Individualized Head-Related Transfer Functions: Efficient Modeling and Estimation from Small Sets of Spatial Samples

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Griffin D. Romigh
B.S., Biomedical Engineering, Wright State University
M.S., Electrical and Computer Engineering, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

December 5, 2012

Acknowledgments

This work has been shaped by many great individuals who have unselfishly shared their passions and knowledge with me as I have struggled to absorb and appreciate all that I have been introduced to here at Carnegie Mellon, back at the Air Force Research Labs, and at home. I would like to to thank several of these individuals in particular.

First, I'd like to thank my committee members whose guidance and encouragement have helped advance this work since its infancy as a fuzzy idea I brought with me to CMU. Dr. Marios Savvides brought life and interest to my pattern recognition class, and helped me recognize and appreciate the good, the bad, and the ugly. Dr. Bhiksha Raj inspired the first iteration of this work in his machine learning for signal processing course, and opened my eyes to the power of signal processing as a tool for discovery. Dr. Douglas Brungart gave me my first real job several years ago as a freshman at Wright State; He perpetually pushed me with projects well outside my limited knowledge base, and is undeniably responsible for a great amount of my success. And finally Prof. Richard Stern, my advisor, and perhaps the one person at CMU willing to stick with a kid foolish enough to change his major, state, and marital status all his first year of graduate school. He has patiently provided me with invaluable amounts of practical guidance and instruction while giving me the space and confidence to pursue my own ideas, something for which I am continually grateful.

Second, I would like to thank my colleagues and mentors at the Battlespace Acoustics Branch; to Dr. Brian Simpson and Dr. Nandini Iyer who along with Doug have raised me up as part of the 6.1 team I will soon join officially, and to Dr. Todd Nelson and Robert McKinley, who help guide me through the chaos that comes along with my unique funding and employment situation. Also, I would like to acknowledge the Science, Mathematics and Research for Transformation (SMART) scholarship program for funding me through my degree program as well as the financial and facilities support provided by the 711th Human Performance Wing of the Air Force Research Labs.

Lastly, I would like to thank my family. My parents, Libbie and Terry Romigh, who instilled in me an appreciation for knowledge, but much more than that, they taught me that a strong work ethic, creativity, and resourcefulness were the tools that give knowledge its real power. My grandparents, Larry and Evelyn Romigh, and James and Marlyn Stacey, who have given me great pride for where I came from. My older siblings, Adam, Taylor, and Drew for making sure I, as the baby, never got it too easy, and for always giving me an example to aspire to. And lastly, to my wife, Emily Romigh, who inspires me to be more, and extracts me from my thoughts when I am lost to the world around me.

Abstract

This dissertation develops and evaluates a novel way of modeling and estimating individualized head-related transfer functions (HRTFs) from *a priori* information and a limited number of acoustic measurements. Head-related transfer functions represent and describe the acoustic transformations caused by interactions of sound with a listener's head, shoulders, and outer ears which give a sound its directional characteristics. Once this transformation is measured, it can be combined with any non-directional sound to give a listener the perceptual illusion that the sound originated from a designated location in space. The ability to manipulate these spatial auditory cues has many applications for both immersive and informational spatial auditory displays (SADs). Unfortunately, high fidelity SADs require the use of individualized head-related transfer functions that are measured for the specific end user of the spatial auditory display. These measurements are typically impractical for widespread commercialization, in large part because traditional HRTF measurement techniques require several hundred spatial locations to be measured around a listener, a procedure which requires expensive and complicated equipment, and which can take upwards of two hours.

This dissertation introduces a novel way to represent a spatially continuous HRTF with a relatively small number of parameters via a spherical harmonic decomposition. This representation is shown to be perceptually equivalent to a full HRTF in terms of the resulting localization accuracy, while providing a convenient form for making HRTF comparisons across individuals and spatial locations. With this new framework it is shown that HRTFs from a large group of individuals can be modeled effectively with a multivariate normal distribution, and that this underlying distribution can be used to provide *a priori* information for individualized HRTF estimation allowing accurate estimates from as few as 12 spatially distributed locations.

The new representation is also shown to be particularly well suited for the separation of individual and non-individual components of the HRTF. Analysis of the underlying distributions characterizing the HRTF, as well as perceptual evidence, indicates that only the spatial variation in an HRTF which corresponds to the vertical and front-back dimensions of a sound source location need to be individualized. This finding forms the basis of what is referred to as the *sectoral HRTF model*. The sectoral HRTF model is shown to be valid in terms of localization accuracy and provides a way to capture the individual components in an HRTF with only nine parameters at a single frequency. The relationship between these parameters and the spatial variance along the vertical and front-back dimensions is also exploited to allow relatively accurate HRTF estimation from acoustic measurements that are limited to a single plane. Together, these developments provide significant advancements in the ease with which individualized HRTFs are collected, as well as a wealth of new information relvent to efficient HRTF modeling.

Contents

List of Figures

Chapter 1

Introduction

Evolving in a world which is dark for half of their lifetime, humans have developed the innate ability to survey their environment by harnessing even the most minute vibrations in the world around them. This ability to detect, localize, and track sounds has long served as a perceptual anchor, rooting us to our surroundings in the midst of both darkness and distraction. From the creak of a floorboard late at night, to the scream of the subway on the daily commute, our spatial hearing capability continues to provide an instinctive and natural avenue through which we are continually gathering information. While most people today will get to old age without fully appreciating this ability, spatial hearing has long captured the attention of psychologists and engineers who seek to understand and exploit its power.

For more than a decade, the auditory analog to the visual computer monitor has been the focus of much research. This technology, the spatial auditory display (SAD), has the potential to serve as a more intuitive and natural channel through which humans can interface with the new wired world around us. Slated for applications as diverse as attitude displays for pilots, navigational aids for the blind, and augmented reality for entertainment systems, spatial auditory displays hold a concrete place in the future of our ubiquitous computing and informational systems. Unfortunately, potential uses for spatial auditory displays have outgrown the technology itself, leaving an information gap standing in the way of widespread commercialization of SAD-based technologies. At the heart of the issue is the need to personalize spatial auditory displays to the individual characteristics of the auditory systems of the actual users of SADs. Current technology that has been developed to accomplish this personalization requires a multitude of complex laboratory-grade acoustic measurements, which are generally infeasible even for most well-equipped laboratories.

The set of measurements needed for a personalized spatial auditory display are collectively known as the head-related transfer function (HRTF)[1]. Head-related transfer functions capture the physical transformations that are imparted on a sound by interactions with the head, shoulders, and outer ear, as it travels from a specific location in space and arrives at the ear canal. This directionally-specific transformation can later be applied to any monaural sound source to give it the physical characteristics needed for accurate perceptual localization. Since an individual's head, shoulders, and outer ears are distinctively different from person to person, the resulting HRTFs vary between individuals. Because of this, the use of SADs based on another individual's HRTF, or some type of average HRTF, often result in poor localization along with the percept appearing to be "inside" the subjects head rather than externalized.

The aim of this dissertation is to provide new insights and capabilities towards the personalization of spatial auditory displays based on harnessing our current understanding of perceptual spatial auditory phenomena, as well as the exploitation of the physical similarities that are inherent in the HRTFs different individuals. Because we currently lack a complete understanding of how perceptual localization cues manifest themselves physically in the HRTF, the current work has progressed by combining experiments designed to extend our current scientific knowledge of spatial auditory phenomena with the development of techniques that provide practical engineering solutions aimed at simplifying the collection

[1] There is some inconsistency in the literature concerning whether the term "HRTF" refers to the transfer function in general, as a function of the three-dimensional coordinates of the sound source, or whether "HRTF" refers to a measurement taken from a specific sound source location. We use the the term "HRTF" in the former sense (*i.e.* in reference to the general response from all source locations) and we introduce the term "sample HRTF" to refer to a measurement taken from one specific sound source location.

of individualized HRTFs. The dissertation is broken into three independent papers, each one adding to our knowledge of HRTF structure, and each one providing a relevant and practical approach to simplifying the estimation of individualized HRTFs.

Following a general review of previous work leading up to the present dissertation, the first paper describes a novel continuous HRTF representation based on a spherical harmonic decomposition. This representation is based on a number of underlying perceptual phenomena related to spatial hearing, and is shown to provide a convenient form for use in future HRTF modeling and estimation techniques. The representation is assessed in terms of ability to describe and predict the physical features contained in the HRTF and the perceptual cues implied by it.

The second paper introduces a new way to think about individualized HRTFs, representing them as a single sample from an underlying distribution representing the HRTFs of all individuals. An estimation technique that enables individualized-HRTF estimation from a small number of spatial locations is developed by exploiting available *a priori* information about the underlying HRTF distribution contained in previously-recorded HRTFs of other individuals. This paper also discusses a way of accurately estimating the parameters of the underlying distribution, and provides a perceptual validation of the overall process. Additional information concerning this estimation process is provided in Chapter 6 which addresses several practical implementation concerns related to the estimation strategy.

The third paper presents the sectoral HRTF model, which is a novel model for individualized HRTFs that separates the individualized and non-individualized components of an HRTF. This further limits the number of parameters which are needed to describe an HRTF, and leads to a new method designed to estimate an HRTF from a set of measurements taken on a single plane, which could greatly simplify traditional HRTF collection systems. The study includes a perceptual validation of both the assumptions of the sectoral model and the estimation strategy itself.

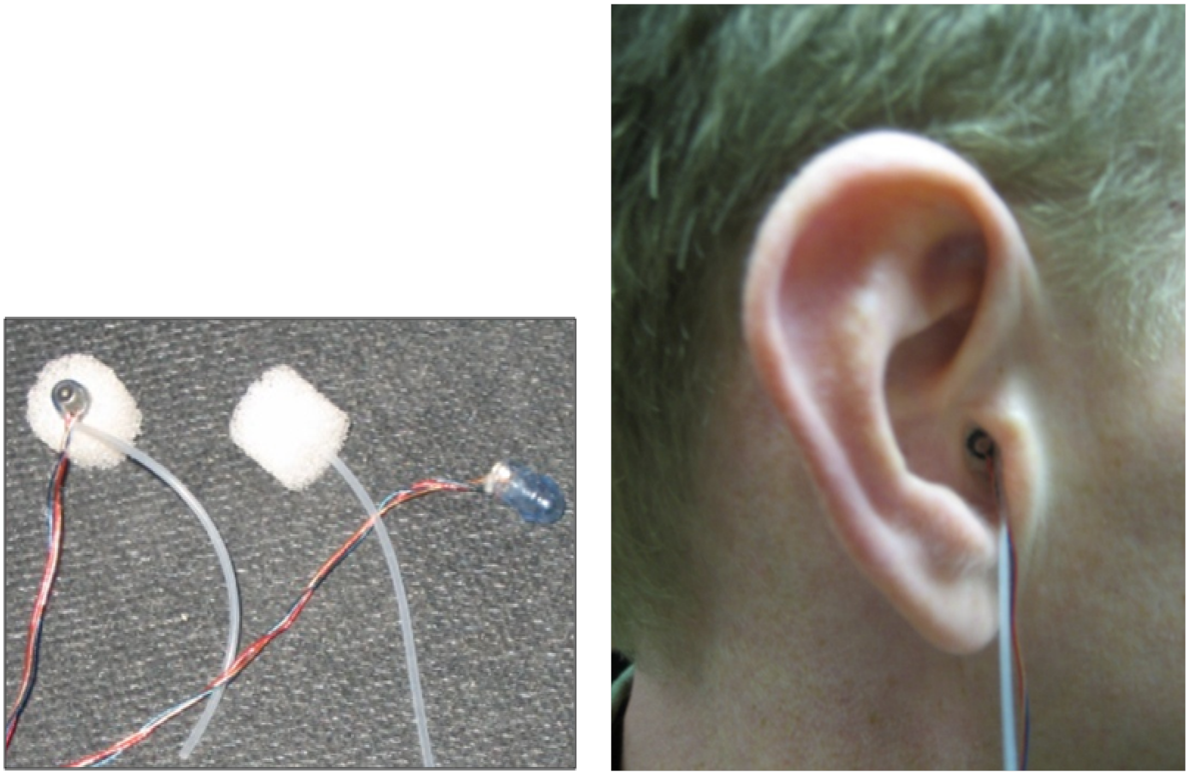
Finally, the last section of the dissertation provides a summary of the novel contributions of the dissertation, suggests several new research avenues that are initiated by the present work.

Chapter 2

Background

2.1 Spatial Hearing and the Head Related Transfer Function

The need for individualized HRTFs has been the greatest impediment for SAD commercialization due to their highly idiosyncratic nature. HRTFs are traditionally modeled as a linear time-invariant systems with a corresponding transfer function, the HRTF. Strictly speaking, the HRTF for a single ear is a complex function of both frequency and three-dimensional space. The dependence on distance is negligible after roughly one meter, allowing us to focus on a range-independent far-field HRTF [1]. Also, humans appear to be sensitive only to HRTF changes that correspond to a spatial separation of 5 to 10 degrees [2], meaning that the continuous HRTF can be approximated reasonably well by a few hundred spatial samples at a fixed radius. The HRTF measured for a source at a single location, hereafter referred to as a *sample HRTF*, can be acquired using any number of traditional system identification techniques. The general process involves playing a test signal with known characteristics from a loudspeaker placed at the desired spatial location. A recording of the test signal is then made using two miniature microphones placed inside the subject's ear canals, as shown in Fig. 2.1. By comparing the test signal to the resulting recording, an estimate of the



(a) (b)

Figure 2.1: Miniature binaural in-ear microphones (Knowles FG 3329) used for HRTF recording. a) Microphone and eardam assebly. b) Microphone inserted into ear canal.

system can be generated under LTI assumptions. This general process is well studied and documented [3, 4, 5, 6].

While the technology to capture and replicate the spatial auditory cues contained in a head-related transfer function has been around for over two decades, two important open questions linger in spatial auditory research:

- What are the physical HRTF features responsible for spatial auditory perception?

- How do these features vary across individuals?

Understanding the physical cues necessary for accurate spatial auditory perception has been a goal of researchers since the early work of Lord Rayleigh and his duplex theory of sound localization [7]. This theory centers on the fact that differences in the sounds arriving at our two ears enable us to determine a sound source's location. Specifically, interaural timing differences (ITDs) at low frequencies and interaural level differences (ILDs) at high frequencies increase as a sound source moves from the midline to the side of the body, and thus can be used as lateral localization cues [8]. This theory has been routinely verified for lateral localization judgments but fails to explain localization in the vertical or front-back dimensions where the interaural cues become ambiguous at any constant lateral angle. The trace of this ambiguity for any lateral angle is referred to as a cone of confusion. This breakdown of perceptual auditory cues into lateral cues and "intraconic" cues has lead to the use of the interaural-polar coordinate system presented in Figure 2.2 to describe most spatial auditory phenomena.

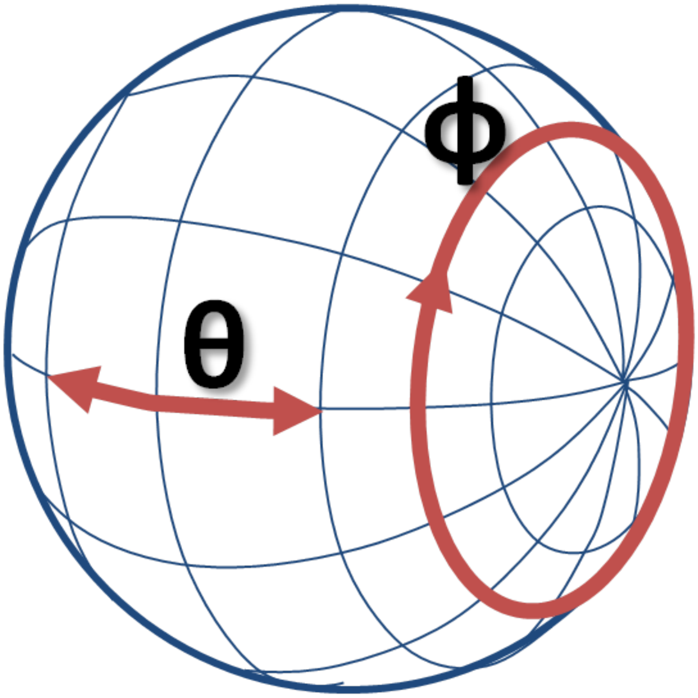


Figure 2.2: Interaural Polar coordinate system. The parameter θ represents the lateral direction and ϕ represents the intraconic (polar) dimension.

While the ITDs and ILDs that are the basis for localization along the lateral dimension have been studied for decades and are relatively well understood, the physical cues responsible for intraconic localization are harder to analyze. Most researchers agree that these localization judgments are based on features of the monaural magnitude spectra [9, 10, 11, 12], and that spectral differences between the two ears are not exploited [13]. Both monaural spectra are used however, but they are perceptually weighted to favor the ipsilateral ear. The contralateral spectrum continues to affect perception until the level difference at the two ears is consistent with a sound source at approximately 30 degrees in lateral angle [14]. The exact physical qualities of the spectra that contribute to the localization judgment are less well understood. Some researchers have proposed that intraconic judgments are based on finding the location that corresponds to a spatial maximum of the HRTF at the center frequency of the stimulus. This "covert peak" idea stems from work with narrow band sources and is hard to generalize to broadband sounds [15]. Other researchers have noted a prominent spectral notch around 8 kHz that increases in frequency as the elevation of the sample HRTF location increases, and point to this as a key feature for elevation discrimination [16, 17]. Even studies aimed at determining the relative importance of spectral peaks versus spectral notches have been unsuccessful at determining their perceptual impact [18]. Evidence also supports the idea that the levels in specific frequency bands are used to obtain front-back and vertical information [1]. The varied theories on spectral cue structure lead to the conclusion that the underlying mechanism is likely complex and certainly not yet well understood.

Despite the fact that the exact physical structure of HRTF cues is unknown, several studies do shed some light on how the underlying cues must differ amongst individuals. One of the first perceptual studies to explicitly show the need for individualized HRTFs was that of Wenzel et al., who studied the localization performance with non-individualized HRTFs [19]. An important conclusion was that localization suffered greatly in the intraconic dimension while lateral localization performance was close to normal. The most dramatic errors occurred when subjects perceived the source to be in the opposite front-back hemisphere compared to where it was presented. The rate of occurrence for these front-back reversals increased dramatically in the non-individualized condition. Similar results from Brungart et al. are recreated in Fig. 2.3 showing the same degradation in performance. In this study, subjects were asked to localize white noise bursts of short duration which had been filtered with a sample HRTF of their own (Individualized), another subject (Non-Individualized), or an acoustic mannequin (Kemar). Average absolute angular errors in their localization judgments are plotted in terms of total, lateral, and polar (intraconic) error components, as well as, the percentage of trials which resulted in a front-back reversal.

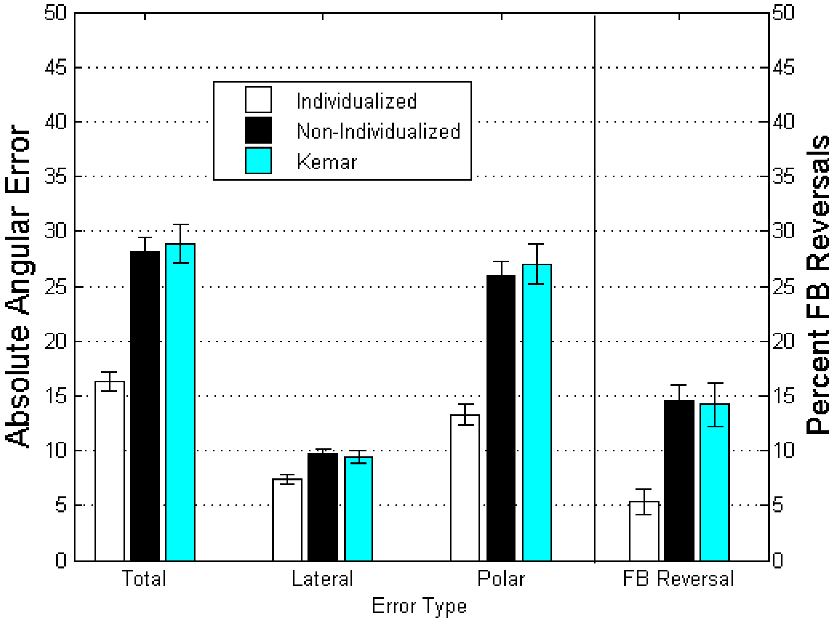


Figure 2.3: Results from a typical perceptual localization study showing performance difference for individualized, average (Kemar) and non-individualized HRTFs.

This study also showed that localization performance could be improved with non-individualized HRTFs by accentuating the spectral variation in the intraconic dimension [20]. While this method never attained the performance equivalent to an individualized condition, it did illustrate that spectral variance in the intraconic dimension might play a key role for vertical and front-back localization judgments. It also showed that some part of the underlying perceptual cue structure must be general in nature and not dependent on individual differences.

2.2 Modeling Head-Related Transfer Functions

Gaps in our current understanding of the perceptually-relevant physical HRTF structures make efficiently modeling HRTFs a significant challenge. Because of this a large number of techniques has been developed to compactly represent sample HRTFs. Most HRTF collection techniques result in relatively inefficient sample HRTFs in the form of long-duration ($\approx$ 50-ms) finite impulse response (FIR) filters [21]. The most traditional of these techniques involves fitting the FIR filter with some time of infinite impulse response (IIR) filter as in [22, 23] and [24] where sample HRTFs could be represented with as few and 10 poles and 10 zeros. Huopaniemi [24] also extended the traditional FIR and IIR filter types by warping the frequency axis to better reflect the frequency resolution of the human auditory system. These frequency-warped filters were shown to provide fidelity similar to traditional filters in a subjective evaluation.

Additional simplifications came along as a consequence of the discovery that listeners are largely insensitive to monaural phase information [25], allowing perceptually-equivalent minimum-phase FIR representations. Because most of the energy in minimum-phase filters is compacted into the early or low-order coefficients [26], these minimum-phase filters could be significantly truncated without affecting localization. Senova *et al.* [27], found that localization accuracy was preserved for minimum-phase-filter durations as short as 320 microseconds. Since the minimum-phase representation is completely described by its magnitude [26], several sample-HRTF models were also designed exclusively around the HRTF magnitude. Kulkarni and Colburn [12], showed that the log magnitude could be accurately represented with as few as 32 Fourier series coefficients. Optimum compression of the magnitude was also investigated by Kistler *et al.* [28] and Martens [29], who investigated principal component analysis (PCA) as a way to model sample HRTFs. With this technique HRTFs could be represented by as few as six components.

While most of the modeling work has focused on simplifying the sample HRTFs, several other methods have focused on modeling the spatial patterns in an HRTF. Jenson [23], investigated fitting the parameters of an auto-regressive moving average sample-HRTF model with von Mises basis functions across spatial locations. Von Mises basis functions are a special form of a symmetric multi-variate normal distribution which has been defined on the surface of a sphere. Also working on a time domain representation, Evans [30] investigated expanding the minimum-phase FIR filter form of the sample HRTFs onto a set of real spherical harmonics, an orthogonal spherical basis. Evans ultimately showed that this expansion was less efficient than expanding the magnitude and phase of the HRTF onto spherical harmonics which allowed 90% of the energy contained in an entire HRTF set to be captured with as few as 64 parameters per frequency. Zotkin *et al.* [31] and Duraiswami *et al.* [32], investigated a similar technique which used the complex spherical-harmonic basis to represent the complex frequency response of the HRTF one frequency at a time.They also provided theoretical bounds which required up to 169 parameters to accurately represent the HRTF at 8 kHz, well above that predicted by Evans. Talagala and Abhayapala [33], also investigated a related technique for expanding the complex spatial frequency response which involved expanding the complex HRTF onto Legendre polynomials in the lateral dimension and a Fourier basis in the intraconic dimension. This technique seemed to result in much more modeling error than the related spherical-harmonic technique, but it may represent a more perceptually-valid decomposition due to its spatial decoupling of lateral and intraconic features. Another efficient spatial representation is that of Xie [34], who performed a spatial PCA on the HRTF magnitudes one frequency at a time, and found that more than 98% of the spatial variance could be accounted for by using only 35 basis functions.

