THREE-DIMENSIONAL LOCALIZATION OF A CLOSE-RANGE ACOUSTIC SOURCE USING BINAURAL CUES

by

PAUL THOMAS CALAMIA, B.S.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

May 1998

THREE-DIMENSIONAL LOCALIZATION OF A CLOSE-RANGE ACOUSTIC SOURCE USING BINAURAL CUES

APPROVED BY
SUPERVISORY COMMITTEE:

Supervisor: _____

_____

In loving memory of Lois Calamia.

Acknowledgments

The work required for this thesis would not have been possible without the help of a large number of people. Many thanks must be offered.

First and foremost, I would like to thank Dr. Elmer Hixson for taking me on as a student, allowing me the freedom to work on a project completely of my own choosing, and providing a tremendous amount of support, equipment, advice, and time. I would also like to thank Dr. Dennis McFadden for taking the time to read and critique my thesis.

Because the work for this thesis was entirely unfunded, much of the equipment used during the measurement phase was borrowed. I am indebted to Dr. Craig Champlin of the Speech and Hearing Center for the use of the KEMAR and all the associated parts, and to Robert Hofkamp and the others at TGI, Ltd. who arranged for the loan of the Tannoy near-field monitors. Thanks also must go to Dr. William Gardner at MIT who provided the code used to generate the MLS, Charley Randall for the DAQ card, and Perry Durkee and Marty Ringuette for their help in designing and building the speaker cradle.

For somewhat less tangible things, I would also like to thank Dr. Richard Duda of San Jose State University who gave me the idea for the project, and Dr. Russell Pinkston who showed great interest in my work and introduced me to the fascinating field of computer music.

A select few people deserve thanks that are infinitely harder to express. Tom Kite and Eric Rosenberg have come to be two of my closest friends, and both have given me tremendous amounts of support (academic, emotional, and otherwise), as well as fond memories of some truly absurd escapades and social commentary. Evan Bronstein, Lon Goldstein, Scott Romanoff, Rob Stein, and Eric Streisand have always been there for me to lean on, my three years in Austin being no exception.

I don't have the words to adequately express my love and gratitude to Kimberley Jordan. No one has ever had such an impact on me. No one has ever given me so much time, so much love, so much support, so much motivation. I could not have finished this thesis without her.

None of this would have been possible without the love and support that my family has given me. The gratitude and love I feel for them is only overshadowed by my shock and sadness that my mother did not live to see this work come to fruition.

Thank you all.

May 8, 1998

THREE-DIMENSIONAL LOCALIZATION OF A CLOSE-RANGE ACOUSTIC SOURCE USING BINAURAL CUES

PAUL THOMAS CALAMIA, M.S.E.
The University of Texas at Austin, 1998

Supervisor: ELMER L. HIXSON

Sound localization is the process through which a listener is able to discern the apparent location of an acoustic source without the aid of sight. Auditory cues to source position fall into two major categories: monaural and binaural. The former are extracted from the information present at a single ear, while the latter result from some form of neural comparison or correlation between the signals at both ears. The focus here is on binaural cues, specifically the interaural time and level differences (ITD and ILD), and whether they contain sufficient information to accurately localize a close-range acoustic source in three dimensions. A model has been developed, based on work published by others, which compares ITDs and ILDs generated by a source in an unknown location with those generated by sources in known locations. The binaural cues used in the model were extracted from close-range head-related transfer functions (HRTFs) measured on a KEMAR manikin. A nearest-neighbor approach is used within the model to estimate the azimuth, elevation, and range of the unknown source.

Table of Contents

Chapter 1

Introduction

Today, the auditory system is primarily relied upon for its use in communication. People relay information to one another through speech, music, and other sounds which are received by the ears and processed by the brain. But hearing, and in particular the ability to identify the location of the origin of a sound, was once of significant concern for survival. Despite the changes brought about by evolution, the same mechanisms needed in the past for eluding a predator or locating prey are now used to locate an on-coming vehicle with a siren, and provide the feeling of being enveloped by the sounds of an orchestra in a concert hall. Spatial hearing may no longer aid the average person in matters of life and death, but it remains the subject of much academic interest. With the advent of virtual reality and the demand for realistic three-dimensional audio through headphones, its role as the subject of continued research seems quite safe.

This chapter serves as an introduction to the work done for this thesis, and a brief summary of pertinent background information. Section 1.1 covers the basics of sound localization. Section 1.2 describes the head-related transfer function and its use in sound localization research. Section 1.3 contains

the statement of purpose for this work, and a brief outline of the remaining chapters.

1.1 Sound Localization

Sound localization is the process through which a listener perceives the apparent spatial position of an acoustic source. The auditory system is provided with a variety of cues which lead to this perception in ways not entirely understood. Under certain circumstances, a human listener can differentiate between sound sources that are in front and behind, above and below, to the right and to the left. With varying degrees of accuracy, a listener can often estimate the particular location of a source in space.

The study of sound localization must consider at least two major issues: the nature of possible localization cues, and the way in which these cues are processed and interpreted by the auditory system and the brain. To this end, researchers have measured and analyzed the acoustic information available to a listener, studied the anatomy of the auditory system, quantified and qualified human (and animal) localization abilities, and modeled these abilities with mathematical formulae and computer programs. Blauert [1] offers a comprehensive discussion of these and many other facets of sound localization. Middlebrooks and Green [29] and Wightman and Kistler [42] provide excellent reviews of the pertinent literature and summaries of auditory localization cues and cue utilization.

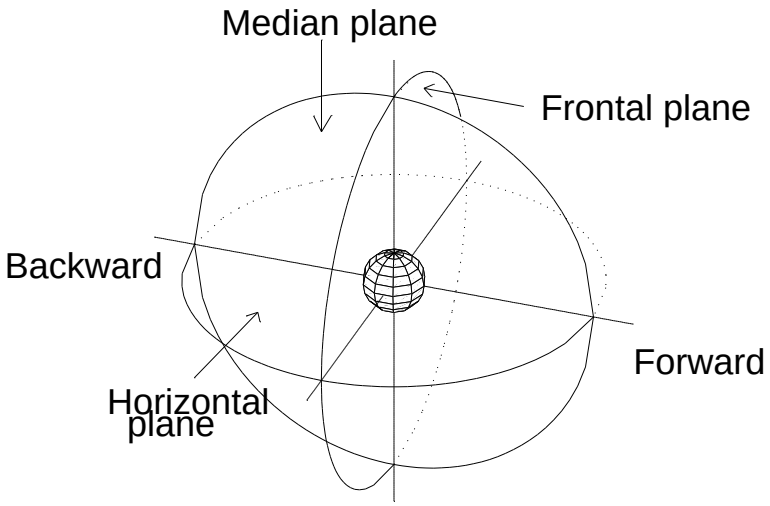


Figure 1.1: Three normal planes which define the space around a listener relative to the center of the interaural axis. The sphere represents the listener's head. Adapted from [15].

1.1.1 Auditory Space

Three normal planes are used to describe the auditory space surrounding a listener. All are referenced to the interaural axis, an imaginary line joining the two ears. The median plane is a vertical plane through the center of the interaural axis; the frontal plane is the vertical plane which contains the interaural axis; the horizontal plane is the horizontal plane which contains the interaural axis. The three planes are depicted in Figure 1.1. In the context of sound localization, positions in auditory space are described using a three-component, head-centered coordinate system. Two of the components, azimuth and elevation, differ in their exact definitions depending on the coordinate system used, although the former usually describes a displacement from the median plane, and the latter a displacement from the horizontal plane. Range, the third component, is always defined as the linear distance from the center of

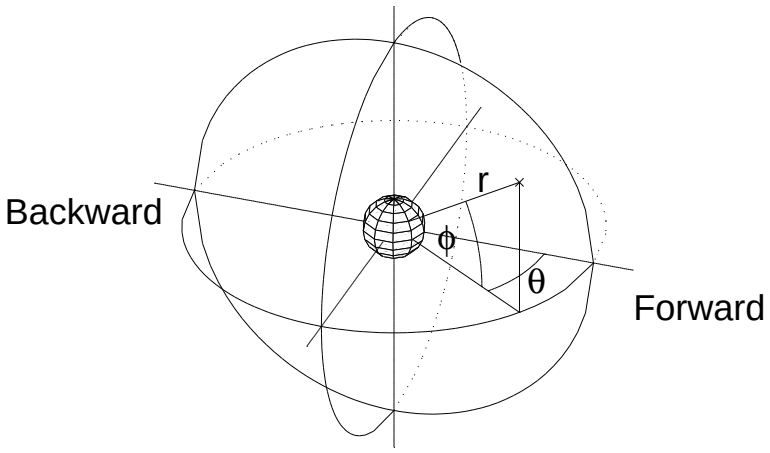


Figure 1.2: Coordinate system with azimuth θ, elevation ϕ, and range r. The source position is marked with an x. See also Figure 1.3.

the listener's head to the sound source in question.

The coordinate system we chose for this thesis is depicted in Figures 1.2 and 1.3. The azimuth angle θ is the angular displacement from the median plane (as measured in the horizontal plane), with positive angles to the right and negative angles to the left. The elevation angle ϕ measures the vertical displacement from the horizontal plane, with positive angles above and negative angles below. Range r is the distance from the center of the interaural axis to the source. This system is particularly useful for positioning a source during localization-related measurements. With a source fixed at a distance r and elevation ϕ, a listener or other receiver (e.g., a microphone) can be rotated horizontally through $360\,^{\circ}$ to sample the desired azimuths. This system is also convenient because the definitions of the angular components are rather intuitive. If one considers a sphere of radius r centered on the midpoint of the interaural axis of a listener, any points on the sphere with the same azimuth fall on the same longitude line, and any points with the same

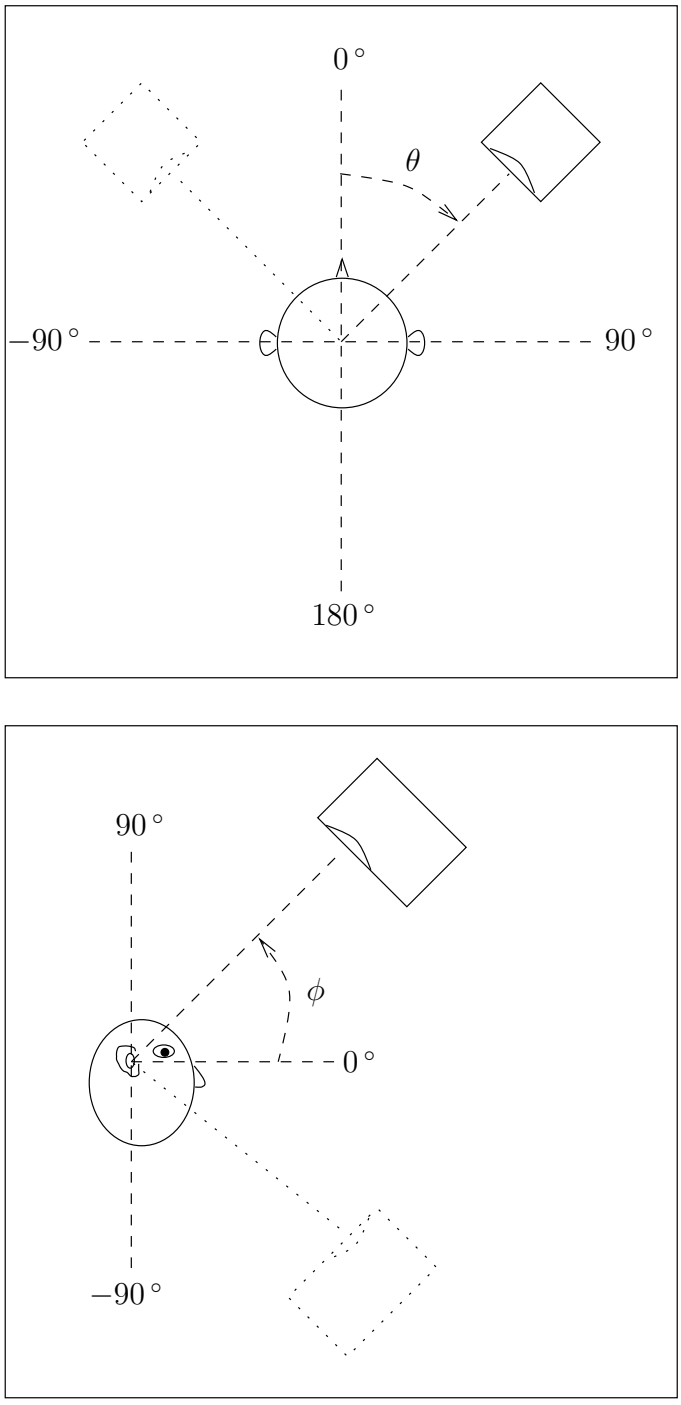


Figure 1.3: Azimuth θ as seen from above, and elevation ϕ as seen from the side. See also Figure 1.2.

elevation fall on the same latitude line. Examples of other coordinate systems can be found in [5] and [15], some of which are considered more appropriate for the analysis of localization cues. Fortunately, coordinate transformations exist which provide a one-to-one mapping of position from one coordinate system to another, so it is possible to exploit the features of multiple systems without loss of location-specific information.

1.1.2 Cues to Source Location

Auditory cues to source location can be classified in numerous ways. In particular, three somewhat overlapping distinctions are usually made: whether a cue is monaural or binaural, whether it is relative or absolute, and whether it is exploited in the time or frequency domain [28, 42]. Monaural cues are extracted from the information present at a single ear, while binaural cues result from some form of neural comparison or correlation between the signals at both ears. Relative cues arise from the comparison of the signals from a source to those from a known condition (e.g., source spectrum or location), and thus require some *a priori* knowledge of the spectrum and/or location of the source. Absolute localization cues are independent of the source spectrum, and can be used without any previous information. Time-domain cues contain salient temporal information, while frequency-domain cues contain location-specific spectral information. This thesis is primarily concerned with absolute, binaural cues, processed in both domains. Both Middlebrooks and Green [29] and Wightman and Kistler [42] offer excellent reviews of the various cues.

The use of binaural cues by the auditory system has received much attention in the literature (e.g., see [2], [4], [9], [11], [17], [21], [36]). In par-

ticular, the interaural time difference (ITD) and the interaural level difference (ILD) have been the subject of much research, as they are considered to be the two most important binaural cues to source location [42]. The ITD is the difference in arrival time at the two ears of the signal from a single source; it is primarily due to the difference in path lengths from the source to the two ears imposed by the head. The ILD is the difference in sound pressure level at the two ears. When a sound source is relatively far from the head ($r \geq 1$ m), the path length difference from the source to the two ears is small when compared with the distance from the source to the listener's head. Reduction in sound pressure amplitude due to spreading between the two ears is negligible, and the ILD results from diffraction about the head. An acoustic shadow is cast over the far ear, reducing the sound pressure level there and thereby causing the ILD. When the source is within 1 m of the head, however, the path length difference is no longer negligible, and the ILD is augmented by the additional attenuation incurred along the path to the ear farther from the source. Lord Rayleigh provided the first theory of sound localization based on the interaural time and intensity differences [36]. His "duplex" theory predicts that the ITD can provide an unambiguous localization cue for low-frequency sources, and the ILD is most useful for localizing high frequencies. Modern research has shown this theory to be incomplete, but it continues to serve as a cornerstone in the study of sound localization.

Of the studies of binaural localization, those which attempt to provide indications of azimuth and/or elevation are the most common; distance localization has received far less attention. Human localization is least accurate for distance [29], and the primary cues to distance tend to be relative and

monaural. Given a source far from the head, it is thought that its range is perceived by the ratio of direct-to-reverberant energy received by the ears, a quantity which decreases with distance; the increase in attenuation of high frequencies by air with increasing distance; and the decrease in sound pressure level with increasing distance due to spreading. If a listener is given only one presentation of a source of unknown spectral content, none of these cues is of much use. A comprehensive review of distance cues can be found in [44].

Of particular interest for this thesis are the absolute cues to the distance of sources located within one meter of the head, which we refer to as close-range sources.[1] At such small distances, reverberation can be ignored if the time scale of analysis is sufficiently short, and high-frequency attenuation by the air is negligible. Changes in sound pressure level with distance are still available as cues, but only with multiple presentations. The most likely distance cue is thus the interaural level difference. As is described above, the ILD which results from a close-range source is due to an acoustic shadow cast over the far ear by the head, as well as the attenuation of the signal travelling to the far ear due to spreading over the larger path length. The distance dependence of the ILD at close range has been modeled mathematically [4, 10], but its salience as a cue is yet to be determined.

[1]Sources within 1 meter of the head are often referred to as "near-field" sources in the literature. This is actually a misuse of the term near-field, which generally describes the area within one wavelength of an acoustic source [32].

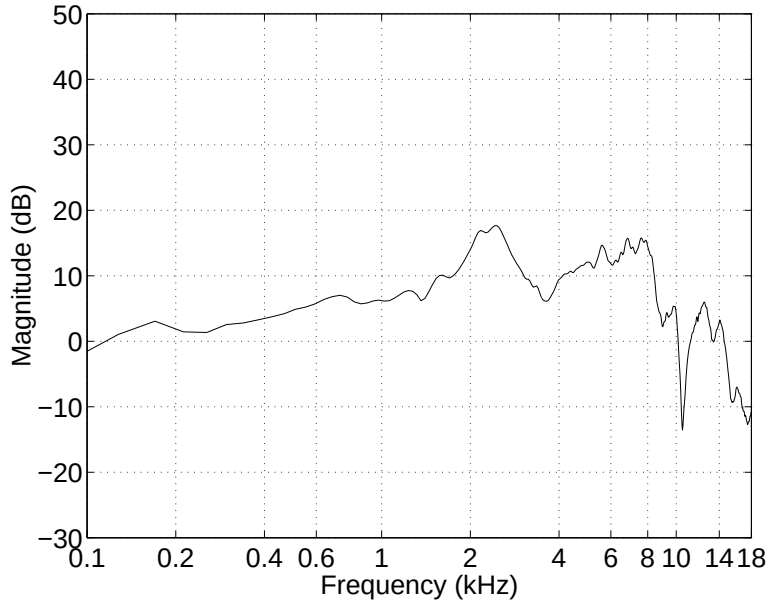


Figure 1.4: A sample HRTF measured on a manikin. Notice the resonance peaks near 2.5 and 7 kHz, and the pinna notch near 10 kHz.

1.2 Head-Related Transfer Functions

Many of the cues used for sound localization are embodied in the head-related transfer function (HRTF). The HRTF has been defined in a number of ways (see [3]), but in general is a frequency-domain function which describes the filtering effects of the head and outer ear on acoustic signals. For this thesis, we use the following definition: the head-related transfer function is the ratio of the sound pressure measured at the ear drum divided by the free-field sound pressure measured at the center of the head with the head removed. Reflections, diffraction, and resonances that occur within the outer ear, as well as reflections from the shoulders and torso and diffraction about the head, cause changes in the spectrum of incoming sound which are both location and frequency dependent. An example of an HRTF can be seen in Figure 1.4. The

peaks near 2.5 and 7 kHz are due to resonances within the outer ear. The notch near 10 kHz is the result of destructive interference from reflections within the outer ear [34].[2] Binaural localization cues, particularly interaural time and level differences, can be extracted from the ratio of the right and left HRTFs for a specific source position. This ratio, known as the interaural transfer function, and its use in localization are described in detail in Chapter 3.

The HRTF varies from individual to individual, and has been measured on human listeners as well as anthropomorphic manikins designed to have average human dimensions. Because the HRTF is a function of azimuth, elevation, range, *and* frequency, comprehensive sets of measurements are difficult to make. Measurements described in the literature usually contain data evenly sampled over a head-centered sphere of fixed radius, thus eliminating the possibility of analyzing the range dependence. However, the HRTF's dependence on range drops off as distance increases, so measurements made sufficiently far from the head are considered to be representative of all distances beyond. As a result, close-range HRTFs have been particularly neglected. We hope to add to the small body of research in this area with this thesis.

1.3 Thesis Outline

The ultimate goal of this thesis is to create a computer-based sound localization model which can estimate the azimuth, elevation, and range of a close-range

[2]This notch, known as the "pinna notch," is considered to be one of the primary cues to source elevation.

acoustic source, given as input binaural cues similar to those available to the human auditory system. In order to obtain this goal, it is first necessary to understand the nature of the auditory cues to each component of the source position, particularly those that occur when the source is within one meter of the head. To this end, an extensive set of close-range head-related transfer functions has been collected, and the direction- and distance-dependent cues within them have been analyzed and extracted for use in the localization model. Chapter 2 covers the HRTF measurements. Chapter 3 discusses the extraction of localization cues from the HRTFs and the analysis of these cues. Chapter 4 describes a three-dimensional localization model. Chapter 5 summarizes the results of the measurements, analysis, and modeling.

