A Computational Model of Spatial Hearing

by

Keith Dana Martin

B.S. (distinction), Cornell University (1993)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1995

© Massachusetts Institute of Technology 1995. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
May 18, 1995

Certified by . . . .
Barry L. Vercoe
Professor of Media Arts and Sciences
Thesis Supervisor

Certified by . . . .
Patrick M. Zurek
Principal Research Scientist
Research Laboratory of Electronics
Thesis Supervisor

Accepted by . . . .
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

A Computational Model of Spatial Hearing
by
Keith Dana Martin

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 1995, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

The human auditory system performs remarkably at determining the positions of sound sources in an acoustic environment. While the localization ability of humans has been studied and quantified, there are no existing models capable of explaining many of the phenomena associated with spatial hearing.

This thesis describes a spatial hearing model intended to reproduce human localization ability in both azimuth and elevation for a single sound source in an anechoic environment. The model consists of a front end, which extracts useful localization cues from the signals received at the eardrums, and a probabilistic position estimator, which operates on the extracted cues. The front end is based upon human physiology, performing frequency analysis independently at the two ears and estimating interaural difference cues from the resulting signals. The position estimator is based on the maximum-likelihood estimation technique.

Several experiments designed to test the performance of the model are discussed, and the localization blur exhibited by the model is quantified. A "perceptual distance" metric is introduced, which allows direct localization comparisons between different stimuli. It is shown that the interaural intensity difference (IID) contains sufficient information, when considered as a function of frequency, to explain human localization performance in both azimuth and elevation, for free-field broad-band stimuli.

Thesis Supervisor: Barry L. Vercoe
Title: Professor of Media Arts and Sciences

Thesis Supervisor: Patrick M. Zurek
Title: Principal Research Scientist
Research Laboratory of Electronics

Acknowledgments

First, I am grateful to Barry Vercoe for his generous support of this work. His Machine Listening (a.k.a "Synthetic Listeners and Performers", "Music and Cognition", and "Bad Hair") group at the MIT Media Laboratory is a truly wonderful working environment. I can not imagine a place where I would be more free to choose my own direction with such fantastic support.

Second, I am grateful to Pat Zurek for agreeing to serve as a co-advisor on this thesis. I am particularly thankful for his willingness to listen to my naive blathering about spatial hearing, and for his patience in correcting my misconceptions. I only wish that my own motivation and ego had not kept me from following his advice more closely.

I wish to thank all of the members of the Machine Listening Group. Particularly, I wish to thank Bill Gardner for his collaboration on the KEMAR measurements, Dan Ellis for his unfailing willingness to debate issues in perceptual modeling and signal processing, and Eric Scheirer for giving me comments on some particularly horrible drafts of various papers, including this thesis. I must also thank Nicolas Saint-Arnaud for being an understanding officemate over the last two years (which couldn't have been easy). Thanks are also due to Adam, Andy, Judy, Judy, Matt, and Mike, for making the work environment at the Media Lab more enjoyable in general.

For financial support, I am indebted to the National Science Foundation, whose graduate fellowship has supported me during my tenure at MIT.

I give special thanks to my parents, for always encouraging me to do great things (even though the list of "great things" specifically omitted playing in a rock 'n roll band). Without their years of encouragement, it is unlikely that I would ever have survived through four years at Cornell, and I would certainly never have arrived at MIT.

Finally, my deepest thanks go to Lisa, for her emotional support over the last eight years. Without her commitment to our relationship and to both of our careers, neither of us could possibly be as happy as we are today.

Contents

Chapter 1

Introduction

1.1 Motivation

Sound is an extremely useful medium for conveying information. In addition to explicit semantic information such as the meaning of words used in speech, the acoustic signals reaching our eardrums carry a wealth of other information, including emotional content conveyed by pitch contour and voice "tone," and physical information such as the size and constituent materials of the sound producer.

Information about the locations of sound sources relative to the listener is also contained in the signals received at the eardrums. The ability to extract and use this information is vital to the survival of many animals, some of which use the information to track prey, and others to avoid being preyed upon.

The methods by which position information is extracted from eardrum signals are not well understood. Thus, the research described in this thesis is principally driven by the question: "What cues are useful for determining sound source position, and how might they be extracted and applied to position estimation?"

There is currently a great deal of interest in "virtual sound environments" and simulation of "acoustic spaces." Much research is devoted to the construction of systems to deliver sound events to the eardrums of listeners. These systems range from mundane (designing a public address system for an office building), to practical (simulating the experience of being in a concert hall long before the hall is built, and potentially saving millions of dollars by detecting acoustic flaws early in the architectural design process), to fantastic (generating "holographic" audio in applications ranging from "virtual reality" to advanced acoustic information displays). With all of these applications, it is desirable to be able to predict the perception of listeners – for example, to determine the intelligibility of speech presented over the office building's public address system, to determine whether a concert hall design will have a good "spatial impression" for performances of particular styles of music, or to test the effectiveness of various algorithms and coding schemes used in an acoustic display. A computational model of spatial hearing will be useful as a quantitative tool for testing designs in a domain where qualitative human judgment is a more expensive and less precise norm.

Yet another important application of a spatial hearing model lies in the study of *auditory scene analysis*, the process by which a listener makes sense of the sound environment. The fundamental problems in auditory scene analysis include the formation and tracking of acoustic sources, and the division of a received sound signal into portions arising from different sources. A spatial hearing model may provide useful cues for grouping of sound events, as common spatial location is a potential grouping cue.

1.2 Goals

There are several immediate goals in the construction of a spatial hearing model. First, it is desirable to design a system that extracts useful localization cues from signals received at a listener's eardrums.[1] Ideally, the extraction of cues should be robust for arbitrary source signals in the presence of interference from reverberant energy and from "distractor" sound sources. Additionally, it is desirable that the extracted cues be the same (or at least similar to) cues extracted by the human auditory system.

The model should combine information from various cues in some reasonable manner. It may be desirable to combine cues in a manner that results in a suboptimal position estimate if the behavior of the resulting model is similar to behavior exhibited by humans in localization tasks. Finally, it is desirable that the cue extraction and position estimation be interpretable in terms of well-understood signal processing, pattern recognition, and probabilistic estimation theories.

The main goal of this thesis is to demonstrate a spatial hearing model that can, by examination of the eardrum signals, successfully determine the position of an arbitrary source signal presented in free-field. The model should be able to estimate both azimuth (horizontal position) and elevation (vertical position) with precision similar to humans.[2] The model should be extensible to the localization of multiple sound sources in both free-field and reverberant environments. To that end, an intermediate goal of the current project is to provide a mechanism which simulates the "precedence effect" that is exhibited by human listeners [40].

1.3 An outline of this thesis

This is chapter one, the **Introduction**, which describes the motivation and goals driving the current research.

Chapter 2, **Background**, gives a brief overview of the relevant results in spatial hearing research, particularly in regard to the "cues" used for localization. It then presents capsule reviews of recent computational models of localization.

Chapter 3, **Form of the Model**, presents a overview of the proposed model, along with summary descriptions of the component parts. It then presents the individual pieces of the model in sequence, describing the front end of the model from a description

[1] By signals at the eardrum, we mean acoustic signals measured anywhere in the ear canal, since the transformation performed by the ear canal is not direction-dependent. When interaural differences are considered, however, it becomes important that the *same* measurement position be used for both ears, and the eardrum is used as a convenient reference point.

[2] The precise meanings of azimuth and elevation depend, in general, on the specific coordinate system. A description of the coordinate systems used in this thesis may be found in Appendix A.

of eardrum signal synthesis to a detailed description of interaural-differences estimation. Finally, it describes the form of the probabilistic position estimator and details its construction.

Chapter 4, **Noise, Bias, and Perceptual Distance**, considers deterministic and non-deterministic "noise" in the model. It then introduces a metric for "perceptual distance," which is used throughout Chapter 5.

Chapter 5, **Comparisons with Human Performance**, presents the results of several experiments designed to test the output of the model. It describes the localization performance of the model for noise signals and for impulsive stimuli, and quantifies the notion of "localization blur." Finally, it describes the results of two experiments designed to test the existence of a "precedence effect" in the model.

Chapter 6, **Conclusions and Future Work**, ties up by summarizing the success and failure of various aspects of the project and highlights some directions for future research.

Chapter 2

Background

2.1 Introduction to spatial hearing

The ability of human listeners to determine the locations of sound sources around them is not fully understood. Several cues are widely reported as useful and important in localization, including *interaural intensity differences* (IIDs), *interaural arrival-time differences* (ITDs), spectral cues derived from the frequency content of signals arriving at the eardrums, the ratio of direct to reverberant energy in the received signal, and additional dynamic cues provided by head movements. Excellent overviews of the cues used for localization and their relative importance may be found in [1] and [18].

2.1.1 The "duplex" theory of localization

Near the turn of the century, Lord Rayleigh made a series of observations that have strongly influenced how many researchers think about localization [24]. Rayleigh observed that sound arriving at a listener from sources located away from the median (mid-sagittal) plane would result in differences in the signals observed at a listener's ears. He noted that the sound received at the far (contralateral) ear would be effectively shadowed by the head, resulting in a difference in the level, or intensity, of the sound reaching the two ears. Rayleigh correctly noted that the intensity difference would be negligible for frequencies below approximately 1000 Hz, where the wavelength of the sound is similar to, or larger than, the distance between the two ears.

Rayleigh also noted that human listeners are sensitive to differences in the phase of low frequency tones at the two ears. He concluded that the phase difference between the two ears caused by an arrival-time delay (ITD) might be used as a localization cue at low frequencies.

The combination of the two cues, IID at high frequencies and ITD at low frequencies, has come to be known as the "duplex" theory of localization, and it continues to influence research in spatial hearing.

