrew Sekey and Allen Gersho (1992). “An Objective Measure for Predicting Subjective Quality of Speech Coders”. In: *IEEE Journal on Selected Areas in Communications* 10.5, pp. 819–829. DOI: 10.1109/49.138987.

Warusfel, Olivier (2003). *Listen HRTF database.* URL: http://recherche.ircam.fr/equipes/salles/listen/ (visited on 07/21/2017).

Watanabe, Kanji et al. (2007). “Estimation of interaural level difference based on anthropometry and its effect on sound localization”. In: *Journal of the Acoustical Society of America* 122.5, pp. 2832–2841. DOI: 10.1121/1.2785039.

Watson, Christopher J.G. et al. (2017). "The generalization of auditory accommodation to altered spectral cues". In: *Scientific Reports* 7.1, pp. 1–8. DOI: 10.1038/s41598-017-11981-9.

Wendt, Florian, Matthias Frank and Franz Zotter (2014). "Panning with height on 2, 3, and 4 loudspeakers". In: *2nd International Conference on Spatial Audio.*

Wenzel, Elizabeth (1995). "The relative contribution of interaural time and magnitude cues to dynamic sound localization". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.* DOI: 10.1109/aspaa.1995.482963.

Wenzel, Elizabeth and S.H. Foster (1993). "Perceptual consequences of interpolating head-related transfer functions during spatial synthesis". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 102–105. DOI: 10.1109/ASPAA.1993.379986.

Wenzel, Elizabeth et al. (1993). "Localization using nonindividualized head-related transfer functions". In: *Journal of the Acoustical Society of America* 94.1, pp. 111–123.

Wierstorf, Hagen, Alexander Raake and Sascha Spors (2013). "Localization in Wave Field Synthesis and higher order Ambisonics at different positions within the listening area". In: *DAGA*. Meran.

Wiggins, Bruce (2004). "An investigation into the real-time manipulation and control of three-dimensional sound fields". PhD thesis. URL: http://derby.openrepository.com/derby/handle/10545/217795.

– (2007). "The generation of panning laws for irregular speaker arrays using heuristic methods". In: *AES 31st International Conference.* London.

– (2017). "Analysis of binaural cue matching using mbisonics to binaural decoding techniques". In: *4th International Conference on Spatial Audio.* Graz.

Wiggins, Bruce, Iain Paterson-Stephens and Pieter Schillebeeckx (2001). "The analysis of multi-channel sound reproduction algorithms using HRTF data". In: *AES 19th International Conference.*

Wightman, Frederic L. and Doris J. Kistler (1992). "The dominant role of low-frequency interaural time differences in sound localization". In: *Journal of the Acoustical Society of America* 91.3, pp. 1648–1661. DOI: 10.1121/1.402445.

Wightman, Frederic L. and Doris J. Kistler (1997). “Monaural sound localization revisited”. In: *Journal of the Acoustical Society of America* 101.2, pp. 1050–1063. DOI: `10.1121/1.418029`.

– (1999). “Resolution of front-back ambiguity in spatial hearing by listener and source movement.” In: *Journal of the Acoustical Society of America* 105.5, pp. 2841–2853. DOI: `10.1121/1.426899`.

Xie, Bosun (2013). *Head-related transfer function and virtual auditory display.* 2nd ed. J Ross Publishing.

Xie, Bosun and Yang Liu (2014). “Analysis on the timbre of Ambisonics recording by circular and spherical microphone array using a binaural loudness model”. In: July, pp. 1–7.

Yang, Liu and Xie Bosun (2015). “Subjective evaluation on the timbre of horizontal ambisonics reproduction”. In: *International Conference on Audio, Language and Image Processing*, pp. 11–15. DOI: `10.1109/ICALIP.2014.7009747`.

Yao, Shu Nung, Tim Collins and Peter Jančovič (2015). “Timbral and spatial fidelity improvement in ambisonics”. In: *Applied Acoustics* 93. DOI: `10.1016/j.apacoust.2015.01.005`.

Zaunschirm, Markus, Christian Schörkhuber and Robert Höldrich (2018). “Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint”. In: *Journal of the Acoustical Society of America* 143.6, pp. 3616–3627. DOI: `10.1121/1.5040489`.

Zotkin, Dmitry N et al. (2006). “Fast head-related transfer function measurement via reciprocity”. In: *Journal of the Acoustical Society of America* 120.4, pp. 2202–2215. DOI: `10.1121/1.2207578`.

Zotter, Franz and Matthias Frank (2012). “All-round Ambisonic panning and decoding”. In: *Journal of the Audio Engineering Society* 60.10, pp. 807–820.

– (2018). “Ambisonic decoding with panning-invariant loudness on small layouts (AllRAD2)”. In: *144th Convention of the Audio Engineering Society.* Milan, Paper 9943.

Zotter, Franz, Matthias Frank and Christian Haar (2015). “Spherical microphone array equalization for Ambisonics Franz”. In: *Fortschritte der Akustik, DAGA.*

Zotter, Franz, Matthias Frank and Hannes Pomberger (2013). “Comparison of energy-preserving and all-round Ambisonic decoders”. In: *Fortschritte der Akusitk - DAGA*. January.

Zotter, Franz, Matthias Frank and Alois Sontacchi (2010). “The virtual T-Design Ambisonics-rig using VBAP”. In: *1st EAA-EuroRegio*.

Zotter, Franz, Hannes Pomberger and Markus Noisternig (2010). “Ambisonic decoding with and without mode-matching: a case study using the hemisphere”. In: *International Symposium on Ambisonics and Spherical Acoustics*.

– (2012). “Energy-preserving Ambisonic decoding”. In: *Acta Acustica united with Acustica* 98.1, pp. 37–47. DOI: `10.3813/AAA.918490`.