UNIVERSITY OF MIAMI

AN IMPROVEMENT TO ANTHROPOMETRY-BASED HEAD AND TORSO HRTF SYNTHESIS MODELS FOR LOCATIONS NEAR THE FRONTAL MEDIAN PLANE

By
Richard S. Juszkiewicz

A THESIS

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science in Music Engineering Technology

Coral Gables, Florida
May 2007

UNIVERSITY OF MIAMI

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Music Engineering Technology

AN IMPROVEMENT TO ANTHROPOMETRY-BASED HEAD AND TORSO HRTF SYNTHESIS MODELS FOR LOCATIONS NEAR THE FRONTAL MEDIAN PLANE

Richard S. Juszkiewicz

Approved:

Dr. Colby N. Leider Assistant Professor of Music Engineering	Dr. Edward Asmus Associate Dean of Graduate Studies
Kenneth C. Pohlmann Professor of Music Engineering	Dr. James D. Shelley Assistant Vice President for Academic and Research Computing

JUSZKIEWICZ, RICHARD S. (M.S., Music Engineering)
An Improvement to Anthropometry-Based Head and Torso HRTF Synthesis Models for Locations Near the Frontal Median Plane. (May 2007)

Abstract of a thesis at the University of Miami.

Thesis supervised by Professor Colby N. Leider.
No. of pages in text. (127)

Due to the recent proliferation of portable media devices, headphones (and earbuds) are becoming the primary means through which people experience recorded phenomena. In the absence of processing, headphone listeners typically perceive sounds as coming from inside of their heads rather than from the surrounding space. Head-related transfer function (HRTF) based algorithms attempt to rectify this issue; however, their need to be personalized for every individual through expensive, and often impractical, methods prevents these implementations from being effective. Such a need has produced an extensive body of research focused on linking the perceptually significant features of HRTFs to anthropometry. The ultimate goal of such work is to synthesize a complete set of personalized HRTFs strictly from morphological measurements. Recent research has produced an anthropometry-based head and torso (HAT) model that accurately approximates the effects that those body parts have on an incident sound. These HAT-based synthesis models produce very convincing lateral localization effects, and a weak sense of elevation far away from the median plane, but they lack the primary elevation cues that are caused by the external ear (pinna). The work presented herein adds pinna-based elevation cues to an existing HAT model that are most effective near the median plane--an area where the HAT's torso-based elevation cues are particularly poor. The aforementioned cues are created by modeling the known resonances and the primary reflections of the

external ear using digital filters whose parameters are determined from an individual's anthropometry. The eventual result of cascading an existing HAT model with the introduced pinna model is the creation of customized HRTFs. Objective results are provided and indicate that the proposed synthesis method approximates the frequency response of measured HRTFs better than a simple HAT model. Psychoacoustic validation reveals that the model is effective at creating an accurate sense of elevation near the median plane for 67% of the subjects tested. This proves the hypothesis for certain cases and leaves room for future improvements.

DEDICATION

To my mom--the provider of the music gene, to my dad--the provider of the engineering gene and to both of them for their undying and tireless support and encouragement.

ACKNOWLEDGMENTS

This work would not have been possible if it was not for the many kind souls that helped me along the way. My greatest amount of gratitude is extended to the people at the University of California Davis and the University of Maryland (the other UM) for graciously and promptly replying to all of my e-mails with insightful answers to every one of my questions.

I would also like to thank all of the people that participated in my exhausting listening test, especially my fellow Music Engineering students who could have spent that hour of their lives doing something far more interesting.

To everyone that I have crossed paths with in my academic life: thank you for sharing in the learning process (even though I doubt any of you will ever read this).

Lastly, I would like to thank my grammar consultant for all of the erudite and condescending discussions about the English language that we have shared over the years; they have come in handy. Hopefully there are not too many embarrassing errors in the pages that follow.

TABLE OF CONTENTS

TABLE OF FIGURES

INTRODUCTION

Spatial hearing is an integral part of human life. Our ability to localize sounds in dark environments, for example, aids in survival. Understanding how the brain interprets auditory cues in order to determine the spatial origins of sounds is a constantly evolving academic field. Recent advances in the recording of head-related transfer functions (HRTFs) have expedited the learning process for researchers; however, there are still plenty of unanswered questions. The ultimate goal of such research is to understand exactly how humans hear so that computers can be programmed to emulate the phenomena, thus creating a virtual acoustic space that can be utilized in a plethora of applications.

OBJECTIVE

The objective of this work is to further the current state of anthropometric HRTF synthesis research with the ambition that eventually it will be possible for everyone to have their HRTFs computed without having to endure an arduous measurement process. Personalized HRTFs are necessary because current binaural synthesis implementations are based upon generic HRTFs and are ineffective at creating a truly accurate virtual acoustic space. This ineffectiveness is due to interpersonal anatomical differences which cause HRTFs to vary drastically among listeners. To achieve the desired goal, an existing Head and Torso (HAT) model is implemented and cascaded with a novel pinna model that is most effective at producing convincing elevation cues where the HAT model is not. It is also of interest to ensure that the algorithm developed is very computationally efficient so that, if

desired, it can be run in real time. Finally, a listening test will be conducted to quantify and validate the alleged improvements.

STRUCTURE

Chapter One of this paper introduces many of the fundamentals of the anatomy, the physics, the psychoacoustics and the physiology of spatial hearing that are necessary to understand before proceeding with the literature survey and implementation. Chapter Two provides an overview of HRTFs and a brief survey of the prior art in the field that is relevant to the objective. Chapter Three details the design and implementation of the proposed solution and Chapter Four provides some objective results. Chapter Five explains the listening test used to subjectively evaluate the implementation and Chapter Six analyzes its results. Chapter Seven draws conclusions from both sets of results and suggests future work.

1

SPATIAL HEARING

The source of an acoustical stimulus, musical or otherwise, can be thought of as a point in three-dimensional space. Humans are able to perceive a sound as coming from a particular point in this space courtesy of the brain. Physiologically, the localization of a sound begins at a so-called "nucleus" in the midbrain known as the lateral superior olive. It is in this section of the brain where temporal, loudness and spectral cues are interpreted via a cross-correlation operation that is performed on the auditory signals received at the left and right ears [18]. The results of these calculations are then sent to higher levels of the brain for further analyses. This chapter focuses on explaining the aforementioned cues, among others, that are used by the brain during the localization process; however, before explaining any of these cues, it is first necessary to establish the coordinate system that is used throughout this work.

1.1 COORDINATE SYSTEM

The interaural-polar coordinate system and terminology used herein to reference the location of a sound can be seen in Figure 1. Much of the literature in the field of spatial hearing uses the same coordinate system; therefore, it is only logical to remain consistent. The origin of this coordinate system is referred to as the interaural center and is the midpoint of an imaginary line (through the head) that terminates at the entrances to the left and right ear canals. A sound originating from the left of the origin will have a negative azimuth ranging from 0 to -90°. Likewise, a sound from the right possesses a positive azimuth with a range of 0 to 90°. Unlike the azimuth, which only has a complete range of

180°, elevation is a full 360° coordinate. An elevation of 0° is directly in front of the listener, -90° is directly below the listener, 90° is directly above the listener and 180° is directly behind the listener. In this coordinate system, it is elevation that differentiates front from back, not azimuth. The location of a sound is referenced using (θ,Φ), where θ indicates the azimuth angle and Φ denotes the elevation angle. The side of the head in which the sound source is closest is known as the ipsilateral side; the opposite hemisphere is termed the contralateral half.

There are three distinctly named planes in this coordinate system: the median plane, the horizontal plane and the vertical plane. The median plane results from a constant azimuth of 0°. The half of the median plane in front of the listener is aptly referred to in this work as the frontal median plane. The horizontal plane consists of all azimuths at elevations of 0° (in front) and 180° (behind). The frontal plane results from the head being divided vertically about the center. Both elevation and azimuth vary in this plane. Each of these planes, as well as the ipsilateral/contralateral convention, can be seen in the bottom half of Figure 1.

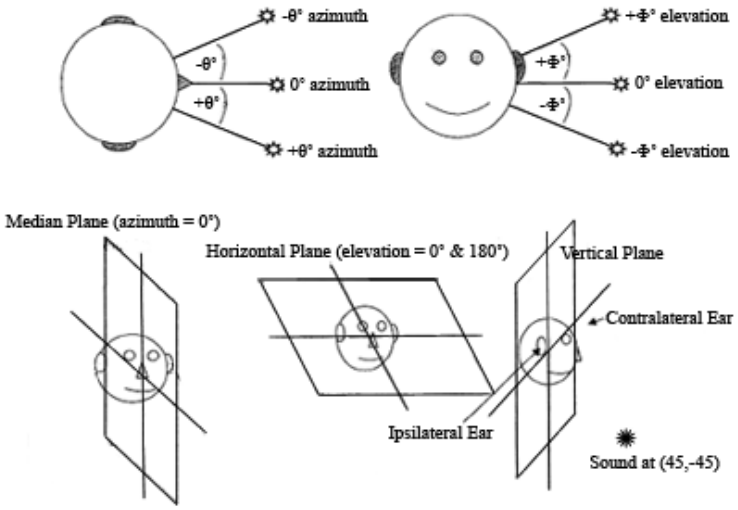

Figure 1. The coordinate system (top) and terminology used throughout this paper (bottom), from [13].

1.2 Interaural Time Difference (ITD)

One of the results of the cross-correlation operation performed by the brain is known as the interaural time difference (ITD). It is intuitively defined as the difference in arrival times of a sound's wavefront at the two ears [18]. Since both ears are isolated from one another by the head, a sound will arrive earlier at the ear in which it is closest. As an extreme example, imagine that a sound originates at (90,0)--a point to the immediate right of a listener. Because of its location, it will arrive at the listener's right ear first. After its arrival at the ipsilateral ear the sound is then diffracted around both the front and back of the head before it arrives at the left ear. The amount of time it takes to travel to the contralateral ear is the ITD. Similarly, if a sound is located to the direct left of a listener at (-90,0) the ITD will be the same as it was in the prior example; however, due to inhibitory and excitatory signals sensed by the Lateral Superior Olive (LSO), the brain knows the ear at which the sound first arrived and is therefore able to correctly interpret the ITD. Both of the aforementioned cases produce the maximum possible ITD. The third obvious ITD example occurs when a sound propagates from directly in front of a listener's face (0,0). If a symmetric head is assumed, then the ITD will be zero because the sound will arrive at both ears at the same time. All other possible ITD values are between these two extremes.

Above approximately 1.5 kHz, the accuracy of the ITD falters. In such cases, wavelengths are shorter than the diameter of the head and aliasing occurs. The phase difference of the signals arriving at the left and right ears no longer corresponds to a unique spatial location. For example, if the wavelength of a sound is equal to 1/3 of the diameter of the listener's head, its arrival time at the contralateral ear can be one, two, or three cycles delayed from when it arrived at the ipsilateral ear. Each cycle of delay corresponds to a

different ITD and, in turn, a different azimuth. In such cases, the brain gets confused and does not know which location is correct. A visual example of the aliasing that occurs when the ITD is greater than a wavelength is shown in Figure 2. Because of this frequency dependence, it is obvious that the brain cannot sufficiently localize a sound in the horizontal plane solely from the ITD; at least one additional cue is necessary.

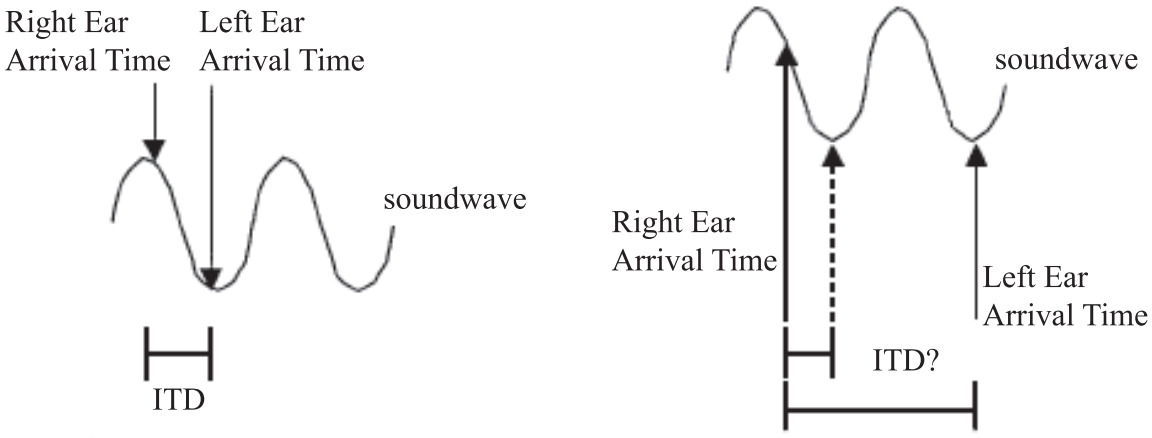

Figure 2. An example showing an ITD of less than a wavelength (left) and greater than a wavelength (right), after [13].

1.3 Interaural Intensity Difference (IID)

Another result of the cross-correlation operation that is performed by the brain is the Interaural Intensity Difference (IID). Fortunately, the IID is most accurate at the frequencies in which the ITD fails. For frequencies above 3 kHz, the listener's head has a substantial shadowing effect on incident waves. As a result, the intensity of the high-frequency portion of the sound that ultimately arrives at the contralateral ear is less than that of the ipsilateral ear. This interaural amplitude difference is the IID. Frequencies below 3 kHz are not attenuated enough by a listener's head because their wavelengths are greater than the diameter of the head; therefore, reliable IID cues are not available to the brain at low frequencies. For example, if a small paperback book is placed in front of a

loudspeaker's tweeter, the high-frequency sounds emanating from that driver will be attenuated to the listener on the other side of the book. If that same book is placed in front of a 15" subwoofer, the amplitudes of the bass frequencies produced will be unaffected by the presence of the book. In spatial hearing the head has the same effect on sound that the paperback book does--it attenuates high frequencies and does not alter low frequencies.

The complementary behavior of these two cues is the foundation of the Duplex Theory that was proposed by Lord Rayleigh over a century ago [11]. He postulated that when localizing sounds in the horizontal plane the ITD and IID complement each other very well but not perfectly. As a result, human localization of sounds is weakest in the horizontal plane for the band of frequencies between about 1.5 and 3 kHz and very accurate for all other frequencies in the audible range. In fact, of the three spatial dimensions, the horizontal plane is by far the most sensitive to humans. It possesses an impressive difference limen of 1° (approximately 10μs) for discerning the left/right orientation of sounds near the median plane--a value that is smaller than the time between samples at 44.1 kHz. For sounds away from the median plane, the difference limen increases to around 3°, and in the rear horizontal plane the just noticeable difference is equal to approximately two times its corresponding value in the frontal hemisphere. The just noticeable difference for sounds in the horizontal plane is also frequency dependent [11], as was previously mentioned, thus making it a complex phenomenon that is dependent upon many variables.

1.4 Spectral Cues

If the head is modeled as a rigid and symmetric sphere, then the ITD and IID cues explained by Rayleigh's Duplex theory are equal for sounds at a constant azimuth with different elevations. This locus of points where the ITD and IID are constant is referred to

in the literature as the "cone of confusion" for ITD or the "tori of confusion" for IID [11]. A visual example of a cone of confusion is provided in Figure 3.

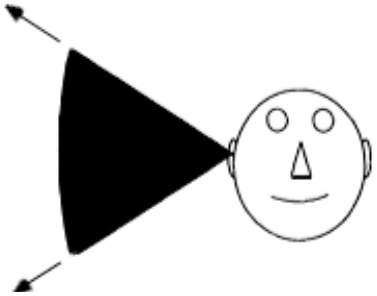

Figure 3. The cone of confusion [13].

In reality, the human head is not a perfect sphere and not symmetric about the ears, so the IID and ITD do vary slightly with elevation, but these are not dominant enough cues for the brain to determine vertical localization; another mechanism is needed. In such cases, the brain uses spectral cues to perceive the heights of sounds. As is the case with the ITD and IID, these spectral cues are also closely tied to human anatomy. The body part most responsible for the perception of elevation is the outer ear (pinna); the torso plays a supporting role. The complexities of the folds of the pinna reflect, diffract, shadow and disperse sound waves before and after they initially reach the tympanic membrane [11], all of which are phenomena that contribute to notches and resonances in the Fourier domain. In this sense, elevation perception is primarily a monaural phenomena: one ear can be plugged and the elevation of a sound can still be accurately perceived [19]. Since the pinna and its individual parts are very small, only frequencies above 3 kHz are affected by it. This could lead one to the logical, yet incorrect, conclusion that in order for a sound to be perceived as having height it must be relatively wideband and contain a good amount of high-frequency energy. This is not true because humans do possess the ability to detect the elevations of low-frequency sounds [1] due to the torso's effect on the vertical localization process.

When a sound is located above the head, the torso acts as a reflector and modifies the low-frequency content of the wave. When a source is located significantly below the head, the torso shadows the sound. Since the torso is significantly larger than the head, it shadows a greater number of low frequencies. These spectral modifications due to the torso, as weak as they may be, are also believed to be used by the brain as elevation cues--especially for locations away from the median plane [1]. Experiments have been performed on human subjects and acoustic mannequins (e.g., KEMAR) that prove the effects of the pinna and torso on elevation localization [1, 4, 7, 8, 11, 17, 24, 27].

Due to the interaural-polar coordinate system used in this work, the elevation coordinate also discriminates front from back. This is fitting because the pinna (along with some torso reflections) contributes to the localization cues that are used by the brain to differentiate between front and back. When a sound is in front of a listener, the entire outer ear is flared open towards the sound which allows it to enter the ear canal with relatively few large obstructions. In the contrary case, when a sound is behind a listener, the opening of the ear canal is not directly exposed and the flare of the pinna acts as a low-pass filter which greatly attenuates most of the frequency spectrum above 2 kHz. Although spectral cues from various body parts aid the brain in distinguishing between front and back, the most important perceptual indicators for such a distinction are visual cues and very small head movements. In fact, visual cues help the brain in all cases; however, they are most effective, and often necessary, in the front/back situation.

As is the case with localization in the horizontal plane, the just noticeable difference for localizing sounds in the median plane is dependent upon more than one variable. It varies due to spectral content and familiarity of the sound. In the median plane the

difference limen is approximately 17° for continuous speech by an unfamiliar person, 9° for continuous speech by a familiar person and 4° for white noise [11].

1.5 Range Dependence

The third and final dimension of directional hearing that remains to be discussed is that of depth. Humans are able to perceive how far away a sound is by analyzing the arrival times of its reverberations. Large amounts of reverberation occur for sounds that originate far away from the listener.

This dependence of distance perception on reverberation can be proven in an anechoic chamber where only the direct sound reaches the listener because there are no room reflections. In an experiment explained by Pierce [14], two loudspeakers are set up in an anechoic space at the same azimuth but with one being closer to the listener than the other. For the test, a sound is played from one of the two speakers and the listener is asked to identify the speaker from which they heard the sound. The intensities are controlled so that the sound coming from the further speaker is equally as loud at the listening position as the sound that is played from the closer speaker. After listening to many sounds from each speaker the subject always chose the closer speaker when attempting to identify the source of the sound. This is because in an anechoic environment the brain has no reverberation cues to determine which sound is further away. Cues in this dimension are dependent upon room dimensions rather than anthropometry but they are of interest if a sound played through headphones is to be "externalized;" however, it is often the case that increasing externalization reduces azimuth and elevation accuracy [12] so, as with all engineering solutions, a compromise must be found.

For humans, localizing in this dimension is the least accurate of the three. There is no real difference limen because depth localization is more of a relativistic endeavor. One sound can be perceived as closer than another sound but quantifying that difference is very difficult.

It is worth mentioning that the distance of a source does not have an effect on the ITD as long as the source is greater than five times the distance of the head's radius away from the listener [16]. This is useful to know when recording HRTFs--a topic that is explained in the next chapter.

1.6 CONCLUSION

It is obvious from the discussions in this chapter that the cues used by the brain for horizontal and vertical localization are very listener-dependent and can be linked to anthropometry. Some of the main contributing anatomical parts are the head, the torso, the shoulders, the neck and the pinnae. The remainder of this project will expand upon this observation to improve upon systems that generate customized filters from anthropometry for use in spatial sound-synthesis algorithms.

2

HEAD-RELATED TRANSFER FUNCTIONS (HRTFS)

Filters that capture the modifications imposed upon a sound by the human body have been around for quite some time [11]; however, due to recent technological advances, measuring these so-called head-related transfer functions (HRTFs) has become easier and more accurate. The analysis of these filters by researchers has facilitated the understanding of binaural human localization. Commercially, HRTFs are used in many virtual-reality and surround sound products. This chapter offers a brief overview of HRTFs, and it lays the necessary groundwork for the subsequent chapter where a novel method for synthesizing personalized HRTFs is explained.

2.1 MEASUREMENTS

As with any type of data acquisition method, much effort is required in order to isolate the parameters of interest. When measuring HRTFs the intention is to investigate the way in which the frequency spectrum varies with stimulus location; therefore, the primary variables are azimuth and elevation. All other parameters--such as room reflections, head and body motion, stimulus range, location of subject, microphone locations, and stimulus used--must be held constant. As a result, a common design of HRTF recording facilities involves the subject standing or sitting as still as possible in an anechoic chamber. Laser pointers are often used to align the subject's interaural center with the origin of the chamber's predetermined coordinate system [25]. Small microphones placed either at the entrances of a subject's blocked ear canals (blocked-meatus method), or at their eardrums,

record sounds that are played out of small speakers at a number of uniformly separated spatial locations. The speakers are all a constant radial distance away from the center of the subject's head; that distance is typically at least one meter. This is done to ensure that the ITD is not affected by range [16], as explained in Chapter One.

Recordings are also taken at the center of the chamber's origin in the absence of the subject. The two recordings are then divided in the frequency domain so that the alterations due to the presence of the subject remain, thus resulting in a set of HRTFs for that particular person. Time domain windowing is then applied to the recorded signals in order to truncate their impulse responses. This process ensures that the effects of the body are isolated from other potential reflections.

2.2 Applications

Once a set of HRTFs is known for a given individual, a monaural sound filtered by the left and right HRTFs of a desired location can be presented to the subject through ear buds or headphones and the sound will appear to the listener as if it is coming from that location in space. An equivalent method convolves the temporal representations of the left and right HRTFs (the head-related impulse responses, HRIRs for short) with the monaural signal [21].

Many uses for such technology are immediately obvious, especially regarding the portable media player that has become ubiquitous in contemporary society. A 5.1 soundtrack to a movie or a video game can be filtered with the appropriate HRTFs and the listener can enjoy the surround-sound experience via ear buds. HRTFs are not only of interest to the entertainment industry--the United States government also uses them to train soldiers via virtual-reality systems. Additionally, HRTFs have the potential to be used in the

cockpits of airplanes to create auditory cues for pilots. One final interesting application of HRTFs involves the sonification of multidimensional data sets.

When listening to audio through stereo ear buds (or headphones), sounds often appear as if they are coming from directly inside of the listener's head, or from just outside of their ears (depending on the location of the particular sound in the mix), because each channel's signal is fed directly to its respective ear without any crosstalk. This usually leads to listener fatigue--a phenomenon that is caused by the cognitive dissonance that the brain experiences when listening in such a way. In short, this unnatural psychoacoustic effect confuses the brain and eventually it grows tired of figuring out the sources of the sounds. Audio filtered with HRTFs relieves the brain of such confusion so that users can listen to their portable media devices longer; this is yet another practical application for HRTF technology.

2.3 SHORTCOMINGS

While all of the aforementioned uses seem promising, there are a few problems that prevent HRTF technology from becoming prolific. The first, and most important, problem arises because the human form is different for every individual on the planet. In much the same way that everyone possesses a unique fingerprint, we also all have different sizes and shapes of every other body part (especially the pinna), and this causes HRTF measurements to be exclusive to the individual from which they were measured. Through experiencing acoustic stimuli in the natural environment, an individual's brain is able to finely tune the accuracy of its sound localization ability through trial-and-error by effectively compensating for the effects that their anatomy has on incident sounds. This lifelong, and constantly

adapting, learning process essentially creates personalized HRTFs that are stored in an individual's brain.

The effects of listening to sounds filtered with non-individualized HRTFs have been studied extensively. It has been experimentally demonstrated that listening in such a way causes a definite increase in vertical localization error and front/back confusion [31]. Localization accuracy in the horizontal plane is also compromised when using non-personalized HRTFs but not nearly as much as elevation localization is. In essence, the experience is equivalent to hearing with someone else's ears. Since recording HRTFs is such an arduous and expensive process, it is not feasible to expect everyone to undergo the procedure just so that they can reap the benefits of the technology. This provides further evidence that a practical and accurate anthropometry-based HRTF synthesis method is of interest.

Additionally, there are a couple of problems that arise when recording the HRTFs of human subjects. Typically, small speakers are used when playing the stimuli and even smaller microphones are used in the recording process; this prevents the low-frequency response of HRTFs from being accurate. If larger speakers and better microphones are used, even the best anechoic chambers still have a difficult time absorbing very low frequencies which means that such a situation may also corrupt the low-frequency response of recorded HRTFs. Although some researchers believe that there are meaningful localization cues in these erroneous low frequencies, the perceptual significance of the frequency region below 500 Hz still remains enigmatic.

Elevations below -45° are also problematic when recording HRTFs because it is difficult to place speakers at such low elevations, especially if the subject is seated--which is often the case. When a subject is seated, sounds produced at low elevations will reflect off

of his/her knees. These additional reflections corrupt the accuracy of the recordings at such locations.

One final problem that currently plagues HRTF synthesis systems is that of interpolation. Most analytical HRTF data is measured at spatial intervals of 4° (or more); therefore, interpolation is needed when panning effects are desired or else discontinuities will be heard by the listener. Interpolating between measured HRTFs is not a trivial process. Many simple implementations that interpolate linearly in the time or frequency domain result in localization error [25], and more complex interpolation algorithms are difficult to compute in real time. Structural feature-based algorithms show some promise since they bridge the gap between interpolation and synthesis; however, if a personalized HRTF modeling algorithm existed, then it would be possible to calculate the HRTF at any desired spatial location in real time and eliminate the need for interpolation.

It can be concluded from this section that a method to synthesize personalized HRTFs from anthropometry is not only of great interest to the research community but to commercial industries as well. If such a method were developed, it would also provide some insight into, and perhaps eliminate, the previously mentioned low-frequency and low-elevation problems. It is the aim of this work is to improve existing anthropometry-based HRTF synthesis systems to eradicate the localization problems that arise when non-personalized HRTF recordings are used in binaural synthesis systems.

2.4 The CIPIC Database & Anthropometry

The database of HRTFs that is used throughout this project was recorded at the CIPIC Interface Laboratory at the University of California Davis [6]. It contains HRTFs for 45 subjects at 1250 unique points (25 azimuths and 50 elevations), all of which were

recorded at a radius of one meter away from the origin of the interaural coordinate system using the blocked-meatus technique that was briefly described in the first section of this chapter.

Fortunately, this database also contains a large amount of anthropometric measurements for most of its subjects. Figures 4 and 5 show the morphological dimensions for the upper body and the pinna, respectively, that are provided with the CIPIC database. In addition to the measurements shown in Figures 4 and 5, the following measurements are also available: height, seated height, head circumference and shoulder circumference. All of these anatomical measurements, and the terms used to name them, comprise the anthropometric basis that will be used in the remainder of this work.

The measurements in the right half of Figure 4 are only provided for one side of the body while the measurements in Figure 5 are provided for each ear. Figure 6 shows a labeled diagram of the human pinna that contains some additional anatomical terms that are not covered in Figure 5. The cymba (concha) and the cavum (concha) are treated as a single entity in this paper and referred to simply as the concha.

In Figure 4 the measurements corresponding to x_1 through x_{13}, respectively, are head width, head height, head depth, pinna offset down, pinna offset back, neck width, neck height, neck depth, torso top width, torso top height, torso top depth, shoulder width and head offset forward. The pinna measurements in Figure 5 labeled d_1 through d_8 are cavum concha height, cymba concha height, cavum concha width, fossa height, pinna height, pinna width, intertragal incisure width and cavum concha depth. The two angles provided (θ_1 and θ_2) are the pinna rotation angle and the pinna flare angle, respectively. Even though all of these measurements are available, it is of interest to use as few of them as possible to synthesize perceptually accurate personalized HRTFs.

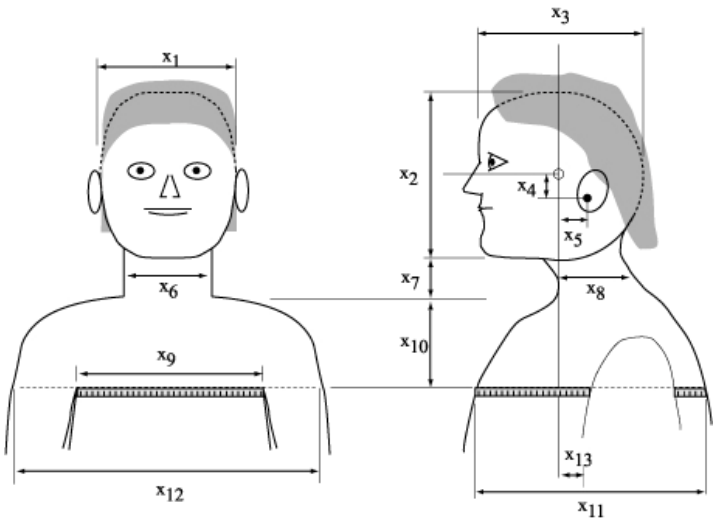

Figure 4. CIPIC head and torso measurements [6].

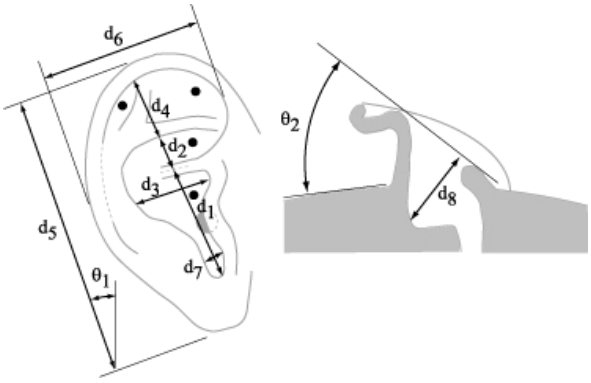

Figure 5. CIPIC pinna measurements [6].