Chapter 2

Measuring the Head-Related Transfer Function at Close Range

As is mentioned in Chapter 1, very little data have been collected on the HRTF at small distances from the head. For the purpose of this thesis, a set of close-range HRTF measurements was taken, upon which the majority of the analysis in Chapter 3 is based. This chapter serves as an overview of the measurement setup. Section 2.1 covers the various components used in the measurements. Section 2.2 provides a brief introduction to the use of maximum-length sequences (MLS) in transfer-function measurements. Section 2.3 discusses the actual measurement process. Section 2.4 covers the system calibration. Section 2.5 describes the spatial locations covered by the measurements. Section 2.6 covers the free-field, system-response measurements.

2.1 HRTF Measurement Setup

2.1.1 Measurement Overview

The general methodology for acquiring the HRTF measurements is described by Gardner and Martin [12]. This methodology was developed for far-field HRTFs, and thus had to be adapted for measurements taken at close-range. What follows is a description of the measurement process, with particular

emphasis on the aspects which are specific to close-range measurements. The procedure involves using a maximum-length sequence to create broadband noise through a loudspeaker, recording the response to the noise at the ear of a manikin, and manipulating the response into a head-related impulse response (HRIR). The HRIR can be converted to its frequency domain equivalent, the HRTF, using a Fourier transform.

2.1.2 Hardware

All HRTFs were measured for a Knowles Electronics Manikin for Acoustic Research (KEMAR) model DB-4004, an anthropomorphic manikin designed to facilitate binaural recordings. The KEMAR consists of an adult-sized hollow torso and a hollow head with removable pinnae. The inside of the head is designed to hold an inner ear simulator and a microphone on each side. For this thesis, measurements were taken on the right side only (see Section 2.6), using a Brüel and Kjær (B&K) 4133 $\frac{1}{2}-$inch microphone, a Knowles Occluded-Ear Simulator, model DB-4005, and a Knowles Right Pinna, model DB-056 (see Figure 2.1). The microphone was attached to a B&K 2669C pre-amplifier, and powered with a B&K 2807 power supply. The KEMAR was mounted on a turntable capable of full 360 ° rotation. The angular position of the turntable (and the KEMAR) could be monitored through the output of a sine-cosine potentiometer. Figure 2.2 shows the KEMAR mounted on the turntable. All measurements were made in an anechoic chamber with inner dimensions: 2.17 m (height) $\times$ 2.47 m (width) $\times$ 3.64 m (depth).

The choice of an appropriate loudspeaker to be used as an acoustic source was limited by the close range of the measurements and the need to

Figure 2.1: A rear view of the inside of the KEMAR's head, fitted with a B&K $\frac{1}{2}$-inch microphone and a Knowles Occluded-Ear Simulator at the right ear.

Figure 2.2: The KEMAR mounted on the turntable.

reproduce an audio bandwidth of approximately 20 Hz to 20 kHz. Consider the geometry of a standard two-way loudspeaker, in particular the physical separation of the woofer and the tweeter. At listening or measuring positions far from the speaker, the separation of the woofer and tweeter is negligible, and the acoustic source can be considered to be at the geometric center of the two drivers without introducing gross error. At shorter measuring distances, however, the separation of the high- and low-frequency sources is no longer negligible (see Figure 2.3), and the uncertainty of the elevation of the source becomes significant. Thus, in order to maintain accuracy in positioning the acoustic source relative to the KEMAR, it was necessary to choose a loudspeaker which could reproduce acoustic information across the full audio bandwidth from a single location in space.

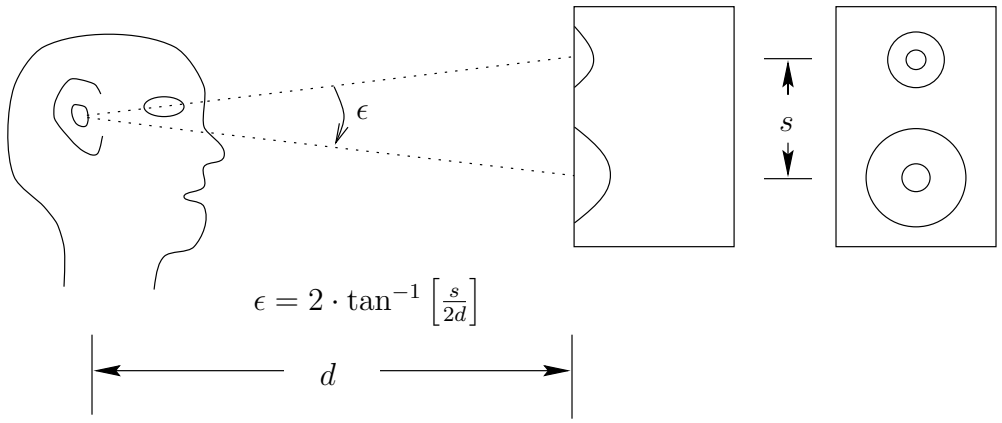


Figure 2.3: A conventional two-way loudspeaker. The separation of the woofer and the tweeter introduces an error ϵ when measuring the elevation angle.

The loudspeaker chosen for these measurements was a Tannoy System 600 Near-Field Reference Monitor. This particular speaker has a published frequency response of ± 3 dB from 52 Hz to 20 kHz. It uses a Dual Concentric ^{TM} design which allows for accurate point-source reproduction at close range over the desired bandwidth by placing a high-frequency waveguide (tweeter) at the center of the woofer (see Figure 2.4).[1] The full set of specifications can be found in the user's manual [39].

Control of the measurement system was supplied a by Packard Bell 75 MHz Pentium computer. Output to the loudspeaker and input from the microphone and the potentiometer on the turntable were accomplished with a National Instruments AT-MIO-16E-2 data acquisition (DAQ) board. The analog input of this board is capable of 12-bit resolution at a maximum sampling frequency of 500 kHz; the analog output also has 12-bit resolution, at a maximum sampling frequency of 600 kHz. For this experiment, the sampling

[1] Dual Concentric is a trademark of TGI, Ltd.

Figure 2.4: The Tannoy loudspeaker. Note the concentric woofer and tweeter.

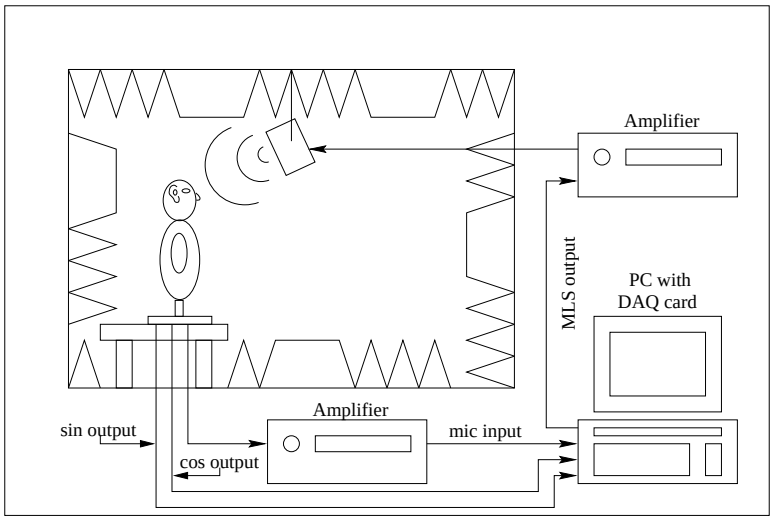


Figure 2.5: HRTF measurement setup.

frequency at both the input and the output was $f_s = 43478.26$ Hz.[2] Three single-ended input channels were used: one connected to the microphone and one connected to each output of the sine-cosine potentiometer. One single-ended output channel was connected to the loudspeaker. The DAQ board was controlled by a program written in C (see Section 2.1.3).

A Tigersaurus 210/A power amplifier was used to amplify the output of the DAQ board to drive the loudspeaker. The gain of the Tigersaurus was held constant throughout all of the measurements to maintain a fixed source level. A Hewlett-Packard 450A pre-amplifier was inserted between the microphone and the DAQ board to add 20 dB of gain before sampling. The entire measurement setup is depicted in Figure 2.5.

[2]A sampling frequency of 44.1 kHz would have been ideal, as it is a standard in digital audio technology, and thus would allow use of the measurements in existing audio systems without resampling. However, the AT-MIO-16E-2 is limited to sampling periods (T_s) of integral microsecond length, so $T_s = 23$ μs, corresponding to $f_s = 43478.26$ Hz, was chosen.

2.1.3 Software

Two types of software were used during the experiment. The first was a set of C-language functions which generate maximum-length sequences (MLS), provided by Dr. William Gardner at the Massachusetts Institute of Technology. These were used to create a 16383-point MLS using a 14-bit shift register. See Sections 2.2 and 3.1 for details on the use of maximum-length sequences in impulse-response measurements.

The second type of software was used to control the data acquisition during the experiment; it was written in C using National Instruments NI-DAQ libraries (version 4.6.1). These libraries allow for direct control of data acquisition hardware from a DOS programming environment.[3] In particular, the DAQ capabilities of the program include controlling simultaneous analog output and input for sending the MLS to the speaker while sampling the microphone's response, and simultaneous two-channel analog input for measuring the angular position of the turntable. The program also generates the appropriate data file name from the azimuth, elevation, and distance associated with a particular measurement, and writes the data to disk for subsequent processing.

All data processing during the analysis phase was done in Matlab. The details of the analysis are given in Chapter 3.

[3] All data acquisition was done in a DOS environment because latencies associated with Windows reduced the maximum sampling rate of the DAQ board to well below the required rate.

2.2 Impulse-Response Measurements Using Maximum-Length Sequences

Since the HRTF can be recovered from the head-related impulse response (HRIR) using a Fourier transform, it is possible to measure the HRTF indirectly using any one of a variety of impulse-response measurement techniques. We chose to use the method described by Gardner and Martin [12], which uses a maximum-length sequence (MLS). An MLS is a pseudo-random, periodic, binary sequence which is generated recursively in a shift register using exclusive-or logic in a feedback loop. Ignoring the DC component, its spectrum is flat up to $f_s/2$, where f_s is the frequency at which it is sampled during digital-to-analog conversion. Thus, when used to drive a loudspeaker, it generates broad-band noise whose bandwidth is controlled by the sampling frequency.

Impulse-response measurements are made by exciting a system with an MLS and measuring the response to the excitation. The impulse response of the system (in this case, the HRIR) is found by cross-correlating the measured system response with the original MLS. The Fourier transform of the impulse response is the frequency response of the system (here the HRTF). Detailed discussions of the use of maximum-length sequences for impulse-response and transfer-function measurements can be found in Rife [40] and Rife and Vanderkooy [37]. Golomb [13] provides an in-depth mathematical treatment of the theory of shift register sequences. The exact procedure used to find the HRIR and HRTF using the KEMAR's measured response to the MLS is described in Section 3.1.

2.3 Measurement Process

2.3.1 Speaker Placement

The loudspeaker was mounted in a cradle and suspended by a cable from a hook in the ceiling of the anechoic chamber. The length of the cable could be adjusted to place the speaker at any desired height. Guide wires attached to the speaker cradle were used to set the tilt of the speaker and control its rotation. For each set of measurements, a multi-step process ensured accurate positioning of the loudspeaker with respect to the KEMAR for both distance and elevation. For a given distance d and elevation angle ϕ, the following procedure was used (refer to Figure 2.6):

1. The speaker was suspended at the correct height, which was found by adding the vertical components h_1, h_2, and h_3;

2. The forward (or rearward) tilt of the speaker was adjusted with a magnetic angle locator to equal the elevation angle ϕ, and was secured with a guide wire attached to the bottom (or top) of the cradle;

3. The KEMAR was rotated to face away from the speaker (azimuth = 180 °), and the top of the head was removed to expose the interaural axis;

4. The position of the KEMAR was adjusted so that the center of the interaural axis was a distance d from the center of the loudspeaker cone, as measured using a pre-cut wooden dowel of the desired length.

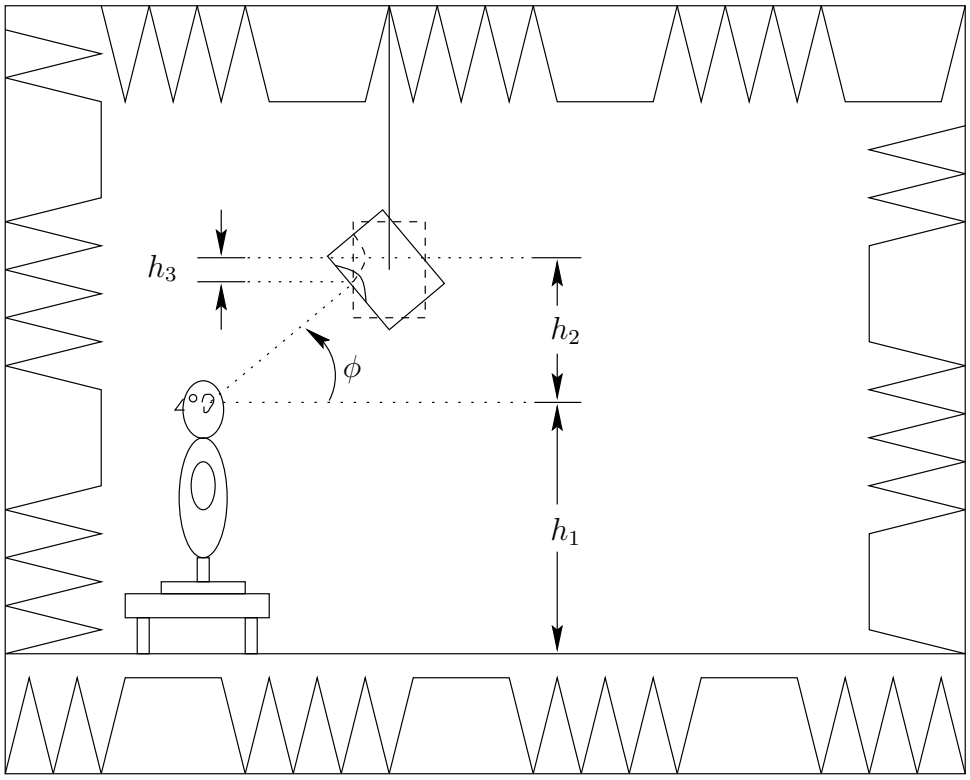


Figure 2.6: Three vertical components of the loudspeaker position. Note that: $h_3 < 0$ when $\phi > 0$, and $h_3 > 0$ when $\phi < 0$; h_1 is constant; h_2 is a function of d and ϕ.

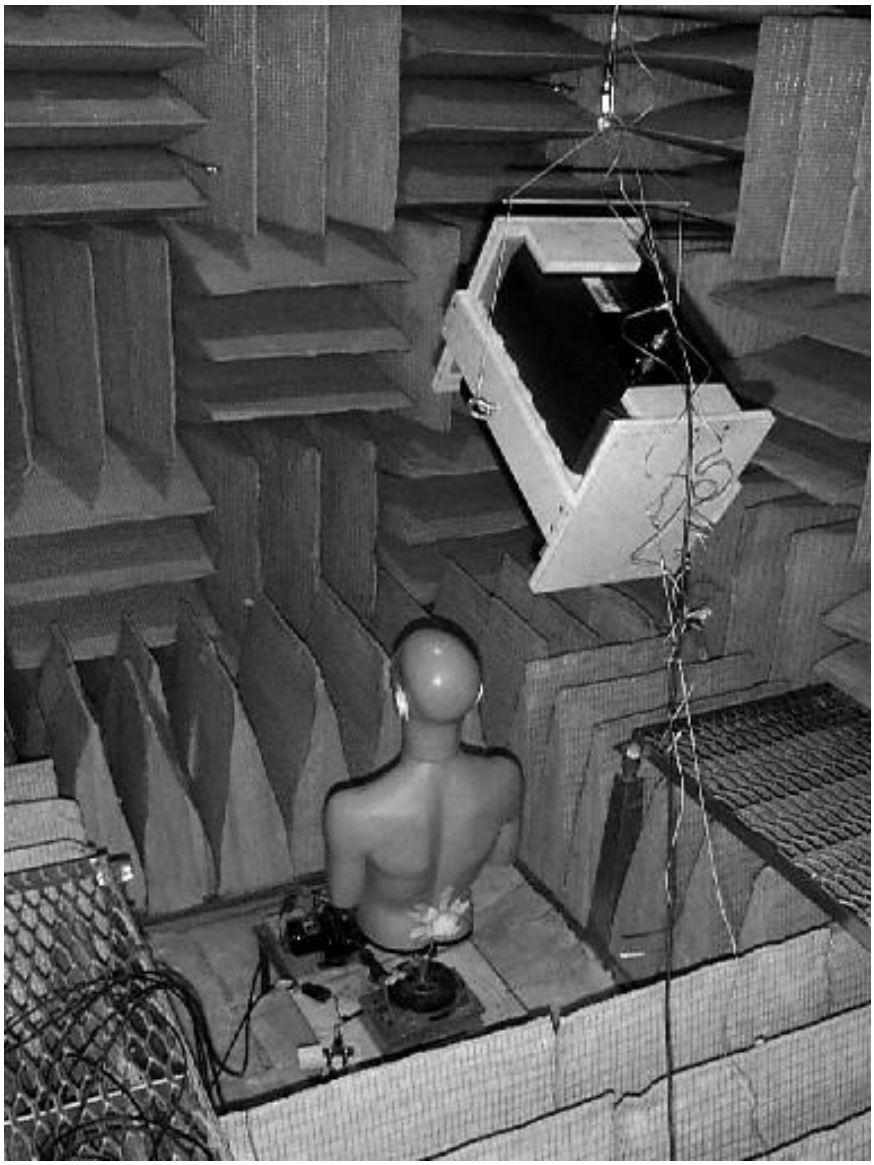


Figure 2.7: Measurement setup in the anechoic chamber.

At this point the top of the head was replaced, and measurements were taken as the KEMAR was rotated through a full 360°. A photograph of the setup can be seen in Figure 2.7.

2.3.2 Data Collection

Once the speaker and the KEMAR were properly located, the anechoic chamber was sealed, and the data collection software was started. The program first queried the user for the elevation and distance of the current setup, both of which were constant throughout a given set of measurements and were used in naming the data files. For each measurement position within the set, the following procedure was used. The controlling program first calculated the azimuth of the loudspeaker relative to the KEMAR by measuring the voltage from the sine-cosine potentiometer on the turntable. It then began simultaneous input and output operations: two copies of the 16383-sample MLS were sent through the DAQ card to the loudspeaker, while the pressure response at the microphone was sampled. In all, 33500 samples were recorded and written to the hard drive of the computer for each position.[4] After each measurement was made, the turntable was rotated to the next position, and the process was repeated.

[4]The acoustic delay between the speaker and the microphone, as well as the internal signal delay of the DAQ board cause a time lag between the output of the sequence from the DAQ board and the arrival of the response signal from the microphone. A 33500 sample buffer ensured capture of the entire response to the MLS despite these delays.

2.4 Calibration

2.4.1 Data Acquisition

The microphone was calibrated with a B&K $\frac{1}{2}$-inch calibrator, Model 4231, which delivers a 94 dB SPL, 1 kHz tone. The microphone response to the calibrator was found to be within 0.5 dB of 94.0 dB before and after each measurement set. The entire data-acquisition system (computer, DAQ card, amplifiers, speaker, and microphone) was tested using a Hewlett-Packard digitizing oscilloscope and a Stanford Research Systems 2-channel network signal analyzer. The sampling period was verified to be 23 μs with the oscilloscope by monitoring the length of a single pulse within the MLS at the output of the DAQ board. The spectrum of the MLS signal was tested at four points within the system: before the digital-to-analog converter (DAC) in the DAQ board; between the DAC and the loudspeaker; between the microphone and the analog-to-digital converter (ADC) in the DAQ board; and after the ADC. The spectrum was of course altered by the non-flat frequency responses of the microphone and speaker, and showed the expected aperture loss from the zero-order hold operation of the DAC. These deviations from a flat response are repeatable and correctable (see Section 2.6). The changes in the responses of the data acquisition components were determined to be negligible over the course of the measurements.

2.4.2 Turntable Calibration

Calibration of the turntable and the angle measurement portion of the data-acquisition software was accomplished as follows. With the KEMAR removed, the turntable was rotated through a full 360 °, in 5 ° increments. At each

stopping point, the angular reading from the position monitoring/controlling software was compared with a reading from a protractor on the base of the table. All readings from the software were within one degree of those from the protractor. Before each set of measurements, the KEMAR was rotated until the software displayed an azimuth of 0°, the alignment with 0° on the table was confirmed, and the 0° mark on the table was visually aligned with the 0° axis of the loudspeaker. The actual azimuth of the speaker relative to the KEMAR is believed to have been within one degree of the desired azimuth.

