The Morphologically Informed Perceptual Enhancement of Spatial Audio

Laurence Jon Hobden

Doctor of Philosophy

University of York

Electronic Engineering

January, 2018

Abstract

With the proliferation of immersive technologies in virtual reality (VR), broadcasting and home entertainment, estimating an individual's head-related transfer functions (HRTFs) conveniently and with a satisfactory perceptual performance is of great topical interest. To facilitate this, a deep understanding of how the head and pinnae form the acoustic cues used to perceive spatial sound is required. This thesis presents the refinement of a powerful research tool, morphoacoustic perturbation analysis (MPA), for advancing knowledge in this field.

To simplify analysis a novel method is presented for smoothing HRTFs based on an equivalent rectangular bandwidth (ERB) criterion. The approach is first evaluated using an auditory localisation model and these results are validated by means of listening tests. It is shown that ERB smoothing achieves perceptual transparency using fewer parameters than a similar constant-bandwidth approach. Furthermore, it simplifies the structure of the HRTFs, since additional perceptually irrelevant features are discarded.

It has been well established that the boundary element method can satisfactorily generate HRTFs based on a three-dimensional (3D) mesh of a listener's head and pinnae shape. A proof-of-principle for MPA, upon which this thesis builds, has successfully inverted this process, making it possible to identify the morphological regions of the head and pinnae responsible for creating an HRTF feature. However, first-generation MPA suffered from significant weaknesses including low mesh resolution, restricted frequency range, and topological issues created by the mesh slicing approach used. In this work these issues are addressed through the use of optimised spherical mesh mapping and spherical harmonic deformations. The theory, implementation and validation of the new method is described to the point where the creation of a full MPA database capable of probing in depth the relationship between human morphology and HRTFs is now possible.

Contents

List of Tables

List of Figures

Acknowledgements

Thanks are primarily due to my supervisor, Tony Tew, for his help and support in my research. Without his willingness and enthusiasm to discuss my work, I would not have got as far as I did. No matter how entrenched in the minutiae I might have got, he was always there to help me take a step back and consider the bigger picture. I would especially like to thank him for his patience whilst I tried to balance finishing my PhD with working full time; without his encouragement and guidance it would have been an insurmountable task.

Many thanks also to Damian Murphy for acting as my thesis advisor on top of acting as my supervisor throughout my undergraduate degree. It's been a long ten years!

Thanks are also due to all those in the Audio Lab: thanks for all the puzzles, the cakes, the Q Factors, the Nerf gun battles, the day trips, the games nights, and the many visits to the Seahorse (and other pubs...). Thanks especially go out to Kat Young for their amazing work on generating the KEMAR mesh, which this research would have been a lot more challenging without, and for our numerous discussions on geometry and BEM simulations.

I am ever grateful to Michael Capp at Meridian for acting as my industrial supervisor and also giving me my first taste of self-led research whilst a placement student at Meridian during my undergraduate degree; I doubt I would be here if it weren't for him.

Thanks should also go to my parents for supporting me throughout all my studies, particularly in the final months of writing, for providing support and sustenance during my numerous "writing retreats" at their house. I definitely wouldn't be here if it weren't for them.

Declaration

I hereby declare this thesis is entirely my own work and all contributions from outside sources, through direct contact or publications, have been explicitly stated and referenced. I also declare that some parts of this work have been presented previously, at conferences and in journals. These publications are listed as follows:

- **Investigating head-related transfer function smoothing using a sagittal-plane localization model**, L. J. Hobden and A. I. Tew, paper and poster presentation at *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, WASPAA 2015, New Paltz, United States, October 2015 (Hobden and Tew, 2015) - included in appendix A.

This work has not previously been presented for an award at this, or any other, University.

Chapter 1

Introduction

Without music, life would be a mistake.

Twilight of the Idols

FRIEDRICH NIETZSCHE

1.1 Background

> *"Imagine that you are on the edge of a lake and a friend challenges you to play a game. The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?"*

The above analogy by Bregman (1994) illustrates rather well the complexity and enormity of the task carried out by our ears in localising sounds. When presented as above, the problem seems impossible to solve, but it is a close analogy with the human hearing system: the lake is the air that surrounds us, the two channels are our two ear canals and the handkerchiefs our eardrums. Based on the two signals arriving at our eardrums, our brain does a remarkable job of estimating the location of not only singular static sound sources but also multiple, simultaneous moving sound sources.

The manner by which humans localise sounds in the horizontal plane are chiefly through detecting the difference in the time of arrival of a sound at the two ears (the interaural time difference (ITD)) and the difference in the amplitude of theses two versions of the sound (the interaural level difference (ILD)). This has been understood for well over a century (Rayleigh, 1907). However, these cues are not unambiguous throughout the whole of three-dimensional (3D) space. There arise multiple positions in 3D space that produce similar interaural cues,

specifically around the so-called "cones of confusion" (see section 2.2.3). In these situations, the spectral filtering imposed on sounds entering the ears via the complex morphology of the outer ears, or pinnae, as well as by the head and torso, becomes important. The spectral modifications are used as directional cues (Shaw and Teranishi, 1968; Teranishi and Shaw, 1968; Blauert, 1970; Hebrank and Wright, 1974; Blauert, 1997). They are captured in an individual's head-related transfer functions (HRTFs) in the frequency-domain and in their corresponding head-related impulse responses (HRIRs) in the time-domain.

Filtering a sound with an individual's HRTFs and playing the sound back over headphones can create very realistic virtual spatial sound sources (Wightman and Kistler, 1989b). This has numerous potential applications in not only an artistic setting (see section 2.3), but also in practical situations such as virtual auditory displays (VADs) for aeronautics (Bronkhorst *et al.*, 1996) or the visually impaired (Loomis *et al.*, 1998; Marston *et al.*, 2007).

However, the measurement of an individual's HRTFs is a time consuming process that requires specialist equipment. The use of generic or non-individualised HRTFs is one option, but it often results in poor elevation perception, internalisation and an increase in front-back confusions (Wenzel *et al.*, 1993). An individual's HRTFs can also be simulated acoustically using acoustic modelling approaches such as the boundary element method (BEM) (Kahana and Nelson, 2006, 2007); although this requires a detailed surface mesh description of the individual's head which is likewise time consuming and requires specialist equipment. Deepening our understanding of the acoustic roles played by the human head and pinnae in spatial hearing is of great scientific interest and can be expected to simplify the synthesis of perceptually valid, individualised HRTFs through acoustic simulations based on morphological measurements.

1.2 Statement of Hypotheses

The long term goals of this research are:

- to deepen scientific understanding about the role of human head and pinna morphology in spatial hearing, for example to identify the most perceptually salient acoustic features in HRTFs and reveal how they are created; and
- to simulate perceptually valid individualised HRTFs from a viable number of simple morphological measurements, for example to promote the availability of high quality spatial audio in consumer products and services.

In advancing towards attaining these goals, the aims of this research are:

1. to simplify individualised HRTFs whilst preserving their ability to render sound with a high degree of perceptual integrity ; and
2. to create an accurate and efficient tool for studying the morphological origins of HRTF acoustic features.

Relating to the above aims, the following hypotheses, are postulated:

1. By considering the frequency selectivity of the human hearing system, the spectral detail of HRTFs can be significantly simplified, whilst maintaining perceptual transparency.
2. By identifying and overcoming the deficiencies in first-generation morphoacoustic perturbation analysis (MPA) (Thorpe, 2009; Tew *et al.*, 2012) it is possible to develop a second-generation MPA tool with an improved performance.

1.3 Thesis outline

To prepare the ground and provide context for the work carried out in this thesis, chapter 2 reviews the relevant literature. Underpinning every aspiration of

this research is the human hearing system, so the chapter begins by outlining the structure of the human hearing system, including an overview of studies investigating the auditory filters, of particular import when considering HRTF simplification. Acoustic localisation cues, both interaural and spectral, and the acuity of human sound localisation can all suffer if simplification is excessive, or the accuracy of simulated HRTFs compromised, and these are reviewed next. A survey of the rapidly evolving world of spatial audio applications reveals the plethora of potential applications of this research and places it in a wider context. The long-standing challenges of HRTF acquisition and estimation are a strong motivation for both research goals and are considered next. Previous work on the relationship between auditory cues and human morphology has been influential in the development of the first-generation MPA tool. These are reviewed in light of this, but also to inform the evaluation of MPA and identify its strengths and weaknesses compared with previous approaches.

Chapter 3 reports improvement to the smoothing algorithm presented by Kulkarni and Colburn (1998). The algorithm is adapted to smooth the HRTF magnitude spectrum based on the frequency selectivity of the human hearing system. The algorithm is first validated using an auditory model and then by perceptual listening tests. The results of both the auditory model and the listening tests suggest that a further increase in smoothing is possible by considering the frequency selectivity of the human hearing system, whilst preserving perceptual transparency. This ensures that only the most perceptually salient cues are retained for MPA, as well as providing insight into the level of spectral detail that is audible.

Development and improvement of MPA (Thorpe, 2009; Tew *et al.*, 2012) is described in chapter 4. Several significant weaknesses of first generation MPA are tackled, including: low mesh resolution, restricted frequency range and topological issues arising from the mesh slicing approach used. A new method of optimised spherical mapping is reported that allows the application of spherical harmonic deformations to the head mesh with minimal distortion.

Whilst creation of the second-generation MPA database has not been possible within the scope of this research, chapter 5 reports steps that have been taken towards creating the database. This includes formulation of the spherical harmonic deformations to be used, their normalisation, verification of the accuracy of the proposed BEM solver, validation of both the limit of linearity and the associated signal-to-noise ratio (SNR), and definition of the directions included within the database.

Finally an overview of the contributions of this work is given in chapter 6. In addition, the hypotheses stated in the preceding section are revisited in light of the work completed and potential further work is discussed.

Chapter 2

Spatial hearing and spatial audio

> Do not go where the path may lead, go instead where there is no path and leave a trail.
>
> RALPH WALDO EMERSON

This chapter offers a review of spatial hearing and spatial audio to provide context for this work. The physiology of the human hearing system is described along with the various cues utilised by humans to localise sounds within three-dimensional (3D) space. The range of spatial audio systems that manipulate these auditory cues to create sonic illusions are also discussed to give examples of the wider applications of this research. Head-related transfer functions (HRTFs) are introduced along with current state of the art acquisition techniques for measuring and estimating HRTFs. Finally previous studies examining the relationship between human morphology and HRTF features are discussed to present specific background to the work presented in the following chapters.

2.1 The auditory system

Before discussing auditory cues it is important to understand the physiology of the human hearing system. Therefore this section first discusses the peripheral

auditory system from the outer to inner ear. Then the frequency limits of human hearing are presented as these limits have important ramifications in terms of the acoustic simulations used in later chapters. Finally the characteristics of the auditory filters are described to provide a foundation for the presented work on HRTF simplification.

2.1.1 Structure of the human ear

The human peripheral auditory system is divided into three parts: the outer, middle and inner ear (figure 2.1). Together they are responsible for converting sound waves travelling through the air and into the outer ear to mechanical vibration within the middle and inner ear and then to neural signals which are sent to the brain for analysis. The neurology of human hearing is beyond the scope of this work, but the anatomy and physiology of the peripheral auditory system are of significant importance and so will be reviewed here.

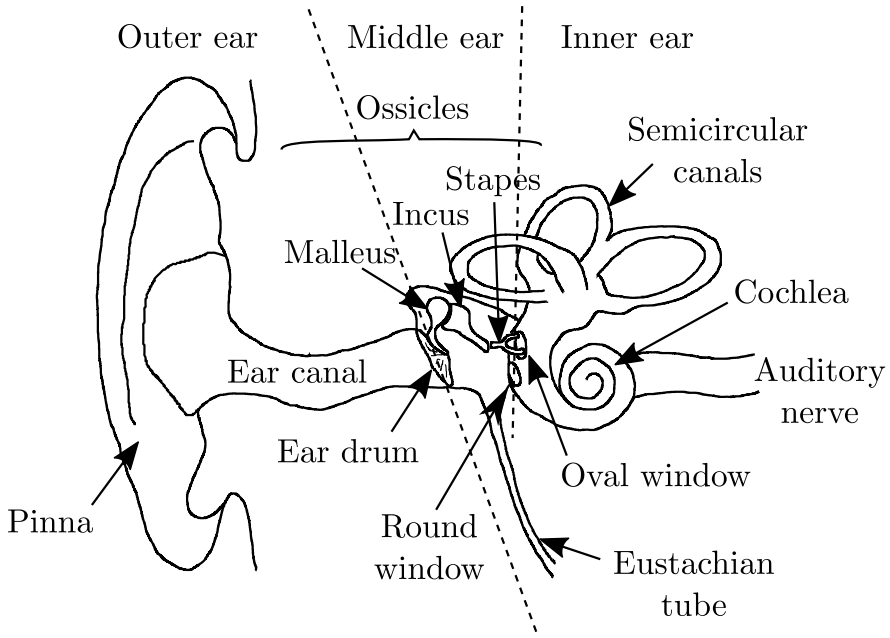


Figure 2.1: The peripheral auditory system: the outer, middle and inner ear. After Fastl and Zwicker (2007, p. 24, Fig. 3.1).

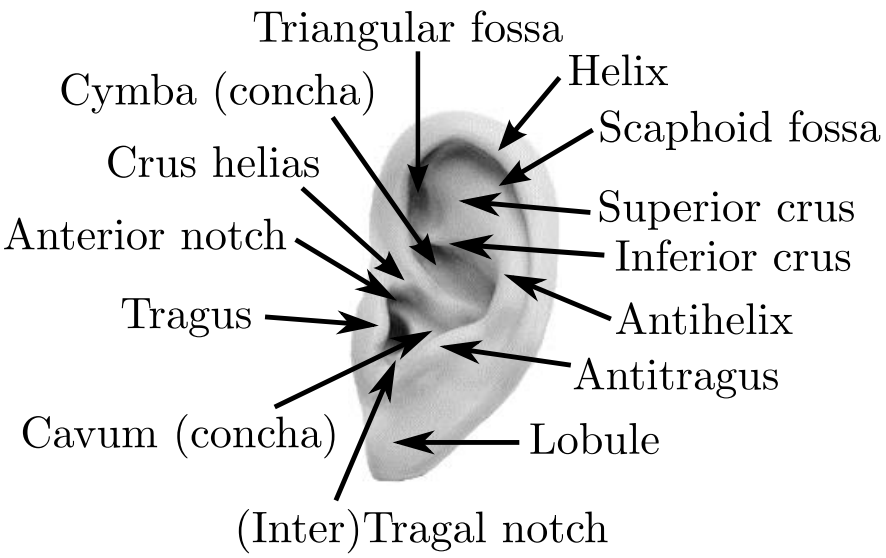


Figure 2.2: Naming conventions of the parts of the human pinna[1].

2.1.1.1 The external ear

The external ear consists of the *pinna* (or *auricle*, plural *pinnae* or *auricles*) and the *ear canal* (or *meatus*). It captures sound waves travelling through the air and transmits them to the *eardrum* (or *tympanic membrane*) where they enter the middle ear. The behaviour of the external ear is entirely passive (i.e. it does not generate any sound energy, nor does it "react" to sound) and is consequently governed by acoustic laws.

The pinnae are responsible for imparting the majority of the fine spectral detail used in localisation on the sound entering the ear (this will be discussed in more detail in section 2.2.4.1). One of the aims of this research is to study how pinna morphology is involved in contributing to these spectral cues. It is therefore necessary to clarify the nomenclature that will be used in throughout this research to refer to parts of the pinna. Figure 2.2 shows the naming conventions that will be used throughout this work.

The large cavity is referred to as the *concha* and is divided by the *crus helias* into an upper (*cymba*) and lower (*cavum*) portion. The fold around the top edge of the pinna is called the *helix* and the cavity under it is known as the *scaphoid*

[1]Image of pinna edited from `http://www.gras.dk/media/catalog/product/cache/1/image/500x500/9df78eab33525d08d6e5fb8d27136e95/k/b/kb0066_web.jpg`

fossa. The raised section between the concha and helix is known as the *antihelix*. The antihelix has two *crura* that form a y-shape around the *triangular fossa*. The *crus* above the *triangular fossa* is referred to as the *superior crus* and the one below is the *inferior crus*. The fleshy section at the very bottom of the pinna is called the *lobule* and the edge between the lobule and cavum concha is known as the *antitragus*. The *tragus* is the prominence that protudes over the ear canal. Above it is the *anterior notch* and below it the *intertragal notch*.

The ear canal is a slightly curved tube 20–30 mm long with a diameter of 7–8 mm (Blauert, 1997; Xie, 2013; Pulkki and Karjalainen, 2015). It is entirely lined with skin but the walls of the outer third are cartilaginous whereas the walls of its inner two-thirds (closest to the eardrum) are bony. As it is open to the air at one end (low acoustic impedance) and closed by the eardrum at the other end (high acoustic impedance), acoustically it acts as a cylindrical tube closed at one end. It therefore exhibits a quarter-wavelength resonance that amplifies frequencies in the region of 3–4 kHz (Wiener and Ross, 1946). This resonance is generally omitted from the study of spectral auditory localisation cues (section 2.2.4) as it has been shown to be independent of direction and is therefore unlikely to play any active part in sound localisation (Mehrgardt and Mellert, 1977; Middlebrooks *et al.*, 1989; Møller *et al.*, 1995a; Algazi *et al.*, 1999). At the end of the ear canal lies the eardrum, an almost circular membrane that is about 0.1 mm thick and has an area of about 66 mm^2 (Blauert, 1997; Xie, 2013) which converts the sound waves arriving at it to mechanical vibrations that are passed to the middle ear.

2.1.1.2 The middle ear

The middle ear is an air-filled cavity that acts as an impedance-matching device between the air in the outer and middle ears and the fluid in the inner ear (Moore, 2013; Xie, 2013; Pulkki and Karjalainen, 2015). Connected to the eardrum are a collection of the smallest bones in the human body: the *ossicles*. The *malleus* (Latin for *hammer*) is connected at one end directly to the eardrum and at the

other end to the *incus* (Latin for *anvil*) which is in turn connected to the *stapes* (Latin for *stirrup*). The stapes is then connected via the oval window to the inner ear. Leading from the middle ear, to the nasopharynx, is the *eustachian tube*. It is normally closed but is occasionally opened, predominantly during yawning or swallowing, to equalise pressure between the middle ear and the exterior during changes in altitude.

Due to the very different acoustic impedances of air and the fluid inside the cochlea (about 1:3000 (Pulkki and Karjalainen, 2015)), if the sound waves were to impact directly on the oval or round window of the cochlea, most of the sound energy would be reflected (Moore, 2013; Xie, 2013; Pulkki and Karjalainen, 2015). However, there are two functions of the middle ear that vastly improve energy transmission. Firstly the ossicles act as a lever mechanism which offers a minor improvement. Secondly, offering a far larger improvement, is the ratio of the surface area of the eardrum to the surface area of the oval window (approximately 30:1 (Xie, 2013; Pulkki and Karjalainen, 2015)). The effect of both these mechanisms is to convert a small pressure with large velocity in the air to a large pressure with low velocity in the fluid of the inner ear. The middle ear is most efficient in the frequency range 500 Hz–5 kHz (Aibara *et al.*, 2001) and effectively

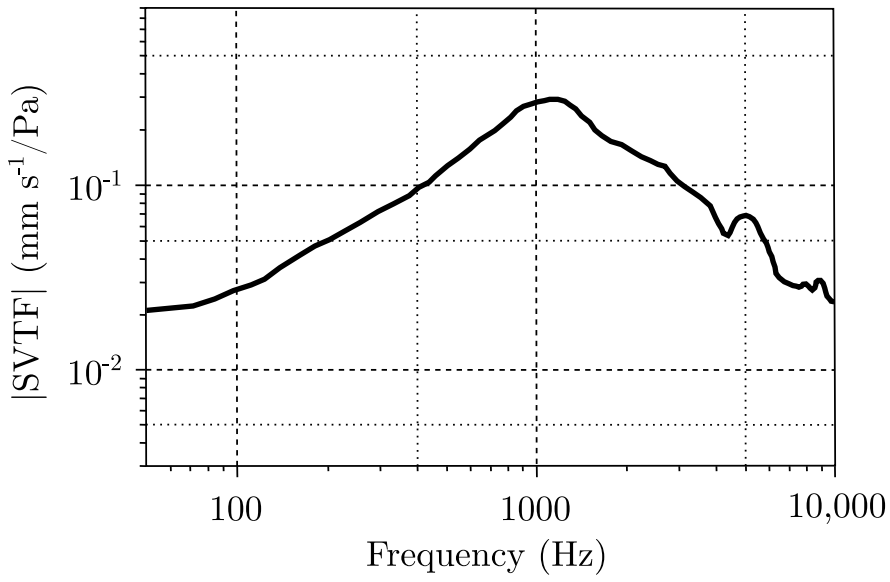


Figure 2.3: Transfer function of the middle ear: the ear canal sound pressure to stapes footplate velocity transfer function (SVTF). After Aibara *et al.* (2001). The plotted curve is the average of three repeated measurements for 11 different ears.

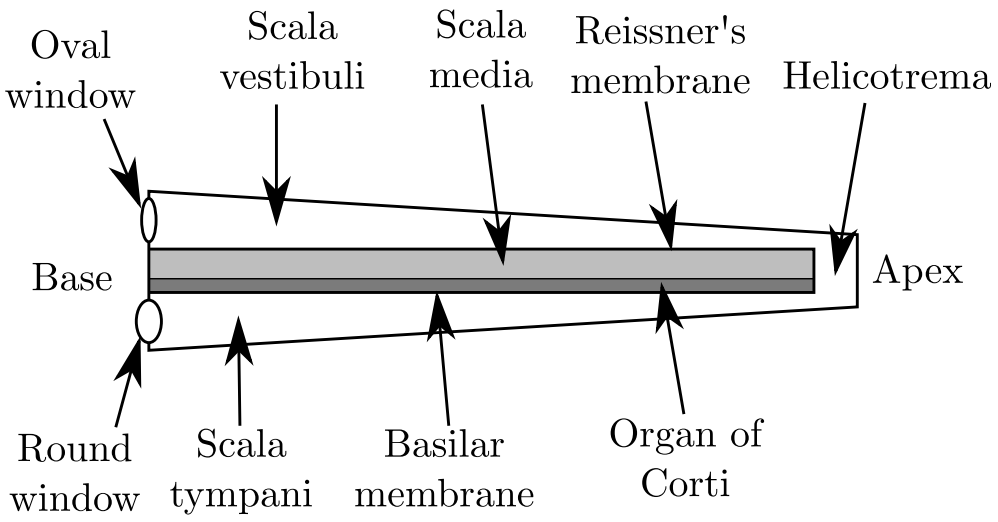
acts as a bandpass filter: attenuating low and high frequencies, as shown in the ear canal sound pressure to stapes footplate velocity transfer function (SVTF) (figure 2.3).

The middle ear also has a further function, known as the *acoustic reflex*. This is a reaction of the muscles attached to the ossicles to loud sounds (reported levels vary from 50-90 dB (Blauert, 1997; Pulkki and Karjalainen, 2015)), whereby the muscles contract, reducing the sensitivity of the ear. However, it is not a particularly efficient protection system as it is too slow to react to impulsive sounds (with a latency of tens to hundreds of milliseconds) and is also only effective below approximately 2 kHz (Pulkki and Karjalainen, 2015). Despite this, it has been suggested that the acoustic reflex is used subconsciously to protect the hearing system from sounds that the brain can anticipate such as self-generated speech/shouting or the sound when running on hard surfaces (Geisler, 1998; Møller, 2000, 2012; Pulkki and Karjalainen, 2015).

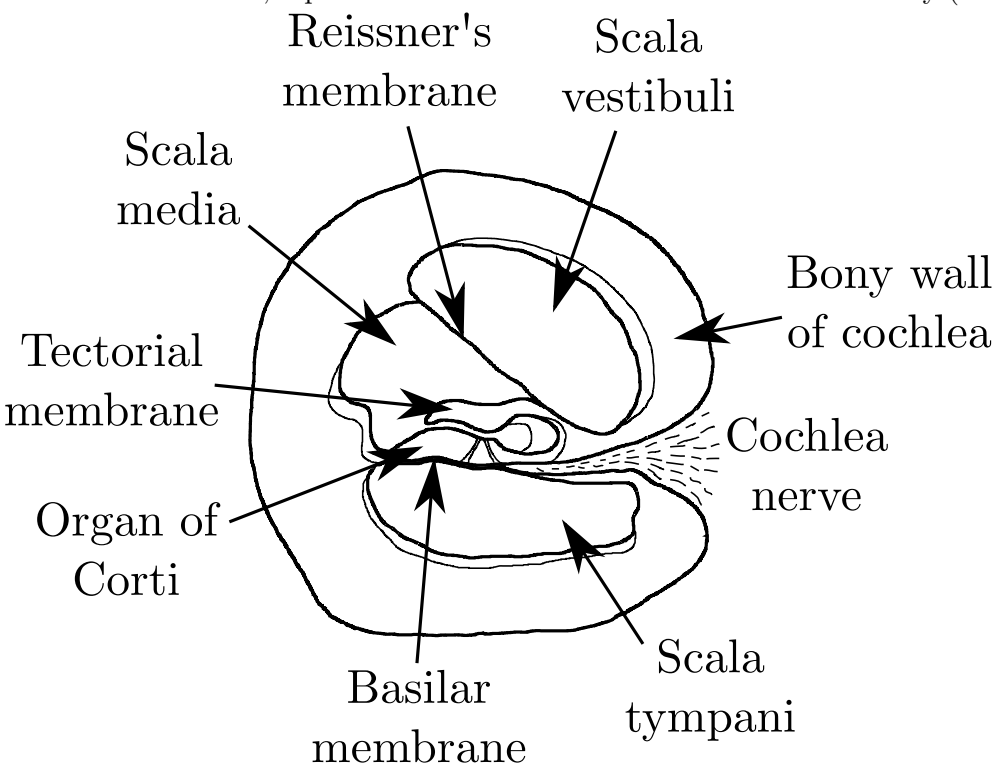
2.1.1.3 The inner ear

The inner ear consists of the *cochlea* and the *semicircular canals* and connects the middle ear to the *auditory nerve*, converting mechanical vibrations to electrical signals. The semicircular canals are used in the sense of balance but play no active part in hearing and so will be discussed no further (Pulkki and Karjalainen, 2015).

The cochlea is a bony-walled, spiral-shaped, fluid-filled tube approximately 35 mm long with about 2.75 turns. However, for simplicity it is often depicted "unwound" as a linear tube as in figure 2.4a. The wide end, nearest the middle ear, is known as the *base* (or the *basal* end) and the narrow end is known as the *apex* (or the *apical* end). Lengthways it is divided into three chambers: the *scala vestibuli* (*vestibular canal*), the *scala media* (*cochela duct*) and the *scala tympani* (*tympanic canal*). At the basal end there are two vibroacoustic connections to the the middle ear: the *oval window* connects the middle ear to the scala vestibuli and the *round window* connects the middle ear to the scala tympani. At the apical

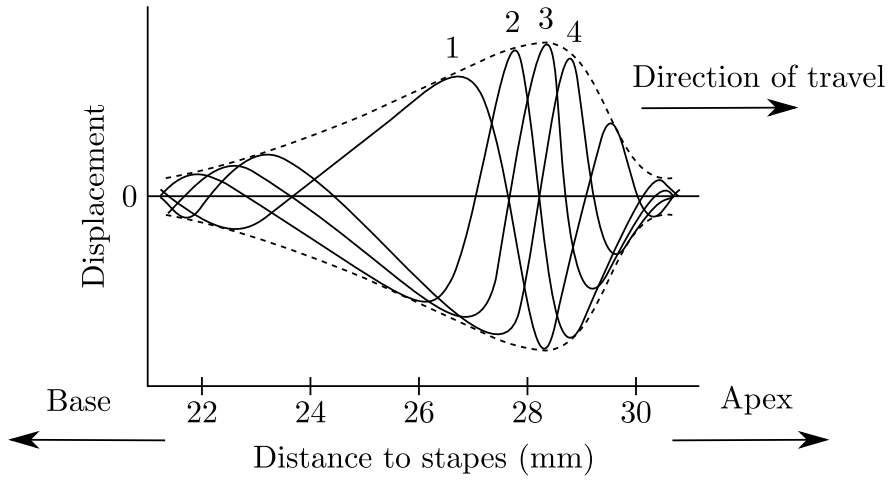


(a) Structure of cochlea, represented as an unwound linear tube. After Kelly (2002).

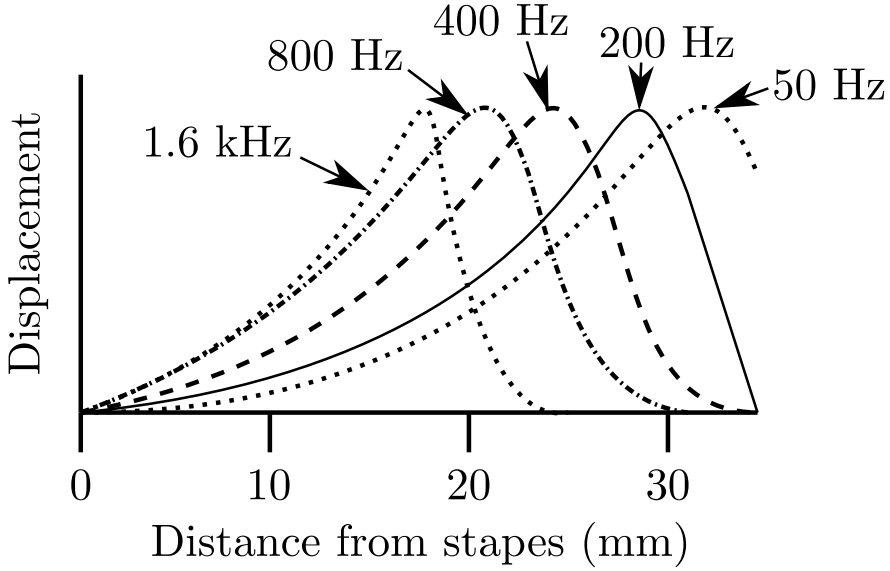


(b) Cross section of the cochlea.

Figure 2.4: The structure of the human cochlea. See text for description. After Encyclopædia Britannica (2018).



(a) The displacement of the basilar membrane for a travelling wave invoked by a 200 Hz tone at four time instants 1–4. The dashed line shows the envelope created by the peaks of the waveform. The wave grows slowly as it travels along the basilar membrane until it reaches a maximum at a given distance from the stapes before it decays rapidly. After Moore (2013, p. 26, fig. 1.9).



(b) The envelopes of basilar membrane displacement for various low frequency tones. After Xie (2013, p. 6, fig. 1.6). Original data from von Békésy (1960).

Figure 2.5: Envelopes of vibrations on the basilar membrane.

end the scala vestibuli is connected to the scala tympani via the *helicotrema* to allow the *perilymph* (the fluid that fills the scala vestibuli and scala tympani) to pass between the two. The scala media is filled with *endolymph* and is separated from the scala vestibuli by the *Reissner's membrane* and from the scala tympani by the *basilar membrane* (figure 2.4b).

As vibrations pass from the middle ear to the inner ear and the stapes presses inwards on the oval window, the round window presses outwards in opposite phase, equalising pressure within the cochlea. This creates a pressure difference across the basilar membrane that generates a travelling wave that passes from base to apex (figure 2.5a). The point along the basilar membrane at which this standing wave reaches its maximum amplitude is frequency dependent: high frequencies reach their maximum nearer the base, whilst low frequencies travel further towards the apex (figure 2.5b). This is because the base of the basilar membrane is stiff and narrow, whereas the apex is soft and wide. It is interesting to note that this is in opposition to the narrowing of the cochlea tube.

The *organ of Corti* runs along the basilar membrane and is responsible for converting vibrations on the basilar membrane to neural signals. There are two groups of hair cells in the organ of Corti: the *inner hair cells* and the *outer hair cells* (figure 2.6). There are approximately 12,000 outer hair cells arranged in three to five rows and a single row of approximately 3500 inner hair cells. On the tips of the hair cells are smaller hair-like structures known as *stereocilia*. Above the stereocilia, in the scala media, lies a gelatinous structure: the *tectorial membrane*.

The endolymph that fills the scala media is rich in potassium K^+ ions whereas the perilymph that fills the scala tympani is relatively low in potassium. This results in a potential difference across the basilar membrane and the organ of Corti. As the basilar membrane vibrates, the stereocilia bend against the tectorial membrane which causes tiny channels in their ends to open and K^+ to flow down the stereocilia. This causes the potential difference to modulate with the vibrations

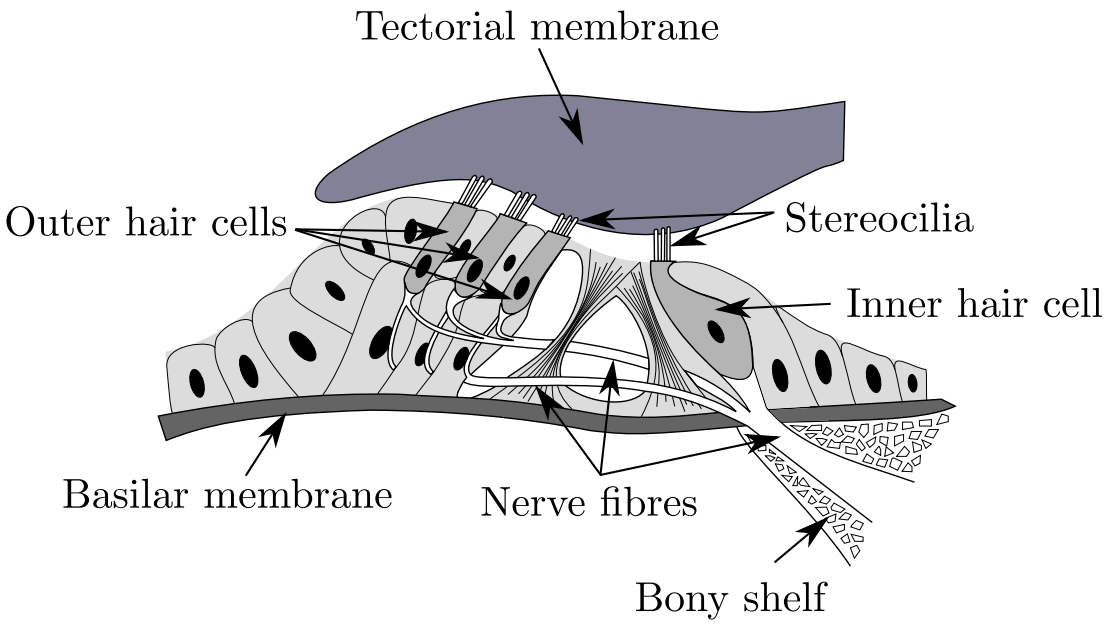


Figure 2.6: The surrounding structure of the organ of Corti. See text for description. After Open University (2018).

of the basilar membrane which in turn triggers the release of neurotransmitters and the firing of neural signals along the auditory nerve.

2.1.2 Frequency limits of human hearing

The frequency range of human hearing is often quoted as 20 Hz–20 kHz (Fastl and Zwicker, 2007; Yost, 2000). However, the upper limit of human hearing varies greatly with age. Some children are able to hear tones as high as 20 kHz (Moore, 2013), but most adults show a steady roll-off in sensitivity above 15 kHz. The degradation of hearing sensitivity with age (presbyacusis (Moore, 2013, p. 62)) affects higher frequencies more, but subject-to-subject variation is also more pronounced at higher frequencies than low (Stelmachowicz *et al.*, 1989).

Takeda *et al.* (1992) measured the upper frequency limits of 6105 ears from subjects aged 5–89 years old. Pure tones were presented over headphones at a level of 75 ± 10 dBSPL and the frequency of the tone reduced until the subject perceived a tone. This was repeated three to five times, depending on the variation in the results, and the limit was taken as the median value. The median results of their study are presented in figure 2.7. They found that the upper frequency limit of

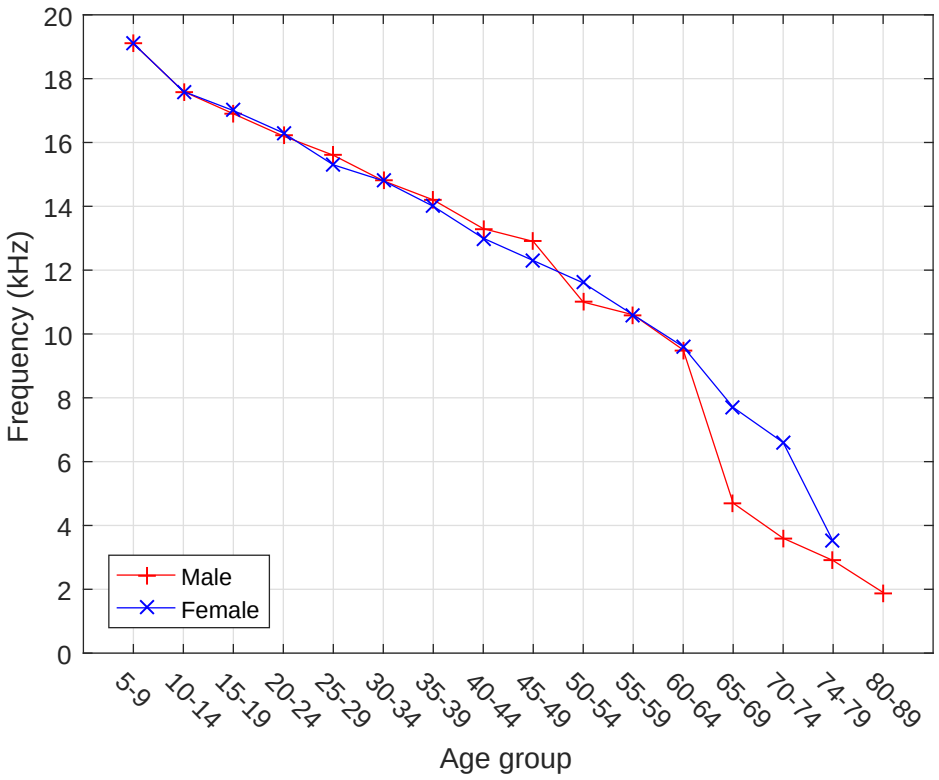
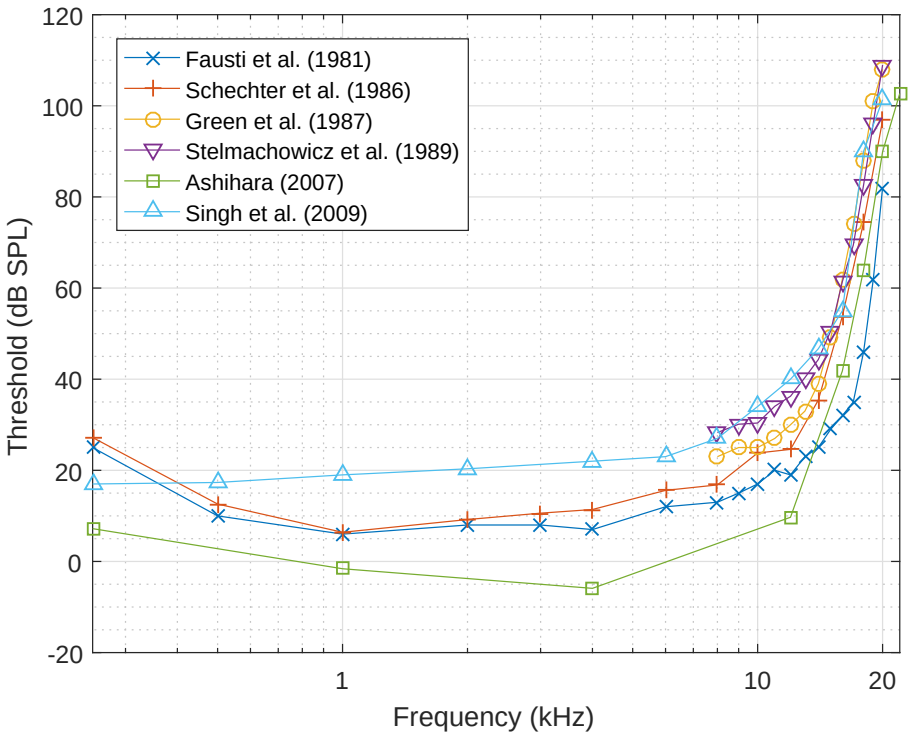


Figure 2.7: Upper frequency limits of hearing for different age groups as reported by Takeda *et al.* (1992)

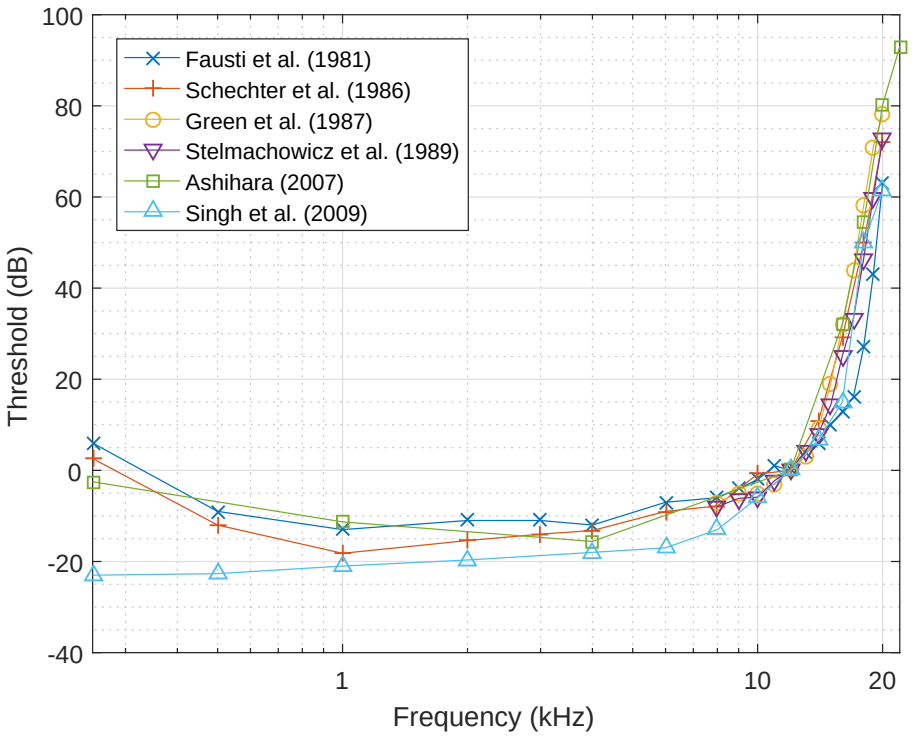
hearing decreased steadily with age up until approximately 45–50 years old, after which it rapidly decreased. They also found that the variation between subjects in upper limit increased with age but found little variation between sexes; except above approximately 60 years old.

Whittle *et al.* (1972) tested the audibility of low frequency sounds and found that frequencies as low as 3.15 Hz could be heard but needed to be presented continuously and at extremely high levels (>120 dBSPL). Johnson and von Gierke (1974) suggested that below 16 Hz the sound itself is not "heard" rather the harmonic distortions in the middle ear are detected.

A number of studies (Fausti *et al.*, 1981; Schechter *et al.*, 1986; Green *et al.*, 1987; Stelmachowicz *et al.*, 1989; Ashihara, 2007; Singh *et al.*, 2009) have investigated the variation in hearing thresholds with frequency and figure 2.8a reports their results. Fausti *et al.* (1981) studied subjects with both normal hearing and those who had been exposed to noise. The curves in figure 2.8 represent the 15 subjects



(a)



(b)

Figure 2.8: Hearing thresholds reported by a number of studies: (a) raw data and (b) normalised to 0 dB at 12 kHz.

aged 18–27 years old with normal hearing from their study. Schechter *et al.* (1986) used a similar procedure to Fausti *et al.* (1981) to measure 157 subjects aged 6–30 years old. The curves in figure 2.8 represent the average across all the age groups they tested.

Green *et al.* (1987) and Stelmachowicz *et al.* (1989) used a high frequency audiometer designed by Stevens *et al.* (1987) to measure the hearing thresholds of 37 18–26 year olds (Green *et al.*, 1987) and 240 10–60 year olds (Stelmachowicz *et al.*, 1989). The curves in figure 2.8 for Stelmachowicz *et al.* (1989) are the results for the 160 subjects aged 10–19 years old.

Ashihara (2007) tested 16 subjects aged 19–25 years old using loudspeakers rather than headphones (used in all the other studies) and Singh *et al.* (2009) measured 50 subjects with normal hearing aged 10–70 years old and both these results are also reported in figure 2.8.

The results from all these studies, as shown in figure 2.8a, vary quite widely. This is probably due to differences in the procedures employed, as suggested by Green *et al.* (1987). However, figure 2.8b shows the results normalised to 0 dB at 12 kHz (a frequency common to all the studies) and it can be seen that, once aligned, the results agree fairly well.

Figure 2.9 shows the average variation in hearing threshold with frequency calculated from the aligned results of the above studies and then normalised to 0 dB at 1 kHz (a common reference point, e.g. the phon). This represents the average of approximately 556 subjects aged 6–70 years old. A common approach to defining the limits of human hearing is to give the range of frequencies audible at 60 dBSPL (Heffner and Heffner, 2007). The 60 dBSPL upper frequency limit calculated from the data in figure 2.9 is 17.86 kHz, which agrees well with other reported values (Jackson *et al.*, 1999; Heffner and Heffner, 2007). The significant reduction in sensitivity of the human hearing system at high frequencies likely explains why little evidence of localisation cues above 16 kHz has been found (section 2.2.4.3).

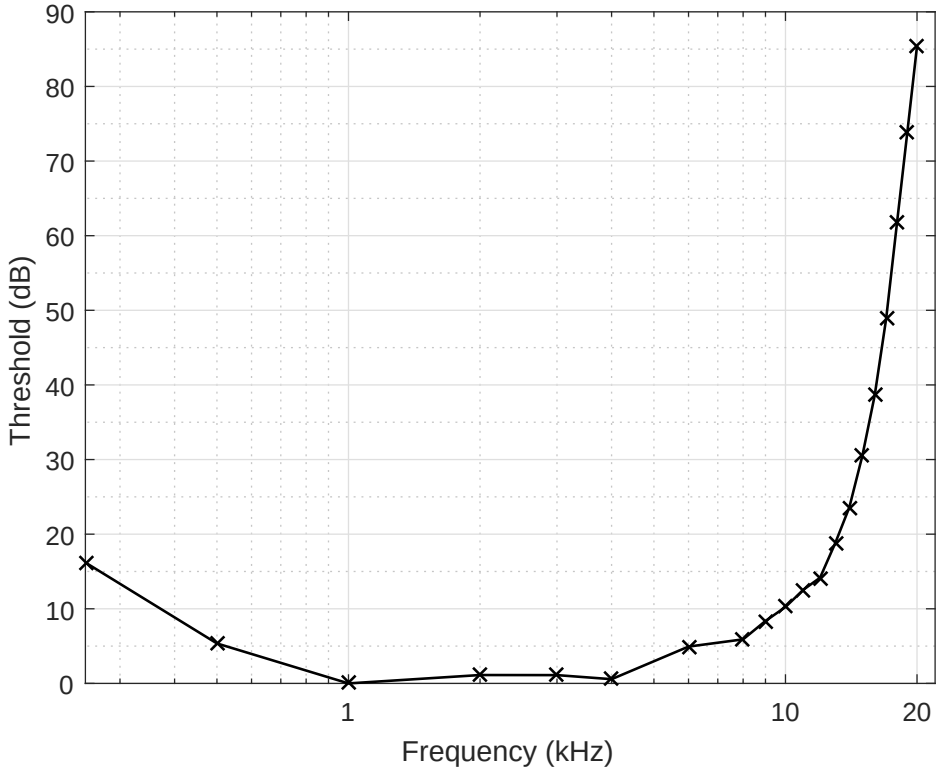


Figure 2.9: Average variation of hearing threshold with frequency calculated from a number of studies (Fausti *et al.*, 1981; Schechter *et al.*, 1986; Green *et al.*, 1987; Stelmachowicz *et al.*, 1989; Ashihara, 2007; Singh *et al.*, 2009) and normalised to 0 dB at 1 kHz.

2.1.3 Auditory filters

2.1.3.1 Auditory filter bandwidth

Each of the inner hair cells reacts most strongly to a different frequency, but also reacts to other frequencies around it. This region of sensitivity is known as the critical bandwidth of the auditory filter and it has been shown to increase with frequency (Moore, 2013; Patterson and Moore, 1986).

There are a number of methods for determining the width of the critical bandwidth and most are based on masking, i.e. the response of the hearing system to one stimulus in the presence of another stimulus, and the idea of the power spectrum model (Moore, 2013). Consider the detection threshold of a sinusoidal signal masked by a narrow band noise signal. The noise is centred on the signal

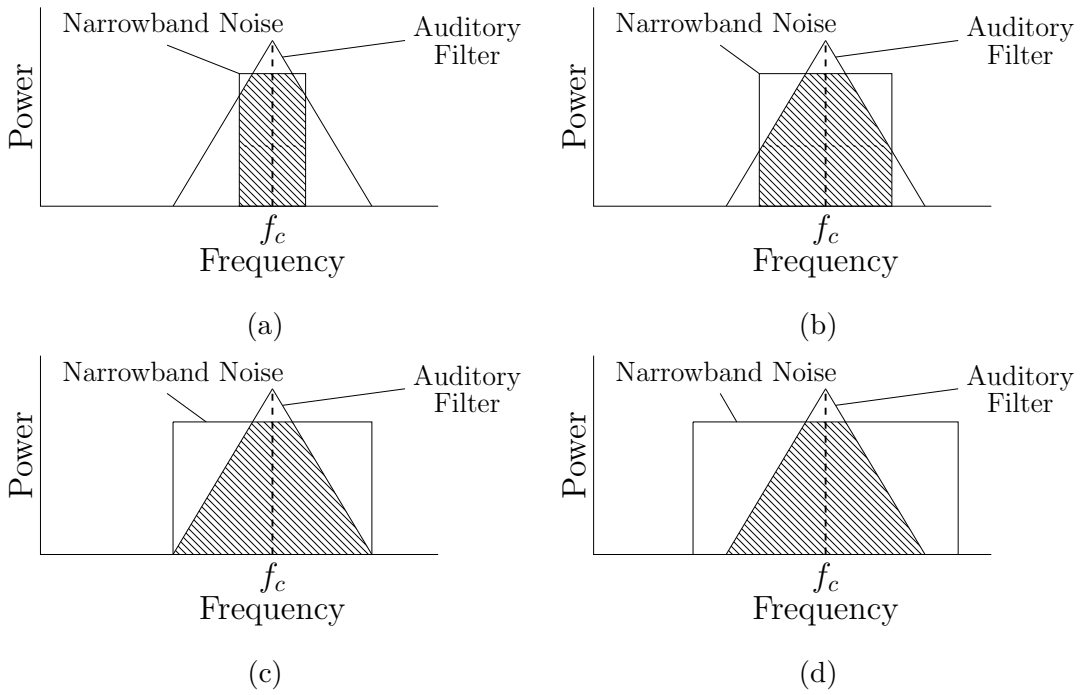


Figure 2.10: Power spectrum model. (a) and (b): as the bandwidth of the narrow band noise centred on f_c increases, so does the noise (dashed section) passing through the auditory filter. Hence, the masking level threshold of the tone at f_c steadily increases. (c): once the bandwidth of the noise reaches the critical band of the auditory filter no more noise can pass through the auditory filter. (d): therefore as the bandwidth of the noise continues to increase, the masking level threshold now remains stable.

frequency f_c, and the noise power density is held constant whilst the bandwidth is increased and so the total noise power increases. The detection threshold of the sinusoidal signal increases as the bandwidth of the noise is increased up to a point after which the threshold remains the same. The assumption is that when listening to a sinusoidal signal in noise one auditory filter is used: one that has a centre frequency closest to the frequency of the signal. Most of the noise is filtered out and only the components that fall within the auditory filter mask the signal. Therefore, as the bandwidth of the noise increases, the amount of noise passing through the auditory filter only increases whilst the bandwidth of the noise is less than the filter bandwidth (figure 2.10).

Early methods of determining the critical bandwidth of the auditory filters consisted of finding the "knee" point of various psychoacoustic tuning curves using a variety of detection tasks (see Fastl and Zwicker (2007) for an in depth description of them). The most common was to use a reference signal of narrow band

noise and to measure the perceived loudness of another narrow band noise signal with varying bandwidth but constant sound pressure level and the same centre frequency as the reference signal. The perceived loudness is measured as a function of bandwidth and is constant up to a point after which it increases steadily (Fastl and Zwicker, 2007). The "knee" point of the loudness curve is taken as the critical bandwidth for the given centre frequency. The critical bandwidths measured using this approach were named Bark bandwidths by Fastl and Zwicker (2007) and are estimated by the following equation (Fastl and Zwicker, 2007):

$$\Delta f_{Bark} = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69} \tag{2.1}$$

where Δf_{Bark} and f_c are both in Hz. As can be seen in figure 2.11, critical bandwidth is roughly constant (approximately 100 Hz) below 400 Hz after which it increases logarithmically with frequency, with a width of several kHz at higher frequencies.

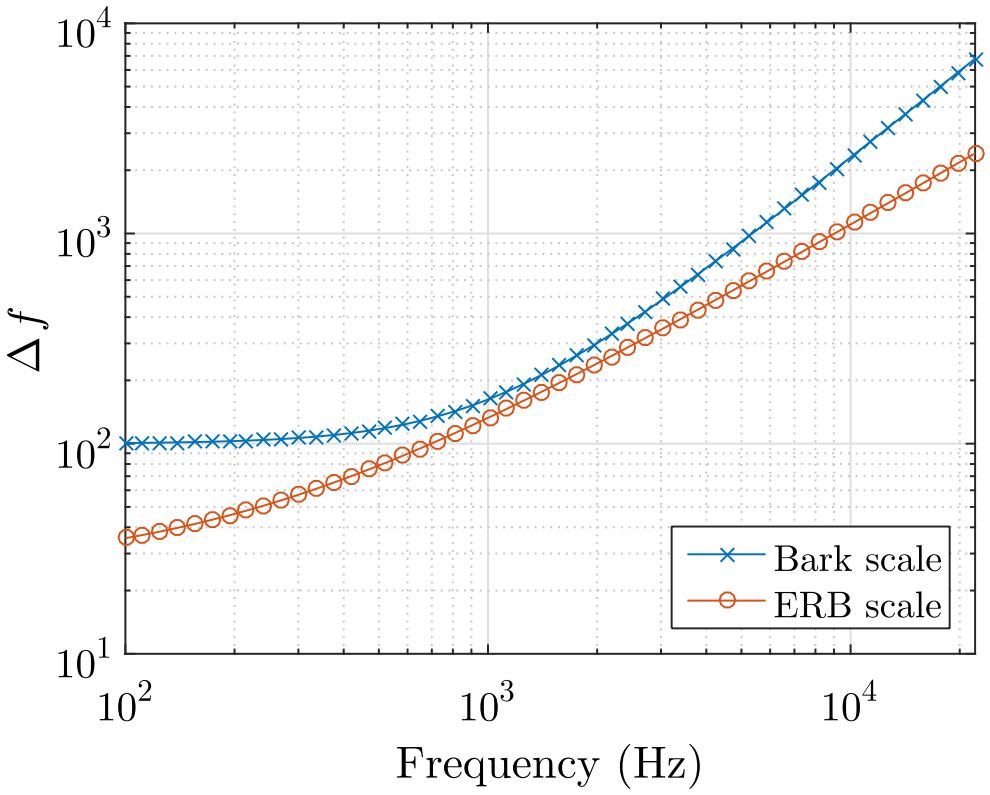


Figure 2.11: Comparison of the Bark and equivalent rectangular bandwidth (ERB) scales for auditory filter bandwidth.

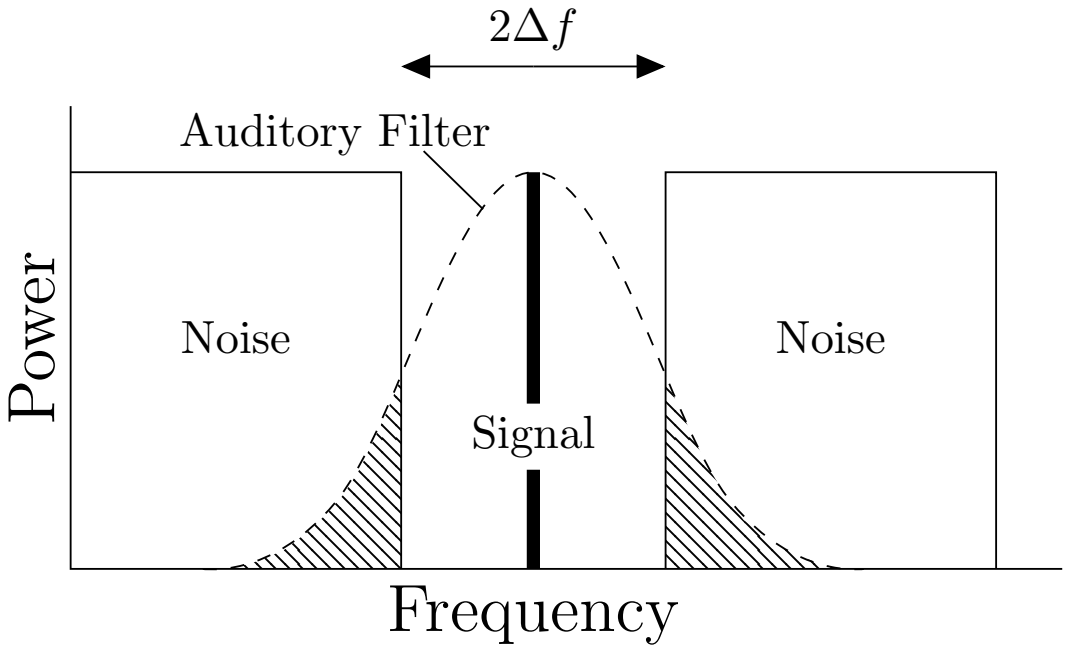


Figure 2.12: Notched noise method of auditory filter width measurement. See text for explanation. After Moore (2013, p. 74, Fig. 3.3).

More recently the Bark scale has generally been rejected in favour of the equivalent rectangular bandwidth (ERB) scale (Glasberg and Moore, 1990). Rather than using a single masker the ERB scale is measured using notched noise which has the advantage that auditory filters above and below the frequency under test cannot be used to detect the signal: a phenomenon called "off-frequency listening" (Moore, 2013; Pulkki and Karjalainen, 2015). In the notched noise method (Patterson, 1976) a sinusoidal signal is masked by narrow band noise maskers either side of it (figure 2.12) and the detection threshold of the signal is measured as a function of the notch width ($2\Delta f$). This is then used to approximate the shape of the auditory filter. The equation for the ERB, Δf_{ERB}, is:

$$\Delta f_{ERB} = 24.7 + 0.108 f_c \tag{2.2}$$

where Δf_{ERB} and f_c are both in Hz. Note this is a slight reformulation of the standard equation for ERB where f_c is usually in kHz. As can be seen in figure 2.11 the bandwidth Δf_{ERB} varies approximately logarithmically with frequency over a larger range than the Bark scale.

2.1.3.2 Variation of auditory filter shape with level

The shape of the auditory filter varies with sound pressure level for a given frequency (Glasberg and Moore, 2000) as shown in figure 2.13. As the sound level increases the auditory filter broadens significantly, especially for the low-frequency skirt. The reason for this variation is generally explained in terms of the auditory filter consisting of an active portion (the cochlea) that results in the tip of the filter and a passive portion (the basilar membrane) that results in the tails of the filter (Moore, 2013). At low levels the active part dominates and this results in a sharp peak in the auditory filter. As the sound level increases the gain of the active part decreases and the passive portion dominates, widening the skirts of the filter.

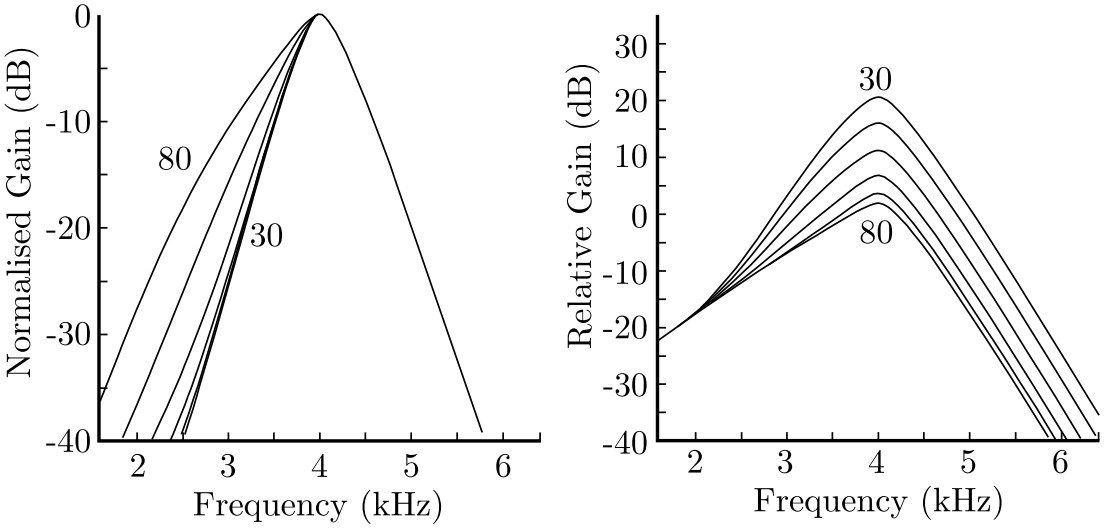


Figure 2.13: Shape of auditory filter centred at 4 kHz with variation of sound level from 30–80dB SPL. The left panel shows the filters normalised to an amplitude of 0dB at the filter's tip. The right panel shows their relative amplitude with a peak gain of 0 dB for a sound level of 80 dB. After Glasberg and Moore (2000).

2.2 Acoustic sound localisation cues and localisation acuity

There are a number of acoustic cues used to localise sounds and they are represented in an individual's head-related transfer function (HRTF) or head-related impulse response (HRIR). They are generally divided into two main categories: interaural and spectral. Sometimes interaural cues are referred to as binaural cues as they rely on the signals arriving at both ears and spectral cues as monaural cues as they work under monaural listening. However, the use of interaural spectral difference (ISD) has also been investigated (Jin *et al.*, 1999, 2004) and whilst their importance is unconfirmed, it is possible that spectral cues operate under both binaural and monaural listening. Therefore in this work, to avoid ambiguity, cues will be classified as either interaural or spectral rather than binaural and monaural, and whilst this work focuses on spectral cues, both interaural and spectral cues will be reviewed here.

2.2.1 Auditory space

Auditory space is defined as "The totality of all possible positions of auditory events" where "auditory events" are any auditorily perceived phenomena, with or without physical origin (Blauert, 1997). The origin in auditory space is generally defined as the point equidistant between the two eardrums (Moore, 2013; Blauert, 1997; Pulkki and Karjalainen, 2015); unless a subspace of auditory space is being studied e.g. just the region around the pinnae (Kahana and Nelson, 2007). In terms of Cartesian coordinates the x axis passes through the head from left to right, the y axis passes through the head from back to front and the z axis passes through the head from bottom to top. Figure 2.14 illustrates the three anatomical planes that divide up the auditory space. Firstly the *median* (or *mid-sagittal*) plane divides the auditory space into left and right portions and in Cartesian space is the y–z plane. Secondly the *frontal* (or *coronal*) plane divides

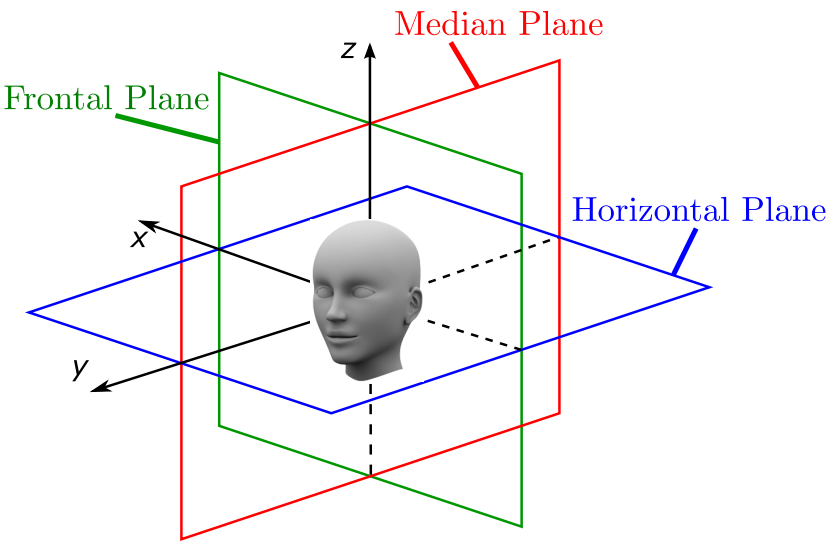


Figure 2.14: Anatomical planes[2].

the auditory space into back and front portions and is the x–z plane in Cartesian space. Finally the *horizontal* (or *transverse*) plane divides the auditory space into top and bottom portions and in Cartesian space is the x–y plane. Any one of the infinite number of planes parallel to the median plane is referred to as a sagittal plane.

2.2.2 Coordinate systems

Due to the head being roughly spherical it is perhaps more intuitive to use spherical coordinate systems rather than Cartesian coordinates. The coordinates in spherical coordinate systems are range/distance (r), azimuth (θ) and elevation (φ). Whilst there has been research into how head-related transfer functions vary with range, e.g. (Duda and Martens, 1998; Kan *et al.*, 2006, 2009; Spagnol, 2015), source distance perception will not be considered in this work and therefore the location of a sound will only be defined in terms of azimuth and elevation.

There are two spherical coordinate systems used in spatial audio research and each has its own benefits. Figure 2.15 illustrates the interaural-polar coordinate system which has been used by some authors, most notably in the creation of

[2]Image of head from `http://www.secondpicture.com/tutorials/3d/modeling_human_head_in_3ds_max.png`

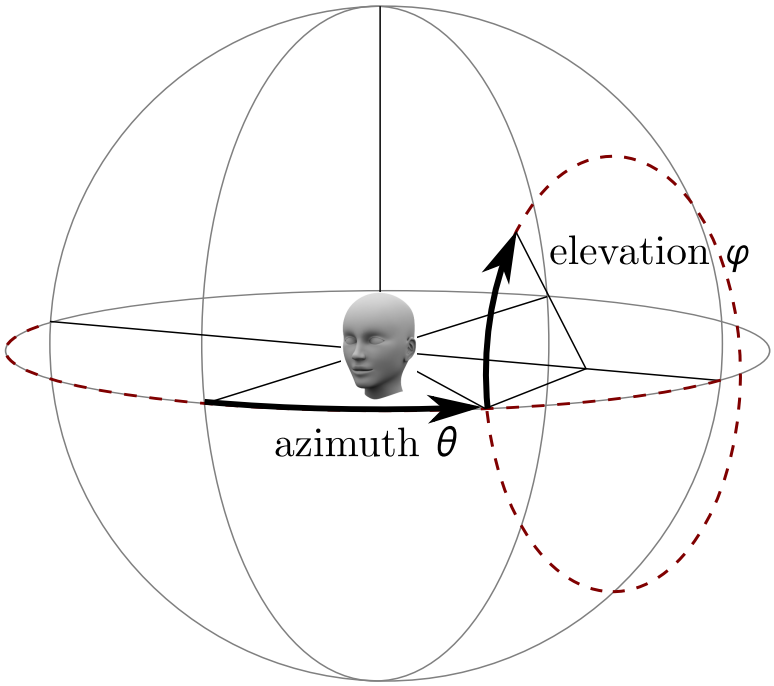


Figure 2.15: The interaural-polar spherical coordinate system. The dotted lines indicate the possible ranges of azimuth and elevation. After Thorpe (2009).

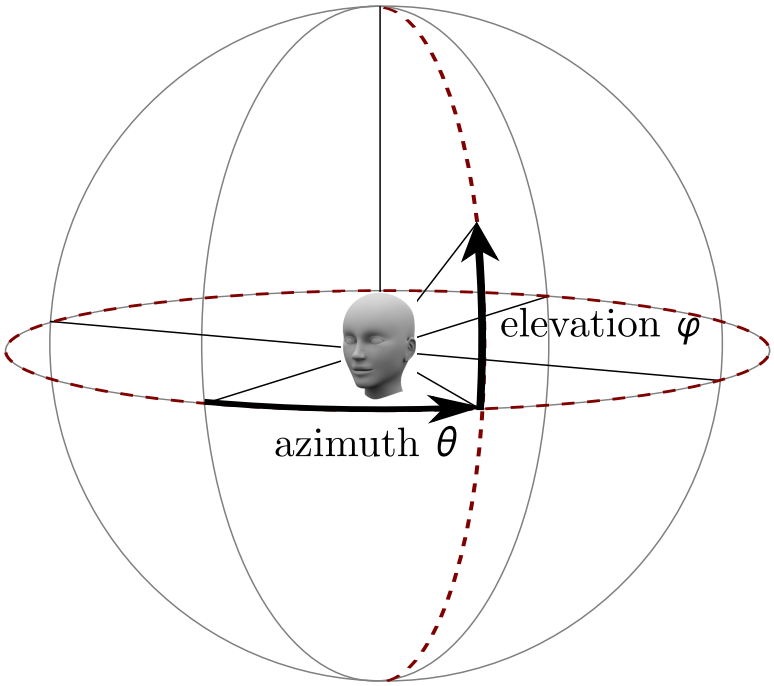


Figure 2.16: The vertical-polar spherical coordinate system. The dotted lines indicate the possible ranges of azimuth and elevation. After Thorpe (2009).

the Center for Image Processing and Integrated Computing (CIPIC) database (Algazi *et al.*, 2001d). Azimuth (θ) varies from –180° (directly behind) to 180° (also directly behind) with 0° lying on the median plane and elevation (φ) varies from –90° (directly below) to 90° (directly above). One of the advantages of the interaural-polar coordinate system is that fixing azimuth and varying elevation traces a section through a "cone-of-confusion" (see section 2.2.3) along a sagittal plane.

The most widely used coordinate system relating to binaural acoustic measurement is the vertical-polar spherical coordinate system (figure 2.16). Azimuth (θ) varies from –180° (directly behind) to 180° (also directly behind) with 0° being directly in front and generally negative azimuths lying in the left hemisphere. Elevation (φ) varies from –90° (directly below) to 90° (directly above). In this system fixing azimuth and varying elevation traces an arc along a plane that passes through the z-axis. This research will use the vertical-polar coordinate system due to the fact that it is simple to convert between Cartesian and spherical coordinates.

2.2.3 Interaural cues — ILD and ITD

Interaural cues occur due to disparity in the audio signals reaching the two ears. The interaural time difference (ITD) — figure 2.17 — arises due to differences in the path length for sound to reach the ipsilateral (near-side) ear compared to the contralateral (far-side) ear; this results in a delay in the signal reaching the contralateral ear. The interaural level difference (ILD) — figure 2.18 — occurs due to shadowing of the contralateral ear by the head; the sound level at the contralateral ear is attenuated.

The ITD and ILD have been well understood for over a century, having been first analysed by John Strutt — Lord Rayleigh — who proposed the "duplex theory" of sound localisation (Rayleigh, 1907). The duplex theory is based on the idea that the ITD and ILD work across different frequency ranges. The

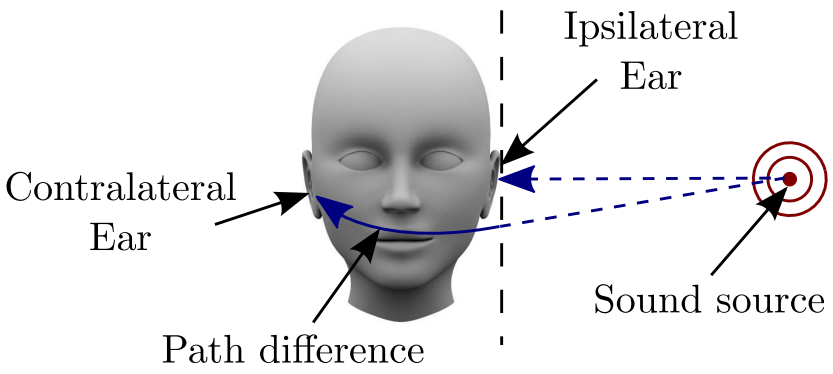


Figure 2.17: Interaural time difference (ITD). After Thorpe (2009).

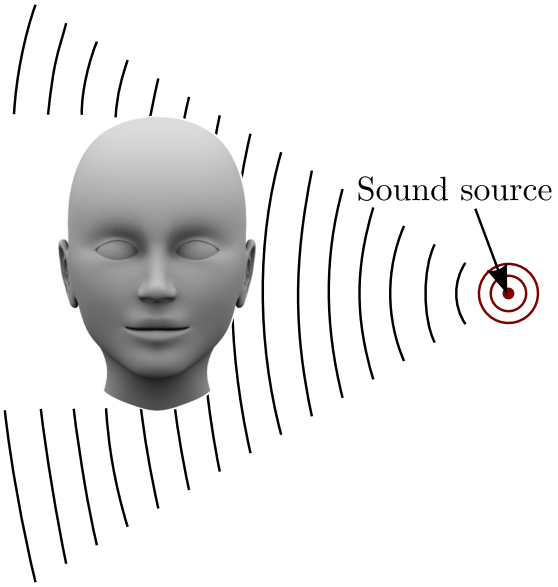


Figure 2.18: Interaural level difference (ILD). After Thorpe (2009).

ITD is also referred to as the interaural phase difference (IPD) as for pure tone stimuli it is detected as phase differences in the signals arriving at the two ears (Blauert, 1997; Moore, 2013). For frequencies below approximately 700–800 Hz these phase differences are completely unambiguous, as the half-period of the sound is less than the maximum ITD (Blauert, 1997). For higher frequencies the phase difference introduced by the ITD is more than 180° which introduces ambiguity and for frequencies greater than 1.5–1.6 kHz the phase differences are completely ambiguous (Blauert, 1997; Moore, 2013; Pulkki and Karjalainen, 2015). In contrast the ILD is only dominant at frequencies above approximately 4 kHz where the wavelength is significantly smaller than the head and hence it shadows the contralateral ear. At lower frequencies the sound waves diffract round the head and so the difference in level of the sound arriving at the two ears is far

smaller. Stevens and Newman (1936) used pure tone stimuli to test the duplex theory and found that there was a dip in azimuthal localisation performance between 2 and 4 kHz. This corresponds to the frequency range that is too high for ITD cues but too low for ILD cues.

There are a number of approaches to estimating ITD from HRTFs that Katz and Noisternig (2014) split into three "families" according to their basic premise. The first family uses onset detection to calculate the main peak of the left and right HRIRs and estimates the ITD as the time difference between the main peaks (Kuhn, 1977). The second family estimates the ITD as the time lag for maximum cross correlation between left and right HRIRs (Kistler and Wightman, 1992). The final family uses excess phase components to estimate the overall group delay and translates this into a time delay (Jot *et al.*, 1995; Minnaar *et al.*, 2000). Katz and Noisternig's (2014) investigation suggests that the variation in calculated ITDs using the different methods can exceed just-noticeable-differences (JNDs) for ITD by several times. However, whilst the work begins to analyse the differences between the methods, they conclude that further work is required to analyse fully the importance of the differences and which method is optimal. The estimation of ILD is less problematic and is generally calculated as the logarithmic ratio of energy in the left and right HRTF, either averaged across all frequencies or in separate frequency bands (Gaik, 1993).

Interaural cues are widely accepted as the predominant cues for azimuthal (left–right) localisation. However, they are ambiguous within the surface of a series "cones of confusion" (Mills, 1972) where the paths to the two ears are the same length (figure 2.19). Because the path lengths are the same, the values for ITD and ILD remain substantially constant and due to this ambiguity the the ear is unable to accurately identify the location of the sound source. It has therefore been suggested that the ear relies on spectral cues to disambiguate and accurately locate sound sources in 3D space.

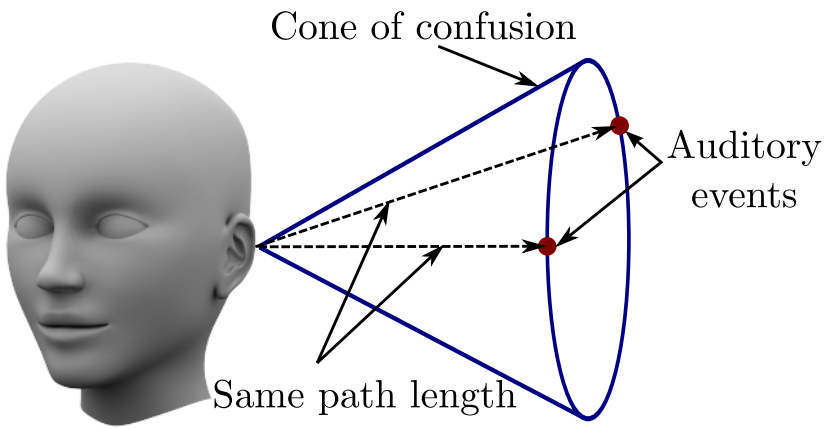


Figure 2.19: Cone of confusion where the interaural cues are the same. The two auditory events indicated produce the same ITD and ILD.

2.2.4 Spectral cues

The morphology of the human body imparts direction-dependent filtering on the signals reaching the ears. This is believed to be the main cue in resolving ambiguity in interaural cues. Batteau (1967) initially proposed time-domain filters for the pinnae cues based on two delayed reflections from the concha and helix acting as cues for azimuth and "altitude" (i.e. elevation) respectively. However, the features of the pinnae are relatively small compared to the wavelength of audible sound and accordingly time-domain, reflection-based models cannot effectively model the complex effects of the pinnae which consists of diffraction as well as reflection (Lopez-Poveda and Meddis, 1996; Blauert, 1997; Xie, 2013). Therefore, spectral localisation cues are now generally studied in the frequency domain, where they can be more effectively modelled (Shaw and Teranishi, 1968; Teranishi and Shaw, 1968; Shaw, 1974b, 1997; Hebrank and Wright, 1974).

In the time-domain the direction-dependent filters of an individual's morphology are known as a head-related impulse responses (HRIRs) and their corresponding frequency-domain representation are known as head-related transfer functions (HRTFs). Sometimes HRTFs are split into the direction-dependent components, the so called directional transfer functions (DTFs), and the direction-independent component, the common transfer function (CTF) Middlebrooks and Green (1990); Middlebrooks (1999). Commonly, the CTF is calculated by com-

bining the average of the magnitude spectra of the HRTFs for all available directions with the corresponding minimum phase spectrum of that average, calculated via the Hilbert transform Middlebrooks and Green (1990). The DTFs are then calculated by dividing the complex HRTF for each direction by the complex CTF.

It is believed that, in addition to the pinnae, the morphology of the human head and torso also contribute to the HRTFs, especially for elevation discrimination (Avendano *et al.*, 1999a; Algazi *et al.*, 2001a; Jin *et al.*, 2013; Kirkeby *et al.*, 2007). However, these cues operate in a different frequency range to the pinnae cues and so generally head-and-torso cues and pinnae cues are studied independently and as such will be reviewed separately here.

2.2.4.1 Pinna cues

The contribution of the pinna to the HRTF is commonly referred to as the pinna-related transfer function (PRTF) and can be split into pinna resonances and pinna reflections (Satarzadeh *et al.*, 2007; Spagnol *et al.*, 2010).

The cavities of the pinnae, predominantly the concha, act as resonant chambers which are excited to different degrees depending on the frequency and direction of the sound source. Shaw (1997) reports six modes of resonance in the human concha, see table 2.1, measured using a point source and probe microphone. The frequencies and "angles of incidence" (equivalent to elevation in the vertical-polar spherical coordinate system) reported in the table are those for which the

Mode	Frequency	Angle of Incidence	Type
1	4.2 kHz	—	omnidirectional monopole
2	7.1 kHz	68°	vertical dipole
3	9.6 kHz	73°	vertical dipole
4	12.1 kHz	-6°	horizontal dipole
5	14.4 kHz	7°	horizontal dipole
6	16.7 kHz	7°	horizontal dipole

Table 2.1: Concha modes reported by Shaw (1997). Angle of incidence is equivalent to angle of elevation in the vertical-polar spherical coordinate system (section 2.2.1).

resonances are maximal. These results have been corroborated by other studies (Geronazzo *et al.*, 2010; Spagnol *et al.*, 2010).

Destructive interference generated by reflections from surfaces of the pinnae cause notches in the PRTF. Hebrank and Wright (1974) identified a notch between 4 and 8 kHz as a key cue for elevation discrimination which has been identified by other studies (Butler and Belendiuk, 1977; Mehrgardt and Mellert, 1977; Kahana and Nelson, 2007; Iida, 2008). They suggested a "single-delay-and-add" mechanism consisting of a reflection from the back wall of the concha was responsible for generating the notch. Spagnol *et al.* (2013) studied spectral notches in the HRTFs of the CIPIC database. They found that all but two subjects out of twenty, who only exhibited two, had three spectral notches in their HRTFs.

Morimoto (2001) used pinna occlusion to investigate the contribution of the ipsilateral and contralateral pinna spectral cues to localisation in sagittal planes. They found that while both ears contribute to vertical localisation, the contribution of the contralateral ear decreases as the sound source moves away from the median plane until, for sagittal planes more than 60° from the median plane, the contralateral ear no longer contributes measurably.

Section 2.5 will give a deeper insight to the morphological origin of the resonances and notches of the PRTF.

2.2.4.2 Head and torso cues

Musicant and Butler (1984) found that in their localisation tasks, even with the pinnae occluded, their subjects retained the ability to localise 4 kHz low-pass noise stimuli and suggested the cues might originate from the torso. A number of other studies have demonstrated that, as well as the pinnae, diffraction of sound around the head and reflections from the shoulders and torso also contribute to the HRTFs (Avendano *et al.*, 1999a; Algazi *et al.*, 2001a; Jin *et al.*, 2013; Kirkeby *et al.*, 2007).

Algazi *et al.* (2001a) also found that subjects retained the ability to estimate the elevation of low frequency sounds, where pinnae cues are generally thought to be absent, surprisingly well; especially for sources located well outside the median plane. They also found that simple spherical head and ellipsoidal torso geometric models estimated the low-frequency head and torso cues very well which has been corroborated in a follow up study (Algazi *et al.*, 2002).

Jin *et al.* (2013) investigated the effect of the torso during the generation of the Sydney-York morphological and recording of ears database (SYMARE database) database. They compared simulated HRIRs for head-only and head-and-torso meshes and found an absence of known key features in the spatial frequency response surfaces (SFRS) of the head-only meshes. They also calculated the correlation between head-only and head-and-torso SFRS for frequencies up to 5.6 kHz (figure 2.20). It can be seen that below approximately 800 Hz there is little difference, then between 1–3 kHz there is a significant decrease in correlation and then the correlation steadily increases again from 3–5 kHz. They suggest that the significant differences between head-only and head-and-torso SFRS in the 1–3 kHz range is due to the fact that the distance between the shoulders and ears is in the order of the wavelength within this frequency range.

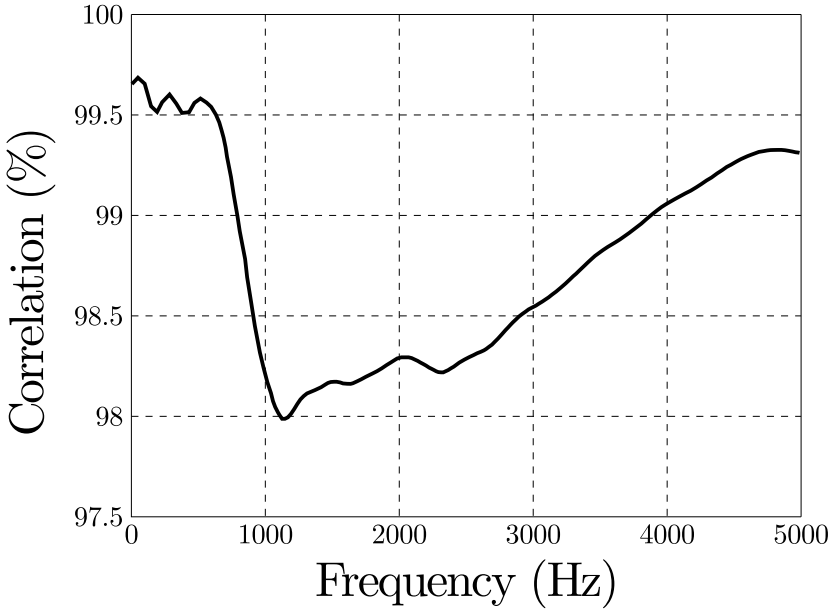


Figure 2.20: Correlation between spatial frequency response surfaces (SFRS) for head-only and head-and-torso meshes. After Jin *et al.* (2013).

Kirkeby *et al.* (2007) investigated the effect of the torso using the ultra weak variational formulation finite element method. They used a head-and-torso simulator (HATS) and simulated 21,872 directions over the frequency range 50 Hz–24 kHz. They found that the shoulder reflections can be separated in time from the rest of the HRIR for sources directly above the listener and that the impedance of the torso has a significant effect for sources directly below. They also concluded that the spectral cues generated by the torso for resolving front-back confusions were relatively weak.

2.2.4.3 Operating frequency ranges

A number of studies have used selective pinna cavity occlusion, in which the shape of the cavity is modified or even removed by filling with putty (or similar material), to investigate the frequency ranges of spectral cues. Gardner and Gardner (1973) found that increased occlusion led to a degradation in median plane localisation acuity and that localisation was best for broadband or high frequency stimuli in the frontal region. Musicant and Butler (1984) investigated the effect of pinna occlusion on localisation of different stimuli: broadband, 1 kHz low-pass, 4 kHz low-pass and 4 kHz high-pass noise. They found that occlusion affected localisation most when high frequencies were included in stimuli but had little effect for low frequency stimuli; leading to their conclusion that pinnae cues are high frequency in origin.

Hebrank and Wright (1974) examined the localisation of band-limited noise to investigate spectral cues and concluded that the cues required for accurate median-plane localisation are contained in the frequency range 3.8–16 kHz. They also studied specific localisation cues and suggested that:

- frontal directions are indicated by a one octave notch between 4 and 8 kHz and a boost of frequencies above 13 kHz
- directions above are indicated by a 1/4 octave peak between 7 and 9 kHz

- directions behind the listener are indicated by a small peak between 10 and 12 kHz with frequencies cut above and below it
- increases in elevation in the frontal direction are indicated by an increase in the lower cut off frequency of the one octave "frontal" notch

These conclusions agree with earlier work by Blauert (1970) using 1/3-octave band noise.

Asano *et al.* (1990) investigated the role of micro and macroscopic variation of high and low frequencies in median plane localisation. They concluded that microscopic details below 2 kHz and macroscopic high frequency details are key cues for front-rear discrimination. They also concluded that macroscopic details above 5 kHz are important elevation cues.

2.2.5 Localisation acuity

2.2.5.1 Localisation acuity - azimuth

Stevens and Newman (1936) carried out the first thorough investigation of horizontal localisation acuity using pure tone stimuli and their mean results, across frequency, are reported in table 2.2. They found that localisation error did not vary greatly, with frequency, from these means except for a dip in acuity between 2 and 4 kHz — as mentioned in section 2.2.3 and shown in the mean, across directions, plotted in figure 2.21.

Source Azimuth	Horizontal Localisation Error
0°	4.6°
15°	13°
30°	15.6°
45°	16.3°
60°	16.2°
75°	15.6°
90°	16°

Table 2.2: Horizontal localisation error as reported by Stevens and Newman (1936). The values reported are the means across frequency.

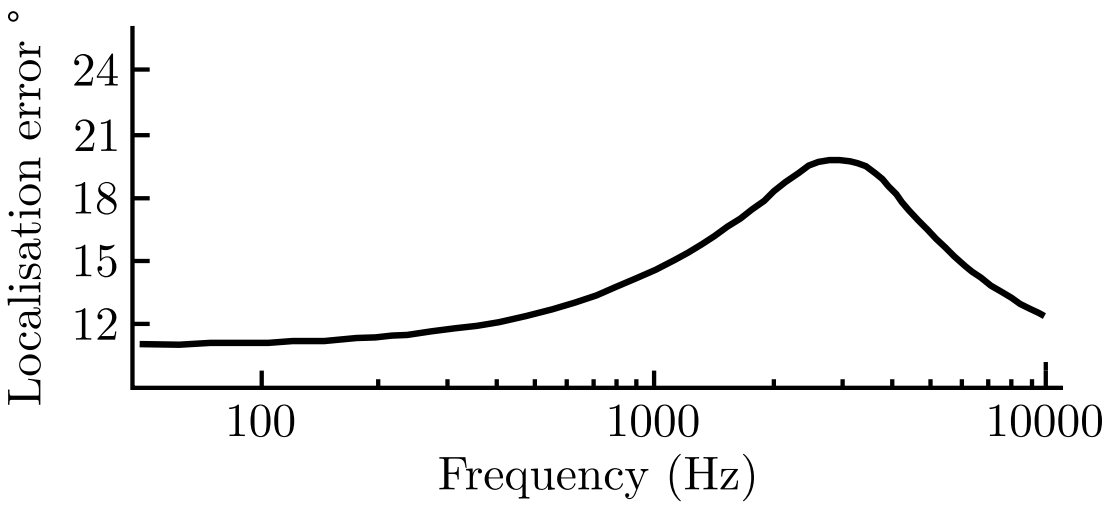


Figure 2.21: Frequency dependency of horizontal localisation error as reported by Stevens and Newman (1936). The plotted curve is the mean across all directions. After Stevens and Newman (1936).

Mills (1958) coined the term minimum audible angle (MAA) as a measure of the minimum detectable difference in horizontal location of two identical sound sources. Mills found that the average MAA for 500 Hz and 1 kHz stimuli varied from approximately 1° for straight ahead to in excess of 10° at extreme source directions (>75°) and there was also a dip in acuity in a similar frequency region to Stevens and Newman. Subsequent investigations (Perrott, 1984; Perrott and Saberi, 1990) have come to conclusions similar to the earlier studies: the azimuthal MAA can be as small as 1° directly in front and are maximum at extreme source azimuth angles.

The minimum audible movement angle (MAMA) is defined as the minimum angle a moving source must move to be discernible from a stationary source or a source moving in the opposite direction. It has been investigated in a number of studies (Perrott and Musicant, 1977; Grantham, 1986; Perrott and Tucker, 1988; Chandler and Grantham, 1992). Investigations have found that the MAMA is only slightly larger than the MAA (approximately 5°) for slow-moving stimuli (Grantham, 1986) but they increase with increasing velocity of the stimulus, increasing frequency of stimulus and decreasing stimulus duration. Perrott and Tucker (1988) found that the MAMA is smaller for lower frequency stimuli (below 1 kHz) and Chandler and Grantham (1992) found that they are also smaller

Source Elevation	Vertical Localisation Error
0°	±9°
30°	±10°
74°	±13°
112°	±22°
153°	±15°

Table 2.3: Vertical localisation error of familiar speech in the median plane as reported by Damaske and Wagener (1969)

for larger bandwidth stimuli. Recently Brimijoin and Akeroyd (2014) found that the MAMA is approximately 1–2° smaller when the subject is moving compared to when the source is moving.