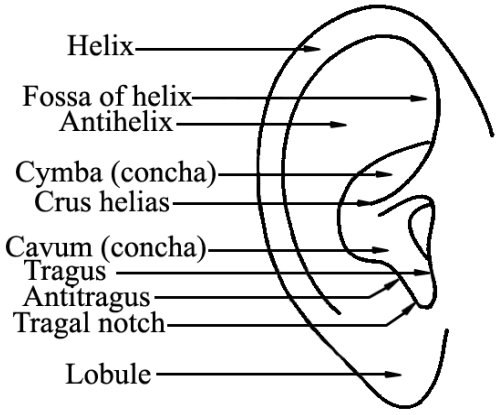

Figure 6. The anatomy of the human pinna, after [30].

2.5 Structural Decomposition

Much of the intricate behavior of HRTFs can be ascribed to the complex interaction between the separate effects of the head, torso, shoulders and pinnae. All of these body parts affect the frequency response of HRTFs in different, but often overlapping, frequency bands. In the interest of modeling HRTFs, it is easier to account for the effects of each body part separately and then cascade the responses of those individual filters together to form a complete HRTF. It has been shown in [4] that this approach results in HRTFs that closely resemble their measured counterparts.

In order to prove their HRTF decomposition hypothesis, the authors of [4] used a KEMAR with removable pinnae. This allowed the response of the KEMAR sans pinnae to be measured, thus isolating the combined contribution of the head and torso to the HRTF. It also made it possible to measure the response of the removed pinna on an "infinite plane" at the same set of spatial locations as the head and torso. The results of these measurements are shown in the top half of Figure 7. These individual responses were then cascaded together in series as shown in Figure 8, where H_{hat} represents the contribution of the head and torso and H_p denotes the contribution of the pinna. The resulting composite HRTF was virtually identical to the HRTF measurements taken with the pinnae-equipped KEMAR. Plots of both of these responses can be seen in the bottom half of Figure 7.

In Figure 7 the HRTFs are displayed as an image. This is a common and convenient way to display multiple HRTFs in one plot. These images can be thought of as spectrograms in way that they display three variables in two dimensions. Since there are four main quantities (frequency, magnitude, azimuth and elevation), and only three can be displayed as variables, one must be held constant. The static quantity is always either an elevation angle or an azimuth angle. In the case of Figure 7, elevation is varied along the x-

axis and the azimuth angle is held constant at 25°. This means that each column of the image corresponds to an HRTF at an azimuth of 25° and the elevation indicated by the x-axis coordinate. The vertical axis plots frequency with the DC component at the top of the image and the Nyquist value at the bottom of the image. The amount of brightness represents the magnitude (in decibels) of the HRTF at a given frequency and spatial location. The scales relating brightness to decibels can be seen on the right side of Figure 7.

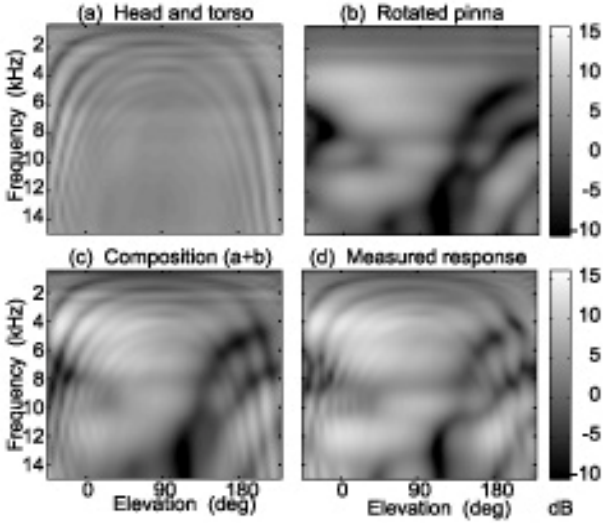

Figure 7. Individual response of the head and torso (a), individual response of the pinna (b), composition resulting from cascading (a) and (b) according to the diagram in Figure 8 (c), and the measured HRTF response (d). All images are for a cone of confusion at 25° [4].

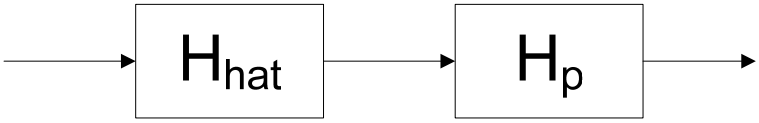

Figure 8. The equivalent single path resulting from cascading the head and torso filter with the pinna filter, after [4].

It can be observed from Figure 7 that the response of the head and torso is relatively simple when compared to that of the pinna. It is also much more uniform which should make it very predictable and easy to model. The response of the pinna is more sporadic and

much less predictable. This could be problematic when trying to model all of the complex interactions that occur due to the pinna; this concern will be addressed in the next section.

It is worth mentioning that since the images in Figure 7 are from a KEMAR with symmetric anthropometry and simplified pinnae, human measurements are expected to be slightly more complex. Regardless of that fact, the HRTF decomposition investigation presented in this section proves that it is possible to create an HRTF by cascading models of its contributing body parts together in series. Even though this idea has been proven conceptually, it has never been put to practical use in an HRTF synthesis system. The synthesis method described in Chapter Three will take advantage of this very valuable discovery by using it as one of the implementation's fundamental concepts.

2.6 Perceptually Significant Features

Based upon the discussions from Chapter One, the results of past experiments and the fundamentals of psychoacoustics, it is evident that much of the spectral detail of HRTFs is perceptually irrelevant to the human localization process. To maintain accurate localization in the horizontal plane, it is only necessary to model the ITD and IID; this has been known for quite some time [11]. More recently, it has been established that a considerable amount of smoothing of the filters' frequency spectra has little to no detrimental effects on horizontal plane localization [23]. This is evidence that a very low-order filter can be used to model the shadowing effects that occur due to the head without compromising perceptual localization in the lateral direction.

For elevation localization isolating the perceptually salient features is somewhat more difficult but still possible. In [7], HRTFs were smoothed using auto-regressive moving average (ARMA) models and the effects of said operation on median plane localization were

studied. From Chapter One it is understood that the primary cues for elevation are due to the pinna and occur at frequencies above approximately 3 kHz. The results of the listening tests performed in [7] using HRTFs smoothed at varying degrees prove that small peaks and troughs in the frequency domain are perceptually irrelevant above 4 kHz and that the brain uses macroscopic patterns of a sound's spectrum in this frequency region to determine elevation.

Another result from Asano's smoothing experiments proves that if the spectra are smoothed below 4 kHz front/rear judgment is compromised greatly [7]. Microscopic peaks and dips below 2 kHz were isolated and established as contributors to perceptually resolving if a source is located in the front or back of a listener. Macroscopic patterns in the high-frequency region were also identified to contribute to front/back discrimination so, once again, fine spectral details can be ignored in this range without consequence.

Figure 9 shows the full resolution HRTF of subject 28 in the CIPIC database at (0,0) and Figures 10, 11, 12 and 13 show examples of the same HRTF at four successive smoothing levels. Figure 13 demonstrates the maximum amount of smoothing that can be done without compromising elevation localization. Note that although the entire bandwidth of the HRTF is smoothed in these figures, acceptable localization performance is only obtained when the smoothing is limited to frequencies above 4 kHz.

In [29], it is explained that all frequencies above 15 kHz are perceptually irrelevant in the localization process for a majority of listeners. Additionally, CIPIC's HRTF measurements are believed to be inaccurate at such high frequencies, yet this inexactness does not affect localization; therefore, it is safe to assume that the frequencies above approximately 15 kHz in the figures below are perceptually irrelevant.

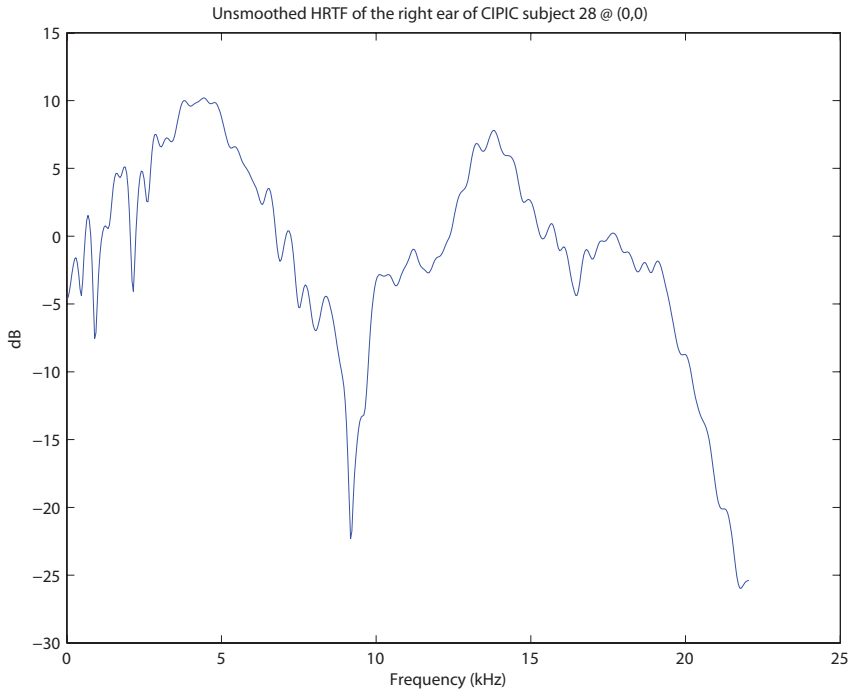

Figure 9. Frequency response of the unsmoothed HRTF of CIPIC subject 28 at a location of (0,0). Image based upon data from [6].

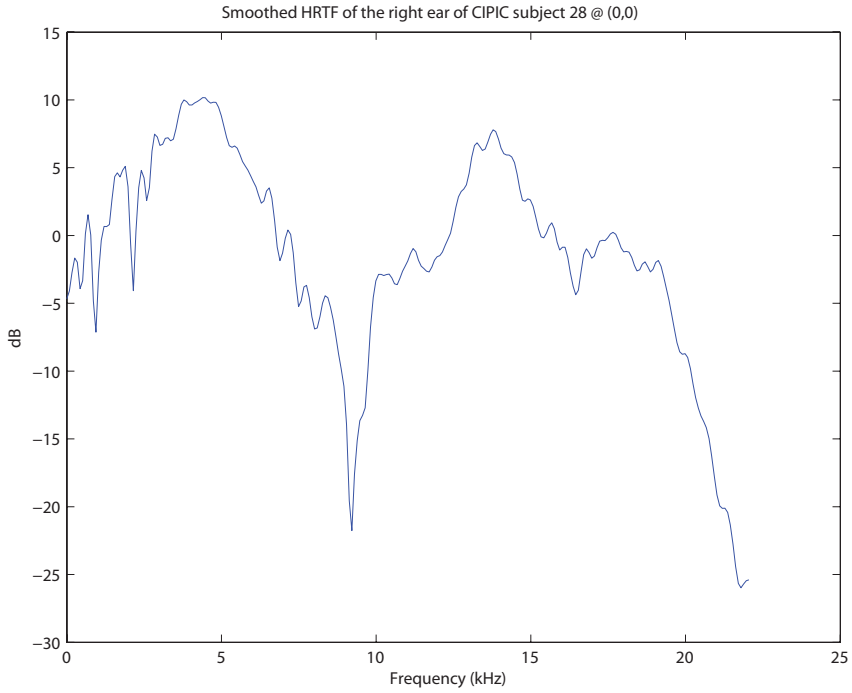

Figure 10. Frequency response of the HRTF of CIPIC subject 28 at a location of (0,0) plotted at a frequency resolution of 256 non-redundant frequency bins. Image based upon data from [6].

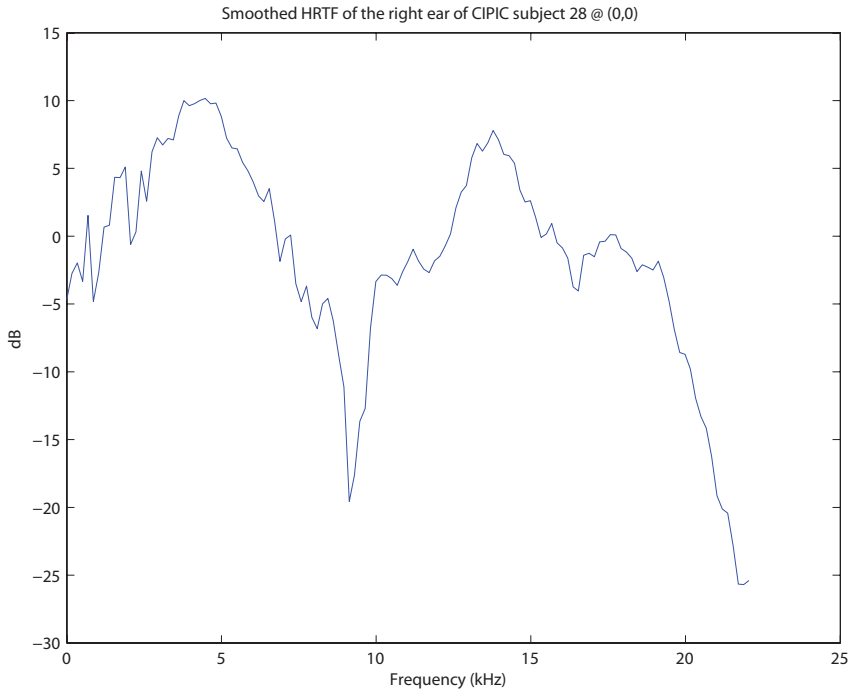

Figure 11. Frequency response of the HRTF of CIPIC subject 28 at a location of (0,0) plotted at a frequency resolution of 128 non-redundant frequency bins. Image based upon data from [6].

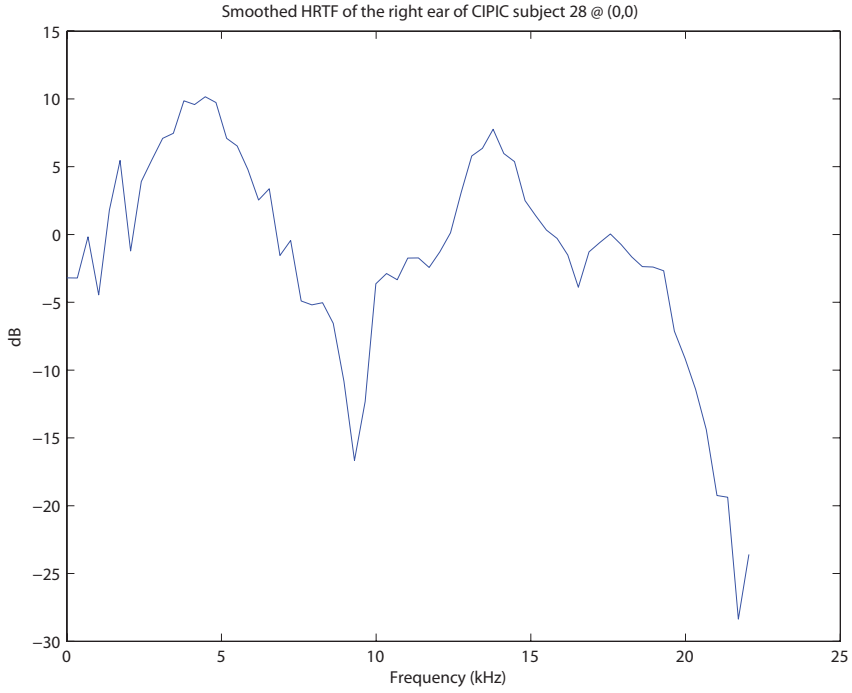

Figure 12. Frequency response of the HRTF of CIPIC subject 28 at a location of (0,0) plotted at a frequency resolution of 64 non-redundant frequency bins. Image based upon data from [6].

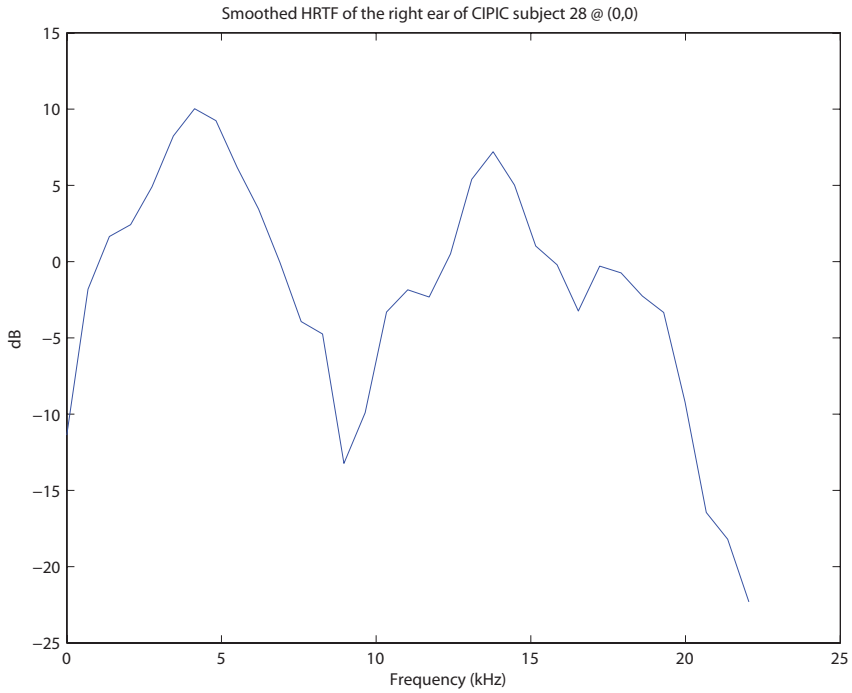

Figure 13. Frequency response of the HRTF of CIPIC subject 28 at a location of (0,0) plotted at a frequency resolution of 32 non-redundant frequency bins. Image based upon data from [6].

All of the aforementioned phenomena can be explained by simple psychoacoustic principles. The human auditory system is known to have a high frequency resolution at low frequencies and a low frequency resolution at high frequencies; this is evident from the bandwidths of the auditory system's critical bands [14]. This explains why the presence of only macroscopic changes in frequencies greater than 4 kHz is needed for accurate spatial perception while preservation of the microscopic patterns in the low-frequency region is necessary to maintain proper localization. Also, the upper limit of hearing for most adults is rarely above 16 kHz; this explains the lack of attention that needs to be paid to the highest recorded frequencies of HRTFs. Knowing and understanding the conclusions drawn in this section, from prior work, is very important to the objective and will be exploited in the HRTF synthesis implementation explained in the next chapter.

2.7 THE CONTRALATERAL HRTF

In the median plane, the HRTFs for the left and right ears are very similar (for obvious reasons); however, as a source moves away from the median plane the HRTF for the contralateral ear becomes more complex than its simpler ipsilateral counterpart. Figure 14 demonstrates this phenomenon by displaying the HRTFs for both ears at three different cones of confusion that correspond to azimuths of 10°, 30° and 65°. In this figure, each subsequent azimuth angle is further away from the median plane than the prior one.

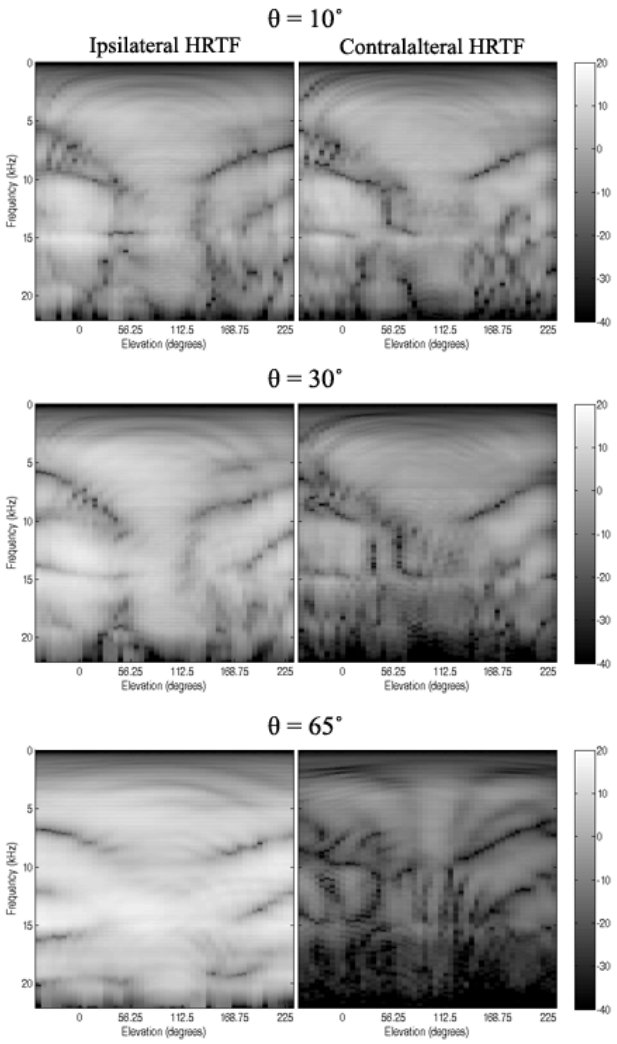

Figure 14. The ipsilateral and contralateral responses for cones of confusion of θ=10°, θ=30° and θ=65°. Image based upon data from [6] (Subject 3).

The top third of Figure 14 presents the HRTFs for each ear at an azimuth of 10°. At this azimuth angle the ipsilateral and contralateral HRTFs are very similar. This is expected because a θ of 10° is very close to the median plane. The next azimuth angle of θ=30°, which is shown in the middle third of the figure, reveals the contralateral response increasing in complexity and the ipsilateral response increasing in simplicity. This is evident when they are compared to the plots from θ=10° and when compared to each other (i.e. the relative complexity difference between them is increasing). Finally, in the bottom third of the image, for an azimuth of 65°, the contralateral response is far more intricate than the much simpler ipsilateral response.

The increasing complexity of the contralateral response is due to the high-frequency shadowing of the direct sound by the head which results in the creation of small magnitude secondary waves that arrive at the contralateral ear. Other noteworthy contributors to the complexity of the contralateral HRTF away from the median plane are time domain bright spots and complex interference patterns [9]. These intricate characteristics present a great problem to the task at hand because modeling the contralateral HRTF is seemingly a far more complicated problem than modeling the ipsilateral HRTF. Fortunately, work has been done that simplifies the process without introducing excessively large amounts of error.

In [9], an intuitive method is derived to model the contralateral HRTF from the frequency response of a spherical head and the ipsilateral HRTF. This means that an HRTF synthesis algorithm only needs to model the ipsilateral HRTF and then the contralateral filter can be derived from it based on

$$H_c(\omega,\theta,\phi) = F(\omega,\theta,\Phi)H_i(\omega,\theta,\Phi), \qquad (1)$$

where ω represents frequency, θ represents the azimuth angle, $H_i(\omega,\theta,\Phi)$ is the ipsilateral response, $H_c(\omega,\theta,\Phi)$ is the contralateral response and $F(\omega,\theta,\Phi)$ is the transformation function.

The transformation function can be calculated using

$$F(\omega,\theta,\phi) = e^{-jD(\theta,\phi)} \frac{S_c(\omega,\theta)}{S_i(\omega,\theta)}, \tag{2}$$

where $S_c(\omega,\theta)$ and $S_i(\omega,\theta)$ are functions that account for head shadow effects and other individual characteristics of the contralateral and ipsilateral responses, respectively. An appropriately designed spherical head model based upon anthropometry can approximate these two functions. The first term on the right hand side of (2) is to account for the time delay between the ipsilateral and contralateral ears. It can be calculated using

$$D(\theta,\phi) = T_c(\theta,\Phi) - T_i(\theta,\Phi), \tag{3}$$

where $T_c(\theta,\Phi)$ is the arrival time at the contralateral ear and $T_i(\theta,\Phi)$ is the arrival time at the ipsilateral ear. These two *T* terms can be approximated using the aforementioned spherical head model. It is worth noting that $D(\theta,\Phi)$ is dependent upon both azimuth and elevation which agrees with what was established in the first chapter. The reason for the slight dependence of ITD on elevation is due to ear offsets which will be addressed in the following chapter.

This simplistic contralateral model was evaluated both objectively and subjectively by the authors of [9], and it produced the results that they expected to witness. The greatest amount of error was found to occur at the interaural poles and at the elevation extremes (below front and below back) while the least amount of error was experienced in and around the median plane. It was concluded that the approximated contralateral HRTFs performed

within an average of 5° when compared to the measured contralateral responses. Such a relatively large amount of error is not ideal for an HRTF synthesis system so, in an attempt to reduce said error, a slightly modified version of this model is introduced in Chapter Three.

This chapter has introduced HRTFs, their uses, and their features by surveying the literature that is most relevant to the objective of this work. In subsequent chapters, when the task at hand is more directly addressed, the information presented in this chapter will be put to practical use.

3

IMPLEMENTATION

While it has been demonstrated that a complete set of individualized HRTFs can be computed by solving the wave equation for every boundary condition presented by the surface of the human body, this method is both computationally very expensive and analytically beyond reach [4]. As computer hardware evolves, this technique may eventually replace acoustic recordings of HRTFs; however, such approaches are still far ahead of current technology, thus a computationally efficient method for synthesizing HRTFs is of great interest.

In this chapter the details of current HAT synthesis models are explained and then a method to improve upon them is introduced. The new method exploits the prior art introduced in Chapter Two and uses digital filters whose parameters are determined from anthropometry to model the response of the human pinna at elevations in and around the median plane. The most notable novel contributions of this work are in the details of the pinna model, particularly with the inclusion of elevation cues, and in the way that the components of the entire HRTF model are connected.

3.1 DESIGN & PROPOSED SOLUTION

The filter-cascading technique of Section 2.5 is used as the underlying basis for the synthesis method described in this chapter. The first part of the design uses an existing Head and Torso (HAT) model to account for the ITD, the IID and subtle torso elevation cues. The anthropometric parameters of the HAT model are head height, head width, head depth, torso height, torso depth, torso width, neck height, left pinna offset back, left pinna

offset down, right pinna offset back and right pinna offset down. It outputs two filters: one for each ear.

The left and right pinna responses at (0,0) are then modeled using a series of digital filters and delays that are designed to approximate the first-order acoustic characteristics of the human pinna. By making use of the fact that the pinna-related transfer function (PRTF) can be smoothed greatly without compromising localization (from Section 2.6), it is apparent that only a few filters are necessary to model the PRTF accurately. The anthropometry inputs to the median plane pinna model are concha depth, concha width, crus helias distance and pinna width (distance to helix). The latter three measurements are taken at a series of elevations.

The focus of the pinna modeling work in this project is limited to creating elevation cues in and around the frontal median plane. The rear hemisphere is ignored because front/back confusion still sometimes exists even when sounds are filtered and played back for a listener using their own HRTFs. Although evidence does exist showing that subtle head movements and the presence low-frequency spectral detail play important roles in resolving the ambiguity, this problem remains largely unsolved [7]. Additionally, when an incident sound is located behind a listener's head, the mechanisms that cause spectral notches are not clear.

Azimuths far away from the median plane are not dealt with in this work because doing so would require gathering three-dimensional anthropometry data from the pinnae, and such acquisitions are impossible to do from digital images. Conceptually proving that the proposed pinna modeling algorithm is effective near the median plane will lay the ground work for future researchers to extend its functionality to all azimuths and elevations in the frontal hemisphere if a three-dimensional model of the ear is available.

Figure 15 shows a block diagram that outlines how all of the modules of the design connect at the highest level of abstraction. A unit step of length 256 samples is inputted into both the median plane pinna model and the HAT model which results in the outputs of each of these components being the impulse responses of the given system. Both models output a separate impulse response for each ear. The other two outputs from the HAT model in Figure 15 are the head-shadowing filters that are necessary if the azimuth of the incident sound is not 0°. In such a case, a conditional statement determines which ear is on the contralateral side of the head and then filters the response of each output of the median plane pinna model accordingly. Since the ITD is already accounted for in the HAT response it is not necessary to include time delays with the filtered pinna responses. This method of approximating off-the-median-plane pinna responses is a modified version of the algorithm introduced in Section 2.7. The final step of the algorithm involves cascading the left and right pinna responses to the left and right HAT responses, respectively, in accordance with the signal flow diagram previously introduced in Figure 8.

Modifications to the contralateral HRTF algorithm from Section 2.7 are used because the HAT model outputs the unique low-frequency responses of each ear; therefore, it is not necessary to use the entire ipsilateral HRTF to approximate the entire contralateral response. Also, since the pinna model only calculates the left and right ear responses at median plane elevations, whenever the incident sound is located off of the median plane the left and right pinna responses needs to be filtered with the appropriate head-shadowing filters in order to maintain accurate horizontal localization. In subsequent sections it will be seen that this method is most effective at azimuths close to the median plane. Once the azimuth increases to a value that corresponds to a primary reflector that is no longer two

dimensional this algorithm loses accuracy. Further details regarding the effectiveness of this method are discussed with the objective results in the next chapter.