2.4.3 Temperature, Humidity, and Sound Speed

The temperature and relative humidity in the anechoic chamber were measured before and after all measurements each day. The speed of sound (c) was calculated using the formula found in [8]. We found c to be within 0.5 m/s of 346 m/s for all measurement periods.

2.5 Measurement Scope

In order to analyze the changes of both monaural and binaural localization cues with range, measurements had to be taken around the KEMAR at a variety of distances. Since the focus of this thesis is close-range localization, we chose distances of 0.25 m, 0.50 m, 0.75 m, and 1.00 m. We originally hoped to include 0.125 m as well, but at this distance the rotation of the KEMAR was impeded by the loudspeaker. At each distance, the HRIR was measured for a number of azimuths and elevations. For the 0.50, 0.75, and 1.00 m distances, elevations from -40° to 90° in 10° increments were used.

Measurements at 0.25 m were taken from -10° to 90° in 10° increments.[5] At all elevations and distances, measurements were taken through a full 360° around the KEMAR. Because the KEMAR is symmetric about the median plane, measurements taken in one ear at azimuths θ and $-\theta$ are equivalent to measurements taken in both ears at azimuth θ. The 360° of monaural measurements made in the right ear thus allow for binaural analysis in the right hemisphere. The azimuth increment was dependent on elevation. Our goal was to sample the spherical space around the KEMAR approximately uniformly, and using a constant azimuth increment would have resulted in oversampling at the higher elevations. Thus, increments were chosen such that adjacent measurement positions at each elevation fell on great-circle arcs separated by 5° in the horizontal plane. Table 2.1 contains the azimuth increment for each elevation. In all, 2656 HRIRs were recorded.

2.6 System Response Measurements

In addition to the HRIR measurements, the impulse response of the data-acquisition system itself was measured at 0.25 m, 0.50 m, 0.75 m, 1.00 m. This was done by removing the microphone from the KEMAR, and mounting it on a stand. At each of the four distances, the microphone was positioned on the 0° axis of the loudspeaker, and the impulse response of the system at that distance was measured using the same process as that for the HRIRs. These measurements served two purposes. First, all of the HRIRs are "colored" by

[5]It was impossible to rotate the KEMAR through 360° with the loudspeaker at elevations below -10° (and at a distance of 0.25 m) without contact between the loudspeaker and the KEMAR's shoulder.

[6]Only one measurement was taken at the 90° elevation position.

Elevation (degrees)	Azimuth Increment (degrees)
-40	6.43
-30	6.00
-20	5.00
-10	5.00
0	5.00
10	5.00
20	5.00
30	6.00
40	6.43
50	8.00
60	10.00
70	15.00
80	30.00
90	N/A[6]

Table 2.1: Dependence of azimuth increment on elevation, adapted from Gardner and Martin [12]

the frequency response of the system. If the impulse response of the system itself is known, an inverse filter can be designed to remove the effects of the system, leaving only the true HRIR. See Section 3.1.3 for a description of the design and implementation of the inverse filters. Note that this filtering is not necessary when analyzing interaural cues, as the same inverse filter will be applied to both the left and right ears, and thus the effect will be divided out. Second, as is discussed in Chapter 1, we define the HRTF as the sound pressure at the ear divided by the sound pressure at the center of the head with the head not present. The response without the head is simply the free-field response measured by the microphone at the correct distance from the sound source, which is precisely the same as the response of the data-acquisition system itself. Thus any one of the measurements made with the KEMAR can

be divided by the system response measurement made at the same distance to yield the actual monaural HRTF.

Chapter 3

Data Analysis

This chapter covers the analysis of the data collected with the KEMAR manikin. The primary focus is on the range dependence of localization cues, and the agreement of the KEMAR data with mathematical predictions. Section 3.1 describes the method used to extract the HRIRs from the measurements discussed in Chapter 2. Section 3.2 describes a mathematical model used to simulate monaural and binaural localization cues. Sections 3.3 and 3.4 cover monaural and binaural localization cues, respectively, which are extracted from the KEMAR data. Emphasis is placed on the binaural cues because they are used in the localization model discussed in Chapter 4.

3.1 Data Reduction

3.1.1 Extracting the HRIR from the Measured Data

The first step in the data-analysis process was to convert the measurements taken with the KEMAR into head-related impulse responses. As is discussed in Section 2.3, this was done by cross-correlating the MLS with the system response to the MLS. The exact method used for this work was taken from Gardner and Martin [12]. Given an MLS of length N, the first N samples of

the recorded system response were discarded, and two copies of the second N samples were concatenated. This sequence (of length $2N$) was cross-correlated with the original MLS; the correlation was implemented with discrete Fourier transform multiplication. A 512-point impulse response was extracted from the result of the correlation starting at the 62^{nd} sample, since the internal delay of the acquisition system was measured to be 61 samples (see Section 3.1.2). Each 512-point impulse response was then windowed with a 512-point Hanning window, normalized to the maximum HRIR value, multiplied by 2^{15}, and stored as 2-byte short integers.

3.1.2 Calculation of the System Delay

The internal system delay was calculated from the system impulse-response measurements which were made with the microphone removed from the KEMAR, as described in Section 2.6. In each recorded data file, the waveform of the system impulse response is preceded by a series of zero-valued samples, which correspond to a delay between the onset of the data collection and the processing of the response by the software. This time lag is made up of two components: an acoustic delay proportional to the distance between the speaker and the microphone, and the system delay. The length of the acoustic delay can be calculated by dividing the distance between the loudspeaker and the microphone by the speed of sound. The system delay is then found by subtracting the acoustic delay (in samples) from the total time lag. For each of the four system impulse-response measurements, the internal system delay

was found to be 61 samples.[1]

3.1.3 Inverse Filters

As is mentioned in Section 2.6, the data-acquisition system itself has a particular frequency response that is imposed on the measured HRTFs. The effects of the system frequency response can be removed with a filter whose frequency response is the reciprocal of the system frequency response. Once the system impulse response was found at each of the four distances, the corresponding inverse filter for each distance was generated with the following process.

1. The system impulse response was windowed with a Hanning window, and the system frequency response was found by applying the Fourier transform;

2. The log-magnitude of the frequency response was inverted at each point to create a frequency response equal to the reciprocal of the system response;

3. The gain of the inverted frequency response was clipped at a maximum of 30 dB;

4. An inverse Fourier transform was applied to the inverted, clipped frequency response to generate an impulse response;

5. The inverse impulse response was windowed with a Hanning window.

[1]The assumption that the system delay is constant at 61 samples was later found to be false. See Section 3.4.1 for details.

Once the filters were created, each HRIR measurement was convolved with the appropriate inverse filter to remove the effects of the measurement system.

3.2 Mathematical Predictions

3.2.1 Overview

Numerous attempts have been made to predict interaural time and level differences (ITDs and ILDs) mathematically [1, 17]. Unfortunately, many models assume for simplicity that the waves impinging on the head are planar. The predictions made with these models tend to be quite good for sources far from the head (where the curvature of the wavefronts at the head is negligible), but break down for close-range sources where the wavefronts are significantly curved. Thus, any model suitable for close-range HRTF, ITD, or ILD predictions must employ a source radiating spherical waves.

Reasonably accurate predictions of monaural HRTFs and binaural phenomena can be made with a mathematical model which approximates the head as a rigid sphere and assumes an external point source radiating spherical waves [3, 10, 35]. One can find the pressure at any point on the surface of the sphere due to the external point source; judicious choice of measurement points to represent the ears allows for the modeling of HRTFs, ITDs, and ILDs.

The pressure at a given point on a rigid sphere from an external point source radiating spherical waves (from [35], derived in [32]) is given by:

$$P_s(r, a, f, \theta) = \frac{\rho c U_0}{2\pi a^2} \times \left\{ \sum_{m=0}^{\infty} \left(m + \frac{1}{2} \right) P_m(\cos\theta) \frac{h_m(kr)}{h'_m(ka)} \right\} e^{j(2\pi f t - \frac{\pi}{2})} \tag{3.1}$$

where

ρ	=	density of air
c	=	speed of sound in air
f	=	frequency of the point source vibration
k	=	wave number $= 2\pi\lambda$ where $\lambda =$ wavelength
a	=	radius of the sphere
r	=	distance from the center of the sphere to the source
θ	=	angle between the ray from the center of the sphere to the source and the radius from the center of the sphere to the measurement point on the surface
U_0	=	volume velocity of the source
$P_m(\cos\theta)$	=	Legendre polynomial in $\cos\theta$
$h_m(kr)$	=	Spherical Hankel function in kr
$h'_m(ka)$	=	derivative of the Spherical Hankel function in ka

Note that there is no dependence on elevation (ϕ) because of symmetry, so all results are interpreted for $0\,^\circ$ elevation. Calculation of a monaural HRTF requires dividing the pressure at the ear by the free-field pressure at the center of the head. The free-field pressure at a distance r from a spherically radiating point source is given by:

$$P_{ff}(r,f) = \frac{\rho c k U_0}{4\pi r} e^{j(2\pi f t - kr + \frac{\pi}{2})} \tag{3.2}$$

with constants and variables as defined for (3.1).

We used a head radius $a = 8.75$ cm and a speed of sound $c = 346$ m/s to make predictions for source distances $r = 0.25, 0.50, 0.75, 1.00,$ and 1.40 meters.[2] The density of air and the source volume velocity divide out in

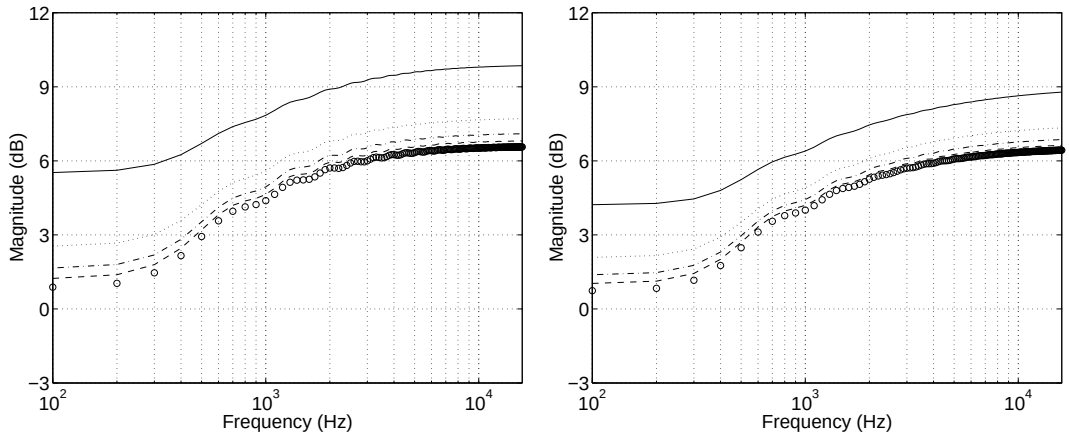
[2] 8.75 cm is traditionally used as the radius when modeling the head as a sphere (see [1], [10]), based on the "average" head size. It is also one-half of the front-to-back diameter of the KEMAR head, and thus is reasonably appropriate for comparison to our measured data.

both monaural (P_s/P_{ff}) and binaural ($P_{s,right}/P_{s,left}$) calculations. The "ears" were placed at positions of 100 ° and 260 ° (with 0 ° at the front of the modeled head), as recommended by Blauert [1] and Duda [10].

3.2.2 Monaural HRTF Predictions

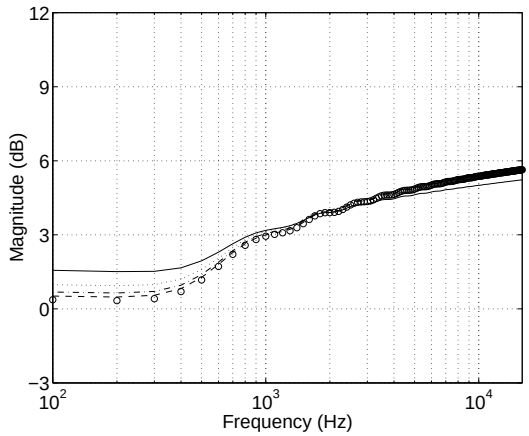
The monaural HRTFs predicted by the spherical head model can be seen in Figure 3.1. Each curve was generated by solving for $P_{s,right}/P_{ff}$ (see (3.1) and (3.2)) at the given azimuth and distance at 100 Hz intervals from 100 Hz to 16 kHz. The graphs shown cover azimuths of $-90\,^\circ \leq \theta \leq 90\,^\circ$ in 30 ° increments. Symmetry of the sphere dictates that the plots for $90\,^\circ \leq \theta \leq 270\,^\circ$ are the same as the ones shown.

There are a number of distance-dependent trends which can be seen in Figure 3.1. First, the magnitude of the low-frequency portion of the HRTF increases relative to that of the high-frequency portion as distance from the head decreases. For example, at $\theta = 90\,^\circ$ (Figure 3.1 (a)), the difference between the magnitude at 16 kHz and 100 Hz decreases monotonically from 5.7 dB at 1.40 m to 4.4 dB at 0.25 m. On the contralateral side ($\theta = -90\,^\circ$), where there is significant high-frequency shadowing, the difference between the low-frequency magnitude and the high-frequency magnitude increases monotonically from 3.2 dB at 1.40 m to 5.5 dB at 0.25 m. At both azimuths, the majority of the increase occurs in the transition from 0.5 m to 0.25 m. This increased low-frequency content (relative to high-frequency content) with decreasing distance may be perceived as a low-pass filtering of the source as it approaches the listener, which may in turn serve as a cue to distance. This result is somewhat contrary to expectations: because the relative attenuation

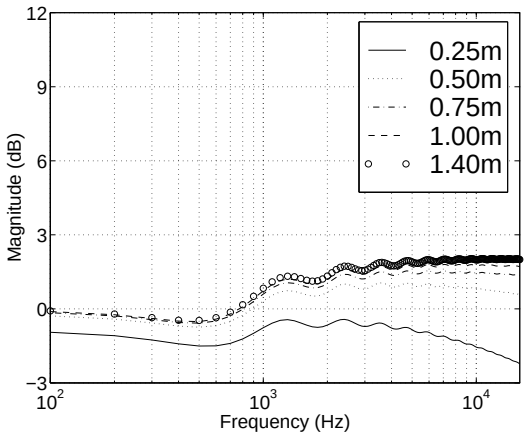


(a) Azimuth = 90 °.

(b) Azimuth = 60 °.

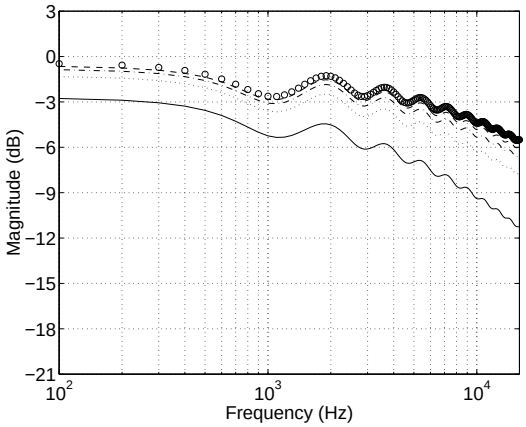


(c) Azimuth = 30 °.

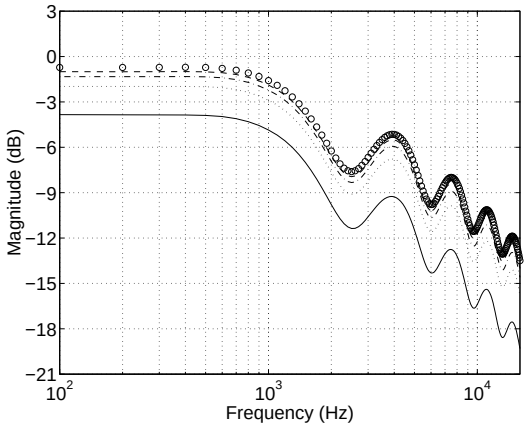


(d) Azimuth = 0 °.

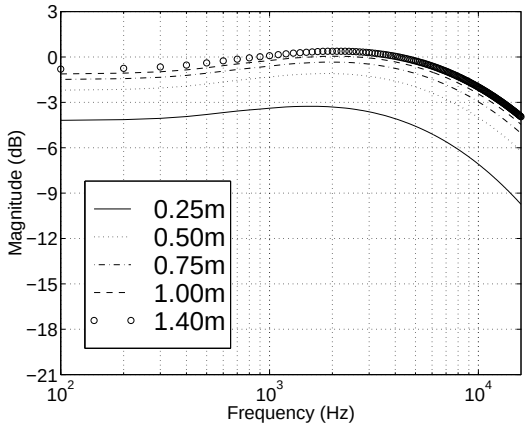
Figure 3.1: Sphere model predictions of HRTFs.



(e) Azimuth = $-30\,^\circ$.



(f) Azimuth = $-60\,^\circ$.



(g) Azimuth = $-90\,^\circ$.

Figure 3.1 (continued): Sphere model predictions of HRTFs.

of high frequencies by air increases as distance increases, *more* distant sources are also perceived as low-pass filtered (see [29]). This apparent contrast was noted by Coleman [7], who commented,

> ...the frequency spectrum may play a dual role in auditory depth perception with relatively greater high-frequency content signaling a *closer* source at distances greater than several feet but signifying a more distant sound source when the source is close to the observer.

The second distance-dependent feature of the modeled HRTFs is the overall magnitude. The HRTF measured at the near ear *decreases* in overall magnitude with increasing source distance, while that measured at the far ear *increases* with increasing source distance. This effect is most pronounced when the source is near an azimuth of $90\,^\circ$ or $-90\,^\circ$, and it is due to the $1/r$ dependence of pressure from the spherical source. It is most easily explained for sources at $\theta = 90\,^\circ$ or $-90\,^\circ$. When the source distance is doubled (for example) from r to $2r$, the free-field pressure measured at the position of the center of the head (P_{ff}) will decrease by 6 dB, since the distance to the source is measured from that point. The distance from the ipsilateral ear (here $\theta = 90\,^\circ$) to the source will increase from $(r - a)$ to $(2r - a)$. Since $(2r - a)/(r - a) > 2$ (assuming both $r, a > 0$), $P_{s,right}$ will decrease by more than 6 dB, with the overall effect of the magnitude of the ipsilateral HRTF ($P_{s,right}/P_{ff}$) decreasing. Given this same doubling of source distance, the distance from the source to the contralateral ear (here $\theta = -90\,^\circ$) will increase from $[(r - a) + \pi a]$ to $[(2r - a) + \pi a]$. Since $[(2r - a) + \pi a]/[(r - a) + \pi a] < 2$,

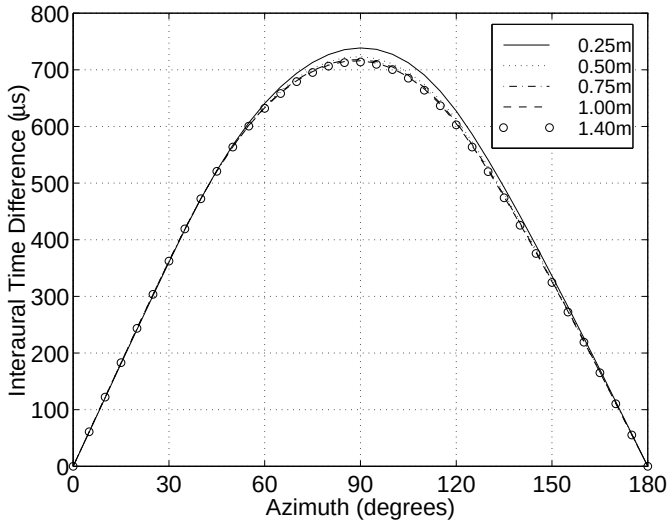


Figure 3.2: Interaural time differences predicted with the spherical head model.

$P_{s,left}$ will decrease by less than 6 dB, with the overall effect of the magnitude of the contralateral HRTF ($P_{s,left}/P_{ff}$) increasing. Use of this trend as a cue to distance would require some *a priori* knowledge of the azimuth of the source since the trend itself is azimuth dependent.

3.2.3 Interaural Time Difference Predictions

The predictions of interaural time differences for $\phi = 0\,^\circ$ elevation can be seen in Figure 3.2. ITD values were calculated for $0\,^\circ \leq \theta \leq 180\,^\circ$ in $5\,^\circ$ increments.[3] Each curve was created by averaging the ITDs from 200 Hz to 1.5 kHz at 100 Hz intervals .[4]

The data clearly show that the ITD has a very weak dependence on

[3]Left/right symmetry of the sphere limits the necessary calculations to azimuths $0\,^\circ \leq \theta \leq 180\,^\circ$.