While the above methods have been shown to be successful in efficiently modeling both sample HRTFs and the spatial patterns in HRTFs independently, still other methods have investigated ways of representing both the spectral and spatial patterns of HRTFs simultaneously. Adams *et al.* [35] and Huang *et al.* [36] each proposed ways to model HRTFs using a state-space representation. In this format, sample HRTFs at multiple locations could be modeled simultaneously and efficiently through the use of simplification techniques that had been used in control theory such as balanced model truncation. Grindlay and Vasilescu [37], on the other hand, performed the tensor equivalent of PCA on the three-dimensional complex HRTF matrix accounting for frequency, spatial location, and individualization. This technique was shown to produce smaller amounts of modeling error when compared to the application of traditional PCA to sample HRTFs. One of the most thoroughly-investigated spectro-spatial decompositions is that proposed by Zhang *et al.* [38] which involves expanding the complex HRTF onto a set of orthogonal basis functions that combine complex spherical harmonics in the spatial domain and complex bessel functions in the spectral domain. This technique provides a spatially- and spectrally-continuous representation [39], as well as a theoretically-based upper limit on the spatial bandwidth required to represent an HRTF [40]. The choice of basis was also shown to be near optimal in terms of modeling error [41].

While the above modeling methods provide an abundance of options for compactly representing HRTFs, most of the previous methods do not lend themselves easily to further analysis. The ideal HRTF model would provide a spatially-continuous representation which is characterized efficiently with a small number of perceptually-related parameters. Current methods which map the HRTFs onto spectral basis functions or into state vectors mask the band-by-band nature of the peripheral auditory system and lead to model parameters which are hard to relate back to auditory phenomena. For example, determining how a specific PCA coefficient might affect the spectral notch at 8 kHz is far from intuitive. In a similar way, arbitrarily-oriented spatial bases can provide similar confusions. It is also important to note that the goal of most of the above modeling techniques was the accurate reconstruction of a measured HRTF, despite the fact that the results of perceptual studies have shown that a large amount of the detail contained in an HRTF is perceptually irrelevant. This means that a large amount of the modeling work may be over-fitting the measured HRTFs from a perceptual standpoint, or potentially fitting the perceptually-irrelevant components. Finally, there are no formal HRTF models in the literature which describe the individual differences in an HRTF in an interpretable fashion. The continuous-HRTF representation presented in Paper 1 describes our work toward attaining the this type of ideal HRTF model, while the work of Papers 2 and 3 describe how the inclusion of individual differences in HRTFs can be modeled using this new representation.

2.3 Spatial Sampling and HRTF Interpolation

A fully three-dimensional spatial auditory display requires sample HRTFs at any arbitrary position in space. Despite this, traditional HRTF measurement schemes only capture the the HRTF at a finite number of spatial sampling locations. Conventional HRTF collection systems accomplish this feat by using one of two physical measurement setups. In the first setup shown in Fig. 2.4a, a large fixed microphone or speaker array is used to rapidly collect the sample HRTFs at all the required spatial locations sequentially. In the second setup shown in Fig. 2.4b, a small speaker array (or in theory, a microphone array) can be sequentially moved to account for all spatial locations. A traditional collection with the first method can be accomplished rather quickly in five to ten minutes [21, 42]. Unfortunately, this type of physical setup has a high equipment cost that is often prohibitive. The second setup costs much less for the equipment, but can require up to two hours to collect a single HRTF [31]. Since the speed and equipment required for an HRTF measurement are directly related to the number of spatial samples needed, the more common way to reduce measurement expense is to develop methods for HRTF interpolation that would require fewer spatial samples.

While Ajdler [2] showed that measured HRTFs are theoretically needed every five degrees in azimuth, practical interpolation strategies have been developed which exceed this bound. The simplest of these techniques involves performing a local linear interpolation of the sample HRTFs from nearby sample points. This technique has been shown to accurately reproduce interpolated HRTFs [20], even when as few as 150 HRTF samples are used which

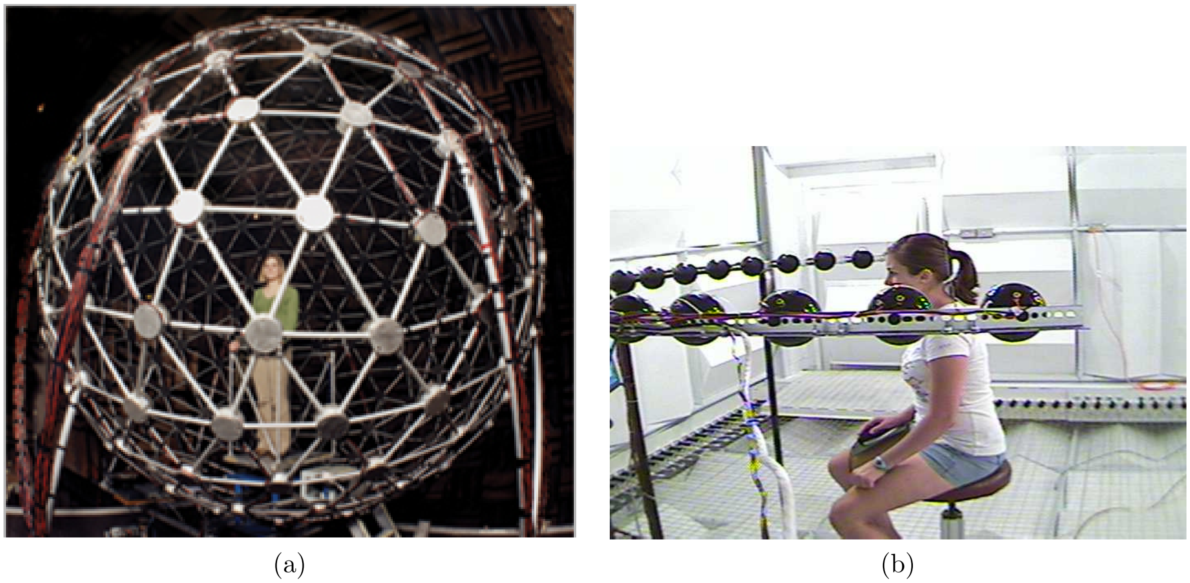

(a) (b)

Figure 2.4: Two exemplar HRTF measurement facilities at the Air Force Research Labs in Dayton, OH. a) Spherical speaker array. b) Single arc speaker array.

are approximately 20 degrees apart [43]. A direct modification of this approach was studied by Sodnik *et al.* [44], and Kentaro and Ando [45], who first performed PCA on the HRTFs and subsequently interpolated the PCA coefficients. Xie [34], also used PCA but in the spatial domain which resulted in HRTFs that were indistinguishable from the original at 65% of the tested locations while only using 73 distributed measurements.

One down side to these approaches is their discrete nature, meaning that a separate computation is required for each new sample point. A more attractive approach would be one which fits the HRTF sample data with a parametric function, allowing for direct computation of an HRTF at any spatial location. Chen *et al.* [46], extended the PCA interpolation approach by interpolating the PCA coefficients using a parametric thin-plate spline. Carlile *et al.* [47] evaluated this technique and found that it performed at least as well as the traditional linear nearest-neighbor (LNN) approach. A similar parametric approach was investigated by Torres and Petraglia [48] who expanded a wavelet representation for HRTFs onto a parametric function. This method was shown to produce smaller reconstruction error than the LNN technique. Enzner *et al.* [49] introduced a radically new interpolation method based on a psuedo-continuous HRTF measurement. In this technique, a continuous HRTF measurement on the horizontal plane was extrapolated to the other elevations. Unfortunetly, this method is incapable of providing front-back localization cues by design.

Another group of parametric interpolation strategies is based on spherical-harmonic expansion. While the spherical-harmonic modeling techniques discussed above by Zhang *et al.* and Duraiswami *et al.* could technically be considered interpolation strategies since they take discrete measurements to a continuous representation, these methods utilized well over a thousand measurement locations to obtain the indicated level of performance, which far exceeds typical interpolation requirements. The method initially introduced by Evans in 1998 [30] used a dense specially-designed Gaussian quadrature measurement grid (648 locations) which enabled reliable estimation of the spherical-harmonic expansion. Most of the work dealing with spherical-harmonic interpolation since then has revolved around issues related to fitting regularization when similar specially-designed measurement grids are not available. Zhang *et al.* [50] introduced a new measurement grid which is more efficient that the Gaussian quadrature and provides the same computational stability. Alternatively, Zotkin *et al.* [31] and Huang *et al.* [51] independently showed that truncated singular value decomposition could be used to provide regularization for less dense measurement grids as well as measurement grids with no spatial samples below a certain elevation.

In addition to these naive interpolation strategies, two groups have proposed methods which use information from other individuals HRTFs to aid interpolation. Guillon *et al.* [52] showed that very low-order spherical-harmonic representations could be used in a cluster analysis to pick a best matched to one of a set of fully-represented non-individual HRTFs that had been collected previously. Alternatively, Lemaire *et al.* [53] used a neural network trained on non-individual HRTFs to help predict HRTF values at unsampled locations. Both methods produced good computational error performance, but were not validated perceptually. This type of method does show promising results, and it provides evidence that non-individual HRTFs can be used to aid interpolation.

All in all, the current interpolation strategies leave much room for improvement. The most efficient methods in terms of the number of spatial samples required are discrete in nature, which means that interpolation will be measurement grid specific. While the spherical thin-plate spline method provided some efficiency with a continuous representation, the underlying functions which result are PCA coefficients, making the representation harder to analyze and relate back to perceptual factors. Very few of the proposed interpolation strategies have gone through any perceptual validation, and fewer still have been shown to provide perceptually-equivalent localization performance. Another shortcoming is that only one method has been proposed which greatly limits the number of locations at which one must sample the HRTF. In other words, although fewer measurements may be needed to implement a particular technique, these measurements are assumed to be distributed across the entire three-dimensional sphere. The one method which did limit measurement distribution relied on an HRTF model which assumed that the HRTF was front-back symmetric, making the resulting HRTF unusable for any front-back localization task. Paper 2 discusses how non-individual HRTFs can be used to regularize spherical-harmonic expansion so that the same continuous interpolation can be estimated from small numbers of distributed sample HRTFs. Paper 3 then goes on to examine how separating out individual and non-individual components in an HRTF permits the restriction of sample HRTFs to a single plane of measurements.

2.4 Personalization of Non-Individual HRTFs

While effective interpolation provides one avenue to simplify individualized HRTF collection, another promising track of research revolves around estimating HRTFs from completely non-acoustic information, which is referred to as HRTF personalization. Silzle [54] and others proposed methods in which subjects listened to non-individualized HRTFs and subjectively selected the one which best suited them based on “naturalness”, or localization performance. These methods have had limited success, but show some improvement in subjective evaluations. Some of the poor performance of these methods may be due to differences in listeners' internal concept of "naturalness" or because of the response errors associated with any perceptual test.

Several additional HRTF personalization methods attempt to relate physical HRTF features to anthropometric measurements of a listeners head size, pinna angle, pinna height, neck height, etc. Wantanabe *et al.* [55] showed that anthropometric measurements could be used to estimate a listener's individual ILD function, which improved accuracy in a horizontal-plane localization task. Both Satarzadeh *et al.* [56] and Hu *et al.* [57] showed that anthropometric measurements could be related to the parameters of structural HRTF models and used to improve modeling accuracy. Middlebrooks [58] also showed that anthropometric measurements could be used to predict a frequency-scaling factor which could be used to improve non-individual localization performance. In contrast to these estimation techniques, Zotkin *et al.* [59] used anthropometric measurements estimated from listener photos to select a best-fit non-individual HRTF from a database, resulting in improvements in the subjective evaluation of the HRTF. This method is similar to that used by Xu *et al.* [60] who used a clustering-based technique to improve non-individual HRTF performance.

While the overall performance of these methods has never reached the performance obtained with fully-individualized HRTFs, it does provide an interesting framework for HRTF estimation that should theoretically be much more fruitful than current results would suggest. One problem with many of these methods is the lack of a simple HRTF representation which characterizes all of the perceptually-relevant HRTF features using a small number of parameters. Paper 1 addresses this missing science by providing a compact and perceptually-valid HRTF representation. Personalization techniques could also benefit from more detailed knowledge of exactly how HRTFs differ among individuals, which is currently scarce. Papers 2 and 3 both provide new information about individual differences in HRTFs, and propose a model which efficiently represents these differences. While the current dissertation limits its investigation to the simplification of acoustic estimation of individualized HRTFs, it is also expected that the models and techniques developed here would also enhance HRTF personalization performance.

2.5 Evaluation of Estimated HRTFs

The purpose of any HRTF is to create the percept of an auditory object located in space. Despite the fact that perception is fundamental to the utility of any characterization of the HRTF, perceptual evaluation of an estimated HRTF is often replaced with evaluations based on a more arbitrary "goodness-of-fit" metric that can lead to overly conservative representations. Lee and Lee [61] compared the perceptual performance of several of these goodness-of-fit metrics and found that methods based on spectral errors appear to best match perceptual performance. While Duraiswami and Raykar [62] also investigated the geometric distance along a learned manifold for evaluating relevant HRTF errors, this method was never compared to perceptual performance in any way. The apparent disconnect between common measures based on numerical error and measures based on perceptual impact implies that any HRTF model-estimation technique needs to be validated with a psychophysical evaluation. These studies generally fall into two broad categories: localization and discrimination.

In a localization task, a subject is presented with an auditory stimulus generated using a modeled HRTF and asked to indicate its perceived direction. Ideally, the spatial distribution of a subject's directional responses with a modeled HRTF will mirror the distribution achieved with a full HRTF. Any quantitative differences in directional response characteristics can then be used for evaluation. Common error metrics for this type of evaluation include absolute localization error, which can then be broken down into directional components, as well as the relative frequencies of certain types of errors such as up-down and front-back reversals, and errors greater than 45 degrees. In a similar manner, a discrimination task involves rendering a sound source with both a full HRTF and a modeled HRTF, corresponding to identical spatial locations, and determining the extent to which the subject can detect the difference.

Both methods provide useful information about an HRTF's fidelity but address fundamentally different questions. A localization task determines whether or not the effects of modeling an HRTF impact the perceived location of a sound source, while a discrimination task determines whether or not modeling effects are perceivable in general. While the difference is sometimes subtle, a sound source rendered with a modeled HRTF could theoretically cause easily distinguishable artifacts while still eliciting an identical spatial percept. In this case, a particular model could be deemed to be perceptually equivalent to a full HRTF in a localization task, but not in a discrimination task. Both methods are used widely in the literature and the best evaluation is sometimes application dependent.

Chapter 3

Paper 1: A Continuous HRTF Representation for Modeling and Estimation

3.1 Introduction

With the advent of human-computer interfaces it has become apparent that the exploitation of human spatial hearing ability could be useful in a number of applications. Generally called a spatial auditory display (SAD) or virtual auditory display (VAD), this technology relies on processing a single-channel sound to cause it to appear perceptually as though it had originated from a designated point in space when played back over headphones. This differs from the traditional "inside the head" percept that occurs with traditional headphone playback. By imparting spatial information to the original sound source, an SAD can be used to provide spatial information to a user for applications such as virtual and augmented reality, navigational aids for the blind, and attitude displays for pilots [63, 64, 65]. The spatial processing is accomplished using a head-related transfer function (HRTF), which characterizes the natural transformations a sound undergoes as it travels from a point in space and arrives at the ear canal. Several researchers have shown that this measurement contains all of the necessary information for accurate sound location [21, 66, 67]

A fully three-dimensional spatial auditory display requires an HRTF for all locations in space. Despite this, traditional HRTF measurement schemes capture the the HRTF at only a finite number of spatial sampling locations. Because of this, sample HRTFs at non-sampled locations must be estimated from the measured samples via some sort of interpolation technique [43]. Recently, a great deal of interest has been given to the use of spherical harmonic expansion as a way to interpolate HRTFs. Theoretically, these methods can be used to represent a continuous HRTF by a relatively small number of expansion coefficients [39]. This makes a SH based representation for HRTFs a promising tool for the development of future HRTF estimation techniques.

To our knowledge there have not been any perceptual studies conducted to investigate the impact of the spherical harmonic expansion on localization accuracy. While several of the authors have shown excellent results in terms of reconstruction errors, it is not clear how these errors relate to perceptual performance. Also, since truncating a spherical harmonic expansion inherently introduces spatial smoothing, a perceptually-validated SH truncation order needs to be determined. While theoretical truncation limits have been calculated for certain types of spherical harmonic expansions, these values have been determined analytically and do not necessarily represent perceptually-relevant limits. The fact that some amount of *spatial* smoothing occurs as the by-product of several frequently used *spectral* simplifications implies that a perceptually-sufficient truncation order might be considerably lower than those determined based on reconstruction error.

This paper introduces a method for modeling HRTFs which combines the advantages of spherical harmonic based expansion with other traditional HRTF modeling ideas. Relevant background information is provided in Section 3.2 dealing with HRTF modeling and SH-based interpolation. Section 3.3 details the innovations provided in the proposed method, and computational and perceptual analyses are described in Section 3.4 which were designed to determine the perceptually-relevant truncation order.

3.2 Background

3.2.1 Modeling the sample HRTF

Typical HRTF measurement techniques provide head-related impulse responses (HRIR) at a finite number of spatial locations in the form of relatively long (of order 1024-2048) with a sampling rate of 44100. Senova *et al.* [27] showed that these filters can be converted to minimum phase and severely truncated while still maintaining perceptual equivalence. In this context, and for the entirety of the paper, the term *sample HRTF* is used to indicate the HRTF evaluated at a single location in space to distinguish it from the HRTF as a whole. While the FIR form of a sample HRTF is convenient for measurement and for FFT-based implementation, it is not an overly efficient representation. One alternative which provides a more compact representation of these filters is the fitting of an infinite impulse response model as in [22, 23]. Frequency warping has also been applied to these conventional filter types to produce even more efficient sample HRTF representations which take advantage of the diminished frequency resolution of human hearing at higher frequencies [24]. Less traditional approaches have also been applied to individual sample HRTFs such as truncated Fourier series expansion of the log-power spectum as in [12]. Here, Kulkarni *et al.* showed that subjects were unable to distinguish between sounds rendered with measured sample HRTFs and sample HRTFs which had a large amount of spectral smoothing. Several "optimal" representations have also been proposed for the sample HRTF such as principal component analysis (PCA) in [29, 28] and Karhunen-Loeve expansion [46]. Together, these studies provide strong evidence that a typically-measured HRTF contains a significant amount of perceptually-irrelevant detail.

In addition to the techniques designed to simplify an HRTF one spatial sample at a time, several techniques have also been developed for modeling an HRTF in the spatial dimension.

Two recent studies have introduced techniques which derive ideal spatial basis functions from previously-collected HRTFs. Huang *et al.* [36] and Xie [34] proposed methods based on Independent Component Analysis and spatial PCA, respectively. These methods offer some of the most efficient spatial representations but suffer from the fact that the basis functions can change depending on the training set. They also require that the entire collection of HRTFs are measured at all the same spatial locations, making comparisons across facilities and methodologies difficult. Another particularly interesting group of methods focus on the expansion of an HRTF onto data-independent spatial basis functions, spherical harmonics. Spherical harmonic based methods offer the advantage that the representations are continuous functions of space and all of the information describing an HRTF is captured in the expansion coefficients. This can make comparisons across different subjects and measurement grids much more straightforward.

Evans was one of the first to investigate spherical harmonic expansion of HRTFs [30]. He showed that efficient representations were possible by expanding either individual time samples from a set of HRIRs, or individual frequencies of both the magnitude and phase responses. Specifically, Evans showed that 90% of the energy in an HRTF could be captured with as few as 64 expansion coefficients per frequency (a 7^{th}-order SH model). When compared to the 300 plus spatial measurements typically contained in a measured HRTF set, this represents a significant amount of savings. Zotkin and Huang both investigated ways of expanding the complex frequency response onto the set of complex spherical harmonics [31, 51]. This technique was adapted by Zhang who chose to expand the complex frequency response onto both complex spherical harmonics in the spatial domain and Bessel functions in the frequency domain [39, 40]. This method produced similar results in terms of efficiency and provides a way to extend the HRTFs to other ranges. A disadvantage of this approach is that the resulting representation makes further analysis difficult because the coefficients are complex and hide any frequency-based features.

3.2.2 Real Spherical Harmonic Expansion

While HRTFs are traditionally viewed as functions of frequency at a single location in space (the sample HRTF), spherical harmonics work on an HRTF's spatial pattern one frequency at a time (sometimes called a spatial frequency response surface). Any continuous square-integrable function can be expanded onto a set of orthonormal basis functions via the generalized Fourier Series. With one-dimensional signals, an infinite set of weighted sinusoids forms this orthonormal basis. For functions defined on the sphere the role of the orthonormal basis is played by the set of real spherical harmonics. The "real" designation is used to disambiguate the real basis from the set of complex spherical harmonics. For most applications, real and complex spherical harmonic representations are interchangeable (with care); the difference being analogous to the relationship between sinusoids and the complex exponential in the traditional one-dimensional Fourier Series [68].

Real spherical harmonic functions $Y_{nm}(\phi, \theta)$ take the form of Eq. 3.1. The spherical harmonic basis function of a certain order n and mode (degree) m form a continuous function of the spherical angles $\{-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}\}$,$\{-\pi \leq \phi \leq \pi\}$. Spherical harmonic basis functions can be defined for any positive order $\{0 \leq n \leq \infty\}$, but in most applications the order is limited to some finite order P. For each spherical harmonic order n there are $2n+1$ individual basis functions which are designated by the mode number $\{-n \leq m \leq n\}$. This means that for a Pth order spherical harmonic representation there are a total of $(P+1)^2$ basis functions.

$$Y_{nm}(\phi, \theta) = \begin{cases} \frac{(2n+1)}{4\pi} P_n^m(\cos(\frac{\pi}{2} - \theta)) & \text{if } m = 0 \\ \frac{(2n+1)}{2\pi} \frac{(n-|m|)!}{(n+|m|)!} P_n^m(\cos(\frac{\pi}{2} - \theta)) \cos(m\phi) & \text{if } m > 0 \\ \frac{(2n+1)}{2\pi} \frac{(n-|m|)!}{(n+|m|)!} P_n^{|m|}(\cos(\frac{\pi}{2} - \theta)) \sin(m\phi) & \text{if } m < 0 \end{cases} \tag{3.1}$$

In Eq. 3.1, P_n^m represents the associated Legendre polynomial of order n and degree m. Associated Legendre polynomials are typically defined in terms of the traditional Legendre polynomials as given in Eq. 3.2.

$$P_n^m(x) = (-1)^m (1-x^2)^{m/2} \frac{d^m}{dx^m} P_n(x) \tag{3.2}$$

Where, $P_n(x)$ is given by Rodrigues' formula:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^m}{dx^m} [(x^2-1)^n] \tag{3.3}$$

Equation 3.4 describes spherical harmonic synthesis, where an arbitrary continuous spatial function $f(\phi, \theta)$ can be formed by the sum of the set of weighted Pth order spherical harmonics. The set of weights, C_{nm}, are known as the spherical harmonic coefficients and carry information about the how function $f(\phi, \theta)$ varies across space. These coefficients can be obtained by the inverse relationship given in the analysis Equation 3.5, where * represents complex conjugation.

$$f(\phi, \theta) = \sum_{n=0}^{P} \sum_{m=-n}^{n} Y_{nm}(\phi, \theta) C_{nm} \tag{3.4}$$

$$C_{nm} = \int_0^{2\pi} \int_0^{\pi} f(\phi, \theta) Y_{nm}^*(\phi, \theta) \sin\theta d\phi d\theta \tag{3.5}$$

This type of analysis requires the integration of a continuous function over the whole spherical angle. When only spatial samples of the underlying continuous function are available (as is often the case for HRTF measurements), an alternative analysis procedure is needed. Given S discrete spatial samples taken at the coordinates $\{\phi_i, \theta_i\}_{i=1}^{S}$, a discretized version of Equation 3.5 can be formulated utilizing an additional measurement-grid-specific set of weighting coefficients w_i as in Equation 3.6.

$$C_{nm} = \sum_{i=1}^{S} w_i f(\phi_i, \theta_i) Y_{nm}^*(\phi_i, \theta_i) \tag{3.6}$$

In practice, how to determine the appropriate set of weights w_i for an arbitrary measurement grid is an open research question and only a small number of valid grid-weight combinations are known to exist. Since most HRTF measurement systems were not designed to use one of these ideal grids, an estimation technique must be used to acquire the SH coefficients. A common way to accomplish this task is by forming a system of linear equations using the discretized version of Equation 3.4 repeated S times, one for each spatial location $\{\phi_i, \theta_i\}_{i=1}^{S}$. This system is given in matrix form by Equation 3.7.

$$\mathbf{f} = \mathbf{Y}\mathbf{c} \tag{3.7}$$

$$\begin{aligned} \text{where } \mathbf{f} &= [f(\phi_1, \theta_1), f(\phi_2, \theta_2), \cdots, f(\phi_S, \theta_S)]^T \\ \mathbf{c} &= [C_{00}, C_{1-1}, C_{10}, C_{11}, \cdots, C_{PP}]^T \\ \mathbf{Y} &= [\mathbf{y_{00}}, \mathbf{y_{1-1}}, \mathbf{y_{10}}, \mathbf{y_{11}}, \cdots, \mathbf{y_{PP}}] \\ \text{and} & \\ \mathbf{y_{nm}} &= [Y_{nm}(\phi_1, \theta_1), \cdots, Y_{nm}(\phi_S, \theta_S)]^T \end{aligned}$$

In this form, provided we have more spatial samples than coefficients, we can use conventional linear algebra techniques to find the unique least squares estimate of the coefficient vector **c** according to Equation 3.8.