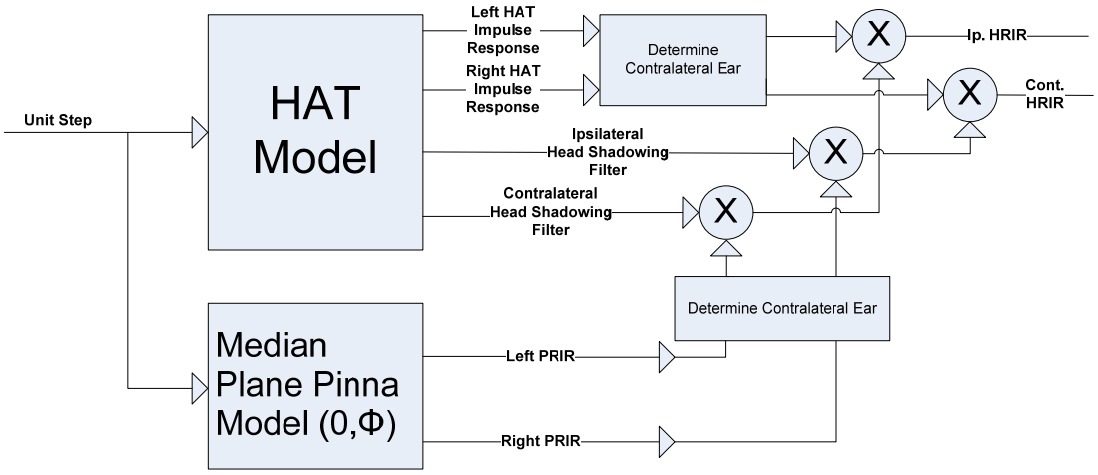

Figure 15. The block diagram for the system at the highest level of abstraction.

3.2 HEAD AND TORSO (HAT) MODEL

The HAT model used in this implementation is one that approximates both the head and torso as rigid spheres--giving the simplified shape of a snowman to the human anatomy. Work has been done to approximate both body parts as ellipsoids with great results [3, 15], but it has been shown that by adding three additional features to spherical models equally accurate results can be produced with less computational overhead.

One of the main problems with modeling complexly shaped body parts as simple and symmetric geometric forms is that an unnatural bright spot is introduced in the frequency response at an observation angle of 180°. This bright spot occurs at 180° because the waves travel equal distances around the symmetric shape and add in phase at the observation point on the opposite side. Due to the asymmetry of human heads, an actual bright spot exists around 155° in measured HRTFs. This observed bright spot is also less prominent that the bright spot exhibited by simple geometric models [3]. By keeping this in

mind, a spherical filter model can be designed that closely approximates the actual response of human body parts.

The second concern that must be addressed when designing a spherical filter model to approximate the response of the head is that of ear placement. It has been shown that the ITD is slightly dependent upon elevation and that this may be used as a secondary cue for low-frequency elevation detection [1]. Head models that take into account ear offsets can accurately approximate the ITD's slight elevation dependence. All of the subjects in the CIPIC database have ears that are located lower than the midpoints of their heads. This observation explains the increase in ITD that is witnessed when a source is elevated and away from the median plane because, in such cases, the sound must travel a greater distance around the head before reaching the contralateral ear than it would if it was in the frontal horizontal plane. For simplicity, most traditional spherical models of the head place the ears along the horizontal axis that goes through the center of the sphere; however, a truly accurate spherical model of the head must take ear offsets into account.

There is one final problem that must be solved when using simple spheres to model complexly shaped human body parts and that is how to calculate the radius of the spheres from anthropometry. Figure 4 indicates that three measurements (depth, width and height) characterize the size and shapes of both the head and the torso. Any one of these three parameters can be used by itself to determine the radius of the sphere, but a thorough method should incorporate all three measurements into the radius calculation.

Chapter One established a minimum difference limen of 1° in the horizontal plane; this value corresponds to an allowable error in head radius of 1%. It is therefore evident that when using a spherical head to approximate the ITD and the IID that is not strategically based upon the head shape trends shown across a large number of human subjects there will

be localization error. Calculating the radius of the spherical torso is not as important of a problem because the size of the torso does not have as direct of an effect on perceptually significant features of the HRTF as the head radius does; this will be explained in more detail later in this chapter.

Most spherical models of the head use an average value of 8.75cm as the sphere's radius and hope it produces sufficient results. While this approach may work for those whose head sizes are close to the mean, large localization errors will result if a subject's head size is much larger or smaller than the average size. This would not be a problem if the standard deviation of head sizes across the human population was very small, but an analysis in [2] indicates that head size was found to deviate from the mean by about 30%--which is a considerable amount. Also in [2], a few methods of calculating spherical head radius were tested. The worst performing method proved to be simply averaging the three measurements of half head width, half head height and half head length. Doing so resulted in a large overestimation of ITD. The most effective method was found by using the weighted average equation

$$a_e = w_1 X_1 + w_2 X_2 + w_3 X_3 + b, \tag{4}$$

where the w terms represent the weights of X_1, X_2 and X_3 which are half head width, half head height and half head length, respectively. a_e denotes the resulting head radius and the bias term b is a constant offset.

A regression analysis was run on the data for all of the subjects in the CIPIC database and the ideal values of the weights in (4) were found to be: $w_1 = .51$, $w_2 = .019$, $w_3 = .18$ and b = 3.2cm. When this radius calculation method is used with the provided weights, it results in an error that is acceptable at all locations except for those of high elevations. If the ears are offset properly according to the subject's anthropometry, then the

elevation error is significantly reduced. The implementation discussed in this section uses the head radius calculation in (4) in conjunction with offset ears and a specially designed shadowing filter to obtain the best possible localization performance from the HAT model.

The parameter of torso radius is far less important than that of head radius. In [3], it is stated that using the geometric mean of the half height, half width and half depth of the upper torso is sufficient for accurately modeling the radius of the spherical torso. The reasons for the lack of attention to detail regarding the torso radius calculation will be justified in the subsequent sections.

3.2.1 Acoustic Filter Model of a Rigid Sphere

Since the shapes of the human head and the torso are approximated as rigid spheres in the HAT implementation, it is of utmost importance to explain the process of modeling the acoustic response of a sphere. Figure 16 shows a top-down view of the sphere and the parameters necessary to model its acoustic response: an observation point (P), a radius (r) and an angle of incidence (θ).

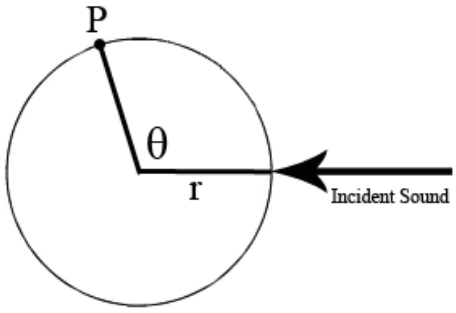

Figure 16. A top-down view of a sphere and its parameters.

If an auditory stimulus is played at a given observation angle away from the observation point, the presence of the sphere will affect certain frequencies of the sound before it reaches the observation point. When this problem is solved numerically for a fixed

observation point with observation angles ranging from 0 to 180°, high frequencies are boosted when θ is less than 105° and attenuated when it is greater. The maximum boost of 6dB occurs at an angle of incidence of 0°; the maximum cut of 20dB occurs at 150°. At 105° the response is relatively flat; therefore, there is no boosting or attenuating of any frequencies at such an observation angle. At 180° the response is also flat--this is due to the bright spot that has already been explained.

A filter designed by Brown & Duda in [12] approximates the numerical solution mentioned above while attenuating the bright spot because the presence of a flat response at 180° causes very unnatural sounding results. The transfer function of their filter is

$$H(s,\theta,r) = \frac{\alpha\tau s + 1}{\tau s + 1}. \tag{5}$$

The parameters of this first-order filter consist of a time constant τ and an asymptotic high-frequency gain α which is a function of θ. The time constant can be calculated using

$$\tau = \frac{r}{2c}, \tag{6}$$

where c is the constant value of the speed of sound (343 m/s) and the variable r is the sphere's radius. Since the denominator of (5) is only dependent upon the radius of the sphere, the filter's pole remains at a fixed location for a given listener. This facilitates real time implementation.

The formula to calculate $\alpha(\theta)$ is

$$\alpha(\theta) = \left[1 + \frac{\alpha_{\min}}{2}\right] + \left[1 - \frac{\alpha_{\min}}{2}\right]\cos\left[\frac{\theta}{\theta_{\min}}\pi\right], \tag{7}$$

and it introduces two new parameters: $\alpha_{\min}$ and $\theta_{\min}$. The values of these two terms are calculated such that the response of the filter matches up well to the previously discussed

numerical solution while also attenuating the bright spot at 180°. With values of $\alpha_{min} = .1$ and $\theta_{min} = 150°$ a flat response results at an angle of incidence of 77.5° (θ_{flat}). This value is much less than the angle of 105° that was observed in the numerical solution, but this filter results in a much more natural sounding response and a more realistic bright spot when compared to the numerical solution. The response of the filter model at 180° has an identical envelope to the curve that corresponds to an angle of incidence of 120°. This provides much more attenuation at the artificial bright spot than the numerical results. A plot of the response curves at observation angles from 0° to 180° at 15° increments for a sphere with a radius of 8.75cm can be seen in Figure 17.

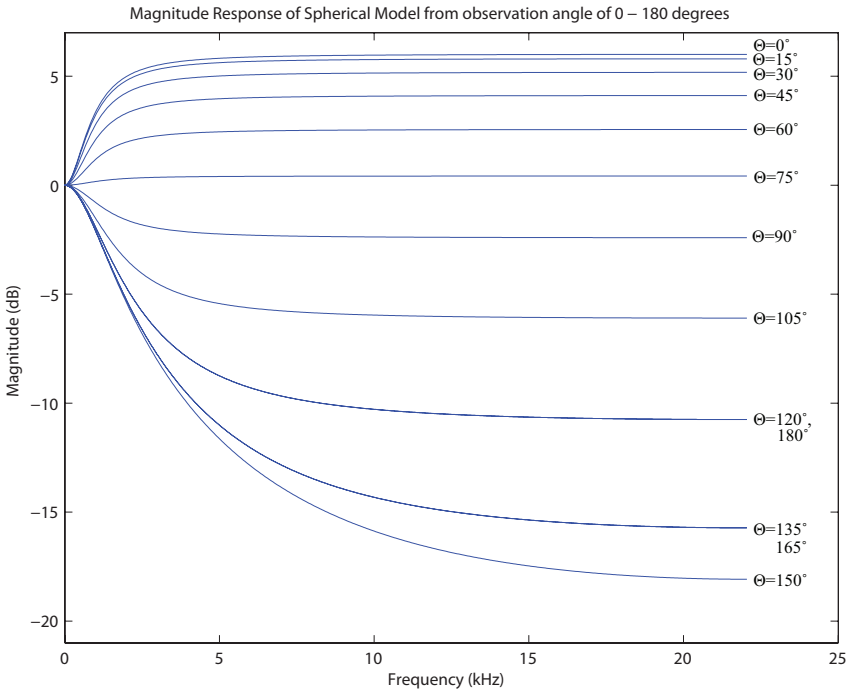

Figure 17. The frequency response of (5) at observation angles from 0° to 180° at increments of 15° for a sphere with a radius of 8.75cm. Figure created using an algorithm found in [12].

The second part of the spherical filter model is the delay module. It is cascaded with the previously discussed shadowing filter as shown in Figure 18. The time delay is determined as the time difference between when a sound arrives at the observation point and when it would arrive at the center of the sphere if the sphere were not present. Its formula,

$$\Delta T = \begin{cases} -\frac{r}{c}\cos\theta & if\ 0 \le |\theta| < \frac{\pi}{2} \\ \frac{r}{c}\left[|\theta| - \frac{\pi}{2}\right] & if\ \frac{\pi}{2} \le |\theta| < \pi \end{cases} \tag{8}$$

is a result of an elementary ray tracing analysis that assumes the source is infinitely distant. It is a function of the observation angle and the sphere's radius and is independent of frequency. It is implemented using an FIR all pass section; therefore, it does not contribute to the magnitude or phase response of the shadowing filter with which it is cascaded. A brief analysis of (8) reveals that the result will be negative when $|\theta| < \frac{\pi}{2}$; this will be dealt with in the next section.

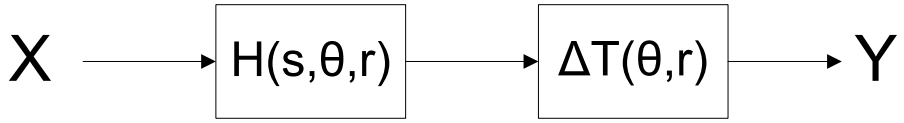

Figure 18. Structure of spherical filter model--the shadowing filter cascaded with a time delay.

3.2.2 Spherical Head Model

Using the filter model introduced in the previous section it is relatively simple to model the ITD and IID effects that occur due to the head. The IID is modeled using the filter equation in (5), and the ITD is modeled using the time delay equation in (8). To correctly approximate the response of the head, two observation points must be used: one

for each ear. Primitive head models place the ears at two points directly in line with the center of the sphere, but anthropometry studies indicate that all humans' ears are offset vertically and horizontally from the center of their head. Also, the left and right ears are almost always located at different points on each side of the head. Since ear locations have a substantial effect on the ITD, a model that is adaptable to all possible scenarios of ear offsets will be the most accurate.

The head model used in this implementation receives the vertical and horizontal distances that the ears are offset from the center of the head as parameters and uses these locations as observation points on the sphere. The ear offset measurements can be seen for the left side of the head as x_4 and x_5 in Figure 4. Details on exactly how they are measured on human subjects will be discussed in Chapter Five. Unfortunately, the CIPIC database only has the pinna offset data for one side of the head, so an unlikely symmetry must be assumed for the objective results and the ITD analysis.

To account for the negative time delay that may arise when using this model, a constant of r/c, which is equal to head radius divided by the speed of sound, is added to every calculated delay time. This has no effect on the ITD because the ITD is the difference in the arrival times of a sound at the two ears (observation points), and since the arrival times at both ears are both offset by the same constant, the relative delay is not affected.

The CIPIC database contains measured ITDs for each subject at every spatial location contained in the database. This makes it possible to input the head dimensions and ear offsets to the head model and compare the resulting ITDs to the experimental data. The ITD is most dependent upon ear location and elevation at locations away from the median plane. With this in mind, cones of confusion corresponding to azimuths greater than 40° off the median plane (in either direction) are used to compare the modeled ITDs with the

experimental ITDs. To analyze error the maximum deviation between the measured and modeled ITDs is tabulated for each subject and then averaged over all subjects in the CIPIC database. The results of this calculation indicated an average maximum error of 2.9658 samples at an azimuth of 55°. This worse-case scenario corresponds to an error of 68μs at a sampling frequency of 44.1 kHz--a value that is very near the difference limen for elevated sounds that originate away from the median plane. The same error analysis was performed across all azimuths with elevation held constant at 0°, and the average maximum error was less than two samples; this is less than the difference limen for horizontal localization away from the median plane.

It is worth noting that since the CIPIC database only provides ear offsets for one side of the head that the error values will improve when ear offsets for both sides of the head are used. Plots of four noteworthy ITD cases are shown in Figures 19, 20, 21 and 22.

Figure 19 shows a case in which the ITD is approximated very well for all elevations at a constant azimuth of -55°. The maximum error in this case is approximately two samples and is located at an elevation of 100°. The rest of the elevations away from 100° have errors of less than one sample.

In Figure 20 the contrary case is presented: one in which the ITD is not modeled as accurately. This example demonstrates that even in the case where the modeled curve does not closely follow the measured one, the maximum error is still only three samples. Figure 21 is for the same subject that was used in Figure 20 except the angle of incidence is on the other side of the head. The shape of the modeled ITD follows the measured ITD closer in this case, but the maximum error at an elevation of 100° is greater. This provides evidence that using separate offset values for each ear will improve the ability of the head model to approximate the ITD.

In Figure 22 azimuth is varied along the x-axis instead of elevation, and the calculated ITD can be seen to match up almost exactly to the measured ITD except at extreme azimuth values. The only azimuths that result in errors of greater than one sample are those near the interaural poles. The difference limen is 3° at such locations, so an error of three samples is acceptable.

In the physical world, as discussed in the first chapter, the ITD value actually varies with frequency: it is two times larger at low frequencies than it is at high frequencies. The non-linear group delay of the IIR shadowing filter precisely accounts for this frequency dependence [12].

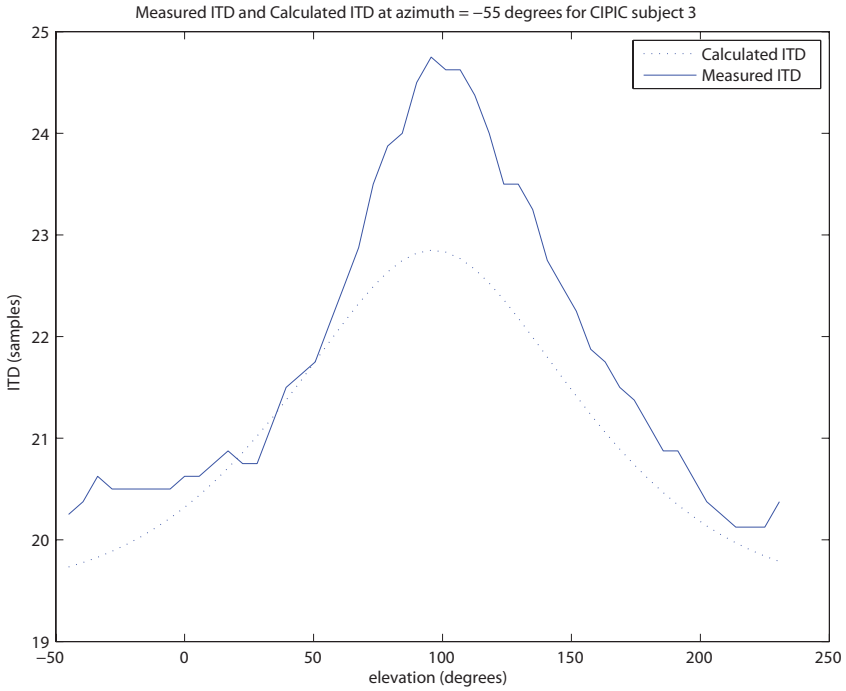

Figure 19. A plot of the ITD calculated using the algorithm described in this chapter (after [12]) and the actual measured ITD from the CIPIC database [6]. This plot is for a cone of confusion around an azimuth of -55° for CIPIC subject 3.

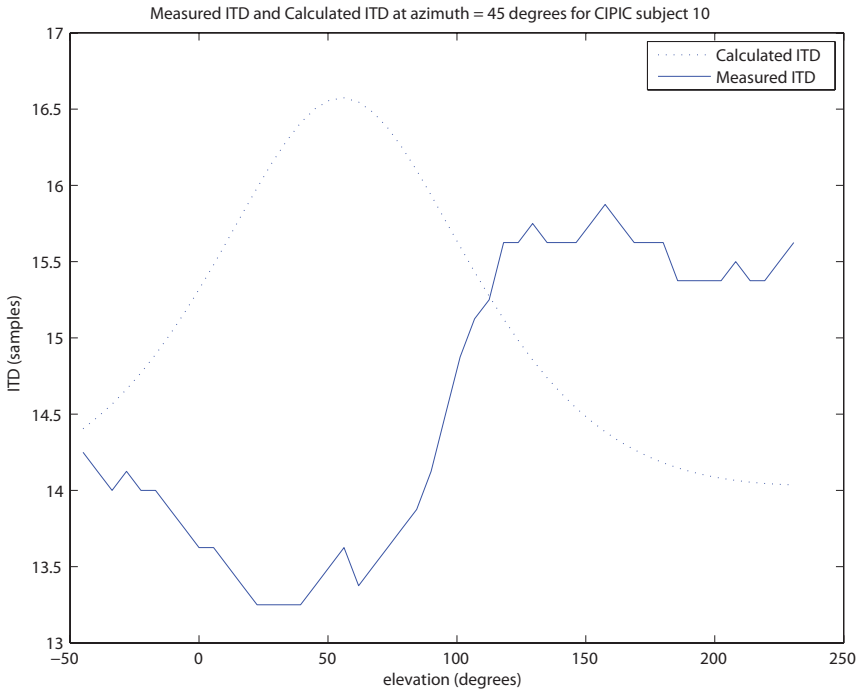

Figure 20. A plot of the ITD calculated using the algorithm described in this chapter (after [12]) and the actual measured ITD from the CIPIC database [6]. This plot is for a cone of confusion around an azimuth of 45° for CIPIC subject 10.

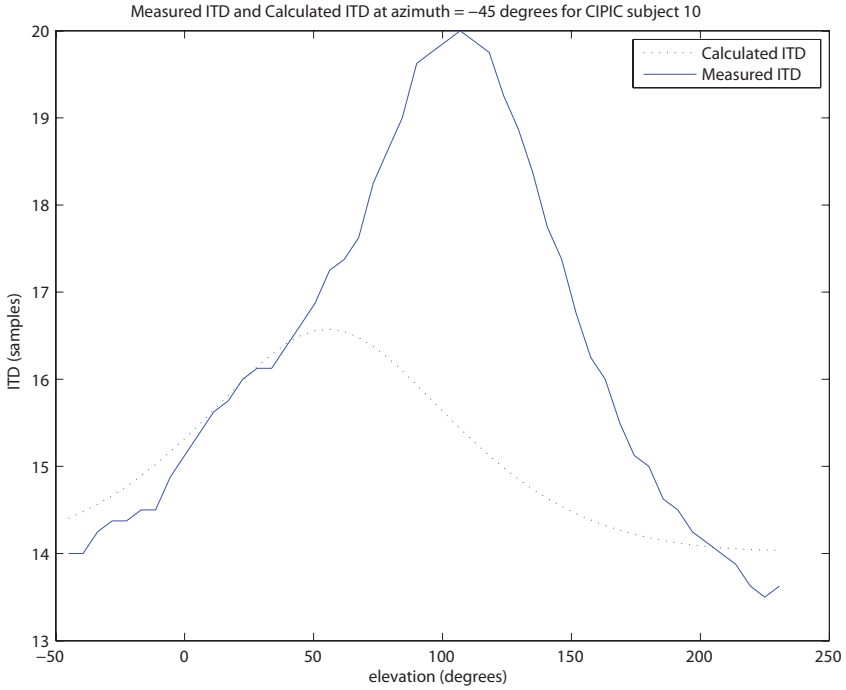

Figure 21. A plot of the ITD calculated using the algorithm described in this chapter (after [12]) and the actual measured ITD from the CIPIC database [6]. This plot is for a cone of confusion around an azimuth of -45° for CIPIC subject 10. This is the same plot as Figure 20 but for the left ear instead of the right ear.

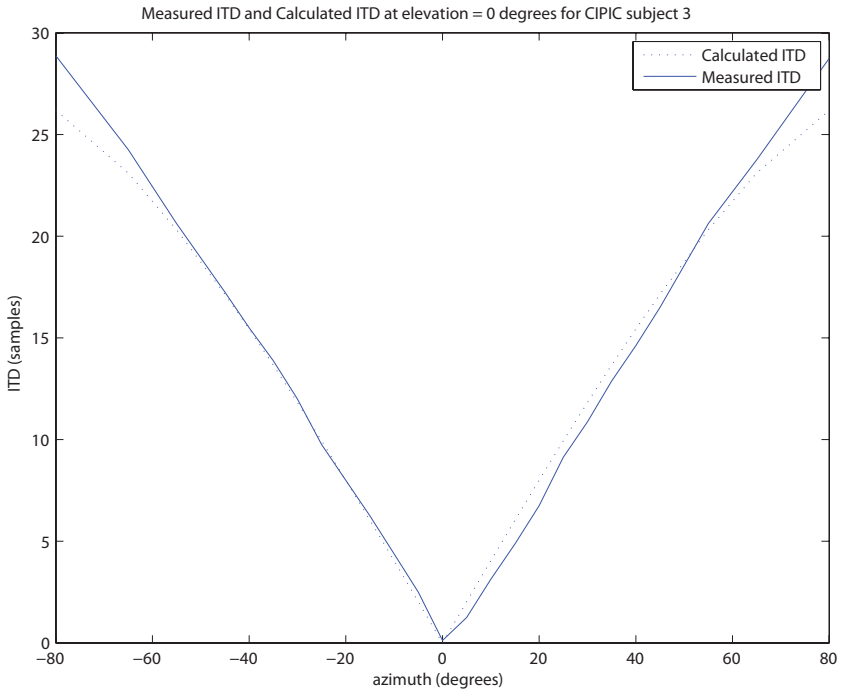

Figure 22. A plot of the ITD calculated using the algorithm described in this chapter (after [12]) and the actual measured ITD from the CIPIC database [6]. This plot is for a varying azimuth at an elevation of 0° for CIPIC subject 3.

The combination of the shadowing filter and time delay accurately approximate the effects that the head has on incident sounds before they arrive at the eardrum. In most cases the ITD error that results from using the model with symmetric ear offsets is at or below the difference limen for localization in the horizontal plane. In the cases when it is not, accounting for non-symmetric ear offsets will rectify the problem. In the experimental data acquired for the listening test that is explained in the Chapter Five, ear offsets for each side of the head will be measured, thus improving upon the errors observed in this section.

3.2.3 Spherical Torso Model

The head model presented in the previous subsection only accounts for half of the HAT model. In order for low-frequency elevation and subtle externalization effects to be added to the model, the torso must be accounted for properly. In the HAT model, the head

model previously described rests atop the spherical torso with an invisible cylindrical neck "coupling" the two spheres. The torso effects must take into account the contributions of all three body parts that are involved: the head, the neck and the torso. While the neck does not play a specific role in reflecting sound, the distance that it separates the head from the shoulders will be seen to be very crucial. The same spherical model that was described in Section 3.1.1 can be used to model half of the torso's role. For reflection, the other half of the torso's contribution, a different modeling algorithm is needed.

Once again, the wave equation was solved numerically for the boundary conditions presented by the complete HAT model in order to gain an understanding of how the additions of a neck and torso affect the acoustic properties of an incident sound. The computational results were analyzed in [5], and the two major behavior modes of the torso were identified as shadowing and reflection. Whether or not an incident sound is located within the set of rays that extend from the ear and are tangent to the torso determines the behavior of the torso. This imaginary group of rays creates what is referred to as the torso shadow cone. Figure 23 shows a three-dimensional view of a torso shadow cone (as it exists on a snowman) with respect to the right ear. There is also a torso shadow cone from the left ear that has similar properties.

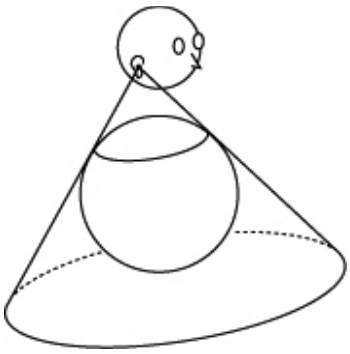

Figure 23. The torso shadow cone for the HAT model drawn with respect to the right ear [5].

If a sound originates inside of this cone, then it will be shadowed by the torso in the same manner that the head shadows a sound. For incident sounds that originate outside of the torso shadow cone, which is the case for the majority of spatial locations, the torso's contribution is as a reflector. In such cases, the incident sound arrives directly at the ear and is also reflected off of the torso and received by the ear at a later time. The act of reflection attenuates the direct sound and delays its arrival by the amount of time that the reflection takes. The reflected version of the sound is then summed with the direct sound as it arrives at the tympanic membrane. In the signal-processing world, this delay-and-add operation is the fundamental behavior of a classic FIR (or inverse) comb filter. The small and periodic notches in the frequency response that result from this reflection process are believed to be used by the human auditory system as spectral cues for determining the elevations of low-frequency sounds [1]. Figure 24 illustrates a two-dimensional representation of when these two cases, along with head shadowing, come into and out of play for all possible locations of an incident sound in the very revealing vertical plane.

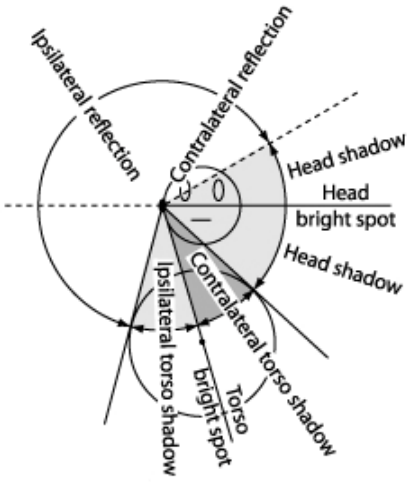

Figure 24. The vertical plane with respect to the subject's right ear and the conditions that result from specific angles in said plane [5].

Determining if an incident sound is in the torso shadow cone or the torso reflection zone is the fundamental conditional of the HAT algorithm. In order to decide which case is satisfied, it is necessary to calculate two vectors: the vector from the ear to the center of the torso ($\vec{d}$) and the unit vector representing the direction of the incident sound ($\vec{s}$). The angle between these two vectors (ζ) will determine if the torso shadows or reflects an incoming sound. Figure 25 shows a two-dimensional representation of the locations of these two vectors with respect to the ear and torso. The case shown in Figure 25 is one in which the source vector is tangent to the torso; therefore, it represents the transitional point from torso shadowing to torso reflection, or vice versa. The same transitional point exists on the contralateral side of the torso thus forming the boundary conditions of torso shadow cone with respect to the given ear. In Figure 24, a two-dimensional view of the entire torso shadow cone for the right ear can be seen as the combination of the areas labeled *ipsilateral torso shadow* and *contralateral torso shadow*. Likewise, the boundaries of the torso shadow cone with respect to the left ear are calculated in the same fashion.

The angle labeled $\zeta_{\min}$ in Figure 25 exists on each side of the torso even though it is only shown for the ipsilateral side. It is the threshold angle that creates the boundaries of the torso shadow cone. Using simple vector geometry and certain anatomical dimensions, $\zeta_{\min}$ can be computed for each side of the torso. The anthropometry dependent parameters needed to calculate $\zeta_{\min}$ are ear offset back, ear offset down, neck height and torso radius. A source that results in an angle of incidence in between the two calculated $\zeta_{\min}$ values is determined to be in the torso shadow cone. Sounds from any other angles of incidence are in the torso reflection zone.