2.1.2 The spherical head model and "cones of confusion"

To a first approximation, binaural difference cues (IIDs and ITDs) can be explained by modeling the human head as a rigid sphere, with the ears represented by pressure sensors at the ends of a diameter. Many researchers have solved the acoustic wave equation for such a configuration with sound sources at various positions, and while the resulting solution is rather complicated, some useful approximations have been made. One such approximation, for ITDs, is presented by Kuhn [14]:

$$\mathrm{ITD} \simeq \begin{cases} \frac{3a}{c}\sin\theta_{\mathrm{inc}} & \text{at low frequencies} \\ \\ \frac{2a}{c}\sin\theta_{\mathrm{inc}} & \text{at high frequencies} \end{cases}, \tag{2.1}$$

where a is the radius of the sphere used to model the head, c is the speed of sound, and θ_{inc} is the angle between the median plane and a ray passing from the center of the head through the source position (thus, $\theta_{\mathrm{inc}} = 0$ is defined as "straight ahead"). There is a smooth transition between the two frequency regions somewhere around 1 kHz.

Similar expressions may be obtained for the IID, but they generally have more complicated variations with frequency. Because of the influence of the pinna (the outer "flange" of the ear), the shape of the head, and their variation across listeners, simple approximations for the IID are not as applicable as those made for the ITD [14].

When the IID and ITD are modeled in this way, Woodworth's "cones of confusion" arise [37]. For a given set of interaural measurements (IID and ITD), there exists a locus of points for which those measurements are constant. A cone, with axis collinear with the interaural axis, is a fair approximation of this surface. Each cone may be uniquely identified by the angle θ that its surface makes with the median plane. The two degenerate cases are $\theta = 0°$, when the cone becomes the median plane, and $\theta = 90°$, when the cone collapses to a ray (the interaural axis). With the symmetry present in the spherical head model, there are no cues available to resolve positional ambiguity on a given cone of confusion.

2.1.3 Head-related transfer functions

It is generally accepted that the cues used for localization are embodied in the free-field to eardrum, or *head-related*, transfer function (HRTF) [18]. An HRTF is a measure of the acoustic transfer function between a point in space and the eardrum of the listener. The HRTF includes the high-frequency shadowing due to the presence of the head and torso, as well as directional-dependent spectral variations imparted by the diffraction of sound waves around the pinna. The head and pinna alter the interaural cues in a frequency-dependent manner with changes in sound source position. An example of this is shown in Figure 2-1, which shows the variation of the IID with frequency and elevation along the 60° cone of confusion for a KEMAR dummy-head microphone [11].

It has been shown that HRTF data can be used effectively to synthesize localization cues in signals intended for presentation over headphones. With care, it is even possible

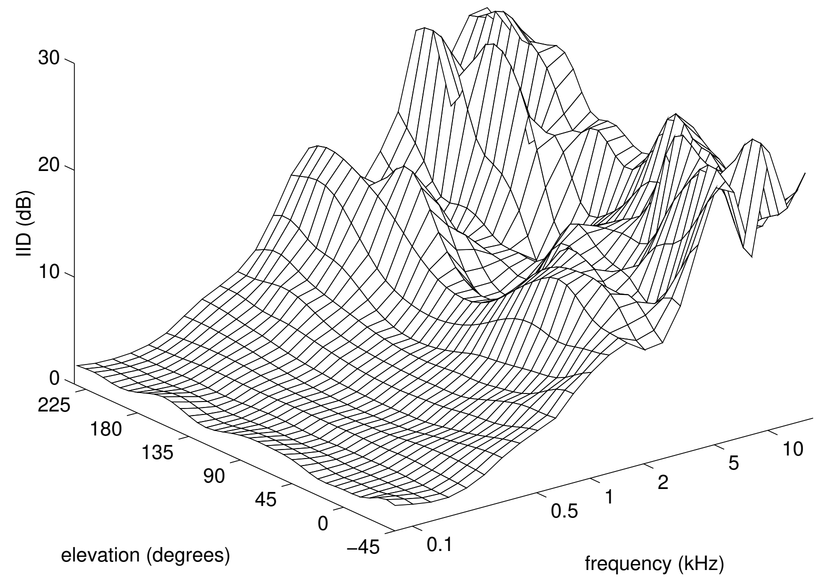

Figure 2-1: Variations of the IID with frequency and elevation along the 60° cone of confusion (IID data derived from measurements of a KEMAR dummy-head microphone [11]).

to achieve *externalization*[1] of synthesized sound sources with headphone presentation. Wightman and Kistler have shown that the information contained in the HRTFs is sufficient to explain localization performance in both azimuth and elevation [34, 35].

HRTFs vary considerably from person to person, but the types of distortions imparted by the head and pinna follow some general patterns, so meaningful comparisons may be made. Quantitative data on "average" HRTFs and differences between listeners are presented by Shaw in [27].

At a given position, the IID and ITD can be extracted as functions of frequency from the complex ratio of the left and right ear HRTFs (i.e., the *interaural spectrum*). This topic will be discussed in detail in Section 3.3.3.

2.1.4 Monaural and dynamic cues

In addition to the interaural cues described above, there are several other useful localization cues, including monaural spectral cues, distance cues, and dynamic cues imparted by head movements. These cues are reviewed briefly in [18].

Spectral cues are in general related to the "shape" of the HRTF functions, possibly including the location of "ridges" and "notches" in the monaural spectrum. These cues may be extracted by making assumptions about the source spectrum. If the source spectrum is known, it can be "divided out" of the received spectrum. Alternatively, if

[1]By externalization, we mean that the sound source is apparently located *outside* the head.

the source spectrum is not known, but can be assumed to be locally flat or of locally constant slope, the shape of the HRTF can be estimated. A computational model based on spectral cues is presented by Zakarauskas and Cynader in [38].

Localization cues for distance are not particularly well understood. The variation of signal level with distance is one potential cue, but it is useful only in regard to changes or with known source signals. The direct-to-reverberant energy ratio has also been cited as a potential localization cue, a result that stems from the fact that the reverberation level is constant over position in an enclosed space, while the direct sound energy level decreases with increasing source-to-listener distance. The ratio of high-to-low frequency energy is another prospective cue, since air attenuates high frequencies more rapidly than low frequencies over distance. To our knowledge, no models have been constructed that make use of these localization cues.

Because a slight shift in the position of the head (by rotation or "tilt") can alter the interaural spectrum in a predictable manner, head movement may provide additional cues to resolve positional ambiguities on a cone of confusion. Such cues have been dismissed by some researchers as insignificant for explaining localization performance in general, but they certainly can play a role in resolving some otherwise ambiguous situations [18].

2.1.5 The precedence effect

In a reverberant environment, "direct" sound from a sound source arrives at a given position slightly before energy reflected from various surfaces in the acoustic space. For evolutionary reasons, it may have been important for localization to be based on the direct sound, which generally reveals the true location of the sound source [40].

Regardless of the reason of origin of the precedence effect, it is true that localization cues are weighted more heavily by the auditory system at the onsets of sounds. This temporal weighting has been observed in many contexts, and the phenomenon has been known by many names, including the "precedence effect", the "Haas effect," the "law of the first wavefront," the "first-arrival effect," and the "auditory suppression effect" [40].

The precedence effect, as it has just been described, is a relatively short term effect, with changes in weighting operating on a time scale of milliseconds. In contrast to this, it has been shown that localization perception changes on the time scale of hundreds of milliseconds, and localization is fairly robust in the presence of reverberation, which operates on a time scale of *thousands* of milliseconds [40].

The mechanism that causes the precedence effect to occur is not well understood, but it has been demonstrated that it is not the result of forward masking, and that it is not due to the suppression of binaural cues alone. It has also been shown that transients (i.e., sharp changes in energy level) must be present in the sound signal for the precedence effect to occur. It is not clear how the precedence effect operates across different frequency bands [40].

Quantitative measures of the precedence effect are presented by Zurek in [39], and a model has been proposed which might explain the experimental evidence [40]. The basic form of the model is shown in Figure 2-2.

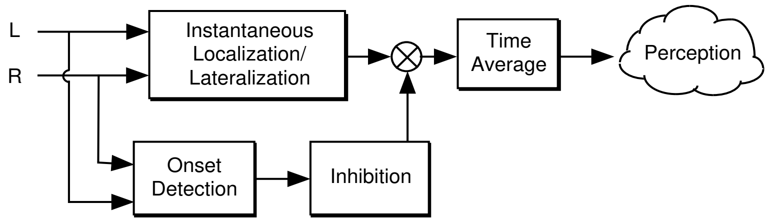

Figure 2-2: A possible model for the precedence effect, as presented by Zurek [40].

2.2 Previous work on localization models

A number of spatial hearing models have been developed. Notably, most of them fall into the "one-dimensional localization" category. These models typically estimate only a single parameter, the subjective "lateralization" of a stimulus. In this context, lateralization means "left to right position" inside the head. Recently, a few models have been constructed that are better described as "two-dimensional localization" models. These models seek to estimate both the azimuth (horizontal position) and elevation (vertical position) of sound sources. In this section, we briefly describe some of these existing models.

2.2.1 Lateralization models ("1D" localization)

Many of the early lateralization models are reviewed in [6]. Of the early models, perhaps the most influential was the neural model proposed by Jeffress ([13]), which sketches out a physiologically plausible method of extracting interaural time cues from eardrum signals. Jeffress's model was the precursor of many cross-correlation-based models, including an early one presented by Sayers and Cherry in 1957 [26].

One of the many cross-correlation-based models is the one considered by Blauert and Cobben, which employs a *running*, or *short-time* cross-correlation that operates on the outputs of band-pass channels intended to model the frequency-analysis function of the cochlea [2]. Their simple model of lateralization based on interaural time differences was extended by Lindemann to deal with interaural intensity differences and non-stationary signals, including a model of the precedence effect [16]. Gaik further extended the model by incorporating a notion of "natural" combinations of IIDs and ITDs [9]. Most recently, Bodden demonstrated that the model could be used for source separation in a multiple-speaker speech context for sources located in the horizontal plane [3].

Another important lateralization model based on cross-correlation is the one proposed by Stern, Zeiberg, and Trahiotis [29, 31, 28, 32]. Their model addresses the issue of combining binaural information across frequency. As in the Blauert/Cobben model, the cross-correlation operator is applied to the outputs of band-pass filters. Stern *et al.* propose that the ambiguity inherent in the periodicity of narrow-band cross-correlation functions might be resolved by a "straightness" weighting of peaks across frequency.