2.2.5.2 Localisation acuity - elevation

Blauert (1997) reports the results of a number of studies into vertical elevation localisation acuity in the forward direction: Blauert (1970) himself found that vertical localisation error for unfamiliar speech was 17°; Damaske and Wagener (1969) reported 9° for continuous familiar speech; Wettschurek (1971) reported 4° for white noise. The results that Damaske and Wagener (1969) found for different angles in the median plane, as reported by Blauert (1997), are shown in table 2.3.

Perrott and Saberi (1990) investigated elevation localisation acuity using an array of speakers that could be rotated through a number of oblique planes. They found that for planes 10°–60° from the horizontal there was little variation in MAAs from the 0.97° recorded for straight ahead in the horizontal plane. It was not until the array was very close to the vertical (80° or greater) that there was any great deviation — 1.8° for 80° and 3.65° for 90° as shown in figure 2.22.

2.3 Spatial audio

Initially sound reproduction equipment consisted only of one channel (monophonic) and spatial attributes were significantly restricted. However as early as

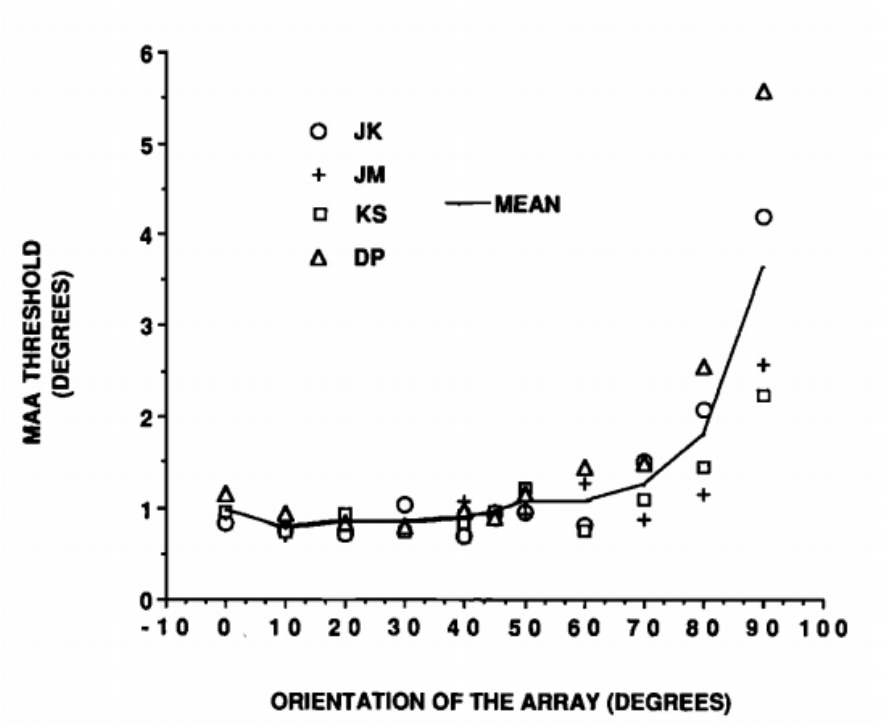


Figure 2.22: Minimum audible angle (MAA) as a function of loudspeaker array orientation as reported by Perrott and Saberi (1990). Image from Perrott and Saberi (1990).

1881 (Rumsey, 2001) people were experimenting with using more channels to create a more realistic sound scape and two-channel stereo has been the standard for audio since the 1960s. Some more recent spatial audio systmes are capable of recreating a 3D soundfield that gives the perceptual impression of a real soundfield experienced under normal listening conditions. Developments in spatial audio research since then have largely been split into two factions: improving the psychoacoustic accuracy of two channel audio and using an increased number of channels to improve realism. Whilst this research is focussed on the former, it may well have implications for the latter, and so advancements in both will be summarised here.

2.3.1 Stereo loudspeaker reproduction

The earliest documented stereophonic sound transmission is generally accepted to be Clement Ader's demonstration at the Paris exhibition in 1881 (Rumsey, 2001). He used multiple telephone pickups mounted across the front of the stage

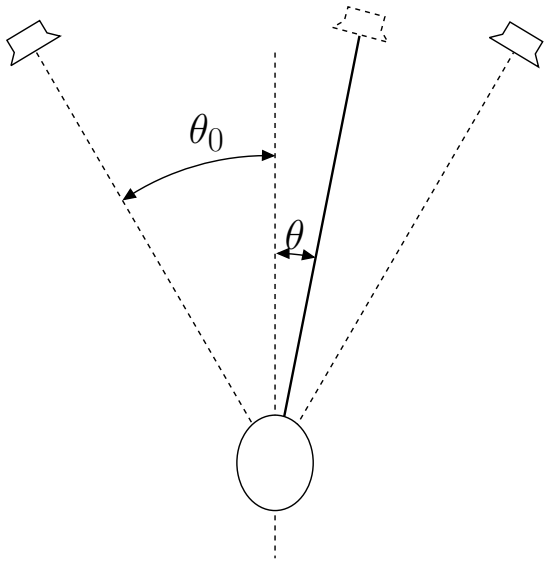


Figure 2.23: Stereo loudspeaker listening configuration. θ_0 is normally 30° and θ is the angle of the virtual source.

of the Paris Opera to transmit signals to pairs of telephone receivers that listeners could listen through at the nearby exhibition. However, it was not until the 1930s that modern stereophonic sound was patented by Blumlein (1933).

A virtual sound source can be placed anywhere between the loudspeakers using amplitude panning, which consists of changing the relative gains of a signal fed to the two loudspeakers. The standard arrangement places the loudspeakers and listener at the vertices of an equilateral triangle, as shown in figure 2.23. Since the sound from each loudspeaker arrives at both ears there is crosstalk between the signals, i.e. sound from the left loudspeaker reaches the right ear and vice versa. This manifests as phase differences below about 1 kHz that are consistent with ITD cues and level differences above 2 kHz that are consistent with ILD cues (Blauert, 1997). For a virtual source located at angle θ from the centre the relative gains (where g_1 is the gain of the ipsilateral loudspeaker and g_2 the gain of the contralateral loudspeaker) are governed by the tangent law:

$$\frac{\tan\theta}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{2.3}$$

Normalisation is then applied to make the perceived loudness of the virtual source

independent of position:

$$\sum_{n=1}^{N} g_n^2 = 1 \tag{2.4}$$

Some restrictions of standard two-channel stereo are that the virtual sources remain stable only for a relatively small "sweet-spot" approximately equidistant from the two loudspeakers. Movement away from this position results in the nearest loudspeaker "pulling" the virtual source towards it. Another downside is that virtual sources panned to the centre of the loudspeaker pair suffer from a certain amount of timbral colouration due to comb-filtering. Furthermore central sources are perceived using the wrong HRTFs which may also contribute to a degradation in their realism.

Some of these problems can be overcome by including an extra centre loudspeaker. The principle of stereo has a long history: beginning initially with the work of Steinberg and Snow (1934) at Bell laboratories in the 1930s, later becoming commonplace in cinematic surround sound systems before seeing a brief resurgence in the early 1990s (Gerzon, 1992a,b). The additional channel in a stereo set-up has a number of benefits: wider range of listening positions, more stable stereo image and less timbral effect on centre sound sources due to the real centre source. However, outside of home cinema set-ups, including a centre loudspeaker in home loudspeaker set-ups is generally inconvenient and three channel stereo systems have faced little commercial success.

2.3.2 Stereo headphone reproduction

When stereo signals meant for playback over loudspeakers are played back over headphones there is no crosstalk between the channels: the left channel only goes to the left ear and the right channel only to the right ear. This means that there are none of the ITD cues present that are normally experienced with stereo loudspeaker listening, only the ILD cues. In practice this gives a relatively similar spatial image; however, it results in in-head localisation of virtual sources panned between the left and right channels and little sense of "space".

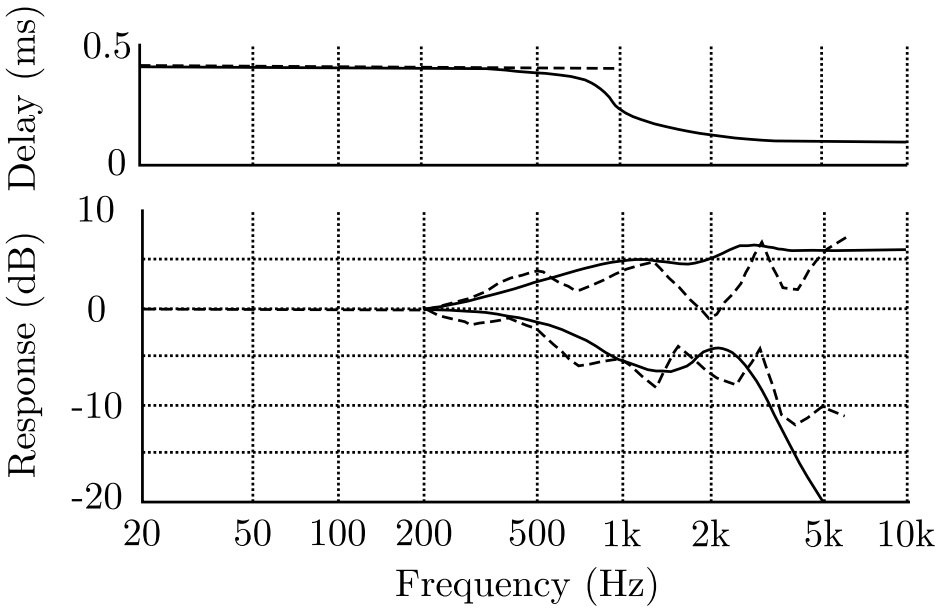


Figure 2.24: Bauer's filter design for loudspeaker signal playback over headphones. The upper panel shows the delay introduced to the crossfeed between channels. The lower panel shows the relative responses of the near ear (upper lines) and far ear (lower lines). Dashed lines show design objectives, solid lines the actual filter responses. After Bauer (1965).

Bauer (1961a,b, 1965) suggested introducing crosstalk between the left and right channels in headphone listening using analogue circuitry to mimic the delay and level differences experienced under loudspeaker listening. His design objectives were based on the measurements of Wiener (1947) which can be seen, along with the circuitry's actual performance in figure 2.24. He also suggested reverse processing for playing binaural measurements over loudspeakers (section 2.3.6).

2.3.3 Surround sound loudspeaker systems

Introducing additional loudspeakers around the listener to increase the sense of envelopment has a history stretching back to the 1970s when a number of four-channel quadrophonic set-ups were proposed. However, these systems suffered from poor spatial quality and experienced little commercial success due to a lack of a standardised quadraphonic system and incompatibility between the different systems. In the late 1970s Dolby used a four channel surround system with three channels across the front plus a mono surround channel for the films "A Star Is

Born" and "Star Wars" (Davis, 2003). This eventually developed into the 5.1 surround sound system which was standardised as an ITU recommendation in 1992 (ITU, 1992).

The 5.1 loudspeaker system is specifically a 3-2-1 system as it consists of three channels across the front (a standard two-channel left/right pair arranged ±30° plus a centre loudspeaker) two surround channels arranged ±100°–120° plus a dedicated low frequency effect (LFE) channel (figure 2.25). The dot in 5.1 indicates that the LFE channel is reduced bandwidth. The 5.1 system is limited in that it does not support true 360° spatial imaging and provides accurate spatial images only in the frontal region. To this end the surround channels are normally used for ambient sounds rather than creating specific virtual sources.

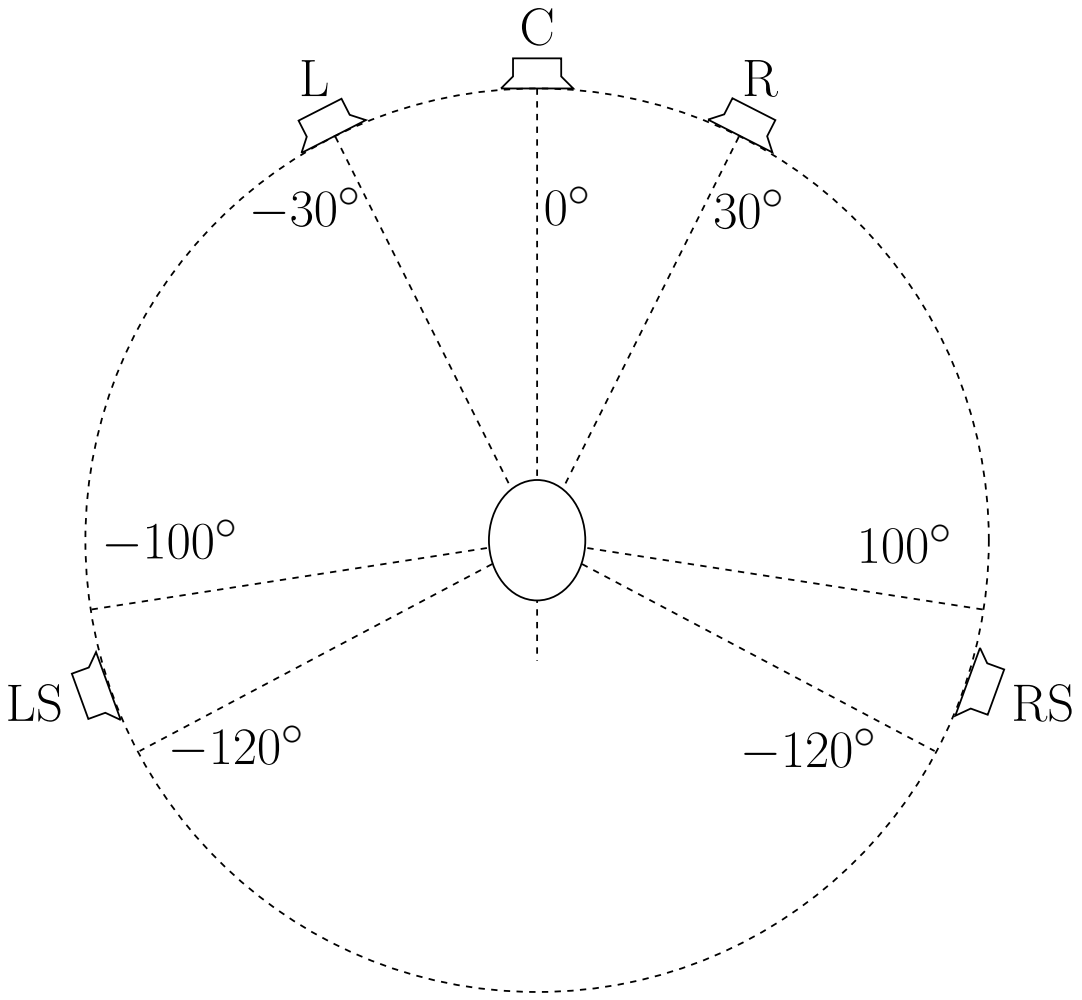


Figure 2.25: Surround loudspeaker listening configuration as specified in ITU-R BS.775 (ITU, 1992).

Various suggestions of additional locations for surround loudspeakers to improve spatial imaging have been made, e.g., 7.1 or 10.2 systems (Rumsey, 2001).

The panning in horizontal multi-channel loudspeaker set-ups is generally carried out by pair-wise amplitude panning. A common formulation for this is known as vector base amplitude panning (VBAP) (Pulkki, 1997). The distance from the loudspeakers to the listening position are represented as unit-length vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ and the direction of the virtual source is represented by unit vector $\boldsymbol{p}$, a weighted sum of the loudspeaker vectors:

$$\boldsymbol{p} = g_1\boldsymbol{l}_1 + g_2\boldsymbol{l}_2 \tag{2.5}$$

The gains g_1 and g_2 are calculated as follows:

$$\boldsymbol{g} = \boldsymbol{p}^T\boldsymbol{L}_{12}^{-1} \tag{2.6}$$

Where $\boldsymbol{g} = [g_1 \ g_2]^T$ and $\boldsymbol{L}_{12} = [\boldsymbol{l}_1 \ \boldsymbol{l}_2]$ and some sort of normalisation is applied to the gains, such as $||\boldsymbol{g}|| = 1$.

VBAP can also be extended from pair-wise horizontal panning to panning between triplets of loudspeakers for three dimensional loudspeaker set-ups (Pulkki, 1997). Thus each virtual source signal is applied to, at most, three loudspeakers. Appropriate gains are applied to each channel to allow the image to be panned anywhere within the triangle formed by the loudspeakers. Its formulation is similar to two-dimensional (2D) VBAP but the virtual source is represented by a 3D Cartesian unit vector $\boldsymbol{p} = [p_1 \ p_2 \ p_3]^T$ which is a combination of the loudspeaker location vectors $\boldsymbol{l}_1,\ \boldsymbol{l}_2,$ and $\boldsymbol{l}_3$:

$$\boldsymbol{p} = g_1\boldsymbol{l}_1 + g_2\boldsymbol{l}_2 + g_3\boldsymbol{l}_3 \tag{2.7}$$

The gain vector $\boldsymbol{g} = [g_1\ g_2\ g_3]$ can then be found:

$$\boldsymbol{g} = \boldsymbol{p}^T \boldsymbol{L}_{123}^{-1} = [p_1\ p_2\ p_3] \begin{bmatrix} l_{11} & l_{12} & l_{13} \\ l_{21} & l_{22} & l_{23} \\ l_{31} & l_{32} & l_{33} \end{bmatrix}^{-1} \tag{2.8}$$

2.3.4 Binaural recordings and rendering

The prime aim of binaural audio is to create an accurate impression of spatial sounds. The simplest binaural recording systems involve a pair of spaced microphones placed at the location of the ears in a source environment. The recorded

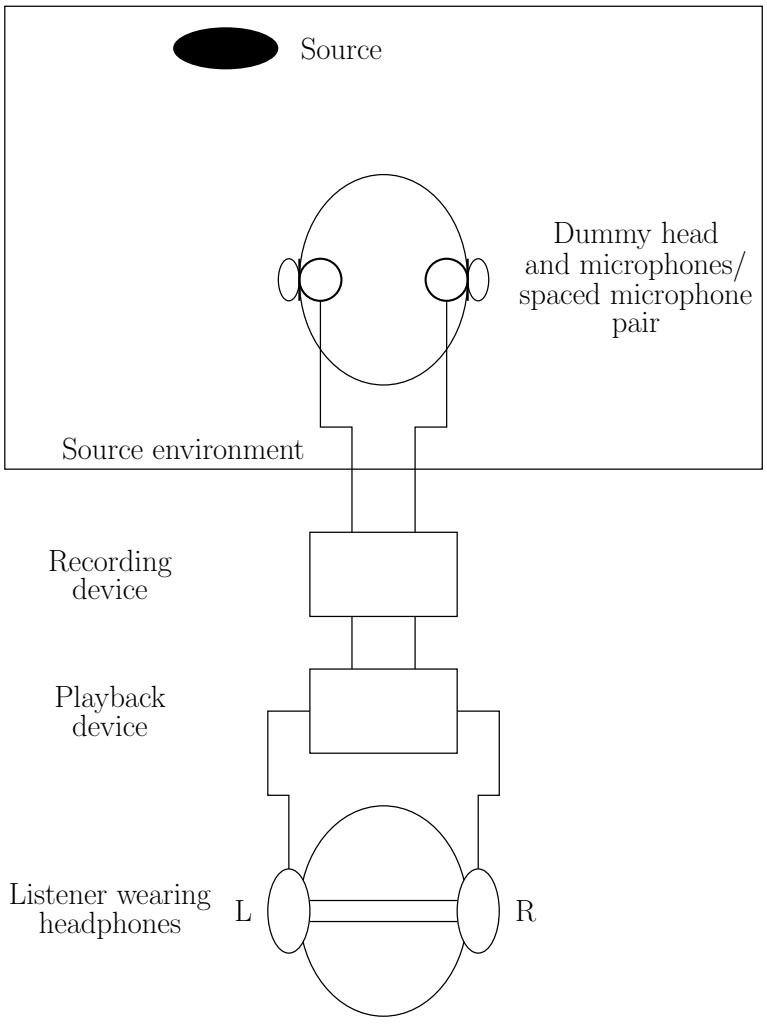


Figure 2.26: Set-up for binaural recording and playback. After Rumsey (2001).

Figure 2.27: The Knowles Electronics manikin for acoustic research (KEMAR) is an example of a head-and-torso simulator (HATS) for binaural recording and analysis.

signals are then played back to a subject over headphones (or loudspeakers — section 2.3.6) with each microphone signal only going to the corresponding headphone (figure 2.26). More effective binaural recording systems use a HATS (figure 2.27) or place the microphones in the ears of an actual listener.

A common alternative to binaural recording is binaural rendering. In binaural rendering a monophonic source signal x is convolved with the HRIR pair $[HRIR_L(\theta, \varphi), HRIR_R(\theta, \varphi)]$ for the desired direction:

$$\begin{aligned} y_L &= x * HRIR_L(\theta, \varphi) \\ y_R &= x * HRIR_R(\theta, \varphi) \end{aligned} \tag{2.9}$$

where θ and φ represent the desired direction and $*$ is the convolution operator.

There are a number of ways to acquire the HRIRs to be used and these are outlined in section 2.4.

Binaural recording and rendering works best when the HRTFs or ears of each individual listener are used. Pinnae vary in shape and hence individual HRTFs also exhibit variation. The use of non-individualised HRTFs results in erroneous spectral cues and has been shown to degrade localisation performance significantly, leading to poor elevation accuracy and increases in front-back reversals, as well as timbral colouration, image instability and reduced externalisation (Wenzel *et al.*, 1993; Møller *et al.*, 1996). Some work has suggested that listeners can, in some senses, learn to use HRTFs that are not their own (Hofman *et al.*, 1998). Alternatively, pattern matching can be used to find the best match for listeners' ears to a database of measured HRTFs (Zotkin *et al.*, 2002). Despite some promising results using the database-matching approach, individualised HRTFs produce the most reliable performance.

An important consideration when making binaural recordings or measuring HRTFs are the microphone positions. When using a HATS it is common practice for the microphones to be placed at the position of the ear drum (Gardner and Martin, 1995). This is of course impossible when making recordings using real subjects and an alternative is to use probe microphones (microphones with attached probe tubes), which are placed as close as possible to the ear drum (Wightman and Kistler, 1989a). Disadvantages include the risk of damage to the ear with incorrect positioning of the microphones and probe microphones can suffer from poor frequency responses and low sensitivity. Furthermore ear canal recordings include not only the directional response of the outer ear but also the largely directionally independent resonance of the ear canal (Algazi *et al.*, 1999), which may or may not be of importance. An alternative is the blocked-meatus approach (Møller, 1992; Møller *et al.*, 1995b), in which diaphragms of miniature microphones are placed at the ear canal entrance and the canal is sealed with either foam or putty. This approach has been validated (Algazi *et al.*, 1999) and has become the favoured method of binaural recordings due to its convenience.

Another consideration when playing back binaural recordings or renderings over headphones is how to remove the response of the headphones. Møller has sated that the aim of headphone equalisation for playback of blocked-meatus recordings and renderings is to produce a flat frequency response at the entrance to the ear canal (Møller, 1992). Møller *et al.* (1995a) suggest the use of free air equivalent coupling (FEC) headphones for playback of blocked-meatus recordings. The recording of the headphone responses is normally carried out at the time of recording the HRTFs, using the same microphones. The inverse of the measured response can then be used as a filter to remove it from the playback system. However, the changes in position which typically occur during replacement of the headphones can have measurable, perceptually significant, effects on the measured responses (Toole, 1984; Paquier and Koehl, 2010). Therefore headphone equalisation methods that use multiple measurements to provide a "an upper variance limit" of multiple headphone transfer function measurements have been proposed (Masiero and Fels, 2011).

Under normal listening conditions listeners generally do not keep their head still, instead using head movements to improve localisation and resolve ambiguities. In standard binaural listening systems there is no compensation for this as the HRTFs used are static. However, some systems (both headphone and crosstalk cancellation — section 2.3.6) use head tracking to update the HRTFs in real time and this can reduce front-back confusions (Begault *et al.*, 2001) as well as improve elevation perception (Rao and Xie, 2005; Zhang and Xie, 2012).

2.3.5 Virtual loudspeaker systems

Virtual loudspeaker systems theoretically allow signals meant for playback over a physical loudspeaker system to be played back over stereo headphones using HRTFs (Laitinen and Pulkki, 2009). Each of the loudspeaker signals is convolved with the HRIR pair for its given direction. Figure 2.28 shows this principle for the standard $\pm 30°$ stereo loudspeaker configuration. The signals for the left (y_L)

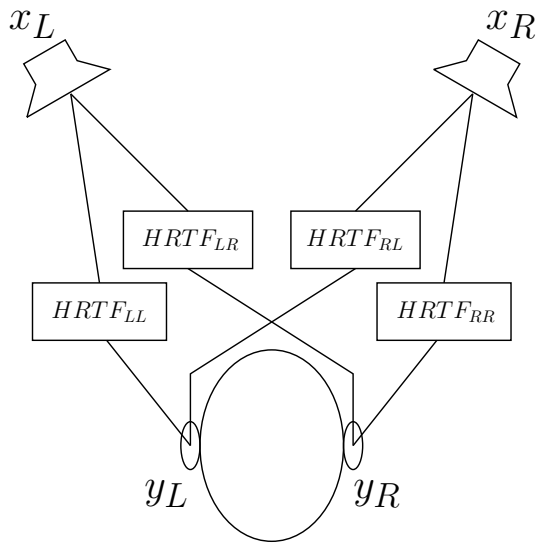


Figure 2.28: Virtual loudspeaker reproduction over headphones.

and right (y_R) headphones are calculated as follows:

$$y_L = x_L * HRIR_{LL} + x_R * HRIR_{RL}$$
$$y_R = x_R * HRIR_{RR} + x_L * HRIR_{LR} \tag{2.10}$$

where $HRIR_{LL}$ and $HRIR_{RR}$ are the ipsilateral HRIRs and $HRIR_{LR}$ and $HRIR_{RL}$ are the contralateral HRIRs and $*$ is the convolution operator. As well as the limitations identified in the previous section the additional limitation of virtual loudspeaker systems is that the HRTFs are normally measured under anechoic conditions and therefore the resulting signals lack the reflections present under normal listening conditions. This can result in a reduction of externalisation with sound images often perceived as coming from within the listener's head. However, this can be corrected using binaural room impulse responses (Møller, 1992) or measured room impulse responses.

2.3.6 Crosstalk cancellation

Crosstalk cancellation (Kirkeby *et al.*, 1998b), also known as *transaural* processing, can be considered the inverse problem of stereo headphone reproduction (section 2.3.2). When playing binaural signals meant for headphone playback

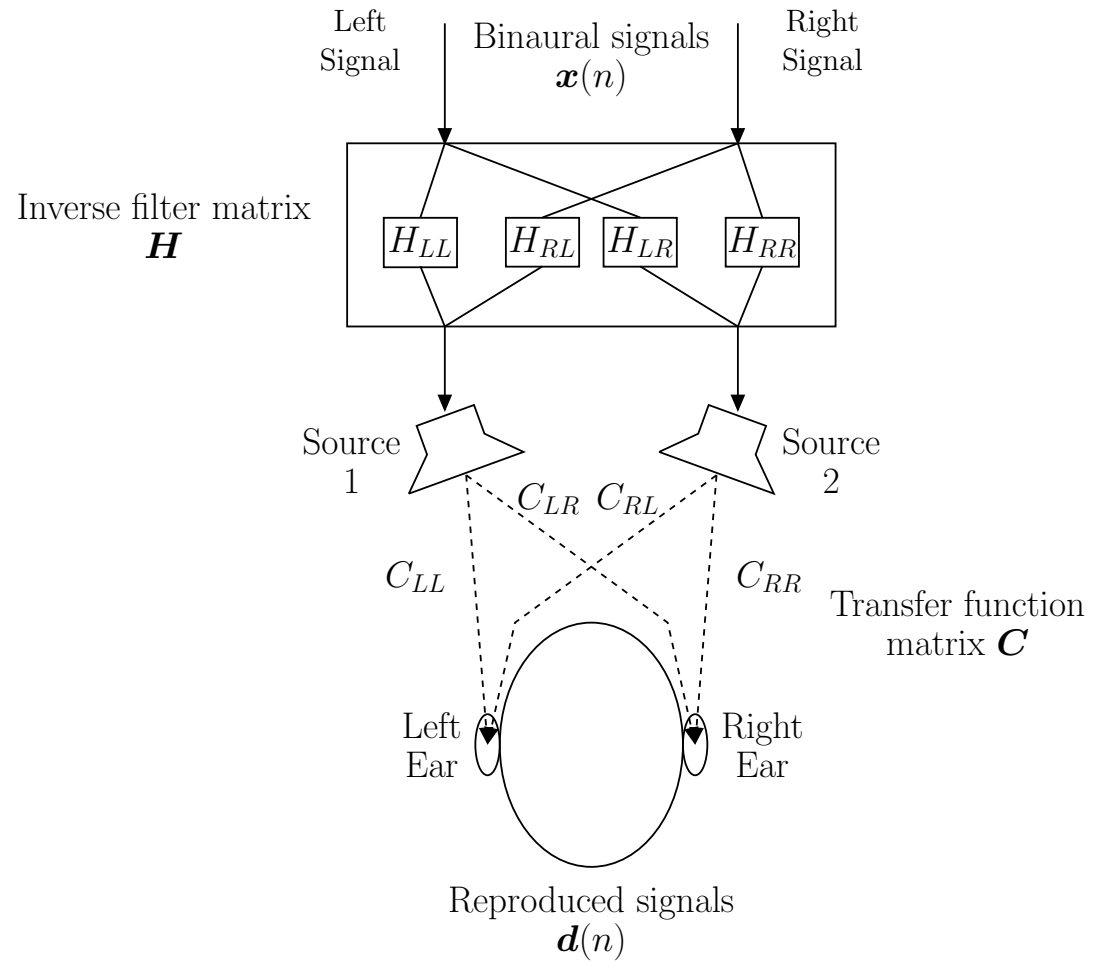


Figure 2.29: Crosstalk cancellation system for playback of binaural signals over loudspeakers. After Kirkeby *et al.* (1999).

over a pair of loudspeakers there needs to be compensation for the "crosstalk" that occurs, i.e., the fact that each channel's signal reaches both ears. A crosstalk cancellation scheme is shown in figure 2.29.

If the binaural signals are represented as a 2 by 1 vector $\boldsymbol{x}(n)$ and the signals present in the ear canals of the listener are $\boldsymbol{d}(n)$ then in the z domain (Pulkki *et al.*, 2011):

$$\boldsymbol{d}(z) = \boldsymbol{C}(z)\boldsymbol{H}(z)\boldsymbol{x}(z) \tag{2.11}$$

where:

$$\boldsymbol{C}(z) = \begin{bmatrix} C_{LL}(z) & C_{LR2}(z) \\ C_{RL}(z) & C_{RR}(z) \end{bmatrix} \tag{2.12}$$

contains the response of the loudspeakers measured in the listener's ear canals

and:

$$\boldsymbol{H}(z) = \begin{bmatrix} H_{LL}(z) & H_{LR}(z) \\ H_{RL}(z) & H_{RR}(z) \end{bmatrix} \tag{2.13}$$

contains responses for inverse filtering to minimise the crosstalk.

Under ideal listening conditions $\boldsymbol{x}(n) = \boldsymbol{d}(n)$, which can only occur if $\boldsymbol{H}(z) = \boldsymbol{C}(z)^{-1}$. The problem with simply inverting the measured response of the loudspeakers is that this can often produce filters with high gain, especially outside the working frequency range of the measurement/playback system. Kirkeby *et al.* (1998a, 1999) suggested applying frequency-dependent regularisation to the inversion problem to generate an optimal inverse filter matrix $\boldsymbol{H}_{opt}$:

$$\boldsymbol{H}_{opt} = \left[\boldsymbol{H}^T(z^{-1})\boldsymbol{H}(z) + \beta B(z^{-1})B(z)\boldsymbol{I}\right]^{-1} \boldsymbol{H}^T(z^{-1})z^{-1} \tag{2.14}$$

where β is a small positive gain factor and $B(z)$ is a frequency-dependent shape factor. $B(z)$ should be large for frequencies that should not be boosted and small for frequencies that require no regularisation.

Crosstalk cancellation systems can provide very good three-dimensional virtual sources, even behind the listener (Schroeder, 1975; Gardner, 1997). However, the listening area is impractically restrictive and even a movement of 1-2 cm can destroy the binaural effect (Pulkki and Karjalainen, 2015). Crosstalk systems ideally require non-reverberant surroundings and work best when individual HRTFs are used for the loudspeaker-to-ear canal responses. They therefore suffer from the same restrictions as other HRTF-dependent systems mentioned in previous sections.

2.3.7 3D loudspeaker systems

An alternative approach, rather than 3D VBAP, to 3D loudspeaker reproduction is Ambisonics (Gerzon, 1973; Gerzon and Barton, 1992). First order Ambisonics is based on four signals: three orthogonal figure-of-eight components X (front-back),

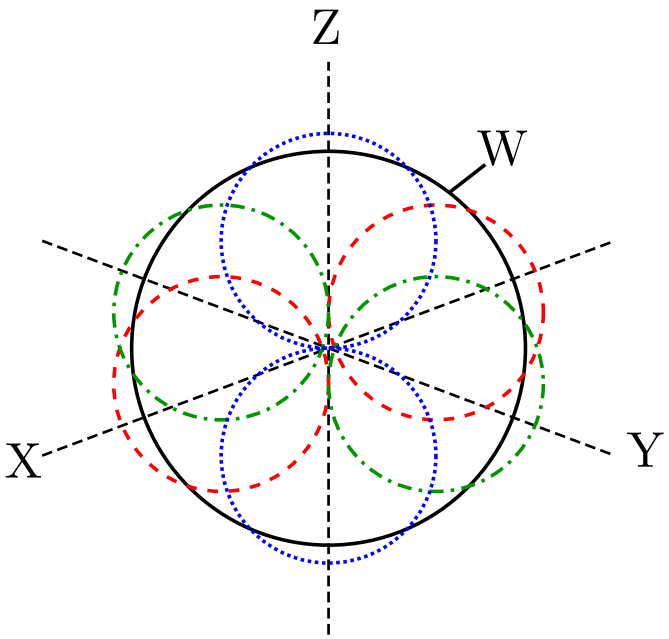


Figure 2.30: Ambisonics B-format components: X (red dashed), Y (green dash-dot), Z (blue dot) and W (solid black).

Y (left-right), Z (up-down) and an omni-directional component W. These are shown in figure 2.30 and are based on zeroth and first degree spherical harmonics (section 5.1.2, particularly figure 5.2).

A virtual source signal S at azimuth θ and elevation φ can then be represented by summing the following signal components (Rumsey, 2001):

$$\begin{aligned} X &= S\cos\theta\cos\varphi \\ Y &= S\sin\theta\cos\varphi \\ Z &= S\sin\varphi \\ W &= \frac{S}{\sqrt{2}} \end{aligned} \tag{2.15}$$

The spatial accuracy of virtual sources created using first order Ambisonics is fairly limited (up to 45° blur (Bertet *et al.*, 2007)), as is the listening volume for which the sound stage is stable. These limitations can be improved by increasing the degree of the spherical harmonics used as components: this is known as *higher order Ambisonics* (Moreau *et al.*, 2006; Daniel *et al.*, 2003; Daniel and Moreau, 2004). Bertet *et al.* (2007) found that on average fourth order Ambisonics

halved the absolute localisation error of virtual sources over first order systems. A limitation of higher order Ambisonics is that it is increasingly difficult to construct microphones capable of recording the required signals. However, the signals can be synthesised and microphones that derive the required signals from a series of omni-directional capsules mounted on a sphere have been developed (Bertet *et al.*, 2007; Jin *et al.*, 2014).

Recently there have been a number of developments in three dimensional loudspeaker systems. NHK Science and Technical Research Laboratories[3] have developed a 22.2 system which is the ITU standard audio format for ultra-high definition television and consists of three layers of loudspeakers as well as two subwoofers (ITU, 2009).

Dolby Laboratories[4] developed a hybrid channel/object-based three dimensional loudspeaker system called Dolby Atmos (Dolby Laboratories, 2014). In object-based audio a channel can carry the audio for an "object" along with metadata about that object such as the desired location of the object in 3D space, or the audio content of the object, e.g. speech, ambience, music etc. These objects are then rendered at playback time using their associated metadata and knowledge of the target speaker set-up. The default configuration for the 128 channels used in Atmos is 10 channels carrying a 9.1 "bed" and a further 118 channels for audio objects. Dolby Atmos systems have been installed in a number of cinemas and the first film released in the Atmos format was Pixar's Brave in 2012. Since then it has been developed for use in the home with traditional 5.1 and 7.1 loudspeaker systems plus additional height loudspeakers and video games with Atmos audio have also been released[5]. The additional height loudspeakers can either be dedicated, ceiling mounted, loudspeakers or upward-firing loudspeakers integrated with the existing loudspeakers (Dolby Laboratories, 2016).

Auro-3D (Auro Technologies, 2015) is a loudspeaker system developed by Auro

[3] `https://www.nhk.or.jp/strl/index-e.html`
[4] `https://www.dolby.com/us/en/index.html`
[5] `http://blog.dolby.com/2015/05/dolby-atmos-coming-to-star-wars-battlefront/`

Technologies[6]. Initially a channel based 9.1 system (a 5.1 system with four added height loudspeakers (Theile and Wittek, 2011)), it evolved into an 11.1 system consisting of a set of five head level loudspeakers along with an additional five height level loudspeakers and a loudspeaker positioned centrally in the ceiling (the so called "voice-of-god" loudspeaker). The first film to be released in Auro-3D format was Lucasfilm's Red Tails in 2012. Since then Auro-3D has been developed into AuroMax, an object-based three-dimensional loudspeaker system with a minimum 20.1 loudspeaker set-up.

2.3.8 Wave field synthesis (WFS)

Wave field synthesis (WFS) (Berkhout *et al.*, 1993) uses a large number of loudspeakers, usually arranged in a horizontal array, to produce artificial wavefronts. It is based on Huygens' principal that any point on a wavefront can be considered a secondary source and therefore any wavefront can be represented as a superposition of elementary wave sources. Each loudspeaker in the array is treated as one of these wave sources and a virtual source is placed behind the loudspeaker, by synthesising convex or planar wavefronts or in front of the loudspeaker array by producing concave wavefronts.

The strength of WFS is that, because the actual wavefront of the virtual source is being generated, the source's location remains largely independent of listener position within the soundfield generated by the loudspeaker array. Its downsides are that it is sensitive to the acoustics of the room in which it is installed in and it is relatively expensive due to the high number of loudspeakers required. Furthermore the systems usually suffer from spatial aliasing due to the fact that to reproduce the wave field across the whole area within the array the loudspeakers need to be at most half a wavelength from each other (Pulkki and Karjalainen, 2015).

[6] `http://www.auro-3d.com`

2.3.9 Measures of spatial audio quality

There are a number of attributes to be considered when assessing spatial audio systems and recently there has been work to standardise this. Lindau *et al.* (2014) developed the spatial audio quality inventory (SAQI). Their aim was to generate a "complete and relevant" lexicon for spatial audio assessment. The SAQI includes 48 verbal descriptors along with corresponding scales divided into nine groups:

- Difference — measure of similarity between sources
- Timbre — e.g. high/low/mid frequency emphasis/attenuation, comb filter colouration
- Tonalness — e.g. perception of pitch within sound, Doppler effect
- Geometry — e.g. perceived horizontal/vertical location, front-back confusions
- Room — e.g. level of reverberation, sense of envelopment
- Time behaviour — e.g. pre/post-echoes, change in sequence of auditory events
- Dynamics — e.g. perceived loudness, dynamic compression
- Artefacts — e.g. pitched/impulsive/noise artefacts, distortion
- General — e.g. loss of clarity, naturalness

Zacharov *et al.* (2016a,b) have developed another, slightly more concise collection of spatial audio quality attributes. They were generated through a mixture of data mining, data analysis and discussion with an expert panel. Some of the definitions are based on the SAQI but with definitions and scales modified to "improve clarity and create an absolute scale". Furthermore some definitions were changed, particularly envelopment which was altered to include the effect of direct sound sources as well as reverberation. In addition *internality* is used instead of externalisation since internal localisation is less natural. Their lexicon

consists of twelve attributes, including four from subdivision of other attributes:

Group	Quality		Scale
Spatial extent	Depth		shallow–deep
	Width		narrow–wide
	Envelopment	Horizontal	not enveloping–completely enveloping
		Vertical	not enveloping–completely enveloping
Localisation	Distance		near–far
	Internality		distinctly internal–indistinct–distinctly external
	Localisability		imprecise–precise
Spatial/ Timbral	Clarity		unclear–clear
Environment	Reverberance		dry–highly reverberant
		Level of reverberance	low–high
		Duration of reverberance	short–long

Table 2.4: Definition of spatial attributes in the lexicon of Zacharov *et al.* (2016a)

In addition to the list of attributes they also developed an updated sound wheel based on earlier work (Pedersen and Zacharov, 2015). The sound wheel consists of a large number of spatial and timbral sound descriptors arranged into categories and groups. For example *depth*, *width* and *envelopment* are grouped under *spatial*

extent which in turn is part of the *spatial* category along with *environment* and *spatial localisation*.

2.4 HRTF acquisition and estimation

As mentioned in the previous section the use of non-individualised HRTFs can produce undesirable effects, yet the acoustic measurement of individual HRTFs is a lengthy process that requires specialist equipment. Therefore the possibility of estimating or simulating individualised HRTFs more efficiently through indirect means is highly attractive. This section firstly reviews the most common approaches to HRTF measurement, including the difficulties associated with HRTF measurement. Then past work on estimating and simulating HRTFs is presented as an attractive alternative to HRTF measurement.

2.4.1 HRTF measurement

The acoustic measurement of HRTFs requires specialist equipment and can be a time-consuming process when measuring a large number of directions. The impulse responses from a sound source to an individual's ears are measured under anechoic conditions for a number of directions. The subject is either kept static (KAIST, 2016; Tohoku University, 2016) or is seated on a rotatable chair (IRCAM, 2003; HFC, 2016; Riederer, 2012). As discussed in section 2.3.4 the microphone is generally placed at the entrance to the ear canal and either a single movable speaker (IRCAM, 2003; SADIE, 2016) or a static (HFC, 2016; Riederer, 2012) or movable (KAIST, 2016; Tohoku University, 2016; Algazi *et al.*, 2001d) hoop of speakers is used to playback some sort of measurement signal.

There are a number of choices for the stimulus signal when measuring HRTFs. Maximum length sequences (MLS) are pseudo-random signals that provide much better signal-to-noise ratio (SNR) than using pure impulses (Rife and Vanderkooy, 1989). Golay codes (Zhou *et al.*, 1992) produce the same SNR as MLS but the

length of the code is a power of two which makes fast Fourier transform (FFT) analysis quicker. More than one measurement can also be used to further improve SNR. However, problems can arise if the measurement system is time variant, e.g. if the subject moves slightly or convection currents exist, because then multiple systems are effectively being measured simultaneously. Another disadvantage of MLS measurements is that they suffer distortion if there are minor nonlinearities in the measured system, e.g. harmonic distortion in the loudspeakers. On the other hand, logarithmic swept sine measurements (Farina, 2000), are resilient to nonlinearities. The results of harmonic distortion are separated in time and can therefore be gated out. Logarithmic swept sine signals have the further advantage that multiple measurements can be carried out using slightly offset sweeps (Majdak *et al.*, 2007). Use of the multiple exponential sweep method has resulted in reported measurement times of as little as five minutes (HFC, 2016).

2.4.2 HRTF databases

There a number of HRIR/HRTF databases available when individualised HRTFs are either not needed or their measurement is not viable within time and financial constraints.

The CIPIC database (Algazi *et al.*, 2001d) is probably the most widely known. HRIRs of 45 subjects were measured for 1250 directions using Golay codes and a number of anthropometric measurements were also made for each subject. The inclusion of anthropometric measurements has led to its use in a number of studies correlating HRTF features and the physical dimensions of the pinna and other morphological measurements (section 2.5).

The Institut de Recherche et Coordination Acoustique/Musique (IRCAM) Listen database (IRCAM, 2003) includes HRIR measurements for 51 subjects for 187 different directions using swept sine technique.

The Austrian Academy of Sciences Acoustic Research Institute (ARI) (Acoustics

Research Institute, 2011) HRTF database is the largest database, consisting of 1550 directions for more than 110 subjects. It differs from the other databases in that all the HRTFs have had the CTF removed, yielding DTFs. The CTF is the average magnitude spectrum of the HRTFs across all directions for a particular subject, combined with its minimum phase spectrum; hence, the DTFs contain only the directionally dependent information. DTFs from the ARI database are used in the Baumgartner *et al.* (2013) sagittal plane localisation model — see section 3.4.

The universities of Sydney and York worked together on the creation of the Sydney-York morphological and recording of ears database (SYMARE database) database (Jin *et al.*, 2013). The database consists of acoustic HRIR measurements for 61 subjects, high resolution meshes of the head and torso obtained from magnetic resonance imaging (MRI) scans and HRIRs calculated from the meshes using the fast fast multipole boundary element method (FM-BEM).

2.4.3 Structural models

Structural models treat each anatomical part of the auditory system, and its corresponding auditory cues, as independent sections that can be combined into an overall model. Brown and Duda (1998) proposed a comprehensive structural model for estimating HRIRs combining head shadow, ITD, shoulder reflections and pinnae resonances. They found it performed reasonably well against individual HRTFs, although testing was limited. Algazi *et al.* (2001c) investigated structural decomposition of HRTFs by taking measurements of isolated pinnae and a pinna-less head and torso model and comparing recombined HRTFs to HRTF measurements of a torso and head with pinnae. Their results indicate that the effects of the head, torso and pinnae could be successfully decomposed and recombined. Despite these promising results, structural models fail to account for acoustic interactions between parts of the model and this is likely to be extremely important when modelling the complex structures of the pinnae.

2.4.4 Mathematical models

There are several mathematical models that represent parts of the auditory periphery as simpler geometric shapes. The most common mathematical model in HRTF simulation is the spherical-head model of ITD (Woodworth and Schlosberg, 1962):

$$ITD = \frac{a}{c}(\theta + \sin\theta) \tag{2.16}$$

where a is the radius of the spherical-head, c is the speed of sound and θ is the angle of incidence of the sound (figure 2.31). Algazi *et al.* (2001b) used the anthropometric measurements of the CIPIC database to propose an equation for calculating the optimum radius, a, to use in equation 2.16 in terms of weighted head height, depth and width measurements:

$$a = w_1 X_1 + w_2 X_2 + w_3 X_3 + b \tag{2.17}$$

where $w_1 = 0.51$, $w_2 = 0.019$, $w_3 = 0.18$, $b = 3.2cm$ and X_1, X_2 and X_3 are the head half width, head half height and head half length in cm, respectively.

A number of studies have explored or built on the spherical-head model. Avendano *et al.* (1999b) used the model to create a transformation for calculating

[7] `http://interface.cipic.ucdavis.edu/images/research/pathgeo.gif`

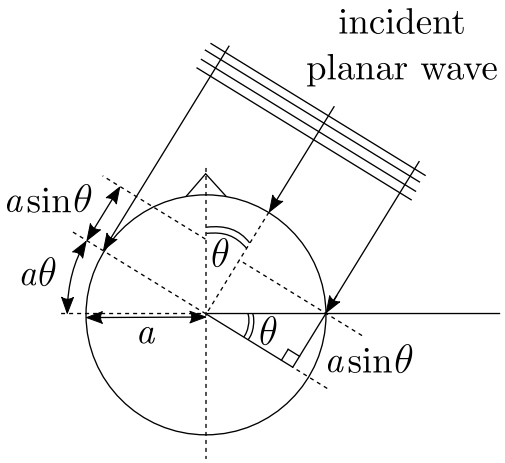


Figure 2.31: Spherical head model for interaural time difference (ITD) developed by Woodworth and Schlosberg (1962). After[7].

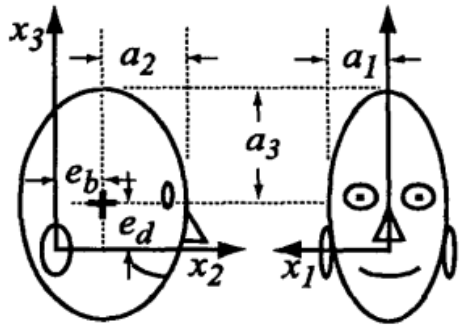


Figure 2.32: Ellipsoidal head model investigated by Duda *et al.* (1999). Image from Duda *et al.* (1999).

the contralateral HRTF from the ipsilateral HRTF. Duda *et al.* (1999) improved the spherical-head model by representing the head as an ellipsoid, rather than a sphere, with the ears positioned slightly down and towards the rear of the ellipsoid (figure 2.32). This allowed the elevation dependence of the ITD to be modelled, which is not possible with the spherical head model and resulted in a small improvement in ITD error — 15 μs vs. 22 μs. However, there is no direct analytical solution to calculating the path difference and so Duda *et al.* (1999) had to calculate the ITD via constrained minimisation of the problem which represents a significant increase in complexity and computation compared to the spherical model. Interestingly, by the time Algazi *et al.* (2002) (which included the authors of the previous study) combined simple models of the head and torso they had returned to the spherical-head model. This suggests that the increased accuracy of the modelled was outweighed by the increase in complexity. They combined the spherical-head model with both a spherical and ellipsoidal torso to attempt to model both interaural and torso cues — e.g. shoulder reflections. They found that the inclusion of either torso model significantly improved the accuracy of the modelled HRTFs and that the ellipsoidal torso showed the closest correspondence with the measured HRTFs of a pinna-less Knowles Electronics manikin for acoustic research (KEMAR).

Teranishi and Shaw (1968) created simple geometric models of the outer ear (figure 2.33). They initially used a shallow cylinder to model the concha and

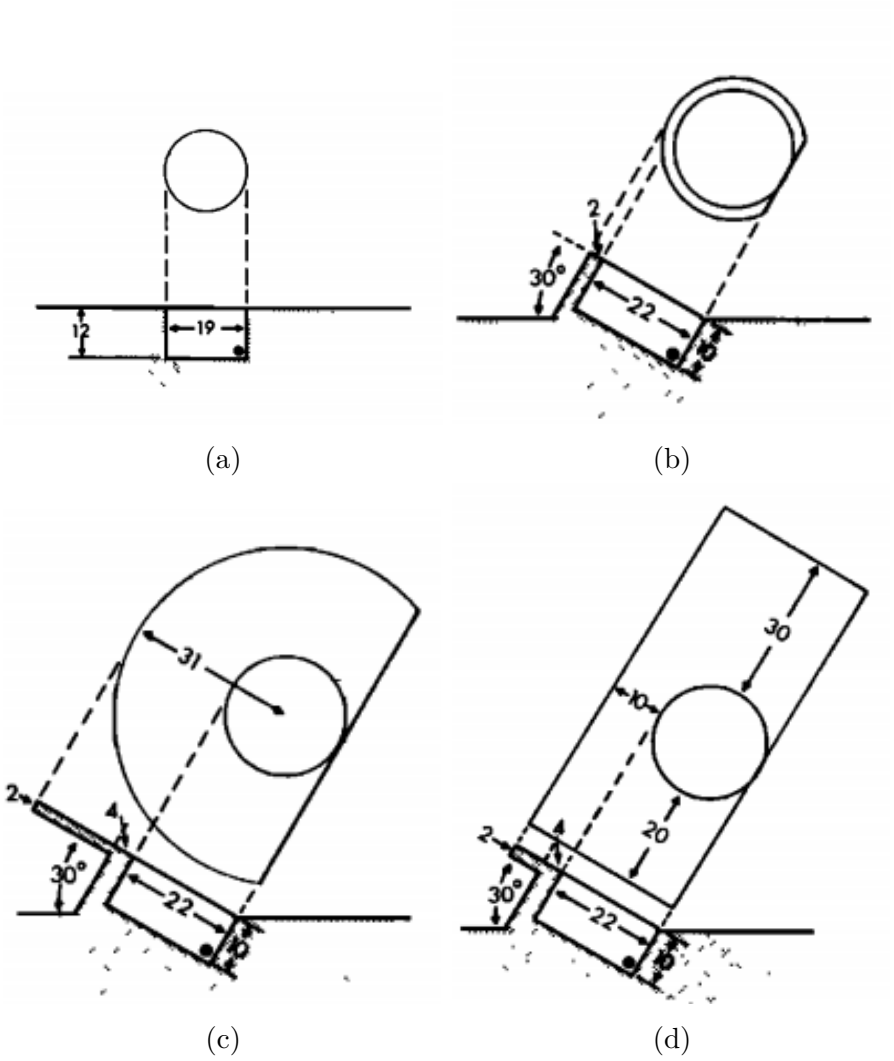


Figure 2.33: Geometric models investigated by Teranishi and Shaw (1968). a): simple cylindrical concha. b): tilted cylindrical concha. c) and d): cylindrical concha with segmented circular pinna and rectangular pinna respectively. Dimensions shown are in millimetres. The lower half of each diagram shows a cross section of the model, whilst the upper half shows the model face on. The dotted lines connect the same point in each view. Images from Teranishi and Shaw (1968).

found that its first resonant mode behaved in a similar manner to measured concha responses but lacked significant variation with direction. They added directionality and extended the model's agreement with acoustic measurements up to approximately 7 kHz by tilting the cylinder and adding simple models of the pinna: one rectangular and one segmented circular. Interestingly the rectangular pinna model performed better than the segmented circular model, despite the segmented circular model sharing more in appearance with a real pinna. The authors suggested that the reason for this may have been that the segmented

circular pinna was only asymmetric in one dimension whilst the rectangular model exhibited asymmetry in both dimensions.

Lopez-Poveda and Meddis (1996) created a physical model of the concha using a spiral-shaped diffraction and reflection model which was compared to the "single-delay-and-add" approximation by Hebrank and Wright (1974). They found that their model predicted the centre frequencies of the pinnae notches more accurately than the single-delay-and-add approximation. However, there was no perceptual testing and it is unlikely that these simplified empirical approaches can accurately describe the complex acoustic function of the whole pinna.

2.4.5 Statistical models

A number of statistical analyses have been applied to HRTFs. Raykar *et al.* (2003) used linear prediction (LP) analysis to decompose the CIPIC database HRTFs into specific features which they then mapped to specific anthropometric measurements (Raykar *et al.*, 2005). They suggested that the peaks and notches of the HRTFs could be reliably predicted from measurements of the pinnae.

A number of studies have used principal components analysis (PCA) to analyse HRTFs, e.g. (Martens, 1987; Kistler and Wightman, 1992; Middlebrooks and Green, 1992; Hwang and Park, 2007; Breebaart, 2013). PCA applies statistical analysis to a weighted sum of HRTFs or HRIRs so that they can be represented by a set of frequency or time domain basis functions.

Given an $N \times M$ matrix, $\mathbf{X}$, consisting of M observations (directions and/or subjects) of N variables (frequency or time samples), an $N \times 1$ mean vector, $\mathbf{u}$, is first calculated:

$$u[n] = \frac{1}{M} \sum_{m=1}^{M} \mathbf{X}[n, m] \tag{2.18}$$

This is then subtracted from $\mathbf{X}$:

$$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{u}.\mathbf{h} \tag{2.19}$$

where $\mathbf{h}$ is an $M \times 1$ vector of ones. The covariance matrix, $\mathbf{C}$, of $\hat{\mathbf{X}}$ is then calculated:

$$\mathbf{C} = \frac{1}{M-1}\hat{\mathbf{X}}.\hat{\mathbf{X}}^* \tag{2.20}$$

where * represents the conjugate transpose of the matrix. The basis vectors, $\mathbf{p}_i$ $(i = 1, 2, \ldots n)$ are then the first n eigenvectors of the covariance matrix, $\mathbf{C}$, corresponding to the n largest eigenvalues. The weights of the basis functions are finally calculated as follows:

$$\mathbf{W} = \mathbf{P}^*.\hat{\mathbf{X}} \tag{2.21}$$

where $\mathbf{P} = [\mathbf{p}_1\ \mathbf{pc}_2\ \ldots\ \mathbf{pc}_n]$. If $n = N$ then full reconstruction of the original data is possible. However, in PCA investigations the aim is usually to evaluate the number of basis functions required to account for a prescribed proportion of the variation in the data. Kistler and Wightman (1992) found that the first five basis functions accounted for 90% of the variation in their HRTFs.

Middlebrooks and Green (1992) compared the basis functions for two sets of HRTFs measured under different experimental set-ups and found that the basis functions were very similar. However, they acknowledged that this is unsurprising as PCA removes the mean from the data and the variation in experimental set up is likely to be eliminated in this step.

Hwang and Park (2007) investigated HRIR customisation through subjective weighting of PCA basis functions. They found that twelve basis functions were required to account for 95% of the variation between individuals and elevations in the median plane HRIRs of the CIPIC database. Their subjective testing allowed subjects to tune the weights of three most dominant of the twelve basis functions, whilst the weights of the other nine were set to their average value. They found that this decreased front-back reversals and improved elevation perception compared to using HRIRs straight from the database, however their testing was limited to two subjects.

One of the weaknesses of traditional PCA when applied to HRTFs is that the

PCA representation is not continuous, i.e. HRTFs for directions not originally measured cannot be generated as no PCA weights exist. To combat this, Xie (2012) proposed spatial principal components analysis (SPCA) where PCA is carried out in the spatial domain rather than the frequency or time domain and a set of spatial basis functions are derived. SPCA allows HRTFs with a high directional resolution to be derived from a limited set of measured direction. They reported that HRTFs from 493 source directions can be reconstructed from 73 measured directions.

Evans *et al.* (1998) used surface spherical harmonicss (SSHs) as the basis functions for representing measured HRTFs, which also allowed for directional interpolation. Initially applied to time-domain HRIRs using SSHs up to degree 17, root mean square (RMS) error between measured and recreated HRIRs varied greatly with direction and at its maximum was 20%. The authors highlighted that even the 17^{th} degree SSHs had relatively high amplitude, that this might be inherent in the time-domain nature of the HRIRs and hence higher degree SSHs might be needed for representing HRIRs. Therefore, they applied the same approach to the magnitude and phase spectra of the HRTFs. This was found to be more efficient, with 90% of variance within the measured HRTFs contained, on average across all frequency bins, within the first seven SSH degrees, and within the first four degrees within the frequency range 5–7.5 kHz (Evans *et al.*, 1997). The worst case error with 17 degrees of SSHs was 1.2%, compared to 20% when applying the approach to HRIRs. Furthermore, considering the HRTFs rather than the HRIRs allowed the study of HRTF features based on frequency bin and direction. For instance, the authors found that the amplitude of the fundamental SSH, which is omnidirectional (section 5.1.2), exhibited a peak around 4 kHz, indicating that frequencies in this range enter the ear at the highest amplitude, regardless of direction. It is possible that this is a manifestation of the ear canal resonance, which typically appears around 4 kHz (Wiener and Ross, 1946), and is largely direction independent (Algazi *et al.*, 1999). An additional advantage of Evans *et al.*'s (1998) approach is that the SSHs are mathematically defined orthogonal

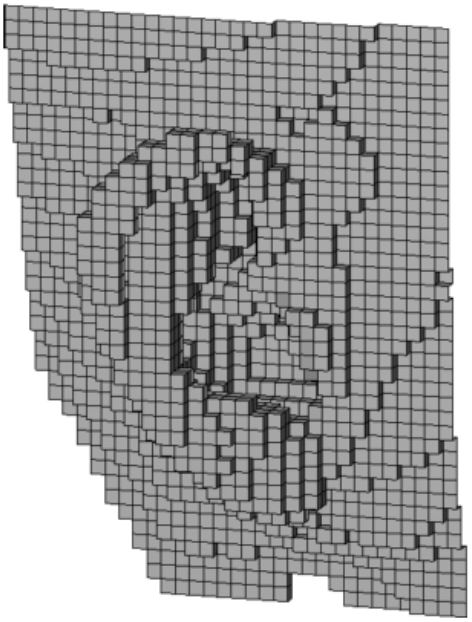
basis functions, not calculated directly from the measured HRTFs as with other approaches, which makes comparisons across different HRTF data sets valid.

2.4.6 Acoustic simulations

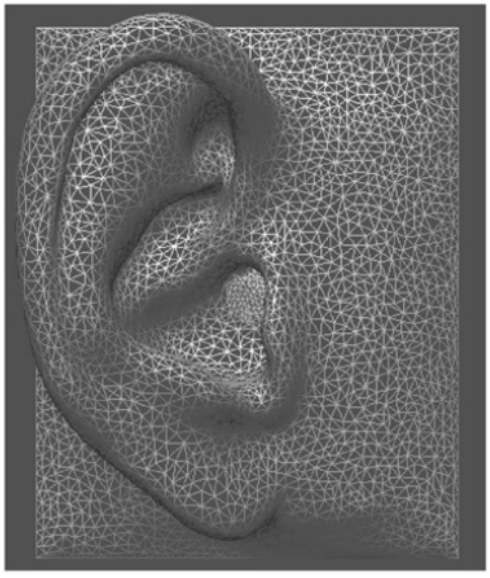
Computational acoustic simulations, e.g. using the boundary element method (BEM) (Katz, 2001a), the finite element method (FEM) (Kahana *et al.*, 1999) and the finite difference time domain (FDTD) method (Xiao and Liu, 2003) have been used to generate HRTFs. All these approaches require accurate 3D morphological models of the head and torso and lengthy simulations so currently do offer a huge advantage in terms of time and equipment over acoustically measuring HRTFs. However, they are well suited to investigating pinna morphology (Tew *et al.*, 2012; Mokhtari *et al.*, 2010) due to the ease with which the pinnae morphology can be altered and new HRTFs generated. Acoustic simulations have also been used to investigate near-field HRTFs (Rui *et al.*, 2013), which are difficult to measure acoustically.

The FDTD method consists of descretising the simulation space into a grid of voxels (figure 2.34a), each of which has an associated density, speed of sound and absorption (Mokhtari *et al.*, 2007). The partial differential equations (PDEs) that govern sound wave propagation are then solved, see Xiao and Liu (2003) for a deeper explanation. FDTD simulations have been shown to generate accurate HRTFs for both humans (Mokhtari *et al.*, 2007) and KEMAR (Mokhtari *et al.*, 2009).

The finite element method (FEM) requires the entire acoustic domain to be discretised into a polygonal mesh made up of smaller elements (Petyt *et al.*, 1976) and the PDEs are solved for each element; thus simplifying the problem. The solutions for each element are then combined to create an approximate solution for the whole system. The FEM has been used to generate HRTFs (Muraoka *et al.*, 2007; Ma *et al.*, 2015) and investigate torso cues (Kirkeby *et al.*, 2007) as well as the response of the ear canal and ear drum (Volandri *et al.*, 2014).



(a) Voxellated representation required for finite difference time domain (FDTD) simulation. Image from Mokhtari *et al.* (2010).



(b) Vertex/planar representation required for boundary element method (BEM) simulation. Image from Kahana and Nelson (2007).

Figure 2.34: Comparison of discretised representations required for acoustic simulations.

However, like the FDTD method, the disadvantage of the FEM is that the whole acoustic system requires modelling and discretising.

In contrast to the FDTD method and FEM, the BEM only requires the surface of the object under simulation to be discretised into a mesh consisting of vertices and the planes that connect them (figure 2.34b). The PDEs of the wave equation are reformulated as boundary integral equations that are then solved to generate the acoustic field around the object. Simulations are run to calculate the pressure on the mesh surface for individual frequencies and then the results are combined. The strength of BEM lies in the need to only solve the wave equation on the boundary of the acoustic domain: reducing a 3D problem to 2D. This makes BEM ideally suited to acoustic scattering problems such as generating HRTFs where it is assumed that the surface of the pinnae is important in determining HRTFs, not the internal structure.

Weinrich (1984) was the first to attempt BEM simulations of the human ear but the computational power restrictions of the time meant he could only simulate up to 1.7 kHz using a very simplified model of the concha. Continual improvements in computing have led to more and more detailed meshes being simulated across increasingly wide frequency ranges (Katz, 2001a,b; Kahana and Nelson, 2006, 2007).

More recently the fast multipole method (FMM) (Gumerov and Duraiswami, 2009; Kreuzer *et al.*, 2009; Gumerov *et al.*, 2010) has been paired with the BEM to create the fast multipole boundary element method (FM-BEM). The FM-BEM allows simulations to be carried out across the whole audible frequency range at a time cost of as little as 150-300 s per frequency point (Huttunen *et al.*, 2013). The FMM expands the Green's functions used in the BEM using multipole expansion to allow clustering of sources which can result in a reduction of complexity from $O(N^2)$ to $O(N)$ or $O(N \log N)$ yielding significant time reductions.

Further reductions in the time required for acoustic simulations have resulted from the use of "cloud-computing" (Huttunen *et al.*, 2013; Kärkkäinen *et al.*,

2013; Huttunen *et al.*, 2014). This relies on off-loading the heavy computational requirements of acoustic simulations to processing as a service (PaaS) and software as a service (SaaS) servers such as Amazon's EC2 cloud (Amazon Web Services, 2017). These servers offer high specifications (up to 244 GiB RAM, 36 cores and 48 TB storage[8]) and allow multiple simulations to be run simultaneously, significantly reducing computation time.

Ziegelwanger *et al.* (2014, 2016) have investigated the possible reductions in processing resulting from non-uniform discretisation of the head mesh. They compared the FM-BEM HRTFs generated by a reference high resolution mesh, lower resolution uniformly discretised meshes, meshes with elements concentrated around the source position using power functions and meshes with elements concentrated around the source position using raised-cosine functions (figure 2.35). They found that the raised-cosine graded mesh yielded the best numerical accuracy even compared to the high resolution reference mesh and that it yielded equal perceptual performance compared to the high resolution mesh with 13,000 elements compared to 100,000: a reduction of almost 90% computation.

One continuing limitation of all acoustic simulation techniques is the requirement for accurate head and torso models. The most reliable methods for generating the required models to date have been laser scanning (Kahana *et al.*, 1999; Katz, 2001a; Kahana and Nelson, 2006, 2007) or MRI scanning (Jin *et al.*, 2013). However, not only are they time consuming and require specialist equipment, MRI scanning especially, requires much post-processing to turn the scans into 3D models. Recently there has been research into more practical acquisition methods for acoustic models. Huttunen *et al.* (2014) compared three approaches for generating meshes:

(a) 52 simultaneous photographs taken from different directions

(b) a commercial 3D facial scanner

(c) video recorded with a mobile phone combined with structure-from-motion

[8]Correct as of 25/11/2016 according to: `https://aws.amazon.com/ec2/`

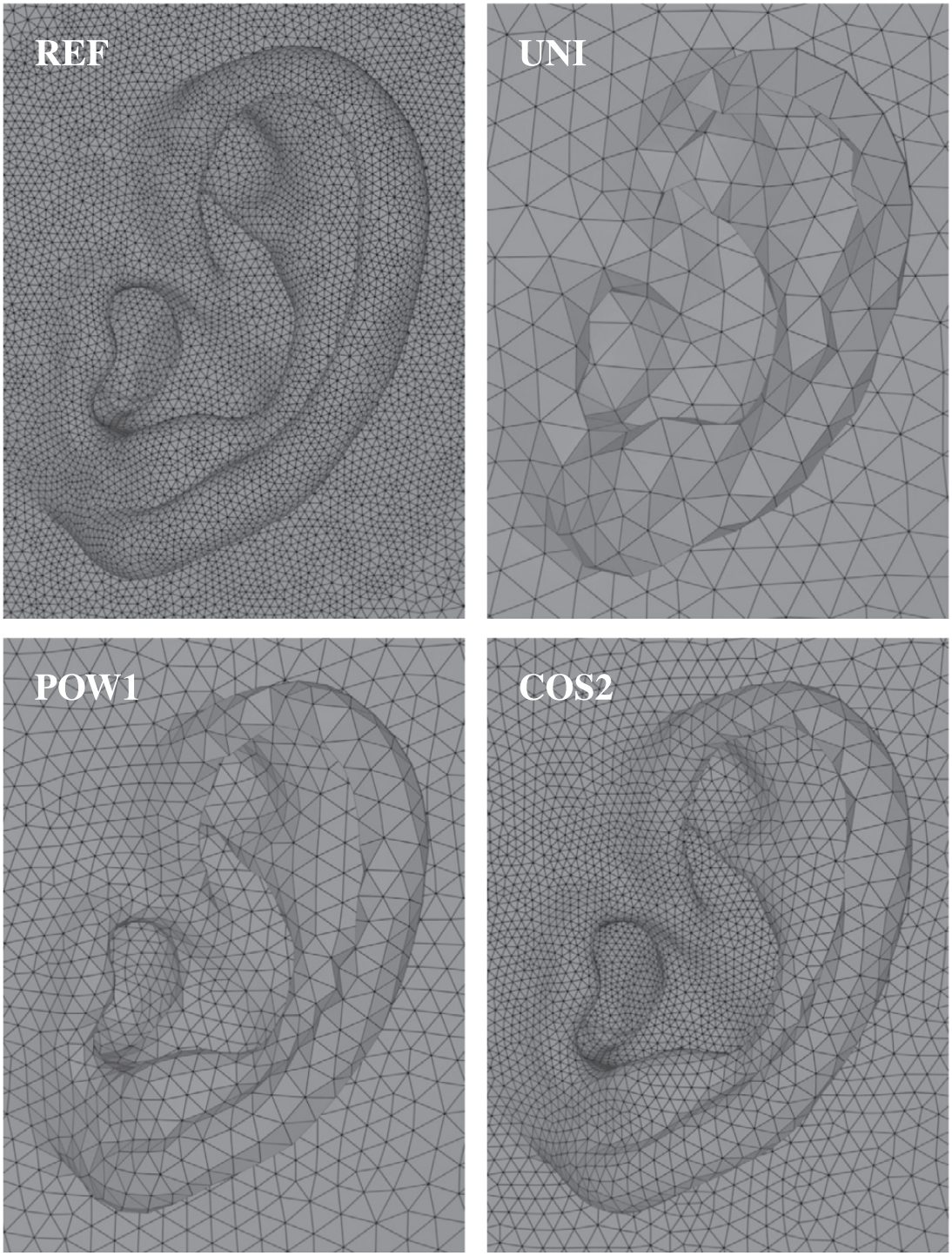


Figure 2.35: Different meshes compared by Ziegelwanger *et al.* (2016). REF — high resolution reference mesh (111,362 elements). UNI — uniform mesh (17,023 elements). POW1 — linearly graded mesh (12,041 elements). COS2 — raised-cosine graded mesh (13,633 elements). Images from Ziegelwanger *et al.* (2016).

software

They did not carry out any thorough verification of their results but the mesh generated from the photos appears to be of high quality. Bonacina *et al.* (2016) investigated using low-cost infrared (IR) stereo-vision equipment to estimate 3D models. They found it resulted in a average error of 2 dB in the HRTFs generated using FM-BEM compared to a reference high accuracy laser-scanned mesh.

2.4.7 HRTF smoothing

If HRTFs are to be estimated it is important that the estimation is as efficient as possible. Since the frequency selectivity of the human ear varies with frequency due to the variation in bandwidth of the auditory filters, not all the spectral detail in an HRTF may be perceptually salient (Carlile *et al.*, 2005). This has important ramifications in terms of the efficiency of HRTF estimation or when studying the morphological origin of HRTF features (chapters 4–6). A number of studies have investigated spectral smoothing to try and discern what level of spectral detail is required by the ear and all agree that HRTFs can be smoothed to some extent.

Kulkarni and Colburn (1998) took the Fourier transform of the HRTF log-magnitude spectrum and systematically discarded higher frequency coefficients to steadily increase smoothing. This approach is based on cepstral techniques (Bogert *et al.*, 1963; Childers *et al.*, 1977; Oppenheim and Schafer, 2004), specifically liftering (Juang *et al.*, 1987). They compared four real, free-field sources in the horizontal plane (0°, 45°, 135° and 180°) to virtual sources presented by tube-phones. The magnitude spectra of the HRTFs used to generate the virtual sources were constructed using 512 (full resolution), 256, 128, 64, 32, 16 and 8 coefficients and the phase spectrum calculated as the minimum phase spectrum of the resulting magnitude spectrum. A two interval, two-alternative, forced-choice discrimination task was used to find the percentage of the time that the listeners could successfully discriminate between the real and virtual sources for different levels

of smoothing. They found that when retaining 32 coefficients or more all the subjects performed within the bounds of chance for all directions tested. For 16 coefficients they found that all subjects performed within chance for the two rear directions (135° and 180°) but for 45° two of the subjects performed at above chance and for 0° three of the subjects performed at better than chance.

Breebaart and Kohlrausch (2001) presented a smoothing approach based on the gammatone filter bank (Patterson *et al.*, 1987). The gammatone filter order was decreased from 3 to 0.5 and subjects were asked to rate the difference between stimuli filtered with unaltered HRTFs and smoothed HRTFs. Non-individualised HRTFs were used and the difference scale consisted of three possible responses: no audible difference, a small audible difference and a large audible difference. HRTF phase smoothing was also tested, both without magnitude smoothing and in combination with magnitude smoothing. They found that there was no difference for order 1 and above for just magnitude smoothing or just phase smoothing, and no difference for order 2 and above when applying smoothing to both the magnitude and phase spectrum.

Senova *et al.* (2002) used HRIR truncation in order to smooth HRTFs. HRIRs were truncated from 20.48 ms (1024 filter taps at the sampling rate of 50 kHz used in the study) to 0.32 ms (16 filter taps) and localization error between real and virtual sources was measured. HRIRs were measured for a total of 354 directions approximating 10° resolution over the sphere and directions were chosen psuedorandomly during testing. Results varied quite significantly between the three subjects tested. One subject only showed a significant decrease in localisation performance when the HRIRs were truncated to 0.32 ms, whilst the other two subjects showed a significant decrease in localisation performance for HRIRs of length 2.56 ms and 5.12 ms. They suggested that the low spectral detail required by the first subject was due to their poor baseline localisation performance.

Pec *et al.* (2013) investigated HRTF simplification using the stationary wavelet transform (SWT) (Fowler, 2005) and PCA. The HRTFs were warped onto the Bark scale using bilinear conformal mapping (Smith and Abel, 1999) and smoothing was applied through three levels of SWT decomposition or retention of the first six principal components. The smoothed and warped HRTFs were then implemented as warped finite impulse response (FIR) or warped infinite impulse response (IIR) filters (Karjalainen *et al.*, 1997) and compared to reference FIR HRIRs. Subjective comparisons of both timbre and location of virtual sources were made and it was found that SWT smoothing slightly outperformed PCA smoothing for noise sources, but performed similarly for speech and synthesised organ. For both smoothing approaches it was found that warped and smoothed filters of 24 coefficients outperformed the reference unsmoothed HRIRs.

Rasumow *et al.* (2014) combined complex fractional-octave HRTF smoothing (Hatziantoniou and Mourjopoulos, 2000) with phase simplification (replacing the phase response above 5 kHz with linear phase). They used a three-interval three-alternative forced choice paradigm in combination with a one-up, one-down staircase method to find the threshold for smoothing for four directions in the horizontal plane: 0°, 90°, 225° and 315°. They found that for the majority of their subjects the threshold was between one octave and two thirds of an octave (two to four ERBs) but the lowest threshold was one fifth of an octave (approximately one ERB).

Recently, Hassager *et al.* (2014, 2016) used a gammatone filter bank with variable bandwidth, rather than order, to investigate the effect of binaural room impulse response (BRIR) smoothing on externalisation. The direct and reverberant portions of the BRIR were smoothed independently using bandwidth factors between 0.316 and 64 ERBs and subjects were asked to rate externalisation of virtual sources on a scale of 1 (for sounds perceived inside the head) to 5 (for sounds perceived as coming from the physical loudspeaker) for two directions in the horizontal plane (0° and 50°). They found little effect on externalisation for smoothing the reverberant portion of the BRIR, whilst smoothing the di-

rect portion with bandwidths broader than one ERB had a significant effect on externalisation.

Xie and Zhang (2010) investigated whether the ipsilateral and contralateral HRTFs could be smoothed to lesser/greater degrees. They smoothed the HRTFs above 5 kHz with a moving frequency window of varying ERB widths. For directions in the horizontal and lateral planes they smoothed either both the HRTFs or just the ipsilateral or contralateral HRTF. For both conditions the ipsilateral HRTF could be smoothed with a bandwidth of 2 ERBs compared to the 3.5 ERBs for the contralateral HRTF. In the median plane they only investigated smoothing both the HRTFs and found that they could be smoothed with a bandwidth of 2 ERBs which agreed with their results for the horizontal and lateral planes.

2.5 The relationship between auditory cues and morphology

This section presents a review of prior work on the morphological origin of spectral HRTF cues. The relationship between HRTF features and morphology is of importance not only for deepening understanding of human sound localisation, but to facilitate the synthesis of perceptually valid individualised HRTFs from a reduced set of morphological measurements: one of the long term goals of this work. This section first presents work based on acoustic measurements of the resonant behaviour of the cavities of the human pinnae before reviewing work on relating spectral notches in the HRTF spectrum to reflecting surfaces within the pinnae. Finally prior investigations utilising acoustic simulation techniques are presented, including morphoacoustic perturbation analysis (MPA), the improvement of which is one of the key contributions of this work.

2.5.1 Acoustic measurements

Pioneering acoustic measurements were carried out by Edgar Shaw and his colleagues over a number of years (Shaw and Teranishi, 1968; Teranishi and Shaw, 1968; Shaw, 1974b,a, 1975, 1979, 1997) in order to investigate the morphological features responsible for a number of spectral details identified in measured HRTFs. Shaw and Teranishi (1968) made measurements of a rubber replica ear, mounted in a metal, gelatin-filled container, fixed to a rigid mounting plane (figure 2.36). A dynamic driver was fitted to a 30 cm long, 1 cm diameter tube that could provide a constant source intensity, ±1 dB, across the replica ear from 1–15 kHz and measurements were made in both the horizontal and frontal planes

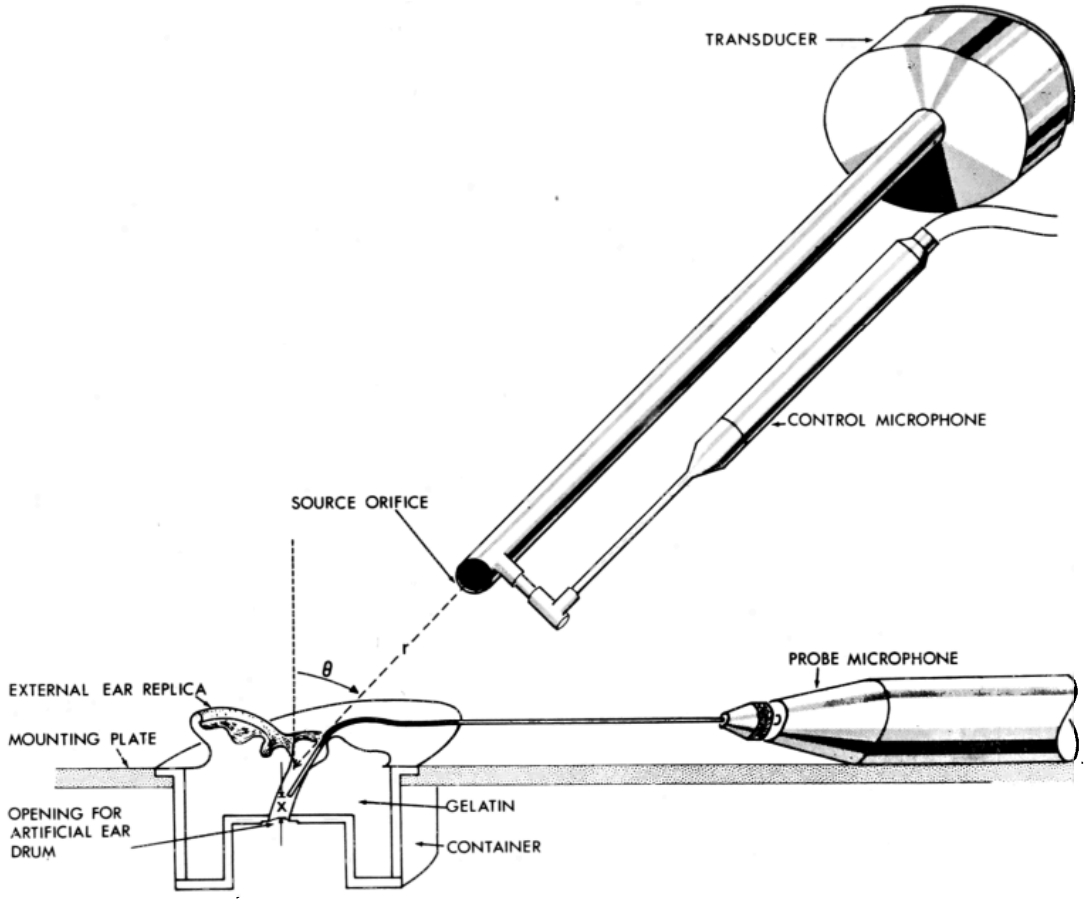


Figure 2.36: Cross-section of experimental setup used by Shaw and Teranishi (1968). Measurements were made in the replica external ear using a probe microphone with an acoustic source at distance r and angle of incidence θ, measured with respect to the mounting plane and the centre of the ear-canal entrance. Image from Shaw and Teranishi (1968).

using a probe microphone under both open-meatus and blocked-meatus conditions (section 2.3.4). The measurement of both conditions facilitated the division of spectral features into those introduced by the pinna and those introduced by the ear canal.

Comparison of the measured responses under open- and blocked-meatus conditions, with the source aligned with the interaural axis, indicated that the open-meatus measurements showed five resonant peaks, $M1$–$M5$, whilst the blocked-meatus measurements only three, $C1$–$C3$. $M1$ and $M4$ were only observed in the open-meatus measurements, $M3$ and $C2$ were observed at the same frequency in both measurements, whilst $M2$ corresponded with $C1$ and likewise $M5$ with $C3$, although shifted slightly higher in frequency under open-meatus conditions. It was suggested that $M1$, which appeared around 3 kHz, was the quarter-wavelength resonance of the ear canal, which agrees with other studies (Wiener and Ross, 1946; Algazi *et al.*, 1999), given that Shaw and Teranishi (1968) acknowledge that the replica ear canal had a larger cross-sectional area than typically seen in real ears. $M2/C1$ was attributed to a quarter-wavelength depth resonance in the concha whilst $M3/C2$ was suggested to be the result of transverse mode resonance within the concha. $M4$ and $M5/C3$ were theorised to arise from a single resonance in the concha around 12 kHz.

The authors also carried out measurements of the real ears of six subjects, again under both open- and blocked-meatus conditions. In the real measurements they identified seven features: F_1–F_3, f_1–f_3 and f_v as shown in figure 2.37. Shaw and Teranishi (1968) suggested correspondence between the peaks F_1, F_2 and F_3 and $M1$, $C1$ and $M3$ respectively. The minima f_1 and f_3 were found to be consistent with quarter- and three-quarter-wavelength standing waves in the canal, whilst f_2 was consistent with a half-wavelength standing wave. The authors highlighted the notch f_v appearing around 8 kHz due to the fact it appeared in both the open- and blocked-meatus measurements. They found that it was largely independent of azimuthal angle but highly dependent on elevation and hence it is probable that it is the "pinna spectral notch" identified in other studies as a key cue for elevation

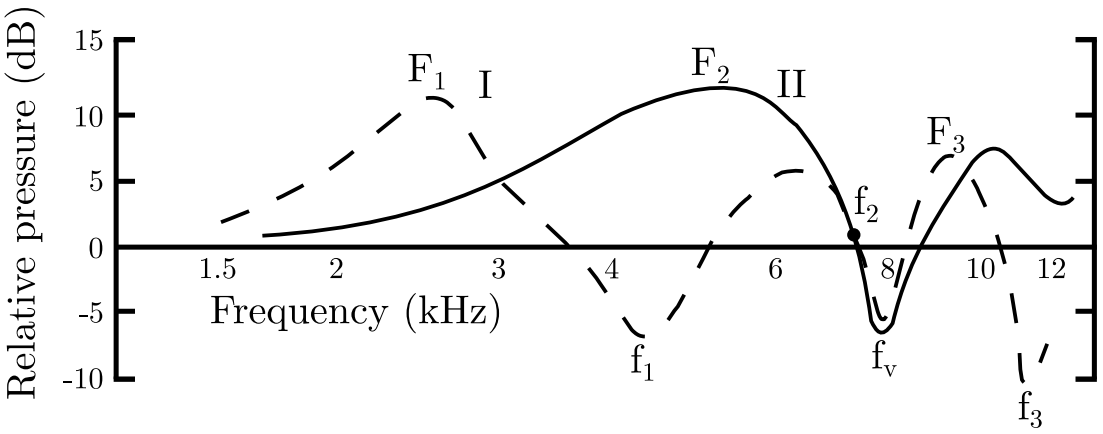


Figure 2.37: Real-ear responses measured by Shaw (1975) under open-meatus (I) and blocked meatus (II) conditions. F_1, F_2 and F_3 are the response maxima, f_1, f_3, and f_v the minima, and f_2 a tangential intersection. The curves represent the average values of the identifiable features from six subjects. After Shaw (1975).

discrimination (section 2.2.4.1). Whilst it was suggested that the notch was an interference pattern of some sort there was no suggestion of the morphological origin.

In later work Shaw (1975) focused on highly acute measurements of the modal response of the concha in which six modes were identified, as shown in figure 2.38. The measurements were made using a probe microphone and either a point source or a source that approximated plane waves. The results shown represent the average of ten subjects.

The first mode is excited equally regardless of the angle of incidence of the incoming sound and is attributed to a quarter-wavelength depth resonance within the concha (Teranishi and Shaw, 1968; Shaw and Teranishi, 1968). Shaw (1997) split the other modes into two groups based on how their nodal patterns divide the concha: "vertical" (modes 2 and 3) and "horizontal" (modes 4, 5 and 6). This grouping also corresponded with the angles of incidence for which the modes were maximally excited.

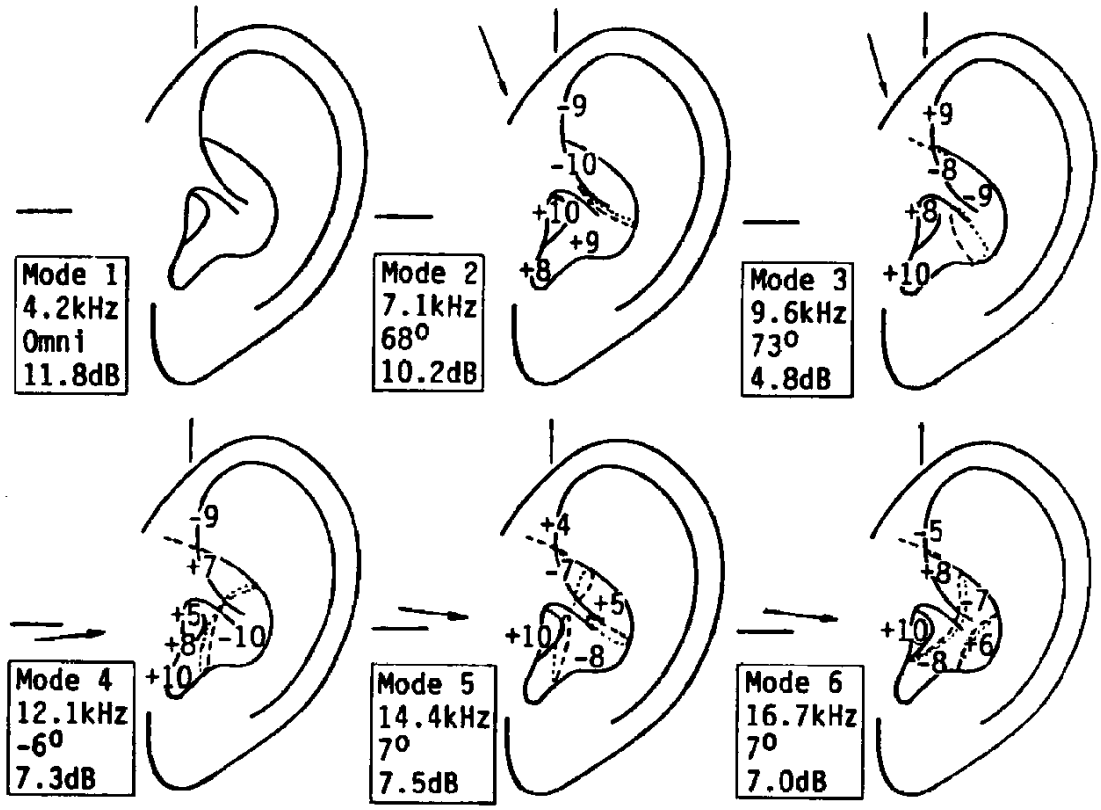


Figure 2.38: Visual representation of pinna modes identified by Shaw and Teranishi (1968) as reported in table 2.1. Arrows indicate the angle of incidence that excites each mode the most. Numbers are relative values of sound pressure whilst the signs (±) indicate phase. Dashed and dotted lines divide the concha into different nodal surfaces. Labels give the modal number, its frequency, the angle at which it is maximally excited and its magnitude. Image from Kahana and Nelson (2006).

