Efficient and Compact Representations of Head Related Transfer Functions

---

Author:
Joseph Sinker

Supervisors:
Prof. J. Angus
Prof. T. Cox

@00233333
School of Computing, Science, and Engineering
College of Science and Technology
University of Salford, Salford, UK

Submitted in Partial Fulfilment of the Requirements of the Degree of Master of Science by Research, September 2014

Contents

List of Figures

Acknowledgements:

Firstly, I would like to thank Prof. Jamie Angus for her continued support and guidance throughout the project; allowing sufficient room for me to make progression through my own critical analysis whilst ensuring I did not stray too far from the path to completion.

Secondly, I would like to thank Prof. Trevor Cox, whose input throughout the length of the project was invaluable.

Thirdly, I would like to thank my fellow postgraduate students, who have acted as a sounding board, offering valuable suggestions and criticism throughout the project.

Finally, I would like to thank my partner Charlotte, and my parents, Ian and Amanda, who have offered me strength and support, without which I could not have completed this work.

Abstract

These days most reproduced sound is consumed using portable devices and headphones, on which spatial binaural audio can be conveniently presented. One way of converting from conventional loudspeaker formats to binaural format is through the use of Head Related Transfer Functions (HRTFs), but head-tracking is also necessary to obtain a satisfactory externalisation of the simulated sound field. Typically a large HRTF dataset is required in order to provide enough measurements for a continuous virtual auditory space to be achieved through simple linear interpolation, or similar.

This work describes an investigation into the use of alternative compact and efficient representations of an HRTF dataset measured in the azimuthal plane. The two main prongs of investigation are the use of orthogonal transformations in a decompositional approach, and parametric modelling approach that utilises techniques often associated with speech processing. The latter approach is explored through the application of a linear prediction derived all-pole model method and a pole-zero model design method proposed by Steiglitz and McBride [Steiglitz and McBride, 1965]. The all-pole model is deemed to offer superior performance in matching the measured data after compression of the HRTF set through computer simulation results, whilst a preliminary subjective validation of the pole-zero models, that contrary to theoretical driven expectations, performed considerably worse in computer simulation experiments, is conducted as a pilot study.

Consideration is also given to a method of secondary compression and interpolation that utilises the Discrete Cosine Transform applied to the angular dependent components derived from each of the approaches. It is possible that these techniques may also be useful in developing efficient schemes of custom HRTF capture.

Chapter 1

Introduction

The general public listen to audio and spatial audio content in a variety of ways; sometimes this listening occurs in the home using traditional stereo or multi-channel loudspeaker setups. However, a large amount of this content is consumed on portable media devices such as smartphones, tablets, and digital media players, all of which commonly deliver audio content over headphones. Headphone listening may well account for a majority of the listening experience of many users. This trend is echoed by the recent decisions of major broadcast companies to move some traditional television and radio programming to online only platforms, clearly illustrating a reliable and foreseeably sustainable demand for content accessible from devices other than the traditional television or kitchen radio. Therefore, there is an increasing, and urgent, need to create effective and immersive experiences for headphone listeners utilising a wide range of devices.

Furthermore, in recent years, media production facilities have also expressed increased tendency toward open plan, or 'transparent', workspaces, which may contain a multitude of occupants, many of which may be tasked with the production of audio content for various delivery platforms. This in turn illustrates an increased value in the accurate simulation of different listening environments or various loudspeaker formations, without the need for the physical space required to house conventional loudspeaker setups, let alone the space and accuracy of placement required to utilise higher order formations for spatial platforms such as ambisonics. This increase in value is also prompted by the seemingly exponentially growing number of 'budget' producers of audio and video content, that come along with the

ever falling cost of enabling software and technologies, and that lack the typically necessary equipment to trial audio material across multiple or even a single correctly realised reproduction system(s).

Stereo headphones present a convenient and well realised platform for the delivery of spatial audio content. Headphones lend themselves particularly well to portability and use in multi-person environments. The acoustic signature of a listening environment, or a specific loudspeaker setup, is characterised by the relationship of the sound incident on each of a listener's two ears from each of the sound sources present in the auditory space. Head-related transfer functions (HRTFs), describe the associated acoustic signal incident on each ear as a function of source location. Using a set of HRTFs measured at a specific listener's ears, at the ears of a generalised mannequin of a head and possibly torso, or even via consideration of an analytical head model such as a sphere, positional cues can be synthesised for any number of discrete audio signals. As is imposed by the physical form of a pair of headphones, the resulting audio scene is reproduced through two discrete channels, feeding directly into the left and right ears individually. Commonly referred to as Binaural Stereo, this technique is the only effective method of rendering spatial audio content to a listener wearing headphones.

Binaural stereo audio is a well documented spatial audio technique, with implementations on a wide range of systems and devices. However, the majority of current implementations make use of large databanks of head-related transfer functions or head-related impulse responses, in order to represent the auditory space around a listener's head in as much detail as possible. For each possible location for which a sound can be synthesised, a pair of HRTFs or corresponding head related impulse responses (HRIRs) must be stored. Considering that HRIRs are commonly between 256 and 2048 samples long, it is clear that for accurate reproduction purposes, a large number of HRTF/HRIR elements must be stored within the system.

It is therefore desirable to be able to represent the data required to create, or recreate, a virtual auditory space in a more efficient or compact form, without the loss of the significant directional information that allows the listener to interpret the location of the various sources within the scene.

The works described in this thesis comprise of an investigative exploration of techniques that can be used to achieve a more efficient means of 'handling' the HRTF data required to achieve adequate coverage of a virtual auditory space. Previous approaches are broken down into three main categories; decompositional, filter modelling, and interpolation led, and are discussed at some length. Following discussion, the thesis presents and discusses the results and implications of the application, and in some instances, the extension of a decompositional approach and two filter modelling approaches applied to a set of HRIR measurements made in the azimuth plane. Objective analysis is performed for all three methods through simulation of results with a subjective analysis of the most promising of the three methods conducted in parallel.

The remainder of this thesis will be structured in the following manner; the introduction is followed by a literature review section in which the reader will be led through a summary and consideration of previous works pertaining to the thesis topic of efficient and compact HRTF representations; highlighting the use and validation of techniques that will be adapted to form a large portion of the works described by the thesis. The thesis will then detail the application and experimentation of several methods of HRTF compression both adapted, and in some cases, extended from the techniques outlined in the literature review, this section will take the form of a series of subsections describing the various approaches conducted, each with a methodology, results, and brief ongoing discussion structure. After which a comparison and general discussion of all the results will be presented, leading into the final conclusions and suggestions of further work. Relevant theory sections regarding techniques pertinent to the works of the thesis are given throughout the thesis where appropriate, though an approximately undergraduate level knowledge of acoustics and signal processing is assumed.

Chapter 2

Literature Review

A broad spectrum of works have already been conducted in the field of head-related transfer functions and their optimal representations, however the ongoing efforts of many authors to develop new methods tells that the question of efficient HRTF representation is still an open one. Past works have approached the problem from various angles but are commonly led by either the aim to compress the HRTF by some means, or alternatively to employ a robust means of HRTF interpolation. This section will attempt to summarise previous works on the topic, beginning with a brief introduction to the concept of the HRTF, then progressing to highlight important commonalities and differences in the methods and works of previous authors in the field that will go on to steadily influence the investigative works described in the latter sections of this thesis.

2.1 Localisation Cues

Spatial audio is the general term for audio that manipulates psychoacoustic cues to give the illusion of virtual sound sources positioned three dimensionally round a listener's head. Spatial audio can be realised through a variety of reproduction systems ranging from two channel systems such as a stereo loudspeaker setup or headphones, to high order ambisonics arrays with many tens of loudspeakers.

The two simplest examples are virtually identical in nature; a stereo loudspeaker pair, and a

pair of headphones, the only difference in terms of signal is the inclusion of a cross coupling network between the two loudspeakers and the listener, whereas in the case of the headphones the two audio channels are presented discretely to each of the listener's ears.

The localisation of a source within a space is a result of acoustic cues generated by the difference between a sounds arrival at each of a listener's two independent ears. A listener's ears are typically spaced between 18cm and 23cm apart, considering this spatial separation it is clear that the sound incident on each ear will differ depending on the ear's proximity to the source and other factors [Howard and Angus, 2009].

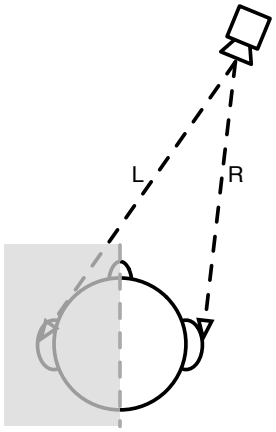


Figure 2.1.1: Spatial cue formation

2.1.1 ITD & ILD

Figure 2.1.1 illustrates the formation of the two vital cues for creating a spatial impression of a source; 'Interaural Time Difference' and 'Interaural Level Difference', hereafter referred to as ITD and ILD respectively for brevity. The two dashed lines represent the average acoustic path from the source to the listener's left and right ears respectively, it is clear that path L is significantly longer than path R as the sound must travel an additional length around the listener's head to reach the left ear. This path length difference gives rise to a phase (or time) difference between the sound incident on the left and right ears. The path length difference also gives rise to a level difference, obviously the ear closest to the sound source will

be subject to a higher acoustic pressure due to the laws of spherical divergence, this effect is compounded by a more dominant effect; the 'acoustic shadow' cast by the hard skull of the listener attenuating the sound incident on the occluded ear [Everest and Pohlman, 2009].

ITDs provide the dominant spatial cue for low frequencies below approximately 1500Hz [Zölzer, 2011]. Above this approximate limit the wavelength of sound is shorter than the spacing of the ears, subsequently the phase differences between the two ears become ambiguous and the ILD becomes the dominant cue.

Considering the case of the sound source positioned laterally at 90°, i.e. at minimum distance to one ear and maximum distance to the other, a rudimentary maximum value of the inter-aural time delay can be calculated as approximately $670\mu s$, assuming a 23cm ear spacing [Woodworth, 1938].

2.1.2 Cone of Confusion

Considering simple geometry it is evident that there exists a cone extending from each ear about the interaural axis, for which a source placed anywhere on its surface will exhibit the same ITD and ILD (due to distance) cues. The so called Cone of Confusion is a well documented psychoacoustic pitfall, and is a common source of front-rear confusions. Figure 2.1.2 illustrates the geometry of the Cone of Confusion about the listener's head.

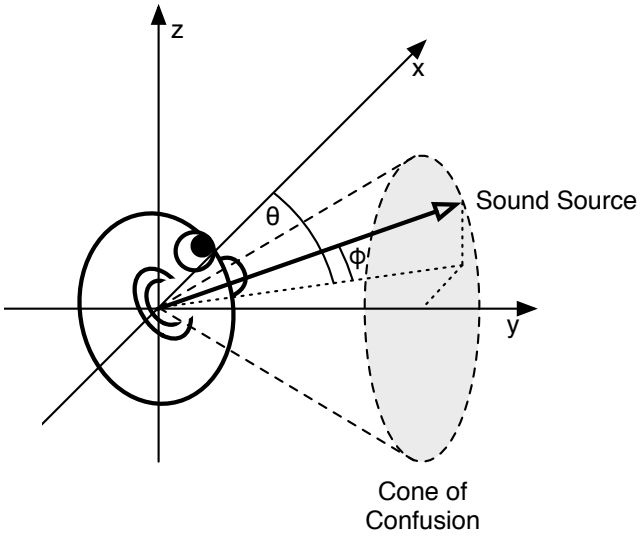


Figure 2.1.2: Cone of Confusion

Ambiguities along the cone of confusion are resolved in two ways [Howard and Angus, 2009]. The first method uses the filtering effects of the ear itself; sound incident on the outer ear is reflected into the ear canal by the grooves and ridges of cartilage that make up the pinna and outer ear structure. These reflections from the pinnae are delayed, if only by a small amount, but the delay is significant to result in comb filtering of the sound incident on the ear drum. The amount of delay varies depending on the angle of arrival of the sound in both azimuth and elevation, with additional filtering effects present in sounds emanating from rear positions due to transmission path through the pinnae. Due to the small order of length of these pinna substructures, the filtering cues occur at high frequencies approximately above 5kHz [Zölzer, 2011]. It is significant to note that this method of ambiguity resolution is unique to each individual listener; the structure of grooves and ridges of the pinnae vary from person to person, and as such, each person is accustomed to the unique 'acoustic fingerprint' of their own pinnae. This individuality can significantly impact the successful externalisation of binaural stereo audio synthesised using non-individualised HRTFs, particularly for angles where front-rear confusions often occur [Begault et al., 2001].

The second method of resolving directional ambiguities is the act of head movement, when a listener hears a sound of interest it is common for said listener to turn their head towards that sound, often attempting even to place the sound directly in front of the head at which the ITD and ILD cues will be equalised. The act of moving the head serves to alter the direction of sound arrival at the ears, this change in direction is dependent on the source position relative to the listener and will therefore serve to resolve the ambiguity. Movement of the head is an important factor to be considered in the attempted externalisation of binaural audio over headphones; if the auditory scene moves with the listener's head then the listener is highly likely to lose the illusion of the audio emanating from elsewhere than the headphones themselves, this is referred to as internalisation. Systems can be designed to compensate the angle used as a criteria for HRTF selection in real-time by tracking the movement of the listener's head by some means.

2.2 Binaural Stereo

Binaural stereo is a spatial audio scheme in which two-channel audio is presented discretely to each of the listener's ears through headphones [Wightman and Kistler, 1989]. Binaural audio can be captured by making recordings with a microphone positioned close to the entrance of each ear canal of a listener or a dummy head, ideally as close as possible. This method of microphone placement attempts to capture the sound incident on each ear separately, thus capturing the all important ILD and ITD cues between the two recorded channels, and ensuring they are preserved in headphone reproduction of the recording.

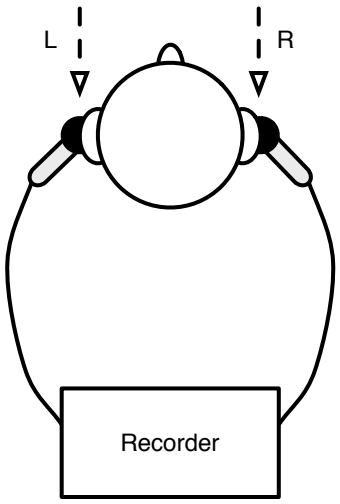


Figure 2.2.1: Binaural recording

Due to the naturally occurring variation in head and pinnae shapes between listener's, individual listeners are accustomed to hearing a specific set of locational cues unique to themselves. A binaural recording made with a specific head, be it real or artificial will achieve varying degrees of success of 3D reproduction across a multitude of listeners [Begault et al., 2001] [Wenzel et al., 1993].

2.3 Head-Related Transfer Functions

The Head Related Transfer Function (HRTF) describes the relationship between the sound emanating from a source in a spatial location and the sound incident at the open end of left or right ear canal (as specified). A pair of HRTFs, one for each ear, can be used to simulate sound emanating from the location described by the two HRTFs in question, as the HRTF encapsulates all of the ITD, ILD, filtering, and shading cues caused by reflections from the head, torso, and pinnae etc.

The HRTF of a listener or dummy head can be measured for any source angle using the binaural stereo recording method; a broadband stimulus such as a Dirac delta pulse yields an impulse response measurement at each ear that encapsulates the HRTF information pertaining to the source direction measured. Other, more practical stimuli such as a broadband

sine sweep may be used, as such the known stimulus signal must be deconvolved from the measured response at each ear to obtain the corresponding impulse response measurements.

HRTFs are often presented as twin sets of discrete responses representing a full, or sometimes limited, sweep of source angles around the head in both azimuth and elevation. Commonly denoted as $H_L(f, \theta, \phi)$ & $H_R(f, \theta, \phi)$ when presented in the frequency domain, where f denotes frequency and θ and ϕ denote angle of azimuth and elevation respectively. The transfer functions are sometimes given in the time domain in the form of a Head Related Impulse Response, denoted as $h_L(t, \theta, \phi)$ & $h_R(t, \theta, \phi)$, where t denotes time. The HRTF is simply the Fourier Transform of the HRIR, and thus the HRIR is the Inverse Fourier Transform of the HRTF.

2.4 Minimum Phase Assumption

Perhaps the best place to begin the analysis of the literature is with the discussion of the minimum phase assumption often adopted in an attempt to simplify the HRTF compression problem.

A system exhibits minimum phase characteristics if both the system and its inverse are causal and stable. In the z-domain this translates to the system having no poles or zeros on or outside the unit circle; poles outside the unit circle imply feedback gain of more than unity, hence the system would become unstable, zeros outside the unit circle, though stable in the original system, translate to unstable poles in the inverse of the system.

The inverse of a system $H(z)$ can be thought of as a corresponding system $H^{-1}(z)$ that exactly rectifies the effect of the original filter, such that:

$$H(z)H^{-1}(z) = 1 \tag{2.4.1}$$

Letting $h_I(k)$ be the impulse response of inverse system $H^{-1}(z)$ in the discrete time domain this corresponds to:

$$h(k) * h_I(k) = \delta(k) = \begin{cases} 0, k \neq 0 \\ 1, k = 0 \end{cases} \tag{2.4.2}$$

First presented by Mehrgardt & Mellert [Mehrgardt and Mellert, 1977], it was found that HRTFs can be approximated to be minimum phase systems. That is, the excess phase component that results from the subtraction of a minimum phase version of an HRTF from it's original phase response has been shown to be approximately linear [Huopaniemi et al., 1999]. This minimum phase assumption implies that the HRTF can be decomposed into two sections [Oppenheim and Schaeffer, 1975]; the first is an angle-dependent frequency-independent delay line or all pass section, the second is the minimum phase filter section.

$$H(e^{jw}) = H_{ap}(e^{jw})H_{min}(e^{jw}) \tag{2.4.3}$$

Where H is the HRTF, and H_{ap} and H_{min} are the associated all pass and minimum phase components of H.

This is somewhat intuitively evident given the typical structure of a HRIR, an example of which is shown in figure 2.4.1; a presumed minimum phase sequence is preceded by an onset delay of nominally zero valued samples.

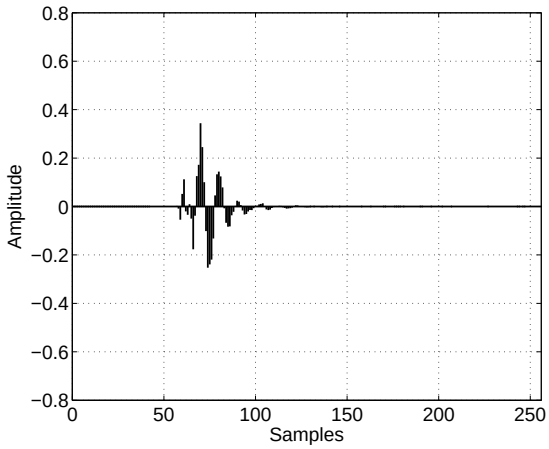


Figure 2.4.1: A typical HRIR

This assumption has been tested both objectively and subjectively, and deemed to have no significant undesired effects by several authors [Kistler and Wightman, 1992] [Kulkarni, 1995] [Kulkarni et al., 1999] [Nam et al., 2008].

The minimum phase assumption has been utilised in a wealth of works as it allows the excess delay component of the HRTFs, corresponding to the ITD, to be removed from consideration. The remaining minimum phase component of the HRTF is particularly convenient to work with as the minimum phase characteristic of the component implies that only the log-magnitude of the filter need be considered as the phase component is unique and obtainable via the Hilbert transform of the log magnitude response [Kulkarni and Colburn, 2004] [Oppenheim and Schaeffer, 1975].

It has also been highlighted that in addition to the reduction of components to compress or model, the minimum phase assumption provides an important time domain characteristic; for a minimum phase impulse response the energy is optimally concentrated in the beginning of the response, i.e. from the initial sample. Not only does this allow for shorter filter lengths, with fewer taps, to achieve the same magnitude response, but also this implies that minimum phase filters are far superior in the implementation of dynamic interpolation [Huopaniemi et al., 1999].

This removal of the need to preserve non-minimum phase information during the attempted transformation or modelling of HRTFs is an attractive property, however not all approaches have utilised this assumption. Chen. et al [1995] for example implement a means of HRTF compression considering the complex output of the Fourier transform of measured HRIRs, and Evans et al. [1998] perform a parallel analysis on both the magnitude and unwrapped phase components.

The works described in the latter sections of this thesis will adopt the minimum phase assumption of the HRTF, concentrating on the compression and efficient representation of the minimum phase component.

2.5 ITD Extraction

The topic of extraction of the interaural time differences from measured data follows closely from that of the minimum phase assumptions, as the ITD must be reintroduced to the modelled or compressed minimum phase component for synthesis. A number of different means of ITD extraction have been contrasted in prior works [Busson et al., 2005], [Lindau, 2010] [Minnaar and Plogsties, 2000].

It is noted by Mills [1958] that the threshold of detection for changes in ITD is approximately $10\mu s$ in optimal conditions. This fact must be taken into consideration as a common sample rate of 44100Hz has an inter-sample time step of approximately $23\mu s$, subsequently it is pertinent in the interest of accurate ITD extraction to first upsample the measured impulse responses or use a peak detection scheme.

2.5.1 Spherical Head Model

Perhaps the simplest method of extraction of the inter-aural time difference for HRIR reconstruction is not extraction, but in fact a model based approach. A simple model for the ITD can be derived from the spherical head assumption [Woodworth, 1938].

$$ITD_\theta = \frac{d}{2c}(\theta + sin\theta) \tag{2.5.1}$$

Where d is the distance between the ears, often assumed to be 18cm, θ is the azimuthal angle, and c is the speed of sound.

This model is reasonably robust due to its physical nature, and gives a good approximation of the ITD in the azimuthal plane [Busson et al., 2005], however it is HRTF measurement independent, and will not provide accurate reproduction of individualised data.

2.5.2 IACC and IACCe Method

Presented by Kistler & Wightman [1992] the Inter-Aural Cross Correlation method models the ITD for a given angle as the time, or lag, for which the maximum value of the cross correlation function of the corresponding left and right ear impulse responses occurs. This approach is based upon the assumption that the auditory system utilises the cross correlation of signals present at the left and right ears in order to retrieve spatial information and localise the sound source [Busson et al., 2005].

Minnaar & Plogsties report that the IACC method consistently overestimated the ITD by as much as $30\mu s$ approaching the inter-aural axis [Minnaar and Plogsties, 2000], suggesting that the technique yields more accurate results if the left and right impulse responses are instead cross correlated with their respective minimum phase components. The ITD is then equal to the difference between the centroids of the left and right cross correlation functions.

Busson et al. suggest that the technique can be improved by instead computing the cross correlation of the signal envelopes of the corresponding left and right impulse responses for any given angle [Busson et al., 2005]. Dubbed the IACCe method, it was shown to perform well in perceptual testing.

It has also been remarked that this method may produce inaccuracies at angles approaching or on the inter-aural axis due to the relatively low signal to noise ratio of the contralateral

impulse response and the possible lack of coherence between the ipsilateral and contralateral impulse responses for these angles [Busson et al., 2005].

2.5.3 Leading Edge Detection Method

Sandvad & Hammershøi propose a method for ITD extraction known as the Leading Edge Detection method, the ITD is calculated as the difference between the times at which each of the left and right impulse responses reaches a threshold value [Sandvad and Hammershoi, 1994]. The threshold value is defined separately for each of the left and right impulse responses as a percentage of the peak value in the left and right impulse responses respectively. This method assumes that the initial portion of the HRIR consists purely of zeros after which the HRTF filter taps begin, i.e the min phase HRTF is preceded by a simple linear phase component.

Busson. et al found that this method successfully predicted the ITD the most closely when compared to psychoacoustic values alongside methods including the IACCe [Busson et al., 2005]. Somewhat conversely, Minnaar. et al remark that the method underestimates the ITD for angles between 90° and 110° [Minnaar and Plogsties, 2000], suggesting instead that it is appropriate when used in conjunction with a phase-based method to determine the inter-aural group delay of the excess phase components.

2.5.4 Phase Methods

Minnaar and Plogsties introduce a method of ITD extraction based upon phase analysis [Minnaar and Plogsties, 2000]. The ITD can be calculated by first evaluating the group delay of the excess phase component of the HRTF for each ear, the ITD is extracted as the inter-aural difference of the left and right group delay at 0Hz.

A number of different techniques have been employed by several authors in order to evaluate the group delay of excess phase component [Busson et al., 2005] [Minnaar and Plogsties, 2000] [Katz and Noisternig, 2014]. These methods are regarded as numerically robust, as they are not impacted by the limitations of the inter-sample time step in the same manner as

the other methods. They can however be made less reliable due to high-pass filtering effects introduced by the frequency response of measurement equipment [Estrella et al., 2010].

2.6 Decompositional Approach

An approach to achieving HRTF compression adopted by several authors is that which is based on the decomposition of the measured dataset into orthogonal subspaces. This can be achieved through the application of techniques commonplace in various disciplines such as the statistical Principal Component Analysis (PCA), the signal processing Karhunen-Loeve Theorem (KLT), or the image processing Singular Value Decomposition (SVD). All three techniques are built upon the efficient decomposition of data into a compressed, more efficient form, achieved through an orthogonal transformation. As such there exist applications in which the three methods are interchangeable, but this is not true for all applications.

The similarities and more importantly, the differences between the PCA, KLT, and SVD, are delineated in detail by Gerbrands [1981], who sought to alleviate the confusion surrounding the choice between the three techniques. Through detailed analysis of the three techniques it is revealed that in the case of a single vector or an n by m matrix in which the m columns are regarded as m realisations of a random stochastic process that the PCA and KLT are in fact identical, apart from a possible shift of the coordinate system origin. If the column covariance matrix of the PCA and KLT is calculated from the m realisations then the identical PCA and KLT are also the same as the SVD, however this similarity only holds true in the application of the techniques to a single matrix $[X]$ of m realisations. In the case of two-dimensional image processing, if the image $[X]$ is considered to be a single realisation of a two-dimensional random process then the covariance matrices for the KLT and PCA techniques will be incorrectly calculated as they should be computed from a number of realisations of that process, i.e. multiple images. It can be concluded that in the case of image processing the correct technique to be used is the deterministically defined SVD. For other applications concerning the realisations of a one dimensional random process the statistically defined PCA and KLT are appropriate.

Principal component analysis is a statistical technique used to reduce the dimensionality of a

multi-dimensional data set [Jolliffe, 2005]. Given a set of observations of possibly correlated variables, PCA transforms the data into a set of values in orthogonal basis referred to as principal components (PCs). The transformation is designed such that the first PC explains the largest amount of variance within the data set, the second PC explains the second largest amount of variance, and so on.

The output of the principal component analysis is the original data transformed into a series of basis vectors and associated weight vectors. The weight vectors describe the contribution of each of the basis vectors required to recreate the original data. Not only are the basis vectors an orthogonal series, but they are also uncorrelated with the weight vectors [Chen et al., 1995]; when considering an HRTF dataset this can be translated to the separation of frequency and angle. The basis vectors describe the principal spectral shapes in decreasing importance, and the weight vectors describe the variation in the basis vector contribution with respect to angle.

PCA attempts to convert a data set into its most efficient form, in which each subsequent component or variable contains only new information, this new information is always accountable for a smaller amount of total variance than that of the preceding component or variable.

The following equations detail the process of conducting a principal component analysis across the log magnitude spectra of an HRTF measurement suite [Kistler and Wightman, 1992]:

Firstly the log magnitude spectra are arranged in a matrix and empirical mean of the data is calculated:

$$u_j = \frac{1}{n}\sum_{i=1}^{n} X_{k,j} \tag{2.6.1}$$

Where u_j is the mean spectrum, $X_{i,j}$ is the matrix of the log magnitude spectra, and i and j are indexes such that $X_{k,j}$ is an n by m matrix where $k = 1, 2, ...n$ and $j = 1, 2, ...m$; n is the total number of spectra and m is the number of frequency bins in each spectrum.

The mean is then subtracted from the original data:

$$D_{k,j} = X_{k,j} - u_j \tag{2.6.2}$$

In the case of an HRTF data set, the subtraction of the mean leaves a set of 'Directional Transfer Functions' or DTFs. DTFs contain only information that is directionally unique, as artefacts common to all directional measurements, such as ear canal resonances, are removed with the subtraction of the 'mean spectrum'.

The next step is the computation of the covariance matrix S, where the covariance of a given pair of frequencies is defined as:

$$S_{i,j} = \frac{1}{n}(\sum D_{k,i} D_{k,j}) \tag{2.6.3}$$

$$\text{for } i, j = 1, 2..., m$$

Where again n is the total number of transfer functions, m is the total number of frequencies, and $D_{k,i}$ is the log magnitude at the ith frequency of the kth DTF.

The basis vectors are the eigenvectors of the covariance matrix S, the lowest 'order' of which correspond to the largest eigenvalues, q.

The weights W_k corresponding to the contribution of each basis vector to a given DTF is given by:

$$W_k = C' d_k \tag{2.6.4}$$

Where C is a matrix, of which the columns are the basis vectors and d_k is the k^{th} DTF magnitude vector.

And hence the DTF magnitude vector is equal to a weighted sum of the basis vectors:

$$d_k = C W_k \tag{2.6.5}$$

Once the analysis has been conducted the original data set can be fully reconstructed through a weighted sum of the total number of basis functions. However, a partial reconstruction of the original data can also be created from the weighted sums of any number of the lowest 'order' principal components, this allows for a compromise between the total amount of original variance explained by the reconstruction, and the greatly reduced expression of the original data set. This was documented by Kistler & Wightman, who found that approximately 90% of the variance of their 5300 HRTF dataset (2 ears of 10 subjects measured at 265 locations) could be expressed with a reconstruction based upon only the first 5 principal components [Kistler and Wightman, 1992]. Similar levels of compression have been achieved when the technique is extrapolated to much larger datasets, such as the CIPIC database of 56250 HRTF pairs (45 subjects measured at 1250 locations), for which Wang et al found that approximately 92% of the variance in the HRTF magnitudes was captured by the first 10 basis functions [Wang et al., 2008].

Chen et al [1995] applied similar techniques; utilising the discrete Karhunen-Loeve expansion to decompose the complex valued Fourier transform of measured HRIRs. The resulting complex valued eigentransfer functions (EFs) are a set of orthogonal frequency dependent functions, by projecting each EF onto the measured data the accompanying weight functions are derived. The weight functions are termed spatial characteristic functions (SCFs) as they are functions of only spatial location. 99.9% of the variance is captured by the first 12 EFs for the measured KEMAR data used in the work, though this is a larger number of basis vectors than was reported by Wightman & Kistler [1992], it is important to note that the technique proposed by Chen et al captures both the magnitude and phase components of the HRTFs.

The PCA (or similar) based decomposition of HRTFs allows for the implementation of a logical interpolation technique based upon manipulation of the angle dependent weight vectors or SCFs, as above. Chen et al. [1995] and Carlile et al. [2000] both propose a continuous functional representation of the HRTF achieved through the process of fitting a continuous piecewise function to the discrete, spatially sampled, weight vectors. Chen et al. utilise

a series of thin plate splines to fit the real and imaginary components of each of the Spatial Characteristic Functions derived from the decomposition of the complex valued Fourier components, whereas Carlile et al. opt to fit a series of spherical thin plate splines to the principal component weights derived from the decomposition of the frequency domain magnitude components of the HRTF dataset used.

Both studies found the interpolation to be reasonably robust; Chen et al. report average percent mean squared errors of less than 1% over most of the frontal and ipsilateral regions, with larger errors occurring in contralateral and lower elevation regions [Chen et al., 1995]. Carlile et al. conclude that by reducing the number of measurement positions retained in a series of models, all of which are constructed from a single measured superset of data, that a high fidelity recreation of a continuous auditory space can be achieved with as few as 150 evenly distributed recorded HRTF positions [Carlile et al., 2000].

The increased error of interpolated data at contralateral positions can be attributed to the large inter-aural level difference due to head shadowing [Chen et al., 1995].

Evans et al. propose a trio of methods akin to the above discussed PCA methods, that are based upon the decomposition of the HRIR and HRTF into the weighted sum of surface spherical harmonics [Evans et al., 1998]. The surface spherical harmonics are a set of basis functions which are orthogonal on the surface of a sphere, and as they are continuous, the derived spherical harmonic representations of the HRTF are also. The method is applied in both the time and frequency domain. Firstly in the time domain on a sample by sample basis of the HRIR. Then again in the time domain on HRIRs with the variable onset delay, representing the ITD, normalised in sacrifice to alleviate undesired reconstruction effects, and finally in the frequency domain on a frequency bin by frequency bin basis of the HRTF magnitude and unwrapped phase. The frequency domain model is reported to be superior after a comparative objective based analysis of results from each of the three variations of the technique; the time domain analysis yielding 'pre-echo' effects in the un-normalised case, likely due to the high amplitude unshadowed measurements included in the analysis that have shorter onset delay than the shadowed measurements.