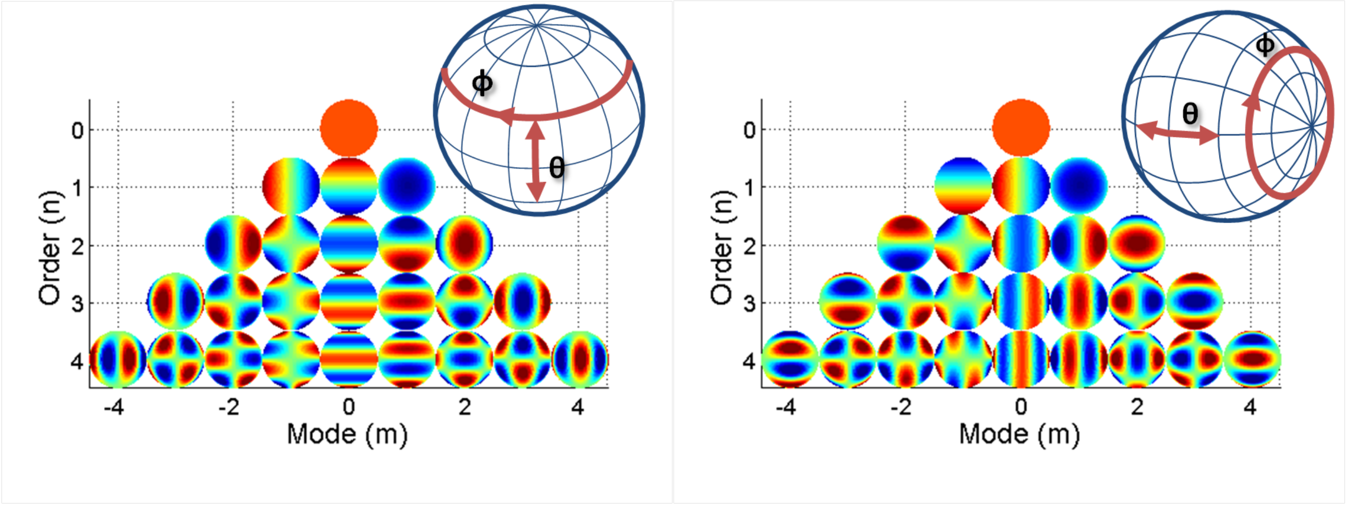
$$\hat{\mathbf{c}} = (\mathbf{Y^T Y})^{\mathbf{-1}} \mathbf{Y^T f} \tag{3.8}$$

In practice, due to noise in the measurements, this method usually requires twice as many spatial samples as the number of SH coefficients to obtain satisfactory results and may need additional regularization if entire portions of the sphere are not sampled; a scenario that often arises due to the limitations in typical HRTF measurement setups at low elevations [31].

3.3 A New SH Based Representation

While all of the spherical harmonic based techniques mentioned in Section 3.2 make good candidates for an interpolation strategy, we propose a new representation which provides similar performance in terms of efficiency but results in a set of spherical harmonic coefficients that we feel are more easily interpretable. The proposed HRTF representation is loosely based on the technique of Evans with a few modifications to better align with our current understanding of how HRTFs are interpreted perceptually. The first modification results from Kulkarni and others who showed that humans are relatively insensitive to the monaural phase structure of a sample HRTF and tend to use only the differences in phase between the two ears for localization[28, 25]. Under this assumption, the difference in phase delay between the ears can be extracted from the individual monaural phase responses and averaged across low frequencies to get a single inter-aural time delay value (ITD). The full HRTF can then later be reconstructed from the magnitude response alone under minimum phase assumptions. This differs both from Evans, who interpolated the entire phase response, and the several authors who have focused on the complex frequency response which incorporates the entire phase response. This removal of the percepually irrelevent phase information allows us to represent the relevent timing information contained in an HRTF much more efficiently, since only one number (the ITD) needs to be represented for each location. The second modification to Evans' technique is to express the magnitude in decibels rather than on the linear scale in order to better reflect perceptual scaling, a decision shared by Guillion et al [52]. The choice to represent an HRTF magnitude in the decibel rather than linear scale has a long history in HRTF interpolation strategies [43] and also ensures that the zeroth-order SH coefficients represent the diffuse-field, or spatial-average, HRTF. Under these modifications the function **f** in Equation 3.8 can represent a set of ITD values or the decibel-magnitude of all spatial samples at a single frequency.

The angles parameterizing the spherical harmonic functions described above can, in general, have any arbitrary orientation with respect to the listener. In most literature, spherical



(a) Vertical Polar

(b) Interaural

Figure 3.1: Real spherical harmonic basis functions for orders 0 though 4. (a) Conventional, vertical-polar coordinate system. (b) Interaural coordinate system.

harmonics are oriented so that θ represents an angle of elevation while ϕ represents the azimuth angle about the upward axis. This orientation is depicted in Figure 3.1a along with the SH basis functions which result. As can be seen, basis functions in which $m = 0$, the *zonal harmonics*, represent variation purely along the up-down or elevation direction while basis functions in which $|m| = n$, the *sectoral harmonics* represent variation almost exclusively in the azimuthal direction. The other basis functions, known as *tesseral harmonics*, capture variation along both directions and are generally less easy to analyze. While this orientation might be useful for some applications, we choose to rotate the conventional orientation to match the perceptually-rooted angles of the interaural-polar coordinate system. This coordinate system is characterized by a θ angle in the lateral direction, and a ϕ angle which represents the angle around the interaural axis as shown in Figure 3.1b. This distinction is important in spatial auditory literature because its coordinates closely correspond to the way auditory cues are processed to make localization judgments. A person's inter-aural cues for timing and intensity are able to differentiate lateral position but not along the set of locations at any particular lateral angle. The judgments within these "cones of confusion" seem to be made independently using mostly monaural spectral cues [15]. For this reason, the two angles are referred to as the lateral angle and intraconic angle. The first 25 basis functions which result from this orientation are shown in Figure 3.1b. As can be seen, the zonal harmonics are now aligned to capture left-right or lateral variation, while the sectoral harmonics are aligned to capture variation along the intraconic direction.

3.4 Determination of Representation Order

3.4.1 Computational Analysis

To gain a better insight into the effect this representation may have on perception, we can first examine the effects of the representation on several of the physical characteristics of an HRTF known to be important for localization. In a manner similar to one-dimensional Fourier Series expansion, higher-order SH coefficients reflect the amount of energy in higher "frequency" components of the underlying function. In this context, frequency can be thought of as the rate of spatial variation as the spherical angles change. Because of this relationship, lowering the SH order of a representation causes a spatial smoothing of the underlying function in a similar way to lowpass filtering of a one-dimensional signal. Depending on how much critical information is removed by the smoothing process, severe order truncation may be possible. To provide additional information about the potential for information loss, Figure 3.2 shows how the energy in the 14^{th}-order representation is divided among the lower order representations. The percentage of total energy is given as a function of representation order at 6 different frequencies. Each line represents the average value for a 1/3-octave band of frequencies centered on the indicated particular frequency. At all frequencies, at least 90% of the energy in the higher-order representation is contained in the fourth-order representation and nearly 100% of the energy is contained in the eight-order representation. While the percentage of energy cannot be directly linked to perceptual impact, it does imply that most of the physical cues likely lie in the low orders.

In Fig. 3.3 the HRTF magnitude of three subjects (each row is a different subject)

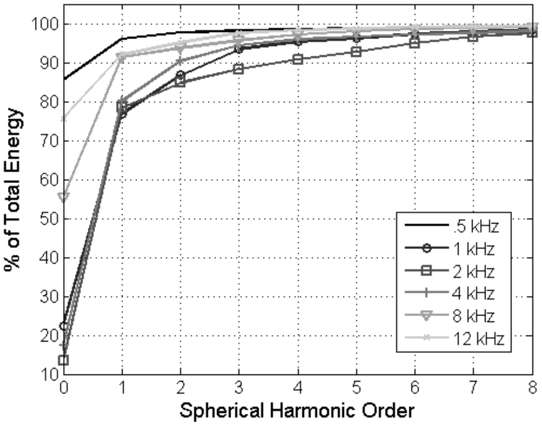


Figure 3.2: Percentage of the total energy in a 14^{th} order SH representation (decibel-magnitude) as a function of SH order for 1/3-octave bands at the indicated center frequencies.

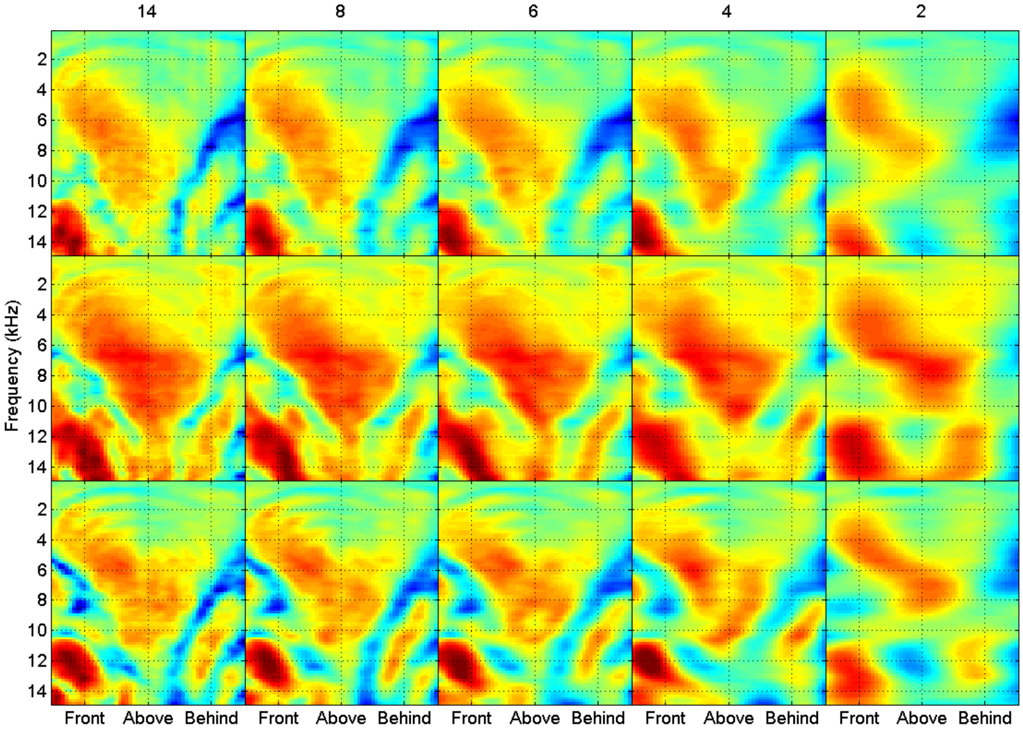
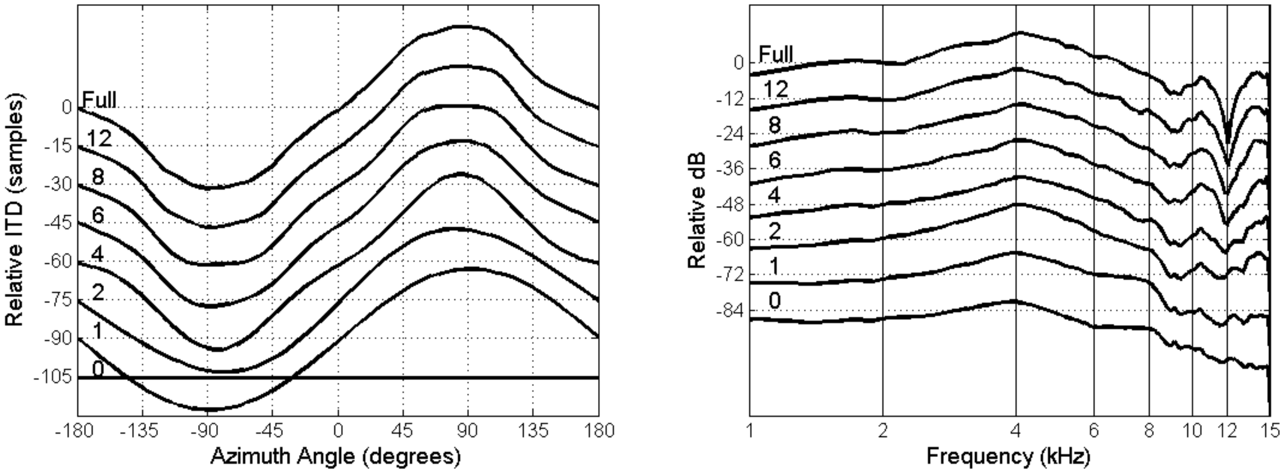


Figure 3.3: Smoothed HRTF magnitude (in dB) as a function of the angle along the median plane for three subjects (rows) and five truncation levels (columns). Red color indicates high magnitude, blue color indicates low magnitude.

are plotted as a function of the angle around the median plane for 5 different levels of smoothing. Here, it can be seen that the HRTFs loose more and more spatial detail as the



(a)

(b)

Figure 3.4: (a) ITD values (in samples) along the horizontal plane for various Spherical harmonic representation levels. (b) HRTF magnitude at zero degrees azimuth and zero degrees elevation for various Spherical harmonic representation levels.

the SH representation order is decreased. In this way, truncation of the SH representation provides a way to smooth the spatial variation of the measured HRTF magnitude.

A similar smoothing can be seen when ITD values are expanded onto SH. Figure 3.4a shows the sample ITD values for one subject taken along the horizontal plane for various SH orders. The ITDs where calculated as the difference between the slope of the best linear fit to the unwrapped phase response at each ear for the frequencies from 300 Hz to 1500 Hz. It can be seen that reducing the SH order smooths out the local variations of the measured ITD function. Dramatic change to the measured ITD function does not occur, however, until the very low orders of zero or one. This agrees with the established assumption that the ITD function can be modeled effectively with a simple sinusoidal model [1].

When the SH order is reduced for all frequencies, this has the effect of taking each sample HRTF closer and closer to the spatial average HRTF (the sample HRTF acquired by averaging the sample HRTFs from all measured locations). Since most spectral features of an HRTF are localized in space, this spatial averaging results in much fewer spectral variations as well. This effect can be seen in Figure 3.4b. Here, the sample HRTF corresponding to the left ear for the location directly in front of the subject is plotted with various orders of SH representation. As can be seen, the prominent spectral notch near 12-kHz is slowly eliminated as the SH order is decreased. It is important to note that the zeroth order representation is exactly the average sample HRTF and would result in an HRTF that is constant across space.

3.4.2 Perceptual Evaluation

The above representation can be made arbitrarily simple in terms of the number of coefficients describing each frequency by truncating the expansion. As seen in Section 3.4.1, this has the side effect of loosing both spatial and spectral detail in the measured HRTF. Because there is currently no way to determine the perceptual impact of these modifications from computational analysis alone, a perceptual localization task was also conducted to determine the representation order necessary to preserve localization accuracy.

HRTF Collection

The perceptual evaluation was conducted at the Auditory Localization Facility (ALF), part of the Air Force Research Labs in Wright Patterson Air Force Base, Ohio. The facility consists of seven-foot-radius geodesic sphere housed in a large anechoic chamber. Each vertex of the sphere contains a loudspeaker (Bose Acoustimass) and a cluster of four LEDs. The chamber also contains a 6-DOF tracking system (Intersense IS900) capable of simultaneously tracking a subject's head position and the position of a small hand-held "wand" pointing device. The system is such that real-time visual feedback can be given to the subject about the orientation of the wand or their head by lighting up the LED cluster which corresponds most closely to the orientation direction. During HRTF collection subjects were asked to stand in the center of the sphere with their head oriented toward a designated speaker location. Before each set of test stimuli were presented, the position and orientation of the subjects head was recorded and the corresponding location was modified to correspond to its position relative to the head.

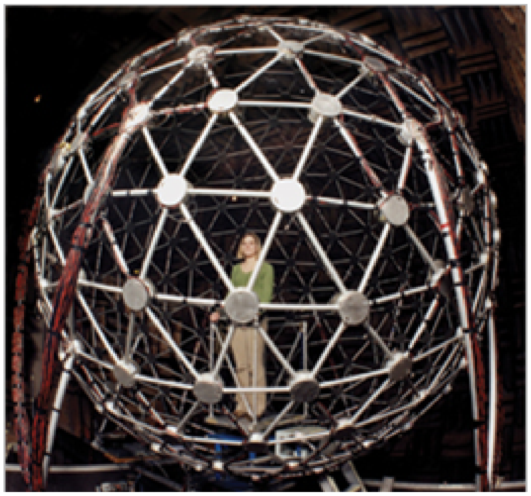


Figure 3.5: Auditory Localization Facility, Wright Patterson Air Force Base, Ohio

The test stimulus consisted of a train of seven periodic chirp signals which swept from 100 Hz to 15 kHz in the span of 2048 points at a 44.1-kHz sampling rate. This 325-ms chirp train was prefiltered to remove any differences in the frequency response between speakers, and presented with the stimuli from 15 other speaker locations with a 250 ms inter-stimulus interval. Binaural recordings were made of the response to each signal. Raw HRTFs were calculated by averaging the response of the five interior chirps of each train and stored as HRIRs (the inverse Discrete Fourier Transform (DFT) of the HRTF). This procedure was repeated until all 277 loudspeaker positions had been measured. A similar technique was also employed to calculate a set of custom headphone correction filters. In this case the test signal was presented overhead phones and recorded with the in-ear binaural microphones. The resulting correction filters were then used to correct the HRTF measurements for the headphone presentation.

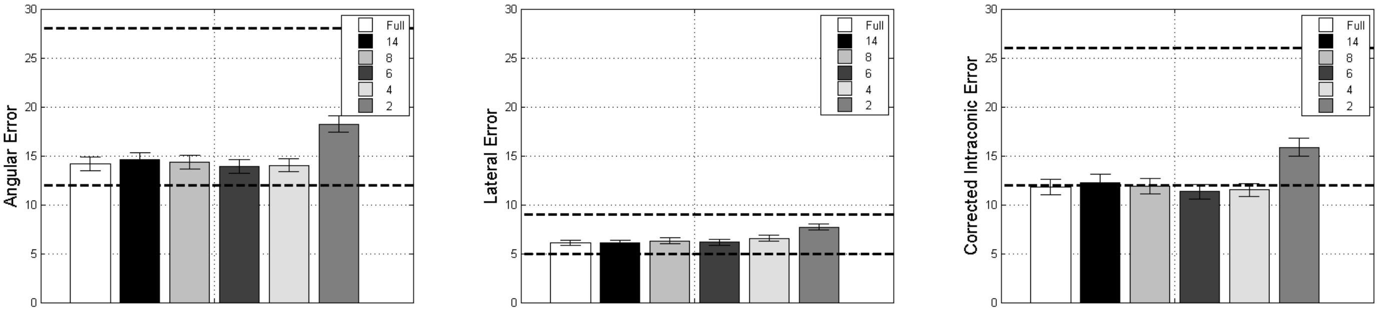
The raw 2048-sample HRIRs were windowed by applying a 401-sample Hanning window centered on the strongest peak of each HRIR to reduce the effects of any residual reflections within the ALF. ITD values were extracted from the raw HRIRs by comparing the best linear fit to the phase response of each ear between 300 Hz and 1500 Hz. The windowed HRIRs were then corrected for the response of the headphones and converted to minimum phase before being truncated to 256 taps with a rectangular window. The ITDs were reintroduced by delaying the contralateral minimum-phase HRIR by the ITD value.

Experimental Task

At the beginning of each 30-minute experimental session, an HRTF and headphone correction were measured using the procedure outlined above. This overall process from microphone fitting to the end of collection took approximately 5-6 minutes after which the subject was asked to complete three 60-trial blocks of a localization task. On each trial the subject was presented with a short stimulus and asked to indicate the perceived direction by orientating the tracked wand toward the perceived location and pressing a response button. The correct location was then presented to the subject by illuminating the LEDs on the actual speaker location, which was then acknowledged via a button press. Subjects were then required to reorient toward the zero-zero direction before they could initiate the start of the next trial by again pressing the button.

Stimulus Generation

All of the stimuli were a 250-ms burst of white noise which had been band-passed between 500 Hz and 15 kHz and windowed with 10-ms onset and offset ramps. The stimuli was convolved with an HRTF and presented to the subject through a pair of custom earphones. All target locations corresponded to one of 245 speaker locations which are above -45 degrees in elevation. Low elevations were excluded from testing because of interference from the subject platform contained in the ALF facility. The HRTFs for all trials within one 60-trial block were generated using the spherical harmonic smoothing technique discussed above for a specific spherical harmonic order. A baseline condition (labeled Full) was also included in the study which consisted of the original processed HRTF with no spatial processing.



(a) (b) (c)

Figure 3.6: (a) Average total angular response error in degrees for all of the tested SH representation orders (b) Average lateral response error. (c) Average intraconic response error corrected for the targets lateral position. Upper and lower dotted bold lines represent non-individual and free-field localization, respectively, from a previous study.

Results

Results of the localization study are shown in Fig. 3.6. Here, average localization error across subject and location is shown for all of the tested HRTF representations. The results in Fig. 3.6a indicate a total angular error of approximately 14 degrees for all of the HRTF types tested with the exception of the second order representation. This 2^{nd}-order representation provided a four-degree increase in total angular error when compared with the Full HRTF. A one-way analysis of variance showed that the HRTFs tested resulted in significantly different means for average angular error ($p < .0001$). A post-hoc Tukey honestly significant difference (HSD) test showed that only the 2nd-order representation provided performance that was significantly different from the baseline full HRTF at a 95% confidence level. Figures 3.6b and 3.6c show the decomposition of localization errors in terms of their lateral and intraconic components. Very little difference in lateral error is seen for any of the tested HRTFs. In this case, most of the difference in performance appears to be intraconic errors where the 2nd-order model provides approximately four degrees worse performance than the baseline condition.

Another important metric for evaluating the success of an HRTF model is the percentage

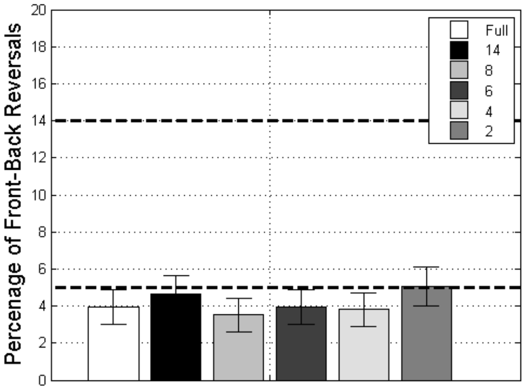


Figure 3.7: Percentage of trials in which the target was localized to be in the opposite front-back hemisphere as where it was intended.

of trials in which the subject commits a front-back reversal. This type of error is often encountered in virtual localization studies, and can be extremely detrimental in practical implementations of an SAD. In this study, a front-back reversal occurs when a subject's response is in the opposite front-back hemisphere compared to the correct target location that results in an error of more than thirty degrees along the front-back dimension. This second criteria keeps conventional mislocalizations along the frontal plane from being categorized as a front-back reversal. Figure 3.7 shows the percentage of trials with a front-back reversal for each of the tested HRTF representations. As can be seen, all of the tested HRTFs resulted in a front-back confusion on 4% to 5% of the trials.