Lateralization models typically have several limitations. Most lateralization experiments are performed with "unnatural" sounds, using stimuli presented on headphones, with either uniform interaural spectra or laboratory-generated distortions of the interaural spectrum. These stimuli never occur in natural listening contexts, so the application of the experiments (and thus the models) to true "localization" is questionable.[2] Also, since most lateralization models deal only with statistically stationary (loosely, "steady-state") input signals, few attempt to model the precedence effect.

There are a few models that fall into the "1D" localization category without explicitly being lateralization models. Among these is the model proposed by Macpherson, which uses a simplified model of the cochlea as a front-end for estimating azimuth and "image width" of sources in the "front half" of the horizontal plane [17]. Estimates are made from measures of the IID and ITD in various frequency bands. The model is designed to work with impulsive and continuous signals, but uses different algorithms for the two types of signals, and includes no mechanism to determine which type of signal it is operating upon. Macpherson includes a simple model of the precedence effect which operates on impulsive signals only.

2.2.2 Azimuth and elevation ("2D" localization)

Few efforts have been made to model localization in more than one dimension. Three such models, however, are the spectral-cue localization model proposed by Zakarauskas and Cynader [38], and the models based on interaural differences proposed by Wightman, Kistler, and Perkins [36] and by Lim and Duda [15].

The model proposed by Zakarauskas and Cynader is based on the extraction of monaural cues from the spectrum received at the eardrum. They describe two algorithms: one which depends on the assumption that the source spectrum is locally flat, and one which depends on the assumption that the source spectrum has locally constant slope. Both algorithms proceed to estimate the HRTF spectrum that has modified the source spectrum. The model was able to localize sound sources in the median plane with 1° resolution, which is much better than human performance.

In [36], the authors summarize a model wherein IID spectrum templates were constructed from HRTF data. Gaussian noise was added to these templates and the results fed to a pattern recognition algorithm in an attempt to determine if sufficient information was available to distinguish position based on the IID spectrum alone. The authors report that discrimination was possible, and that as the available bandwidth of the signal was reduced, localization ability was reduced in much the same manner as in humans.

The model proposed by Lim and Duda uses the outputs of a simple cochlear model to form measurements of the IID and ITD in each frequency band. The ITD measurements are used to estimate the azimuth of the sound source, and then the IID measurements are used to predict elevation. For impulsive sound sources in an anechoic environment, the model is capable of 1° horizontal resolution and 16° vertical resolution, which is comparable to the resolution exhibited by humans. The model has not yet been generalized to non-impulsive sound sources.

None of the models described in this section have attempted to model the precedence effect.

[2] It has been suggested that laboratory generated distortions of IIDs and ITDs which do not arise in the "real world" may result in sound events that appear to occur *inside* the head [8, 9].

Chapter 3

Form of the Model

3.1 Design Overview

3.1.1 Localization cues used in the current model

As described in Chapter 2, there are several types of localization cues available. At present, monaural spectral cues are ignored, because there already exists a localization model based on them [38]. It is suggested that a spectral-cue localization model will provide an essential complement to the current model. Cues based on head movements are dismissed because of their relative non-importance ([18]) and because their use in a model precludes the use of prerecorded test signals. Distance cues, such as the direct-to-reverberant level ratio, are also ignored. Primarily, distance cues are ignored because HRTF data are assumed not to vary much with distance (regardless, HRTF data is not currently available as a function of azimuth, elevation, *and* distance). It is proposed that IIDs and ITDs provide sufficient cues for localization of sound sources away from the median plane. Thus, this thesis concentrates on interaural differences as the primary cues for localization.

For purposes of this thesis, the IID is defined as the difference (in dB) of the signal intensities measured at the two eardrums in response to a sound at a particular point in space. Similarly, the ITD is defined as the difference in arrival time between the signals at the two eardrums. In this thesis, IID and ITD are not single values for a given spatial location. Rather, they vary as a function of location and *frequency*. This distinction is important because the effects of head-shadowing and pinna diffraction are frequency dependent. Henceforth, the frequency dependent IID and ITD will be denoted the *interaural spectrum*.

In this thesis, we further divide the ITD into two components, the *interaural phase*, or fine-structure, delay (IPD) and the *interaural envelope*, or *gating*, delay (IED). It has been shown that humans are sensitive to both types of delay [41]. The distinction is necessary because the two types of delay may vary independently in eardrum signals.

3.1.2 Block summaries

The model described in this thesis can be broken into several layers, as shown in Figure 3-1. The layers are grouped into two sections: (1) the model's "front end," which transforms the acoustic signals at the two eardrums into measures of interaural differences, and (2) a statistical estimator which determines the (θ,ϕ) position most likely to have given rise to the measured interaural differences.

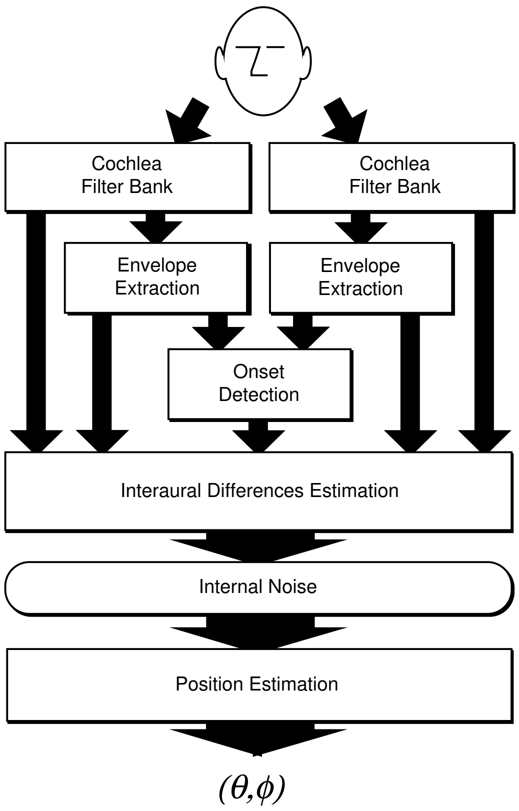

Figure 3-1: Block diagram of the proposed position estimator.

The structure of the model is conceptually simple. At the input, the eardrum signals are passed through identical filter banks, which model the time-frequency analysis performed by the cochlea. Energy envelopes are estimated for each channel by squaring and smoothing the filter outputs. The envelope signals pass through onset detectors, which note the time and relative intensity of energy peaks in each signal. This information is used by the interaural difference estimator to model the precedence effect. The interaural difference estimators use the outputs of the filter bank, the envelope estimators, and the onset detectors to extract the IID, IPD, and IED at onsets in each frequency band. A "spatial likelihood" map is then generated from the interaural differences (which may be corrupted by internal "perceptual noise"), based on probability distributions for interaural differences derived from HRTF data and human psychoacoustic data. The global maximum of the likelihood map, which corresponds to the maximum likelihood position estimate, is interpreted as the "perceived sound source location."

3.2 The Front End

3.2.1 Eardrum signals

The model described in this thesis operates upon acoustic signals measured at the eardrums of a listener. As it is quite difficult to make recordings of such signals, a common approach is to use an acoustic mannequin with microphones placed at the eardrums. To a good approximation, the acoustic system from a point in space to an eardrum microphone is linear, making measurement of the appropriate impulse responses and synthesis by convolution a reasonable approach to the generation of eardrum test signals.

KEMAR HRTF measurements

Until recently, there were no publicly available impulse-response measurements of free-field to eardrum systems. To satisfy the data requirements of the current research, measurements were made of a Knowles Electronics Manikin for Acoustic Research (KEMAR). The KEMAR, fitted with a torso and two different model pinnae with eardrum microphones, was placed on a turntable in MIT's anechoic chamber. A small two-way speaker (Radio Shack model Optimus-7) was mounted, by way of a boom stand, at a 1.4 m radius from the center of the dummy-head. The speaker was positioned by hand at each of 14 elevations (in $10°$ increments from $-40°$ to $90°$ in the "latitude/longitude" coordinate system). At each elevation, KEMAR was rotated under computer control to each azimuth measurement position (measurement positions were separated by approximately $5°$ great-circle arcs). Measurements were made using bursts of pseudo-random noise ("maximum-length sequences") which have the property that their autocorrelation is non-zero only at zero lag (and at multiples of the sequence length for periodically repeating signals). This property is important, as it allows the impulse response of the system to be obtained by cross-correlation of the measured response with the original noise signal [25, 33]. The measurement process is presented in greater detail in [10] and [11]. In total, 710 positions were measured.

Test signal synthesis

In the absence of reverberation and nearby acoustically reflective objects, the acoustic signals at the eardrums consist entirely of the direct sound from the sound source. Eardrum signals for such free-field sources may therefore be synthesized by convolution with *head-related impulse-responses* (HRIRs) as mentioned above.

For the experiments described in this thesis, the KEMAR HRIR measurements were used to synthesize test signals. Only the measurements of the left pinna were used, and right ear impulse responses were found by symmetry. Since the test signals required for the experiments in this thesis are for free-field signals, no more complicated synthesis techniques, such as ray-tracing or reverberation, are required.

Several types of input signals were used to test the localization model. Bursts of Gaussian white noise of various lengths, and short impulsive signals, were used to test the localization ability of the model. Also, the signals used by Zurek to quantify the precedence effect were recreated. The results of these experiments are described in Chapter 5.

3.2.2 Cochlea filter bank

There are several issues involved in the implementation of a filter bank designed to model the frequency analysis performed by the cochlea. These issues include the frequency range and resolution of the filters and their "shape."

As many of the important localization cues occur in the frequency range above 4-5 kHz, the range covered by the filter bank is very important. The filter bank currently employed has center frequencies ranging from 80 Hz to 18 kHz, which covers most of the audio frequency range. The filters are spaced every third-octave, yielding 24 filter channels in total. The third-octave spacing models the approximately logarithmic resolution of the cochlea above 500 Hz.