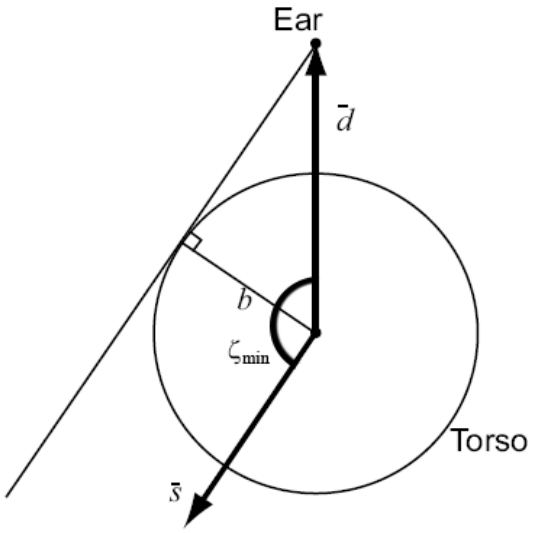

Figure 25. The transitional point between torso shadowing and torso reflection. $\vec{d}$ represents the vector from the ear to the center of the torso and $\vec{s}$ represents the vector from the center of the torso to the source. The angle ζ_{min} between these two vectors is the transitional angle between the torso shadow cone and the torso reflection zone. Image is a modified version of one found in [5].

The first mode of behavior to be analyzed is that of torso shadowing. When an incident sound is in the torso shadow cone, waves must travel around the torso before reaching the ear. This rather complex phenomenon results in waves traveling many different paths of varying distances before reaching the ear at an assortment of angles. In order to simplify this multifaceted problem and obtain only the perceptually critical results, the behavior is approximated by assuming that the spherical torso shadows the sound before it arrives at the head at an effective angle of θ_H. Figure 26 illustrates this behavior. Such an approximation results in a torso shadowing filter based on the spherical filter model from Section 3.1.1 (without the time delay component) being cascaded with the head model designed in Section 3.1.2. The block diagram in Figure 27 outlines the entire filter model for the torso shadowing case.

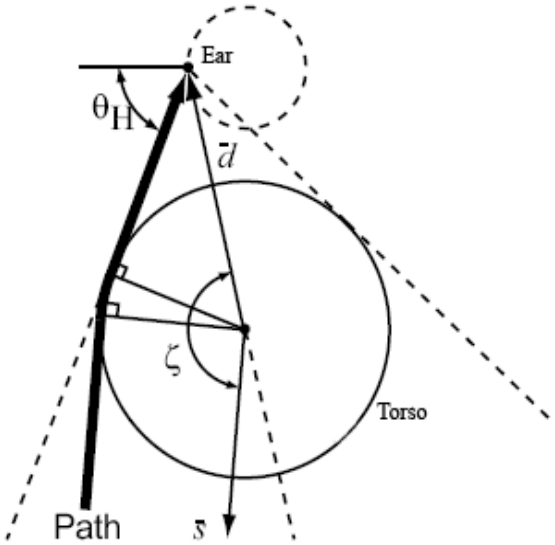

Figure 26. The behavior when an auditory source is located inside of the torso shadow cone, after [5].

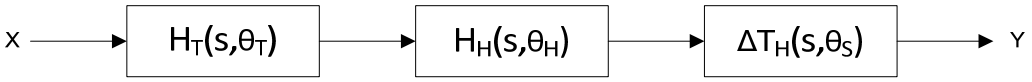

Figure 27. Block diagram of the torso shadow sub-model, after [5].

The input parameters passed into the spherical model function that are necessary to determine the torso shadowing filter $H_T(s,\theta_T)$ are the radius of the torso and the observation angle θ_T. Intuitively, one would be inclined to use the angle ζ from Figure 26 as θ_T, but this leads to a discontinuity at the torso shadow boundary condition because the response will be heavily shadowed just inside of the torso shadow cone and nearly flat just outside of the torso shadow cone. To rectify this discontinuity, the θ_T used is actually one that is interpolated between π and θ_{flat} (from Section 3.1.1), as described in [5].

For the head-shadowing filter $H_H(s,\theta_H)$, it is assumed that the waves shadowed by the torso leave at the point of tangency to the torso and travel directly to the ear. This can be seen as the bold line labeled *path* in Figure 26. A discontinuity in θ_H is encountered

when the source crosses from the ipsilateral torso shadow zone to the contralateral torso shadow zone (see Figure 24); however, in this case, the results due to the discontinuity agree with the frequency response curves of the numerical data, so it does not need to be remedied. The angle θ_H that is inputted into the spherical model to determine the head-shadowing filter $H_H(s,\theta_H)$ is calculated using a ray tracing analysis that is explained thoroughly in the appendix of [5]. Further explanation of the θ_H calculation is omitted from this work because it is not exceptionally relevant to the objective.

If θ_H is used to determine the time delay portion of the head model $\Delta T_H(s,\theta_S)$, the acceptable discontinuity discussed in the previous paragraph leads to a drastic jump in perceived position which is not realistic. In [5], the authors indicate that good-sounding results can be obtained by taking θ_S as being equal to the angle between the source and the ear (the observation angle) as if the torso is not present to interfere with the wave. The same approach used in this implementation.

The behavior of empirical HRTFs at the lowest elevations that result in torso shadowing is somewhat ambiguous and relatively unknown due to the limitations addressed in Section 2.3; however, solving the boundary conditions numerically for a snowman model produces results consistent with the aforementioned torso shadowing algorithm. Informal listening tests by this paper's author also reaffirm the algorithm's performance.

The torso's other, more perceptually relevant, effect on incident waves is as a reflector. As was briefly explained earlier, sources that are located outside of the torso shadow cone are modeled as arriving at the ear in twice. The initial arrival is of the direct wave and the second arrival is of the wave after it has been reflected off of the torso. Mathematically this results in the temporal domain equation

$$y(t) = x(t) + \Gamma x(t - t_d), \tag{9}$$

where Γ is the reflection coefficient, t_d is the time delay incurred from reflection, *x(t)* is the original signal and *y(t)* is the signal resulting from the sum of the direct and reflected signals.

In the frequency domain, *y(t)* exhibits periodic peaks and notches that are characteristic of classic inverse comb filter behavior. The reflection coefficient controls the depth of said notches by determining how much of the reflected signal is added back to the original. In practice, reflection is a highly complex process that depends on a variety of factors; however, the aforementioned comb filter attempts to capture the perceptually relevant behavior by modeling the first-order reflections that are known to occur. A block diagram of the reflection portion of the HAT implementation is shown in Figure 28.

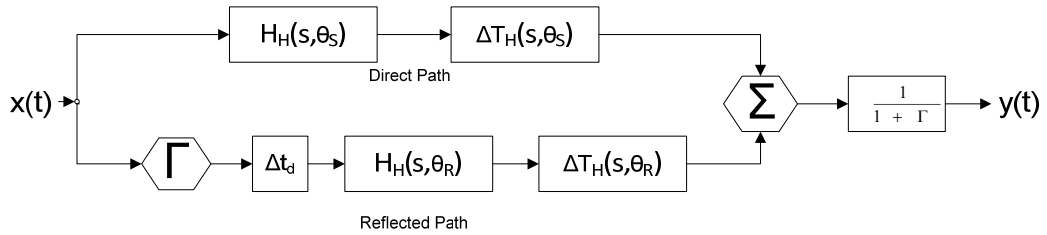

Figure 28. Block diagram of the torso reflection case, after [5].

The upper half of the block diagram in Figure 28 represents the direct path, one in which the incident sound is only affected by the head before it reaches the ear. The angle of incidence (θ_S) that is used in the calculation of the head-shadowing filter and time delay component of the direct path is the angle between the ear and the source (the observation angle).

The bottom path of Figure 28 illustrates the reflection case. As a sound is reflected off of a medium, it is scattered in all directions, and eventually, an attenuated version of the

original sound will arrive back at the eardrum. This attenuation is accounted for by the reflection coefficient Γ.

The arrival time of the reflected sound at the tympanic membrane is delayed from the direct sound due to the extra distance that the reflected waves must travel before reaching the eardrum; this time delay is seen as Δt_D in the block diagram in Figure 28. Also, since the act of torso reflection alters the path of the incident ray, the reflected sound will arrive at the ear with an angle of incidence that is different from the direct path's angle of incidence; this is denoted in Figure 28 as θ_R. Lastly, the $\frac{1}{1+\Gamma}$ term of Figure 28 is included as a normalization factor to prevent the summed signal from clipping.

The formulas for calculating the reflected sound's delay time (Δt_D) and the reflected sound's angle of incidence (θ_R) are derived using ray tracing analysis algorithms that are explained in great detail in the appendix of [5]. In short, the length of the reflected path is computed and then divided by the speed of sound to obtain the time delay Δt_D. Once the path of the reflected sound is known, the angle of incidence in which the reflected sound arrives at the ear (θ_R) can be calculated quite easily. Figure 29 illustrates the relevant geometry that is used when calculating these two parameters. It can be seen in Figure 29 that the anthropometric quantities required for these calculations are ear offset down, ear offset back, torso radius and neck height. In Figure 29 the angle labeled β corresponds to the angle of incidence θ_R, and the sum of the paths labeled $f \cos 2\psi$ and f equals the distance traveled by the reflected sound.

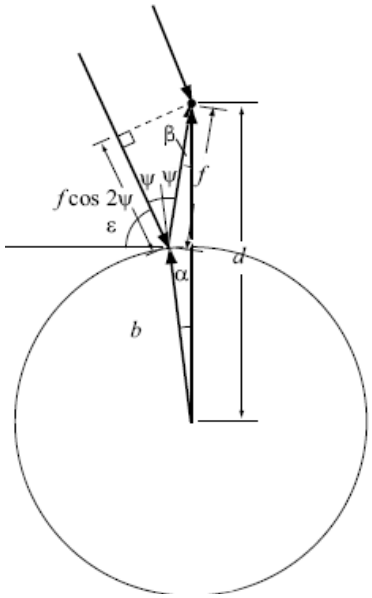

Figure 29. Ray tracing analysis used to calculate the reflected sound's time delay and the length of the reflected path, after [5].

As previously mentioned, the torso radius is not an especially important parameter in the HAT model because it does not solely determine anything except for the amount of torso shadowing; therefore, crudely approximating it as the geometric mean of the three torso dimensions does not have a significant perceptual effect on the results. However, the height of the neck does prove to be a very critical parameter because it accounts for the majority of the reflection distance in the torso reflection case.

When the results of this portion of the model are compared with numerical results, it was found that a reflection coefficient between 0.3 and 0.4 fit the modeled data to the numerical data very closely [5]. An image displaying the HRTFs around a 25° cone of confusion that result from inputting the anthropometry of a KEMAR into the HAT synthesis model is shown in Figure 30. This image can be compared to Figure 7a which shows the measured head and torso response of an actual KEMAR. Due to a higher sampling rate and a wider range of elevations, there is more data present in the image in

Figure 30 than there is in image in Figure 7a; however, a meaningful comparison between the two is still possible.

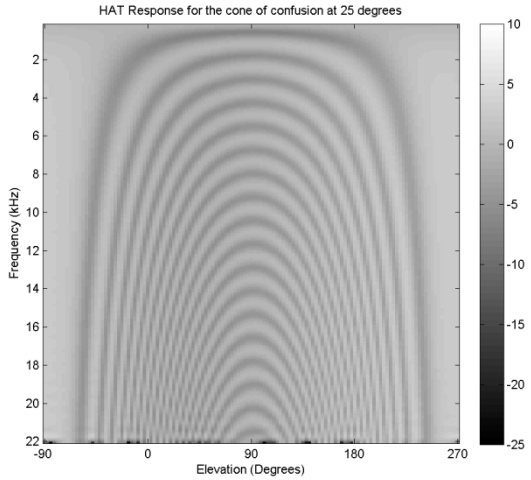

Figure 30. The frequency response of the HAT model for the anthropometry of a KEMAR at a cone of confusion of 25°.

One obvious difference between the plots in Figure 30 and Figure 7a is immediately evident: the spectral notches are present at all frequencies and elevations in Figure 30, but in Figure 7a there is a lack of well-defined notches at high frequencies for elevations between 0 and 180°. This is because reflection is a complex phenomenon and the actual value of Γ is not constant; it is dependent upon both frequency and orientation. In the physical world, at very low frequencies (when the wavelength is larger than the radius of the torso) the waves are not affected by the torso. At high frequencies the abnormalities in the shape of the human torso scatter high frequency waves thus reducing the amount of reflection that occurs at said frequencies. This scattering is very much frequency and orientation dependent. Perceptually, the notches in the response above about 5 kHz are meaningless which leads to two possible solutions to the problem of modeling the scattering phenomenon: the reflection coefficient can be made frequency dependent and allow only

frequencies below 5 kHz to be reflected or Γ can be made both frequency and orientation dependent to model the response that is visible in Figure 7a.

The former solution is much simpler to implement; it can be done by replacing the constant value of Γ in (9) with a simple low-order low-pass filter. Filtering the reflected wave in such a way allows only the low frequencies to be added to the direct wave, thus resulting in a frequency dependent reflection coefficient. One obstacle that must be dealt with is the group delay of the filter that is used as the reflection coefficient. The filter must have a constant group delay and the delay must be accounted for properly or else the amount of samples that the reflected signal is delayed will increase. Such an increase will result in the spectral notches being more frequent and at different locations which produces an incorrect model of the reflection process.

Since infinite impulse response (IIR) filters possess group delays that vary with frequency, they are not ideal to be used for the reflection coefficient in this case. The linear phase nature of FIR filters produces group delays equal to half of their orders for all frequencies. This makes them perfect candidates for the desired frequency dependent reflection coefficient. Since the group delay is always known for FIR filters, the direct signal can be delayed by the group delay before the reflected signal is added to it. This ensures that the two signals that are added together (the direct and reflected signals) are offset from each other only by the desired time delay (Δt_D), and not by of the sum of Δt_D and the reflection coefficient's group delay. The plot in Figure 31 shows the result of using a simple sixth-order Hamming window FIR filter with a non-orientation dependent cutoff frequency of 5 kHz as Γ.

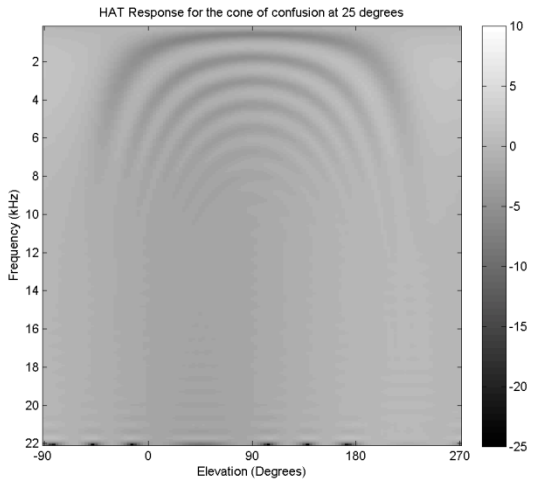

Figure 31. The frequency response of the HAT model for the anthropometry of a KEMAR with a frequency dependent reflection coefficient at a cone of confusion of 25°.

To model the more realistic case where the reflection coefficient is also orientation dependent, it is necessary to add one more step to the algorithm that created the image in Figure 31. This step involves parabolically varying the cutoff frequency of the reflection coefficient with orientation. The magnitude response resulting from the frequency and orientation dependent reflection coefficient is shown in Figure 32.

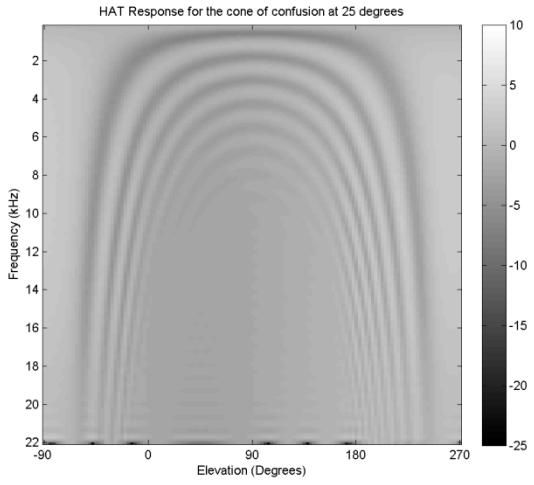

Figure 32. The frequency response of the HAT model for the anthropometry of a KEMAR using a frequency and orientation dependent reflection coefficient at a cone of confusion of 25°.

To test the perceptual performance of the HAT model, the author of this paper inputted his anthropometry into the algorithm and conducted an informal listening test. The experiment revealed that the HAT model is severely lacking in creating a sense of elevation in and around the median plane. This is backed up objectively by the fact that the frequency response of the HAT model does exhibit the notches and resonances that occur due to the pinnae which have been previously established as the primary cues used by the brain for resolving elevation. This lack of elevation perception at and near a cone of confusion of 0° is present because ITD and IID cues are nearly equal to zero, thus meaningless, and because torso reflections do not differ much between the left and right ears in the median plane, so the spectra of the signals that arrive at each ear are nearly identical. In other words, there are no definitive localization cues caused by the head or the torso in the median plane; therefore, at best, a weak sense of elevation is created.

For incident sounds originating around the vertical plane a slightly better sense of elevation and externalization is created for the contained locations that are far away from the median plane. This was expected because the notches caused by reflection are different at each ear and this binaural spectral difference aids in resolving elevation. It was also observed that the horizontal plane localization performance of the HAT model is very accurate; this was also expected.

Listening to sounds filtered with the HAT model is essentially what the world would sound like if humans didn't have external ears. Since humans do possess external ears, this model is insufficient for spatial sound synthesis if a true sense of elevation is desired. In the next section, the pinna model that is added to the HAT implementation is introduced. The primary intention of the pinna model is to improve elevation effects at and around the frontal median plane.

3.3 PINNA MODEL

D.W. Batteau was one of the first researchers to delve into the details of the pinna's role in the localization process [10]. He hypothesized that the abnormal convolutions of the outer ear cause multiple reflections of incident sound waves and that the difference in arrival times (at the tympanic membrane) of the direct and reflected waves varies with the elevation of a source, thus providing a vertical localization cue. Such delays produce spectral notches that are similar to those created from torso reflections but with much shorter delay times.

There were two main criticisms of Batteau's delay-and-add hypothesis, the first of which is that the human auditory system may not be able to perceptually resolve the short delays that result from the reflection of sounds off of the small and close quartered anatomical parts of the pinna. This skepticism can be easily understood because the delays that are caused by pinna reflections are less than 100μs which is a very short amount of time. The other criticism of his theory is that it oversimplifies a very complex phenomenon.

The former objection to Batteau's theory was investigated by Wright, Hebrank and Wilson [32]. They performed an experiment that delayed white noise by amounts of time varying between 10μs and 300μs and then added it back to the original signal to emulate the reflection process that was thought to occur in the pinna. These synthesized sounds were then presented to human subjects over headphones in an attempt to figure out the shortest amount of delay that is perceivable. Their experimental results indicate that humans can easily interpret delays as short as 20μs (approximately one sample at 44.1 kHz) provided that the amplitude of the delayed signal is at least 67% of the original signal's amplitude. The amplitude ratio of the delayed signal to the original signal is analogous to a comb filter's reflection coefficient, and it can be modeled as such.

In the same experiment, the just noticeable difference for said delays was also investigated. It was found that a change of 7μs is readily detectible as long as the reflection coefficient is equal to or greater than 0.70. For an amplitude ratio of 0.40 the difference limen was found to be 12μs. These delay times correspond well to the just noticeable difference values for elevation localization that were introduced in Section 1.4.

In a later experiment Hebrank and Wright loosely linked dimensions of the concha (refer to Figures 5 and 6 for the anatomy of the pinna) at various angles of incidence to the locations of the first deep spectral notches in HRTFs [20]. They also identified spectral trends in measured HRTFs as elevation increased that agreed with their concha measurements. In short, as elevation increases, the distance that the reflected wave travels to the wall of the concha decreases (due to the shape of the concha), and this results in an increase in the frequency location of the first deep spectral notch. Figure 33 illustrates the change in reflection distance with an increase in elevation on an image of a human pinna.

Figure 33. Relationship of the concha shape to elevation angle, taken from subject 1 of this work's subjective testing.

The large black dots on the image represent the reflection distance traveled by a sound originating at location (0,0). The small white dots illustrate the shorter distance traveled by a sound emanating from spatial location (0,Φ). For elevations below (0,0), it can be seen in Figure 33 that there may be an additional reflection off of the crus helias. This will be addressed and discussed in more detail in Section 3.3.4.

The center of the coordinate system superimposed onto the image of the pinna in Figure 33 is located at the approximate origin of the interaural-polar coordinate system. In theory, this is the location of the microphone when recording HRTFs with the auditory meatus blocked. This location is different for every subject and is especially difficult to locate if the entrance to the ear canal is not visible in the image. The angle Φ above the abscissa in Figure 33 is equal to the angle that an elevated sound first arrives at the ear canal. In the cases investigated in this work, which are limited to locations near the median plane, the arrival angle is equivalent to the interaural-polar elevation. This is true because the incident sound creeps around the head and continues to travel in the incident direction until it reaches the eardrum. At azimuths greater than approximately 30° (the actual angle varies from person to person and is dependent upon head size) this assumption is no longer valid for the ipsilateral ear because there is much less head shadowing (if any) at such angles of incidence. Also, the contralateral response becomes much more complex at azimuths far away from the median plane. These reasons explain why the model introduced herein is most effective for azimuths close to the median plane where there is a considerable amount of head shadowing before the incident sound ultimately arrives at each ear.

The labeled points on Figure 33 indicate that the initial arrival point at the ear canal varies with elevation. For certain high and low elevations the sound will arrive at the ear canal initially through the upper and lower tragal notches located above and below the

tragus, respectively. Due to the great variation in pinna shapes and dimensions across the population these stitches are more prominent on some ears than they are on others.

The beginning of the distance measurement for location (0,0) is seen to be a point on the tragus rather than directly over the entrance to the ear canal because, since the tragus protrudes over the ear canal entrance, there is an additional distance that the reflected sound must travel underneath the tragus before eventually arriving at the ear canal entrance. This is accounted for by starting the measurement for (0,0), and other neighboring locations, slightly before the edge of the tragus. A similar method is used in [29]. These observations are consistent with the conclusions drawn by Hebrank and Wright in [20].

While studying the spectral trends of their HRTF measurements Hebrank and Wright also noticed that the gains of certain resonant peaks vary with elevation. In addition, as elevation changed they observed that some spectral peaks also shifted in center frequency in a manner similar to how the center frequencies of the pinna spectral notches vary with elevation. It is believed that the mere presence or absence of peaks at specific frequencies is a perceptual cue for resolving elevation. Other researchers have also verified the importance of spectral peaks in elevation localization [11, 24, 30]. This leads one to believe that elevation cues are related to spectral notches *and* spectral peaks.

The resonant characteristics of the human ear that cause the spectral peaks observed by Hebrank and Wright have been thoroughly investigated by E.A. Shaw [30]. He concluded that while each human ear has distinctive patterns of response there are some characteristics that are common to all ears. These common properties are linked to the normal modes of the concha. Shaw measured these normal modes under both the free field and blocked-meatus conditions. His work with the blocked-meatus condition is more

pertinent to the objective of this work because it is the most common method used to measure HRTFs.

Once he identified the normal modes of the human ear Shaw then built a physical model of the outer ear that exhibited all of said modes. A summary of his findings for the blocked-meatus condition is shown in Figure 34. The measured data presented in this figure was later verified numerically in [22]. Note that the origin of Shaw's coordinate system is in close agreement to the one used in this work that was explained in a previous paragraph and shown in Figure 33.

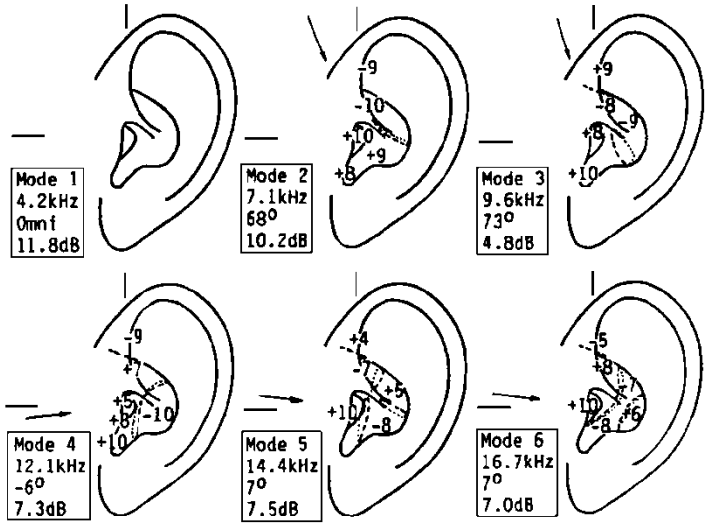

Figure 34. The normal modes of the human ear as identified by Shaw [30]. The resonant frequency, response level and angle of maximum excitation are indicated for each mode.

As Figure 34 shows, there are six major modes of the human pinna. The first of these modes is excited from all directions and is the most pronounced of the six resonances. The next two modes can be viewed as a vertical pair of transverse resonances that are best excited at high elevations. This is in agreement with the findings of Hebrank and Wright, among others, where it was established that sounds with large amounts of spectral energy in the region around 8 kHz are perceived as being elevated. The last three modes are often

referred to as a horizontal triplet because they are most excited at elevations near the horizontal plane. Perceptually, the third mode of this triplet (mode six) is irrelevant because the spectral cues interpreted by the brain during the localization process are all below 16 kHz; however, the presence of a resonant peak at 16 kHz does affect the envelope of the frequencies in the perceptually significant range. Because of this fact, it is necessary to model the sixth pinna mode.

Recently, additional credibility has been given to both Batteau's delay-and-add theory and Hebrank and Wright's concha reflection hypothesis. Raykar et al. [28] devised a signal processing method to extract the notch locations from the HRTFs of subjects in the CIPIC database. The corresponding delay times can be calculated from the extracted notches which, in turn, can be used to compute the reflection distances associated with the identified pinna notches. In [28], this was done for a series of uniformly spaced elevations on the median plane and the resulting distances were plotted on an image of the subject's pinna. The plotted points traced the outline of the subject's concha wall. One such example is shown in Figure 35. This provides further evidence supporting the two aforementioned theories.

It was believed by Raykar et al. that the presence of each notch was independent of the others. This is evidenced by the three distinct groupings of dots on the pinna image in Figure 35. Based on the known spectral modifications that occur when a signal is delayed and added back to itself, Raykar et al.'s assumption is in fact an incorrect one because a single reflection can account for multiple periodic notches. This phenomenon also exists with torso reflection and was demonstrated in Section 3.2.3. It is also worth noting that in Figure 35 the crus helias accounts for one of the reflections at low elevations which is a theory that was introduced earlier.

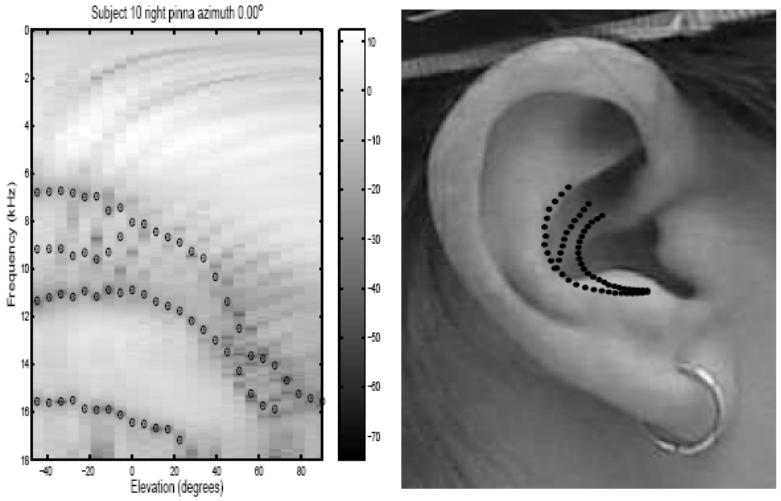

Figure 35. The median plane HRTF for CIPIC Subject 10 (left) and the corresponding distances on the subject's pinna (right). Image taken from [28].

Soon after the work of Raykar et al. Satarzadeh was able to accurately model the PRTF from anthropometry at the location (0,0) for a majority of the subjects in the CIPIC database [29]. He ran a perturbation analysis on the physical model outlined in Shaw's work in order to discover which anthropometric parameters affected the PRTF the most. His pinna model incorporated the work of Shaw, the delay-and-add theory of Batteau, and a fair amount of novel work.

The pinna model presented in this work combines methods by all of the aforementioned authors along with some original work in order to add more realistic elevation cues to the previously discussed HAT model at locations at and around the frontal median plane.

3.3.1 Extracting PRTFs

The PRTFs shown throughout the remainder of this section are obtained from the HRIR by removing the onset time, windowing to eliminate the effects of the torso and then transforming into the frequency domain. The window used is a .9ms half Hann window. A

length of 0.9ms is used because it allows for the effects of the pinna to be isolated. If the window is any longer, torso effects will make their way into the response.

A major side-effect of time domain windowing is that it smoothes the details of the frequency spectrum; however, this is acceptable because it has been proven that a majority of the spectral detail in HRTFs is perceptually irrelevant for the frequencies affected by the pinna. Another side effect that sometimes occurs is an unpredictable shift in zeros and poles. This does not seem to be of concern in this case because the PRTFs obtained from windowing possess characteristics that closely agree with the work of Shaw and other researchers. Other, more sophisticated, methods for extracting the PRTF have been developed, but the simple windowing method described herein works well enough for analysis purposes.

3.3.2 RESONANCES AT (0,0)

In this section, the modeling of the PRTF's resonances at the spatial location of (0,0) will be discussed. Section 3.2.4 will extend the model to be applicable for a range of frontal elevations in and around the median plane.

The resonances of the external ear that are outlined by Shaw can be modeled using three separate band-pass filters. Conventional band-pass filters are determined by three design parameters: gain, quality factor (3dB bandwidth) and center frequency. A simple second-order IIR equalization filter [26] with the standard bi-quad filter formula of

$$H(z) = \frac{\left(\frac{G_0 + G\beta}{1+\beta}\right) - \left(\frac{2(G_0\cos(\omega_0))}{1+\beta}\right)z^{-1} + \left(\frac{G_0 - G\beta}{1+\beta}\right)z^{-2}}{1 - 2\left(\frac{\cos(\omega_0)}{1+\beta}\right)z^{-1} + \left(\frac{1-\beta}{1+\beta}\right)z^{-2}} \qquad (10)$$

is used to model each the three most perceptually important pinna resonances. The parameter ω_0 is the filter's center frequency in radians/second, G is the filter's gain and G_0 is its DC offset. The remaining variable in (10) can be calculated using

$$\beta = \left(\sqrt{\frac{G_B^{\,2} - G_0^{\,2}}{G^2 - G_B^{\,2}}} \right) \tan\left(\frac{\Delta\omega}{2} \right), \tag{11}$$

where $\Delta\omega$ is the bandwidth of the filter in radians/second and G_B is equal to 3dB less than the gain of the filter (G).