2.5.2 Reflection models

A number of investigations into the morphological origin of auditory cues focus on the relationship between reflecting surfaces within the pinna and notches within the pinna-related transfer function (PRTF) (Raykar *et al.*, 2005; Satarzadeh *et al.*, 2007; Spagnol *et al.*, 2010, 2013; Spagnol and Avanzini, 2015). To facilitate this the pinna is treated as a simple reflection model as suggested by Batteau (1967) and shown in figure 2.39. The measured signal $y(t)$, i.e. the HRIR, is modelled as the sum of a direct incident wave $x(t)$ and a delayed, reflected copy of the incident wave:

$$y(t) = x(t) + ax(t - t_d(\varphi)) \tag{2.22}$$

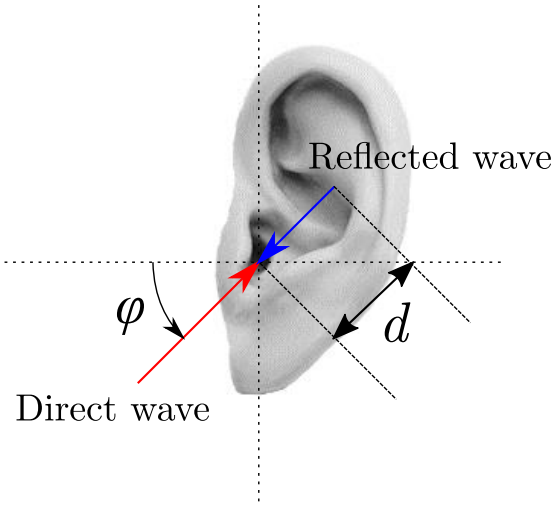


Figure 2.39: Simple reflection model for pinna spectral notches. The direct wave (red arrow) at an incident angle of φ is reflected by the wall of the concha. This reflected wave (blue arrow) travels an extra distance of $2d$, arriving $\frac{2d}{c}$ later, where c is the speed of sound. After Raykar *et al.* (2005).

where a is a reflection coefficient, φ the elevation angle of incidence and $t_d(\varphi)$ the time delay of the reflected wave, given by:

$$t_d(\varphi) = \frac{2d(\varphi)}{c} \tag{2.23}$$

where c is the speed of sound and $d(\varphi)$ is the distance from the entrance of the ear canal to the reflecting surface. This delay causes notches in the HRTF spectrum with centre frequencies of:

$$f_n(\varphi) = \frac{2n+1}{2t_d(\varphi)} = \frac{c(2n+1)}{4d(\varphi)} \quad n = 0, 1, 2 \ldots \tag{2.24}$$

and hence the centre frequency of the first spectral notch is:

$$f_0(\varphi) = \frac{c}{4d(\varphi)} \tag{2.25}$$

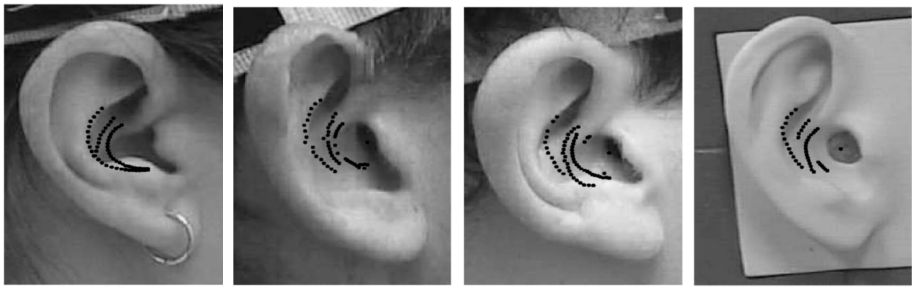
Conversely, given the centre frequency, f_c, of a notch, extracted from the HRTF, the distance of the reflecting surface from which it arises is:

$$d(\varphi) = \frac{c}{4f_c(\varphi)} \tag{2.26}$$

Raykar *et al.* (2005) used this equation to calculate the distance of the reflecting surface responsible for pinna spectral notches that they extracted from the HRTFs of a number of CIPIC database subjects using the approach introduced in their earlier paper (Raykar *et al.*, 2003). The variation of the distance with elevation was then plotted on the photographs of the CIPIC subjects' pinnae available in the database. Examples for three subjects and a KEMAR ear are shown in figure 2.40. The plotted contours matched some features of the pinna morphology, particularly the rear wall of the concha (figure 2.40a), however some of the distance contours did not seem to match any obvious reflecting surfaces, and all the distance contours were concentrated within the concha.

The reason for this was identified by Satarzadeh *et al.* (2007). Equation 2.24 assumes that the reflecting surface is hard, and therefore the reflected wave does not change phase upon reflection. In this case the notches appear when the distance difference $2d$ is equal to half a wavelength, as per equation 2.24. However, if the phase of the reflected wave is reversed upon reflection, which might happen as the wave passes over the rim of the pinna, due to the impedance difference between the pinna itself (high impedance) and the air space behind it (low impedance), then notches will appear at full-wavelength delays. Therefore equation 2.24 becomes:

$$f_n(\varphi) = \frac{n+1}{t_d(\varphi)} = \frac{c(n+1)}{2d(\varphi)} \quad n = 0, 1, 2 \ldots \tag{2.27}$$



(a) (b) (c) (d)

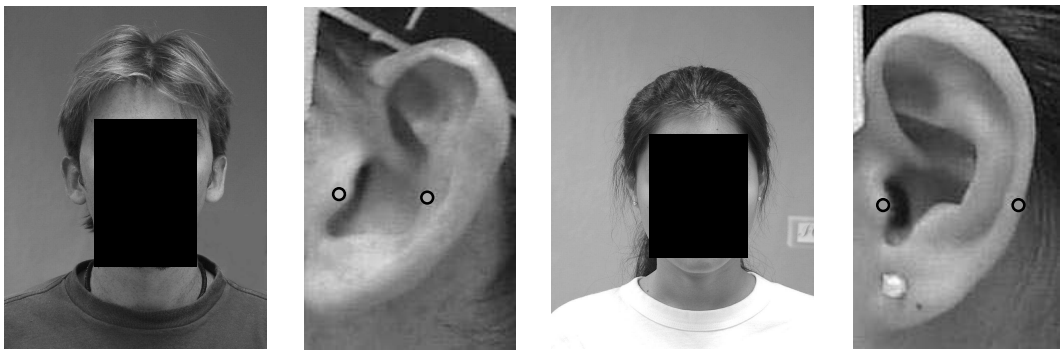
Figure 2.40: Distances, calculated by Raykar *et al.* (2005), corresponding to pinna spectral notches for different elevation angles in the median plane, marked on pinna photos of (a) subject 10, (b) subject 27, (c) subject 134 and (d) subject 165 of the CIPIC database. Images from Raykar *et al.* (2005).

and correspondingly, equation 2.26 becomes:

$$d(\varphi) = \frac{c}{2f_c(\varphi)} \tag{2.28}$$

Satarzadeh *et al.* (2007) used both equation 2.26 and equation 2.28 to calculate distances from the frequencies of the pinna spectral notches of a number of CIPIC database subjects. They found that the distances usually indicated a reflection point on the back wall of the pinna or on the rim of the helix. They suggested that the criterion for choosing between the two cases was the flare of the pinnae, i.e. the angle between the head and the pinnae. They reasoned that a large flare (figure 2.41a) resulted in the rear wall of the concha being the primary reflecting surface (figure 2.41b), whilst a smaller flare (figure 2.41c) left the helix as the reflecting surface (figure 2.41d). Interestingly, they only considered a single delay-and-add model and did not suggest that both morphological features might be the source of spectral notches for the same subject.

Spagnol *et al.* (2010, 2013) also considered both in and out of phase reflected waves as the origin of pinna spectral notches, but extended their work to consider multiple morphological features working in tandem to produce different pinna spectral



(a) (b) (c) (d)

Figure 2.41: Reflecting surfaces in the pinna responsible for pinna spectral notches, as identified by Satarzadeh *et al.* (2007). They suggested that large pinna flare, (a), corresponded with the rear wall of the pinna as the reflecting surface, (b), whilst smaller pinna flare, (c), corresponded with the helix as the primary reflecting surface, (d). Photographs are: (a) frontal view of CIPIC subject 27, (b) pinna view of same subject, (c) frontal view of CIPIC subject 18, (d) pinna view of same subject. Images from Satarzadeh *et al.* (2007).

notches. They extracted pinna spectral notches from the measured HRTFs of the 20 CIPIC database subjects for whom photographs and anthropometric data is available. The extracted notches were grouped into three notch "tracks" T_1–T_3, increasing in frequency. A track was defined as a series of notches whose centre frequency varied with elevation, were deeper than 5 dB and remained within the frequency range 4–16 kHz, where pinna cues are generally found (section 2.2.4.3). They considered five contours, C_1–C_5, of the pinna as the possible reflecting surfaces responsible for the notches: the helix border, the helix inner wall, the concha border, the antihelix/concha wall and the crus helias (figure 2.42).

Each notch track was translated into two distance contours using equations 2.26 and 2.28 and a scoring function was derived based on the Euclidean distance between each distance contour and the contours C_1–C_5, which were hand-drawn onto the pinna photos. The assumption was made that each notch track originated from a single morphological contour within the pinna and so each track was associated with a contour based on the best score. They found that for all their subjects C_4 and C_3 corresponded most closely with the centre frequencies

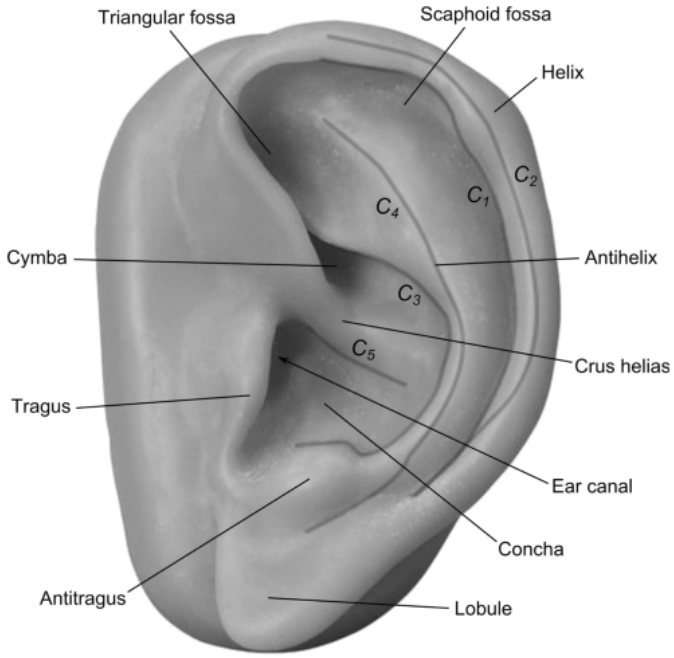
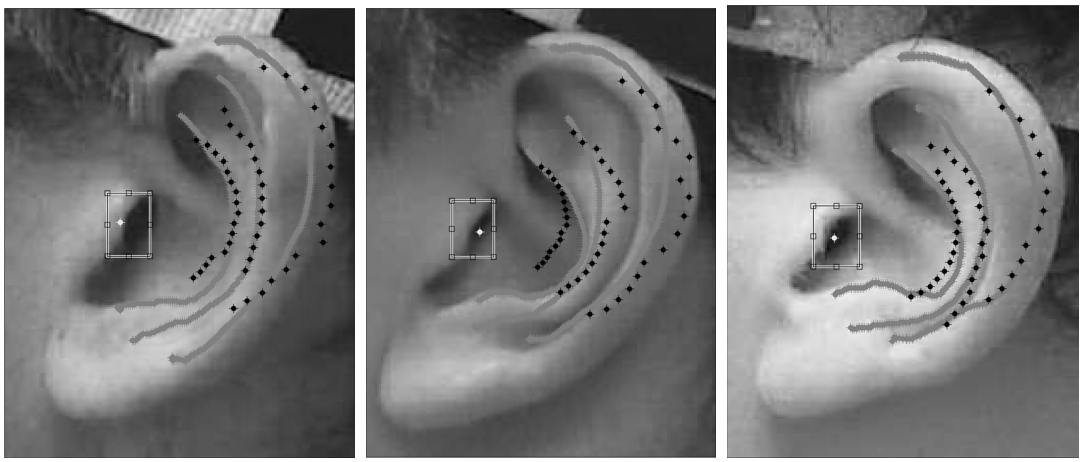


Figure 2.42: Pinna contours considered for generation of spectral notches by Spagnol *et al.* (2013). C_1: helix border; C_2: helix inner wall; C_3: concha border; C_4: antihelix/concha wall; C_5: crus helias. Image from Spagnol *et al.* (2013).



(a) (b) (c)

Figure 2.43: Optimal ray-traced reflection surface contours calculated by Spagnol *et al.* (2013). The grey lines are hand-drawn pinna contours, whilst the black points represent distance contours of the pinna spectral notches most closely matched with the surface contours. (a) subject 027, (b) subject 050 and (c) subject 134. Images from Spagnol *et al.* (2013).

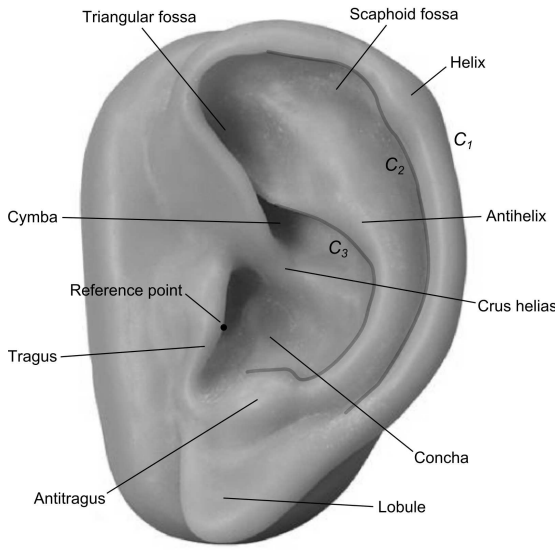


Figure 2.44: Pinna contours considered for generation of spectral notches by Spagnol and Avanzini (2015). Note difference between contour C_1 in this study and contour C_2 as used by Spagnol *et al.* (2013) (figure 2.42). Contour C_3 is the same in both studies and contour C_2 in this study and contour C_1 in Spagnol *et al.* (2013) are the same, the index has just changed. Image from Spagnol and Avanzini (2015).

of notches T_2 and T_3 respectively. Whereas approximately 50% of their subjects showed correspondence between the centre frequency of notch T_1 and contour C_1, whilst the other 50% showed correspondence with contour C_2. Figure 2.43 shows the optimally matched surface and distance contours for three of their subjects. Subject 027 (figure 2.43a) exhibited a final optimal score close to the median, subject 050 (figure 2.43b) had the second worst score and subject 134 (figure 2.43c) the third best. They suggested that one reason for errors might be the assumption that each notch is the result of a single reflection path, since very few, if any, surfaces in the pinna are completely flat. Their model does not consider the multiple reflection paths that would, for instance, occur from a concave surface. They also highlighted that it is impossible to accurately trace the pinna contours on a 2D photograph and that 3D models might be needed.

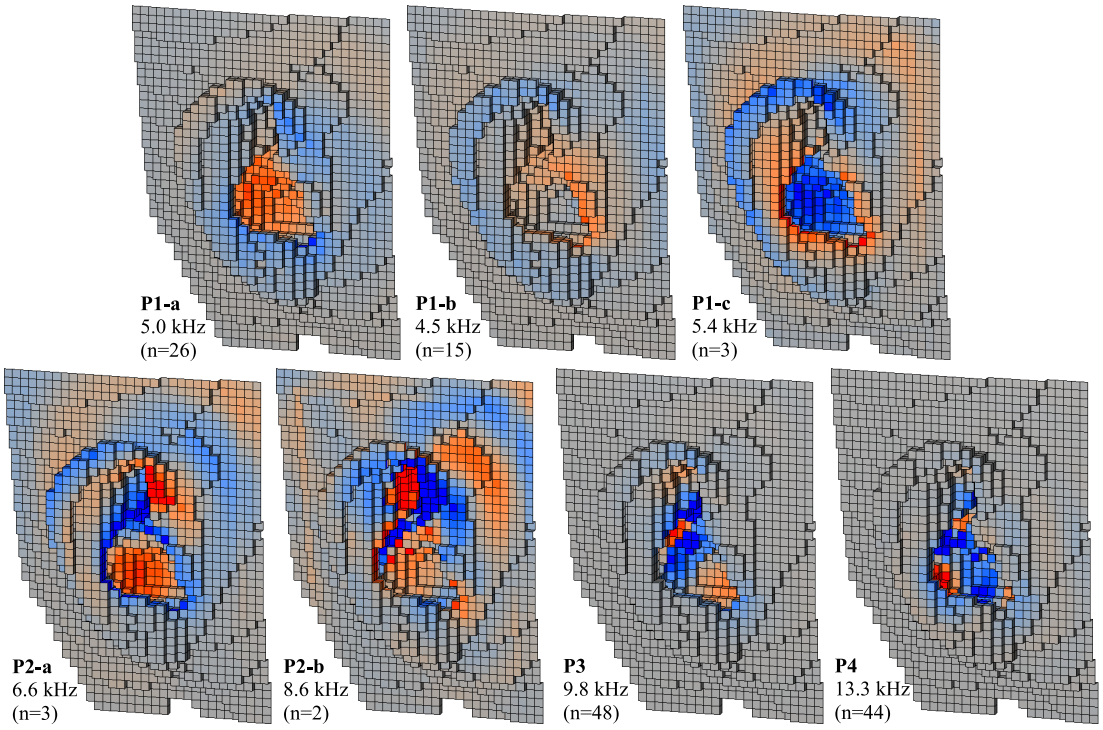
Spagnol and Avanzini (2015) also extracted three notch tracks, N_1–N_3, from the HRTFs of the CIPIC database. They then used a linear regression model to relate thirteen anthropometric measurements from the CIPIC database to the centre frequencies of the pinna notches. The thirteen measurements were the ten anthropometric parameters of the pinna available as part of the database, plus three extracted pinna contours, C_1–C_2, as shown in figure 2.44. They found that eight of the thirteen available measurements were able to predict the centre frequencies of the first pinna notch with a reasonable degree of accuracy. However, they found that the thirteen measurements considered were insufficient to predict the second or third notches.

2.5.3 Acoustic simulations

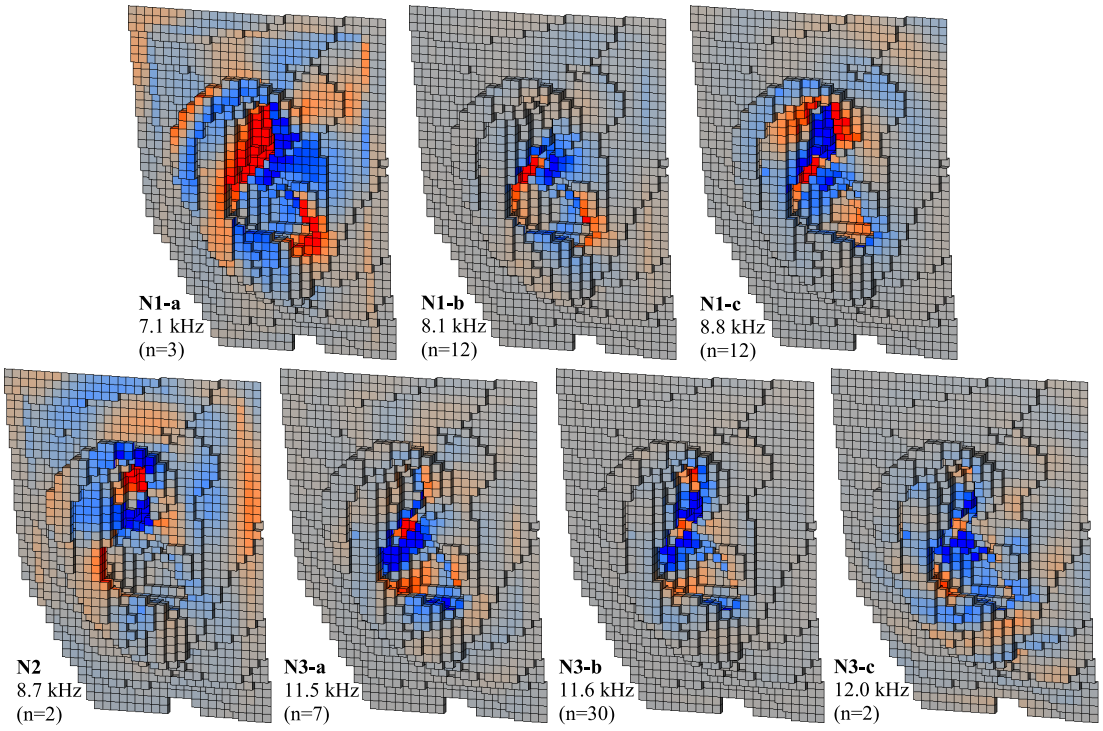
Acoustic simulation techniques (section 2.4.6) such as the BEM and FDTD method are strong tools for investigating the relationships between morphology and HRTF features due to the relative ease with which the meshes can be altered and simulations re-run.

Mokhtari *et al.* (2010) used finite difference time domain (FDTD) simulations of the DB60 KEMAR pinna to explore the spectral effect of the morphological features of the pinnae. A mesh of the pinna and an adjacent patch of the head was generated using 2 mm voxels and initial baseline simulations were run for 45 locations at a radius of 1 metre. A peak-picking algorithm was used to extract the centre frequencies and amplitudes of the peaks and notches in the baseline PRTFs and then single additional voxels were systematically added to each point on the surface of the pinna to create "micro-perturbations" and the simulations re-run. The frequency shifts in features introduced by each micro-perturbation were then calculated and plotted on the surface of the pinna for a given feature (figure 2.45). Their results generally agreed with previous studies (Shaw, 1997; Kahana *et al.*, 1999; Kahana and Nelson, 2006) and indicated that the first pinna resonance was the result of a depth resonance of the concha, the second and third resonances were the result of vertical modes and the fourth a more complex resonance involving both horizontal and vertical components. Their analysis of the pinna notches also corroborated other studies, with the first pinna notch attributed to elevation-dependent cancellation due to reflections from the rear wall of the concha.

Mokhtari *et al.* (2011) used the same micro-perturbation technique to study the mechanisms responsible for the first pinna notch in more detail. They carried out analysis for 25 elevations in the frontal median plane in steps of 5.625° from -45° to +90°, matching the CIPIC database. Again their results agree that the elevation dependent shift in the notch frequency is largely controlled by out-of phase reflections cancelling out the direct sound entering the ear. However, as shown in figure 2.46, their results also showed that the reflecting surface was spread across localised areas of the pinna, rather than a single reflection point, as Spagnol *et al.* (2013) also suggested. They also identified an additional area of sensitivity around the tragus which they attributed to the fact that the tragus lies in direct path from frontal directions and so will introduce diffraction and delay the direct sound. To test their hypothesis they carried out additional analysis for



(a)



(b)

Figure 2.45: Pinna sensitivity maps created by Mokhtari *et al.* (2010) for (a) peaks and (b) notches in pinna-related transfer function (PRTF). Cold/grey/warm colours indicate negative/zero/positive changes in centre frequency and the maximum colour saturation representing a change of 0.29% in centre frequency. The text indicates the feature, its average centre frequency and the number of voxels that contributed to its variation. Images from Mokhtari *et al.* (2010)

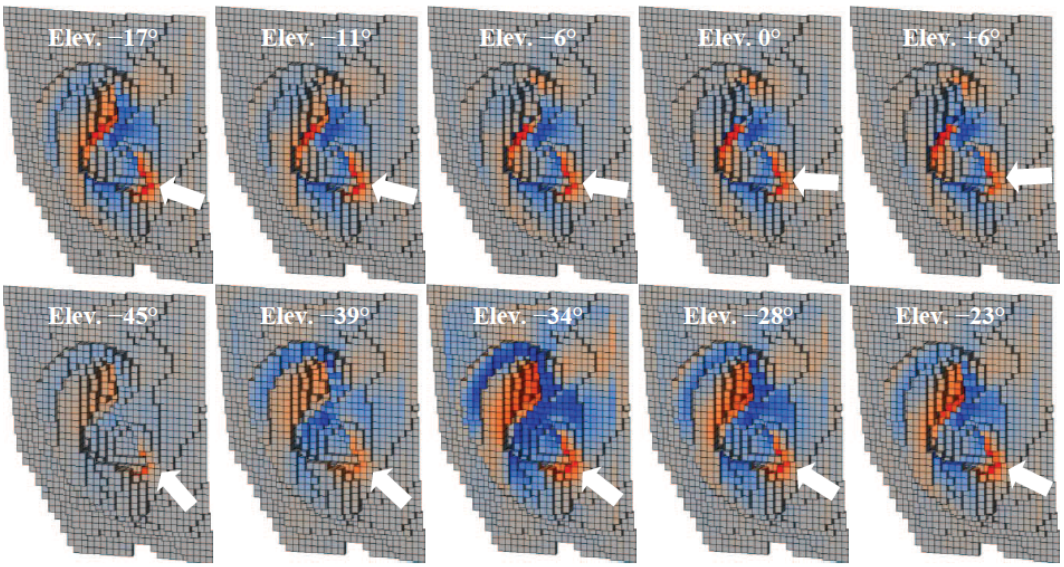


Figure 2.46: Pinna sensitivity maps for various elevations in the frontal median plane, as reported by Mokhtari *et al.* (2011). Arrows indicate angle of incidence. Colour map as per figure 2.45, with maximum saturation representing a change of 0.36% in centre frequency. Image from Mokhtari *et al.* (2011).

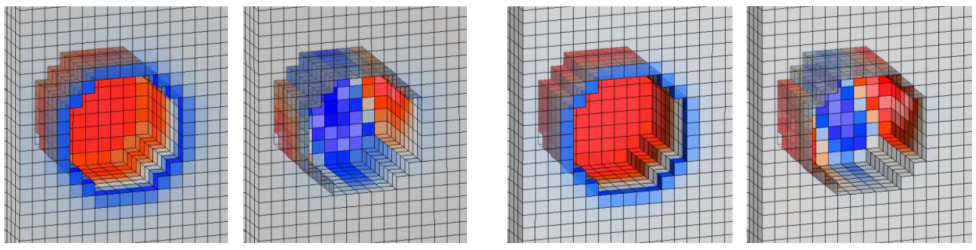
two directions away from the median plane (+30° and +65°) and found that the area of sensitivity around the tragus disappeared.

In contrast to their earlier micro-perturbation technique, Takemoto *et al.* (2012) extracted the peaks and notches of the PRTFs and used sinusoidal excitation signals at the given frequency of each feature to record the steady-state pressure distributions within the pinna. Their analysis of the resonant peaks agreed with the earlier work of Shaw (1997) and Kahana and Nelson (2006) in that the first three peaks of the PRTFs were the first, second and third normal modes of the pinna respectively. Their analysis of the first pinna notch categorised the pressure distribution patterns into three types dependent on elevation. The first type appeared for sound sources below the horizontal and consisted of an anti-node in the upper pinna cavities of the pinna, that varied with elevation, partnered with a node within the concha. The second and third types both consisted of anti-nodes in the triangular fossa and cymba as well as a node in the concha, however type 2 appeared for elevated sound sources in the frontal hemisphere, whilst type 3 appeared for elevated sound sources in the rear hemisphere.

In later work, Mokhtari *et al.* (2013) suggested that their previous micro-perturbation

method was both time consuming, due to the numerous simulations required, and error-prone, due to the processing of the PRTFs required to extract the PRTF features. They therefore proposed a new approach based on acoustic radiation pressure. Acoustic radiation pressure is defined as the time-average force exerted by a sound field on a unit surface. By carrying out a single FDTD simulation of the unperturbed pinna, and splitting the results into potential and kinetic energy densities based on both the pressure and velocity results, the acoustic sensitivity to a single voxel perturbation can be calculated. They validated their approach using a simple cylindrical model of the concha and compared it to results obtained using their earlier approach (Mokhtari *et al.*, 2010, 2011). The two approaches did not produce identical results (figure 2.47), but did show similar trends and the authors considered the approach to be accurate enough to analyse the peaks in the PRTF of a human pinna, for which they found similar results to previous studies. However, there was no suggestion as to the suitability of their new approach to analysing notches in the PRTF.

The acoustic radiation pressure approach was developed further by Mokhtari *et al.* (2014) to include not only pressure distributions, but also velocity vectors indicating the movement of air between the cavities of the pinna. The velocity vectors were calculated by applying PCA to the particle velocity results of the FDTD simulations to isolate the direction and strength of oscillating vectors. These vectors could then be plotted on the sensitivity maps to indicate exchange



(a) (b)

Figure 2.47: Comparison of sensitivity maps of a cylindrical concha model for (a) the pertubation method and (b) the acoustic radiation pressure method as reported by Mokhtari *et al.* (2013). The left hand panel of each pair is the sensitivity map for the first peak in the PRTF and the right hand panel is for the second peak. The colour map is as previous figures. Image from Mokhtari *et al.* (2013).

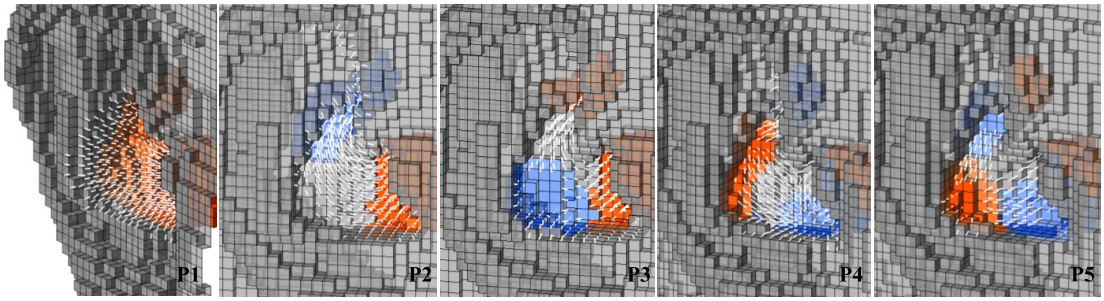
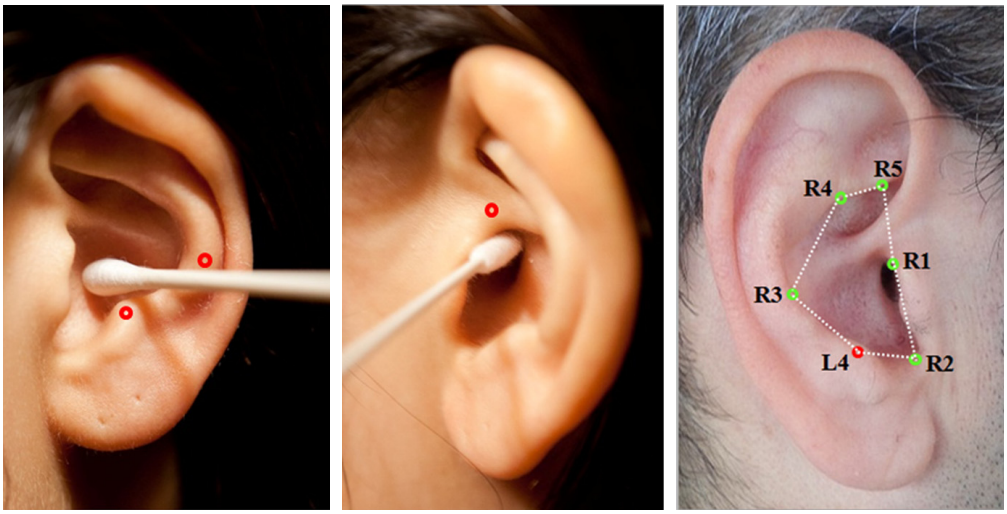


Figure 2.48: The first five modes of the pinna, showing both pressure distributions (colour map as before) and velocity vectors (white lines) as reported by Mokhtari *et al.* (2014). Image from Mokhtari *et al.* (2014).

of acoustic energy between adjacent anti-nodes as shown in figure 2.48. Again analysis was limited to resonant peaks in the PRTFs, specifically the first three modes of the pinna. Whilst the second and third mode exhibited similar pressure distributions and velocity vectors, applying thresholding to isolate pressure and velocity quantities with the same phase, suggested that the second pinna mode was the result of two independent resonances in the concha and scaphoid fossa whereas the third mode was the result of a single acoustically coupled resonance consisting of energy exchange between the two cavities.

By applying linear regression modelling to FDTD simulations and morphological measurements from the mesh geometry, Mokhtari *et al.* (2015) attempted to estimate the frequency and amplitude of the first pinna resonance: the so-called concha-depth resonance. Two different measures of concha depth were considered as well as an estimate of the diameter of the concha aperture. It was found that the frequency of the first PRTF peak was best estimated (correlation coefficient of $r = 0.84$) by the longest measurement of concha depth from the base of the rear of the cavum concha to the surface of the antitragus and antihelix (figure 2.49a). The amplitude of the resonance was best estimated ($r = 0.83$) by the shortest measure of concha depth, i.e. the lateral distance from the ear canal entrance to the side of the cheek near the anterior notch (figure 2.49b), as well as the equivalent diameter of the concha aperture (figure 2.49c). These results suggest that the resonant frequency is therefore related to the quarter-wavelength of the longest depth of the concha and that the amplitude of the resonance varies



(a) (b) (c)

Figure 2.49: Practical anthropometric measurements suggested by Mokhtari *et al.* (2015) for estimation of the frequency and amplitude of the first pinna resonance. (a) suggested depth measurement from base of the concha to the antitragus and antihelix for estimation of resonant frequency. (b) suggested depth measurement from ear canal entrance to the anterior notch for estimation of the amplitude of the resonance. (c) suggested approximation of concha aperture area by a six-sided polygon, also used for estimation of the amplitude of the resonance. Images from Mokhtari *et al.* (2015)

based on energy lost to the surrounding air from the concha aperture.

As part of their verification of the BEM Kahana and Nelson (2006) recreated the experiments of Shaw (1974b) using a mesh generated from a laser scanned DB-65 KEMAR pinna. Figure 2.50 shows their simulation results of the first six modes of the pinna for comparison with the results of Shaw as shown in figure 2.38. They found that their results showed similar modal patterns to those found by Shaw, although, as reported in table 2.5, there were some differences between the frequencies of the modes and the directions at which they were maximally excited.

Tao *et al.* (2003b) proposed the use of the BEM in their differential pressure synthesis (DPS) method for determining the acoustic effect of changes in morphology. The basis of DPS is that BEM simulations are run to create a database of the acoustic changes caused by applying a series of orthogonal deformations to a template mesh. Then the acoustic influence of any change in shape that can be represented as a sum of the orthogonal deformations, can be simply calculated

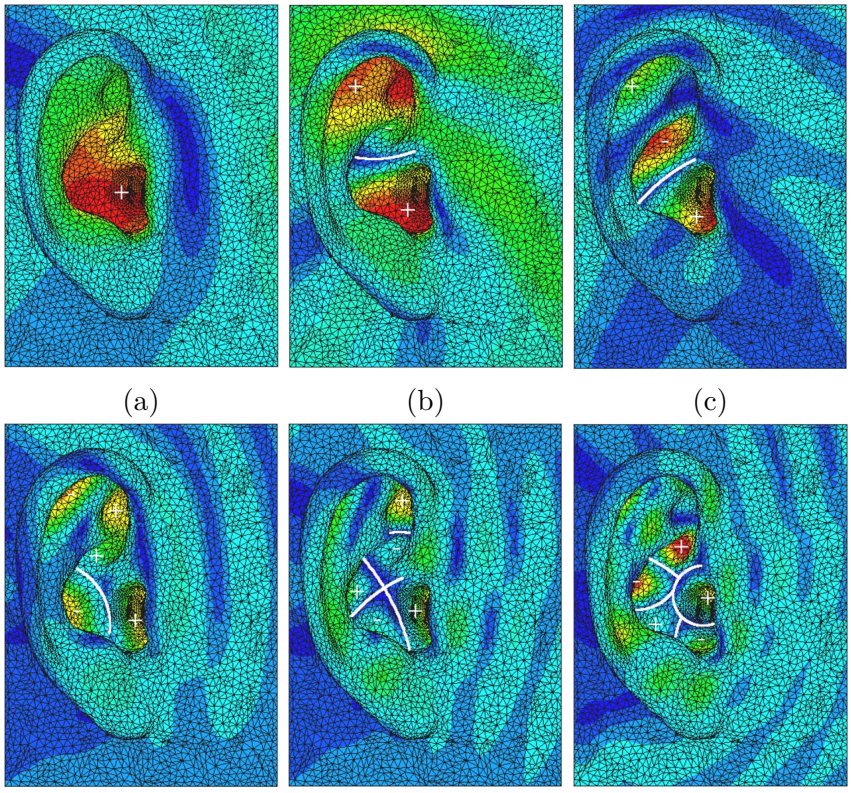


Figure 2.50: First six modes of the pinna found by Kahana and Nelson (2006) using boundary element method (BEM) simulations of the DB-65 KEMAR pinna. The plots correspond to the following frequencies: (a) 4.2 kHz, (b) 7.2 kHz, (c) 9.5 kHz, (d) 11.6 kHz, (e) 14.8 kHz and (f) 18 kHz. Colours indicate the magnitude of the surface pressure and the phase is indicated by ± signs. Images from Kahana and Nelson (2006).

Shaw (1974b)		Kahana and Nelson (2006)	
Frequency	Direction	Frequency	Direction
4.2 kHz	Omni	4.2 kHz	Omni
7.1 kHz	68°	7.2 kHz	60°
9.6 kHz	73°	9.5 kHz	94°
12.1 kHz	-6°	11.6 kHz	0°
14.4 kHz	7°	14.8 kHz	4°
16.7 kHz	7°	18 kHz	-16°

Table 2.5: Comparison of the frequencies and directions of maximum excitation for the first six modes of the pinna as reported by Shaw (1974b) and Kahana and Nelson (2006).

as the appropriately weighted sum of the acoustic changes of each deformation. In two dimensions, DPS uses radial harmonic deformations based on the Fourier series but with the polar angle θ replacing x, whilst in three dimensions surface spherical harmonicss (SSHs) are used. In a partner paper (Tao *et al.*, 2003a), DPS was used to investigate simplification of a pinna-less KEMAR head. The mesh was initially represented by SSHs up to degree 34 (figure 2.51), the harmonics gradually discarded and the pressure error introduced by the simplification calculated. It was found that discarding all harmonics above degree eleven introduced no more than 5% RMS pressure error. However, the approach is not directly applicable to the full head and pinnae because SSHs can only be used to decompose shapes that can be represented by a unique function $r(\theta, \varphi)$ and within the complex morphology of the pinnae, this is not the case.

Therefore Hetherington and Tew (2003a) proposed the use of the elliptic Fourier transform (EFT) (Park and Lee, 1987) to decompose radially sliced contours of an isolated pinna into Fourier components. However, the EFT only works on closed contours and so the slices had to be traversed start to end and then back again in order to force them to be closed. Nevertheless this still resulted in discontinuities which added unrequired energy to the spectra of the contours. In order to combat this Hetherington and Tew (2003b) applied the EFT to a slice of the whole head

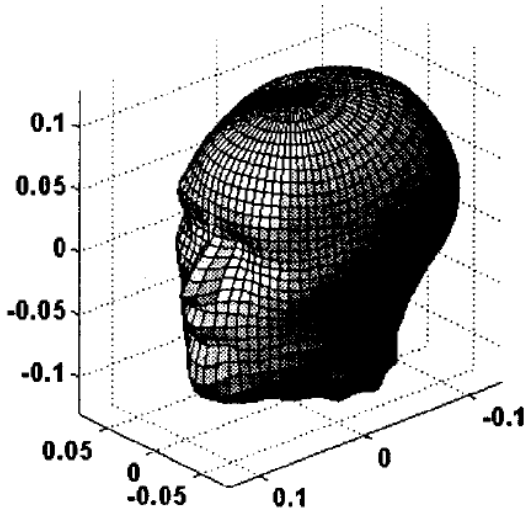
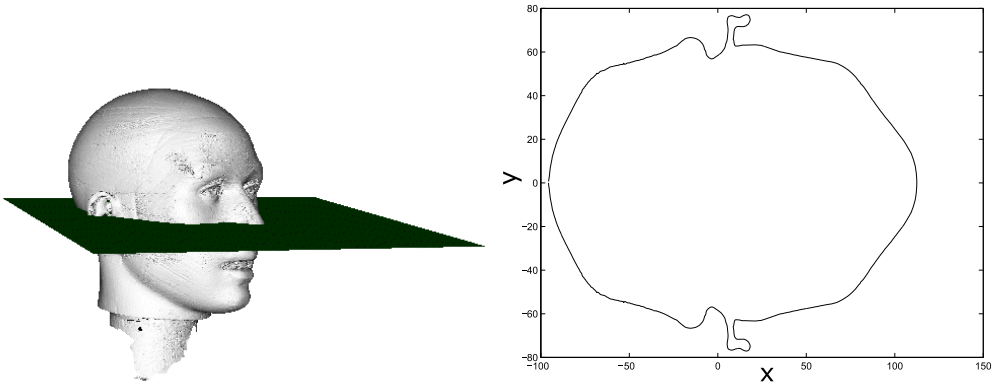


Figure 2.51: Pinna-less KEMAR head, represented by surface spherical harmonics (SSH) up to degree 34, used by Tao *et al.* (2003a). Image from Tao *et al.* (2003a).



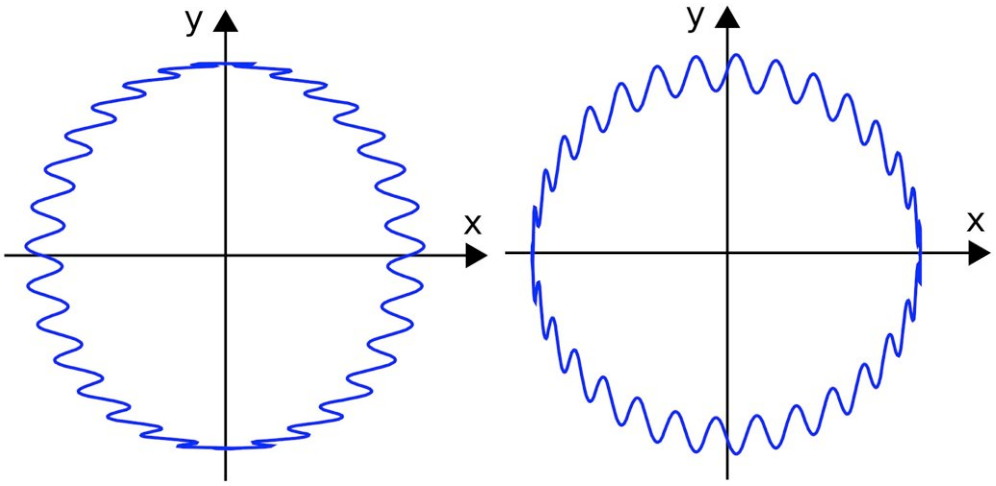
(a) (b)

Figure 2.52: Plane, (a), used to take exemplar contour, (b), for application of the elliptic Fourier transform (EFT) by Hetherington and Tew (2003b). Images from Hetherington and Tew (2003b).

and pinnae (figure 2.52a), which resulted in a closed contour (figure 2.52b). They suggested that the appropriate choice of radial slicing axis within the pinna would allow the whole head to be decomposed into contours and the EFT to be used instead of SSHs in DPS.

However, when the EFT was employed in DPS it was found that, due to the rapidly changing slope of the head contours, especially within the pinnae areas, the deformations introduced by the EFT were not evenly distributed along the surface of the head. This is due to the fact that the perturbations introduced by the EFT result in sinusoidal harmonics on either the x- or y-component of each contour slice. Therefore, when the gradient of the contour slice is close to horizontal, in the case of the x-component, or vertical, in the case of the y-component, the spatial frequency of the deformations is significantly compressed as shown in figure 2.53.

To overcome these shortcomings, Thorpe (2009) proposed the use of "elliptic surface harmonic deformations" to generate perturbations that act perpendicularly to the contour slice. For a deeper explanation see section 4.1, but briefly, the contour slices are mapped to a 2D surface on the x-y plane in a 3D Cartesian coordinate system and 2D Fourier harmonics are applied to alter the z coordinate. This change in z coordinate is then mapped back to the original contour slices to displace each point perpendicularly to the contour surface. The database of



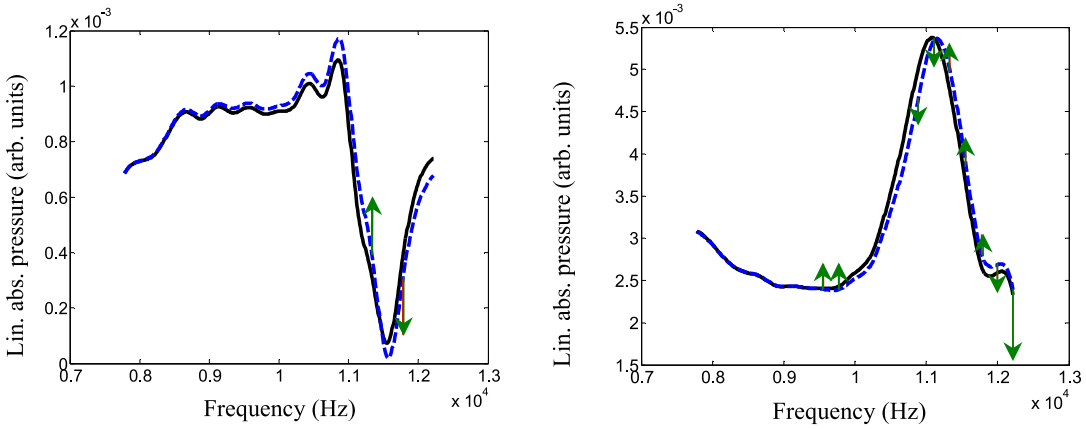
(a) (b)

Figure 2.53: Example of EFT perturbations on a contour slice of a sphere. (a) When harmonics are applied to the x-component they are compressed when the contour gradient is close to horizontal. (b) Conversely, when the harmonics are applied to the y-component they are compressed when the gradient is close to vertical. Images from Thorpe (2009).

acoustic pressure changes for each harmonic deformation is then calculated as per DPS.

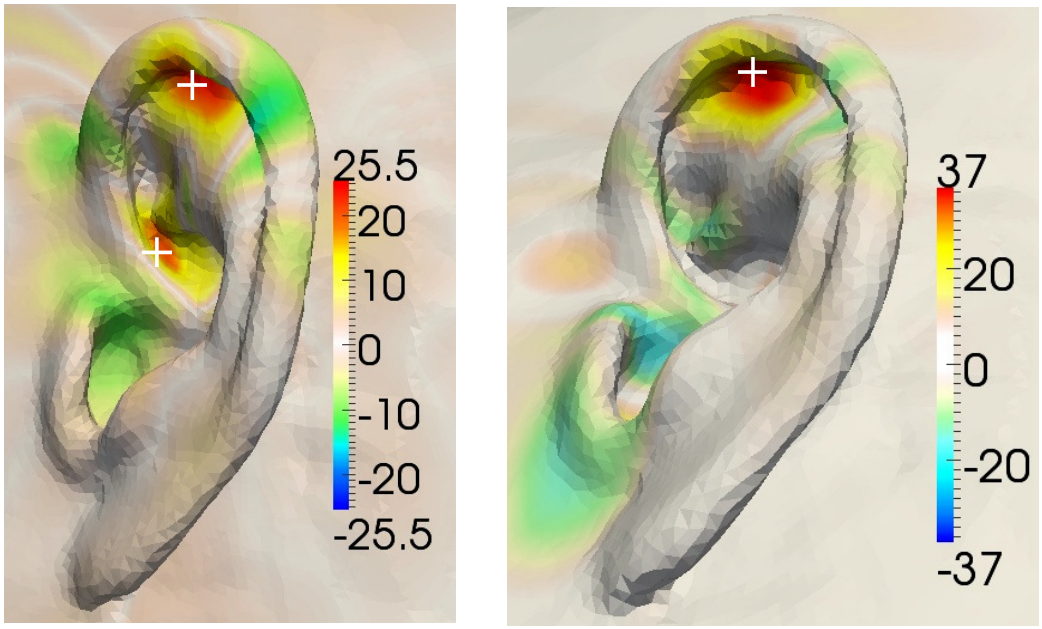
Thorpe (2009) then describes using the DPS database in reverse to map acoustic changes in the HRTF back to the pinnae morphology responsible in an approach christened morphoacoustic perturbation analysis (MPA). Prior to this, the term morphoacoustics (or sometimes morpho-acoustics) had no formal definition, but had occasionally been used when linking the shape of a physical object to its acoustical properties (Kolla and Coumes, 1983; Frey *et al.*, 2007; Conti *et al.*, 2016). Morphoacoustics is formally defined in Tew *et al.* (2012) as "the study of relationships and interactions between the morphology (shape) of an object and its acoustic properties."

Tew *et al.* (2012) applied MPA to a notch and peak in an HRTF (figure 2.55). The notch was identified for a location behind the head ($\theta = -166°, \varphi = -2°$) and a morphing vector, $\boldsymbol{m}$, was generated, corresponding to an increase in centre frequency of the notch (indicated by green arrows in figure 2.54a). By taking the dot product of the morphing vector and the matrix of pressure changes a set of weights is generated corresponding to how closely the pressure changes match



(a) (b)

Figure 2.54: Notch (a) and peak (b) investigated by Tew *et al.* (2012). Black lines show the original HRTF, the blue dashed lines show the HRTF plus the weighted sum of acoustic changes introduced by the harmonic deformations and the green arrows indicate the morphing vector. Images from Tew *et al.* (2012).



(a) (b)

Figure 2.55: Pinna sensitivity maps generated by Tew *et al.* (2012) for the notch (a) and peak (b) shown in figure 2.54. Red areas indicate that an outward movement of the surface in those areas contributes to the given change in HRTF. Images from Tew *et al.* (2012).

the morphing vector. Since the harmonic deformations are small enough for a linear relationship to hold between them and the acoustic changes they cause, the principle of superposition applies. Hence, the same weights can be used to sum the harmonic deformations and generate a sensitivity map of the pinna highlighting areas that contribute to the change in the HRTF (figure 2.55a). The same approach was applied to a peak in the HRTF (figures 2.54b and 2.55b) for a direction above the head ($\theta = 82°, \varphi = 76°$) and both results were verified by applying putty to the pinna of a KEMAR and carrying out acoustic measurements. However, as discussed further in chapter 4, there are problems with the first generation of MPA due mainly to large variations in contour length, which result in, amongst other things, the erroneous coloured streaks visible in figure 2.55. Therefore a key contribution of this work will be the development of a new method of applying harmonic deformations to a head mesh that does not suffer from the shortcomings of the earlier approach.

2.6 Summary

This chapter has provided a review of the literature relevant to this work. Firstly, the human auditory system was described in terms of the physiological structure of the human ear, from the pinna to the auditory nerve, the limits of human hearing and the nature of human auditory filters. The latter two are of particular interest to this research in terms of the new HRTF smoothing algorithm discussed in chapter 3.

Then the current understanding of human sound localisation cues and localisation acuity was discussed. This included both the interaural localisation cues, the interaural time difference (ITD) and interaural level difference (ILD), which have been well understood for over a century, and the spectral cues, which are primarily associated with the pinnae and still not fully understood. One of the long term aims of this research is to further the understanding of spectral localisation cues.

Next, the field of spatial audio was reviewed in order to provide wider context for the scope of this research. Whether developing loudspeaker-based or headphone-based spatial audio systems, a deeper understanding of the relationship between morphology and localisation cues is key.

A review of the state of art approaches to HRTF measurement and estimation served to highlight the problems associated with current approaches to both. This highlights the relevance of the possibility of estimating individualised HRTFs from a reduced set of morphological measurements — one of the longer term goals of this research.

Finally, prior work on relating human morphology to auditory cues was described in order to provide context for the novel contributions of this work, described in chapter 4, to the improvement of morphoacoustic perturbation analysis (MPA) as an important tool within this field.

Chapter 3

Head-related transfer function smoothing

> The distance between insanity and genius is measured only by success.
>
> BRUCE FEIRSTEIN

Acoustic measurement of head-related transfer functions (HRTFs) reveals them to be complicated signals containing many slopes, peaks and troughs. Previous research has shown that not all these features play a role in spatial hearing (see section 2.4.7). An improved method for removing many of the superfluous features is the subject of this chapter. To this end, a new smoothing algorithm for simplifying HRTF features whilst retaining the perceptually salient ones is presented. In the longer term, outside the scope of this thesis, the features which remain after smoothing will be studied to identify their morphological origins using the novel morphoacoustic perturbation analysis (MPA) approach presented in later chapters.

Development of the smoothing algorithm is first described and it is then evaluated using an auditory localisation model. For comparison purposes, an established HRTF smoothing method is also passed through the model. The model results

are validated by means of subjective listening tests.

3.1 Concepts and motivation for HRTF smoothing

As mentioned in section 2.4.7, it has been shown that HRTFs contain spectral variation that is imperceptible and a number of studies have investigated this (Kulkarni and Colburn, 1998; Breebaart and Kohlrausch, 2001; Senova *et al.*, 2002; Xie and Zhang, 2010; Pec *et al.*, 2013; Rasumow *et al.*, 2014; Hassager *et al.*, 2014, 2016). Since the morphological origin of HRTF features is to be investigated with a view to improving the efficiency of individualised HRTF synthesis, it makes sense to consider only those HRTF features that are perceptually relevant.

Kulkarni and Colburn's (1998) approach of discarding higher frequency Fourier coefficients of the Fourier transform of the HRTF magnitude spectrum is neat in that it also offers a method of data reduction by storing just the coefficients rather than the head-related impulse response (HRIR). However, their HRTFs were expressed on a linear frequency scale, whereas the human auditory system adheres to an approximately logarithmic one (section 2.1.3.1). Furthermore, they tested locations on the horizontal plane only, where interaural rather than spectral cues generally dominate. Senova *et al.* (2002) tested a wide selection of directions within the auditory sphere in a localization test, but their approach of truncating HRIRs to smooth the HRTFs also smooths on a linear frequency scale.

Breebaart and Kohlrausch (2001) and Hassager *et al.* (2016) used gammatone filter banks with varying orders and bandwidths respectively to smooth HRTFs, which is more ecologically valid. However, the former also only tested in the horizontal plane and used the same set of non-individualised HRTFs throughout, whilst the latter focused only on externalisation. Pec *et al.* (2013) incorporated perceptual considerations by warping their HRTFs onto the Bark scale before smoothing. However, the equivalent rectangular bandwidth (ERB) scale seems to be favoured nowadays (section 2.1.3.1) and the bilinear conformal mapping they used to warp their HRTFs cannot accurately map to the ERB scale (Smith

and Abel, 1999). The work of Rasumow *et al.* (2014) smoothed HRTFs based on constant *relative* bandwidth (fractional octave) rather than constant *absolute* bandwidth, which again approximates the frequency selectivity of the human ear better, but they also limited testing to the horizontal plane.

This chapter covers the development and verification of a new, perceptually motivated smoothing algorithm which seeks to address the limitations identified in previous studies. Based on the Fourier approach of Kulkarni and Colburn (1998), the smoothing algorithm is evaluated using an auditory model and subjective listening tests, both of which include directions outside the horizontal plane. Sensitivity of the listening tests is maximised by asking participants to identify any discernible difference between the smoothed and unsmoothed HRTFs.

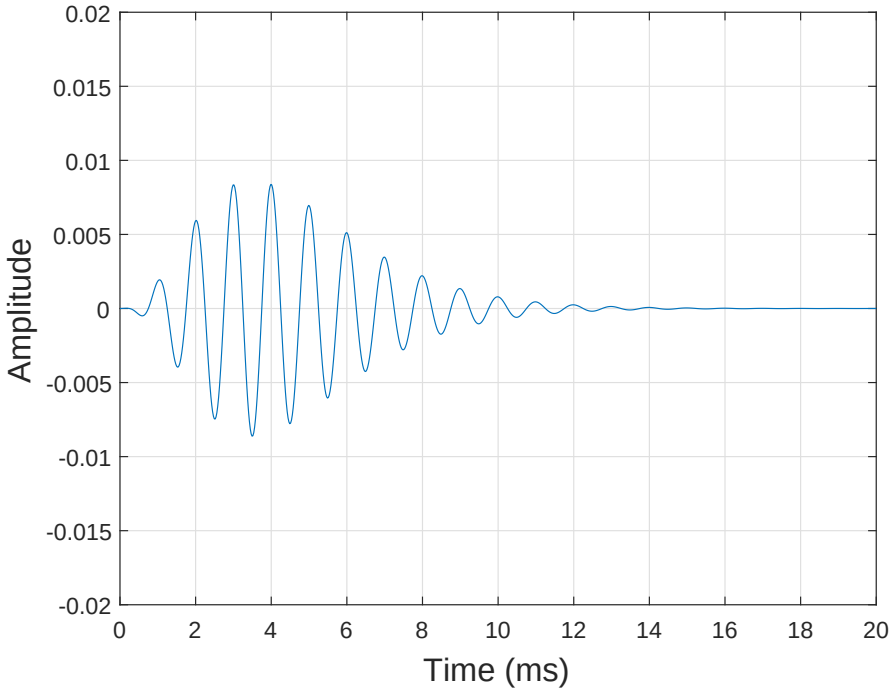
3.2 Initial ideas

For the reasons outlined above, perceptual validity was a priority when designing the new smoothing algorithm and this was promoted by basing it on the frequency selectivity of the human ear. The initial algorithm that was considered calculated the average energy in the HRTF within each of the 24 critical bands of the Bark scale (Fastl and Zwicker, 2007) and then interpolated the HRTF back to the original frequency points. However, this algorithm provided no easy way of varying the amount of smoothing applied to the HRTFs. Therefore the gammatone filter bank (Patterson *et al.*, 1987) was considered next.

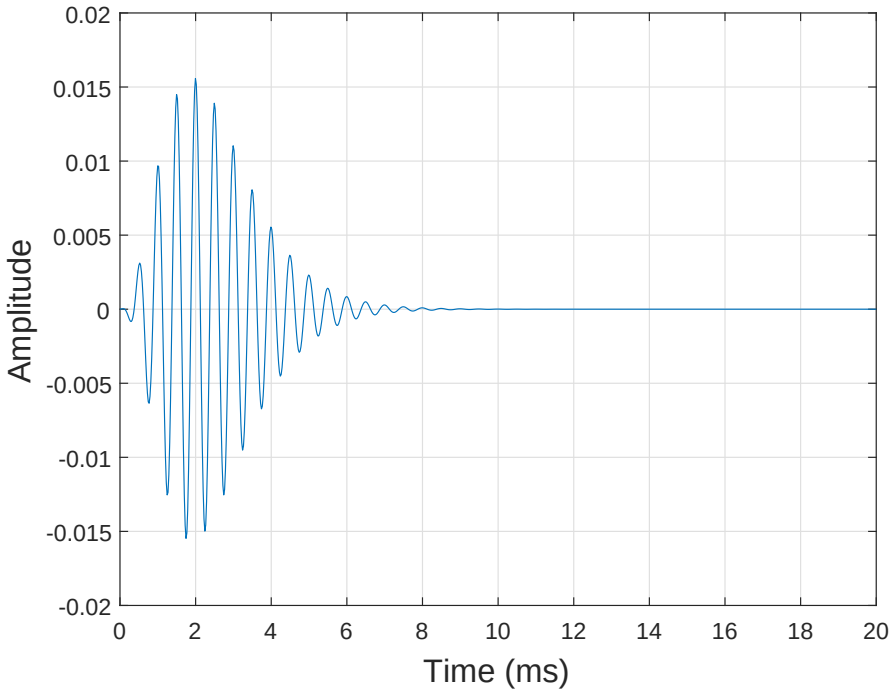
The impulse response of a gammatone filter is:

$$g(t) = at^{n-1}cos(2\pi f_c t + \varphi)e^{-2\pi bt} \tag{3.1}$$

where n is the order of the filter, a is the gain, b its bandwidth, f_c the centre frequency and φ is the phase shift. Patterson *et al.* (1992) showed that a 4^{th} order gammatone filter impulse response is a good fit to the shape of the human auditory filter derived by Patterson and Moore (1986). They recommend that b



(a) 1kHz centre frequency



(b) 2kHz centre frequency

Figure 3.1: Gammatone filter impulse responses, generated using the Auditory Toolbox (Slaney, 1998).

is set to 1.019 times the ERB as defined by Glasberg and Moore (1990):

$$ERB(f_c) = 24.7\left(\frac{4.37 f_c}{10^3} + 1\right) \tag{3.2}$$

where f_c is in Hz and $ERB(f_c)$ is in Hz. Figure 3.1 shows a comparison between the impulse responses of gammatone filters of centre frequencies 1 kHz and 2 kHz. As the centre frequency increases, the impulse response grows shorter in length and larger in amplitude.

Slaney (1993) developed an efficient implementation of the gammatone filter bank as part of his Auditory Toolbox[1] for MATLAB (Slaney, 1998). Slaney uses an adaptable version of equation 3.2:

$$ERB(f_c, EarQ, minBW, n) = \left(\left(\frac{f_c}{EarQ}\right)^n + minBW^n\right)^{\frac{1}{n}} \tag{3.3}$$

where n, $EarQ$ and $minBW$ are the filter order, filter selectivity at high frequencies and the required low frequency minimum bandwidth, respectively. The values used are those suggested by Glasberg and Moore (1990):

$$n = 1$$

$$EarQ = 9.26449$$

$$minBW = 24.7$$

The centre frequency of the nth gammatone filter in an N channel filter bank is calculated as follows:

$$f_n = -A + (f_h + A)\exp(p\left(-\log\left(f_h + A\right) + \log\left(f_l + A\right)\right)/N) \tag{3.4}$$

where $A = EarQ \times minBW$, f_l is the lowest frequency and f_h the highest frequency in the filter bank. Figure 3.2 shows a bank of 15 gammatone filters covering the 200 Hz – 22.05 kHz range, generated by the Auditory Toolbox.

[1] available from `https://engineering.purdue.edu/~malcolm/interval/1998-010/`

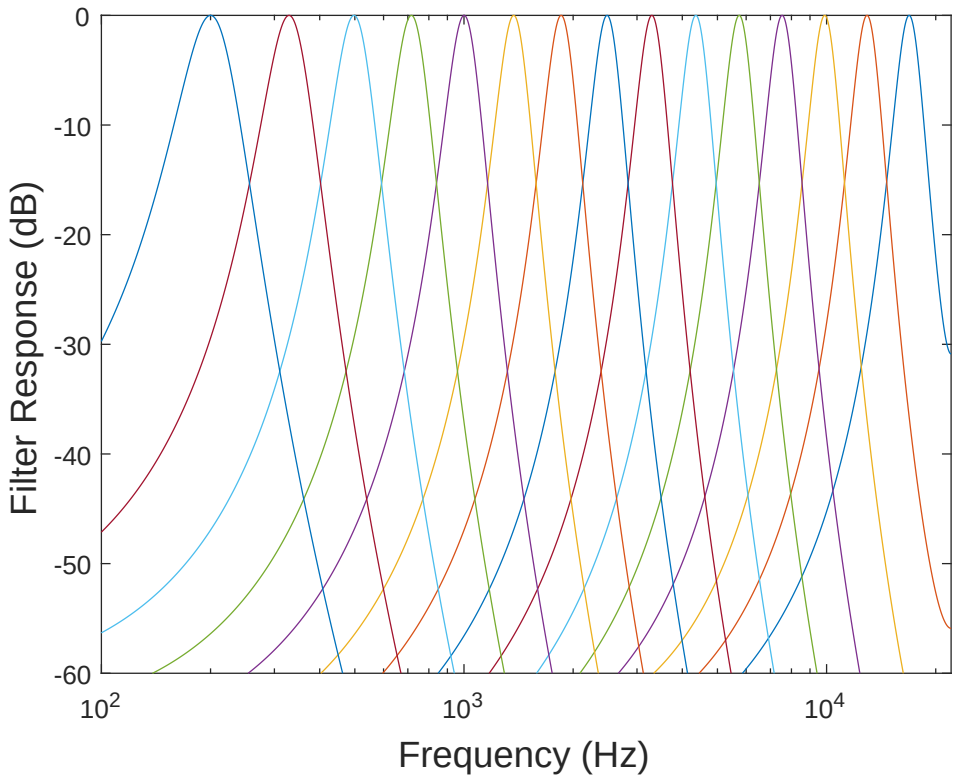


Figure 3.2: 15 channel gammatone filter bank generated by the Auditory Toolbox.

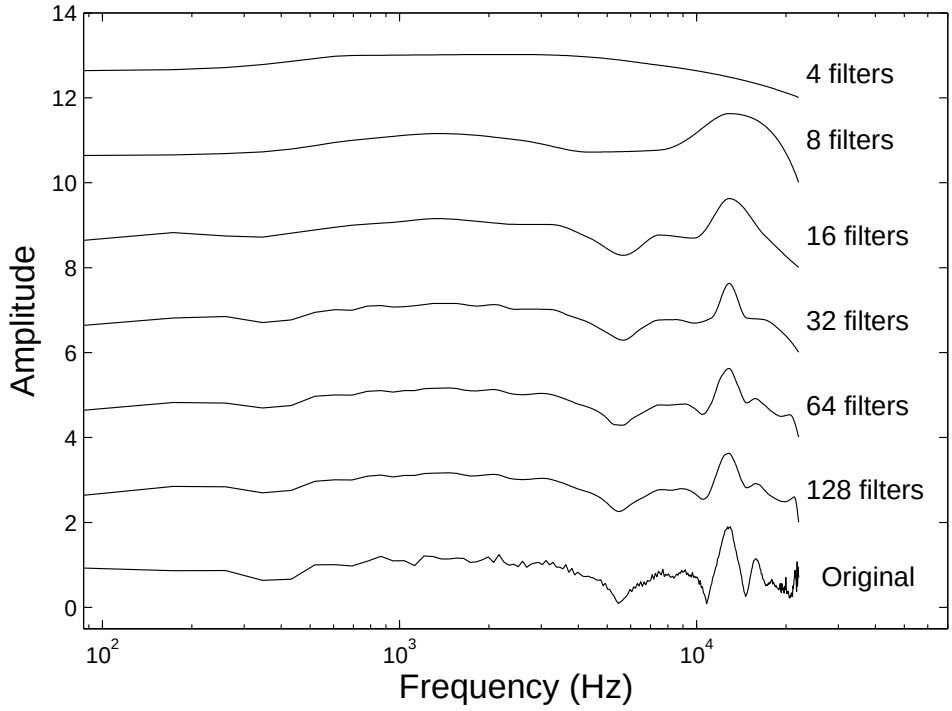


Figure 3.3: Demonstration of gammatone filter bank smoothing algorithm on exemplar HRTF.

By varying the number of channels, N, in the gammatone filter bank, the degree of HRTF smoothing can be altered (figure 3.3). The number to the right of each curve is the number of channels in the gammatone filter bank.

The gammatone filter bank smoothing algorithm consists of the following steps and uses the Auditory Toolbox gammatone filter bank implementation:

1. Calculate the centre frequencies of N gammatone filters using equation 3.4.
2. Calculate the coefficients for the gammatone filter bank.
3. Pass the HRIRs through the gammatone filter bank.
4. Calculate the root mean square (RMS) energy in the output of each gammatone filter.
5. Scale according to the maximum RMS energy that can pass through each filter, i.e. the area under each of the curves in figure 3.2, to give the HRTF magnitude at each of the N gammatone filter centre frequencies. This eliminates level discrepancy due to the varying width of the gammatone filters.
6. Interpolate the smoothed HRTF to calculate its magnitude spectrum at the N frequency points used in the original HRTF.
7. Combine the modified magnitude spectrum with the original HRTF phase spectrum.

Following this procedure revealed problems with the gammatone filter bank smoothing approach. The fixed minimum bandwidths of the gammatone filters asymptotically limits the degree of spectral resolution achievable, such that increasing the number of channels never increases spectral resolution as far as the spectral resolution of the unsmoothed HRTF. Hence, passing HRTFs through the smoothing algorithm will always result in some smoothing. The algorithm also results in attenuation of peaks and notches due to the fact that the gammatone filter bank acts like a variable-window-size moving average filter. Both these effects can be

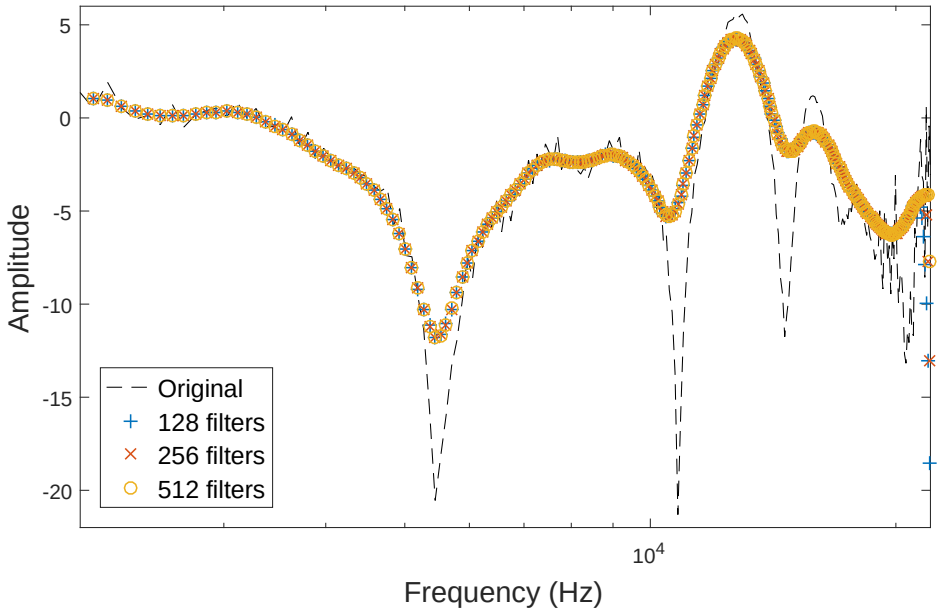


Figure 3.4: Demonstration of the gammatone filter bank smoothing algorithm, showing that its maximum spectral resolution falls short of the original resolution.

seen in figure 3.4. There is very little difference between the HRTFs smoothed with 128, 256 and 512 channel gammatone filter banks — the small deviations are to do with the variation in centre frequencies rather than with variations in spectral resolution.

It was found that even at the full spectral resolution of the gammatone filter bank, a difference in timbre of HRTF filtered broadband white noise could be heard compared to full resolution HRTFs. To provide the required continuum in smoothing, from full resolution HRTF reconstruction downwards, another algorithm was developed and this is covered in the following section.

3.3 Smoothing algorithm

The novel smoothing algorithm finally adopted is a development of Kulkarni and Colburn's (1998) cepstral-based approach in which the HRTF magnitude spectrum is described using Fourier coefficients. However, before calculating the Fourier coefficients, the HRTF is interpolated and the magnitude spectrum resampled on a frequency scale based on the equivalent rectangular bandwidth (ERB)

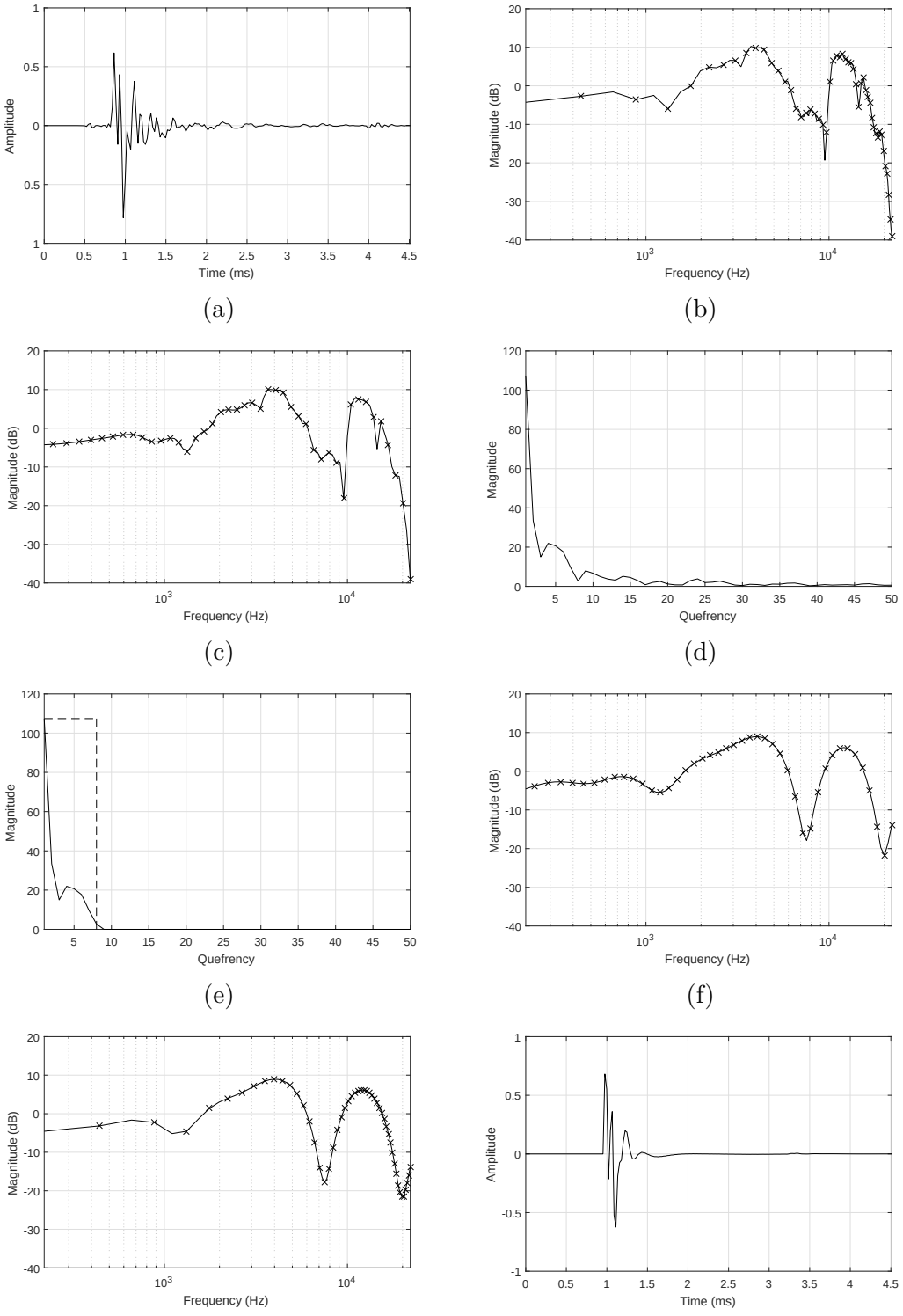


Figure 3.5: Visualisation of the steps of the new HRTF smoothing algorithm. See text for description.

(Glasberg and Moore, 1990). This maximises the amount of perceptually relevant spectral information retained for any given number of Fourier coefficients. The steps in this algorithm are as follows (visualisation is given in figure 3.5):

1. Perform an M-point fast Fourier transform (FFT) on the original HRIR (figure 3.5a) to create an HRTF with Hermitian symmetry.

2. To reduce computation, retain only the distinct first $M/2 + 1$ points, i.e. up to $f_s/2$, of the log-magnitude HRTF, sampled at linear frequency points f_{lin} — figure 3.5b.

3. Calculate $M/2+1$ auditory frequency points, f_{aud}, on the ERB scale up to $f_{\text{s}}/2$.

4. Resample the original spectrum at f_{aud}, using shape-preserving piecewise cubic interpolation — figure 3.5c.

5. Calculate the FFT of the resampled log-magnitude spectrum — figure 3.5d. This represents the cepstrum.

6. Apply the required degree of smoothing by retaining only the N lower cepstral coefficient values (for frequency points $k = 1 \ldots N$). Set any remaining coefficients ($k = N + 1 \ldots M/2 + 1$) to zero (figure 3.5e). This is liftering. Reflect the N lower-order cepstral coefficients about $f_{\text{s}}/2$ to create a symmetrical cepstrum.

7. Compute the inverse FFT to generate the smoothed HRTF spectrum, sampled at the auditory frequency points — figure 3.5f.

8. Resample the modified log-magnitude spectrum at the original linear frequency points, f_{lin}, using shape-preserving piecewise cubic interpolation — figure 3.5g.

9. Reflect the magnitude spectrum about $f_{\text{s}}/2$ to produce a magnitude spectrum with even symmetry.

10. Perform real cepstrum minimum-phase reconstruction (Pei and Lin, 2006)

to compute the minimum-phase HRTF. Combine this with a pure delay equal to the ITD (Kistler and Wightman, 1992), calculated using the cross correlation technique (section 2.2.3) — figure 3.5h.

The ERB-based frequency points f_{aud} were calculated using equation 3.4 with the following values:

$$\begin{aligned} N &= M/2 \\ f_{\text{l}} &= 0 \\ f_{\text{h}} &= f_{\text{s}}/2 \end{aligned} \tag{3.5}$$

and the values suggested by Glasberg and Moore for $EarQ$ and $minBW$ have been used.

Figure 3.6 shows the smoothing effect of Kulkarni and Colburn's algorithm on an exemplar HRTF compared to the algorithm used in this study; the number to

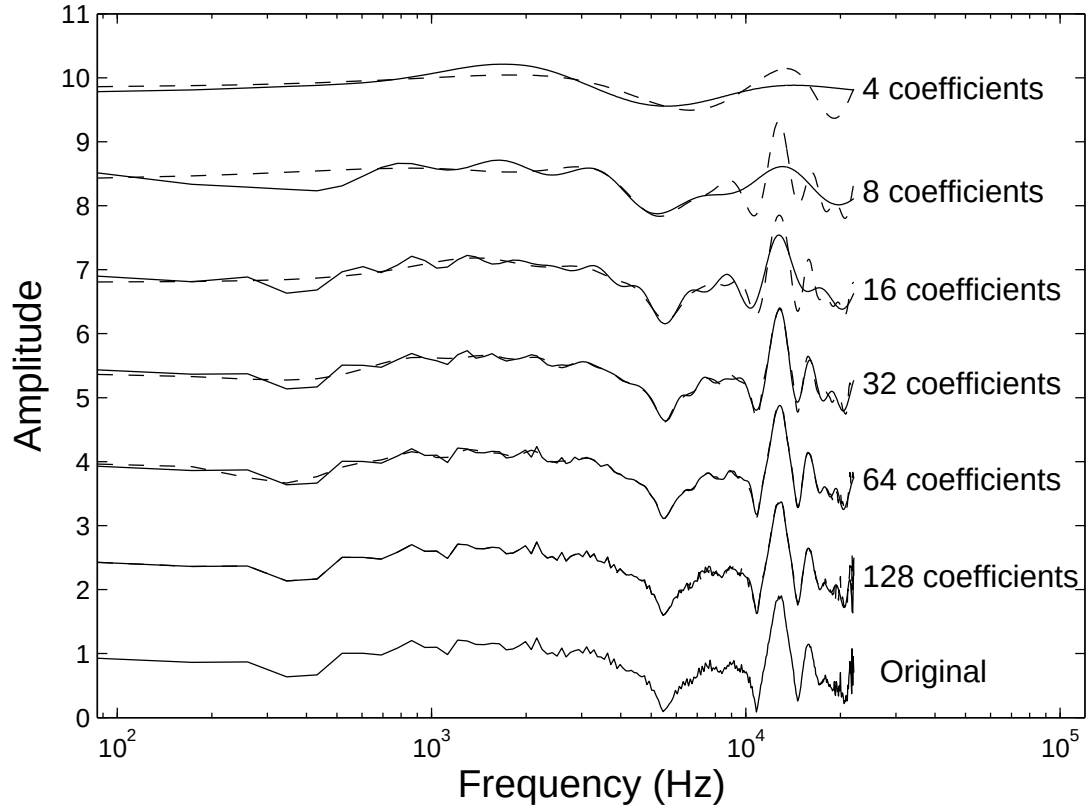


Figure 3.6: Comparison of FFT smoothing algorithms. Dashed line - Kulkarni and Colburn's approach. Solid line - proposed approach.

the right of each pair of plots is the number, M, of coefficients retained. It can be seen that, because the new algorithm smooths on an ERB-based frequency scale, high frequency features tend to be smoothed more and low frequency features less than occurs using Kulkarni and Colburn's smoothing with the same number of coefficients.

3.4 Auditory model simulations

3.4.1 The localisation model

Sagittal planes (SPs) are planes parallel to the median plane, where spectral cues tend to dominate localisation. Therefore, the sagittal plane (SP) localisation model developed by Baumgartner *et al.* (2013) is well suited to investigating the effect of spectral smoothing. Figure 3.7 outlines the model. Briefly, directional transfer functions (DTFs) are applied to the audio inputs in a manner that emulates incoming sounds to the peripheral hearing system. Now impressed with acoustic directional cues, the audio signals are passed through a gammatone filterbank (GT) and a simple inner-ear hair cell model (IHC) comprising half-wave rectification and a low pass filter. The resulting data are compared with a similarly processed internal "template" created using a reference set of DTFs during an initial learning phase for the model. The spectral distance of the audio processed using the DTF left/right pair under test from the same audio processed

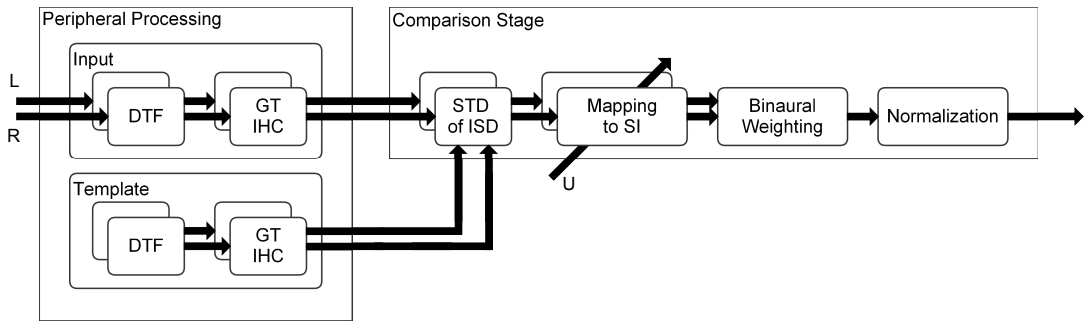


Figure 3.7: Sagittal plane (SP) localisation model developed by Baumgartner *et al.* (2013). After Baumgartner *et al.* (2013).

by all the template DTFs in turn is computed to estimate the perceptual similarity in each case. The final output from the model is a distance metric which indicates the probability that a human listener would select this polar angle (i.e. elevation angle within an SP) for the direction of the sound. To assess the effect of HRTF smoothing on localisation, the input DTFs can be altered, leaving the template DTFs unchanged and the distance metric recomputed in each case. The comparison stage consists of five steps:

1. For each frequency band output of the peripheral processing stage, the inter-spectral differences (ISDs) between the input sound and every template angle are calculated.

2. The spectral standard deviations (STDs) of the inter-spectral differences are found.

3. The STDs of the ISDs form the argument to a Gaussian function with a mean of zero and a standard deviation dependent on a listener-specific localisation uncertainty, U, to produce similarity indices (SIs).

4. Left and right SIs are combined by weighting ipsilateral and contralateral SIs according to the lateral angle of the SP being considered (e.g. 50/50 for the median plane), forming a binaural SI.

5. The binaural SI sum is normalised to unity.

The comparison stage results in a discrete probability density function or probability mass vector (PMV) for each target polar angle. These PMVs can be combined to form a matrix for each SP, as shown in figure 3.8. For each column (target polar angle) the estimated probability of a listener with these DTFs selecting a particular response angle is proportional to the image's darkness.

From the PMVs, the polar localisation error and quadrant error can be calculated to facilitate analysis of changes in a listener's performance. The quadrant error is calculated as the sum of the PMV entries for which the response-target difference is greater than ninety degrees. The polar error is calculated as the discrete

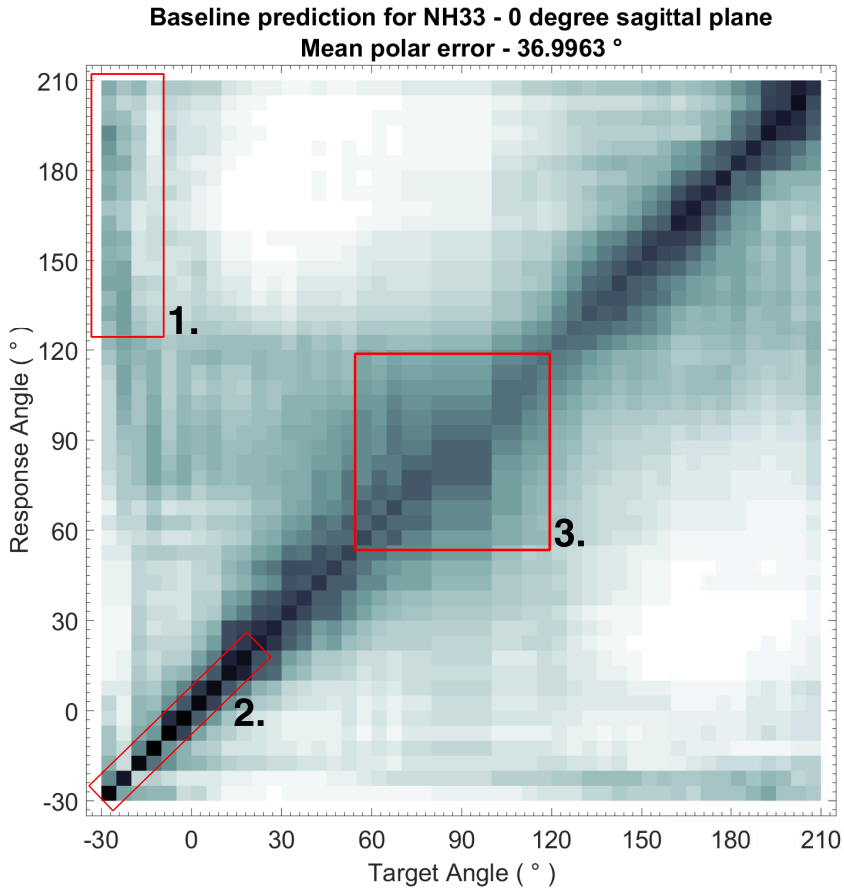


Figure 3.8: Exemplar output of SP localisation model. Subject NH33 (from the Acoustics Research Institute (2011) database) using their own DTFs in the 0° (median) SP. Showing areas of quadrant error (1), low polar error (2) and high polar error (3).

expectancy within the local polar range (less than ninety degrees). In the matrix representation in figure 3.8, quadrant errors manifest as dark areas in the top left and bottom right quadrants (1), low polar error manifests as concentrated areas of darkness around the diagonal (2) and high polar error manifests as large areas of similar brightness (3). It can be seen that, as expected, the darkness along the leading diagonal spreads, indicating a reduction in localisation acuity, as the target polar angle increases from the horizontal (0° polar angle) to above the listener (90° polar angle).

Figure 3.9 shows how the PMVs change when another listener's DTFs are used as the input DTFs. It can be seen that there is a far greater spread in the darkness along the diagonal, signifying an increase in polar error. Furthermore there is a larger spread of darkness in the top left and bottom right quadrants signifying

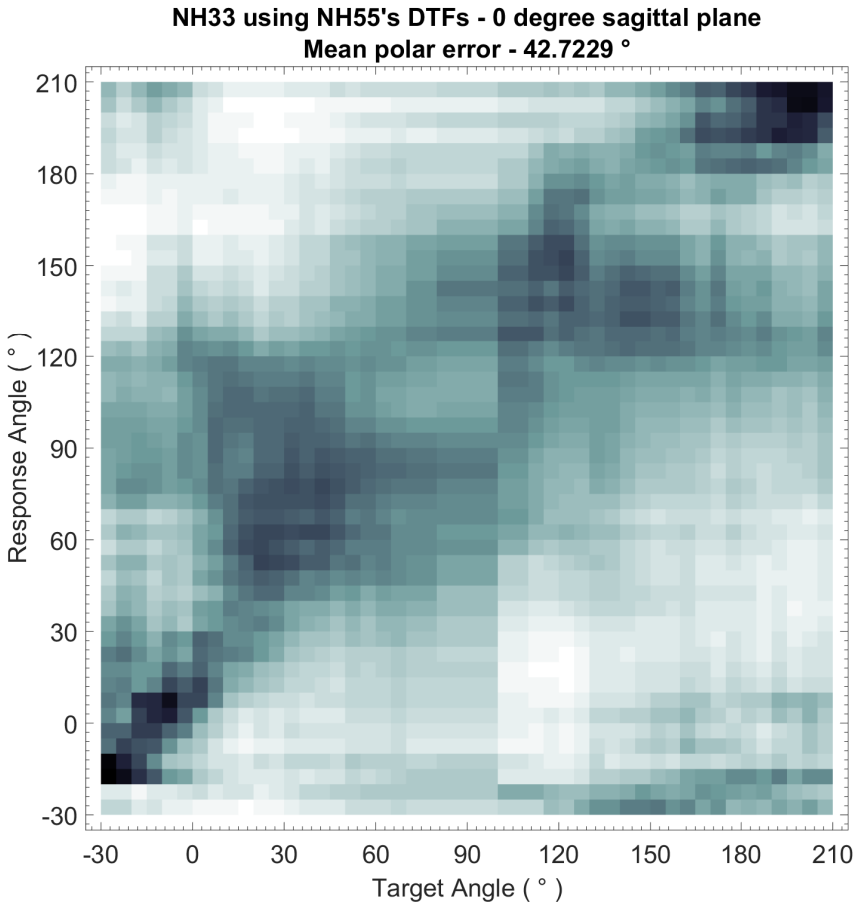


Figure 3.9: Exemplar output of SP localization model using non-matching sets of DTFs (subject NH33 (from the Acoustics Research Institute (2011) database) template DTFs used in conjunction with the input DTFs for subject NH55 in the 0° (median) SP).

an increase in quadrant error, which is to be expected using non-individualised HRTFs.

3.4.2 Simulations

The Auditory Modelling Toolbox (AMT) (Søndergaard and Majdak, 2013) is a collection of MATLAB scripts for auditory research. It includes the SP localisation model by Baumgartner *et al.* (2013), described in the previous section, and it is this version which was used for the work described in this chapter. All 17 DTFs provided in the AMT were used. These DTFs originated from the Acoustic Research Institute (ARI) HRTF database (Acoustics Research Institute, 2011) and further details about which subjects' DTFs are in the AMT can be found in

Baumgartner *et al.* (2013).

The `baumgartner2013` function in the AMT takes as two of its inputs the "input" and "template" DTFs. The input DTFs were progressively smoothed using the algorithm outlined in section 3.3 and the polar error compared to the values obtained when the unsmoothed DTFs were used. This comparison was also conducted using Kulkarni and Colburn's algorithm to smooth the DTFs, so that the extent of the anticipated improvement of our smoothing algorithm could be assessed.

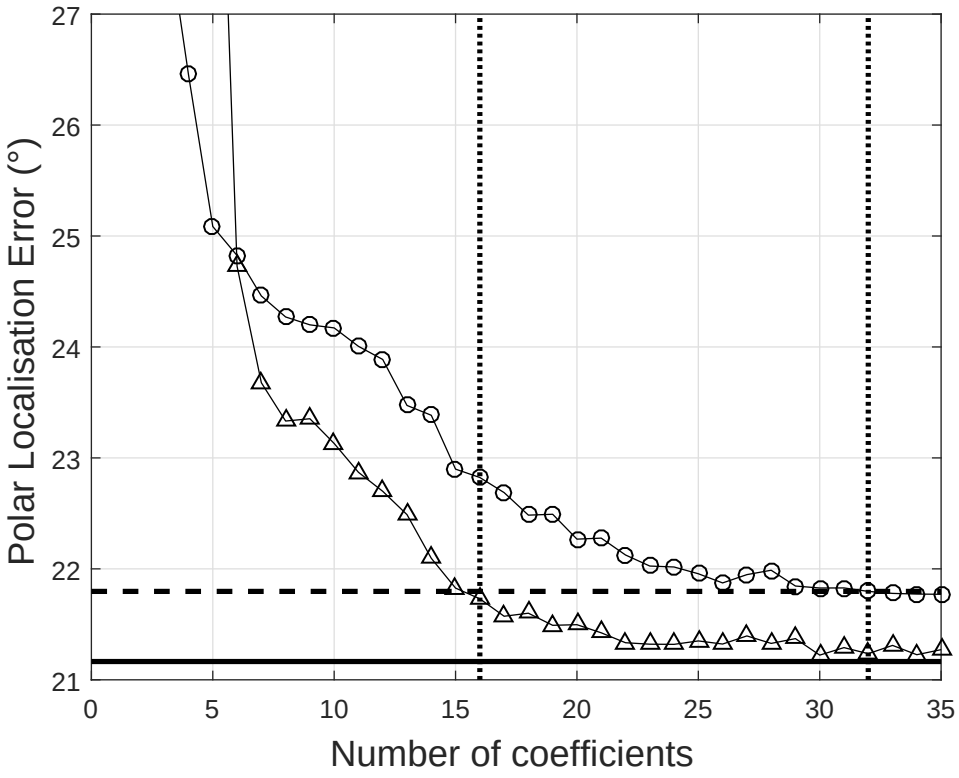


Figure 3.10: Polar localisation error against number of coefficients retained for proposed algorithm (triangles) compared to Kulkarni and Colburn's algorithm (circles). The RMS errors have been averaged across DTF sets for a target angle of 0° in the median plane. The horizontal dashed line crosses the Kulkarni and Colburn (1998) curve at the threshold of perceptual difference (32 coefficients), indicated by the right-hand vertical dotted line. The left-hand vertical dotted line shows that 16 coefficients are required to achieve the same performance using the proposed algorithm. The lower horizontal solid line shows the baseline performance with no smoothing.