[4]The 200 - 1500 Hz bandwidth was chosen to match that used when extracting ITDs from the KEMAR data.

range, particularly when the source is far from the interaural axis. However, when the source azimuth is near 90°, there is a small increase in the ITD with decreasing distance, particularly when the source is within 0.50 m. The ITD at 90° increases only 9 μs when the source is moved from 1.40 m to 0.50 m (from 714 μs to 723 μs), but 16 μs when the source is moved from 0.50 m to 0.25 m (from 723 μs to 739 μs). Brungart [3] reports similar results, and has shown an additional (and larger) increase in the ITD near 90° when the source is moved from 0.25 m to 0.125 m. Based on the facts that the time difference is dependent on the path length difference between the two ears, and the ears are fixed, the small increase in ITD must be due to the fact that the curvature of the wavefronts impinging on the head increases as source distance decreases.

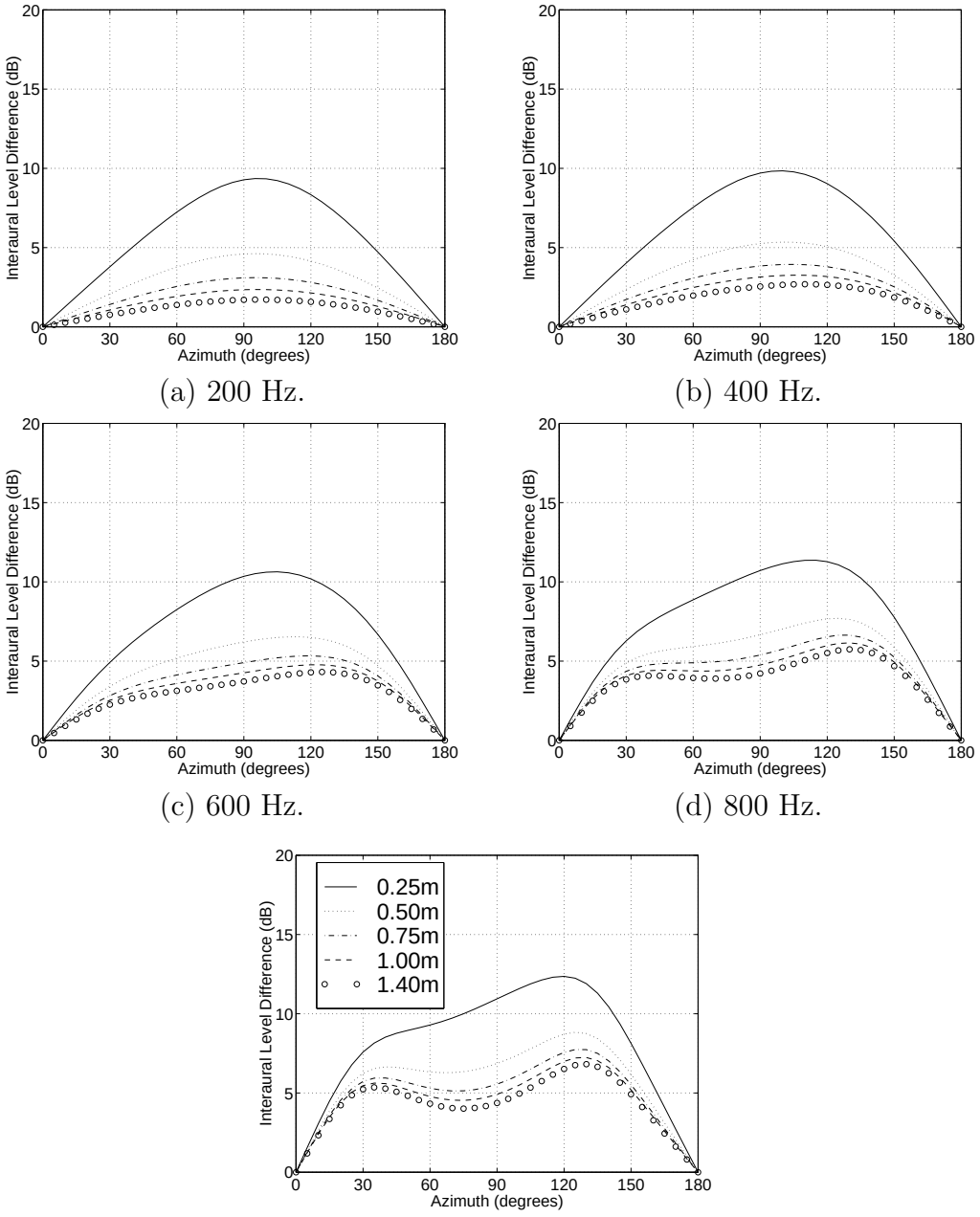
The predicted ITDs are not symmetric about 90° because the "ears" were not modeled at opposite ends of a diameter. The ITDs drop off more quickly to the rear of the listener because the ears are closer to the back of the head than to the front. The azimuth range in which the ITD at 0.25 m is visibly greater than the others is approximately $50° \leq \theta \leq 150°$, which is centered around the 100° ear position rather than 90°.

Based on the results of the sphere-model simulations, the interaural time difference does not appear to be useful as a cue to source distance, except perhaps when the source is extremely close to the head and nearly aligned with the interaural axis. With the correct coordinate transformation, however, the ITD can be used effectively as a cue to azimuth. We exploit this fact in the localization model described in Chapter 4.

3.2.4 Interaural Level Difference Predictions

The ILD predictions from the spherical head model are shown in Figure 3.3. They were generated with (3.1) by solving for $P_{s,right}/P_{s,left}$ for azimuths $0\,^\circ \leq \theta \leq 180\,^\circ$ in $5\,^\circ$ increments, and for frequencies 200 Hz to 2.0 kHz in 200 Hz increments and 2.0 kHz to 10.0 kHz in 2.0 kHz increments. Again, the symmetry of the spherical model restricts the measurements to the horizontal plane ($\phi = 0$), and dictates that the ILDs for $-180\,^\circ \leq \theta \leq 0\,^\circ$ will be the same as those shown. At low frequencies, the ILDs predicted by the spherical model increase monotonically from 0 dB as the source moves toward the measurement ear from the front of the head ($0\,^\circ \leq \theta \leq 100\,^\circ$), then decrease monotonically as the source passes the ear and moves to the rear of the head ($100\,^\circ \leq \theta \leq 180\,^\circ$). The peak is at $100\,^\circ$ rather than $90\,^\circ$ due to the placement of the ears within the model. When the wavelength of the sound impinging on the head is much greater than the radius of the head [as it is in Figure 3.3 (a) – (c)], diffraction by the head is minimal, and the ILD is due mainly to the additional attenuation due to spreading incurred by the signal travelling to the far ear. As the source moves closer to the head, the ratio of the distance to the far ear and the distance to the near ear increases since the path length difference remains constant; this results in an increase in ILD. The path length difference can actually exceed the distance to the near ear when the source is within 0.5 m, thus giving rise to the large ILDs when the source is 0.25 m from the head.

As the source frequency is increased, two trends become evident. First, the overall magnitude of the ILD increases. This is due to an acoustic shadow



(a) 200 Hz.

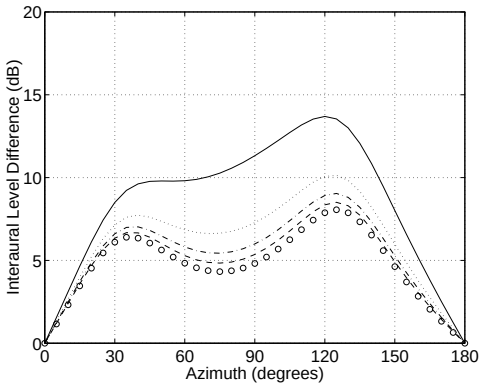
(b) 400 Hz.

(c) 600 Hz.

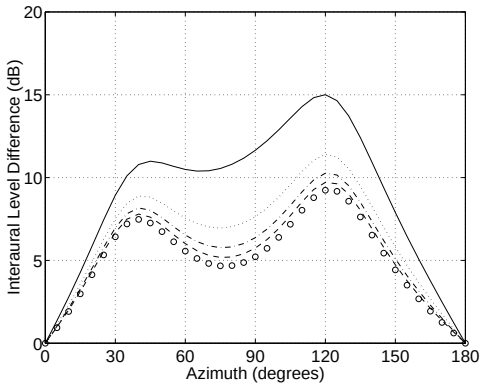
(d) 800 Hz.

(e) 1000 Hz.

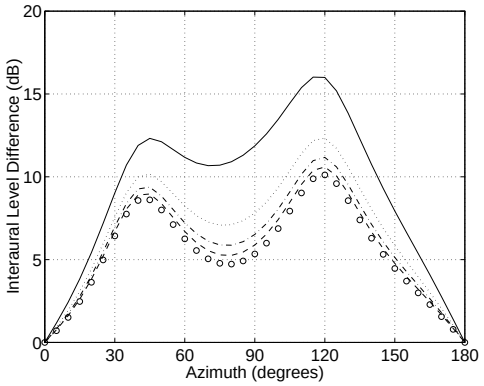
Figure 3.3: Sphere model predictions of ILDs: 200 Hz – 1000 Hz.



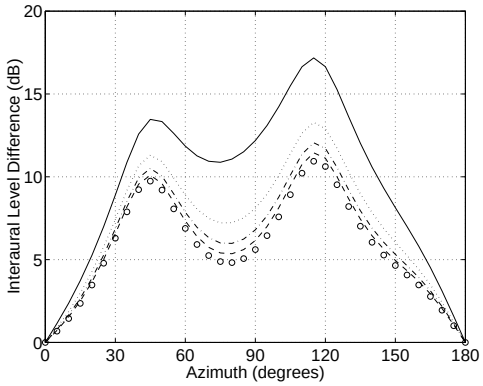
(f) 1.2 kHz.



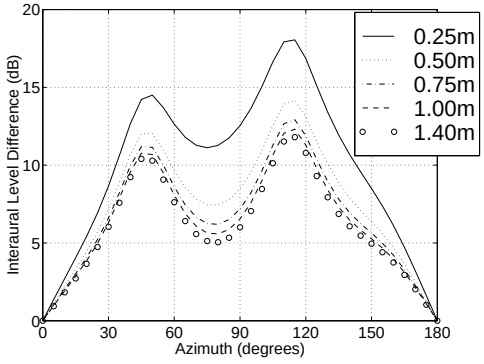
(g) 1.4 kHz.



(h) 1.6 kHz.

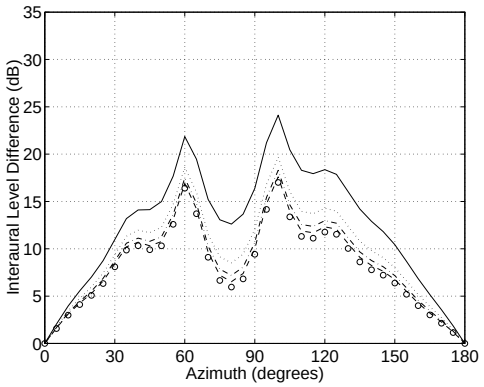


(i) 1.8 kHz.

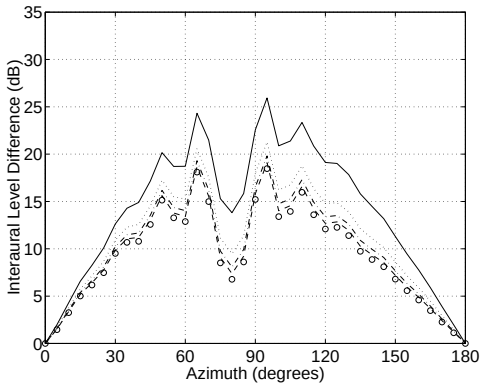


(j) 2.0 kHz.

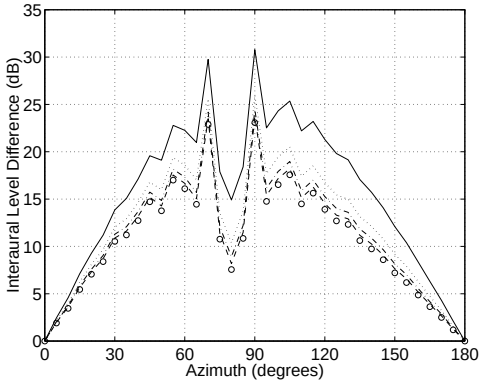
Figure 3.3 (continued): Sphere model predictions of ILDs: 1.2 kHz – 2.0 kHz.



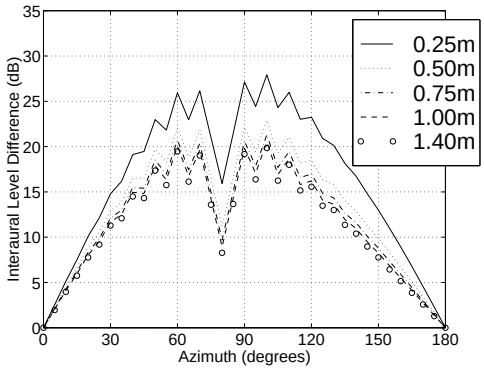
(k) 4.0 kHz.



(l) 6.0 kHz.



(m) 8.0 kHz.



(n) 10.0 kHz.

Figure 3.3 (continued): Sphere model predictions of ILDs: 4.0 kHz – 10.0 kHz. Note the change in scale from (a) - (j).

cast by the head which reduces the pressure at the far ear, and pressure doubling on the near side due to the rigid surface of the sphere.[5] Second, the ILDs at all distances have a local minimum in the vicinity of $\theta = 80\,^{\circ}$. Constructive interference of the diffracted waves traveling around the sphere creates an acoustic "bright spot" on the contralateral side of the sphere: the diffracted waves arrive roughly in phase at the far ear, causing an increase in pressure (relative to lower frequencies). This increase in pressure ($P_{s,left}$) causes the dip in the ILD.

A further increase in source frequency results in additional changes to the ILDs. The notch near the interaural axis described above deepens, reaching a maximum depth of approximately 15 dB at 8.0 kHz (see Figure 3.3 (m)). The notch is centered around an azimuth of $80\,^{\circ}$. When the source is positioned at $\theta = 80\,^{\circ}$, the far ear (modeled at $\theta = 260\,^{\circ}$) is at the opposite end of a diameter of the sphere from the point where the sound impinges on the head. All paths around the sphere to the far ear are thus of equal distance (1/2 of the circumference, or πa where a is the radius), resulting in a maximum of constructive interference of the diffracted waves and a local minimum in the ILD. At 4.0 kHz and above, additional local maxima and minima can be seen in the ILD curves. These are caused by the more complex diffraction which occurs when the radius of the sphere is equal to or greater than the wavelength of the sound from the source. The patterns in the high-frequency ILD curves are consistent across all measurement distances (albeit reduced in magnitude). Despite the irregularities in shape, the ILD curves show a trend of increasing

[5] See [32] for an in-depth discussion on the physical acoustics of diffraction about a rigid sphere.

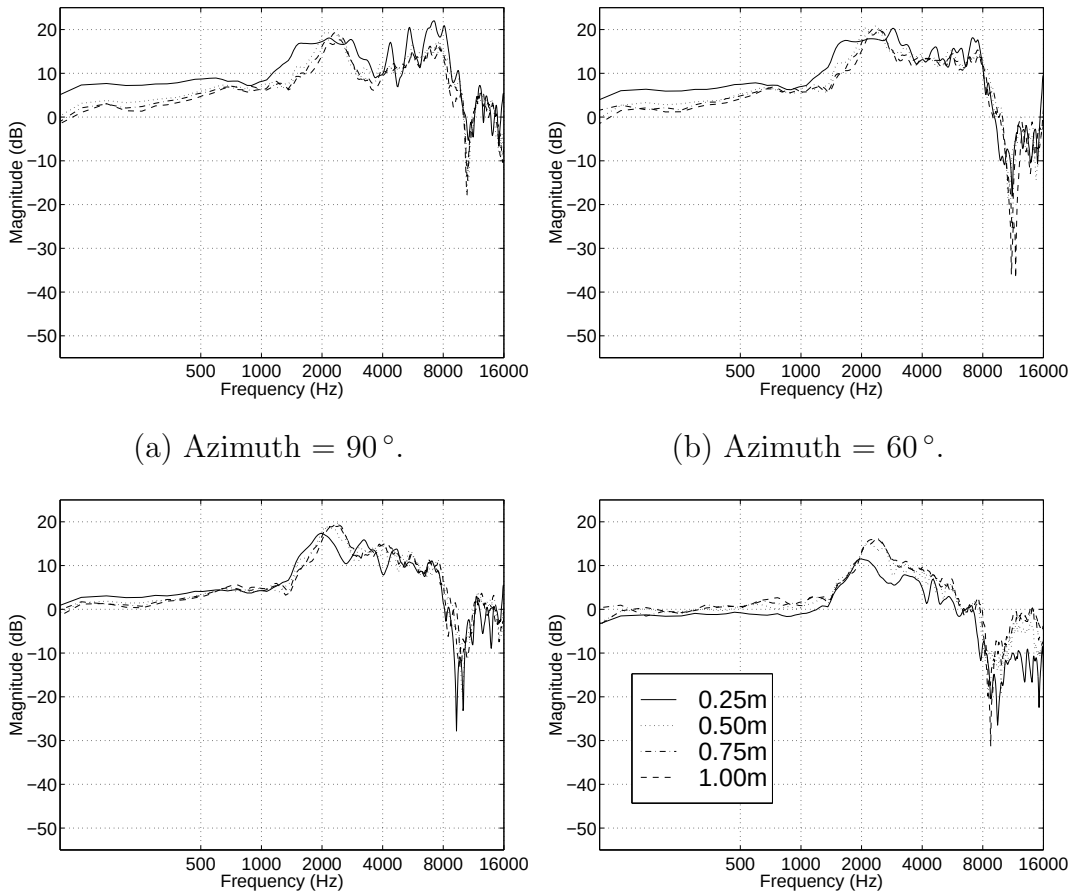
overall magnitude with increasing frequency.

The feature common to the predicted ILD at all azimuths and frequencies is a monotonic increase in magnitude with decreasing source distance. This strong dependence on range, which is most pronounced when the source is within 0.5 m of the head, should serve as an excellent localization cue for close-range sources.

3.3 Monaural Cues from the KEMAR Data

A small sample of the 2656 HRTFs measured with the KEMAR is shown in Figure 3.4. In particular, these match the spherical model HRTFs (see Figure 3.1) in azimuth and elevation (with the addition of the data from $\theta = 180\,^{\circ}$ in Figure 3.4 (h)). The two distance-dependent trends noted for the modeled HRTFs are also present in the KEMAR HRTFs. First, the measured HRTFs show an increased low-frequency content relative to high-frequency content as the source moves closer to the head. Second, the magnitude (particularly at low frequency) of the HRTF decreases with increasing source distance for the ear nearer the source, and increases with increasing source distance for the ear further from the source.

Because our goal is to exploit binaural rather than monaural cues in our localization model, an in-depth analysis of the changes in the monaural HRTF with distance has been omitted. Despite this fact, two features of the KEMAR HRTFs should be noted here. The first is the pinna notch, which is not present in the modeled HRTFs because the model makes no attempt to simulate the outer ear. Certain features of this notch, in particular the slope



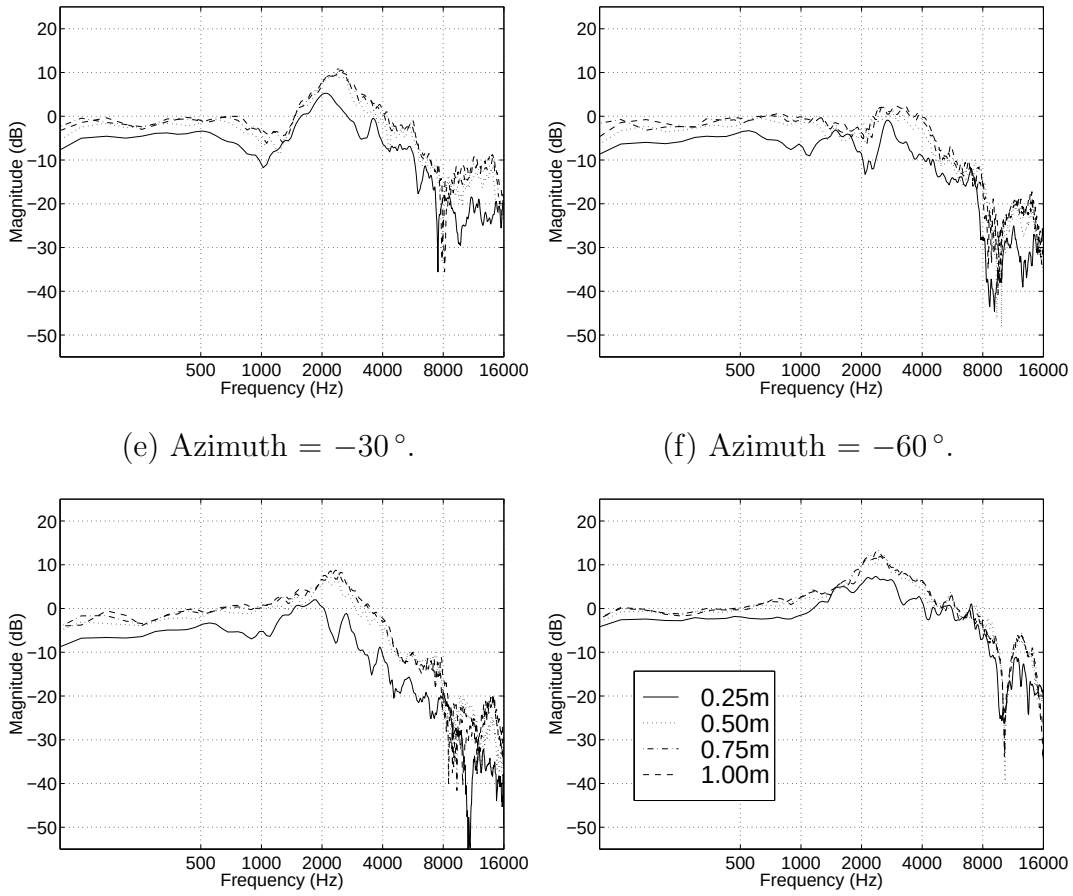
(a) Azimuth = 90 °.