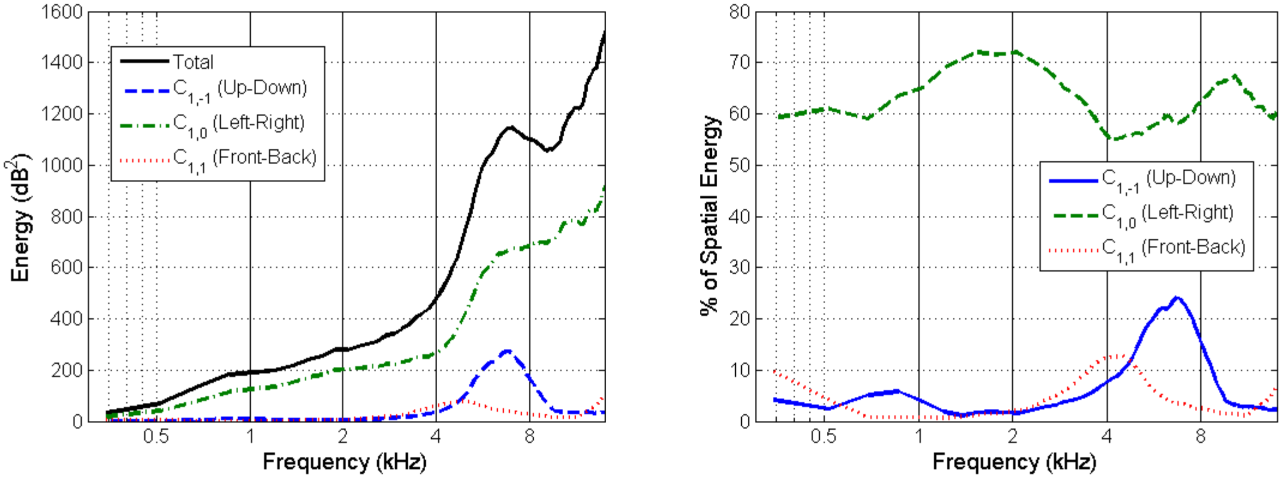
The bold dotted lines in Figures 3.6 and 3.7 show results from a previous localization study conducted using similar methods [20]. The upper lines correspond to performance when subjects were asked to localize a virtual sound source which had been rendered with a non-individualized HRTF and the lower lines correspond to localization performance with free-field sounds. In all cases, the localization errors seem comparable to the free-field performance acquired in the previous study, and even though the 2nd-order model performed significantly worse than baseline, its performance is still far from that of non-individualized HRTFs.

3.5 Discussion

Results of the localization task show that accuracy consistent with a conventional individualized HRTF is possible with the above SH representation of order greater than 4. A comparison of these results to the information presented in Figure 3.3 suggests that a significant amount of spatial smoothing can occur before localization accuracy is affected. In fact, while the 2nd-order representation resulted in noticeably poorer localization performance, it is still significantly better than the localization performance observed in previous studies with non-individualized HRTFs. This result may not be so surprising if we think about it in the context of the spectral smoothing incurred. Similar amounts of smoothing of log magnitudes have been shown to be adequate for sample HRTFs by Kulkarni and Colburn [12]. Nonetheless, the present results provide additional support to the idea that only gross spectral features are necessary for adequate localization.

The above perceptual results seem to indicate that a significant amount of the perceptually-relevant localization cues are contained in the first few spherical harmonic coefficients (at least with the proposed representation). While no such analysis was included in his work, Evans originally suggested that the 3 first-order SH coefficients might be particularly important for understanding localization cues [30]. These coefficients capture variation which is aligned purely in the up-down, left-right, and front-back dimensions for mode numbers of negative-one, zero, and one, respectively. This means that if the HRTF at a particular frequency is well modeled by one of these coefficients, that frequency could be a reasonable cue for the corresponding directional judgment. Figure 3.8a shows the energy (value squared) of these three coefficients for frequencies from 200 Hz to 14 kHz, as well as, the total amount of spatial energy. Here we use the term *spatial* energy to mean the sum of the squared coefficient values for all coefficients from the first through fourteenth orders. The zeroth order coefficient was left out because it represents a spatial constant and therefore contains no directional information.

The overall shape of the total spatial energy function agrees with what is expected con-



(a)

(b)

Figure 3.8: (a) Energy contained in each of the three first-order SH coefficients for frequencies from 200 Hz to 14 kHz. (b) Percentage of the total spatial energy contained in each of the three first-order SH coefficients for frequencies from 200 Hz to 14 kHz.

sidering that head shadow at high frequencies dominates the spatial variation in an HRTF. Also apparent from the total energy function is the dramatic increase in spatial variance after 4 kHz. The region above 4 kHz has long been cited as the most important region for spectral localization cues. Clearly, most of the total spatial energy at each frequency is contained in the $C_{1,0}$ or left-right coefficient. This can be seen more clearly in Figure 3.8b which shows the energy in each coefficient as a percentage of the total spatial energy. This agrees with the many perceptual results showing that the left-right level cue is by far the most robust perceptual cue. The $C_{1,-1}$ or up-down coefficient shows a significant increase in energy around 7 kHz while the $C_{1,-1}$ or front-back coefficient shows increases around 4 kHz and 14 kHz. This would indicate that these frequency regions would be the most useful for making the corresponding directional judgments. These regions also correspond well to Blauert's "boosted bands" which he claimed might be particularly important for "overhead" and front-back judgments [1]. The relative amount of energy in these coefficients also might provide some explanation for the relative robustness of left-right and up-down judgments compared to front-back judgments if we consider that percentage of spatial energy could be interpreted as a signal-to-noise ratio for making a judgment along a certain direction. While these first-order cues are convenient to analyze because of their simple spatial patterns, they clearly do not explain all of the perceptual features contained in an HRTF since the perceptual results indicate that at least a fourth-order representation is needed.

A 4th-order representation contains 25 coefficients per frequency and an additional 25 for the ITD. This means that the 256-bin HRTFs used in the evaluation above can be completely represented with as few as 3225 numbers (using the even symmetry of the magnitude spectrum). This represents a savings of over 95% when compared to the original 256-tap FIR form. This amount of simplification could be valuable to future prediction and estimation efforts since it limits the number of free parameters that would have to be estimated for each new HRTF. This could be particularly important for efficient HRTF personalization methods utilizing more easily collected personal information like anthropometric measurements and subjective evaluations.

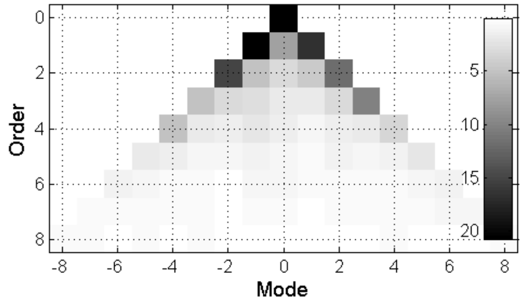


Figure 3.9: Inter-subject variance for the spherical harmonic coefficients of orders 0 through 8 averaged over the frequency region from 4 kHz to 14 kHz.

Lastly, the interpretability of this representation could help inspire new methods for HRTF personalization. One potentially useful observation is highlighted by Fig. 3.9 which shows the inter-subject variance of each coefficient for orders zero through eight averaged over the frequency region from 4 kHz to 14 kHz. Here, a dark color represents a coefficient with high variance, while a light color implies low variance. Clearly, the amount of inter-subject variance decreases with order as you might expect from the results discussed above. More interesting is the fact that almost all of the variation is contained in the sectoral coefficients, the coefficients in which $|m| = n$. Since these coefficients represent variation almost exclusively in the intraconic direction (see Fig. 3.1b), this would imply that the spatial variation along this intraconic direction is what sets one person's HRTF apart from another's. This is an insight that could help to further refine methods aimed at the personalization or estimation of an HRTF by focusing on this intraconic variation.

3.6 Conclusion

This paper introduced a novel HRTF representation based on the spherical harmonic expansion of an HRTF. The representation differers from traditional spherical harmonic based techniques becuase it uses the expanded HRTF decibel-magnitude and ITD, rather than its full representation in either real-and-imaginary or polar form; the representation also uses rotated spherical harmonic basis functions which allows it to capture perceptually relevent information in a more easily interpretable form. Both computational and perceptual results indicate that with a 4th-order spherical harmonic expansion of the proposed type, localization accuracy is preserved to the level of performance achieved with a full individualized HRTF despite that fact that a significant amount of spatial and spectral smoothing takes place with the spherical harmonic representation. The resulting representation was also shown to be related to a number of underlying perceptual phenomena which might make it a good vehicle for future work related to HRTF modeling and estimation.

Chapter 4

Paper 2: Bayesian Estimation of Individualized Head Related Transfer Functions

4.1 Introduction

High fidelity spatial auditory displays (SAD) require the use of individualized head-related transfer functions. This creates the need for a fast, efficient, method for attaining individualized measurements for the future end-users of these SADs. Traditional technologies for making these measurements require a large number of spatial measurements to be taken across a three-dimensional sphere of locations around the user. Facilities capable of making these types of measurements quickly typically contain dedicated spherical speaker arrays so that all of the spatial samples can be taken without the need to move the subject or the array [21]. More commonly, a small circular or semi-circular speaker array is rotated to create a similar spherical sampling pattern. This setup requires much less equipment, but can significantly slow down the overall measurement process due to frequent array rotations.

Due to the relatively high time and equipment costs associated with these traditional approaches, several alternative measurement procedures have been proposed which aim to parallelize the overall process by measuring multiple locations simultaneously. Majdak *et al.* [69], showed that several HRTFs could be measured simultaneously by using spectral-time asynchrony. Here, test stimuli are arranged so different spectral regions are measured at different locations concurrently. Zotkin [42] proposed another strategy based on the reciprocity principle of acoustics. Here, the position of the microphones and speakers in the traditional measurement setup is switched. A tiny speaker (driver) is placed inside the subject's ear canal and plays a sound which is picked up by a spherical array of microphones arranged around the listener. This technique has the advantage that sample HRTFs for all of the measurement locations can be attained for one ear simultaneously. Both methods reduce the time needed for a full HRTF measurement, but still require the same amount of measurements and equipment as traditional approaches. Also, to our knowledge no perceptual localization testing has been conducted to validate these collection techniques.

Another approach for expediting the collection of an individualized HRTF is by reducing the number of locations at which measurements are made. In this way both the time and equipment necessary to make an individualized HRTF are reduced. A large number of these interpolation strategies have been proposed for HRTFs. While most focus on the interpolation of already dense sets of spatial measurements, several investigated the performance when limiting the number of spatial measurements. Martins *et al.* [43], showed that accurate localization was maintained when simple linear nearest neighbor interpolation technique was applied to locations as far apart as 15-20 degrees in azimuth and elevation. This is similar to results achieved by Carlile *et al.* [47] for both nearest neighbor and spherical thin plate spline based approaches. This corresponds to between 150-625 spatial locations depending on how they are distributed.

The above interpolation methods are naive in the sense that no *a priori* information about HRTFs is used when doing the interpolation. Several authors have also proposed methods which use information contained in previously recorded HRTFs from other subjects to help improve the interpolation. Jenison [23], and Lemaire *et al.* [53], both proposed strategies based on neural networks trained to predict HRTFs based on a small number of spatial measurements. Guillion *et al.* [52], used a clustering-based approach to map a set of spatial features derived from a sparse set of measurements to a closely matching dense HRTF. This method was shown to outperform the spline-based naive interpolation method in terms of reconstruction error. Xie [34] also showed accurate reconstruction performance using a principal-component-analysis-based approach which focused on representing variance in the spatial domain. Xie also showed that with as few as 74 spatial measurements approximately 65% of the estimated HRTFs tested were indistinguishable from the measured HRTFs. While several of these techniques show promising results in terms of reconstruction or modeling error, no explicit localization studies have been conducted to determine the exact number of spatial measurements that are required to achieve accurate localization.

The present works builds off previous HRTF measurement and interpolation methods to enable the estimation of a continuous individualized HRTF from a small number of spatially distributed measurements. Section 4.2 first provides a description of the continuous HRTF representation utilized in the method before the Bayesian estimation strategy is laid out in Sec. 4.3. Details regarding the acquisition and processing of *a priori* HRTF information are described in Sections 4.4 and 4.5. Finally, computational and perceptual evaluations of the proposed estimation technique are provided in Sections 4.6 and 4.7. Results indicate that perceptually-valid individualized HRTFs can be estimated from as few as 12 spatial measurements.

4.2 The Spherical Harmonic Representation

An important first step in the estimation process is to provide a parameterized continuous representation of an HRTF. In our method, this process begins with an HRTF inn the form of a 256-tap minimum phase FIR filter for each each and a corresponding ITD value at each spatial location location. Details regarding the extraction of this ITD are given in Sec. 4.5. The log-power spectrum at each frequency and the ITDs are then expanded onto a set of real spherical harmonics. Spherical harmonics are a set of real orthonormal spherical basis functions defined by Equation 4.1. Y_{nm}, the spherical harmonic basis function of order n and mode m, is given in terms of the two interaural coordinates $\{-\frac{\pi}{2} \leq \theta \leq \frac{\pi}{2}\}$ and $\{-\pi \leq \phi \leq \pi\}$. Here θ represents the lateral or right-left angle, and ϕ refers to the angle circling the interaural axis (the imaginary axis running between the two ears) such that the coordinates (0,0) represent directly in front of the individual, and (0,π) is directly behind. $P_n^{|m|}$ represents the associated Legendre polynomial of order n and degree m. Legendre polynomials are defined and calculated recursively as in [31] and [68]. As can be seen in Fig. 4.1, higher order harmonics contain more and more spatial variation the 25 basis functions of order zero through four are shown.

$$Y_{nm}(\phi,\theta) = \begin{cases} \frac{(2n+1)}{4\pi} P_n^m(\cos(\frac{\pi}{2}-\theta)) & \text{if } m = 0 \\ \frac{(2n+1)}{2\pi}\frac{(n-|m|)!}{(n+|m|)!} P_n^m(\cos(\frac{\pi}{2}-\theta))\cos(m\phi) & \text{if } m > 0 \\ \frac{(2n+1)}{2\pi}\frac{(n-|m|)!}{(n+|m|)!} P_n^{|m|}(\cos(\frac{\pi}{2}-\theta))\sin(m\phi) & \text{if } m < 0 \end{cases} \tag{4.1}$$

By expanding the HRTF onto a set of spherical harmonics all of the subject dependent information is conveniently represented by a set of expansion coefficients C_{nm} at each frequency and an additional set of expansion coefficients which describe the ITD at all spatial locations. Previous studies have shown that a 4^{th}-order representation is sufficient to preserve accurate localization, meaning that 25 coefficients per frequency are needed to describe the magnitude as well as an additional 25 coefficients to describe the expanded ITD [70]. Traditionally, for an arbitrary measurement grid these coefficients are calculated using a least squares fit to the discretized SH basis functions using the system of linear equations described in Equation 4.2.

$$\mathbf{f} = \mathbf{Y}\mathbf{c} \tag{4.2}$$

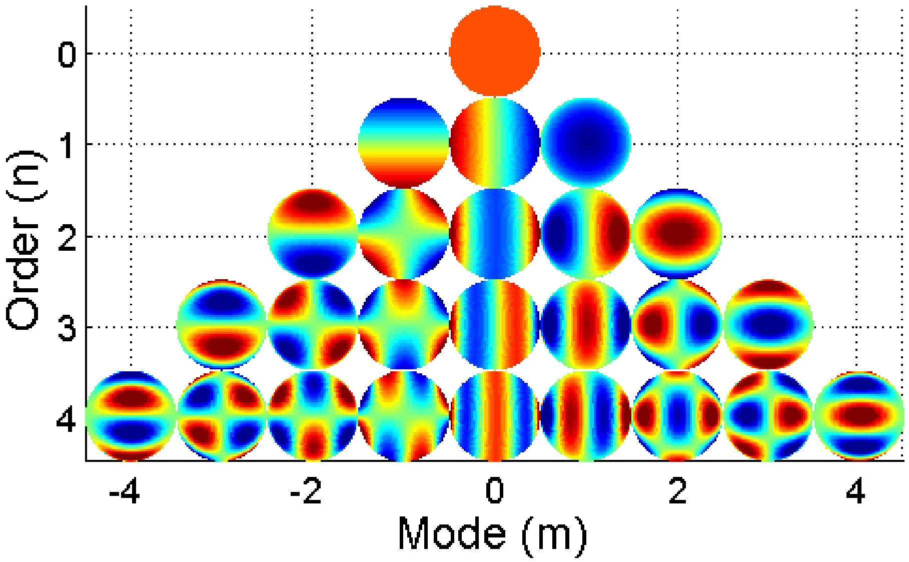


Figure 4.1: Real spherical harmonic basis functions for orders 0 though 4.

$$\text{where } \mathbf{f} = [f(\phi_1, \theta_1), f(\phi_2, \theta_2), \cdots, f(\phi_S, \theta_S)]^T$$

$$\mathbf{c} = [C_{00}, C_{1-1}, C_{10}, C_{11}, \cdots, C_{PP}]^T$$

$$\mathbf{Y} = [\mathbf{y_{00}}, \mathbf{y_{1-1}}, \mathbf{y_{10}}, \mathbf{y_{11}}, \cdots, \mathbf{y_{PP}}]$$

and

$$\mathbf{y_{nm}} = [Y_{nm}(\phi_1, \theta_1), \cdots, Y_{nm}(\phi_S, \theta_S)]^T$$

In this case f is the arbitrary symbol given to a vector of HRTF features at all measured locations in space. In the current method this vector could be a set of decibel-magnitudes at a single frequency or the set of all ITDs. $\mathbf{Y}$ is a matrix containing the discretized SH asis functions, and $\mathbf{c}$ is a vector containing the set of SH expansion coefficients. Additionally, (ϕ_i, θ_i) is used to represent the coordinates of the i^{th} spatial sample from a total of S available measurements,and P represents the SH truncation order. The expansion coefficients can be solved for, given a set of spatial measurements provided in $\mathbf{f}$, by using the conventional least squares solution shown below in Eq. 4.3.

$$\hat{\mathbf{c}} = (\mathbf{Y^TY})^{-\mathbf{1}}\mathbf{Y^Tf} \tag{4.3}$$

For many conventional measurement grids this system is poorly conditioned due to induced correlations in the discretized basis functions. Also, as the number of measurement locations is decreased below the number of SH coefficients, Eq. 4.2 becomes under determined. This means that there are an infinite number of possible solutions and some type of regularization must be utilized to provide a single answer. The least-squares solution provided in Eq. 4.3 provides the single solution from the infinite set which results in the minimal norm. In practice, this technique often results in significant degradation when there are less than two times as many measurements as coefficients. While Zotkin [31] and Huang [51] both proposed methods to overcome these limitations based on truncated singular value decomposition, both of these methods failed to take advantage of the *a priori* information contained in previously measured HRTFs to predict the HRTFs at missing locations. Instead, Sec. 4.3 describes a method which regularizes the expansion by providing an *a priori* distribution for the underlying spherical harmonic coefficient vector, $\mathbf{c}$. This provides the solution to Eq. 4.2 which is most probable given the set of discrete measurements available for that individual.

4.3 Bayesian HRTF Estimation

The Bayesian estimation strategy is based on the idea that every individualized HRTF represents a single sample from some underlying HRTF distribution. If we decompose an HRTF at a single frequency (or ITD) as described above, we can model the underlying HRTF with a multi-variate normal (MVN) distribution on the coefficient vector $\mathbf{c}$. In other words, given some mean coefficient vector $\mathbf{m_c}$ and covariance matrix $\mathbf{R_{cc}}$, the HRTF coefficients are assumed to be distributed as $\mathbf{c} : \mathcal{N}(\mathbf{m_c}, \mathbf{R_{cc}})$. Incorporating this distribution into Eq.

4.2 and adding a term representing white Gaussian measurement noise, $\mathbf{n} : \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$, this gives us a traditional linear system MVN model provided in Eq. 4.4.

$$\begin{aligned} \mathbf{f} &= \mathbf{Y}\mathbf{c} + \mathbf{n} \qquad (4.4) \\ \mathbf{c} &: \mathcal{N}(\mathbf{m_c}, \mathbf{R_{cc}}) \\ \mathbf{n} &: \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}) \end{aligned}$$

Now, the distribution for a single HRTF can be written as:

$$\mathbf{f} : \mathcal{N}(\mathbf{Y}\mathbf{m_c}, \mathbf{Y}\mathbf{R_{cc}}\mathbf{Y^T} + \sigma^{\mathbf{2}}\mathbf{I}) \qquad (4.5)$$

This provides a way to regularize the Eq.4.2. Using the above distribution, we can compute the minimum mean square estimate for the coefficient vector $\mathbf{c}$ given the set of HRTF measurements. This is done by using Equation 4.6 where $\mathbf{f}$ is taken to be the set of measurements for one subject at a single frequency (or the ITD fuction).

$$\begin{aligned} \hat{\mathbf{c}} &= E[\mathbf{c}|\mathbf{f}] \\ &= \mathbf{m_c} + \mathbf{R_{cc}}\mathbf{Y^T}(\mathbf{Y}\mathbf{R_{cc}}\mathbf{Y^T} + \sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}(\mathbf{f} - \mathbf{Y}\mathbf{m_c}) \qquad (4.6) \\ &= \mathbf{m_c} + \mathbf{c_I} \qquad (4.7) \end{aligned}$$

This equation is rewritten as in Equation 4.7 to highlight the fact that with this estimation strategy, a subject's estimated set of coefficients is found by taking the average set of coefficients m_c and adding an "innovations" term, c_I which represents the difference between the average HRTF and spatial samples measured for that particular subject.

4.4 Model Training via the Expectation Maximization Algorithm

The previous section provides a way of estimating of a set of SH coefficients given that the distribution of the coefficients is already known. The problem of acquiring the prior distribution can be thought of as a classical estimation problem because the underlying parameters are believed to be fixed but unknown. Given a significant number of samples of this random coefficient vector, we can use classical maximum-likelihood (ML) estimation to predict the parameters of the prior distribution. With no way of directly measuring SH coefficients, this presents a chicken-or-egg scenario since we need priors to estimate the coefficients, but coefficients are needed to estimate the priors. Fortunately, this type of problem can be handled recursively by using a variant of the Expectation-Maximization (EM) algorithm. For this application, the EM algorithm consists of four steps given below.

Expectation Maximization Algorithm

1. Initialization: Assign arbitrary values to the hyperparameters $\mathbf{R_{cc}}, \mathbf{m_c}$.
2. Expectation: Obtain estimated values for the coefficients based on the current hyper-parameters via Eq. 4.6.
3. Maximization: Obtain new ML estimates of the hyperparameters, based on the current estimates of the coefficients.
4. Recursion: Repeat steps 2 through 4 until changes in estimates fall below a given criterion.

The EM algorithm is a well-motivated and frequently-used heuristic, but as such, there is no formal guarantee of convergence to the true parameter values. It does however come with the guarantee that each iteration will result in a likelihood no less than the previous, meaning more iterations will never result in a less-likely estimate. It is important to note that theoretically this iterative algorithm is only necessary until a significant number of individualized HRTFs have been included in the calculation, at which point inclusion of additional subjects will not significantly change the prior distribution. When this critical number of subjects is reached the iterative method can be replaced by the simple Bayes MMSE technique of Equation 4.6. Ideally, the number of subjects included in the database would be significantly larger than the number of hyper-parameters you are trying to estimate. An additional interesting and useful benefit of this iterative method is that it allows for the incorporation of HRTFs measured on different and potentially arbitrary grids, something which will allow HRTFs from multiple facilities and setups to be included in the estimation process if desired.

4.5 HRTF Database Collection

In order to evaluate the performance of the proposed model, a large database of HRTFs was collated from HRTFs that had been recorded over the previous three years. A total of 54 subjects' HRTF were included; all recorded at the Auditory Localization Facility (ALF) of the Air Force Research Labs in Dayton, OH. This facility consists of a 7-foot radius geodesic sphere housed in a large anechoic chamber as pictured in Figure 4.2. The collection method utilized in this facility has been shown to produce HRTFs which maintain the localization abilities of human subjects with free field stimuli [21].

During an HRTF collection, a test stimulus is played from each of the 277 loudspeakers located at vertices of the sphere. The test stimulus consisted of a train of seven periodic chirp signals each swept from 200 Hz to 15 kHz in the span of 2048 samples at a 44.1 kHz sampling rate. This 325-ms chirp train was prefiltered to remove any differences in the frequency response between speakers and presented to the subject. Binaural recordings were

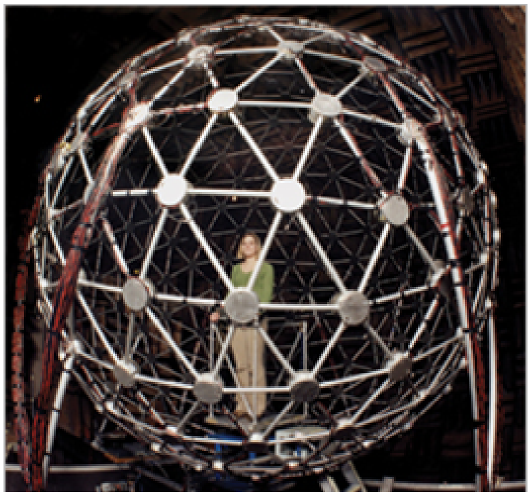


Figure 4.2: Auditory Localization Facility, Wright Patterson Air Force Base, Ohio

made of each stimulus and raw HRTFs were calculated by averaging the response of the five interior chirps of each train and stored as HRIRs (the inverse Discrete Fourier Transform of the HRTF). This procedure was repeated until all 277 loudspeaker positions had been measured. Before the onset of each stimulus presentation, the position of the subject's head was recorded and later used to calculate a head-relative location for storage.