All filters are fourth order IIR, with repeated conjugate-symmetric poles at the center frequency (CF) and zeros at D.C. and at the Nyquist frequency. The pole moduli are set such that the Q of each filter (equal to the ratio of the center frequency to the -3 dB bandwidth) is 8.0, and each filter is scaled such that it has unity gain at its CF. The constant Q nature of the filter bank is a loose fit to the *critical bandwidths* reported for the human ear [19].

The "shape" of the filters is a result of the design technique, and is not intended to be a close match to cochlear tuning curves.

3.2.3 Envelope processing

There are several reasonable approaches to extracting the envelope of a band-pass signal. Three possible techniques are the Hilbert transform method, rectification-and-smoothing, and squaring-and-smoothing.

The Hilbert transform is a phase-shifting process, where all sinusoids are shifted by 90° of phase (e.g., sine to cosine). By squaring the signal and its Hilbert transform, summing the two signals, and taking the square root, the envelope signal may be extracted.

Alternatively, the physiology of the cochlea suggests a half-wave rectification and smoothing algorithm [22]. This approach has been used by many authors (e.g., [2]), but the non-linearity of the half-wave rectifier implied by the cochlear physiology is difficult to analyze in terms of frequency domain effects and has thus not been used in the current model.

A third alternative is the "square-and-smooth" method that is adopted in the current model.

Assuming that we are interested in a channel that is band-pass in character with some center frequency ω_k (e.g., the vibration of a point on the basilar membrane), then the signal in that channel may be modeled as an amplitude modulated sinusoidal carrier:

$$x_k(t) = A(t)\cos(\omega_k + \phi).$$

With this model, the function $A(t)$ is the envelope of the signal $x_k(t)$ and is defined to

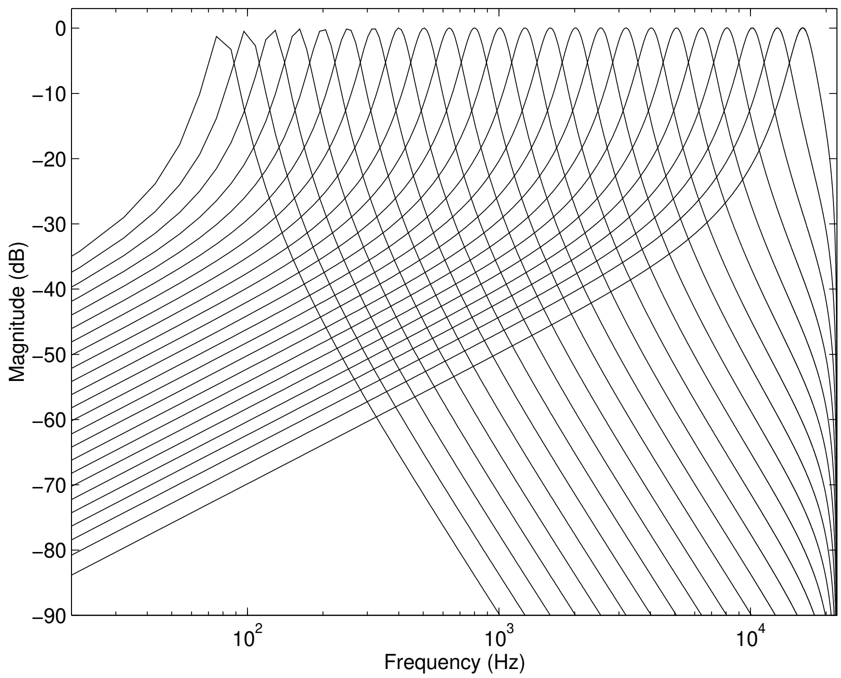

Figure 3-2: Frequency response of the filter bank used to model the frequency-analysis performed by the cochlea. Center frequencies are spaced in third-octave steps, with constant center-frequency-to-bandwidth ratio ($Q = 8.0$).

be low-pass in nature.

The squared signal is given by:

$$x_k^2(t) = A^2(t)\cos^2(\omega_k + \phi). \tag{3.1}$$

The Fourier transform of this signal is equal to the convolution of the Fourier transforms of the $A^2(t)$ term and the $\cos^2(\omega_k + \phi)$ term. Since the Fourier transform of the $\cos^2(\omega_k + \phi)$ term is the sum of three Dirac δ-functions and $A^2(t)$ is low-pass in nature, as long as the bandwidth of the $A^2(t)$ term is less than ω_k, the resulting spectrum can be low-pass filtered at cutoff ω_k, yielding the spectrum of $A^2(t)$. Further, if $A(t)$ is nonnegative, it may be recovered exactly by taking the square root of the resulting signal. Physically, it is impossible to realize a "brick-wall" low pass filter, but with care, the recovered $\hat{A}(t)$ is a very reasonable approximation of $A(t)$. This concept is demonstrated in Figures 3-3 and 3-4.

In a discrete-time system, such as the implementation used in this research, aliasing is a potential problem in this processing. However, if we define BW to be the lowest frequency above which $A(t)$ contains no energy, then as long as $BW + \omega_k \leq \pi$ where 2π is the sampling frequency, there will be no aliasing. To a good approximation, this is the case for all of the filters used in this research.

There is physiological evidence that the human ear is only capable of following changes in the envelope of a signal up to an upper frequency limit. In this model,

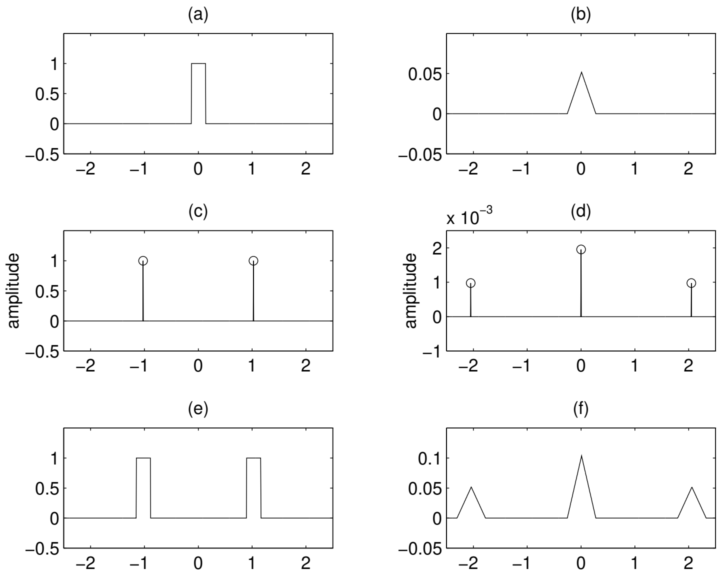

Figure 3-3: Example of Fourier spectra of $x(t)$ and $x^2(t)$ for a simple amplitude modulation spectrum. All horizontal axes are in units of frequency divided by ω_k. (a) Spectrum of $A(t)$, (b) Spectrum of $A^2(t)$, (c) Spectrum of $\cos(\omega_k t)$, (d) Spectrum of $\cos^2(\omega_k t)$, (e) Spectrum of $x(t) = A(t)\cos(\omega_k t)$, (f) Spectrum of $x^2(t) = A^2(t)\cos^2(\omega_k t)$. It is possible to arrive at the spectrum in (b) by appropriate filtering of the spectrum in (f).

the limit is set to 800 Hz, which is the value chosen by Blauert and Cobben [2]. In the implementation, this limit is imposed by not allowing the cutoff frequencies of the smoothing filters to exceed 800 Hz.

3.2.4 Onset detection

Why onsets?

In a realistic localization scenario, the energy received at the eardrums comes from reverberation and "distractor" sound sources as well as directly from the "target" source. Sharp peaks in energy that may be attributed to the target source generally provide small time windows with locally high signal-to-noise ratios for estimating attributes of the source signal (e.g., interaural differences).

As mentioned previously, there is evidence that onsets are particularly important portions of the signal for purposes of localization. This is demonstrated by their importance in relation to the precedence effect [40].

Extraction method

The model searches for energy peaks in the various filter bank channels by looking for maxima in the envelope signals. These "onsets" are found by a simple peak-picking algorithm coupled with a suppression mechanism. In the current implementation, the envelope signals at the two ears are summed before onsets are detected; thus, onsets are

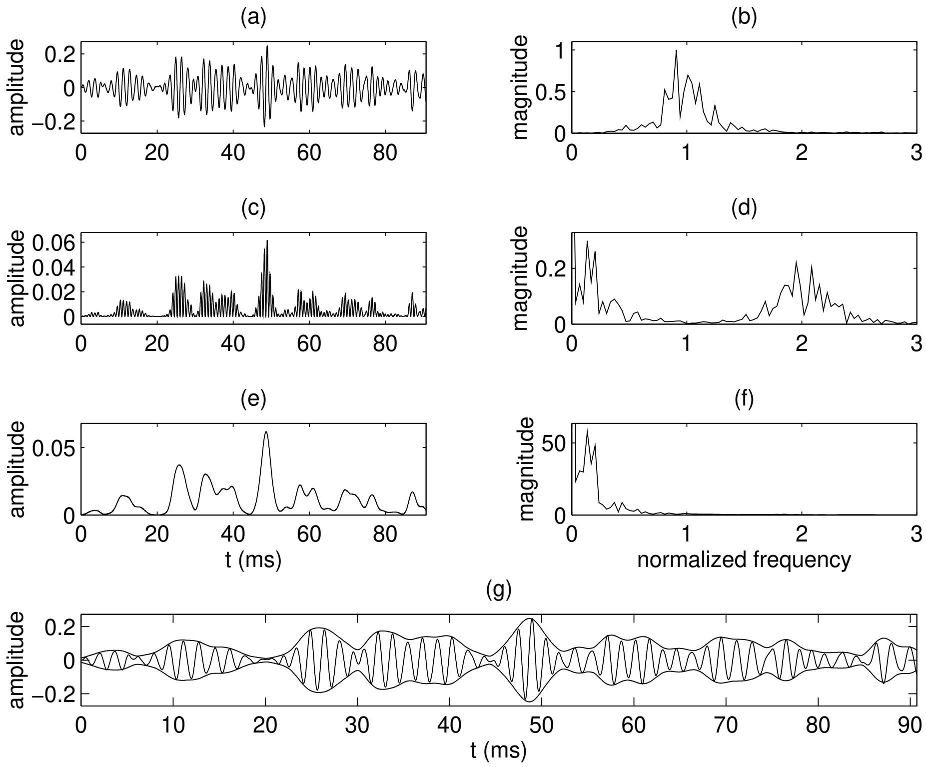

Figure 3-4: Example of the envelope extraction process for a band-pass signal. The normalized frequency axis is in units of frequency divided by ω_k. (a) Time waveform of the bandpass $x_k(t)$, (b) Spectrum of $x_k(t)$, (c) Time waveform of $x_k^2(t)$, (d) Spectrum of $x_k^2(t)$, (e) Time waveform of $A^2(t)$, (f) Spectrum of $A^2(t)$. (g) Time waveform of the bandpass $x_k(t)$ with $A(t)$ and $-A(t)$ superimposed. Since the frequency axis is normalized and the filter bank used in this research is constant Q, the above plots are typical of all frequency channels, with an appropriate scaling of the time axis.

not independently found at the two ears. The model proposed by Zurek (see Figure 2-2) suggests that a sharp onset should suppress localization information (including other onsets) occurring over approximately the next 10 ms [40]. In the current model, a small backward-masking effect has been implemented to ensure that a false-alarm is not triggered by a small onset when it is immediately followed by a much larger one.