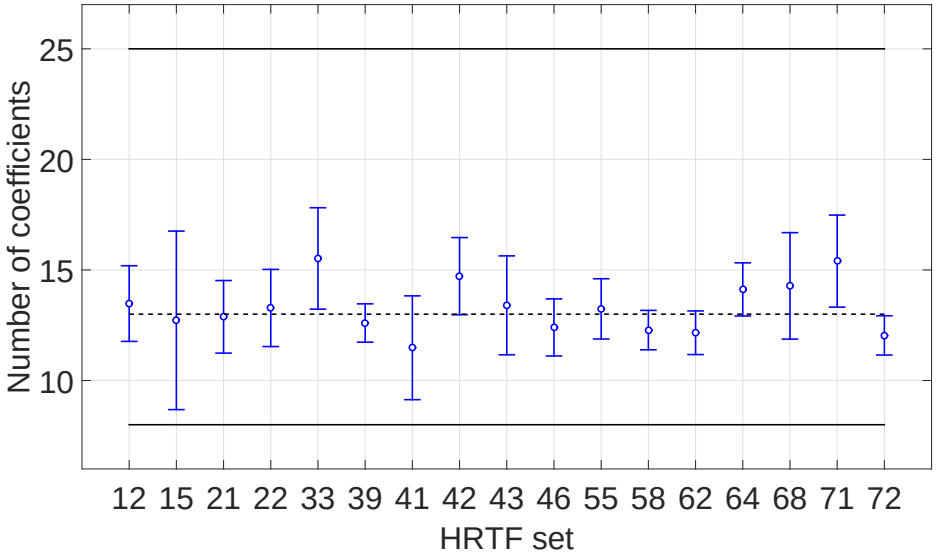


Figure 3.11: Number of coefficients needed for equal performance with Kulkarni and Colburn's algorithm using 32 coefficients. The mean across SPs (circles) and the ± one standard deviation limits (whiskers) are shown for each DTF set. The dashed line is the mean across DTF sets. The solid lines show the minimum and maximum values.

3.4.3 Results

Figure 3.10 shows the mean polar error across all 17 DTF sets for a target angle of 0° (straight-ahead) in the median plane for both Kulkarni and Colburn's algorithm and the new algorithm. Kulkarni and Colburn (1998) reported that in their localisation task, whilst performance was seldom affected with fewer than 32 coefficients, three out of the four participants did exhibit a reduction in performance for 16 coefficients in the frontal direction. Therefore, the localisation performance results obtained from the model using Kulkarni and Colburn's algorithm with 32 coefficients was used as a threshold to assess the new algorithm.

Figure 3.11 shows the number of coefficients needed, for each DTF set, for the new algorithm to equal the performance of Kulkarni and Colburn's algorithm with 32 coefficients. The means and one standard deviation limits across all SPs are presented.

As can be seen, the mean number of coefficients, across all subjects and all SPs, required for the new smoothing algorithm is 13, more than a twofold saving on the

number of coefficients required by Kulkarni and Colburn's smoothing algorithm for the same localisation performance.

Figure 3.12 shows how the number of coefficients needed varies across SPs. The means and one standard deviation limits across all subjects are presented. Little variation is evident across SPs and an analysis of variance (ANOVA) showed no significant effect of SP on performance ($F = 0.9, p = 0.60$). The simulation results indicate that on average across SPs and DTF sets fewer than half the number of coefficients are required to meet the performance threshold set by Kulkarni and Colburn's algorithm.

It is interesting to note that in figure 3.10 there is a region below 6 coefficients where Kulkarni and Colburn's algorithm outperforms the new algorithm. It is likely that this behaviour is a result of the relatively sudden loss of sufficient spectral accuracy across all frequencies in the new perceptually motivated algorithm as the number of coefficients is reduced. In Kulkarni and Colburn's algorithm the loss of spectral accuracy with fewer and fewer coefficients occurs more gradually, with spectral resolution becoming insufficient at lower frequencies sooner

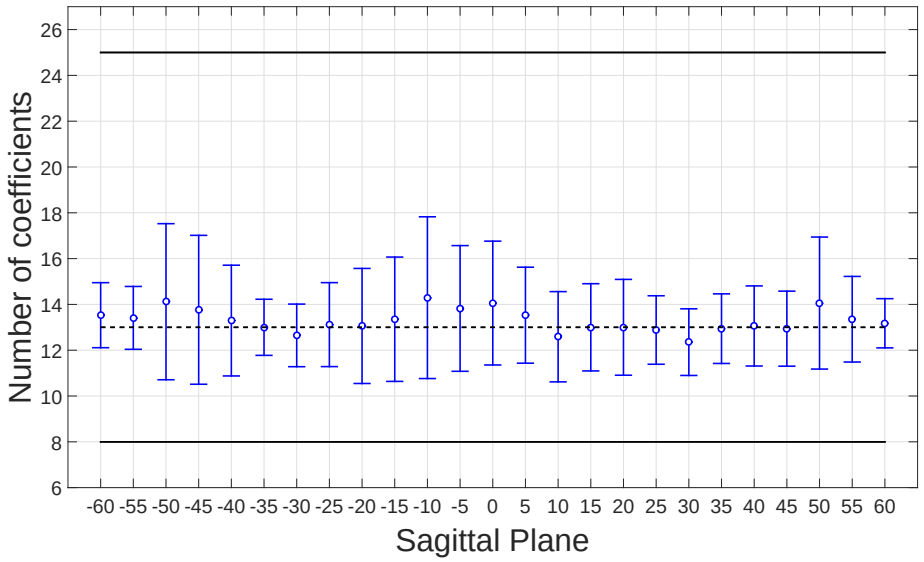


Figure 3.12: Number of coefficients needed for equal performance with Kulkarni and Colburn's algorithm using 32 coefficients. The mean across subjects (circles) and the ± one standard deviation limits (whiskers) are shown for each SP. The dashed line is the mean across subjects. The solid lines show the minimum and maximum values.

than at higher frequencies. Also of note is the relatively large baseline (∼21–22°) polar error. Whilst this seems very large, it is an average across the whole auditory space and Baumgartner *et al.* have shown their model on average tends to actually underestimate polar error — see figure 7 in Baumgartner *et al.* (2013).

It should also be mentioned that the model estimates localisation error, whereas, with the exception of Senova *et al.* (2002), previous studies (Kulkarni and Colburn, 1998; Breebaart and Kohlrausch, 2001; Hassager *et al.*, 2014) have assessed discriminability. In terms of finding a perceptual limit for smoothing, discrimination would appear to be both a more sensitive test and more easily perceived than asking a participant to isolate and assess sound localisation. The next section (section 3.5) details subjective discrimination listening tests carried out to corroborate the model simulation findings.

3.5 Listening tests

3.5.1 HRTF measurements

Individualised HRTFs, rather than non-individualised HRTFs drawn from a database, were used for the listening test due to their more reliable performance (section 2.3.4). Each participant's HRTFs were measured in the anechoic chamber at the University of York using an RME Fireface UFX audio interface. Four concentric driver KEF speakers (two E301s and two HTS3001s) were mounted in an arc on a custom made stand at elevations of +20°, +10°, 0° and -10°. The speakers were driven from the interface by Behringer A-500 power amplifiers. The participant was seated on a chair on a custom-made turntable at a distance of 1.5 m from the centre of the speaker arc. The turntable allowed azimuthal increments of 5° from -90° (left) to +90° (right). The chair had a headrest to reduce head movements and a pair of lasers, one mounted on the headrest and the other on a cap they wore. These were used to ensure the participant kept their head facing in the correct direction as the turntable was rotated. Black fabric was hung around the

walls of the anechoic chamber at head height to facilitate lining up of the lasers.

Sennheiser microphone capsules (KE 4-211-2 [2]) were mounted in drilled out foam earplugs and logarithmically swept sine signals of duration 5 seconds, and sweeping upwards in frequency from 20 Hz to 20 kHz, were used to record the participant's HRIRs at a sample rate of 44.1 kHz. Headphone transfer functions for each participant were measured while they wore Beyerdynamic DT990 headphones to facilitate later removal of the microphone and headphone responses. These headphones were chosen for the listening tests due to their performance in a study of free air equivalent coupling (FEC) (see section 2.4.1) headphones (Møller *et al.*, 1995a). The measured HRIRs were processed and equalised as described below using MATLAB to produce the 256–point HRIRs used in the listening test.

The measured HRTFs included the actual HRTFs, the response of the microphones and the response of the loudspeakers:

$$HRTF_{measured} = HRTF_{actual} \times M \times L \tag{3.6}$$

To obtain the participant's HRTFs alone, i.e. $HRTF_{actual}$, the response of the microphones and loudspeakers needed to be removed. The response of the loudspeakers L was measured using a reference microphone with a nominally flat frequency response (Earthworks M30) placed at the position of the centre of the head, with the participant absent.

Due to the intention to synthesise sounds over headphones, the response of the headphones had to be considered. The headphone transfer functions (HpTFs) were measured using the same microphones used for measuring the HRTFs, giving the expression for the recorded transfer function, $HpTF_{measured}$:

$$HpTF_{measured} = HpTF_{actual} \times M \tag{3.7}$$

This was combined with the measured loudspeaker responses L and inverted

[2] `http://datasheet.octopart.com/KE-4-211-2-Sennheiser-datasheet-13035495.pdf`

to produce an equalisation filter that removed the effects of the microphones, loudspeakers and headphones:

$$EQ = \frac{1}{HpTF_{measured} \times L} \tag{3.8}$$

Multiple headphone measurements were made to allow generation of the headphone equalisation filters in the manner outlined in Masiero and Fels (2011). Firstly the magnitude of the filter was calculated as:

$$|HpTF| = \mu + 2\sigma \tag{3.9}$$

where μ is the mean of all the measured headphone transfer function magnitudes and σ is the standard deviation. This magnitude response was smoothed with a 1/6-octave moving average filter and the phase response was calculated as the minimum phase spectrum of the magnitude spectrum.

Frequency dependent regularisation (Kirkeby *et al.*, 1999) was applied to limit the gain of the equalisation filters outside the operating frequency range and also in the region of inverted notches. This was achieved by introducing a frequency-dependent regularisation term into the equalisation filter:

$$EQ[f] = \frac{1}{(HpTF_{measured}[f] \times L[f]) + \beta B[f]} \tag{3.10}$$

where β is a regularisation constant and B is a frequency-dependent shape function set to 1 inside the passband and $1/\beta$ outside. Inside the passband the regularisation term ($\beta B[f]$) is small and has little effect, whereas in the stopband (where $HpTF_{measured} \times L$ is very small) the regularisation term limits the gain of the equalisation filter to 1. In this case, β was set to 0.0001 and B was set to $1/\beta$ below 100 Hz and above 20 kHz. B was shaped using segments of a raised cosine curve up to 1 between 100 Hz and 150 Hz and from 20 kHz down to 18 kHz. The maximum gain of of the regularised equalisation filters was set to 12 dB.

3.5.2 Changes to the smoothing algorithm

When it came to smoothing the measured HRTFs, a problem was found with how the smoothing algorithm interacted with the steep high frequency roll-off of the measurement system. Figure 3.13 shows an example of this. The dashed line shows the original, unsmoothed, HRTF and the solid line shows the HRTF with 16 coefficients retained during smoothing. It can be seen that just before the roll off around 18 kHz the smoothed HRTF is quite distorted, whereas the roll off itself is quite accurately reproduced. This is undesirable because, not only is the roll-off unlikely to be a perceptual cue, it is also above the perceptually significant upper limit of human hearing (see section 2.1.2). The reason that this distortion was not apparent during the auditory model simulations of section 3.4 is that the Baumgartner *et al.* (2013) auditory model uses DTFs from the ARI database. DTFs do not contain the system response because the average of all

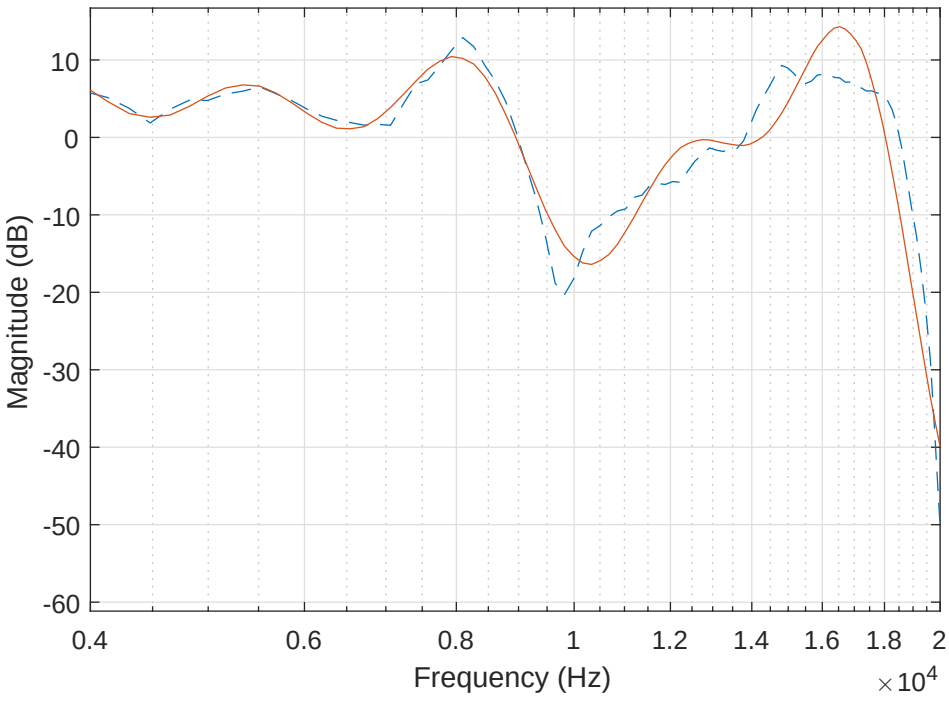


Figure 3.13: Distortion in smoothed HRTF due to the steep roll off in the measurement system. The dashed line is the original HRTF, the solid line is the smoothed HRTF. The peak around 16–17 kHz is quite distorted, whilst the steep roll off around 18 kHz is quite accurately reproduced.

the measurements, the common transfer function (CTF) (section 2.2.4), has been subtracted from them. The absence of the CTF removes not only the common response component in the HRTFs but also the response of the measurement system. Figure 3.14 shows the CTF for all the measured HRTFs. The roll-off of the measurement system above approximately 18 kHz can be clearly seen.

Therefore, prior to smoothing, the measured HRTFs' magnitude spectra were windowed (figure 3.15) to ensure that spectra coefficients are concentrated on representing the relevant HRTF features and not the system response. The windowing was carried out in the log-magnitude domain. A raised cosine window was generated from 16 kHz up to 18 kHz and, since a roll-off discontinuity exists also at low frequencies, another window was generated from 300 Hz down to 200 Hz. Both windows were applied to the log-magnitude HRTF spectra to window them up/down to 0 dB outside of the pass-band and therefore discard the system roll-offs at low and high frequencies. This means that the HRTF magnitude spectra coefficients are concentrated on representing the relevant HRTF features and not the system response.

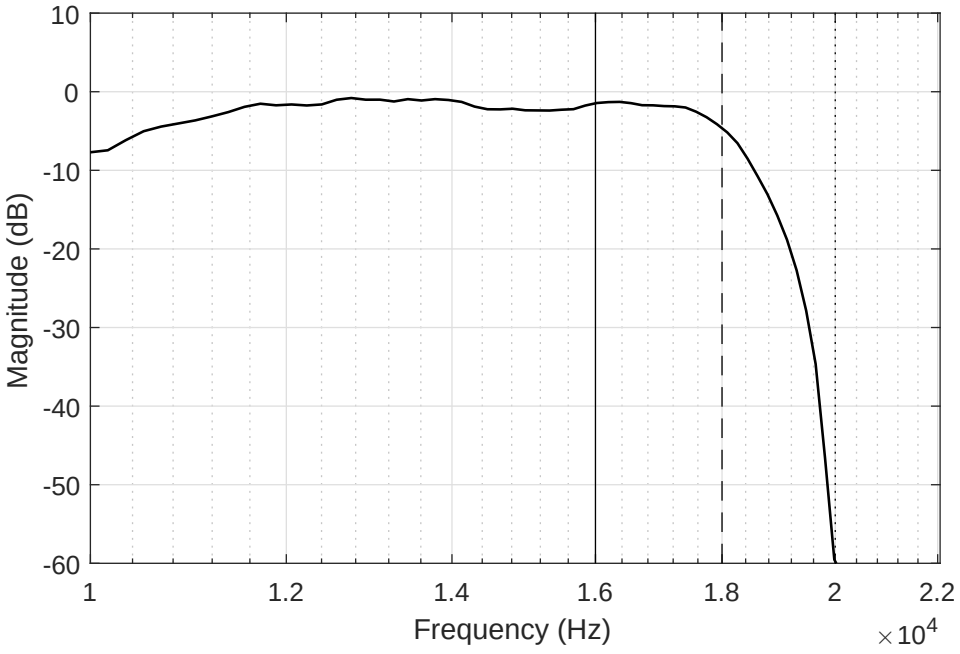


Figure 3.14: Measurement system roll-off at high frequencies in log-magnitude CTF of measured HRTFs. Vertical lines are at: 16 kHz (solid), 18 kHz (dashed) and 20 kHz (dotted).

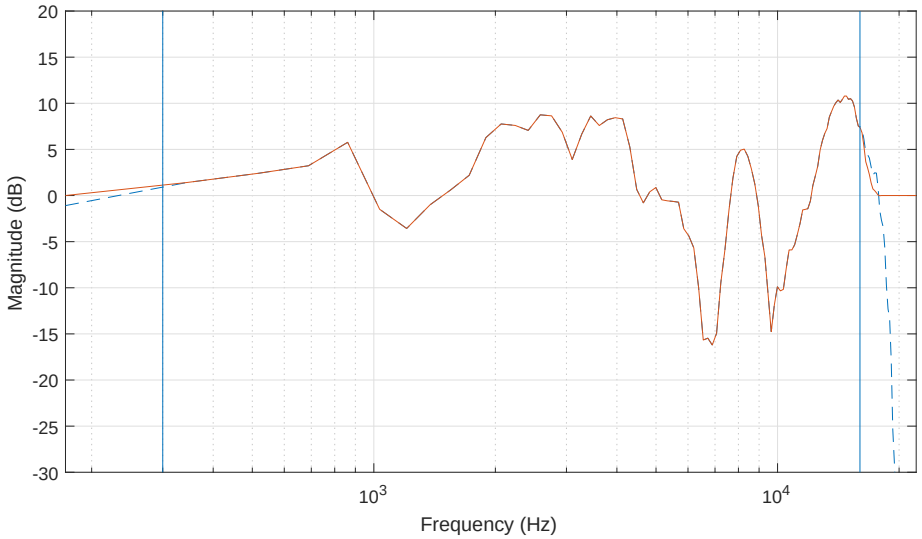


Figure 3.15: Example of windowing applied to HRTFs for the listening tests. Windowing to 0 dB in the ranges 300 Hz down to 200 Hz and 16 kHz up to 18 kHz (continuous line) reduces spectral magnitude artefacts due to low frequency and high frequency roll-offs (dashed line).

Figure 3.16 compares the mean squared error between smoothed and unsmoothed HRTFs, using the new algorithm, between 300 Hz and 16 kHz, i.e. the unwindowed section, for all measured HRTFs with and without the windowing described above. When retaining 30 coefficients or more there is little difference between windowing and not windowing the HRTFs prior to smoothing. However, between 10 and 20 coefficients, i.e. in the range of the perceptual limit for the new algorithm suggested by the auditory model, there is a substantial increase in accuracy afforded by windowing (over 10 dB at some frequencies).

The result confirm that the windowing described above is beneficial. Windowing was applied prior to smoothing and was implemented for both the new algorithm and Kulkarni and Colburn's algorithm. Treating both algorithms similarly in the final listening test ensured that the perceptual comparison was conducted purely on the basis of whether the HRTF was smoothed on a linear or ERB frequency scale.

Another change that was made for the listening test was that, rather than using minimum-phase HRTFs, linear phase HRTFs were used. This is because they have been shown to perform as well as original phase HRTFs and have been shown to

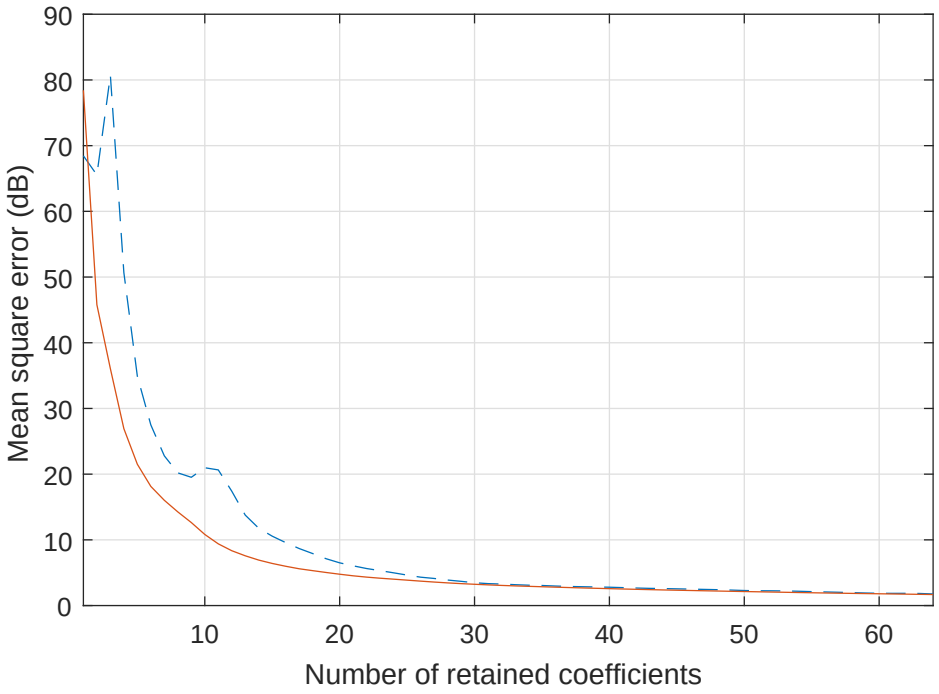


Figure 3.16: Mean squared error against number of retained coefficients for windowed (solid line) and non-windowed (dashed line) smoothed HRTFs.

outperform minimum phase HRTFs (Kulkarni *et al.*, 1999). To generate linear phase HRTFs the HRTF magnitude spectrum is combined with a zero-phase spectrum and the inverse Fourier transform is taken. The resulting HRIR is then circularly shifted by half the number of samples to create a causal filter that is symmetric about the centre sample (figure 3.17) and the contralateral HRIR in a pair is then delayed by the frequency-independent interaural time difference (ITD) calculated using the cross-correlation technique (see section 2.2.3). In the listening tests both the unsmoothed and smoothed HRTFs were converted to linear phase so that the only difference between them was their magnitude response.

3.5.3 Test procedure

The aim of psychophysical procedures is to uncover the relationship between physical stimuli and psychological perception. In psychoacoustics this is the relationship between sound events, the acoustic stimuli that enter the listener's

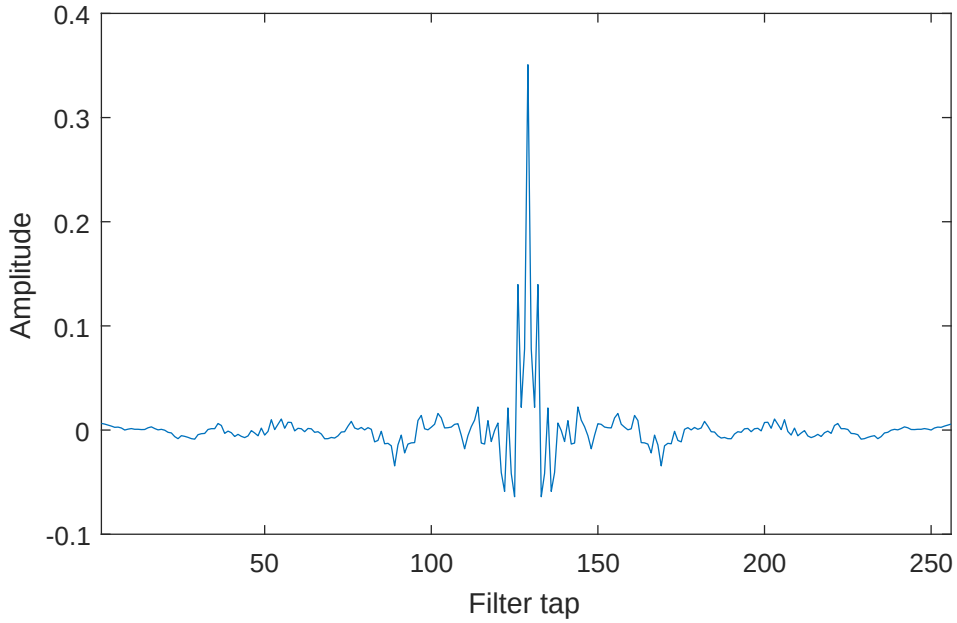


Figure 3.17: Linear phase HRIR. Linear phase HRIRs are causal filters that are symmetric about the centre filter tap.

auditory system, and auditory events that they perceive (Blauert, 1970). One family of psychoacoustic procedures are those that measure the thresholds of hearing and these procedures can be split broadly into two types of task. The first are *detection tasks* that aim to find *absolute* thresholds of hearing; for example, the required level of a sound event to elicit an auditory event. The second are *discrimination tasks* that aim to find the *difference* thresholds of hearing; for example, the smallest change in a sound event that elicits a change in the perceived auditory event (Pulkki and Karjalainen, 2015). This study is concerned with the latter. In either case, psychoacoustic threshold tasks are designed to generate a mapping from a continuous scale representing the probability of responding correctly at a given stimulus level, to a binary yes/no scale. In discrimination tasks a "correct" response is generally the ability to pick which stimulus in a set is the odd one out and commonly the aim is to find the stimulus level at which the listener can reliably answer correctly. The mapping function in threshold tasks is known as the psychometric function and approximates to a sigmoid function, although in reality it is asymptotic to 0% and 100% due to human error (figure 3.18). The aim of these listening tests is to find the level of HRTF smoothing

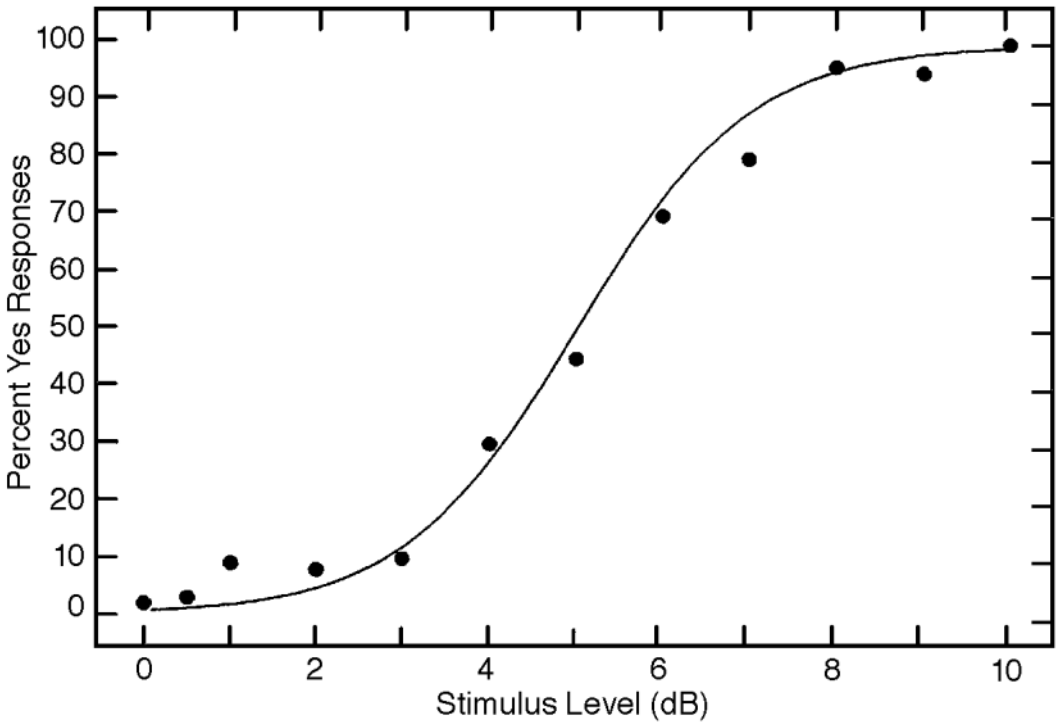


Figure 3.18: Example psychometric function relating stimulus level to percentage of correct responses. The dots show individual data points to which the psychometric function (line) has been fitted. Figure from Leek (2001)

at which a listener can discriminate between a sound filtered with unsmoothed HRTFs and a sound filtered with smoothed HRTFs.

There are a number of aspects to psychophysical procedures, such as a listening test, and in the following section the taxonomy of Marvit *et al.* (2003) is used. A listening test *procedure* consists of a *paradigm* and a *method*. The paradigm consists of two parts: firstly the *mode of stimulus presentation* which defines how many *intervals* (stimuli) are presented at each step of the procedure (*trial*) and what their content is, and secondly the *task* which defines what the listener must do when the stimuli are presented. The method also has two parts: the *measurement strategy* that defines the rules with which stimuli are chosen from trial to trial and the *datum definition* which defines how the results are derived from the stimuli and listener responses. Additionally, the measurement strategy can be broken down into three rules: (1) the *starting rule* which defines the stimulus used in the starting trial, (2) the *progression rule* which governs what stimulus is presented next based on previous stimuli and the participant's corresponding

responses and (3) the *stopping rule* which defines when to stop the procedure. In this study a four-interval, two-alternative forced choice paradigm was used in combination with an adaptive 3-down-1-up staircase method.

The method of stimulus presentation for each trial consisted of a series of four 500 ms white noise bursts (with a 20 ms raised cosine fade-in and fade-out) separated by 500 ms of silence (figure 3.19). The broadband white noise was generated with a flat magnitude response between 120 Hz and 16 kHz using a raised cosine roll-off down to 100 Hz and up to 18 kHz, as shown in figure 3.20. The phase response was randomised and the inverse FFT of the combined phase and magnitude response taken to generate the noise burst. The same white noise stimulus was used for each trial so that the randomised phase could not contribute to perceived differences. One of the middle two stimuli was filtered with the smoothed HRTFs (from now on referred to as the smoothed stimulus), the other three were filtered with the unsmoothed HRTFs (from now on referred to as the unsmoothed stimulus), giving two possible presentation orders: AABA and ABAA where A is the unsmoothed stimulus and B is the smoothed stimulus. This is a four-interval, two-alternative forced choice paradigm and has been used in a number of prior studies (Bernstein and Trahiotis, 1982; Kulkarni *et al.*, 1999; Kulkarni and Colburn, 2004; Shub *et al.*, 2008); although Bernstein and Trahiotis referred to it as a "2-cue, 2-interval forced choice". This paradigm ensures there is always a reference anchor of an unsmoothed stimulus either side of the smoothed

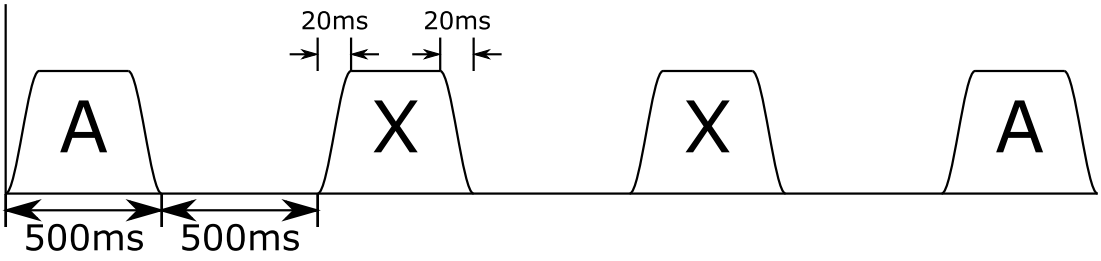


Figure 3.19: Method of stimulus presentation used in the listening test. In each trial the participant was presented with a series of four 500 ms white noise bursts with 20 ms raised cosine fade-in and fade-out, separated by 500 ms of silence. One of the middle two bursts was filtered with the smoothed HRIR whilst the other three were filtered with the unsmoothed reference HRIR, giving two possible presentation orders: AABA and ABAA.

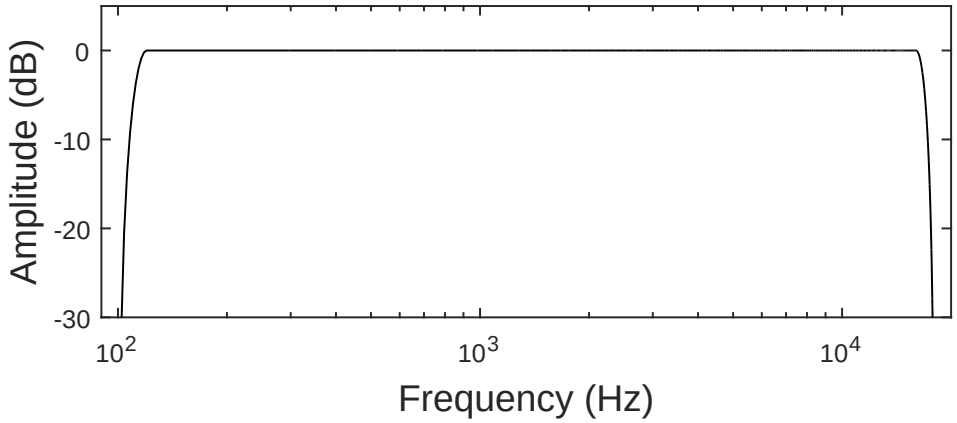


Figure 3.20: Defined frequency response of the white noise used in the listening tests.

one, which was found, informally, to aid reliable detection of the odd one out. As the aim of this study was to identify *any* discernible difference between the smoothed and unsmoothed stimulus, the listener's task was simply to pick "Which of the middle two samples is the odd one out?". If they were unsure they were instructed to guess. Using a broad question such as this should avoid bias towards a particular difference such as timbre or perceived direction. The listening test was created in MATLAB — see figure 3.21 for the graphical user interface (GUI).

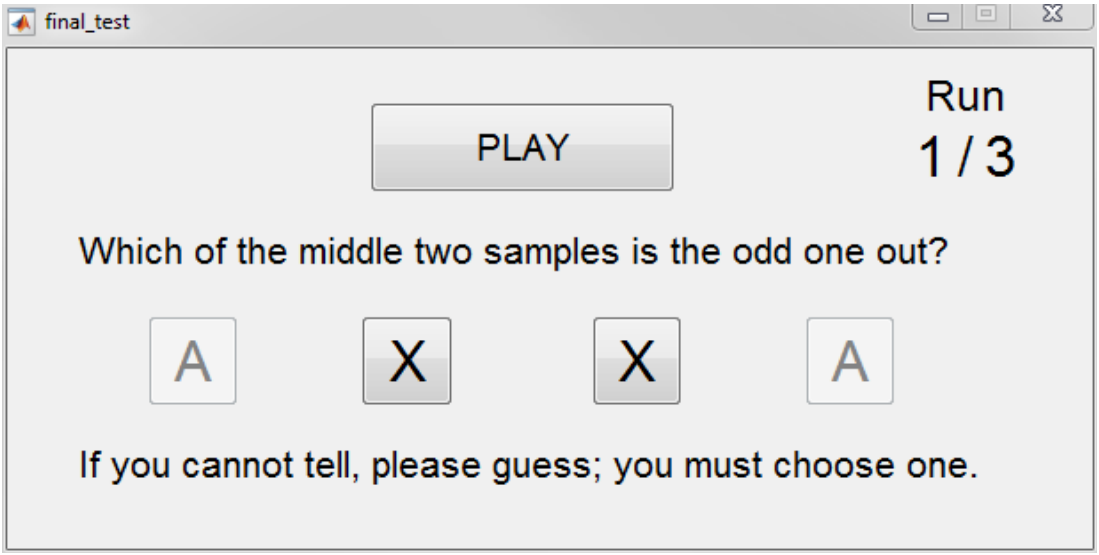


Figure 3.21: MATLAB GUI created for the listening test. The participant was presented with four bursts of HRTF-filtered white noise (three filtered with unsmoothed HRTFs and one with a smoothed HRTFs) and asked to pick which of the middle two bursts was the odd one out. If they were unsure they were instructed to guess.

The four-interval, two-alternative forced choice paradigm was combined with an adaptive staircase method (Levitt, 1971) in a similar manner to Shub *et al.* (2008). However, the staircase method used in this study had a 3-down-1-up progression rule, rather than the 2-down-1-up used by Shub *et al.* (2008), which means that the listener had to correctly identify the odd one out three times in succession for the smoothing level to be decreased (number of coefficients increased), whereas the level of smoothing was increased (number of coefficients decreased) after only one incorrect response. The 3-down-1-up method targets the 79% level of the psychometric function, that is, the level at which the listener answers correctly 79% of the time. This is compared to the 70% level for 2-down-1-up and only 50% for 1-down-1-up (Levitt, 1971). In staircase methods a *reversal* is defined as a change from correct to incorrect response, or vice versa, a *run* is a series of successive correct or incorrect responses between reversals and the *track* is all the responses from the starting trial to the finishing trial (Marvit *et al.*, 2003). Figure 3.22 shows a simulated track of responses that follows a 3-down-1-up progression rule. Red '+' signs indicate correct responses and yellow 'o' signs indicate incorrect responses. For this simulation the response was randomised for 15 coefficients or more and always correct for less than 15 coefficients, a significant simplification of actual subjective responses, but adequate for demonstrating how a track might look.

The number of coefficients for the starting trial was set to two; i.e. the maximum level of smoothing, aside from retaining just the level offset of the HRTF. This ensured that the odd one out was clearly audible and that the participant was engaged from the beginning of the test. The coefficient step size between each trial was not constant, rather it was decreased as the test progressed; a common approach in adaptive methods (Kidd *et al.*, 1989; Alves-Pinto and Lopez-Poveda, 2005; Shepherd and Hautus, 2007; Shub *et al.*, 2008). Initially the step size was set at five coefficients, until the first incorrect response, when it was decreased to three, then two on the next reversal and finally one. The test finished after 15 reversals at a step size of one, giving 16 runs at the smallest step size. The

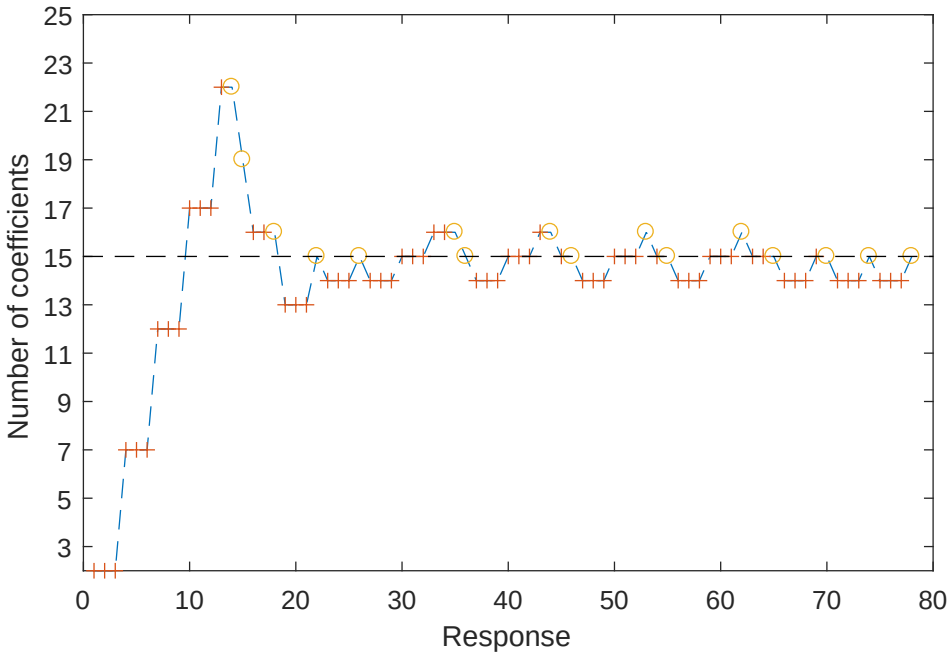


Figure 3.22: Simulated example of the adaptive 3-down-1-up procedure used in listening test. Responses were randomised for 15 coefficients and above and always correct below. Red '+' signs indicate correct responses, yellow 'o' signs indicate incorrect responses. The dashed line indicates the final level calculated as the average of the midpoints of the last eight runs, rounded up to the nearest integer.

datum definition was to calculate the final level of smoothing as the average of the midpoints of the last eight runs, rounded up to the nearest integer value.

The same three directions were tested for both Kulkarni and Colburn's algorithm and the proposed algorithm: (0°azimuth, 20°elevation), (90°azimuth, 0°elevation) and (-30°azimuth, -10°elevation). These directions were chosen to test a range of azimuths and elevations within the measured ranges whilst keeping the number of listening test sessions from escalating. For the listening test the headphones (Beyerdynamic DT990s) were driven by a MOTU Ultralite-mk3 audio interface with its output level calibrated to 70 dBSPL A-weighted, using unfiltered white noise. The calibration was carried out with the headphones placed on a dummy head and the probe microphone of a sound pressure level meter placed in the ear canal of the dummy head. After filtering with HRTFs this resulted in an average presentation level of 65 dBSPL A-weighted, which was found to be a comfortable listening level and falls within the range of presentation levels for listening tests as reported by Xie (2013). Informal pilot tests indicated that each track took

approximately 15–20 minutes to finish and that after about 30 minutes of testing, i.e. half way through the second track, ear fatigue set in and the test became increasingly difficult. Therefore to ensure that this did not happen, the testing was carried out in six separate test sessions; one for each direction-algorithm combination. The order in which the direction-algorithm combination sessions were carried out was randomised between subjects.

Eight unpaid participants from the Audio Lab at the University of York took part in the listening tests. All reported normal hearing and all had previous experience in listening tests. They each had their own HRTFs measured and processed, as outlined in section 3.5.1, and the listening tests took place in the listening space of the Audio Lab (average background noise level of 42 dBA). The participants were not disturbed for the duration of each session but they were allowed to take breaks whenever they wished.

3.5.4 Results

During the listening tests, every time the participant responded, and after multiple checks for correct responses, the current number of coefficients was stored. This resulted in a staircase vector of values as shown in figure 3.23. The final number of coefficients (dashed black line) was taken as the average of the midpoints of the last eight runs (blue 'x's), rounded up to the nearest integer. This resulted in two sets of 24 values (three directions $\times$ eight participants), one for each smoothing algorithm. These values represent the number of coefficients that it is necessary to retain when smoothing HRTFs using the corresponding algorithms, in order that white noise filtered with the smoothed HRTFs can be discriminated 79% of the time from white noise filtered with unaltered HRTFs.

Firstly a one-sample Kolmogorov-Smirnov test was carried out on each data set to test for normal distributions. The two data sets were pre-processed by subtracting their means and dividing by their standard deviations and the Kolmogorov-Smirnov test was then carried out on this pre-processed data. The one-sample

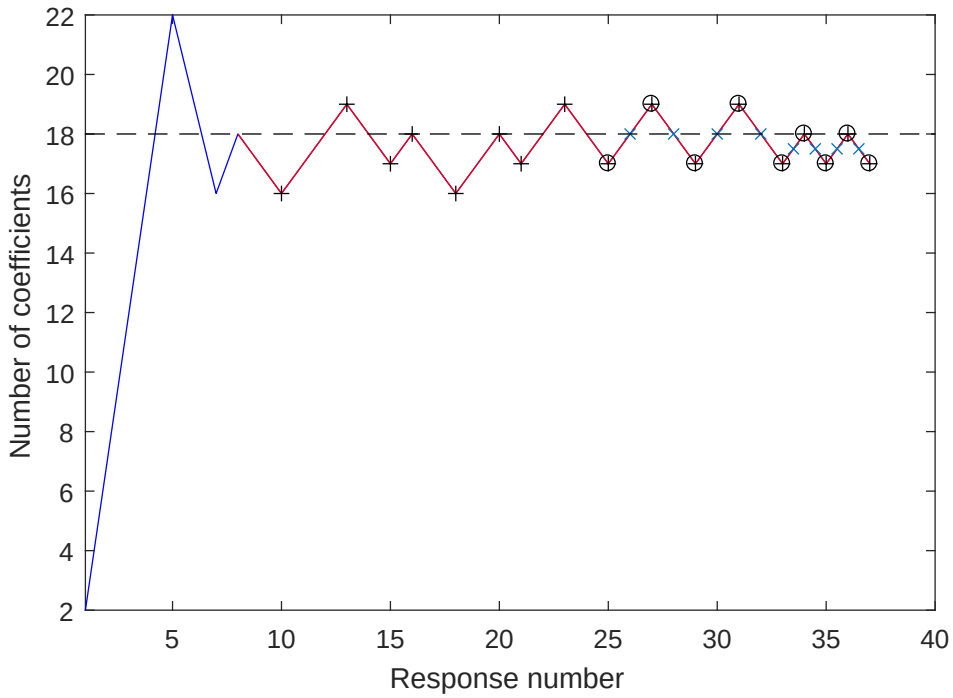
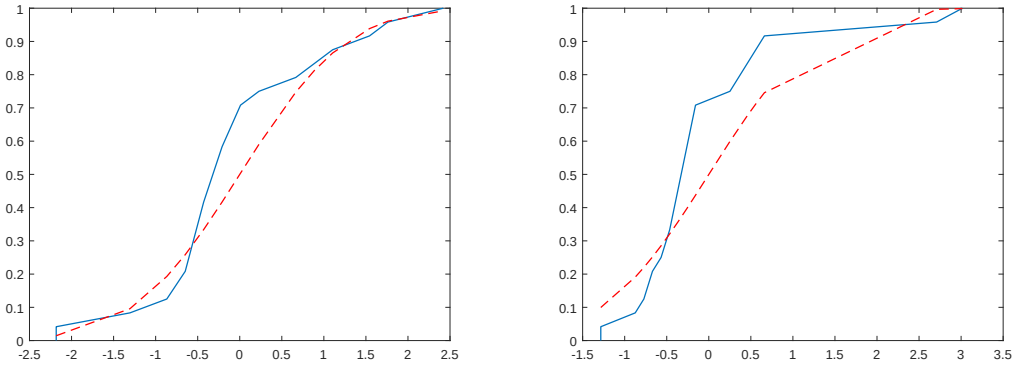


Figure 3.23: Exemplar listening test run. The portion of the track at the smallest step size is highlighted red whilst all the reversals at the smallest step size are marked as black '+' signs. Those in black circles represent the bounds of the last eight runs. The final smoothing level (black dashed line) was calculated as the average of the midpoints (blue 'x's) of the last eight runs, rounded up to the nearest integer.

Kolmogorov-Smirnov test tests the null-hypothesis that the data is from a normal distribution with a given mean and standard deviation. The p-value for Kulkarni and Colburn's algorithm was 0.0477 whilst the p-value for the new algorithm was 0.2330, indicating that the null hypothesis could not be rejected for the new algorithm at the 5% significance level but was rejected for Kulkarni and Colburn's algorithm. Figure 3.24 shows comparisons of the empirical cumulative distribution functions (CDFs) and standard normal CDFs for the results of both algorithms. It can be seen that the empirical results for the new algorithm match the normal distribution fairly well, whilst the empirical results of Kulkarni and Colburn's algorithm do not. Since the one-sample Kolmogorov-Smirnov test indicated that the results of Kulkarni and Colburn's algorithm do not come from a normal distribution, non-parametric statistical tests, which require no assumptions about the probability distribution of the data, will be carried out, rather than traditional parametric statistical tests, which assume the data is from a



(a) New algorithm (b) Kulkarni and Colburn's algorithm

Figure 3.24: Comparison of empirical (solid blue line) and standard normal (dashed red line) cumulative distribution functions (CDFs) for the listening test results of (a) the new algorithm and (b) Kulkarni and Colburn's algorithm.

normal distribution.

To determine whether the two data sets: namely, the results of Kulkarni and Colburn's algorithm and the results of the new algorithm, came from the same distribution, a two-sample Kolmogorov-Smirnov test was carried out. A p-value of 0.0029 indicates that the null hypothesis, that the two data sets come from the same continuous distribution, can be rejected. Therefore a two-sided Wilcoxon rank sum test was carried out on the results for Kulkarni and Colburn's algorithm and the new algorithm. A p-value of 0.0010 indicates that the null hypothesis that the distributions for the two algorithms have equal medians can also be rejected and therefore there is a difference in the median number of coefficients required for each algorithm. Furthermore a left-tailed rank sum test gave a p-value of 5.2220e-04, indicating that the median number of coefficients required for the new algorithm is lower than the median for Kulkarni and Colburn's algorithm. This can be seen in figure 3.25, which shows box plots of the results for both the new smoothing algorithm and Kulkarni and Colburn's algorithm. The median number of coefficients for the new algorithm is 17, compared to 22 for Kulkarni and Colburn's algorithm.

It is noteworthy that the two extreme outliers in the results of Kulkarni and Colburn's algorithm (marked as red '+'s) are from different participants, but for

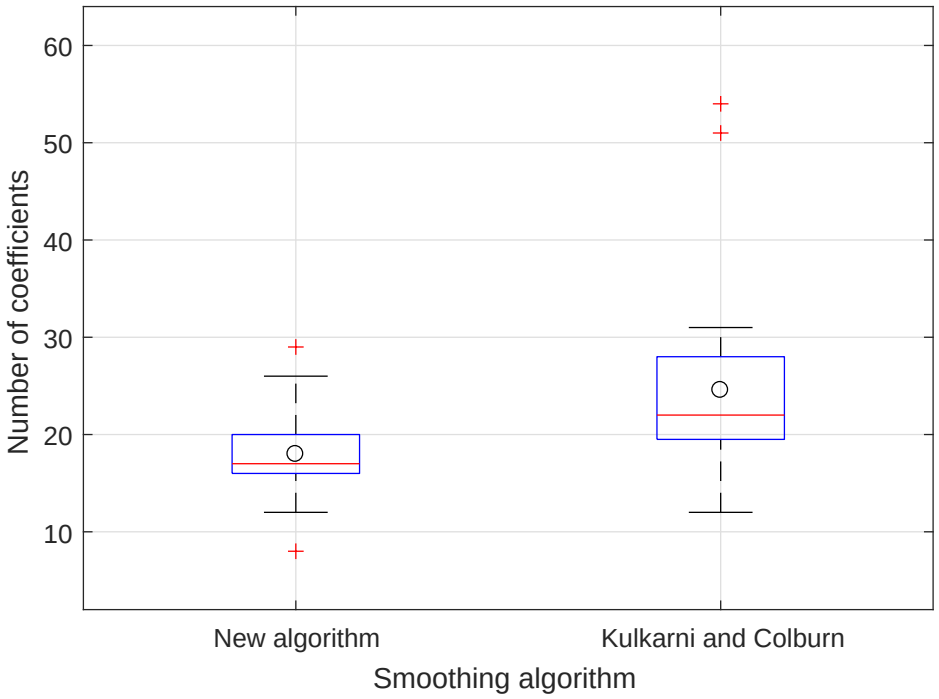


Figure 3.25: Box plot of the listening test results for both HRTF smoothing algorithms tested. The red lines represent the median result for each algorithm, the black 'o's the mean and the top and bottom of the boxes the 75th and 25th respectively. Red '+'s indicate outliers and the whiskers extend to the most extreme data points not considered outliers.

the same direction: (0°azimuth, 20°elevation). Figure 3.26 shows the results track for one of the two participants. The participant answered incorrectly a number of times at smoothing levels greater, i.e. fewer coefficients, than the finishing level, suggesting that the number of coefficients at the final level might be too high. However, there is also a series of what would have been, after triple checking, 24 correct responses leading up to the final threshold which is unlikely to happen by chance if the participant were guessing. Furthermore, figure 3.27 shows the squared error between the unsmoothed HRTF of the left ear for the direction (0°azimuth, 20°elevation) and the corresponding HRTF smoothed using Kulkarni and Colburn's algorithm with 50 coefficients retained for all participants. This is below the indicated final level for both the outlier participants, i.e. the error should be imperceptible, and it can be seen that even so, the outliers (highlighted red and blue) have much larger errors than the other participants; around 500 Hz for the blue curve and 1 kHz for the red. So it is possible that the outlier

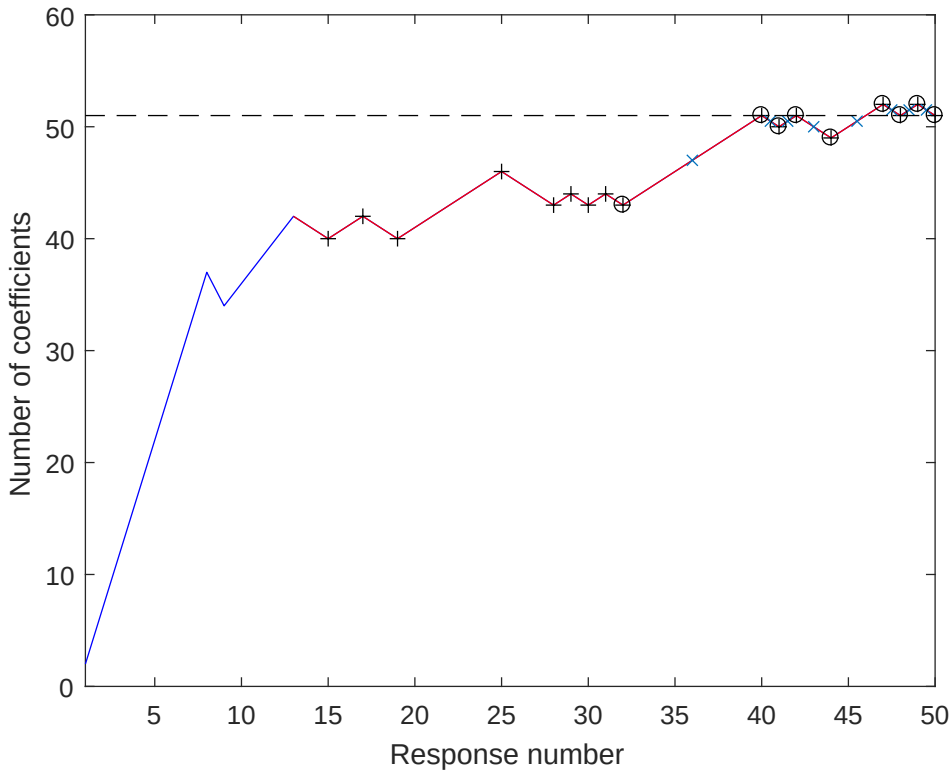


Figure 3.26: Listening test results track for one of the outliers in the results of Kulkarni and Colburn's algorithm.

results are correct and not anomalous. Therefore they have been retained in for the following analyses.

As a measure of difference between the two data sets the Vargha and Delaney (2000) *A measure* was calculated. In this case the A measure indicates, on average, how often Kulkarni and Colburn's algorithm required more coefficients than the new algorithm. A value of 0.5 would indicate equal performance, a value less than 0.5 would indicate that on average Kulkarni and Colburn's required fewer coefficients and a value greater than 0.5 would indicate that on average the new algorithm required fewer coefficients. The "size" of the effect increases the further the value is from 0.5. Vargha and Delaney (2000) classify a "small" effect as an A value of 0.56, a "medium" effect as a value of 0.64 and a "large" effect as a value of 0.71. The A measure for the two data sets in this study was 0.7760, indicating a large improvement of the new algorithm over Kulkarni and Colburn's algorithm.

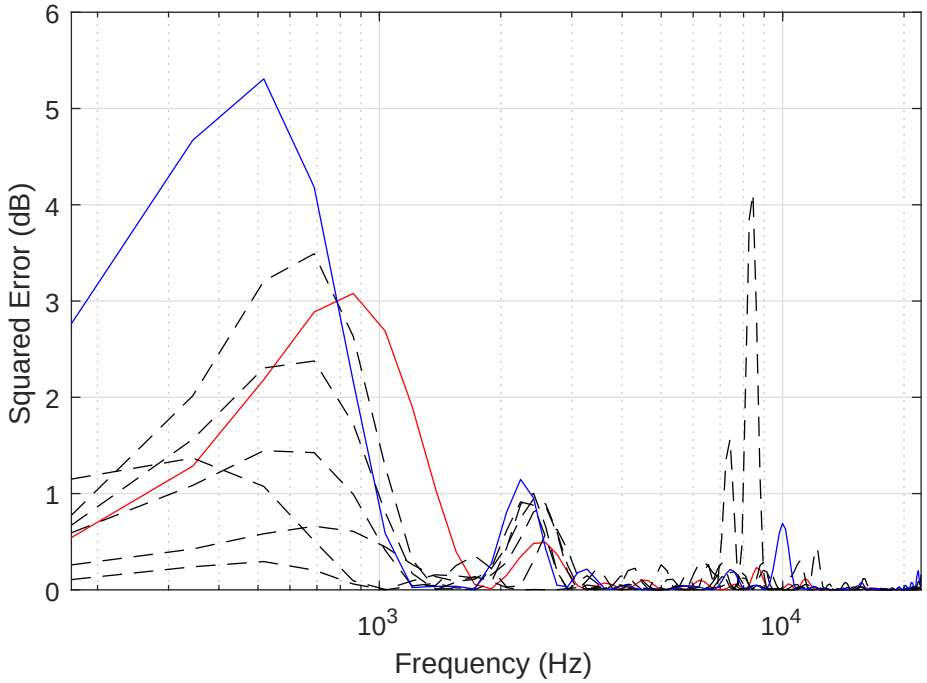


Figure 3.27: Squared error between the smoothed and unsmoothed HRTFs retaining 50 coefficients for the direction (0°azimuth, 20°elevation) using Kulkarni and Colburn's algorithm. Each of the curves represents the error for an individual listener's HRTFs. The curves for the two outliers in the listening test results are highlighted red and blue. It can be seen that the red curve shows significantly higher error levels between 1 and 2 kHz, whilst the blue curve shows significantly higher error below 1 kHz.

Smoothing Algorithm	Directions (-30°,-20°) vs (0°,20°)	Directions (-30°,-20°) vs (90°,0°)	Directions (0°,20°) vs (90°,0°)
Kulkarni and Colburn	0.0308	0.1963	0.0093
New algorithm	0.3899	0.1520	0.2681

Table 3.1: P-values for two-sided Wilcoxon rank sum tests carried out on the listening test results for different directions.

To investigate the influence of direction on the results, two-sided Wilcoxon rank sum tests were carried out on the three possible pairs of directions for each of the algorithms and the p-values are reported in table 3.1. The null hypothesis of equal medians could not be rejected for any of the pairs of directions for the new algorithm, indicating that direction has no effect on the required number of coefficients for the new algorithm. However, for Kulkarni and Colburn's algorithm the null hypothesis of equal medians was rejected for two of the three pairs of directions, indicating that direction does have an effect on the number of coeffi-

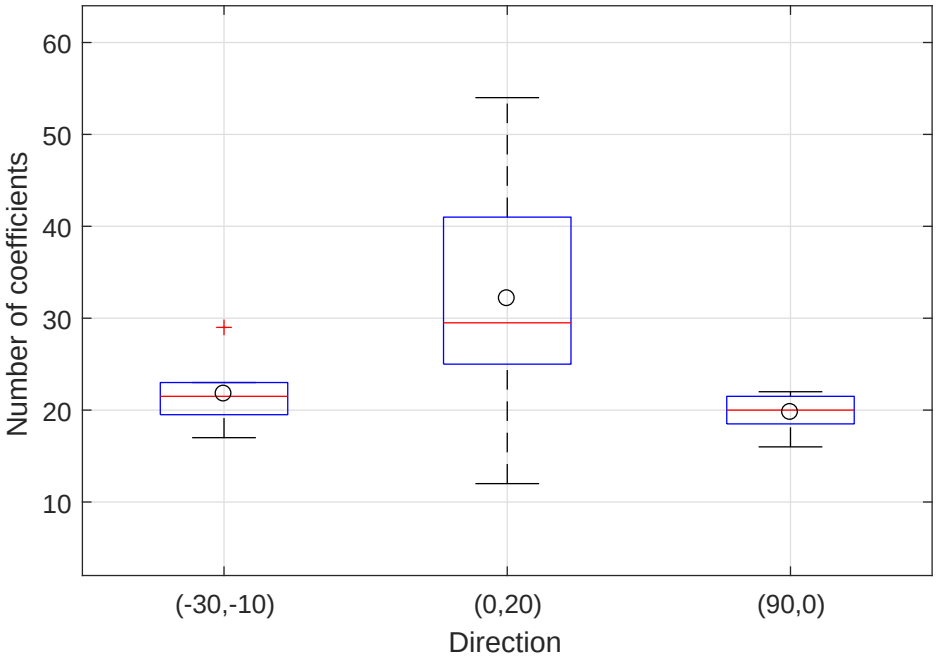


Figure 3.28: Box plot of listening test results for Kulkarni and Colburn's algorithm across the three directions tested.

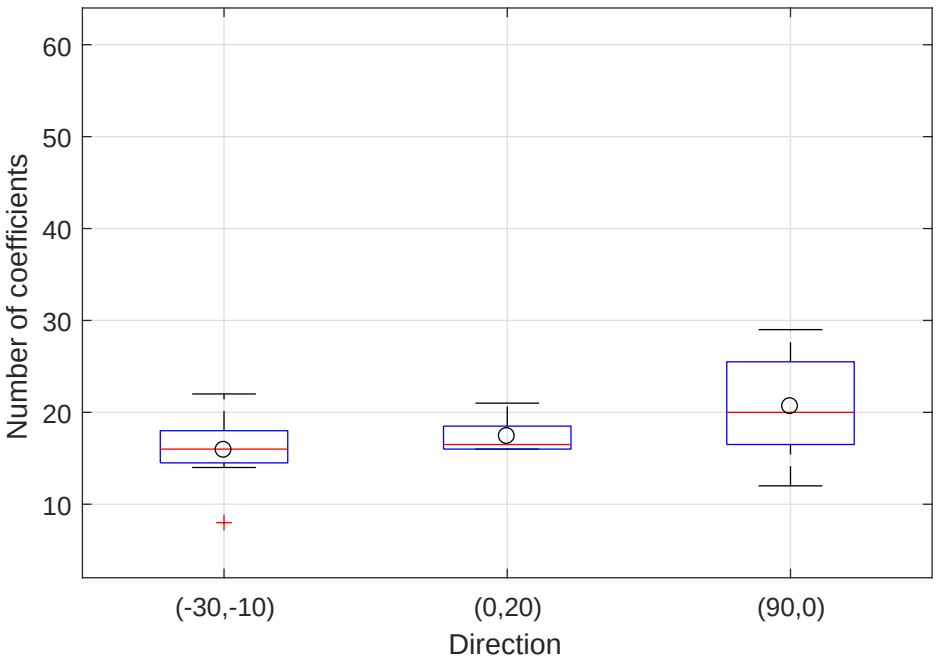


Figure 3.29: Box plot of listening test results for the new algorithm across the three directions tested.

cients required for their algorithm. Figure 3.28 shows boxplots of the results of Kulkarni and Colburn's algorithm for the three directions tested. It can be seen that the results for (0°azimuth, 20°elevation) differ dramatically from the other two directions as indicated by the rank sum test p-values. This is due to the two outliers previously identified. Conversely, as shown in figure 3.29, there is little variation between the three directions in the results for the new algorithm.

3.6 Summary

This chapter has presented a new, perceptually-motivated HRTF smoothing algorithm that allows perceptually transparent simplification of the complex spectral features of HRTFs. The algorithm smooths by discarding Fourier coefficients used in the synthesis of the HRTF magnitude spectrum based on an ERB criterion, rather than on a linear frequency scale. This increases the perceptually salient information retained during progressive simplification of the HRTF magnitude spectrum. The smoothing algorithm has been evaluated using both an auditory localisation model and subjective listening tests.

The perceptual results reported by Kulkarni and Colburn (1998) formed the basis of an objective evaluation of the smoothing algorithm using the sagittal plane (SP) auditory localisation model of Baumgartner *et al.* (2013). The results indicate that the proposed smoothing algorithm strongly outperforms Kulkarni and Colburn's (1998) algorithm. The auditory model suggests that an equivalent level of localisation performance can be achieved by the new algorithm using approximately half the number of Fourier coefficients. Whilst the listening test results also support the hypothesis that the new algorithm outperforms Kulkarni and Colburn's algorithm, using more sensitive perceptual discrimination listening tests, the advantage is less marked. The results of the listening tests suggest that more spectral coefficients are required than was predicted by the results of the model.

The dependence of Kulkarni and Colburn's listening test results on direction is of particular interest. For the direction (0°azimuth, 20°elevation) the new algorithm requires a median value of only 18 coefficients to perform similarly to Kulkarni and Colburn's algorithm using 30 coefficients. This direction is at the largest elevation, where spectral cues may be expected to increase in importance. The better performance of the new algorithm may, therefore, be attributable to its inherent ability to support perceptually salient spectral detail using fewer coefficients.

Figures 3.28 and 3.29 indicate that the closer the source direction is to the horizontal plane, the more similar the performance of the two algorithms. This again supports the notion that spectral details reduce in importance close to the horizontal plane.

One possible reason for the difference in results between the auditory model simulations and the listening test results is that the model by Baumgartner *et al.* (2013) compared the two algorithms at elevations across the whole of each SP and the results reported represent the average across all elevations and all SPs. It is therefore possible that the results from the model are dominated by elevations greater than those tested in the listening tests, since -10° to +20° represents a very small portion of a sagittal plane. The results of the listening test suggest that the performance differs more at elevations further from the horizontal plane, and so this may explain why the auditory model indicated a larger difference in performance.

Further investigation is needed to establish the degree of direction independence of the new smoothing algorithm, as suggested by these results. The level of acceptable smoothing indicated by this study may be used as the starting point for the adaptive staircase procedure used in the listening tests, which should reduce the duration of each run, thus allowing the testing of more directions. However, in future listening tests it might be advisable to use pink noise, rather than white noise. This is because pink noise has equal power per constant percentage

bandwidth, rather than equal power across all frequencies. Therefore, to the human ear, which as highlighted in this chapter operates on an approximately logarithmic frequency scale, pink noise sounds more evenly distributed across all frequencies. The additional high frequency energy in white noise makes it more audible than the high frequency content in pink noise, leading to it sounding "hissy". Therefore, it is possible that the white noise used in this study may have caused the subjects to focus on differences at high frequencies, whereas using pink noise may allow them to differentiate equally across all frequencies.

Despite the additional work identified above, being able to apply a fixed degree of smoothing across an entire HRTF, as suggested by these results, is a potentially valuable simplification which will assist with the wider aims of this research.

Chapter 4

Morphoacoustic perturbation analysis (MPA)

> Any sufficiently advanced technology is indistinguishable from magic.
>
> ARTHUR C. CLARKE

This chapter presents developments made in morphoacoustic perturbation analysis (MPA) (Thorpe, 2009; Tew *et al.*, 2012). MPA is a powerful tool for studying the morphological origin of head-related transfer function (HRTF) features. However, first generation MPA suffered from a number of weaknesses, mainly due to how the harmonic deformations were applied to the head mesh. In this chapter new methods for mapping head meshes to a sphere for the application of spherical surface harmonics are discussed, together with novel techniques for optimising the spherical head mesh in order to minimise distortion of the surface harmonics when mapped back to the head mesh.

4.1 Principles of MPA

Morphoacoustic perturbation analysis (MPA) was developed by Thorpe (2009) and reported by Tew *et al.* (2012) as a tool for studying the morphological origin of HRTF features. The principles of MPA are to apply harmonic deformations to a mesh description of the human head and use acoustic simulations to calculate the HRTFs for both the template head mesh and each of the harmonically deformed head meshes.

This results in a database of the harmonic deformations and their corresponding ΔHRTFs (the acoustic pressure changes introduced by application of each morphological shape deformation). From the database, the effect of an arbitrary shape perturbation can be investigated by appropriate weighting and summation of harmonic deformations and their corresponding ΔHRTFs.

The database can also be used in reverse to find the morphology that introduces particular changes in the HRTF pressures. This can, for example, be used to

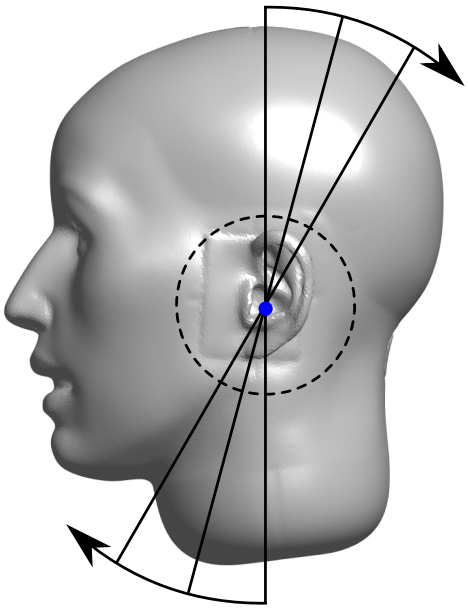


Figure 4.1: Radial slicing of the mesh head used in first generation morphoacoustic perturbation analysis (MPA) to create slice contours for the application of harmonic deformations. The dotted line shows an idealised cross-slice contour. After Tew *et al.* (2012).

investigate the morphology responsible for controlling the centre frequency of an HRTF notch or peak (Tew *et al.*, 2012) as explained in section 2.5.3.

First generation MPA uses "elliptic surface harmonic deformations" to perturb the surface of the Knowles Electronics manikin for acoustic research (KEMAR) mesh. The first step in applying these harmonic deformations is the same as the earlier work of Hetherington and Tew (2003a,b) and consists of radially slicing the head mesh (figure 4.1) around an axis that passes through the concha of each pinna. The axis must be placed carefully so that no disconnected "island" contours, such as those shown in figure 4.2, are generated.

After the head mesh is radially sliced, the S slice contours are uniformly sampled at P points. Figure 4.3 shows an example of the radial slicing axes for $S = 6$ slices and the resulting six slice contours, uniformly sampled at $P = 8$ points. It should be noted that first generation MPA assumed symmetry of the head and so each slice contour only traversed one half of the head. Of particular note here is that the resulting contours vary in length due to the differing morphology of the pinnae, and therefore the locations of the sampling points, indicated by 'x's in figure 4.3, vary greatly from contour to contour. For example, the fourth sample

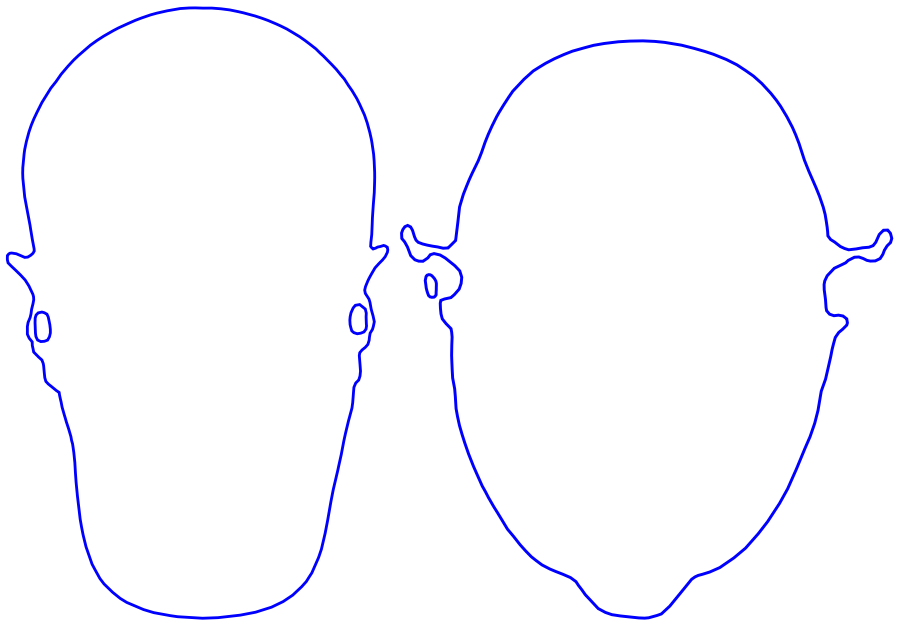


Figure 4.2: Example "island" contours from incorrect placement of radial slicing axis in first generation MPA.

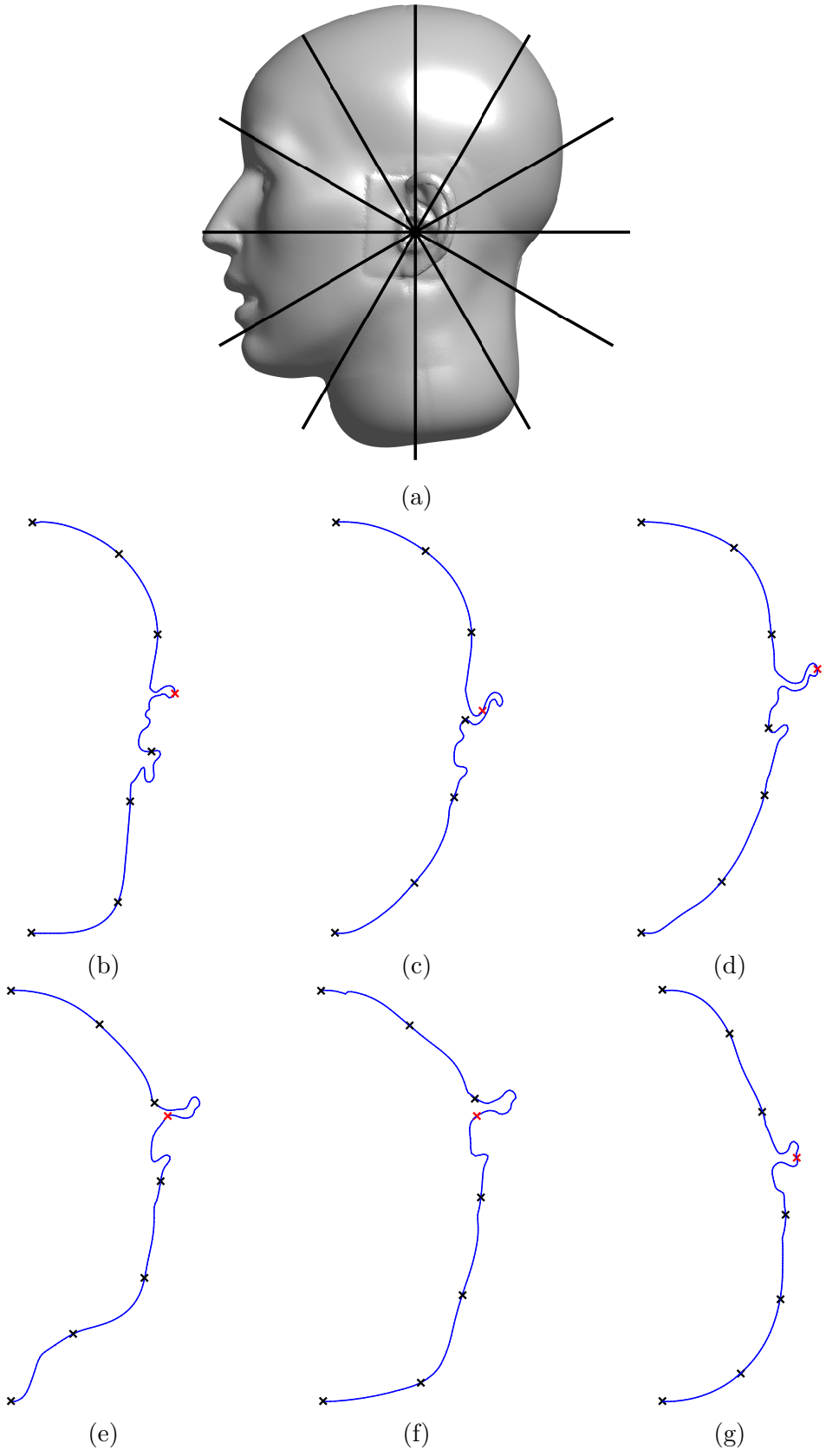


Figure 4.3: Example of radial slicing axes (a) and resulting $S = 6$ slice contours (b)–(g), sampled at $P = 8$ uniformly spaced points (black 'x's). The fourth sample point of each slice has been highlighted red to show the drift across a surface feature of the head mesh from one slice to another.

point from the top of each contour (highlighted red to aid visualisation) moves from the edge of the helix in contour (b) to the rear of the pinna in contour (c), then back to the edge of the helix in contour (d), before moving to the rear wall of the concha in contour (e). This results in distortion of the harmonic deformations across the surface of the head mesh, which is a significant limitation of the slicing approach, and will be discussed further in section 4.2.

After radial slicing and sampling of the S slice contours at P points, each contour is mapped to a two-dimensional (2D) surface on the x-y plane in a three-dimensional (3D) cartesian coordinate system (figure 4.4) where the p^{th} point of the s^{th} slice contour is mapped to the point:

$$(x, y, z) = (s, p, 0) \tag{4.1}$$

This results in a $P \times S$ surface, as shown in figure 4.5a, to which small displacements, $q(s, p)$ $(0 \leq s \leq S-1,\ 0 \leq p \leq P-1)$, are applied to displace the sample points along the z-axis of Cartesian space. The change in z coordinate is then mapped back to the slice contour to perturb the surface perpendicularly to the contour.

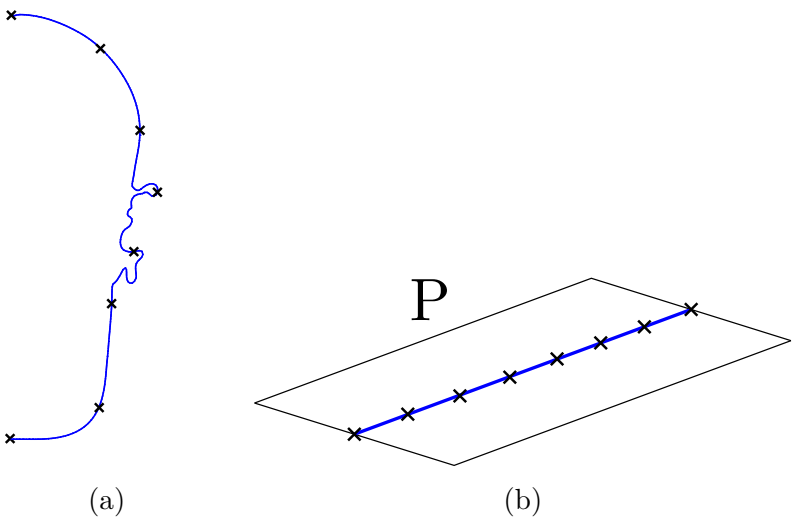


Figure 4.4: Exemplar mapping of a slice contour to the 2D rectangular plane for application of harmmonic deformations. The p^{th} point of the s^{th} slice contour is mapped to the point $(x, y, z) = (s, p, 0)$ on the plane.

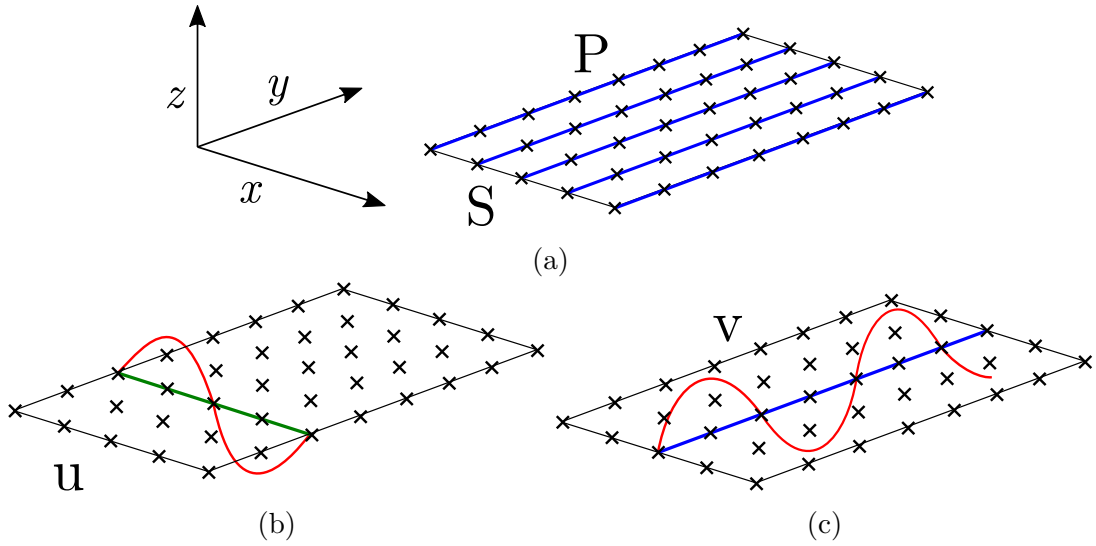


Figure 4.5: Application of in-slice and cross-slice harmonics to an exemplar 2D plane. (a) The resulting 2D plane from $S = 6$ slice contours each sampled at $P = 8$ points. (b) Exemplar cross-slice harmonic deformation u. (c) Exemplar in-slice harmonic deformation v.

The displacements, $q(s,p)$, are synthesised and analysed via application of the 2D Fourier transform:

$$q(s,p) = \sum_{u=0}^{S-1}\sum_{v=0}^{P-1} Q(u,v) e^{2\pi j\left(\frac{su}{S}+\frac{pv}{P}\right)} \tag{4.2}$$

where u $(0 \leq u \leq S-1)$ and v $(0 \leq v \leq P-1)$ are the cross-slice (figure 4.5b) and in-slice (figure 4.5c) harmonic numbers, respectively. The surface harmonic coefficients $Q(u,v)$ are given by:

$$Q(u,v) = \frac{1}{SP}\sum_{u=0}^{S-1}\sum_{v=0}^{P-1} q(s,p) e^{-2\pi j\left(\frac{su}{S}+\frac{pv}{P}\right)} \tag{4.3}$$

where $q(s,p)$ and $Q(u,v)$ form a Fourier transform pair.

The exponential form of the Fourier transform in equation 4.2 can be separated into real (cosine) and imaginary (sine) parts:

$$q(s,p) = \sum_{u=0}^{S-1}\sum_{v=0}^{P-1} Q(u,v)\cos\left[2\pi\left(\frac{su}{S}+\frac{pv}{P}\right)\right]\ldots \tag{4.4}$$

$$+\sum_{u=0}^{S-1}\sum_{v=0}^{P-1} jQ(u,v)\sin\left[2\pi\left(\frac{su}{S}+\frac{pv}{P}\right)\right] \tag{4.5}$$

and since the deformations, $q(u, v)$, are real, by exploiting Hermitian symmetry and making the following substitutions:

$$A_{u,v} = Q(u, v) + Q(u, v)^* = 2|Q(u, v)| \cos \angle Q(u, v) \tag{4.6}$$

$$B_{u,v} = j\left(Q(u, v) - Q(u, v)^*\right) = -2|Q(u, v)| \sin \angle Q(u, v) \tag{4.7}$$

equation 4.4 becomes:

$$q(s, p) = \sum_{u=0}^{S/2-1} \sum_{v=0}^{P/2-1} \left\{ A_{u,v} cos\left[2\pi\left(\frac{su}{S} + \frac{pv}{P}\right)\right] + B_{u,v} sin\left[2\pi\left(\frac{su}{S} + \frac{pv}{P}\right)\right]\right\} \tag{4.8}$$

Finally a new variable is introduced:

$$C(\sigma, u, v) = \begin{cases} A_{u,v} & \text{for } \sigma = 0 \\ B_{u,v} & \text{for } \sigma = 1 \end{cases} \tag{4.9}$$

as per the differential pressure synthesis (DPS) formulation of Tao *et al.* (2003b). This allows the change in pressure introduced by an arbitrary perturbation to be expressed as:

$$\Delta p(\theta, \phi, f) \approx \sum_{\sigma=0}^{1} \sum_{u=0}^{S/2-1} \sum_{v=0}^{P/2-1} \frac{\partial p_{\theta,\phi,f}}{\partial C(\sigma, u, v)} C(\sigma, u, v) \tag{4.10}$$

where θ and ϕ are the azimuth and elevation, respectively, of a pressure source relative to the head, and f is the acoustic frequency. This approximation is only valid if the perturbations $q(s, p)$ are small enough for the relationship between the amplitude of the shape variation and the corresponding pressure variation to be substantially linear. Tew *et al.* (2012) suggest a maximum contour displacement, q_{max}, of 0.1 mm. These pressure changes are grouped into a vector, $\Delta \boldsymbol{p}(\theta, \phi, u, v, \sigma)$, termed the ΔHRTF ("delta HRTF"), representing the changes over the frequency range of interest.

Boundary element method (BEM) simulations are run for the unperturbed head

mesh, as well as head meshes to which surface harmonic deformations $c(s, p, u, v, \sigma)$:

$$c(s, p, u, v, 0) = q_{max} \cos\left[2\pi\left(\frac{su}{S} + \frac{pv}{P}\right)\right] \tag{4.11}$$

$$c(s, p, u, v, 1) = q_{max} \sin\left[2\pi\left(\frac{su}{S} + \frac{pv}{P}\right)\right] \tag{4.12}$$

for all combinations of s, p, u, v and σ, have been applied. The resulting ΔHRTFs and the associated harmonic deformations, $c(s, p, u, v, \sigma)$, constitute the MPA database.

Once the database is complete, the acoustic change introduced by arbitrary small perturbations of the head mesh can be calculated. This is done by first representing the arbitrary perturbation as a weighted sum of the harmonic deformations, $c(s, p, u, v, \sigma)$, and then calculating the corresponding weighted sum of their relative ΔHRTFs.

Alternatively the database can be used in reverse to investigate the morphological origin of HRTF features. Figure 4.6 shows an HRTF notch investigated using MPA by Tew *et al.* (2012). An increase in the centre frequency of the notch would result in an increase in pressure on the low frequency slope of the notch and a decrease in pressure on the high frequency slope, as indicated by the green

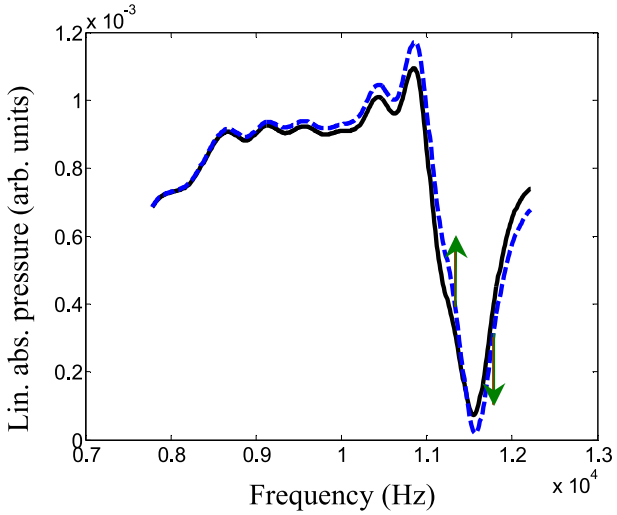


Figure 4.6: Notch investigated by Tew *et al.* (2012) using MPA. The black line shows the original HRTF, the blue dashed line the template HRTF plus the weighted sum of the acoustic changes introduced by the harmonic deformations and the green arrows indicate the morphing vector $\boldsymbol{m}$. Image from Tew *et al.* (2012).

arrows. These pressure changes are defined in the so-called morphing vector, $\boldsymbol{m}$, and the dot product of $\boldsymbol{m}$ with the ΔHRTFs is taken to create a set of weights:

$$w(\theta, \phi, u, v, \sigma) = \boldsymbol{m} \cdot \Delta \boldsymbol{p}(\theta, \phi, u, v, \sigma) \tag{4.13}$$

The weights are used to sum the harmonic deformations to create a displacement of the template mesh at each of the nodes:

$$\mu(s, p) = \sum_{\sigma=0}^{1} \sum_{u=0}^{S/2-1} \sum_{v=0}^{P/2-1} w(\theta, \phi, u, v, \sigma) c(s, p, u, v, \sigma) \tag{4.14}$$

The displacements represent how much a change in morphology at that node contributes to the specified change in the HRTF spectrum. The surface of the pinna can then be coloured according to the magnitude of μ as shown in figure 4.7, with warm colours representing displacements outwards and cool colours displacements inwards with respect to the surface of the template mesh.

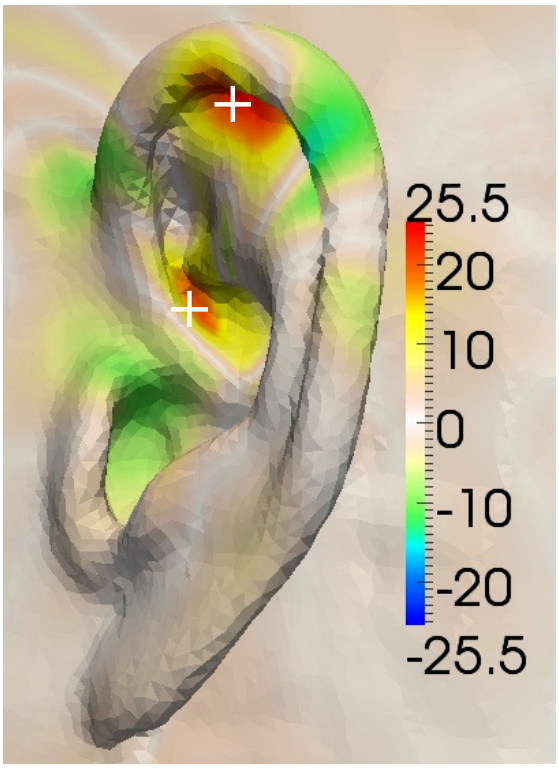


Figure 4.7: Pinna sensitivity map for the notch investigated by Tew *et al.* (2012) shown in figure 4.6. Warm colours represent outward movement of the pinna surface in the generation of the notch, whilst cold colours indicate inward movement. Image from Tew *et al.* (2012).