Evans et al. conclude, similarly to the other authors mentioned in this section, that a large HRTF dataset can successfully be decomposed into a series of basis functions and corresponding weight functions, in this case a parallel pair of the first 17 surface spherical harmonics and their derived weight functions for both the magnitude and phase components [Evans et al., 1998]. However it is noted that although the surface spherical harmonic method proposed yields greater consistency in recreation of measured data, it does not perform as robustly as Chen et al's EF and SCF based analysis when interpolation error is considered.

Comparatively, the surface spherical harmonic based decomposition does not offer as great an efficiency as other techniques mentioned in this section, however it is most appropriately compared to the approach proposed by Chen et al. [1995] as neither of these methods rely on the minimum phase assumption, opting instead to directly encode the phase in the prescribed methods. It is remarked by Evans et al. that for applications concerned with storage efficiency, the Karhunen-Loeve expansion based method proposed by Chen et al. may be considered more appropriate [Evans et al., 1998].

Like the surface spherical harmonic led approach proposed by Evans et al. [1998], Zhang et al. propose a decomposition approach not based on optimality such as the PCA and KLE approaches, but by instead using non-measurement specific functions with mathematical definition as the bases [Zhang, 2009]. A continuous two dimensional model of the azimuthal plane is constructed,a Fourier-Bessel series is used to reproduce the spectral variation in measured data, and a Fourier series is used in tandem to reproduce the corresponding spatial variation. Empirical data is used to guide the choice of orthonormal function as the basis function for the spectral variation, however even so, the basis functions are independent of the empirical data, and all subject or measurement dependent differences are encoded in the model coefficients. In validating the model, Zhang et al. directly compare their modelling technique to those previously conducted using the statistical based PCA [Kistler and Wightman, 1992] and KLE [Chen et al., 1995] methods, by re-implementing them on the same 2D dataset.

Although providing the least (approximate) error of reconstruction for the magnitude spectra, the PCA approach is dismissed as it does not attempt to encode or predict the phase

components of the HRTFs [Zhang, 2009]. The KLE method is reported to give marginally superior error in the reconstruction of both the magnitude and phase data for a little over 300 less model parameters than the total of 4900 used in the proposed Fourier-Bessel / Fourier model. It is concluded that the proposed model's measurement data independence and continuous nature, without the need for interpolation (spline fitting), is advantageous over the aforementioned KLE techniques.

The potential efficiency of the PCA based approach is furthered by Wang et al., who propose that following a PCA decomposition of the CIPIC HRTF database, the principal component weights may be expressed further more efficiently using a vector codebook technique [Wang et al., 2008]. The codebook technique achieves compression via a technique known as vector quantisation, in which a given vector is approximated by the nearest matching vector in a designed vector codebook, allowing the input vector to be recorded as a single value representing the closest matching codebook index. Wang et al. report that the error introduced by the quantisation can be considered negligible; 7.23% compared to the 6.71% average error already present in the unquantised PCA reconstruction.

2.7 FIR and IIR Modelling

Another arm of approach to the problem of HRTF compression comes not from a basis of decomposition, but rather the consideration of the HRTF as an implementable digital filter.

The simplest of such implementations is to utilise the HRIR itself; the samples of a given HRIR, or any IR for that matter, represent the taps or coefficient weights of a finite impulse response (FIR) digital filter [Zölzer, 2011]. A rudimentary form of HRTF compression can be realised by truncating, or windowing, the measured HRIRs to reduce the number of filter taps, usually referred to as the order of the filter, used to represent each component of the HRTF dataset.

Sandvad & Hammershøi found through experimentation that for the purpose of HRIR truncation, there was no sufficient justification to use any window type other than rectangular

[Sandvad and Hammershoi, 1994], though rectangular windowing can result in frequency domain oscillation or ripple known as the Gibbs phenomenon, the nature of the HRIR filters, more specifically their lack of frequency domain discontinuities, allows for simple rectangular windowing to be used with negligible undesired effects being introduced to the frequency domain response of the filters. The use of alternative window designs, such as the Hamming window, typically selected as an alternative to rectangular windowing in an attempt to negate the influence of the Gibbs phenomenon, significantly smooths the frequency domain response of the filter, which could possibly translate to the loss of pertinent spectral characteristics contained in the 'detail' of the response.

Several authors have conducted works which suggest that the fine detail lost in HRTF smoothing or HRIR truncation is perceptually unimportant. Senova et al. [2002] found that the psychoacoustic performance of truncated HRIRs only began to perform poorer than free field loudspeaker signals for IR lengths of between 0.32 and 5.12ms. Through the use of a gammatone filterbank designed to mimic the spectral filtering of the human cochlea, Breebaart and Kohlrausch [2001] show that HRTF phase and magnitude spectra do not need a higher spectral resolution than that of the filterbank of the peripheral auditory system. More specifically they show that a first order gammatone filterbank with bandwidths of one equivalent rectangular band sufficiently describes the phase and magnitude spectra. In particular the high frequency content of the HRTF has been shown to be of little importance for both the ipsilateral and contralateral ears, with the least detriment to psychoacoustic perception occurring for the contralateral Xie and Zhang [2010]. As such it can be considered that the truncation of measured HRIRs may provide a simple means of reducing the number of stored elements in an HRTF/HRIR dataset without significantly altering the psychoacoustic perception of the data. Furthermore, simple HRIR truncation could be used in conjunction with further methods of compression to improve the overall efficiency of the system.

More advanced approaches consider the modelling of HRTFs or HRIRs as alternative filter types, a common starting point of which is the infinite impulse response filter (IIR). IIR filters offer numerous advantages over their FIR counterparts, the most useful of which is their efficiency in approximating, or even matching, filter designs that would require com-

paratively high order FIR implementations, in far fewer coefficients. This is due to the IIR filter's feedback coefficients in the denominator of the transfer function, which can create a more pronounced response with superior efficiency to the FIR in terms of processing power required for implementation.

The IIR approximation of a given system, such as the HRTF for a given direction, can be derived by modelling the time domain system output, the HRIR, as the output of an auto-regressive moving average (ARMA) system [Farhang-Boroujeny, 1999]. For an ARMA system the output sample $y(k)$ is defined as the weighted sum of all previous input and output samples, this can be expressed as:

$$y(k) = -\sum_{i=1}^{n} a_i y(k-i) + \sum_{i=0}^{m} b_i x(k-i) \tag{2.7.1}$$

and yields the transfer function:

$$H(z) = \frac{\sum_{i=0}^{m} b_i z^{-i}}{\sum_{i=0}^{n} a_i z^{-i}} = \frac{B(z)}{A(z)} \tag{2.7.2}$$

Where x is the record of input samples, m and n are the orders of the numerator and denominator respectively, and a_i and b_i are the coefficients or tap weights.

The poles of the model, the locations of which are described by the denominator of the transfer function; the a coefficients, make up the auto-regressive component of the system, and in the case of the HRTF, translate to the acoustic resonances in the sound path between the source and the ear. The zeros of the model, the locations of which are described by the numerator of the transfer function; the b coefficients, make up the moving-average component of the system, and in the case of the HRTF, translate to the anti-resonances and reflections in the sound path between the source and the ear [Asano et al., 1990].

The order of the numerator and denominator (m and n) dictate the number of poles and zeros in the system model, i.e. the number of a and b coefficients. The coefficients of the system model are determined such that they minimise the quadratic expression of the error

between the model and the system to be modelled, first proposed by Kalman [Kalman, 1958]:

$$E^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(w)A(w) - B(w)|^2 dw \tag{2.7.3}$$

Where $A(w) = A(z)|_{z=e^{jw}}$ and $B(w) = B(z)|_{z=e^{jw}}$.

Kulkarni & Colburn detail two low order model approximations made using IIR filters; an all-pole model and a general pole-zero model [Kulkarni and Colburn, 2004]. The all pole case is derived as an implementation of the autocorrelation method for linear prediction as described by Makhoul [1975], whereas the pole-zero case is derived from a weighted-least-squares variation of the modified least-squares problem proposed by Kalman [1958]. The pole-zero derivation uses a technique known as the Steiglitz-McBride iteration [Steiglitz and McBride, 1965] to minimise the quadratic error function presented by Kalman, and obtain the optimum coefficients of the IIR filter model. In order to simplify the modelling process the mean spectrum was computed and subtracted from all HRTFs, thus leaving the DTFs, as described in the initial steps of the PCA procedure implemented by Wightman & Kistler [1992]. The removal of the mean, or common, spectral components of the data is believed to likely reduce the order of the derived IIR models needed to sufficiently recreate the measured dataset, as only the spatially dependent variances are encoded in such an approach. Small scale subjective evaluation across three subjects found that a model with just 6 poles and 6 zeros, was largely indistinguishable from original measurements, the 25 pole all-pole model performed similarly well but it does not provide as much representational efficiency as the pole-zero formulation.

Prior to the work of Kulkarni & Colburn, Asano et al. adopted a similar pole-zero model restricted to fit HRTFs in the horizontal plane only [Asano et al., 1990]. Using the ARMA model to define the general form of the transfer function and the quadratic form of error minimisation as proposed by Kalman shown in equation 2.7.3 to determine the pole and zero locations to be used. Asano et al. solve the minimisation using both the measured impulse response and its covariance sequence in a method described by Mullis & Roberts [1976].

Kulkarni & Colburn provide an insightful comparison into the differences in the approaches, and subsequently results, of the two studies. Asano et al. report that comparatively high model orders (40 poles and 40 zeros) were required to approach the same psychophysical performance as was found with the measured HRTFs in an absolute localisation task; clashing with the finding of Kulkarni & Colburn, that as low an order as 6 poles and 6 zeros is almost indistinguishable from the empirically obtained data. The difference in findings is explained as a combination of two likely sources of inconstancy between the studies. Firstly, the model proposed by Asano et al. is based on the Kalman estimate algorithm alone, whereas the method proposed by Kulkarni & Colburn utilises an iterative weighting procedure ([Steiglitz and McBride, 1965]) to achieve an optimal fit between the modelled and measured HRTF log magnitude functions. Secondly, Asano et al. modelled the HRTFs directly, whereas Kulkarni & Colburn modelled the DTFs calculated as the HRTFs of a dataset less the empirical mean. It is likely that the modelling of the more complex HRTFs that include not only the spatially dependent characteristics but the common-to-all spatial independent characteristics as well, means that the efficiency of the modelling process is reduced due to the need to fit poles and zeros to these addition spectral characteristics [Kulkarni and Colburn, 2004].

Ramos & Cobos present a parametric model of the HRTF based on a low order IIR implementation achieved as a chain of second order sections of conventional shelving and peak audio filters [Ramos and Cobos, 2013]. The minimum phase component of the HRTF is modelled via an iterative process for which the central frequency w_i, log-gain G_i, and the quality factor Q_i of the shelving ad then successive peak filters is defined in order. A random optimisation is used to vary the parameters of each section until the error of each designated error frequency zone is minimised, upon completion a global post optimisation process is performed in which the second order sections are ordered in decreasing frequency order and the random optimisation is performed again, however this time it is performed for groups of 3 adjacent second order sections simultaneously. It was found through both objective analysis and subjective evaluation, using the MUSHRA recommendation, that as little as 6 second order sections could be used to recreate a given HRTF with reasonable accuracy; outperforming alternative Yule-Walker and Prony methods for frequencies below 3kHz and performing slightly worse at frequencies above 10kHz. Ramos & Cobos conclude that not only does the proposed method allow for a reduction in HRTF database size, through the

transformation of HRIR samples to 3-parameter filter section parameters, but that the filter parameters also represent a convenient means of performing a nearest neighbour interpolation scheme.

2.8 Interpolation Led Approaches

An alternative means of improving the efficiency of HRTF storage, or in particular the reduction in measurement redundancy, such that less measurements need be performed, is to deduce a means of interpolating a dense measurement scheme from a comparatively sparse one.

Nishimura et al. propose an interpolation algorithm based upon spatial linear prediction [Nishimura et al., 2009]. The method requires that a single set of measurements are made with a high spatial resolution, this resolution defines the interpolation points for additional datasets. The high resolution data is used to calculate the optimum filter coefficients w, that satisfy the system of simultaneous equations associated with the theory of linear prediction, against the complex Fourier coefficients of the HRTF set. The coefficients obtained can be used to interpolate HRTF sets measured at a lower spatial resolution as far as the limit imposed by the spatial resolution of the dataset used to calculate the coefficients. The method attempts to exploit the observed periodicity of the measured HRTFs in the azimuthal plane, Nishimura et al. report a significant reduction in interpolation error compared to simple linear time domain alternative methods, as well as an increased rate of correct judgement of the rotation of a virtual sound source in unofficial listening tests. It is concluded that the interpolation method might be expanded by interpolating the coefficient set, which would allow interpolation of coarse datasets up to higher spatial resolutions than the current limit of the resolution of the dataset used to derive said coefficients.

The test HRTFs were obtained for all directions (360°) by applying one of four methods to a subsampled dataset with angular resolution of 15° [Nishimura et al., 2009]. The four interpolation methods comprised: the proposed method with a tap length of 2, the proposed method with a tap length of 6, a simple linear interpolation in the time domain, and the

same time domain interpolation with a correction to equalise impulse response arrival time, as was suggested as the outcome of investigation by Matsumoto & Yamanaka [2004]. The aforementioned investigation [Matsumoto and Yamanaka, 2004] considered the differences of accuracy of three interpolation methods, with and without arrival time correction. The arrival times of the interpolated responses were calculated linearly from the difference between the cross correlation of the left and right ear impulse responses for the two adjacent measured angles. The methods considered were simple linear interpolation, spline interpolation, in which a piecewise mathematical polynomial is fitted to the data to achieve a continuous functional representation, and a method based on the Discrete Fourier Transform (DFT). The DFT method consists of arranging all HRIRs as the columns of a matrix and taking the DFT of each row, the output of each DFT is then transformed by adding zeros to the center of the array, finally the inverse DFT is taken of each of the transformed arrays resulting in a larger matrix of spatially oversampled HRIRs. Arrival time correction coupled with the simple linear interpolation case was shown to yield the greatest interpolation accuracy of the six cases for most angles, and arrival time correction improved the accuracy of all three methods, the results were assessed by comparing the signal to deviation ratio of the interpolated and measured HRIRs.

In an attempt to better consider the interpolation and the often neglected range effects for close sources, Duraiswami et al. approached the HRTF through a scattering analysis [Duraiswaini, 2004]. In considering that the sound field captured by the HRTF arises from the scattering of sound from a source caused by the torso, head, and pinnae of a listener, it can be shown that the HRTF is expressible in terms of a series of multipole solutions to the Helmholtz equation. Under the principle of reciprocity, which states that source and receiver are interchangeable in a complex audio scene in terms of observed signal, the ears are considered to be sources and as such the multi-path sound measured at the ear microphone from the speaker can be assumed to the be the multi-path sound at the speaker location, assuming the idealised point source speaker were in the ear. Measured points extracted from the HRTF dataset are used to solve a system of linear equations and define a set of coefficients, after which, the acoustic field of the virtual auditory scene can be evaluated at any desired location outside of the sphere encapsulating the sound sources in the scene.

Due to the physical nature of the method in question [Duraiswaini, 2004], in particular the modelled source-encapsulating sphere, the scattering analysis method yields impressive results when compared to the analytical model of a spherical head as used by Duda & Martens [Duda and Martens, 1998], the objective analysis of the scattering analysis model also fair well when fitted to a set of KEMAR data. The evolution of the predicted potential field at varying distances offers logical physical arguments including the growth of the HRTF magnitude in the direction corresponding to the direct path, and the enlargement of the shadowed magnitude region as the source approaches the head. It is noted that although highly promising, the reported results are based on objective analysis alone, and performance of the method should be investigated perceptually.

Some approaches to HRTF interpolation have been designed to take large numbers of, or even all of, the measurement points within a dataset; although such approaches have yielded increased interpolation accuracy over methods that consider only a small subset of measurement positions, they demand comparatively high computational expenses. This increased computational expense can become troublesome when attempting to render a complex auditory scene containing multiple sources, even more so if the sources are to be rendered as moving through the virtual auditory space.

A relatively straightforward approach to HRIR interpolation is to apply the bilinear interpolation method; considering a subset of the four closest points for which measured data is available. Given an HRIR dataset containing information for all points on the measurement grid defined by the fixed angular intervals θ_{grid} for azimuth and ϕ_{grid} for elevation. An interpolated HRIR for a desired direction (θ, ϕ) can be evaluated as:

$$h_i(k) = (1 - c_\theta)(1 - c_\phi)h_a(k) + c_\theta(1 - c_\phi)h_b(k) + c_\theta c_\phi h_c(k) + (1 - c_\theta)c_\phi h_d(k) \qquad (2.8.1)$$

where $h_a(k)$,$h_b(k)$,$h_c(k)$, and $h_d(k)$ are the HRIRs of the four adjacent measurement points, and c_θ and c_ϕ are the normalised relative angular positions, calculated as:

$$c_\theta = \frac{\theta \ mod \ \theta_{grid}}{\theta_{grid}} \qquad c_\phi = \frac{\phi \ mod \ \phi_{grid}}{\phi_{grid}} \qquad (2.8.2)$$

Bilinear interpolation of four adjacent measurement points has been shown to exhibit smoother variation with change in angle than similar methods, such as bilinear interpolation of three adjacent measurement points [Gamper, 2013]. This property yields smoother interpolation, without discontinuities, and is an attractive facet in the rendering of moving sources.

Gamper proposes an alternative method of subset selection and subsequent interpolation based on the vectorisation of the HRTF measurement grid [Gamper, 2013]. Assuming a source distance of at least 1m for all measured points and desired points (to be interpolated), the distance effects of the HRTF can be assumed to be negligible, thus the directions to all measurement points and desired points can be described as unit vectors. Gamper's proposed method is based on the assumption that an HRTF estimate for the desired source direction s can be constructed as a linear combination of three measurement directions h_1, h_2, h_3 that form an enclosing convexly curved triangle around s on the unit sphere. In the interests of both speed and computational efficiency a triangulation of the unit sphere is performed as the algorithm is initialised; during which the surface of the unit sphere is mapped as a series of non overlapping triangles constructed from triplets of measurement points, the results of which are stored. A Delaunay triangulation is used to maximise the minimum angle of all the angles of the triangles in the triangulation, this has been shown to be advantageous for interpolation by other authors. To further increase runtime efficiency, the inverse of each measurement point triplet, required to calculate the contribution gain of each measurement vector to obtain the desired direction, is calculated and stored during the same initialisation process.

Gamper draws comparison between the proposed vector based amplitude panning method of interpolation weight calculation, with an inverse distance weighting, and a bilinear interpolation of three measurement points. It is shown that the interpolation algorithm performs comparably to, if not better than the other methods used in the objective tests. In particular, the proposed method yields smooth variation of interpolation weights with changes in both azimuth and elevation, allowing for the convenient rendering of moving sources without the negative effects imposed by interpolation discontinuities [Gamper, 2013].

2.9 Conclusion

It is evident from the literature discussed in this section that given an interest in both efficient representation and interpolation of the HRTF dataset then an interpolation led method is largely unsuitable, though an interpolation based approach can be thought of as offering a means of compression through the reduced number of measurements that need be stored this level of compression is minimal in comparison to that which is achievable via a decompositional or filter modelling approach.

Both the decompositional and filter modelling approaches have apparent strengths and some similar weaknesses, such as the recurring increased error at contralateral positions, however no one method seems to be identifiable as optimal in so far as to offer maximal compression for minimal loss in reconstruction/model error. As such the experimental works described in this thesis will approach the initial compression problem from both a decompositional and parametric filter modelling standing. In order to take advantage of possible underlying cyclic features of the HRTF and also to simplify the problem to a more appropriate project length, the methods will be investigated using an HRTF dataset limited to the azimuthal plane only.

Restricting the analysis to the horizontal plane somewhat limits the exploration of spectral compression methods, as many of the psychoacoustically significant spectral variations are an effect of the influence of the asymmetrical pinnae on changes in source elevation. However this limitation should not affect the objective analysis of the compression methods or interpolation scheme presented in the remainder of this work. Furthermore, human ability to detect source direction is known to be most acute in the horizontal plane as such it can be considered to be of principal interest, particularly in the context of interpolation.

The investigative works will begin with a decompositional analysis and partial reconstruction of the measured magnitude spectra in both the linear and logarithmic domain, followed by a parametric modelling approach broken into two halves; an all-pole filter approximation made using linear predictive coding techniques, and a pole-zero filter approximation using an implementation of the Steiglitz-McBride iteration.

Further to the compression-centric methods, a means of secondary compression and possible convenient interpolation of the angle dependent decompositional or parametric model components, in the spirit of Wang et al. [Wang et al., 2008], will be investigated.

Chapter 3

Preparation for Experimental Works

3.1 TU Berlin HRTF Analysis

This section of the report gives a brief analysis of the dataset used as the basis of the investigative works described in the following sections. The aim of this section is to highlight the key features of the HRTF and HRIR in the azimuthal plane, and familiarise the reader with the measured dataset.

For this project, the HRIR measurement dataset chosen has been provided by TU Berlin [Wierstorf et al., 2011], the set is made freely available under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license.

The impulse responses were measured in an anechoic chamber, with lower frequency limit 63Hz, using a KEMAR mannequin at a range of four different loudspeaker distances; 0.5m, 1m, 2m, and 3m. The excitation signal was a 5.3s linear sine sweep with a 6dB per octave low-shelf emphasis below 1kHz, resulting in an amplification of 20dB for low frequencies.

The loudspeaker was positioned at ear-level, and a high precision stepper motor was used to rotate the mannequin in order to obtain measurements in increments of one degree in the azimuthal plane.

The dataset has been compensated for inaccuracies in the measurement procedure; firstly

the minimum ITDs were determined between measurement angles $-90°$ and $90°$, the dataset was then rotated by $2° - 3°$ to align the minimum to $0°$, secondly the ILD between the two ears was corrected by adjusting the gain of the left and right HRIRs to achieve an ILD of 0dB at $0°$. The loudspeaker transfer function was also compensated between 100Hz and 10kHz by the design and application of inverse FIR filters.

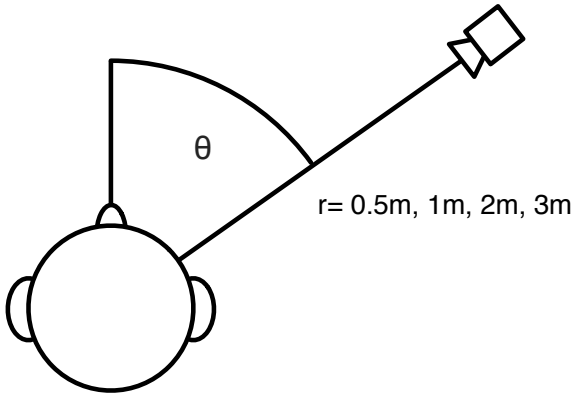
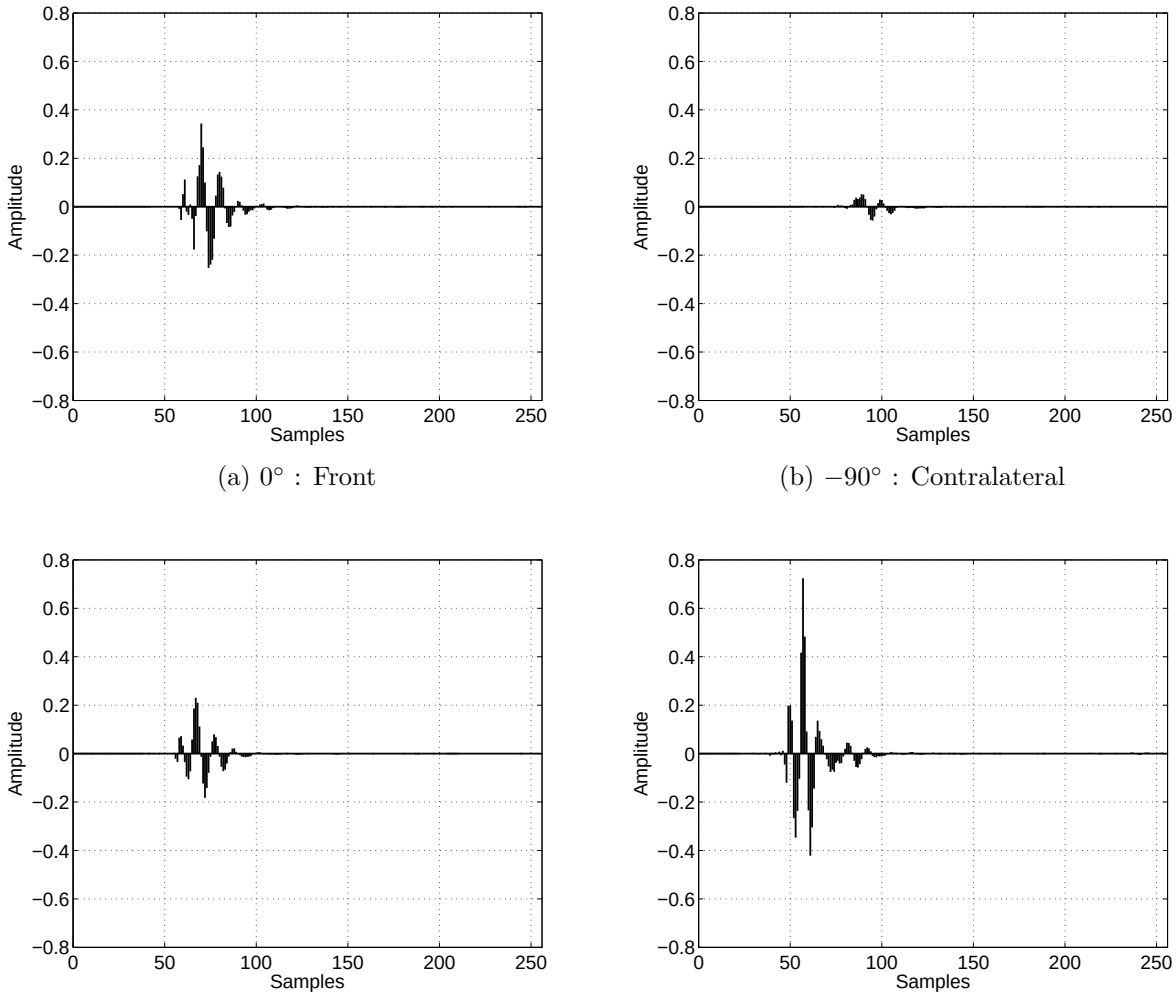


Figure 3.1.1: TU Berlin measurement scheme

The co-ordinate nomenclature used in the TU Berlin dataset and adopted in this project, is defined such that the azimuth angle θ describes the placement of the loudspeaker or source, anti-clockwise around the head. Thus the angle denoted in Figure 3.1.1 is $-60°$.



(a) 0° : Front

(b) −90° : Contralateral

(c) ±180° : Rear

(d) 90° : Ipsilateral

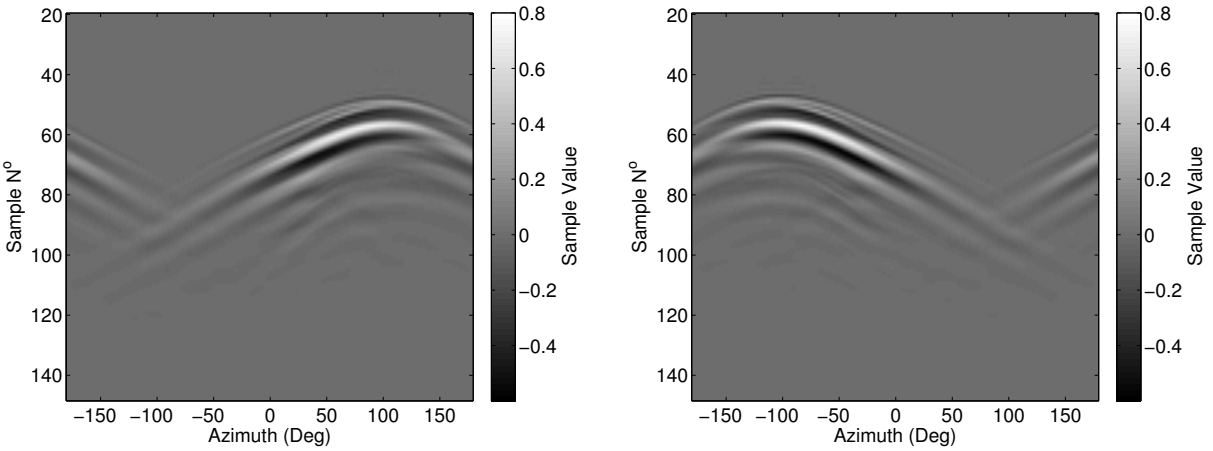
Figure 3.1.2: Measured HRIRs for cardinal directions with respect to left ear

Figure 3.1.2 shows the first 256 samples of the measured head related impulse responses corresponding to the four cardinal directions, the front, contralateral, ipsilateral, and rear positions as measured at the left ear. Several important characteristics of the HRIR in the azimuthal plane can be approximated simply by comparing these four impulse response plots. Comparing subfigures 3.1.2a and 3.1.2c, illustrates the slight level difference between the front and rear measurements that occurs due to the shading and filtering effects of the pinnae, without introducing a significant arrival time difference between the two as the two

measurement positions are approximately equidistant from the either ear. Comparing subfigures 3.1.2b and 3.1.2d clearly conveys the difference in sound path between the contralateral and ipsilateral ear respectively. The ipsilateral position clearly yields maximal excitation out of the four cardinal directions, as well as the shortest onset/arrival time, conversely the contralateral position is characterised by minimal excitation due to the acoustic shading of the head positioned directly between the speaker and, in this case occluded, ear, and the longest onset/arrival time of all four pictured responses.

All four of the HRIRs presented in figure 3.1.2 share a common onset delay before the variable onset delay due to measurement position. The common onset delay occurs due to the distance between the loudspeaker and the dummy head used to measure the data; the data was measured with a constant distance of 1m between the centre of the loudspeaker and the centre of the mannequin. In post-processing the length of the onset delay of all 1m measurements was trimmed to that equivalent to a measurement distance of 0.5m between the centre of the loudspeaker and the centre of the mannequin [Wierstorf et al., 2011]. At the sample rate of 44100Hz a distance of 0.5m should correspond to a common onset delay of approximately 64 samples, however due to the distance from the centre of the mannequin to the microphone transducers at the ear canals, this value is slightly too large. A better estimate is to address the case of the ipsilateral measurement position; assuming an ear spacing of 0.18m, true of the KEMAR design, the expected onset delay due to measurement distance of (0.5-0.09) 0.41m is approximately 53 samples. This seems congruent with the measured impulse response of the ipsilateral position in subfigure 3.1.2d, which seemingly exhibits an onset delay of similar order to the approximated value.

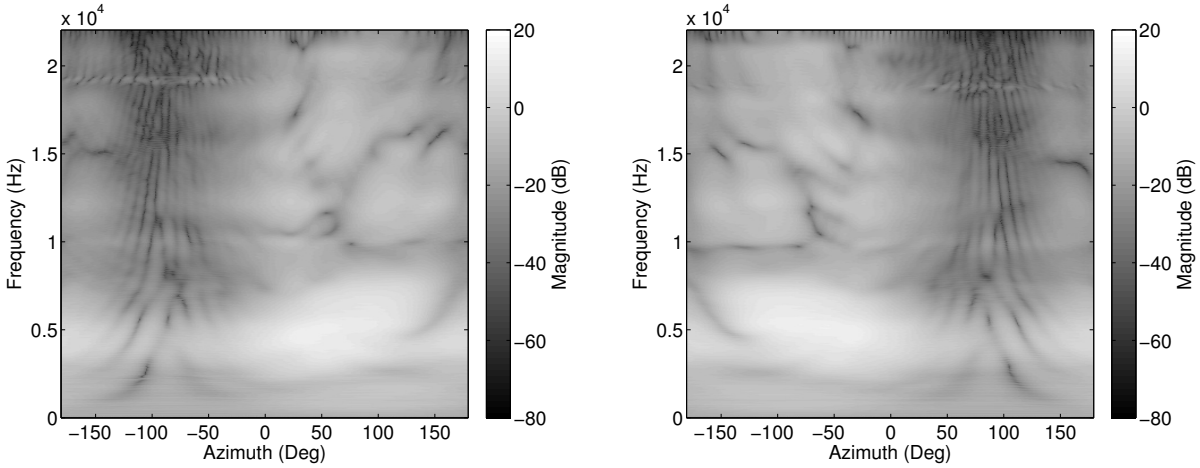
The overall variation in the amplitude and onset/arrival time of the measured HRIRs for each ear with respect to source angle is better depicted in figure 3.1.3.