(b) Azimuth = 60 °.

(c) Azimuth = 30 °.

(d) Azimuth = 0 °.

Figure 3.4: KEMAR HRTFs.



(e) Azimuth = $-30\,^{\circ}$.

(f) Azimuth = $-60\,^{\circ}$.

(g) Azimuth = $-90\,^{\circ}$.

(h) Azimuth = $180\,^{\circ}$.

Figure 3.4 (continued): KEMAR HRTFs.

of its low-frequency side and the frequency at maximum depth, are generally considered to serve as the main cue to source elevation, particularly in the median plane where binaural cues are greatly reduced (see Han [16]). The notch affects the interaural level difference, and thus plays an indirect role in binaural localization; see Section 3.4.2 for more detail. The second notable feature of the KEMAR HRTFs in Figure 3.4, particularly those measured at a distance of 0.25 m, is the apparent comb-filtering in the high frequencies. This is due to multiple reflections between the head and the loudspeaker cabinet during the measurement, and is greatly reduced when the source is at larger distances. Brungart [3] has developed an alternative measurement procedure which utilizes a much smaller source, and thus may eliminate this comb filtering.

3.4 Binaural Cues from the KEMAR Data

3.4.1 Interaural Time Differences

There are a number of ways to extract interaural time differences from HRTF data. One is to use a left/right pair of HRIRs as input to a cochlear model, and extract the ITD from various frequency bands by cross-correlating the left and right output of the model for each band (see [21] and [25]). Since the ITD shows little variation with frequency for a given source location, a first-order approximation of the ITD may also be obtained by simply finding the time lag associated with the maximum of the cross-correlation of the left and right HRIRs (see [42]). By defining the interaural transfer function (ITF) as the ratio of the right and left HRTFs (for a particular source position), one can calculate the ITD from the derivative of the unwrapped phase (group delay)

of this ITF (see [3] and [9]). For this thesis, the last of the three methods was chosen. In particular, the ITD is obtained from the slope of the best linear fit to the unwrapped phase of the ITF, using a bandwidth of 200 Hz to 1.5 kHz.[6]

The ITDs calculated for source elevation $\phi = 0\,^\circ$ and all measurement distances are shown in Figure 3.5. The jagged nature of the curves is due to jitter in the data-acquisition system. As is explained in Section 3.1.2, the DAQ system delay was calculated to be 61 samples, and thus the first 61 samples of each HRIR were discarded. Unfortunately, the system delay was not constant, but in fact varied over a small but significant range (see Figure 3.6).[7] If the system delay differs by n samples between the left- and right-ear measurements used to calculate a particular ITD, the ITD will be artificially increased or decreased by 23 μs (since the sampling period $T_s = 23$ μs). The effect of even a 1 sample difference is not negligible, in particular near the midline ($0\,^\circ \leq \theta \leq 20\,^\circ$ and $160\,^\circ \leq \theta \leq 180\,^\circ$) where the ITD is less than 200 μs.

Because there is no way to know the exact system delay for a given measurement, the jitter cannot be removed from the ITD data. In order to compensate, we smoothed all of the ITD curves (see Figure 3.8) with a polynomial best fit.[8] The ITD values used in the localization model described

[6]This bandwidth was chosen because the ITD is considered to be a useful cue only at low frequencies. The interaural phase difference implied by the ITD is not detectable above 1.5 kHz, and thus the ITD ceases to be an unambiguous cue to localization at higher frequencies [5].

[7]The data for Figure 3.6 were obtained by connecting the output channel of the DAQ board directly to the input channel. A sampled step function of magnitude 1 was sent through the signal path, and the index of the first non-zero sample seen at the input was stored. The figure shows the results for 2656 trials.

[8]The order of the polynomial varied with elevation, since the number of measurements

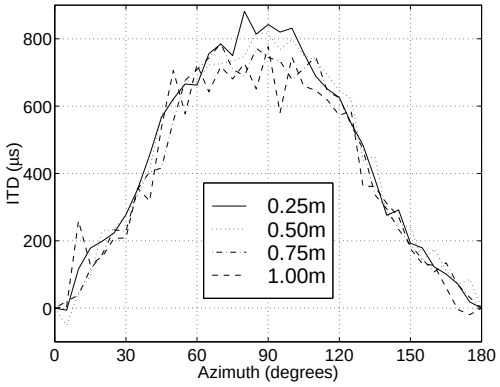


Figure 3.5: Measured interaural time differences at $\phi = 0\,^\circ$.

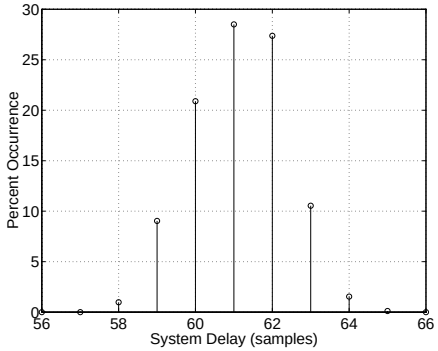


Figure 3.6: Jitter in the delay of the data acquisition system around 61 samples: 1 sample = 23 μs.

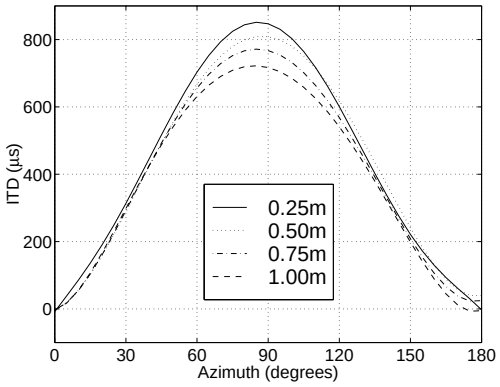
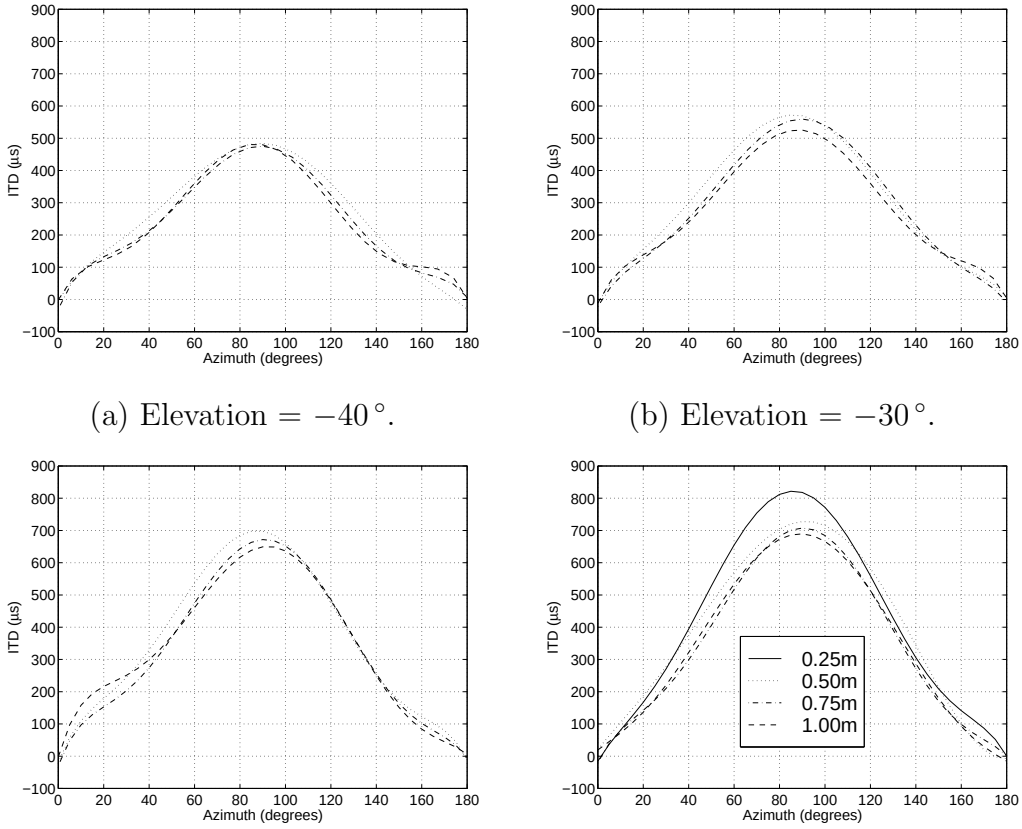


Figure 3.7: Smoothed interaural time differences at $\phi = 0\,^\circ$. See also Figure 3.8.



(a) Elevation = $-40\,^\circ$.

(b) Elevation = $-30\,^\circ$.

(c) Elevation = $-20\,^\circ$.

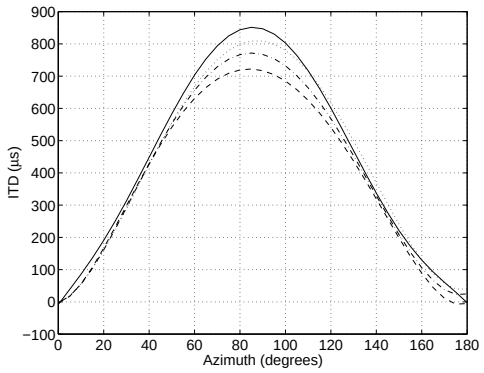
(d) Elevation = $-10\,^\circ$.

Figure 3.8: Interaural time differences extracted from the KEMAR data. Panes (a)–(c) show only three curves because no measurements were made below $\phi = -10\,^\circ$ at 0.25 m.

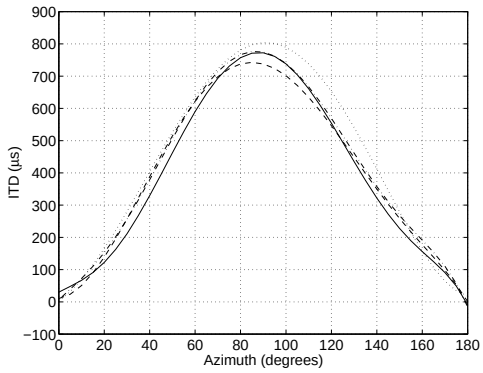
in Chapter 4 are taken from the smoothed data.

The smoothed ITDs extracted from the KEMAR data are in relatively good agreement with the predicted values. In particular, those from elevations $-10\,^\circ, 30\,^\circ, 40\,^\circ, 50\,^\circ, 60\,^\circ$, and $70\,^\circ$ [Figure 3.8 panes (d) and (h)–(l)] clearly show that the ITD is not a function of range for sources more than 0.5 m from

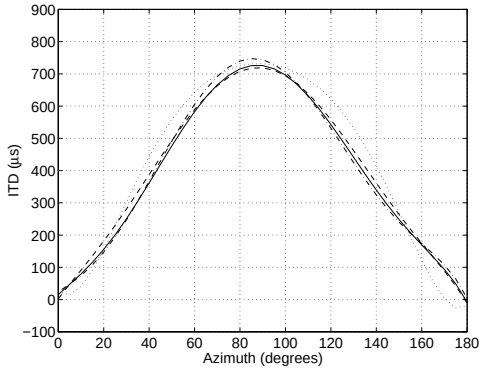
decreased as the elevation moved away from $0\,^\circ$.



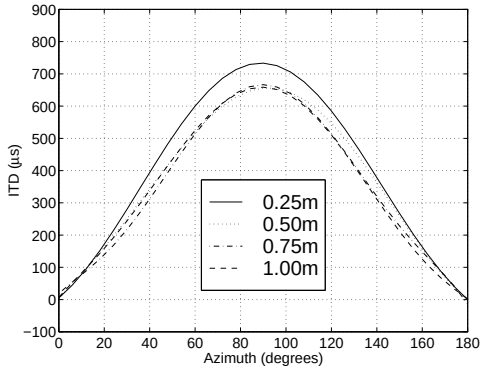
(e) Elevation = 0 °.



(f) Elevation = 10 °.

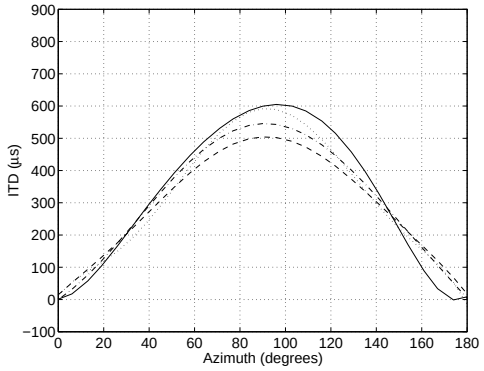


(g) Elevation = 20 °.

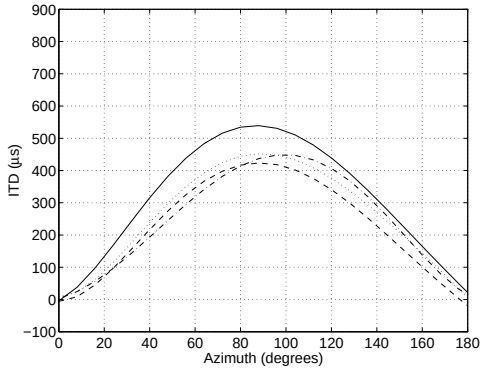


(h) Elevation = 30 °.

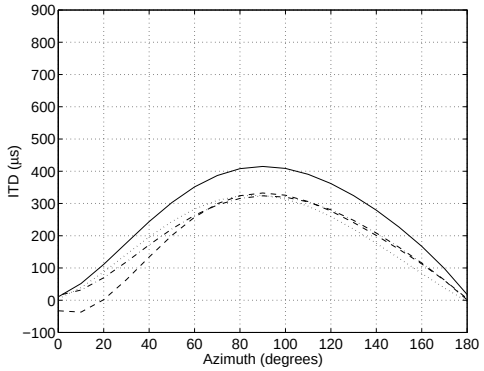
Figure 3.8 (continued): Interaural time differences extracted from the KEMAR data.



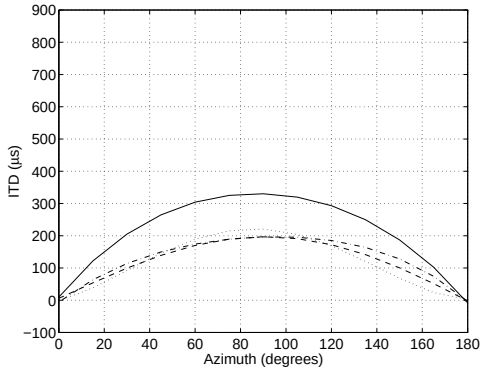
(i) Elevation = 40 °.



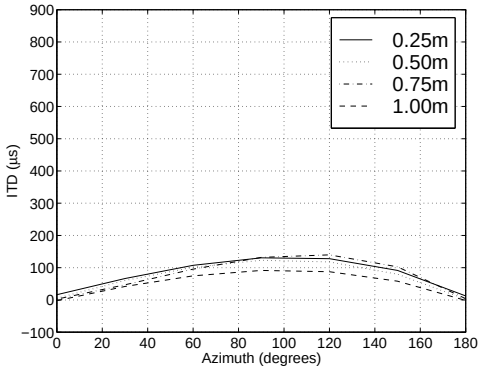
(j) Elevation = 50 °.



(k) Elevation = 60 °.



(l) Elevation = 70 °.



(m) Elevation = 80 °.

Figure 3.8 (continued): Interaural time differences extracted from the KEMAR data.

the head, and that the ITD increases slightly when the source is moved from 0.5 m to 0.25 m. Panes (a)–(c) also show that the ITD is not dependent on range for sources outside a 0.5 m radius. Despite the lack of measurements at 0.25 m for elevations below $-10\,^\circ$, it would be logical to expect a small increase in the ITD for sources inside 0.5 m, based on extrapolation from measurements made at positive elevations. The actual ITDs for elevations $0\,^\circ, 10\,^\circ$, and $20\,^\circ$ are somewhat more jagged than the others, and the expected trends at these elevations have been obscured by the smoothing.

The slight asymmetry around $90\,^\circ$ in both the measured and predicted ITD plots is due to the ears being set slightly back from the center of the head.[9] The curves from the measured data are slightly more narrow than those from the predicted data (see Figure 3.2) because the KEMAR head is not perfectly spherical. The interaural axis is somewhat shorter than the front-to-back "diameter" of the head (14.5 cm vs. 17.5 cm), so the measured ITDs are smaller than the predicted ITDs near the median plane. The discrepancy in the peaks of the 0.25 m ITDs (predicted = 739 μs, measured = 852 μs)[10] can be attributed to small errors introduced by the smoothing process and in the positioning system, as well as facial features of the KEMAR (e.g., the nose) which increase the interaural path length. Overall, the analysis of the data extracted from the KEMAR measurements leads to the same conclusion as the analysis of the spherical head model predictions: the interaural time difference is not strongly dependent on the distance of an acoustic source from

[9]This agreement between the predicted and measured values indicates that $\theta = 100\,^\circ$ and $\theta = 260\,^\circ$ are good azimuthal locations for the "ears" in the sphere model.

[10]The difference of 113 μs is equivalent to a .039 m difference in path length, given a speed of sound equal to 346 m/s.

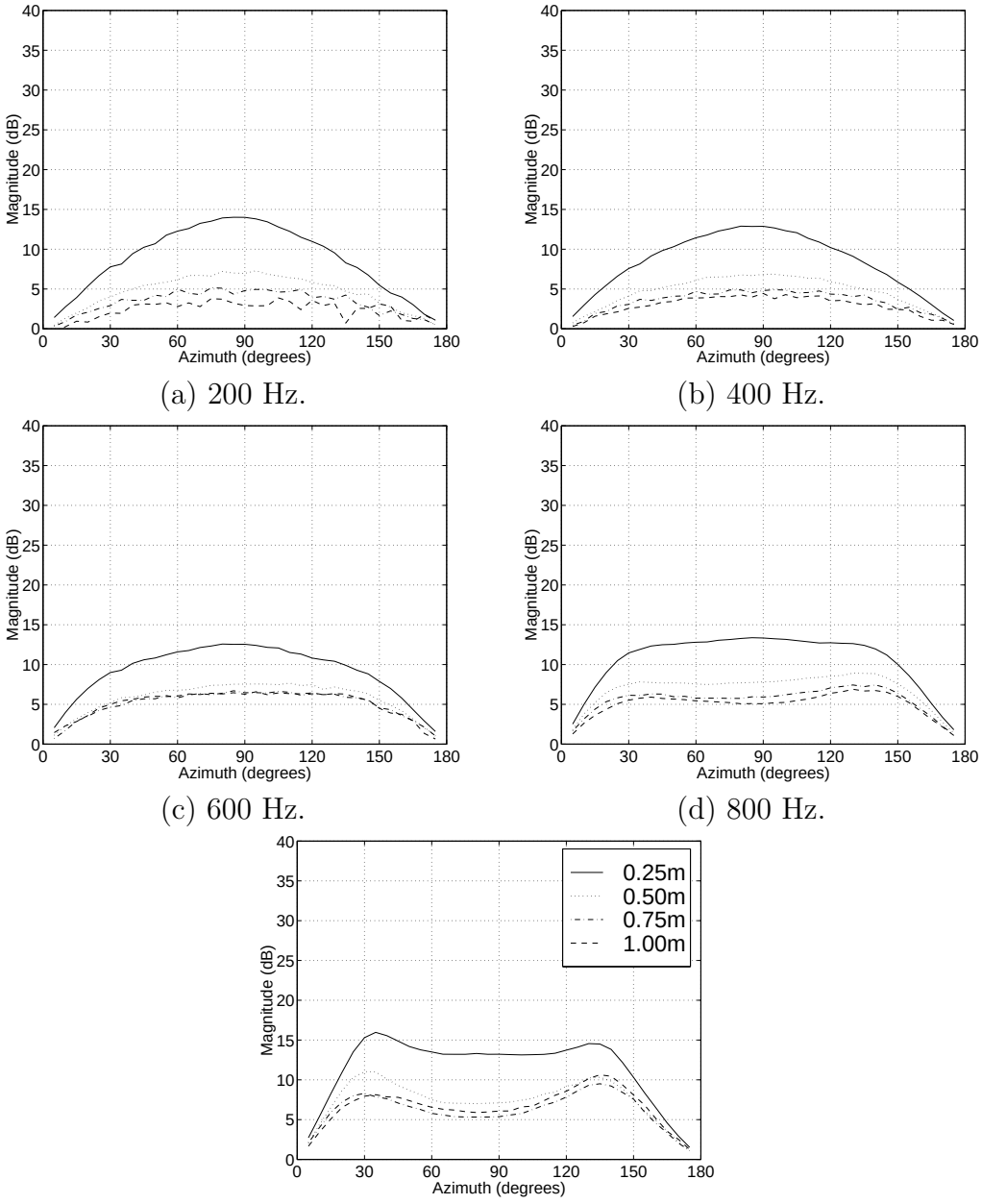
a listener, and thus it is not a useful cue in distance localization.