4.2 Limitations of previous work

A strength of the radial slicing approach used in first generation MPA, and outlined in the previous section, is that slices are concentrated in the pinna, making it inherently easy to achieve relatively high resolution in this region. This is beneficial because the pinna is the principal means of creating spectral spatial cues and therefore the area of greatest interest for studying the spectral origin of HRTF cues using MPA. High resolution in the pinnae occurs naturally in the cross-slice deformations when the axis of rotation is placed in the conchae, but radial slicing provides no equivalent improvement in resolution in the radial (in-slice) direction. Instead, greater radial resolution is achieved by use of nonlinear sampling of the contours, with higher sampling rates concentrated in the pinna regions.

The key limitation of the slicing approach, however, is the way in which contours vary in length from one slice to another due to the complex morphology of the pinnae, as briefly mentioned in section 4.1. For example the horizontal contour that passes through the tragus and the back of the pinna is much shorter than the vertical contour that passes through the lobule and the top of the pinna. Therefore, rather than the idealised cross-slice contour shown in figure 4.1, real cross-slice contours appear more like that shown in figure 4.8. This may become problematic when the lengths of neighbouring contours differ substantially. When two such contours are uniformly sampled, two points close to each other on the 2D surface may not be close in 3D space. Figure 4.9 shows a visualisation of this problem. The two solid lines show two potential radial slicing planes and the coloured dots represent sampling points on the two corresponding contours (the two planes have been placed further apart than neighbouring contours would normally be in order to aid visualisation). The green highlighted pair of sample points have identical p index values and so are adjacent on the 2D S-P surface (see figure 4.5). They also lie close to each other on the head mesh surface. This leads to a direct correspondence between the wavelength of spatial deformations

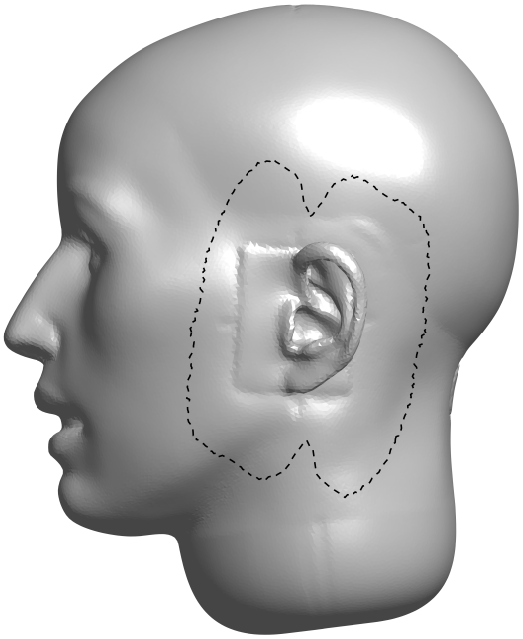


Figure 4.8: Example of a real cross-slice contour.

applied in the S-P plane and the wavelength of these deformations on the head mesh surface. The situation with the pair of red points is different, however. These also have identical p index values but, due to differences in the contour lengths of these two slices, they are not adjacent on the head mesh surface.. This leads to a "sheering" effect when the harmonic deformations are mapped to the head and results in a wide variation in deformation wavelength on the head surface for a given harmonic number. Therefore many more harmonics are required to achieve satisfactory spatial and acoustic resolution over the entire head surface. In addition, the harmonic deformations on the surface of the head mesh suffer distortions as shown in figure 4.10.

Applying 2D normal projections of a set of 3D harmonic deformations to the surface mesh would, in principle, avoid the problems of sheer, slicing and island creation. In effect, independent perturbation of each vertex within a volume containing the surface mesh may be controlled using three sets of 2D perturbation surfaces in the x-y, y-z and z-x planes, respectively. A vertex can be displaced in any 3D direction by means of a weighted sum of the 2D perturbations. The projection of the 3D movement in the direction of the surface normal in this

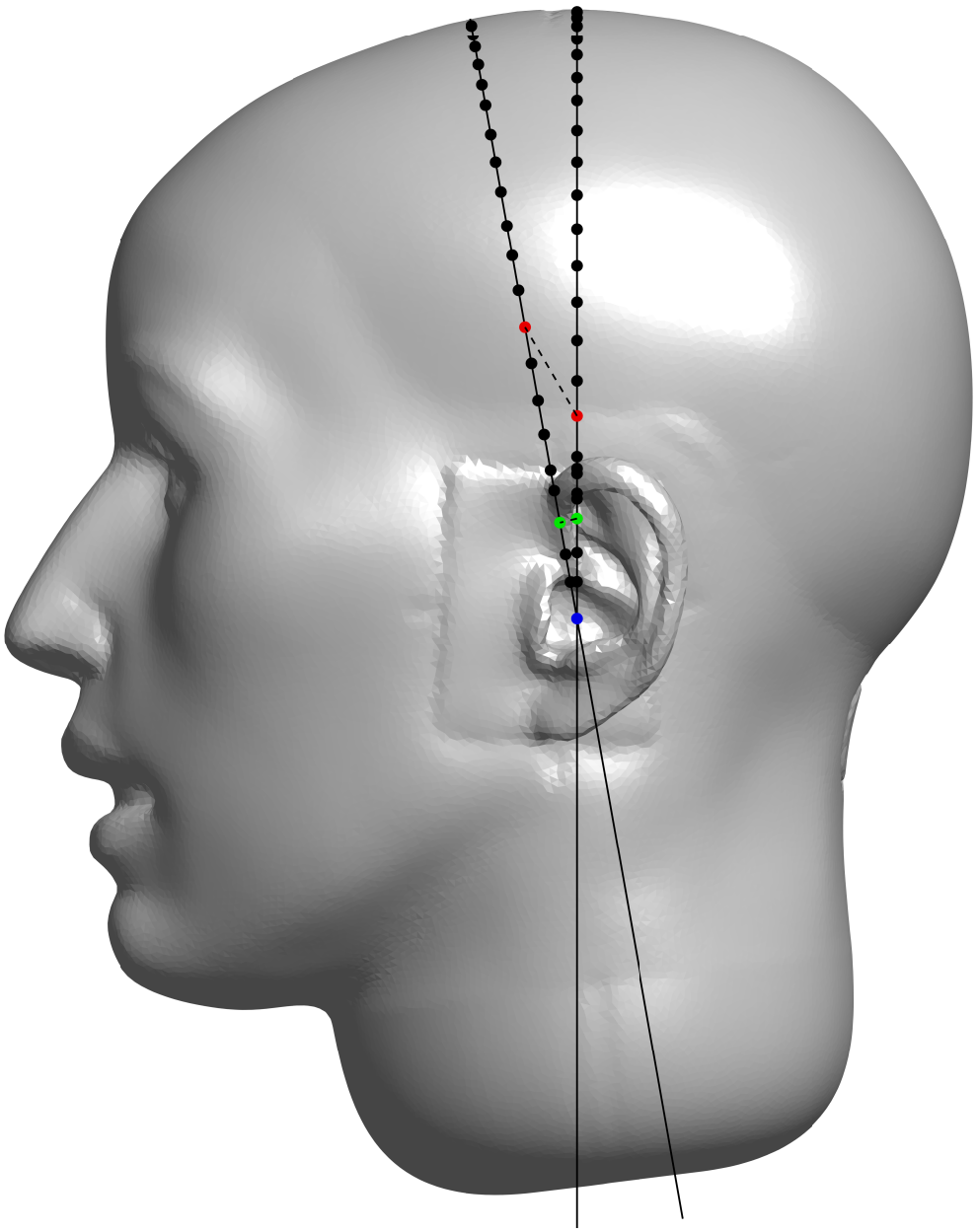
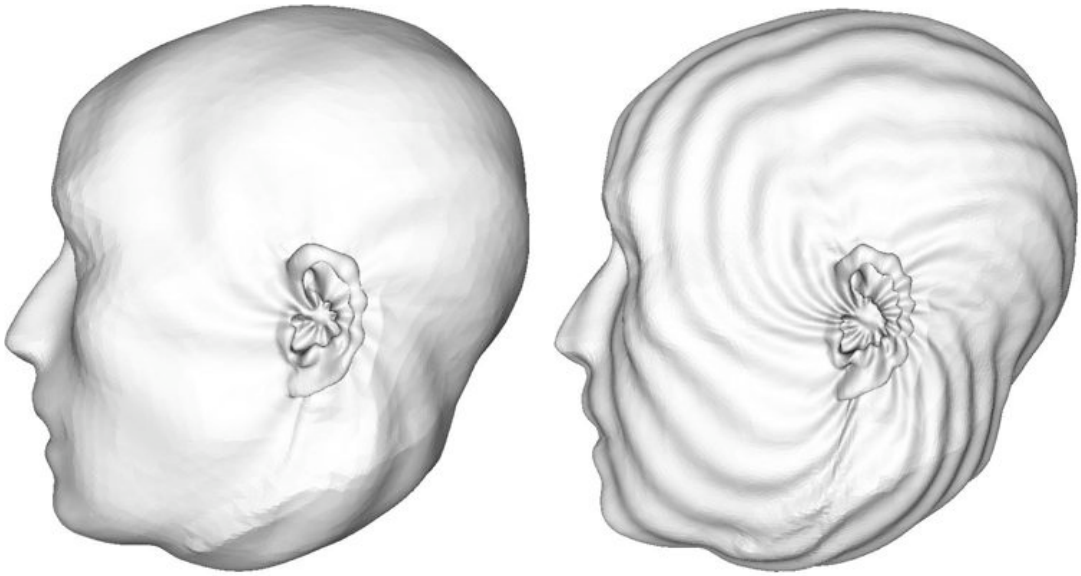


Figure 4.9: Visualisation of the main limitation of first generation MPA. The straight lines represent two slicing axes. The blue dot lies at the origin of the slicing axes and the black circles represent equally-spaced sample points along each slice contour. The green highlighted points show an example of where sample points that are close together in 3D space are also close together in the S-P 2D plane (figure 4.5). The red highlighted sample points show an example where sheering has occurred. Here, adjacent points on the S-P surface (i.e. with the same p index value on neighbouring slices) are not adjacent on the original head mesh surface.



(a) (b)

Figure 4.10: Exemplar harmonic deformations on the surface of the head mesh as generated by first generation MPA. The sheering effect due to the variation in contour length can be seen in the distortions of the harmonic deformations in (b): ideally the harmonic deformations would form a perfect spiral. Images from Thorpe (2009).

region is used to apply a deformation to the vertex.

Although this approach is attractive for the reasons outlined above, it is fatally computationally intensive. This increase in computation is firstly due to the extra dimensionality of the harmonics required, compared to 2D surface harmonic deformations. Secondly, the computational burden is exacerbated by the nature of the pinna surface, which folds back on itself to produce a thin lamina (the cartilage). For clarity, this phenomenon is visualised in one dimension in figure 4.11. MPA requires the ability to apply arbitrary displacements to the surface of the head mesh (figure 4.11b). These are generated by means of weighted sums of orthogonal functions. If the surface is wrapped back on itself (figure 4.11c), such as occurs in the complex morphology of the pinna, then when the deformations are 2D and applied over the surface of the mesh, the displacements wrap with the surface (figure 4.11d). This creates the desired structure of harmonic deformations. However, when 3D deformations are applied (figure 4.11e) then, due to the close proximity of the two regions of the surface in 3D space, extremely high orders of 3D harmonics would be required to produce arbitrarily different defor-

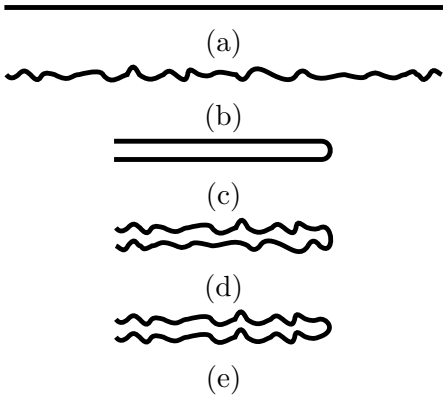


Figure 4.11: Visualisation of the problem with applying MPA deformations to the head mesh in 3D. Consider a cross-section of a flat surface (a). The fundamental principle of MPA is to apply arbitrary deformations to the surface of the mesh (b) using weighted sums of orthogonal functions. Now consider when the surface is wrapped back on itself (c). If the harmonic deformations are 2D across the surface then they also wrap with the surface (d). However, applying 3D harmonic deformations will, in general, result in (e), where the deformations on each side of the laminar are similar due to their proximity. It would therefore require extremely high order harmonics to produce different deformations on each side of the surface.

mations on either side of the lamina. This would increase the size of the MPA database to a degree which would make it impossible to create in a reasonable time.

A novel contribution of the work reported in this chapter is the development of a method for mapping the vertices of a head mesh to the surface of a sphere whereupon surface spherical harmonic deformations are applied. This inherently eliminates sheering and, by also keeping the harmonic deformations in a 2D domain, holds the computational load in check. Furthermore, the distribution of the head mesh vertices on the surface of the sphere can be optimised to encourage a close relationship between harmonic number and acoustic resolution. This establishes a simple trade-off between MPA database size and the acoustic resolution achievable.

4.3 Mesh description of head and pinnae

Preliminary work on improvements to MPA was carried out on the head meshes in the Sydney-York morphological and recording of ears database (SYMARE

database) (Jin *et al.*, 2013), which include meshes valid for maximum acoustic frequencies of 4, 8, 12, 16 and 20 kHz. However, the final database was generated for a head-only KEMAR mesh. This is because the KEMAR is designed to represent, in some sense, the average human head and torso morphology (Burkhard and Sachs, 1975), compared to the individual meshes of the SYMARE database. Furthermore KEMAR is readily available for verification and extension of the results of this study. The KEMAR mesh was generated from computer aided design (CAD) data of the G.R.A.S. KEMAR[1] and the CAD mesh was processed and optimised for BEM simulations as outlined in Young *et al.* (2016). The head of the KEMAR was separated from the torso and the neck was rounded in a similar manner to the head-only meshes of the SYMARE database. Surface rounding reduces the acoustic artefacts which would be caused by diffraction at the sharp edges of a truncated neck. Figure 4.12 shows the head mesh used to generate the database. A widely applied guideline for the boundary element method states that the length of no mesh edge should exceed one sixth of the acoustic wavelength at the highest frequency of analysis (Katz, 2001a,b; Kahana and Nelson, 2006, 2007). On this basis the mesh adopted in this work is valid for simulation by the BEM for frequencies up to 20 kHz.

The mesh $\mathcal{M} \in \mathbb{R}^3$ is described topologically by a set of n vertices $\mathcal{V}$, m triangular faces $\mathcal{F}$ and $3m$ edges $\mathcal{E}$. Topologically the vertices $\mathcal{V}$ are an abstract set of indices:

$$\mathcal{V} = \{1, 2, \ldots, n\} \tag{4.15}$$

and geometrically each vertex is defined by a triplet of Cartesian coordinates in 3D space:

$$V = (\mathbf{v}_i)_{i=1}^{n} = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}_{i=1}^{n}, \quad \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \in \mathbb{R}^3 \tag{4.16}$$

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

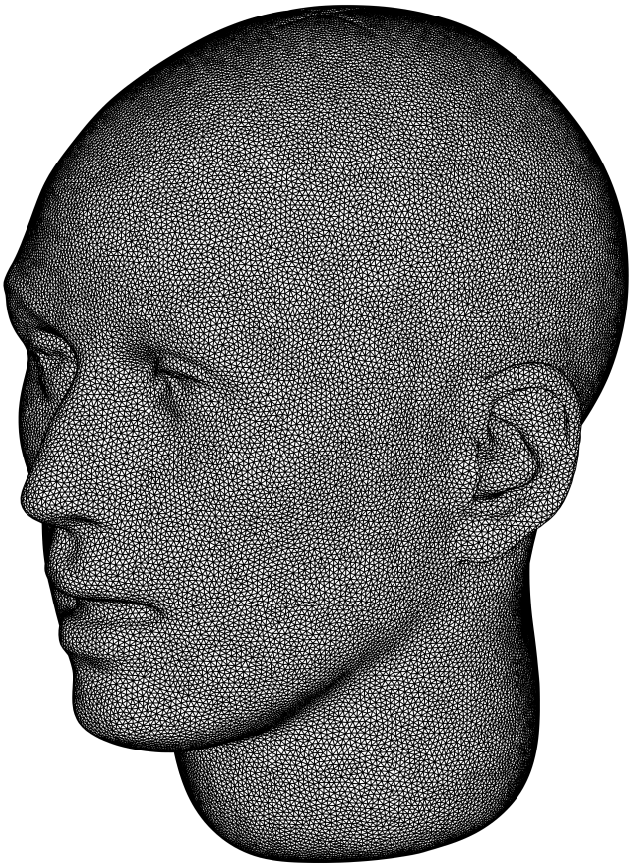


Figure 4.12: KEMAR head-only mesh used for MPA database generation. The edges of the truncated neck have been rounded to minimise acoustic diffraction effects.

The faces $\mathcal{F}$ are triplets of vertex indices:

$$\mathcal{F} = (f_i)_{i=1}^{m}, \quad f_i \in \mathcal{V} \times \mathcal{V} \times \mathcal{V} \tag{4.17}$$

and the edges $\mathcal{E}$ are pairs of vertex indices:

$$\mathcal{E} = (e_i)_{i=1}^{3m}, \quad e_i \in \mathcal{V} \times \mathcal{V} \tag{4.18}$$

For a correctly oriented mesh, i.e. all the surface normals point in the same direction with respect to the mesh surface, the set of edges is assumed to be symmetric:

$$(i, j) \in \mathcal{E} \iff (j, i) \in \mathcal{E} \tag{4.19}$$

This means that each pair of vertex indices appears in the edge matrix twice:

once with the vertex indices in one order and once in reverse. This creates an additional condition that:

$$(i,j,k) \in \mathcal{F} \iff (i,j)(j,k)(k,i) \in \mathcal{E} \tag{4.20}$$

and it is further assumed that there are no isolated edges:

$$\forall (i,j) \in \mathcal{E}, \quad \exists\, k, \quad (i,j,k) \in \mathcal{F} \tag{4.21}$$

4.4 Path length relaxation (PLR) mapping method

The process of mapping the KEMAR head and pinnae mesh onto a sphere is non-trivial, primarily due to the complex folds and cavities present in the pinnae. Not only must the topology of the mesh be preserved, but the original length of each edge should be disturbed as little as possible during mapping, so as to maintain a close relationship between harmonic deformation number and the final acoustic resolution of MPA. In this section we describe the development of a novel technique termed "path length relaxation (PLR)" for addressing these requirements.

In PLR mapping, each vertex in the head mesh is mapped to an azimuthal angle φ and an elevation angle θ in the spherical coordinate system commonly used in mathematics. The azimuthal angle φ is the angle in the x-y plane, measured in a right-hand sense from the x-axis to the point of intersection of the perpendicular from the point P. The elevation angle θ is the positive angle between the z axis and the line OP, as shown in figure 4.13.

Section 4.4.1 describes the mapping of the azimuthal angle φ for each vertex. It is the angle by which the x-y plane must be rotated about the y-axis to make it intersect with the vertex in question. The mapping of the elevation angle θ for

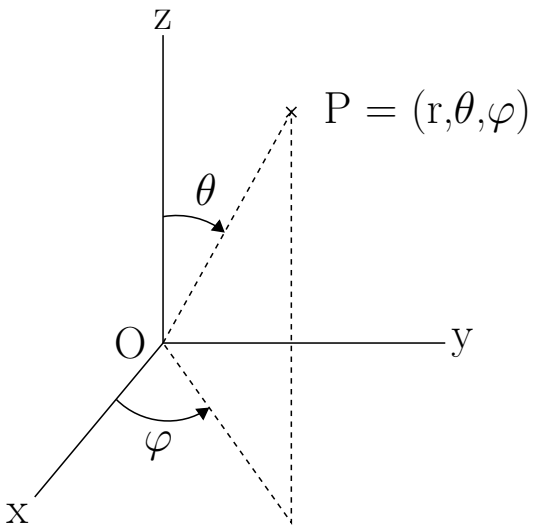


Figure 4.13: Spherical polar coordinate system used for mapping the head mesh vertices onto the unit sphere.

each vertex, as reported in section 4.4.2, is based on the distance travelled across the surface of the head mesh from a specified origin in the pinna to the vertex in question. This can be visualised by moving the pinna origin to the "north pole" on the sphere; then the azimuthal angle on the sphere is simply the azimuthal angle of the vertex in the head mesh, and the elevation angle is based on the distance travelled downwards across the surface of the mesh to reach the vertex (figure 4.14).

This is a natural adaptation of the slicing method of Thorpe (2009) and Tew *et al.* (2012) but avoids slicing and resampling the head mesh; each vertex is mapped directly to the surface of the sphere. However, it is still prone to unequal path lengths for neighbouring points due to sudden changes in morphology along adjacent paths. Therefore section 4.4.4 covers development of a "force minimisation" approach to reduce the distortion of the edge lengths in the spherical point set.

Finally, whilst PLR works well for the low resolution (4 kHz) SYMARE database mesh used to expedite development of the technique, section 4.4.5 details limitations found when extending it to full resolution head meshes.

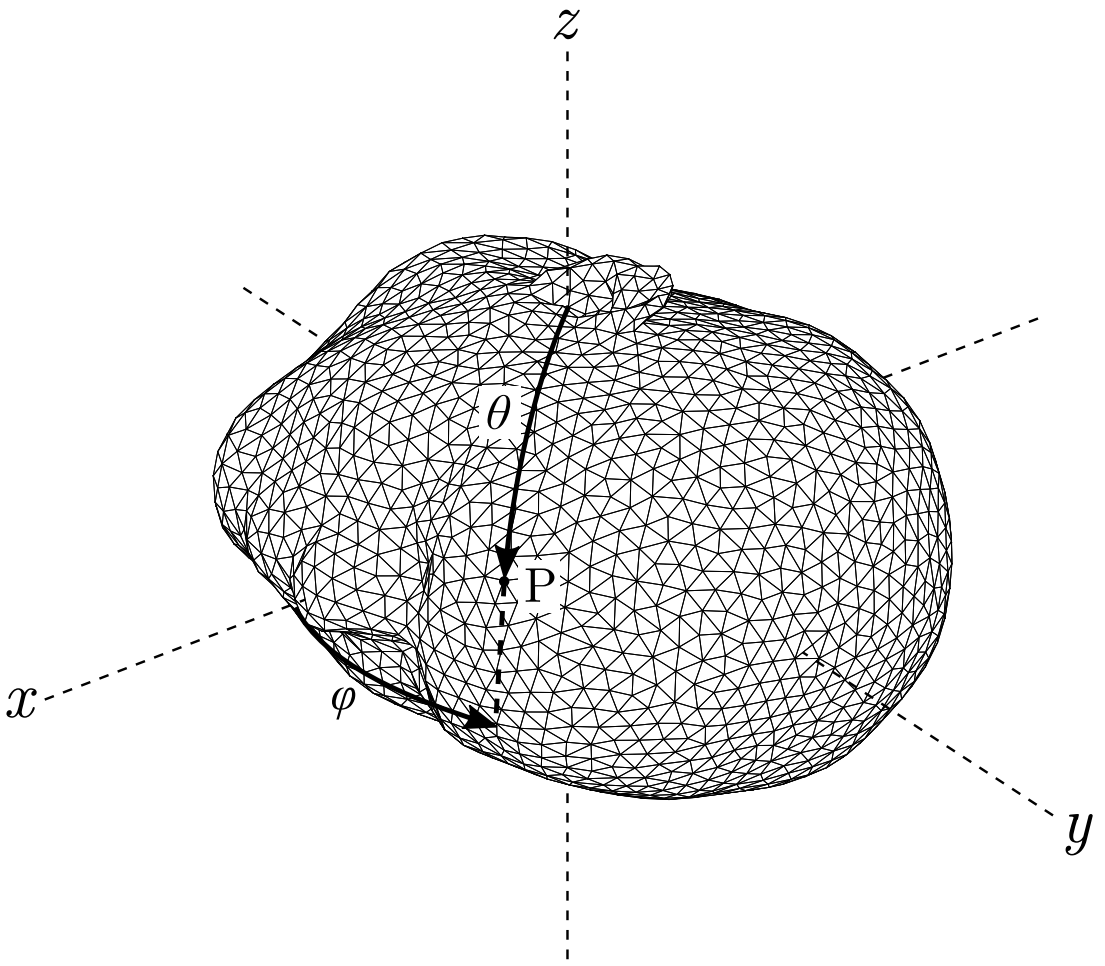


Figure 4.14: Visualisation of azimuthal φ and elevation θ angles used in path length relaxation (PLR) mapping.

4.4.1 Azimuthal angle φ mapping

The mapping of the azimuthal angle φ for PLR is simply a standard cartesian to polar coordinate conversion but with the y and z coordinates interchanged:

$$\varphi_i = \tan^{-1}\left(\frac{z_i}{x_i}\right) \tag{4.22}$$

where i is the index of the vertex being mapped. The azimuthal angle is therefore the effective rotation of the x-y plane about the y-axis to intersect the vertex in question (figure 4.15).

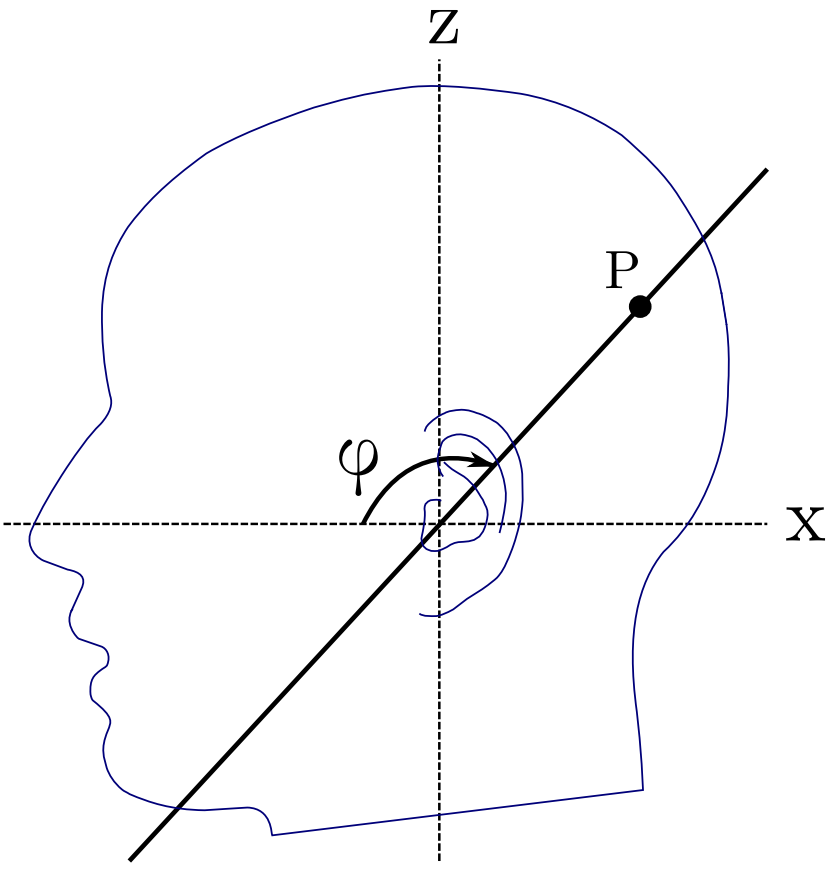


Figure 4.15: Visualisation of mapping of azimuthal angle φ.

Algorithm 4.1 Pseudocode of algorithm for calculating the distance used to calculate the elevation angle θ for a given vertex.

```
1: calculate equation of vertex plane
2: set currentFace to origin face
3: for all edges of currentFace do
4:     calculate intersection point of vertex plane and edge line
5:     check whether intersection point lies within bounds of edge
6: end for
7: find intersection point in correct direction based on vertex angle φ
8: update currentPoint
9: add distance from origin to currentPoint to dist
10: while currentPoint ≠ vertex do
11:     find other face that shares edge that currentPoint lies on
12:     update currentFace
13:     for all edges of currentFace do
14:         calculate intersection point of vertex plane and edge line
15:         check whether intersection point lies within bounds of edge
16:     end for
17:     update currentPoint to intersection point that is not currentPoint
18:     add distance moved to dist
19: end while
```

4.4.2 Elevation angle θ mapping

For PLR the elevation angle θ is calculated using the path length from an origin point on the pinna, along the surface of the head mesh, to the vertex being mapped. Note that, rather than using the shortest (geodesic) distance, this is the length of the contour created by the intersection of the rotated x-y plane (rotated by azimuth angle φ, to intersect the vertex in question) with the surface of the mesh.

The algorithm for calculating this distance for a given vertex is given in algorithm 4.1. In line 7 the process for finding the first intersection point in the correct direction is to calculate the vector $\boldsymbol{v}_o$ from the origin to each of the intersection points and choose the correct intersection point based on the following criteria:

if $\varphi == 0$ **then**

choose intersection point for which the x component of $\boldsymbol{v}_o$ is positive

else if $\varphi == \pi$ **then**

choose intersection point for which the x component of $\boldsymbol{v}_o$ is negative

else if $0 < \varphi < \pi$ **then**

choose intersection point for which the z component of $\boldsymbol{v}_o$ *is positive*

else

choose intersection point for which the z component of $\boldsymbol{v}_o$ *is negative*

end if

Figure 4.16 shows an exemplar path used for calculating this distance for a given vertex (indicated by □) on an arbitrary mesh. The origin point is marked with ◯, the path is indicated by the dashed line, the vertex's plane is outlined by a dotted line and the intermediate intersection points with face edges are marked using ×s.

To calculate the intersection points of the vertex plane and a face's edges, the equation of the rotated x-y plane is required. The general form of the equation of a plane in 3D-space is:

$$ax + by + cz + d = 0 \tag{4.23}$$

where the constants a, b and c define the vector normal to the plane $\boldsymbol{n}$ and, once

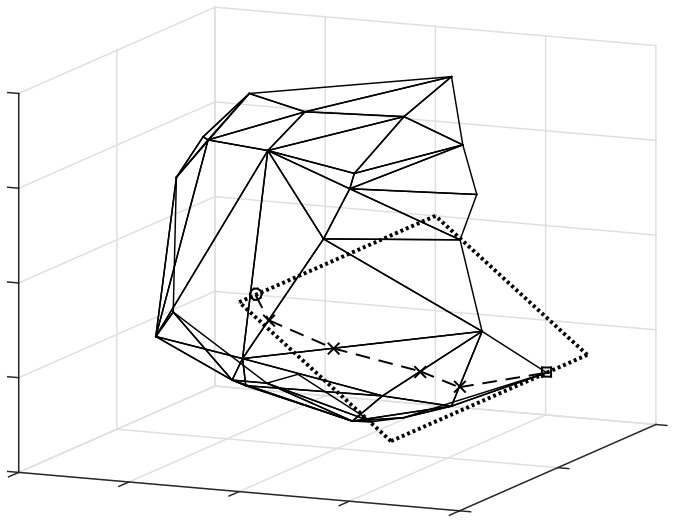


Figure 4.16: Exemplar path used for mapping elevation angle θ for a given vertex. ◯ origin, --- path, × intermediate points, □ vertex and ⋯ the vertex's plane.

these have been found, the constant d can be calculated from any point on the plane. The vector normal $\boldsymbol{n}$ can be calculated from three points on the plane (P, Q and R) as the cross product of the two vectors $\boldsymbol{PQ}$ and $\boldsymbol{PR}$:

$$\boldsymbol{n} = [a \quad b \quad c] = \boldsymbol{PQ} \times \boldsymbol{PR} \tag{4.24}$$

Then the constant d is calculated from a point on the plane $V_0 = (x_0, y_0, z_0)$ as:

$$d = -(ax_0) - (by_0) - (cz_0) = -\boldsymbol{n} \cdot V_0 \tag{4.25}$$

and any of the three points P, Q or R can be used for V_0. For calculating the vertex plane equation the three points used are:

$$\begin{aligned} P &= [x_i \quad y_i \quad z_i] \\ Q &= \left[0 \quad y_i + \frac{l}{2} \quad 0\right] \\ R &= \left[0 \quad y_i - \frac{l}{2} \quad 0\right] \end{aligned} \tag{4.26}$$

where $[x_i \quad y_i \quad z_i]$ are the coordinates of vertex i and l is the distance from the vertex to the y-axis perpendicular to the y-axis as shown in figure 4.17. The use of these three points ensures that the triangle they define is not too acute and thus the cross product is not too small. Figure 4.18 shows a visualisations of the rotated x-y plane for an exemplar vertex on an arbitrary deformed sphere.

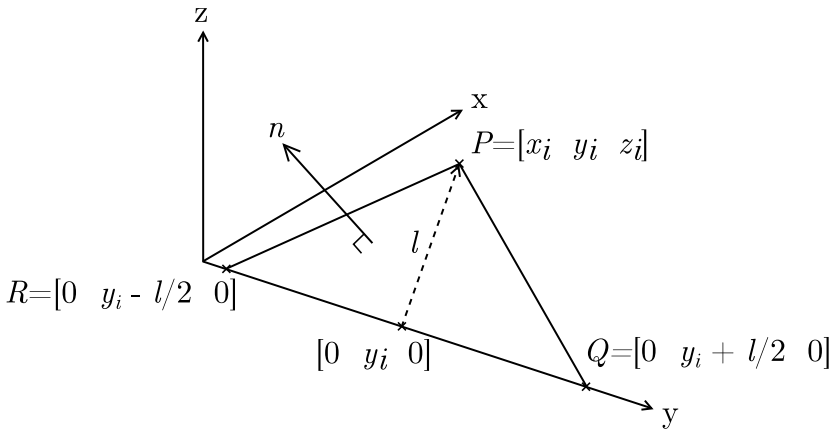


Figure 4.17: Visualisation of the points P, Q and R used for calculating vertex plane equations.

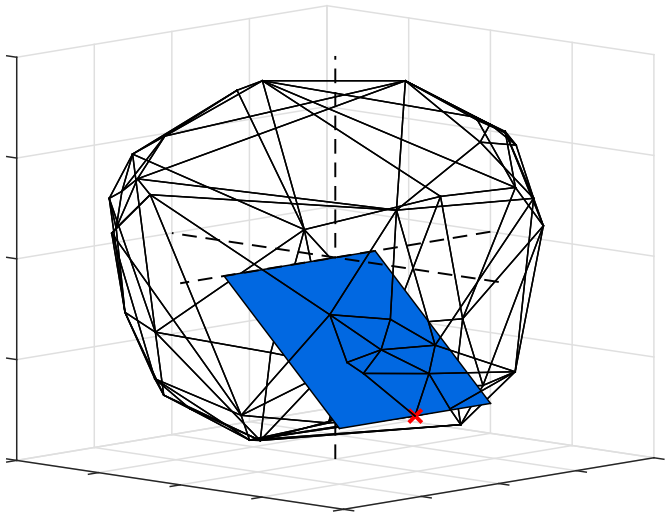


Figure 4.18: Visualisation of a vertex plane for an exemplar vertex of an arbitrary deformed sphere. The red cross indicates the corresponding vertex.

The equations of the face edges are also required. The equation of a line in 3D space is defined by a point P_0 and a direction vector $\boldsymbol{v}$:

$$P_0 = (x_0, y_0, z_0) \quad \text{and} \quad \boldsymbol{v} = [a, b, c] = P_1 - P_0 \tag{4.27}$$

where P_0 and P_1 are two points on the line; in this case the edge vertices. In standard form the equation is:

$$\frac{x - x_0}{a} = \frac{y - y_0}{b} = \frac{z - z_0}{c} \tag{4.28}$$

In parametric form the equation is:

$$\begin{aligned} x &= x_0 + at \\ y &= y_0 + bt \\ z &= z_0 + ct \end{aligned} \tag{4.29}$$

Or more concisely:

$$P(t) = P_0 + \boldsymbol{v}t \tag{4.30}$$

Once the equation of the plane and the line have been calculated, the dot product of the line direction vector and the plane normal ($\boldsymbol{v} \cdot \boldsymbol{n}$) is calculated. If it is nonzero then the line and plane are non-parallel and intersect at some point $P(t_i)$. To calculate t_i the parametric descriptions of the line are substituted into the equation of the plane:

$$a_p x + b_p y + c_p z + d_p = 0 \tag{4.31}$$

where:

$$\begin{aligned} x &= x_0 + a_l t_i \\ y &= y_0 + b_l t_i \\ z &= z_0 + c_l t_i \end{aligned} \tag{4.32}$$

and (x_0 , y_0 , z_0) is a point P_0 on the line, (a_l , b_l , c_l) are given by the line direction vector $\boldsymbol{v}$, (a_p , b_p , c_p) are given by the plane normal $\boldsymbol{n}$ and d_p is given by $-\boldsymbol{n} \cdot V_0$ where V_0 is a point on the plane. This gives the equation:

$$a_p(x_0 + a_l t_i) + b_p(y_0 + b_l t_i) + c_p(z_0 + c_l t_i) + d_p = 0 \tag{4.33}$$

and multiplying out gives:

$$a_p x_0 + a_p a_l t_i + b_p y_0 + b_p b_l t_i + c_p z_0 + c_p c_l t_i + d_p = 0 \tag{4.34}$$

Then grouping the t_i terms gives:

$$t_i(a_p a_l + b_p b_l + c_p c_l) + a_p x_0 + b_p y_0 + c_p z_0 + d_p = 0 \tag{4.35}$$

and rearranging gives:

$$t_i = \frac{-(a_p x_0 + b_p y_0 + c_p z_0 + d_p)}{a_p a_l + b_p b_l + c_p c_l} \tag{4.36}$$

or more concisely:

$$t_i = \frac{-(\boldsymbol{n} \cdot P_0 - \boldsymbol{n} \cdot V_0)}{\boldsymbol{n} \cdot \boldsymbol{v}} = \frac{\boldsymbol{n} \cdot (V_0 - P_0)}{\boldsymbol{n} \cdot \boldsymbol{v}} \tag{4.37}$$

Then t_i can be substituted back into equation 4.30 to find the point of intersection:

$$P(t_i) = P_0 + \boldsymbol{v} t_i \tag{4.38}$$

Finally to check whether the intersection point lies on the line segment between points P_0 and P_1 the point $P(t_i)$ is put back into equation 4.30 and solved for t:

$$t = \frac{P(t_i) - P_0}{\boldsymbol{v}} \tag{4.39}$$

If $0 \leq t \leq 1$ then the point $P(t_i)$ lies between points P_0 and P_1. If $t < 0$ then $P(t_i)$ lies beyond P_0 whilst if $t > 1$ then $P(t_i)$ lies beyond P_1.

Once the distances to all vertices have been calculated they are normalised into the range $0 \leq \theta \leq \pi$ to give the elevation angle for mapping to the sphere.

4.4.3 Mapping of low resolution SYMARE database mesh

For speed and simplicity the PLR mapping approach was initially tested on the lowest resolution (4 kHz) head and ears mesh of the first subject in the SYMARE database (Jin *et al.*, 2013) (figure 4.19).

It was found that there was no suitable origin within the pinnae where there was a direct path along the surface of the mesh to all the vertices: specifically problems were found with finding paths to vertices in the other pinna. This is similar to the problem of "islands" found in the slicing approach used in the original MPA scheme (section 4.1). However, there were suitable origins within the pinnae to map each half of the head independently. Therefore the decision was made to validate the PLR algorithm using this half-head topology. Figure 4.20 shows one half of the head mesh (the left half) with the origin face highlighted in red. Once

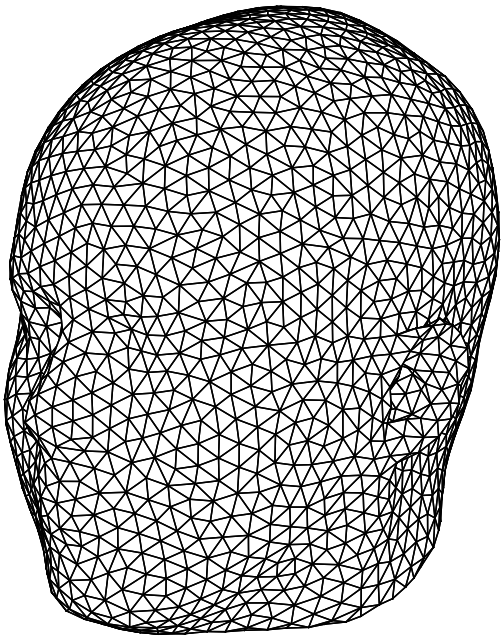


Figure 4.19: Head and ears of the 4 kHz mesh for the first subject in the SYMARE database used to validate the PLR mapping approach.

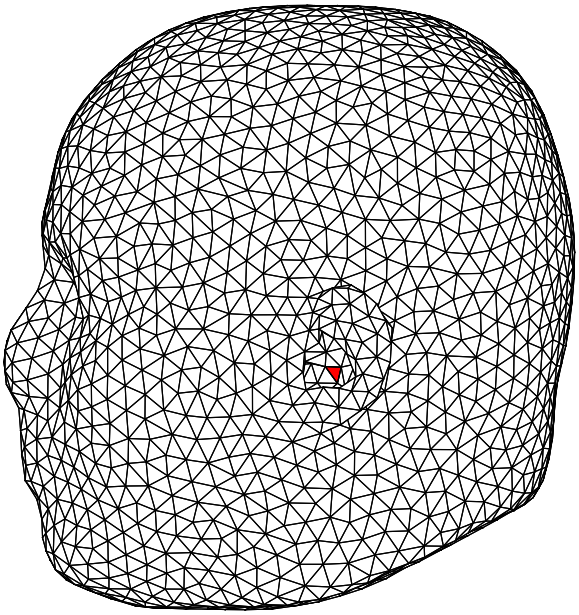
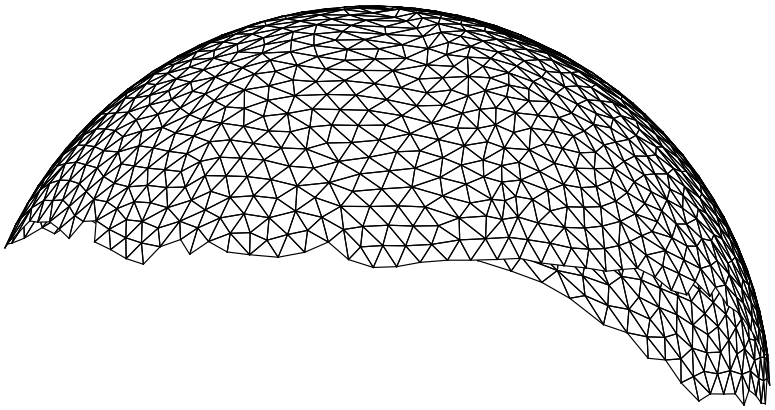
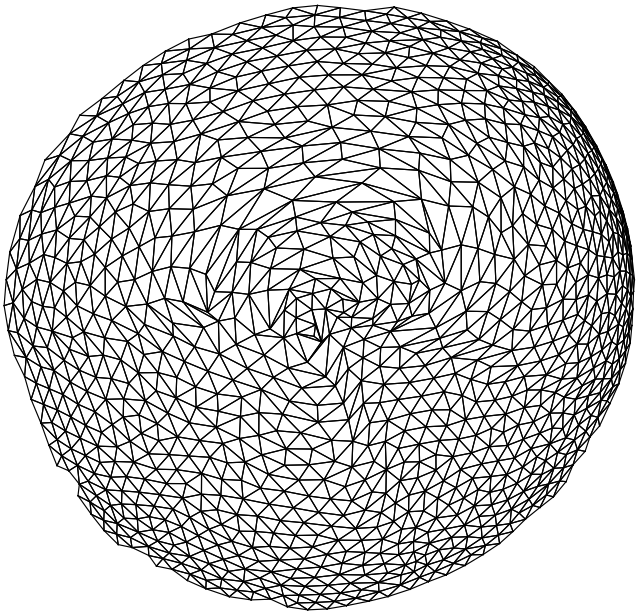


Figure 4.20: Left half head of the 4 kHz mesh for the first subject in the SYMARE database, with origin face highlighted.



(a)



(b)

Figure 4.21: 4 kHz half-head mesh for the first subject in the SYMARE database after mapping to a hemisphere using the PLR mapping approach. (a) shows the variable vertex distances from the pinna origin at the top of the sphere. (b) is the same sphere as (a) but rotated to show the distortions in the pinna area.

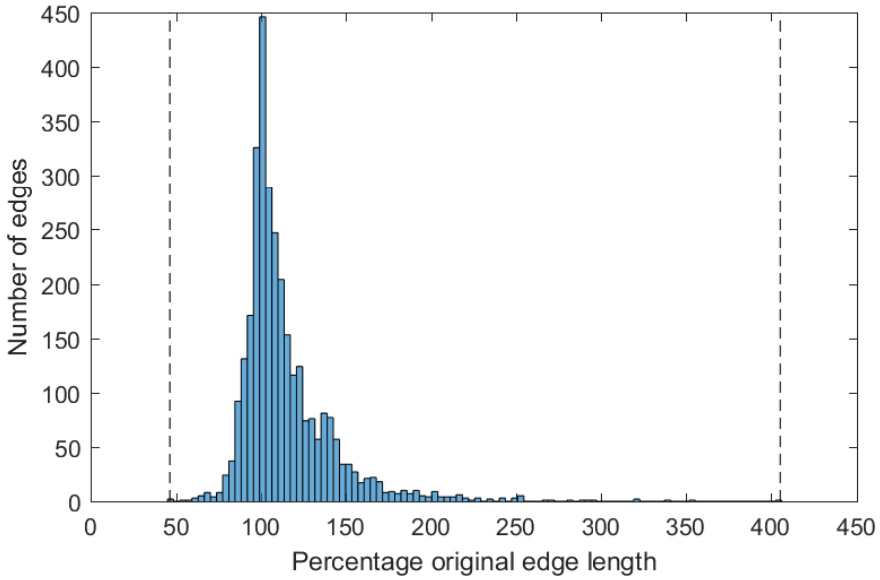


Figure 4.22: Distribution of edge length distortions for the 4 kHz half-head mesh for the first subject in the SYMARE database after mapping to a hemisphere using PLR mapping. The dashed vertical lines show the extents of the distribution.

the origin face had been selected, the mesh was shifted so that the centre of the origin face lay on the y-axis and the vertices of the half head were mapped as outlined in sections 4.4.1 and 4.4.2, but with the elevation angle restricted to $0 \leq \theta \leq \frac{\pi}{2}$ in order to map to just the upper hemisphere. The results of this mapping are shown in figure 4.21.

In figure 4.21a the large variation in surface path lengths is demonstrated by the irregular lower boundary of the mesh on the upper hemisphere and the problems that this introduces can be seen visually in figure 4.21b. Rather than the approximately equilateral triangles of the original head mesh, the faces on the sphere are heavily distorted; especially within the pinna, where the morphology changes rapidly. This results in distortion of the mesh edge lengths on the sphere.

The distribution of edge length distortions for a SYMARE database 4 kHz mesh after mapping to the sphere using PLR is shown in figure 4.22. The ratio between the largest length distortion factor and the smallest length distortion factor is 8, i.e. for the radius of sphere used some of the edges are half their original length, whilst others are more than four times their original length. If the spherical harmonic deformations were to be applied to the spherical mesh in figure 4.21

the harmonic deformations would likewise be heavily distorted and would require many more harmonics to achieve an acceptable resolution of deformations in the pinnae. Therefore a method of "force minimisation" was developed to reduce the distortion of edges/faces on the sphere.

4.4.4 Force minimisation

The second step in PLR is a "force minimisation" optimisation step in order to minimise distortion of the spherical head mesh. Force minimisation treats each distorted edge connected to a given vertex as a force acting on it. The magnitude of the force for each connected edge depends on how much the edge is distorted relative to its original length on the head mesh. The direction of the force acting on the vertex is then simply the sum of the contributing forces from each edge and the vertex is moved so as to reduce this force. This is repeated iteratively to eventually minimise the edge length distortions. The order in which the vertices are moved is randomised for each iteration to reduce the likelihood of reaching a local minimum.

Algorithm 4.2 "Force minimisation" algorithm used to reduce the distortion of edge lengths on the sphere.

```
1: for number of iterations do
2:     generate random permutation of vertex indices
3:     for all entries in list of vertex indices do
4:         find vertices connected to current vertex
5:         calculate edge vectors connecting current vertex to those vertices
6:         calculate weights for edge vectors
7:         convert weights to decibel values
8:         calculate force as sum of weighted edge vectors
9:         convert force to surface tangential
10:        scale force
11:        move current vertex in direction of force
12:        project back onto surface of sphere
13:        check edge lengths
14:        check for edge crossings
15:        check angles
16:    end for
17: end for
```

The algorithm is given in more detail in algorithm 4.2. The weights w_i for each edge vector (line 6) are calculated as the ratio of the edge length on the sphere to the edge length on the head:

$$w_i = \frac{|\boldsymbol{e}_{i(sphere)}|}{|\boldsymbol{e}_{i(head)}|} \tag{4.40}$$

where $\boldsymbol{e}_{i(sphere)}$ is the edge vector on the sphere and $\boldsymbol{e}_{i(head)}$ is the vector of the corresponding edge on the head. These weights are then converted to decibel values:

$$w_i = 20 \log_{10}(w_i) \tag{4.41}$$

This means that an edge that is 50% its original length and an edge that is 200% its original length will have equal weights but with opposite signs, i.e. the first edge will repel the vertex, whilst the second edge will attract the vertex.

Once the edge vectors' weights have been calculated they are weighted and summed to generate the "force" $\boldsymbol{f}$ acting on the vertex:

$$\boldsymbol{f} = \sum_i w_i \boldsymbol{e}_i \tag{4.42}$$

Whilst it is probably safe to assume that the resolution of the sphere is such that the faces connected to a given vertex are approximately planar, and as such the summed force $\boldsymbol{f}$ should be approximately tangential to the surface of the sphere, the force is then explicitly converted to act tangentially to the surface of the sphere (line 9 in algorithm 4.2). This is done by first calculating the normal component $\boldsymbol{f}_\perp$ of the force:

$$\boldsymbol{f}_\perp = (\boldsymbol{f} \cdot \boldsymbol{n})\boldsymbol{n} \tag{4.43}$$

where $\boldsymbol{n}$ is the vertex normal, and then subtracting the normal component from

the total force to calculate the tangential component $\boldsymbol{f}_{\parallel}$:

$$\begin{aligned}\boldsymbol{f}_{\parallel} &= \boldsymbol{f} - \boldsymbol{f}_{\perp} \\ &= \boldsymbol{f} - (\boldsymbol{f} \cdot \boldsymbol{n})\boldsymbol{n}\end{aligned} \tag{4.44}$$

Figure 4.23 shows a visual example of how the forces acting on a particular vertex are calculated. This vertex (bold 'x') is connected to five other vertices ('x's) via five edges (dotted lines). The five other vertices are connected one-to-another by five outer edges (dashed lines) to make up five faces (dotted and dashed lines). The distortion of each edge is indicated next to each connected vertex. The coloured arrows show the force exacted by each edge on the vertex in question. The forces for both shortened and lengthened edges act along the direction of the edge, but for edges that are over 100% their original length the force attract

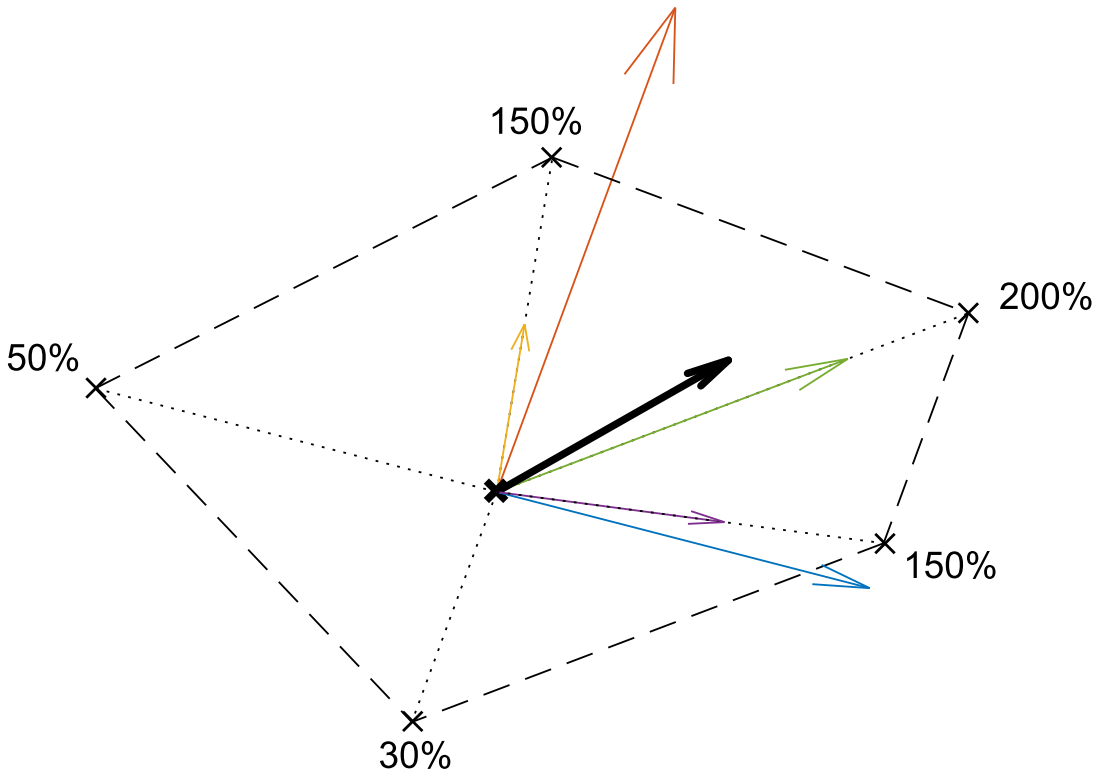


Figure 4.23: Visual example of "forces" acting on a vertex. The coloured arrows show the forces associated with each individual connected vertex and the bold black arrow is their sum. The forces for edges more than 100% their original length act along the edge so as to attract the vertex, whilst the forces for edges less than 100% their original length act in the opposite direction.

the vertex, whilst forces for those edges that are less than 100% their original length act in the opposite direction, i.e. repel the vertex. The bold black arrow indicates the sum $\boldsymbol{f}_{\parallel}$ of the, in this case five, forces.

The tangential force $\boldsymbol{f}_{\parallel}$ gives the direction in which to move the vertex but it needs to be suitably scaled (line 10 in algorithm 4.2) before moving the vertex. This is done by first converting the tangential force $\boldsymbol{f}_{\parallel}$ to a unit vector:

$$\hat{\boldsymbol{f}}_{\parallel} = \frac{\boldsymbol{f}_{\parallel}}{|\boldsymbol{f}_{\parallel}|} \tag{4.45}$$

which can then be multiplied by the desired scaling factor d to change the distance the vertex should be moved:

$$\boldsymbol{d} = d\hat{\boldsymbol{f}}_{\parallel} \tag{4.46}$$

The initial distance (step size) was set to $1/10^{\text{th}}$ of the average length of all the edges connected to the vertex:

$$d = \frac{1}{10n}\left(\sum_{i=1}^{n} |\boldsymbol{e}_i|\right) \tag{4.47}$$

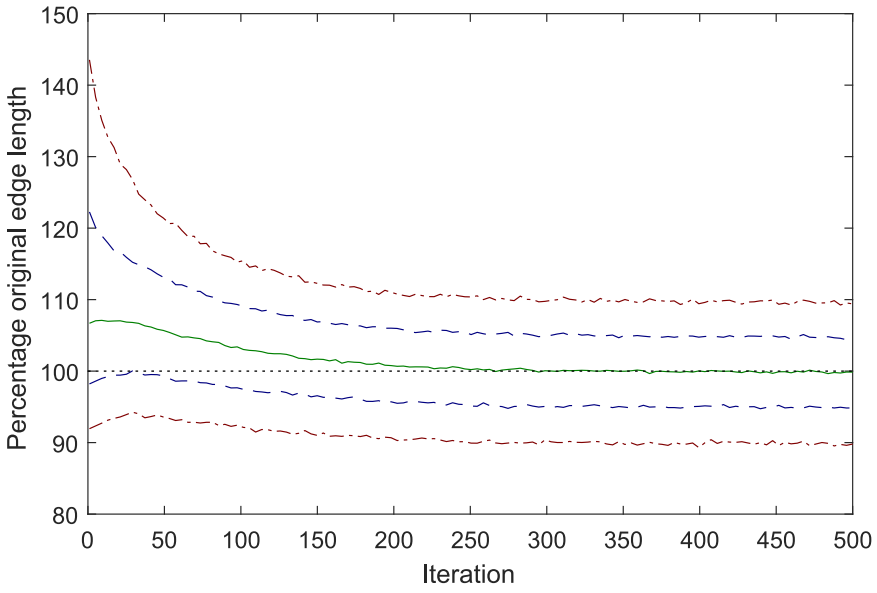


Figure 4.24: Percentiles of the edge length distortion distribution after 500 iterations of PLR at the initial step size. Solid green line: 50^{th} percentile (median), dashed blue lines: $25^{\text{th}}/75^{\text{th}}$ percentiles, red dash-dot lines: $10^{\text{th}}/90^{\text{th}}$ percentiles.

and 500 iterations of PLR were applied.

Figure 4.24 reports the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the edge length distortions over the 500 iterations of PLR at the initial step size. It can be seen that all have levelled out by the 500^{th} iteration. Therefore the step size was decreased to $1/100^{\text{th}}$ of the average length of the edges connected to the vertex and another 500 iterations run. The percentiles across the 500 iteration for this are shown in figure 4.25.

It can be seen that decreasing the step size has little effect on the median (50^{th} percentile) but initially leads to an improvement in all the others after which they plateau. This improvement can be seen in the distributions of edge length distortions. Figure 4.26 compares the distributions of edge length distortions before any PLR (green dash-dot lines), after 500 iterations at the initial step size (red dashed lines) and after a further 500 iterations at the decreased step size (blue solid lines). Whilst the limits of the distributions do not change much after the additional iterations at the smaller step size, a further improvement to the overall shape of the distribution is suggested by the doubling in the number of edges that are the correct length (100%).

Decreasing the step size further to $1/1000^{\text{th}}$ of the average length of the edges connected to the vertex and running another 500 iterations was also tested, however as figure 4.27 shows it resulted in very little improvement in the edge length distortions.

After moving each vertex it was necessary to carry out a series of checks (lines 13–15 in algorithm 4.2). If any of the checks were failed then the vertex was moved back to its original position and the algorithm moved onto the next vertex. The first of these checks was to ensure that no edge's length was ever decreased from more than 50% its original length to less than 50% or increased from less than 200% to more than 200% by the PLR algorithm. This was to stop certain edges being "sacrificed" to improve other edges, i.e. improving the variance at the expense of the maximum/minimum. The limits of 50% and 200% were chosen

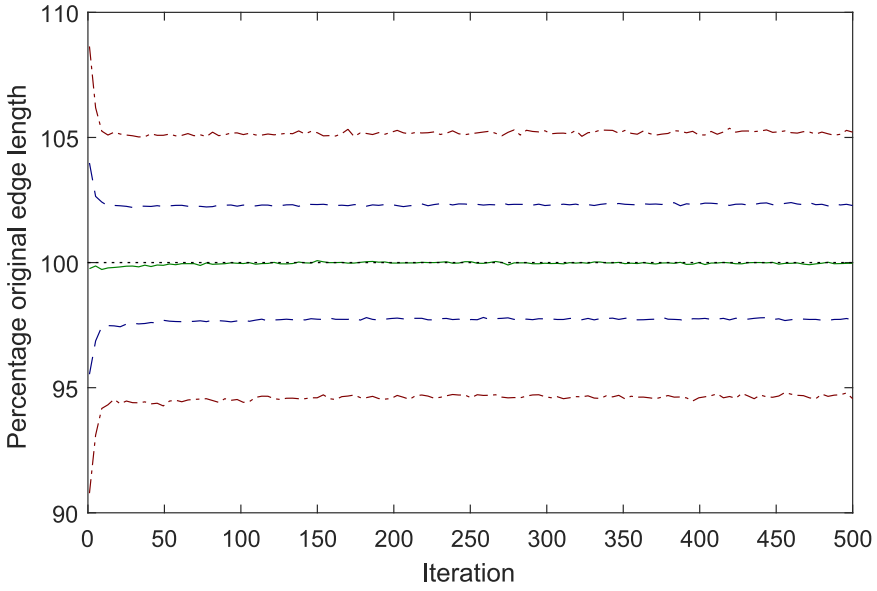


Figure 4.25: Percentiles of the edge length distortion distribution after 500 iterations of PLR at a reduced step size. Solid green line: 50^{th} percentile (median), dashed blue lines: $25^{\text{th}}/75^{\text{th}}$ percentiles, red dash-dot lines: $10^{\text{th}}/90^{\text{th}}$ percentiles.

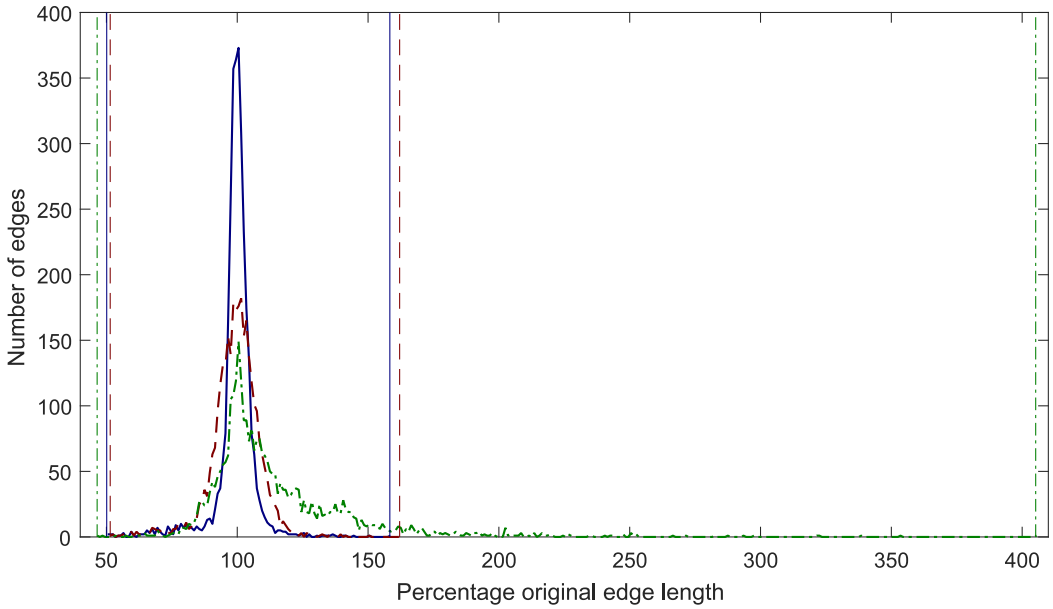


Figure 4.26: Comparison of the distributions of edge length distortions before and after application of PLR to the half-head mesh on the hemisphere. Dash-dot green line: original sphere before PLR, dashed red line: sphere after 500 iterations of PLR at initial step size, solid blue line: sphere after an additional 500 iterations of PLR at reduced step size. Vertical lines show maximum/minimum limits of distributions.

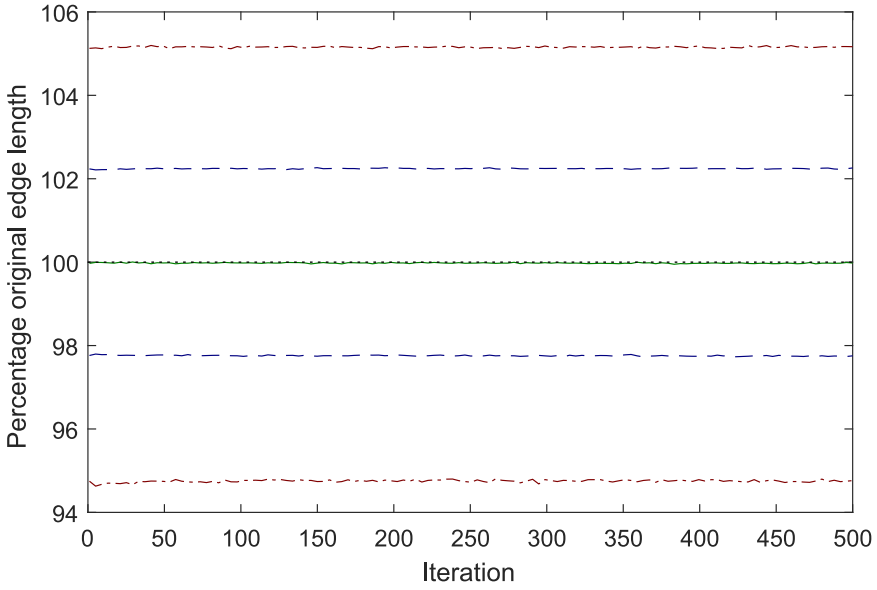


Figure 4.27: Percentiles of the edge length distortion distribution after 500 iterations of PLR at a further reduced step size. Solid green line: 50^{th} percentile (median), dashed blue lines: $25^{\text{th}}/75^{\text{th}}$ percentiles, red dash-dot lines: $10^{\text{th}}/90^{\text{th}}$ percentiles.

based on the fact that BEM-ready meshes generally adhere to a $\lambda/6$ rule and therefore can, at the absolute limit, support a minimum to maximum edge length ratio of 1:3 without violating the $\lambda/2$ requirement for avoiding aliasing.

The second check was to ensure that the vertex never crossed any of the edges ringing it (dashed lines in figure 4.23). To perform this check, after moving the vertex, the normal of each face connected to the vertex was calculated as per equation 4.24. Then the angle between the face normals and the vertex normal was calculated by taking their dot product. If the vertex lies within the bounds of the outer edges then the angles between the vertex normal and the face normals must all be between 0° and 90°. An example using the same vertices as figure 4.23 is shown in figures 4.28 and 4.29.

Figure 4.28 shows the vertex normal (bold black arrow) and the face normals (thin coloured arrows) before movement of the vertex (bold black 'x'): all the normals face in the same direction. Figure 4.29 shows the same set of vertices but after the vertex in question has crossed one of the outer edges. As can be seen one of the face normals now faces in the opposite direction to both the vertex

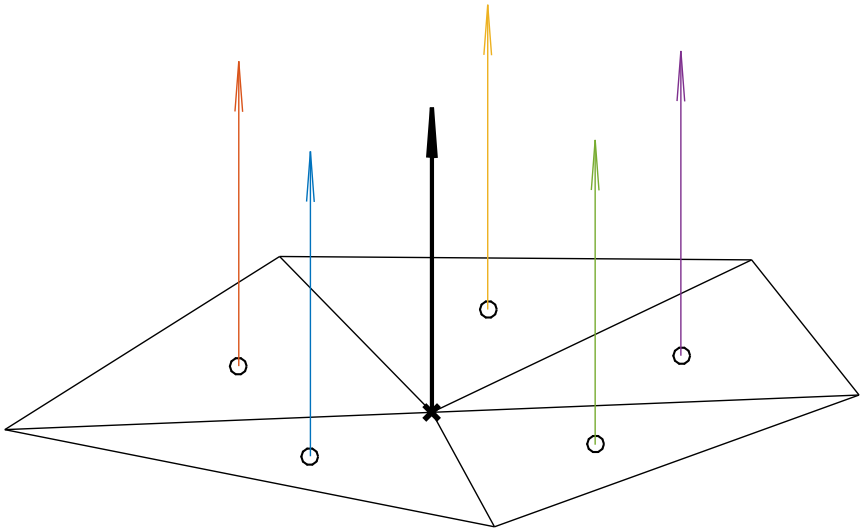


Figure 4.28: Example of vertex normal (bold black arrow) and face normals (thin coloured arrows) before movement of vertex: all normals point in the same direction. Circles indicate the centre of each face.

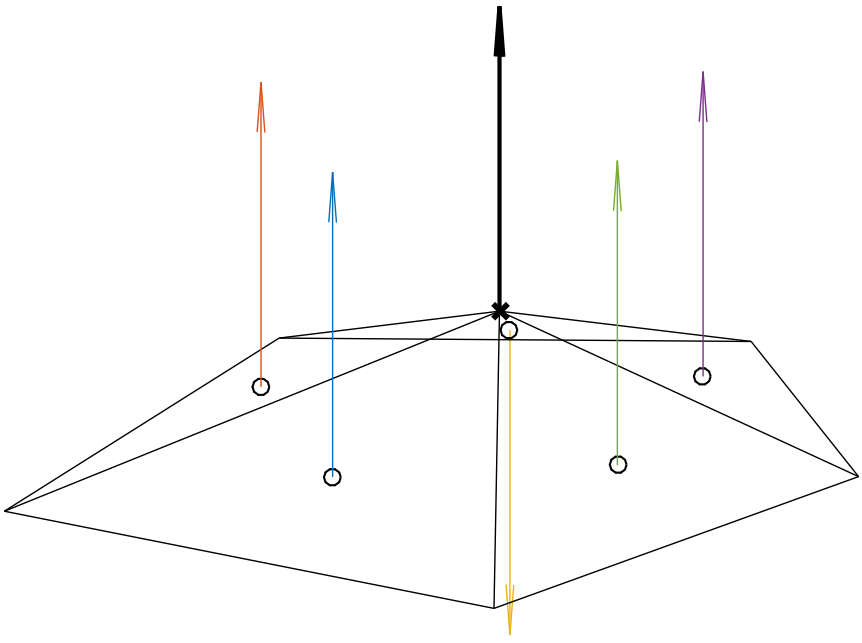


Figure 4.29: Example of vertex normal (bold black arrow) and face normals (thin coloured arrows) after vertex crosses one of the outer edges: one face normal points in the opposite direction to the other normals. Circles indicate centre of each face.

normal and the other face normals. Therefore its dot product with the vertex normal will be outside the range 0° to 90° and a crossing will be detected.

The final check was to ensure that the faces on the hemisphere did not become too distorted, as this could lead to an unwanted artifact similar to the sheer effect described in section 4.1. To ensure this, after moving the vertex, the angles between all the edges around the vertex in question were calculated according to:

$$\cos\theta = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{|\boldsymbol{v}_1| \cdot |\cdot \boldsymbol{v}_2|} \tag{4.48}$$

where $\boldsymbol{v}_1$ is the vector representing one edge and $\boldsymbol{v}_2$ is the vector representing the other. If any of the angles were less than 15° (and had not started at less than 15°) then the vertex in question was moved back to its original position. It may seem that this check is superfluous given the check for edge crossing. However, due to the fact that the vertex and surrounding faces are not perfectly planar, a very small angle is possible without failing the edge crossing check and likewise the edge crossing check can be violated, whilst the angle check is not.

4.4.5 Limitations of PLR method

After verification of PLR on the low resolution (4 kHz) SYMARE database mesh it was tested on the high resolution (20 kHz) KEMAR mesh. Again the head mesh was split into two halves and each half mapped separately. To find a suitable origin face, every face within a certain region of the pinnae was tested to verify whether all other vertices on the half head could be successfully mapped. The search area is highlighted in red in figure 4.30. A number of suitable origin faces were found and one was arbitrarily selected.

Figure 4.31 shows the KEMAR head mesh after initial mapping to the hemisphere. It can be seen that the mesh has successfully mapped to the hemisphere, however if the hemisphere mesh is examined closer it can be seen that some of the mesh edges already overlap one another (red highlighted face in figure 4.32).

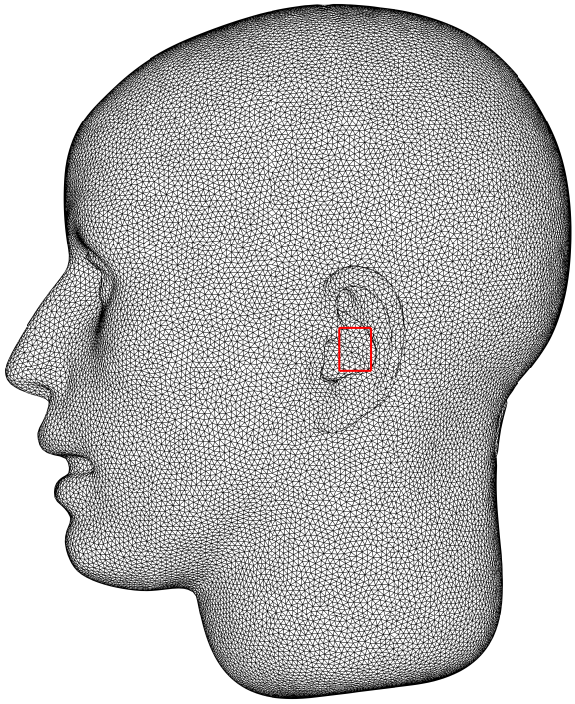
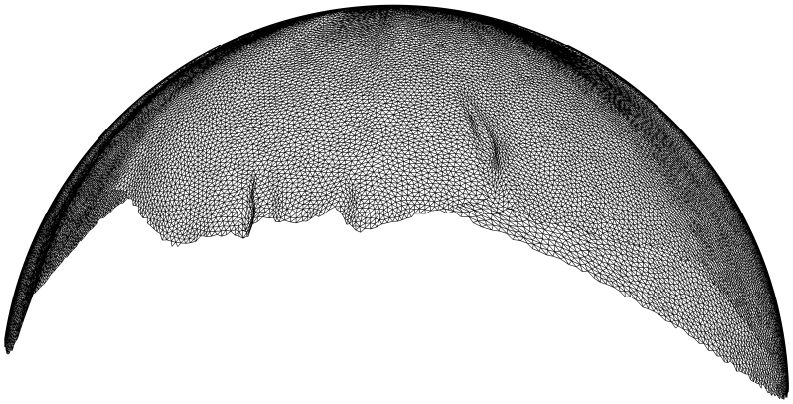


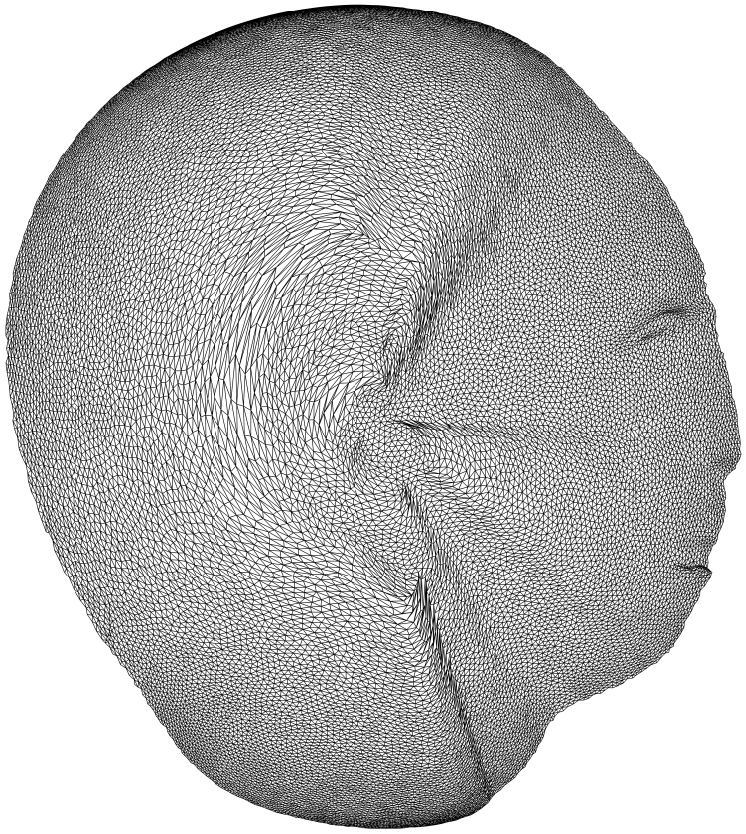
Figure 4.30: Pinna area of KEMAR mesh used to search for suitable origin for PLR (highlighted in red).

Therefore additional processing would be required before force minimisation.

Furthermore the speed of the PLR algorithm made processing the high resolution mesh unviable; especially given the need for successive reduction of the step size. One iteration of the SYMARE database 4 kHz mesh took on average one second whilst one iteration of the KEMAR mesh took on average 42 seconds. Therefore the path length relaxation (PLR) method was rejected in favour of the optimised projection (OP) method detailed in the next section.



(a)



(b)

Figure 4.31: KEMAR head mesh after initial mapping to the hemisphere using PLR. (b) same as (a) but rotated to show the pinna area.

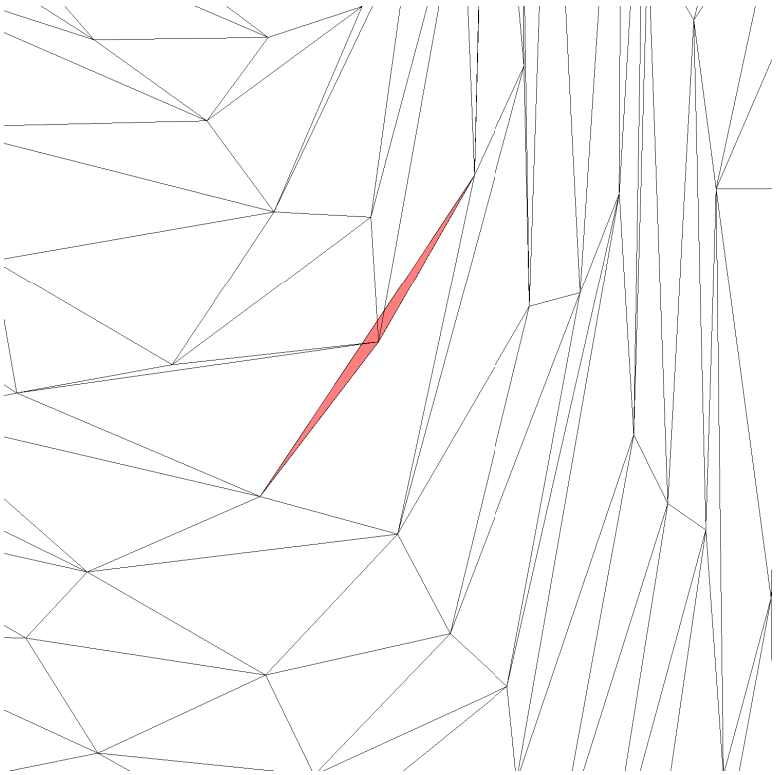


Figure 4.32: KEMAR head mesh after initial mapping to the hemisphere using PLR with overlapped face highlighted (red).

4.5 Optimised projection (OP) mapping method

Due to the limitations of path length relaxation (PLR) identified in the previous section, a series of novel methods of mapping and optimisation, jointly termed optimised projection (OP), were developed. This section reports the development of these methods. The mapping of the head mesh onto the sphere is based on prior methods of projection and iterative spatial averaging to ensure correct orientation of the faces as reported by Peyré (2011a,b,c) and in Numerical Tours (2017a) (section 4.5.1). The method of spatial averaging is then adapted to reduce the edge length distortions on the sphere (section 4.5.3). Section 4.5.4 reports the application of a novel approach, ellipsoidal pinna emphasis (EPE), to increase the significance of the pinnae regions on the sphere and finally EPE is refined into dynamic pinna emphasis (DPE) (section 4.5.5), another novel technique, where emphasis is varied on an edge-by-edge basis to optimally approach a more uniform acoustic wavelength for a given deformation wavelength.

4.5.1 Projection onto the sphere and initial spatial averaging

The first step in OP is to project the head mesh onto a sphere. For each vertex $\mathbf{v}_i$ in the mesh, this is done by dividing its coordinates by its norm and multiplying by the radius r of the sphere:

$$\mathbf{v}_i = r\frac{\mathbf{v}_i}{|\mathbf{v}_i|} \tag{4.49}$$

where r is calculated as the average radius, in spherical coordinates, of all the head mesh vertices.

However, since the head mesh is not genus 0 (i.e. there is not necessarily a direct path from the origin to every vertex of the head mesh without passing through the surface of the mesh) some of the faces are inverted during projection onto the sphere. To check whether a face in inverted, the face normal is calculated and

the dot product formed with the vector from the origin to the face centre. If the dot product is negative then the face has become inverted (this assumes that the mesh was correctly oriented to begin with (section 4.3), which can be verified by ensuring that the two times that each edge appears in the list of edges $\mathcal{E}$, the vertex indices are in opposite orders (equation 4.19)).

If any of the dot products are negative then a smoothing procedure is employed to gradually unfold the surface and flatten it onto the sphere. This is achieved by iteratively recalculating the position of each vertex as the average position of its connected neighbours. This process continues until none of the faces are inverted, signalling that the unfolding process is complete. An algorithm for accomplishing unfolding is described below. An $n \times n$ adjacency matrix A is defined as:

$$A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{4.50}$$

where n is the number of vertices, and $1 \geq i \leq n$ and $1 \geq j \leq n$. Therefore $A_{i,j}$ is only unity if vertices i and j are connected, otherwise it is zero. Since most entries are zero A can be stored as a sparse matrix, which speeds up computation. A is then normalised so that the sum of each row is one. To do this, the connectivity weight vector d is calculated where:

$$d_i = \sum_j A_{i,j} \tag{4.51}$$

i.e. d_i is the number of edges (and thus the number of other vertices) connected to vertex i. This is stored in another sparse, diagonal matrix $D = \text{diag}_i(d_i)$ and the normalised adjacency matrix $\hat{A}$ is then calculated as:

$$\hat{A}_{i,j} = \frac{1}{d_i} A_{i,j} \tag{4.52}$$

or in matrix formulation:

$$\hat{A} = AD^{-1} \tag{4.53}$$

Therefore multiplying the vertex matrix V by the normalised adjacency matrix:

$$V = V\hat{A} \tag{4.54}$$

is equivalent to calculating each vertex's position as the average of the positions of the other vertices connected to it.

To work towards having the same orientation for all the faces on the sphere (i.e. unfolding) an iterative approach is used:

1: *Project vertices onto unit sphere*
2: *Calculate normalised adjacency matrix*
3: *Calculate face normals*
4: *Calculate vector from origin to each face centre*
5: *Calculate dot products of face normals and their corresponding "face centre" vector*
6: **while** *Any dot products less than zero* **do**
7: *Multiply vertex matrix V by normalised adjacency matrix $\tilde{A}$*
8: *Project vertices onto unit sphere*
9: *Calculate face normals*
10: *Calculate vector from origin to each face centre*
11: *Calculate dot products of face normals and their corresponding "face centre" vector*
12: **end while**

It should be noted that there is no theoretical guarantee that the above will work for all meshes (Numerical Tours, 2017b), but the algorithm was tested on all head and pinnae, BEM-compatible meshes (61 subjects at 5 resolutions, giving 305 in all) in the SYMARE database and resulted in no inverted faces. This provides a high degree of confidence that it is suitable for head meshes in general, presumably due to their relatively close similarity to the sphere.

Figure 4.33 shows the initial steps of spherical parameterisation. The starting

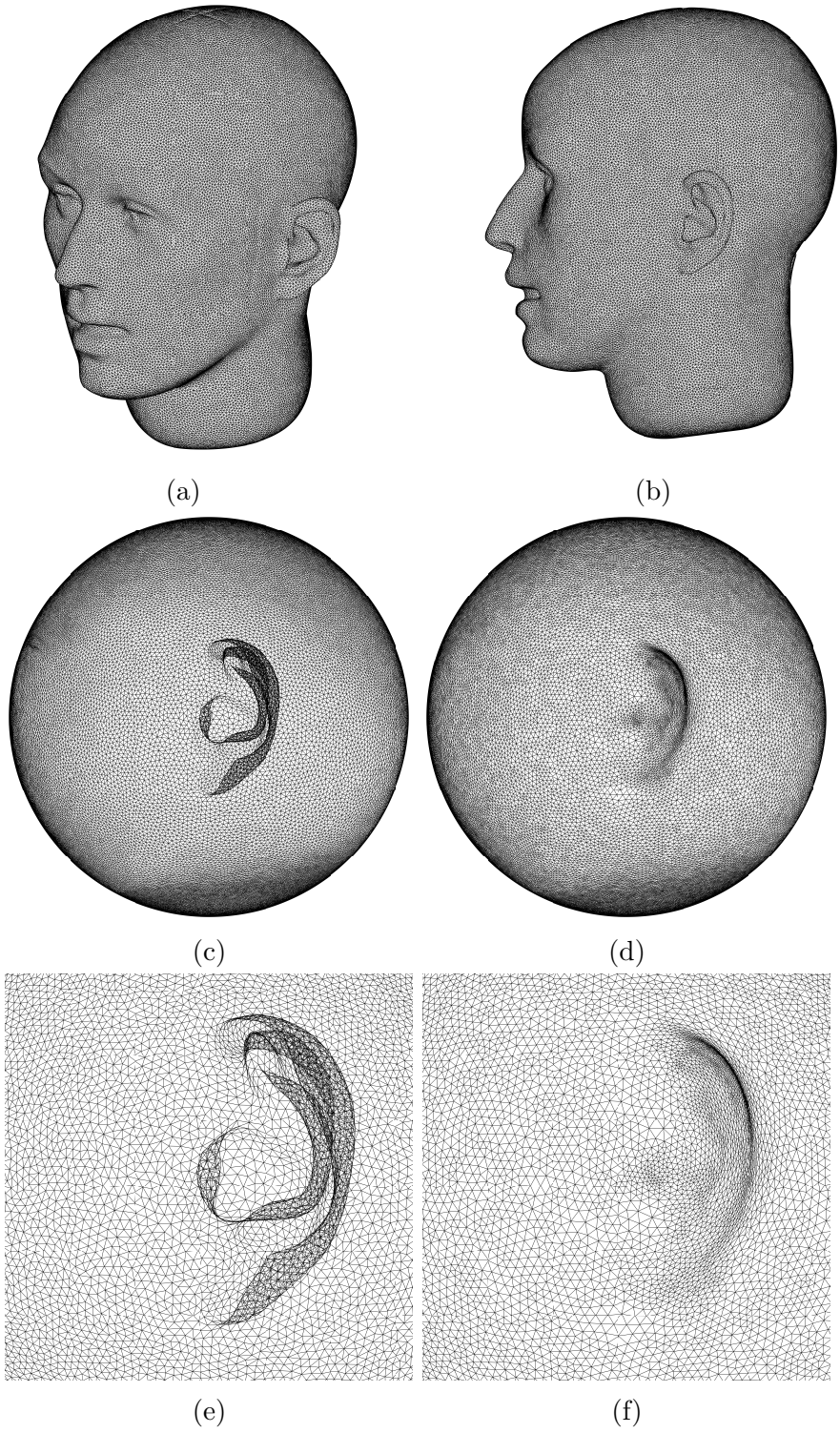


Figure 4.33: Initial spherical parameterisation. (a) initial KEMAR head mesh. (b) initial KEMAR head mesh rotated to focus on pinna. (c) mesh after projection onto the unit sphere. (d) mesh after iterative spatial averaging to unfold the inverted faces. (e) same as (c), but zoomed in on pinna. (f) same as (d), but zoomed in on pinna.

point is the head mesh, viewed from different angles in (a) and (b), which is first projected directly onto a sphere, (c). However, this results in inversion of some of the faces and hence they overlap with other faces on the surface of the sphere, especially in the pinnae (e). Applying the spatial averaging algorithm, described above corrects the orientation of the inverted mesh faces, unfolding them on the surface of the sphere, (d) and (f).

Whilst the initial spatial averaging ensures that no faces are inverted, the edges on the sphere that it produces are, as with PLR (section 4.4), quite distorted; some lengthen to as much as 300% of their original length and some shorten to as little as 3% of their original length (figure 4.34). If measures to reduce this distortion of edge lengths were not taken, a uniform MPA deformation wavelength applied over the spherical mesh would map to a wide variation in deformation wavelengths over the original head mesh. This is undesirable (for an explanation see, for example, section 4.1). Therefore an adaptation to the spatial averaging approach, to reduce the distortions of the sphere edges, was developed.

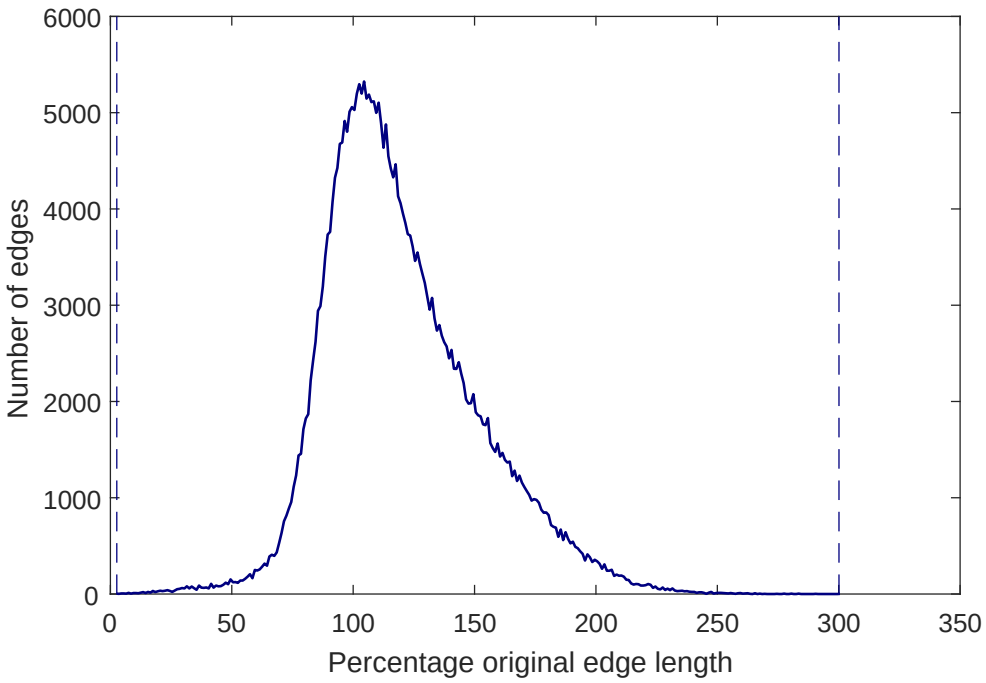


Figure 4.34: Distribution of edge length distortions after initial optimised projection (OP) mapping to the sphere. Some of the edges are as much as 300% their original length, whilst others are less than 3% their original length (the extremes are indicated by the vertical dashed lines).

4.5.2 Metrics for optimisation techniques

The development of any optimisation technique requires a means of evaluating its effectiveness and for comparing its performance with other approaches. In this section a new metric for doing so is described.

In the preceding section, edge length distortion was presented as a measure of performance of the process for mapping the head mesh onto a sphere. Closer study reveals that this metric has certain weaknesses, which will become clearer later in this section. Instead, a new approach was adopted of considering the maximum spatial frequency that can be supported by the spherical mesh when applying deformations to it.