Shaw's work provides average values for each of the necessary parameters that offer good starting points for the design of each filter, but some of these values can be calculated from anthropometry which will produce more personalized results. Shaw explains the first mode of the pinna to be a standard quarter wavelength depth resonance. If the shape of the concha is approximated as a cylinder that is open at one end (the other end is closed due to the blocking of the meatus that occurs during the recording of HRTFs) its resonant frequency can be calculated quite easily from anthropometry. The radius of the cylinder is analogous to the depth of the concha and the length of the cylinder is the width of the concha. Satarzadeh was the first to establish the relationship between the depth resonance of the concha and the resonant frequency of a cylinder [29]. He calculated the center frequency of the depth resonance using

$$f_{depth} = \frac{c}{4(d + .41w)}, \tag{12}$$

where d is the depth of the concha, w is the width of the concha and c is the speed of sound.

The shape of the concha does vary across the population, and it is very rarely an ideal cylinder; however, modeling it as such provides a good approximation to its resonant frequency. In most cases, the center frequency of the depth resonance is approximated

within about 350 Hz of its actual location. If the depth of the concha is unknown and/or unavailable, then the simple formula of $f_{depth} = c/4L$ can be used, where the sole parameter of L is the width of the concha. This method is not desired though because, since it only depends on one parameter, it is less accurate.

In this work, since the CIPIC database's anthropometry measurements include each ear's concha depth for a number of subjects, the concha depth parameter is included in the depth resonance calculation for objective results. For the subjective results reported in Chapter Six, concha depth is not one of the measured parameters. Instead, (12) is used with a constant concha depth of 10mm, a value that corresponds to the average concha depth of the subjects in the CIPIC database. Concha depth is omitted from the anthropometry acquisition for the subjective results because a precise caliper is needed to accurately measure it and an instrument suitable enough for obtaining the parameter could not be found.

Although the gain and quality factor parameters are yet to be linked to anthropometry, both of these values were found to be very consistent across the PRTFs of many subjects. Figure 36 plots the left ear PRTFs of 10 different subjects in the CIPIC database at (0,0), and it provides convincing evidence for the consistency of the gain and quality factor parameters across multiple PRTFs. The frequency axis of the plot only extends up to 6.5 kHz so that the depth resonance is isolated for easier viewing.

Additionally, the plot in Figure 36 loosely confirms Shaw's measurements of the external ear's depth resonance. His values were also obtained from averaging data from ten different subjects. Unfortunately, no statistically significant correlation between concha size (or overall pinna size) and the gain of the depth resonance was found when analyzing the anthropometry of 15 subjects in the CIPIC database. Because of this, a decision was made

to use an average value of 9dB for the gain of the depth resonance in the pinna model at (0,0).

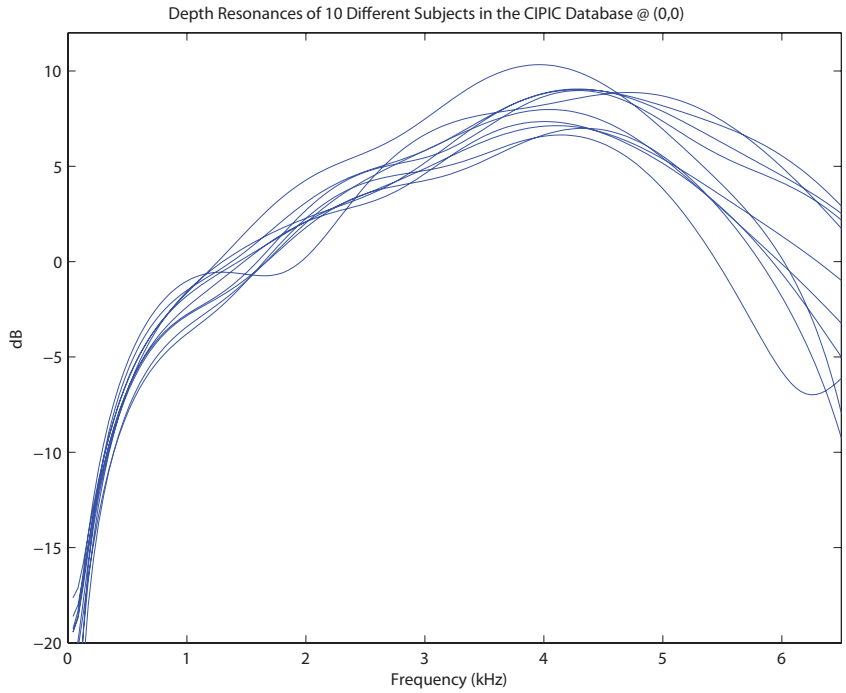

Figure 36. The PRTFs of 10 different subjects in the CIPIC database [6] at (0,0); the plot is limited to 6.5 kHz so that the depth resonance is isolated.

In a similar manner, the quality factor of the band-pass filter used to model the depth resonance was chosen to be 2.5 kHz. This value was discovered to match up well with the bandwidth of the first resonance of many PRTFs. Additionally, the parameter G_0 was set to zero. This creates infinite nulls in the frequency response (also known as zeros) at the DC and Nyquist values of (10) which are necessary to accurately model the behavior of the PRTF at the extremes of the frequency response.

Figure 37 shows the PRTF of the left ear of CIPIC subject 20 plotted alongside the band-pass filter used to model the depth resonance. Subject 20 has a left concha depth of 12mm and a left concha width of 20mm. It can be observed that the band-pass filter used to model the depth resonance accurately approximates the actual response of the depth

resonance in the PRTF for this particular subject. Examining the performance of this method with other subjects in the CIPIC database reveals similar results.

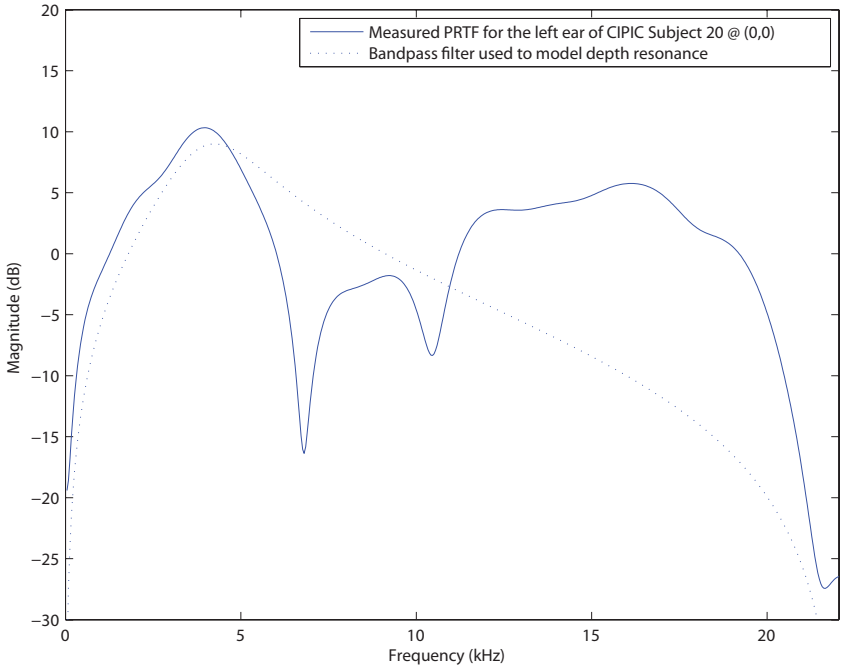

Figure 37. The PRTF for the left ear of CIPIC subject 20 at (0,0) plotted along with the band-pass filter used to model the depth resonance.

The other two modes that are necessary to model at (0,0) are modes four and six. These resonances are evident, respectively, at roughly 12 kHz and 16 kHz in the PRTF of Subject 20 that is shown in Figure 37. Many other subjects exhibit these resonances at very similar locations. These two width resonances are also modeled using the bi-quad filter introduced in (10). Shaw indicates that, at an elevation of -6°, the average gain value of the first width resonance is 7.3dB; however, an analysis of the PRTFs of a large number of subjects in the CIPIC database reveals this value to be 5.1dB at the spatial location of (0,0). In this implementation, the latter value is used for the gain of the band-pass filter that models the first width resonance. Heuristically, a bandwidth of 3 kHz was found to provide a good fit for the shape of the first width resonance of most PRTFs at (0,0). The location of

the center frequency of Mode 4 was found in [29] to line up consistently with a certain peak of the comb filter that is used to model the PRTF's notches. Further details about the calculation of this parameter are provided in Section 3.3.3.

The location of the second width resonance (Mode 6) is technically out of the range of perceptual significance which makes it is easy to ignore; however, this resonance should not be overlooked because modeling it will allow the frequency domain envelope below 16 kHz to be more accurate. Unfortunately, none of the parameters of this resonance have been linked to anthropometry. This is probably because of its seemingly perceptual insignificance.

With no prior art to build upon it was decided that the parameters for this resonance are to be independent of the subject. This proves to be an acceptable approach because the properties of this resonance do not vary much from subject to subject. An acceptable gain value was found to be 6dB, the bandwidth decided upon was 3 kHz, and the location of the center frequency was discovered to be 16 kHz. Figure 38 shows all three of the band-pass filters that are used to approximate the pinna's resonances plotted on the same set of axes as CIPIC subject 20's PRTF. From Figure 38, it can be seen that all three of the band-pass filters match up to the shapes and locations of the resonances for this particular subject's PRTF very well. Once again, the performance was found to be similar across many other PRTFs in the CIPIC database. The next chapter provides results of the entire model for other subjects in the CIPIC database and it can be deduced from most of those examples that modeling the pinna's resonances as band-pass filters is an acceptable solution to the problem.

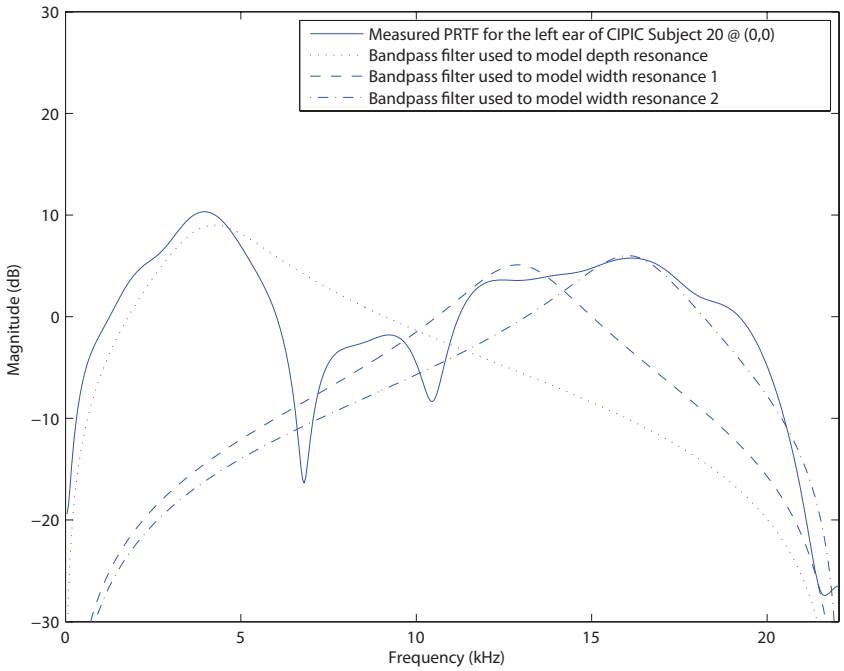

Figure 38. The PRTF for the left ear of CIPIC subject 20 at (0,0) plotted along with the band-pass filters used to model the three resonances.

The first inclination when deciding how to connect all three of these resonances together is to cascade the three band-pass filters in series; however, this does not produce results that match up well to the envelope of the PRTF and it also does not model the physics of the problem accurately [29]. Physically, all three of these resonances act in parallel; therefore, the filters that model them should be connected in such a way. A block diagram illustrating this connection is shown in the next section in Figure 41. The result of the parallel cascading of the three band-pass filters is shown in Figure 39. The parallel connection produces an envelope that closely approximates the spectral peaks of CIPIC subject 20's PRTF. Since the locations of the two width resonances are so close together, each of their gains is boosted due to the presence of the other when they are added in parallel. To counter the overall gain increase that occurs from the parallel connection, the gains of both width resonances are decreased by 1dB prior to the cascading process.

The resonances in most other subjects in the CIPIC database are approximated equally as close; this will be seen in the next chapter. From Figure 39 it is also evident why it is not necessary to model the pinna's fifth mode. This is because when the band-pass filters modeling Modes 4 and 6 are connected in parallel a gain is naturally present at the resonant frequency of Mode 5; therefore, it is not required to specifically account for Mode 5.

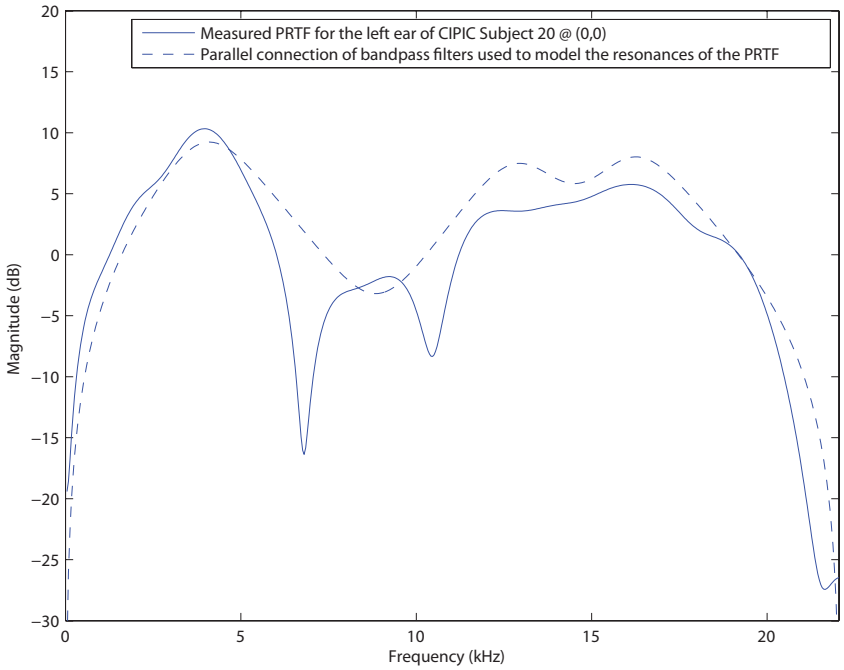

Figure 39. The PRTF for the left ear of CIPIC subject 20 at (0,0) plotted along with the parallel connection of the three band-pass filters shown in Figure 38.

3.3.3 NOTCHES AT (0,0)

To model the spectral notches that are caused by the pinna, a delay-and-add implementation based upon the work of Batteau and subsequent researchers is used. The parallel sum of the resonances modeled in the previous section (and shown in Figure 39) is delayed and then added back to a non-delayed version of itself in the time domain to model the pinna's first-order reflections. This solution makes physical sense and also proves to be analytically accurate as well. The formula used to model the inverse comb filter behavior of

the pinna's reflections is shown in (9). It is the same as the formula used to model torso reflections; however, there are two fundamental differences between the reflection properties of the torso and pinna: the reflection coefficient Γ and the time delay t_d.

In the torso model, the reflection coefficient has a low-pass nature which makes physical sense; in the pinna model, Γ has a band-pass response. Since the largest dimension of the pinna is its height, which is no larger than 60mm, the pinna will not affect frequencies with wavelengths larger than 60mm. Solving

$$f = \frac{c}{\lambda}, \tag{13}$$

which links wavelength (λ) to frequency (f) and the speed of sound (c), for a wavelength of 60mm results in a frequency of 5.7 kHz; this is the approximate minimum frequency that is reflected by the pinna. Because of this, the high pass cutoff of the reflection coefficient's band-pass filter is fixed at 5.7 kHz.

The low-pass portion of the reflection coefficient's band-pass response occurs because very high frequencies are scattered more sporadically than low frequencies; therefore, there is a lesser amount of high frequencies than low frequencies arriving back at the ear canal post-reflection. To model this, a greater attenuation of very high frequencies must be accounted for in the band-pass response of the reflection coefficient.

The upper cutoff frequency of the band-pass response is anthropometry dependent because it varies with the distance that the reflected wave travels. If the reflection distance is shorter, a greater amount of high frequencies will arrive back at the ear canal after reflection because there is less area present for scattering to occur. Also, since the reflection distance varies with elevation, this parameter is also orientation dependent. An examination of the notch depths at higher frequencies in HRTF data confirms this hypothesis. Exactly how the

response of the band-pass reflection coefficient varies with distance and frequency is addressed in the next section.

A plot of two comb filters is shown in Figure 40. The solid line plot was created with a constant reflection coefficient that is not dependent upon frequency; the dashed plot was created using a reflection coefficient with a band-pass response. The latter curve is characteristic of the reflection behavior that could occur in a human pinna. Accounting for the group delay of the reflection coefficient's filter that was explained in the torso reflection section is also necessary in the pinna case.

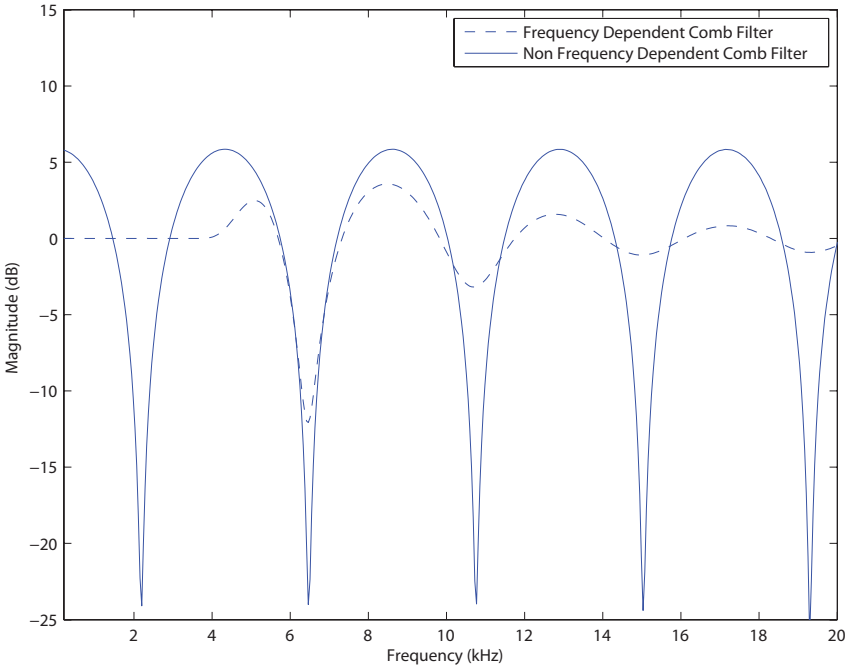

Figure 40. A non-frequency dependent comb filter and a comb filter with a frequency dependence characteristic of a human pinna.

The other parameter used in determining the reflection characteristics of the outer ear is the time delay. It has been hypothesized, and proven under certain cases, that the concha acts as the primary reflector; therefore, its dimensions determine the location of the notches in the spectrum [20, 28]. In [29], the nature of the pinna's primary reflector is more thoroughly investigated. It was found that the amount of pinna flare plays a role in

determining the primary reflector at (0,0). For subjects whose pinnae have considerable flare, the concha is the primary reflector. When the protrusion of the pinna is not substantial (and the pinna is very flat), the helix was found to be the primary reflector, especially at the location (0,0). The pinna protrusion angle is available as the parameter θ_2 in Figure 5, but there is not yet a mathematical way to determine if the pinna flare is substantial or not just from the protrusion measurement. In most cases, visually inspecting the subject's pinna flare from a digital image that was taken with the subject facing the camera has been demonstrated to work quite well in determining the primary reflector for elevations near 0°.

In Satarzadeh's investigation some anomalies were found where neither the helix nor the concha acted as the primary reflector at (0,0). In such cases, this model may fail. This was only the case for approximately 20% of the subjects in the entire CIPIC database; for the remaining subjects, it was found that the primary reflector was well defined as being either helix or the concha. Only the location of (0,0) was investigated in [29] but based on the anatomy of the pinna it is very likely that the primary reflector changes with elevation. This is especially true at high elevations where the concha is almost always the primary reflector because the helix merges with the earlobe at such angles; therefore, it cannot act as a reflector. Following the largest elevation angle shown in Figure 33 demonstrates that at high elevations (large Φ values) the concha is the only available reflector. The possibility of having either the helix or the concha act as the primary reflector at certain elevations will be taken into account during the subjective testing that is explained the Chapter Five.

Another contribution of [29] is the equation

$$t_d = \frac{4 d_r}{c} \tag{14}$$

that links the reflection distance (d_r) to the FIR comb filter's time delay (t_d). The term, c, is once again the speed of sound.

Using (14), it is now possible to explicate the calculation of the first width resonance's center frequency--an explanation that was left out of the previous section. If the concha is determined to be the primary reflector then the location of Mode 3's resonance corresponds very closely to the third peak of the FIR comb filter. A brief examination of the third peak of the comb filter in Figure 40 and of the location of the first width resonance of the PRTF in Figure 39 proves this to be the case. The third peak of the comb filter can be calculated using

$$f_c = \frac{3}{t_d}, \tag{15}$$

where f_c is the center frequency of the resonance and t_d is the time delay calculated in (14). If the helix acts as the primary reflector, or if the concha's width is uncharacteristically large, then the fourth peak of the comb filter corresponds to the first width resonance. Likewise, it can be calculated by replacing the 3 in the numerator of (15) with a 4. In the actual implementation, a conditional statement is used to ensure that the comb filter peak chosen to represent the location of the center frequency of the first width resonance is between 10.5 kHz and 13 kHz.

Equation (14) will result in very short time delays for reflection distances that are less than 15mm which is very common at high elevations. Also, as the reflection distance decreases with increasing elevation, a difference in the reflection distance of less than a millimeter will result in a substantial change in the notch location, which is perceptually critical in elevation perception. For example, a possible reflection distance for an incident sound at 45° is 14mm while the reflection distance at 60° may be 13.4mm. In an analog

system, the difference in the locations of the first deep spectral notches resulting from these two delays is 411 Hz; however, in a digital system, if the time delay corresponding to such a small change is simply multiplied by the sampling rate and rounded to the nearest sample, then the precision of a tenth of a millimeter is lost and both values will correspond to the same time delay. In this example, at a sampling rate of 44.1 kHz, the time delay corresponding to a reflection distance of 14mm is 7.2 samples and for 13.4mm it is 6.89 samples. Both of these values will round to 7 in a system with a resolution of integer samples, thus placing the notch locations for elevations of 45° and 60° at the same frequency which is unacceptable for obvious reasons.

To counter this, the impulse response of the parallel sum of the resonances is up-sampled by a factor of 10, delayed at the higher resolution, down-sampled back to the original rate and then added back to a non-delayed version of itself. This allows for the delay of the comb filter to have a resolution of a tenth of a sample which proves to be necessary when the pinna reflection distance is short. In order to fully reap the benefits of the higher resolution delay, the concha measurements must be very precise. The method in which the measurements are acquired is explained in the Chapter Five.

The connections of all of the modules that comprise the pinna model at (0,0) are shown in Figure 41. Results of this PRTF model at (0,0) for two subjects in the CIPIC database are shown in Figures 42 and 43. Figure 42 is an example of when the model fits extremely well, and Figure 43 demonstrates an example where the model matches relatively well but not as precisely as the prior example. The notches and resonances are in the correct locations but the gains are consistently too large for each of the three resonances. This is because the gains of the resonances are not personalized from anthropometry. Calculating the gains from measurements of the pinna may rectify this problem, but such work has not

yet been done. The location of the first deep spectral notch and gains of the width resonances relative to the depth resonance are arguably the most perceptually important characteristics of the PRTF in the localization process so, with that in mind, the aforementioned gain mismatch may not be very critical. Additionally, since the gains are offset by a constant amount across all of the resonances, the problem can be easily rectified by attenuating the entire spectrum. The subjective results presented in Chapter Six will shed some more light on the seriousness of this gain mismatch problem.

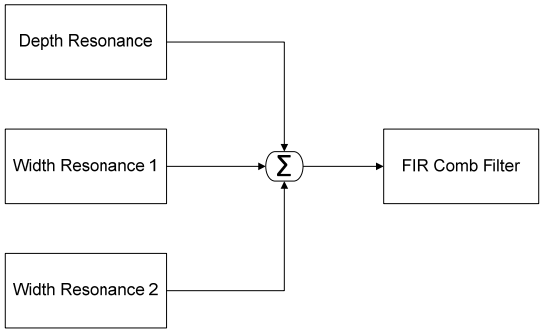

Figure 41. Pinna filter block diagram.

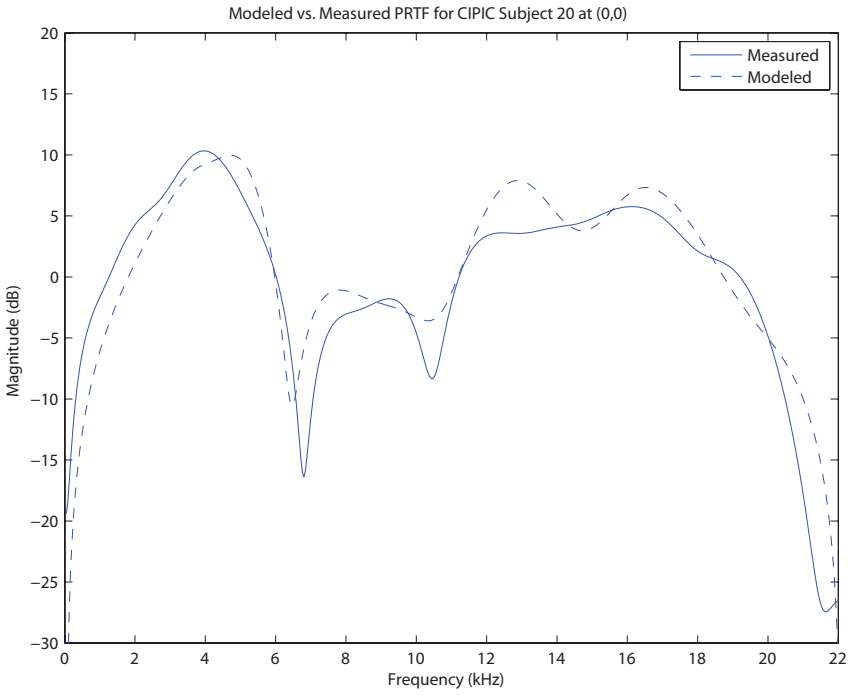

Figure 42. The measured and modeled PRTFs for CIPIC subject 20 at (0,0).

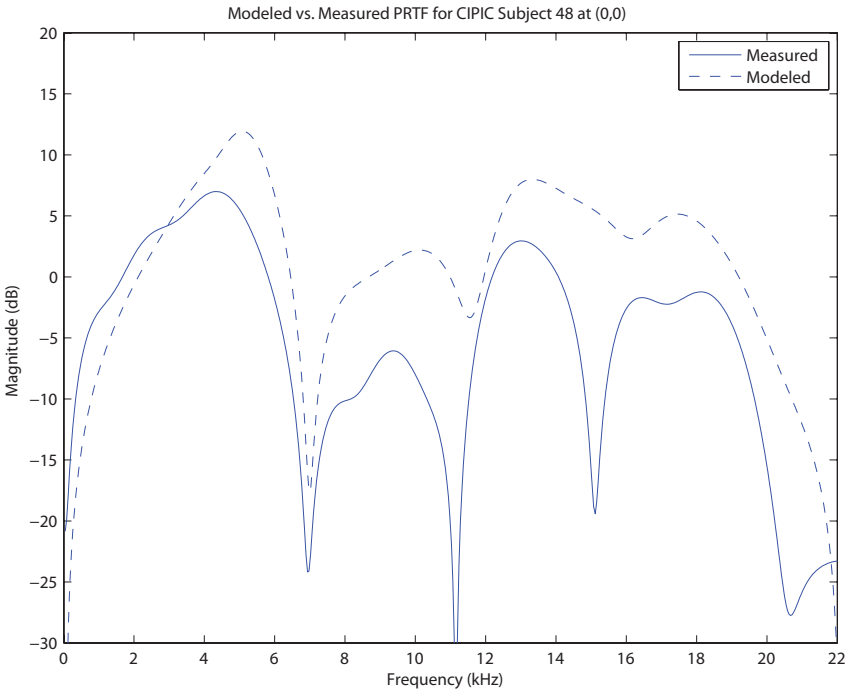

Figure 43. The measured and modeled PRTFs for CIPIC subject 48 at (0,0).

3.3.4 Elevation Dependence

The prior two sections explained how the acoustic response of the pinna can be modeled for a sound located at the point directly in front of the listener. For most subjects, sounds synthesized with the output of the HAT model at (0,0) are perceived as being slightly elevated. Adding a pinna model to the HAT model will 'lower' such sounds, thus making them more accurate; however, for the most part, the HRTF of a straight ahead location is relatively uninteresting.

Many of the parameters discussed in the prior two sections are dependent upon elevation. The gain of the depth resonance varies by about 4dB for elevations below 0°, the gains of the width resonances are largely dependent on elevation, the comb filter's time delay changes with elevation and so does the response of its reflection coefficient. From Figure 34, it is observed that resonant modes two and three are also excited at high elevations.

Their gains and center frequencies are elevation dependent. In Section 3.3.2 neither of the height resonances are modeled as band-pass filters because they are not excited at (0,0).