3.4.2 Interaural Level Differences

There are also a number of ways to extract interaural level differences from the collected data. ILD extraction does differ from ITD extraction, however, in that the ILD shows significant changes with frequency (as one would expect given that the ILD is largely due to diffraction around the head), and thus cannot be collapsed into one value per source position. The use of a cochlear model has been suggested for ILD analysis. The log magnitude difference between the left and right outputs can be found for as many narrow-band channels as are available in the cochlear model (see [21] and [25]). A fast Fourier transform (FFT) method may also be used, where the log-magnitude value of the interaural transfer function at a given frequency is interpreted as the ILD at that frequency (see [9]). This is equivalent to subtracting the log-magnitude of the left HRTF from that of the right HRTF (for a given source position) at the frequency of interest. The latter method is the simpler of the two, but it does supply a finer frequency resolution than is available in the cochlea [21].[11] We chose the FFT method for its ease of use and for more straightforward comparison to the sphere model predictions described in Section 3.2.4.

The ILDs extracted from the KEMAR which match the predicted ILDs in elevation ($\phi = 0\,^\circ$) and frequency are shown in Figure 3.9. The agreement at low frequencies [panes (a)–(e)] is quite good, with the measured ILDs showing

[11] The frequency resolution of the cochlea is limited by the bandwidth of critical bands (see [34]).



(a) 200 Hz.

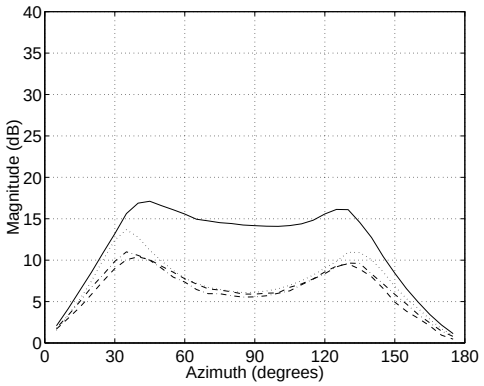
(b) 400 Hz.

(c) 600 Hz.

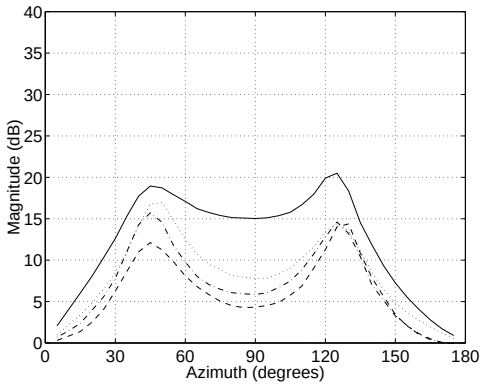
(d) 800 Hz.

(e) 1000 Hz.

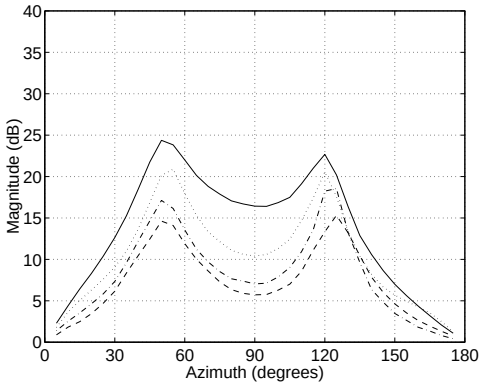
Figure 3.9: ILDs extracted from the KEMAR data: 200 Hz – 1000 Hz.



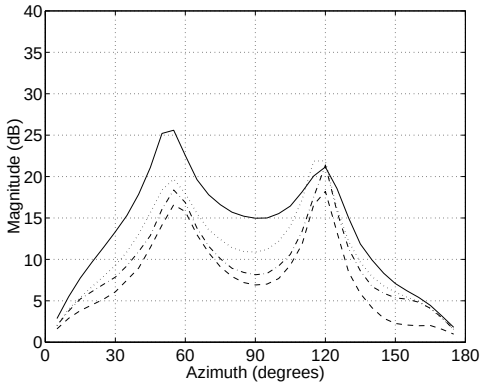
(f) 1.2 kHz.



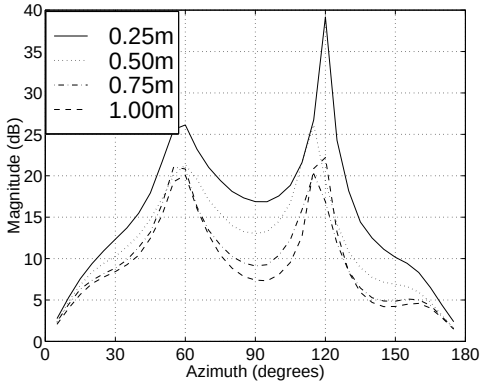
(g) 1.4 kHz.



(h) 1.6 kHz.

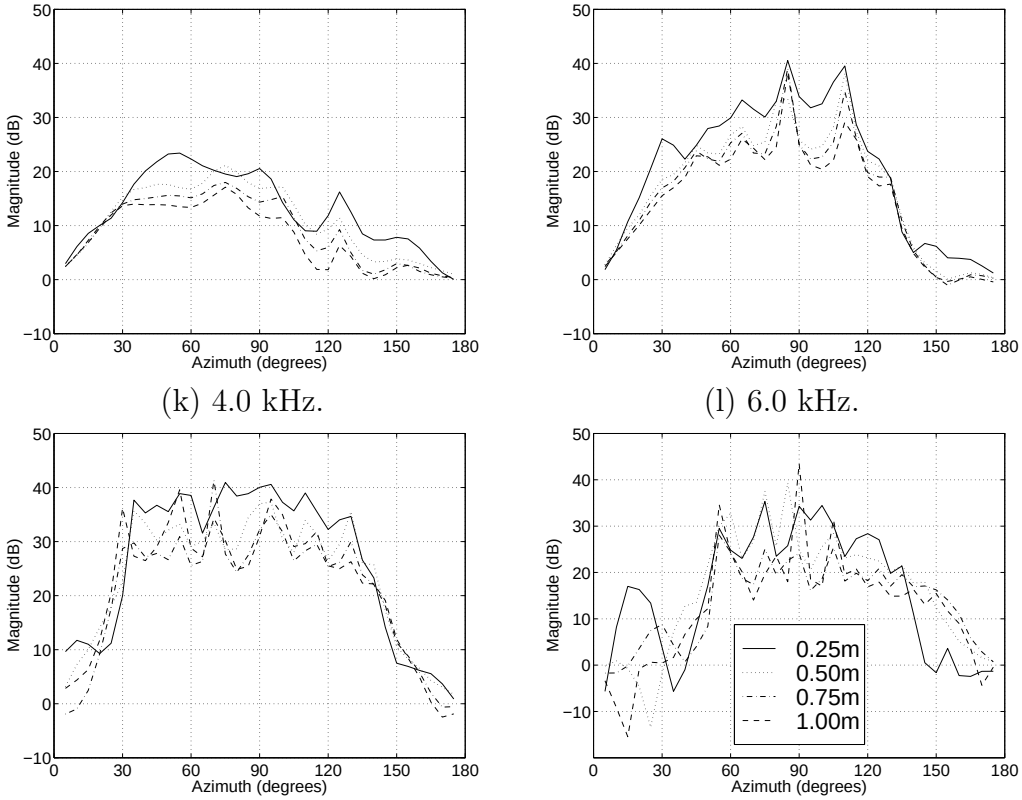


(i) 1.8 kHz.



(j) 2.0 kHz.

Figure 3.9 (continued): KEMAR ILDs: 1.2 kHz – 2.0 kHz.



(k) 4.0 kHz. (l) 6.0 kHz.

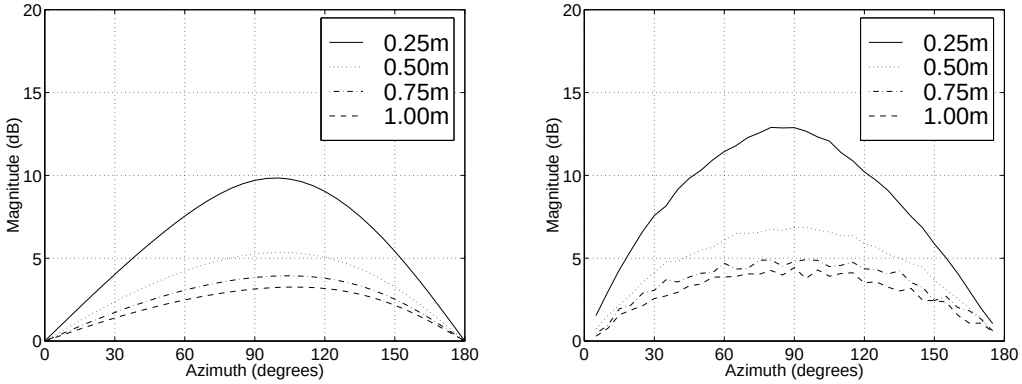
(m) 8.0 kHz. (n) 10.0 kHz.

Figure 3.9 (continued): KEMAR ILDs: 4.0 kHz - 10.0 kHz. Note the change in scale from parts (a) – (j).

the expected dependence on both azimuth and distance. At and below 1.0 kHz, the measured values exceed the predicted values at all azimuths and distances, with the greatest deviation at the 0.25 m distance. This small discrepancy (less than 3 dB) is most likely due to the KEMAR's neck and torso, which add to the acoustic shadow cast over the far ear. Some of the predicted ILD curves are reproduced in Figure 3.10 to allow for easier comparison.

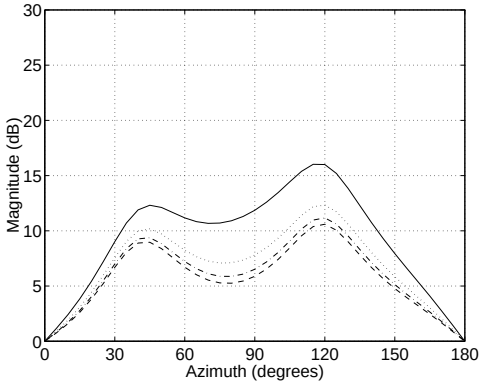
The measured ILDs begin to show greater deviations from the predictions above 1.0 kHz [see Figure 3.9 (f)–(j)]. The expected dip near the interaural axis does develop as expected, but its center is shifted closer to $\theta = 90\,^\circ$, rather than $\theta = 80\,^\circ$ as predicted by the model. The local maxima surrounding the dip move from $\theta = 40\,^\circ$ and $120\,^\circ$ at 1.2 kHz to $\theta = 50\,^\circ$ and $110\,^\circ$ at 2.0 kHz in Figure 3.3 and from $\theta = 40\,^\circ$ and $130\,^\circ$ to $\theta = 60\,^\circ$ and $120\,^\circ$ in Figure 3.9. The peak ILDs from the model for frequencies from 1.2 kHz to 2.0 kHz always appear when the source is toward the rear of the sphere, near $\theta = 120\,^\circ$. The peak in the measured data moves from $\theta \approx 120\,^\circ$ at 1.4 kHz to $\theta \approx 50\,^\circ$ at 1.6 and 1.8 kHz, then back to $\theta \approx 120\,^\circ$ at 2.0 kHz. The exact reason for this movement of the peak is not known. The radius of the head is approximately one quarter of a wavelength at 1.0 kHz (and one half of a wavelength at 2.0 kHz), and thus the deviations of the head from a perfect sphere begin to have an effect at these frequencies.

At 4.0 kHz and above, the measured ILDs differ significantly from the predicted values. Since the wavelength associated with 4 kHz is approximately equal to the radius of the head, the modeling assumption that the head is spherical is inappropriate at higher frequencies. Features of the KEMAR, such

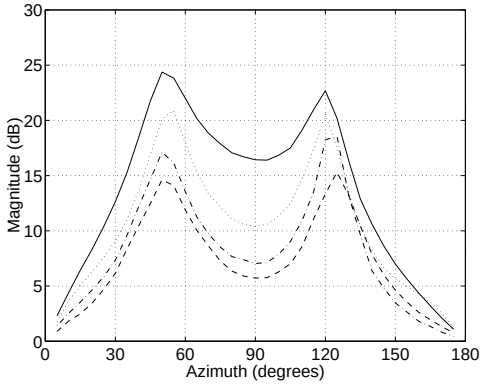


(a) Predicted ILDs: 400 Hz.

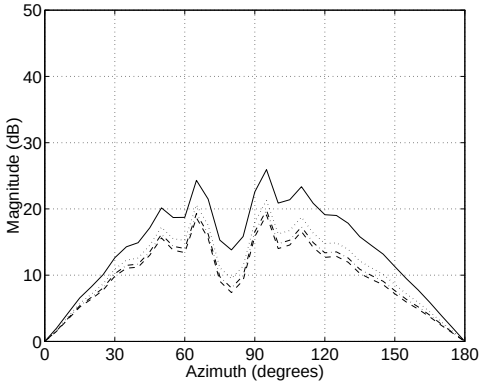
(b) Measured ILDs: 400 Hz.



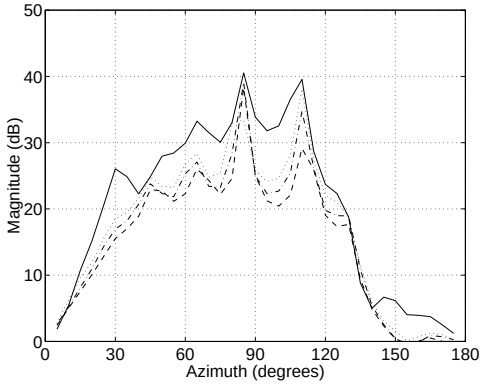
(c) Predicted ILDs: 1600 Hz.



(d) Measured ILDs: 1600 Hz.



(e) Predicted ILDs: 6000 Hz.



(f) Measured ILDs: 6000 Hz.

Figure 3.10: Comparison of predicted and measured ILDs at 400, 1600, and 6000 Hz.

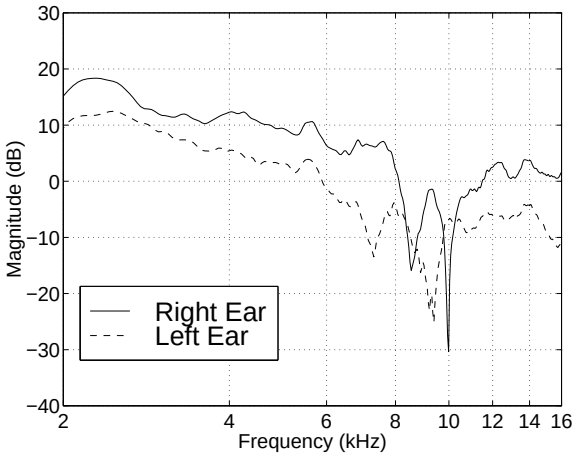
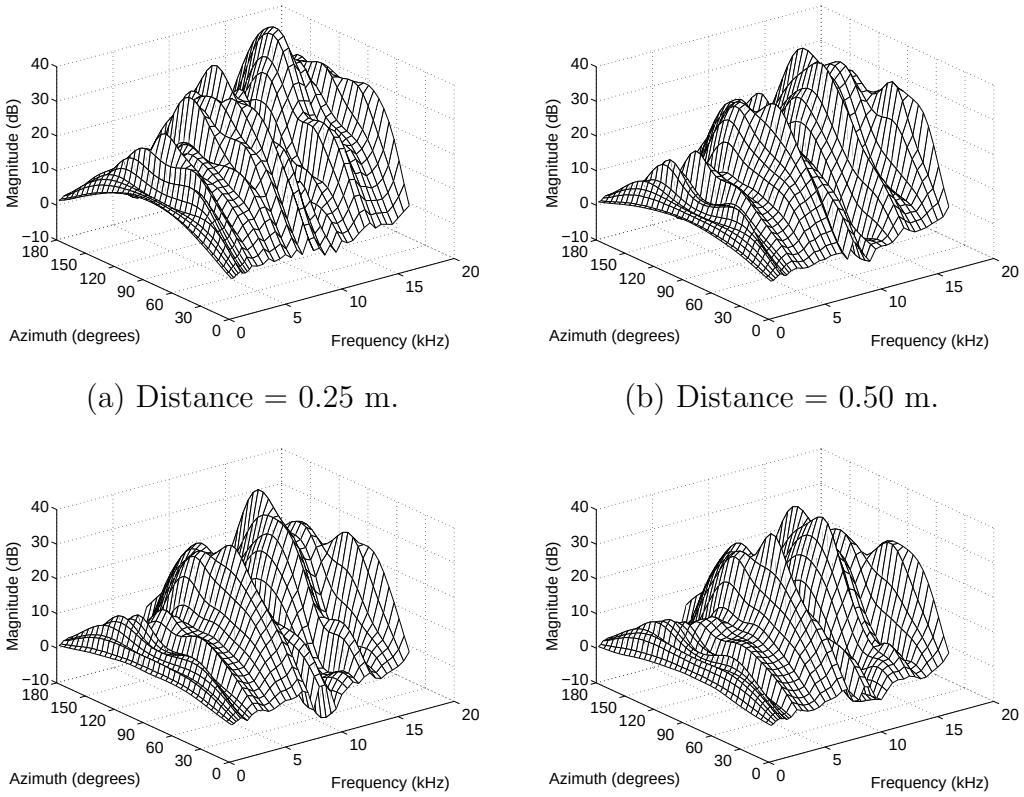


Figure 3.11: Left and right HRTFs measured at 1.00 m for a source at 15 ° azimuth and 0 ° elevation. Note the pinna notch at 10.0 kHz in the right HRTF, which leads to a negative ILD at this frequency. See text for details.

as the nose, are no longer small compared to the wavelength of the impinging sound, and cause the ILDs to become less symmetric.

The 10.0 kHz curves in Figure 3.9 (n) are of particular interest because of the *negative* ILDs which were found for $\theta = 15\,^\circ$ at 1.00 m, $\theta = 25\,^\circ$ at 0.50 m, and $\theta = 35\,^\circ$ at 0.25 m. Since $0\,^\circ \leq \theta \leq 180\,^\circ$ implies a source on the right side of the KEMAR and the ILDs were all calculated as $P_{s,right}/P_{s,left}$ [see (3.1)], a negative ILD is counter-intuitive, as it implies a higher pressure at the *far* side of the head. These values can actually be explained by the plots of the monaural HRTFs in Figure 3.11. The maximum depth of the pinna notch in the HRTF of the right ear is at 10.0 kHz, where the magnitude is actually below that of the left HRTF, causing the negative ILD. It should be noted here that the elevation-specific information provided by the pinna notch, a monaural cue, is embodied in the interaural level difference, a binaural cue. Figure 3.11 only shows the HRTFs specific to the negative ILD at 1.00 m



(a) Distance = 0.25 m.

(b) Distance = 0.50 m.

(c) Distance = 0.75 m.

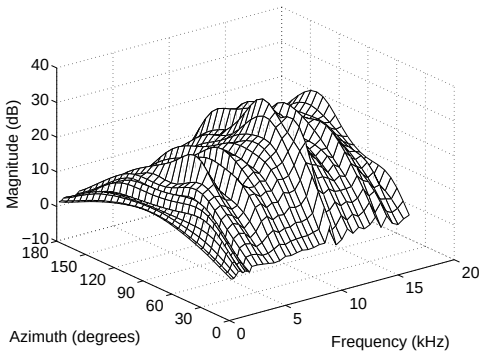
(d) Distance = 1.00 m.