The raw 2048-sample HRIRs were windowed by applying a 401-sample Hanning window centered on the strongest peak of each HRIR to reduce the effects of any residual reflections within the ALF facility. ITD values were extracted from the raw HRIRs by comparing the best linear fit to the phase response of each ear between 300 Hz and 1500 Hz. The windowed HRIRs were then converted to minimum phase before being truncated to 256 taps with a rectangular window.

4.6 Computational Evaluation

To evaluate the proposed estimation technique, HRTFs from 44 of the 54 subjects in the above database were used to estimate the values of the hyperparameters $\mathbf{m_c}$ and $\mathbf{R_c c}$ according to the EM-based method discussed in the previous section for a 6th-order SH representation. HRTFs of the ten remaining subjects, were then used as a test set to evaluate the

Bayesian estimation technique. For each HRTF, the coefficients of a 6th order representation were estimated from a reduced set of the 274 available locations via both the Bayesian and conventional least squares techniques as presented above. The sample locations were picked to be approximately equally distributed on the sphere, and varied from one HRTF to the next. The mean square error (MSE) between the coefficients estimated using the reduced set and the coefficients found using all 274 locations is plotted in Figure 4.3 as a function of the number of samples used in the estimation.

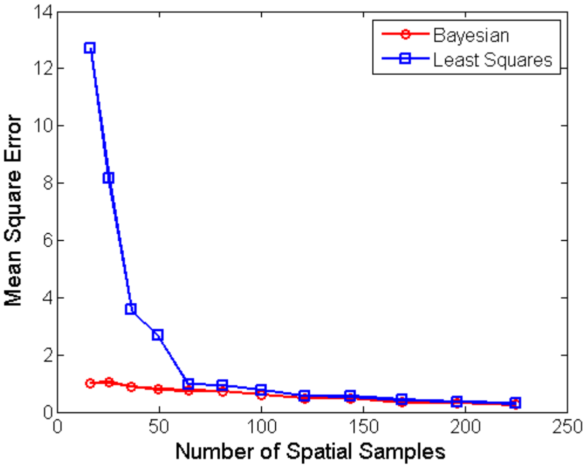


Figure 4.3: MSE for coefficient estimation technique averaged frequency and subject.

Note here, that for a 6th-order model, there are 49 coefficients. As can be seen, the least squares approach begins to degrade significantly as you lower the number of available spatial samples towards the theoretical limit for a unique solution. In contrast, the mean square coefficient error using the proposed Bayesian technique remains quite stable, and shrinks linearly as the number of spatial samples increases. These early results indicate that the Bayesian estimation technique may be capable of accurately estimating the SH coefficients with as few spatial samples as the number of coefficients in the model, or less.

Figure 4.4 shows estimated HRTFs for threes subjects (one subject per row) taken along the median plane when a decreasing number of spatial measurements were used, as indicated by the column headings. Here, it can be seen that the HRTF begins to lose individuality and become more and more like the subject-average HRTF (zero measurements) as the number of spatial samples is reduced. Also apparent is the “noisy” characteristic of the estimated HRTFs when only a few measurements are used due to the frequency-by-frequency form of the estimation. While this degradation appears dramatic in the visualization of Fig. 4.4, it is likely that much of this spectral variation is undetectable due to the frequency resolution limitations of the peripheral auditory system.

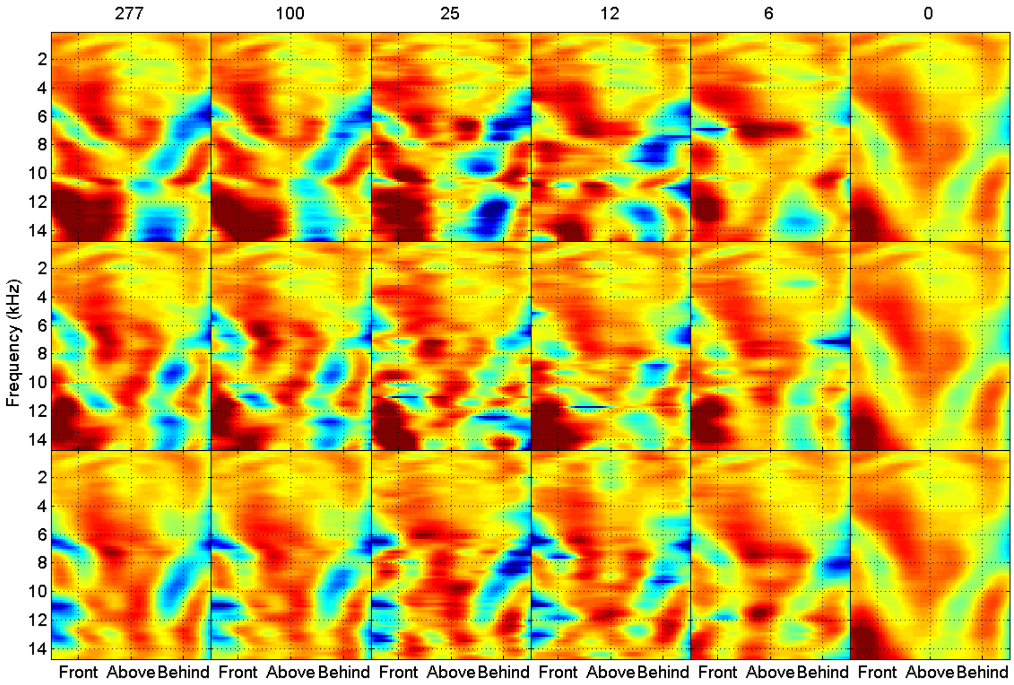


Figure 4.4: A 4^{th}-order HRTF magnitude (in dB) for three subjects (rows) plotted as a function of the angle around the median plane. Each column represents the estimated HRTF using the number of spatial measurements indicated above the column. Red indicates high magnitude, while blue indicates low magnitude.

4.7 Perceptual Evaluation

As with any HRTF estimation technique, a true evaluation of its effect on localization can only be accomplished perceptually. In order to validate the Baysian HRTF estimation technique and determine the minimal number of spatial measurements that are required to maintain localization performance, the following perceptual localization experiment was conducted.

4.7.1 Experimental Task

The perceptual localization task was also conducted at the ALF, the location where the database of Sec. 4.5 was collected. The six subjects (4 males, 2 females) participated in 40 60-trial blocks over the course of 2 to 3 weeks. At the beginning of each trial the subject was asked to align his or her head with a speaker located at the front of the facility. The subjects received feedback about the position of their head via an LED indicator slaved to a 6 DOF head tracker (Intersense IS900). The indicator consisted of a clusters of LEDs centered at each speaker location which were illuminated individually to indicate the direction that the subject was facing. The subject could initiate a trial by pressing a button on a hand-held wand. Each trial consisted of a single stimulus and the subject was asked to respond to the perceived direction of the stimulus by orienting the wand to point in the indicated direction. Real-time visual feedback about their pointing direction was also provided via the LED clusters in a similar manner to the head-position feedback, however, in this case the indicator illuminated a cluster based on the orientation of the wand. When the judged location was illuminated, the subject would have to press a trigger to record the response at which time the LED cluster of the correct location was illuminated. The subject had to acknowledge the correct response with an additional button press before re-centering and beginning the next trial.

4.7.2 Stimulus Generation

The test stimulus for all trials consisted of a random segment of white noise which had been bandpass filtered between 300 Hz and 15 kHz. Segments with a 250 millisecond duration were used on 11 out of 12 trials while 10 second segments were used for the remaining. The stimuli were created using Matlab and processed via SLAB, a PC-based open source spatial-audio system capable of real-time rendering [71] and presented through a pair of Beyerdynamic

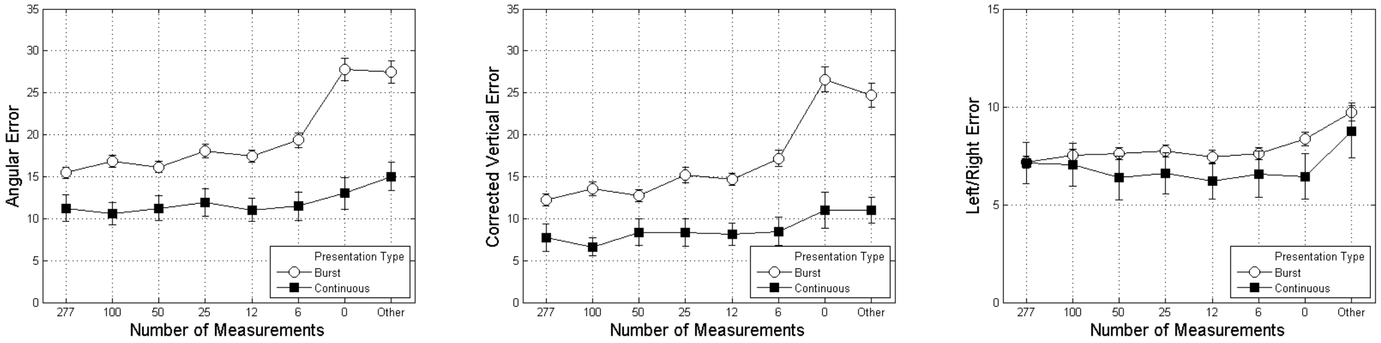
DT990 headphones. SLAB processed the stimuli with one of many different types of HRTFs depending on the experimental condition. The HRTFs used in the experiment consisted of a "Full" HRTF estimated from all 274 measurement locations, 5 non-individual HRTFs from the other 5 subjects, and HRTFs which had been estimated using a reduced set of 100, 50, 25, 12, 6, or 0 measurement locations. The zero location condition consisted of the average HRTF from the database and was therefore identical for all subjects.

4.7.3 Estimating the HRTF from small sample sets

HRTFs for 6 additional test subjects were recorded in order to evaluate the model perceptually. The HRTFs for these six subjects (4 Males, 2 Females) were not included in the training database but were collected in an identical manner. For an N-location test HRTF, N locations were chosen randomly from the available measurement locations. An algorithm modeled after electron repulsion was then used to iteratively repel the selected locations away from each other. At each iteration if the repulsion resulted in a location which was closer to a new measurement location the selected location was replaced by the nearer. This process was repeated until the repulsion did not result in any changes to the selected locations. The repulsion algorithm was used to ensure that the sample locations were roughly distributed across the sphere. Using the N measurements, an estimated HRTF was calculated using the Bayesian estimation technique of Section 4.3. In this case, the hyperparameters of the model, $\mathbf{m_c}$ and $\mathbf{R_{cc}}$ had been trained using all 54 subjects of the HRTF database described in Section 4.5 via the algorithm of Section 4.4. For each N-location HRTF, 5 different randomly chosen measurement sets were tested.

4.7.4 Results

Figure 4.5a shows the average absolute localization error for both stimulus durations and all of the tested HRTF estimation types. Results for the burst condtion show an average angular error of approximately 15 degrees for the 277-measurement HRTF. This error steadily



(a) (b) (c)

Figure 4.5: Absolute localization errors (in degrees) averaged over subject and spatial location in terms of (a) Total angular error, (b) Error in the vertical-polar dimension, and (c) Error in the lateral dimension.

increases as the number of measurement locations is lowered until it reaches approximately 20 degrees for the six-location HRTF. A dramatic increase in total angular error is seen for both the average HRTF (the zero-measurement HRTF) as well as when a non-individual (other) HRTF is used. Not surprisingly, the overall effect of the number of measurements on total angular error in the burst condition is significant ($p < 0.001$), however upon a *post hoc* Tukey honestly significant difference (HSD) test, measurement numbers as low as 12 resulted in performance which was not significantly different than the baseline 277-measurement case at a 95% confidence level.

A slightly different trend is seen for the continuous stimuli. Here, the baseline performance of approximately 11 degrees is maintained as the number of measurement locations is decreased until the two non-individualized conditions. The number of HRTF measurements also results in a significant difference for continuous stimuli ($p < 0.001$)), however upon a *post hoc* Tukey HSD test only the two non-individualized conditions were significantly different at a 95% confidence interval.

Figures 4.5b and 4.5c represent the same localization errors but in this case broken down into the components which lie along or across cones of confusion, respectively. The predominant error dimension for all of the HRTFs tested was the vertical error. Here the term "corrected vertical error" is used to signify that these errors are weighted by the cosine of the lateral angle to reflect the fact that these vertical errors contribute less and less to the total angular error as the lateral angle increases in magnitude. As with total angular error, the number of measurements can be reduced to near 12 locations before performance starts to noticeably suffer. The fact that 6 measurement locations seem to be sufficient for lateral localization indicates that it is the vertical cues that begin to suffer first as the number of measurements are reduced.

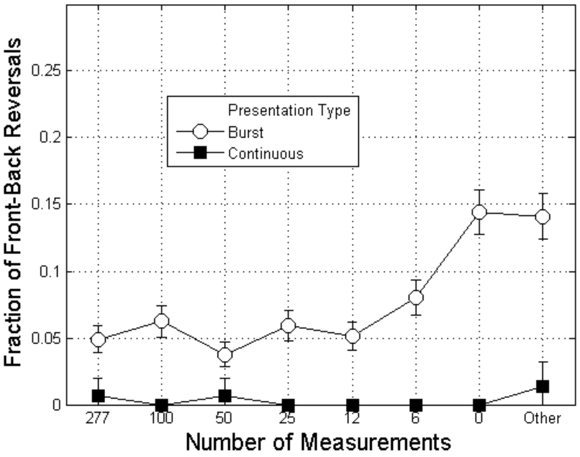


Figure 4.6: Fraction of trials which resulted in a Front-Back Reversal averaged across subject and spatial location for each estimated HRTF type.

Another useful metric of analyzing localization responses is the proportion of trials in which the subject perceived the sound source to be in the opposite front-back hemisphere from the actual source location. These front-back reversals are particularly prevalent when using virtual audio with non-individual HRTFs [19]. Figure 4.6 shows the proportion of trials with a front-back reversal for each of the experimental conditions. As expected front-back reversals were extremely rare ($< 2\%$ of trials) for long-duration stimuli. For the short duration stimuli, HRTFs with more than 12 measurement locations showed little to no increase in the rate of reversals compared to the baseline condition of 277 measurements. As the number of measurements was decreased to 6, the number of front-back reversals increased slightly. When non-individual HRTFs were used the percentage of reversals increased an additional 10% above baseline to around 15% of trials.

4.8 Discussion

The computational results clearly show that the use of non-individual HRTFs as a source of *a priori* information can help make accurate predictions of the spherical harmonic coefficients. It is the belief of the authors that the particular choice of HRTF representation played a significant role in the success of the technique at maintaining both computational and perceptual performance at low numbers of measurements. In particular, the efficiency of this particular method for spherical harmonic expansion severely limits the number of coefficients needed to represent an HRTF and therefore limits the amount of new information with needs to be collected. Perhaps the most interesting conclusion is that these results provide further concrete evidence that a significant amount of information contained in an HRTF is constant across individuals; a conclusion hinted at earlier by Brungart *et al.* [20]. While it is unclear exactly which HRTF structures are common, Brungart's previous results combined with the perceptual results discussed here seem to indicate that accurately capturing the variation within the cones of confusion may be the key distinction between individual and non-individual performance.

While the computational results may be insufficient in general for evaluating the perceptual effect of the estimation technique, the perceptual results also indicate a strong adherence to the computational evaluation. The computational results indicated that only a relatively small amount of error was introduced to the estimation process even when reducing the number of measurements significantly, but it was unclear what perceptual impact this small amount of error introduced. The perceptual results clearly show that fairly good localization performance can still be achieved with as few as 12 measurement locations. HRTFs estimated from only 12 measurements performed significantly better than either non-individualized HRTF. This implies that a large savings in terms of measurement locations can be accomplished without greatly affecting localization performance. Reducing the number of measurement locations from 277 to 12 could potentially cut both measurement time and equipment costs by over 95%.

As it stands, this method, while successful, still represents initial progress towards streamlining the overall HRTF collection process. Future work might include providing additional tuning of the model itself. This could come in the form of a larger cross-facility HRTF database, or potentially a more compact, focused database which only contains training HRTFs of subjects that are similar to the test subject. This cluster identification could be accomplished via easily collectible anthropometric data as in [58], or from the limited set of HRTFs themselves as in [52]. Additional work could also analyze the impact of particular measurement locations. Weighting ipsilateral locations so that they contribute more to the estimation process may provide a way of counteracting the decrease in SNR at contra-lateral locations. Additionally, finding a set of “optimal” measurement locations could provide better performance than the random selection used here.

4.9 Conclusion

This paper introduced a novel HRTF estimation technique based on the inclusion of *a priori* information from the HRTFs of other individuals. HRTFs were first represented by a set of SH coefficients and a distribution of individualized HRTFs was found by modeling a database of HRTFs with a multivariate normal distribution. This distribution was then exploited via traditional Bayesian estimation techniques to enable accurate estimation of new individualized HRTFs from a small number of spatially distributed measurements. It was shown that this technique provided a decrease in reconstruction error compared to traditional naive approaches and maintained adequate perceptual performance when as few as 12 measurement locations were utilized.

Chapter 5

Paper 3: The Sectoral Model of Individualized Head Related Transfer Functions

5.1 Introduction

Over the last few decades, a number of new and interesting technologies have been developed to take advantage of the natural human ability to determine the direction from which a sound arrives. These technologies, broadly referred to as spatial auditory displays (SADs), use a person's innate localization ability to provide useful directional information through the auditory channel. In order to provide three-dimensional directional information, these spatial auditory displays rely on head-related transfer functions (HRTFs); measurements which capture the acoustic filtering caused by the head, shoulders, and outer ears. These HRTFs encode all of the information necessary to perform auditory localization and can be combined with any broadband source to create the precept of that sound originating from a specific location in 3-dimensional space.

Head-related transfer functions are intricate functions of space, frequency, and individual characteristics, which means that acquiring an HRTF for a particular individual typically involves obtaining spatial sample measurements at a large number of locations distributed around a 3-dimensional sphere. This process can be prohibitively expensive both in terms of the physical equipment necessary for HRTF measurement, and the time needed to make the large number of measurements. While several methods have been proposed which seek to simplify the process by minimizing the number of spatial HRTF measurements that need to be taken, relatively little attention has been given to simplifying where the spatial measurements need to be taken. This paper introduces a novel model for individualized HRTFs which separates the spectral cues that are individual in nature from those that are consistent across all individuals. Because the individual information is assumed to be contained entirely in the HRTF variation along the vertical and front-back dimensions, the model allows individualized HRTFs to be estimated from measurements taken on a single sagittal plane, the plane which separates the body into right and left sections. Section 5.2 provides a summary of the relevant work related to HRTF modeling and estimation. Motivation for the new model is provided in Sec. 5.3 along with its formal description. Section 5.3 goes on to describe the process by which the individualized model parameters can be estimated from only locations along a single sagittal plane. Finally, Sec. 5.5 describes the perceptual validation of both the model and the estimation technique.

5.2 Background

5.2.1 Spatial Hearing

The general principles behind sound localization have been known for some time. It is well understood that differences in the signals arriving at the two ears are used to determine a sound source's direction of arrival in the lateral dimension [1]. More specifically, lateral localization is dominated by gross interaural (between the ears) level differences (ILDs) at high frequencies, and interaural time differences (ITDs) at low frequencies [8]. It has also been shown that the frequency-dependent timing information of the signals to the individual ears is perceptually irrelevant to sound localization, and can be replaced with a constant group delay between the two ears referred to as the interaural time difference [25]. Unfortunately, due to the symmetry of the human head, the two interaural cues are roughly constant at locations with the same lateral (left-right) angle. These contours of constant interaural cues are referred to as cones of confusion because the interaural cues cannot be used to distinguish directions within any one cone. This separation of lateral and "within-cone" localization cues leads naturally to the interaural-polar coordinate system shown in Fig. 5.1. Here, the lateral coordinate θ originates from in front of the listener and goes to positive and negative 90^o to the right and left of the listener, respectively. The intraconic coordinate, ϕ, starts at 0^o on the frontal-horizontal plane are goes to $\pm 180^o$ on the rear-horizontal plane, where positive 90^o is directly above.

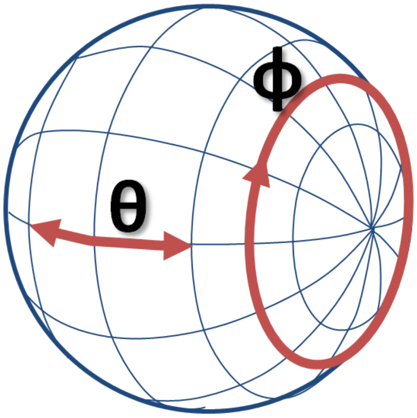


Figure 5.1: The interaural coordinate system. θ represents the lateral or right-left dimension, while ϕ represents the intraconic dimension. (Sometimes referred to as polar, vertical, or sagittal dimension)

While interaural cues alone are not sufficient to localize accurately in the intraconic dimension, humans remain relatively accurate at sound localization within this dimension through the use of monaural (one-ear) spectral cues [4, 66]. While it is unclear the exact mechanism which is used perceptually to interpret these spectral cues, it does appear that humans are only sensitive to braod spectral features [12], which could include both spectral resonances (peaks) [15], and spectral nulls (notches) [17].

Regardless of the exact nature of the localization cues, several authors have shown that all of the perceptually-relevant localization cues can be captured in a measurement known as the head-related transfer function, which can be used to impart directional characteristics to a monaurally sound source directional characteristics [4, 66, 67]. Head-related transfer functions are traditionally measured by placing a set of small binaural microphones in the ear canals of a person and recording a known test stimulus which is played from a large number of spatial locations via a loudspeaker. The resulting recordings can then be compared to the original signal in order to calculate the direction-specific head-related transfer function [5].

5.2.2 Idiosyncrasy in HRTFs

It has been well documented that when a person listens to sounds rendered with someone else's HRTF, localization accuracy begins to suffer. This is believed to be due to individual differences in HRTFs which arise due to the physical differences in size, shape, and orientation of a person's head, shoulders, and outer ears. Non-individual HRTFs seem to perform particularly bad in the cones of confusion. Wenzel *et al.* [19] showed that listeners tended to perceive non-individualized sounds as originating from near the horizontal plane independent of their intended elevation. They also reported a tendency of subjects to localize sounds to the opposite front-back hemisphere from where they were presented. This increase in so-called front-back reversals and poor elevation performance was also described by Brungart and Romigh [20]. Figure 5.2 shows results from their study comparing the performance when using individualized HRTFs, non-individualized HRTFs, and HRTFs measured on an acoustic manikin (KEMAR). Here, localization accuracy is reported in terms of average absolute localization error and the percentage of front-back reversals. Total error refers to the total angular error between the desired direction and the response direction. Lateral and polar error refer to the components of this total error which lie along the interaural axis and along a cone of confusion, respectively. As can be seen, localization performance using either type of non-individualized HRTF is severely degraded. As alluded to earlier, this performance degradation seems to be almost exclusively due to errors within a cone of confusion. These results indicate that most of the inter-subject differences in HRTF structure are likely to be those physical features which act as perceptual cues for within-cone localization judgments.

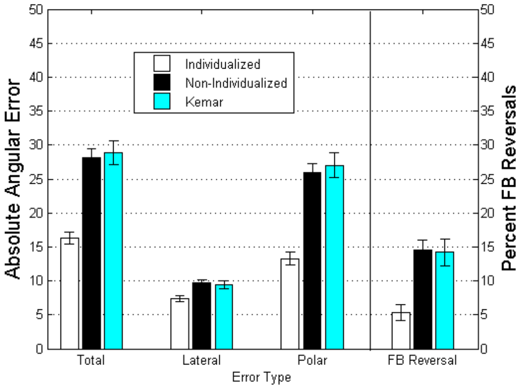


Figure 5.2: Localization accuracy for three different types of HRTFs.

In addition to the non-individual localization results, Brungart and Romigh [20] also showed that localization within the cones of confusion could be improved when using non-individual HRTFs by increasing the magnitude of spatial variance within each cone individually. While performance obtained using non-individual HRTFs never reached that obtained using individualized HRTFs, the study showed a key relationship between the spatial variation of the spectrum around a single cone of confusion, and the perceptual cues that are used to make those spatial judgments. If we combine these results, it seems likely that only the intraconic spatial variation within an HRTF is highly individualized. It is this concept that is exploited in the sectoral HRTF model described in Sec. 5.3.