It actually turns out that as elevation increases, and the reflection distance decreases, one of the peaks of the comb filter matches up well with the envelope change that occurs with the increase in elevation. There is not a distinctly defined resonance at high elevations, but instead there is an extension of the depth resonance that results in a boosting of the frequencies between 7 and 10 kHz. Since the value of the reflection coefficient in this range (at short reflection distances) is relatively large, the comb filter's second peak can account for the envelope changes that occur at high elevations. Physically, this makes sense, and it also avoids the need to add additional resonances to the model in Figure 41. The envelope change between 7 and 10 kHz is best demonstrated visually; therefore, Figure 44 is included to show the PRTFs of CIPIC subject 20 at three different positive elevations. In this figure, the manner in which the width of the depth resonance increases with an increase in elevation can be seen easily.

Some other important observations regarding the elevation trends that occur in the PRTF can be made from Figure 44. The most obvious of these observations is the increase in the frequency location of the first spectral notch. At 0° the first notch is located at 6.8 kHz, at 30° the first notch is located at 9 kHz and at 60°the first, and only, notch is located at 11.33 kHz. In this example the notches are spaced apart somewhat regularly, which makes them predictable, but this is rarely the case with most human subjects. As a counter example, the notch locations for the same elevations of CIPIC subject 48 are: 7 kHz, 8.3 kHz and 10 kHz. These locations comprise a distribution that is less uniform than that of subject 21. Examining more than three elevations for any subject will further prove the point that the elevation notches are seldom spaced apart uniformly. This shift in the

frequency location of the first notch with elevation that is observed in Figure 44 is expected based on the prior art that was introduced earlier [20, 28]. The depth of the first spectral notch, and other subsequent notches, also varies with elevation. This makes sense based on the principles of physics because the reflection distance decreases with an increase in elevation which results in less scattering of high frequency waves.

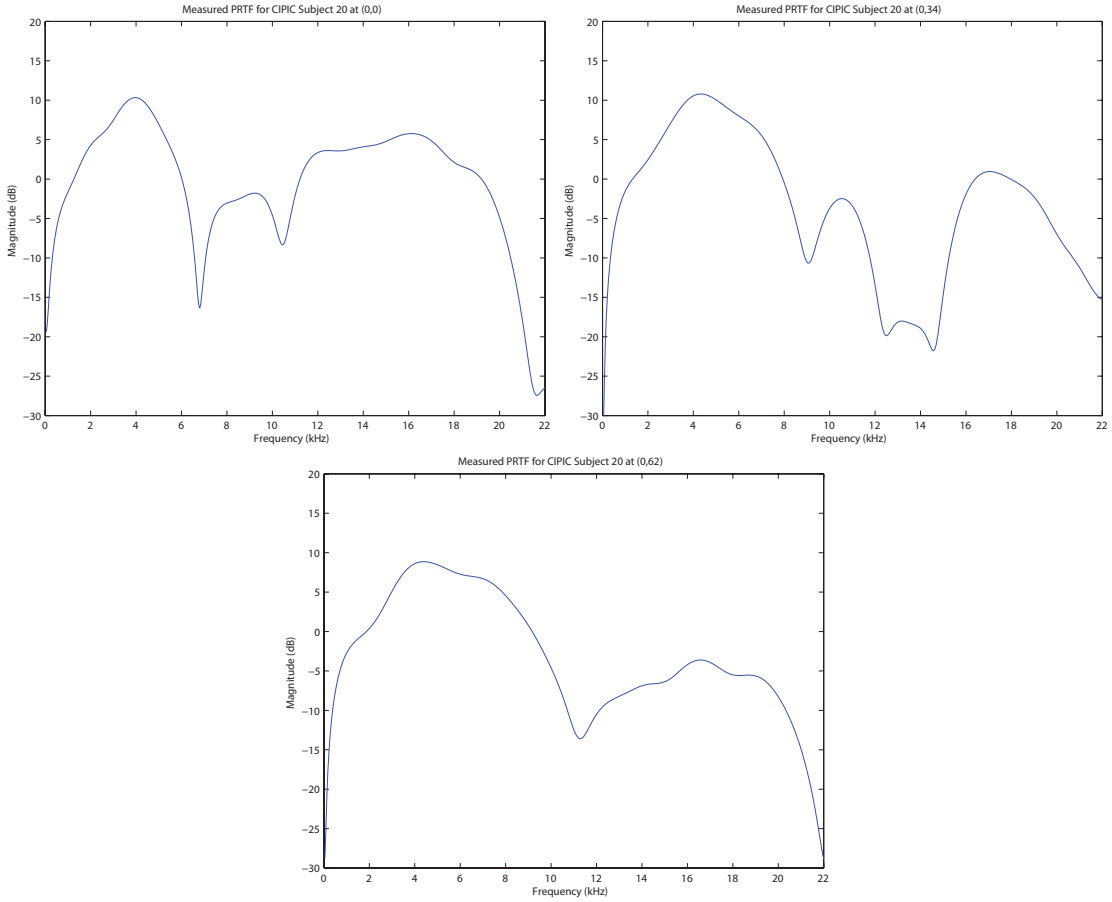

Figure 44. The median plane PRTFs of CIPIC subject 20 at elevations of 0° (top left), 34° (top right) and 62° (bottom).

Another significant observation that can be made from the plots in Figure 44 is that the gains of the width resonances decrease as elevation increases until they are close to zero at an elevation of 60°. Although negative elevations are not shown in Figure 44, the gains of the width resonances also decreases in a similar manner as elevation declines from 0 to -45°.

All of the trends mentioned in the prior paragraphs must be accounted for in a PRTF model if accurate elevation localization in the median plane is desired.

Shaw's work indicates that both of the width resonances are excited most at elevations near 0° which agrees with the observations mentioned in the previous paragraph; therefore, as the elevation of an incident sound strays from 0° the gains of the width resonances are affected. The behaviors of the gains of the width resonances were examined for five subjects in the CIPIC database and nearly linear trends were found to be common across all subjects. Two curves (one for each width resonance) that plot elevation against gain were created based on the behaviors observed in the gain analysis. The values of each gain at any elevation between -45° and 65° can be found by interpolating onto these curves. Figure 45 shows the two curves that are used to determine the gains of the width resonances at any elevation between -45° and 65°.

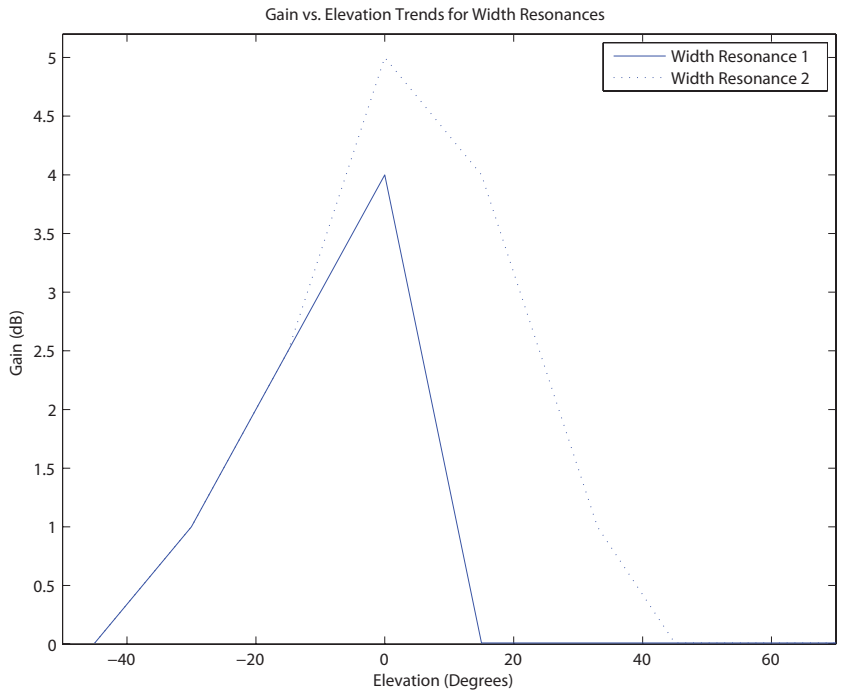

Figure 45. The elevation dependence of the width resonance gains.

Since this method for calculating the gains of the width resonances is not based upon anthropometry, it is obviously not personalized; however, it does seem to accurately model the behavior for most subjects in the CIPIC database. Naturally, there will be some cases in which this method is not as effective; the perceptual effects of these anomalies still remain unknown. The subjective testing in subsequent chapters may provide some insight into this.

Perhaps the most important elevation trend to model is the way in which the reflection coefficient changes with reflection distance. This is crucial because for every notch in a comb filter there is a corresponding resonance, and if the response of the reflection coefficient remains constant above the high-pass cutoff, then these peaks will be very large which is unrealistic. Attenuating the formants after the first notch is critical in accurately modeling the PRTF.

It has already been established that the reflection coefficient is of a band-pass nature with the high-pass cutoff fixed at 5.7 kHz, but the location of the low-pass cutoff and the response between the two cutoffs varies drastically with reflection distance. The response of the reflection coefficient is a multi-variable problem since the gain of the band-pass filter depends on frequency and reflection distance. Ideally, an equation that models the reflection properties of the pinna could be computed either from data or from the physics of the problem; however, since there is more than one variable, establishing such an equation is very difficult. This is why a hybrid method is used in this work.

The algorithm created to solve this problem is based upon the behavior of the reflection coefficient at various concha reflection distances for five subjects in the CIPIC database. Gain values at various points are averaged across the five subjects at a series of reflection distances and a set of curves are created from this gathered data. These curves are stored in memory as part of the algorithm along with the curves to calculate the gains of the

width resonances. When the algorithm is running, two-dimensional interpolation is performed using the curves that are stored in memory and the given reflection distance in order to calculate the gains of certain frequencies of the reflection coefficient. The band-pass filter that is used as the reflection coefficient is then designed in real time based upon the previously calculated frequency response. In Figure 46 the response of the reflection coefficient is shown for three concha reflection distances. This figure shows that the curves based upon data from CIPIC subjects model the known physics of the problem fairly well.

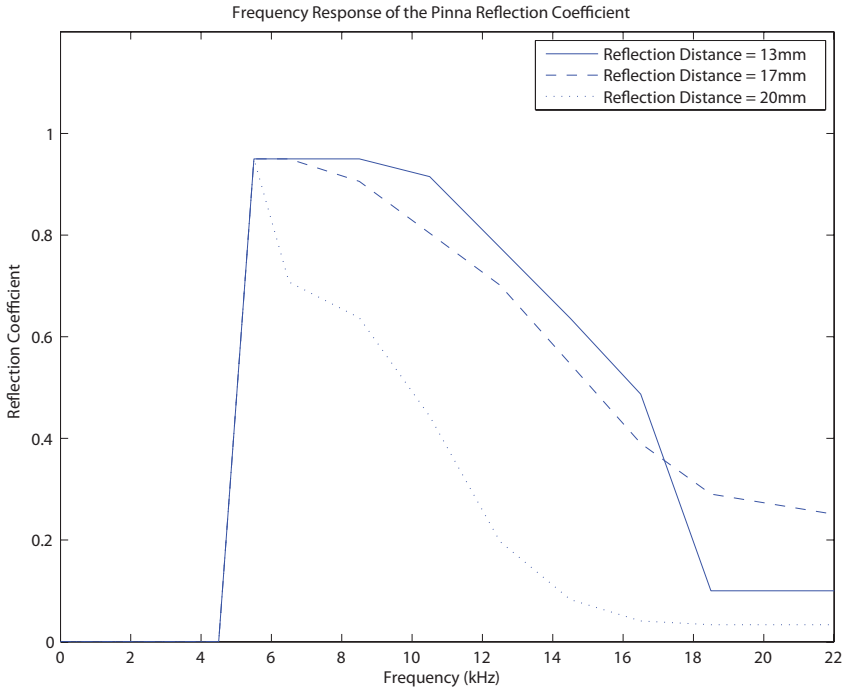

Figure 46. The frequency response of the pinna reflection coefficient at three different reflection distances.

One final phenomenon that must be modeled is the additional reflection off of the crus helias that occurs at low elevations. Accounting for multiple reflections adds a great amount of complexity to the model and is difficult to accurately realize mathematically. In this implementation, the additional reflection is modeled by cascading an extra FIR comb filter onto the end of the model shown in Figure 41. Based on the inputted anthropometry,

a conditional is used to determine if a crus helias reflection exists. If it does, then the time delay of the second comb filter is calculated from the measured distance to the crus helias reflection point. Some of the listeners used in subjective testing possessed crus helias reflections at elevations as high as 0° while others did not exhibit any secondary crus helias reflections. The pinna displayed in Figure 35 exhibits crus helias reflections up to around -7°.

The results of the PRTF model at median plane elevations above and below 0° are left out of this section to avoid redundancy because in the next chapter objective results for the entire HRTF model are shown for various CIPIC subjects at certain spatial locations.

3.4 CONTRIBUTIONS

The main contribution of this work is the addition of anthropometry-based pinna elevation cues to the model introduced in [29]. Using a reflection coefficient of a band-pass nature that is dependent upon reflection distance is also novel. The incorporation of pinna Mode 6 into the model and accounting for a secondary crus helias reflection at low elevations are other noteworthy additions.

A few additions are also made to the HAT model. Non-symmetric ear offsets and a reflection coefficient that is dependent upon both frequency and orientation are included to model the contributions of those body parts more accurately. These subtle modifications to the HAT model augment the unique contribution of this work and improve the performance of the model.

Finally, the method in which all of the modules of the implementation are connected in Figure 15 forms an HRTF synthesis algorithm (for frontal locations near the median

plane) that is based solely on anthropometry--something that, to the author's knowledge, has never been done in a computationally efficient manner.

4

OBJECTIVE RESULTS

Although no established criteria exist for objectively quantifying HRTF models, it is possible to compare the frequency response of each model to that of the original and analyze the similarities and differences. The primary HRTF model being investigated in this section is created by cascading the pinna model with the HAT model (in accordance with Figure 15), creating what is henceforth referred to as the PHAT model. The PHAT model is then plotted on same set of axes as the original HAT model and the measured HRTF. This allows for all three of the filters to be simultaneously viewable and compared easily, thus providing a decent way to objectively evaluate the validity of this work's hypothesis.

Without access to the pinna photos from the CIPIC database (due to privacy issues), most of the reflection distances had to be inferred from the measured HRTF plots in order to test the hypothesis of this thesis. This prevents a completely thorough objective investigation of this work's hypothesis from being possible; however, it is still possible to conceptually prove the idea that a HAT model can be improved by augmenting it with a well designed pinna model. All of the inferred reflection distances agree with the theories of past researchers in that the concha reflection distance generally decreases as the height of a sound increases.

The main part of this work that may suffer from not having access to the CIPIC pinna images is not being able to investigate exactly how the measured anatomical reflection distances correspond to the notches in the measured HRTFs. The biggest challenge with measuring the anatomical reflection distances exists in trying to figure out the correct

location to place the center of the coordinate system on the pinna image. More insight to this process may have been gained if the CIPIC pinna images were available. Past researchers, Shaw included, seemed to have used the same logic that this work uses (which is explained in Section 3.3) when identifying the center of the coordinate system on the pinna. With that in mind, not having access to the CIPIC pinna images may not be as much of a handicap as originally expected.

Since the reflection distances that are inferred are only done so at eight elevations per subject, the modeled HRTFs cannot be viewed as an image--a method that was used as a convenient way to view multiple HRTFs at the same time in previous sections. Instead, standard magnitude response plots will be shown for some select cases that demonstrate flaws or strengths of the PHAT and HAT models.

The first such example is shown in Figure 47. It is for CIPIC subject 10 at (0,-45). This is a good example of an HRTF that possesses a secondary crus helias reflection at a low elevation. This additional reflection can be seen as the notch around 10 kHz. It combines with the second notch created from the concha reflection to form one continuous notch between 10 and 12 kHz. The curve of the PHAT model follows that of the measured HRTF closely between 4 and 16 kHz--the most important frequency range in human elevation perception. For this particular subject, since the elevation of -45° is in the torso shadow cone, there is no torso reflection. This is why the curve representing the HAT model is very simple--it only attenuates high frequencies and exhibits no reflections. It is obvious that the PHAT model follows the curve of the measured HRTF more closely than the HAT model does. This provides evidence that the addition of a pinna model to a HAT model will result in improved elevation perception, and this is a trend that will be witnessed throughout this section.

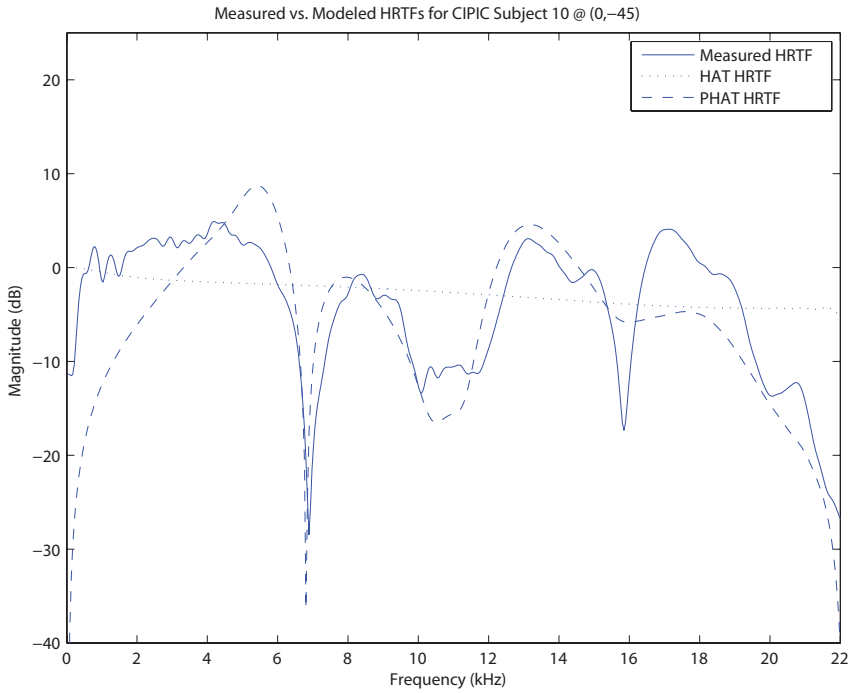

Figure 47. The measured HRTF for CIPIC subject 10 at (0,-45) plotted with the results of the HAT and PHAT models.

Figure 48 shows the next case that will be examined: CIPIC subject 20 at (0,-17). Torso reflection occurs for this subject and the HAT model curve displays this; however, the HAT model still does not match the measured HRTF plot very closely. The envelope of the PHAT model's curve matches well at both of the prominent notches but the gains at the first two resonances are attenuated. This is because when cascading the HAT model with the pinna model, the subtle torso reflection notches in the HAT model match up with the locations of the center frequencies of the depth resonance and the first width resonance of the pinna; this results in an attenuation at those frequencies in the final HRTF model. As a result, the entire envelope of the PHAT model is essentially shifted downward by 8dB for all perceptually relevant frequencies. To counter this, a uniform gain of 8dB can be added across the spectrum to the output of the PHAT model (which is equivalent to turning up the

volume of the headphones) to create a more exact match to the measured HRTF without introducing any undesired subjective effects.

The additional notches at approximately 7 and 10.5 kHz in this example may be related to a crus helias reflection but without access to the pinna images it is impossible to be certain. Even without accounting for a potentially extra reflection, the PHAT model still performs far better than the HAT model when they are both compared to the measured HRTF.

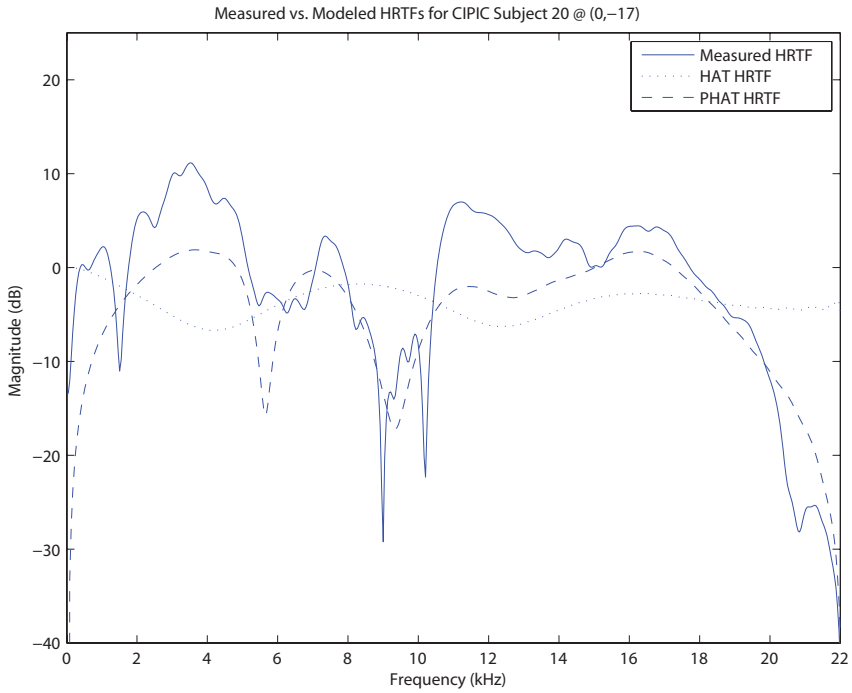

Figure 48. The measured HRTF for CIPIC subject 20 at (0,-17) plotted with the results of the HAT and PHAT models.

Examples from four subjects are displayed in Figure 49 for the spatial location of (0,0). Subject 20's plot is shown in the top left portion of the figure, subject 10's is in the top right, subject 33's is in the bottom left and subject 48's is in the bottom right. The PRTFs and pinna model results of subjects 20 and 48 at (0,0) are shown in the previous

sections, so it makes sense to show the results of the final model in this section. The HRTFs were accurately predicted by the PHAT model for both of these subjects.

Subject 33 is included to show an example of a case when the helix acts as the primary reflector. In such cases the reflection distance is considerably greater than when the concha acts as the primary reflector, but this isn't always the case. Because of the greater reflection distance, the resulting comb filter possesses more notches and resonances than the plots of subjects 20 and 48. The measured HRTF is in agreement as it matches up well to the output of the PHAT model. The first notch is in the correct location and the gains of the width resonances are nearly equal.

The PHAT model for subject 10 does not perform as well at (0,0) as it does for the prior three examples. The location of the first modeled notch is in the correct location; however, the second notch is not modeled, and there is a null at the first width resonance that is not present in the analytical data. The second notch can possibly be linked to a crus helias reflection but once again without access to the pinna images of the CIPIC subjects this is impossible to tell for certain. It is possible to test this theory by adding an additional reflection that possesses a notch at 11.2 kHz to the PHAT output of subject 10 at (0,0). The resulting plot is shown in Figure 50, and it is slightly more accurate than the one in Figure 49; however, the first width resonance is still somewhat attenuated. As was the case with the prior examples discussed in this section, the PHAT model once again out performs the HAT model in all four of these cases; this provides further evidence that modeling one or two pinna reflections in combination with the resonances of the external ear can accurately approximate the response of the pinna and improve upon a simple HAT model as a viable way to synthesize HRTFs.

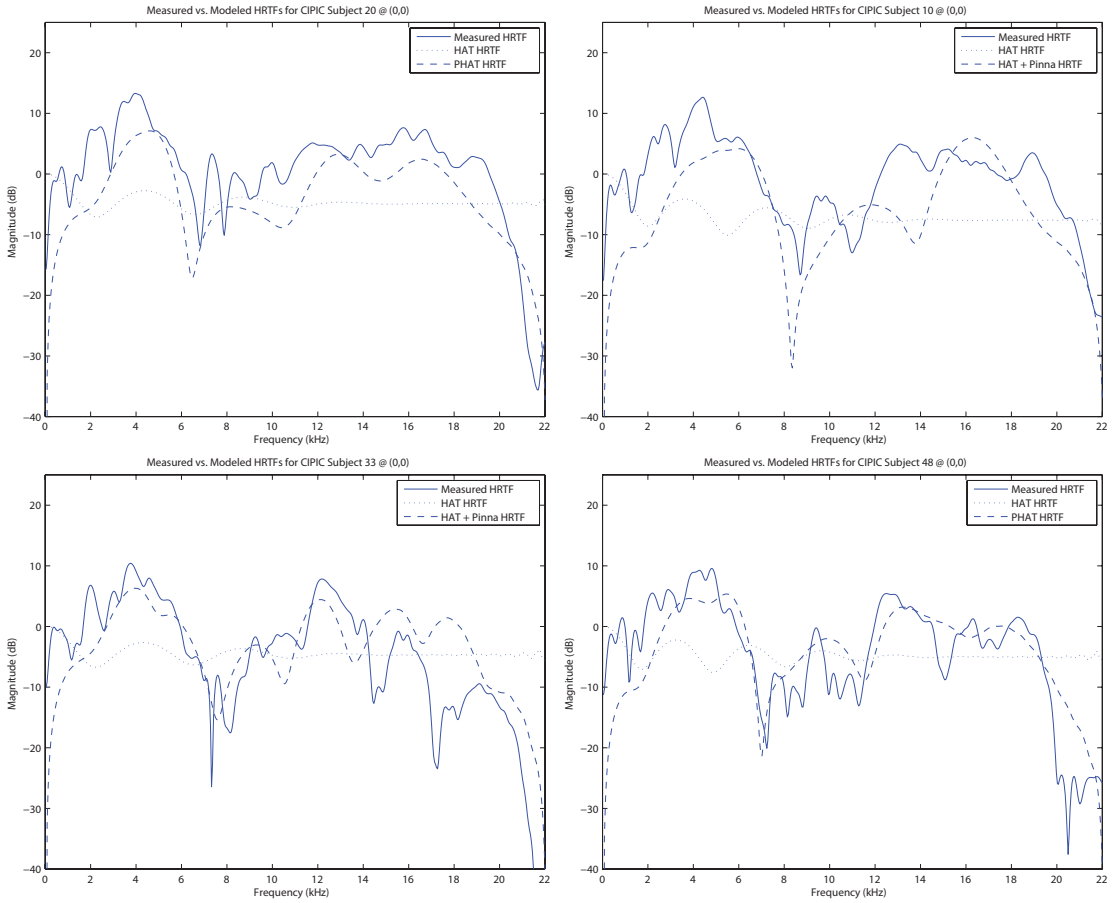

Figure 49. The measured HRTFs for CIPIC subjects 20 (top left), 10 (top right), 33 (bottom left) and 48 (bottom right) at (0,0) plotted with the results of the HAT and PHAT models.

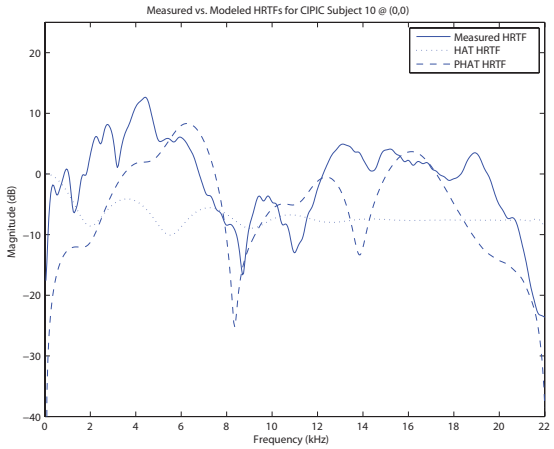

Figure 50. CIPIC subject 10 at (0,0) with a crus helias reflection.

For the median plane elevation of 17°, two examples are provided in Figure 51: CIPIC subject 48 (left) and CIPIC subject 20 (right). The first notch in subject 48's PHAT model matches well while the second notch is about 1 kHz too high. This interferes with the first width resonance; therefore, perceptual inaccuracies may result even though the first width resonance theoretically isn't as excited at this elevation as it is at 0°. The envelope of the PHAT model for subject 20 fits more closely than that of subject 48. The location of the first notch is 500 Hz greater in the PHAT model than it is in the measured HRTF. This also affects the location of the second notch since they are periodically related. If the location of the first notch is in fact the dominating spectral cue in elevation localization, as it has been hypothesized by prior researchers, then the 500 Hz error in this example will result in a perceived elevation slightly greater than 17°. In these two cases, even though the PHAT model doesn't perform as well as it does at other elevations, it still augments the accuracy of the HAT model.

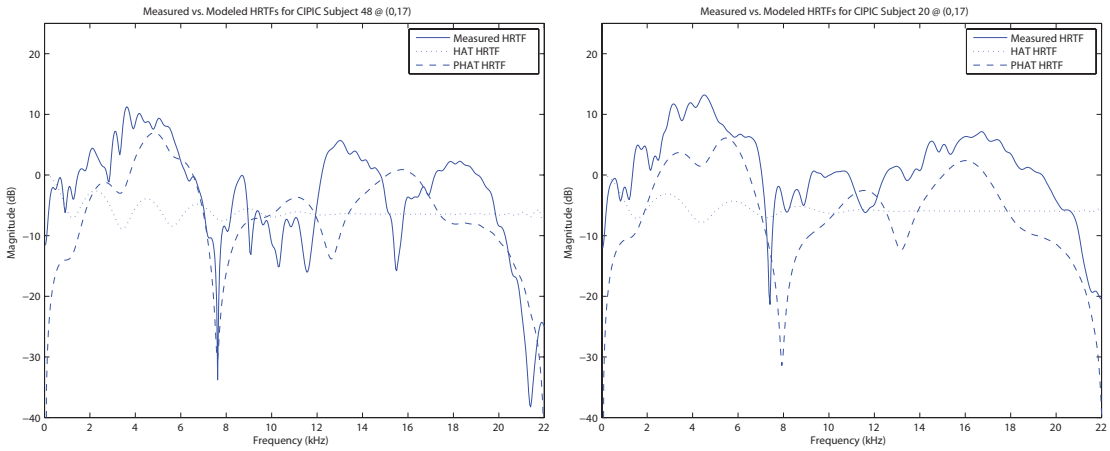

Figure 51. The measured HRTF for CIPIC subjects 48 (left) and 20 (right) at (0,17) plotted with the results of the HAT and PHAT models.