Figure 3.12: Smoothed KEMAR ILD surfaces at -10° elevation.

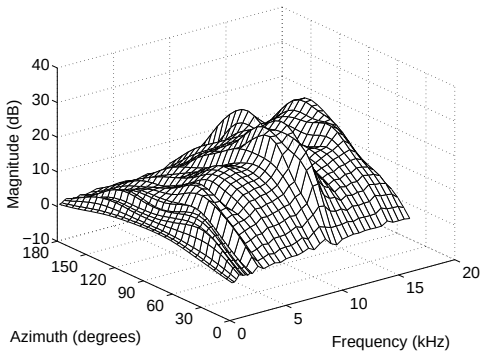
mentioned above.

Despite the deviations of the measured ILDs from the predicted ILDs, it is clear from both that the ILD has a strong dependence on range. Particularly at low frequency, the ILD increases in magnitude at all azimuths as the source distance decreases. Since ILDs are, for the most part, independent of the source spectrum, they may be the most salient, absolute auditory cue to source distance.

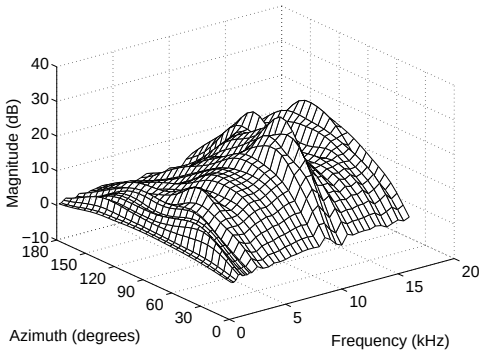
The ILDs shown in Figure 3.9 are all taken from the data measured at



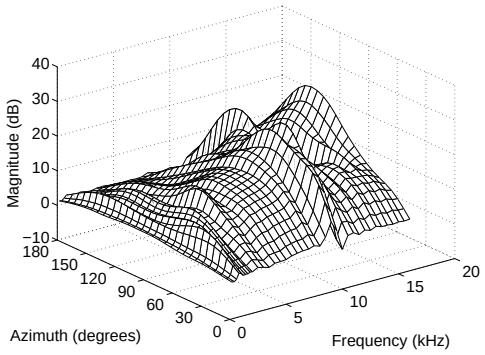
(e) Distance = 0.25 m.



(f) Distance = 0.50 m.

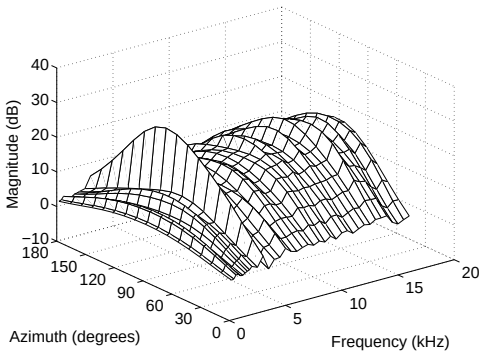


(g) Distance = 0.75 m.

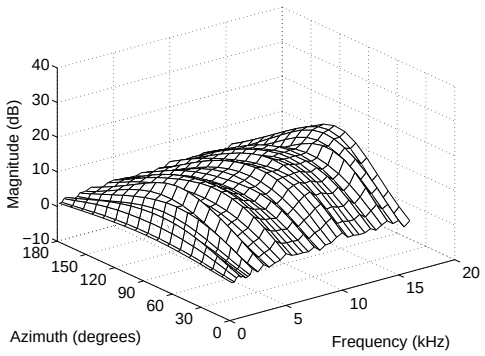


(h) Distance = 1.00 m.

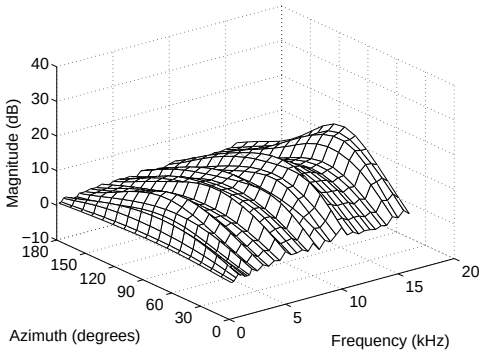
Figure 3.12 (continued): Smoothed KEMAR ILD surfaces at 30 ° elevation.



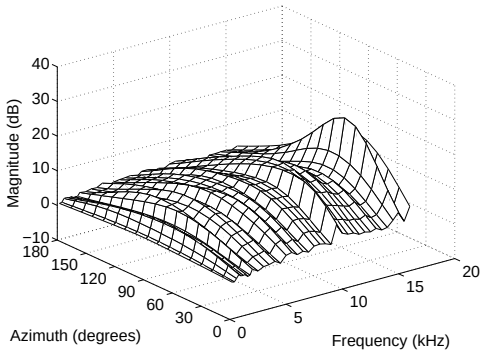
(i) Distance = 0.25 m.



(j) Distance = 0.50 m.



(k) Distance = 0.75 m.



(l) Distance = 1.00 m.

Figure 3.12 (continued): Smoothed KEMAR ILD surfaces at 60 ° elevation.

$0\,^\circ$ elevation to allow for direct comparison with the sphere model predictions. A detailed discussion of the features of the ILD at numerous frequencies for all fourteen elevations used in our measurements is impractical for this thesis. Duda [9] provides an excellent discussion of the dependence of the ILD on elevation (albeit in a different coordinate system). It is important, however, to note that the distance dependence of the ILD exists at *all* elevations. Figure 3.12 contains three-dimensional plots of the ILD surface ($0\,^\circ \leq \theta \leq 180\,^\circ$, 100 Hz $\leq f \leq$ 16000 Hz) at numerous elevations and all distances.[12] The increase in magnitude with decreasing distance is apparent at each elevation shown.

[12]The data in Figure 3.12 have been smoothed with a Gaussian kernel to make the distance trends more clear.

Chapter 4

Three-Dimensional Localization Model

This chapter summarizes our efforts to create a three-dimensional localization model for close-range acoustic sources. Section 4.1 presents a number of previous localization models. Section 4.2 describes our model, a three-dimensional extension of the one used by Duda [9] and Lim and Duda [21] which employs a nearest-neighbor estimation procedure. The results we obtained using our measured KEMAR data with the model are given in Section 4.3.

4.1 Previous Localization Models

4.1.1 One-Dimensional Models

Numerous attempts have been made to create a model which can accurately estimate the location of an unknown acoustic source. The most common of these are "one-dimensional" and focus on lateralization, the ability to localize only the azimuth of the source position.[1] A few are described below.

Jeffress [18] proposed a lateralization model with a simple neural network which responds to an interaural time difference. Other cross-correlation-

[1]It is common in the literature to refer to each component of the source position (azimuth, elevation, range) as a "dimension," despite the fact that changes in azimuth or elevation alone can bring about changes in more than one dimension of space.

based algorithms, e.g., those described by Blauert and Cobben [2], Lindemann [22], and Gaik [11], have also been used in lateralization models, some of which exploit ILDs as well as ITDs. One of the few one-dimensional localization models concerned with range is given by Hirsch [17]. He presents an equation (4.1) which uses the interaural time difference, the interaural intensity difference, and the average of the intensities found at the two ear drums to find the distance from a listener to a sound source at long range.[2]

$$r = \frac{2c\Delta t}{\Delta I / I_{avg}} \tag{4.1}$$

$$\begin{aligned} c &= \text{speed of sound in air} \\ \Delta t &= \text{interaural time difference} \\ \Delta I &= \text{interaural intensity difference} \\ I_{avg} &= \text{average intensity at the two ears} \end{aligned}$$

4.1.2 Multi-Dimensional Models

The literature also contains many two-dimensional models which attempt to estimate the azimuth and elevation of an unknown source. Zakarauskas and Cynader [45] describe one of the few models which exploits only monaural spectral cues. Given a source whose spectrum has a locally constant slope, they apply first- and second-order difference transforms to the spectrum. The source azimuth and elevation are estimated by comparing the results from the transforms to results from the transforms applied to HRTFs from known positions. Wightman et al. [43] create interaural spectral templates for known

[2]Greene [14] later showed that the expected uncertainty dr/r in using this equation can be as high as 45%.

source locations, and use a pattern-recognition scheme to find a best match between ILDs from an unknown source position and one of the templates. Martin [25] generates a spatial likelihood map for an unknown source position from interaural time, level, and phase differences (based on their probability distributions) and uses the global maximum of the map to estimate the azimuth and elevation. Lim and Duda [21] employ a nearest-neighbor approach to localize a source in both azimuth and elevation using ITDs and ILDs. Their model is described in more detail in Section 4.2 below, and is the basis of our three-dimensional model.

At the time of writing, we know of no other model which attempts to estimate the azimuth, elevation, *and* range of a sound source at an unknown location.

4.2 A Three-Dimensional Localization Model Using Binaural Cues

As is mentioned above, the localization model we implemented is based on the two-dimensional model described in [9] and [21] but has been extended to include range as a third dimension. For the two-dimensional case using only ILDs, Duda presents the following statistical estimation method. He first defines the *true* interaural level difference $A(\omega, \theta, \phi)$ as the log ratio of the right and left HRTFs measured for azimuth θ and elevation ϕ, and the *measured* interaural level difference for an unknown source, $A_m(\omega)$. The two are related by the expression

$$A_m(\omega) = A(\omega, \theta, \phi) + N \tag{4.2}$$

where N is a normally distributed noise term representing noise at the ears not related to the source in question. By Bayes' rule

$$p[\theta, \phi | A_m(\omega)] = \frac{p(A_m|\theta, \phi)p(\theta, \phi)}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(A_m|\theta, \phi)p(\theta, \phi)d\theta d\phi} \tag{4.3}$$

and the probability that the source is at the location given by the pair (θ, ϕ) given its measured interaural spectrum is maximized when the difference between A_m and A is minimized. To implement this concept, the model is given a set of vectors, each of the form

$$A(\omega_1, \cdots, \omega_n, \theta_i, \phi_j) = [I(\omega_1)\ I(\omega_2)\ \cdots\ I(\omega_n)] \tag{4.4}$$

for p known azimuths $\theta_1, \cdots, \theta_p$ and q known elevations $\phi_1, \cdots, \phi_q$. $I(\omega_k)$ is the log-magnitude ILD measured at frequency ω_k. When presented with $A_m(\omega_1, \cdots, \omega_n)$ [of the same form as (4.4)] for an unknown source, the model compares the data from the unknown source with the data from each known position in the set using

$$\delta(\theta_i, \phi_j) = \sqrt{\sum_{a=0}^{n} [A_m(\omega_a) - A(\omega_a, \theta_i, \phi_j)]^2} \tag{4.5}$$

Finding the pair (θ_i, ϕ_j) which minimizes δ yields the estimated location of the source.

This model is easily extended to include the ITD as well as the ILD (see [21]). The vector A for each known source position can be made of the form

$$A(\omega_1, \cdots, \omega_n, \theta_i, \phi_j) = [\tau_{i,j}\ I(\omega_1)\ I(\omega_2)\ \cdots\ I(\omega_n)] \tag{4.6}$$

where $\tau_{i,j}$ is the measured ITD for a source at position (θ_i, ϕ_j).[3] (4.5) is still valid, and $\delta(\theta_i, \phi_j)$ contains the difference in ITD, as well as the difference in ILD, between the unknown source and a source at (θ_i, ϕ_j).

The extension of the model from two to three dimensions does not require any changes to the algorithm itself. As is discussed in Section 3.4.2, the ILD has a strong dependence on range. Specifically, the ILD generated by a source at a given position has been shown to increase monotonically as source distance decreases, particularly at lower frequencies. An example can be seen in Figure 4.1. As the source distance decreases from 1.00 m to 0.25 m, the shape of the ILD remains roughly constant while the magnitude increases across the frequency spectrum shown. This indicates that, given estimates of azimuth and elevation, it should be possible to determine the range of a source based on the magnitude of the ILD spectrum. Since the model described above has been shown to accurately estimate azimuth and elevation, and the vector of interaural differences contains ILD magnitudes which allow for discrimination between sources at two distinct positions differing only in distance, it is logical to assume that the model can localize distance as well. Thus, given a set of vectors each of the form

$$A(\omega_1, \cdots, \omega_n, \theta_i, \phi_j, r_k) = [\tau_{i,j,k}\ I(\omega_1)\ I(\omega_2)\ \cdots\ I(\omega_n)] \tag{4.7}$$

describing source positions which vary in azimuth, elevation, *and* distance, and a similar vector describing a source at an unknown location, we can use the above model to estimate all three components of the source position.

[3]Since the ITD shows little dependence on frequency, only one value ($\tau_{i,j}$) is needed for the source position (θ_i, ϕ_j).

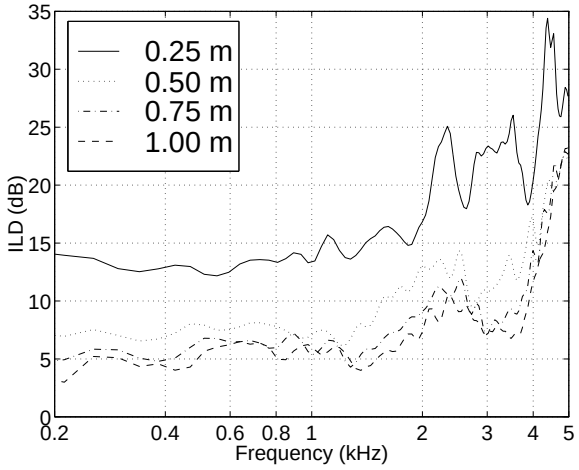


Figure 4.1: ILDs at $\theta = 90\,^\circ, \phi = 0\,^\circ$ for four distances. Note the monotonic increase in ILD with decreasing source distance.

Our implementation of the model utilizes the close-range HRTF measurements described in Chapter 2. We measured 2656 HRTFs, 102 of which are unusable in the model because they correspond to positions in the median plane ($\theta = 0\,^\circ, 180\,^\circ$) where the interaural differences are necessarily zero. This leaves 1227 pairs of HRTFs corresponding to 1227 positions in the right hemisphere around the head. At this point, it is convenient to introduce the terminology used by Duda to describe the sets of vectors, those for known source positions and of unknown positions, used with the model. The vectors associated with known source positions will be referred to as the training set, and those associated with unknown positions the test set.[4] The two sets were created as follows. Starting with the $0\,^\circ$ azimuth position at each elevation and each distance, we chose every other measurement position. For each chosen po-

[4]This terminology comes from the literature pertaining to neural networks. The model is effectively "trained" with data from known source locations, and "tested" with data from unknown locations.

sition, a vector was added to the training set with the following values: the ITD $\tau_{i,j,k}$ taken from the smoothed ITD values (see Section 3.4.1); the ILD values at 21 logarithmically spaced frequencies from 200 Hz to 4.0 kHz extracted from the log-magnitude ratio of the right and left HRTFs measured at (θ_i, ϕ_j, r_k). For the points not chosen in this process, data were added to the training set by averaging the vectors of the left and right neighbors of that position. For example, the vector representing $\theta = 5\,^\circ, \phi = 0\,^\circ, r = 0.50$ m contains ITD and ILD values which are averages of the values from $\theta = 0\,^\circ, \phi = 0\,^\circ, r = 0.50$ m and $\theta = 10\,^\circ, \phi = 0\,^\circ, r = 0.50$ m. The training set contains a total of 1227 vectors. The test set is made up of vectors containing the *measured* ITD and ILD values for the positions represented by *average* values in the training set. For example, the vector representing $\theta = 5\,^\circ, \phi = 0\,^\circ, r = 0.50$ m in the *training set* contains average values as described above, while the vector representing that same position in the *test set* contains measured ITD and ILD values extracted from the appropriate HRTFs. The test set contains 612 vectors. The positions in the two sets can be seen in Figure 4.2. An 'x' on the sphere represents a point for which the corresponding vector in the training set contains measured ITD and ILD values. An 'o' on the sphere represents a point for which the corresponding vector in the training set contains ITD and ILD values found by averaging its left and right neighbors, and also represents a point in the test set.

Results were generated by using all 612 test set vectors as input to the model, one at a time.[5] For each vector, the model calculated δ for all 1277

[5]Since the interaural cues used in the vectors in the test set are extracted from the ratio

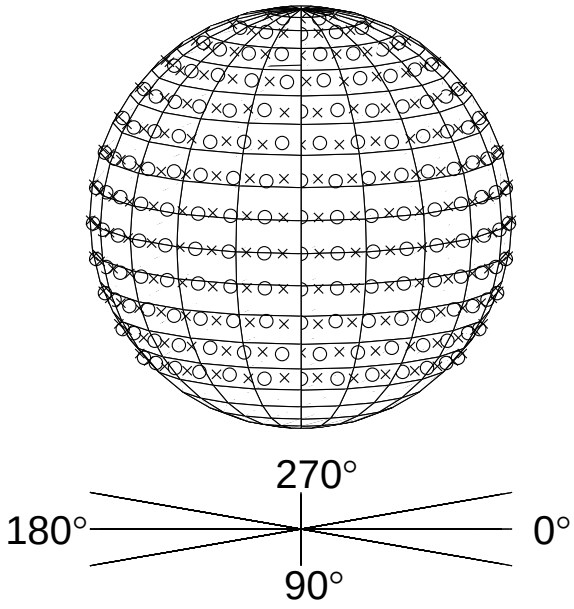


Figure 4.2: Spatial positions used in the localization model, as seen from $\theta = 90\,^\circ, \phi = 0\,^\circ$. See text for details. This figure shows only the points at one distance; the pattern of x's and o's is the same for all distances.

positions in the training set using (4.5), and returned the position (θ_i, ϕ_j, r_k) corresponding to the minimum δ. Three points about the execution of the model should be noted here. First, since each position represented in the test set is also represented in the training set, it is possible for the model to localize each position with zero error. Second, any localization error is necessarily quantized to the increments used in the HRTF measurements (see Chapter 2). Third, the evaluation of (4.5) requires that all of the components of $A_m(\omega)$ and $A(\omega, \theta_i, \phi_j, r_k)$ have the same units to allow for valid distance calculations. This was accomplished by multiplying the ITD values ($\tau_{i,j,k}$) by a weighting factor $b = \frac{1dB}{50\mu s}$, also known as a 'time-intensity trading ratio,'

of the right and left HRTFs at each point, and the HRTFs are found by applying a Fourier transform to the HRIR, the test source is implicitly assumed to be an impulse.

to convert from microseconds to decibels. The value for b was obtained from results of studies which investigated the ILD value necessary to counteract a perceived lateralization due to a specific ITD. See [5] and [30] for details.

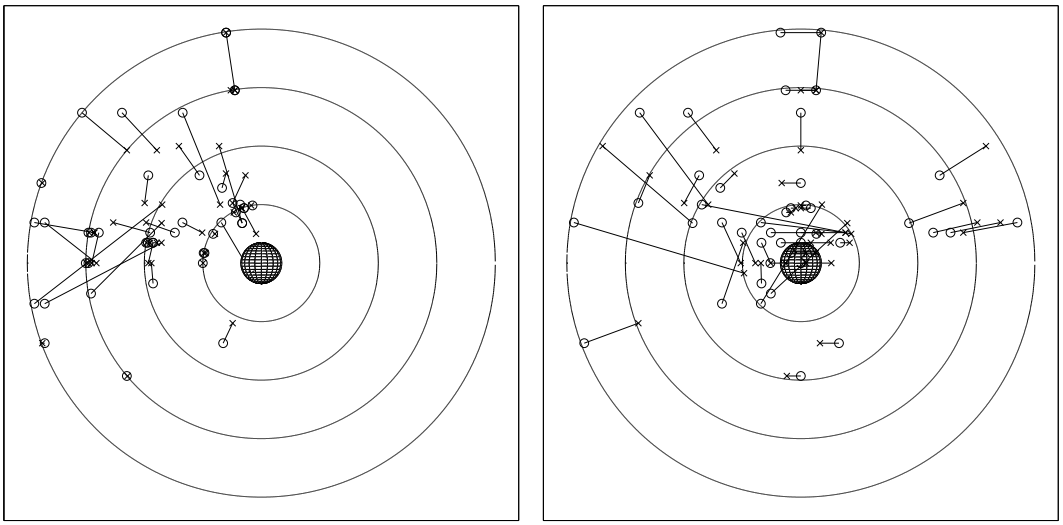
4.3 Results

4.3.1 Performance of the Model

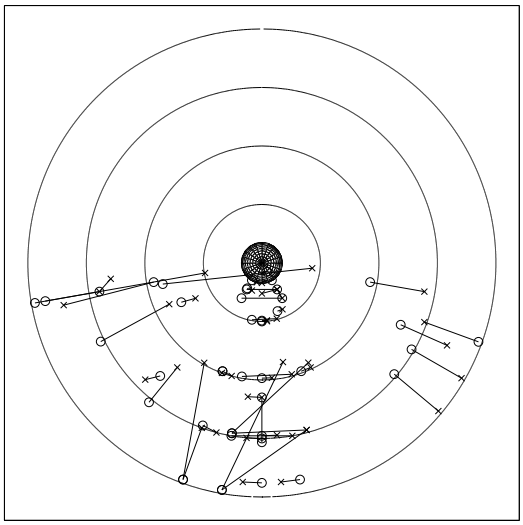
The results from our localization model are quite good. Only 45 out of the 612 source positions in the test set were incorrectly localized, for a success rate of nearly 93 percent. The mislocated points are shown in Figure 4.3.