Figure 3-5 shows an example of onsets detected in an envelope signal which has been extracted from a channel with a center frequency of 800 Hz.

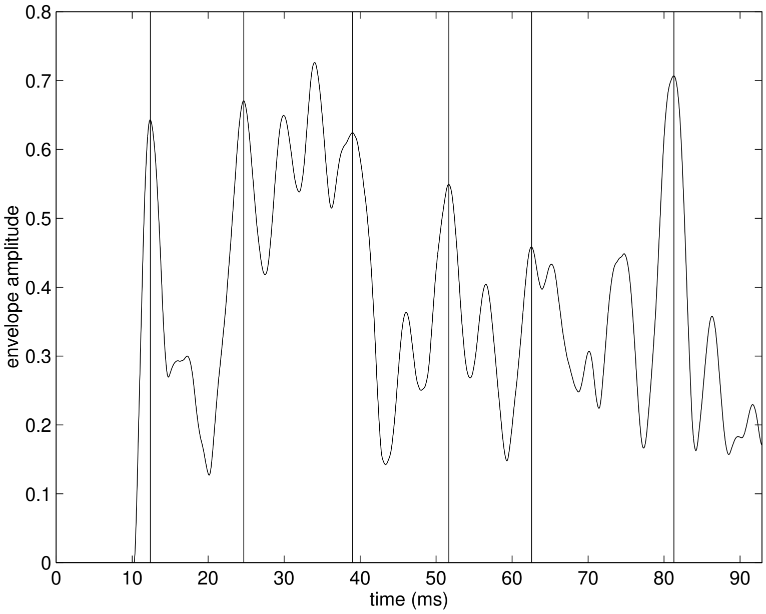

Figure 3-5: Example of onsets extracted from an envelope signal in a channel with a center frequency of 800 Hz. Some of the local maxima (and, in fact, the global maximum) have been suppressed in this example.

3.2.5 Interaural differences estimation

Interaural intensity differences

In the current model, the IID is simply formed by the log ratio of the envelope signals:

$$\mathrm{IID}_k(t) = 20\log_{10}(E_{R,k}(t)/E_{L,k}(t)), \tag{3.2}$$

where $E_{L,k}(t)$ and $E_{R,k}(t)$ are the envelopes of the kth band-pass channel in the left and right ears respectively. The IID function is not smoothed further in any way. Rather, there is a smoothing inherent in the precedence effect model, which is discussed in Section 3.3.1.

With the current IID formulation, an interesting artifact occurs that is due to the coupling of IID and ITD in natural listening situations. IID and ITD tend to co-vary for single sound sources in free-field. This fact follows directly from the form of natural HRIRs, which dictate that an energy burst will excite the ipsilateral (near-side) ear more intensely and sooner than the contralateral (far-side) ear. Similarly, an energy offset will cause the excitation at the ipsilateral ear to drop off before the excitation drops off at the contralateral ear (it should be noted that this time lag is in general smaller than 1 ms). Because of this effect, a sharp energy peak in the input signal will cause fluctuations in the output of the IID estimator. This effect can be seen in Figure 3-6, which shows the IID trace for the same signal as was used to generate Figure 3-5. This effect will be described further in Section 4.1, which treats the issue of measurement noise.

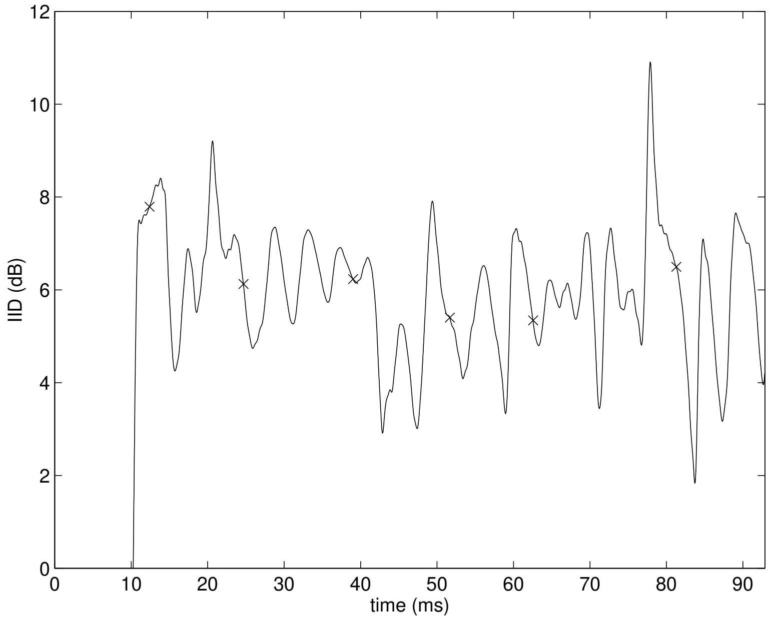

Figure 3-6: Example showing variation of the IID function over time for a time-varying input spectrum. Each ($\times$) marks the position of an onset detected in the signal (see Figure 3-5).

Interaural phase delays

The IPD is estimated by a running cross-correlator similar to those proposed by Blauert [2] and Lindemann [16]:

$$\mathrm{IPD}(t) = \underset{\tau}{\operatorname{argmax}} \int_{-\infty}^{\infty} L_k(t' - \frac{\tau}{2}) R_k(t' + \frac{\tau}{2}) w(t' - t) dt'. \tag{3.3}$$

Here, $L_k(t)$ and $R_k(t)$ are the signals in the kth channel of the left and right ear filter banks respectively, $w(t)$ is a window function, and τ is limited to the range $-1 < \tau < 1$ ms, which includes the range of IPDs encountered in natural listening conditions. The window used by Blauert is a decaying exponential with a time constant of a few milliseconds [2]. In the current model, we use a window of the form:

$$w(t) = Ate^{-t/\tau}, \tag{3.4}$$

with an effective "width" of 4–5 ms.

In the current implementation, the running cross-correlation is sampled in τ. The sampling is performed in approximately 45 μs increments (which corresponds to two sample periods at a 44.1 kHz sampling rate) over the $-1 < \tau < 1$ ms range. The peak in the cross-correlation function is then interpolated by a quadratic fit to the three points surrounding the global maximum. This process breaks down due to under-sampling (of τ) for filters with center frequencies that are larger than approximately 1.5 kHz. Since humans are incapable of fine structure analysis above 1.5 kHz ([19]), this breakdown is acceptable, and the IPD is not evaluated in channels with center frequencies larger than 1.5 kHz.

The cross-correlation of a narrow-band signal is nearly periodic in τ, with period equal to the inverse of the carrier frequency of the signal. This near-periodicity, which is present in the cross-correlations of band-pass signals in the cochlear filter-bank, makes extraction of the interaural delay a difficult task, with "period" errors common. The effect of introducing a time-limited window into the cross-correlation is to reduce the size of cross-correlation peaks at lags distant from zero. Since interaural delays tend to fall in range $-1 < \tau < 1$ ms, we may largely ignore cross-correlation peaks at lags outside of this range.

Interaural envelope delays

The IED estimation is identical to the IPD estimation, except that the envelope signals, $E_{L,k}(t)$ and $E_{R,k}(t)$, are substituted for the filter bank signals in Equation 3.3. In general, the two window functions $w(t)$ used for the IPD and IED estimators may be different, but the same window is used presently for the sake of simplicity. Since the envelope signals are low pass in nature, under-sampling of the cross-correlation function is not an issue.

3.3 Position Estimation

3.3.1 Precedence effect model

The precedence effect model is based upon the one proposed by Zurek (see Figure 2-2 and [40]). We assume that the inhibition (or suppression) described in the model applies to the localization cues, rather than an actual position estimate. Further, we make the assumption that the inhibition operates independently in the various filter-bank channels. It is unknown whether these assumptions are valid.

The onset detector described in Section 3.2.4 provides a time-stamped list of "onsets" detected in each filter channel. It is assumed that each detected onset causes a sampling of the interaural difference signals derived from the appropriate filter channel. The sampling function (or window) is positioned in time such that the main lobe is centered around each onset's time-stamp. Each interaural measurement is given by the inner product of the window function with the appropriate running interaural difference signal (as described in Section 3.2.5). The window function is scaled such that it integrates to unity, which ensures that the measured value is correct for a constant interaural difference signal.

The form of the sampling function is chosen to approximately match the inhibition curve sketched by Zurek (see Figure 3-7a). The window takes the form in Equation 3.4, with τ set to make the effective window "width" approximately 2–3 ms. The sharp drop in sensitivity is assumed to be due to the shape of the sampling function (as shown in Figure 3-7b), while the limitation imposed by the onset detector (that onsets occur no more frequently than one every 10 ms) is assumed to account for the return of sensitivity approximately 10 ms following an onset.