(a) Left Ear (b) Right Ear

Figure 3.1.3: Measured HRIRs

Figure 3.1.3 shows a portion of the measured head related impulse responses for all measured angles at each ear. For illustrative purposes only 128 samples of the impulse responses have been plotted, starting at sample number 20 and ending with sample number 148. Both the inter-aural time and level differences can be seen clearly across the angular variation; the highest contrast areas represent maximal excitation in the measured sound field. It is evident that the largest amplitudes occur in the ipsilateral regions for each ear, occurring around $+90°$ for the left ear in figure 3.1.3a and around $-90°$ for the right ear in figure 3.1.3b. The image plots also show the variation in onset delay of the HRIRs with respect to the variation in source angle, the uppermost region of excitation on each of the two plots, corresponding to the shortest onset delay, occurs as expected at the ipsilateral source position of $+90°$ and $-90°$ for subfigures 3.1.3a and 3.1.3b, reaching a corresponding minimum at $-90°$ and $+90°$ respectively.



(a) Left Ear (b) Right Ear

Figure 3.1.4: Measured HRTFs

Figure 3.1.4 shows the magnitude spectra of the measured head related transfer functions for each ear obtained as $20log_{10}$ of the 2048 point Fourier transform of the measured HRIRs. The lightest regions between the axes of angle and frequency represent the larger magnitudes of the measured responses, the largest magnitudes occur around the ipsilateral source positions between approximately 2kHz and 8kHz for each ear. The darkest regions depict the areas of lowest magnitude, considering the contralateral angle positions of each ear and moving South to North along the frequency axis shows the effects of head shading and how such shading varies with frequency. Lower frequencies at the contralateral source positions exhibit low order diffraction patterns, characterised by the semi-periodic minima that occur as a result of the superposition of the sound waves that split to travel around either side of the head, incurring a significant path difference that results in cancellation of the acoustic pressure. At higher frequencies the magnitude at these contralateral positions is significantly lower due to both increased air absorption, and the increased directivity of high frequency sound (less diffraction).

As the basis of the conducted works is concerned with the efficient representation or compression of the HRTF, it is desirable to be able to increase the efficiency of representation before the application of more advanced techniques. As is suggested by the representations

of the HRIR measurement set in figures 3.1.2 and 3.1.3, a fundamental means of compression may be achieved through the truncation of the measured impulse responses to include the important region of excitation, including the variable onset delay. The measured impulse responses are of length 2048 samples, however upon simple visual inspection it seems that the latter portion of the responses contains little to no aptitude variation. To ensure that no significant information is lost in truncation, the energy in the impulse response should be considered as a basis for determining the truncation value.

The energy decay curve was defined by Schroeder [1965] as a means of measuring and defining the reverberation time of a space using the impulse response. The energy decay curve (EDC) is defined as the reverse integral of the squared impulse response at time t and describes the total signal energy remaining in the impulse response at that time:

$$EDC(t) \equiv \int_t^\infty h^2(t)dt \tag{3.1.1}$$

In order to select an appropriate truncation length, the EDC of the impulse response corresponding to the contralateral source position measured at the left ear is plotted. Logically the contralateral impulse response should have the slowest decay of energy, partially due to its minimal total energy, and partially due to the almost exclusive presence of lower frequency diffracted frequencies at the occluded ear.

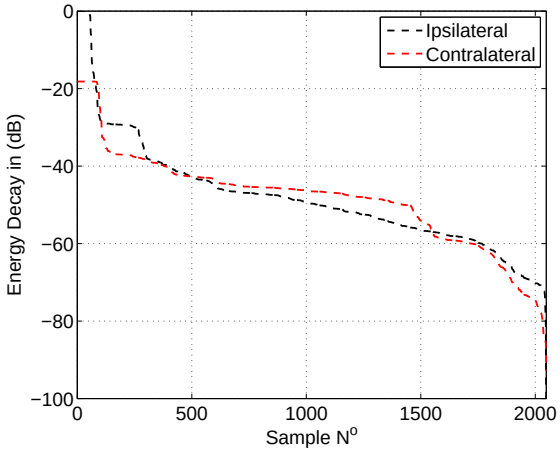


Figure 3.1.5: EDCs of ipsilateral and contralateral positions

Figure 3.1.5 shows the energy decay curve of the ipsilateral and contralateral impulse responses measured at the left ear, for illustrative purposes both curves have been normalised to the total energy of the ipsilateral HRIR. Considering the values of each curve at the first sample shows that the contralateral HRIR has approximately 20dB less total energy than the ipsilateral, this is of course due to the previously mentioned effects of head shadowing. Both curves converge around -40dB approximately 500 samples into the responses, after which they exhibit a similar decay over the remaining IR length. In the latter region the curves do diverge up to a maximum of ∼5dB, the similarity between the curves in this region suggests that it is likely that this region is dominated by noise in the measured HRIRs. The relatively gentle decay is an artefact of the reverse Schroeder integration technique used to calculate the curves, the HRIRs are reversed and then summed from beginning to end, as such the noise floor in the measurements would yield a steady yet likely shallow gradient in the curve as the cumulative energy in the noise with respect to time increases. The ∼5dB deviation between the curves in the latter 1500 sample region is most likely due to the lower signal to noise ratio of the contralateral HRIR; as less of the total energy in the signal is due to the sound emanating from the source on the other side of the head, more of the total energy in the signal is due to the noise, subsequently the relative contribution of the noise dominated region of the EDC will be greater than that of the ipsilateral counterpart.

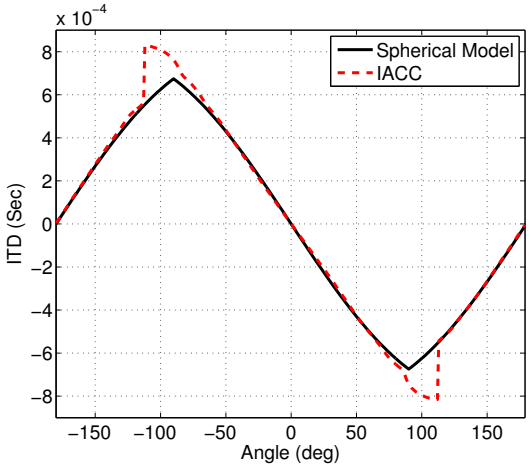
The fact that the energy decay curves shown in figure 3.1.5 both seem to exhibit a noise dominated response after approximately 500 samples suggests that the HRIRs can be truncated to a 512 sample length with no significant loss of any pertinent binaural cues. Taking the 512 and 2048 point Fourier transforms of the truncated and full length HRIRs respectively was found to yield no discernible truncation effects such as the Gibb's phenomenon.

3.2 ITD Extraction

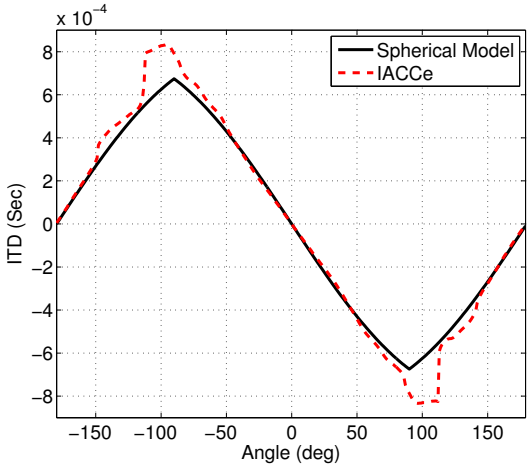
For the purposes of the works described in this thesis: the application and possible extension of discussed decompositional and parametric modelling techniques to the TU Berlin HRIR dataset, the measured responses will be assumed to comply with the aforementioned minimum phase assumption. As such the remainder of this thesis will deal mainly with the compression or efficient representation of the magnitude components of the HRTF given

that the key ITD information can be synthesised as required simply by zero padding of the processed minimum phase impulse responses.

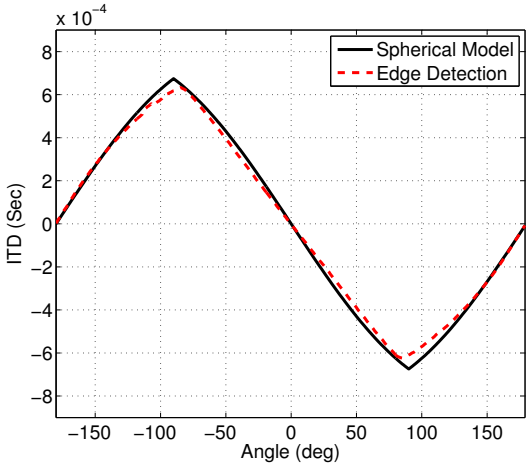
This section of the thesis presents a brief comparison of a number of previously identified techniques for the extraction of the ITD applied to the TU Berlin dataset, and leads to the selection and justification of the method that is used in later reconstructions.



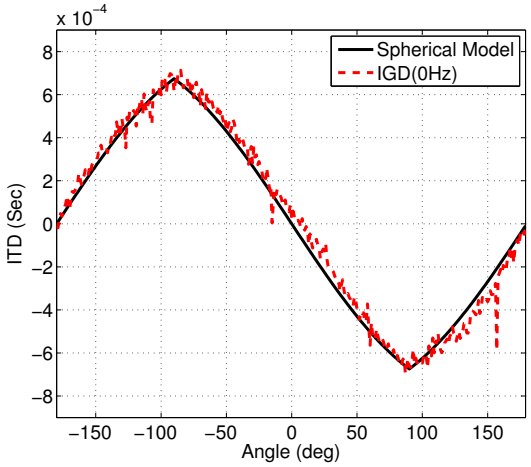
(a) IACC



(b) IACCe



(c) Edge Detection



(d) IGD_{0Hz}

Figure 3.2.1: Comparison of ITD extraction methods

Figure 3.2.1 shows the ITD with respect to angle extracted from the measured Tu Berlin data using each of the four methods. The resulting ITD curve for each method has been plotted against the reference curve of the spherical head model ITD prediction, the spherical head model provides a useful measurement independent reference for the Tu Berlin data as it is known to give a good estimation of ITD in the azimuthal plane, but also because the Tu Berlin data was measured using the KEMAR mannequin and as such should fit well with the approximation of a spherical head.

For the first three methods; the IACC, IACCe, and Edge Detection, the measured impulse responses are first up sampled by a factor of 20, in order to alleviate errors in the ITD time calculation in seconds that is imposed by the measurement sample rate of 44100Hz.
The IACC derived ITD curve shown in figure 3.2.1a was calculated by first computing the cross-correlation function between the left and right ear measurements for each angle using the *xcorr.m* function in MATLAB, and then finding the lag value, in samples, for which the maximum of each function occurs. These sample values are then divided by the sample rate of the original measurement and the up sampling factor to obtain ITD values in terms of seconds.

The ITD curve derived using the IACCe method as shown in 3.2.1b is computed using the same method as the IACC, the only difference is that the cross-correlation of the envelopes of the left and right ear impulse responses is taken rather than of the raw signals. The impulse response envelopes are calculated as the magnitude of the Hilbert analytical signal for each of the left and right signals respectively using the *hilbert.m* function.

The ITD curve that is the result of the Edge Detection method is shown in figure 3.2.1c, it is computed simply by defining for each pair of HRIR measurements; a pair of threshold values, 15% of the peak value in the left and right measurements, for each angle. The ITD in samples at each angle is then equal to the difference between the sample numbers at which the left and right measurements first exceed the corresponding threshold value. As with the IACC methods the ITD values in samples are then divided by the sample rate and upsampling factor to obtain the corresponding values in seconds.

Figure 3.2.1d shows the ITD curve derived from the phase based IGD method. It is computed ideally in three steps for every angle; firstly by calculating the group delay of the measured HRTF pair and its corresponding minimum phase representation, secondly by subtracting them from one another to obtain the group delay of the excess phase components for both the left and right ear impulse responses, then finally by evaluating the group delay at 0Hz, the difference between the two evaluations is the ITD in samples. However in practice it is usually not possible to obtain a meaningful evaluation of the group delay at 0Hz for real measurements, due to the limitations of the physical transducer, to overcome this the group delay has instead been evaluated at the first adjacent frequency bin, 172Hz, assuming that for low frequencies the group delay is constant [Minnaar and Plogsties, 2000].

Comparing the four curves it is clear that the IACC and IACCe methods both suffer from significant discontinuity for angles close to the inter aural axis, though the discontinuity is somewhat smoothed in the IACCe case at the apparent cost of further more subtle discontinuities at angles slightly further to either side of the inter aural axis. These discontinuities have been explained by other authors as a result of the comparatively low signal to noise ratio of the contralateral impulse responses and the subsequent lack of coherence this may cause between the ipsilateral and contralateral measurements. The Edge Detection method yields an ITD curve very close to that of the spherical head model however it exhibits a minor skewing that results in a slight under prediction of the values near the interaural axis, this effect may be due to the physical shape of the KEMAR mannequin head not being a perfect sphere. The phase method matches the overall shape of the spherical head model curve very closely however it suffers from consistant small scale fluctuations that are likely caused by noise in the measured HRIR.

Given the consideration of the computed ITD curves shown in figure 3.2.1, for the TU Berlin dataset the Edge Detection method seems to be the most appropriate of the four. The comparatively smooth behaviour of the ITD curve and lack of discontinuities is congruent with the expected ITD behaviour. This is further evidenced by the goodness of fit of the derived curve with that of the estimation belonging to the spherical head model, which as previously stated is known to achieve a good approximation of the ITD in the azimuthal plane. The Edge Detection method has been shown to match well with psychoacoustically derived ITD

values [Busson et al., 2005], and importantly, the method is also computationally inexpensive, especially when compared to the phase method that requires the use of comparatively complex routines to evaluate the various phase components at each position.

3.3 Error Metric

In order to be able to effectively interpret the relative merit of HRTF compression methods, specifically reconstruction techniques, a consistent error metric must be adopted. One such metric that appears quite commonly in the field of HRTF compression is the mean squared error (MSE) between reference and reconstructed or modelled spectra defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i - X_i)^2 \tag{3.3.1}$$

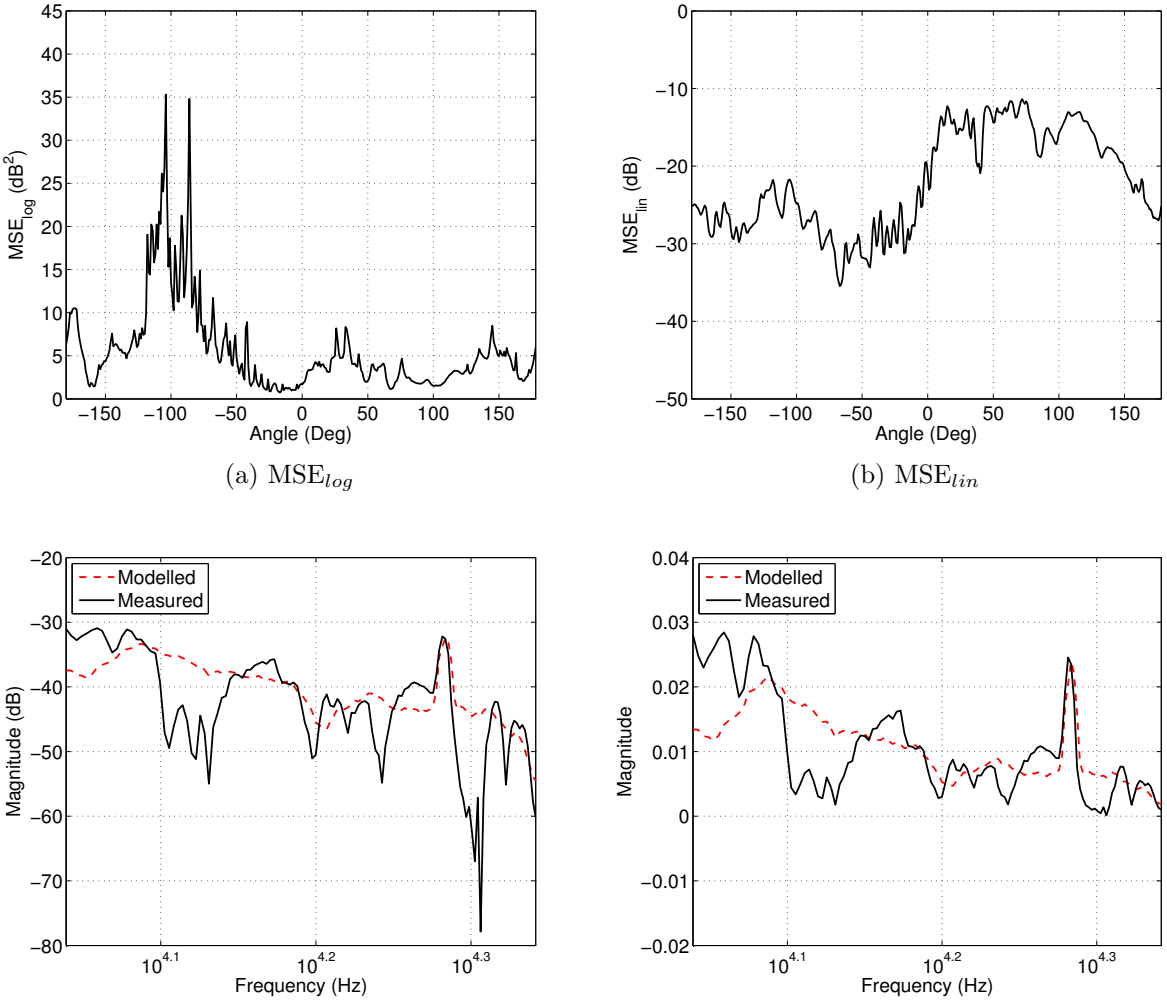
The mean squared error is a parameter often used in statistical disciplines to asses the quality of an estimator or series of predictions with reference to measured data. The MSE can be thought of as a means of costing the goodness of fit between two vectors such as two frequency spectra, and as such provides a convenient and consistent means of comparing different prediction models, or in the case of the current works, of comparing different reconstruction and even interpolated reconstruction methods.

A potential criticism of the mean squared error is that due to the squaring of every term, larger errors are effectively weighted more heavily than smaller errors. This can lead to marred interpretations of overall error trends when the MSE value is calculated for data containing outliers. However in the application of MSE as an error metric between a measured and modelled/reconstructed HRTF this criticism can be considered somewhat advantageous; the effective weighting of larger errors means that the MSE calculated between two spectra will be dominated by the extremes of the data, i.e. the spectral peaks. This means that the use of MSE in this context is actually quite apt, as the auditory system can also be considered to be peak dominated; peak amplitude frequencies cause maximal excitation on the basilar membrane and lower amplitude frequencies, particularly those that occur adjacent to the area of maximum excitation on the membrane can become masked. Thus suggesting that the peak frequencies can be considered the most important in the characterisation of a complex sound consisting of multiple frequency components.

MSE has been utilised in many previous works in the field, sometimes a variation of the MSE such as the percentage mean error (PMSE) [Wang et al., 2008] is developed, but the

key components, and limitations, of the MSE metric remain, they are merely reported in alternate units.

As such it can be deemed that MSE provides a somewhat psychoacoustically weighted error metric that will serve to highlight error regions that are most likely to introduce tangible subjective listening misjudgements. However, careful consideration should be given to the application of MSE as an error metric when considering in particular; the error between frequency spectra. There exists an inherent ambiguity in the definition of the MSE error, pertaining to wether the MSE should be calculated between the logarithmic (dB) or linear spectra of the measured and modelled systems, as the results can vary substantially not only in scale but also in meaning.



(a) MSE_{log} (b) MSE_{lin}

(c) Log Freq. Response (d) Lin Freq. Response

Figure 3.3.1: Comparison of logarithmic and linear MSE calculations

An example of the effect that domain in which the MSE is calculated is shown in figure 3.3.1; MSE_{log} (3.3.1a) is calculated using equation 3.3.1 where $\hat{X}$ and X are the measured and modelled logarithmic spectra (in dB) respectively, whereas MSE_{lin} (3.3.1b) is calculated again using equation 3.3.1 but where $\hat{X}$ and X are the measured and modelled linear spectra, the resulting MSE_{lin} values are then converted the log domain using $10log_{10}(X)$.

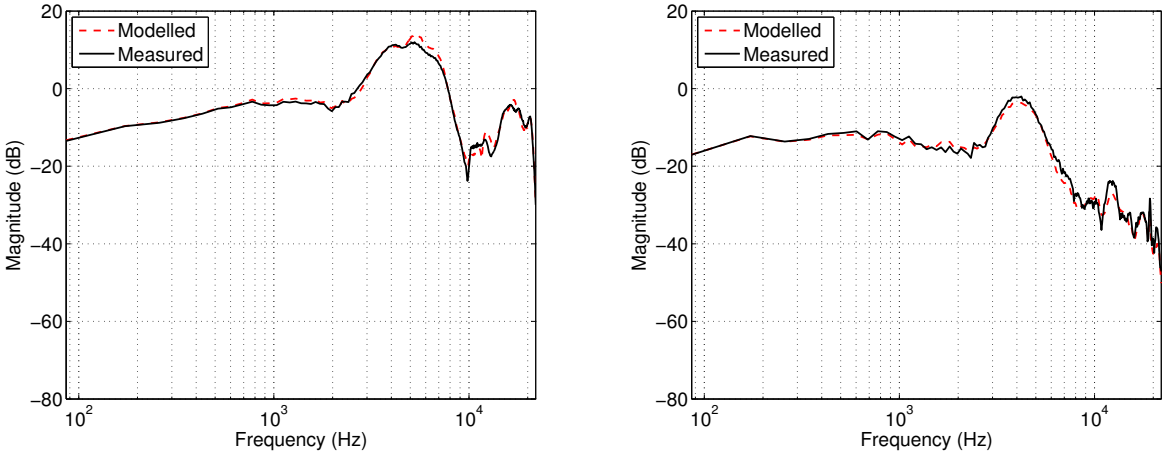
Figures 3.3.1c and 3.3.1d illustrate the difference between a portion the measured and mod-

elled frequency spectra at a single angle selected for demonstrative purposes. The frequency spectra are shown in the log domain in 3.3.1c and the linear domain in 3.3.1d. The modelled results used to create these demonstrative figures are obtained using the log-domain PCA based decompositional approach.

The resulting distributions for MSE calculated in the linear and logarithmic domains over angles -180 to +180 differ substantially, not simply in terms of scaling but more importantly in terms of meaning; the MSE_{log} distribution shows a clear increase in error in the contralateral hemisphere, whilst the MSE_{lin} distribution shows the opposite.

Certain peaks and nulls are aligned in both distributions however many are not, and peak error values seem far more pronounced in figure 3.3.1a. This is almost certainly due to large increases calculated in MSE_{log} that occur due to the underprediction of significant notches that arise in the logarithmic frequency responses of many measurement positions at high frequencies. Figure 3.3.1c shows a portion of the high frequency responses of a measured and modelled HRTF; notice the grossly underpredicted notch that occurs at approximately $10^{4.3}$Hz, sharp notches such as these are common among the higher frequencies and of course occur at frequencies of low magnitude, however these detailed high frequency notches are not crucial to the adequate reconstruction of spectral based locational cues [Humanski and Butler, 1988] [Morimoto, 2001] and as such it is inconvenient that they skew the MSE calculated across the spectrum. The relative magnitude of these notches is an artefact introduced by the conversion of the linear magnitude frequency response to the logarithmic domain, as the log domain is unbounded in the negative direction, i.e. extending infinitely, the finite linear range between one and zero maps from zero to minus infinity upon conversion. This means that positive linear values approaching zero become much much smaller and as such high frequencies of low linear magnitude are expressed in the log domain as sharp notches. Therefore it may be more appropriate to calculate the MSE in the linear domain, which can then be expressed logarithmically, for as it can be seen in figure 3.3.1d, in the linear domain the low magnitude of the high frequency components and more importantly the difference between two low magnitude components is not numerically exaggerated as it is in the log domain.

Upon studying the measured and modelled spectra at the angles which pertain to a minimum and maximum in the MSE_{lin} a further characteristic of the linear MSE calculation becomes apparent.



(a) Maximum MSE_{lin} (b) Minimum MSE_{lin}

Figure 3.3.2: Comparison of maximum and minimum MSE_{lin} cases

Figure 3.3.2 shows the measured and modelled logarithmic spectra at the angles for which the maximum and minimum MSE_{lin} values can be seen in figure 3.3.1b; approximately $+75°$ and $-65°$ respectively. According to the MSE_{lin} calculation there is a difference between the error of the two cases of approximately 25dB, however it can clearly be seen that both cases exhibit a similar performance upon visual inspection of figure 3.3.2. The larger magnitude of the ipsilateral HRTFs means that the same relative error performance as may be present in a corresponding contralateral model will be reported as larger due to the larger amount of total energy in the ipsilateral measurements.

To abate this inconsistency the linear MSE calculation can be improved by normalising the MSE at each angle by the total amount of energy in the measured spectra for that angle, thus yielding an error metric that reports the relative MSE performance between modelled spectra made for different angles. This metric is referred to as the normalised means squared error (NMSE) and is defined as:

$$NMSE = \frac{\sum_{i=1}^{n}(\hat{X}_i - X_i^2)}{\sum_{i=1}^{n}(\hat{X_i}^2)} \tag{3.3.2}$$

The works described in this thesis are assessed, when appropriate using the NMSE metric.

Chapter 4

Decompositional Approach

This section of the thesis details the application of a decompositional approach to the compression of the TU Berlin azimuthal HRIR dataset. The decompositional technique used is the principal component analysis according to the description given by Kistler & Wightman [1992]. It is noted that this description is congruent with other PCA definitions offered by alternative authors [Jolliffe, 2005] [Martens, 1987], as well as Chen et al.'s description of the discrete Karhunen-Loeve transform [Chen et al., 1995].

The terminology used to describe the output elements of the principal component analysis varies slightly from the aforementioned works. In this section, and for the remainder of the thesis, the eigenvectors of the covariance matrix will be described as the basis vectors, the corresponding weightings will be described as the weight vectors, and the term principal component will be used to describe a singular pair of vectors consisting of a basis vector and its corresponding weight vector. Common to all previous author terminologies, the order of principal components will be described such that the first principal component is the basis and weight vector pair that explains the largest proportion of the variance of the total data set, i.e. the eigenvector (basis vector) corresponding to the largest eigenvalue.

The principal component analysis and subsequent decomposition is performed on both the linear and logarithmic magnitude components of the TU Berlin HRTF data, obtained by the Fourier transform of the measured HRIR data.

The principal component analysis was performed using the software package MATLAB [The MathWorks Inc., 2014] according to the following procedure:

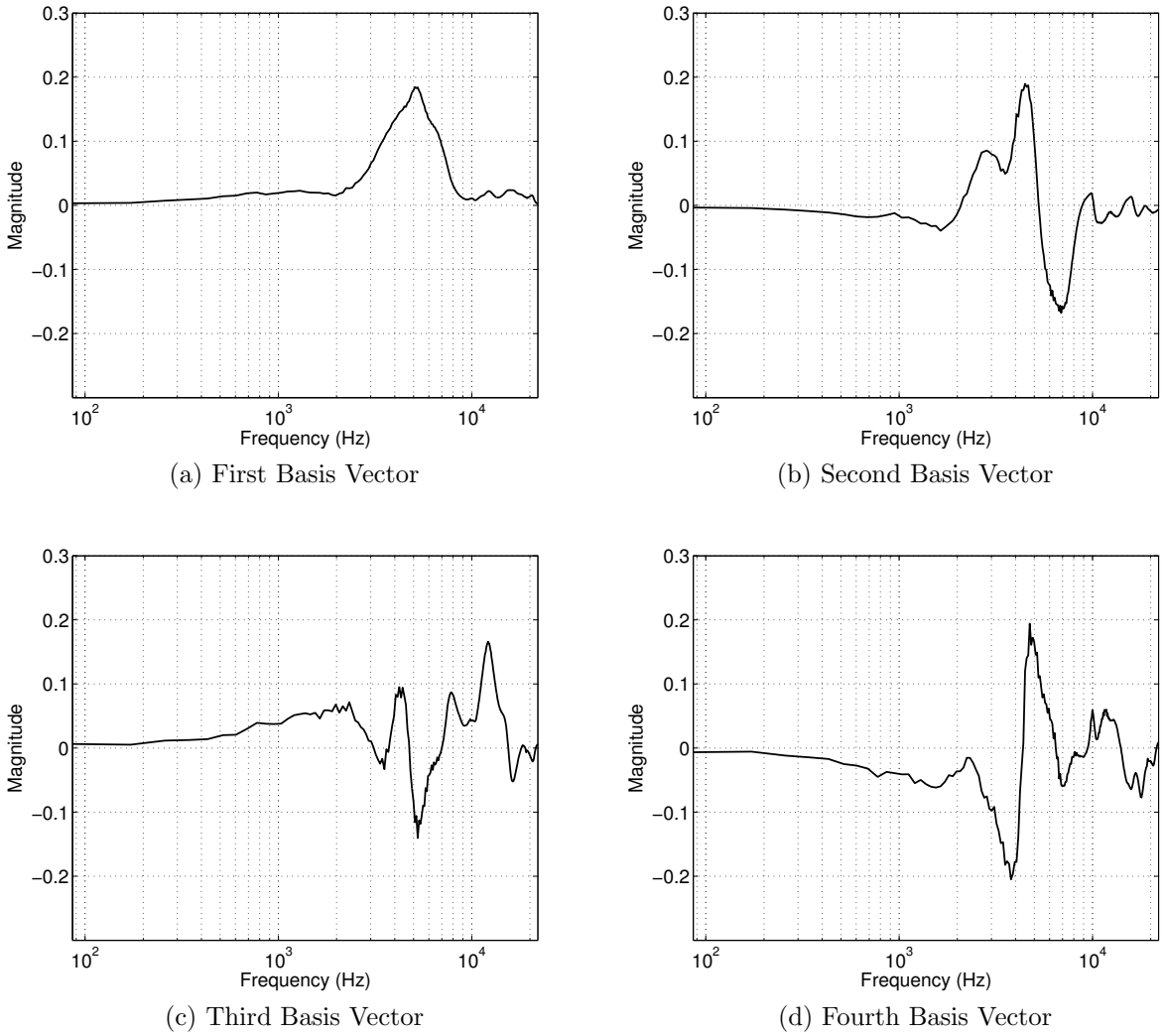
First the magnitude spectra are arranged in a matrix such that each column corresponds to a Fourier transform frequency bin, and each row corresponds to a measurement position, or angle. The mean value of each frequency bin calculated across all angles is calculated to form a row vector equal to the average spectrum of the measurement set. This average spectrum is then subtracted from all rows, i.e. each of the measured magnitude spectra, to leave the original data less the mean; the measured direct transfer functions (DTFs). The *cov.m* function in MATLAB is used to compute the covariance matrix of the DTF dataset, and the *eig.m* function is called to extract the eigenvectors and corresponding eigenvalues of the matrix. The output of *eig.m* is a matrix, each column of which is an eigenvector, and a diagonal matrix of corresponding eigenvalues in ascending order; the *fliplr.m* function is used to reverse the order of the eigenvalues and corresponding eigenvectors such that the largest eigenvalue and eigenvector are contained in the first column of the two matrices. Finally the weights are computed according to equation 2.6.5 with *inv.m* used to perform the inversion of the eigenvector matrix.

It has already been stated that the eigenvalues of the covariance matrix correspond to the amount of total variance of the data set that is explained by the principal component corresponding to that eigenvalue. However, the eigenvalues themselves explain little other than the order of importance of the PCs; a more useful representation of the eigenvalues is to calculate the percentage of total variance explained by each PC. This can be calculated simply as:

$$V_i = \frac{\lambda_i}{\sum_{n=1}^{N} \lambda_n} \times 100\% \tag{4.0.1}$$

where V_i is the percentage explained by the i^{th} principal component, λ_i is the i^{th} eigenvalue, and N is the total numbers of principal components and subsequently eigenvalues.