The literature on sound-localization models contains numerous methods for quantifying localization error. Some authors choose to report the average angular error for each dimension (azimuth, elevation) separately (e.g., see [21]). Others report the angle subtended by the great circle which includes the actual and estimated position (e.g., see [45]). Unfortunately, most of the methods are difficult to apply to the results of a localization model which includes range. An angular error ϵ at a distance $r = 0.50$ m represents a smaller separation between the actual and estimated positions than does the same error at a distance $r = 1.00$ m. Great circle arcs cannot be found between points on spheres of different radii. In addition, given our coordinate system, an angular error ϵ at an elevation $\phi = 80\,^\circ$ represents a smaller separation between the actual and estimated positions than does the same error at an elevation $\phi = 0\,^\circ$.

Because of these difficulties, we have chosen to report three types of localization error:



(a) Front view ($\theta = 0\,^\circ$, $\phi = 0\,^\circ$). (b) Side view ($\theta = 90\,^\circ$, $\phi = 0\,^\circ$).



(c) Top view ($\theta = 0\,^\circ$, $\phi = 90\,^\circ$).

Figure 4.3: Source positions incorrectly localized by the model. For each case, a line connects the actual source position, marked with an 'o', to the estimated position, marked with an 'x'. The sphere in the center represents the listener's head, and the concentric circles represent the four measurement distances.

1. ϵ_L: the linear distance between the actual and estimated source positions divided by the linear distance from the center of the listener's head to the actual source position (r);

2. ϵ_I: the number of measurement *increments* (azimuth, elevation, and range) between the actual and estimated source positions;[6]

3. ϵ_A: the angle subtended by the great circle arc which includes the actual and estimated position, *only for cases when the ranges of the actual and estimated source positions are equal.*

ϵ_L provides an error value as a percentage of the actual source distance. ϵ_I can be calculated for each dimension (notated $\epsilon_{I,\theta}$, $\epsilon_{I,\phi}$, and $\epsilon_{I,r}$) to separately examine errors in azimuth, elevation, and range. ϵ_A can (and should) be calculated for each measurement distance to see if the model's accuracy is a function of distance. Error measurements of each type can be found in Table 4.1. Averages were taken over all 612 unknown locations tested.

An examination of the data in Table 4.1 reveals two particularly obvious points. First, the model made significantly more and larger errors in azimuth than in elevation or range. This is most likely due to the fact that the increments in azimuth are generally smaller than those in elevation or range. Second, the model made significantly fewer errors for sources at $r = 1.00$ m (3) than for sources at $r = 0.25, 0.50$, or 0.75 m (8,6, and 10 errors, respectively). Because we used a constant angular spacing rather than a constant arc

[6]See Table 2.1 for the azimuth increments. One elevation increment equals 10 °. One range increment equals 0.25 m.

Error Type	Number	Average	Units
ϵ_L	45	2.22	percent
$\epsilon_{I,\theta}$ (azimuth)	28	0.14	increments
$\epsilon_{I,\phi}$ (elevation)	13	0.03	”
$\epsilon_{I,r}$ (range)	18	0.04	”
$\epsilon_A,\ r = 0.25\ m$	8	0.43	degrees
$\epsilon_A,\ r = 0.50\ m$	6	0.43	”
$\epsilon_A,\ r = 0.75\ m$	10	0.70	”
$\epsilon_A,\ r = 1.00\ m$	3	0.13	”

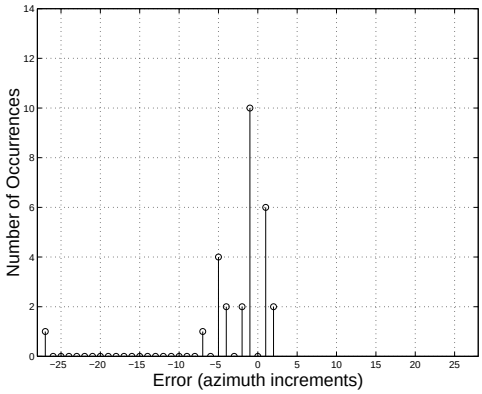
Table 4.1: Average localization errors.

length at the four measurement distances, the distance between measurement positions increases as the distance increases. Thus, since adjacent positions at $r = 1.00$ m are farther apart than adjacent positions at $r = 0.25$ m (for example), the interaural cues are more distinct for adjacent positions at larger distances and the model is less prone to error at larger distances.

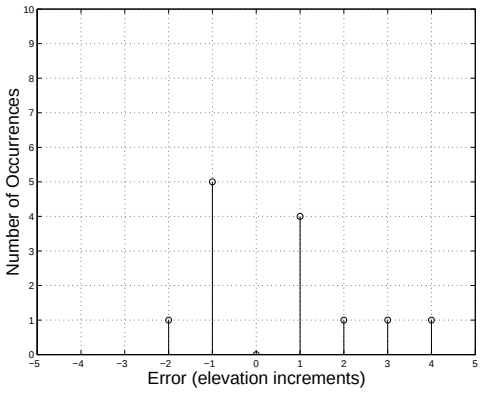
Distributions of each single-dimension increment error are shown in Figure 4.4. A negative error indicates the estimated source dimension is less than the actual dimension; a positive error indicates the estimated source dimension is greater than the actual dimension. For example, comparison of an estimated position of $\theta = 5\,^\circ, \phi = 0\,^\circ, r = 0.50$ m, with an actual position of $\theta = 0\,^\circ, \phi = 30\,^\circ, r = 1.00$ m yields a positive azimuth error (1 increment), and negative elevation and distance errors (-3 and -2 increments, respectively).

4.3.2 Comparison with Human Localization Abilities

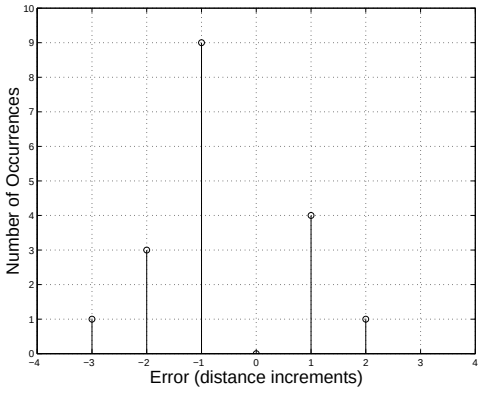
Direct comparison of the model's performance with measured human localization abilities is somewhat difficult due to the variability in localization exper-



(a) Azimuth error.



(b) Elevation error.



(c) Distance error.

Figure 4.4: Distributions of single-dimension increment errors.

iments. Because human localization accuracy is a function of source position and spectral content, experiments are often limited to small sections of auditory space or to sources with particular spectral properties. For example, localization of azimuth is often measured only for sources restricted to the horizontal plane, while localization of elevation is often measured only for sources restricted to the median plane. Source spectra vary from single frequencies to broadband noise.

Despite the disparity in experimental conditions, four major trends in human localization are easily identified. First, localization performance is best for broadband signals and degrades with decreasing bandwidth. Second, accuracy is best for frontal sources near the intersection of the median and horizontal planes, and decreases as the source moves toward the interaural axis and/or toward the extremes of elevation. Third, estimates of azimuth are more accurate than estimates of elevation. Fourth, estimation is the least accurate for range, with source distance generally underestimated [1, 5, 29, 44]. Given ideal conditions, human localization accuracy is on the order of 1° in azimuth and 10° in elevation. Accuracy in distance estimation, particularly for close-range sources, is not well documented. Given the azimuth increment of 5° in the horizontal plane, the 10° elevation increment, and the 0.25 m distance increment used in our measurements, the incremental errors $\epsilon_{I,\theta}, \epsilon_{I,\phi}$, and $\epsilon_{I,r}$ shown in Table 4.1 correspond to average measured errors of 0.7° in azimuth, 0.3° in elevation, and .01 m in range, respectively.

Chapter 5

Summary and Conclusions

This thesis was undertaken with three goals in mind: to measure the head-related transfer function at close range; to analyze the distance dependence of interaural localization cues extracted from the HRTF measurements; and to create a computer-based model capable of accurate, three-dimensional localization of a close-range sound source. The remainder of this chapter summarizes the results and conclusions drawn from the work done toward these goals.

5.1 Measurements

As is discussed in chapter 2, we have collected a set of 2656 close-range head-related transfer functions. The HRTFs were measured in the right ear of a KEMAR manikin at distances of 0.25, 0.50, 0.75, and 1.00 m. The measurements cover a full 360° of azimuth at elevations from −40° to 90° in 10° increments. The measurement process was adapted from the work of Gardner and Martin [12], and uses a maximum length sequence as a broad band noise source. In order to avoid source location ambiguities in the measurements, we used a loudspeaker with a concentric woofer and tweeter which served effectively as a point source at short distances.

Two problems with the measurements became clear during the analysis phase. First, the delay intrinsic to the computer-based data-acquisition system used to measure the HRTFs varied, causing difficulties in the analysis of interaural time differences (see Sections 3.1.2, 3.4.1, and 5.2). It is still not clear whether this was solely the result of not disabling interrupts within the acquisition software, or whether the DAQ card used was somehow defective. Since similar measurements have been made by others without this problem being reported, it is clear that this problem can be and should have been avoided. Second, the proximity of the loudspeaker to the manikin at the closer measurement distances combined with the relatively long time scale of the MLS resulted in standing waves between the speaker and the KEMAR. These standing waves caused a comb filter-like effect in the HRTFs which can be seen as periodic peaks and notches in the high-frequency region. Since the interaural level differences used in our model are generally lower in frequency than the comb filtering, this seems to have caused no adverse effects on the localization performance of the model. Brungart [3] describes a measurement system which is immune to this problem due to a smaller sound source, and thus may be more suitable for close-range HRTF measurements.

We did hope that the HRTF measurements would be useful in future localization research, and could be used in perception experiments with human listeners. At this point it is not clear whether either is possible because of the problems described above.

5.2 Analysis

Chapter 3 describes the extraction of interaural time and level differences from the HRTF measurements, a comparison of these cues to predicted values, and an analysis of the distance dependence of these cues. ITDs were calculated by finding the slope of the unwrapped phase of the interaural transfer function; ILDs were calculated as the log-magnitude of the interaural transfer function. The measured ITDs are in relatively good agreement with those predicted by a spherical head model, and show only a weak distance dependence. This dependence is most likely due to the increase in curvature of a spherically spreading wave with decreasing distance from the sound source. The measured ILDs are generally larger than the predicted values, but the predicted dependence on azimuth is quite clear in the measured data, particularly at frequencies below 2 kHz. Disagreement at higher frequencies was not unexpected; modeling the head as a rigid sphere is only valid for low frequencies whose wavelengths are larger than the radius of the head. The measured ILDs show a strong distance dependence at close range: a monotonic decrease in distance results in a monotonic increase in ILD.

5.3 Modeling

The localization model described in Chapter 4 is an extension of the one described by Duda in [9] and Lim and Duda in [21]. It uses a nearest-neighbor estimation procedure to localize an unknown source in azimuth, elevation, and range. Given a vector containing a single ITD and ILD values for 21 logarithmically spaced frequencies from 200 Hz to 4 kHz for a source at an un-

known location, the model finds the difference between this vector and similar vectors from known source locations. The estimated azimuth, elevation, and range correspond to the vector of known position with the minimum difference. The model was provided with 1277 vectors from known locations in the right hemisphere at distances of 0.25, 0.50, 0.75, and 1.00 m. It was tested with 612 unknown source locations, also in the right hemisphere and at the same distances, and correctly localized nearly 93 percent of them. The average error in localization, found by dividing the linear distance between the actual and estimated positions by the distance of the actual position from the head was 2.2 percent. It is clear from this that there is enough information in the ITD and low-frequency ILD to accurately localize a source in three dimensions.

5.4 Future Work

The most obvious demand for future work that arises from this thesis comes from the fact that our localization model does not compare well with human abilities. Our model localizes more accurately than does the average human, particularly in range. It does not seem to suffer the front-back confusions that are reported in many studies of human listeners. Part of this is clearly due to the fact that the modeled source locations (both known and unknown) are restricted to a quantized set of positions, while natural sources can be located anywhere. Also, the model assumes an anechoic listening environment and a single, stationary, impulsive source, all of which are ideal conditions for localization. While the human cochlea must process spectral information in bands, our model exploits single-frequency interaural level differences. The human head and torso are also significantly more asymmetric than the KE-

MAR. Many of these factors can and should be incorporated into the model to help understand human localization in three-dimensional space.

Despite the problems described in Section 5.2 above, it would be worthwhile to test our measured HRTFs on human listeners. It is possible that the jitter in the ITD will not have an adverse effect on three-dimensional audio generated with our HRTFs and delivered over headphones. If this is the case, the measurements could be valuable for use in the virtual simulation of small environments such as automobiles or airplane cockpits.

Bibliography

[1] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization.* MIT PRESS, Cambridge, 1983.

[2] J. Blauert and W. Cobben. Some consideration of binaural cross correlation analysis. *Acustica*, 39(2):96–104, 1978.

[3] D.S. Brungart. *Near-Field Auditory Localization.* PhD thesis, Massachusetts Institute of Technology, 1998. Supervised by Dr. N.I. Durlach.

[4] D.S. Brungart and W.R. Rabinowitz. Auditory localization in the near field. Third International Conference on Auditory Display, 1996.

[5] S. Carlile, editor. *Virtual Auditory Space: Generation and Applications.* R.G. Landes Company, Austin, 1996.

[6] P.D. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin.*, 60:302–315, 1963.

[7] P.D. Coleman. Dual role of frequency spectrum in determination of auditory distance. *J. Acoust. Soc. Am.*, 44(2):631–632, 1968.

[8] O. Cramer. The variation of the specific heat ratio and the speed of sound in air with temperature, pressure, humidity, and CO_2 concentration. *J. Acoust. Soc. Am.*, 93(5):2510–2516, 1993.

[9] R.O. Duda. Elevation dependence of the interaural transfer function. In R. Gilkey and T. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, Mahwah, 1997.

[10] R.O. Duda and W.L. Martens. Range–dependence of the HRTF for a spherical head. Submitted to the 1997 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics.

[11] W. Gaik. Combined evaluation of interaural time and intensity differences: Psychoacoustic results and computer modeling. *J. Acoust. Soc. Am.*, 94(1):98–110, 1993.

[12] W. Gardner and K. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical Report 280, MIT Media Lab Perceptual Computing, 1994.

[13] S.W. Golomb. *Shift Register Sequences*. Aegean Park Press, Laguna Hills, 1982.

[14] D.C. Greene. Comments on 'Perception of the range of a sound source of unknown strength'. *J. Acoust. Soc. Am.*, 44(2):34, 1968.

[15] E.R. Hafter and C. Trahiotis. Functions of the binaural system. In Malcolm J. Crocker, editor, *Encyclopedia of Acoustics*. John Wiley and Sons, New York, 1997.

[16] H.L. Han. Measuring a dummy head in search of pinna cues. *J. Audio Eng. Soc.*, 42(1/2):15–36, 1994.

[17] H.R. Hirsch. Perception of the range of a sound source of unknown strength. *J. Acoust. Soc. Am.*, 43(2):373–374, 1968.

[18] L.A. Jeffress. A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, 41:35–39, 1948.

[19] L. Kinsler, A. Frey, A. Coppens, and J. Sanders. *Fundamentals of Acoustics*. John Wiley and Sons, New York, 1982.

[20] J.C.R. Licklider. A duplex theory of pitch perception. In E. D. Schubert, editor, *Psychological Acoustics*. Dowden, Hutchingson, and Ross, Inc., Stroudsburg, 1979.

[21] C. Lim and R.O. Duda. Estimating the azimuth and elevation of a sound source from the output of a cochlear model. In *Proceedings of the 28th Asilomar Conference on Signals, Systems, and Computers*, pages 399–403, 1994.

[22] W. Lindemann. Extension of a binaural cross–correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6):1608–1622, 1986.

[23] R.Y. Litovsky and D.H. Ashmead. Development of binaural and spatial hearing in infants and children. In R. Gilkey and T. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, Mahwah, 1997.

[24] E. Lopez-Poveda and R. Meddis. A physical model of sound diffraction

and reflections in the human concha. *J. Acoust. Soc. Am.*, 100(5):3248–3259, 1996.

[25] K.D. Martin. A computational model of spatial hearing. Master's thesis, Massachusetts Institute of Technology, 1995.

[26] S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *J. Acoust. Soc. Am.*, 61(6):1567–1576, 1977.

[27] D.H. Mershon. Phenomenal geometry and the measurement of perceived distance. In R. Gilkey and T. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, Mahwah, 1997.

[28] D.H. Mershon and J.N. Bowers. Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, 8:311–322, 1979.

[29] J.C. Middlebrooks and D.M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42:135–159, 1991.

[30] B.C.J. Moore. *Introduction to the Psychology of Hearing*. University Park Press, Baltimore, 1977.

[31] P.M. Morse. *Vibration and Sound*. Acoustical Society of America, 1995.

[32] P.M. Morse and K.U. Ingard. *Theoretical Acoustics*. Princeton University Press, 1968.

[33] A.V. Oppenheim and R.W. Schafer. *Discrete–Time Signal Processing*. Prentice Hall, Englewood Cliffs, 1989.

[34] J.O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press, London, 2nd edition, 1988.

[35] W. M. Rabinowitz, J. Maxwell, Y. Shao, and M. Wei. Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere. *Presence*, 2:125–129, 1993.

[36] Lord Rayleigh. On our perception of sound direction. *Phil. Mag.*, 13:214–232, 1907.

[37] D.D. Rife and J. Vanderkooy. Transfer–function measurements using maximum–length sequences. *J. Audio Eng. Soc.*, 37(6):419–444, 1989.

[38] T. Sone, Y. Suzuki, S. Takane, and K. Suzuki. Distance perception in sound localization and its control by simulation of head-related transfer functions. In *Proceedings of the 14th ICA, L-7-4*, 1992.

[39] Tannoy, Ltd. *Reference Monitors System 600 User Manual.*

[40] J. Vanderkooy. Aspects of MLS measuring systems. *J. Audio Eng. Soc.*, 42(4):219–231, 1994.

[41] G. von Békésy. *Experiments in Hearing.* McGraw – Hill, New York, 1960.

[42] F.L. Wightman and D.J. Kistler. Factors affecting the relative salience of sound localization cues. In R. Gilkey and T. Anderson, editors, *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum Associates, Mahwah, 1997.

[43] F.L. Wightman, D.J. Kistler, and M.E. Perkins. A new approach to the study of human sound localization. In W.A. Yost and G. Gourevitch, editors, *Directional Hearing*. Springer–Verlag Inc., New York, 1987.

[44] P. Zahorik. Auditory distance perception: A literature review. In partial fulfillment of Univ. of Wisconsin - Madison Dept. of Psychology prelim. exam. requirement, 1996.

[45] P. Zakarauskas and M. S. Cynader. A computational theory of spectral cue localization. *J. Acoust. Soc. Am.*, 94(3):1323–1331, 1993.

[46] E. Zwicker and H. Fastl. *Psychoacoustics: Facts and Models*. Springer–Verlag Inc., Berlin, 1990.

Vita

Paul Thomas Calamia was born in New Brunswick, NJ, on August 5, 1970, to Thomas and Lois Calamia. After graduating from East Brunswick High School in 1988, he matriculated at Duke University in Durham, NC. He graduated cum Laude in 1992 with a Bachelor of Science Degree in mathematics, and began working for the First Boston Corporation (now Credit Suisse First Boston). In 1993 he accepted a position as an acoustical engineer with Wyle Laboratories in Arlington, VA, and in 1995 entered the Engineering Acoustics Program at the University of Texas at Austin.

Permanent address: 12 South Drive
East Brunswick, New Jersey 08816

This thesis was typeset with LaTeX[‡] by the author.

[‡]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.