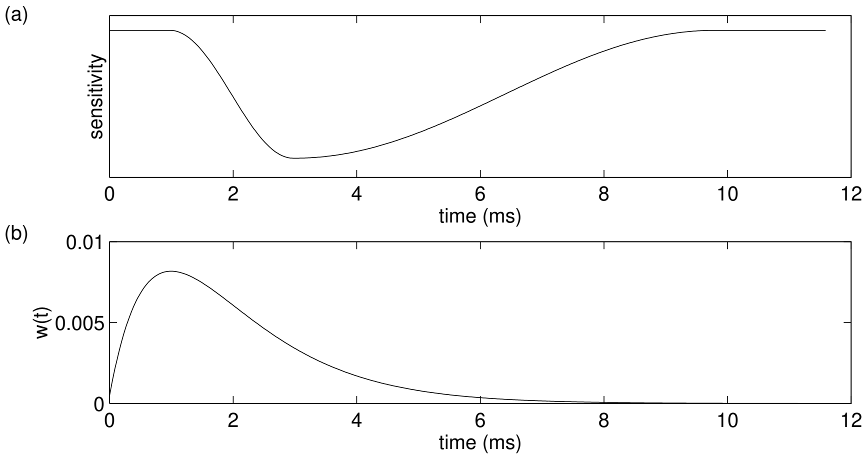

Figure 3-7: (a) Inhibition curve for localization information suggested by Zurek. (b) Sampling window used in the model (of the form $w(t) = At\exp(-t/\tau)$).

Some results of experiments testing the adequacy of this implementation are presented in Chapter 5.

3.3.2 Development of the position estimator

If we make the restriction that the model has no *a priori* knowledge of the input spectrum, we might assume, for purposes of constructing a model, that the "average" spectrum is locally "white" (i.e., that the probability distribution of input signal energy is constant over frequency). We can then argue that the interaural differences for a source at a given position will vary in a noise-like manner about some mean as changes in the input spectrum interact with the fine structure of the HRTFs within any given band-pass channel. We further make the doubtful, but nonetheless common, assumption that the noise is additive, Gaussian, and zero-mean in nature.

With these assumptions, we may write the probabilistic formulation:

$$\mathbf{P} = \overline{\mathbf{P}}_{\theta,\phi} + \mathbf{v}, \tag{3.5}$$

where

$$\mathbf{v} \sim \mathcal{N}(\mathbf{0}, K). \tag{3.6}$$

Here, $\mathbf{P}$ is a vector consisting of the various interaural measurements under consideration, $\overline{\mathbf{P}}_{\theta,\phi}$ is the expected value of $\mathbf{P}$, given that the sound source is at the location (θ, ϕ), and $\mathbf{v}$ is the additive noise vector, which is Gaussian with covariance matrix K. With n interaural difference measurements, $\mathbf{P}$, $\overline{\mathbf{P}}_{\theta,\phi}$, and $\mathbf{v}$ are $(n \times 1)$ and K is $(n \times n)$. The probability density function for $\mathbf{P}$, given (θ, ϕ), is then Gaussian with mean $\overline{\mathbf{P}}_{\theta,\phi}$ and covariance K:

$$\mathbf{P} \sim \mathcal{N}(\overline{\mathbf{P}}_{\theta,\phi}, K), \tag{3.7}$$

$$p(\mathbf{P}|\theta, \phi) = \frac{1}{(2\pi)^{n/2}|K|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{P} - \overline{\mathbf{P}}_{\theta,\phi})^T K^{-1}(\mathbf{P} - \overline{\mathbf{P}}_{\theta,\phi})\right\}. \tag{3.8}$$

If we also assume that the noise in each interaural difference measurement is independent of the noise in the other measurements, then the matrix K is diagonal:

$$K = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix}. \tag{3.9}$$

Here, σ_k^2 is the variance of the noise in the kth measurement (noise in the model is examined in depth in Chapter 4).

With this formulation, we may write a simplified equation which represents the likelihood that the measurements arose from a source at location (θ, ϕ):

$$\mathcal{L}_{\theta,\phi} = p(\mathbf{P}|\theta, \phi) = \frac{1}{(2\pi)^{n/2} \prod_{k=1}^{n} \sigma_k} \exp\left\{-\frac{1}{2}\sum_{k=1}^{n} \frac{(P_k - \overline{P}_{\theta,\phi,k})}{\sigma_k^2}\right\}. \tag{3.10}$$

It is worth noting at this point that σ_k is assumed to be constant over all (θ, ϕ) positions, which makes the term multiplying the exponential a constant over all positions. It is therefore quite straightforward to compare the "likelihoods" at different positions, based on the observed measurements.

By calculating $\mathcal{L}_{\theta,\phi}$ at a number of different positions, it is possible to generate a *spatial likelihood map*, which shows the *relative* likelihood that the sound source is at each position. Examples of spatial likelihood maps are given in Chapter 5.

3.3.3 Template extraction from HRTF data

In order to calculate the likelihood function for a particular position, the mean values for the interaural differences for a source at that position are required. Reapplying the input signal assumptions that we used to derive the probabilistic model, we may extract the expected mean values of the interaural differences from HRTF data by making some assumptions about the types of input signals that the system is likely to encounter.

The assumptions that we choose to make are that the average spectrum is "white" (i.e., the probability distribution for input signal energy is constant over all frequencies of interest). When the time-structure of the input signal is required in the template calculations, we assume that the signal is *impulse-like* (i.e., it is time-limited, with short duration), and we in fact use the Dirac δ-function to represent the input.

Interaural intensity differences

We may consider the envelope extraction process to derive the IID template function. For a single free-field sound source at position (θ, ϕ), the signals received at the eardrums are given by:

$$\begin{aligned} L(t) &= x(t) * h^L_{\theta,\phi}(t) && (3.11) \\ R(t) &= x(t) * h^R_{\theta,\phi}(t), && (3.12) \end{aligned}$$

where $L(t)$ and $R(t)$ are the signals received at the left and right eardrums respectively, $x(t)$ is the source input signal, and $h^L_{\theta,\phi}(t)$ and $h^R_{\theta,\phi}(t)$ are the left and right ear HRIRs respectively. In these equations and those that follow, $*$ denotes the convolution operator.

The squared envelope signals extracted from these input signals are then given by:

$$\begin{aligned} E^2_{L,k}(t) &= (L(t) * h_k(t))^2 * h^E_k(t) && (3.13) \\ E^2_{R,k}(t) &= (R(t) * h_k(t))^2 * h^E_k(t), && (3.14) \end{aligned}$$

where $E_{L,k}(t)$ and $E_{R,k}(t)$ are respectively the left and right side envelope signals in the kth band-pass channel, $h_k(t)$ is the impulse response of the kth band-pass filter, and $h^E_k(t)$ is the impulse response of the smoothing filter used in the envelope extraction process in the kth channel.

We may then form the following expression for the IID:

$$\mathrm{IID}_{k,\theta,\phi}(t) = \int_{-\infty}^{\infty} w(t'-t) 10 \log_{10}\left(\frac{E^2_{R,k}(t')}{E^2_{L,k}(t')}\right) dt', \qquad (3.15)$$

where $w(t)$ is the window function described in Section 3.3.1. We must be careful to ensure that $E_{L,k}(t)$ and $E_{R,k}(t)$ are always non-zero, but due to the IIR nature of the filter bank, once the filters have been initially excited, this condition is generally satisfied. Regardless, a simple threshold mechanism would suffice to correct this problem.

Assuming an impulse-like input signal (a Dirac δ-function in the limit), we may form the IID template function:

$$\mathrm{IID}_{k,\theta,\phi}(t) = \int_{-\infty}^{\infty} w(t'-t) 10 \log_{10}\left(\frac{(h^R_{\theta,\phi}(t') * h_k(t'))^2 * h^E_k(t')}{(h^L_{\theta,\phi}(t') * h_k(t'))^2 * h^E_k(t')}\right) dt'. \qquad (3.16)$$

Interaural phase delays

For a known input signal, the IPD is given by Equation 3.3. If we again assume that the input is an impulse, we may construct the IPD template as follows:

$$\hat{L}_{k,\theta,\phi}(t) = h_k(t) * h^L_{\theta,\phi}(t) \tag{3.17}$$

$$\hat{R}_{k,\theta,\phi}(t) = h_k(t) * h^R_{\theta,\phi}(t) \tag{3.18}$$

$$\mathrm{IPD}_{k,\theta,\phi}(t) = \underset{\tau}{\operatorname{argmax}} \int_{-\infty}^{\infty} \hat{L}_{k,\theta,\phi}(t' - \frac{\tau}{2})\hat{R}_{k,\theta,\phi}(t' + \frac{\tau}{2})w(t' - t)dt' \tag{3.19}$$

Interaural envelope delays

In the same fashion as we constructed the IPD template, we may construct a template for the IED:

$$\hat{E}^{L^2}_{k,\theta,\phi}(t) = (h_k(t) * h^L_{\theta,\phi}(t))^2 * h^E_k(t) \tag{3.20}$$

$$\hat{E}^{R^2}_{k,\theta,\phi}(t) = (h_k(t) * h^R_{\theta,\phi}(t))^2 * h^E_k(t) \tag{3.21}$$

$$\mathrm{IED}_{k,\theta,\phi}(t) = \underset{\tau}{\operatorname{argmax}} \int_{-\infty}^{\infty} \hat{E}^L_{k,\theta,\phi}(t' - \frac{\tau}{2})\hat{E}^R_{k,\theta,\phi}(t' + \frac{\tau}{2})w(t' - t)dt' \tag{3.22}$$

3.3.4 Spherical interpolation

We would like to calculate spatial likelihood maps that are much more densely sampled in azimuth and elevation than the KEMAR HRTF data. Without a dense sampling, it is difficult to measure localization errors quantitatively, because many errors will be due to the measurement position quantization. There are two possible approaches to solving this problem: (1) interpolation of the spatial map derived from interaural difference templates at KEMAR measurement positions only, and (2) interpolation of the interaural difference templates themselves. If the interaural difference templates are relatively smooth functions of position (which seems to be the case), then the second method is clearly preferable.