4.1 PCA of Linear HRTF Magnitudes



(a) First Basis Vector

(b) Second Basis Vector

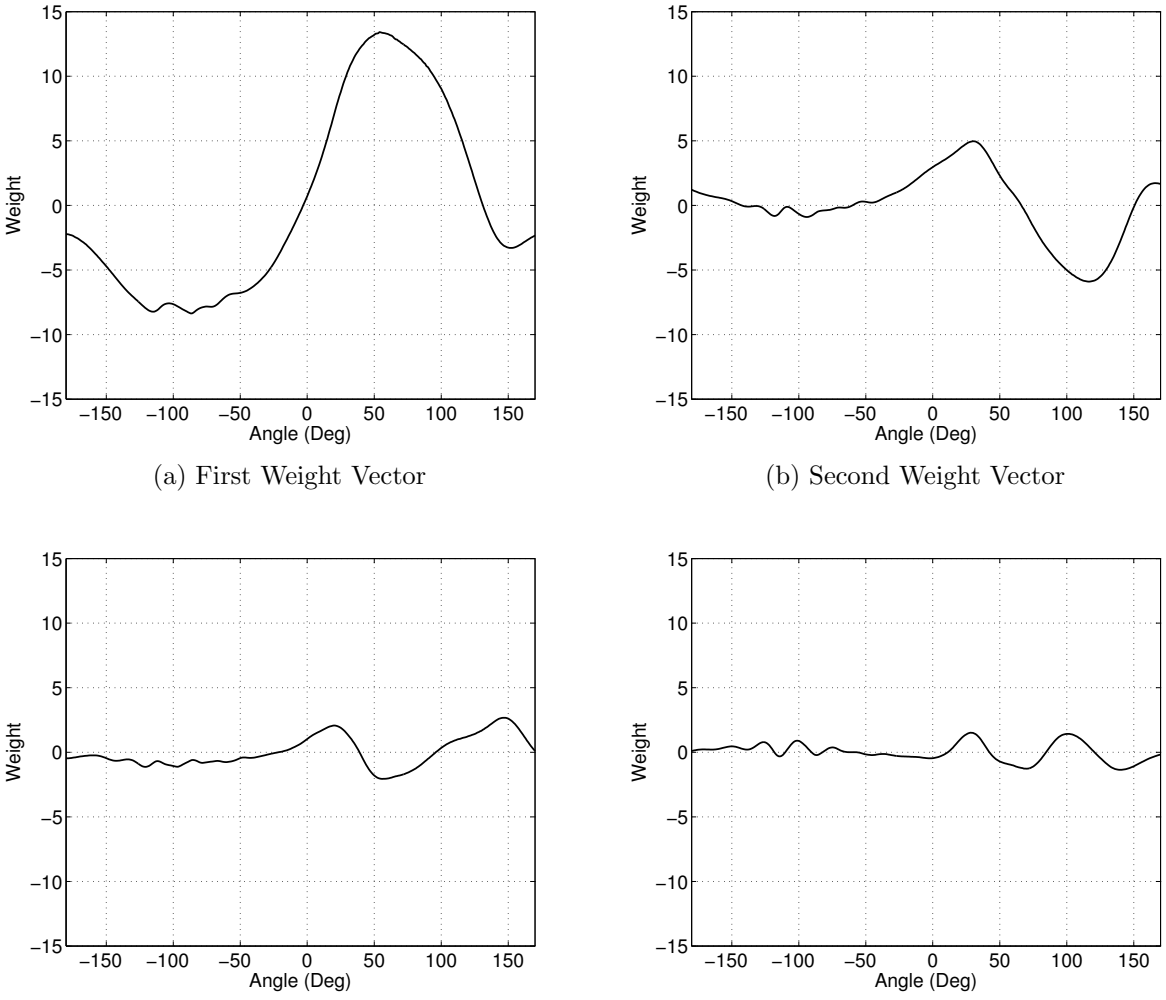
(c) Third Basis Vector

(d) Fourth Basis Vector

Figure 4.1.1: Linear magnitude basis vectors

Figure 4.1.1 shows the first four basis vectors in principal component order, i.e. in order of descending variance explained. Note that all four basis functions are approximately or at least very close to zero for frequencies below approximately 1-3kHz, this occurs due to the fact that there is little directionally dependent variation at theses lower frequencies, likely due to the diffraction that occurs allowing the low frequency waves to 'bend' around the

comparatively small diameter of the head. It is clear that the first four basis vectors that account for 85, 11, 2, and less than 1 percent of the total variance respectively, are dominated by the directionally dependent high frequency variations occurring from around 3kHz and up.



(a) First Weight Vector

(b) Second Weight Vector

(c) Third Weight Vector

(d) Fourth Weight Vector

Figure 4.1.2: Linear magnitude weight vectors

Figure 4.1.2 show the first four weight vectors that correspond to the first four basis vectors shown in figure 4.1.1. The weight vectors describe relative contribution of each of the basis vectors to the original HRTF for all measured angles (or positions). All four weight vectors

exhibit similar ripple-like behaviour over the contralateral angle region of approximately -150 to -50 degrees, once again this is likely due to the lower signal to noise ratio of the contralateral HRIRs. The first principal component weight vector shown in figure 4.1.2a shows a general formulation such that the contribution of the first basis vector, that accounts for $\sim$85% of the total variance, is negative in the contralateral hemisphere and mostly positive in the ipsilateral hemisphere. This trend is slightly skewed by the rear positions (-180 and +180) that are also negative, and the curve crosses the x axis to become positive approximately 5° from the frontal position (0°) towards the contralateral side. This skewing is likely due to the orientation of the KEMAR ear canals, which lie slightly off the inter-aural axis. It can further be seen that there is a difference in the magnitude of the positive and negative peaks of first weight vector; the negative peak, in the contralateral hemisphere, has smaller magnitude than that of the positive peak, which lays within the ipsilateral hemisphere. The smaller magnitude of the minimum peak is due to the reduced level at the (left) ear for contralateral source positions, i.e. head shadowing.

The implications of the trend described in the first weight vector in figure 4.1.2a are somewhat difficult to interpret given that the corresponding basis vector represents a linear spectral shape, but the most evident features do seem consistent with the remarks made by Kistler & Wightman during the initial analysis of their log magnitude PCA components [Kistler and Wightman, 1992].

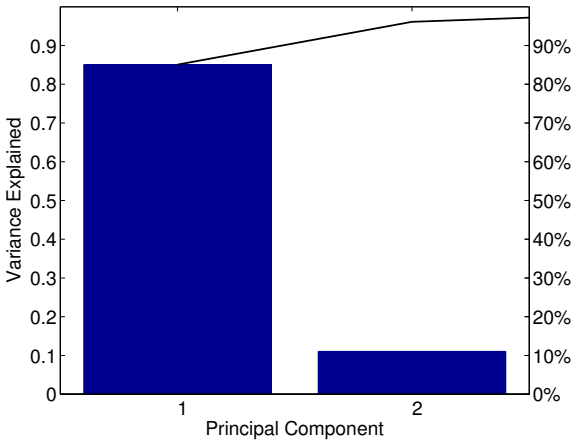


Figure 4.1.3: Pareto plot : Linear magnitude PCA

Figure 4.1.3 illustrates the amount of variance explained by each of the principal components until 95% of the total variance has been cumulatively explained. The bars represent the amount of variance explained by each principal component and the black line depicts the cumulative total of variance explained by the principal components combined. For the PCA applied to the linear magnitude HRTF data 95% of the variance of the data set is captured by the first two principal components; that is two basis vectors defining key spectral shapes, and a corresponding two weight vectors defining the angular variation in the contribution of these spectral shapes.

This result in itself suggests that a high level of compression can be achieved using this method, the reduction from 360 HRTFs of length 256 frequency bins each to two basis vectors of length 256 frequency bins, and two weight vectors of length 360 angular positions, is a significant reduction. In fact this result suggests that the method of applying the PCA to the linear magnitude components of the HRTFs outperforms the methods described in previous works [Kistler and Wightman, 1992] [Wang et al., 2008], both of which suggest more than two principal components are required to recapture as little as 90% of the variance of other datasets. This is most likely due to the fact that both the dataset used by Kistler & Wightman and the CIPIC database used by Wang et al. feature measurements made across several different listeners as opposed to solely the KEMAR mannequin as in the TU Berlin

data. It is to be expected that greater variance will be introduced into the measurement suite for every additional head measured and thus the principal component analysis would infer seemingly less efficient results.

It is also key to discuss that although this method yields conveniently smooth and arguably appealing basis and weight vector definitions, they mask a critical obstacle that becomes apparent in the attempted reconstruction of the data using a reduced number of principal components. Due to the low magnitude of certain high frequency components in the measured HRTFs, some of which approach zero, that occur predominately at the contralateral measurement positions; reconstructions at some of these angles do not exceed zero at said frequencies. The information required to recreate such frequencies at contralateral angles may well be contained in high order principal components; components that may not be included in the reconstruction. It is also possible that the total magnitude of these more subtle spectral fluctuations at higher frequencies are staggered across many of the principal components, unlike the macroscopic spectral fluctuations that are captured by low order principal components that account for a greater proportion of the total variance. Though at larger magnitude frequency components the loss of a small amount of detail in terms of absolute magnitude may not be a concern, at higher 'notch' frequencies for which the total magnitude is of similar order to the magnitude inaccuracies introduced by the reduction of the number of principal components used in the reconstruction, this presents an issue.

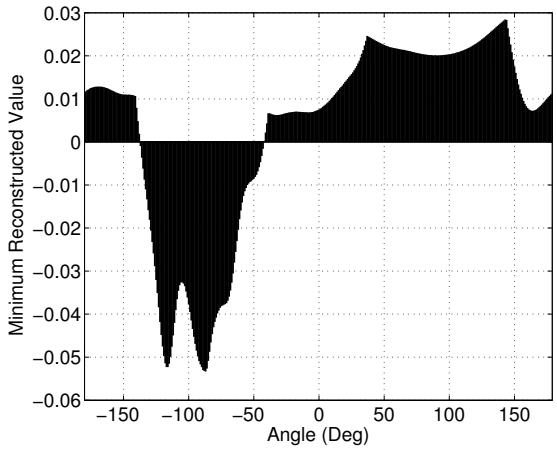


Figure 4.1.4: Minimum reconstructed values of linear magnitude PCA

Figure 4.1.4 shows the minimum reconstructed value for each azimuth angle using just 2 principal components. Clearly there exists a wide region centred on the contralateral azimuth for which the principal components used to reconstruct the data do not carry enough of the measured magnitude information to ensure that the reconstruction clears zero. Although the reconstructed negative components are of small magnitude themselves and in other disciplines such a reconstruction may be suitable and valid, however in an acoustical sense they simply cannot be ignored. A negative linear magnitude implies a complex logarithmic magnitude (dB) as:

$$|X|_{dB} = 20log_{10}(|X_{lin}|) \tag{4.1.1}$$

For which $log_{10}(-x)$ is undefined; considering $y = log_{10}(x)$, as the base (10) is positive, the base raised to the power of y must be positive for any real value of y.

A possible means of overcoming this would be to 'correct' all negative values to an almost negligibly small positive value, doing so would avoid the aforementioned numerical complications however it is possible that the relative increase of 'corrected' notch frequency magnitudes to the more correctly captured peak frequency magnitudes could serve to distort

the binaural cues contained within he reconstructed HRTF.

The simplest means of avoiding the issue however, is to ensure that enough principal components are used in the data reconstruction so that the minimum reconstructed value at each angle is greater than zero, as it is in the unprocessed measured HRTF data.

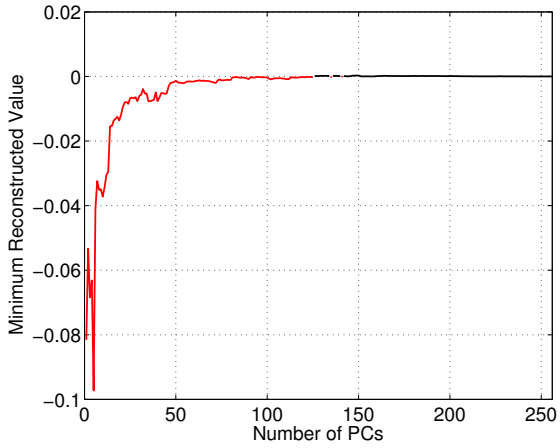


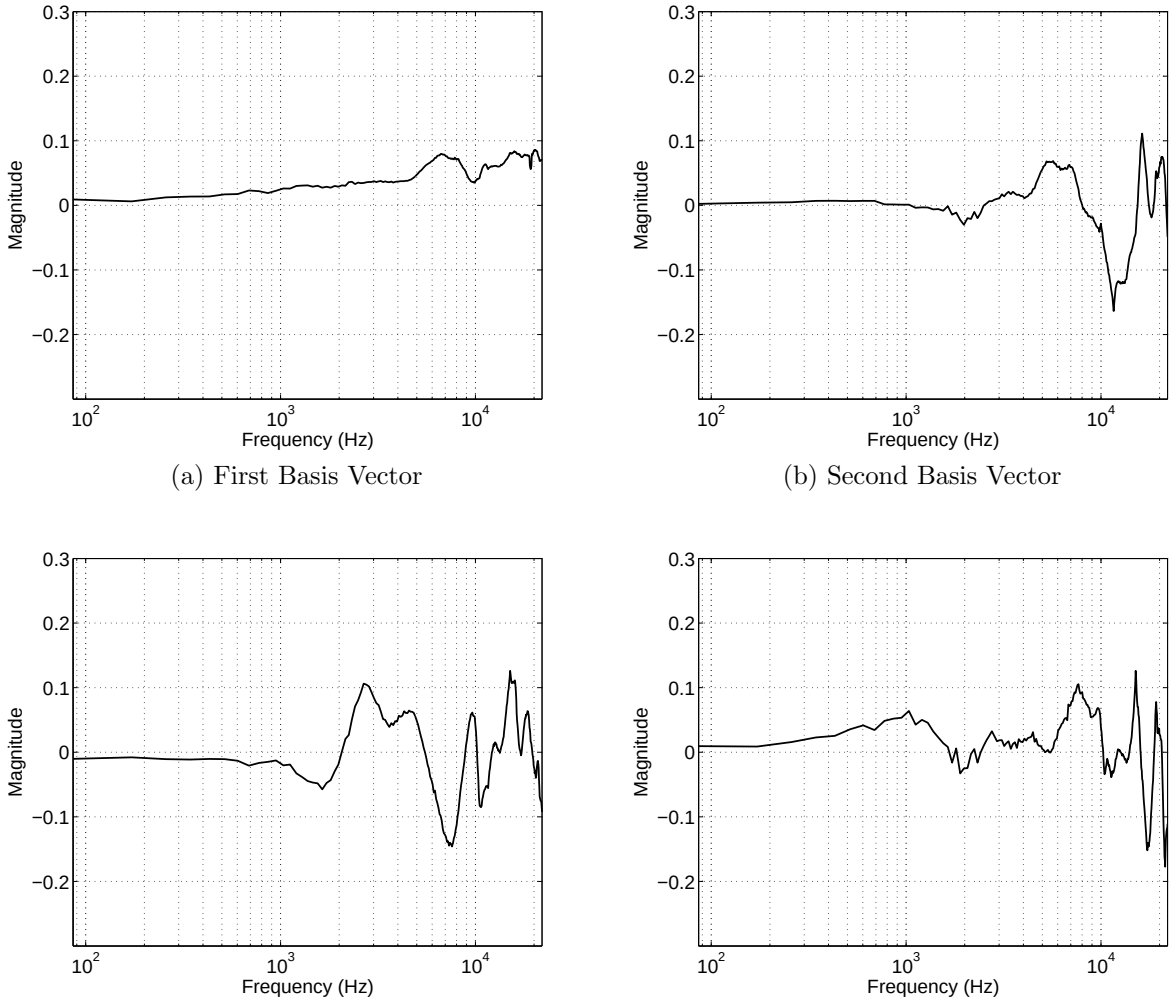
Figure 4.1.5: Minimum reconstructed values against number of PCs used in reconstruction

Figure 4.1.5 shows the number of PCs required for the minimum reconstructed value to clear zero; the red region of the curve indicates a minimum reconstructed value of less than or equal to zero, whereas the black region indicates a minimum reconstructed value greater than zero. It can be seen that a minimum of between 125 and 150 principal components are required to provide a wholly positive reconstruction of the linear spectra at all angles.

Although still a viable means of HRTF dataset compression, due to the numerical complication that arise from the small magnitude values at high frequencies in contralateral angles,and the subsequent need for a large number of principle components to be retained in reconstruction, it is likely that this is far from an optimal representation of the data.

4.2 PCA of Logarithmic HRTF Magnitudes

A similar approach, that was adopted by several previous authors is to apply the principal component analysis to the log magnitude values of the HRTFs, i.e. the spectra in dB. In converting the linear magnitude spectra into their log counterparts ($20log_{10}(|X|)$) before conducting the principal component analysis, the numerical issue of attempting to take the logarithm of a negative reconstructed value is bypassed completely as the HRTFs are already expressed in terms of the desired and more commonly used decibel scale.



(a) First Basis Vector

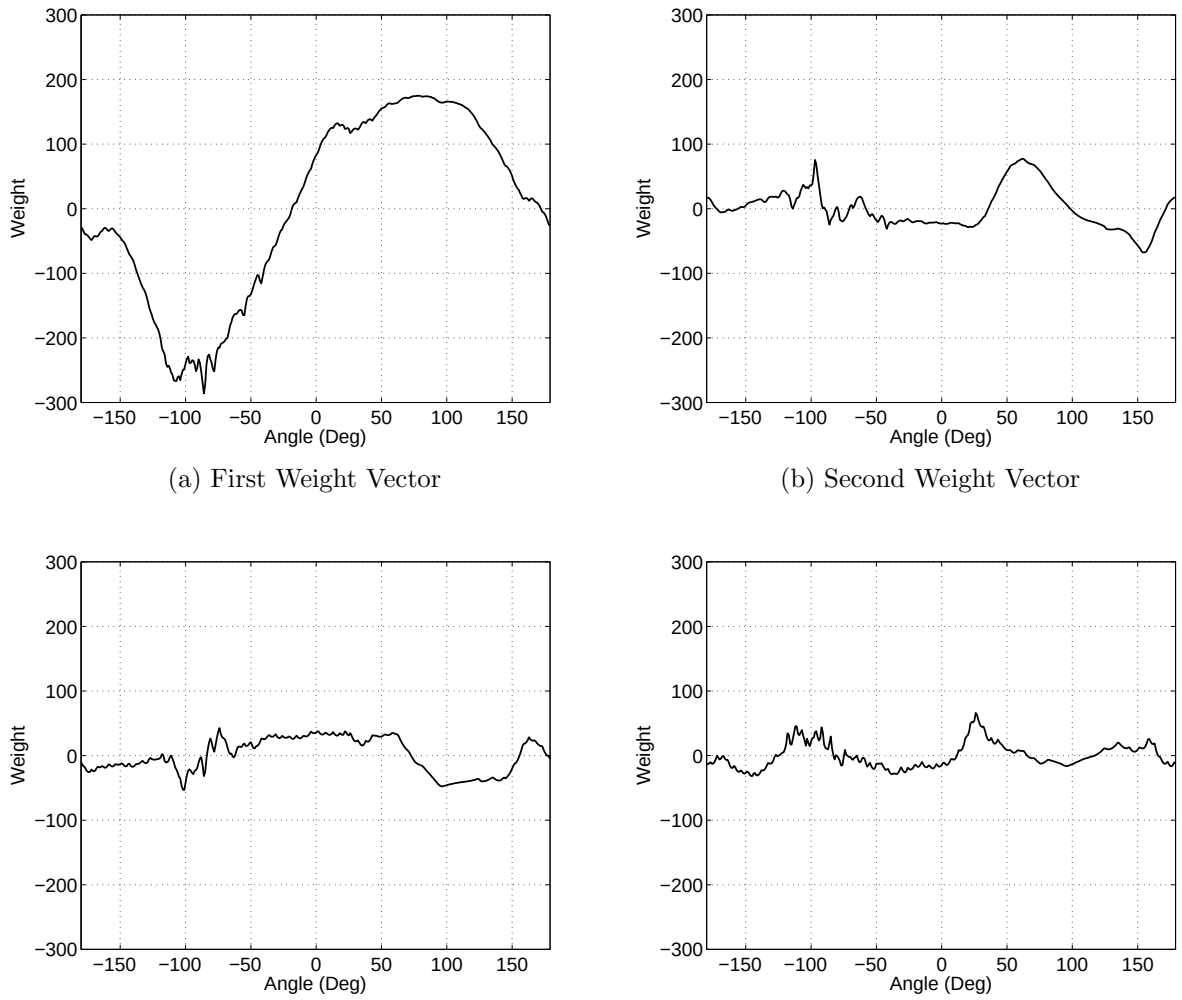
(b) Second Basis Vector

(c) Third Basis Vector

(d) Fourth Basis Vector

Figure 4.2.1: Logarithmic magnitude basis vectors

Figure 4.2.1 shows the first four basis vectors derived from the PCA of the HRTF log magnitudes. Similarly to the first four basis vectors of the linear magnitude PCA (figure 4.1.1), the second and third basis vectors are fairly constant at zero below approximately 1kHz, however the first and fourth are not. The first basis vector is of a fairly constant shallow positive gradient below approximately 5kHz above which it exhibits a slight peak and small notch before seemingly settling somewhat into the higher frequencies (10kHz and above). The first four basis vectors account for approximately 85, 4, 3 ,and 2 percent of the total variance explained and clearly reflect an important directionally dependent fluctuation in the magnitudes of higher frequency components. This was identified similarly for the linear magnitude basis vectors, and the log magnitude basis vectors reported by other authors [Kistler and Wightman, 1992] [Martens, 1987].



(a) First Weight Vector

(b) Second Weight Vector

(c) Third Weight Vector

(d) Fourth Weight Vector

Figure 4.2.2: Logarithmic magnitude weight vectors

Figure 4.2.2 shows the first four weight vectors corresponding to the first four principal components of the log magnitude PCA. As with the linear PCA weight functions in figure 4.1.2 the first weight function is perhaps the most easily interpreted in a physical sense; the first weight vector in figure 4.2.2a exhibits a similar shape to the first linear magnitude weight function shown in 4.1.2a, albeit on a much larger scale, this suggests that the PCA of both the linear and logarithmic HRTF magnitudes yields the same first principal component. However it is easier to observe the nature of the first component in the log magnitude analysis; the

first basis vector is effectively of a shallow positive gradient from roughly 200Hz and beyond, with some slight variation at higher frequencies, this spectral shape coupled with the weight vector which exhibits mostly negative values in the contralateral hemisphere and mostly positive values in the ipsilateral hemisphere, implies an overall de-emphasis of high, and to a lesser extent mid, frequencies in the contralateral hemisphere and an increased emphasis of the same frequencies in the ipsilateral hemisphere. The first weight vector reaches minimum and maximum values along the inter-aural axis, that is at $\pm 90°$, and as such these values give rise to the largest emphasis and de-emphasis of the high frequencies. The nature of the first principal component in both the linear and logarithmic magnitude PCA offers a conveniently interpretable definition of a primarily angular-dependent phenomenon of the HRTF; high frequency content is de-emphasised in the contralateral hemisphere, maximally at the contralateral ear, caused by the lack of diffraction due to the order of the diameter of the head and the subsequent acoustic shadow cast by the hard skull of the listener at these frequencies. A similar trend can be seen in the first principal component derived in the works of other authors [Kistler and Wightman, 1992] .

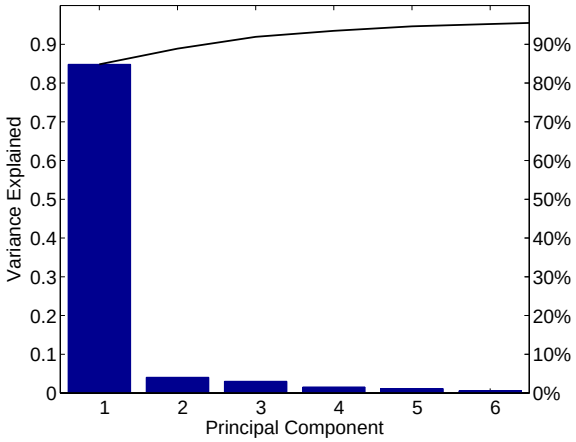


Figure 4.2.3: Pareto plot : Logarithmic magnitude PCA

Varying further from the similarities with the linear magnitude PCA, figure 4.2.3 shows that the first six principal components are needed in order to explain 95% of the total variance of the log magnitude data. The first principal component accounts for approximately 85% of the total variance in the data with the remaining five only accounting for less than roughly

4% each. The amount of variance explained by the first component is almost identical to the amount of variance explained by the first component of the linear magnitude PCA, shown in figure 4.1.3. This suggests that the first principal component derived in both analyses represents the same characteristic of the measured HRTF, namely the de-emphasis of high frequency components of the HRTFs located in the contralateral hemisphere. The need for another five components to capture the remaining 10% of the total variance, and to account for 95% of the total variance, in contrast to the one additional component required in the linear magnitude analysis, reflects the visible differences in the smoothness of the weight vectors. The increased amount of small magnitude fluctuations present in the logarithmic magnitude weight vectors are likely caused by the exaggeration of some 'microscopic' details in the measured spectra, that become increasingly statistically significant when the HRTF spectra are analysed on the numerically much larger, negatively unbounded, logarithmic (dB) scale.

4.2.1 Reconstruction Performance

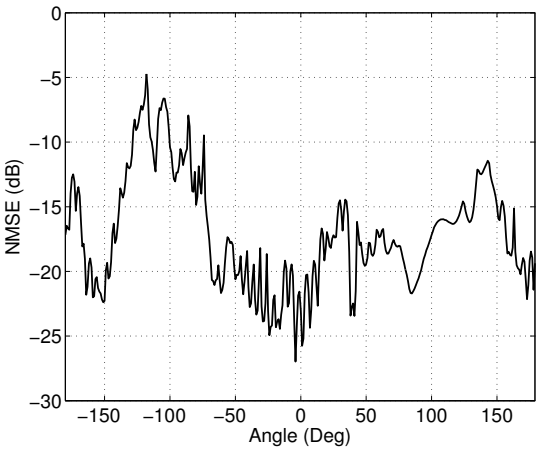
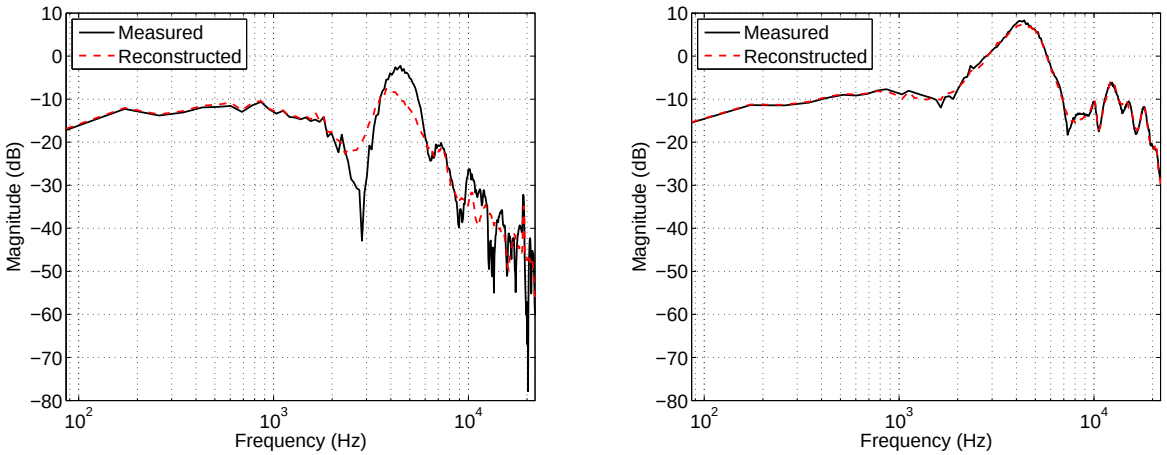


Figure 4.2.4: Normalised mean squared error of 6 PC reconstruction

Figure 4.2.4 shows the NMSE calculated between the measured HRTF magnitude spectrum and the counterpart spectrum reconstructed using only the first six principal components for all measured angles (-180° to 180°). The figure shows that the six component reconstruction

performs similarly well for the majority of all angles, though with a significant increase in error around the contralateral and rear positions. Note that the maximum error does not occur at the contralateral angle $-90°$ but slightly further towards the rear of the head, the off-interaural-axis occurrence of the two error peaks towards the rear and the descending two peaks that fall slightly toward the front of the head with reference to he contralateral angle, indicate a possible source of error that may stem from a specific HRTF characteristic that is exhibited significantly only in a relatively small range of angles either side of the contralateral position. The asymmetry of the series of peaks is likely due to the asymmetry of the pinna about the interaural axis. Though in isolation this figure tells little besides the increased error in proximity to the contralateral position, it serves a reference for comparison when considering the error introduced in the interpolation discussed in the following section.

As the NMSE for each angle is calculated as a single value representing the average squared error across all frequencies, it is perhaps beneficial to investigate the source of the increased NMSE reported around the contralateral angle. This should shed light onto the nature of the increase NMSE in particular it should aid in identifying if the error stems from a particular frequency range, or an inherent flaw in the decompositional method at these angles.



(a) Maximum Error Reconstruction ($-118°$) (b) Minimum Error Reconstruction ($-4°$)

Figure 4.2.5: Maximum and minimum error reconstructions using 6 PCs

As stated above, the maximum value of the NMSE between the measured and reconstructed data was calculated at azimuthal angle -118°; figure 4.2.5a shows the magnitude spectra of both the measured and 6 PC reconstructed HRTFs at this angle. In contrast, figure 4.2.5b shows the measured and reconstructed magnitude spectra of an angle taken from a region of minimum NMSE. The reconstruction performs similarly well for frequencies below approximately 1kHz, where little directional dependence is expected, however the performance differs quite substantially at higher frequencies. In particular the maximum error case of the six component reconstruction, shown in figure 4.2.5a seems to over predict the magnitude of the dip just between 1kHz and 3kHz and the broad double peak centred around approximately 4kHz, this over-prediction of these two features appear to contribute somewhat to the increased error, however there also exists a severe notch in the measured response at approximately 19kHz that is entirely omitted from the reconstruction, the difference between the reconstructed response and the measured notch is on the order of 20dB, and will certainly have impacted the average of the squared error over the whole spectrum. As such extreme high frequency variation occurs only in a subset of the measured responses, centred around the contralateral angle, it can be assumed that the information needed to reconstruct them is contained in higher order PCs, as statistically they account for only a small amount of the total variance of all measured HRTFs.

The limited component reconstruction is less successful in capturing some of the more subtle high frequency detail above 10-15kHz, however such high frequencies are often regarded as less important in the preservation of localisation cues and have even been omitted from analysis and evaluation in the works of some other authors [Kistler and Wightman, 1992] [Evans et al., 1997] [Zhang and Abhayapala, 2009]. Subsequently the lack of reconstruction of the sharp high frequency notch or notches that occur upwards of 15kHz for some contralateral positions, such as in figure 4.2.5a, should not necessarily be considered to imply the need for more principal components in the reconstruction.