At high elevations it has been previously established that the bandwidth of the depth resonance increases due to its interaction with the comb filter that models the primary

reflection. Figure 52 shows an example of the performance of the PHAT model at an elevation of 45° for CIPIC subject 10. The width of the depth resonance is accurate in the PHAT model which results in more energy in the frequency band between 6 and 9 kHz--the range that is home to the first spectral notch (and therefore substantially less energy) at lower elevations. The torso reflections from the HAT model combine with the pinna model to accurately approximate some of the microscopic peaks of the measured HRTF below 4 kHz. The overall gain of the PHAT model is uniformly less than the measured HRTF, which was seen previously in the (0,-17) example, and once again, it is of little concern because it can be easily rectified. At an elevation as high as 45°, the first width resonance is not excited at all; this is evident in both the PHAT model's curve and the plot of the measured HRTF.

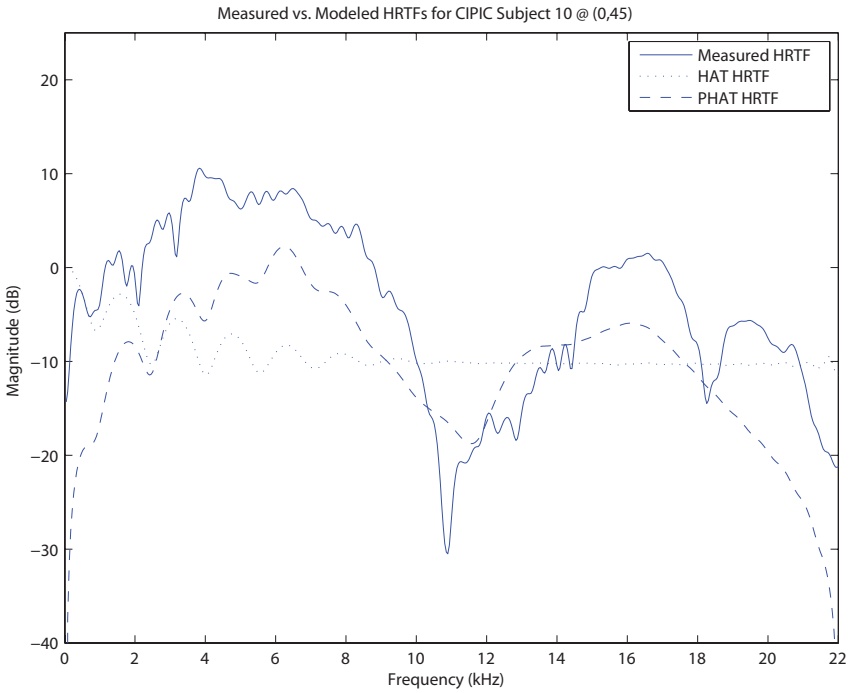

Figure 52. The measured HRTF for CIPIC subject 10 at (0,45) plotted with the results of the HAT and PHAT models.

The final median plane example is shown in Figure 53 for CIPIC subject 20 at an elevation angle of 62°. For this subject, the concha reflection distance is actually slightly

greater than it was for the subject shown in the previous example at 45°; this results in its first notch being at a lower frequency. For the 60° case that is shown in this example, the measured HRTF is approximated very well by the PHAT model with the exception of the gain of the depth resonance. Whether or not this has perceptual effects will be revealed in the listening test that is explained in the subsequent chapters.

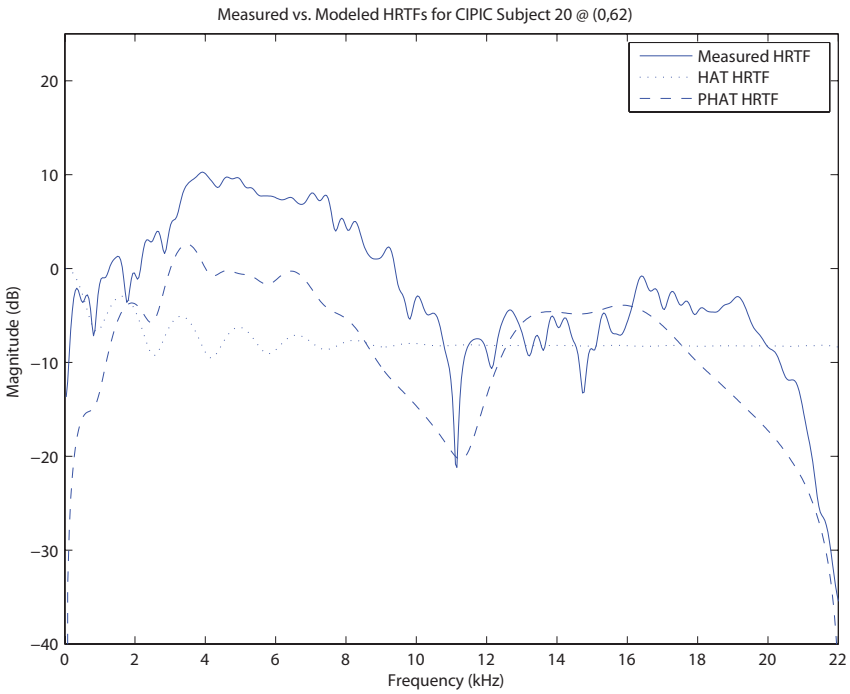

Figure 53. The measured HRTF for CIPIC subject 20 at (0,62) plotted with the results of the HAT and PHAT models.

The final cases examined are those that demonstrate the performance of the model for locations that deviate slightly from the median plane. At an azimuth of 20° off of the median plane in either direction, the model performs well for the ipsilateral ear and relatively poorly for the contralateral ear. At 45° away from the median plane, the ipsilateral ear's primary reflector is somewhere inside of the concha; this makes the phenomenon three-dimensional, and the necessary three-dimensional data is not provided, nor can it be easily acquired. Also, at such an azimuth, the presence of secondary waves at the contralateral ear

is substantial, and the arrival angles of these waves at the ear canal are not necessarily equal to the source's elevation angle. This is contrary to what the PHAT model assumes, and this explains why it performs poorly at azimuths very far from the median plane.

In Figure 54 some arbitrary examples of the PHAT model's performance on the ipsilateral side of the head are shown for an azimuth of 20°. The model is most accurate at elevations at and above -17°. Examples of two subjects (10 and 48) at -34° are shown, respectively, in the top right and lower left sections of Figure 54. The notches in the PHAT models of these examples line up well with the measured HRTFs, but the depths of the high-frequency notches in the measured HRTFs are much deeper than those of the model. At higher elevations the model matches much more closely. This can be seen in the top right and bottom right plots of Figure 54. The top right plot shows the response of the model for CIPIC subject 20 at an elevation of -17°. It matches well with the exception of the depths of the second and third notches. Lastly, an example showing the accuracy of the model at very high elevations in the ipsilateral hemisphere (62°) is shown in the bottom right plot of Figure 54. This is the best fitting example of the four shown.

Figure 55 contains examples for the same subjects and locations that are shown in Figure 54, but they are for the contralateral ear. These results are much less accurate than their ipsilateral counterparts due to reasons that have been previously explained. At an azimuth of 20° on the contralateral side of the head, the effects of the reasons explained earlier are not as prevalent as they are at an azimuth of 45°; however, they still do exit, and this is the reason for the relatively poor objective performance of the model in such cases. There is also an overestimation of the head shadow from the spherical model on the contralateral side of the head, and this may result in horizontal localization inaccuracies. Comparing the PHAT model plots in Figures 54 and 55 for frequencies above 4 kHz reveals

that the only difference between the ipsilateral and contralateral cases is the amount of head shadowing/boosting that is present. The envelope shapes of the PRTF are otherwise identical. The work on approximating the contralateral HRTF in [9] from which the approach used here is based had similar results--the amount of localization error increased as the incident sound moved closer to the interaural poles. The perceptual effects of this off-the-median-plane approximation method at an azimuth of 20° are investigated in the subsequent chapters.

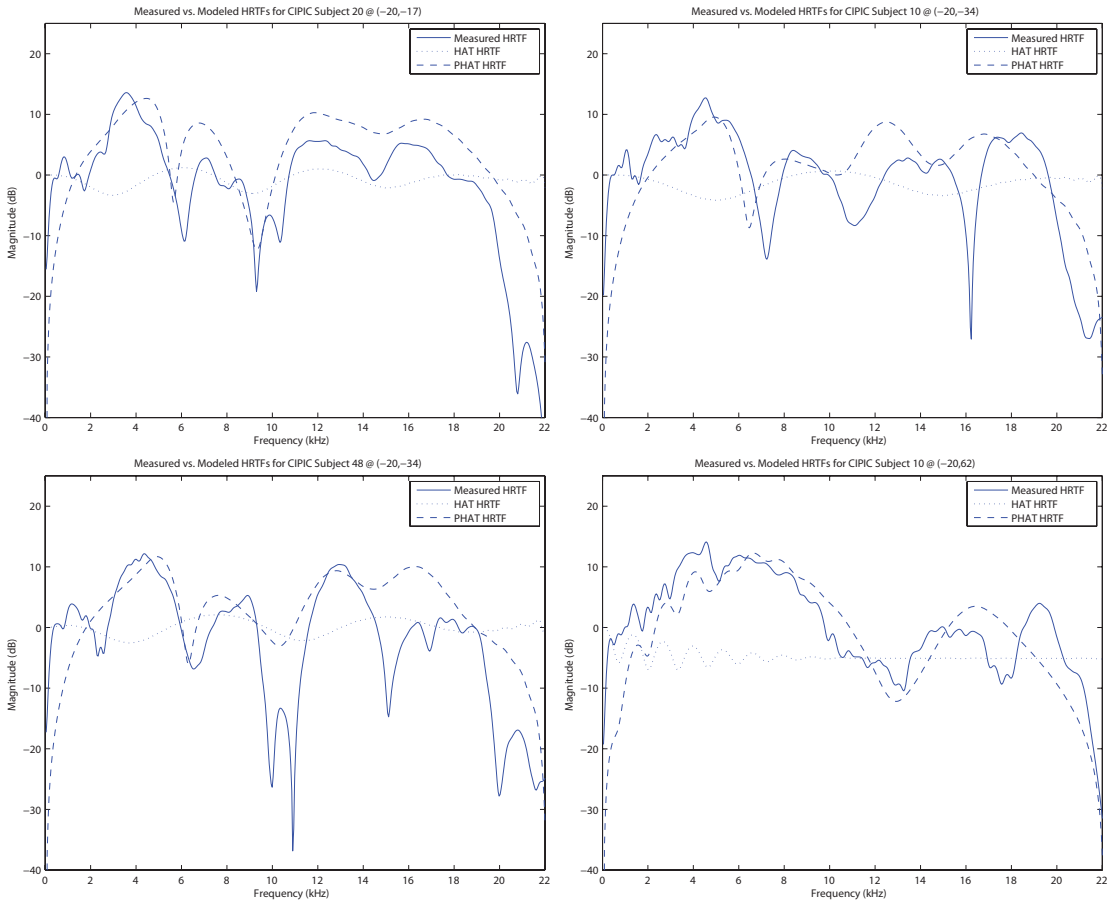

Figure 54. Examples of the model for the left ear at an azimuth of -20° for various subjects and elevations. The top left plot is of CIPIC subject 20 at -17°; the top right plot is CIPIC subject 10 at -34°; the bottom left plot is CIPIC subject 48 at -34°; the bottom right plot is CIPIC subject 10 at 62°.

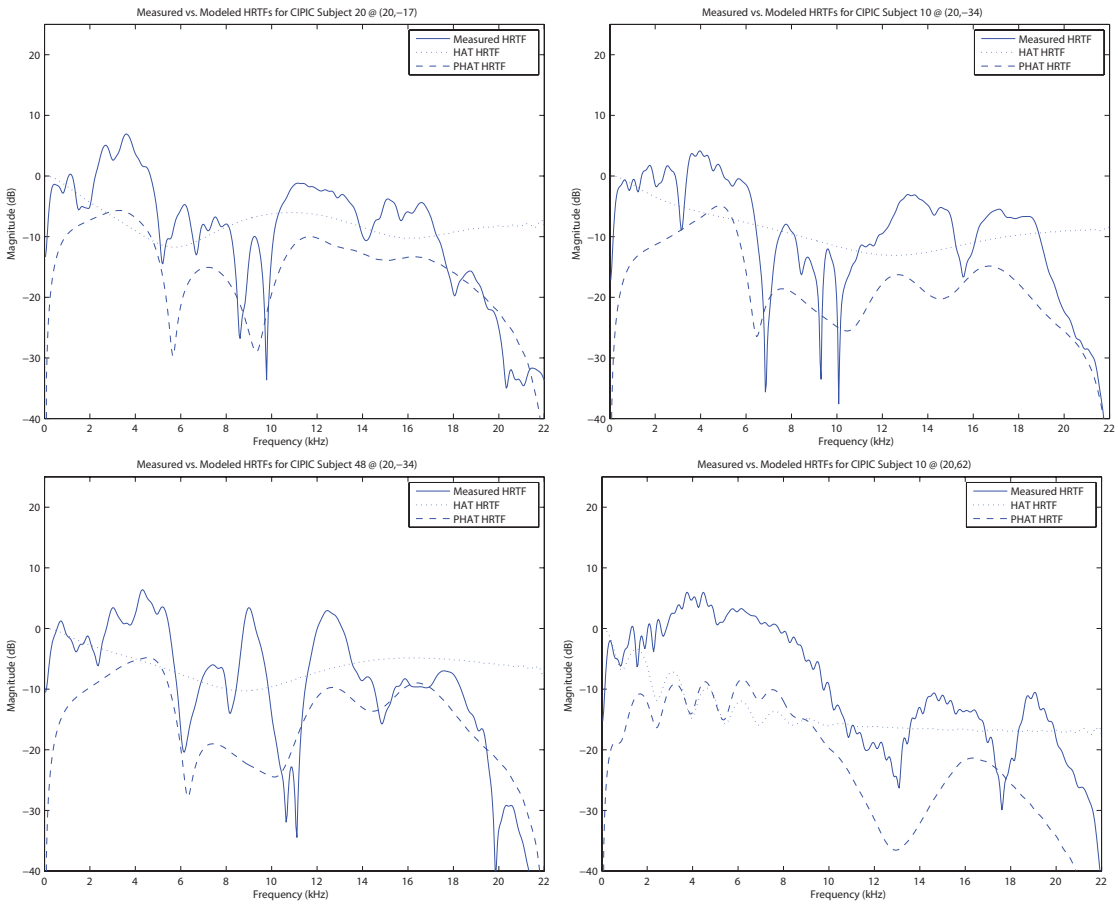

Figure 55. Examples of the model for the left ear at an azimuth of 20° for various subjects and elevations. The top left plot is of CIPIC subject 20 at -17°; the top right plot is CIPIC subject 10 at -34°; the bottom left plot is CIPIC subject 48 at -34°; the bottom right plot is CIPIC subject 10 at 62°. This is the same as Figure 53 with the exception of the azimuth angle.

There are two known cases where the PHAT algorithm does not perform exceptionally well. The first of these cases is when the reflection distance for an elevation near 0° results in a notch that is located near the center frequency of first width resonance. This was mentioned earlier in the (0, 17) example. In the physical world, when the reflection interacts with first width resonance, the resonance may dominate. Preventing the notch from dominating can be accounted for in the PHAT model by adjusting the response of the reflection coefficient at the frequency that corresponds to the width resonance. A simple conditional can be used to compare the locations of the width resonances and the notches at

certain elevations near 0°. If a common notch and resonance location is detected, then the reflection coefficient can be adjusted accordingly. Another cause of this problem could be due to an inaccurate estimate of the primary reflection distance.

Another problem, that is more serious than the case just explained, is that if the reflection distance at high elevations is less than 10.2mm the model will fail perceptually. The reason that the model breaks down at such short reflection distances is because the comb filter's first notch (which at most other reflection distances occurs before the high pass cutoff of the reflection coefficient's band-pass response and is therefore 'blocked') creeps in at around 5.5 kHz for a reflection distance of 10.2mm and at increasingly higher frequencies with shorter reflection distances. The presence of this additional notch also prevents the envelope of the PHAT model from possessing high energy content in the 6 to 9 kHz region (which is necessary in the perception of high elevations) because it no longer accurately augments the bandwidth of the depth resonance. Fortunately, none of the CIPIC subjects possessed reflection distances of less than 10.2mm at the highest elevation examined (62°). Physically, if the primary reflection distance is in fact less than 10.2mm, then the model does accurately approximate what would happen to the frequency response. This fact, in combination with the fact that none of the CIPIC subjects possessed reflection distances of less than 10.2mm at the highest elevation examined, leads to the conclusion that this may be an issue with the measuring the primary reflection distance from digital images. More details regarding this problem, along with its subjective effects, are discussed in Chapter Six.

It is worth mentioning that the PHAT algorithm is capable of being run in real time on a modern CPU or DSP. The coefficients of the three third-order band-pass filters used to approximate the resonances of the pinna require approximately 25 arithmetic operations per filter to be calculated. The delay used in the comb filter is always less than 50 samples,

and the up and down sampling done to gain a precision of a tenth of a sample is always done on an impulse response of 256 samples; therefore, the entire comb filter portion of the algorithm is not CPU intensive. The filter used as the pinna's reflection coefficient is on the order of 40-80 taps, which is somewhat high, but in most cases at least at quarter of the multipliers are zeros which simplifies the computation. With a buffer of at most 100 samples, which corresponds to 2.2ms of lag time, any modern processor can run the PHAT algorithm in real time.

The algorithm was not tested in C++; however, in MATLAB (a notoriously slow resource hog), all of a given subject's test files for the listening test were generated in mere seconds. This leads one to believe that the algorithm will run in real time in an efficient programming environment such as C or C++. If the algorithm does turn out to be too taxing for a given processor, then the length of the outputted impulse responses can be cut in half to 128 samples. This will effectively cut the amount of required processing in half.

The memory requirement for this algorithm is less than conventional HRTF-based binaural synthesis implementations. This is because only 60 anthropometry measurements per subject need to be stored in memory (for the eight elevations used in the listening test). Most other binaural synthesis algorithms require 256 filter coefficients per spatial location per subject to be stored in memory. This is far greater than the minuscule memory requirement of the PHAT model.

Based on the results shown in this chapter, it is evident that by modeling the first-order resonances and reflections of the external ear and cascading the results with a HAT model that HRTFs can be accurately approximated using only anthropometric measurements as input parameters. Even though the exact reflection distances are unknown because access to the pinna images was not possible, it was still conceptually proven that the

proposed model performs closer than the HAT model to measured HRTFs in every case examined.

5

LISTENING TEST

While Chapter Four did indicate promising results for a limited number of examples from the CIPIC database, a more revealing assessment of the implementation is only possible through a listening test. The design of such a test is explained in this chapter and the results are analyzed in the next chapter.

5.1 OVERVIEW

The listening test is conducted in three stages, two of which require the subject to be present. At the first session, multiple images are taken of each subject. In the second stage, the morphological measurements required by the HAT and PHAT models are extracted from each subject's images by the test's principle investigator. The obtained anthropometry is then inputted into each model and custom HRTFs are generated for a specific set of spatial locations. White noise is then filtered with all of the generated HRTFs and saved to a computer's hard drive. The third portion of the test involves each subject listening to the sounds filtered with their custom HRTFs over headphones and localizing them.

5.2 ANTHROPOMETRY ACQUISITION

The first step to acquiring the anthropometry that is necessary for the PHAT and HAT algorithms is to take seven digital images of each subject. A ruler is present in each image to provide the necessary scale. For the head and torso measurements, a 24" steel ruler with gradations of 1/16 of an inch is used. The ruler also contains the equivalent metric quantities, but the measurements for the head and torso parameters are done in inches and

then converted to meters within the program. For the pinna measurements, a ruler graduated at every millimeter is used. At first this was thought to be insufficient due to the precision needed for the reflection distances, especially at high elevations, so a ruler with a resolution of 1/4 of a millimeter was tested. It turns out that when using the ruler feature of the *imtool* function in MATLAB's image processing toolbox to extract the measurements the results from images featuring each ruler (1mm and 0.25mm) were in very close agreement. This is because the MATLAB function gives the measurements to the nearest hundredth of a pixel. If 10mm is measured on the ruler graduated at every millimeter and the resulting number of pixels is divided by 10, then number of pixels per millimeter can be obtained. When measuring one millimeter on the ruler with graduations every 0.25mm, the results were less than a pixel different from those of the 1mm ruler. Once the proportion is set up to relate 1mm to the equivalent number of pixels, the pinna measurements, in theory, can have a precision of greater than a hundredth of a millimeter using either ruler which is sufficient for this experiment.

The required measurements of the head are depth, width, height and left and right pinna offsets (down and back). To obtain these measurements three snapshots of the head are taken in classic mug shot fashion with the tops of the subject's shoulders at the bottom of the image. The first picture is of the subject facing forward, and the next two images are of the subject facing left and right. The degree of pinna flare (to determine the primary reflector) can be examined from the photo of the subject facing forward. Neck height, which is another required parameter, can be obtained from any of the three images. Head depth can be measured from both sideways images, and head height can be measured from all three photos. The ability to measure certain distances in multiple images allows for some degree of error control in the measurements. For most of the subjects, the head

measurements that could be obtained from more than one image rarely deviated by more than 0.5cm. Head width and pinna offsets can only be measured from one image which makes these measurements most susceptible to error. All of the head measurements are taken in accordance with those shown in Figure 4 and explained in Section 2.4.

Only two images of the subject's upper body are necessary to obtain the required torso parameters of torso height, depth and width: one in which the subject is facing straight ahead and one in which the subject is facing either right or left. Neck height, which is arguably the most important parameter in modeling torso reflections, can also be measured from either torso image (in addition to the previously mentioned head images); this prevents a large amount of error from being present in this critical measurement.

One snapshot is required for each pinna. A close up of the subject's pinna is taken from the side of the head while the subject holds the aforementioned small metric ruler that is graduated every millimeter alongside his/her outer ear. The coordinate system, as shown in Figure 33 and explained in Section 3.3, is then superimposed onto the image using Adobe Photoshop. The concha width, pinna width and crus helias distance (when applicable) are then measured at the following elevation angles: -45°, -30°, -15°, 0°, -15°, 30°, 45°, 60°.

Figure 56 shows an example of the pinna measurement process. In the zoomed out image, 55.13 pixels equals 10mm, and based on that proportion, the subject's concha width at 0° is 15.645mm. The rest of the pinna measurements are done in the same way.

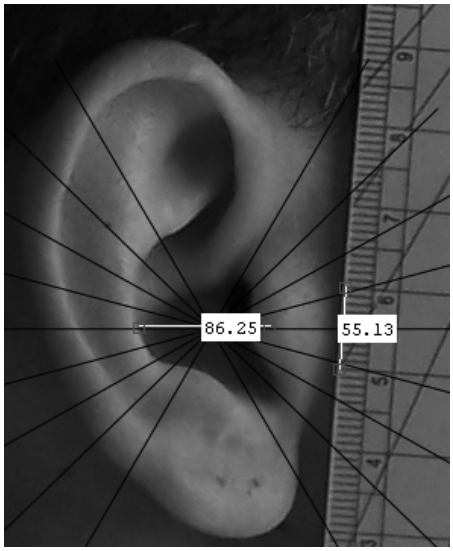

Figure 56. An example of the anthropometry acquisition process for subject 1's right pinna at 0°.

5.3 TEST DESIGN

After acquiring all of the necessary anthropometry from the images, two sets of customized HRIRs are generated for each subject. The first set is synthesized using the HAT model, and the other set is created from the PHAT model. If it is uncertain whether the primary reflector for a given subject is the concha or the helix, then two separate filter sets are generated using the PHAT model for certain elevations. At elevations above 15°, it is assumed that the concha is always the primary reflector due to the lack of definition of the helix rim towards the bottom of the pinna.

The HRIRs of each set are generated at the following spatial locations: (0,-45), (0,-30), (0,-15), (0,0), (0,15), (0,30), (0,45), (0,60), (20,-45), (20,-30), (20,-15), (20,0), (20,15), (20,30), (20,45) and (20,60). These locations are chosen because they are all at and around the median plane which is where the PHAT algorithm has been objectively shown to be most effective. In addition to the files for each individual location that are created, files with a series of increasing or decreasing elevations are created by concatenating a series of

individual elevation files together. The nature and purpose of these files is explained later in this section.

It is well documented that wide bandwidth stimuli are easiest for the human auditory system to localize [11, 14, 18, 24, 33]; therefore, it is logical to use white noise as the stimulus. Each test sound is sampled at 44.1 kHz and is 4.5 seconds in length. It consists of three 500ms bursts of noise separated by 500ms of silence and then concludes with 1.5s of continuous noise. The sound is then filtered with all of the generated HRIRs for each subject. These filtered sounds are saved as .wav files onto a laptop's hard drive and randomly labeled *xx_yy_z.wav* with *xx* representing the subject number, *yy* representing the file number and *z* being 1 or 2 to denote the method used (1 = HAT, 2 = PHAT). A record is kept by the principle investigator that relates the subject number and file number to the subject's name and the corresponding spatial location, respectively.

Without access to expensive head motion trackers, some creativity is required to design a sound localization listening test. One main problem with said limitation is that head motion by the subject will skew the results because spatially synthesized sounds presented through headphones that do not appropriately account for motion tracking allow for the location of the sound to move with the subject's head. To deal with this problem in the listening test, a very small mirror is placed on the wall directly in front of the listener and the listener is instructed to ensure that their eyes are visible in the mirror at all times while listening to the test sounds. Having the subject make sure that their eyes are always visible in the mirror also serves a second function--it guarantees a common interaural center across all of the subjects. A chair with an adjustable height is provided so that the subject can adjust it until they see their eyes in the mirror.

On the wall directly in front of the seated subject is a large grid. The labeled points on the grid correspond to every 5° of median plane elevations between -50° and 65°. The previously mentioned mirror is located at an elevation of 0°. The rest of the angles labeled on the grid are calculated using simple trigonometry. The base of the right triangle used in the angle calculation is the distance from the wall to the center of the seated subject's head. This distance is fixed at .5m to prevent high-elevation angles from being on the ceiling and low elevations from being on the floor. An illustration of the test conditions is shown in Figure 57. The distance denoted by *y* in this figure is equal to .5tanΦ, and it is calculated at 5° intervals for the aforementioned range.

Grado Labs Prestige Series Headphones (SR225) are connected to the sound board of the laptop that contains the database of generated test files. Apple's iTunes software is used as the media player for the .wav files. Each subject's test files are loaded into separate playlists in iTunes and the subject is given complete control over the sounds that are played. iTunes is set to continuously loop the selected sound that is playing because the duration of each individual location file is only four seconds, and it may take longer than that for the subject to accurately localize the sound.

At the beginning of each subject's playlist is the increasing and decreasing elevation files created from the HAT and PHAT methods. The increasing elevation file is created by concatenating the files at elevation angles of 0°, 15°, 30°, 45° and 60°. The similarly structured decreasing elevation file concatenates files with elevation angles of 0°, -15°, -30° and -45°. This is done using the files from the HAT and PHAT methods resulting in 4 total files: PHAT increasing, PHAT decreasing, HAT increasing and HAT decreasing. The subject starts the test by listening to these files and indicating whether the PHAT or the HAT methods give them a better sense of uniformly increasing and decreasing elevation.

Starting with this task provides a bit of informal training for the subject by familiarizing them with the nature of the sounds.

The rest of the test involves localizing individual sounds. To do this, subjects are provided with a laser pointer that they are instructed to hold in line with the entrance to one of their ear canals and point in the direction from which they perceive the sound. Since the synthesized sounds lack a great amount of externalization, the subject is told that the sounds may appear to be located just outside of their head and to point the laser in the vertical direction from which the sound appears to be located on their head. They are then instructed to make note of the location on the grid that the laser is pointing so that they can write that elevation angle on the answer sheet along with the corresponding file name. This is repeated for all sounds, and then the subject is thanked for their participation and dismissed.

The test was conducted in a dead room in the Weeks Recording Studio at the University of Miami's Coral Gables campus so that no outside sounds were presented to the subject. Upon arrival to the test the subject was asked to read the test instructions, and the procedure was demonstrated to them by the principle investigator. The instruction and answer sheets provided to each subject prior to the test can be found in Appendix A.

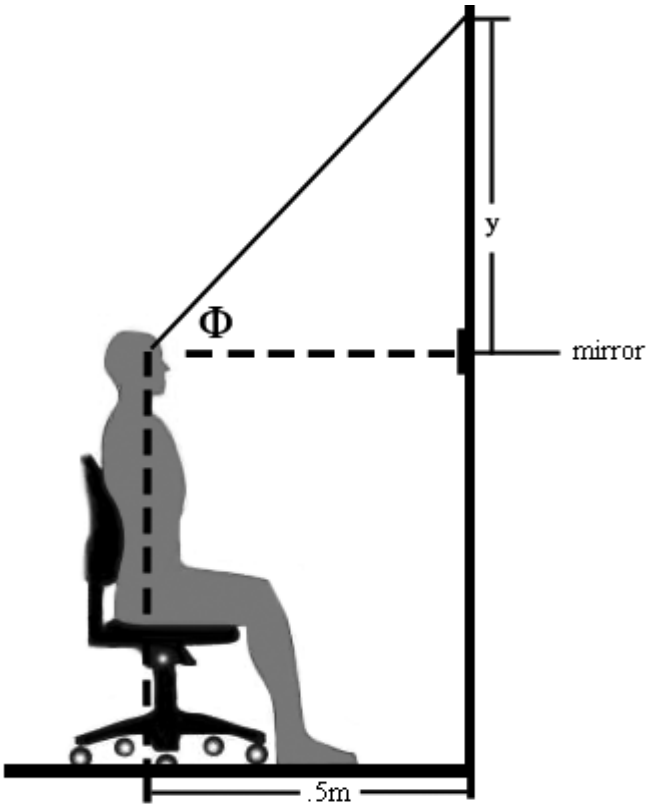

Figure 57. An illustration of the listening test conditions.