Interpolation of functions in spherical coordinate systems is a difficult problem, however. The KEMAR HRTF data set is not sampled regularly (i.e., the sample points are not located at the vertices of a polyhedron); therefore, the interaural difference templates are also irregularly sampled. There are many possible approaches to the "spherical interpolation" problem, including minimum mean-square error (MMSE) interpolation with some basis function set defined under spherical coordinates.

The MMSE method is a well-known functional approximation technique using basis functions. First, we define an inner product in our coordinate system:

$$< x(\theta,\phi), y(\theta,\phi) >= \iint\limits_R x(\theta,\phi)y(\theta,\phi)dA, \tag{3.23}$$

where $< x(\theta, \phi), y(\theta, \phi) >$ is the inner product of two functions, x and y, defined in our coordinate system, and the integral is evaluated over the range of possible (θ, ϕ) values. It should be noted that we have stated this problem in terms of spherical coordinates, but it is equally applicable to problems in other coordinate systems. The only change in Equation 3.23 is the form of the dA term (and possibly the number of integrals). Henceforth, we will use the angle bracket notation introduced above to represent the inner product.

Next, we define a set of basis functions, $\Psi_k(\theta, \phi)$, $k = 1 \ldots N$, which will be used to approximate the function $d(\theta, \phi)$:

$$\hat{d}(\theta, \phi) = \sum_{k=1}^{N} w_k \Psi_k(\theta, \phi). \tag{3.24}$$

With this formulation, the weights w_k which minimize the mean-squared error:

$$E = \iint\limits_R \left[d(\theta, \phi) - \hat{d}(\theta, \phi)\right]^2 dA \tag{3.25}$$

are given by the solution of:

$$\begin{bmatrix} < \Psi_1, \Psi_1 > & \cdots & < \Psi_1, \Psi_N > \\ \vdots & \ddots & \vdots \\ < \Psi_N, \Psi_1 > & \cdots & < \Psi_N, \Psi_N > \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} < \Psi_1, d > \\ \vdots \\ < \Psi_N, d > \end{bmatrix}. \tag{3.26}$$

In this thesis, we are interested in finding smooth approximations to the interaural difference functions. Since the interaural differences are only measured at discrete (θ, ϕ) positions, we approximate the continuous integral in Equation 3.23 by the following discrete summation:

$$< x, y > \simeq \sum_k x_k y_k A_k. \tag{3.27}$$

Here, k indexes the various measurement positions, x_k and y_k are the values of $x(\theta, \phi)$ and $y(\theta, \phi)$ at position k, and A_k is the "area" represented by measurement position k. For the training positions used to estimate the interaural difference surfaces (the positions are separated by roughly 10° great circle arcs), we have $A_k \simeq K$, for all k. Thus, the A_k term may be ignored in Equation 3.27.

For our basis functions, we define a set of "spherical Gaussian" functions of the form:

$$\Psi_k(\mathbf{p}) = \exp\left[-\frac{1}{K}\Phi^2(\mathbf{p} - \mathbf{p}_k)\right], \tag{3.28}$$

where $\Phi(\mathbf{p} - \mathbf{p}_k)$ is defined as the great circle arc (in degrees) between positions $\mathbf{p}$ and $\mathbf{p}_k$. We define $\mathbf{p}$ and $\mathbf{p}_k$ to be positions on a sphere, each indexed by azimuth and elevation in a convenient spherical coordinate system. In Equation 3.28, K is a smoothing factor (chosen in this thesis to be 120 degrees-squared). With this formulation, the basis function is unity at its *kernel position*, $\mathbf{p}_k$, and falls off by a factor of $1/e$ for each $\sqrt{120} \simeq 11°$ distance increment. In Appendix A, Figure A-2 shows contours of equal great-circle distance for various kernel positions in two different spherical coordinate systems. The contours are also equal-value contours for the "spherical Gaussian" basis functions at those kernel positions (and in fact, the contours are drawn at $1.0\sqrt{K}^\circ, 1.5\sqrt{K}^\circ$, and $2.0\sqrt{K}^\circ$ distances with $K = 120$).

The choice of kernel positions is important, and the optimal choice may depend on the characteristics of the surface to be approximated. In this thesis, we choose the kernel positions to be spaced by approximately 10° great circle arcs. According to the results in [5], this spacing is on the verge of under-sampling the data. Unfortunately, a more dense sampling is problematic with our limited data set, as will be discussed shortly.

The MMSE interpolation method works extremely well for approximating noise-free functions with some "smoothness" constraints. The interaural difference data, however, has no well-defined smoothness constraints and is certainly not noise-free (noise may be present both in the original HRTF measurements and in the interaural difference template calculations). The results in [5] suggest that the surfaces do have an underlying smoothness or regularity, but our data may be under-sampled to some degree.

With a data set such as this, it is important to force the MMSE approximation to "under-fit" the data (by under-fit, we mean that the approximation surface will not pass through all of the data points exactly). This under-fitting is ensured in part by choosing the number of basis functions to be fewer than the number of measurement positions (thus making the number of degrees of freedom [DOFs] in the approximation smaller than the number of DOFs in the measurement data). We also choose the basis functions to be relatively smooth functions of position (in our case, by choosing a reasonably large value of K in Equation 3.28). If the basis functions are too smooth, the inner product matrix in Equation 3.26 will become badly singular, making it impossible to find a reasonable set of basis function weights w_k.

While implementing the MMSE interpolation for this thesis, there was some difficulty in fitting both hemispheres of the interaural data simultaneously (i.e., it was difficult to find a set of kernel positions and a smoothness factor K which yielded non-singular inner-product matrices for the interaural difference data). Since the interaural difference data is highly symmetric (the function value at a given position is the negative of its reflection across the median plane), a reasonable approach is to define kernel positions in one hemisphere only. We may then fit the interaural data in that hemisphere only and exploit symmetry to find the interaural differences in the other hemisphere.

Chapter 4

Noise, Bias, and Perceptual Distance

There are two types of "noise" to be considered in the current model: "measurement noise" in the interaural differences (caused by the measurement method and by ignorance of the input signal) and "perceptual noise" which may be added to corrupt the interaural difference measurements or the final position estimate, in order to better approximate human performance. There are also biases inherent in the current model, which are caused by the interpolation of the interaural difference templates. At the end of this chapter, we will consider the implied "perceptual distance" between various signals, based on a particular choice of variance estimates.

It should be noted that the measurement noise discussed in this thesis is not truly noise in a probabilistic sense; the variation of the interaural cues with changes in the input signal is completely deterministic. From the point of view of the position estimator, however, the variation of the cues appears "noise-like," and may certainly be characterized in terms of its distribution and variance. Henceforth, we will use the term "measurement noise" to describe this deterministic variation and the term "perceptual noise" to describe true randomness in the system.

4.1 Measurement noise

As mentioned in Section 3.2.5, IID and ITD tend to co-vary for sound sources in free-field. This coupling of intensity differences and delay is responsible for some of the measurement noise in the IID signal. A simple example of this is given in Figure 4-1, which shows the envelope responses and resulting IID signal for an amplitude-modulated tone which has been filtered by a simple HRTF model.

The measured IID also varies with the particular input signal spectrum. This follows from the fact that the left and right ear HRTFs are generally not the same "shape" (i.e., they do not differ only by a multiplicative constant). This result is easily seen by considering the pathological case of pure sinusoidal input signals. Regardless of the "shape" of the cochlear band-pass filter, the IIDs generated by sinusoidal stimulation will be dependent on the frequency of the sinusoid because the amplitude ratio of the

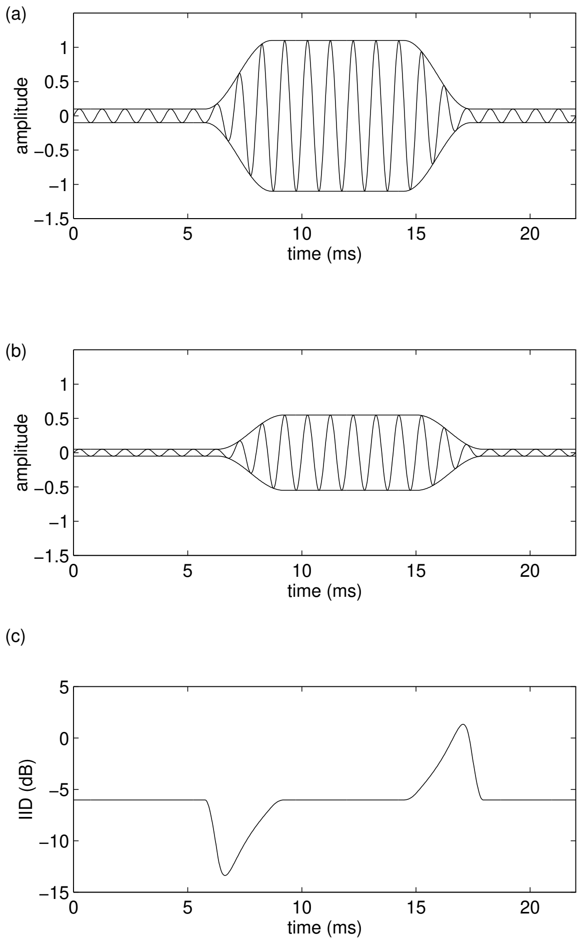

Figure 4-1: Result of stimulating a simplified HRTF model (a simple delay of 0.5 ms and a scaling by a factor of 0.5 of the contralateral ear response) with an amplitude-modulated sinusoid. (a) Envelope and fine structure at the ipsilateral ear. (b) Envelope and fine structure at the contralateral ear. (c) Interaural intensity difference estimated from the received signals. The ipsilateral ear is excited sooner and more intensely than the contralateral ear, causing a sharp rise in the magnitude of the measured IID from the "steady-state" value of -6 dB. Similarly, the excitation drops off sooner at the ipsilateral ear, causing a dip in the measured IID magnitude.

two HRTFs is not constant over frequency.