4.3 Interpolation of Weight Vectors

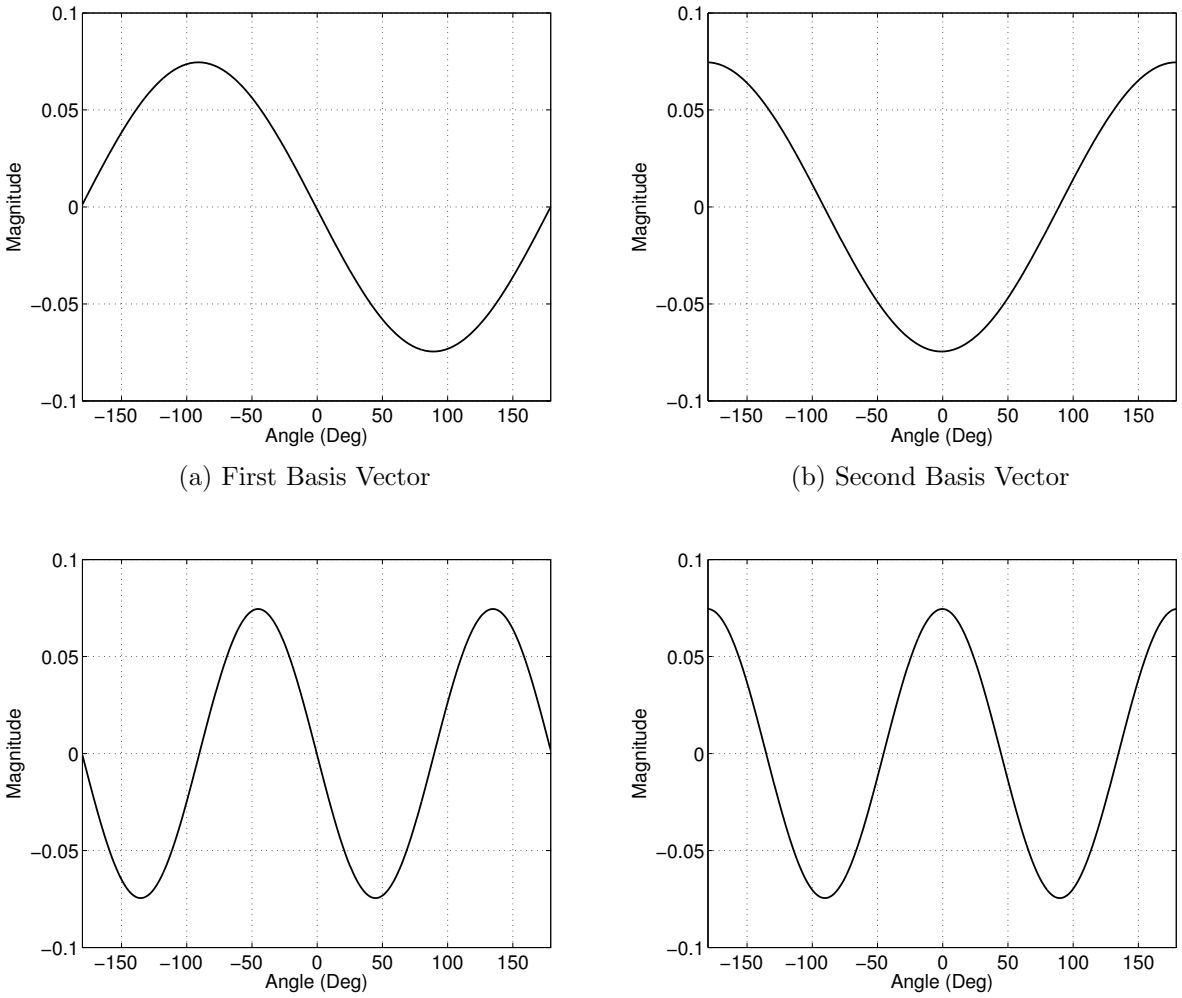
An advantage of the decompositional approach to HRTF compression is that it separates the positional and frequency dependent variations in the dataset. Such an approach offers

a logical and somewhat intuitive means of unambiguous HRTF interpolation through the interpolation of only the spatial decompositional components, i.e., the weight vectors. Each weight vector can be interpolated to effectively up-sample the measurement positions of the dataset by an interpolation factor as and when higher spatial resolution is required, however a stronger method sees each weight vector translated to a functional form that is continuous by definition; therefore allowing for HRTF interpolation with comparatively little run-time computation, as the functional representation can be evaluated for any desired measurement angle, whereas the former method requires computation of an interpolated vector through an interpolant that is dependent on the requirement for positional accuracy etc in the virtual auditory scene.

4.3.1 One-Dimensional PCA/KLT

In addition to the application of a PCA/KLT to a 2-dimensional dataset, as was demonstrated in the decomposition of the HRTF dataset into a series of frequency dependent basis vectors and spatially dependent weight vectors, the PCA/KLT can also be applied to a 1-dimensional set of data. In this case the 1-dimensional data is a single weight vector that will be decomposed into a series of orthogonal basis vectors and singular set of weights that describe the relative contribution of each basis vector to the reconstruction of the original weight vector. In this application the previous method must be altered slightly; instead of the eigenvectors and eigenvalues of the covariance matrix estimated from the matrix of input observations, the eigenvectors and corresponding eigenvalues are taken from the auto-correlation function of the input vector expressed in toeplitz matrix form.

In order to alleviate confusion, for the remainder of this section the formerly conducted principal component analysis of the log-magnitude spectra will be referred to as PCA-1, and the further 'sub' principal component analysis of the weight vectors derived in PCA-1 will be referred to as PCA-2.



(a) First Basis Vector

(b) Second Basis Vector

(c) Third Basis Vector

(d) Fourth Basis Vector

Figure 4.3.1: Basis vectors of first weight vector decomposition

Figure 4.3.1 shows the first four basis vectors derived from the first PCA-1 weight vector during the PCA-2 process. The principal component analysis is designed such that the derived basis vectors are an orthogonal series of optimally efficient basis functions, thus it is convenient to observe that the optimal basis vectors for the reconstruction of the PCA-1 weight functions appear to be sinusoidal/cosinusoidal. Note that although only the PCA-2 basis vectors derived for the first PCA-1 weight vector are presented, the PCA-2 basis vectors derived for all PCA-1 weight functions exhibit this characteristic.

Considering once more the PCA-1 weight vectors shown in figure 4.2.2, the overall variation of the weight values in the y axis is large or macroscopic in the first weight vector, as such one might expect to find that only a small number of low order sinusoidal components are needed to capture most of the variance in the vector. However for the successive weight vectors, the overall variation of the weight values tends towards becoming increasingly small or microscopic, hence one would correctly expect that a greater number of higher order or higher frequency sinusoidal components will be required to recapture the same amount of variance for each vector.

The presence of these underlying sinusoidal and cosinusoidal functions is perhaps not surprising when one considers the composition of the HRTF, less surprising still when considering the prior decomposition of the HRTF into the PCA-1 derived frequency and angular basis vectors. The angular variation of the HRTF in the azimuthal plane is by definition periodic, where one period completes a rotation around the head at a fixed distance, but also consider that, particularly in the case of the dummy head (KEMAR head in the case of the TU Berlin data), the geometry of the head is close to that of a sphere or cylinder. It is rational to expect to identify characteristics of near simple oscillations in the angular variation of the relative weightings of the crucial spectral shapes moving around the head.

An important outcome of the reduction of the PCA-1 weight vectors to the weighted sum of optimally decorrelated basis functions, and the observation that these basis functions so closely resemble simple sine and cosine functions becomes overtly evident when considering interpolation. The sine and cosine functions are continuous, they are mathematically defined for any possible input, therefore an approximate representation of the PCA-1 weight vectors as the weighted sum of continuous functions will also be a continuous function. This is a powerful implication for the compression of an HRTF measurement dataset, as not only can one express the data in a compact form consisting of a small subset of basis and weight vectors, but the weight vectors can further be simplified and expressed as set of weights and frequencies of sine and cosine terms, in addition to being superiorly compact to the original data, this representation is also theoretically functional in terms of measurement angle θ.

4.3.2 The Discrete Cosine Transform

In the same way that the discrete Fourier transform decomposes a finite length sampled signal into a weighted sum of sine and cosine functions oscillating at different frequencies, the discrete cosine transform decomposes a finite length sampled signal into a weighted sum of varied frequency cosine functions only. The DCT is a technique commonly used in data compression applications often in the fields of image and signal processing, that was originally derived as means of approximating the eigenvectors of a of the auto-correlation matrix of an autoregressive (AR(1)) signal block, which pertains to a special case of the Karhunen-Loeve transform [Malvar, 1992]. In this special case, for an AR(1) signal, as the correlation between adjacent samples tends to one, the KLT basis functions become sinusoidal.

The DCT of a finite length sampled sequence x_n is defined as:

$$X_k = \sum_{n=0}^{N-1} x_n cos[\frac{\pi}{N}(n + \frac{1}{2})k] \tag{4.3.1}$$

4.3.3 DCT Approximation of Weight Functions

Taking the N-point DCT of one of the PCA derived weight vectors yields a new vector describing the magnitude, or in the spirit of the aforementioned PCA, the coefficient weights of N cosine functions oscillating at orthogonal frequencies, the sum of which ($\sum_{n=0}^{N-1}$) returns the original PCA derived vector. The frequency of each cosine function is prescribed by the number of the frequency bin n, and the value in that bin prescribes the magnitude or weight of that cosine function contained in the original vector.

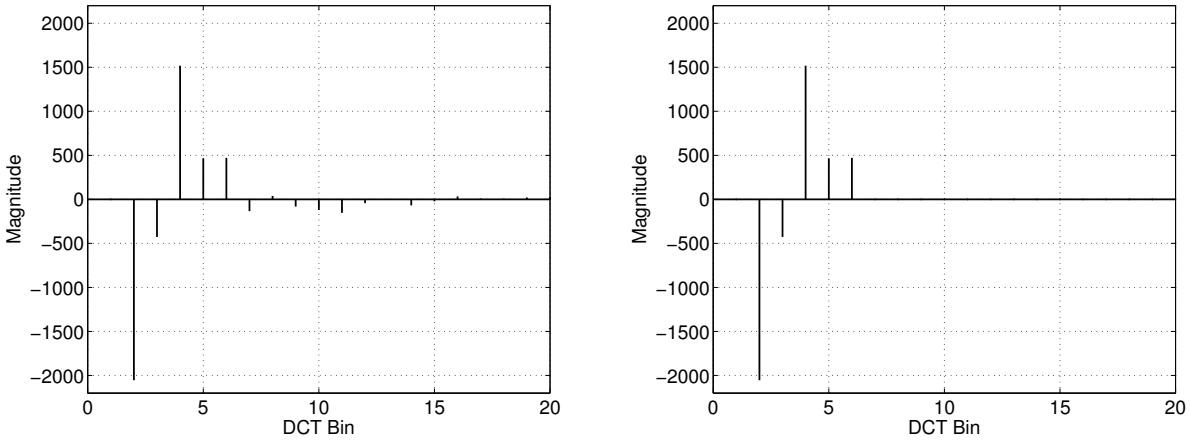
Much in the same way as with the PCA decomposition of the magnitude spectra vectors, it is true that only a subset of the N cosine functions are required in order to capture the total variance of the PCA derived weight vector up to some arbitrary threshold. It can be assumed that the DCT components with the largest magnitudes, that represent a larger contribution in the reconstruction of the original domain signal, are equivalent to the largest eigenvalues in the PCA composition; in other words, that the largest magnitude components of the DCT of series x account for the largest proportions of total variance explained in x.

By first arranging the DCT component magnitudes in descending order, the amount of total variance explained $V\%$ by i orthogonal cosine functions can be calculated as:

$$V_i = \frac{||x_i||}{||x||} \times 100 \tag{4.3.2}$$

Where x_i is the i^{th} largest magnitude cosine function and $||\cdot||$ denotes the 2-norm operation which can be written as:

$$||x_i|| = \sqrt{\sum_{i=0}^{n} x_i^2} \tag{4.3.3}$$



(a) First 20 DCT Components

(b) DCT Components Required to Explain 99% of Variance

Figure 4.3.2: DCT representations of first PCA-1 weight vector

Figure 4.3.2a shows the first 20 DCT components of the first PCA-1 weight vector. In 4.3.2b the DCT components are limited to the 5 components that have the largest magnitude and account for 99% of the variance of the weight vector. In figure 4.3.2b the 'non-essential' DCT components that pertain to the remaining $\sim$1% of the total variance have been set to zero.

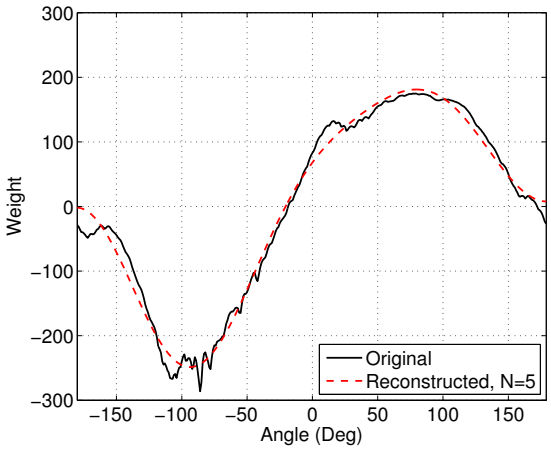


Figure 4.3.3: DCT reconstruction of first weight vector

Figure 4.3.3 compares the first PCA-1 weight vector with an approximated version reconstructed from the 5 largest DCT components only. The reconstructed weight vector is generated by taking the inverse discrete cosine transform of the DCT component vector, recall that although the vector is still 256 elements long, all but the largest 5 elements have been set to zero. The DCT method clearly provides an effective means of compression of the weight vectors in addition to the compression of the whole dataset realised by the initial principal component analysis. The number of DCT components required to capture the desired threshold level of the total variance in each of the weight vectors is dependent on the order of the weight vector, a general trend emerges that the number of DCT components needed increases as the order of the weight vector increases.

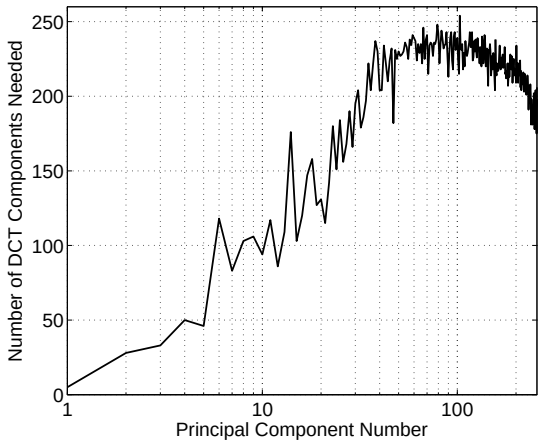


Figure 4.3.4: Number of DCT components vs principal component weight vector order

Figure 4.3.4 shows the number of DCT components required to capture 99% of the variance of each PCA derived weight vector. The figure shows a sharp increase in the number of DCT components required as the principal component number increases, this sharp incline appears to level off at approximately the 50^{th} principal component. The weight vectors themselves are each 256 samples long, evidently the DCT compression method offers little to no gain in efficiency in representing the weight vectors for the latter 200 or so principal components, however as was previously shown in figure 4.2.3 only the first 6 principal components need be considered to capture 95% of the total variance of the original dataset. Returning to figure 4.3.4 it can be seen that for the first 6 principal components the number of DCT components required to express the weight vectors is significantly less than 256, and so the DCT method offer a significant compression of the weight vectors.

The DCT also offers a convenient means of computationally efficient interpolation. Typically both the DCT and the IDCT performed to respectively deconstruct and then reconstruct one of the principal component weight vectors use the same number of evaluation points N. However it is possible to effectively oversample the original data by zero padding the DCT component vector in the reciprocal domain, or oversample the reciprocal domain response by zero padding the original vector. For example, considering an azimuthal HRTF dataset with 180 measurement points, or a spatial resolution of 2°; taking the 180 point DCT of

one of the PCA derived weight vectors of the set will yield a vector describing the DCT component magnitudes with 180 DCT bins. By appending 180 additional zeros to the end of the DCT vector and taking the 360 point IDCT one obtains the original weight vector with an additional 180 samples interleaved with the original 180 samples, this total of 360 samples characterises an interpolation by a factor of 2.

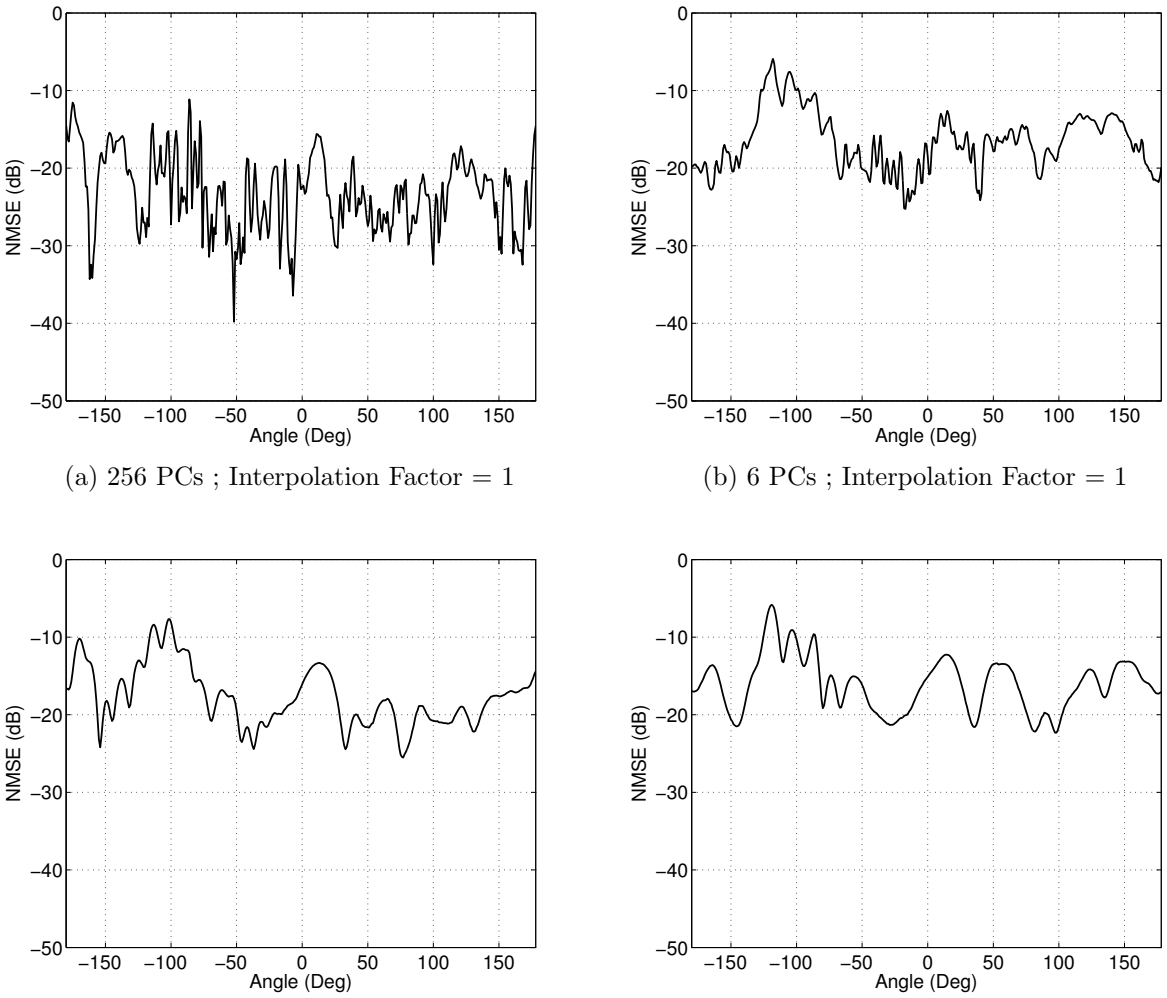
Interpolation via this method results in a reduction of amplitude in the reconstructed weight vector; the reduction is equal to the square root of the interpolation factor. In the example mentioned above, the interpolation factor of 2 results a reconstructed amplitude reduction of a factor of $\sqrt{2}$, in order to correct for the reduction in y axis displacement of the reconstructed weight vector, the values of either the weight vector or the DCT component vector must be compensated by multiplying by the reduction factor.

Thus the PCA weight vectors derived from an HRTF measurement set of an arbitrary number of evenly spaced positions can be zero padded in the reciprocal DCT domain in order to oversample the measurement scheme spatially and obtain weight vectors of any desired length, ultimately allowing for the PCA reconstruction of HRTF data at positions for which no measurement was taken. This of course has intriguing implications regarding the possible applications of convenient interpolation such as upscaling of small measurement sets in order to achieve more accurate reproduction of a virtual auditory space, but also possibly for the implementation of efficient HRTF measurement schemes that could be used to greatly reduce the time taken to capture personalised HRTF data which would result in better externalisation for specific listeners than measurements captured using a dummy head or non-personalised head model.

4.3.4 Interpolation Performance

The performance of the interpolation method can be reported meaningfully by comparing the reconstructed spectra with that of the original measured data in terms of the MSE, in the same way as was used to investigate the uninterpolated data reconstruction using a minimal number of principal components in reconstruction. In order to demonstrate the performance of the proposed interpolation method the measured data is resampled with a

decreased sample rate, such that the resampled data has less measurements positions than the original TU Berlin dataset but retains a constant measurement density in the azimuthal plane. The principal component analysis is conducted using the log magnitude spectra of the resampled dataset, and the derived weight vectors are deconstructed into cosine weights use the DCT as described earlier in this section of the thesis. Pending the IDCT operation to reconstruct the simplified weight vectors, the DCT component vectors are zero padded such that the output of the IDCT operation are reconstructed weight vectors of length 360 elements, corresponding to the 360 measurement points of the original TU Berlin dataset. Using the interpolated weight vectors, the original data can be reconstructed for any number of principal components at an interpolated number of measurement positions. Thus by calculating the MSE between the Tu Berlin measured spectra and the reconstructed spectra, interpolated from the smaller subset of measurements, at each of the original measurement angles an overview of the interpolation performance of the method can be attained.



(a) 256 PCs ; Interpolation Factor = 1

(b) 6 PCs ; Interpolation Factor = 1

(c) 256 PCs ; Interpolation Factor = 10

(d) 6 PCs ; Interpolation Factor = 10

Figure 4.3.5: DCT interpolation performance in MSE

Figure 4.3.5 illustrates the variation in the NMSE between the measured and reconstructed spectra for full (256) and partial (6) PC reconstructions both with and without interpolation. For the purpose of demonstration an interpolation factor of 10 has been chosen, this corresponds to a resampled subset of 36 measurement positions with a inter-measurement spacing of 10°.

Figures 4.3.5c and 4.3.5d show the NMSE of the same 256 and 6 principal component recon-

structions as in figures 4.3.5a and 4.3.5b respectively, however in figures 4.3.5c and 4.3.5d the HRTF dataset has been downsampled by a factor of 10 and then re-interpolated back to the original spatial sample density.

Figure 4.3.5a shows the inherent NMSE introduced to the system by using the DCT based decomposition and reconstruction of the individual weight vectors; as only 99% of the variance of each weight vector is retained upon reconstruction some small amount of error is introduced, even if all 256 principal components are used to reconstruct the log-magnitude spectra.

Figure 4.3.5b shows the NMSE of the data reconstructed using 6 PCs and the DCT reconstructed weight vectors with a resampling ratio of 1:1. The MSE distribution is very similar to that of the 6 PC reconstruction using the unaltered weight vectors shown in figure 4.2.4, performing only fractionally worse at some angles due to the introduction of the small amount of error associated with the use of the DCT reconstructed weight vectors.

Figure 4.3.5c shows the NMSE of the data reconstructed using all 256 principal components, and weight vectors that have been interpolated from 36 to 360 measurement positions using the proposed DCT based interpolation method. The figure shows that the interpolation of the weight vectors by a factor of ten incurs an increase in the NMSE, peak values are increased by between approximately 2-5dB and the whole distribution is seemingly compressed and also somewhat smoothed, such that the ratio between the local minima maxima of the curve has reduced significantly. For some local minima, such as that at approximately -50° in figure 4.3.5a, the interpolation results in an increase in NMSE of ∼20dB, which is certainly significant, however the NMSE distribution for the interpolated results seen in 4.3.5c may well still be within the boundaries of an acceptable HRTF compression system. The curve still retains most of the key features of the uninterpolated performance curve, specifically the increase in NMSE at the contralateral positions, in fact the difference in NMSE between the contralateral and ipsilateral positional ranges is somewhat more pronounced in the interpolated results.

It is interesting to observe the change in the NMSE distribution for the same interpolated

data following the reduction of the number of principal components used in the reconstruction of the HRTF data from 256 to 6. Figure 4.3.5d shows this NMSE distribution, note that although the NMSE has increased by a small amount on the order of 1-2dB for almost all angles, this increase is not as large as may be expected. It is not unreasonable to expect to find that the MSE error distribution for this case would be similar in magnitude to the logarithmic sum of the distributions for the 6 PC - Interpolation Factor 1 (4.3.5b) and the 256 PC Interpolation - Factor 10 (4.3.5c) reconstructions. However it is apparent that the MSE distribution for the 6 PC - Interpolation Factor 10 (4.3.5d) reconstruction exhibits only a slight increase in NMSE evenly across all angles, most noticeably in the local minima, than shown for the 6 PC - Interpolation Factor 1 reconstruction shown in figure 4.3.5b. The maximum occurs at the same position as that of the uninterpolated 6 PC reconstruction shown in 4.3.5b.

It is significant to find that a resampled measurement set, with only one tenth as many measurement points as the original set, can be used to derive a reconstructed dataset with as little increase to the normalised mean square error above the baseline distribution imposed by the reconstruction using only 6 principal components. This finding is most likely explained by considering the difference in the amount of total variance between the measured and downsampled datasets; as the spatially downsampled dataset contains only 36 measured spectra it is expected that the analysis of the frequency spectra and their variation over measurement angle will yield less overall variation simply due to the lack of more subtle changes that occur gradually over small changes in angle. However, given that the data is subject to a principal component analysis which extracts the basis vectors or spectral shapes in descending order of importance, the most crucial variations in the spectra will be captured within the first principal components, these first components are likely to be very similar to if not identical to the first components that are derived from the untampered dataset with full spatial resolution. As both the reconstruction of the uninterpolated full dataset and interpolated downsampled dataset are both limited to use only the first 6 principal components there is very little increase in reconstruction error due to interpolation; the finer spectral details that perhaps pertain to specific angular subsets or regions are captured and contained in higher order principal components and thus the missing detail in the interpolated dataset does not become apparent until a greater number of principal components are used

in reconstruction, this is evidenced by the difference in the NMSE between the full dataset and downsampled/interpolated dataset reconsrcutions using all 256 principal components shown in figure 4.3.5a and figure 4.3.5c respectively.

Chapter 5

Parametric Modelling Approach

This section of the thesis details the application of parametric modelling techniques in a bid to compress the TU Berlin azimuthal HRIR dataset. The section begins with the application of a linear prediction model which is used to generate all-pole variants of the measured HRTFs, this is followed by an advancement to the use of the Steiglitz-McBride iteration, a useful system identification technique, in order to model the measured HRTFs as pole-zero filters of varying order.

5.1 Linear Prediction

Linear predictive coding, often abbreviated to LPC, is a signal processing technique that can be used to express a frequency spectrum in a compressed from. Developed as a means of compression in speech signal processing applications, LPC techniques are derived from an approximate model of the human vocal system.

The technique begins with the basic assumption that the discrete time signal S_n is considered to be the output of a system with unknown output U_n such that:

$$S_n = -\sum_{k=1}^{p} a_k S_{n-k} + G \sum_{l=0}^{q} b_l U_{n-l} \quad , \qquad b_0 = 1 \tag{5.1.1}$$

where a_k,p,b_l,q, and G are the as yet undefined system parameters. Note that this is also the difference equation of an IIR filter, and as such states that the signal S_n is a linear

function of previous input and output samples. This implies that the system output is in fact predictable by the linear combination of previous input and output samples.

The model described by equation 5.1.1 can take three forms depending on the values of the a and b coefficients. If all a coefficients of the model are set to zero then the modelled system will have only zeros at non-zero locations, in the fields of statistics and economics this is referred to as a moving average (MA) model. Conversely, if all the b coefficients are set to zero then the modelled system has only poles at non-zero locations, this is known as an autoregressive (AR) model. The third form has both poles and zeros at non-zero locations, and is known as an autoregressive moving-average (ARMA) model.

For the purposes of LPC the vocal system is modelled simply as a buzzing excitation source representing the vocal folds, at one end of a variable diameter tube, in turn representing the vocal tract. In this model speech is assumed to be an autoregressive process, i.e. an all-pole model representing a series of acoustic resonances formed by the physical cavities and pipes that occur in the vocal tract and mouth.

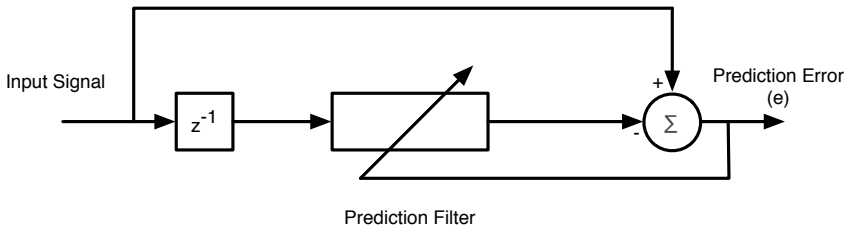


Figure 5.1.1: LPC diagram

Mathematically LPC seeks to minimise the mean squared error of the linear prediction filter with respect to the coefficients of that filter. Theoretically the mean squared error should be calculated as a continuous integral over all time (from $-\infty$ to $+\infty$), however in practice, and in the digital domain, it can be approximated for any given time t from a finite range of samples spaced symmetrically around the time of interest.

Therefore MSE can be approximated as:

$$E_t \approx \frac{1}{n}\sum_{n} e_t^2(n) \tag{5.1.2}$$

Where E_t is the MSE at time t, n is the set of samples around time t, and e_t is the prediction error at time t.

If the predicted signal is a weighted sum of delayed samples of the input signal then it can be written as:

$$Y_t = \sum_{k=1}^{M} a_k S_t(n-k) \tag{5.1.3}$$

Where Y_t is the prediction at time t, a_k is one of M weighting coefficients, and $S_t(n-k)$ is the input signal delayed by k samples.

This can be rewritten in a more compact vector form as:

either

$$Y_t = [a_1\ a_2\ ...\ a_L] \begin{pmatrix} S_t(n-1) \\ S_t(n-2) \\ \vdots \\ S_t(n-L) \end{pmatrix} = a^T S \tag{5.1.4}$$

or

$$Y_t = [S_t(n-1)\ S_t(n-2)\ ...\ S_t(n-L)] \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_L \end{pmatrix} = S^T a \tag{5.1.5}$$

The mean squared error can therefore be rewritten as:

$$e_t^2 = (S_t(n) - a^T S)^2 \tag{5.1.6}$$

Expanding brackets gives:

$$e_t^2 = a^T S\, S^T a - 2S_t(n)S^T a + S_t^2(n) \tag{5.1.7}$$

Attempting to simplify further:

$$SS^T = \begin{pmatrix} S_t(n-1) \\ S_t(n-2) \\ \vdots \\ S_t(n-L) \end{pmatrix} [S_t(n-1)\ S_t(n-2)\ ...\ S_t(n-L)] \tag{5.1.8}$$

Which can also be written as:

$$SS^T = \begin{pmatrix} S_t(n-1)S_t(n-1) & S_t(n-1)S_t(n-2) & ... & S_t(n-1)S_t(n-L) \\ S_t(n-2)S_t(n-1) & S_t(n-2)S_t(n-2) & ... & S_t(n-2)S_t(n-L) \\ \vdots & \vdots & \ddots & \vdots \\ S_t(n-L)S_t(n-1) & S_t(n-L)S_t(n-2) & ... & S_t(n-L)S_t(n-L) \end{pmatrix} \tag{5.1.9}$$

and:

$$S_t(n)S = S_t(n) \begin{pmatrix} S_t(n-1) \\ S_t(n-2) \\ \vdots \\ S_t(n-L) \end{pmatrix} = \begin{pmatrix} S_t(n)S_t(n-1) \\ S_t(n)S_t(n-2) \\ \vdots \\ S_t(n)S_t(n-L) \end{pmatrix} \tag{5.1.10}$$

Due to the fact that the mean squared error is calculated as a sum over n samples it is useful to define the following matrix:

$$R = \frac{1}{n}\sum_n SS^T = \frac{1}{n} \begin{pmatrix} \sum_n S_t(n-1)S_t(n-1) & \sum_n S_t(n-1)S_t(n-2) & ... & \sum_n S_t(n-1)S_t(n-L) \\ \sum_n S_t(n-2)S_t(n-1) & \sum_n S_t(n-2)S_t(n-2) & ... & \sum_n S_t(n-2)S_t(n-L) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_n S_t(n-L)S_t(n-1) & \sum_n S_t(n-L)S_t(n-2) & ... & \sum_n S_t(n-L)S_t(n-L) \end{pmatrix} \tag{5.1.11}$$

And the following vector:

$$P = \frac{1}{n}\sum_n S_t(n)S = \frac{1}{n}\begin{pmatrix} \sum_n S_t(n)S_t(n-1) \\ \sum_n S_t(n)S_t(n-2) \\ \vdots \\ \sum_n S_t(n)S_t(n-L) \end{pmatrix} \tag{5.1.12}$$

MSE can therefore be expressed as:

$$E_t = a^T Ra - 2P^T a + \frac{1}{n}\sum_n S_t^2(n) \tag{5.1.13}$$

Equation 5.1.13 shows that the mean squared error of the prediction filter can be expressed as a quadratic in a. Therefore it is shown that there exists a single global minimum on the MSE curve which represents the optimum set of filter coefficients for the system.

This lack of local minima and single optimum set of filter coefficients are an obvious strength of the linear predictive coding technique and are likely largely responsible for its wide use in speech signal processing applications.