The reader is reminded that an important reason for wishing to map the head mesh onto the surface of a sphere is so that the harmonic deformations required as part of the MPA method can be applied using spherical surface harmonics (section 5.1). The spatial frequency limit of the harmonic deformations that can be supported by the spherical mesh is defined by its longest edge. Specifically, the spherical mesh can support harmonic deformations with a deformation wavelength no shorter than twice the longest edge, otherwise spatial aliasing will occur:

$$\lambda min = 2 \max(|\boldsymbol{e}_{(sphere)}|) \tag{4.55}$$

Furthermore, the most contracted edge on the sphere will suffer the largest reduction in spatial frequency deformation when mapped back to the head mesh. Conversely, the most stretched edge on the spherical mesh will suffer the greatest increase in spatial frequency deformation when mapped back to the head mesh. Whilst the latter is not necessarily problematic, the former is, especially if the reduction in the frequency of spatial deformations means they are unable to fully resolve the features in the acoustic feature map produced by MPA.

For MPA it is necessary to be able to apply harmonic spatial deformations to the head mesh with spatial frequencies equivalent to acoustic wavelengths across the

whole audio frequency range of interest. This gives an upper audio, and hence upper spatial, frequency limit of 16 kHz (section 2.2.4.3) when studying HRTF features. Therefore to assess the suitability of a spherical mesh for deformation by spherical surface harmonics, firstly the minimum valid spatial deformation wavelength on the sphere was calculated as per equation 4.55. Then the length of each edge on the spherical mesh was calculated as a fraction of the longest edge:

$$l_i = \frac{|\boldsymbol{e}_i|}{\max(|\boldsymbol{e}_{(sphere)}|)} \tag{4.56}$$

This was then used to calculate the effective wavelength of λmin on each edge of the head mesh:

$$\lambda min_i = |\boldsymbol{e}_{i(head)}| \left(\frac{360^\circ}{180^\circ l_i} \right) \tag{4.57}$$

This is based on the fact that, as mentioned above, the longest edge in the spherical mesh will have a 180° phase change along it at the minimum valid wavelength. Therefore all the other sphere edges will have less than 180° phase change along them. This, combined with the actual length of the edge on the head mesh gives the shortest valid wavelength for each edge on the head. The shortest valid wavelength can be easily converted to the maximum valid frequency for each edge.

The distribution of maximum valid spatial deformation frequency for the projected and spatially averaged spherical mesh is shown in figure 4.35. It shows that there are still a number of edges that have a maximum valid frequency less than the required 16 kHz and, furthermore, figure 4.36 shows that these edges are concentrated in the pinnae, limiting the resolution achievable by MPA in the region of greatest interest. Hence there is a need for further optimisation of the spherical mesh prior to applying the spherical surface harmonic deformations.

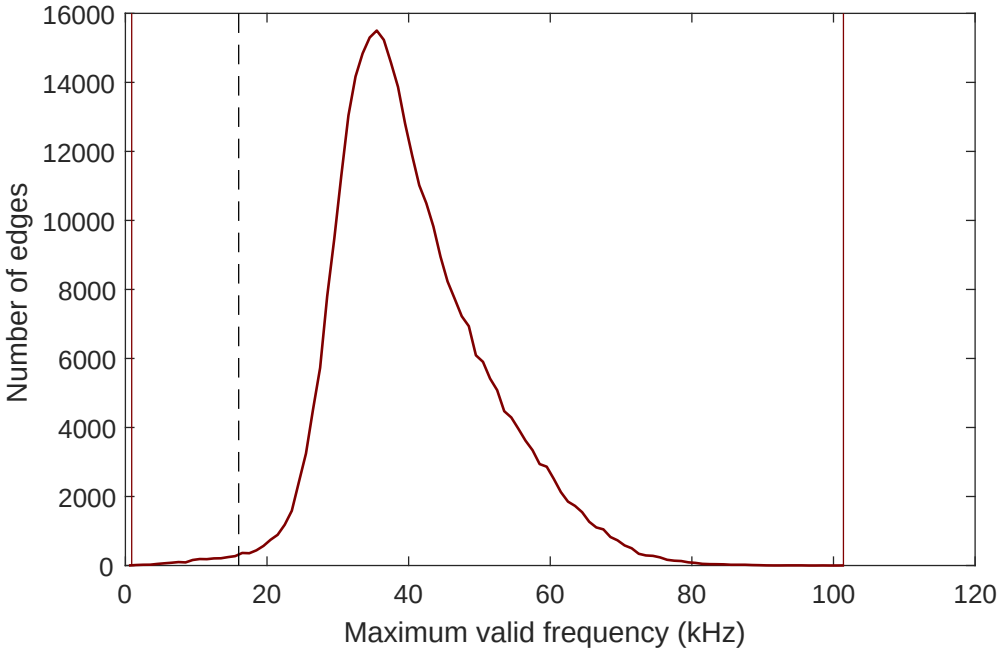
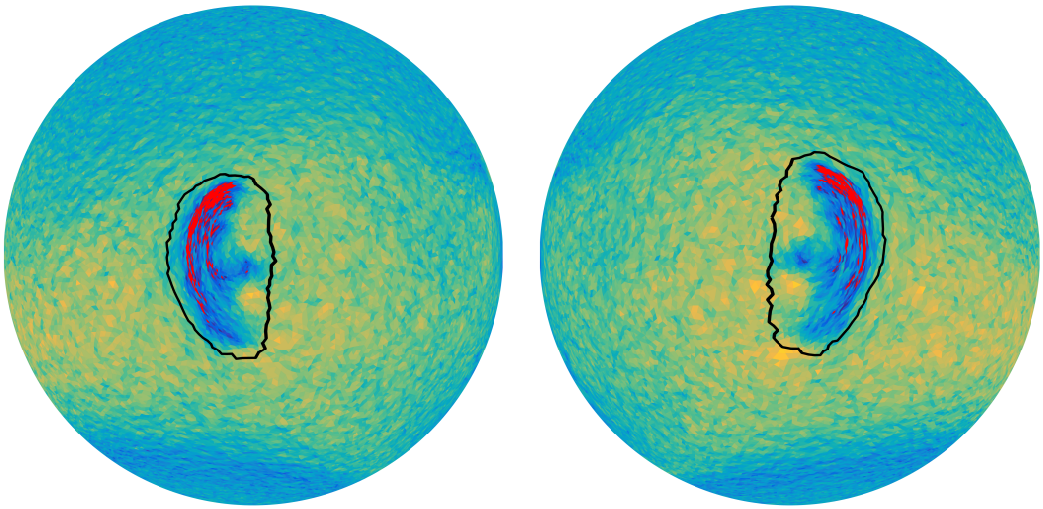


Figure 4.35: Distribution of maximum valid spatial deformation frequency after initial projection and spatial averaging. Solid vertical lines show limits of distribution. Dashed vertical line shows required maximum frequency of 16 kHz. Therefore edges to the left of the dashed line will not have required resolution.



(a) (b)

Figure 4.36: Spherical mesh after initial projection and spatial averaging showing areas of reduced resolution for (a) the right pinna and (b) the left pinnae. Faces highlighted in red have at least one edge that has a maximum valid spatial deformation frequency resolution of less than 16 kHz. All other faces are coloured from 16 kHz (blue) to the maximum valid frequency for the whole sphere (yellow). The bold black outlines show the boundary where each pinna meets the even surface of the head.

4.5.3 Spatial averaging optimisation

The first approach to optimising the spherical mesh was to refine the spatial averaging approach used in the initial generation of the spherical mesh (section 4.5.1). To improve the spatial averaging the adjacency matrix A was redefined as:

$$A_{i,j} = \begin{cases} \frac{|\boldsymbol{e}_{i,j(sphere)}|}{|\boldsymbol{e}_{i,j(head)}|} & \text{if } (i,j) \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases} \tag{4.58}$$

i.e., rather than calculating the position of a vertex as the equally weighted average of the other vertices connected to it, the weight of each of the connected vertices is proportional to the distortion of the edge connecting it to the vertex in

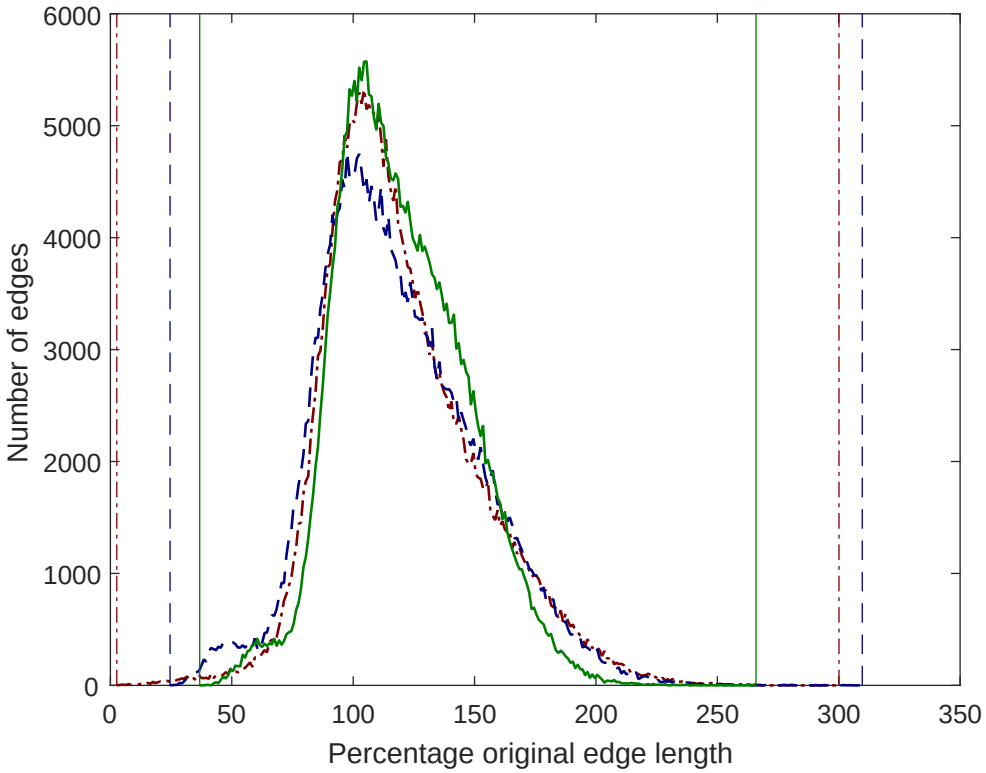


Figure 4.37: Comparison of the distributions of edge length distortions before and after application of spatial averaging with different adjacency matrices. The red dash-dot lines show the distribution after application of the equal weight adjacency matrix until all faces were correctly oriented. The blue dashed lines show the distribution after a further 1000 iterations of the equal weight adjacency matrix. Finally the solid green lines shows the distribution after 1000 iterations of the edge length distortion adjacency matrix. Vertical lines are limits of distributions.

question. Therefore, if the edge is longer than it should be, it is weighted more in the calculation of the average, meaning the vertex in question will move in the direction of that edge, thereby shortening it and reducing the distortion. Whereas, if an edge is shorter than it should be, it is weighted less in the calculation of the average which will lead to the vertex in question shifting away from it, thereby increasing its length, again reducing the distortion.

Figure 4.37 compares the edge length distortion distributions after the initial projection and averaging until all the faces were correctly oriented as outlined in the previous section (red dash-dot lines), after an additional 1000 iterations using the equal weighting adjacency matrix (blue dashed lines) and after 1000 iterations using the edge length distortion adjacency matrix (green solid lines) just described. It can be seen that, whilst the additional iterations of the equal weighting adjacency improve (i.e. shift right) the lower limit of the distribution, they actually worsen (i.e. also shift right) the upper limit. In contrast, the application of the edge length distortion adjacency matrix improves both the upper and lower limits of the distribution (i.e. shifts the lower limit right and the upper limit left). This improvement can also be seen in the resulting spherical meshes as shown in figure 4.38. The equal-weighting adjacency matrix results in "bunching" in the pinna area, whilst using the edge length distortion adjacency matrix results in a more even mesh.

A comparison of the distributions of maximum valid spatial deformation frequency, as per section 4.5.2, before and after the application of the edge length distortion adjacency matrix is given in figure 4.39. The application of the edge length distortion adjacency matrix has significantly reduced the number of edges with a maximum valid frequency of less than 16 kHz. However, there are still a number of edges with a resolution of less than 16 kHz and, as shown in figure 4.40, these are primarily in the pinnae.

The likely cause for this can be seen in figure 4.41, which compares the edge length distortion distributions for the whole spherical mesh and just the pinnae

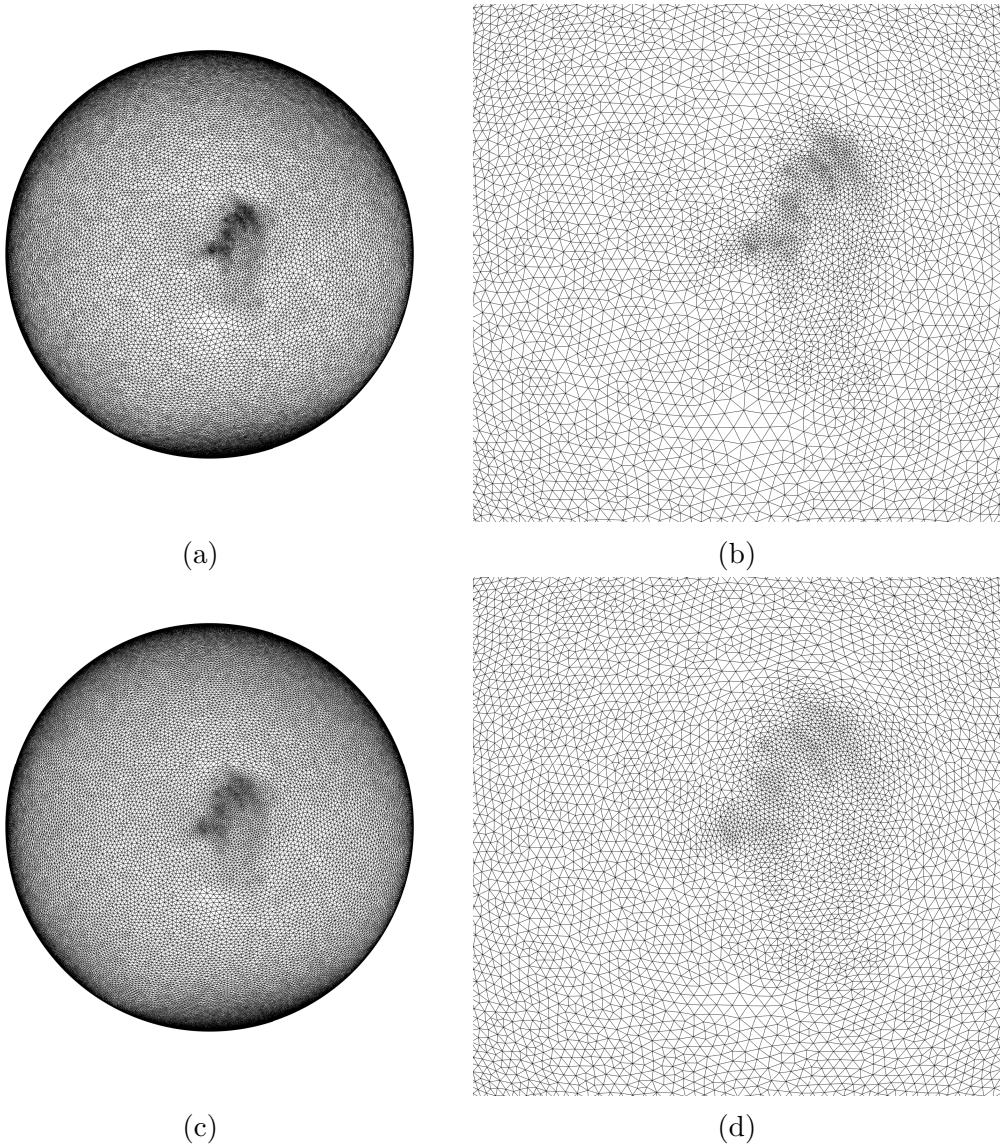


Figure 4.38: Comparison of spherical meshes after application of spatial averaging with different adjacency matrices. (a) and (b) show the sphere after an additional 1000 iterations of the equal weight adjacency matrix. (c) and (d) show the sphere after 1000 iterations of the edge length distortion adjacency matrix. Right hand panels are the same as left but zoomed in on the pinna region.

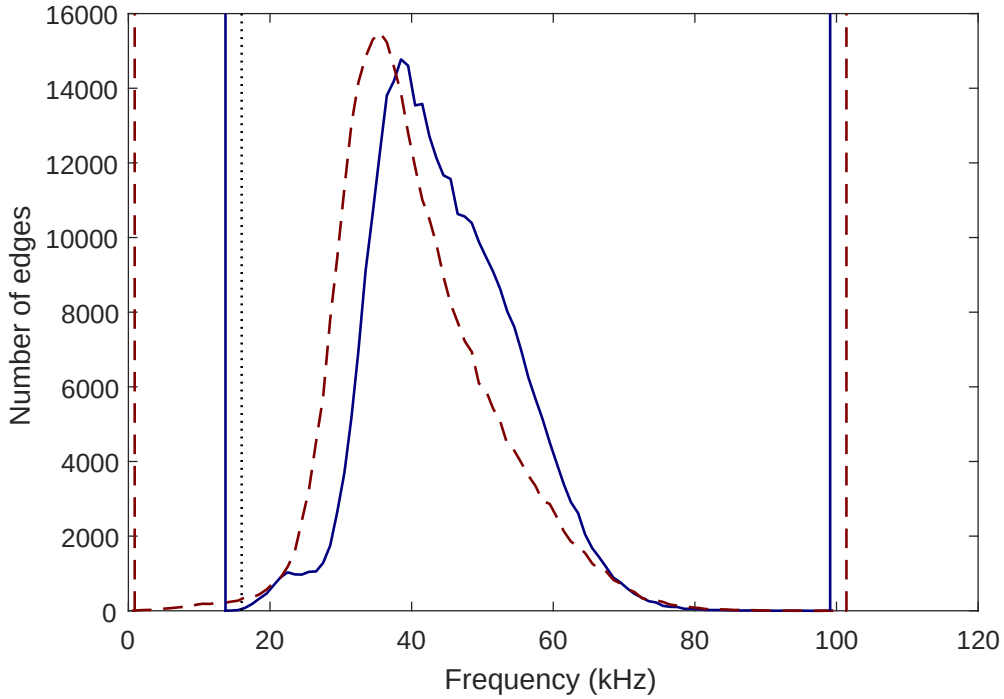
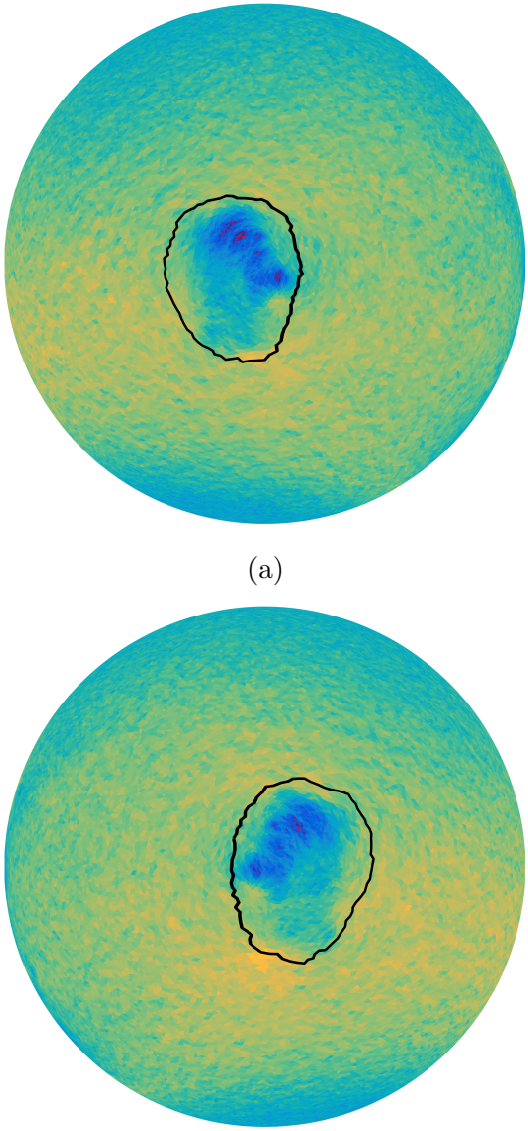


Figure 4.39: Comparison of the distributions of maximum valid spatial deformation frequency before (red dashed lines) and after (blue solid lines) spatial averaging with the edge length distortion adjacency matrix. Solid and dashed vertical lines give limits of the respective distributions. Dotted vertical line shows the required resolution of 16 kHz.

edges. The peak of the distribution for the whole head is around 100%, i.e. the edges are the correct length. However, the peak of the distribution for the pinna edges is around 60% and the whole distribution is skewed towards the lower end for these edges, indicating that they are on average too compressed. To combat this, a method of ellipsoidal optimisation was developed to stretch the edges and faces of the pinnae.



(a)

(b)

Figure 4.40: Spherical mesh after spatial averaging using the edge length distortion adjacency matrix for (a) the right pinna and (b) the left pinna. Faces highlighted red have at least one edge that has a maximum valid spatial deformation frequency resolution of less than 16 kHz. All other faces are coloured from 16 kHz (green) to the maximum valid frequency for the whole sphere (blue). The bold green outlines show the boundary where each pinna meets the even surface of the head.

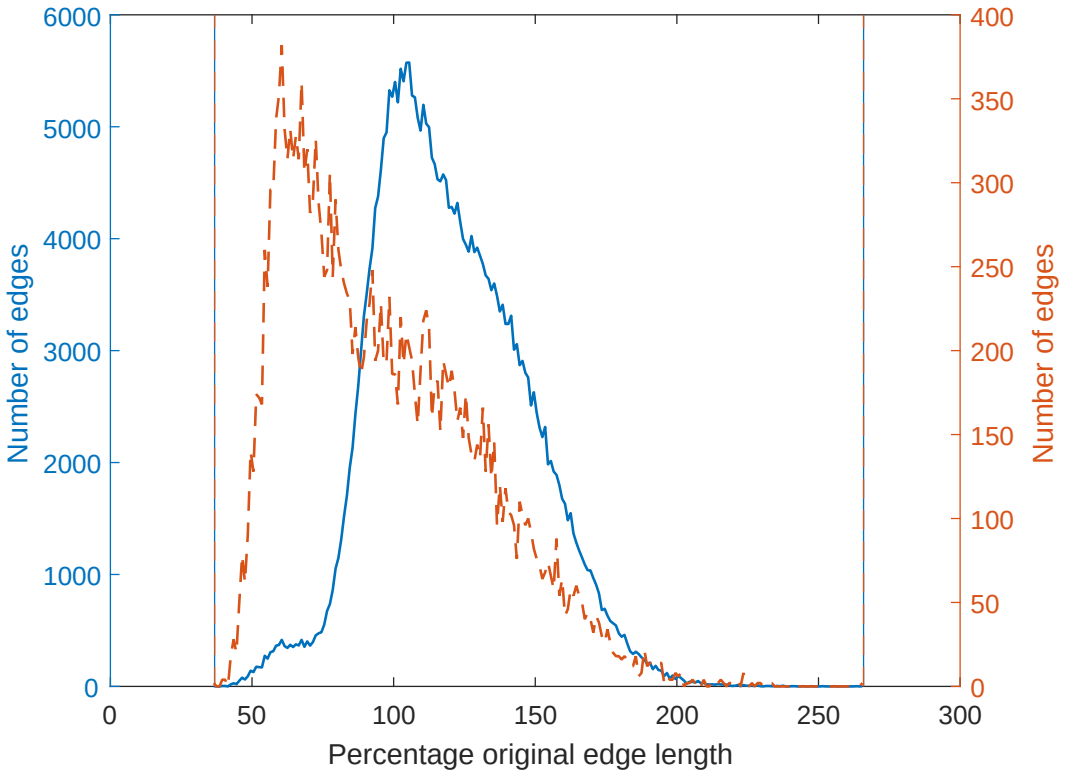


Figure 4.41: Comparison of the distributions of edge length distortions for the spherical mesh (solid blue, left y-axis) and just the pinnae edges (dashed orange, right y-axis), after application of spatial averaging using the edge length distortion adjacency matrix.

4.5.4 Ellipsoidal pinna emphasis (EPE)

The principal aim of ellipsoidal pinna emphasis (EPE) is to enlarge certain regions of the spherical mesh, in this case the pinnae, which are still very compressed, at the expense of less critical regions of the head. To do this, firstly the spherical mesh is rotated to align the area that needs emphasising along the y-axis. Next, the cartesian coordinates are converted into spherical coordinate azimuth φ, elevation θ and radius r via:

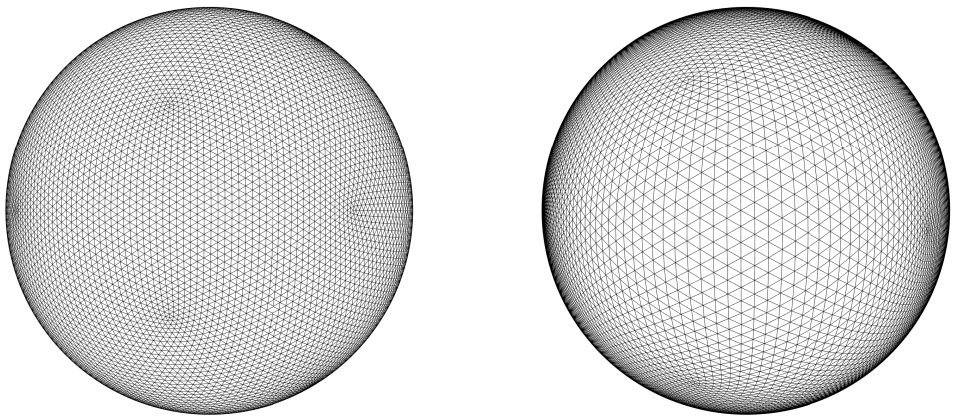
$$\begin{aligned} r &= \sqrt{x^2 + y^2 + z^2} \\ \theta &= \tan^{-1} \frac{z}{r} \\ \varphi &= \tan^{-1} \frac{y}{x} \end{aligned} \tag{4.59}$$

Then the points are projected onto an ellipsoid in cartesian space using different radii; a for the x-axis, b for the y-axis and c for the z-axis:

$$\begin{aligned} x &= a\cos\theta\cos\varphi \\ y &= b\cos\theta\sin\varphi \\ z &= c\sin\theta \end{aligned} \tag{4.60}$$

If the y-axis radius is smaller than the x and z axes radii then if the azimuth and elevation angles are recalculated as per equation 4.59, and projected back onto the sphere in Cartesian space using the original constant radius r, then both the elevation and azimuth angles are stretched in the areas close to the y-axis. This is shown in figure 4.42 for a uniform sphere mesh. For this example, the radii ratio was 1:0.5:1, i.e. the y-axis radius was set to half the x and z axes radii, which were equal.

Figure 4.43 shows a visualisation of the steps of EPE applied to one of the spherical head mesh pinna. Firstly the spherical mesh is rotated to align the focus of emphasis along the y-axis (figure 4.43b). Then EPE is applied via projection



(a) (b)

Figure 4.42: Uniform spherical mesh before (a) and after (b) the application of ellipsoidal optimisation. Note: Each compact dense region in the mesh is due to a reduction in the valency of one face. These are caused by the way in which the spherical meshes were originally generated and are shown later to be inconsequential.

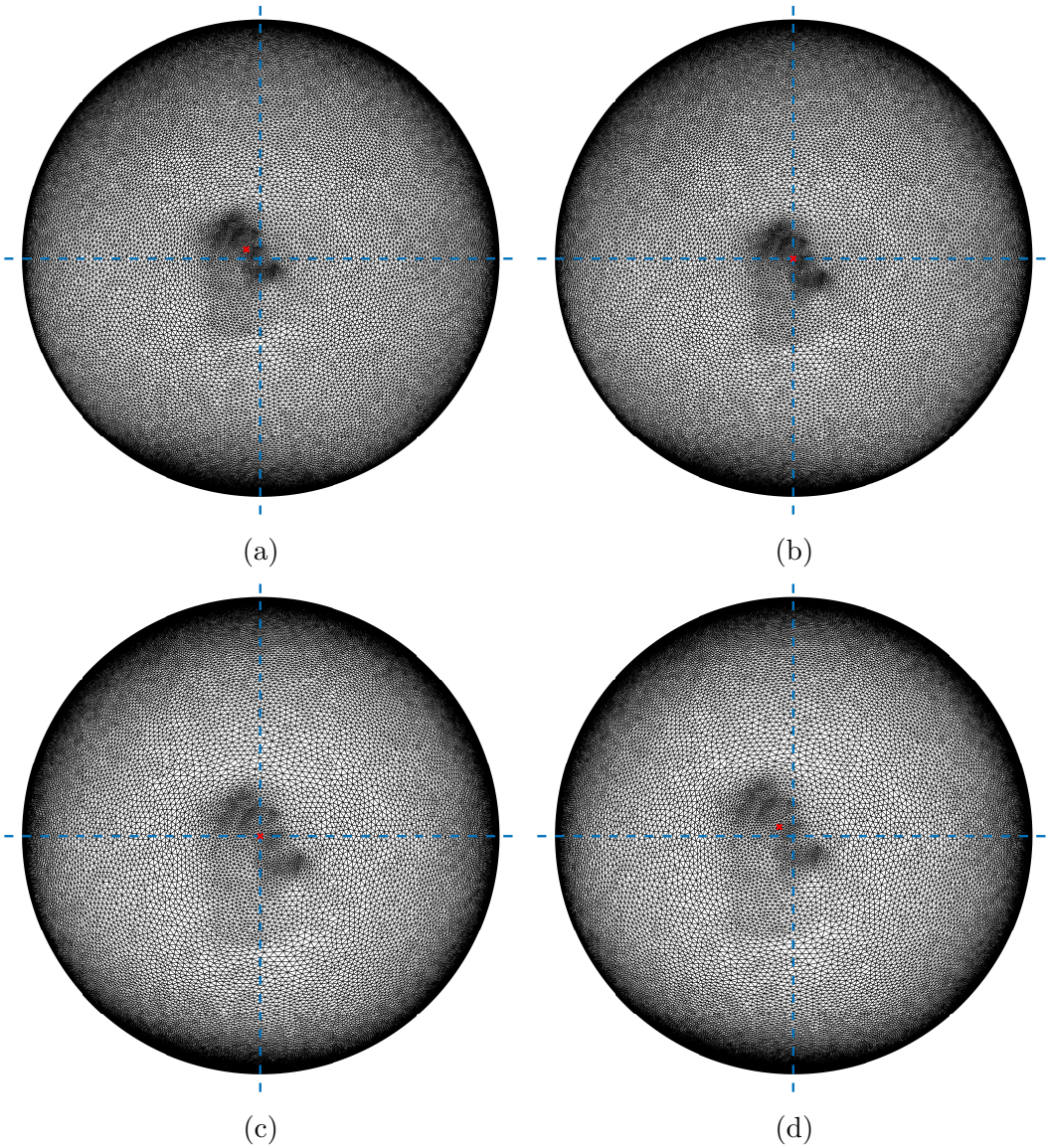


Figure 4.43: Visualisation of ellipsoidal pinna emphasis (EPE). (a) shows the initial spherical mesh before EPE. (b) shows the initial mesh rotated to align the focus of emphasis with the y-axis. (c) shows the mesh after EPE. (d) shows the mesh after rotating the focus back to its original position. The red 'x 'shows the focus of emphasis, the blue dashed lines represent the x and z axes.

to and from the ellipsoid as outlined above, which "stretches" the area around the focus (figure 4.43c). Finally the focus is rotated back to its original position (figure 4.43d).

To apply EPE to both pinnae requires independent processing of the two pinnae and recombination of the respective spherical meshes. This is due to the fact that the two pinnae are not diametrically opposite each other on the sphere. Thus it is not possible to align both pinnae along the y-axis at the same time. The focus of emphasis for each pinna is chosen manually as the approximate centre of the "bunched" area of the pinna (figure 4.43a).

To combine the two spherical meshes, two sagittal planes equidistant from the median plane are selected. Between the two sagittal planes a linear spatial crossfade is applied, i.e. the position of each vertex is calculated as the weighted average of the corresponding vertex from the left and right pinna meshes. The weighting of the left pinna mesh varies linearly from unity on the left sagittal plane to zero on the right sagittal plane. Similarly, the weighting of the right pinna mesh varies from zero on the left sagittal plane to unity on the right sagittal plane. Therefore close to the contour in median plane the weightings of the left and right pinna meshes are approximately equal. Outside the region contained by the sagittal planes the resulting mesh is the pinna mesh corresponding to that side of the head.

The distance between the two sagittal planes is set by the ratio of the y-axis radius to the x and z radii. Figure 4.44 compares the resulting spherical mesh from combining the left and right pinna meshes after EPE with ratios of 1:0.75:1 or 1:0.25:1, and with the crossfade applied between the $\pm 10°$ sagittal planes or the $\pm 30°$ sagittal planes. Using the 10° sagittal planes (figures 4.44a and 4.44c) for either radius ratio leaves clear, unwanted, joining lines, whilst using the 30° sagittal planes works for the 1:0.75:1 ratio (figure 4.44b) but not for the 1:0.25:1 ratio (figure 4.44d). In general the larger the radii ratios, the further apart the sagittal planes need to be. It was found experimentally that using the $\pm 50°$

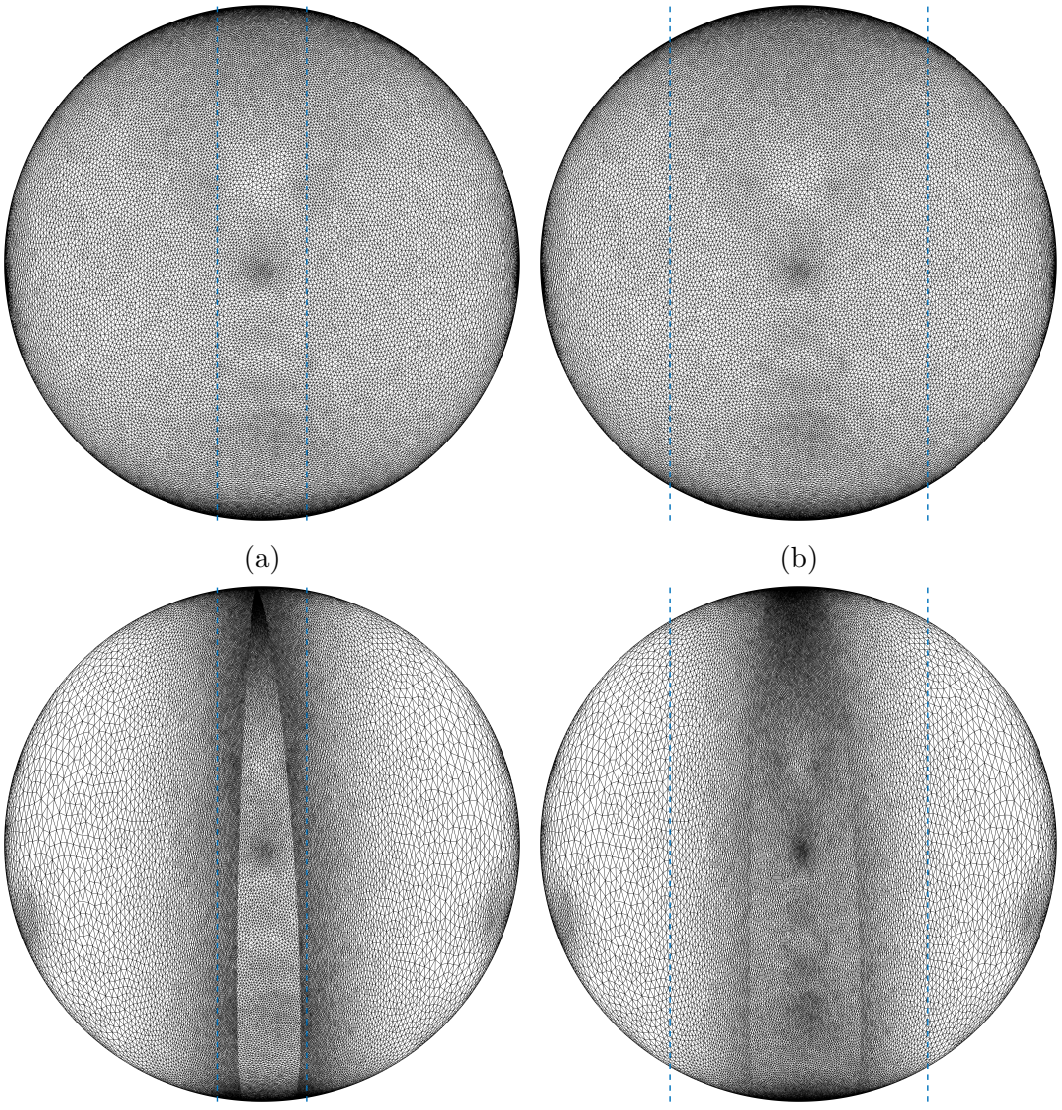


Figure 4.44: Reconstructed EPE meshes for different radii ratios and different spatial crossfading. The top row are for the radii ratios 1:0.75:1 and the bottom row are for the ratios 1:0.25:1. The left hand plots are with crossfade applied between the $\pm 10°$ sagittal planes whilst the right hand plots crossfade between the $\pm 30°$ sagittal planes. Dashed blue lines show the respective sagittal planes.

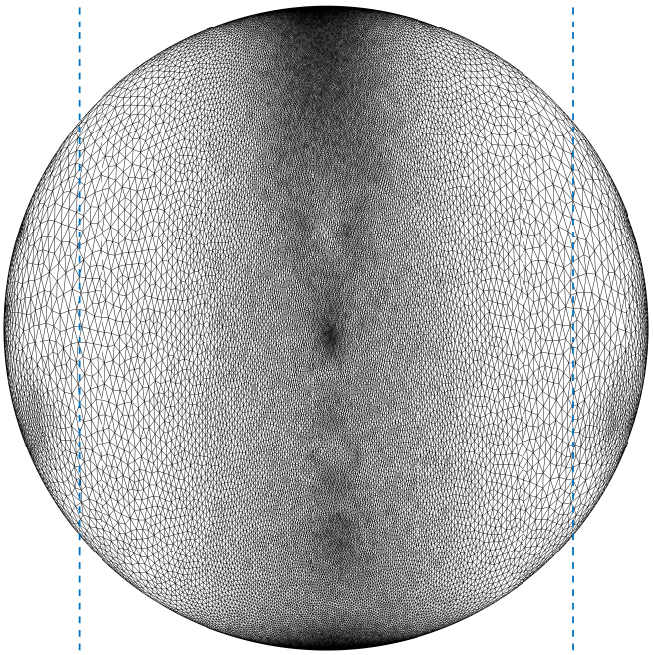


Figure 4.45: Reconstructed EPE mesh for the radii ratios 1:0.25:1 and with crossfade applied between the 50° sagittal planes. Dashed blue lines show the sagittal planes.

sagittal planes produced smoothly transitioning spherical meshes for ratios down to 1:0.25:1 (figure 4.45).

To find the appropriate radii ratios, the radius of the y-axis was progressively reduced until there were no edges within the pinnae with a spatial deformation frequency resolution of less than 16 kHz. The required ratio was 1:0.45:1 and the resulting spherical mesh, after EPE, is shown in figure 4.46. Whilst there are no edges within the pinnae that have a resolution of less than 16 kHz it can be seen that the level of EPE required for this to be so has significantly compressed the edges around the median plane. Furthermore if the maximum valid frequency distributions are examined (figure 4.47) it can be seen that EPE has actually increased the overall number of edges with a resolution of less than 16 kHz. Therefore a new method of local optimisation, dynamic pinna emphasis (DPE) was developed, which is described in the next section.

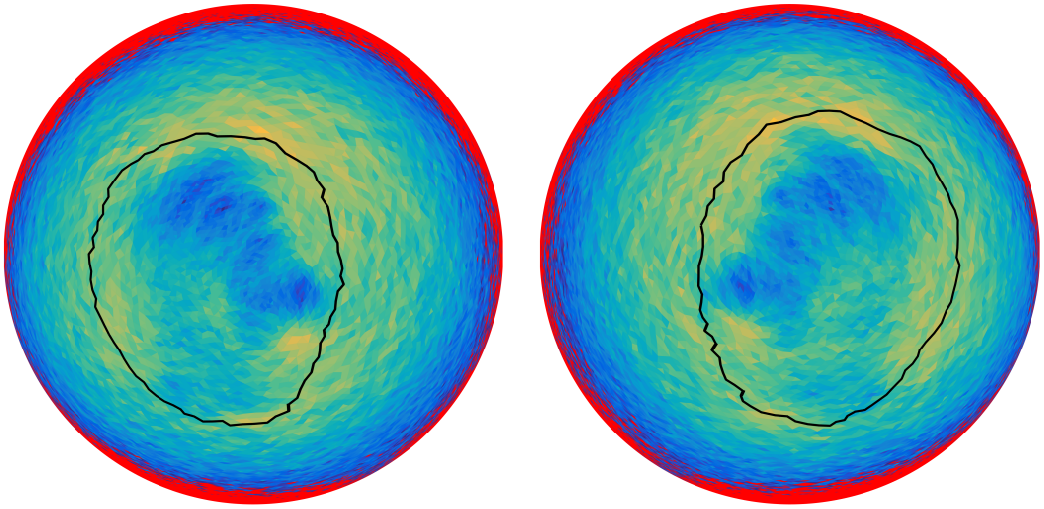


Figure 4.46: Spherical head mesh after applying EPE until no edges in the pinnae have a spatial deformation frequency resolution of less than 16 kHz. None of the pinna edges are under resolution, however many edges around the median plane are.

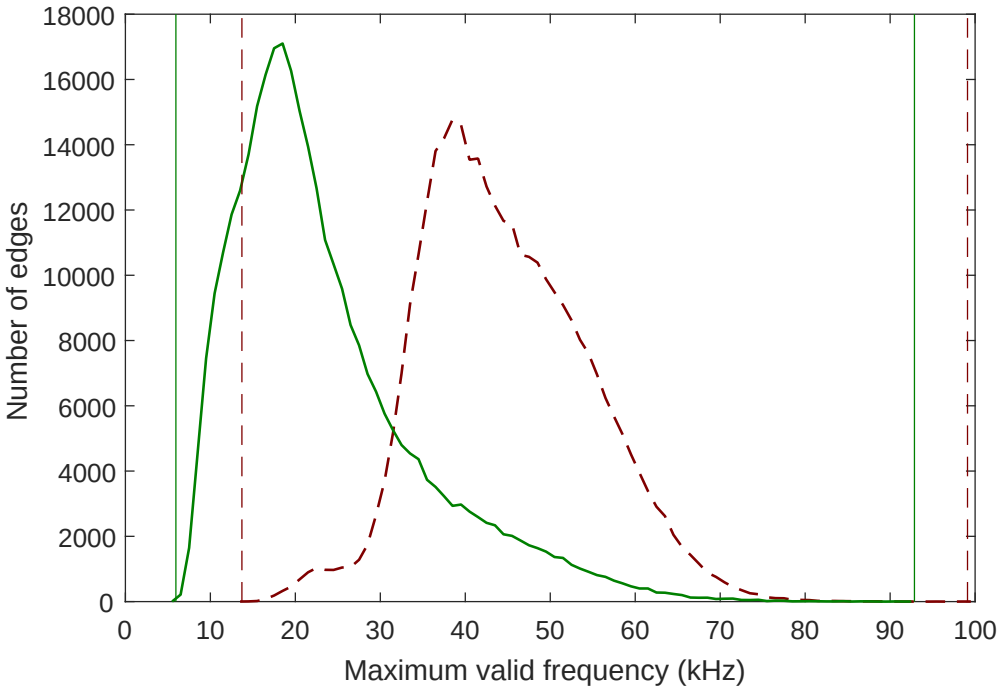


Figure 4.47: Distributions of the maximum valid spatial deformation frequencies for the spherical head mesh before (red dashed lines) and after (green solid lines) application of EPE. Vertical lines give absolute limits.

4.5.5 Dynamic pinna emphasis (DPE)

One of the limitations of the spatial averaging approach is that it works best for meshes of equal valence , i.e. when all the vertices are connected to the

same number of other vertices (Diestel, 2005). The algorithm tends to produce "bunching" in areas of reduced valence. This is suboptimal as the pinna mesh regions contain instances of reduced valence due to their complex connectivity. The basis of dynamic pinna emphasis (DPE) is to apply the spatial averaging to a template spherical mesh to which the distortions of the spherical head mesh have been mapped. Once the template mesh has been optimised it is used to apply localised optimisation to the spherical head mesh. The chief property of the spherical template mesh is its almost uniform valence. Complete uniformity is not achieved because spherical meshes made up of triangular faces are generally created through subdivision of the faces of an icosahedron. Therefore there are always 12 vertices in the spherical mesh that have a valence of five, whilst the rest have a valence of six. The template spheres used in this work were generated using the "GeoSphere" tool in Autodesk 3ds Max®[2].

The steps of DPE are summarised in algorithm 4.3. The spatial averaging in line 9 is carried out as per sections 4.5.1 and 4.5.3, but now the adjacency matrix is defined by the edge weights calculated in lines 2–8. Figures 4.48a and 4.48b show a template mesh with approximately 10,000 vertices, before and after spatial averaging. The faces in both figures have been coloured based on the average

[2]https://www.autodesk.co.uk/products/3ds-max/overview

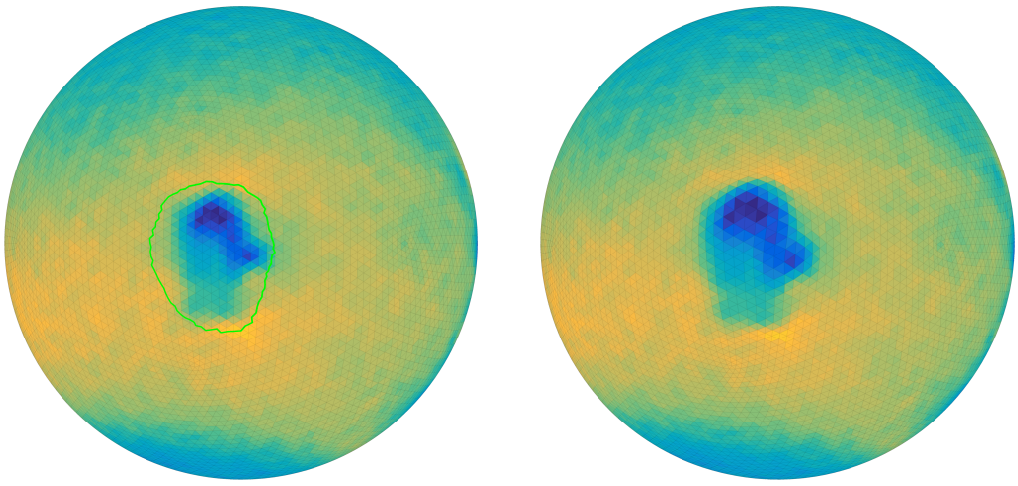
Algorithm 4.3 Dynamic pinna emphasis (DPE) algorithm.

1: *calculate distortions of spherical head mesh edges*
2: **for all** *vertices in template mesh* **do**
3: *find closest vertex in spherical head mesh*
4: *set template vertex weight as average of the distortions of the edges connected to spherical head mesh vertex*
5: **end for**
6: **for all** *edge in template mesh* **do**
7: *calculate the edge weight as average of the weights of its two vertices*
8: **end for**
9: *apply distortion weighted spatial averaging to template sphere*
10: **for all** *vertices in spherical head mesh* **do**
11: *calculate position of vertex relative to original template mesh*
12: *calculate vertex's new position relative to the new, optimised template mesh*
13: **end for**

weight of each face's edges. Blue faces indicate that the edges of the face are, on average, compressed, and yellow faces indicate that the edges of the face are, on average, stretched. Green faces indicate that the edges are the correct length. The weights have been converted to decibel values for visualisation to create a symmetrical linear colour scale, i.e. a weight of 0.5 and a weight of 2 give shades equally distant from the mid shade. As seen in figure 4.48a the pinnae areas are initially heavily compressed. After 100 iterations of spatial averaging, shown in figure 4.48b, the faces in the pinnae regions have increased in area. This emphasis is then mapped to the spherical head mesh.

To calculate the position of each of the spherical head mesh vertices relative to the template mesh (lines 10–13 in algorithm 4.3), barycentric coordinates are used. Barycentric coordinates parameterise a space in terms of a weighted sum of reference points (Ericson, 2005).

To calculate the barycentric coordinates, α, β and γ, of a point within a triangle, consider a triangle consisting of vertices A, B and C. Any point P in the triangle



(a) (b)

Figure 4.48: 10,000-vertex template mesh before (a) and after (b) application of dynamic pinna emphasis (DPE). The faces are coloured based on the average weight of its edges. Blue faces indicate weights signifying compression in the spherical head mesh, yellow faces indicate weights signifying stretch in the spherical head mesh. Green faces represent areas of correct edge length in the spherical head mesh. The bold green lines show the limits of the pinnae in the spherical head mesh.

can be uniquely expressed as:

$$P = \alpha A + \beta B + \gamma C \tag{4.61}$$

where α, β and γ are constants, and $\alpha + \beta + \gamma = 1$. The above equation can be reformulated as:

$$P = A + \beta(B - A) + \gamma(C - A) = (1 - \beta - \gamma)A + \beta B + \gamma C \tag{4.62}$$

Rearranging this gives:

$$P - A = \beta(B - A) + \gamma(C - A) \tag{4.63}$$

which can be written in vector form as:

$$\boldsymbol{v}_3 = \beta \boldsymbol{v}_1 + \gamma \boldsymbol{v}_2 \tag{4.64}$$

where:

$$\begin{aligned} \boldsymbol{v}_1 &= B - A \\ \boldsymbol{v}_2 &= C - A \\ \boldsymbol{v}_3 &= P - A \end{aligned} \tag{4.65}$$

Now, a pair of linear equations can be formed by taking the dot product of both sides of equation 4.64 with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$:

$$\begin{aligned} &(\beta \boldsymbol{v}_1 + \gamma \boldsymbol{v}_2) \cdot \boldsymbol{v}_1 = \boldsymbol{v}_3 \cdot \boldsymbol{v}_1, \text{ and} \\ &(\beta \boldsymbol{v}_1 + \gamma \boldsymbol{v}_2) \cdot \boldsymbol{v}_2 = \boldsymbol{v}_3 \cdot \boldsymbol{v}_2 \end{aligned} \tag{4.66}$$

and, since the dot product is a linear operator, this is equivalent to:

$$\begin{aligned}\beta(\boldsymbol{v}_1 \cdot \boldsymbol{v}_1) + \gamma(\boldsymbol{v}_2 \cdot \boldsymbol{v}_1) &= \boldsymbol{v}_3 \cdot \boldsymbol{v}_1, \text{ and} \\ \beta(\boldsymbol{v}_1 \cdot \boldsymbol{v}_2) + \gamma(\boldsymbol{v}_2 \cdot \boldsymbol{v}_2) &= \boldsymbol{v}_3 \cdot \boldsymbol{v}_2 \end{aligned} \tag{4.67}$$

This pair of equations can be solved using Cramer's rule, which states that in a linear system:

$$\begin{aligned} a_1 x + b_1 y &= c1, \text{ and} \\ a_2 x + b_2 y &= c2 \end{aligned} \tag{4.68}$$

x and y can be found by:

$$\begin{aligned} x &= \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{c_1 b_2 - b_1 c_2}{a_1 b_2 - b_1 a_2}, \text{ and} \\ y &= \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{a_1 c_2 - c_1 a_2}{a_1 b_2 - b_1 a_2} \end{aligned} \tag{4.69}$$

Therefore setting:

$$\begin{aligned} a_1 &= \boldsymbol{v}_1 \cdot \boldsymbol{v}_1 \\ a_2 &= \boldsymbol{v}_1 \cdot \boldsymbol{v}_2 \\ b_1 &= \boldsymbol{v}_2 \cdot \boldsymbol{v}_1 \\ b_2 &= \boldsymbol{v}_2 \cdot \boldsymbol{v}_2 \\ c_1 &= \boldsymbol{v}_3 \cdot \boldsymbol{v}_1 \\ c_2 &= \boldsymbol{v}_3 \cdot \boldsymbol{v}_2 \end{aligned} \tag{4.70}$$

and solving equations 4.69 allows β to be γ calculated, after which α can be calculated as $1 - \beta - \gamma$.

In DPE the reference points, A, B and C, for a spherical head mesh vertex are the vertices of the template mesh face that this spherical head mesh vertex lies within. To find the template mesh face which contains the spherical head mesh vertex, the closest template mesh vertex to the current spherical head mesh vertex is found. Then each of the template mesh faces connected to this vertex is used to calculate the barycentric coordinates α, β and γ. If all three coordinates lie between zero and one, then the spherical head mesh vertex lies within the template mesh face. After optimisation, the new position of the spherical head mesh vertex is calculated as the weighted sum of the optimised template mesh vertex positions:

$$P_{opt} = \alpha A_{opt} + \beta B_{opt} + \gamma C_{opt} \tag{4.71}$$

To avoid errors due to the vertex and face being non coplanar, before calculation of the barycentric coordinates the spherical head mesh vertex is projected onto the same plane as the template mesh face. This is achieved by defining an origin for the plane, which is taken as the centre of the face:

$$O = \frac{1}{3}(A + B + C) \tag{4.72}$$

Next, a vector is defined from the origin O to the point of interest P:

$$\boldsymbol{v} = P - O \tag{4.73}$$

Then the distance from the point of interest to the plane, along the face plane's normal is calculated by taking the dot product of the vector $\boldsymbol{v}$ and the unit normal $\boldsymbol{n}$:

$$d = \boldsymbol{v} \cdot \boldsymbol{n} \tag{4.74}$$

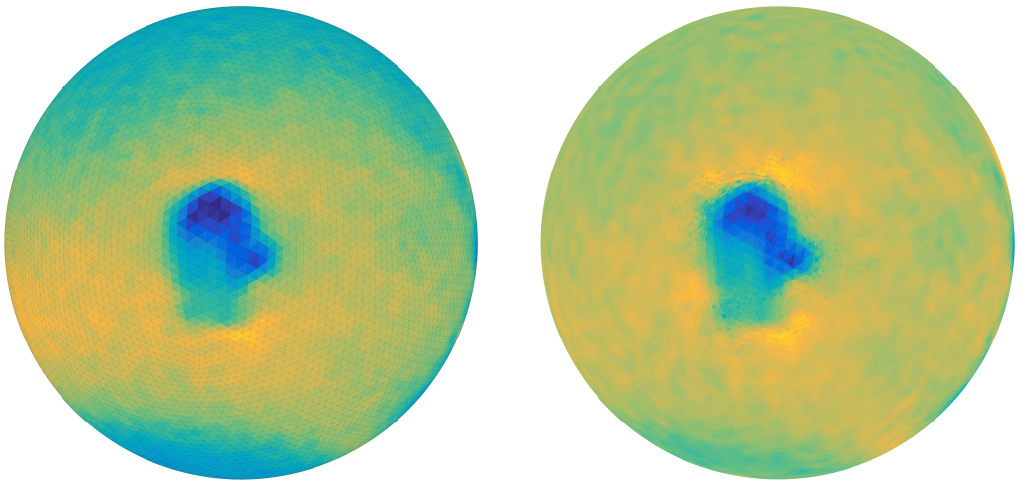
where $\boldsymbol{n}$ is calculated as the normalised cross product of two edges of the face. Finally the point's position on the face plane is calculated by subtracting the product of the distance d and the unit normal $\boldsymbol{n}$ from the point P:

$$P_{plane} = P - d\boldsymbol{n} \tag{4.75}$$

To establish the effectiveness of DPE, after calculating the new positions of the spherical head mesh vertices as outlined above, the ratios of the edge lengths before applying DPE to the edge lengths after applying it were calculated. Then each face on the spherical head mesh was coloured based on the mean ratio of its three edges using the same colour scale as figures 4.48a and 4.48b. Therefore the resulting mesh should resemble figure 4.48b, as the edges in the areas of compression (low weights) in figure 4.48b should have increased in length after application of DPE, and thus the ratio of the length before to the length after should be less than one. And vice versa with areas of stretch.

Figure 4.49 compares the spherical head mesh with its faces coloured based on the above criteria with the template mesh in figure 4.48b. There is a lot of similarity in and around the pinnae. For instance the dark blue area in the antihelix is clearly visible on both meshes, as is the yellow area near the pinna lobule. However, away from the pinna, edge length ratios become less similar; but, as the pinnae are the areas of most interest for MPA, this was deemed to be acceptable.

The maximum spatial deformation frequency resolution of the spherical head



(a) (b)

Figure 4.49: Comparison of the 10,000-vertex template sphere in figure 4.48b (a) and the spherical head mesh (b) with faces coloured based on ratio of edge lengths before DPE to edge lengths after. In (b) blue faces indicate edges that have increased in length, whilst yellow faces indicate shortening of edges.

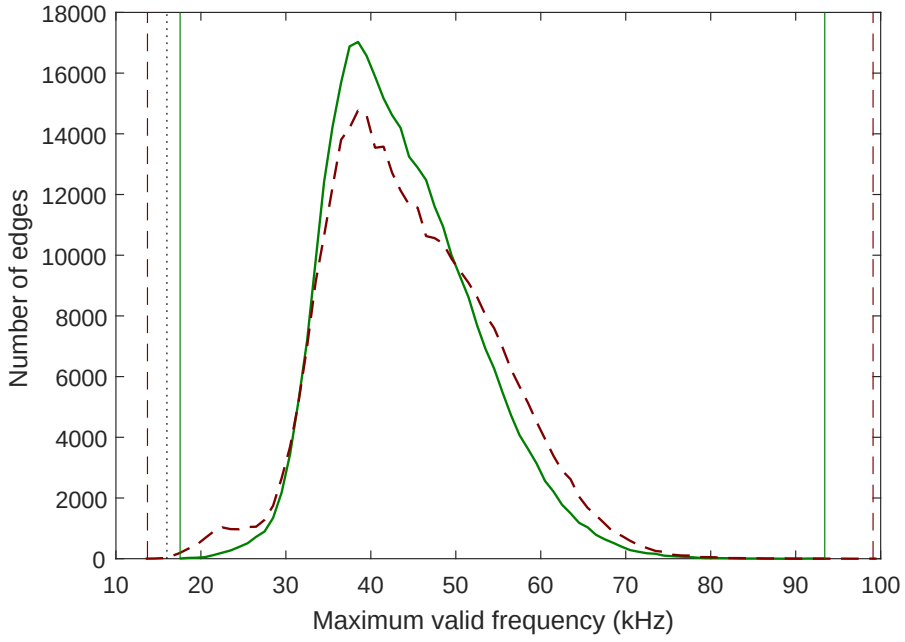


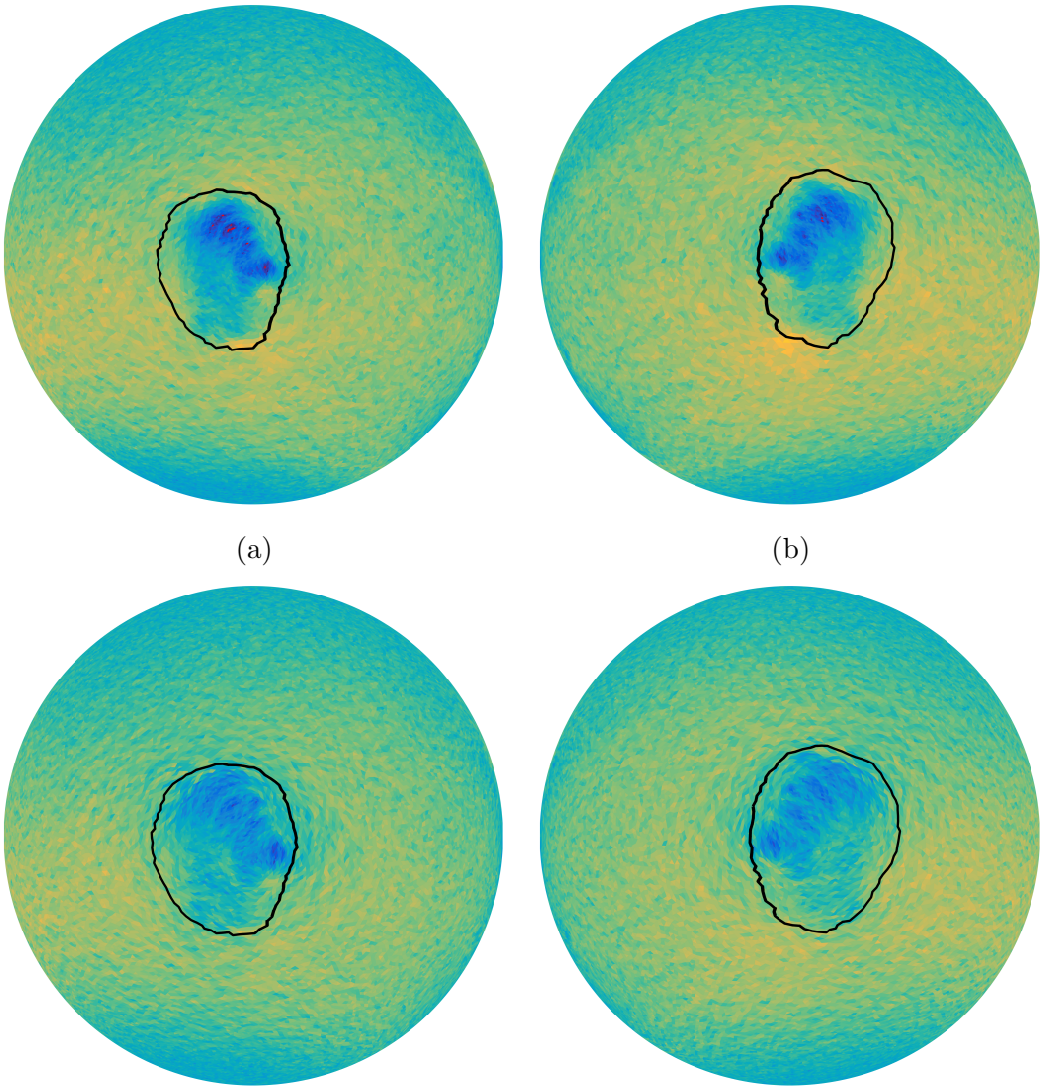
Figure 4.50: Distributions of maximum valid spatial deformation frequency before (red dashed lines) and after (green solid lines) application of DPE using a 10,000-vertex template mesh. Black vertical dotted line shows required minimum resolution of 16 kHz.

mesh before and after application of DPE with the same 10,000-vertex template mesh was also evaluated. The distributions of the maximum spatial frequency resolution are given in figure 4.50 and the corresponding visualisation is given in figure 4.51. The application of DPE has resulted is none of the edges having a maximum resolution of less than 16 kHz: the lower limit of maximum resolution is 17.5 kHz. Furthermore the upper limit of maximum resolution has decreased, meaning that fewer harmonics will be required to attain the required resolution.

The results of DPE so far are based upon using 100 iterations of spatial averaging on a 10,000-vertex template mesh. The effect of increasing and decreasing the number of iterations, as well as increasing the number of vertices in the template mesh was also explored.

Unlike the spatial averaging used in section 4.5.3, for DPE it was not practical to update the adjacency matrix on every iteration. This is because, after spatial averaging of the template mesh, it is necessary to run through each vertex in the spherical head mesh, calculate its position with respect to the original template mesh and then calculate its new position. To update the adjacency matrix on every iteration of the spatial averaging would require recalculation of the spherical mesh vertices' positions on each iteration, rather than just once at the end, as well as recalculation of the template mesh weights. For the 10,000-vertex template mesh 100 iterations using the same adjacency matrix took 63.36 seconds, updating the adjacency matrix on each iteration would take over an hour and a half. For this reason the same matrix, based on the initial edge distortions, was used for each of the 100 iterations.

The initial choice of 100 iterations was arbitrarily chosen to test the principle of DPE. It was therefore necessary to validate whether this choice was optimal. To do this, the adjacency matrix based on the initial distortion was used for all iterations, but every ten iterations (to give a balance between speed and data sampling) the positions of the spherical head mesh were updated and the maximum spatial deformation frequency resolutions (as per section 4.5.2) were



(a) (b)

(c) (d)

Figure 4.51: Visualisation of maximum valid spatial deformation frequency before, (a) and (b), and after, (c) and (d), application of DPE using a 10,000-vertex template mesh. Faces coloured as per previous plots. (a) and (c) show the right pinna, (b) and (d) the left. DPE has resulted in an enlarged, more even distribution of pinna face resolutions. The red (under-resolved) faces, present particularly in the left pinna, have been corrected.

recalculated. This gives an indication of how the spherical head mesh changes over successive applications of the same adjacency matrix. Figure 4.52 shows the upper and lower limits of the maximum spatial deformation resolution over 500 iterations using a 10,000-vertex template sphere. It can be seen that the lower limit of resolution increases with the application of DPE and the upper limit decreases, which is what is needed to reduce the number of harmonic deformations required. It seems from figure 4.52 that by 100 iterations both the upper and lower limits have settled. However, if the *change* in these values is considered (figure 4.53) it can be seen that, whilst the lower limit has stopped increasing by 100 iterations, the upper limit is still decreasing. In fact it is not until 220 iterations that the rate of decrease in the upper limit of maximum spatial frequency becomes smaller than the rate of decrease of the lower limit and hence the difference between upper and lower limits begins to rise again. Therefore, for a 10,000-vertex template mesh, 220 iterations are required for optimum results. It should be noted that the change in the range only decreases by 300 Hz between 100 and 220 iterations, whereas the decrease up to 100 iterations is 9.5 kHz; so the returns are diminishing.

Additionally the effect of the number of vertices in the template mesh was investigated. Increasing the number of vertices in the template mesh should increase the resolution of DPE since it should be possible to apply more localised optimisation. The downside of increasing the number of vertices in the template mesh is the increase in processing time.

Firstly a comparison of the performance of increased resolution template meshes was made using the calculated optimum number of iterations (220) for the 10,000-vertex template mesh. The comparison of the maximum spatial deformation frequency distributions for the original 10,000-vertex template mesh and 50,000-, 150,000- and 300,000-vertex meshes is shown in figure 4.54. It can be seen that with 220 iterations, for the 150,000- and 300,000-vertex template meshes, there is no improvement over the 10,000-vertex mesh, and whilst there is an improvement in the lower limit for the 50,000-vertex mesh, the upper limit has increased by a

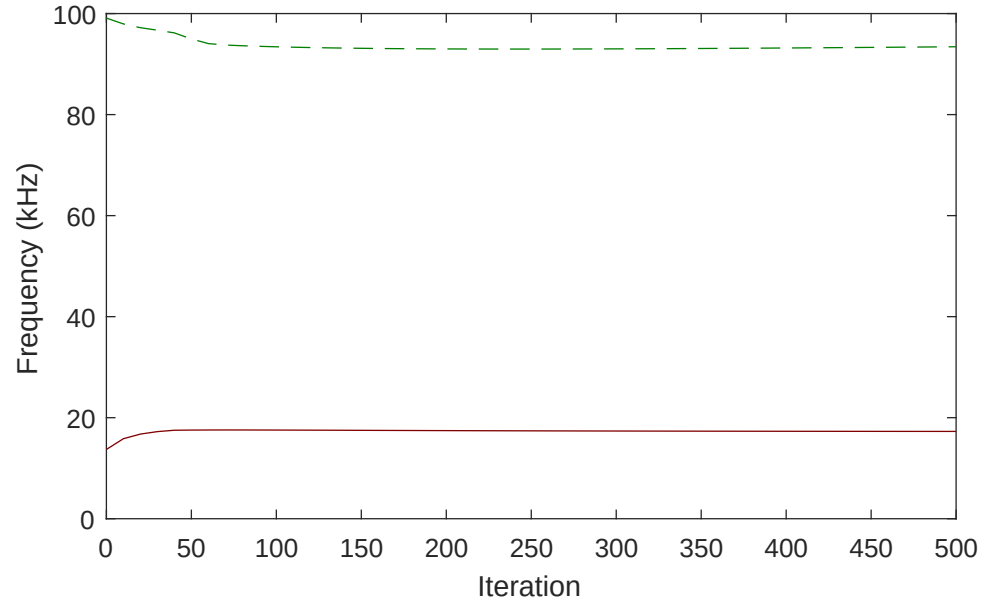


Figure 4.52: Upper (dashed green line) and lower (solid red line) limit of maximum spatial deformation resolution over 500 iterations of DPE using a 10,000-vertex template sphere. Data points are generated every ten iterations.

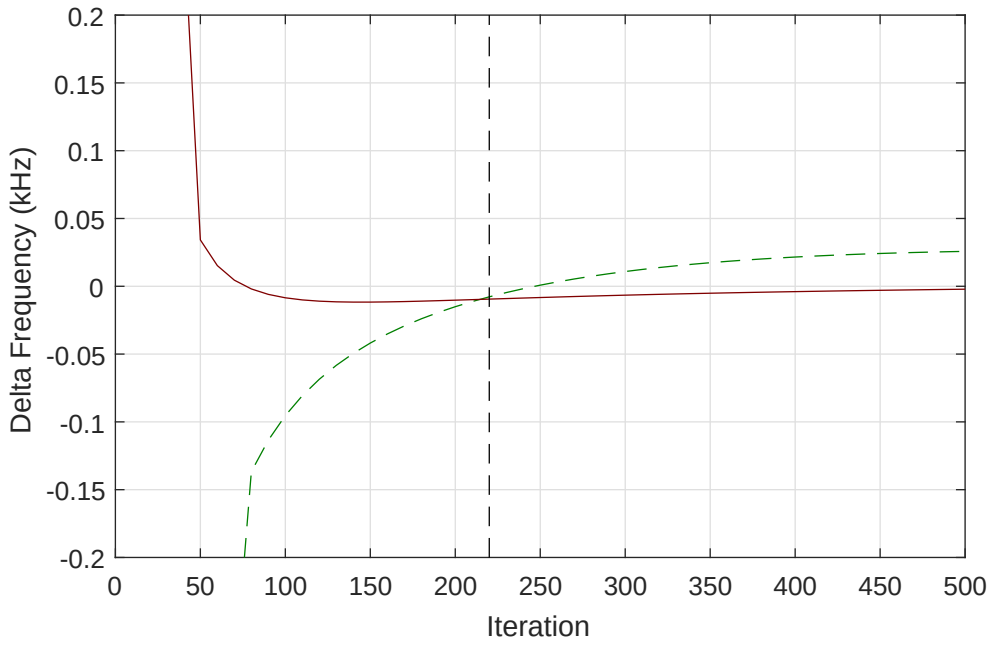


Figure 4.53: Change in the upper limit (dashed green) and lower limit (solid red) of maximum spatial deformation frequency resolution over 500 iterations of DPE using a 10,000-vertex template sphere. Black dashed vertical line shows the point at which the range between these limits is a minimum.

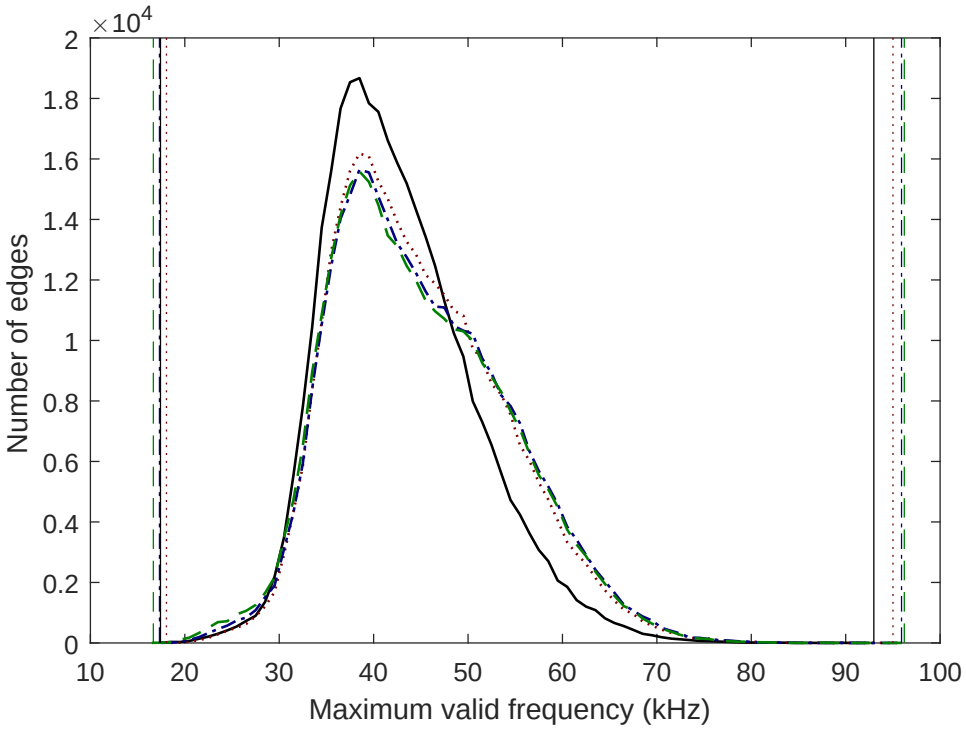


Figure 4.54: Distributions of maximum valid spatial deformation frequency before and after application of DPE using different resolution template meshes with the same number of iterations. The different resolution meshes are: 10,000 vertices (solid black lines), 50,000 vertices (dotted red lines), 150,000 vertices (blue dash-dot lines) and 300,000 vertices (green dashed lines).

greater margin. The 50,000-vertex template mesh would therefore require more harmonic deformations than the 10,000-vertex mesh.

Therefore the same analysis that was carried out on the 10,000-vertex mesh to find its optimum number of iterations was carried out for each of the higher resolution meshes. To keep computation times down to a reasonable level the maximum number of iterations was restricted to 3000 and the edge lengths of the spherical head mesh were calculated every 100 iterations. It was found that the 50,000-vertex mesh required 1100 iterations to minimise the range of maximum spatial deformation frequencies. Neither the 150,000- or 300,000-vertex meshes reached their optimum point within the 3000 iteration limit. Therefore linear extrapolation over the last, approximately linear portion of the curves was applied to estimate how many iterations would be required. An example of this for the 150,000-vertex mesh is given in figure 4.55. This gave values of 3117 iterations for the 150,000-vertex mesh and 4380 iterations for the 300,000-vertex mesh.

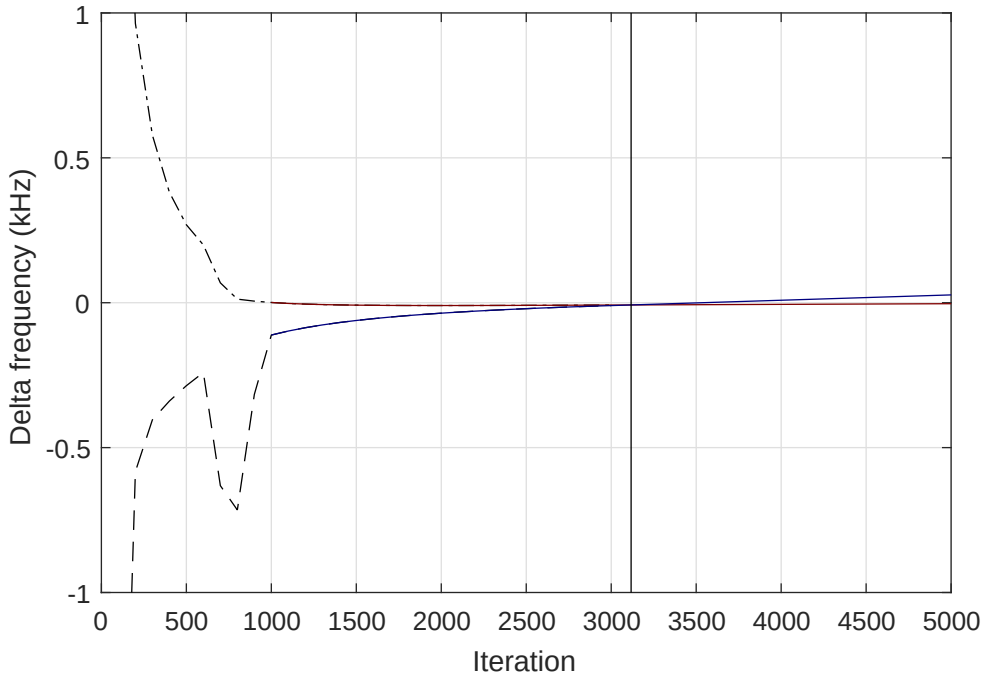


Figure 4.55: Example of extrapolation used to estimate required iterations for the 150,000-vertex template mesh. The change in the upper limit (dashed line) and lower limit (dash-dot line) resolution was recorded for 3000 iterations and linear extrapolation applied based on the last 2000 iterations to estimate where the two curves would cross. The solid red and blue lines show the extrapolated data, including its overlap with the recorded data.

The limits and ranges of the maximum spatial deformation frequencies after DPE, using their respective optimum number of iterations, are given in table 4.1 for 10,000-vertex, 50,000-vertex, 150,000-vertex and 300,000-vertex template spheres. It can be seen that application of the higher resolution template meshes with enough iterations firstly increases the lower limit, but secondly decreases the range of values. This is important for reducing the number of required harmonics to reach the required spatial frequency resolution across the whole head mesh. (Note that the lower/upper limits can be adjusted independently of the range simply by altering the radius of the sphere.)

The possibility of applying DPE with a lower resolution mesh and then applying fewer iterations of a higher resolution mesh was also considered as a potential means for reducing the computation required. The results of applying the 300,000-vertex mesh for 4380 iterations were compared with applying 220 itera-

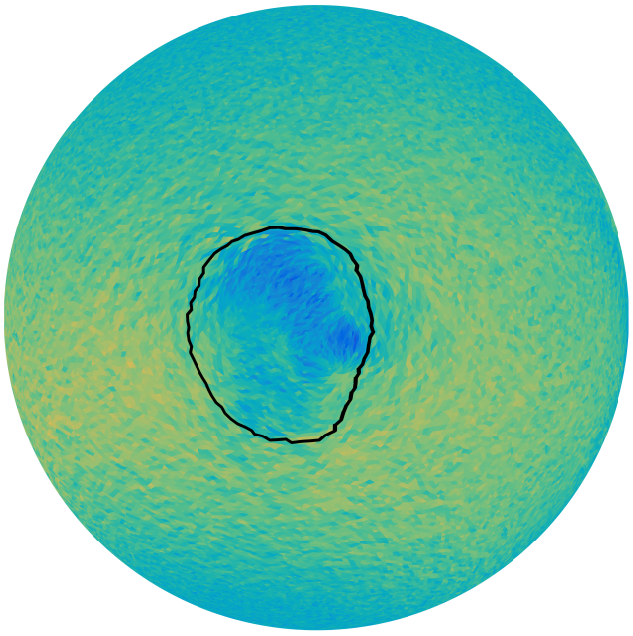
No. vertices	Min value (kHz)	Max value (kHz)	Range (kHz)
10,000	17.42	92.98	75.56
50,000	17.9	92.24	74.34
150,000	18.52	92.13	73.61
300,000	19.21	92.73	73.52

Table 4.1: Limits and ranges of maximum spatial deformation frequency distributions after DPE with different resolution template meshes and the optimum number of iterations for each mesh.

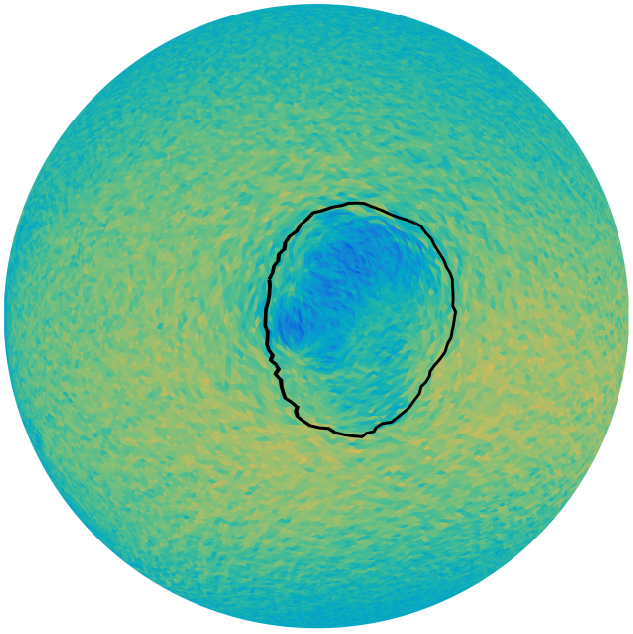
tions of the 10,000-vertex mesh, followed by just 100 iterations of the 300,000-vertex mesh. The two approaches gave very similar results: just applying the 300,000-vertex mesh gave a range of 73.52 kHz, compared to 73.61 kHz for the two meshes. Furthermore, both strategies took about the same time to compute, as the most time-consuming parts of DPE are the weighting of the template mesh (lines 2–8 in algorithm 4.3) and the updating of the spherical head mesh vertex positions (lines 10–13). In applying the two template meshes, both of these steps had to be carried out twice and so the overall saving in spatial averaging iterations was inconsequential.

A further, higher resolution, template mesh consisting of 400,000 vertices (the maximum resolution of GeoSphere producible in 3ds Max®) was also tested; however, it gave negligible improvement over the 300,000-vertex mesh.

Therefore the spherical head mesh finally adopted to generate the MPA database was the result of applying DPEto the 300,000-vertex template mesh until the range of maximum spatial deformation frequencies over all edges was minimised (4380 iterations). This mesh is shown in figure 4.56 with its faces coloured according to the maximum spatial deformation frequency of each edge. Figure 4.57 shows the improvement in the distribution of maximum spatial deformation frequency over the resulting spherical head mesh from section 4.5.3, i.e. after the application of the edge length distortion adjacency matrix. The lower limit is above the required 16 kHz and the maximum limit has been reduced from 99.11 kHz to 92.73 kHz, meaning fewer harmonic deformations will be required



(a)



(b)

Figure 4.56: Final spherical head mesh, after spatial averaging optimisation and DPE, used for generation of the MPA database.

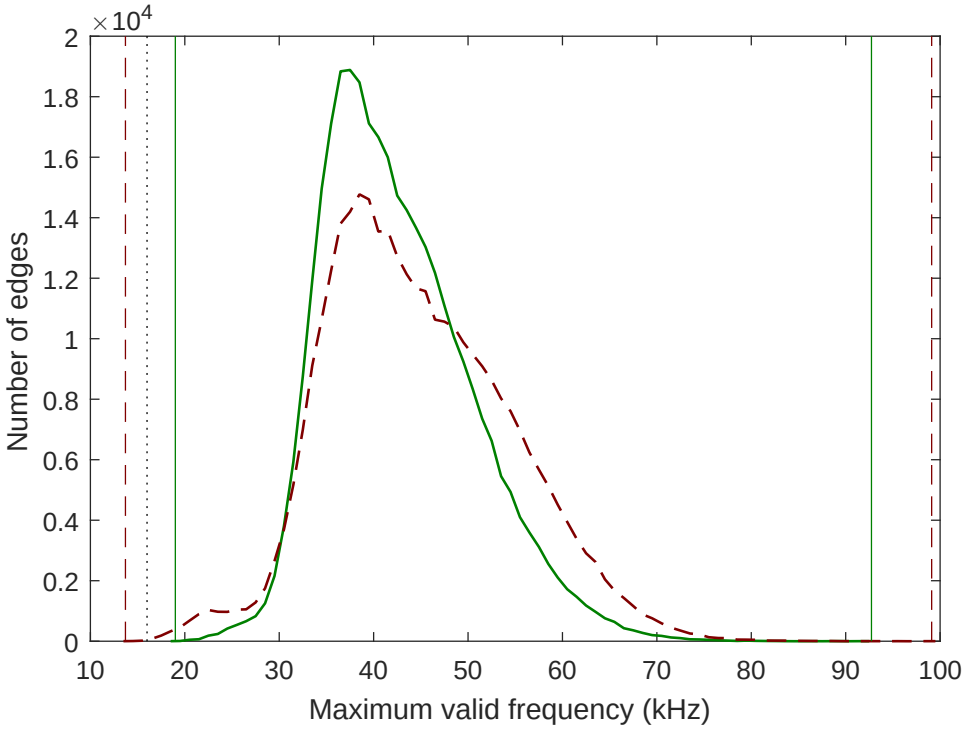


Figure 4.57: Comparison of distributions of maximum valid spatial deformation frequency before (dashed red lines) and after (solid green lines) final DPE.

when generating the MPA database (chapter 5).

4.6 Summary

This chapter has presented work on developing and improving morphoacoustic perturbation analysis (MPA) (Thorpe, 2009; Tew *et al.*, 2012). First-generation MPA suffered from a number significant weaknesses, chiefly created by the approach of slicing the head mesh and mapping the head to a 2D surface for the application of harmonic deformations. These issues have been addressed using several novel approaches based on optimised spherical mapping for the application of surface spherical harmonic deformations.

The first novel approach of mapping the head mesh to the sphere based on angular rotation and the distance travelled along the surface of the mesh is a natural adaptation of the slicing method used by Thorpe (2009) and Tew *et al.* (2012). Whilst path length relaxation (PLR) mapping provides promising results for low

resolution meshes, when applied to full resolution head meshes, computation time became excessive and overlapping of faces occurred.

Therefore a combination of new approaches, collectively termed optimised projection (OP), based on projection of the head mesh onto the surface of the sphere and optimisation of the spherical head mesh were developed. Prior techniques of projection and spatial averaging (Peyré, 2011a,b,c; Numerical Tours, 2017a) allowed the mapping of high resolution head meshes to the sphere without creating any overlapping faces. However, there was significant distortion of the edges on the sphere including, crucially, in the pinnae. Based on the maximum valid spatial deformation frequency supported by each edge in the spherical mesh, it was found that many of the edges could not support the required spatial deformation frequency resolution.

A novel adaptation of equal-weighted spatial averaging, with weights based on the distortion of the edges in the spherical mesh, was shown to reduce the distortion of edge lengths appreciably. However, there still existed edges within the pinnae with a maximum valid spatial deformation frequency resolution of less than 16 kHz, the required resolution for MPA. This was demonstrated to be largely due to relative compression of edges within the pinnae compared to the rest of the spherical mesh.

Hence, the novel method of ellipsoidal pinna emphasis (EPE) was developed to improve the resolution of the pinnae regions. Via rotation of the sphere and projection to and from an ellipsoid, EPE allowed emphasis of the pinnae. However, the effectiveness of EPE was limited by the large amount of emphasis required to ensure adequate resolution within the pinnae, which detrimentally reduced resolution on and close to the median plane.

Finally, another new optimisation approach, dynamic pinna emphasis (DPE), was presented. By applying spatial averaging to a uniform template sphere, with weightings based on the distortion of the spherical head mesh, highly localised optimisation was possible; regardless of the connectivity of the head mesh. This

resulted in a spherical head mesh in which every edge possesses a spatial deformation frequency resolution of over 16 kHz. Additionally, the range of edge distortions was systematically minimised, reducing the number of harmonics required to achieve the desired resolution across the whole mesh. For these reasons, DPE is considered the most suitable technique for optimising a spherical mesh derived from the template head mesh. The resulting mesh exhibits the properties required for efficient application of surface spherical harmonic deformations, a key step in the creation of an MPA database.

Chapter 5

Database creation

> If we knew what it was we were doing, it would not be called research, would it?
>
> ALBERT EINSTEIN

The previous chapter described methods for optimally mapping a boundary element method (BEM)-ready head mesh to a sphere for the application of spherical surface harmonics. This is an essential first step in the production of a morphoacoustic perturbation analysis (MPA) database.

An MPA database consists of two parts, which may be summarised as follows:

1. The set of harmonic deformations. Spherical harmonic deformations up to a specified degree N and order M are applied to the sphere. These deformations are then back-applied to the surface of the template head mesh using the optimised mapping of the template head on a sphere (chapter 4).

2. The set of Δhead-related transfer functions (HRTFs). HRTFs for a chosen set of directions are computed for the template and for each harmonic deformation using the BEM. The spectral difference between the HRTFs for each harmonic deformation and the corresponding HRTFs for the template are stored as ΔHRTFs.

This chapter details the work that has begun on generation of the new MPA database. It starts by focusing on the production of the harmonic deformations (part 1 above). Some background theory is provided on Legendre polynomials, which are an important element in the calculation of spherical harmonics. This is followed by a description of the normalisation technique required for generating deformations with the correct relative amplitudes. The chapter goes on to consider preparations towards the production of ΔHRTFs (part 2 above). These include validation of the BEM solver and definition of the absolute amplitudes of the harmonic deformations required to maintain linearity and to ensure adequate signal-to-noise ratio.

5.1 Harmonic deformations

Provided that the principle of superposition holds, a weighted sum of a rich set of orthogonal surface spherical harmonic components allows arbitrary deformations of the template head mesh to be generated. Both linearity and richness are essential if the MPA technique is to operate correctly. This section explores the process of producing a suitable harmonic series with which to deform the template head.

5.1.1 Legendre polynomials

Legendre polynomials $P_n(x)$ are the solutions to Legendre's differential equation:

$$\frac{d}{dx}\left[(1-x^2)\frac{d}{dx}P_n(x)\right] + n(n+1)P_n(x) = 0 \tag{5.1}$$

which appears in many mathematical and physical situations. The polynomials can be constructed to define a complete set of orthogonal functions over the interval [-1,1] which is where their use in spherical harmonics lies. Arfken and Weber (1995) show that the equation for the polynomials $P_n(x)$ is:

$$P_n(x) = \sum_{k=0}^{n/2}(-1)^k \frac{(2n-2k)!}{2^n k!(n-k)!(n-2k)!}x^{n-2k} \tag{5.2}$$

and can be transformed as follows, for n an integer:

$$P_n(x) = \sum_{k=0}^{n/2}(-1)^r \frac{1}{2^n k!(n-k)!}\left(\frac{d}{dx}\right)^n x^{2n-2k} \tag{5.3a}$$

$$= \frac{1}{2^n n!}\left(\frac{d}{dx}\right)^n \sum_{k=0}^{n}\frac{(-1)^k n!}{k!(n-k)!}x^{2n-2k} \tag{5.3b}$$

The binomial theorem and the extension of the upper limit allow replacement of

the summation by $(x^2-1)^n$:

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n \tag{5.4}$$

which is Rodrigues' formula for the unassociated Legendre polynomials.

The associated Legendre polynomials $P_n^m(x)$ generalise the Legendre polynomials and are solutions to the associated Legendre differential equation. For positive m they are defined as:

$$P_n^m(x) = (-1)^m(1-x^2)^{m/2}\frac{d^m}{dx^m}P_n(x) \tag{5.5}$$

where n is a positive integer and $m = 0, \ldots, l$. Or more explicitly, by combining equations 5.4 and 5.5:

$$P_n^m(x) = (-1)^m\frac{1}{2^n n!}(1-x^2)^{m/2}\frac{d^{n+m}}{dx^{n+m}}(x^2-1)^n \tag{5.6}$$

The associated Legendre polynomials for negative m are then defined by:

$$P_n^{-m}(x) = (-1)^m\frac{(n-m)!}{(n+m)!}P_n^m(x) \tag{5.7}$$

and:

$$P_n^0(x) = P_n(x) \tag{5.8}$$

There are two sign conventions in use, with some authors (Arfken and Weber, 1995) omitting the Condon-Shortley phase term $(-1)^m$ in the definition of the associated Legendre polynomials, instead including it in the definition of the spherical harmonics.