5.2.3 An Efficient HRTF Representation

Because a large amount of the detail present in a typical HRTF recording has been shown to be perceptually irrelevant to sound localization, it is often helpful to simplify an HRTF into a form with less perceptual noise. Romigh *et al.* [72], developed a perceptually-minimal HRTF representation based on spherical harmonic (SH) decomposition which takes advantage of several auditory phenomena including humans' insensitivity to monaural phase information and fine spectral detail, as well as the natural decomposition of localization cues along the lateral and intraconic dimensions. This representation expands the log magnitude spectra (in dB) at a single frequency, and the ITD, onto a set of real spherical harmonic basis functions which have been orientated to reflect the interaural-polar coordinate system.

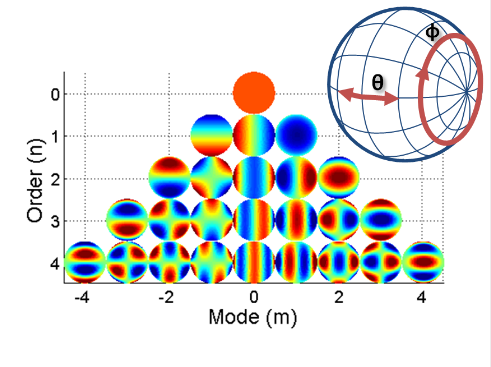


Figure 5.3: Real spherical harmonic basis functions for orders 0 though 4 in the Interaural-polar coordinate system.

The first 25 spherical harmonic basis functions of orders zero through four, are shown in Fig. 5.3. These spatial basis functions are characterized in terms of the spherical harmonic order and mode (degree), as well as the two interaural coordinates ϕ and θ, as described in Eq.5.1.

$$Y_{nm}(\phi,\theta) = \begin{cases} \frac{(2n+1)}{4\pi} P_n^m(\cos(\frac{\pi}{2}-\theta)) & \text{if } m = 0 \\ N_n^m P_n^m(\cos(\frac{\pi}{2}-\theta))\cos(m\phi) & \text{if } m > 0 \\ N_n^m P_n^{|m|}(\cos(\frac{\pi}{2}-\theta))\sin(m\phi) & \text{if } m < 0 \end{cases} \tag{5.1}$$

Here, P_n^m corresponds to the associated Legendre Polynomial of order n and degree m, an orthonormal basis defined recursively and described further in [68] and [30]. N_n^m is a normalization constant included to ensure the orthonormality of the basis functions. The equation describing N_n^m is given in Eq. 5.2.

$$N_n^m = \frac{(2n+1)}{2\pi}\frac{(n-|m|)!}{(n+|m|)!} \tag{5.2}$$

Romigh *et al.* [72] used a weighted combination of these basis functions to describe a spatially continuous HRTF (in dB) at a single frequency as illustrated in Eq. 5.3 where $h(\phi, \theta)$ represents the HRTF and C_{nm} are the SH coefficients or weights. In general the expansion is truncated to a P^{th}-order representation, and the authors showed that a 4^{th}-order representation was sufficient to preserve localization accuracy.

$$h(\phi, \theta) = \sum_{n=0}^{P} \sum_{m=-n}^{n} Y_{nm}(\phi, \theta) C_{nm} \tag{5.3}$$

While this representation itself provides savings in terms of the number of parameters which are need to describe an HRTF [70], and correspondingly the number of spatial samples that are required to estimate an individualized HRTF [72], it is not inherently separated into individual and non-individual components. The following sections describe why this type of decomposition may be possible and how such a decomposition leads to a greatly simplified individualized HRTF measurement technique.

5.3 The Sectoral HRTF Model

The goal of the new HRTF model is to provide a more useful representation for HRTFs which allows the complete separation of individualized information from the information that is general in nature. There are a few important points in the discussion above on individual HRTF cue structure that give us reason to believe such a separation is possible. First, lateral localization cues seem to be fairly consistent across individuals while the cues available for intraconic localization vary greatly. Research also suggests that the lateral and intraconic localization judgments are made fairly independently, implying that the underlying cue structures may also be independent. Additionally, evidence seems to suggest that the physical manifestation of the intraconic perceptual cues seem to be related to the spatial variation of the HRTF in the intraconic dimension while the lateral spectral cues seem to be tied solely to the gross lateral level differences caused by head shadow.

Applying these concepts to the SH-based representation described above, the model coefficients which correspond to lateral variation should be subject independent while the coefficients which capture intraconic spatial variation will be largely individualized. Evidence to support this claim in shown in Fig. 5.4. Here, the average coefficient variance over the frequency range from 4 kHz to 14 kHz is shown as a function of spherical harmonic order and mode. In general, the coefficients with a high amount of inter-subject variance correspond to spherical harmonics where $n = |m|$, sometimes referred to as the sectoral harmonics. If we compare these coefficients to the the corresponding SH basis functions shown in Fig. 5.3, it can be seen that these sectoral coefficients represent spatial variation almost exclusively in the intraconic dimension. One obvious exception is the zeroth order coefficient which represents the spatial-average HRTF, a feature which is known to vary greatly across individuals but believed to carry no directional information. For the purposes of this paper the zeroth-order basis function is included as one of the sectoral harmonics.

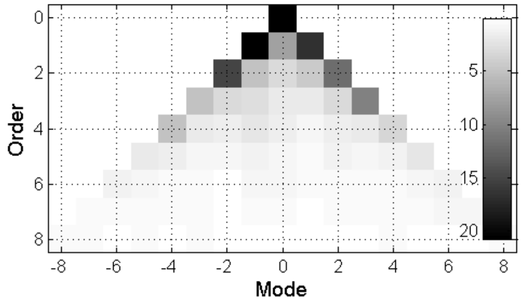


Figure 5.4: Inter-subject variance for the spherical harmonic coefficients of orders 0 through 8 averaged over the frequency region from 4 kHz to 14 kHz.

The above observations imply that, potentially, only the sectoral coefficients need to be individualized. More formally, if we compute the SH-based HRTF representation discussed above and let $\bar{C}_{ij}$ represent the average over all subjects of the coefficient with order i and mode j, we obtain the new SH representation for an individualized HRTF given in Eq. 5.4. Here, H_{Latl} represents the subject-independent portion of the HRTF which is calculated using the average coefficient values of the non-sectoral coefficients, while H_{Sec} consists of the individualized portion of the HRTF calculated exclusively from the sectoral coefficients and the zeroth order coefficient for a particular person.

$$H \approx H_{Lat} + H_{Sec} \tag{5.4}$$

where

$$H_{Lat} = \sum_{n=1}^{P} \sum_{m=-(n-1)}^{n-1} Y_{nm}\bar{C}_{nm} \tag{5.5}$$

$$H_{Sec} = \sum_{n=0}^{P} (Y_{nn}C_{nn} + Y_{n,-n}C_{n,-n}) \tag{5.6}$$

Evidence for the validity of this model can be seen in Fig. 5.5. Here, the HRTFs for three subjects are plotted as a function of the angle around the median plane using their full HRTF, the sectoral model of their HRTF, or the full subject-average HRTF. Here, all HRTFs correspond to a 14th-order SH representation and color is used to indicate the level in decibels; red corresponds to regions of high magnitude, while blue represents regions of low magnitude. As can be seen, the sectoral model is capable of capturing most if not all of the spectral detail within the median plane even in a high-order SH representation.

While the above formalization of the model relies on the SH-based representation presented previously, in theory any decomposition method which separates the lateral and intraconic variation could be used similarly. Possible examples include the lateral and polar decomposition of Brungart and Romigh [20], or the orthogonal representation of Talagala and Abhayapala [33].

5.4 HRTF Estimation from the Median Plane

While the sectoral model provides a way of reducing the number of parameters that describe an individualized HRTF, it is not immediately obvious that such a model reduction leads directly to a simplified HRTF measurement. This simplification stems from the fact that

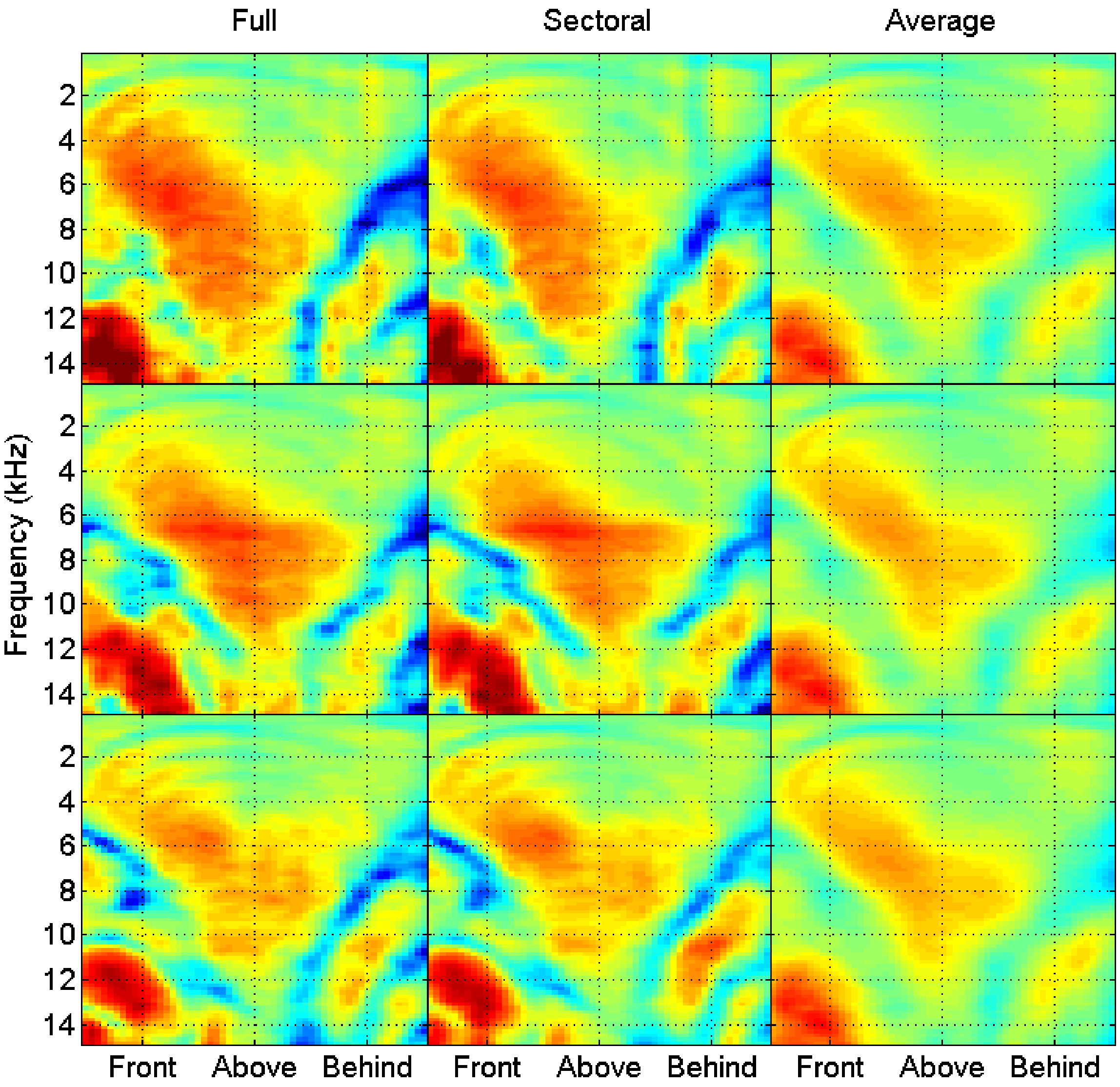


Figure 5.5: The HRTF taken along the median plane for three different subject (rows) using three different representations: the full HRTF, the sectoral HRTF, and the subject-average HRTF

under the assumptions of the model, only the sectoral coefficients need to estimated for each person. Since the corresponding sectoral basis functions are roughly constant in the lateral dimension for low orders (see Fig. 5.3), reasonable estimation of those coefficients should be attainable from measurements taken around a single cone of confusion. In practice, due to error inherent in both the model and any HRTF measurement technique, sampling a cone of confusion near the median plane should result in the most robust estimates. This is due to the nature of the sectoral harmonics at high orders and large values of the lateral angle θ, where the function approaches zero in magnitude.

In order to take advantage of both the weak lateral dependence of the sectoral harmonics and the significant amount of non-individual HRTF information available from various HRTF databases, the technique previously developed in [72] was adapted to estimate the coefficients belonging to the sectoral HRTF model from measurement locations on the median plane. The estimation technique is based on the linear system presented in Eq. 5.7. Here the HRTF (or ITD) **h** at a single frequency can be reconstructed from a linear combination of the SH basis functions given in **Y** via the individualized set of SH coefficients **c**.

$$\mathbf{h} = \mathbf{Y}\mathbf{c} \tag{5.7}$$

$$\begin{aligned} \text{where } \mathbf{h} &= [h(\phi_1, \theta_1), h(\phi_2, \theta_2), \cdots, h(\phi_S, \theta_S)]^T \\ \mathbf{c} &= [C_{00}, C_{1-1}, C_{10}, C_{11}, \cdots, C_{PP}]^T \\ \mathbf{Y} &= [\mathbf{y_{00}}, \mathbf{y_{1-1}}, \mathbf{y_{10}}, \mathbf{y_{11}}, \cdots, \mathbf{y_{PP}}] \\ &\text{and} \\ \mathbf{y_{nm}} &= [Y_{nm}(\phi_1, \theta_1), \cdots, Y_{nm}(\phi_S, \theta_S)]^T \end{aligned}$$

In the above description (ϕ_i, θ_i) represent the coordinates of the i^{th} spatial sample from a total of S available spatial measurements, and P represents truncation order. If we split this representation according to the sectoral HRTF model introduced above, we obtain a term that is dependent only on the sectoral coefficients and a term that is dependent only on the non-sectoral coefficients, as shown in Eq. 5.8. Here, each SH basis vector and their corresponding coefficients are sorted based on whether that particular basis vector corresponds to one of the sectoral harmonics.

$$\mathbf{h} = \mathbf{Y_{Lat}c_{Lat}} + \mathbf{Y_{Sec}c_{Sec}} \tag{5.8}$$

Since the model assumes that only the sectoral coefficients are individualized, we can define a new sectoral-HRTF vector, $\mathbf{h_{Sec}}$ which is the full HRTF with non-sectoral components of the HRTF subtracted off as in Eq. 5.10. This is achieved by substituting in the average values for all of the non-individualized, non-sectoral coefficients $\mathbf{\bar{c}_{Lat}}$.

$$\begin{aligned} \mathbf{h_{Sec}} &\approx \mathbf{h} - \mathbf{Y_{Lat}\bar{c}_{Lat}} \qquad (5.9) \\ &\approx \mathbf{Y_{Sec}c_{Sec}} \qquad (5.10) \end{aligned}$$

With this modification, we can now use the Baysian estimation technique introduced in [72] by substituting in the sub-sampled basis matrix, $\mathbf{Y_{Sec}}$, and the sub-sampled coefficient vectors, $\mathbf{c_{Sec}}$ in place of the original full versions described in the original paper. This substitution results in the minimum mean square estimator for the underlying sectoral coefficients given in Eq. 5.11. Here, $\mathbf{R_{Sec}}$ and $\mathbf{\bar{c}_{Sec}}$ represent the covariance matrix and the average of the sectoral coefficient vector, respectively, across subjects. For more detailed motivation and a description of this technique see Romigh *et al.* [72].

$$\begin{aligned} \mathbf{\hat{c}_{Sec}} &= E[\mathbf{c}|\mathbf{h_{Sec}}] \\ &= \mathbf{\bar{c}_{Sec}} + \mathbf{R_{Sec}Y_{Sec}}^{\mathbf{T}}(\mathbf{Y_{Sec}R_{Sec}Y_{Sec}^{T}} + \sigma^{\mathbf{2}}\mathbf{I})^{-\mathbf{1}}(\mathbf{h_{Sec}} - \mathbf{Y_{Sec}\bar{c}_{Sec}}) \qquad (5.11) \end{aligned}$$

Since this technique could be applied using HRTFs measured along any cone of confusion, we also investigated the effect on modeling error as the cone of confusion that is being sampled is moved farther away from the median plane. Here, modeling error is defined as the amount of spectral distortion, the RMS error of the resulting HRTF spectrum (in decibels) averaged over both ears, location, and individual. This analysis was carried out by estimating each HRTF from the AFRL HRTF database described in [72] from 25 locations contained on a single cone of confusion at the indicated lateral angle. As can be seen in Fig. 5.6, modeling error begins to increase as the lateral angle of the measurement cone is increased past around 15 degrees. This agrees with our previous assertion that the estimation technique for the model parameters is best suited for measurements taken near the median plane and motivates the choice in the perceptual testing of Sec. 5.5 to use median plane locations for the perceptual validation of the proposed methods.

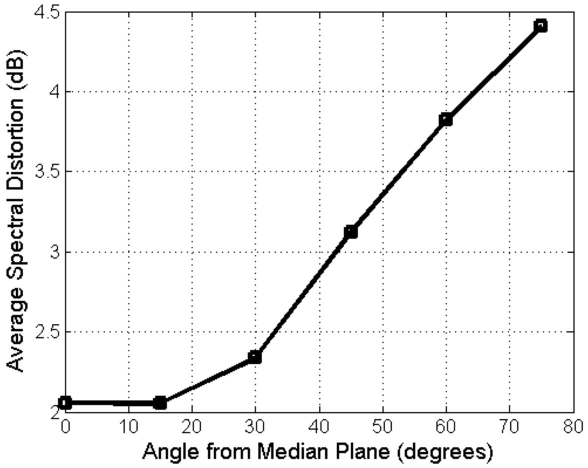


Figure 5.6: Spectral distortion in the resulting HRTF as a function of the lateral angle where the cone of confusion was sampled.

5.5 Perceptual Model Evaluation

To evaluate the performance of any HRTF estimation strategy, the true litmus test involves performing a perceptual localization study where subjects are asked to localize sounds which have been rendered with both fully-measured individualized HRTFs and estimated HRTFs. Effects of the estimation strategy can then be analyzed by comparing localization performance across the HRTF types. The following perceptual evaluation was designed to answer two questions: “What is the impact on localization performance when the sectoral model is used to describe an individualized HRTF?” and “How does localization performance compare to the previously-studied HRTF estimation strategy based on the full 4^{th}-order representation when only a small number of measurements are available?”.

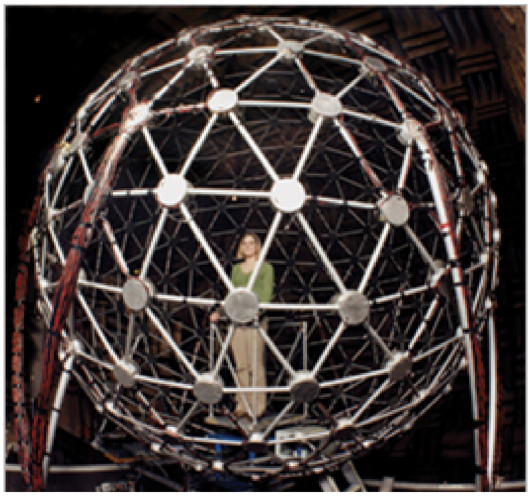


Figure 5.7: Auditory Localization Facility, Wright Patterson Air Force Base, Ohio

5.5.1 Experimental Task

The perceptual evaluation was conducted at the Auditory Localization Facility (ALF) at the Air Force Research Labs in Dayton, Ohio. The task consisted of six subjects (4-males, 2-females) with normal hearing localizing broadband noise stimuli which had been rendered with an estimated HRTF and presented over headphones (Beyerdynamic DT990). Stimuli consisted of 250 millisecond pseudo-random white noise which had been bandpass filtered between 300 Hz and 15 kHz. Stimuli was rendered using SLAB [71], a spatial auditory display software capable of rendering head-tracked virtual stimuli in real-time using customizable HRTFs.

Subjects participated in 85 60-trial blocks over the course of three to four weeks. In each trial, subjects were positioned in the center of the ALF and asked to orient their head toward a center speaker location. Subject could monitor their head position in real time through the use of an LED indicator that moved according to the position of the head. The indicator worked by illuminating an LED cluster on the ALF speaker which most closely corresponded to the subject's orientation. Orientation of both the subject's head and a small hand-held wand were captured using a 6-degree-of-freedom ultrasonic tracking system (Intersense IS900). Once a subject's head was centered he or she could initiate a new trial by pressing a button on the wand. Once the test stimulus was played the subject could indicate the perceived direction of the source by pointing a wand-slaved indicator LED and pressing a button. Feedback of the intended direction was then presented to the subject by illuminating the LED cluster corresponding to the target direction. Subjects were required to acknowledge the correct location with another button press before recentering to start another trial.

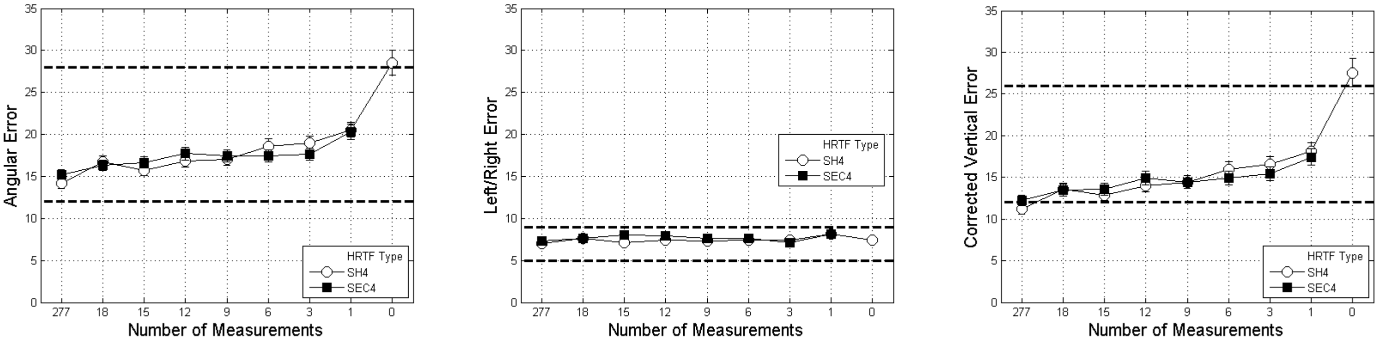
To investigate the effect of the modeling assumptions, a full individualized HRTF collection (277 spherically distributed measurements) were used to estimate both a full 4^{th}-order SH model, and a 4^{th}-order sectoral model introduced in Sec. 5.3. An abundant number of spatially distributed measurements were used to separate the possible effects of using only median plane measurements from the performance of the underlying model assumptions. To test the effects of limiting the measurement locations to the median plane, individualized HRTFs were estimated using two methods. Median-plane sectoral HRTFs were estimated using the adapted Baysian method introduced in Sec. 5.4 from N locations on the median plane where N was one of seven values (1,3,6,9,12,15,18). HRTFs were also estimated using the same number of locations distributed across the surface of the sphere but using the full SH-based Baysian technique developed by Romigh *et al.* [72]. With this combination of HRTFs, direct comparisons can be made between HRTFs derived from median plane locations and HRTFs estimated with the same number of spatially distributed locations. An additional baseline HRTF was also included which consisted of the average HRTF used in the Bayesian technique (which required zero subject-specific measurements).

5.5.2 Results

Results for the perceptual localization task are shown in Fig. 5.8. Figure 5.8a shows the average absolute angular error between the intended location and the subject's directional response. This total angular error is then broken down into its lateral and intraconic (Corrected Vertical) components in Fig. 5.8b and 5.8c, respectively. The bold dotted lines in each figure represent the corresponding errors from a previous study using free-field stimuli (bottom lines) and non-individualized HRTFs (top lines).

Not surprisingly, the total angular error with both models increases as the number of locations is decreased from around 15 degrees with all 277 measurement locations to around 20 degrees with only a single location. Across all conditions the sectoral model seems to perform similarly to that of the full SH model. A slight "flipping" of better performance between the two models can be seen with fewer than 9 measurements, but it does not appear to be statistically significant. Both models resulted in performance similar to free-field performance when all 277 measurement locations were used and significantly better than non-individualized performance even with only a single measurement. The intraconic errors seem to account for most of the performance degradations in the total angular error since the lateral error shows little difference amongst the two models or the number of measurements.