The coefficients can be calculated by equating the derivative of equation 5.1.13 to zero and rearranging to find:

$$Ra = P \tag{5.1.14}$$

Equation 5.1.14 presents a set of linear simultaneous equations that can be solved using a range of conventional methods for solving such systems.

5.1.1 Levinson-Durbin Recursion

The Levinson-Durbin Recursion is an algorithm that is commonly used to form a solution to equations in the form of equation 5.1.14. Written more generally the Levinson-Durbin Recursion solves systems in the form:

$$\overrightarrow{y} = M\overrightarrow{x} \tag{5.1.15}$$

Where $\overrightarrow{y}$ is a known vector, $\overrightarrow{x}$ is an unknown vector to be calculated, and M is a known matrix in Toeplitz form.

The algorithm manipulates the inherent symmetry of the Toeplitz matrix to reduce the order of calculation required to solve the system from n^3 to n^2.

5.1.2 Implementation

The linear prediction model of each HRTF is generated using MATLAB function *levinson.m* to solve the system of linear equations associated with the autocorrelation input method. This is implemented by first using *xcorr.m* to find the autocorrelation sequence of the measured HRIR that is to be modelled, then removing the first portion of the autocorrelation function that pertains to the negative lags, this prepared autocorrelation function of the HRIR is then passed to *levinson.m* along with the desired order of the model. The *levinson.m* function converts the input autocorrelation sequence to symmetric Toeplitz form and returns the a coefficients that satisfy the associated system of linear equations described in equation 5.1.14.

5.1.3 All Pole Model Performance

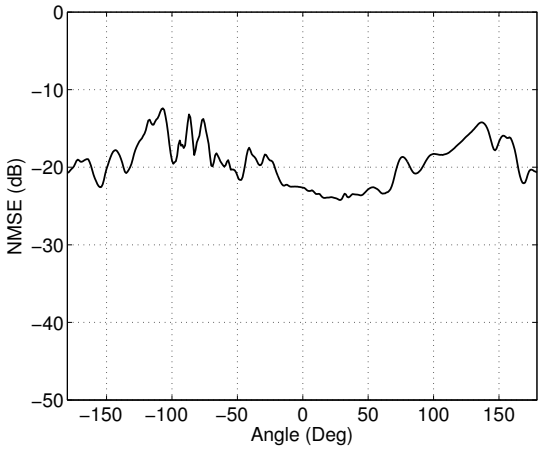
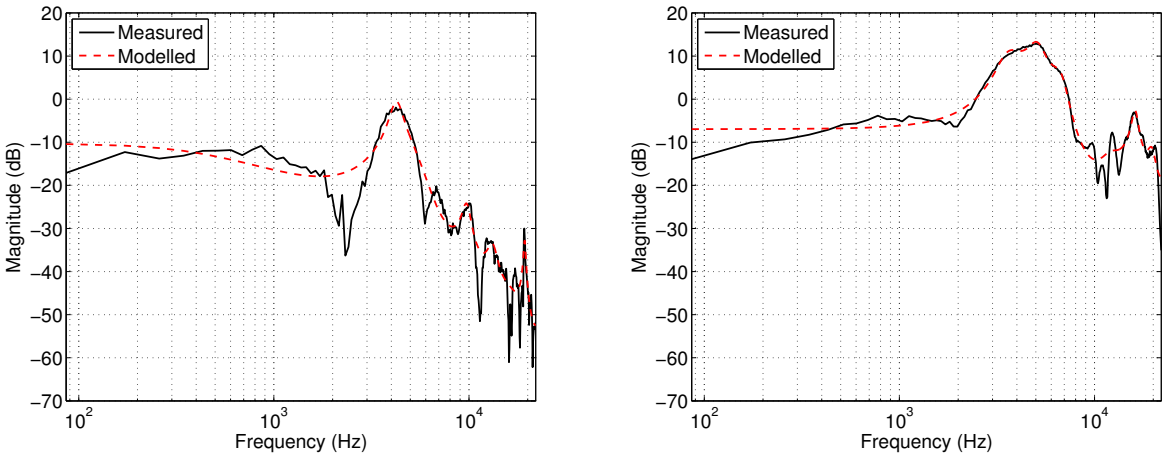


Figure 5.1.2: NMSE of all-pole model order 15

Figure 5.1.2 shows the normalised mean squared error calculated between the measured spectra and the spectra modelled as a 15^{th} order All Pole filter using the linear prediction method. The figure shows a range of a little over 15dB between the angles at which maximum and minimum NMSE occurs. The worst performance of the model occurs in the contralateral hemisphere, with the maximum error arising at -107°, however the model exhibits a similar level of error in the angles close to the ipsilateral position, and as such exhibits an approximate symmetry that is centred not around 0° but seemingly centred around $\sim 25°$. A fairly flat region of minimum NMSE exists between approximately 0° and 50°, covering the centre of the apparent symmetry in the error.

The distribution of the NMSE over angle, shown in figure 5.1.2 appears to be a direct characterisation of the inherent weakness in the all-pole approximation of the HRTF. Notches are a significant feature of the HRTF, at both ipsilateral and contralateral angles sharp notches arise due to destructive interference caused by pinna reflections, as the all-pole filter has no non-zero zeros it is unable to capture the notches in the frequency response of the HRTF.



(a) Maximum NMSE Case: -107° (b) Minimum NMSE Case: 29°

Figure 5.1.3: Maximum and minimum error HRTFs : All-pole order 15

Figure 5.1.3 illustrates the measured and modelled HRTF at the two angles which yield the maximum and minimum NMSE respectively when using a 15^{th} order all-pole filter. The minimum NMSE case occurs at 29°, in the frontal region of the measurement circle, whereas the maximum NMSE case occurs at -107°, towards the rear of the head in the contralateral hemisphere. The increased error of the modelled spectra observed at the maximum NMSE case shown in figure 5.1.3a is clearly a result of the significant notches present in the measured HRTF at this angle, such notches are not present in the frontal HRTFs like the one shown in figure 5.1.3b. For relatively low orders of all-pole filter, the model is unable to capture the spectral notches in a given frequency response and as such is incapable of reproducing the measured spectra with sufficient detail to maintain a consistent level of model error. However it is possible that given a sufficient, comparatively high, number of poles to be allocated, the linear prediction method will seek to approximate notches in the frequency spectra using clustered pole placements.

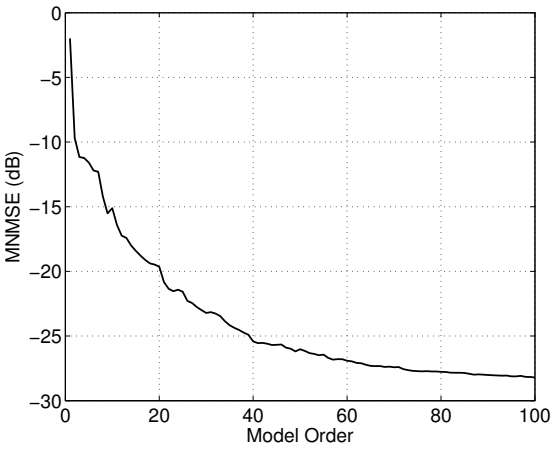
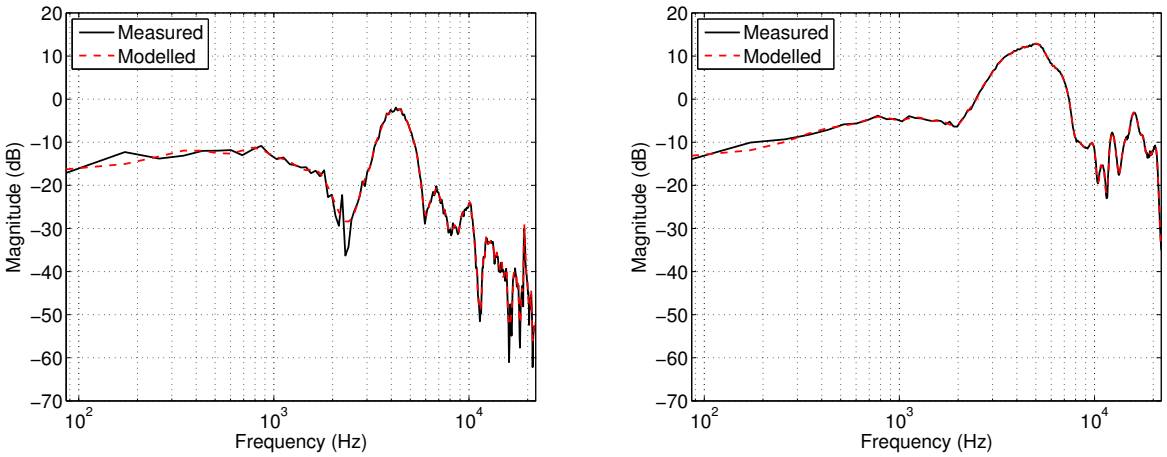


Figure 5.1.4: MNMSE of all-pole model orders 1 to 100

Figure 5.1.4 shows the mean normalised mean squared error of the all-pole linear prediction model as a function of the model order; that is the NMSE of each model order averaged over all angles to obtain a single value per model order expressed in decibels. The data was generated by analysing the performance of a series of linear prediction all-pole filters used to model the Tu Berlin measured HRTF at each angle for all-pole model orders ranging from 1 to 100. It is clear that with the addition of poles the model is better able to capture the spectral detail of the measured HRTF, and based upon the already well performing fit of the peaks in the spectra shown for relatively low orders in figure 5.1.3, it can be reasoned that the increased performance is a result of the improved ability to capture detail surrounding notches that comes with a surplus of available poles. This can be seen directly by observing the increased accuracy of the modelled spectra for the same angles that yield the minimum and maximum NMSE for the order 15 all-pole case.



(a) Prior Maximum NMSE: -107°

(b) Prior Minimum NMSE: 29°

Figure 5.1.5: Prior maximum and minimum error HRTFs of all-pole order 15 modelled with all-pole order 100

Figure 5.1.5 shows the improved performance of the 100^{th} order all-pole model for the angles which correspond to the maximum and minimum NMSE in the 15^{th} order case respectively. Note that the order 100 model still performs worse for the more notched frequency response shown in 5.1.5a, however for both angles it can be seen that the increased number of poles allows the linear prediction method to better approximate the frequency notches.

The overall trend in figure 5.1.4 appears to be similar to that of an exponential decay; the MNMSE reduces rapidly as the number of poles increases from 1 to approximately 7, this is followed by a small region of noticeably shallower gradient, at approximately model order 11 the roughly exponential shape resumes. The curve exhibits slight fluctuations deviating from the ideal exponential curve at lower model order, but exhibits increasing smoothness approaching the higher model orders in the plot. These fluctuations are likely caused at lower model orders due to the number of poles available being less than optimal for the modelling of specific measured HRTF spectra, as such the linear prediction method may focus all available poles on significant peak details, possibly leaving too few poles to capture other lesser peaks, differently depending on the relative optimality of the number of poles used to each of the measured HRTFs. The decrease in NMSE continues with a progressively

shallower gradient as the model order is increased, between model order 80 and 100 the curve exhibits a near smoothness and very shallow negative gradient, seemingly suggesting that the NMNSE is approaching convergence with a lower limit between -25dB and -30dB. If the trend continues beyond model order 100 it is likely that the exponential decay characteristic of the curve will yield an asymptote at approximately -26dB, meaning that the addition of poles after approximately 80 will yield little to no further increase in performance according to the MNMSE criterion.

5.2 K Coefficients

K Coefficients, sometimes referred to as reflection or lattice coefficients are the associated weights of the lattice filter structure. The lattice filter structure is of modular design such that increasing the filter order is achieved simply by adding one extra module to the filter with no changes needing to be made to the existing modules.

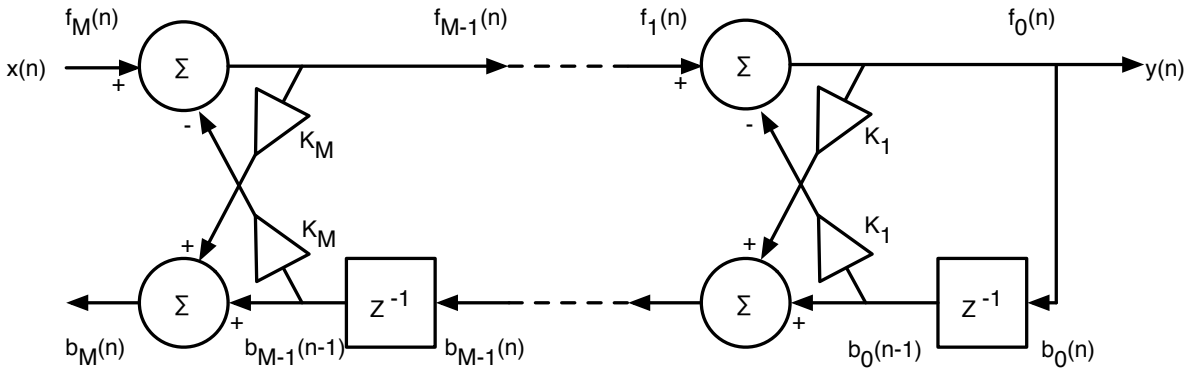


Figure 5.2.1: Lattice filter structure

Figure 5.2.1 shows the general form of an IIR all-pole recursive filter in lattice form, where the dashed lines indicate the separation between modules.

A convenient property of the K coefficient representation of the modelled HRTFs is that they offer a hierarchical representation quality of reconstruction as a function of model order. As the first module of the lattice filter structure pertains to the most significant spectral characteristic, the second module pertains to the second and so on. Thus a single model with

an upper limit of accuracy dictated by the order specified at calculation can be created and adapted for reduced quality simply by reducing the number of K coefficients included in the lattice filter structure at run time.

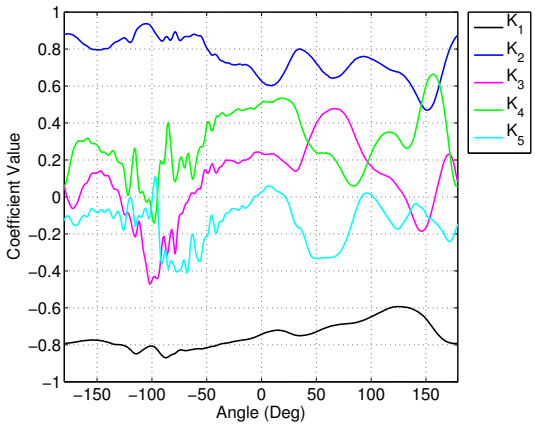


Figure 5.2.2: First 5 K coefficients variation with angle

Figure 5.2.2 shows the variation of the first 5 K coefficients of the 15 pole all-pole model derived using the linear prediction method with respect to angle. The first K coefficient appears to vary the most smoothly with respect to angle, as the K coefficients are an orthogonal series it is possible that this behaviour signifies that the first K coefficient pertains to low order physical phenomenon as a general emphasis or de-emphasis of high frequency content in the measured spectra. The higher order coefficients are less directly interpretable due to their more complex variations, they do however exhibit notable characteristics; coefficients K_{2-5} all show 'noisier' behaviour in the contralateral hemisphere, presumably once again due to the significantly lower signal to noise ratio of the measurements at these positions. Coefficients K_{3-5} are low valued approaching $\pm 180°$, suggesting that they pertain to features that are less prevalent in the spectra measured for rear positions.

Observing that the K coefficient variation with respect to angle is largely continuous leads to the assumption that, in combination with their orthogonal properties, the K coefficients may provide a convenient platform for interpolation between measured angles. The 'one-dimensional' principal component analysis of each of the K coefficient vectors yields an

further orthogonal series of sinusoidal/cosinusoidal basis vectors; suggesting that the DCT based compression and interpolation of these angular dependent K coefficient vectors, as described in detail in Section 4.3.3 of this thesis, may yield promising results.

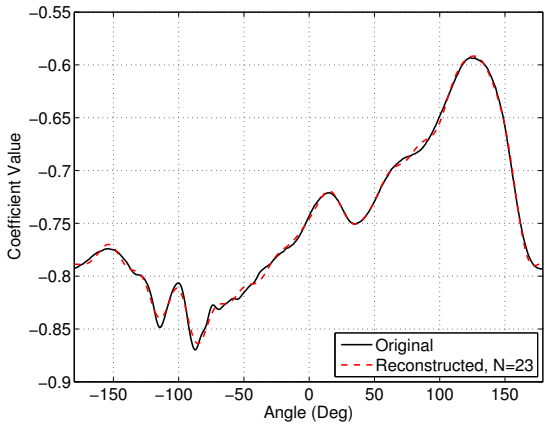


Figure 5.2.3: DCT reconstruction of first K coefficient across measured angles

Figure 5.2.3 shows DCT approximation of the first K coefficient vector; 23 DCT components are required to obtain a representation of the K coefficient vector that retains 99.999% of the total variance of the untampered vector. A larger value than 99% (0.99) is required in the calculation of how many DCT components are to be retained in the reconstruction of the vector. This is because the magnitude of the first significant component is far larger than that of the succeeding components. As the first (and second) K coefficient vector is non-zero mean, or not approaching zero mean, the first DCT component that accounts for the greatest amount of the variance is that which corresponds to a cosine function oscillating at 0Hz, i.e. the DC offset. As the DC offset component accounts for such a large amount of the variance in the K coefficient vector, the DCT retention calculation with a threshold value of only 99% returns only the DC offset component, which is obviously not very useful in preserving the angular variation of the first (and second) K coefficients. Hence a higher threshold value, in this case 99.999%, is chosen to ensure that a sufficient amount of non-constant DCT components are retained.

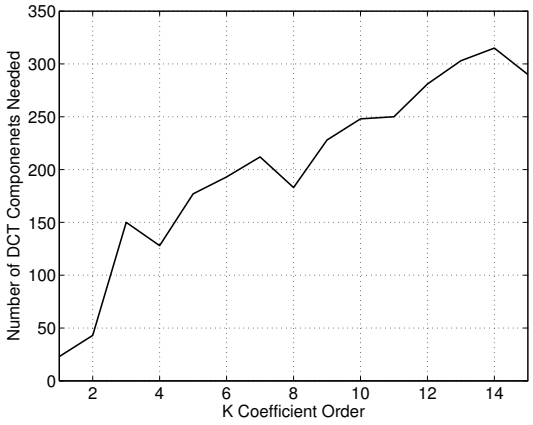
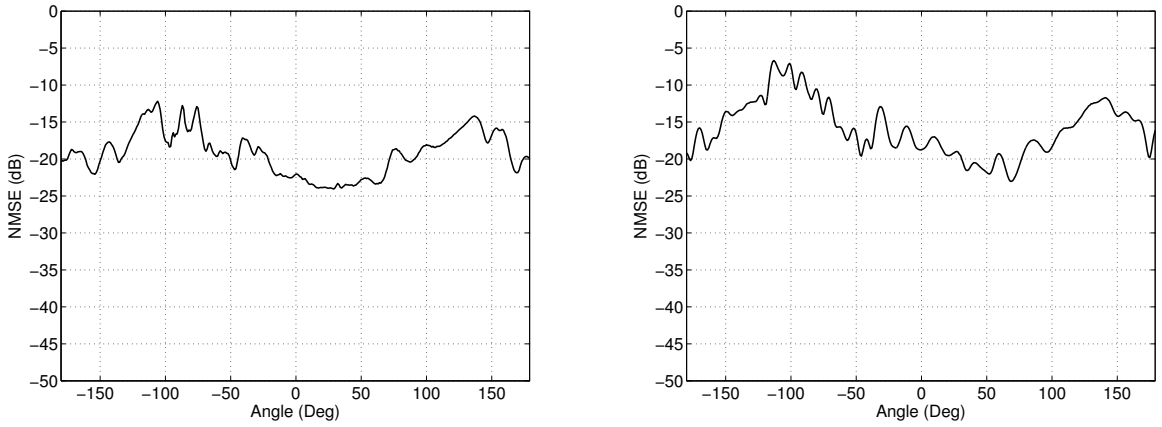


Figure 5.2.4: Number of DCT components vs K coefficient order

Figure 5.2.4 illustrates the number of DCT components required to retain 99.999% of the variance of each of the 15 K coefficient vectors describing angular variation. The figure shows a reasonably linear relationship between the order of the K coefficient vector and the number of DCT components required; the minimum occurs for the first order K coefficient vector for which 26 DCT components are needed, and the maximum occurs for the penultimate (14^{th}) order vector for which 315 DCT components are needed. Though the higher order K coefficients seemingly require comparatively large numbers of DCT components it is worth noting that any number of DCT components below the number of points in the original array, 360, represents an effective compression in the number of elements that must be stored to use the untampered K coefficient vectors, which themselves represent a greatly compressed representation of the original dataset from 2048 to, in this section, 15 elements per single HRTF spectra.

5.2.1 Interpolation Performance



(a) Interpolation Factor: 1

(b) Interpolation Factor: 10

Figure 5.2.5: NMSE performance of order 15 all-pole model using DCT approximation of K coefficents

Figure 5.2.5 shows the NMSE performance of the K coefficient representation of the 15 pole all-pole model derived from both the full 360 measurement dataset and a downsampled subset of 36 measurements at 10° spacing. Both models make use of the DCT decomposition and reconstruction of the angular K coefficient vectors, with an interpolation factor of 1 and 10 in sub-figures 5.2.5a and 5.2.5b respectively. Figure 5.2.5a shows an almost identical NMSE distribution over angle as seen in the direct 15 pole all-pole implementation show in figure 5.1.2, which is unsurprising given the high threshold used in the calculation of the number of DCT components to be retained to reconstruct the K coefficient vectors. The NMSE distribution shown in figure 5.2.5b shows a marked increase in error across all angles; this increase is of largest magnitude in the contralateral region where a peak increase of approximately 5dB can be seen at -115° to -120°. The increased error in the contralateral region seems likely to be a symptom of the increased 'microscopic' variation visible in the contralateral regions of the K coefficient vectors, which are mis-predicted as a consequence of the reduction in the number of DCT components used to reconstruct said vectors.

5.3 Steiglitz-McBride Iteration

The Steiglitz-McBride Iteration is a technique useful for the identification of linear systems using known samples of the system's input and output. The technique models the system as an ARMA process (Pole-Zero filter) and performs an iterative method to approximate the system coefficients.

The technique was derived first considering a common simple linear problem; assuming the input and output records are related by a rational z-transform $\frac{N(z)}{D(z)}$.

Where

$$N(z) = \alpha_0 + \alpha_1 z^{-1} + ... + \alpha_{n-1} z^{-(n-1)} \tag{5.3.1}$$

and

$$D(z) = 1 + \beta_1 z^{-1} + ... + \beta_n z^{-n} \tag{5.3.2}$$

in which α and β are the n numerator and denominator coefficients, and z is the z-transform variable.

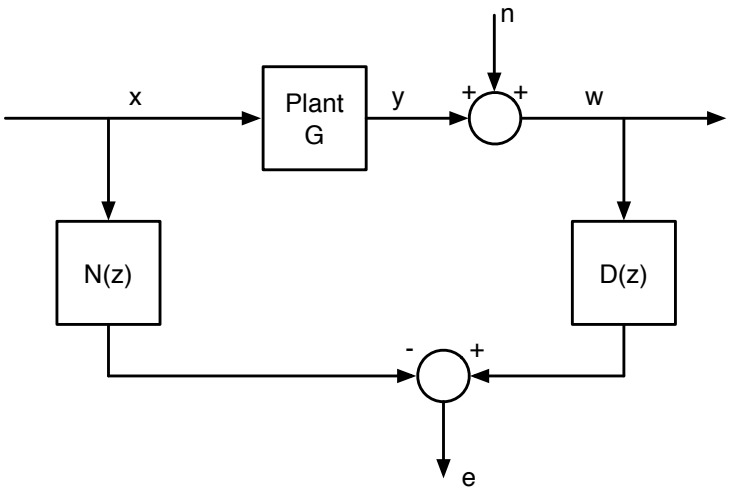


Figure 5.3.1: Simple linear problem

Given the available input and output records, x and w respectively, figure 5.3.1 leads to the

following minimisation task:

$$\sum e_j^2 = \frac{1}{2\pi j} \oint |XN - WD|^2 \frac{dz}{z} = min \tag{5.3.3}$$

Where: $X = X(z) = \sum x_j z^{-j}$; $W = W(z) = \sum w_j z^{-j}$; summations are carried out over record length; and the contour of integration is the unit circle.

For such a minimisation task it can be shown that the solution is

$$\delta = Q^{-1} c \tag{5.3.4}$$

Where

$$\delta = \begin{bmatrix} \alpha \\ -\beta \end{bmatrix} \tag{5.3.5}$$

is the coefficient vector, and Q and c are the appropriate correlation matrix and vector computed from the records of x and w.

However, although this case is easily solved it is not of any particular interest; the error residual does not pertain to a real physical property of the system.

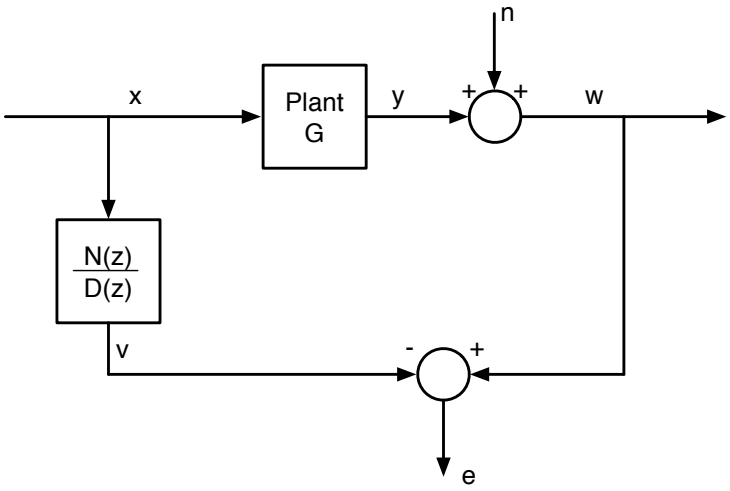


Figure 5.3.2: Complex non-linear problem

Figure 5.3.2 illustrates a more useful system definition, the solution to which is far more meaningful as the error residual is that of the error between the predicted and observed plant outputs. The mean squared error residual minimisation for this system can be written as

$$\sum e_j^2 = \frac{1}{2\pi j} \oint |X\frac{N}{D} - W|^2 \frac{dz}{z} = min \tag{5.3.6}$$

This minimisation represents a complex and highly non-linear regression problem for which the Steiglitz-McBride Iteration technique defines an iterative process of pre-filtering of the input and output records to reduce the problem in complexity to that of the simple case seen in figure 5.3.1.

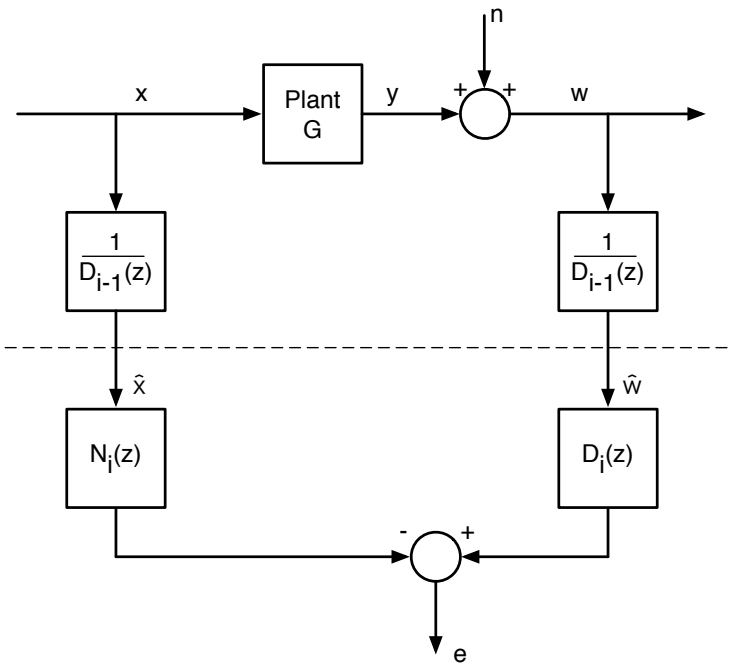


Figure 5.3.3: Iterative method system

First the original minimisation problem pertaining to the simple case shown in figure 5.3.3 is solved using equation 5.3.4 to obtain an initial estimate of $N(z)$ and $D(z)$. This is known as the Kalman estimate, as it was Kalman who first suggested the application of the input output record linear regression analysis that leads to the associated minimisation task [Kalman,

1958]. The estimate of $D(z)$ is then used to pre-filter the input and output records, the pre-filtered input and output records are then used to define the vectors in equation 5.3.4 and the minimisation problem is solved again. The second estimate of the system denominator $D(z)$ is then used to pre filter the original input and output records again and the process is repeated for i iterations. If the denominator coefficients converge as i becomes large then the error of figure 5.3.3 becomes equal to the error of figure 5.3.2 and hence the complex non-linear regression problem has been approximated.

In MATLAB the *stmcb.m* function performs the Steiglitz-McBride iteration in order to obtain the a and b coefficients of a filter with an impulse response approximately equal to the input desired impulse response using as many poles and zeros, or a and b coefficients as specified [The MathWorks Inc., 2014]. If an initial estimate of the denominator coefficients is not given by the user at the time of calling *stmcb.m* then the function utilises *prony.m* to obtain the denominator coefficients for the first iteration of the Steiglitz-McBride process. *prony.m* is an implementation Prony's method, a technique in which an evenly sampled time domain signal is effectively decomposed into a sum of damped complex exponentials, adapted to the application of IIR filter design the method of which is described in detail by Parks & Burrus [Parks and Burrus, 1987].

5.3.1 Pole-Zero Model Performance

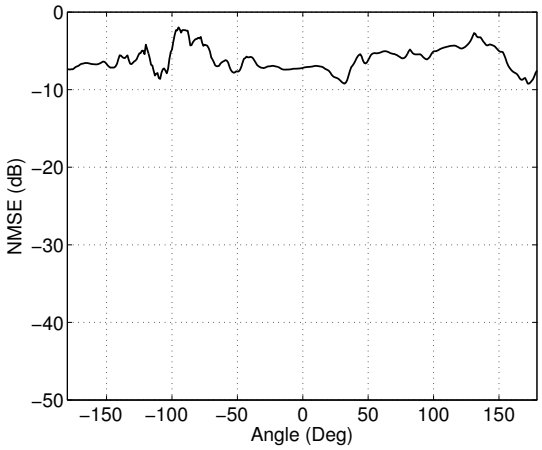
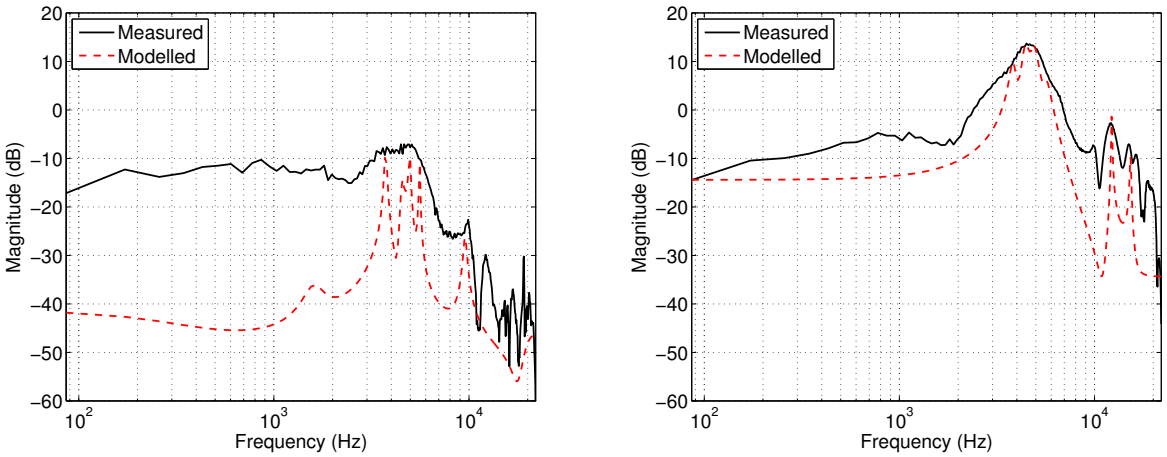


Figure 5.3.4: NMSE performance of 15 pole 15 zero model

Figure 5.3.4 shows the NMSE between the measured spectra and the counterpart spectra obtained using a 15 pole 15 zero model derived using the Steiglitz-McBride method, at all measured angles. The maximum NMSE occurs in close proximity to the contralateral ear at approximately -95°, travelling towards the back of the head from this angle the peak error region is followed by one of three global minima of value between -8dB and -10dB, the other two minima occur at $\sim 35°$ and $\sim 175°$ respectively. The range of NMSE between the minima and maxima is approximately $\sim$ 7dB suggesting that the method yields modelled HRTFs that perform quite consistently across all angles. There appears to be a slight increase in the NMSE of the modelled spectra for ipsilateral positions in the approximate range 50° to 100°.