Arfken and Weber (1995) show on pages 726–727 that the orthogonality integral

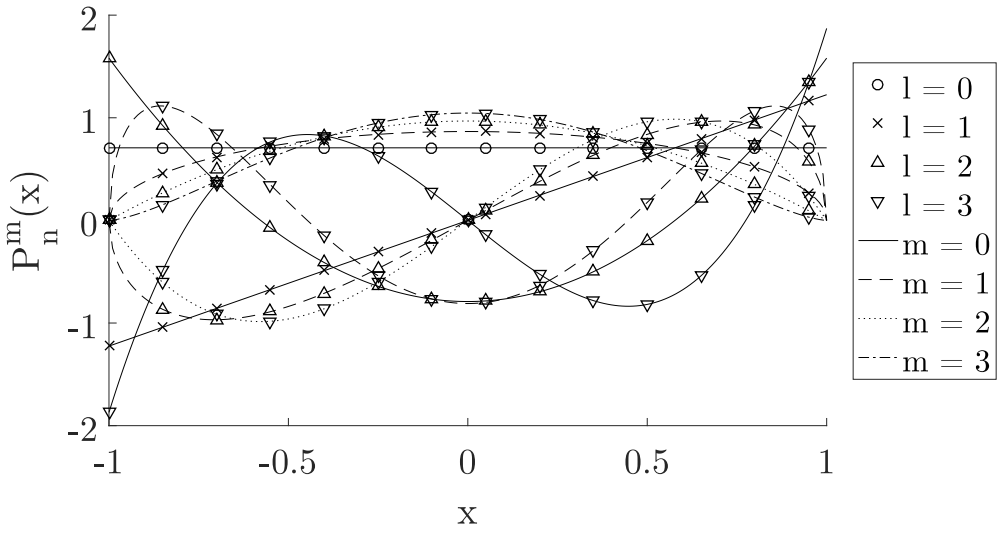


Figure 5.1: Fully normalized associated Legendre polynomials up to degree 3.

for the associated Legendre polynomials is:

$$\int_{-1}^{1} P_p^m(x) P_q^m(x) dx = \frac{2}{2q+1} \cdot \frac{(q+m)!}{(q-m)!} \delta_{p,q} \tag{5.9}$$

which is used later in the definition of spherical harmonics.

The first few associated Legendre Polynomials (shown both in terms of x and with $x = \cos\theta$) are:

$$\begin{aligned}
P_0^0(x) &= 1 && = 1 \\
P_1^0(x) &= x && = \cos\theta \\
P_1^1(x) &= -(1-x^2)^{\frac{1}{2}} && = -\sin\theta \\
P_2^0(x) &= \frac{1}{2}(3x^2-1) && = \frac{1}{2}(3\cos^2\theta - 1) \\
P_2^1(x) &= -3x(1-x^2)^{\frac{1}{2}} && = -3\sin\theta\cos\theta \\
P_2^2(x) &= 3(1-x^2) && = 3\sin^2\theta \\
P_3^0(x) &= \frac{1}{2}x(5x^2-3) && = \frac{1}{2}\cos\theta(5\cos^2\theta - 3) \\
P_3^1(x) &= -\frac{3}{2}(5x^2-1)(1-x^2)^{\frac{1}{2}} && = -\frac{3}{2}(5\cos^2\theta - 1)\sin\theta \\
P_3^2(x) &= 15x(1-x^2) && = 15\cos\theta\sin^2\theta \\
P_3^3(x) &= -15(1-x^2)^{\frac{3}{2}} && = -15\sin^3\theta
\end{aligned}$$

5.1.2 Spherical harmonics

Spherical harmonics satisfy the spherical harmonic differential equation, which is given by the angular part of Laplace's equation in spherical coordinates. In three dimensions, the spherical harmonic differential equation is given by:

$$\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)+\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}+n(n+1)\right]F=0 \tag{5.10}$$

Where θ is the polar (colatitudinal) coordinate $\theta \in [0,\pi]$ and ϕ is the azimuthal (longitudinal) coordinate $\phi \in [0,2\pi]$. Writing $F=\Phi(\phi)\Theta(\theta)$ gives:

$$\frac{\Phi(\phi)}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta(\theta)}{d\theta}\right)+\frac{\Theta(\theta)}{\sin^2\theta}\frac{d^2\Phi(\phi)}{d\phi^2}+n(n+1)\Theta(\theta)\Phi(\phi)=0 \tag{5.11}$$

Multiplying through by $\sin^2\theta/\Phi(\phi)\Theta(\theta)$ allows the separation of the azimuthal (ϕ) and polar (θ) dependences:

$$\left[\frac{\sin\theta}{\Theta(\theta)}\frac{d}{d\theta}\left(\sin\theta\frac{d\Theta(\theta)}{d\theta}\right)+n(n+1)\sin^2\right]+\frac{1}{\Phi(\phi)}\frac{d^2\Phi(\phi)}{d\phi^2}=0 \tag{5.12}$$

The separated azimuthal dependence becomes:

$$\frac{1}{\Phi(\phi)}\frac{d^2\Phi(\phi)}{d\phi^2}=-m^2 \tag{5.13}$$

which has solutions:

$$\Phi(\phi)=Ae^{-im\phi}+Be^{im\phi} \tag{5.14}$$

and satisfies the orthogonality integral:

$$\int_0^{2\pi}e^{-im_1\phi}e^{im_2\phi}d\phi=2\pi\delta_{m_1,m_2} \tag{5.15}$$

From this the following can be derived:

$$\Phi_n = \frac{1}{\sqrt{2\pi}} e^{im\phi} \tag{5.16}$$

which is orthonormal with respect to integration over the azimuth angle (ϕ) (Arfken and Weber, 1995).

Once the azimuthal dependence has been split off the remaining polar dependence is satisfied by the associated Legendre functions. Therefore:

$$\Theta(\theta) = P_n^m(\cos\theta) \tag{5.17}$$

Normalising the associated Legendre function by equation 5.9 leads to the following orthonormal, with respect to polar angle θ, function:

$$\wp_n^m(\cos\theta) = \sqrt{\frac{2n+1}{2}\frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta), \quad -n \leq m \geq n \tag{5.18}$$

The complex spherical harmonics Y_n^m are defined as the product of equation 5.16 and equation 5.18:

$$Y_n^m(\theta,\phi) = N_n^m P_n^m(\cos\theta) e^{im\phi} \tag{5.19}$$

where N_n^m is the normalization coefficient:

$$N_n^m = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} \tag{5.20}$$

and the complete orthogonality integral becomes:

$$\int_{\phi=0}^{2\pi}\int_{\theta=0}^{\pi} Y_{n_1}^{m_1 *}(\theta,\phi) Y_{n_2}^{m_2}(\theta,\phi)\sin\theta d\theta d\phi = \delta_{n_1,n_2}\delta_{m_1,m_2} \tag{5.21}$$

Using Euler's formula the spherical harmonics can be rewritten in terms of real

and imaginary parts as:

$$Y_n^m(\theta,\phi) = N_n^m P_n^m(\cos\theta)(\cos(m\phi) + i\sin(m\phi)) \tag{5.22}$$

$$\Re(Y_n^m) = N_n^m P_n^m(\cos\theta)\cos(m\phi) \tag{5.23}$$

$$\Im(Y_n^m) = N_n^m P_n^m(\cos\theta)\sin(m\phi) \tag{5.24}$$

In most applications of spherical harmonics the functions under consideration are real-valued and therefore only the real-valued spherical harmonics $Y_{nm}(\theta,\phi)$ are considered. These are defined for integers $n \in [0,\infty]$ and $m \in [-n,n]$ as:

$$Y_{nm}(\theta,\phi) = \begin{cases} \sqrt{2}\Im(Y_n^{|m|}) = \sqrt{\frac{2n+1}{2\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\theta)\sin(|m|\phi) & \text{for } m < 0 \\ \sqrt{\frac{2n+1}{4\pi}}P_n^0(\cos\theta) & \text{for } m = 0 \\ \sqrt{2}\Re(Y_n^m) = \sqrt{\frac{2n+1}{2\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)\cos(m\phi) & \text{for } m > 0 \end{cases} \tag{5.25}$$

Figures 5.2 and 5.3 show two common ways of visualising spherical harmonics. In figure 5.2 the radius of the shape is the real spherical harmonic value for the corresponding azimuth (ϕ) and elevation (θ) pair. Grey indicates a positive value, black a negative value. Only the positive orders are plotted for conciseness: the negative orders have the same shape as the positive but are rotated about the z-axis. This can be seen in figure 5.3 where the polarity of the real spherical harmonics are plotted on the surface of the unit sphere. Again grey indicates a positive value, black a negative value.

Spherical harmonics can be split into three classes based on the relationship between the degree n and order m:

- *Zonal* harmonics are spherical harmonics with $m = 0$ and the visual curves that appear on the surface of the unit sphere divide it into latitudinal sections. These are the first column in figure 5.2 and the middle column in figure 5.3.
- *Sectorial* harmonics are of the form $|m| = n$ and divide the unit sphere into

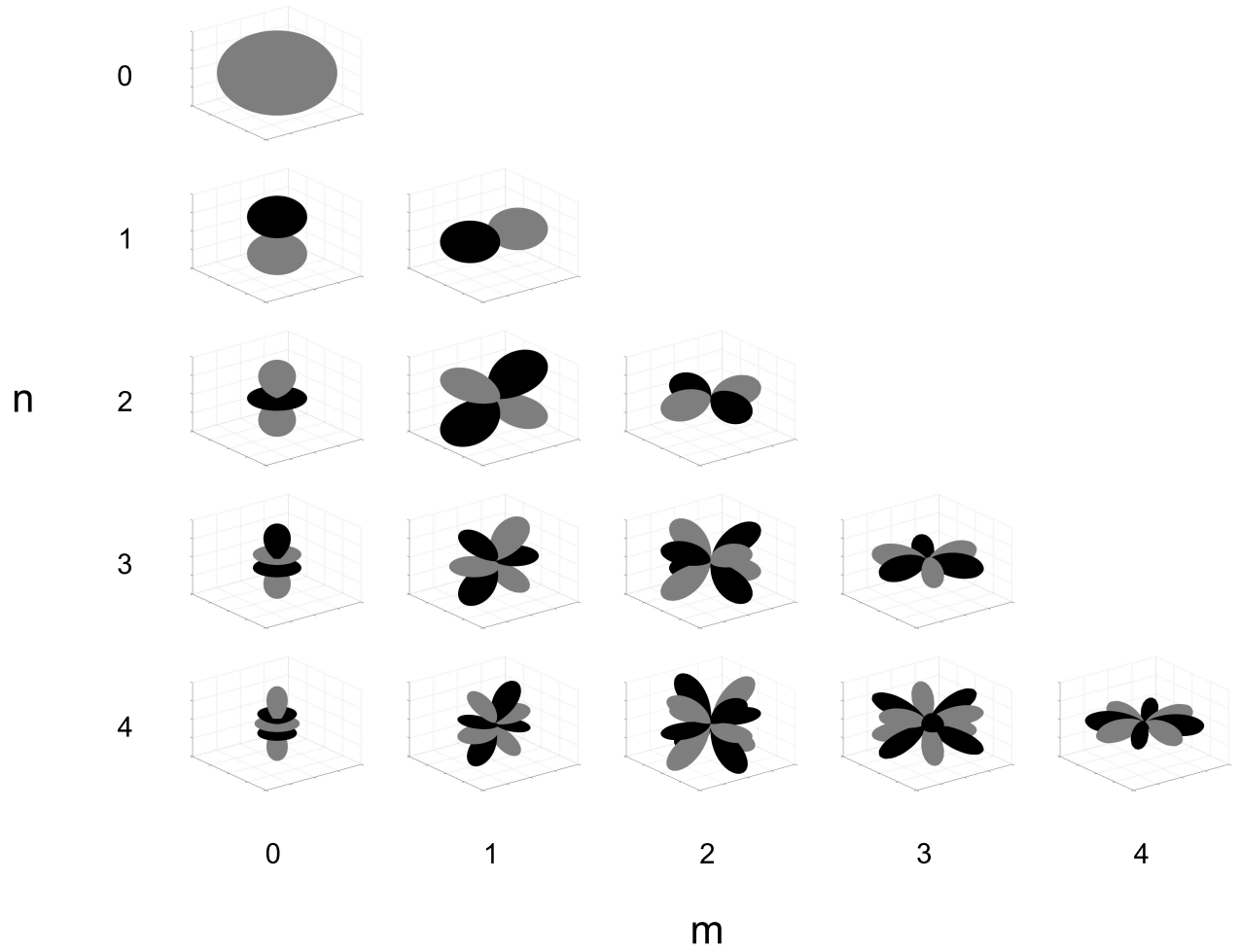


Figure 5.2: Real spherical harmonics for degree $n \in [0, 4]$ and order $m \in [0, n]$. Grey indicates positive values, black negative. Note that only the spherical harmonics for positive values of m are shown because, as demonstrated in figure 5.3, the harmonics for negative values of m are simply rotated versions of the corresponding positive values.

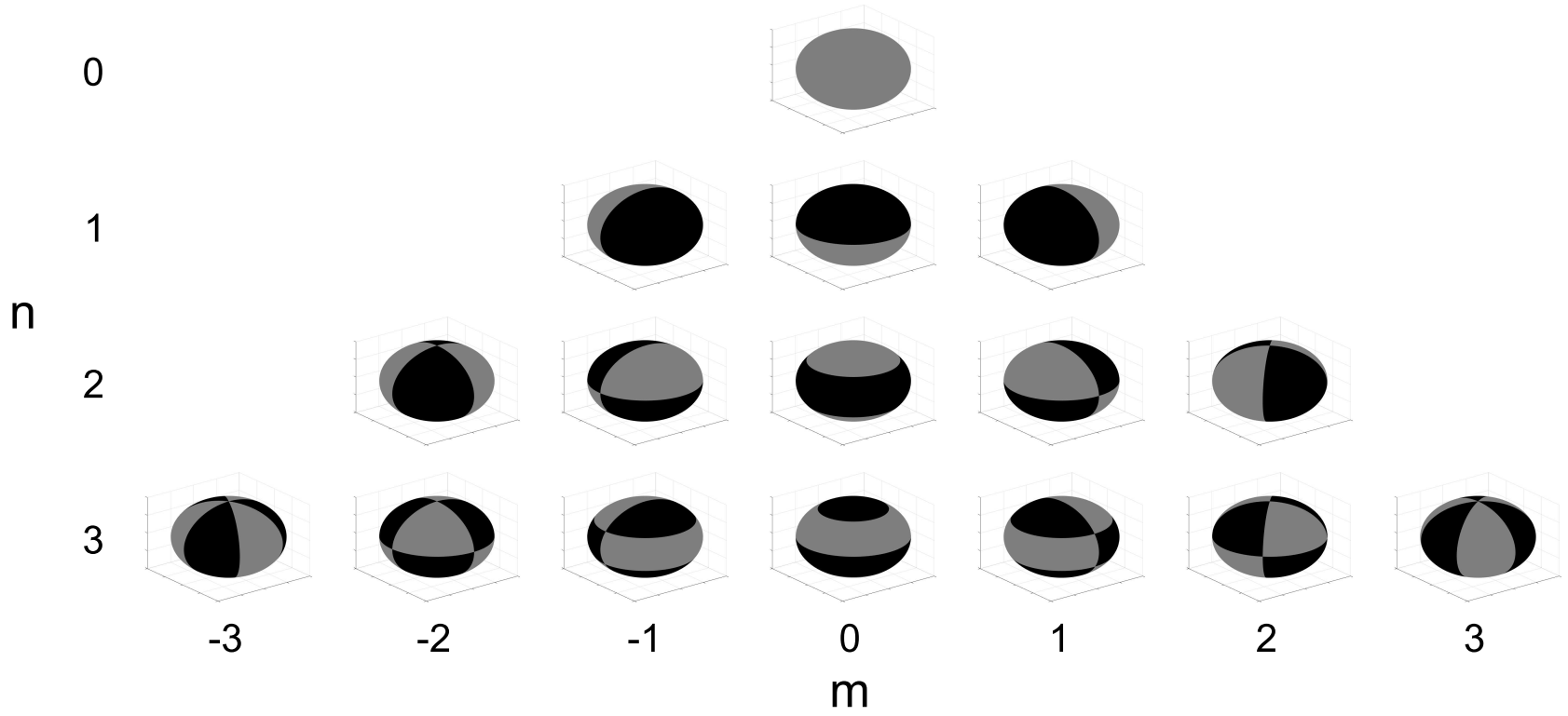


Figure 5.3: Real spherical harmonic polarity patterns for degree $n \in [0, 3]$ and order $m \in [-n, n]$, plotted on the surface of the unit sphere. Grey indicates positive values, black negative.

latitudinal sections. These are the diagonals in figures 5.2 and 5.3.

- *Tesseral* harmonics are all other spherical harmonics where $|m| \neq n$ and divide the unit sphere into blocks of both latitude and longitude.

Any arbitrary spherical function $f(\theta, \phi)$ can be expanded as a sum of spherical harmonics:

$$f(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta, \phi) \tag{5.26}$$

where C_{nm} are the spherical harmonic coefficients and can be computed by projecting the function $f(\theta, \phi)$ onto the basis functions $Y_{nm}(\theta, \phi)$:

$$C_{nm} = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) Y_{nm}(\theta, \phi) \sin\theta d\theta d\phi \tag{5.27}$$

As with other harmonic series, the expansion is only exact if it is infinite, and in practice the series is truncated to some upper limit N and only an approximation is found:

$$\hat{f}(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} C_{nm} Y_{nm}(\theta, \phi) \tag{5.28}$$

5.1.3 Spherical harmonic deformations

There are various normalisation techniques for spherical harmonics, depending on their intended use. The normalisation coefficient given in equation 5.20 normalises the spherical harmonics so that:

$$\int |Y_n^m|^2 d\Omega = 1 \tag{5.29}$$

However, for the generation of the MPA database the spherical harmonics are required to be normalised so that the morphological deformations are in the region where they are linearly related to the resulting acoustic pressure changes.

This requires normalisation of the maximum value of each spherical harmonic:

$$max(|Y_n^m|) = 1 \tag{5.30}$$

A similar normalisation approach is used in some Ambisonics systems, where it is known as max-normalisation (maxN), or, with an additional $1/\sqrt{2}$ scaling of the W signal, Furse-Malham (FuMa) normalisation (Daniel, 2001; Malham, 2003). However, as Carpentier (2017) notes, there is no closed-form or recursive solution to calculating the normalisation factors for maxN. The maxN normalisation coefficients, up to degree 3, given by Malham (2003) were merely calculated by inspection. Carpentier (2017) reports conversion factors from fully normalised spherical harmonics, for which there exist recursive approaches to generation (Holmes and Featherstone, 2002), but only up to spherical harmonic degree 16. The generation of the MPA database requires much higher degree spherical harmonics than this.

As shown in equation 5.25 the general form for the real-valued spherical harmonics is:

$$Y_{nm}(\theta, \phi) = \begin{cases} N_n^m P_n^{|m|}(\cos\theta)\sin(|m|\phi) & \text{for } m < 0 \\ N_n^m P_n^m(\cos\theta)\cos(m\phi) & \text{for } m \geq 0 \end{cases} \tag{5.31}$$

where N_n^m is the chosen normalisation coefficient. Therefore to normalise the maximum value of the spherical harmonic the maximum value of P_n^m is needed; since $\sin(|m|\phi)$ and $\cos(m\phi)$ are bounded to $\pm$ one:

$$N_n^m = max(|P_n^m(\cos\theta)|) \tag{5.32}$$

For generation of the MPA database the maximum value of $P_n^m(\cos\theta)$, up to degree 100, was calculated by evaluating the associated Legendre polynomials for 1,000,000 linearly spaced points between minus one and plus one (the bounds of $\cos\theta$). It was found experimentally that increasing the number of points from

10,000 to 100,000 resulted in changes of the maximum value of up to 3.5% whilst increasing the number of points from 100,000 to 1,000,000 resulted in a maximum change of 0.08%. Therefore it was deemed that 1,000,000 points was adequate for calculating accurate maximum values, since the changes were all less than 0.1%.

This allows the maximum value of the spherical harmonics to be normalised to any value. Section 5.3 considers verification and selection of the maximum size of deformations used in the database creation.

5.2 Experimental validation of BEM results

MPA uses BEM simulations (section 2.4.6) to calculate the HRTFs for the template and perturbed heads and the decision was made to use Mesh2HRTF (Ziegelwanger *et al.*, 2015a) as the BEM solver. Despite the existence of published results of simulations using Mesh2HRTF (Ziegelwanger *et al.*, 2013, 2014, 2015b, 2016), the decision was made to carry out experimental validation of the software to check that it was correctly configured. The acoustic scattering of a rigid sphere was chosen for this validation as there is a long-standing analytical solution (Strutt, 1877; Rayleigh, 1904; Duda and Martens, 1998). When using the analytical solution as the ground truth, the code in appendix B of Duda and Martens (1998) was used.

For the simulations Autodesk 3ds Max® was used to generate a high resolution spherical mesh. The mesh consisted of 294,546 vertices and 112,500 faces, with a maximum edge length of 2.064 mm. Assuming a $\lambda/6$ rule, this gives a maximum valid simulation frequency of 27.6 kHz. The radius of the sphere was set to 10.9 cm which is the radius of the bowling ball Duda and Martens (1998) used to experimentally validate their range-dependent analytical solution. The principle of reciprocity was used, with the source created by applying a velocity boundary condition to the face of the sphere through which the positive x-axis passed (i.e. 0° azimuth) and receivers placed at 5° steps around the horizontal plane at a

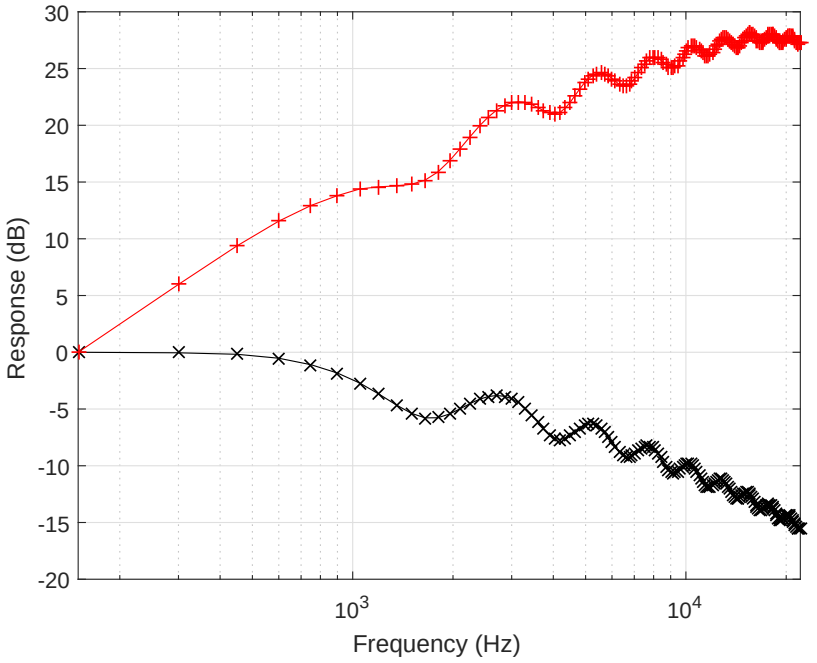


Figure 5.4: Raw simulation results of acoustic scattering by a rigid sphere (red '+' signs) compared to analytical solution (black 'x's) for a source at 145° azimuth and 1 m radial distance showing a 20 dB per decade increase in simulation response. Responses have been aligned to 0 dB at 150 Hz to aid visualisation.

radial distance of 1 m. BEM simulations were carried out in 150 Hz steps from 150 Hz to 22.05 kHz.

The simulation results obtained directly from the solver exhibited a 20 dB per decade increase in their response compared to the analytical solution (figure 5.4). The authors of Mesh2HRTF were consulted and this scaling of the responses with frequency was attributed to the relationship between the velocity potential, ϕ, produced by the equation system in the Mesh2HRTF BEM solver and the sound pressure, p, calculated in post-processing via the equation:

$$p = i\rho\omega\phi \tag{5.33}$$

where ρ is the density of air, i the imaginary unit and ω the circular frequency. Therefore the simulation results were multiplied by a -20 dB per decade scaling factor to align them more closely with the analytical solution, as shown in figure 5.5.

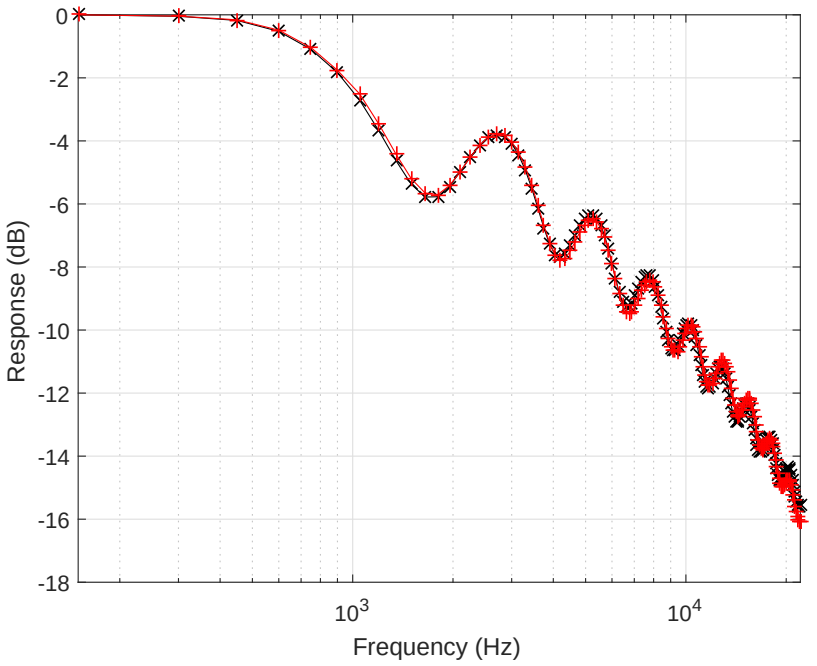


Figure 5.5: Simulation results of acoustic scattering by a rigid sphere, multiplied by a -20 dB per decade scaling (red '+' signs) compared to analytical solution (black 'x's) for a source at 145° azimuth and 1 m radial distance. Responses have been aligned to 0 dB at 150 Hz to aid comparison.

However, after application of the frequency-dependent scaling, there remained an overall level offset between the simulation responses and the analytical solutions as shown in figure 5.6. This overall level difference is likely due to the area of the mesh face used as the source, as this affects the excitation energy in the simulation (see section 5.3.1 for further discussion of this effect). The numerical difference between simulated and analytical pressures at 1 kHz in the direction 0° azimuth and 0° elevation was found to be 4.93 dB and a positive shift of this magnitude was applied to all simulation results (figure 5.7).

After application of the frequency dependent scaling and the overall level shift the simulation results gave good correspondence with the theoretical results. The overall mean squared error across all directions and frequencies was 0.0393 dB, with a maximum error of 1.95 dB at 13.2 kHz for a source at 175° azimuth. Figure 5.8 displays the magnified region of figure 5.7 in which the largest errors occur, and figure 5.9 shows the variation of the squared error across all frequencies and directions. It can be seen that the larger errors are concentrated towards higher

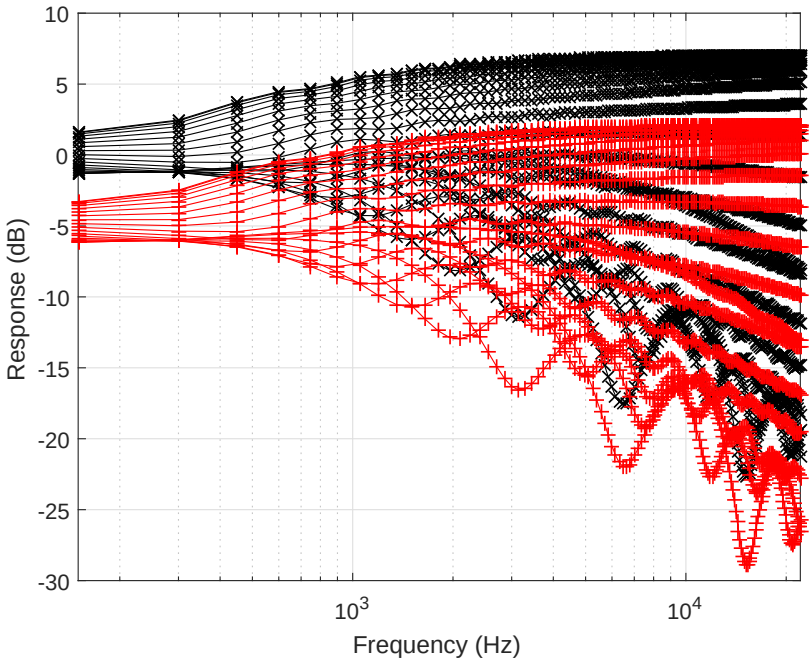


Figure 5.6: Comparison of simulation (red '+' signs) and theoretical results (black 'x's) showing overall level offset.

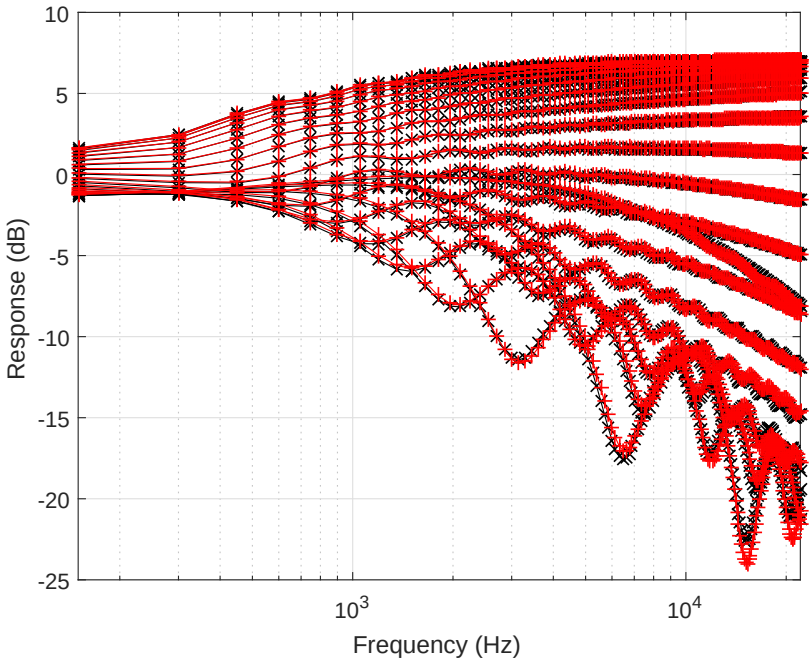


Figure 5.7: Comparison of simulation (red '+' signs) and theoretical results (black 'x's) with +4.93 dB offset applied to all simulation results.

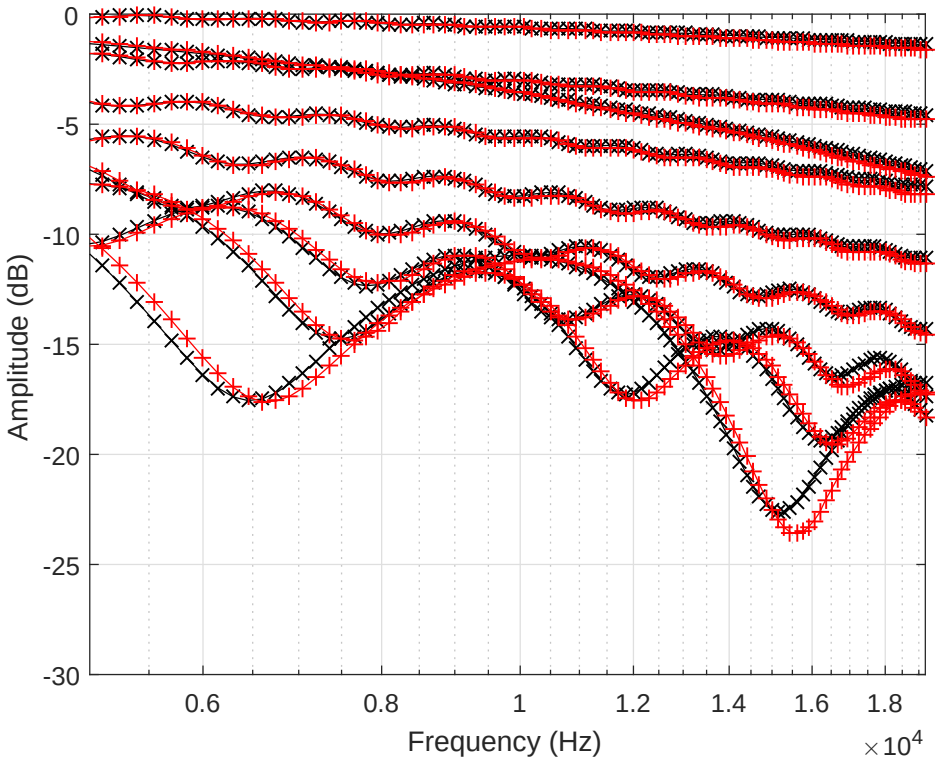


Figure 5.8: As figure 5.7, but zoomed in on region of largest error between simulation and theoretical results.

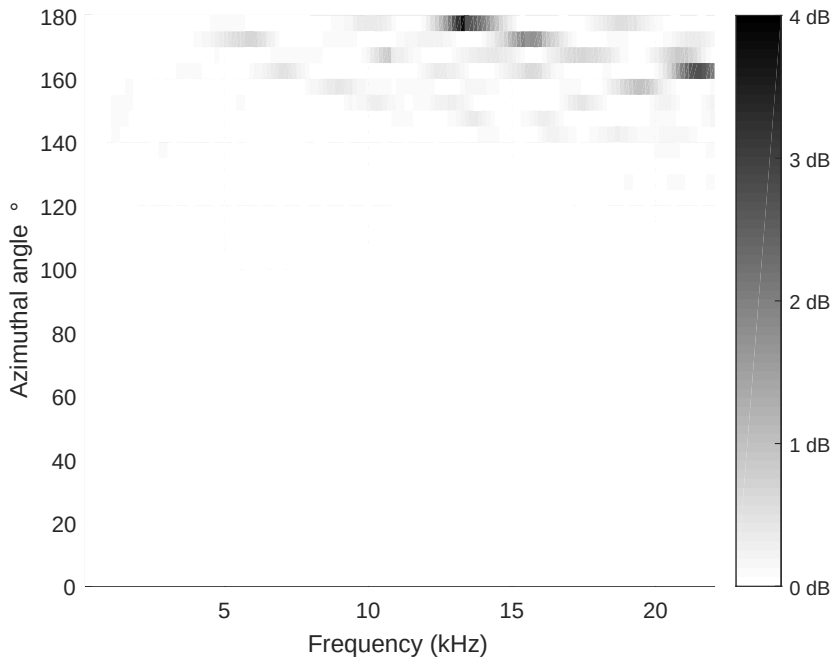


Figure 5.9: Squared error between simulation and theoretical results across azimuth angle and frequency.

frequencies and directions to the rear of the sphere where pressure levels are at their lowest and diffraction effects strongest. It should be noted that for these

directions and frequencies, the sound pressure levels are already very low and therefore errors appear larger in a decibel sense. In an absolute pressure error sense, the errors are comparable across all directions and frequencies. Therefore, it was decided that the overall mean squared error of 0.0393 dB was acceptable and the decision was made to adopt Mesh2HRTF as the main BEM computation tool in this research.

5.3 Amplitude of deformations

The MPA database stores the effects on an HRTF for the template head mesh of independently applying $N + M$ harmonic deformations to the mesh. Valid use of the database depends on each acoustic change being linearly related to the harmonic deformation which caused it. In these circumstances, the principle of superposition holds and the overall acoustic change caused by a weighted sum of harmonic deformations may be calculated by summing the acoustic changes for each harmonic deformation using the same set of weights. Hence, for MPA a critical trade-off exists between keeping the amplitude of the harmonic deformations small enough to ensure a satisfactory degree of linearity between a shape change and the corresponding acoustic change, yet large enough that the acoustic changes are above the noise floor of the BEM solver (Thorpe, 2009; Tew *et al.*, 2012). Therefore it was necessary to evaluate, firstly, the linearity of the relationship between deformations and acoustic changes, and secondly, the noise floor of the BEM solver.

5.3.1 Linearity

To assess the degree of linearity between deformations and acoustic changes the same spherical mesh was used as in section 5.2. This mesh was chosen because, as in section 5.2, the results could easily be compared to the analytical solution. Initially the harmonic deformations were kept simple by increasing the radius of

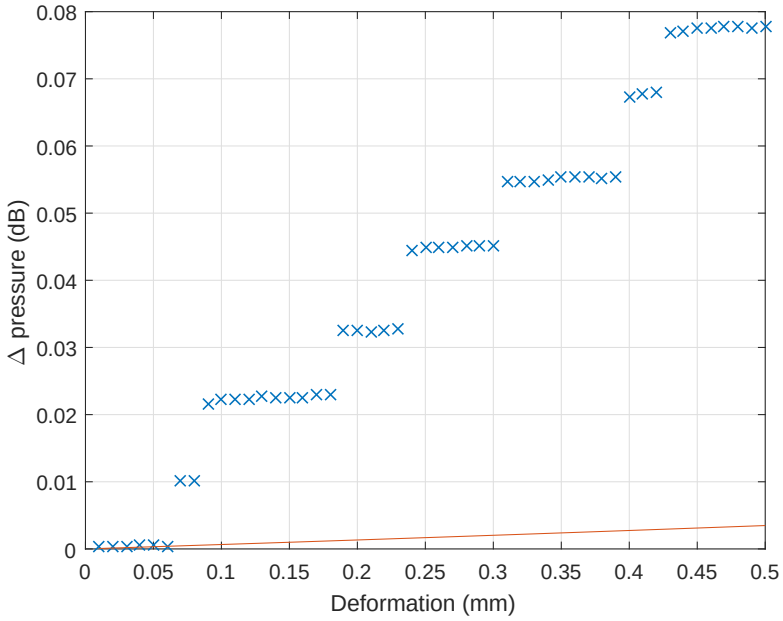


Figure 5.10: Comparison between simulation results (blue 'x's) and analytical results (solid orange line) for the change in pressure (Δ pressure) introduced by increasing the radius of a sphere.

the sphere in 0.01 mm steps from 0 mm to 0.5 mm. This is equivalent to a zeroth degree, zeroth order spherical harmonic. Figure 5.10 shows the initial simulation results ('x's) compared to the analytical solution (solid line). The results are plotted as delta pressures, relative to the results of the original, undeformed, spherical mesh.

It can be seen that the simulation results exhibit a succession of steps which increasingly deviate from the analytical results. However, as shown in figure 5.11, each step more closely approximates the analytical solution if it is shifted to align with the first data point of each step. After scrutinising the Mesh2HRTF code it was found that the original code did not read in the mesh coordinates at full resolution, it only read in six decimal places, which, considering a 10.9 cm sphere with coordinates expressed in metres offers a maximum of only six significant figures (s.f.s). Figure 5.12 shows the effect of increasing the precision with which Mesh2HRTF reads in the mesh. Increasing the precision from six to eight s.f.s (i.e. by 100 times) results in a marked reduction in step size. As expected, improving it by a further 100 times, from eight to ten s.f.s yields a

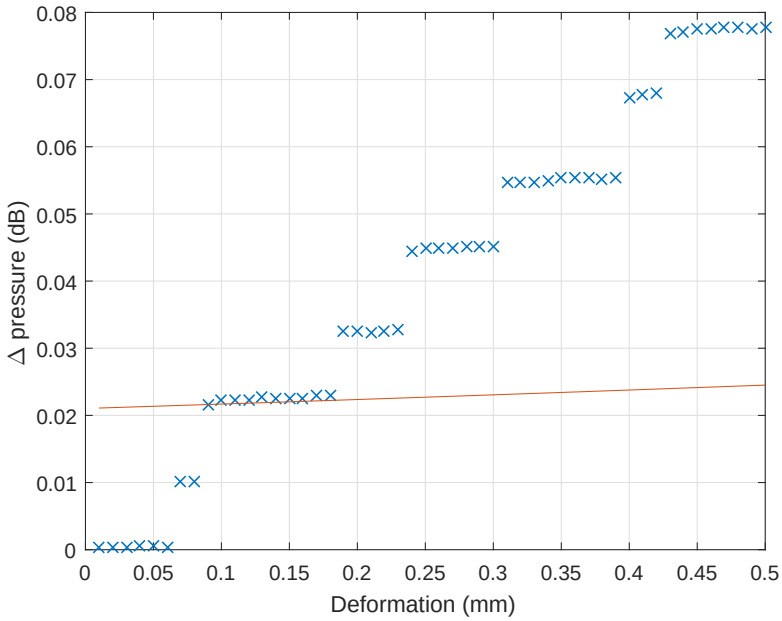


Figure 5.11: Comparison between simulation results (blue 'x's) and analytical results (solid orange line) for the change in pressure (Δ pressure) introduced by increasing the radius of a sphere. The analytical results have been shifted to align with the first point of the third "step" to demonstrate that each subsequent point in the step closely follows the analytical solution.

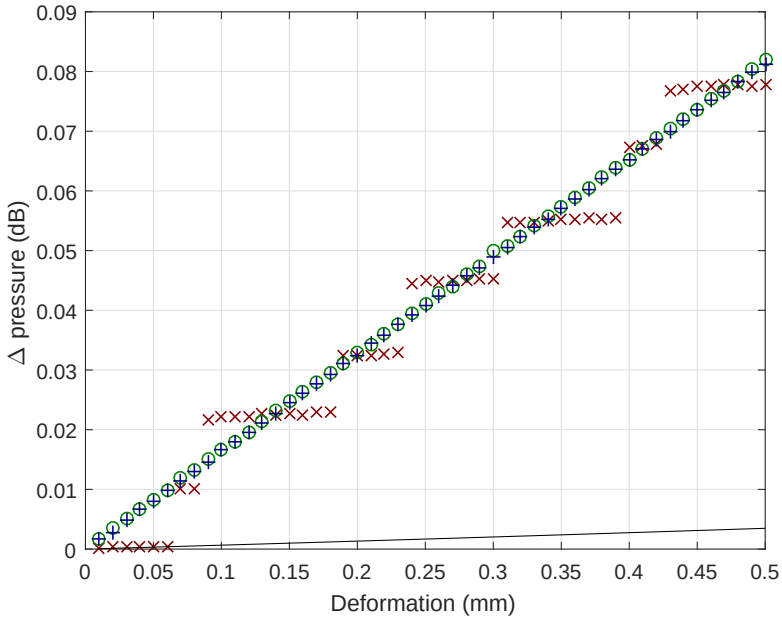


Figure 5.12: Simulation results with spherical mesh coordinates imported into Mesh2HRTF to six (red 'x's), eight (green 'o's) and ten (blue '+'s) decimal places. The black line shows the analytical solution for reference.

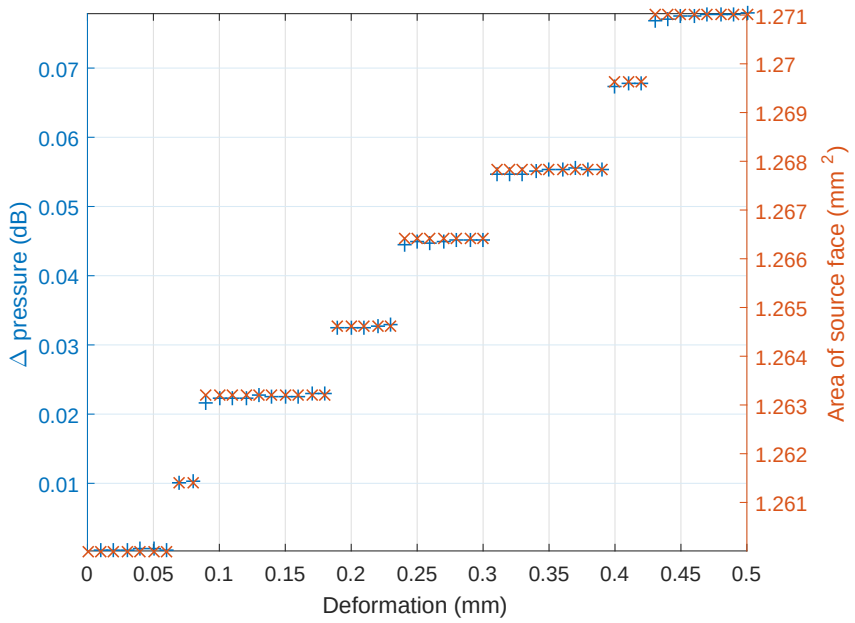


Figure 5.13: Comparison between Δ pressures (blue '+'s, left axis) and area of the source face (orange 'x's, right axis) with spherical mesh coordinates imported into Mesh2HRTF to six decimal places demonstrating how the steps in the Δ pressures correspond to changes in the area of the source face.

correspondingly smaller reduction in error from this cause. Whilst increasing the resolution removes the steps from the simulation results, it still doesn't remove the linearly increasing deviation from the analytical solution with increasing radius.

It was then realised that one possible source of this deviation was the fact that a source in Mesh2HRTF is generally implemented as a vibrating face in the mesh using velocity boundary conditions. Therefore, as the radius of the spherical mesh is increased the effective size of the source also increases, making the source become gradually more energetic. This could explain the increasing deviation of the simulation results from the analytical solution. To test this explanation, the mesh coordinates were read in using a precision of only six decimal places and the simulation results were compared to the area of the source face for each sphere of increasing radius. In figure 5.13 it can be seen that the radius values where steps occur in the simulation results correspond to the step transitions in the area of the source face.

Figure 5.14 shows the results of the linearity test with the Mesh2HRTF code

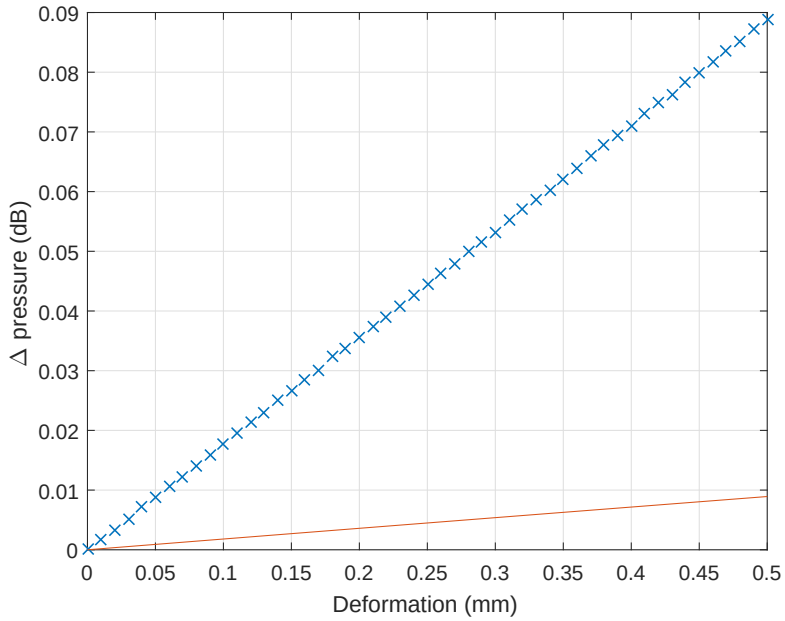


Figure 5.14: Simulation results (blue 'x's) with the spherical mesh coordinates imported into Mesh2HRTF at full resolution, compared to the analytical results (orange solid line) with no compensation for source face area.

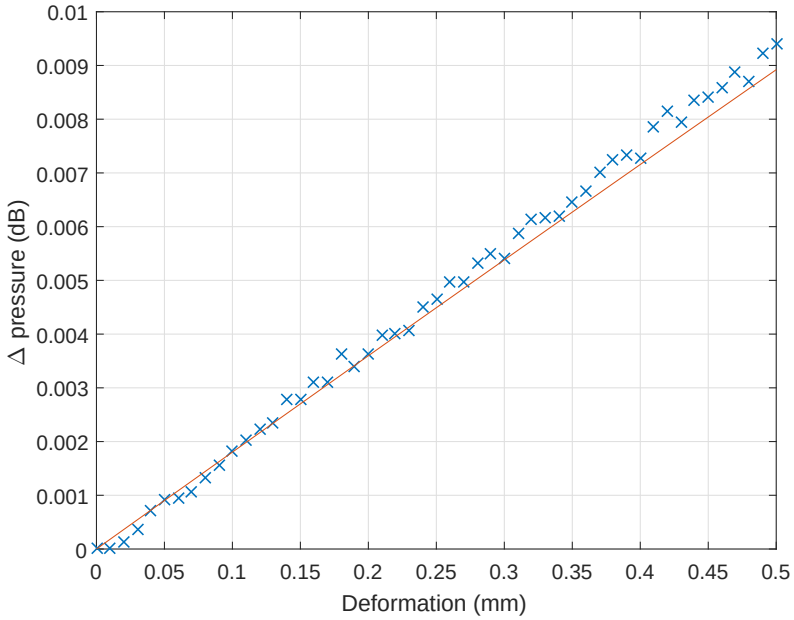


Figure 5.15: Simulation results (blue 'x's) with the spherical mesh coordinates imported into Mesh2HRTF at full resolution, compared to the analytical results (orange solid line) with compensation for source face area.

updated to read in the mesh data at full double precision and figure 5.15 shows the same results but scaled by the ratio of the area of the source face of each deformed sphere to the area of the source face of the undeformed sphere. It can be seen that after scaling by the relative area of the source face there is good correspondence between the simulation and analytical results. Figure 5.15, however, indicates that at least two further sources of error remain. Firstly, there appear to be random variations about the mean slope in the computed delta pressures. Secondly, there appears to be a systematic increase in error between the computed and analytic delta pressures as the radius of the sphere is increased.

The random variations are of sufficient amplitude to cause concern and need further investigation. Figure 5.16 compares simulation results to the analytical solution with the radius of the sphere scaled by factors ranging from 0.25 to 8, representing deformations between -0.0818 and 0.7630 mm (for the 10.9 cm radius sphere). What appears as a systematic deviation in the results in figure 5.15 is revealed in figure 5.16 to be a smoothly varying curve for this larger range of deformation amplitudes. The curve is likely to be an alternative manifestation of the errors in the level and phase of the ripples, just visible in figures 5.5 and 5.7, rather than computation noise. Such errors were deemed to be acceptable in section 5.2.

For this zero-order spherical harmonic deformation, the relationship between deformation amplitude and acoustic pressure change, as shown in figure 5.13 appears to be linear across the whole range of 0.01 mm to 0.5 mm. An analysis of a selection of much higher-order harmonic deformations is needed to establish a limit on deformation amplitude in general. Such an analysis was conducted by Thorpe (2009), who concluded that a maximum deformation amplitude of 0.3 mm is acceptable. This figure was successfully applied in the MPA database described in Tew *et al.* (2012). To provide a further safety margin, we adopt a deformation limit of 0.1 mm in the work presented in this thesis.

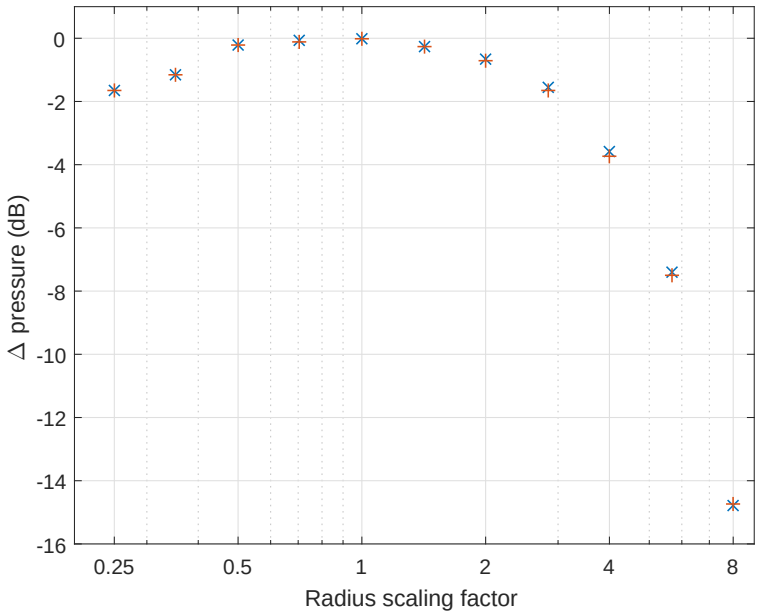


Figure 5.16: Simulation results (blue 'x's) and analytical results (orange '+'s) for much larger deformations.

5.3.2 Signal-to-noise ratio (SNR)

After deciding on a proposed maximum deformation amplitude of 0.1 mm it was necessary to check that the acoustic changes introduced by deformations of this size are well above the noise floor of the BEM solver.

For each spherical harmonic degree, n, from 0–100, a random order, m, where $m \in [-n, n]$, was chosen. The amplitude of the spherical harmonic of this degree and order, normalised to a peak value of 0.1 mm, was calculated for each vertex in the *spherical head mesh* generated in section 4.5. The normal for each vertex in the *original head mesh* was determined by calculating the average of the normals of the faces connected to it. Finally the harmonic deformation was applied by moving each vertex in the head mesh by the amplitude of the deformation associated with the corresponding vertex in the spherical mesh.

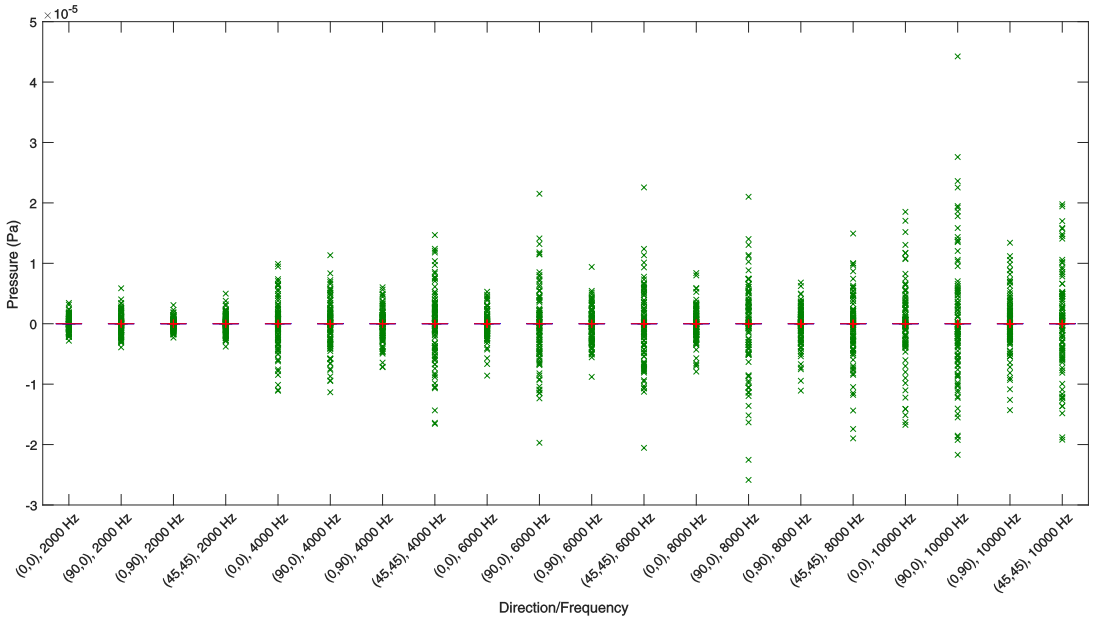
Simulations were run for the 101 deformed meshes for the five frequencies 2, 4, 6, 8 and 10 kHz and four directions (0° azimuth, 0° elevation), (90° azimuth, 0° elevation), (0° azimuth, 90° elevation), (45° azimuth, 45° elevation). These frequencies were chosen to span the frequency range of most interest for pinna

cues (section 2.2.4.3) and the directions were chosen to be very different. However, it would be beneficial to verify the signal-to-noise ratio for additional directions, as four directions is likely not enough. In addition 101 simulations of the template head mesh were run to give an indication of the innate computation noise in the BEM solver.

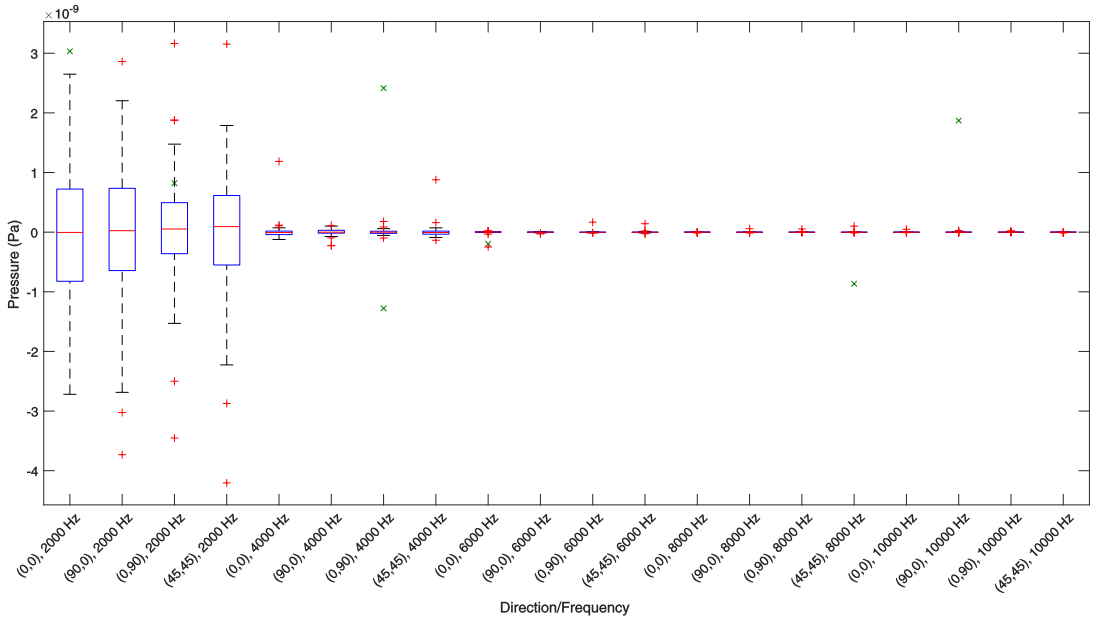
Figure 5.17 shows the results of these simulations. The green crosses represent the delta pressures of the deformed meshes, whilst the box plots represent the distributions of the 101 template mesh results with their mean removed, i.e. the noise in the solver. Figure 5.17a has the y-axis scaled to the delta pressures of the deformed meshes whilst figure 5.17b has the y-axis scaled to accommodate the limits of the box plots of the noise in the system. The figures show that the spread in delta pressures due to the deformations plus computation noise is several orders of magnitude greater than the spread in delta pressures due to computation noise alone.

The histogram in figure 5.18 shows the distribution of the 101 delta pressure values computed at 6 kHz in the direction (45° azimuth, 45° elevation). Its approximately Gaussian shape is indicative of the distributions for the other 100 delta pressure sets in figure 5.17. The fact that a small proportion of the delta pressures in each set lies in the noise floor (as seen in figure 5.17b) is therefore to be expected.

As a further check, a pseudo signal-to-noise ratio (SNR) was calculated by computing the ratio of the variance in the deformation delta pressures to the variance in the template results. This was calculated both as a function of direction and of frequency and the results are given in table 5.1. The SNR exhibits a small variation of approximately 5 dB with direction, but there is a variation of almost 65 dB with frequency. However, all deformations tested exhibited an SNR of over 60 dB, confirming that, for these harmonics, frequencies and directions, and using a peak deformation amplitude of 0.1 mm, computation noise is a negligible source of error.



(a)



(b)

Figure 5.17: Δ pressures (green 'x's) between harmonically deformed head mesh results and the template head mesh results alongside box-plots of 101 repeated simulations of the template head mesh with the average template result removed. (b) same as (a) but with y-axis expanded to show the box-plots.

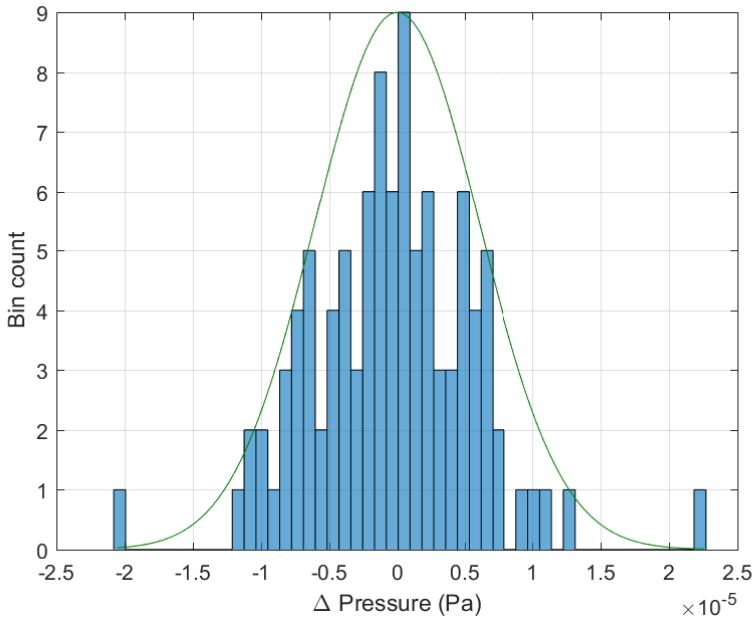


Figure 5.18: Histogram of Δ pressures at 6 kHz for 45° azimuth, 45° elevation, with Gaussian distribution (green curve) plotted on top.

Direction	SNR
(0,0)	77.8688 dB
(90,0)	82.6028 dB
(0,90)	78.3116 dB
(45,45)	82.2657 dB
Frequency	**SNR**
2000 Hz	62.7558 dB
4000 Hz	94.1029 dB
6000 Hz	108.604 dB
8000 Hz	117.512 dB
10000 Hz	127.452 dB

Table 5.1: Signal-to-noise ratio (SNR) for Δ pressures as a function of direction and frequency.

5.4 Generation of evaluation grid

A critical choice which had to be made before the MPA database could be computed concerned the directions in which the HRTFs for the template head and its spherical harmonic deformations would be calculated. This choice is influenced by the nature of the analyses likely to be performed using the database; such as whether studying variations in morphoacoustic properties will be concentrated in

certain directions and at particular distances. Whilst the number of evaluation grid points does not have as large an impact on BEM simulation time as the number of head mesh points, increasing the number of evaluation grid points nevertheless leads to an increase in post-processing required, as well as to an increase in storage required for the results.

The first set of points considered was a 1.5 m radius sphere around the head mesh for the study of far-field HRTF cues. However, rather than constant angular spacing of azimuth and elevation on the surface of the sphere, which gives points whose

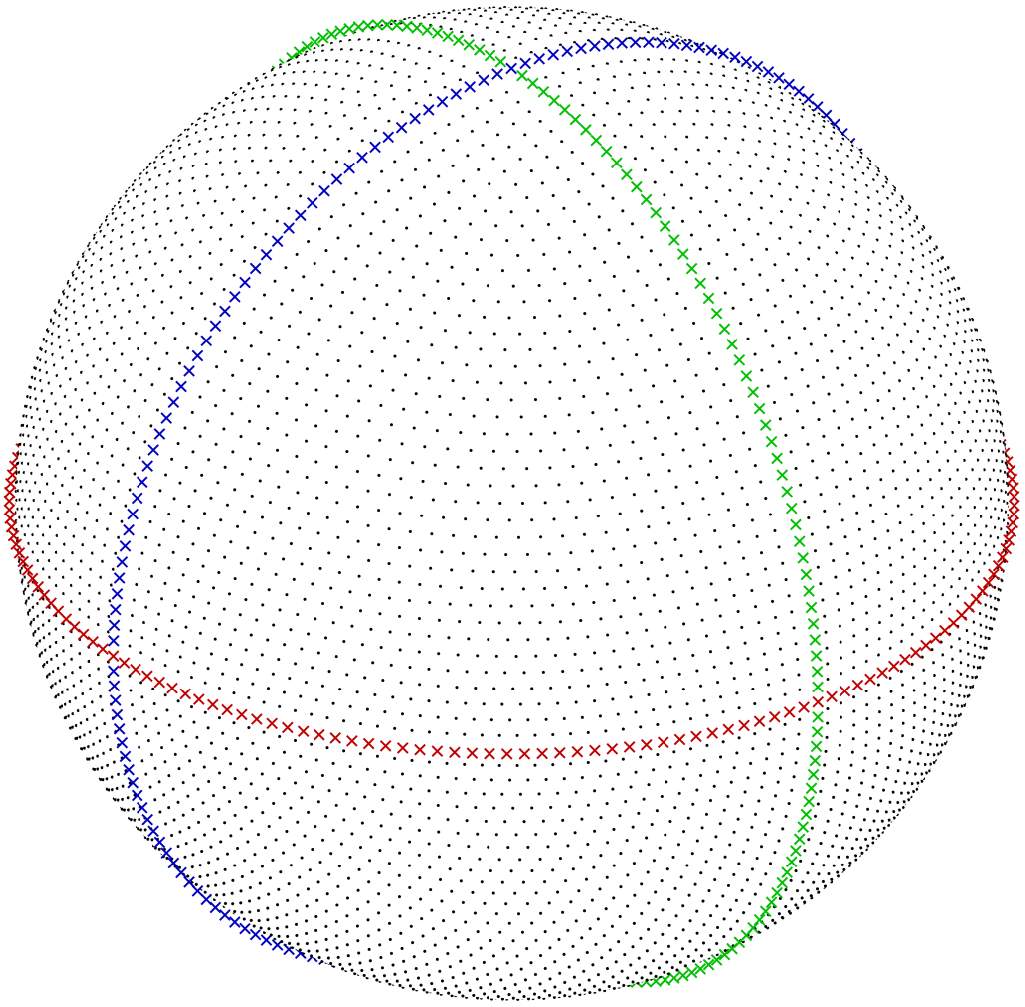
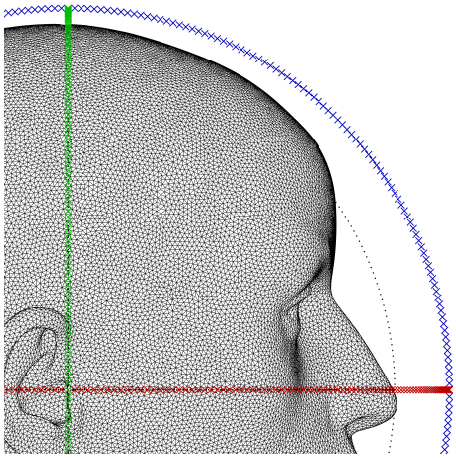


Figure 5.19: Spherical far-field HRTF evaluation grid points for the BEM simulations showing approximately constant linear distribution of the points on the sphere. The points lying on each of the three principal planes are highlighted as coloured 'x's (median (blue), horizontal (red) and frontal (green)) and they have been set slightly proud of the surface of the sphere for increased clarity.

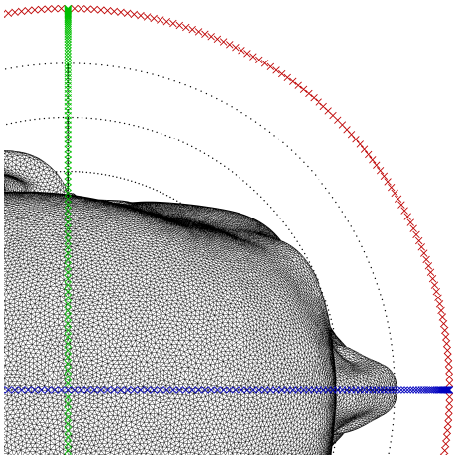
linear separation reduces with increasing elevation, an approximately equal linear distribution of points was generated. The elevation steps were set to a constant 2° whilst the number of azimuthal points for each elevation step was reduced approximately linearly from 360 points at 0° elevation, giving 2° resolution in the horizontal plane, to a single point at +/-90° elevation. This produces constant 2° angular steps in each of the three principal (great circle) planes: median, horizontal and frontal. Figure 5.19 shows the spherical point cloud with each of the three principal planes' points highlighted as coloured 'x's.

To facilitate future studies of the morphological origin of acoustical distance cues, three additional sets of concentric rings of evaluation points were generated; one set for each of the three principal planes. The angular step around each ring was set at 2° and the radii in a set were incremented in steps of 5 cm. The radius of the innermost ring for each plane was set at the first multiple of 5 cm which lay fully outside the head mesh and the radius of the outermost ring was set at 3 m. The rings of radius 1.5 m were removed as these points were already included in the spherical evaluation grid. Figure 5.20 shows the innermost rings of each of the three planes and figure 5.21 shows the complete set of rings for all three planes.

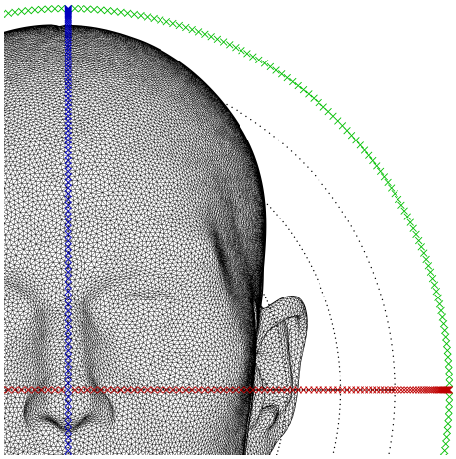
Finally the evaluation points of the sphere and the three planes were combined into a single set of evaluation points. Due to the absence of a torso from the Knowles Electronics manikin for acoustic research (KEMAR) head mesh, the accuracy of HRTFs for source directions at increasingly negative elevations will deteriorate. A more thorough analysis is required to verify exactly what the minimum valid elevation for the mesh is, but points below -45° are thought to be almost certainly invalid and so were removed from the evaluation grid. The final set of 32,739 evaluation grid points is shown in figure 5.22.



(a)

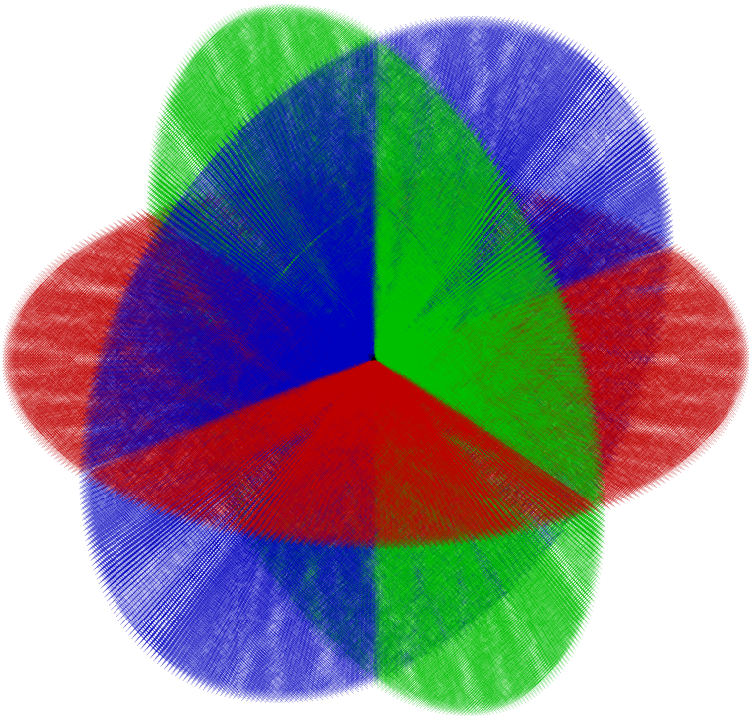


(b)

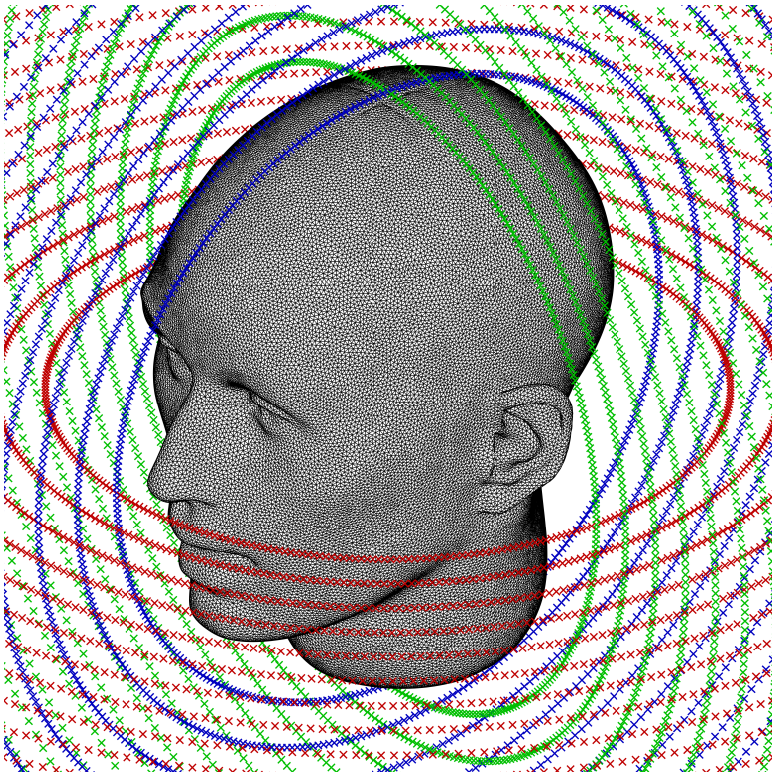


(c)

Figure 5.20: Innermost BEM simulation HRTF evaluation grid points for the three principal planes: (a) median, (b) horizontal and (c) frontal. The innermost rings used are highlighted as coloured 'x's, whilst the black points pass through the head mesh and so were discarded. Each plot is zoomed in on a single quadrant of auditory space to facilitate visualisation.



(a) BEM simulation HRTF evaluation grid points for the three principal planes.



(b) As (a) but zoomed in on head mesh.

Figure 5.21: BEM simulation HRTF evaluation grid points for all three principal planes: median plane points (blue), horizontal plane points (red) and frontal plane points (green).

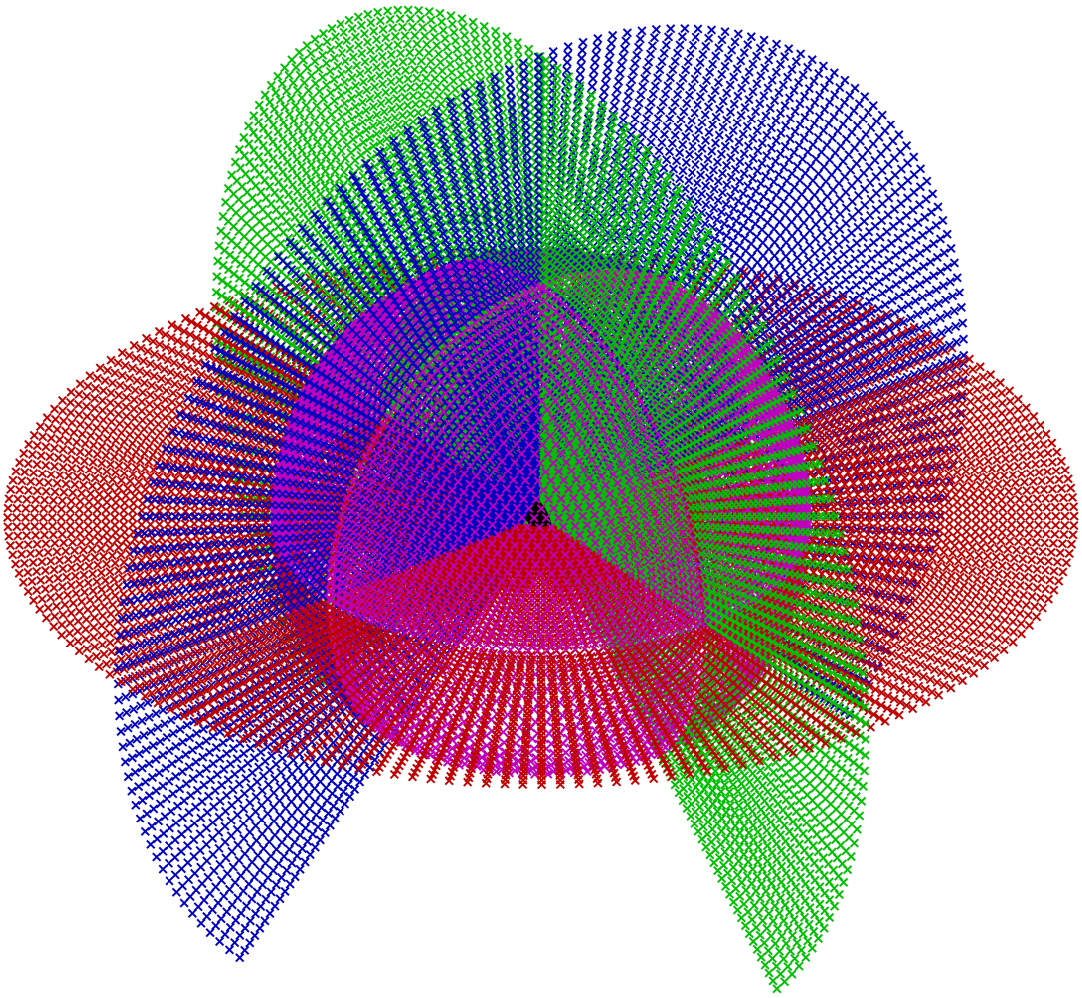


Figure 5.22: The final set of BEM simulation HRTF evaluation grid points. The points are coloured as follows: spherical far-field points (magenta), median plane points (blue), horizontal plane points (red) and frontal plane points (green).

5.5 Summary

Chapter 4 presented novel improvements to first generation morphoacoustic perturbation analysis (MPA) and this chapter has reported the work based on these improvements that has been carried out in preparation for generating a second generation MPA database.

Firstly, the formulation of the spherical harmonic deformations for application to the optimised spherical head mesh resulting from chapter 4 was presented in section 5.1. Because the deformations of interest are real-valued, only the real-

valued spherical harmonics are used. The spherical harmonics were normalised to a maximum absolute value of one, so that the desired peak deformation amplitude is simply achieved by multiplying the harmonic functions by this figure. This is similar to the max-normalisation (maxN) normalisation technique used in Ambisonics (Daniel, 2001; Malham, 2003). However, because no closed-form or recursive solution exists for this normalisation technique, the normalisation factors had to be calculated empirically.

The performance of the BEM solver chosen for generating the MPA database, Mesh2HRTF (Ziegelwanger *et al.*, 2015a), was experimentally validated in section 5.2. The example of acoustic scattering by a rigid sphere was chosen as the case study for validation, due to the long-standing availability of analytical solutions (Strutt, 1877; Rayleigh, 1904; Duda and Martens, 1998). Some errors in the order of 2 dB were seen at high frequencies and extreme contralateral angles, where the pressure levels were very low. However, the overall mean square error between the analytical results and the simulation results was 0.0393 dB and, therefore, the accuracy of the Mesh2HRTF solver was deemed to be adequate for generation of the MPA database.

Once the accuracy of the Mesh2HRTF solver had been verified, a suitable amplitude for the harmonic deformations was explored in section 5.3. There are two key considerations in choosing the deformation amplitude: the amplitude of the deformations must be small enough to maintain linearity in the changes in pressure associated with the deformations, but large enough that the changes in pressure are above the noise floor of the BEM solver.

The linearity of the pressure changes was investigated using the spherical scattering example. Initially it was found that there was a limitation in the accuracy that Mesh2HRTF reads in meshes, which introduced quantisation errors in the results. However, the source code was adapted to read in the meshes at a much higher precision. After this improvement, and with suitable scaling to account for the changes in the acoustic source size due to the deformations, the linearity

simulation results generally agreed closely with the analytical results. However, there was a source of random error identified in the simulation results that requires further investigation. The results were nevertheless adequate in that they provided an indication of the limit of linearity. Previous analysis by Thorpe (2009) suggested a maximum deformation of 0.3 mm should be used to maintain linearity and the results presented here agreed, however, a maximum deformation limit of 0.1 mm was chosen to provide some further improvements in linearity.

Once the maximum deformation limit had been decided it was necessary to ensure that the changes in pressure associated with a deformation limit of 0.1 mm lay significantly above the noise floor of the BEM solver. Simulations were run for five different frequencies, distributed across the operating frequency range of pinna cues, and four directions for a randomly chosen order $m \in [-n, n]$ for each spherical harmonic degree n from 0-100. The changes in pressure were then compared to the variation in simulation results for 101 repeated simulations of the template mesh. It was verified that the deformation amplitude limit of 0.1 mm provided adequate signal-to-noise ratio (>60 dB). However, further work is necessary to validate this for additional directions and frequencies as time constraints meant it was only possible to test five frequencies and four directions in this study.

Finally, in section 5.4, the evaluation grid for the simulations, i.e. the HRTF directions within the database, was defined. In total 32,739 directions were chosen. This included points on a 1.5 m sphere around the head with an approximately uniform 2° spacing in both azimuth and elevation as well as points on each of the median, frontal and horizontal planes at 2° angular steps and 5 cm radial distance steps. These directions were chosen in order to cover the main directions of interest for studying the morphological origins of HRTF features.

Despite the further work required to identify the source of random error identified in the investigation of linearity, and additional verification of adequate SNR for additional directions within auditory space, this chapter has presented a number of important steps made towards the generation of a second generation MPA

database.

Chapter 6

Conclusions

> Equipped with his five senses, man explores the universe around him and calls the adventure Science.
>
> EDWIN POWELL HUBBLE

This chapter firstly presents a summary of the key novel contributions of this thesis, then the initial hypothesis is revisited in light of the results presented, and finally potential future work is identified.

6.1 Summary of work completed and key contributions

This work has led to several distinct contributions in the form of novel experimental techniques and results relating to the smoothing of perceptually valid head-related transfer functions, and a new approach to the development of a powerful tool for extending current understanding of the role of ear morphology in spatial hearing. The work carried out and the key outcomes are discussed in more detail below.

6.1.1 Head-related transfer function (HRTF) smoothing

Chapter 3 introduced a new algorithm for smoothing the magnitude spectra of head-related transfer functions (HRTFs). Previously published approaches to HRTF simplification have either smoothed on a linear scale, whereas the human ear adheres to an approximately logarithmic frequency scale, or have been limited by their perceptual testing. The algorithm presented is based on the principle of cepstral smoothing ("liftering"), as used by Kulkarni and Colburn (1998). However, rather than expressing the Fourier spectrum (the "cepstrum") of the HRTF magnitude spectrum on a linear "quefrency" scale, it is expressed on an equivalent rectangular bandwidth (ERB)-based quefrency scale. Therefore when the cepstral coefficients are gradually discarded, the HRTF magnitude spectrum is smoothed in a perceptually motivated manner; i.e. lower frequencies are smoothed more gradually than higher frequencies.

The algorithm was initially validated using a sagittal plane auditory localisation model developed by Baumgartner *et al.* (2013). The model takes as its input a set of template HRTFs, in this case the unsmoothed HRTFs, and a set of input HRTFs, in this case the smoothed HRTFs. The model then returns the likely polar localisation error introduced by using the input HRTFs, rather than the

template HRTFs to spatialise a sound. To compare the performance of the new algorithm versus Kulkarni and Colburn's algorithm, the estimate of localisation error introduced by their algorithm when retaining 32 cepstral coefficients (the minimum number required for perceptual transparency in their subjective testing) was initially calculated. Then the number of coefficients required by the new algorithm to equal or improve that error was calculated. The results using the auditory model indicated that on average across all subjects and sagittal planes that only 13 coefficients were required by the new algorithm to match the performance of Kulkarni and Colburn's algorithm with 32 coefficients.

Following this initial verification of the new algorithm using the auditory model, a subjective perceptual listening test was carried out by eight subjects from the Audio Lab at the university of York to corroborate the findings. The listening test combined a four-interval, two-alternative forced-choice paradigm with an adaptive 3-down-1-up adaptive staircase method to find the minimum number of coefficients needed for the listeners to hear no discernible difference between their smoothed and unsmoothed individualised HRTFs. Three directions were tested, (0°azimuth, 20°elevation), (90°azimuth, 0°elevation) and (-30°azimuth, -10°elevation), which were chosen to test a wide range of azimuths and elevations within the directions for which the subjects' HRTFs were measured. The test was repeated for both Kulkarni and Colburn's algorithm and the newly proposed algorithm.

Results of the listening tests suggested that the average number of coefficients, across all subjects and directions required by the new algorithm for perceptual transparency of the smoothed HRTFs is 17, compared to the 22 coefficients required by Kulkarni and Colburn's algorithm. This differs from the results of the auditory model, which suggested a far larger improvement with the new algorithm, but reasons for this difference were suggested. In the listening tests, for the direction 0°azimuth, 20°elevation, the required number of coefficients for the new algorithm was 18, compared to 30 for Kulkarni and Colburn's, which agrees with the model results more closely. This is of note for two reasons. Firstly it

was the highest elevation tested and the spectral detail of HRTFs are known to be a key elevation cue and secondly it was the only direction in the median plane, where, again, spectral cues are known to dominate. It was therefore suggested that the difference in performance between the two algorithms is greatest away from the horizontal plane and that the auditory model results, which were averaged across all elevations, were dominated by this, whereas the listening test only tested a small number of directions, relatively close to the horizontal plane where the difference between the two algorithms is smaller.

In summary, this work has demonstrated both the benefits and some limitations associated with the psychoacoustically-informed smoothing of HRTFs. Reducing the complexity of HRTFs allows them to be stored more efficiently and modelled using fewer parameters, but the gains observed vary with sound source direction. This simplified representation also has the potential to make the synthesis of high quality individualised HRTFs easier from morphological descriptors of ear shape, which is the chief motivation for the latter sections of this thesis.

6.1.2 Improvements to morphoacoustic perturbation analysis (MPA)

Chapter 4 presented improvements to the generation of a morphoacoustic perturbation analysis (MPA) database. The principles of MPA are to generate a database consisting of a template head mesh and a series of deformed versions of the template mesh in which orthogonal harmonic perturbations have been applied. The database also contains the acoustic changes associated with the HRTFs for each harmonic deformation compared with the corresponding template mesh HRTFs, calculated via boundary element method (BEM) simulations. This resulting database can then be used to investigate the morphological origins of acoustic features in HRTFs.

First generation MPA, as developed by Thorpe (2009) and reported by Tew *et al.* (2012), suffered from a number of limitations. In first generation MPA the har-

monic deformations applied to the template mesh during the creation of the MPA database are applied by first slicing the head mesh into contours, before mapping the contours to a two-dimensional (2D) plane and applying sine and cosine deformations using the Fourier transform. The first limitation of this technique is that the slicing axis must be placed carefully in order to avoid "island" contours within the pinnae. The second, and far greater, limitation of first generation MPA is that that each contour must be represented by an equal number of sample points. Since the slice contours vary greatly in length due to the complex morphology of the human pinnae, significant distortion of the harmonic deformations along the surface of the head often occurs. This leads to the requirement of additional higher frequency harmonics in order to obtain a spatial deformation wavelength with an adequate resolution. In addition, the distortions in the harmonic deformations on the head introduce erroneous results (artefacts) to the MPA analysis.

To combat this, a series of novel approaches to applying harmonic deformations to the template head mesh were presented. All the approaches shared the objective of mapping the head mesh to the surface of a sphere, where 2D spherical surface harmonics can be used for generating the deformations required for the MPA database. Each approach also aimed to minimise the distortion between the harmonics generated on the sphere and the corresponding harmonics on the head mesh. How this was achieved represents a key difference between the approaches.

The first approach, named path length relaxation (PLR), shared similarities with the slicing approach used in first generation MPA. However, rather than slicing the head mesh and resampling each slice contour, the x-y plane is rotated about the y-axis to intersect each vertex in turn. The vertex is then mapped to the sphere in a spherical coordinate system where the azimuth angle is the rotation of the x-y plane required to intersect with the vertex in question and the elevation angle is proportional to the distance from an origin in the pinna to the vertex, along the intersection of the plane and the surface of the head mesh.

Once the vertices of the head mesh have been mapped to the sphere a "force-minimisation" approach is applied. Each edge connected to a given vertex is treated as a force acting upon the vertex where the force is proportional to the distortion of the edge, i.e. an edge that is too short will "push" a vertex, whereas an edge that is to long will "pull" it. The distortions of the edges on the sphere are successively reduced by randomly cycling through all the vertices in the mesh and shifting each vertex based on the sum of the forces acting on it. The size of the shifts applied to the vertices is gradually reduced as the optimisation process progresses.

Whilst PLR worked well for lower resolution meshes, it was found to be unsuitable for full-resolution head meshes. The reasons for this were twofold. Firstly, it was found that the initial mapping to the sphere resulted in overlapping edges for the higher resolution mesh due to the complex morphology of the pinnae. These overlapping edges were difficult to handle. Secondly for the higher resolution mesh the force-minimisation approach was found to be prohibitively slow due to the large number of vertices and the requirement for numerous iterations through all the vertices at gradually decreasing step sizes.

Therefore another new approach was proposed, the optimised projection (OP) mapping method. Based on existing approaches to spherical parameterisation, the head mesh is initially projected on the surface of the sphere and then spatial averaging is applied to ensure that there are no overlapping edges. The novel contribution of this approach was to then adapt the spatial averaging so that rather than using equal weighting in the averaging, weights based on the distortion of the edges are used. This was found to reduce the edge distortions considerably. However, spatial averaging is most effective on uniform valence meshes, i.e. meshes where each vertex has the same number of edges connected to it. The complex morphology within the pinnae results in highly non-uniform valence and this reduced the effectiveness of spatial averaging, resulting in compressed pinnae regions.

In order to reduce the compression of the pinnae areas another novel method termed ellipsoidal pinna emphasis (EPE) was presented. The principle of EPE is to map the spherical mesh to and from an ellipsoid. By aligning the pinnae regions with the y-axis and reducing the radius of the ellipsoid along the y-axis, mapping to and from the ellipsoid stretches the pinnae areas, reducing the compression seen from spatial averaging. However, it was found that to achieve significant improvements in the pinnae, the edges around the median plane were detrimentally compressed.

Therefore, a final new approach, dynamic pinna emphasis (DPE), was developed. This takes advantage of the fact that spatial averaging works best for uniform valence meshes. A template, uniform valence spherical mesh is generated and the location of each vertex in the spherical head mesh is pre-calculated in barycentric coordinates. Then weighted spatial averaging of the uniform mesh is carried out as per OP. The weights for spatial averaging are calculated by first setting the weight of each vertex in the uniform mesh to the distortion associated with the closest vertex in the spherical head mesh, then the weight for each edge in the uniform mesh is calculated as the average weight of the two vertices at either end of it. Once the spatial averaging has been applied to the uniform mesh the new positions of the template head mesh vertices are calculated from their barycentric coordinates. In order to provide finer and finer optimisation, the resolution of the template uniform mesh is gradually increased.