Figure 5.9 shows the percentage of trials in each condition which resulted in a front-back reversal. Here, a front-back reversal is any error which was greater than 30 degrees in front-back-relative angular error and the perceived direction was in the opposite front-back hemisphere as the actual target direction. As can be seen, the number of front-back reversals increases slightly as the number of measurement locations is decreased before almost doubling for the non-individualized condition (zero measurements). Also apparent is the "flipping" of performance between the two models with fewer than 9 measurements. This is likely due to the fact that the sectoral model has fewer coefficients (9 versus 25 for a 4^{th}-order model) which need to be estimated and should therefore remain more stable with fewer measurements.



(a) (b) (c)

Figure 5.8: (a) Average total angular response error in degrees for all of the tested SH representation orders (b) Average lateral response error. (c) Average intraconic response error corrected for the targets lateral position. Upper and lower dotted bold lines represent non-individual and free-field localization, respectively, from a previous study.

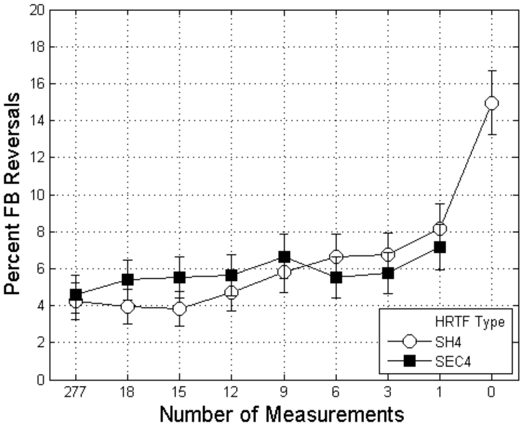


Figure 5.9: Percentage of trials resulting in a front-back reversal for each of the HRTF test conditions.

5.6 Discussion

Results of the localization study when all available measurements were used (N=277) clearly show that the sectoral model does a good job of capturing all of the perceptually-relevant localization cues. While there does appear to be a slight increase in average localization error when the measurements are confined to the median plane, measured performance seems to remain quite good when there are a sufficient number of median plane samples, especially when compared to the non-individual performance. While some initial non-perceptual evaluations suggested that perhaps the sectoral model performance would degrade as the target direction moved further away from the median plane, Fig. 5.10 shows that performance was fairly constant in this regard across the two models. Here, corrected vertical error is plotted as a function of the lateral target angle for the two 18-measurement conditions and the non-individual (zero-measurement) condition.

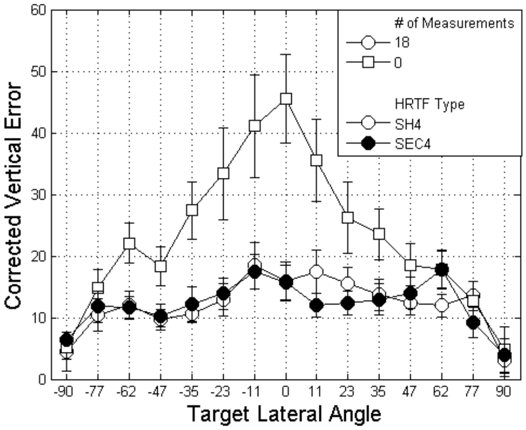


Figure 5.10: Corrected vertical (intraconic) error plotted as a function of the lateral target angle.

Clearly, there is no systematic difference between performance obtained using the two models despite the fact that one model utilized measurements distributed around the whole sphere while the other model utilized measurements obtained only from the median plane. We also note the large increase in intraconic error for the non-individualized HRTF near the median plane. This provides further evidence that capturing individualized median-plane variation might be particularly important.

Even when all 18 of the available median- plane locations were used, the resulting reduction in overall equipment cost compared to the full spherical measurement is significant. While reduction in terms of the number of spatial measurements required seems comparable to the previous method of Romigh *et al.*, the fact that only a single speaker arc needs to be used with the sectoral method provides substantial additional improvement in terms of ease of measurement and equipment costs. Only needing a single plane of measurements opens up the possibility that future measurement systems could literally “hang on a wall” provided that the performance attained here, using median-plane locations, extends to other sagittal planes.

While the sectoral model was utilized in this investigation to simplify acoustic HRTF measurement, it could also be used to help improve the performance of a number of existing HRTF personalization strategies. One prominent candidate for this strategy would be the methods aimed at extracting relationships between anthropometric data and HRTFs such as [59]. The fact that only nine parameters are needed for the sectoral model at each frequency (plus the ITD) could greatly simplify the regression necessary for that type of customization. Also notable is the fact that existing methods aimed at parallelizing HRTF measurement could be combined with this method including both the reciprocity technique proposed in [42] and the spectro-temporal asyncrony technique used in [69].

5.7 Conclusion

The results of this study showed that an individualized HRTF can be modeled quite effectively by assuming that only the spatial variation within the intraconic dimension is truly individualized. When this concept was applied to the spherical-harmonic-based HRTF representation of Romigh *et al.* it resulted in a model where only the SH coefficients corresponding to the sectoral harmonics needed to individualized. This model was shown to result in perceptually-equivalent localization to that of the full SH-based model and similar to previously reported free-field localization performance when an abundance of spatial measurements were available to estimate the sectoral coefficients. This result means that as few as nine parameters per frequency can be used to describe an individualized HRTF, which could greatly simplify the way we model and estimate individual differences in HRTFs.

The study also showed that the simplified form of the sectoral HRTF model could be exploited to simplify HRTF collection to a point where measurements only on the median plane were required without a significant loss in localization performance. This performance should theoretically also hold for sampling distributions confined to other sagittal planes, provided that a significant number of samples is obtained within the plane and that the plane is sufficiently close to the median plane. These results indicate that a large savings could be made in the way we typically measure HRTFs since spherical or movable speaker arrays would no longer be necessary.

Chapter 6

Additional Considerations for Bayesian HRTF Estimation

6.1 HRTFs from Different Databases

In the previous evaluation of the Bayesian HRTF estimation technique, the HRTFs used to train the model priors were ideal in many respects. First of all, all of the HRTFs were recorded in the same facility with the same measurement technique and similar measurement grids. This means that most of the differences between HRTFs were in fact due to the inter-subject dependencies we are attempting to model along with a small amount of random measurement noise. In a more practical scenario, the prior models would be trained separately using HRTFs from a dedicated facility like the one used in our investigation because sufficiently high resolution HRTFs are easily attainable. The actual estimation technique would then be used by secondary facilities interested in avoiding the costs associated with traditional measurements. This means that the HRTFs used to train the model and those being estimated could potentially contain differences related to the acoustics of the facility, positioning of the in-ear microphones, and artifacts of the measurement stimulus, all of which would affect the accuracy of the model. The extent to which these external factors affect the model performance can be analyzed through the use of other publicly-available HRTF databases.

To examine the potential complications of using the Bayesian estimation technique across HRTF measurement facilities, we compared performance when using our in house HRTF collection (AFRL) to two external publicly-available databases, the CIPIC database [73], and the ARI database [87]. The CIPIC database is a collection of 200-tap HRIRs corresponding to 45 individuals measured at 1250 locations. The ARI database consists of a collection of 54 1024-tap HRIRs corresponding to 1550 measurement locations. The corresponding HRTFs of both databases are equalized for the response at the center of the head. Both measurement grids, as well as the grid used in the AFRL HRTFs are shown in Fig. 6.1 from the right side (top row) and above (bottom row) of the subject.

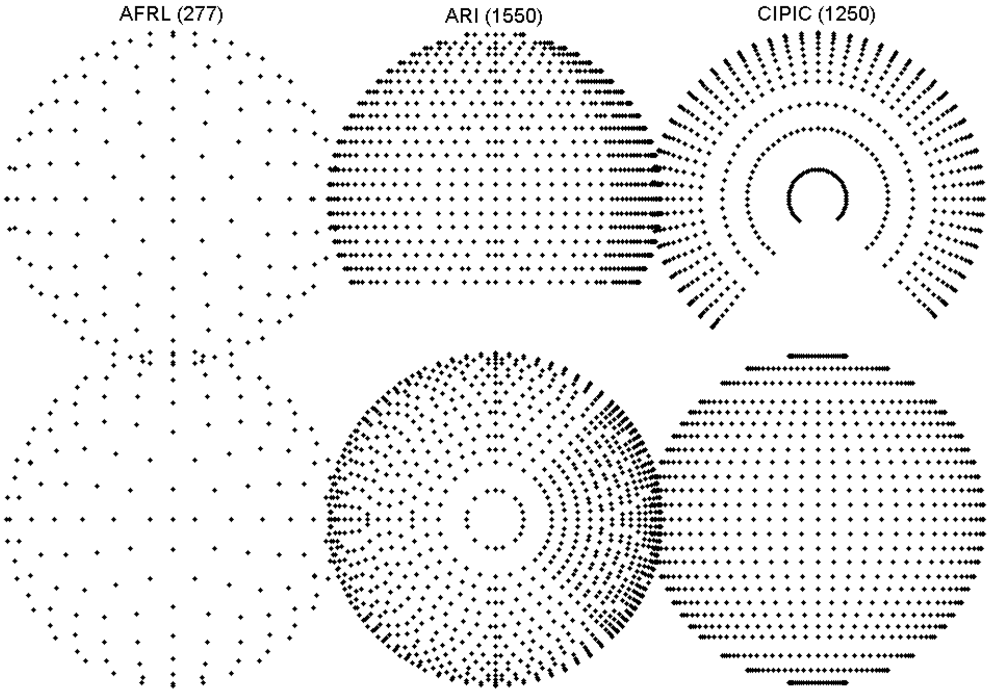


Figure 6.1: Measurement grids for the three publically-available databases used in this analysis.

One important distinction between the external HRTFs and the AFRL HRTFs is the fact that both sets of external HRTFs were equalized for the free-field microphone response, while the HRTFs from the AFRL database are not compensated and therefore also contain the response of the measurement microphones (Our HRTFs are compensated before playback when the headphone correction is made). Additionally, the measurement grids used in the three databases are dramatically different. The AFRL grid is particularly well suited for the SH representation used in the estimation technique because it includes measurement locations completely surrounding the subject, and the measurement location are evenly distributed throughout the sphere. On the other hand, both external databases contain a bottom region without any measurement locations. The ARI database contains locations above -30 degrees in elevation and has sampling with even spacing in the vertical-polar coordinate system, while the CIPIC database contains locations only above 45 degrees in intraconic angle and is sampled evenly in the intraural-polar coordinate system.

6.1.1 Resulting Distributions

One of the first questions to answer is what effect the differences in database HRTFs have on the parameters of the Bayesian HRTF estimation technique; namely, the inter-subject mean and inter-subject variance of the SH coefficients. To answer this question, SH representations were estimated for each database separately using the EM-based algorithm presented above. Figure 6.2 shows the average value of the zeroth order coefficient (along with it's 95% confidence interval) as a function of frequency for the three databases. As can be seen, the largest difference is between the AFRL database and the two external databases. Since the zeroth-order coefficient represents the spatial-average HRTF (or common transfer function, CTF), this coefficient should contain all of the differences between compensation techniques present in the HRTFs.

If we look at the same plot for the three first-order coefficients in Fig. 6.3, it can be seen that only the coefficient in Fig. 6.3a shows a large amount of difference between the different databases, however systematic differences appear for all three coefficients. Because

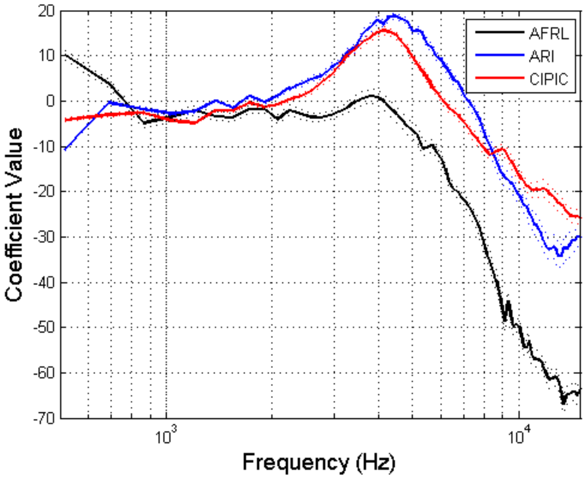
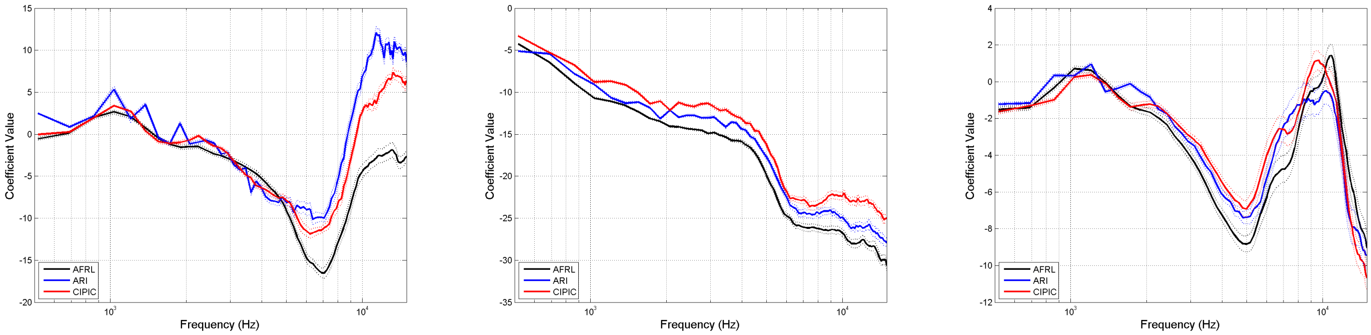


Figure 6.2: Average zeroth-order coefficient.

the differences in HRTFs in principle should only show up in the zeroth-order coefficient, it is believed that these distinctions arise as artifacts of the SH fitting process with the different measurement grids. This is analyzed further in the following section. For all of the coefficients and frequencies analyzed, the resulting inter-subject variances seemed roughly consistent across the databases.



(a) (b) (c)

Figure 6.3: Average first-order coefficient values and their corresponding 95% confidence intervals. (a) $C_{-1,1}$ (Up-Down), (b) $C_{0,1}$ (Left-Right), (c) $C_{1,1}$ (Front-Back)

6.1.2 Effects of measurement grid on coefficient distributions

In order to determine the extent to which the differences we observed across databases were due to artifacts related to the measurement grids or due to actual differences in the HRTFs, we repeated the above analysis with AFRL HRTFs that were interpolated so that they corresponded to the measurement locations of the ARI and CIPIC grids, individually. These resampled HRTFs were then used to re-estimate the corresponding coefficient average and variance of the resulting SH representations across individuals. The resulting estimates for the zeroth order coefficient are presented in Fig. 6.4. Here, there is no discernible difference between the three measurement grids implying that the differences originally seen in this coefficient are likely to be the result of the HRTF calculation differences between the databases rather than the differences in measurement grid.

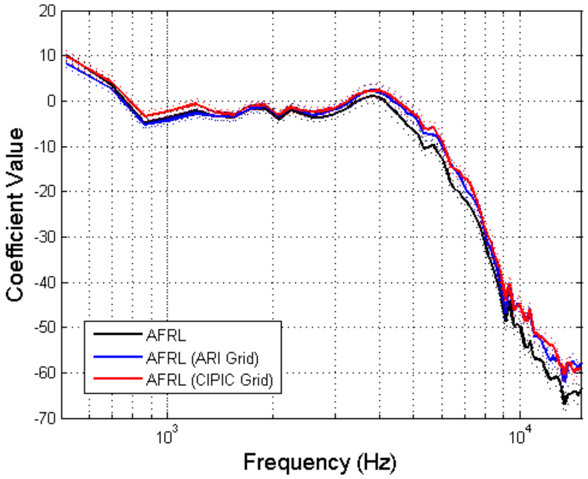


Figure 6.4: Average zeroth-order coefficient values when HRTFs were estimated from AFRL HRTFs which had been interpolated to correspond to the measurement grids seen in the ARI and CIPIC databases.

Figure 6.5 shows the same analysis for the first order coefficient corresponding to the up-down spatial variation. The large differences seen with the actual databases in again seen for the interpolated AFRL databases. While not a complete re-creation of the curves seen in Fig. 6.3a, the elevated coefficient values at high frequencies when one of the external measurement grids is used seems to indicate that the differences in the mean value of the spatial coefficients (coefficients which capture spatial variation, SH orders above 0) seen between the databases are likely the result of estimation error do to the measurement grids rather than differences in the underlying HRTFs themselves.

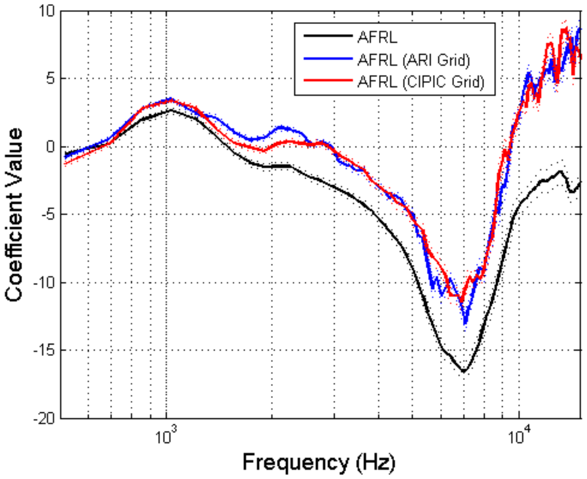


Figure 6.5: Average first-order up-down coefficient values when HRTFs were estimated from AFRL HRTFs which had been interpolated to correspond to the measurement grids seen in the ARI and CIPIC databases.

Fitting spherical harmonic representations to functions with large gaps in the measurement grid is known to cause errors in the resulting representation. It is believed that these errors occur because the corresponding SH basis functions begin to become correlated when entire regions of the sphere are left unsampled. Evidence of this can be seen in Fig. fig4:BasisCorr. Here, the correlation coefficient between each of the 225 basis functions in a 14^{th}-order model is shown for each of the database measurement grids when the spherical harmonics are defined according to the vertical-polar (top row) or interaural-polar (bottom row) coordinate systems. The color of the element of row i and column j in each box represents the absolute value of the correlation coefficient for the i^{th} and j^{th} basis functions; dark corresponds to a high absolute correlation). As can be seen, the AFRL measurement grid results in SH basis functions which are largely uncorrelated. To rule out the potential influence of having fewer measurements, the rightmost column of Fig. 6.6 shows a randomly distributed sampling grid with 1550 locations. As expected, this grid shows zero correlation

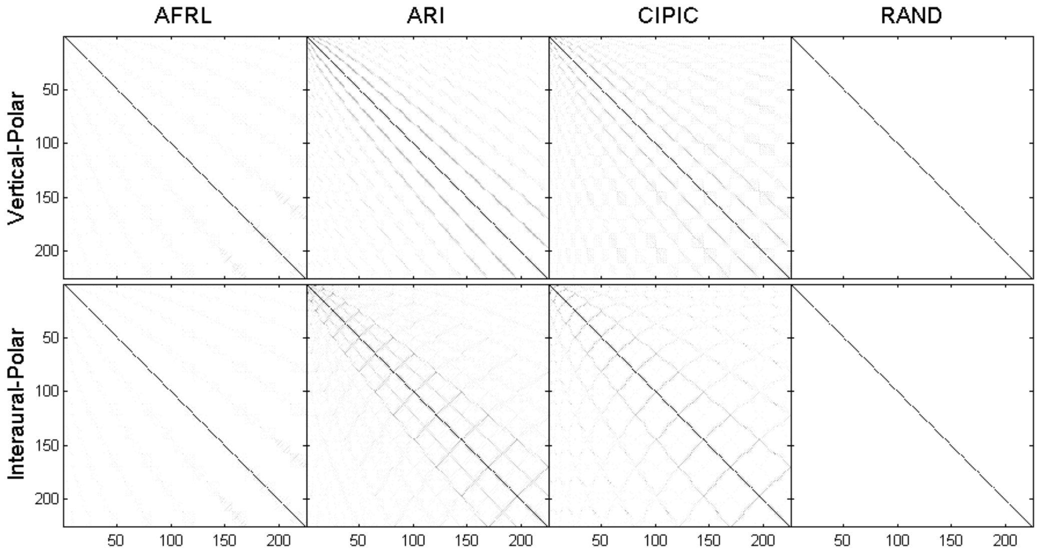


Figure 6.6: Correlation coefficient matrices between each of the 225 basis functions in a 14^{th}-order representation for each of the database measurement grids when the spherical harmonics are defined according to the vertical-polar (top row) or interaural-polar (bottom row) coordinate systems.

between basis functions implying that the results seen with the AFRL grid are due to spatial arrangement of its locations rather than the smaller number of locations. Both the CIPIC and ARI databases show systematic correlations between their corresponding basis functions for both coordinate system.

6.1.3 Effects of Training Database on the Resulting HRTFs

While results from the above sections clearly show that changing HRTF databases can lead to different distributions, it is still unclear how these differences in the underlying distributions affect the resulting estimated HRTFs. In order to examine the effect of training the Bayesian HRTF model using external HRTF databases, we estimated a small set of five AFRL HRTFs that were purposefully withheld from the AFRL database using the models trained using the two external databases. Figure 6.7 shows the resulting average spectral distortion of the resulting 14^{th}-order HRTFs when the number of spatial samples is reduced from a full set of 277 down to only 25. It can be seen that for every number of measurements, the spectral distortion is greater for the external databases than for the the AFRL database. Interestingly, performance remains poorer even when all 277 locations are used in the estimation process. This implies that differences in the underlying training database are cable of causing errors which cannot be overcome with additional individualized measurements.

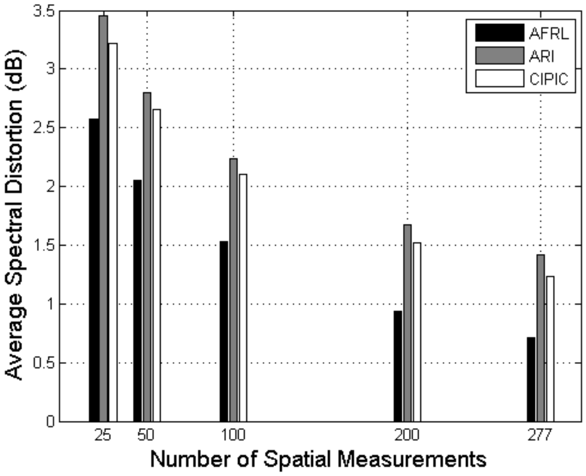


Figure 6.7: Spectral distortion as a function of number of spatial measurements used in the estimation process when the prior distributions are modeled with HRTF from the three HRTF databases.

6.2 Effects of Bottomless Measurement Grids

While the results of the previous section clearly show that training the model with external HRTF databases can lead to errors in the resulting HRTFs, it does not address the problem of a bottomless measurement grid when the model has already been trained with full HRTFs. To examine this issue, the model trained with the original set of AFRL HRTFs was used to estimate several individualized 4^{th}-order HRTFs using 150 spatial samples which had been randomly selected from the upper region of the sphere. The definition of this upper region was varied to include less and less of the sphere by increasing the angle from the bottom of the sphere which got excluded from 0 degrees to 90 degrees in terms of both elevation angle and intraconic angle.

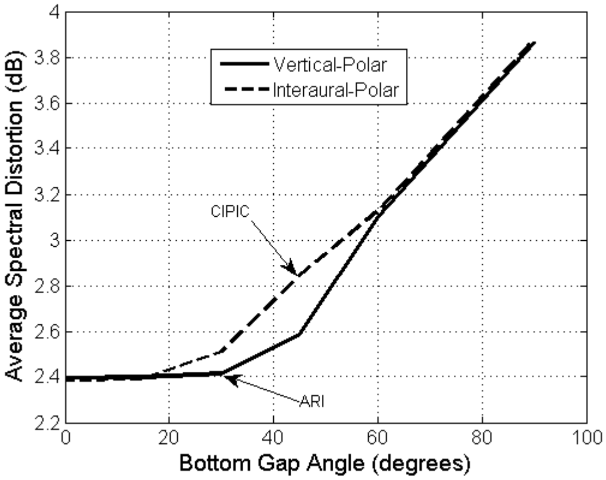


Figure 6.8: Spectral Distortion as a function of bottom gap size.

Figure 6.8 shows the resulting average spectral distortion as a function of the two bottom gap angles. It can be seen that both gaps begin to produce an increased amount of spectral distortion when the angle becomes greater than 20 degrees. For the angles between 20 degrees and 60 degrees, it appears that removing samples based on elevation produced a smaller amount of spectral distortion. Also indicated on the plot is the location on the two curves which correspond to the bottom gaps seen in the ARI and CIPIC measurement grids. These results show that the estimation technique is robust to small gaps in the bottom of the measurement grid, but begins to produce additional spectral distortion when the bottom gap is increased above about 20 degrees.