(a) Maximum NMSE Case: -95°

(b) Minimum NMSE Case: 32°

Figure 5.3.5: Measured and modelled spectra at angles of worst and best performance

Figure 5.3.5 shows the measured and modelled HRTF spectra for the two angles that yield the maximum and minimum NMSE respectively for a 15 pole 15 zero filter approximation of the system. The maximum NMSE case occurs close to the contralateral ear at -95° whereas the case for which the model performs best is located at 32°, in the frontal region, off-axis in the direction of the ipsilateral ear. Figure 5.3.5a illustrates the worst performance case of the 15 pole 15 zero model, the figure exhibits two prominent characteristics that likely pertain to the primary sources of the increased error in the approximation; firstly the model underestimates the low frequency range of the HRTF by a significant amount, below approximately 3kHz the modelled magnitudes are consistently under predicted by as much as 30dB, and secondly the model attempts to approximate the wide peak in the measured response between approximately 3kHz and 7kHz as a series of poles of seemingly high quality factor, leading to an increased prediction error in the frequency regions immediately to either side of these pole locations as the consecutive narrow width peaks form an almost combed response in this region. With these two issues aside it can be seen that the modelled HRTF in figure 5.3.5a does manage to retain some of the key features of the measured HRTF, such as the secondary peak at 10kHz and the global minima defined by the high frequency notches between 10kHz and 20kHz.

Figure 5.3.5b shows the best performance case of the 15 pole 15 zero model, the modelled HRTF spectrum at this frontal angle is clearly a far superior approximation of the measured data than that of the contralateral worst case. The best case model still under predicts the low frequency components below approximately 2kHz to 3kHz, however the difference in magnitude is much smaller, at maximum approximately 10dB. The figure also shows that the model is superior in capturing the overall shape of the main peak region of the measured response located between approximately 2kHz and 10kHz, though the series of peaks in the modelled response still appear to be of a narrow bandwidth, the error in this region is reduced by the lack of such severe notches in-between the modelled peaks in close proximity to one another seen in the maximum NMSE case. The modelled response in figure 5.3.5b also accurately captures the location of the secondary and even tertiary peaks located in the high frequency region upwards of 10kHz, however the model error in this region will still be significant again due to the exceedingly narrow bandwidth of the modelled peaks. For both the maximum and minimum NMSE cases, the modelled spectra contain a similar error region, between approximately 7kHz and 11kHz for which the modelled response retains the correct shape of the measured frequency response but under predicts it's magnitude by ∼15-20dB.

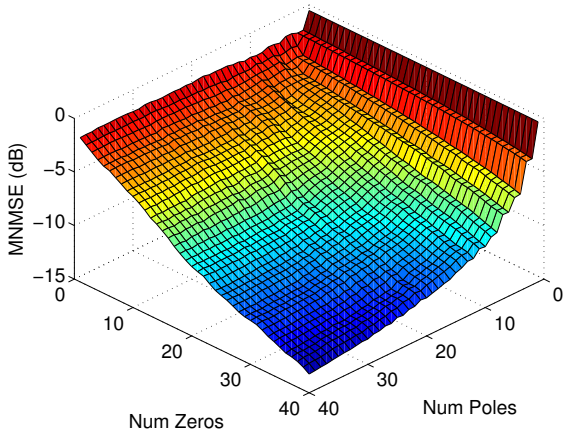


Figure 5.3.6: Mean normalised mean squared error

In order to observe the relationships between the number of poles and zeros in the Steiglitz-

McBride models and the performance of said models in terms of the error between the measured and modelled HRTF spectra, a simulation of the methods performance with all combinations of 1-40 poles and 1-40 zeros measured at all 360 angles was conducted. Figure 5.3.6 shows the error surface derived from the simulated results, where for each model containing nP poles and nZ zeros a single value of mean normalised mean squared error (MNMSE) has been calculated as the mean of the NMSE values computed over all 360 angles for which measured and modelled HRTF spectra available. The figure shows that for a model consisting of any number poles (up to 40) and a single zero the performance will improve little after the addition of the first 10 poles. Similarly, a single pole with any number of 1 to 40 zeros yields a constant high MNMSE value. The surface exhibits slight troughs of local minima running North-West to South-East, in parallel with the 'Num Zeros' axis, suggesting that certain numbers of poles are somewhat optimal, regardless of the number or zeros in the model, however these troughs or channels become less prevalent at higher numbers of zeros. For higher numbers of zeros above approximately 20, it can be seen that the reduction of poles yields a seemingly exponential increase in MNMSE and subsequent curvature of the error surface in the South-West to North-East direction. In contrast, for seemingly all numbers of poles the increase in number of zeros leads to an approximately linear decrease in the MNMSE, the gradient of which becomes steeper as the number of poles is also increased. Both of these characteristics lead the surface to a global minimum at the highest model order, which pertains to a model consisting of 40 poles and 40 zeros, and achieves a minimal value of approximately -14dB MNMSE.

Chapter 6

Pilot Study: Subjective Validation of Pole-Zero Models

Given the inherent ambiguity surrounding error criteria and the lack of a well developed psychoacoustically derived performance metric, the works described in this thesis were designed to include a subjective experiment in an aim to better understand the real implications of the HRTF compression performance described in a means by the normalised mean squared error metric. Ideally the experiment would asses the psychoacostic performance of all three of the methods applied to the Tu Berlin data; Principal Component Decomposition, Linear Prediction All-Pole Modelling, and Steiglitz-McBride, however the size of such an experiment designed to evaluate not only all three methods, but a range of model or reconstruction orders for each, would be too large to rest within the scope of the works conducted. Hence only a single method was selected as the basis for subjective experimentation as a pilot study. The pilot experiment was designed and run in parallel with the computer simulation and objective analysis of the three methods in the latter stages of the works, at the time of the design of the experiment the Steiglitz-McBride Pole-Zero Modelling approach was deemed the most promising in terms of retention of key spectral detail for the minimum number of stored elements. As such the experiment was designed to evaluate the performance of Steiglitz-McBride HRTF approximations of varying order.

The objective analysis of the three methods for the compression of measured HRIRs in the prior sections of this thesis show that each technique offers differing degrees of objective

success. However it is difficult to predict the performance of such compressed HRIR or HRTFs from objective evaluation alone, and there is no clear way to determine the influence of the amount of spectral detail retained or discarded by differing model or reconstruction orders on the psychoacoustic interpretation of the filters, therefore subjective validation of the modelled HRIRs or HRTFs is desired.

A subjective validation offers an additional arm of investigation into the plausibility of using the described methods to express measured HRIR or HRTFs in a more compact form. The methods yield varying reductions in the amount of data required to be stored as well as a convenient means of real time implementation, but it is unclear as to whether or not the methods are capable of retaining enough key components of the spectra, unique to each measured HRTF, in order to sufficiently preserve the ILD and pinnae filtering cues that allow the listener to localise a sound filtered by the HRTF as coming from a specific direction.

6.1 Experimental Design

The aim of the experiment is to asses the effect of the order of the IIR models used to approximate the Tu Berlin measured HRIR dataset, more specifically to asses the the effect of the model order on how well the subjective impression of source location is preserved. By testing a series of IIR HRTF models of decreasing order, one should be able to observe any deterioration in the subjective localisation as a result of this decrease.

The experiment in question was originally designed as a test of absolute positional judgement; participants were to be played a sound which had been filtered by a modelled or measured HRTF, then asked to report the apparent position of the sound on a circle of fixed distance around the listener's head in the azimuthal plane. The test was to be conducted using an automated graphical user interface constructed and run in MATLAB. In order to provide more reliable results and alleviate high precision errors in reported location the participant was asked to report the perceived location of each sound using a clock face paradigm; selecting two options from drop down menus to render a phrase in format 'D H', one to indicate the descriptor 'D', and the other to denote the hour 'H'. The descriptors available in the

list were "Just before", "Exactly", "Just after", and "Half past", the hours available are of course one to twelve, as such a rendered phrase might read "Exactly 10", or "Just before 3". Using this syntax to record the perceived locations effectively divides the continuous circle of possible angles around the head into 48 possible positions, each covering a 7.5° subset of angles. Localisation quantisation of the perceived angles in this was included as a measure to improve the reliability of results and reduce the amount of error introduced by the perhaps unnecessary precision invited by an open response simply in terms of an absolute angle; it should be noted that 7.5° pertains to approximately twice the perceptual resolution of the human ear in the azimuthal plane. As a part of the GUI the participant would receive visual feedback regarding the selected position in realtime; this took the form of a simple diagram of a clock face with a representation of the listener pictured at the centre, a single 'clock hand' was imposed over the diagram to illustrate the position currently described by the selected descriptor and hour variables in the GUI. The visual feedback was included to attempt to ensure that the participant fully understood the paradigm and also to aid in the visualisation of the source location such that a relatively accurate description could be made using the available inputs.

After preliminary implementation, the original experimental design was retired for fear that it relied too heavily on each participant's ability to judge the localisation of sound sources in absolute values. Instead, the experiment was redesigned using a simple A/B test configuration that asked participants to report answers in relative terms only.

As stated, the redesigned experiment took the form of a series of A/B comparison tests, for each test A and B represent a sound filtered with a measured HRTF and a sound filtered with a corresponding modelled HRTF in a random order. The experiment is conducted using an automated GUI constructed and operated in MATLAB, the interface consists of a pair of sliders, one horizontal and one vertical, a pair of buttons labelled 'A' and 'B', and a third button labelled 'Submit'. The participant is asked to move the two sliders, one to mark the position of sound B relative to sound A, as if the position of sound A always provided the anchor for the centre of the slider, and the second to describe the timbre of sound B with reference to sound A, again as if the timbre of sound A always provided the anchor for the centre of the second slider. The timbre slider is labelled at either end with a descriptor;

'Brighter' at the maximum vertical position and 'Darker' and the minimum vertical position. The participant is able to play samples A and B as many times as needed before choosing to submit the values of the sliders and proceed to the next test.

6.1.1 Subjects

A total of ten subjects participated in the experiment, the group consisted of eight males and two females. The age of the participants ranged from 20 to 50, though it should be noted only a single participant was over the age of 30. All ten participants claimed to have no known hearing impairments, however audiometric testing was not performed for the purposes of the experiment in question. 7 of the 10 participants can be considered to be trained listeners, having participated in numerous unrelated listening experiments previously.

6.2 Experimental Stimuli

The stimuli used in the experiment are filtered bursts of pink noise, pink noise was selected due to its inverse energy distribution over frequency which yields a constant energy per octave. Pink noise is common in acoustical measurement procedures, in particular those that are concerned with listening as the -3dB per octave downward slope and subsequent relatively greater proportion of energy at lower frequencies characteristic of pink noise is known to approach the way in which the human ear subjectively perceives sound [Everest and Pohlman, 2009].

Seven model cases were chosen for testing, five different order IIR models generated using the Steiglitz-McBride method, and two reference cases. The IIR model orders were selected as points of interest on the MNMSE error surface shown in figure 5.3.6, more specifically they were selected as points of interest along the diagonal intersection with the surface along which the number of poles and zeros is always equal. The decision was made to keep the number of poles and zeros equal for the models used in the test as the overall trend in figure 5.3.6 does show the normalised mean squared error to decrease proportionally with the increase in the total number of poles and zeros, albeit at different rates, and the inclusion

of each pole and zero pair added both to the denominator and numerator costs little or no more than the inclusion of just one pole or one zero to either the denominator or numerator respectively, computationally speaking. The five model orders chosen were 30, 22, 16, 10, and 4; the remaining two model cases consisted of a positive and negative reference intended to aid in the comparison of results. The positive reference case was the perfect model, i.e. the measured HRIR implemented as a full length FIR filter, whereas the negative reference case represents the worst possible model, i.e. no minimum phase component filter, simply an ITD introduced between the left and right audio channels.

Alongside the seven model cases, ten test positions were selected; five positions limited to within a single frontal quadrant of the azimuthal plane: 0°, 15°, 30°, 60°, and 90°. Each of the five positions is also mirrored in the front rear axis to obtain a total of ten measurement positions at ±0°, ±15°, ±30°, ±60°, and ±90°. Test positions were limited to the frontal hemisphere of the azimuthal plane as the resolution of the human ear performs optimally in this range, this limit was also imposed in order to keep the test length reasonably short, hopefully avoiding effects of participant fatigue, the mirror doubling of positions was influenced by the desire to obtain robust statistics without the need to repeat discrete test cases, and also influenced in part to reflect common positions that may be exploited in virtual implementation of a common loudspeaker formation.

Overall, seven model cases tested at ten positions yields a total of 70 A/B comparisons per participant. The 70 test samples were rendered prior to testing as stereo wav files, in the interest of maximum efficiency only left ear data is used to construct the samples, the right ear data is assumed to be the mirror image of the left ear data and is treated as such. For the IIR model cases the wav files were generated by using the appropriate HRTF model coefficients to filter the pink noise sample and generate the left and right stereo components at each test angle, following this the ITD is added to the lagging channel by appending the correct amount of sample zeros according the ITD curve extracted using the edge detection method in figure 3.2.1c. For the positive reference case, the appropriate measured HRIR pair was used as the b coefficients of an FIR filter to render the left and right channels of the filtered pink noise sample. Finally for the negative reference case the pink noise sample was simply rendered to left and right channels with the appropriate ITD added to the lagging

channel again as a series of zero value samples dictated by the extracted results in figure 3.2.1c.

Testing was conducted using a single laptop computer running a copy of MATLAB, a Focusrite Scarlett 2i2 USB Audio Interface, and a pair of Beyerdynamic DT770 Pro headphones. The frequency response of the audio interface is specified by the manufacturer as being maximally flat between 20Hz and 20kHz to within ±0.2dB and is considered to have no effect on the test audio that could affect the results. Testing was conducted in a nominally quiet environment, but due to time constraints the room used was not specially treated acoustically, however the closed back circumaural design of the DT770 headphones provide sufficient isolation, quoted as approximately 18dBA by the manufacturer [beyerdynamic GmbH & Co. KG, 2014], to assume that the background noise level in the testing room was low enough to have negligible effect on the test results.

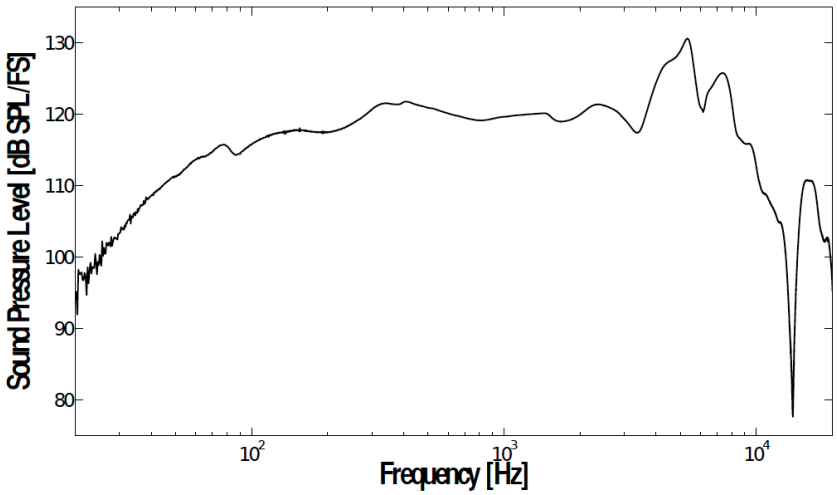


Figure 6.2.1: DT770 Pro frequency response [Man and Reiss, 2013]

Figure 6.2.1 illustrates the frequency response of the DT770 Pro headphone as measured using the KEMAR mannequin by Man & Reiss [Man and Reiss, 2013]. The response has been averaged over 3 left ear and 3 right ear measurements made using a swept sine excitation signal. The figure shows the response to be essentially flat between approximately

100Hz and 3kHz, with a primary and secondary peak at approximately 6kHz and 8.5kHz respectively. A significant notch occurs in the response at approximately 16kHz, this most likely corresponds to a cavity resonance that arises from the circumaural closed-back design of the headphones.

The pink noise sample used to excite each of the 70 test files was normalised to -18dBFS, measurements were performed using a B&K Head and Torso Simulator in order to calibrate the presentation level of the experimental stimuli. The processed wav files were not normalised as the adjustment of the relative level of the HRTF at different angles interferes with the broadband ILD cues that exist between, for example, frontal and rear positions, in which the ITD is ambiguous and the spectral differences and broadband ILD provide the dominate source of localisation cue. Thus instead, the A-weighted RMS level in dBFS of each of the processed wav files calculated, the wav file corresponding to the maximum was used to set the maximum A-weighted SPL at the ear. A 94dB SPL at 1kHz calibration tone was used as a reference such that the relationship between the internal dBFS level and the output SPL level of the system could be calculated. The master level of the USB audio interface was calibrated such that the maximum A-weighted RMS dBFS sample yielded a sound pressure level of 75dBA at each ear; a comfortable listening level within the bounds of the Lower Action Level of the Physical Agents Directives for Noise [European Parliament, 2003].

6.3 Experimental Methodology

The methodology of the experiment was quite simple given the use of an automated GUI. Each participant was first given an information sheet regarding the experiment that outlined the nature of the project and the experiment within which they were to participate, a copy of which can be found in Appendix B.1 of this report. Before beginning the test each participant was also asked to read and sign a formal declaration of consent to participate in the experiment and for the data collected to be used in the project works anonymously, again a copy of this consent form can be found in Appendix B.2 of this document.

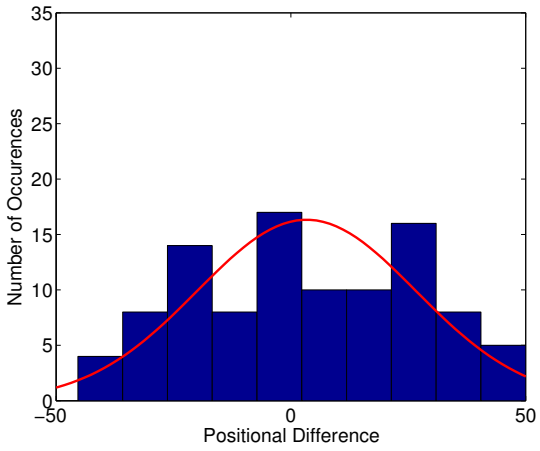
Following the signing of the consent document each participant was then instructed to put

on the headphones, adjusting the headband size to allow for a comfortable fit. For each participant, upon instantiation, the test GUI randomises both the order of the tests from 1 to 70 as well as the allocation of the reference and non-reference sample to the A and B stimuli slots for each of the 70 tests. The orders of all random permutations were recorded by the system such that the test results could be re-ordered consistently for data processing.

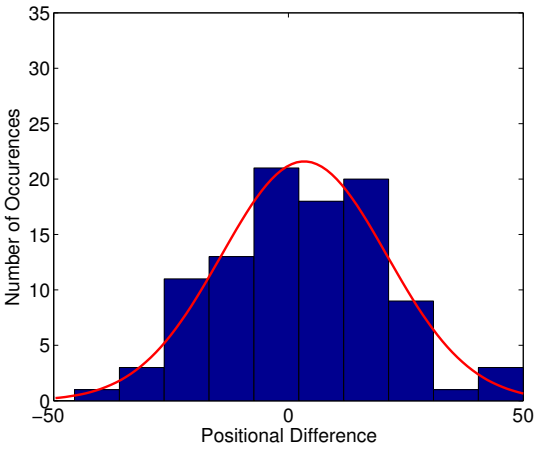
Upon completion of the experiment participants were given the opportunity to offer informal feedback on the test and test method. Significant remarks were noted anonymously such that over the course of the experiment common remarks could be recognised and brought forward to reconsideration of the experimental design.

6.4 Experimental Results

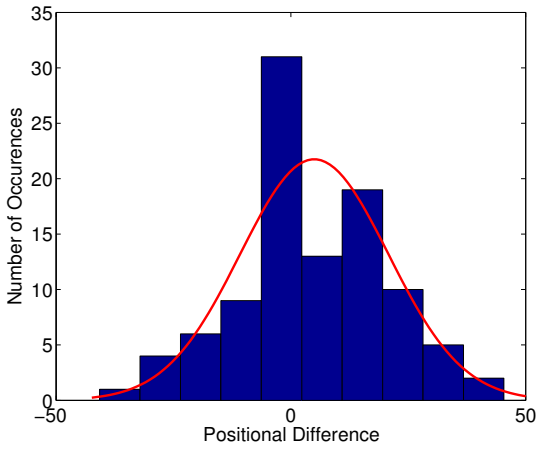
Several variables were measured during the experiment, including the positional differences for each A/B pair, the timbral differences, the time taken before submitting the slider values for each pair, and the number of times both A and B were played for each test, however the variable of most interest was the reported positional differences as they provide the crucial information as to how well the spatial impression of the modelled HRTFs were retained in each model case. Subsequently this is the focus of the results reported in this section.



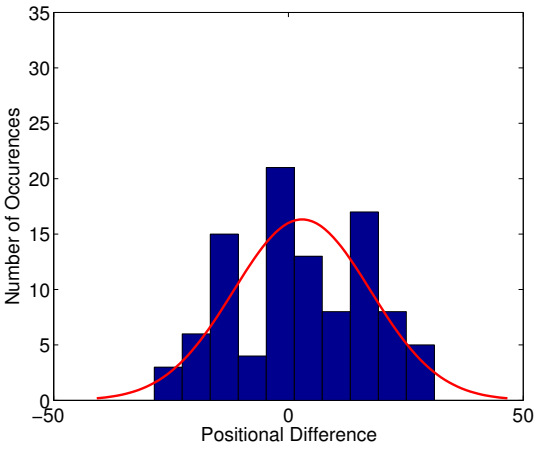
(a) Order: 0



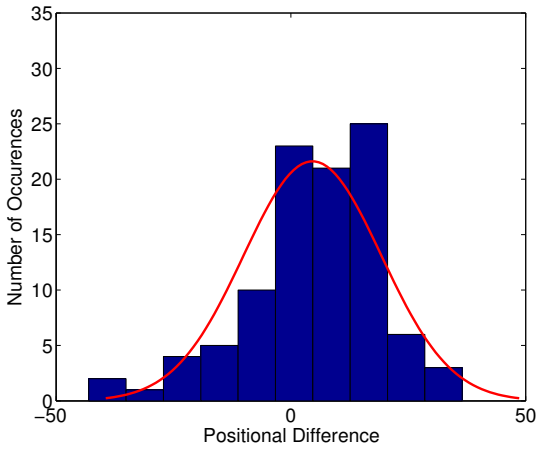
(b) Order: 4



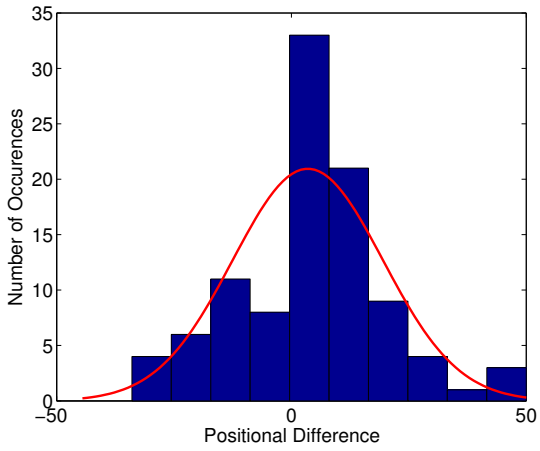
(c) Order: 10



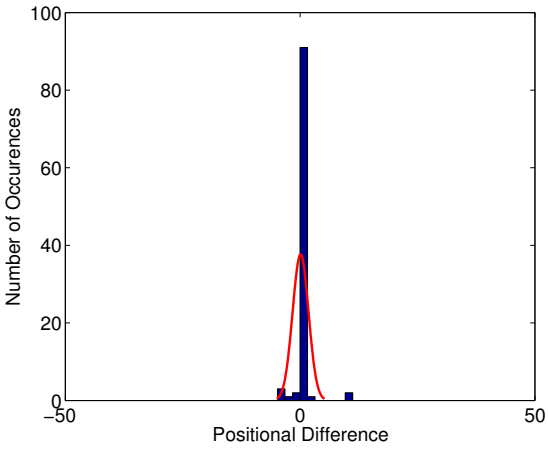
(d) Order: 16



(e) Order: 22



(f) Order: 30



(g) Order: 'inf'

Figure 6.4.1: Histograms of reported positional differences per model

Figure 6.4.1 shows the number of occurrences of grouped reported positional difference values for each model order across all participants and positions. An important characteristic of the various model histograms is that they all appear to exhibit an approximately normal distribution, as is evidenced by the distribution fit superimposed in red over each of the individual plots. The varying widths of the normal fit distributions describe the amount of variance from the mean of the reported values; figures 6.4.1a and 6.4.0g show the histogram plots of the '0 order' ITD only model and the 'perfect' reconstruction model respectively, as may be

expected these two model cases exhibit the widest and narrowest normal distribution fits of all model cases. The fact that all of the distribution fits center on approximately 0 suggest that there was no consistent source of directional bias introduced by either the experimental stimuli or method. An important advantage of the approximately normal distribution of the collected experimental data for each model is that it allows for convenient analysis and comparison through commonplace statistical techniques.

Of particular interest is the standard deviation of the positional difference data obtained for each model. The standard deviation is a measure of the average deviation of the the measured data from the mean of that data, and is defined as follows:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}} \tag{6.4.1}$$

Where σ is the standard deviation of the data group, x is the data group, μ is the mean of the data group, and N is the number of elements in the data group x.

As the measured data for the positional differences are approximately zero-mean, the standard deviation of the data pertaining to each of the 7 model cases provides a suitably robust measure of the perceived variation in the localisation of the modelled HRTF compared to the measured HRTF.

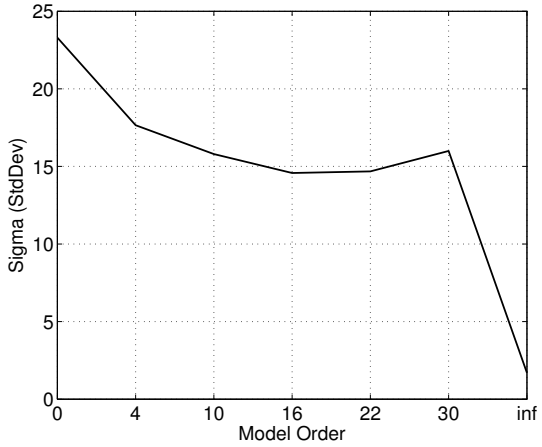


Figure 6.4.2: Sigma of reported positional differences against model case

Figure 6.4.2 shows the standard deviations of each of the 7 model cases, each derived from the 100 model specific observations collected during the experimental run; 10 participants remarking upon the positional differences of 10 positions for each model. The 0^{th} and 'infinite' order models yield the largest and smallest values of standard deviation respectively; this is an expected result as the two model cases were designed to yield a worst and best case reference in terms of observed positional deviations. The values for the 5 IIR model cases lay in-between the best and worst cases as expected, however the standard deviation observed for IIR model orders 22 and 30 provide an unexpected result. The experimental hypothesis states that the increase in IIR model order should yield a decrease in the observed variation in positional difference as reported by the experimental participants, i.e. the higher the model order, the closer the standard deviation should be to that of the 'infinite' order model case. The 4^{th}, 10^{th}, and 16^{th} order models seem to exhibit an approximately linear trend that agrees with this hypothesis, however the 22^{nd} order model exhibits approximately the same standard deviation as the 16^{th} order model and the 30^{th} order shows an increase in the standard deviation to approximately the same value as the 10^{th} order model. It is possible that the increase in standard deviation of the higher two model orders is telling of a limit for which the addition of further poles and zeros to the IIR model actually causes the model to effectively become overdetermined. It is also possible that the seemingly narrow spread of the standard deviations for the 5 IIR model cases suggests that additional experimental data should be gathered before the experimental hypothesis and null hypothesis can be remarked upon with sufficient confidence.

The ANOVA test, short for Analysis of Variance, is a statistical model useful for testing the means of three or more groups of observations or variables for statistical significance. The procedure serves to identify a single probability value p, under the null hypothesis that all samples from all data groups are drawn from populations with the same mean; i.e. that the IIR model order has no significant impact in the mean value of reported positional differences of the A and B samples. However the 1-way ANOVA test merely remarks that at least one group has a statistically significantly different mean value, a more convenient method of analysis is to use the 1-way ANOVA data to perform a multiple comparison test. The multiple comparison test provides information regarding which pairs of group means are significantly different and which are not, whereas the ANOVA test only returns an indication

of whether or not all group means are statistically similar.

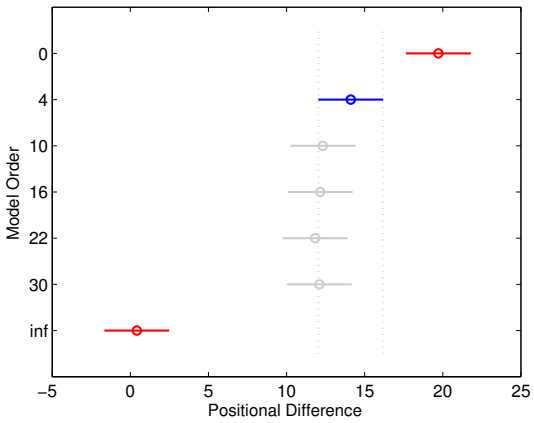


Figure 6.4.3: Multiple comparison of model means

Figure 6.4.3 illustrates the findings of the multiple comparison procedure; the circle markers represent the means of the different data groups listed along the Y-axis, the pair of symmetrical lines protruding from each of the mean values represent the confidence intervals of each group mean. In the figure the 4^{th} order model data group is selected, as such the mean of this model is highlighted in blue, the remaining IIR model groups exhibit confidence intervals that overlap with that of the selected group and as such are coloured grey to indicate that the means of these groups are not significantly different from the 4^{th} order model group. Conversely the 0^{th} and 'infinite' order model groups are highlighted in red to denote that their means can be considered significantly different from that of the 4^{th} order group, or any of the IIR model groups in fact, as none of the confidence intervals extend sufficiently to overlap with the two extreme group means. The confidence intervals may be reduced in size given a larger population from which the statistics are generated, that is to say that further subjective testing would likely reduce the size of the confidence intervals sufficiently that the confidence intervals of some overlapping groups may no longer overlap. However the fact that the confidence intervals of the 0^{th} and 4^{th} order model groups do not overlap may be considered to suggest that a sufficient population was sampled to somewhat confidently remark on the experimental findings.

Chapter 7

Discussion

This chapter of the thesis provides a comparison and critical discussion of the results and implications of the application of the described methods of HRTF compression and interpolation to a measured dataset. This section also highlights possible flaws in the experimental procedure forming a foundation from which the concluding remarks and designation of further works may be drawn.

7.1 Compression

Broadly speaking, the results presented in sections 4 and 5 show that both the decompositional and parametric modelling based approaches are able to achieve a compressed representation of the measured HRIR dataset with differing levels of accuracy in reconstruction or modelling performance. Under the umbrellas of these two approaches a total of four methods of HRIR/HRTF compression have been investigated: under the decompositional approach; the application of the principal component analysis to the measured magnitude spectra in the linear domain and the application of the principal component analysis to the measured magnitude spectra in the logarithmic domain, and under the parametric modelling approach; the modelling of measured HRTFs as all-pole filters using linear prediction and the modelling of measured HRTFs as pole-zero filters using the Steiglitz-McBride algorithm.

To summarise the implications of the findings regarding the level of compression achiev-

able through each technique at the cost of NMSE performance: The linear PCA technique achieves a reduction from 360 measurements each of 512 samples to as few as two 256 element basis vectors and two corresponding 360 element weight vectors whilst retaining 95% of the total variance, however the low values of the linear magnitudes require a much greater number of basis and weight vector pairs to be used, shown in figure 4.1.5 to be approximately 125 pairs needed to ensure no reconstructed elements are negative. The logarithmic PCA technique requires 6 principal components, that is 6 basis and weight vector pairs, to be retained in order to reconstruct 95% of the variance of the dataset, furthermore as the PCA is applied to the logarithmic magnitudes no additional principal components need be retained to ensure a practical reconstruction is achieved, and as such the logarithmic domain PCA should be considered the superior of the two decompositional methods used. The parametric modelling approaches offer a reduction from 360 measurements each of 512 samples to 360 sets of coefficient weights, the order of which can be adjusted in order to increase or decrease the performance of the model filters; this is also true of the decompositional approach in that the number of principal components used in reconstruction can be adjusted at the benefit or cost of the reconstruction performance. Superior performance was shown to be obtained from as little as 15 coefficients using the linear prediction method to derive all-pole filter approximations of the measured HRTFs (figure 5.1.2) when compared to the logarithmic PCA method using a 6 PC reconstruction (figure 4.2.4). However the same cannot be said of the simulation results for the Steiglitz-McBride based method, which showed a significantly poorer performance, in particular when used to derive the 15^{th} order pole-zero filter approximations of the measured spectra that exhibited higher NMSE for all angles (figure 5.3.4) than that of the counterpart 15^{th} order all-pole models (figure 5.1.2). The unexpectedly poor performance of the Steiglitz-McBride method will be discussed in some detail later in this section.