Application of DPE to the spherical head mesh ensured that 16 kHz spatial frequency resolution was achieved across the whole of the head without aliasing, meaning that HRTF features across the whole operating frequency range can be investigated using MPA. In addition, DPE ensured that the range of edge distortions in the spherical head mesh was minimised in order to reduce the required number of spherical harmonics to give adequate spatial harmonic resolution across the whole head mesh.

The combination of the novel techniques of dynamic pinna emphasis and opti-

mised projection satisfactorily address the key difficulties associated with first generation MPA. The next step will be to apply these new techniques to the creation of an improved MPA database (see sections 6.1.3 and 6.3).

6.1.3 Generation of MPA database

Whist generation of the second generation MPA database has not been possible due to time constraints, some important steps have been taken towards database generation.

Firstly, the formulation of the spherical harmonics used as deformations in the MPA database has been defined as well as the normalisation technique used to limit the spherical harmonic deformation amplitude within the range of linear pressure change. Since the harmonic deformations are purely real-valued, only the real-valued spherical harmonics are required. A technique analogous to the max-normalisation (maxN) normalisation sometimes used in Ambisonics was chosen. This limits the absolute magnitude of any spherical harmonic to unity, so that, via multiplication of all harmonics by the required maximum deformation amplitude, the deformation amplitudes can be limited to the region of linear pressure change. Normalisation factors for the real-valued spherical harmonics up to degree 100 have been calculated. Due to the fact that no closed-form or recursive solution exists for calculating the maxN normalisation factors, they were computed via numerical evaluation of the associated Legendre polynomials.

In addition, the operation of the BEM solver proposed for generating the MPA database, Mesh2HRTF (Ziegelwanger *et al.*, 2015a), has been experimentally verified. The example of acoustic scattering by a rigid sphere was chosen for verification because numerous analytical solutions to the problem exist, allowing the theoretical and simulation results to be compared. It was found that the mean square error between the simulation results and the analytical results across all frequencies and all angles of incidence was 0.0393 dB. Some errors of up to 2 dB were seen at high frequencies and for extreme angles of incidence. However, be-

cause the sound pressure responses at these frequencies and angles were already so small as to be insignificant, it was decided that the overall error was adequate for generation of the MPA database.

Once the BEM solver had been correctly configured and its performance shown to be adequate, two important investigations relating to the generation of the MPA database were carried out. The first was to determine a maximum amplitude for the harmonic deformations applied to the template mesh. The amplitude must be set to ensure a sufficiently linear relationship exists between the applied deformations and their associated acoustic pressure changes. Again the example of acoustic scattering of a sphere was used, and the maximum deformation amplitude of 0.3 mm suggested by Thorpe (2009) was verified. Once this had been decided the second consideration was whether the acoustic changes associated with harmonic deformations of this amplitude were significantly above the noise floor of the BEM solver. It was verified that for the four directions and five frequencies considered that the signal-to-noise ratio (SNR) was greater than 60 dB. However, to improve confidence that this level of performance applies generally, the SNR should be evaluated for additional directions. In view of the very satisfactory performance observed, the maximum amplitude of all harmonic deformations was reduced from 0.3 mm to 0.1 mm to provide improved database linearity.

The final piece of preparation work for generation of the database was to consider the evaluation grid for the BEM simulations — the evaluation grid defines the directions of the HRTFs within the database. The various application areas of the MPA database were therefore considered and the final evaluation grid reflects this. The final grid consists of 32,739 points, spread across the surface of a 1.5 m sphere around the head, as well as across the three main anatomical planes: the median, frontal and horizontal planes.

6.2 Restatement of hypotheses

As stated in section 1.2 the key aims of this research were:

1. to simplify individualised HRTFs whilst preserving their ability to render sound with a high degree of perceptual integrity ; and
2. to create an accurate and efficient tool for studying the morphological origins of HRTF acoustic features.

In relation to these aims, the following two hypotheses were postulated:

1. By considering the frequency selectivity of the human hearing system, the spectral detail of HRTFs can be significantly simplified, whilst maintaining perceptual transparency.
2. By identifying and overcoming the deficiencies in first-generation morphoacoustic perturbation analysis (MPA) (Thorpe, 2009; Tew *et al.*, 2012) it is possible to develop a second-generation MPA tool with an improved performance.

Chapter 3 looked to answer question 1 by examining what level of spectral smoothing can be applied to HRTFs without perceptual consequences. A novel HRTF smoothing algorithm, based on the discarding of cepstral coefficients (Kulkarni and Colburn, 1998) was developed that considered the frequency selectivity of human ear. Subjective listening tests showed, firstly, that the magnitude spectrum of the HRTF can be significantly simplified without introducing perceptual changes, which agrees with previous studies (Kulkarni and Colburn, 1998; Breebaart and Kohlrausch, 2001; Senova *et al.*, 2002; Xie and Zhang, 2010; Pec *et al.*, 2013; Rasumow *et al.*, 2014; Hassager *et al.*, 2014, 2016). Secondly the listening tests demonstrated that, for the directions tested, smoothing on an auditory frequency scale, rather than a linear frequency one, offers a substantial reduction in the number of coefficients required to represent a perceptually valid HRTF magnitude spectrum. However, only a limited number of directions were tested

subjectively and further work is needed to verify the results across the whole auditory space.

Question 2 was addressed in chapter 4, which contains a number of novel contributions. The main weakness in first generation MPA, as presented by Thorpe (2009) and Tew *et al.* (2012), was the manner in which the harmonic deformations were applied to the template head mesh. The head mesh was sliced radially, the slices were sampled and then mapped to a 2D plane where Fourier harmonics were applied, which were finally mapped back to the head mesh.

This resulted in significant distortions in the harmonic deformations on the head, largely due to the complex morphology of the pinnae. In this work a new approach was developed, based on mapping the head mesh vertices directly to the surface of a sphere, where spherical harmonics can be used to generate the harmonic deformations that make up the MPA database. In addition to the spherical mapping approach, several novel optimisation techniques were presented for reducing the distortions in the spherical head mesh in order to maximise the equivalence of the harmonic deformations on the sphere and on the head mesh. This new approach represents a great improvement over first-generation MPA and whilst the new MPA database has yet to be created, chapter 5 reported additional steps taken towards this goal.

6.3 Further work

Whilst this research has provided novel contributions to both HRTF simplification and the development of a second-generation MPA tool, there is further work that would enhance both areas. This section will review the outstanding work as well as potential, longer term, research avenues going forwards.

The work on the smoothing of HRTF magnitude spectra presented in chapter 3 has indicated that the newly proposed, perceptually-informed smoothing algorithm performs better than algorithms that do not consider the frequency selec-

tivity of the human ear. In addition this research has verified that significant simplifications can be made to HRTFs. However, the subjective testing was limited in the number of directions that were tested. It is desirable to establish the performance of the algorithm across the whole of auditory space. It is feasible that now that an indication of the number of coefficients required by the algorithm has been found, the listening test process could be shortened, allowing the testing of more directions. For example the number of coefficients suggested by this study could be uses as the starting level of the adaptive staircase method, rather than starting from the minimum number of coefficients. In addition, in this further testing it would be perhaps advantageous to use pink noise, rather than white noise as the test stimuli. Since pink noise has equal power per equal percentage bandwidth, it better approximates the frequency selectivity of the human ear. This should lead to an improved evaluation of the smoothing algorithm and would almost certainly reduce listening fatigue during the tests.

Once this additional work on the smoothing algorithm has been completed, and the algorithm validated further, it will provide a powerful tool to use in conjunction with MPA. The simplification of the HRTF magnitude spectra will ensure that only the most perceptually salient HRTF features are studied. Additionally, the proposed algorithm's approach of representing the HRTF as spectral coefficients has significant implications in terms of data storage for the MPA database. In a broader scientific sense, the smoothing algorithm will also provide novel insight into finding the spectral cues that are most perceptually important.

Although chapter 4 has shown significant improvements to first-generation MPA and chapter 5 has documented some steps taken towards the generation of a second-generation MPA database, the database has yet to be created. Therefore, perhaps the most pressing further work is creation of the second-generation MPA database using the new approach for producing harmonic deformations. However, as identified in section 5.5, there are still a couple of questions to be answered before generation of the database is possible. The first question to be answered is what the source of random error found in the linearity tests is. It is believed

that this is a result of the seeding of the BEM solver, but it requires further investigation. The second question is whether the signal-to-noise ratio for the chosen harmonic deformation amplitude is high enough across all frequencies and directions, and not just the relatively limited number already tested.

Once these questions have be answered and the database has been generated, the database should be verified against first-generation MPA. As discussed in section 2.5.3, first-generation MPA was validated by studying the morphological origin of a notch and a peak within the HRTFs of Knowles Electronics manikin for acoustic research (KEMAR) (Tew *et al.*, 2012). The MPA database was applied to the two features and their morphological origin extracted. Then acoustic measurements of KEMAR were carried out, applying putty to the areas suggested by the MPA analysis to verify that the expected changes to the HRTF features were observed. It would therefore make sense to verify the second-generation MPA database using a similar approach.

After creation and validation of the second-generation MPA database, the future research avenues associated with it can be split into two areas: "scientific" and "applied". The scientific future research relates to improving knowledge of human spatial hearing and the applied research relates to improvement of practical spatial audio systems.

There are a number of scientific applications for MPA. As highlighted in section 2.5, whilst there is some understanding of the relationship between human morphology and auditory localisation cues, there is still a lot that is not understood. Therefore, one potential area of future research is the morphological origin of localisation cues within "cones of confusion". As discussed in section 2.2, cones of confusion are the loci where interaural cues do not change and thus the brain must rely solely on spectral localisation cues. It would be of great interest to study how the spectral detail of the HRTFs changes around the cones of confusion, and what the morphological origin of those changes is. Another potential application is to investigate what morphology is responsible for front-back lo-

calisation cues. One common effect of the use of non-individualised HRTFs is front-back confusions (section 2.3.4) and therefore finding what morphology is critical for accurate front-back localisation cues would be of great benefit.

It would also be valuable to expand the second-generation MPA database to contain additional subjects beyond the KEMAR used in this study. Expanding the database to include further subjects would facilitate detailed study of inter-subject variation in the morphological origin of localisation cues. For instance, subjects might exhibit similar patterns of HRTF spectral features, but it is possible that the cues originate from different morphological regions. One possible source of additional subjects for the MPA database is the Sydney-York morphological and recording of ears database (SYMARE database) (section 2.4.2), which contains high resolution BEM-ready head meshes for 61 subjects. This would provide an invaluable pool of information for both the previously mentioned scientific applications and also the applied research covered next.

As stated in section 1.2, one of the long-term aims of this research is the simulation of perceptually valid HRTFs from a practical number of morphological measurements. This represents perhaps the most pertinent applied research stemming from this work. It is not unreasonable to see a point in the future where people can use an app on their smartphone to take pictures of their ears in order to generate individualised HRTFs and indeed software based on this principle is already becoming available (3D Sound Labs, 2017). However, there is still a large body of research required to reach that point. This research can be broadly split into three steps. The first is to identify the most perceptually relevant HRTF features, the second is to discover the morphological origin of those features and the final step is to develop a method for gathering the required information about those morphological features in order to accurately synthesise the HRTFs.

The smoothing algorithm presented in chapter 3 is an important tool for the first step. It will allow the HRTFs to be distilled down to only those spectral features that are perceptually relevant. This reduced set of HRTF features could then

be analysed in terms of importance, firstly via the objective analysis previously mentioned pertaining to front-back localisation cues and cues within cones of confusion, and then through subjective listening tests to corroborate any findings.

The second step will then consist of the application of the second-generation MPA database to investigate the morphological origin of the HRTF features identified in step one. After this application of the MPA database a list of HRTF features and their corresponding morphological origins in order of importance should have been generated. This will represent the spectral HRTF features of most importance for accurate, individualised HRTFs and the morphological features that control those localisation cues.

Finally, investigation into methods for gathering details of the most important morphological features from simple measurements will be required. As mentioned in section 2.4.6, there has been some research into producing the meshes required for acoustic simulations from photographs or videos. Another potential avenue might be that a template ear mesh could be individualised using the most important morphological measurements that would have been previously defined. The gathering of these morphological measurements might be feasible using just 2D photographs taken with a smartphone. Then cloud computing could be utilised to carry out the acoustic simulations in order to generate HRTFs that can be delivered back to the device. This would provide an affordable manner to deliver accurate spatial audio to many people.

Appendix A

Investigating head-related transfer function smoothing using a sagittal-plane localization model

Originally paper and poster presentation at *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, WASPAA 2015, New Paltz, United States, October 2015

INVESTIGATING HEAD-RELATED TRANSFER FUNCTION SMOOTHING USING A SAGITTAL-PLANE LOCALIZATION MODEL

Laurence J. Hobden and Anthony I. Tew*

Audio Lab
Department of Electronics
University of York
Heslington, York, UK
Emails: lh571@york.ac.uk tony.tew@york.ac.uk

ABSTRACT

A new head-related transfer function (HRTF) smoothing algorithm is presented. HRTF magnitude responses are expressed on an equivalent rectangular bandwidth frequency scale and smoothing is increased by progressively discarding the higher frequency Fourier coefficients. A sagittal plane localization model was used to assess the degree of spectral smoothing that can be applied without significant increase in localization error. The results of the localization model simulation were compared with results from a previous perceptual investigation using an algorithm that discards coefficients on a linear frequency scale. Our findings suggest that using a perceptually motivated frequency scale yields similar localization performance using fewer than half the number of coefficients.

***Index Terms*—** head-related transfer function, spatial sound, spectral smoothing, auditory localization model

1. INTRODUCTION

Interaural sound localization cues can be ambiguous on the so-called *cones of confusion* [1] where the interaural cues (interaural time difference (ITD) and interaural level difference (ILD) [2]) are near-constant. Therefore, to accurately localize sounds the ear also depends on spectral variation in a listener's head-related transfer functions (HRTFs) [3, 4]. To exploit these cues to maximum effect in spatial audio applications it is generally agreed that individualized HRTFs are required [5]. However, the acoustic measurement of an individual's HRTFs is a lengthy process which requires specialist equipment. Therefore the possibility of estimating an individual's HRTFs more simply by other means is an active area of research.

There have been a number of investigations into modeling HRTFs: including mathematical models [6, 7], anthropometric models [8], structural models [9] and physical models [10], and these have met with some success. One of the most promising approaches of late has been synthesizing HRTFs using the 3D boundary element method [11, 12] and the finite-difference time-domain method [13]. Although these synthesis methods can yield accurate HRTFs, they are computationally demanding. Measured HRTFs have been shown to contain spectral detail that is not perceptually salient [14] and modeling or synthesizing only the salient features provides desirable computational savings.

*Funding of this work by Meridian Audio Ltd., UK and the University of York, UK is gratefully acknowledged.

A number of studies [15, 16, 17, 18] have investigated spectral smoothing to try and discern what level of spectral detail is required and all agree that HRTFs can be smoothed to some extent without affecting localization. Kulkarni and Colburn [15] took the Fourier transform of the HRTF magnitude and systematically discarded higher frequency components to steadily increase smoothing. This approach has similarities to cepstral techniques [19, 20, 21], specifically liftering [22]. Their HRTFs, however, were expressed on linear frequency scales whereas the human auditory system adheres to an approximately logarithmic frequency scale. Furthermore, they only tested a limited number of locations (four) in their discrimination task, which all lie on the horizontal plane where interaural rather than spectral cues generally dominate. Senova *et al.* [16] tested a wide selection of directions within the auditory sphere in a localization test, but their approach of truncating HRIRs to smooth the HRTFs also smooths on a linear frequency scale. Breebaart and Kohlrausch's smoothing algorithm [17] is more ecologically valid as it uses the gammatone filter bank [23]. However, their discrimination testing is limited to the horizontal (azimuthal) plane and uses the same set of non-individual HRTFs for all test participants. Recently, Hassager *et al.* [18] used a gammatone filter bank with variable bandwidth filters to investigate the effect of HRTF smoothing on externalization. However, they also limited their perceptual testing to the horizontal plane and focused on externalization rather than localization.

This work develops a new perceptually motivated HRTF smoothing algorithm based on Kulkarni and Colburn's [15] approach of describing the HRTF magnitude spectrum using Fourier coefficients. However, before calculating the Fourier coefficients, we interpolate the HRTF and resample the magnitude spectrum on an equivalent rectangular bandwidth (ERB) [24] based frequency scale. This maximises the amount of perceptually relevant spectral information retained for any given number of Fourier coefficients. In this paper the new algorithm is first described and then the Baumgartner [25] sagittal plane auditory localization model is used to assess the new algorithm's performance and results are compared to results obtained using Kulkarni and Colburn's algorithm.

2. METHOD

2.1. Head-related transfer function smoothing

The proposed algorithm is as follows:

1. The FFT of the HRIR is taken to compute the M-point symmetrical HRTF.

2. To reduce computation only the first $M/2+1$ points, i.e. up to $f_s/2$, of the log-magnitude HRTF, sampled at linear frequency points f_{lin}, are considered, as the HRTF exhibits Hermitian symmetry.
3. $M/2+1$ auditory frequency points, f_{aud} are calculated.
4. The spectrum is resampled at f_{aud}, using shape-preserving piecewise cubic interpolation.
5. The FFT of the resampled spectrum is calculated.
6. The N lower spectral magnitude coefficient values (for frequency points $k = 1...N$) are retained and any remaining coefficients ($k = N+1...M/2+1$) are set to zero.

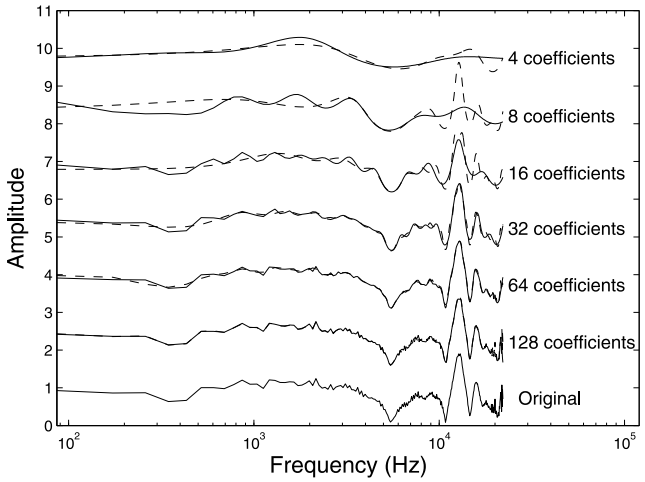


Figure 1: Comparison of smoothing algorithms. Dashed line - Kulkarni and Colburn's approach. Solid line - proposed approach.

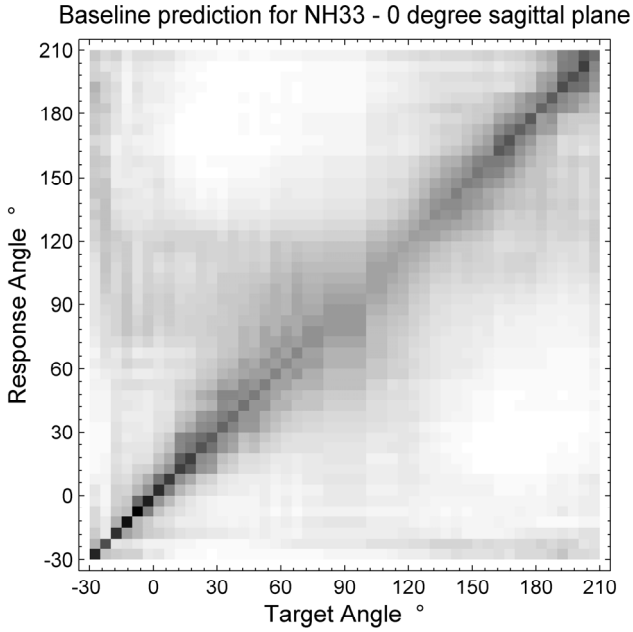


Figure 3: Exemplar output of Baumgartner model [25] using the 0 degree (median) sagittal plane DTFs for set NH33.

7. The inverse FFT is computed.
8. The modified log-magnitude spectrum is resampled at the original linear frequency points, f_{lin}, using shape-preserving piecewise cubic interpolation.
9. The spectrum is reflected about $f_s/2$ to produce a spectrum with even symmetry.
10. Real cepstrum minimum-phase reconstruction [26] is used to compute the minimum-phase HRTF which is combined with a pure delay equal to the ITD [6].

The auditory frequency points calculated in step 3 are based on the equivalent rectangular bandwidth (ERB) scale. This scale gives an approximation to the bandwidths of the auditory filters and is defined as follows [24]:

$$\text{ERB}(f_c) = 24.7\left(\frac{4.37 f_c}{10^3}+1\right) \tag{1}$$

where f_c is in kHz and $\text{ERB}(f_c)$ is in Hz. Slaney [27] developed an adaptable version of equation 1 in his Auditory Toolbox for MATLAB [28]:

$$\text{ERB}(f_c, EarQ, minBW, n) = \left(\left(\frac{f_c}{EarQ}\right)^n + minBW^n\right)^{\frac{1}{n}} \tag{2}$$

where n, $EarQ$ and $minBW$ are the filter order, filter selectivity at high frequencies and the required low frequency minimum bandwidth, respectively. The values used for n, $EarQ$ and $minBW$ are those suggested by Glasberg and Moore [24]:

$$\begin{aligned} n &= 1 \\ EarQ &= 9.26449 \\ minBW &= 24.7 \end{aligned} \tag{3}$$

Slaney also developed the following equation for calculating the centre frequency of the pth auditory filter in an P-channel filter bank:

$$f_p = -A + (f_h + A)\exp(p\,(-\log(f_h + A) + \log(f_l + A))/P) \tag{4}$$

where $A = EarQ \times minBW$, f_l is the lowest frequency and f_h the highest frequency in the filter bank.

Equation 4 was used to calculate the auditory frequency points f_{aud} using the following values:

$$\begin{aligned} P &= M/2 \\ f_l &= 0 \\ f_h &= f_s/2 \end{aligned} \tag{5}$$

The values suggested by Glasberg and Moore were used for $EarQ$ and $minBW$.

Figure 1 shows the smoothing effect of Kulkarni and Colburn's algorithm on an exemplar HRTF compared to the algorithm used in this study; in each case the number of coefficients retained, N, is shown on the right of the figure. It can be seen that for a given number of coefficients our algorithm retains more spectral detail at lower frequencies, where the human ear is more frequency selective, and less at higher frequencies more, where the ear is less frequency selective.

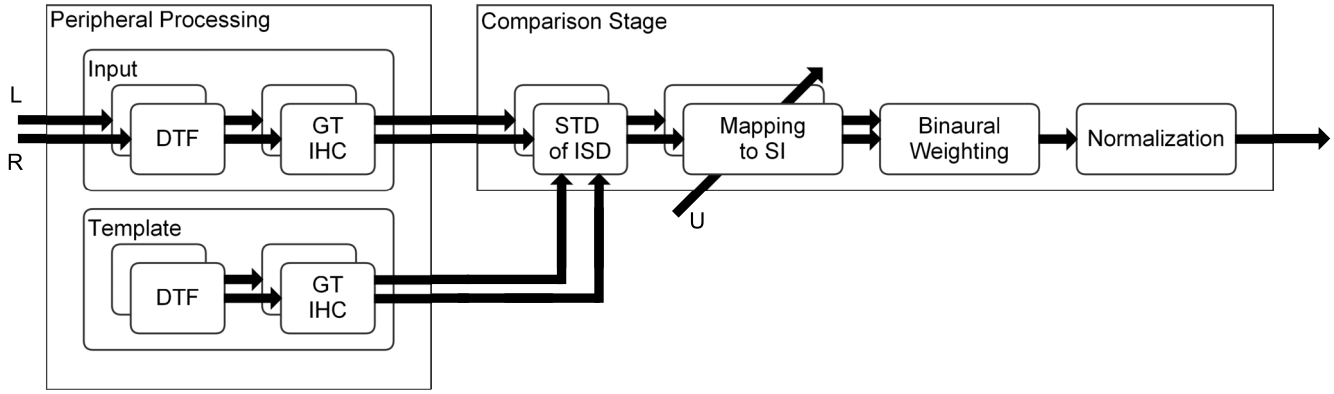


Figure 2: Baumgartner *et al.'s* sagittal plane localization model (after [25]).

2.2. Baumgartner sagittal localization model

Sagittal planes are planes parallel to the median plane, where spectral cues dominate localization. Therefore, the sagittal plane (SP) localization model developed by Baumgartner *et al.* [25] is well suited to investigating the effect of spectral smoothing. Figure 2 outlines the model. Briefly, the incoming sounds are processed to emulate the peripheral hearing system and the resulting data are compared with a similarly processed internal reference or "template". The spectral distance of the incoming processed sound from each entry in the template is computed which results in a discrete probability density function or probability mass vector (PMV) for each target polar angle. See [25] for a deeper explanation.

The PMVs for each sagittal plane can be combined to form a matrix, as shown in figure 3. For each column (target angle) the estimated probability of a subject with these directional transfer functions (DTFs) selecting a particular response angle is inversely proportional to the image's brightness. From the PMVs the polar error and quadrant error can be calculated to facilitate analysis of changes in subject performance. The quadrant error is calculated as the sum of the PMV entries for which the response-target difference is greater than ninety degrees. The polar error is calculated as the discrete expectancy within the local polar range (less than ninety degrees). In the matrix representation in figure 3, quadrant errors manifest as dark areas in the top left and bottom right quadrants, e.g. around 120°–210°for -30°target angle. Low polar error manifests as concentrated areas of darkness around the diagonal, e.g. around 0°for 0°target angle, whilst high polar error manifests as large areas of similar brightness, e.g. around 90°for 90°target angle.

This investigation used the implementation of the SP localization model in the Auditory Modelling Toolbox (AMT) [29], a collection of MATLAB scripts for auditory research. All 17 DTFs provided in the AMT for use with the Baumgartner model were used. These DTFs originated from the ARI HRTF database [30] and further details about which subjects' DTFs are in the AMT can be found in [25].

The `baumgartner2013` function in the AMT takes as two of its inputs the "input" and "template" DTFs. The input DTFs were progressively smoothed using the algorithm outlined in section 2.1 and the polar error compared to the values obtained when the unsmoothed DTFs were used. This comparison was also conducted using the Kulkarni and Colburn algorithm to smooth the DTFs so that the extent of the anticipated improvement of our smoothing algorithm could be assessed.

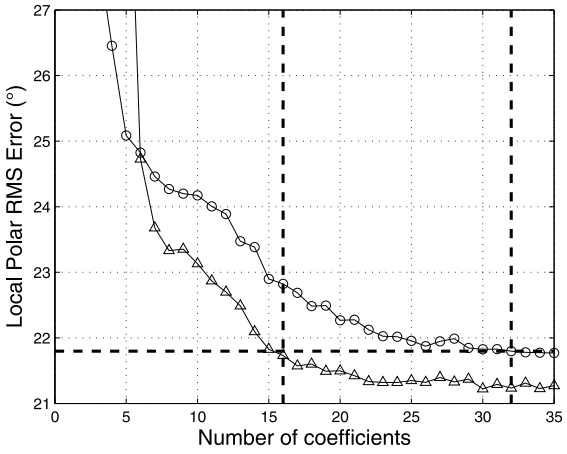


Figure 4: Polar error against number of coefficients retained for proposed algorithm (triangles) compared to Kulkarni and Colburn's algorithm (circles). Average across DTF sets for a target angle of 0°in the median plane. Dotted lines plotted to show comparative performance.

3. RESULTS

Figure 4 shows the mean polar error across DTF sets for a target angle of 0°(straight-ahead) in the median plane for both the Kulkarni and Colburn algorithm and the new algorithm. Kulkarni and Colburn [15] reported that in their localization task, whilst performance was seldom affected with fewer than 32 coefficients, three out of the four subjects did exhibit a reduction in performance for 16 coefficients in the frontal direction. Therefore the localization performance results obtained from the model using the Kulkarni and Colburn algorithm with 32 coefficients was used as a threshold to assess the new algorithm.

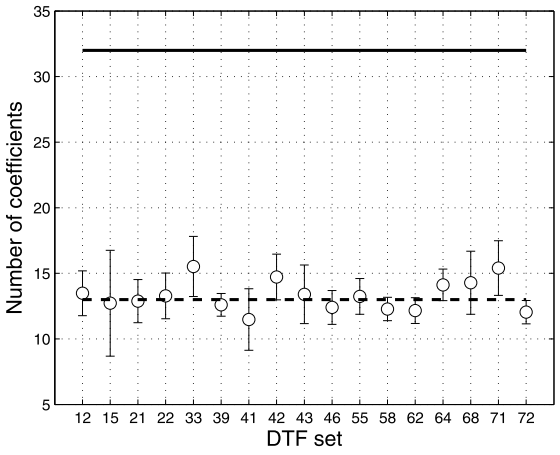


Figure 5: Number of coefficients needed for equal performance with Kulkarni and Colburn algorithm using 32 coefficients. The mean across sagittal planes and the one standard deviation limits are shown for each DTF set. The dashed line is the mean across DTF sets.

Figure 5 shows the number of coefficients needed, for each DTF set, for the new algorithm to equal the performance of Kulkarni and Colburn's algorithm with 32 coefficients. The means and one standard deviation limits across all sagittal planes are presented.

As can be seen, the mean number of coefficients required for the new smoothing algorithm is 13, more than a twofold saving on the number of coefficients required by the Kulkarni and Colburn smoothing algorithm for the same localization performance. An ANOVA carried out on the data showed no significant effect of sagittal plane on performance ($F = 0.9, p = 0.60$).

4. DISCUSSION

The simulation results indicate that on average across sagittal planes and DTF sets fewer than half the number of coefficients are required to meet the performance threshold set by Kulkarni and Colburn's algorithm. However, it is interesting to note that in figure 4 there is a region below 6 coefficients where the Kulkarni and Colburn algorithm outperforms the new algorithm. It is likely that this behaviour is a result of the relatively sudden loss of sufficient spectral accuracy across all frequencies in our perceptually motivated algorithm as the number of coefficients is reduced. In Kulkarni and Colburn's algorithm this loss occurs more gradually, with spectral resolution becoming insufficient at lower frequencies sooner than higher frequencies as the number of coefficients is reduced. Also of note is the relatively large baseline ($\sim$21.5°) polar error. Whilst this seems very large, Baumgartner *et al.* have shown their model tends to underestimate polar error across the whole auditory space — see figure 7 in [25].

It should also be mentioned that the model simulation assesses localization error, whereas, with the exception of Senova *et al.* [16], previous studies [15, 17, 18] have assessed discriminability. In terms of finding a perceptual limit for smoothing, discrimination would seem a better form of assessment. Therefore, whilst the results presented here look promising, a subjective discrimination task would fully corroborate the model simulation findings.

These preliminary results have important ramifications in terms of reducing storage requirements for HRTFs. Reducing the complexity of high quality HRTFs will also facilitate efforts to synthesise them more efficiently and more rapidly from parametric data by reducing the degrees of freedom in the problem. It will also simplify the study of HRTF morphoacoustics [31] and assist in the identification of perceptually significant acoustic contributions to HRTF features.

5. CONCLUSIONS

This paper presents a new algorithm for smoothing HRTFs. It is based on discarding Fourier coefficients with equal separation on a frequency scale derived from the equivalent rectangular bandwidth of the auditory filters. The algorithm was tested using a sagittal plane localization model and the results suggest that fewer than half the number of coefficients - 13 versus 32 - are required compared with discarding coefficients on a linear frequency scale in the manner described by Kulkarni and Colburn [15]. This result has a number of implications in terms of reducing HRTF storage, speeding up HRTF synthesis and studying the perceptual contribution of HRTF features. Subjective perceptual testing is now required to validate the results provided by these localization model simulations.

6. REFERENCES

[1] A. W. Mills, "Auditory Localisation," in *Foundations of Modern Auditory Theory*, J. V. Tobias, Ed. New York: Academic Press, 1972, vol. 2, pp. 303–348.

[2] J. W. Strutt, "On our perception of sound direction," *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[3] J. Blauert, "Sound Localization in the Median Plane," *Acoustica*, vol. 22, pp. 205–213, 1969/1970.

[4] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *J. Acoust. Soc. Am.*, vol. 56, no. 6, pp. 1829–1834, 1974.

[5] E. M. Wenzel, M. A., D. Kistler, and F. L. Wightman, "Localization using non-individualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993.

[6] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647, 1992.

[7] S. Hwang and Y. Park, "Interpretations on principal components analysis of head-related impulse responses in the median plane," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. EL65–71, 2008.

[8] P. Satarzadeh, V. R. Algazi, and R. O. Duda, "Physical and filter pinna models based on anthropometry," May 5—8 2007, presented at 122nd AES Conv., Vienna, Austria.

[9] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 5, pp. 476–488, 1998.

[10] E. A. Lopez-Poveda and R. Meddis, "A physical model of sound diffraction and reflections in the human concha," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3248–3259, 1996.

[11] Y. Kahana and P. A. Nelson, "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models," *J. Sound and Vibration*, vol. 300, no. 3-5, pp. 552–579, 2007.

[12] C. Jin, A. Tew, P. Guillon, N. Epain, R. Zolfaghari, A. van Schaik, C. Hetherington, and J. Thorpe, "Creating the sydney york morphological and acoustic recordings of ears database," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 37–46, 2014.

[13] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato, "Comparison of simulated and measured hrtfs: Fdtd simulation using mri head data," October 5–8 2007, presented at 123rd AES Conv., New York, USA.

[14] S. Carlile, R. Martin, and K. McAnally, "Spectral information in sound localization," in *Auditory Spectral Processing*, M. Malmierca and D. Irvine, Eds. Massachusetts, USA: Academic Press Inc., 2005.

[15] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747–749, 1998.

[16] M. Senova, K. McAnally, and R. Martin, "Localization of virtual sound as a function of head-related impulse response duration," *J. Audio Eng. Soc.*, vol. 50, no. 1/2, pp. 57–66, 2002.

[17] J. Breebaart and A. Kohlrausch, "The perceptual (ir)relevance of hrtf magnitude and phase spectra," May 12–15 2001, presented at 110th AES Conv., Amsterdam, the Netherlands.

[18] H. G. Hassager, T. Dau, and F. Gran, "The effect of spectral details on the externalization of sounds," September 7–12 2014, presented at Forum Acusticum, Kraków, Poland.

[19] B. P. Bogert, M. J. Healy, and J. W. Tukey, "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," in *Proceedings of the symposium on time series analysis*, vol. 15, 1963, pp. 209–243.

[20] D. G. Childers, D. Skinner, and R. Kemerait, "The cepstrum: A guide to processing," in *Proceedings of the IEEE*, vol. 65, no. 10, Oct 1977, pp. 1428–1443.

[21] A. Oppenheim and R. Schafer, "From frequency to quefrency: a history of the cepstrum," *Signal Processing Magazine, IEEE*, vol. 21, no. 5, pp. 95–106, Sept 2004.

[22] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 7, pp. 947–954, 1987.

[23] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, 1987.

[24] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[25] R. Baumgartner, P. Majdak, and B. Laback, "Assessment of sagittal-plane sound localization performance in spatial-audio applications," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 93–119.

[26] S.-C. Pei and H.-S. Lin, "Minimum-phase fir filter design using real cepstrum," *IEEE Trans. Circuits and Systems II: Express Briefs*, vol. 53, no. 10, pp. 1113–1117, 2006.

[27] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Perception Group - Advanced Technology Group - Apple Computer Inc., Tech. Rep., 1993.

[28] ——, "Auditory toolbox, version 2," Interval Research Corporation, Tech. Rep. #1998-010, 1998.

[29] P. L. Søndergaard and P. Majdak, "The auditory modelling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 33–56.

[30] Austrian Academy of Sciences - Acoustics Research Institute, "ARI HRTF Database," accessed September 10 2014, available at http://www.kfs.oeaw.ac.at/hrtf.

[31] A. I. Tew, C. Hetherington, and J. Thorpe, "Morphoacoustic perturbation analysis : principles and validation," April 23–27 2012, presented at Acoustics 2012, Nantes, France.

List of Acronyms

2D two-dimensional.

3D three-dimensional.

AMT Auditory Modelling Toolbox.

ANOVA analysis of variance.

ARI Acoustic Research Institute.

BEM boundary element method.

BRIR binaural room impulse response.

CAD computer aided design.

CDF cumulative distribution function.

CIPIC Center for Image Processing and Integrated Computing.

CTF common transfer function.

DPE dynamic pinna emphasis.

DPS differential pressure synthesis.

DTF directional transfer function.

EFT elliptic Fourier transform.

EPE ellipsoidal pinna emphasis.

ERB equivalent rectangular bandwidth.

FDTD finite difference time domain.

FEC free air equivalent coupling.

FEM finite element method.

FFT fast Fourier transform.

FIR finite impulse response.

FM-BEM fast multipole boundary element method.

FMM fast multipole method.

FuMa Furse-Malham.

GUI graphical user interface.

HATS head-and-torso simulator.

HFC Human Factors Consult.

HpTF headphone transfer function.

HRIR head-related impulse response.

HRTF head-related transfer function.

IIR infinite impulse response.

ILD interaural level difference.

IPD interaural phase difference.

IR infrared.

IRCAM Institut de Recherche et Coordination Acoustique/Musique.

ISD interaural spectral difference.

ITD interaural time difference.

JND just-noticeable-difference.

KAIST Korea Advanced Institute of Science and Technology.

KEMAR Knowles Electronics manikin for acoustic research.

LFE low frequency effect.

LP linear prediction.

MAA minimum audible angle.

MAMA minimum audible movement angle.

maxN max-normalisation.

MLS maximum length sequences.

MPA morphoacoustic perturbation analysis.

MRI magnetic resonance imaging.

OP optimised projection.

PaaS processing as a service.

PCA principal components analysis.

PDE partial differential equation.

PLR path length relaxation.

PMV probability mass vector.

PRTF pinna-related transfer function.

RMS root mean square.

s.f.s significant figures.

SaaS software as a service.

SADIE Spatial Audio for Domestic Interactive Entertainment.

SAQI spatial audio quality inventory.

SFRS spatial frequency response surfaces.

SNR signal-to-noise ratio.

SP sagittal plane.

SPCA spatial principal components analysis.

SSH surface spherical harmonics.

SVTF ear canal sound pressure to stapes footplate velocity transfer function.

SWT stationary wavelet transform.

SYMARE database Sydney-York morphological and recording of ears database.

VAD virtual auditory display.

VBAP vector base amplitude panning.

VR virtual reality.

WFS Wave field synthesis.

References

3D Sound Labs (2017), "Personal HRTFs - 3D Sound Labs", Available online: http://pro.3dsoundlabs.com/category/personal-hrtfs/, (Accessed: 13/03/2019).

Acoustics Research Institute (2011), "ARI HRTF database", (Online) Available: https://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=608&catid=158&Itemid=606&lang=en, (Accessed: 11/03/2017).

Aibara, R., Welsh, J. T., Puria, S., and Goode, R. L. (2001), "Human middle-ear sound transfer function and cochlear input impedance", *Hearing Research*, **vol. 152**, no. 1-2, pp. 100–109.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001a), "Elevation localization and head-related transfer function analysis at low frequencies", *Journal of the Acoustical Society of America*, **vol. 109**, no. 3, pp. 1110–1122.

Algazi, V. R., Avendano, C., and Duda, R. O. (2001b), "Estimation of a spherical-head model from anthropometry", *Journal of the Audio Engineering Society*, **vol. 49**, no. 6, pp. 472–479.

Algazi, V. R., Avendano, C., and Thompson, D. M. (1999), "Dependence of subject and measurement position in binaural signal acquisition", *Journal of the Audio Engineering Society*, **vol. 47**, no. 11, pp. 937–947.

Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002), "Approximating the head-related transfer function using simple geo-

metric models of the head and torso", *Journal of the Acoustical Society of America*, **vol. 112**, no. 5, pp. 2053–2064.

Algazi, V. R., Duda, R. O., Morrison, R. P., and Thompson, D. M. (2001c), "Structural composition and decomposition of HRTFs", presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, New Paltz, United States, October 2001.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001d), "The CIPIC HRTF database", presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2001)*, New Paltz, United States, October 2001.

Alves-Pinto, A. and Lopez-Poveda, E. A. (2005), "Detection of high-frequency spectral notches as a function of level.", *Journal of the Acoustical Society of America*, **vol. 118**, no. 4, pp. 2458–2469.

Amazon Web Services (2017), "Elastic compute cloud (ec2) cloud server & hosting - aws", (Online) Available: `https://aws.amazon.com/ec2/`, (Accessed: 11/03/2017).

Arfken, G. B. and Weber, H. J. (1995), *Mathematical Methods for Physicists*, Academic Press, 4th edn.

Asano, F., Suzuki, Y., and Sone, T. (1990), "Role of spectral cues in median plane localization", *Journal of the Acoustical Society of America*, **vol. 88**, no. 1, pp. 159–168.

Ashihara, K. (2007), "Hearing thresholds for pure tones above 16 kHz", *Journal of the Acoustical Society of America*, **vol. 122**, no. 3, pp. EL52–EL57.

Auro Technologies (2015), "Auromax: Next generation immersive sound system", Available online: `http://www.auro-3d.com/wp-content/uploads/documents/AuroMax_White_Paper_24112015.pdf`, (Accessed: 11/03/2017).

Avendano, C., Algazi, V. R., and Duda, R. O. (1999a), "A head-and-torso model for low-frequency binaural elevation effects", presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 1999)*, New Paltz, United States, October 1999.

Avendano, C., Duda, R. O., and Algazi, V. R. (1999b), "Modeling the contralateral HRTF", presented at *16th International Audio Engineering Society Conference: Spatial Sound Reproduction*, Rovaniemi, Finland, March 1999.

Batteau, D. W. (1967), "The role of the pinna in human localization", *Proceedings of the Royal Society of London. Series B. Biological Sciences*, **vol. 168**, no. 1011, pp. 158–180.

Bauer, B. B. (1961a), "Phasor analysis of some stereophonic phenomena", *Journal of the Acoustical Society of America*, **vol. 33**, no. 11, pp. 1536–1539.

Bauer, B. B. (1961b), "Stereophonic earphones and binaural loudspeakers", *Journal of the Audio Engineering Society*, **vol. 9**, no. 2, pp. 148–151.

Bauer, B. B. (1965), "Improving headphone listening comfort", *Journal of the Acoustical Society of America*, **vol. 37**, no. 6, pp. 1210–1210.

Baumgartner, R., Majdak, P., and Laback, B. (2013), "Assessment of sagittal-plane sound localization performance in spatial-audio applications", in *The Technology of Binaural Listening* (edited by Blauert, J.), chap. 4, pp. 93–119, Springer.

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001), "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source", *Journal of the Audio Engineering Society*, **vol. 49**, no. 10, pp. 904–916.

von Békésy, G. (1960), *Experiments in hearing*, McGraw-Hill, New York, USA.

Berkhout, A. J., de Vries, D., and Vogel, P. (1993), "Acoustic control by wave

field synthesis", *Journal of the Acoustical Society of America*, **vol. 93**, no. 5, pp. 2764–2778.

Bernstein, L. R. and Trahiotis, C. (1982), "Detection of interaural delay in high-frequency noise", *Journal of the Acoustical Society of America*, **vol. 71**, no. 1, pp. 147–152.

Bertet, S., Daniel, J., Gros, L., Parizet, E., and Warusfel, O. (2007), "Investigation of the perceived spatial resolution of higher order ambisonics sound fields: A subjective evaluation involving virtual and real 3d microphones", presented at *30th International Audio Engineering Society Convention: Intelligent Audio Environments*, Saariselkä, Finland, March 2007.

Blauert, J. (1970), "Sound localization in the median plane", *Acustica*, **vol. 22**, pp. 205–213.

Blauert, J. (1997), *Spatial hearing: the psychophysics of human sound localization, Revised Edition*, MIT press, Cambridge, Massachusetts.

Blumlein, A. D. (1933), "Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems", British Patent Specification 394,325.

Bogert, B. P., Healy, M. J. R., and Tukey, J. W. (1963), "The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking", in *Proceedings of the symposium on time series analysis*, vol. 15, pp. 209–243.

Bonacina, L., Canalini, A., Antonacci, F., Marcon, M., Sarti, A., and Tubaro, S. (2016), "A low-cost solution to 3d pinna modeling for HRTF prediction", presented at *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016.

Breebaart, J. (2013), "Effect of perceptually irrelevant variance in head-related transfer functions on principal component analysis.", *Journal of the Acoustical Society of America*, **vol. 133**, no. 1, pp. EL1–6.

Breebaart, J. and Kohlrausch, A. (2001), "The perceptual (ir)relevance of HRTF magnitude and phase spectra", presented at *110th Audio Engineering Society Convention*, Amsterdam, the Netherlands, May 2001.

Bregman, A. S. (1994), *Auditory Scene Analysis: the perceptual organization of sound*, The MIT Press, 1st edn.

Brimijoin, W. O. and Akeroyd, M. A. (2014), "The moving minimum audible angle is smaller during self motion than during source motion", *Frontiers in Neuroscience*, **vol. 8**.

Bronkhorst, A. W., Veltman, J. A. H., and van Breda, L. (1996), "Application of a Three-Dimensional Auditory Display in a Flight Task", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, **vol. 38**, no. 1, pp. 23–33.

Brown, C. P. and Duda, R. O. (1998), "A structural model for binaural sound synthesis", *IEEE Transactions on Speech and Audio Processing*, **vol. 6**, no. 5, pp. 476–488.

Burkhard, M. D. and Sachs, R. M. (1975), "Anthropometric manikin for acoustic research", *Journal of the Acoustical Society of America*, **vol. 58**, no. 1, pp. 214–222.

Butler, R. A. and Belendiuk, K. (1977), "Spectral cues utilized in the localization of sound in the median sagittal plane.", *Journal of the Acoustical Society of America*, **vol. 61**, no. 5, pp. 1264–1269.

Carlile, S., Martin, R. L., and McAnally, K. I. (2005), "Spectral information in sound localization", in *Auditory Spectral Processing*, vol. 70 of *International Review of Neurobiology*, pp. 399–434, Academic Press.

Carpentier, T. (2017), "Normalization schemes in Ambisonic: does it matter?", presented at *142nd Audio Engineering Society Convention*, Berlin, Germany, May 2017.

Chandler, D. W. and Grantham, D. W. (1992), "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity", *Journal of the Acoustical Society of America*, **vol. 91**, no. 3, pp. 1624–1636.

Childers, D. G., Skinner, D. P., and Kemerait, R. C. (1977), "The cepstrum: A guide to processing", in *Proceedings of the IEEE*, vol. 65, pp. 1428–1443.

Conti, A., D'Emidio, M., Macelloni, L., Lutken, C., Asper, V., Woolsey, M., Jarnagin, R., Diercks, A., and Highsmith, R. (2016), "Morpho-acoustic characterization of natural seepage features near the macondo wellhead (ecogig site oc26, gulf of mexico)", *Deep Sea Research Part II: Topical Studies in Oceanography*, **vol. 129**, pp. 53–65.

Damaske, P. and Wagener, B. (1969), "Richtungshörversuche über einen nachgebildeten kopf", *Acustica*, **vol. 21**, pp. 30–35.

Daniel, J. (2001), "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia", PhD, Université de Paris 6, Paris, France.

Daniel, J. and Moreau, S. (2004), "Further study of sound field coding with higher order ambisonics", presented at *116th Audio Engineering Society Convention*, Berlin, Germany, May 2004.

Daniel, J., Moreau, S., and Nicol, R. (2003), "Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging", presented at *114th Audio Engineering Society Convention*, Amsterdam, the Netherlands, March 2003.

Davis, M. F. (2003), "History of spatial coding", *Journal of the Audio Engineering Society*, **vol. 51**, no. 6, pp. 554–569.

Diestel, R. (2005), *Graph Theory*, Springer-Verlag, 3rd edn.

Dolby Laboratories (2014), "Dolby atmos next-generation audio for cinema", Available online: http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-next-generation-audio-for-cinema-white-paper.pdf, (Accessed: 09/03/2017).

Dolby Laboratories (2016), "Dolby atmos home theater installation guidelines", Available online: http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-home-theater-installation-guidelines.pdf, (Accessed: 11/03/2017).

Duda, R. O., Avendano, C., and Algazi, V. R. (1999), "An adaptable ellipsoidal head model for the interaural time difference", in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1999)*, pp. 965–968, Phoenix, United States.

Duda, R. O. and Martens, W. L. (1998), "Range dependence of the response of a spherical head model", *Journal of the Acoustical Society of America*, **vol. 104**, no. 5, pp. 3048–3058.

Encyclopædia Britannica (2018), "Human Ear - Organ of Corti | Britannica.com", (Online) Available: https://www.britannica.com/science/ear/Organ-of-Corti, (Accessed: 19/05/2018).

Ericson, C. (2005), *Real Time Collision Detection*, Morgan Kaufmann Publishers, San Francisco, USA, 1st edn.

Evans, M. J., Angus, J. A. S., and Tew, A. I. (1997), "Spherical harmonic spectra of head-related transfer functions", presented at *103rd Audio Engineering Society Convention*, New York, USA, September 1997.

Evans, M. J., Angus, J. A. S., and Tew, A. I. (1998), "Analyzing head-related transfer function measurements using surface spherical harmonics", *The Journal of the Acoustical Society of America*, **vol. 104**, no. 4, pp. 2400–2411.

Farina, A. (2000), "Simultaneous measurement of impulse response and distortion with a swept-sine technique", presented at *108th Audio Engineering Society Convention*, Paris, France, February 2000.

Fastl, H. and Zwicker, E. (2007), *Psychoacoustics: Facts and Models*, Springer, Berlin, 3rd edn.

Fausti, S. A., Erickson, D. A., Frey, R. H., Rappaport, B. Z., and Schechter, M. A. (1981), "The effects of noise upon human hearing sensitivity from 8000 to 20 000 Hz", *Journal of Acoustical Society of America*, **vol. 69**, no. 5, pp. 1343–1349.

Fowler, J. E. (2005), "The redundant discrete wavelet transform and additive noise", *IEEE Signal Processing Letters*, **vol. 12**, no. 9, pp. 629–632.

Frey, R., Volodin, I., and Volodina, E. (2007), "A nose that roars: anatomical specializations and behavioural features of rutting male saiga", *Journal of Anatomy*, **vol. 211**, no. 6, pp. 717–736.

Gaik, W. (1993), "Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling", *Journal of the Acoustical Society of America*, **vol. 94**, no. 1, pp. 98–110.

Gardner, M. B. and Gardner, R. S. (1973), "Problem of localization in the median plane: effect of pinnae cavity occlusion", *Journal of the Acoustical Society of America*, **vol. 53**, no. 2, pp. 400–408.

Gardner, W. G. (1997), "Head tracked 3-d audio using loudspeakers", presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 1997)*, New Paltz, United States, October 2015.

Gardner, W. G. and Martin, K. D. (1995), "HRTF measurements of a KEMAR", *Journal of the Acoustical Society of America*, **vol. 97**, no. 6, pp. 3907–3908.

Geisler, C. D. (1998), *From sound to synapse: physiology of the mammalian ear*, Oxford University Press, New York, USA.

Geronazzo, M., Spagnol, S., and Avanzini, F. (2010), "Estimation and modeling of pinna-related transfer functions", in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*.

Gerzon, M. A. (1973), "Periphony: With-height sound reproduction", *Journal of the Audio Engineering Society*, **vol. 21**, no. 1, pp. 2–10.

Gerzon, M. A. (1992a), "Optimum reproduction matrices for multispeaker stereo", *Journal of the Audio Engineering Society*, **vol. 40**, no. 7/8, pp. 571–589.

Gerzon, M. A. (1992b), "Psychoacoustic decoders for multispeaker stereo and surround sound", presented at *93rd Audio Engineering Society Convention*, San Francisco, USA, October 1992.

Gerzon, M. A. and Barton, G. J. (1992), "Ambisonic decoders for HDTV", presented at *92nd Audio Engineering Society Convention*, Vienna, Austria, March 1992.

Glasberg, B. R. and Moore, B. C. J. (1990), "Derivation of auditory filter shapes from notched-noise data", *Hearing research*, **vol. 47**, no. 1, pp. 103–138.

Glasberg, B. R. and Moore, B. C. J. (2000), "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise", *Journal of the Acoustical Society of America*, **vol. 08**, no. 5, pp. 2318–2328.

Grantham, D. W. (1986), "Detection and discrimination of simulated motion of auditory targets in the horizontal plane", *Journal of the Acoustical Society of America*, **vol. 79**, no. 6, pp. 1939–1949.

Green, D. M., Kidd, G. K. J., and Stevens, K. N. (1987), "High-frequency audiometric assessment of a young adult population", *Journal of the Acoustical Society of America*, **vol. 81**, no. 2, pp. 485–494.

Gumerov, N. A. and Duraiswami, R. (2009), "A broadband fast multipole accelerated boundary element method for the three dimensional helmholtz equation", *Journal of the Acoustical Society of America*, **vol. 125**, no. 1, pp. 191–205.

Gumerov, N. A., O'Donovan, A. E., Duraiswami, R., and Zotkin, D. N. (2010), "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation", *Journal of the Acoustical Society of America*, **vol. 127**, no. 1, pp. 370–386.

Hassager, H. G., Dau, T., and Gran, F. (2014), "The effect of spectral details on the externalization of sounds", presented at *Forum Acusticum*, Kraków, Poland, September 2014.

Hassager, H. G., Gran, F., and Dau, T. (2016), "The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment", *Journal of the Acoustical Society of America*, **vol. 139**, no. 5, pp. 2992–3000.

Hatziantoniou, P. D. and Mourjopoulos, J. N. (2000), "Generalized fractional-octave smoothing of audio and acoustic responses", *Journal of the Audio Engineering Society*, **vol. 48**, no. 4, pp. 259–280.

Hebrank, J. and Wright, D. (1974), "Spectral cues used in the localization of sound sources on the median plane", *Journal of the Acoustical Society of America*, **vol. 56**, no. 6, pp. 1829–1834.

Heffner, H. E. and Heffner, R. S. (2007), "Hearing ranges of laboratory animals", *Journal of the American Association for Laboratory Animal Science*, **vol. 46**, no. 1, pp. 20–22.

Hetherington, C. and Tew, A. I. (2003a), "Parameterizing human pinna shape for the estimation of head-related transfer functions", presented at *114th Audio Engineering Society Convention*, Amsterdam, The Netherlands, March 2003.

Hetherington, C. and Tew, A. I. (2003b), "Three-dimensional elliptic fourier methods for the parameterization of human pinna shape", in *Proceedings IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), pp. 612–615.

HFC (2016), "Construction of a hrtf test site", (Online) Available: `http://human-factors-consult.de/en/2014/construction-of-a-hrtf-test-site/`, (Accessed: 11/03/2017).

Hobden, L. J. and Tew, A. I. (2015), "Investigating head-related transfer function smoothing using a sagittal-plane localization model", presented at *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2015)*, New Paltz, United States, October 2015.

Hofman, P. M., Van Riswick, J. G. A., and Van Opstal, A. J. (1998), "Relearning sound localization with new ears", *Nature Neuroscience*, **vol. 1**, no. 5, pp. 417–421.

Holmes, S. A. and Featherstone, W. E. (2002), "A unified approach to the clenshaw summation and the recursive computation of very high degree and order normalised associated legendre functions", *Journal of Geodesy*, **vol. 76**, no. 5, pp. 279–299.

Huttunen, T., Tuppurainen, K., Vanne, A., Ylä-Oijala, P., Järvenpää, S., Kärkkäinen, A., and Kärkkäinen, L. (2013), "Simulation of the head-related transfer functions using cloud computing", in *Proceedings of Meetings on Acoustics*, vol. 19.

Huttunen, T., Vanne, A., Harder, S., Paulsen, R. R., King, S., Perry-Smith, L., and Kärkkäinen, L. (2014), "Rapid generation of personalized HRTFs", presented at *55th International Audio Engineering Society Conference: Spatial Audio*, Helsinki, Finland, August 2014.

Hwang, S. and Park, Y. (2007), "HRIR customization in the median plane via principal components analysis", presented at *31st International Audio Engineering Society Conference: New Directions in High Resolution Audio*, London, UK, June 2007.

Iida, K. (2008), "A pair of spectral notches which plays a role as a spectral cue in the vertical localization, and it's application to estimation of sound source elevation from binaural signals", *Journal of the Acoustical Society of America*, **vol. 123**, no. 5, pp. 3456–3456.

IRCAM (2003), "Listen HRTF database", (Online) Available: `http://recherche.ircam.fr/equipes/salles/listen/index.html`, (Accessed: 11/03/2017).

ITU (1992), "Multichannel stereophonic sound system with and without accompanying picture", International Telecommunication Union (ITU) Reccomendation BS.775, Available online: `https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf`.

ITU (2009), "Multichannel sound technology in home and broadcasting applications", International Telecommunication Union (ITU) Reccomendation BS.2159, Available online: `https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BS.2159-6-2013-PDF-E.pdf`.

Jackson, L. L., Heffner, R. S., and Heffner, H. E. (1999), "Free-field audiogram of the japanese macaque (macaca fuscata)", *Journal of the Acoustical Society of America*, **vol. 106**, no. 5, pp. 3017–3023.

Jin, C. T., Corderoy, A., Carlile, S., and van Schaik, A. (1999), "Spectral cues in human sound localization", in *Proceedings of the Neural Information Processing Systems Conference, NIPS*, pp. 768–774.

Jin, C. T., Corderoy, A., Carlile, S., and Van Schaik, A. (2004), "Contrasting monaural and interaural spectral cues for human sound localization", *Journal of the Acoustical Society of America*, **vol. 115**, no. 6, pp. 3124–3141.

Jin, C. T., Epain, N., and Parthy, A. (2014), "Design, optimization and evaluation of a dual-radius spherical microphone array", *IEEE Transactions on Audio, Speech and Language Processing*, **vol. 22**, no. 1, pp. 193–204.

Jin, C. T., Guillon, P., Epain, N., Zolfaghari, R., van Schaik, A., Tew, A. I., Hetherington, C., and Thorpe, J. B. A. (2013), "Creating the Sydney York morphological and acoustic recordings of ears database", *IEEE Transactions on Multimedia*, **vol. 16**, no. 1, pp. 37–46.

Johnson, D. L. and von Gierke, H. (1974), "Audibility of infrasound", *Journal of the Acoustical Society of America*, **vol. 56**, no. S1, pp. S37–S37.

Jot, J.-M., Larcher, V., and Warusfel, O. (1995), "Digital signal processing issues in the context of binaural and transaural stereophony", presented at *98th Audio Engineering Society Convention*, Paris, France, February 1995.

Juang, B.-H., Rabiner, L. R., and Wilpon, J. G. (1987), "On the use of bandpass liftering in speech recognition", *IEEE Transactions on Acoustics, Speech and Signal Processing*, **vol. 35**, no. 7, pp. 947–954.

Kahana, Y. and Nelson, P. A. (2006), "Numerical modelling of the spatial acoustic response of the human pinna", *Journal of Sound and Vibration*, **vol. 292**, pp. 148–178.

Kahana, Y. and Nelson, P. A. (2007), "Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models", *Journal of Sound and Vibration*, **vol. 300**, pp. 552–579.

Kahana, Y., Nelson, P. A., Petyt, M., and Choi, S. (1999), "Numerical modelling of the transfer functions of a dummy-head and of the external ear", presented at *16th International Audio Engineering Society Conference: Spatial Sound Reproduction*, Rovaniemi, Finland, March 1999.

KAIST (2016), "Development of the basic technology of the korean HRTF", (Online) Available: `http://sdac.kaist.ac.kr/project/index.php?mode=completed&act=KHRTF`, (Accessed: 11/03/2017).

Kan, A., Jin, C. T., and van Schaik, A. (2006), "Distance variation function for simulation of near-field virtual auditory space", in *Proceedings IEEE Interna-*

tional Conference on Acoustics, Speech and Signal Processing (ICASSP 2006), pp. 325–328.

Kan, A., Jin, C. T., and van Schaik, A. (2009), "A psychophysical evaluation of near-field head-related transfer functions synthesized using a distance variation function.", *Journal of the Acoustical Society of America*, **vol. 125**, no. 4, pp. 2233–2242.

Karjalainen, M., Harma, A., and Laine, U. K. (1997), "Realizable warped iir filters and their properties", in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 1997)*, pp. 2205–2208.

Kärkkäinen, A., Kärkkäinen, L., and Huttunen, T. (2013), "Practical procedure for large scale personalized head related transfer function acquisition", presented at *51st International Audio Engineering Society Conference: Loudspeakers and Headphones*, Helsinki, Finland, August 2013.

Katz, B. F. G. (2001a), "Boundary element method calculation of individual head-related transfer function. I. rigid model calculation", *Journal of the Acoustical Society of America*, **vol. 110**, no. 5, pp. 2440–2448.

Katz, B. F. G. (2001b), "Boundary element method calculation of individual head-related transfer function. II. impedance effects and comparisons to real measurements", *Journal of the Acoustical Society of America*, **vol. 110**, no. 5, pp. 2449–2455.

Katz, B. F. G. and Noisternig, M. (2014), "A comparative study of interaural time delay estimation methods", *Journal of the Acoustical Society of America*, **vol. 135**, no. 6, pp. 3530–3540.

Kelly, M. C. (2002), "Efficient representation of adaptable virtual auditory space", Phd, Department of Electronics, University of York, York, UK.

Kidd, G., Mason, C. R., Brantley, M. A., and Owen, G. A. (1989), "Roving-level tone-in-noise detection", *Journal of Acoustical Society of America*, **vol. 86**, no. 4, pp. 1310–1317.

Kirkeby, O., Nelson, P. A., and Hamada, H. (1998a), "Digital filter design for virtual source imaging systems", presented at *104th Audio Engineering Society Convention*, Amsterdam, Netherlands, May 1998.

Kirkeby, O., Nelson, P. A., and Hamada, H. (1998b), "Local sound field reproduction using two closely spaced loudspeakers", *The Journal of the Acoustical Society of America*, **vol. 104**, no. 4, pp. 1973–1981.

Kirkeby, O., Rubak, P., Nelson, P. A., and Farina, A. (1999), "Design of cross-talk cancellation networks by using fast deconvolution", presented at *116th Audio Engineering Society Convention*, Munich, Germany, May 1999.

Kirkeby, O., Seppälä, E. T., Kärkkäinen, A., Kärkkäinen, L., and Huttunen, T. (2007), "Some effects of the torso on head-related transfer functions", presented at *122nd Audio Engineering Society Convention*, Vienna, Austria, May 2007.

Kistler, D. J. and Wightman, F. L. (1992), "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction", *Journal of the Acoustical Society of America*, **vol. 91**, no. 3, pp. 1637–1647.

Kolla, V. and Coumes, F. (1983), "Morpho-acoustic and sedimentologic characteristics of the indus fan", *Geo-Marine Letters*, **vol. 3**, no. 2-4, pp. 133–139.

Kreuzer, W., Majdak, P., and Chen, Z. (2009), "Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range", *Journal of the Acoustical Society of America*, **vol. 126**, no. 3, pp. 1280–1290.

Kuhn, G. F. (1977), "Model for the interaural time differences in the azimuthal plane", *Journal of the Acoustical Society of America*, **vol. 62**, no. 1, pp. 157–167.

Kulkarni, A. and Colburn, H. S. (1998), "Role of spectral detail in sound-source localization", *Nature*, **vol. 396**, pp. 747–749.

Kulkarni, A. and Colburn, H. S. (2004), "Infinite-impulse-response models of the head-related transfer function", *Journal of the Acoustical Society of America*, **vol. 115**, no. 4, pp. 1714–1728.

Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (1999), "Sensitivity of human subjects to head-related transfer-function phase spectra", *Journal of the Acoustical Society of America*, **vol. 105**, no. 5, pp. 2821–2840.

Laitinen, M.-V. and Pulkki, V. (2009), "Binaural reproduction for directional audio coding", in *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2009)*, pp. 337–340.

Leek, M. R. (2001), "Adaptive procedures in psychophysical research", *Perception & Psychophysics*, **vol. 63**, no. 8, pp. 1279–1292.

Levitt, H. (1971), "Transformed up-down methods in psychoacoustics", *Journal of the Acoustical Society of America*, **vol. 49**, no. 2B, pp. 467–477.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (2014), "A spatial audio quality inventory (SAQI)", *Acta Acustica*, **vol. 100**, no. 5, pp. 984–994.

Loomis, J. M., Golledge, R. G., and Klatzky, R. L. (1998), "Navigation system for the Blind: Auditory display modes and guidance", *Presence: Teleoperators and Virtual Environments*, **vol. 7**, no. 2, pp. 193–203.

Lopez-Poveda, E. A. and Meddis, R. (1996), "A physical model of sound diffraction and reflections in the human concha", *Journal of the Acoustical Society of America*, **vol. 100**, no. 5, pp. 3248–3259.

Ma, F., Wu, J. H., Huang, M., Zhang, W., Hou, W., and Bai, C. (2015), "Finite element determination of the head-related transfer function", *Journal of Mechanics in Medicine and Biology*, **vol. 15**, no. 5.

Majdak, P., Balazs, P., and Laback, B. (2007), "Multiple exponential sweep

method for fast measurement of head-related transfer functions", *Journal of the Audio Engineering Society*, **vol. 55**, no. 7/8, pp. 623–637.

Malham, D. G. (2003), "Space in music - music in space", MPhil, Department of Music, University of York, York, UK.

Marston, J. R., Loomis, J. M., Klatzky, R. L., and Golledge, R. G. (2007), "Nonvisual route following with guidance from a simple haptic or auditory display.", *Journal of Visual Impairment & Blindness*, **vol. 101**, no. 4, pp. 203–211.

Martens, W. L. (1987), "Principal components analysis and resynthesis of spectral cues to perceived direction", in *Proceedings of the International Computer Music Conference*, pp. 274–281, San Francisco, United States.

Marvit, P., Florentine, M., and Buus, S. (2003), "A comparison of psychophysical procedures for level-discrimination thresholds", *Journal of the Acoustical Society of America*, **vol. 113**, no. 6, pp. 3348–3361.

Masiero, B. and Fels, J. (2011), "Perceptually robust headphone equalization for binaural reproduction", presented at *130th Audio Engineering Society Convention*, London, UK, May 2011.

Mehrgardt, S. and Mellert, V. (1977), "Transformation characteristics of the external human ear.", *Journal of the Acoustical Society of America*, **vol. 61**, no. 6, pp. 1567–1576.

Middlebrooks, J. C. (1999), "Individual differences in external-ear transfer functions reduced by scaling in frequency", *Journal of the Acoustical Society of America*, **vol. 106**, no. 3, pp. 1480–1492.

Middlebrooks, J. C. and Green, D. M. (1990), "Directional dependence of interaural envelope delays", *Journal of the Acoustical Society of America*, **vol. 87**, no. 5, pp. 2149–2162.

Middlebrooks, J. C. and Green, D. M. (1992), "Observations on a principal components analysis of head-related transfer functions", *Journal of the Acoustical Society of America*, **vol. 92**, no. 1, pp. 597–599.

Middlebrooks, J. C., Makous, J. C., and Green, D. M. (1989), "Directional sensitivity of sound - pressure levels in the human ear canal", *Journal of the Acoustical Society of America*, **vol. 86**, no. 1, pp. 89–107.

Mills, A. W. (1958), "On the minimum audible angle", *Journal of the Acoustical Society of America*, **vol. 30**, no. 4, pp. 237–246.

Mills, A. W. (1972), "Auditory localisation", in *Foundations of Modern Auditory Theory* (edited by Tobias, J. V.), vol. 2, pp. 303–348, Academic Press, New York.

Minnaar, P., Plogsties, J., Olesen, S. K., Christensen, F., and Møller, H. (2000), "The interaural time difference in binaural synthesis", presented at *108th Audio Engineering Society Convention*, Paris, France, February 2000.

Mokhtari, P., Takemoto, H., Adachi, S., Nishimura, R., and Kato, H. (2013), "Three-dimensional acoustic sensitivity analysis of pinna geometry at peaks of head-related transfer functions", in *Proceedings of the Autumn Meeting of the Acoustical Society of Japan*, pp. 861–864.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2007), "Comparison of simulated and measured HRTFs: FDTD simulation using MRI head data", presented at *123rd Audio Engineering Society Convention*, New York, United States, October 2007.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2009), "Acoustic simulation of KEMAR's HRTFs : verification with measurements and the effects of modifying head shape and pinna concavity", presented at *International Workshop on the Principles and Applications of Spatial Hearing (IWPASH 2009)*, Miyagi, Japan, November 2009.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2010), "Acoustic sensitivity to micro-perturbations of KEMAR's pinna surface geometry", in *Proceedings of the International Congress on Acoustics (ICA 2010)*, Sydney, Australia.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2011), "Pinna sensitivity patterns reveal reflecting and diffracting surfaces that generate the first spectral notch in the front median plane", in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, pp. 2408–2411.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2014), "Visualization of acoustic pressure and velocity patterns with phase information in the pinna cavities at normal modes", presented at *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, Florence, Italy, May 2014.

Mokhtari, P., Takemoto, H., Nishimura, R., and Kato, H. (2015), "Frequency and amplitude estimation of the first peak of head-related transfer functions from individual pinna anthropometry", *Journal of the Acoustical Society of America*, **vol. 137**, no. 2, pp. 690–701.

Møller, A. R. (2000), *Hearing: its physiology and pathophysiology*, Academic Press.

Møller, A. R. (2012), *Hearing: anatomy, physiology, and disorders of the auditory system*, Plural Publishing.

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

Møller, H., Hammershøi, D., Jensen, C. B., and Sørensen, M. F. (1995a), "Transfer characteristics of headphones measured on human ears", *Journal of the Audio Engineering Society*, **vol. 43**, no. 4, pp. 203–217.

Møller, H., Sørensen, M. F., Hammershøi, D., and Jensen, C. B. (1995b), "Head-related transfer functions of human subjects", *Journal of the Audio Engineering Society*, **vol. 43**, no. 5, pp. 300–321.

Møller, H., Sørensen, M. F., Jensen, C. B., and Hammershøi, D. (1996), "Binaural technique: Do we need individual recordings?", *Journal of the Audio Engineering Society*, **vol. 44**, no. 6, pp. 451–469.

Moore, B. C. J. (2013), *An Introduction to the Psychology of Hearing*, Koninklijke Brill NV, Leiden, The Netherlands, 6th edn.

Moreau, S., Daniel, J., and Bertet, S. (2006), "3D sound field recording with higher order ambisonics – objective measurements and validation of spherical microphone", presented at *120th Audio Engineering Society Convention*, Paris, France, May 2006.

Morimoto, M. (2001), "The contribution of two ears to the perception of vertical angle in sagittal planes", *Journal of the Acoustical Society of America*, **vol. 109**, no. 4, pp. 1596–1603.

Muraoka, H., Takahashi, Y., Haedar, T., Otani, M., and Hirahara, T. (2007), "The head related transfer function simulation by FEM/IEM and reciprocal theorem", presented at *14th International Congress on Sound and Vibration (ICSV14)*, Cairns, Australia, July 2007.

Musicant, A. D. and Butler, R. A. (1984), "The influence of pinnae-based spectral cues on sound localization", *Journal of the Acoustical Society of America*, **vol. 75**, no. 4, pp. 1195–1200.

Numerical Tours (2017a), "Numerical tours - a numerical tour of data science", Available online: `http://http://www.numerical-tours.com/`, (Accessed: 16/09/2017).

Numerical Tours (2017b), "Spherical mesh parameterization", Available online: `http://www.numerical-tours.com/matlab/meshdeform_2_parameterization_sphere/`, (Accessed: 16/09/2017).

Open University (2018), "Hearing 3.2 The anatomy of the cochlea - OpenLearn - Open University - SD329_1", (Online) Available: http://www.open.edu/openlearn/science-maths-technology/science/biology/hearing/content-section-3.2, (Accessed: 19/05/2018).

Oppenheim, A. V. and Schafer, R. W. (2004), "From frequency to quefrency: a history of the cepstrum", *IEEE Signal Processing Magazine*, **vol. 21**, no. 5, pp. 95–106.

Paquier, M. and Koehl, V. (2010), "Audibility of headphone positioning variability", presented at *128th Audio Engineering Society Convention*, London, UK, May 2010.

Park, K. S. and Lee, N. S. (1987), "A Three-Dimensional Fourier Descriptor for Human Body Representation/Reconstruction from Serial Cross Sections", *Computer and Biomedical Research*, **vol. 20**, pp. 125–140.

Patterson, R. D. (1976), "Auditory filter shapes derived with noise stimuli", *Journal of the Acoustical Society of America*, **vol. 59**, no. 3, pp. 640–654.

Patterson, R. D. and Moore, B. C. J. (1986), "Auditory filters and excitation patterns as representations of frequency resolution", in *Frequency Selectivity in Hearing*, pp. 123–177, Academic Press Limited, London.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987), "An efficient auditory filterbank based on the gammatone function", presented at a Speech-Group meeting of the Institute of Acoustics on Auditory Modelling, RSRE, Malvern, UK, December 1987.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992), "Complex sounds and auditory images", in *Auditory Physiology and Perception, proceedings of the 9th International Symposium on Hearing*, pp. 429–446, Pergamon, Oxford.

Pec, M., Bujacz, M., and Strumiłło, P. (2013), "Low-order modelling of head related transfer functions based on spectral smoothing and principal component

analysis", in *Proceedings of the International Conference on Auditory Displays (ICAD 2013)*, pp. 207–212, Łódź, Poland.

Pedersen, T. H. and Zacharov, N. (2015), "The development of a sound wheel for reproduced sound", presented at *138th Audio Engineering Society Convention*, Warsaw, Poland, May 2015.

Pei, S.-C. and Lin, H.-S. (2006), "Minimum-phase fir filter design using real cepstrum", *IEEE Transactions on Circuits and Systems II: Express Briefs*, **vol. 53**, no. 10, pp. 1113–1117.

Perrott, D. R. (1984), "Concurrent minimum audible angle: A re-examination of the concept of auditory spatial acuity", *Journal of the Acoustical Society of America*, **vol. 75**, no. 4, pp. 1201–1206.

Perrott, D. R. and Musicant, A. D. (1977), "Minimum auditory movement angle: Binaural localization of moving sound sources", *Journal of the Acoustical Society of America*, **vol. 62**, no. 6, pp. 1463–1466.

Perrott, D. R. and Saberi, K. (1990), "Minimum audible angle thresholds for sources varying in both elevation and azimuth", *Journal of the Acoustical Society of America*, **vol. 87**, no. 4, pp. 1728–1731.

Perrott, D. R. and Tucker, J. (1988), "Minimum audible movement angle as a function of signal frequency and the velocity of the source", *Journal of the Acoustical Society of America*, **vol. 83**, no. 4, pp. 1522–1527.

Petyt, M., Lea, J., and Koopmann, G. H. (1976), "A finite element method for determining the acoustic modes of irregular shaped cavities", *Journal of Sound and Vibration*, **vol. 45**, no. 4, pp. 495–502.

Peyré, G. (2011a), "THE NUMERICAL TOURS OF SIGNAL PROCESSING", *Computing in Science & Engineering*, **vol. 13**, no. 4, pp. 94–97.

Peyré, G. (2011b), "THE NUMERICAL TOURS OF SIGNAL PROCESSING

PART 2: MULTISCALE PROCESSINGS", *Computing in Science & Engineering*, **vol. 13**, no. 5, pp. 68–71.

Peyré, G. (2011c), "THE NUMERICAL TOURS OF SIGNAL PROCESSING PART 3: IMAGE AND SURFACE RESTORATION", *Computing in Science & Engineering*, **vol. 13**, no. 6, pp. 72–75.

Pulkki, V. (1997), "Virtual sound source positioning using vector base amplitude panning", *Journal Audio Engineering Society*, **vol. 45**, no. 6, pp. 456–466.

Pulkki, V. and Karjalainen, M. (2015), *Communication Acoustics: an Introduction to Speech, Audio and Psychoacoustics*, John Wiley & Sons, Chichester, UK, 1st edn.

Pulkki, V., Lokki, T., and Rocchesso, D. (2011), "Spatial effects", in *DAFX: Digital Audio Effects* (edited by Zölzer, U.), John Wiley and Sons, Chichester, UK.

Rao, D. and Xie, B. (2005), "Head rotation and sound image localization in the median plane", *Chinese Science Bulletin*, **vol. 50**, no. 5, pp. 412–416.

Rasumow, E., Blau, M., Hansen, M., van de Par, S., Doclo, S., and Mellert, V. (2014), "Smoothing individual head-related transfer functions in the frequency and spatial domains", *Journal of the Acoustical Society of America*, **vol. 135**, no. 4, pp. 2012–2025.

Raykar, V. C., Duraiswami, R., Davis, L., and Yegnanarayana, B. (2003), "Extracting significant features from the HRTF", in *Proceedings of the International Conference on Auditory Display (ICAD03)*, pp. 115–118, Boston, United States.

Raykar, V. C., Duraiswami, R., and Yegnanarayana, B. (2005), "Extracting the frequencies of the pinna spectral notches in measured head related impulse responses", *Journal of the Acoustical Society of America*, **vol. 118**, no. 1, pp. 364–374.

Rayleigh, L. (1904), "On the acoustic shadow of a sphere", *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **vol. 203**, no. 359-371, pp. 87–110.

Rayleigh, L. (1907), "On our perception of sound direction", *Philosophical Magazine*, **vol. 13**, pp. 214–232.

Riederer, K. A. J. (2012), "KAR HRTF measurements, analysis, evaluation and implementation", (Online) Available: `http://www.kar.fi/KARAudio/Services/KARAcoustics_3DAudio.html`, (Accessed: 11/03/2017).

Rife, D. D. and Vanderkooy, J. (1989), "Transfer-function measurement with maximum-length sequences", *Journal of the Audio Engineering Society*, **vol. 37**, no. 6, pp. 419–444.

Rui, Y., Yu, G., Xie, B., and Liu, Y. (2013), "Calculation of individualized near-field head-related transfer function database using boundary element method", presented at *134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

Rumsey, F. (2001), *Spatial Audio*, Focal Press, Oxford, UK.

SADIE (2016), "Sadie binaural measurements: Head related transfer function measurements for virtual loudspeaker rendering", (Online) Available: `http://www.york.ac.uk/sadie-project/binaural.html`, (Accessed: 11/03/2017).

Satarzadeh, P., Algazi, V. R., and Duda, R. O. (2007), "Physical and filter pinna models based on anthropometry", Vienna, Austria, May 2007.

Schechter, M. A., Fausti, S. A., Rappaport, B. Z., and Frey, R. H. (1986), "Age categorization of high-frequency auditory threshold data", *Journal of the Acoustical Society of America*, **vol. 79**, no. 3, pp. 767–771.

Schroeder, M. R. (1975), "Models of hearing", .

Senova, M. A., McAnally, K. I., and Martin, R. L. (2002), "Localization of virtual sound as a function of head-related impulse response duration", *Journal of the Audio Engineering Society*, **vol. 50**, no. 1/2, pp. 57–66.

Shaw, E. A. G. (1974a), "The external ear", in *Auditory System: Anatomy Physiology (Ear)* (edited by Keidel, W. and Neff, W. D.), pp. 455–490, Springer Berlin Heidelberg, Berlin, Heidelberg.

Shaw, E. A. G. (1974b), "Transformation of sound pressure level from the free field to the eardrum in the horizontal plane", *Journal of the Acoustical Society of America*, **vol. 56**, no. 6, pp. 1848–1861.

Shaw, E. A. G. (1975), "The external ear: new knowledge", in *Earmoulds and Associated Problems, Proceedings of the Seventh Danavox Symposium, Scandinavian Audiology* (edited by S. Christensen-Dalsgaard), pp. 24–50.

Shaw, E. A. G. (1979), "The elusive connection: 1979 rayleigh medal lecture", presented at *Annual Meeting of the Institute of Acoustics*, Southampton, UK, April 1979.

Shaw, E. A. G. (1997), "Acoustical features of the human external ear", in *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 25–48, Lawrence Erlbaum Associates, Mahwah, United States.

Shaw, E. A. G. and Teranishi, R. (1968), "Sound pressure generated in an external-ear replica and real human ears by a nearby point source", *Journal of the Acoustical Society of America*, **vol. 44**, no. 1, pp. 240–249.

Shepherd, D. and Hautus, M. J. (2007), "The measurement problem in level discrimination", *Journal of the Acoustical Society of America*, **vol. 121**, no. 4, pp. 2158–2167.

Shub, D. E., Durlach, N. I., and Colburn, H. S. (2008), "Monaural level discrimination under dichotic conditions", *Journal of the Acoustical Society of America*, **vol. 123**, no. 6, pp. 4421–4433.

Singh, R., Saxena, R., and Varshney, S. (2009), "Early detection of noise induced hearing loss by using ultra high frequency audiometry", *The Internet Journal of Otorhinolaryngology*, **vol. 10**, no. 2.

Slaney, M. (1993), "An efficient implementation of the patterson-holdsworth auditory filter bank", Tech. Rep. #35, Apple Computer Inc.

Slaney, M. (1998), "Auditory toolbox, version 2", Tech. Rep. #1998-010, Interval Research Corporation.

Smith, J. O. and Abel, J. S. (1999), "Bark and ERB bilinear transforms", *IEEE Transactions on Speech and Audio Processing*, **vol. 7**, no. 6, pp. 697–708.

Søndergaard, P. L. and Majdak, P. (2013), "The auditory modelling toolbox", in *The Technology of Binaural Listening* (edited by Blauert, J.), chap. 2, pp. 33–56, Springer.

Spagnol, S. (2015), "On distance dependence of pinna spectral patterns in head-related transfer functions", *Journal of the Acoustical Society of America*, **vol. 137**, no. 1, pp. EL58–EL64.

Spagnol, S. and Avanzini, F. (2015), "Frequency estimation of the first pinna notch in head-related transfer functions with a linear anthropometric model", in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, Trondheim, Norway.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2010), "Structural modeling of pinna-related transfer functions", in *Proceedings of the International Conference on Sound and Music Computing (SMC 2010)*, pp. 422–428.

Spagnol, S., Geronazzo, M., and Avanzini, F. (2013), "On the relation between pinna reflection patterns and head-related transfer function features", *IEEE Transactions on Audio, Speech and Language Processing*, **vol. 21**, no. 3, pp. 508–519.

Steinberg, J. C. and Snow, W. B. (1934), "Auditory perspective — physical factors", *Transactions of the American Institute of Electrical Engineers*, **vol. 53**, no. 1, pp. 12–17.

Stelmachowicz, P. G., Beauchaine, K. A., Kalberer, A., and Jesteadt, W. (1989), "Normative thresholds in the 8- to 20-kHz range as a function of age", *Journal of the Acoustical Society of America*, **vol. 86**, no. 4, pp. 1384–1391.

Stevens, K. N., Berkovitz, R., Kidd Jr, G., and Green, D. M. (1987), "Calibration of ear canals for audiometry at high frequencies", *Journal of the Acoustical Society of America*, **vol. 81**, no. 2, pp. 470–484.

Stevens, S. S. and Newman, E. B. (1936), "The localization of actual sources of sound", *American Journal of Psychology*, **vol. 48**, no. 2, pp. 297–306.

Strutt, J. W. (1877), *The Theory of Sound*, vol. 1, Macmillan.

Takeda, S., Morioka, I., Miyashita, K., Okumura, A., Yoshida, Y., and Matsumoto, K. (1992), "Age variation in the upper limit of hearing", *European Journal of Applied Physiology and Occupational Physiology*, **vol. 65**, no. 5, pp. 403–408.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., and Iida, K. (2012), "Mechanism for generating peaks and notches of head-related transfer functions in the median plane", *Journal of the Acoustical Society of America*, **vol. 132**, no. 6, pp. 3832–3841.

Tao, Y., Tew, A. I., and Porter, S. J. (2003a), "A study on head-shape simplification using spherical harmonics for HRTF computation at low frequencies", *Journal of the Audio Engineering Society*, **vol. 51**, no. 9, pp. 799–805.

Tao, Y., Tew, A. I., and Porter, S. J. (2003b), "The differential pressure synthesis method for efficient acoustic pressure estimation", *Journal of the Audio Engineering Society*, **vol. 51**, no. 7/8, pp. 647–656.

Teranishi, R. and Shaw, E. A. G. (1968), "External-ear acoustic models with simple geometry", *Journal of the Acoustical Society of America*, **vol. 44**, no. 1, pp. 257–263.

Tew, A. I., Hetherington, C., and Thorpe, J. B. A. (2012), "Morphoacoustic perturbation analysis : principles and validation principles of MPA-FD", in *Proceedings of the Acoustics 2012 Conference*, pp. 867–872, Nantes, France.

Theile, G. and Wittek, H. (2011), "Principles in surround recordings with height", presented at *130th Audio Engineering Society Convention*, London, UK, May 2011.

Thorpe, J. B. A. (2009), "Human sound localisation cues and their relation to morphology", Phd, Department of Electronics, University of York, York, UK.

Tohoku University (2016), "Spherical speaker array system", (Online) Available: `http://www.ais.riec.tohoku.ac.jp/Lab3/sp-array/index.html`, (Accessed: 11/03/2017).

Toole, F. E. (1984), "The acoustics and psychoacoustics of headphones", presented at *2nd International Audio Engineering Society Conference: The Art and Technology of Recording*, California, USA, May 1984.

Vargha, A. and Delaney, H. D. (2000), "A critique and improvement of the "cl" common language effect size statistics of mcgraw and wong", *Journal of Educational and Behavioral Statistics*, **vol. 25**, no. 2, pp. 101–132.

Volandri, G., Carmignani, C., Di Puccio, F., and Forte, P. (2014), "Finite element formulations applied to outer ear modeling", *Strojniški vestnik-Journal of Mechanical Engineering*, **vol. 60**, no. 5, pp. 363–372.

Weinrich, S. G. (1984), "Sound field calculations around the human head", Tech. Rep. No. 37, Acoustics Laboratory, Technical University of Denmark.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993), "Localization using nonindividualized head-related transfer functions", *Journal of the Acoustical Society of America*, **vol. 94**, no. 1, pp. 111–123.

Wettschurek, R. (1971), "Über unterschiedsschwellen beim richtungshören in der medianebene", in *Gemeinschaftstagung für Akustik und Schwingungstechnik*, pp. 385–388.

Whittle, L. S., Collins, S. J., and Robinson, D. W. (1972), "The audibility of low-frequency sounds", *Journal of Sound and Vibration*, **vol. 21**, no. 4, pp. 431–448.

Wiener, F. M. (1947), "On the diffraction of a progressive sound wave by the human head", *Journal of the Acoustical Society of America*, **vol. 19**, no. 1, pp. 143–146.

Wiener, F. M. and Ross, D. A. (1946), "The Pressure Distribution in the Auditory Canal in a Progressive Sound Field", *Journal of the Acoustical Society of America*, **vol. 18**, no. 2, pp. 401–408.

Wightman, F. L. and Kistler, D. J. (1989a), "Headphone simulation of free-field listening. i: Stimulus synthesis", *Journal of the Acoustical Society of America*, **vol. 85**, no. 2, pp. 858–867.

Wightman, F. L. and Kistler, D. J. (1989b), "Headphone simulation of free-field listening. II: Psychophysical validation", *Journal of the Acoustical Society of America*, **vol. 85**, no. 2, pp. 868–878.

Woodworth, R. S. and Schlosberg, H. (1962), *Experimental Psychology*, Rinehard and Winston, New York.

Xiao, T. and Liu, Q. H. (2003), "Finite-difference computation of head-related transfer function for human hearing", *Journal of the Acoustical Society of America*, **vol. 113**, no. 5, pp. 2434–2441.

Xie, B. (2012), "Recovery of individual head-related transfer functions", *Journal of the Acoustical Society of America*, **vol. 132**, no. 1, pp. 282–294.

Xie, B. (2013), *Head-Related Transfer Function and Virtual Auditory Display*, J. Ross Publishing, Plantation, Florida, USA, 2nd edn.

Xie, B. and Zhang, T. (2010), "The audibility of spectral detail of head-related transfer functions at high frequency", *Acta Acustica united with Acustica*, **vol. 96**, pp. 328–339.

Yost, W. A. (2000), *Fundamentals of Hearing: An Introduction*, chap. 10, Academic Press, London, UK, 4th edn.

Young, K., Tew, A. I., and Kearney, G. (2016), "Boundary element method modelling of KEMAR for binaural rendering: Mesh production and validation", in *Proceedings of the Interactive Audio Systems Symposium*.

Zacharov, N., Pedersen, T. H., and Pike, C. (2016a), "A common lexicon for spatial sound quality assessment-latest developments", presented at *IEEE 8th International Conference on Quality of Multimedia Experience (QoMEX 2016)*, Lisbon, Portugal, June 2016.

Zacharov, N., Pike, C., Melchior, F., and Worch, T. (2016b), "Next generation audio system assessment using the multiple stimulus ideal profile method", presented at *IEEE 8th International Conference on Quality of Multimedia Experience (QoMEX 2016)*, Lisbon, Portugal, June 2016.

Zhang, C. and Xie, B. (2012), "Platform for virtual auditory environment real time rendering system", *Journal of the Acoustical Society of America*, **vol. 131**, no. 4, pp. 3269–3269.

Zhou, B., Green, D. M., and Middlebrooks, J. C. (1992), "Characterization of external ear impulse responses using golay codes", *Journal of the Acoustical Society of America*, **vol. 92**, no. 2, pp. 1169–1171.

Ziegelwanger, H., Kreuzer, W., and Majdak, P. (2015a), "Mesh2HRTF: An open-source software package for the numerical calculation of head-related tranfer functions", presented at *22nd International Congress of Sound and Vibration (ICSV22)*, Florence, Italy, July 2015.

Ziegelwanger, H., Kreuzer, W., and Majdak, P. (2016), "A priori mesh grading for the numerical calculation of the head-related transfer functions", *Applied Acoustics"*, **vol. 114**, pp. 99–110.

Ziegelwanger, H., Majdak, P., and Kreuzer, W. (2014), "Efficient numerical calculation of head-related transfer functions", presented at *Forum Acusticum*, Kraków, Poland, September 2014.

Ziegelwanger, H., Majdak, P., and Kreuzer, W. (2015b), "Numerical calculation of listener-specific head-related transfer functions and sound localization: Microphone model and mesh discretization", *The Journal of the Acoustical Society of America*, **vol. 138**, no. 1, pp. 208–222.

Ziegelwanger, H., Reichinger, A., and Majdak, P. (2013), "Calculation of listener-specific head-related transfer functions: Effect of mesh quality", in *Proceedings of Meetings on Acoustics*, vol. 19.

Zotkin, D. N., Duraiswami, R., Davis, L. S., Mohan, A., and Raykar, V. C. (2002), "Virtual audio system customization using visual matching of ear parameters", pp. 1003–1006, Québec City, Canada, August 2002.