6.3 Colored Measurement Noise

The linear system MVN model introduced in the Bayesian estimation technique contains an error term that has uncorrelated elements of equal average power σ^2. While the uncorrelated assumption seems valid for most measurement noise, the “white” assumption is less likely to remain valid in practical measurement setups. Indeed, each element of the noise term $\mathbf{n}$ represents the noise in the measured signal at a specific frequency bin and spatial location. It is well known that the head creates an acoustic shadow at high frequencies which can severely lower the signal-to-noise (SNR) ratio for measurement locations on the opposite side of the head. Because of this, certain frequencies and spatial locations may contain more measurement noise than others.

In order to examine whether the equal average noise power assumption in the model affects performance, the AFRL was trained using two additional models for the noise covariance matrix. The original assumption was that the noise was white. Specifically, the covariance matrix was assumed to be a weighted identity matrix, $\mathbf{R_{nn}} = \sigma_{\mathbf{s}}^{\mathbf{2}}\mathbf{I}$, where σ_s^2 is the average noise power for that particular frequency. Assuming that the uncorrelated assumption holds, a model that would allow for different noise levels at different spatial locations is $\mathbf{R_{nn}} = \mathbf{D}$, where D is used to signify that the matrix is diagonal with potentially unequal elements on the diagonal and different values at different frequencies. For reference, an additional naive noise model was also tested where the covariance matrix was always assumed to be the identity matrix $\mathbf{R_{nn}} = \mathbf{I}$ for all frequencies.

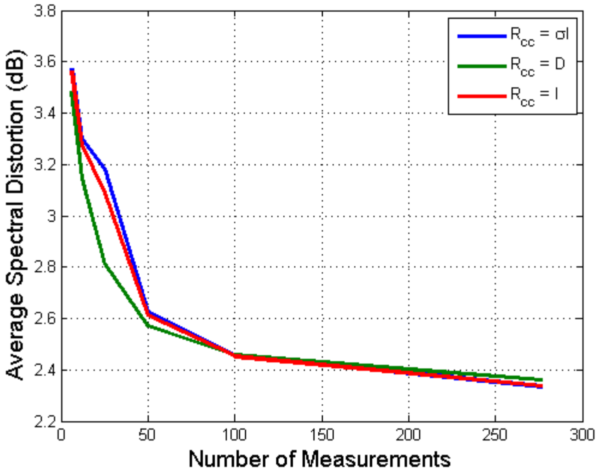


Figure 6.9: Spectral Distortion for three different models of measurement noise as a function of the number of measurement locations

Figure 6.9 shows the resulting spectral distortions when the three noise models were trained using the EM-based algorithm presented previously, along with a varying subset of measurement locations used to estimate 10 individualized HRTFs that were not included in the training set. As can be seen, there is a small advantage to using the diagonal noise model when estimating HRTFs from very few measurements. This advantage seems to disappear as the number of included measurements increases above about 75. Interestingly, there does not appear to be any benefit by using the frequency dependent white noise model, $\mathbf{R_{nn}} = \sigma_{\mathbf{s}}^{\mathbf{2}}\mathbf{I}$, over the completely naive identity matrix. While these results do indicate that estimation performance could possibly be improved by incorporating the spatially-dependent noise model, the overall effect seems small in comparison to the overall benefit provided by the Bayesian HRTF technique.

Chapter 7

Summary of Contributions

This dissertation shows that a significant amount of information about an individualized HRTF can be acquired through the analysis of HRTFs from other individuals. Results of the modeling and estimation clearly show that despite the fact that HRTFs need to be individualized for use in high fidelity spatial auditory displays, only a small portion of the HRTF itself is truly individualized. This development lead to several important contributions to the science of efficient HRTF modeling and estimation, as well as what we understand about the way physical HRTF cues are used perceptually. The major contributions of the dissertation are laid in terms of these areas below, followed by a short section describing potential future research topics that are motivated by the findings of this dissertation.

7.1 Contributions to HRTF Modeling

This dissertation provided a new representation and several key developments that are important for individualized HRTF modeling. The continuous representation presented here built on previous knowledge about the structure of spectral cues in HRTFs to enable the expansion of the entire individualized HRTF onto a set of only 25 perceptually-oriented real spherical harmonic basis functions. This means that the HRTF for a single ear at all locations can be represented simultaneously with as few as $25(K/2+1)$ parameters, where K is the number of discrete frequencies required for each sample HRTF (256 in the perceptual evaluations presented above). It was shown in Chapter 4 that while these parameters are individualized in nature, they can be modeled effectively as belonging to a multivariate normal distribution across subjects. This provides a novel framework for examining individual differences in HRTFs in which a significant amount of the perceptually-irrelevant variation has been removed. By examining the resulting HRTF distribution we were able to provide in Chapter 5 a new model for individual HRTFs in which only a subset of the required parameters need to be truly individualized. This sectoral model states that as few as $9(K/2+1)$ parameters can be used to completely represent the perceptually-relevant individual differences between HRTFs. These contributions are important new findings which may lead to new advances in HRTF analysis and estimation well beyond the estimation strategies introduced in this dissertation.

7.2 Contributions to HRTF Estimation

In addition to the contributions important for HRTF modeling, several advancements were also developed in this dissertation which can greatly simplify the way we collect individualized HRTFs. By taking advantage of the HRTF distribution across subjects, an estimation technique based on Bayesian minimum mean square error estimation was introduced in Chapter 4 that is capable of perceptually-adequate HRTF estimation from as few as 12 spatially-distributed measurement locations. While the included evaluation utilized HRTFs covering the whole sphere which were recording in the same facility, a supplementary section of the dissertation showed that the effects of these ideal conditions only slightly affected modeling error, and could be mitigated by careful choice of measurement grid and measurement procedure. An additional advancement for HRTF collection is presented in Chapter 5 that exploits the spatial structure of the sectoral HRTF model. In this case, it was shown that HRTFs with perceptual accuracy similar to that obtained using full HRTFs could be estimated from measurements taken in a single plane under the assumptions of the sectoral model. This simplification allows for future measurement systems to consist of little more than a fixed semicircular speaker array, something which represents a great deal of simplification over the traditional spherical or movable speaker arrays which are used in traditional HRTF collection.

7.3 HRTF Perceptual Insights Gained

While the main goal of this dissertation was to effectively model and exploit similarities in the HRTFs of different individuals, there were also a few results which provided new insights into how HRTFs might be interpreted perceptually. Results presented in Chapter 3 show that a significant amount of detail contained in a measured HRTF is perceptually irrelevant. While a small amount of spectral smoothing had previously been shown to be irrelevant , the present results demonstrate that a significant amount of spatial smoothing is also well tolerated. This indicates that only gross changes in the spectrum as a function of spatial angle are needed to perceive its directional attributes, and also may provide some evidence that spectral cues are compared across multiple spectral regions when making directional judgments. Additionally, the perceptual results described in Chapter 5 show that even with very little personalization (one physical measurement), localization performance remains significantly better than with non-individual HRTFs. A cursory post-analysis showed that with only a small number of measurements, the zeroth order coefficient was the only parameter that changed significantly. This would imply that proper personalization of the spatial-average HRTF (captured by the zeroth-order coefficient) may be responsible for a significant percentage of the performance difference seen when using individualized versus non-individualized HRTFs. The spatial-average HRTF is typically something which is discarded in HRTF-based localization models, meaning that the previous result may provide new insights for these models.

7.4 Potential Future Avenues of Research

While the present dissertation provides several important contributions to the science and technologies surrounding HRTFs, it also suggests several new research opportunities. The efficiency and validity of the sectoral HRTF model provide the opportunity to improve upon existing HRTF personalization strategies. While this work has focused on personalization through acoustic measurements, several other authors have investigated methods which may provide even greater simplification to individualized HRTF estimation if they are adapted to take advantage of the sectoral model. The greatest potential opportunity is for the personalization of HRTF parameters through anthropometric measurements such as head size and pinna height. These measurements, which can even be estimated from photographs of an individual, have been shown to provide some benefit for HRTF personalization, but existing models of HRTFs had required enormous numbers of free parameters, which made correlating anthropometric measurements to HRTF parameters difficult. Future research may show that the simplified representation provided by the sectoral model could provide the missing link to success using these methods. Another research opportunity comes from the fact that all of our estimation strategies focused on estimating the HRTF independently for each frequency. Due to limitations in spectral resolution of the human auditory system at high frequencies, it is likely that performance could be improved by simultaneously considering the information contained in neighboring frequencies. While it is unlikely that addition of this type of processing would have simplified acoustic measurements beyond what was shown in this thesis, future estimation work which focuses on non-acoustic personalization could significantly benefit from analyzing the covariance of individual frequencies. Finally, the implied perceptual importance of the spatial-average HRTF needs to be examined much more closely. If it proves to be true that the spatial-average HRTF (ATF) dominates the individualized performance benefit, this could significantly change the way we think about the individuality of HRTFs since it would imply that the difference is due largely to non-directional aspects of the HRTF.

Bibliography

[1] J. Blauert, *Spatial Hearing.* The MIT Press, 1997.

[2] T. Ajdler, C. Faller, L. Sbaiz, and M. Vetterli, "Sound field analysis along a circle and its applications to hrtfs interpolation," Audiovisual Communications Laboratory, EPFL, Lausanne, Switzerland, Tech. Rep., 2008.

[3] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. i: Stimulus synthesis," *Journal of Acoustical Society of America*, vol. 85, pp. 858–867, 1989.

[4] ——, "Headphone simulation of free-field listening. ii: Psychophysical validation," *Journal of Acoustical Society of America*, vol. 85, pp. 868–878, 1989.

[5] S. Mehrgardt and V. Mellert, "Transformation of the external human ear," *Journal of Acoustical Society of America*, vol. 61, pp. 1567–1576, 1977.

[6] H. Moller, M. Sorensen, D. Hammershoi, and C. Jensen, "Head-related transfer functions of human subjects," *Journal of Audio Engineering Society*, vol. 43, pp. 300–321, 1995.

[7] J. W. S. aka Lord Rayleigh, "On our perception of sound direction," *Philosophical Magazine*, 1907.

[8] F. L. . Wightman and D. J. Kistler, "The dominant role of low frequency interaural time differences in sound locaization," *Jour*, vol. 91, pp. 1648–1657, 1992.

[9] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *Journal of Acoustical Society of America*, vol. 56, p. 18291834, 1974.

[10] J. Blauert, "Sound localization in the median plane," *Acustica*, vol. 22, pp. 205–213, 1969.

[11] E. H. A. Langendijk and A. W. Bronkhorst, "Contribution of spectral cues to human sound localization," *Journal of Acoustical Society of America*, vol. 112, pp. 1583–1596, 2002.

[12] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747–749, 1998.

[13] F. L. . Wightman and D. J. Kistler, “Monaural sound localization revisited,” *Journal of Acoustical Society of America*, vol. 101, pp. 1050–1063, 1997.

[14] E. A. Macpherson and A. T. Sabin, “Binaural weighting of monaural spectral cues for sound localization,” *Journal of Acoustical Society of America*, vol. 121, pp. 3677–3688, 2007.

[15] J. C. Middlebrooks, “Narrow-band sound localization related to external ear acoustics,” *Journal of Acoustical Society of America*, vol. 92, pp. 2607–2624, 1992.

[16] E. A. G. Shaw, *Localization of Sound: Theory and Applications*, ”, Ed. Amphora, Groton, CT, 1982.

[17] V. Raykar, R. Duraiswami, and B. Yegnanarayana, “Extracting the frequencies of the pinna spectral notches in measured head related impulse responses,” *Journal of Acoustical Society of America*, vol. 118, pp. 364–374, 2005.

[18] R. Greff and B. F. G. Katz, “Perceptual evaluation of HRTF notches versus peaks for vertical localisation,” in *19th International Congress on Acoustics, Madrid, 2-7 September 2007*, 2007.

[19] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization using nonindividualized head-related transfer functions,” *Journal of Acoustical Society of America*, vol. 93, pp. 111–123, 1993.

[20] D. S. Brungart and G. D. Romigh, “Spectral HRTF enhancement for improved vertical-polar auditory localization,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE WASPAA, 2009.

[21] D. S. Brungart, G. D. Romigh, and B. D. Simpson, “Rapid collection of HRTFs and comparison to free-field listening,” in *International Workshop on the Principles and Applications of Spatial Hearing*, 2009.

[22] F. Asano, Y. Suzuki, and T. Sone, “Role of spectral cues in median plane localization,” *Journal of Acoustical Society of America*, vol. 88, pp. 159–168, 1990.

[23] R. L. Jenison, “A spherical basis function neural network for pole-zero modeling of head-related transfer functions,” in *Proc. IEEE ASSP Workshop Applications of Signal Processing to Audio and Acoustics*, 1995, pp. 92–95.

[24] J. Huopaniemi, N. Zacharov, and M. Karjalainen, “Objective and subjective evaluation of head related transfer function filter design,” *Journal of Audio Engineering Society*, vol. 47, pp. 218–239, 1999.

[25] A. Kulkarni, S. K. Isabelle, and H. S. Colburn, “Sensitivity of human subjects to head-related transfer-function phase spectra,” *Jou*, vol. 105, pp. 2821–2840, 1992.

[26] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, A. V. Oppenheim, Ed. Prentice Hall Publishing, 1999.

[27] M. A. Senova, K. I. McAnally, and R. L. Martin, "Localization of virtual sound as a function of head-related impulse response duration," *Journal of the Audio Engineering Society*, vol. 50, pp. 57–66, 2002.

[28] D. J. Kistler and F. L. . Wightman, "A model of head-related transfer functions based on principal componentsa nalysisa nd minimum-phaser econstruction," *Journal of Acoustical Society of America*, vol. 91, pp. 1637–1647, 1992.

[29] W. H. Martens, "Principal components analysis and resynthesis of spectral cues to perceived direction," in *Proceedings of the International Computer Music Conference*, 1987.

[30] M. J. Evans, J. A. S. Angus, and A. I. Tew, "Anlaysing head-related transfer function measurements using surface spherical harmonics," *Journal of Acoustical Society of America*, vol. 104, pp. 2400–2411, 1998.

[31] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized hrtf fitting using spherical harmonics," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics WASPAA '09*, 2009, pp. 257–260.

[32] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, "High order spatial audio capture and its binaural head-tracked playback over headphones with hrtf cues," in *AES 119th Convention, New York, NY, USA*, 2005.

[33] D. S. Talagala and T. D. Abhayapala, "Novel head related transfer function model for sound source localization," *IEEE*, vol. ", p. ", 2010.

[34] B.-S. Xie, "Recovery of individual head-related transfer functions from a small set of measurments," *Journal of Acoustical Society of America*, vol. 132, pp. 282–294, 2012.

[35] N. H. Adams and G. H. Wakefield, "State-space models of head-related transfer functions for virtual auditory scene synthesis," *Journal of Acoustical Society of America*, vol. 2009, p. 3894390, 125.

[36] Q. Huang and K. Liu, "A reduced order model of head-related impulse resoponses based on independent spatial feature extraction," in *IProc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP*, 2009.

[37] G. Grindlay and M. A. O. Vasilescu, "A multilinear (tensor) framework for hrtf analysis and synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2007*, vol. 1, 2007.

[38] W. Zhang, T. D. Abhayapala, R. A. Kennedy, and R. Duraiswami, "Modal expansion of hrtfs: Continuous representation in frequency-range-angle," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing ICASSP 2009*, 2009, pp. 285–288.

[39] W. Zhang, R. A. Kennedy, and T. D. Abhayapala, "Efficient continuous hrtf model using data independent basis functions: Experimentally guided approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 819–829, 2009.

[40] W. Zhang, T. D. Abhayapala, and R. A. Kennedy, “Insights into head-related transfer function: Spatial dimensionality and continuous representation,” *Journal of Acoustical Society of America*, vol. 127, pp. 2347–2357, 2010.

[41] M. Zhang, R. A. Kennedy, and T. D. Abhayapala, “Efficiency evaluation and orthogonal basis determination in functional hrtf modeling,” in *Proc. IEEE Int Acoustics, Speech and Signal Processing (ICASSP) Conf*, 2011, pp. 53–56.

[42] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerovd, “Fast head-related transfer function measurement via reciprocity,” *Journal of Acoustical Society of America*, vol. 120, pp. 2202–2215, 2006.

[43] R. Martin and K. McAnally, “Interpolation of head-related transfer functions,” Air Operations Division Defence Science and Technology Organisation, Tech. Rep., 2007.

[44] J. Sodnik, R. Susnik, and S. Tomazic, “Resolution enhancement of a general hrtf library,” in *Proceedings of ACOUSTICS 2005*, 2005.

[45] M. Kentaro and A. Ando, “Estimation of individualized head-related transfer function based on principal component analysis,” *Journal of Acoustical Society of Japan*, vol. 30, pp. 338–347, 2009.

[46] J. Chen, B. D. V. Veen, and K. E. Hecox, “A spatial feature extraction and regularization model for the head-related transfer function,” *Journal of Acoustical Society of America*, vol. 97, pp. 439–452, 1995.

[47] S. Carlile, C. Jin, and V. van Raad, “Continuous virtual auditory space using hrtf interpolation: acoustic & psychophysical errors.” in *International Symposium on Multimedia Information Processing*, 2000.

[48] J. C. B. Torres and M. R. Petraglia, “HRTF interpolation in the wavelet transform domain,” in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009.

[49] G. Enzner, M. Krawczyk, F.-M. Hoffmann, and M. Weinert, “3d reconstruction of hrtf-fields from 1d continuous measurements,” in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 157–160.

[50] W. Zhang, M. Zhang, R. A. Kennedy, and T. D. Abhayapala, “On high-resolution head-related transfer function measurements: An efficient sampling scheme,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 575–584, 2012.

[51] Q. Huang and Y. Fang, “Interpolation of head-related transfer functions using spherical fourier expansion,” *Journal of Electronics (China)*, vol. 26, pp. 571–576, 2009.

[52] P. Guillon and R. Nicol, “Head-related transfer function reconstruction from sparse measurements considering a priori knowledge from database analysis: a pattern recognition approach,” in *Audio Engineering Society, 125th Convention*, 2008.

[53] V. Lemaire, F. Clerot, S. Busson, R. Nicol, and V. Choqueuse, "Individualized hrtfs from few measurements: a statistical learning approach," in *Proceedings of International Joint Conference on Neural Networks*, 2005.

[54] A. Silzle, "Selection and tuning of hrtfs," in *112th Convention of the Society of Audio Engineering*, 2002.

[55] K. Watanabe, K. Ozawa, Y. Iwaya, Y. Suzuki, and K. Aso, "Estimation of interaural level difference based on anthropometry and its effect on sound localization," *Journal of Acoustical Society of America*, vol. 2832-2841, p. 122, 2007.

[56] P. Satarzadeh, V. R. Algazi, and R. O. Duda, "Physical and filter pinna models based on anthropometry," in *122nd Convention of the Society of Audio Engineering*, 2007.

[57] H. Hu, L. Chen, and Z. yang Wu, "The estimation of personalized hrtfs in individual vas," in *Fourth International Conference on Natural Computation*, 2008.

[58] J. C. Middlebrooks, "Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency," *Journal of Acoustical Society of America*, vol. 106, pp. 1493–1510, 1999.

[59] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "Hrtf personalization using anthropometric measurements," in *Proc. IEEE Workshop . Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 157–160.

[60] S. Xu, Z. Li, and G. Salvendy, "Individualized head-related transfer functions based on population grouping," *Journal of Acoustical Society of America*, vol. 124, pp. 2708–2810, 2008.

[61] K.-S. Lee and S.-P. Lee, "A relevant distance criterion for interpolation of head-related transfer functions," *IEEE*, vol. 1, pp. 0–0, 2010.

[62] R. Duraiswami and V. C. Raykar, "The manifolds of spatial hearing," in *IEEE ICASSP 2005*, 2005.

[63] B. D. Simpson, D. S. Brungart, R. C. Dallman, R. J. Yasky, G. D. Romigh, and J. F. Raquet, "In-flight navigation using head-coupled and aircraft-coupled spatial audio cues," in *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, 2007.

[64] J. M. Loomis, R. L. Klatzky, R. G. Golledge, J. G. . Cicinelli, J. W. Pellegrino, and P. A. Fry, "Nonvisual navigation by blind and sighted: Assessment of path integration ability," *Journal of Experimental Psychology: General*, vol. 122, pp. 73–91, 1993.

[65] E. M. Wenzel, "Localization in virtual acoustic displays," *Presence*, vol. 1, pp. 80–107, 1992.

[66] R. Martin, K. McAnally, and M. Senova, "Free-field equivalent localization of virtual audio," *Journal of Acoustical Society of America*, vol. 49, pp. 14–22, 2001.

[67] E. H. A. Langendijk and A. W. Bronkhorst, “Fidelity of three-dimensional-sound reproduction using a virtual auditory display,” *Journal of Acoustical Society of America*, vol. 107, no. 1, pp. 528–537, January 2000.

[68] M. Poletti, “Unified description of ambisonics using real and complez spherical harmonics,” in *Ambisonics Symposium*, 2009.

[69] P. Majdak, P. Balazs, and B. Laback, “Multiple exponential sweep method for fast measurement of head-related transfer functions,” *Journal of the Audio Engineering Society*, vol. 55, pp. 623–637, 2007.

[70] G. D. Romigh, D. S. Brungart, and R. M. Stern, “A continuous hrtf representation for modeling and estimation,” included as a previous chapter.

[71] J. D. Miller and E. M. Wenzel, “Recent developements in SLAB: A software-based system for interactive spatial sound synthesis,” in *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan*, 2002.

[72] G. D. Romigh and R. M. Stern, “Bayesian estimation of individualized head-related transfer functions,” included as a previous chapter.

[73] V. R. Algazi, R. O. Duda, and C. A. D. M. Thompson, “The CIPIC HRTF database,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001.

[74] V. R. Algazi, R. O. Duda, R. Duraiswami, N. A. Gumerov, and Z. Tang, “Approximating the head-related transfer function using simple geometric models of the head and torso,” *Journal of Acoustical Society of America*, vol. 112, pp. 2053–2064, 2002.

[75] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, “Structural composition and decomposition of hrtfs,” in *Proc. IEEE Workshop the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 103–106.

[76] C. I. Cheng and G. H. Wakefield, “Introduction to head-related transfer functions(hrtfs): Representations of hrtfs in time, frequency, and space,” *Journal of Audio Engineering Society*, vol. 49, pp. 231–249, 2001.

[77] R. O. Duda, “Modeling head related transfer functions,” in *Proc. Conf Signals, Systems and Computers Record of The Twenty-Seventh Asilomar Conf*, 1993, pp. 996–1000.

[78] H. Jo, Y. Park, and Y. sik Park, “Analysis of individual differences in head-related transfer functions by spectral distortion,” in *Proc. ICCAS-SICE*, 2009, pp. 1769–1772.

[79] ——, “Optimization of spherical and spheroidal head model for head related transfer function customization: Magnitude comparison,” in *Proc. Int. Conf. Control, Automation and Systems ICCAS 2008*, 2008, pp. 251–254.

[80] K. I. McAnally and R. L. Martin, “Variability in headphone-to-ear-canal transfer function,” *Journal of the Audio Engineering Society*, vol. 50, pp. 263–266, 2001.

[81] A. W. Mills, "On the minimum audible angle," *Journal of Acoustical Society of America*, vol. 30, pp. 237–246, 1958.

[82] M. Morimoto and H. Aokata, "Localization cues of sound sources in the upper hemisphere," *Journal of Acoustical Society of Japan*, vol. 5(3), pp. 165–173, 1984.

[83] L. L. Scharf, *Statistical Signal Processing*, R. Roberts, Ed. Addison-Wesley Publishing Company, 1991.

[84] F. L. . Wightman and D. J. Kistler, *Factors Affecting the Relative Salience of Sound Localization Cues, Binaural and Spatial Hearing in Real and Virtual Enviroments*, R. Gilkey and T. Anderson, Eds. Lawrence Erlbaum Associates, 1997.

[85] D. N. Zotkin, R. Duraiswami, L. S. Davis, A. Mohan, and V. Raykar, "Virtual audio system customization using visual matching of ear parameters," in *Proc. 16th Int Pattern Recognition Conf*, vol. 3, 2002, pp. 1003–1006.

[86] R. Gilkey and T. Anderson, Eds., *Binaural and Spatial Hearing in Real and Virtual Environments*. Psychology Press, 1997.

[87] Acoustics research institute hrtf database. Website. [Online]. Available: http://www.kfs.oeaw.ac.at/content/view/608/606/