6

SUBJECTIVE RESULTS

Fifteen subjects of varying ethnicities, ages and gender (all of whom reported no documented hearing problems) were gathered to take part in the listening test that is described in the previous chapter. The morphological measurements (pinna and concha dimensions especially) deviated greatly across all of the subjects. This observation somewhat justifies having such a small sample space because many different pinna shapes and sizes are present across all of the participating subjects. Informal interviews conducted with some of the subjects after they completed the test by the principle investigator were, in some cases, as revealing as the data obtained from the test. Details of the findings from the interviews will be addressed later in this chapter--after the numerical results are reported.

The subjective results obtained from the listening test will be analyzed in three different groups. For 47% of the subjects the personalized HRTFs generated from the PHAT model performed very well and were consistently better those generated using the HAT model. This cluster of subjects will be henceforth referred to as Group I. An additional 20% of the subjects were able to very accurately localize the elevations of sounds 20° off of the median plane, but they performed poorly when localizing median plane elevations; these subjects comprise Group II. The results of the remaining 33% of the subjects, referred to as Group III, were very sporadic. The perceived elevations of the HAT and PHAT models for all subjects in Group III exhibited no correlation to the intended locations of the sounds both on and off of the median plane.

Reporting the combined results of all three groups is not revealing and consequently very uninformative; for that reason, they are not included in this analysis. Also omitted from this report are the results of Group III. Detailed reasons explaining the less-than-stellar performance of Group III are given, but accompanying these explanations with numerical results does not offer any additional insight; therefore, they are not presented.

Ideally, if the PHAT model worked perfectly, all of the perceived elevations would equal the intended elevations. In such cases, on a two-dimensional graph with the perceived elevation as the ordinate and the actual elevation as the abscissa, the x-values would always equal the y-values. For example, if the subject perceives a sound that was intended to be elevated by 45° at 45° and this result was plotted on the aforementioned coordinate system, a point would exist at (45,45). Extending this example to all of the tested elevations (-45° to 60° at increments of 15°) and connecting all of the points results in a line with a slope of 1 and a y-intercept of 0. This is the benchmark used in this chapter to gauge the quality of the experimental results.

An average value at every tested location for each model (HAT and PHAT) is calculated from the responses of the subjects in a particular group, and the points are then plotted on the perceived elevation versus intended elevation coordinate system that was introduced in the previous paragraph. A linear regression is then performed on these points which results in a best fit line with a specific slope and y-intercept. How close the line's slope is to one is an indication as to how good the results for that particular case are. Regression lines with slopes close to zero indicate very poor performance. The results are reported for each group for a couple of different cases, and each case is followed by some analysis. After all of the data is presented and explained, a discussion of some of the

potential causes of the poor performance of the PHAT model for the subjects in Group III is provided.

The first case analyzed is that of the Group I at an azimuth of 0°. Figure 58 shows the results of this case for the HAT model (left) and the PHAT model (right). A line with a slope of one is plotted on each graph to provide a reference to the ideal results. The value of the regression line's slope is provided at the top of each plot. In this case, the slope of the regression line for the HAT results is 0.20. This reveals that the elevation performance of the HAT model in the median plane for Group I is very poor. This is expected since the hypothesis of this work is to improve upon the poor elevation performance of HAT models in and around the median plane. The slope of the regression line for the PHAT model's results is 0.73; this indicates much better performance than the HAT model, thus proving the hypothesis of this paper for this case.

In the plot of the PHAT results, it can be seen that most of the plotted points are very near to the ideal line except for the points at the extreme values of -45° and 60°. For two subjects in this group, the reflection distance measured at an elevation of 60° is less than 10.2mm which has been identified previously to be the threshold of the pinna model. The subjective results further justify this theory that the PHAT model breaks down at reflections distances of less than 10.2mm because the two subjects in Group I whose reflection distances are less than the known threshold perceive the 60° sound created by the PHAT algorithm as having a negative elevation. It was discussed that the reason for the breakdown of the algorithm at such small reflection distances is because a notch is introduced near 6 kHz. Looking back at the measured HRTF of CIPIC subject 10 at (0,-45) in Figure 47 reveals a spectral notch at about 6.5 kHz. Based on the plot in Figure 47, and all of the prior art discussed in the pinna model section, it can be concluded that notches at very low

frequencies in the spectrum are interpreted by the human auditory system as being created by sounds at low elevations. The two subjects in this group with reflection distances of less than the PHAT algorithm's known threshold at 60° perceived the 60° test sound as being located at -45° and -15°. This is proof that the algorithm does in fact break down at such small reflection distances and that brain associates sounds with low-frequency notches as having low elevations.

If the responses at 60° for these two subjects are omitted from the results, then the point at 60° in the PHAT plot of Figure 58 will move much closer to the ideal curve. This new plot is shown in Figure 59 to have a slope of 0.87 which is a 20% improvement from the plot in Figure 58. With this in mind, it can be concluded that the PHAT model performs well up to 45° for the subjects in this group, and that the PHAT model performs well as long as all of the pinna dimensions of subject are not below the known threshold of the PHAT algorithm. The definite, and most obvious, conclusion is that the elevation performance of the HAT model is greatly improved with the addition of the pinna model introduced in this work.

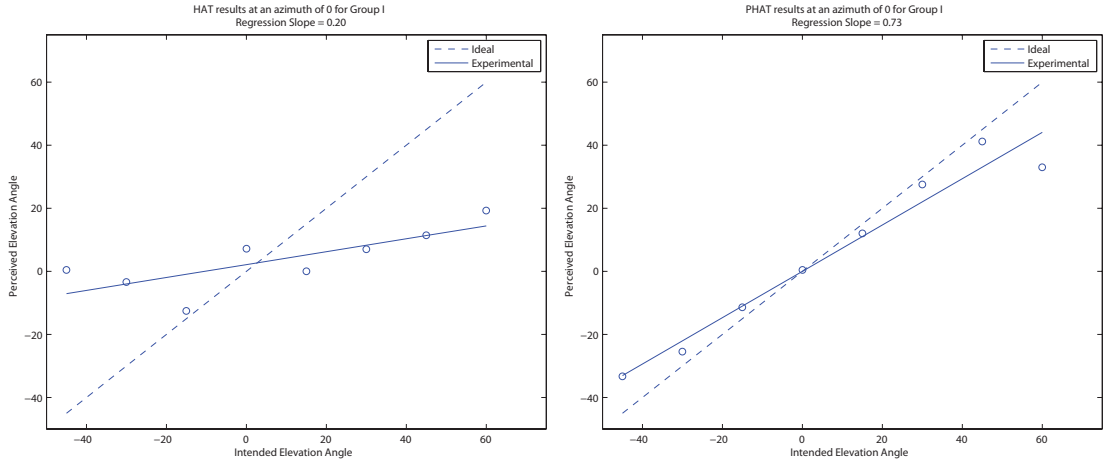

Figure 58. The subjective results for Group I at an azimuth of 0° for the HAT model (left) and the PHAT model (right).

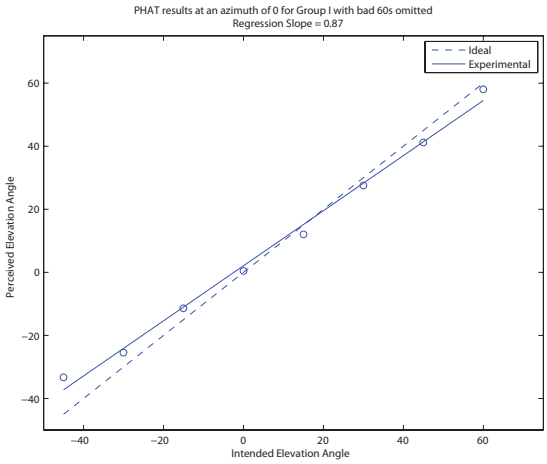

Figure 59. The subjective results for Group I at an azimuth of 0° for the PHAT model with the two explained outlier points at 60° removed.

The next case examined is the same as the prior one except that it is at an azimuth of 20°. Objectively, it was observed that the performance of the PHAT model at such an azimuth is worse than it is in the median plane. It was stated that the overestimation of head shadow on the contralateral side of the head theoretically should result in azimuthal localization inaccuracies, but that the notches were, for the most part, in the correct locations. The subjects of the listening test were only asked to focus on vertical localization, so the potential horizontal localization issues are not examined in this work. Results of this case are shown in Figure 60. These results are very similar to those of the previous case which is expected because the median plane performance for these subjects was very good; however, it is surprising because the analysis of the objective results suggested otherwise. It can be concluded that the depths of the notches are not as perceptually important as the locations of the notches. Once again, the results for the two subjects possessing reflection distances at 60° that are smaller than the threshold for the pinna model skew the results of the PHAT model. Figure 61 shows the results of Figure 60 with the 60° responses of the

two problematic subjects removed. The slope of the regression line in this case is 0.85; this is slightly worse than the median plane results which is expected.

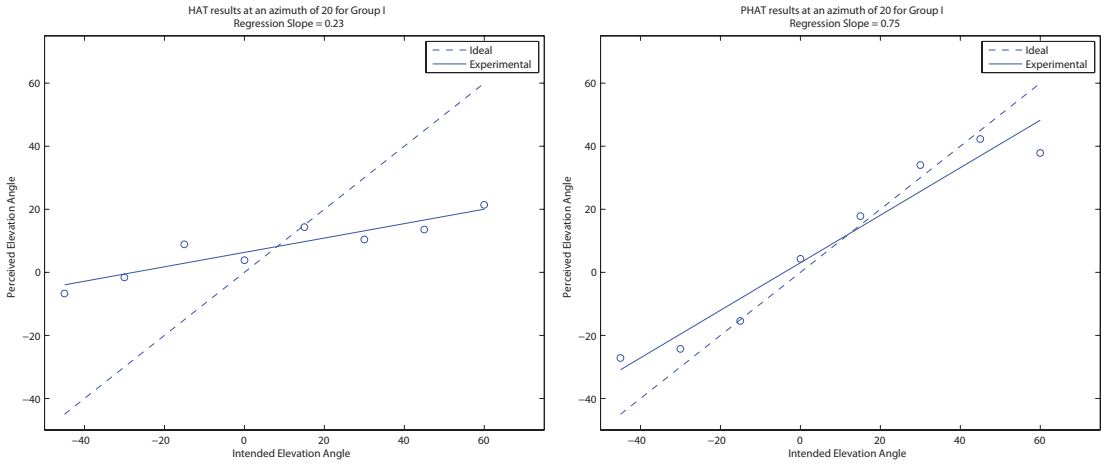

Figure 60. The subjective results for Group I at an azimuth of 20° for the HAT model (left) and the PHAT model (right).

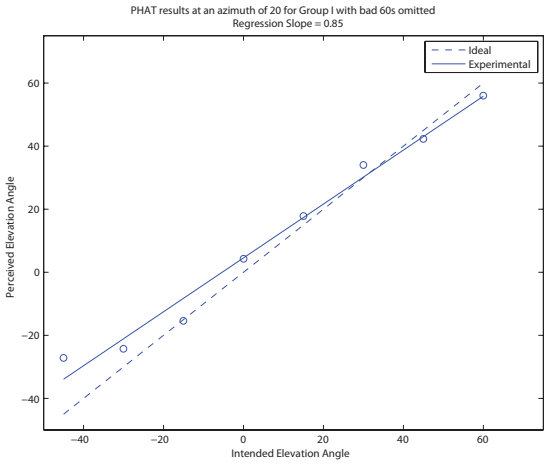

Figure 61. The subjective results for Group I at an azimuth of 20° for the PHAT model with the two explained outlier points at 60° removed.

As previously explained, the subjects in Group II correctly localized the elevations of sounds created with the PHAT model for azimuths off of the median plane, but they failed to accurately localize sounds in the median plane. The results for the subjects in Group II are added to those reported in Figures 60 and 61 to create the plots in Figures 62 and 63. The inclusion of the data from Group II causes the results to worsen. This is expected and

it also ends up with the overall off-of-the-median-plane localization results being worse than the results in the median plane; this is also expected.

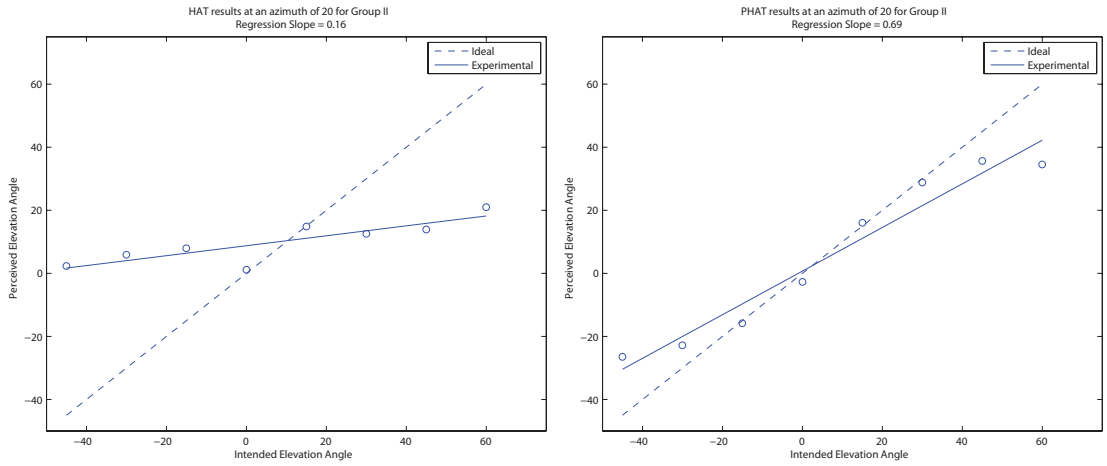

Figure 62. The subjective results for Group II at an azimuth of 20° for the HAT model (left) and the PHAT model (right).

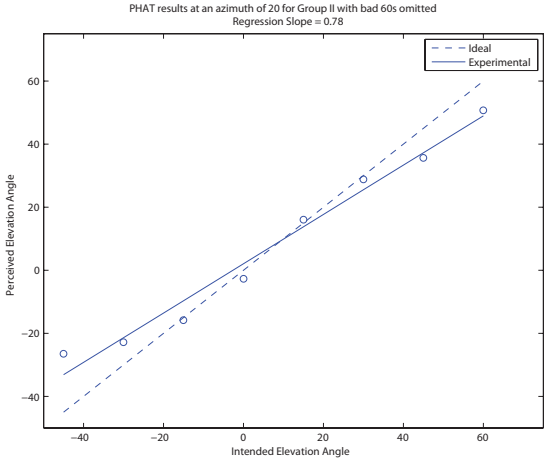

Figure 63. The subjective results for Group II at an azimuth of 20° for the PHAT model with the four explained outlier points at 60° removed.

In all of the aforementioned cases, the PHAT model outperformed the HAT model; this subjectively validates the hypothesis of this work, in most cases, for more than half of the subjects tested. The remaining cases, ones in which the model failed, belong to both azimuths tested in Group III and median plane locations in Group II.

Informal interviews with the subjects that ended up being categorized into Group III reveal that the subjects only heard timbral changes in the sounds and were unable to perceive any sense of elevation from them. While it is true that the timbre of a sound changes with elevation, due to the spectral envelope at every location being different, the subjects were asked to try to ignore such changes when attempting to localize the sounds. All of the subjects in Group III concluded that it was impossible to ignore the timbral changes. The subjects in this group are all musically trained which may have something to do with their perception of the filtered white noise that was used as a test stimulus. Instead of hearing the sound elevate they reported hearing emphases on different pitches/frequencies.

In much the same way that some people do not possess as good of a sense of vision as others, it is possible for some people to not be very proficient at localizing sounds. When using HRTFs in binaural synthesis, the intention is to fool the brain into thinking it is hearing sounds that originated in the physical world; however, if the subject possesses a poor ability to localize sounds in nature, then they will perform poorly on any synthetic HRTF localization tests. The subjects in Group III could be poor real-world localizers which would explain their poor performance in the listening test. A more thoroughly designed listening test would examine each subject's free-field localization prior to the binaural experiment.

Examining the obtained anthropometry of the subjects in Group III leads to another explanation for the poor performance of the model. All of the subjects in Group III have concha measurements that result in reflection distances of less than 10.2mm at elevations as low as 0°. It has already been proven that the model breaks down at such short reflection distances; this explains the sporadic results obtained from the subjects in Group III. This

could mean that the primary reflector for these subjects is somewhere beyond the concha for elevations above approximately -10°; however, without access to their recorded HRTFs, it cannot be known for certain.

An interesting anecdote worth mentioning involves a peculiar otoacoustic sensation that was reported by all three of the subjects that participated in the test on its third and final day. Two of these three subjects were members of Group III and the other is featured in Group II. These subjects had a difficulty verbalizing the experience but one person said that during the silent parts of the sounds it seemed as if there was still some amount of noise present in their ear canal. Another subject described it as noise that was created during the silent parts of the file. This sensation forced the subjects to take frequent breaks in the middle of the test. Short breaks were encouraged by the principle investigator even before this sensation was reported, but these subjects claimed that they absolutely had to take breaks in order to complete the test. They were encouraged to turn down the volume to the headphones which eased the sensation, but it did not completely eliminate it. The fact that the only subjects to experience this sensation all took the test on the same day leads one to believe that something in the test was flawed on that particular day, yet nothing was changed from the prior days of testing. This may or may not have had any effects on the results of the test but it is worth reporting since none of these subjects performed particularly well on the test.

Personal correspondences with Richard O. Duda at the CIPIC Laboratory reveal that median plane psychoacoustic validation has been quite problematic even with recorded HRTFs. With no definitive binaural cues (ITD and IID) existing in the median plane, the sole localization mechanism has always been figured to be spectral cues which are sometimes difficult for some listeners to resolve. This may be due to a lack of thorough training of the

subjects (perhaps using visual cues), the failure of room models to externalize sounds in the median plane, the failure to account for room acoustics in general, an absence of visual and environmental cues, not accounting for listener motion, an unfamiliarity with the source or a myriad of other potential contributing factors. Any of these reasons can potentially explain why the subjects in Group II were able to resolve elevation better for sounds off of the median plane even though the model is not as objectively accurate in such cases.

It is also believed that only hearing a short burst of filtered noise may not be enough to trigger all of the mechanisms that are used by the brain in the localization process, and this may cause variable results in tests that are designed around said stimulus. An interesting follow-up experiment to the work presented herein would test the subjects in Group I again using the same files to see if their results are consistent with those from the first round of testing. Due to time constraints such a test was not possible.

All disparities aside, this chapter proves that the hypothesis of this work was found to be true for more than half of the subjects examined in the psychoacoustic experiment that was outlined in the previous chapter. Examining potential causes of the cases in which the model failed leaves the door open for future work in the field.

7

CONCLUSIONS AND FUTURE WORK

While this thesis does provide good objective results and selectively decent subjective results, there is still much work to be done. To attempt to entirely solve the problem of synthesizing a complete set of HRTFs from anthropometry in such a short amount of time is unreasonable, but hopefully this work has contributed to those that wish to research the field in the future.

The first, and most obvious, aim of future work would be to extend the model to work for all frontal azimuths by accounting for the three-dimensional nature of the concha. A three-dimensional physical model of the concha can perhaps be used in the future to calculate the approximate reflection distance for azimuths greater than 30°. Accounting for diffraction within the concha may also improve the pinna model designed and implemented in this work.

Lack of convincing externalization effects, especially in the median plane, also seems to be an issue that prevents binaural synthesis algorithms from being totally effective. Conflicting opinions in the literature argue that post-HRTF equalization to account for the resonance of the ear canal may be a solution to the externalization problem but a definitive conclusion is yet to be reached. It makes sense that the ear canal resonance is stimulated during headphone listening, so introducing it twice would result in unnatural sounding results, but this is yet to be proven true or false.

A study of the effect of room acoustics on localization may also improve the quality of the algorithm introduced in this work. In a reverberant space, sounds will arrive at the

pinna not only at the initial elevation angle of the incident sound but from many additional angles. The brain may use the arrival angles of the reverberations as secondary cues to determine elevation. Modifying this algorithm to account for the additional arrival angles may improve the subjective results.

Studying the HRTFs of subjects with very small conchas, and generally narrow pinnae, such as those in Group III, to establish a link to the anatomical part that acts as the primary reflector at high elevations in such cases would improve the robustness of the pinna model and would be of great interest. The discrepancy may not be with the primary reflector but, instead, it may be with the fundamental designs of the model which is also of interest. Coming up with a more definitive method for placing the origin of the interaural coordinate system on the ear may also lead to the solution to this problem.

Establishing an anthropometric link to the gains of the resonances will definitely improve the performance of the model. This can be done by computationally constructing Shaw's physical pinna model and perturbing certain parameters to examine how the gains change with anthropometry and elevation.

Lastly, a C or C++ implementation of the algorithm introduced in this work is also of interest to ensure that it can in fact run in real time.

It is obvious from the discussion in this chapter, and this entire body of work, that many problems in the field of spatial hearing and binaural synthesis still remain unsolved. It is hoped that this work contributes, in some way, to the solving of these problems in the future.

REFERENCES

[1] V. R. Algazi, C. Avendano, and R. O. Duda, "Elevation localization and head-related transfer function analysis at low frequencies," *The Journal of the Acoustical Society of America*, vol. 109, pp. 1110, 2001.

[2] V. R. Algazi, C. Avendano, and R. O. Duda, "Estimation of a Spherical-Head Model from Anthropometry," *J. Audio Eng. Soc*, vol. 49, pp. 472-478, 2001.

[3] 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," *The Journal of the Acoustical Society of America*, vol. 112, pp. 2053, 2002.

[4] V. R. Algazi, R. O. Duda, R. P. Morrison, and D. M. Thompson, "Structural composition and decomposition of HRTFs," *Proc. IEEE WASPAA01, New Paltz, NY*, pp. 103-106, 2001.

[5] V. R. Algazi, R. O. Duda, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," *Proc. AES 113th Convention, Los Angeles, CA*, 2002.

[6] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 99-102, 2001.

[7] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *The Journal of the Acoustical Society of America*, vol. 88, pp. 159, 1990.

[8] C. Avendano, V. R. Algazi, and R. O. Duda, "A head-and-torso model for low-frequency binaural elevation effects," *Applications of Signal Processing to Audio and Acoustics, 1999 IEEE Workshop on*, pp. 179-182, 1999.

[9] C. Avendano, R. O. Duda, and V. R. Algazi, "Modeling the contralateral HRTF," *Proc. AES 16th International Conference on Spatial Sound Reproduction, Rovaniemi, Finland*, pp. 313-318, 1999.

[10] D. W. Batteau, "The Role of the Pinna in Human Localization," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 168, pp. 158-180, 1967.

[11] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*: MIT Press, 1997.

[12] C. P. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, pp. 476-488, 1998.

[13] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," *Journal of the Audio Engineering Society*, vol. 49, pp. 231–249, 2001.

[14] P. R. Cook, *Music, cognition, and computerized sound: an introduction to psychoacoustics*: MIT Press Cambridge, MA, USA, 1999.

[15] R. O. Duda, C. Avendano, and V. R. Algazi, "Adaptable ellipsoidal head model for the interaural time difference," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings*, vol. 2, pp. 965-968, 1999.

[16] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, pp. 3048, 1998.
[17] M. B. Gardner and R. S. Gardner, "Problem of localization in the median plane: effect of pinnae cavity occlusion," *The Journal of the Acoustical Society of America*, vol. 53, pp. 400, 1973.
[18] W. M. Hartmann, "How we localize sound," *Physics Today*, vol. 52, pp. 24-29, 1999.
[19] J. Hebrank and D. Wright, "Are two ears necessary for localization of sound sources on the median plane?," *The Journal of the Acoustical Society of America*, vol. 56, pp. 935, 1974.
[20] J. Hebrank and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," *The Journal of the Acoustical Society of America*, vol. 56, pp. 1829, 1974.
[21] J. Huopaniemi and M. Karjalainen, "Review of Digital Filter Design and Implementation Methods for 3-D Sound," *Proceedings of the 102nd Convention of the Audio Engineering Society, Preprint*, vol. 4461, 1997.
[22] Y. Kahana, P. A. Nelson, M. Petyt, and S. Choi, "Numerical modeling of the transfer functions of a dummy-head and of the external ear," *Proceedings of the AES 16th international conference, Spatial sound reproduction, Rovaniemi, Finland*, pp. 330–345, 1999.
[23] A. Kulkarni and H. S. Colburn, "Role of spectral detail in sound-source localization," *Nature*, vol. 396, pp. 747-749, 1998.
[24] J. C. Middlebrooks, and D.M. Green, "Sound Localization by Human Listeners," *Annual Review of Psychology*, vol. 42, pp. 135-139, 1991.
[25] P. Minnaar, J. Plogsties, and F. Christensen, "Directional resolution of head-related transfer functions required in binaural synthesis," *Journal of the Audio Engineering Society*, vol. 53, pp. 919-929, 2005.
[26] S. K. Mitra, *Digital Signal Processing: A Computer-based Approach*: McGraw-Hill, 1998.
[27] A. D. Musicant and R. A. Butler, "The influence of pinnae-based spectral cues on sound localization," *The Journal of the Acoustical Society of America*, vol. 75, pp. 1195, 1984.
[28] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the frequencies of the pinna spectral notches in measured head-related impulse responses," *The Journal of the Acoustical Society of America*, vol. 118, pp. 364-374, 2005.
[29] P. Satarzadeh, "A Study of Physical and Circuit Models of the Human Pinnae," in *Electrical and Computer Engineering*, vol. Master of Science. Davis, CA: University of California Davis, 2006.
[30] E. A. G. Shaw, "Acoustical features of the human external ear," *Binaural and Spatial Hearing in Real and Virtual Environments*, pp. 25-48, 1997.
[31] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *The Journal of the Acoustical Society of America*, vol. 94, pp. 111, 1993.
[32] D. Wright, J. H. Hebrank, and B. Wilson, "Pinna reflections as cues for localization," *The Journal of the Acoustical Society of America*, vol. 56, pp. 957, 1974.
[33] D. N. Zotkin, R. Duraiswami, L. S. Davis, A. Mohan, and V. Raykar, "Virtual audio system customization using visual matching of ear parameters," *Proc. IEEE ICPR*, pp. 1003-1006, 2002.

APPENDIX A

Listening Test Instructions -- Subject *xx* - Name

INTRODUCTION

The purpose of this listening test is to compare two different HRTF synthesis models. The test consists of 35-50 audio files with durations ranging from 4 to 22 seconds. Each file was created by filtering white noise with the left and right HRTFs generated by one of the two models. White noise was chosen as the stimulus because wideband sounds are easiest for the human auditory system to localize. Listening to white noise is far from exciting, so if it ever gets exhausting, boring or redundant, feel free to take a break.

The main variable being tested is that of elevation. Since the synthesized sounds lack a great amount of externalization, most of the sound files will sound as if they are coming from directly in front of your face, from the top of your head, or from slightly inside of your head which makes it somewhat difficult to perceive elevation. The test is designed with this ambiguity in mind.

PROCEDURE

Please adjust the height of your chair so that you can see your eyes in the mirror when looking straight ahead. Also, please make certain that the chair is lined up with the marks on the floor and that your shoulders are parallel to the wall. This should put your face approximately .5m from the mirror, and it ensures that the angles marked on the wall are accurate. You may place the computer on your lap if that is easiest and most comfortable. When putting on the headphones, be sure to put the right and left phones are on the correct ears.

All of the files for this test are loaded into iTunes and there is a playlist for each subject. Please see the top of this sheet to find your subject number and select that playlist. iTunes is set on repeat so that it will continuously loop each file until you are finished with it. When a sound is playing, it is necessary that your eyes remain visible to you in the mirror at all times. Also, when listening, make sure the volume is at a high level--it will make the localization process easier.

The first two files in the playlist consist of five concatenated sounds with each subsequent one being at a higher elevation. There will be three sequences of .5 seconds of noise and .5 seconds of silence followed by one final 1.5 second noise burst for each elevation. The sound for the next elevation will fade into the sound from the previous elevation. For clarity, the file structure for each elevation is shown below:

.5s of noise - .5s of silence - .5s of noise - .5s of silence - .5s of noise - .5s of silence - 1.5 seconds of noise.

The elevation starts out at 0° and increases uniformly to 60°. Your task is to identify which model (HAT or PHAT) creates a better sense of elevation. The next two files do the same thing but with four decreasing elevations starting at 0°. Once again your task is to identify which model creates a better sense of elevation. Please circle your answers on the attached answer sheet. Note that after the last elevation is played in each of these files that the playback will start again at 0° because iTunes is set to repeat each file.

The remaining sounds are of the aforementioned file structure, but they only focus on one elevation location at a time. Your task with these sounds is to identify their perceived elevations. While a sound is playing, place the laser pointer alongside one of your ears and aim it in the vertical direction from which you perceive the sound. When listening to a sound, it is crucial that you remain facing straight ahead so that your eyes are always

visible in the mirror; otherwise, the angles marked on the wall and the location of the played sound will not be accurate.

The lateral direction of the sound is not important, and it is recommended that you aim the pointer to either side of the mirror parallel to the axis that runs up the wall. This will prevent you from potentially blinding yourself by pointing the laser at the mirror if the sound is perceived as being straight ahead. If your arm gets tired from holding the pointer alongside your ear, you can either switch arms or take a break. (For some sounds, which may be perceived as off-center to the right, it is best that you have the pointer on the right side of your head). Once you feel you have identified the sound's location, look up or down at the wall without moving the laser pointer and make note of the elevation angle on the wall where the laser is pointing. Rounding to the nearest 5° is acceptable. The attached answer sheet contains a table; please write the file name in the left column and the perceived elevation in the right column.

No answers are final until the test is complete, so you are able to go back to any previously listened sound and change your answers if you'd like. You can also change the order of the files in the playlist as long as you make sure you write the correct filename on the answer sheet. You can also turn off repeat if you'd like to hear one sound go right into the next sound. It is sometimes easier to get an accurate sense of elevation by comparing the locations of two sounds one right after the other. Another hint that might make localization easier is to listen to the sounds and aim the laser pointer with your eyes closed (as long as your head remains facing directly forward) and then open them to see the angle at which you pointed.

Thank you for your time and effort.

Answer Sheet -- Subject *xx* - Name

Increasing Elevation (circle one)

PHAT HAT

Decreasing Elevation (circle one)

PHAT HAT

Filename	Perceived Elevation