From the fundamental compression standpoint alone it is perhaps quite difficult to objectively define which of either the logarithmic PCA or linear prediction all-pole methods achieves a more efficient representation of the measured dataset, as they have been shown to perform somewhat comparably for a similar number of total stored elements. However when the techniques are extended to include the DCT deconstruction and limited component reconstruction of angle dependent element vectors then it can be shown that, ignoring further

compression achievable via spatially downsampling the measured dataset and assuming the use of the 6 PC reconstruction and the $15th$ order all-pole model, the linear prediction method yields superior compression for slightly improved NMSE performance than that of the logarithmic PCA method.

The performance of the DCT based method of secondary compression of the angular dependent element vectors could be improved. By implementing varying reconstruction thresholds which are dependent on the relative level of total variance captured by the angular dependent vector in question. For example, in the linear prediction method, a high threshold value was required to account for the large amount of variance in the angular variation of the first order K coefficient. However for higher order K coefficient vectors this threshold could have been reduced with little cost to the overall model performance. This is also true of the PCA weight vector DCT decomposition and reconstruction, as both methods see the angular dependent vectors arranged in order of importance, or amount of total variance explained. As such both can be exploited to this end by reducing the DCT reconstruction threshold for each angle dependent vector, be it PCA weight or K coefficient, proportionally to the amount of the total dataset variance it accounts for.

7.2 Interpolation

Continuing the discussion of the application of the DCT to not only compress but also to interpolate the angle dependent features in both methods; the DCT was shown to allow for a compressed expression of the PCA weight vectors and the K coefficient vectors, for the lower order vectors in each method the number of DCT components required was significantly lower than the number of elements in the uncompressed vector with an approximate loss of less than 1% of the variance of the vector. The DCT has been shown to be an effective and robust means of interpolation in other applications by other authors [Agbinya, 1992] [Hsu and Chen, 1997], and the work of these past authors can be considered to support the interpolation results identified in the works of this thesis.

The compression and interpolation results obtained for both methods and 'sub-methods'

show a recurring increase in reconstruction/modelling error of HRTFs at contralateral angles. This phenomenon has been addressed by many authors, and is consistently attributed to the comparatively low signal level at the occluded ear for source positions on or close to the inter aural axis, and the relative complexity of the spectral shapes at these positions, due to diffraction around the head [Kulkarni et al., 1999] [Chen et al., 1995] [Zhang, 2009]. For the purposes of ITD estimation or extraction, it is possible that HRIR measurement could be conducted using a higher level acoustic stimulus, such that the lower level signal measured at contralateral positions is sufficiently clear of the system noise floor. This would attempt to reduce the inaccuracies of the ITD extraction methods that likely occur as a result of the lower signal to noise ratio. However a global level adjustment as described would provide no improvement in the reconstruction/modelling error that occurs due to the increased spectral complexity and relatively lower level of contralateral measurements. Possible improvements may be achievable in the reconstruction or modelling of contralateral positions should they be given a weighted emphasis during processing. In particular for the decompositional process; a boost in the relative level of the contralateral measurements would serve to add 'importance' to them, increasing their share of the total variance of the dataset. Hence during decomposition the more complex spectral shapes, that occurred previously only at relatively low levels, would be better represented by the lower order principal components, or perhaps orthogonal K coefficients. Such a method would likely result in an increased number of principal components or K coefficients required in order to capture the whole variance of the dataset, and of course an appropriate inverse weighting after reconstruction, but should serve to increase the reconstruction or modelling accuracy of the contralateral positions. Chen et al. suggested similarly that the contralateral data could be weighted upon the construction of the PCA covariance matrix, however they did not attempt to implement such an optimisation [Chen et al., 1995].

7.3 Steiglitz McBride Performance

A surprising feature of the simulation results is the consistently poor performance of the Steiglitz-McBride derived pole-zero filter models, which exhibit the largest amount of NMSE of all the implemented techniques for comparable or equivalent model quality or order. The

Steigltiz-McBride derived filters with 15 poles and 15 zeros gave an overall worse NMSE performance across all angles than that of the linear prediction method 15 pole all-pole equivalent, with a difference on the order of 20dB NMSE at some angles. It was expected that the effective addition of the zeros to the all-pole model would yield a reduction of NMSE for all angles, due to the models improved ability to capture and recreate notch details in the measured frequency spectra, however the results suggest this expectation to be erroneous. In previous work Kulkarni & Colburn [Kulkarni and Colburn, 2004] found superior results using a modification of the Steiglitz-McBride method; by introducing a weighting to the quadratic cost function originally proposed by Kalman [Kalman, 1958], they were able to obtain a superior fit of the HRTF spectra in the logarithmic domain. It should also be noted that Kulkarni & Colburn performed the modelling procedure only for frequencies below 15kHz and on the mean-less direct transfer functions as opposed to the measured HRTFs directly. Although these differences should be considered to explain some of the inconsistency between the results obtained by Kulkarni & Colburn and the results found in this work, it is unlikely that they account for all of it. The lack of the logarithmic based weighting function should not have had much of an impact on the high NMSE observed as the error calculation is performed in the linear domain, which is the same as the unweighted cost function at the heart of the Steiglitz-McBride method. The lower number of modelled frequencies and the use of the DTF in Kulkarni & Colburn's work should certainly result in a more optimal placement of poles and zeros for the same model order when compared to the wideband HRTF models, however the wideband models observed in these works consistently exhibit sharp peaks that under predict many of the wider peak features of the measured HRTFs. It is likely that this characteristic has formed as an effect of an element of the Steiglitz-McBride function *stmcb.m* in MATLAB. By default *stmcb.m* uses Prony's method to obtain an initial estimate of the denominator coefficients, which are in turn used to pre-filter the input and output records of the system in the iterative procedure as described by Steiglitz & McBride [Steiglitz and McBride, 1965]. Prony's method is known to perform poorly in the presence of noise [Marple, 1987] and is likely a largely contributing factor to the unexpectedly poor performance of the pole-zero filter models. In testing certain model orders greater than 40 resulted in huge error peaks on the MNMSE surface such as the one shown in 5.3.6, upon further informal examination it was found that the source of the increased error was a gross mis-prediciton of excessively high peaks in one or two of the contralateral models. As the

lower level contralateral positions are the 'noisiest' of the measured positions this seems to support this explanation. A final criticism of the implementation of the iterative method used is that the *stmcb.m* function performs a fixed number of iterations, 5 by default [The MathWorks Inc., 2014], and does not perform a means of checking to see if the model coefficients have converged. It is possible that this blind approach to the number of iterations performed could be responsible in part for the poor performance of the models, however Kulkarni & Colburn found that their, albeit adapted, procedure usually converged within 4 iterations [Kulkarni and Colburn, 2004].

An alternative improvement that could be made to the Steiglitz-McBride method used in these works, besides those implemented by Kulkarni & Colburn [Kulkarni and Colburn, 2004], would be to combine it with the linear prediction method used to generate the all-pole filter models. The results obtained for the all-pole method simulation show a promising fit with the measured data for low model orders. As such the linear prediction method described in section 5.1 could be used to obtain the first estimate of the denominator coefficients for the iterative method. Though still affected by noise, the linear prediction method has been shown to yield comparatively good results for all angles, and as such are assumed to provide a superior estimate of the denominator coefficients, which should in turn yield an improved accuracy in the placement of the numerator terms or zeros, for the same fixed number of iterations.

7.4 Pilot Study

Regarding the subjective experiment designed and run as a pilot study in order to elicit a trend between pole-zero model order and pychoacoustically judged source position. Though the pole-zero models were selected for use in the pilot study under a false prediction made during preliminary simulation results, the results of the subjective experiment still offer some relevant information regarding the influence of pole-zero model order on the perceived location of the stimuli. The most informative results are displayed in figure 6.4.3; It is significant to observe, with statistical confidence, that the ITD only model case performed poorer than any of the pole-zero model cases. Furthermore a possible trend seems to be emerging amidst

the pole-zero case results; that the positional error decreases proportionally to the increase in pole-zero model order. However given the small preliminary sample size it is unclear as to whether the trend would develop further, or whether the two highest model orders do in fact represent cases of overdetermination due to an excess of zero components. This possible trend would of course be better supported or unsupported given further testing to gather a larger sample base from which the statistics are drawn.

During testing a number of remarks made by participants highlighted possible 'weak' elements of the experimental design, which given the purpose of a pilot study, would be appropriate to be considered as a basis of re-design of said elements before extensive testing begins. The remarks addressed two design elements in particular. The first is the use of a relative position slider, which was remarked to be somewhat confusing due to the need for the participants to mentally reposition the slider such that the centre of the slider corresponded to the location of the A stimulus. The second is the apparent ambiguity in the definition of the timbre slider, which is likely due to the shown multi-dimensionality of the term timbre [Plomp and Smoorenburg, 1970] [Schouten, 1968] [Samson et al., 1997]. In hindsight the choice of the term timbre is perhaps inappropriate as the purpose of the slider was to elicit a remark on the similarity of the two stimuli in question, not the complex qualities of the sounds themselves.

Considering the means by which the data has been analysed, and given the exact nature of the experimental hypothesis the positional slider may be simplified to represent a directionally independent measure of the proximity of two stimuli. Such that the maximum slider position would denote that the two stimuli appear to emanate from the same spatial location, and conversely the minimum slider position would denote that the sources are at a maximum distance apart. The ambiguity in the definition of the timbre slider should be somewhat alleviated by the use of an alternate naming scheme. A meaningful simplification of the similarity/dissimilarity between the two stimuli might be obtained by applying a similar unidirectional encoding scheme as the revised proximity slider. In the revised scheme a maximum slider value would denote that the two samples sound identical regardless of their relative positions, and the minimum slider value would denote that the two samples sound maximally different.

A possible improvement to the overall design of the test could see the implementation of a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) design [ITU, 2014]. Choosing to either perform two simultaneous tests to capture the revised proximity and similarity slider information, or possibly discarding the secondary similarity slider information altogether. The MUSHRA design would permit the convenient testing of an expanded stimuli base whilst also serving to provide key results with higher statistical significance than the original A/B test design.

Chapter 8

Conclusions

The aim of the work described in this thesis was to investigate a pair of decompositional and a pair of parametric modelling approaches to the compression of a measured HRIR dataset restricted to the azimuthal plane. Novel application of a secondary compression and convenient interpolation scheme to orthogonal angle dependent components using the Discrete Cosine Transform has been considered also. Overall it can be claimed that this aim has been met; of the four methods investigated the relative strengths and weaknesses have been identified and compared, with consideration given to the underlying causes of differences in the performance of the methods with respect to an adapted error metric.

To summarise the four methods in question, it has been shown that of the two decompositional approaches the PCA approach in the linear domain is inferior to the PCA approach in the logarithmic domain. This is due to the need of a large number of principal components required to obtain a wholly positive reconstructed spectrum. Of the two parametric modelling approaches, the linear prediction method offers vastly superior accuracy in modelling the measured spectra for comparable model orders than those models generated using the Steiglitz-McBride approach. Though the relevant theory would lead to the assumption that the pole-zero models generated using the Steiglitz-McBride method should provide a superior fit of modelled to measured data, several explanations have been offered regarding sources of the unexpectedly poor performance of the method. Furthermore, likely means of improving the implementation of the Steiglitz-McBride method have also been given.

Subjective experimentation in the preliminary form of a pilot study has suggested that, although the Steiglitz-McBride derived pole-zero models performed poorly with respect to the NMSE metric, the models perform better than an ITD only case with regards to the preservation of subjective impression of source location, but perform considerably worse than the measured HRTF. The early results also indicate that more test participants are required in order to confirm, or deny, the presence of a seemingly emergent trend between the pole-zero model order and subjective localisation performance with greater statistical significance.

Further to the compression achieved by the application of either the PCA approach in the logarithmic domain or the linear prediction all-pole method, consideration has also been given to an extension of each of these methods. Utilising the discrete cosine transform one obtains a means of performing a secondary compression and convenient interpolation of the orthogonal angle dependent terms of the efficient representation. For the logarithmic PCA method, these are the weight vectors that describe the contribution of each of the basis vectors to the reconstructed spectrum at each angle. Whereas for the linear prediction method these are the variation of the K coefficients with respect to angle, constructed as vectors. This extension of the HRTF compression methods is shown to yield significant levels of additional compression. These levels are particularly marked for the comparatively low order variation in the vectors pertaining to either the first PCA weights or the low order K coefficients.

In terms of interpolation performance, the validity of the DCT-based interpolation in this application is illustrated by observing the increase in the distribution of NMSE over angle between full and interpolated datasets. Interpolated approaches are realised using a down-sampled dataset of 36 evenly distributed measurements. These subsets are then interpolated back to the 1° accuracy of the measured superset, using the DCT method in conjunction with one of the compression methods. The DCT method of interpolation is shown to incur a moderately low increase in NMSE over the non-interpolated, but identically compressed cases. As such the method can be considered to be an inexpensive and convenient means of interpolating the compressed datasets. However the interpolation method is yet to be compared directly to other interpolation algorithms applied at the same stage in the reconstruction process, and as such should be considered somewhat unverified.

Upon consideration of the extended logarithmic domain PCA and K-coefficient linear prediction methods, it can be concluded that the latter holds greater potential for representational efficiency of a measured HRTF dataset. The method achieves this potential through a combination of factors. Firstly via the reduction from lengthy impulse responses to comparatively low order all-pole filter coefficients. Secondly through the possible reduction in reference data measurement size through convenient interpolation to higher spatial sampling rates. Finally through the expression of each of the orthogonal K coefficient angular variation vectors as a smaller number of DCT components.

Chapter 9

Further Work

Immediate further work on the project will see a full scale subjective investigation conducted. This will include refinement of the experimental design following the preliminary results obtained through the pilot study, as well as expansion of the stimuli base to include the reconstructed or modelled HRTFs of both the log magnitude PCA and linear prediction methods respectively.

The secondary compression stage of the K-coefficient linear prediction method may be improved even further, by realising the reconstruction threshold of each DCT analysis proportionally to the order of the K coefficient vector being analysed. This would result in a reduction of the number of DCT components retained for higher order angular K coefficient vectors, which should contain less pertinent spectral information. Ultimately this modification would serve to further increase compression, with only minor loss of high order detail that accounts for significantly less of the total variance of the measured data.

Furthermore, the current DCT based interpolation scheme could be altered in order to obtain a functional representation of the K coefficient, or even PCA weight, angle dependent vectors. During the DCT analysis, if rather than retaining the DCT output as a vector, to be IDCT'd, the DCT output can be realised as the weights of a series of continuous cosine functions of varying argument. The resulting weighted sum is continuous and therefore can be queried for any angle on a continuous scale. A functional representation of the data such as this is advantageous in applications for which the desired spatial resolution is initially

unknown or varying, as the function requires no adjustment or re-analysis of the original dataset to provide data for any angle.

An additional avenue of investigation that is within short reach from the current state of the works is the orthogonal transformation of the linear prediction a coefficients through either the Fourier or discrete cosine transform. Instead of expressing the all-pole filters in terms of the orthogonal K coefficients that drop out of the Levinson-Durbin solution, an alternative orthogonal transformation of the series of a coefficients for each angle, in particular one with a sinusoidal/cosinusoidal basis, may yield yet further compression and may also uncover simpler or lower order angular variation vectors of the compressed coefficients.

Likely informative results will be gathered from an in-depth comparison of alternative interpolation methods and the DCT interpolation in orthogonal domains as suggested in the works described in this thesis. Considering interpolation methods that operate in the time, frequency, or further orthogonal domains will offer a more contextual validation of the amount of error introduced using the proposed DCT based interpolation method.

Looking further the works will extend to cover further investigation into the somewhat surprising shortcomings of the Steiglitz-McBride method for the derivation of the pole-zero filter models. Further work will attempt to isolate and explore the source of the behaviour of the method in this implementation, beyond the comparison with the more successful results obtained using an extended form of the technique by another author, as previously discussed.

Bibliography

Agbinya, J. (1992). Interpolation using the discrete cosine tranform. *Electronics letters*, 28(20):1991–1992.

Asano, F., Suzuki, Y., and Sone, T. (1990). Role of spectral cues in median plane localization. *The Journal of the Acoustical Society of America*, 88(1):159–68.

Begault, D., Wenzel, E., and Anderson, M. (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 ...*, 49(10).

beyerdynamic GmbH & Co. KG (2014). DT770 Pro Specifications.

Breebaart, J. and Kohlrausch, A. (2001). The perceptual (ir) relevance of HRTF magnitude and phase spectra. *Audio Engineering Society Conference 110*.

Busson, S., Katz, B., and Nicol, R. (2005). Subjective Investigations of the Interaural Time Difference in the Horizontal Plane. *Audio Engineering Society Convention 118*.

Carlile, S., Jin, C., and Raad, V. V. (2000). CONTINUOUS VIRTUAL AUDITORY SPACE USING HRTF INTERPOLATION : acoustic & psychophysical errors. *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pages 220–223.

Chen, J., Van Veen, B. D., and Hecox, K. E. (1995). A spatial feature extraction and regularization model for the head-related transfer function. *The Journal of the Acoustical Society of America*, 97(1):439–52.

Duda, R. and Martens, W. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058.

Duraiswaini, R. (2004). Interpolation and range extrapolation of HRTFs. *Acoustics, Speech, and Signal Processing*, 4:iv–45 – iv–48.

Estrella, J., Lindau, A., and Weinzierl, S. (2010). Individualization of dynamic binaural synthesis by real time manipulation of the ITD. *Audio Engineering Society Conference 128*.

European Parliament (2003). DIRECTIVE 2003/10/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 February 2003 on the minimum health and safety requirements regarding the exposure of workers to the risks arising from physical agents (noise).

Evans, M., Angus, J., and Tew, A. (1997). Spherical harmonic spectra of head-related transfer functions. *Audio Engineering Society Convention 103*, 4571.

Evans, M., Angus, J., and Tew, A. (1998). Analyzing head-related transfer function measurements using surface spherical harmonics. *The Journal of the Acoustical Society of America*, 104(May 2013):2400–2411.

Everest, F. A. and Pohlman, K. C. (2009). *Master Handbook of Acoustics*. McGraw-Hill Companies, Inc., 5th edition.

Farhang-Boroujeny, B. (1999). *Adaptive Filters*. John Wiley & Sons, Ltd., Chichester.

Gamper, H. (2013). SELECTION AND INTERPOLATION OF HEAD-RELATED TRANSFER FUNCTIONS FOR RENDERING MOVING VIRTUAL SOUND SOURCES. *dafx13.nuim.ie*, pages 1–7.

Gerbrands, J. (1981). ON THE RELATIONSHIPS BETWEEN SVD, KLT AND PCA. *Pattern recognition*, 14(4):375–381.

Howard, D. and Angus, J. (2009). *Acoustics and Psychoacoustics*. Focal Press, Oxford, fourth edition.

Hsu, Y. and Chen, Y. (1997). Rational interpolation by extendible inverse discrete cosine transform. *Electronics letters*, 33(21):1–2.

Humanski, R. and Butler, R. (1988). The contribution of the near and far ear toward localization of sound in the sagittal plane. *The Journal of the Acoustical Society of America*, 83:2300–2310.

Huopaniemi, J., Zacharov, N., and Karjalainen, M. (1999). Objective and subjective evaluation of head-related transfer function filter design. *Journal of the Audio Engineering Society*, 47(4):218–239.

ITU (2014). BS. 1534-2 (06/2014) Method for the subjective assessment of intermediate quality levels of coding systems.

Jolliffe, I. (2005). Principal Component Analysis.

Kalman, R. (1958). Design of a self-optimizing control system. *Transactions of the American Society of Mechanical Engineers*, 80:468–478.

Katz, B. F. G. and Noisternig, M. (2014). A comparative study of Interaural Time Delay estimation methods. *The Journal of the Acoustical Society of America*, 135(6):3530–40.

Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *The Journal of the Acoustical Society of America*, 91(3):1637–47.

Kulkarni, A. (1995). On the minimum phase assumption OF HEAD-RELATED TRANSFER FUNCTIONS. *Applications of Signal Processing to Audio and Acoustics , IEEE ASSP Workshop on*, pages 84–87.

Kulkarni, A. and Colburn, H. S. (2004). Infinite-impulse-response models of the head-related transfer function. *The Journal of the Acoustical Society of America*, 115(4):1714.

Kulkarni, A., Isabelle, S. K., and Colburn, H. S. (1999). Sensitivity of human subjects to head-related transfer-function phase spectra. *The Journal of the Acoustical Society of America*, 105(5):2821–40.

Lindau, M. (2010). On the extraction of interaural time differences from binaural room impulse responses. Technical Report September 2010, TU-Berlin.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4).

Malvar, H. S. (1992). *Signal Processing with Lapped Transforms*. Artech House, Inc., Norwood.

Man, B. D. and Reiss, J. (2013). A knowledge-engineered autonomous mixing system. *Audio Engineering Society Convention 135*.

Marple, S. L. J. (1987). *Digital spectral analysis with applications*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Martens, W. (1987). Principal components analysis and resynthesis of spectral cues to perceived direction. *Proceedings of the International Computer Music Conference*, pages 274–281.

Matsumoto, M. and Yamanaka, S. (2004). Effect of Arrival Time Correction on the Accuracy of Binaural Impulse Response Interpolation–Interpolation Methods of Binaural Response. *Journal of the Audio Engineering Society*, 52(1):56–61.

Mehrgardt, S. and Mellert, V. (1977). Transformation characteristics of the external human ear. *The Journal of the Acoustical Society of America*, 61:1567–1576.

Mills, A. (1958). On the minimum audible angle. *The Journal of the Acoustical Society of America*, 45(1905):237–246.

Minnaar, P. and Plogsties, J. (2000). The Interaural Time Difference in Binaural Synthesis. *Audio Engineering Society Convention 108*, 5133.

Morimoto, M. (2001). The contribution of two ears to the perception of vertical angle in sagittal planes. *The Journal of the Acoustical Society of America*, 109(4):1596–1603.

Mullis, C. and Roberts, R. (1976). The use of second-order information in the approximation of discreate-time linear systems. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(3):226–238.

Nam, J., Kolar, M., and Abel, J. (2008). On the minimum-phase nature of head-related transfer functions. *Audio Engineering Society Convention 125*, pages 1–8.

Nishimura, R., Kato, H., and Inoue, N. (2009). INTERPOLATION OF HEAD-RELATED TRANSFER FUNCTIONS BY SPATIAL LINEAR PREDICTION. *Acoustics, Speech and Signal Processing, IEEE International Conference on*, pages 1901–1904.

Oppenheim, A. V. and Schaeffer, R. W. (1975). *Digital Signal Processing*. Prentice-Hall, Inc., London.

Parks, T. and Burrus, C. (1987). *Digital filter design*. Wiley-Interscience New York, NY, USA, New York.

Plomp, R. R. and Smoorenburg, G. F. (1970). Frequency Analysis and Periodicity Detection in Hearing. *The Proceedings of the International Symposium on Frequency Analysis and Periodicity Detection in Hearing*.

Ramos, G. and Cobos, M. (2013). Parametric head-related transfer function modeling and interpolation for cost-efficient binaural sound applications. *The Journal of the Acoustical Society of America*, 134(3):1735–8.

Samson, S., Zatorre, R., and Ramsay, J. (1997). Multidimensional scaling of synthetic musical timbre: Perception of spectral and temporal characteristics. *Canadian Journal of Experimental Psychology*, pages 307–315.

Sandvad, J. and Hammershoi, D. (1994). Binaural auralization Comparison of FIR and IIR Filter representation of HIRs. *Audio Engineering Society Convention 96*, 3862.

Schouten, J. F. (1968). The Perception of Timbre. *Reports of the 6th International Congress on Acoustics*, 6(90):35–44.

Schroeder, M. (1965). New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America*.

Senova, M., McAnally, K., and Martin, R. (2002). Localization of virtual sound as a function of head-related impulse response duration. *Journal of the Audio Engineering Society*, pages 57–66.

Steiglitz, K. and McBride, L. (1965). A technique for the identification of linear systems. *Automatic Control, IEEE Transactions on*, 10(4):461–464.

The MathWorks Inc. (2014). MATLAB Help Files.

Wang, L., Yin, F., and Chen, Z. (2008). HRTF compression via principal components analysis and vector quantization. *IEICE Electronics Express*, 5(9):321–325.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonidividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94(1):111–123.

Wierstorf, H., Geier, M., and Spors, S. (2011). A free database of head related impulse response measurements in the horizontal plane with multiple distances. *Audio Engineering Society Convention 130*, pages 3–6.

Wightman, F. L. and Kistler, D. J. (1989). Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–67.

Woodworth, R. S. (1938). *Experimental Psychology*. Holt, New York, New York.

Xie, B. and Zhang, T. (2010). The audibility of spectral detail of head-related transfer functions at high frequency. *Acta Acustica united with Acustica*, 96(2):328–339.

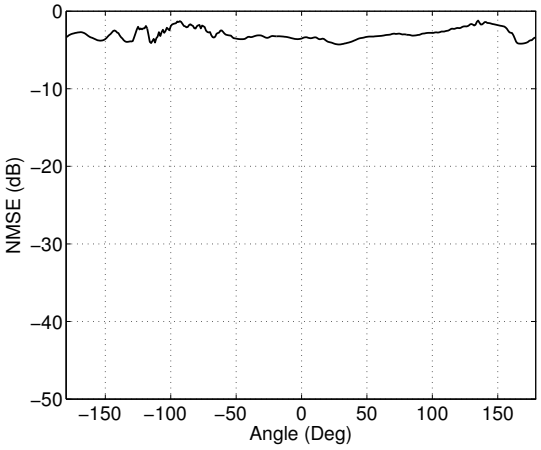
Zhang, W. (2009). Efficient continuous HRTF model using data independent basis functions: Experimentally guided approach. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):819–829.

Zhang, W. and Abhayapala, T. (2009). MODAL EXPANSION OF HRTFS : CONTINUOUS REPRESENTATION IN FREQUENCY-RANGE-ANGLE. *Acoustics, Speech and Signal Processing, IEEE Conference on*, pages 285–288.

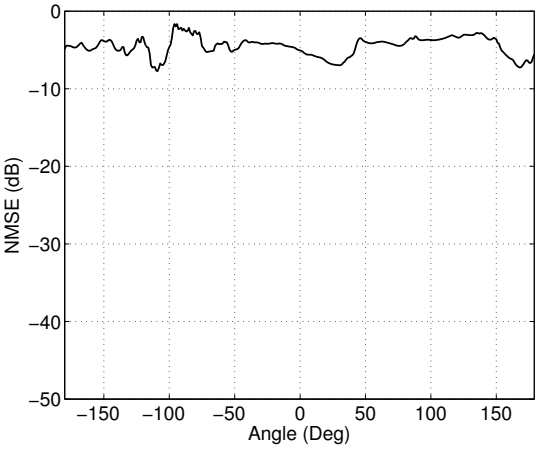
Zölzer, U. (2011). *DAFX: digital audio effects*. John Wiley & Sons, Inc. New York, NY, USA, second edi edition.

Appendix A

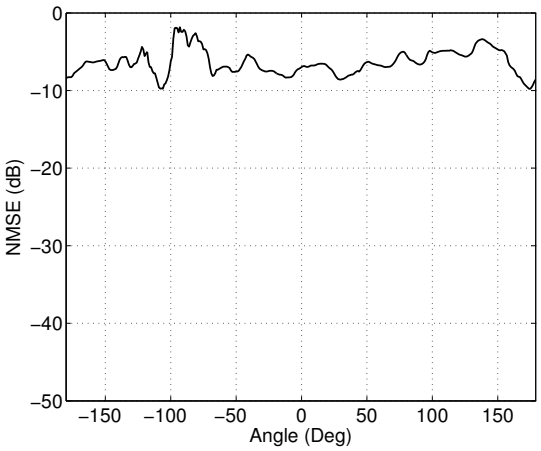
Pole-Zero Model Performances



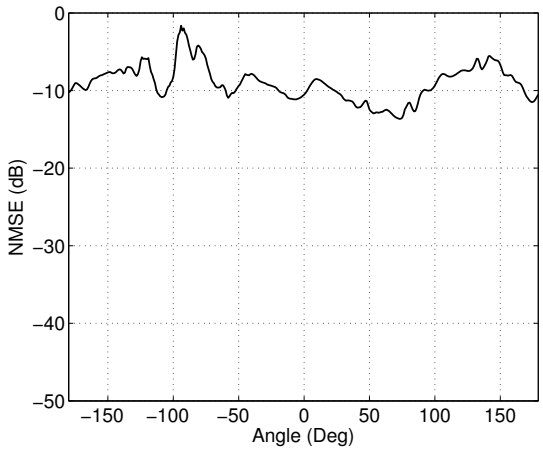
(a) Order: 4



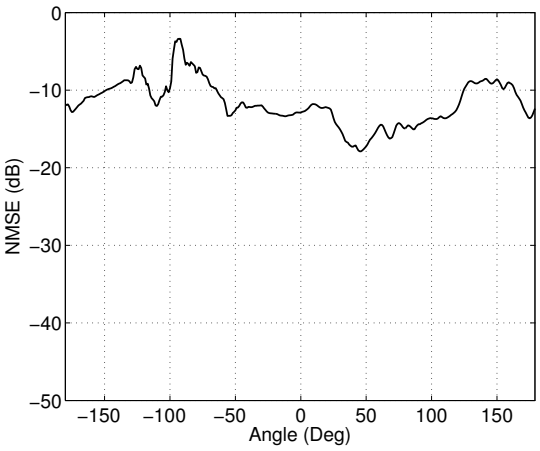
(b) Order: 10



(c) Order: 16



(d) Order: 22



(e) Order: 30

Figure A.0.-1: NMSE of pole-zero model orders used as subjective stimuli

Appendix B

Subjective Testing Information

B.1 Information for Participant

Efficient Representation of Head Related Transfer Functions

You are invited to take part in a research study. Before you decide whether or not you wish to take part it is important for you to understand why the study is being done and what it will involve if you agree to take part. Please read the following information carefully. Ask the researcher if there is anything you don't understand or if you would like more information.

What is the purpose of the study?

To investigate efficient and compact representations of the Head Related Transfer Functions necessary to create 3D audio over headphones.

What will happen to me if I take part?

You will be asked to wear a pair of headphones and listen to a series of 70 A/B comparisons of filtered noise samples. For each comparison you will be asked to indicate on a scale the relative position of one sample compared to the other, you will also be asked to the indicate the difference in timbre of the two sounds on a second scale. The test should take no longer than 30 minutes to complete.

Is there any risk?

There is no risk of harm. All audio has been set to a comfortable listening level well in accordance with the relevant guidance.

Confidentiality – who will have access to the data?

All the data will be held securely and will be treated confidentially. The arrangements for data storage and security comply with the terms of the Data Protection Act. If it helps to clarify the results of the study we may quote some phrases that you say when talking to the researcher. Any quotes will be anonymous – nobody will know you have taken part in the study. If you decide to withdraw from the study for any reason or at any time, any data already collected will be deleted and any paper copies destroyed.

What will happen to the study results?

In any material published from this study, all participants will be anonymous.

You can decide to change your mind and withdraw from the study at any time without having to give a reason for withdrawing.

The researcher conducting the test will be able to answer your questions
(Researcher's name and contact details)

……

……

B.2 Consent Form

CONSENT FORM

Title of Project: Efficient Representation of Head Related Transfer Functions

Participant Identification Number for this trial:

Name of Researcher: _____

Please tick box and sign.

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐

2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. ☐

3. I understand that some things I say to the researcher may be quoted anonymously in project reports. I give permission for anonymous quotes to be used. ☐

4. I agree to take part in the above study. ☐

_____	_____	_____
Name of Participant	Date	Signature
Name of Person taking consent	Date	Signature

When completed, 1 for participant; 1 for researcher file