Both of these effects are present in the IID signal presented in Figure 3-6. It should be noted that by sampling the IID at peaks in the envelope signal (where the derivative is zero, and the IID has had the opportunity to return toward the "correct" value), the IID measurement variance is somewhat reduced. The variance could be further reduced by introducing a correction delay to cancel the interaural delay. In the current model, it is assumed that IID and ITD are evaluated independently; thus, such an approach is dismissed.

As with the IID measurements, we may show that the IPD and IED measurements will vary with changes in the input spectrum. This follows directly from the fact that the phase delay difference and group delay difference between the HRTFs at the two ears are not constant functions over frequency.

4.2 Perceptual noise

As it has been presented, the current model is deterministic. Humans, however, are not. In fact, it is clear that humans are incapable of performing precise, repeatable measurements of physical phenomena. Most computational models are capable of handling a much larger dynamic range of inputs and of performing much more precise measurements than humans. For these reasons, perceptual noise is often a component of such models. In this thesis, we consider "perceptual noise" as a possible component because the measurement noise inherent in the signal processing may not contain enough variance to match human performance.

Generally, perceptual noise is modeled as Gaussian, zero-mean, and additive, with variance estimated from human psychophysical experiments. In the current model, there are two logical places to include perceptual noise: (1) at the outputs of the interaural difference estimators, and (2) at the output of the position estimator. We shall consider only the former.

Often, just-noticeable-difference (JND) measurements are used to gauge the limit of human sensitivity to changes in a physical quantity. Applying this to interaural differences, we might measure the smallest ΔIID which results in a "different" perception from some baseline IID (0 dB is a convenient reference). This ΔIID will be called the just-noticeable interaural intensity difference. Similarly, we might measure the smallest interaural delay $\Delta\tau$ which results in a different perception from the baseline ITD = 0.

In "A Decision Model for Psychophysics," Durlach interprets the significance of such JND measurements [7]. Following his derivation, we first assume that the presentation of a stimulus gives rise to a unidimensional perceptual variable and that the mapping from the physical variable to the perceptual variable is monotonic. With additive perceptual noise, each human "measurement" may be described as a random variable with variance equal to the noise variance and mean equal to some value uniquely determined by the value of the physical property under consideration.

With this formulation, two different values of the physical property will result in two response probability distributions (pdfs) with different means. Assuming a forced-choice paradigm, we might ask a human subject which of two stimuli contained a larger interaural difference. If the two perceptual variables resulting from the stimuli have

Gaussian pdfs with equal variance σ^2, an optimal decision variable (i.e., one that minimizes probability of error) is formed by taking the difference of the two measurements ($\mathbf{Y} = \mathbf{X_2} - \mathbf{X_1}$). If a larger interaural difference leads to a larger mean value of the perceptual variable, then an optimum decision rule is given by:

$$\mathbf{Y} = \mathbf{X_2} - \mathbf{X_1} \quad \underset{\text{Say } S_1}{\overset{\text{Say } S_2}{\gtrless}} \quad 0, \tag{4.1}$$

where "Say S_1" means "report that the larger interaural difference was present in stimulus 1", and "Say S_2" is the converse.

The decision variable is Gaussian with mean $E[\mathbf{Y}] = E[\mathbf{X_2}] - E[\mathbf{X_1}]$ and variance $2\sigma^2$. We may calculate the probability of a correct response as follows:

$$p(C) = p(\text{Say } S_1|S_1)p(S_1) + p(\text{Say } S_2|S_2)p(S_2), \tag{4.2}$$

where S_1 is the event where stimulus 1 contains the larger interaural difference, and S_2 is the converse. Assuming that the two stimuli are equally likely to contain the larger interaural difference, we have:

$$\begin{aligned} p(C) = \; & 0.5 \int_{-\infty}^{(m_2+m_1)/2} \mathcal{N}(m_1 - m_2, 2\sigma^2)dx + \\ & 0.5 \int_{(m_2+m_1)/2}^{\infty} \mathcal{N}(m_2 - m_1, 2\sigma^2)dx \end{aligned} \tag{4.3}$$

where m_2 is the mean value of the perceptual variable for the stimulus with the larger interaural difference and m_1 is the mean value resulting from the smaller interaural difference.

The probability of correct response is monotonically related to the variable:

$$d' = \frac{m_2 - m_1}{\sigma}. \tag{4.4}$$

With this decision model, we consider the interaural JND measurements reported by Hershkowitz and Durlach [12]. For a 500 Hz tone, the authors report the just-noticeable intensity and time-delay differences. Using a forced-choice paradigm, they report the JND as the interaural difference increment which results in a probability of correct response of 0.75, which corresponds to $d' \simeq 1.349$.

To a good first-order approximation over a reasonable range of IIDs and ITDs, the JNDs reported by Hershkowitz and Durlach are constant, with approximately a 0.88 dB JND for IID and approximately 11.7 μs JND for ITD. Since $d' \simeq 1$, these JNDs are nearly the same as the standard deviation, σ, of the perceptual noise. The data in [12] can not be extrapolated to other stimulus frequencies, but as a first approximation, we might assume that the perceptual variance is the same in all frequency channels.

These results may be used to guide the relative weighting of cues in our localization model. In the current model, interaural difference measurements are weighted samples of the time-varying interaural difference signals (as discussed in Section 3.3.1). We make the assumption that each measurement is corrupted by additive perceptual noise with some variance. The perceptual noise variance should be equal to the variance implied by the JND data minus the measurement variance inherent in the signal processing.

In the current implementation, we have not added perceptual noise to the interaural difference measurements. An examination of the outputs of the interaural difference estimators revealed that the measurement variance was at least as large as the variance implied by the JND data. There remains an issue of integration of information over time (i.e., over different "onsets"), which would tend to reduce the over-all variance, but the effect of such integration has not yet been examined.

4.3 Biases

The position-estimate output of the current model is generally biased, a result that stems directly from the interaural difference template calculations and subsequent spherical interpolation. We shall ignore biases caused by normal variations of the input spectrum (e.g., a pathological signal consisting of a line input-spectrum will exhibit interaural difference measurements which may be significantly different from the "average" templates. These irregularities will cause the peak in the likelihood map to shift to a different position.). Rather, we concern ourselves with two types of template extraction errors.

We have already mentioned the near-periodicity and multiple peaks in narrow-band cross-correlation functions (Section 3.2.5). When extracting the interaural phase delays, it is quite possible that the "wrong" peak is chosen on occasion. Clearly, an incorrect template value will create a bias in the output of the position estimator. These biases, however, are difficult to detect and to control.

Significant additional biases are introduced by errors in the spherical interpolation process. As discussed in Section 3.3.4, the interpolation surface is an intentional "underfit" to the data. At the measurement points, this results in slight template shifts, causing small biases in the position estimator output. This problem is exacerbated at intermediate positions, where the template values are purely based upon "nearby" template values rather than physical measurements at the intermediate positions.

The effect of these biases will become clear with the presentation of experimental data in Chapter 5.

4.4 Perceptual distance

The perceptual "distance" between two stimuli is generally not well-defined, except by analogy. One possible definition of perceptual distance is the number of JNDs between the two stimuli, a metric which allows us to quantify the analogies, making direct comparisons between stimuli. Since the d' term introduced in Section 4.2 is directly related to the number of JNDs when the measurement and perceptual noise are zero-mean Gaussian, we will call d' the perceptual distance. We may extend the expression for d' to a multidimensional context:

$$d' = \sqrt{(m_1 - m_2)^T K^{-1} (m_1 - m_2)}, \tag{4.5}$$

where m_1 and m_2 are the multidimensional perceptual-variable means induced by the two stimuli, and K is the covariance matrix of the total noise. This metric is the *Mahalanobis* distance between the two stimuli [30].

With the position estimator used in this model, we may use this definition of perceptual distance to estimate the *localization blur*[1] that will be exhibited by the model at various positions. Figure 4-2 is an example of an equal perceptual-distance contour map for the interaural difference template at 40° azimuth and 0° elevation. It was derived from the interpolated interaural difference templates with a particular choice of "measurement variance."

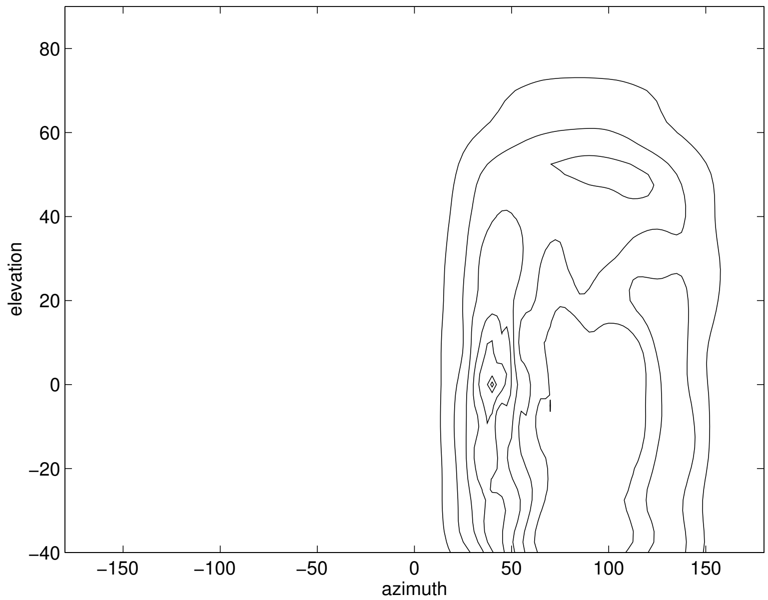

Figure 4-2: Equal perceptual-distance contour map based on interaural difference templates and a particular choice of measurement variance. The zero-distance point in this example is at ($\theta = 40°, \phi = 0°$).

A quantitative discussion of localization blur based upon perceptual distance is presented in Chapter 5.

At this point, we introduce a variance weighting factor α to our position estimator. We rewrite the likelihood function in Equation 3.10:

$$\mathcal{L}_{\theta,\phi} = K \exp\left\{ -\frac{\alpha}{2} \sum_{k=1}^{n} \frac{(P_k - \overline{P}_{\theta,\phi,k})}{\sigma_k^2} \right\}. \tag{4.6}$$

This weighting factor can be viewed as a scaling of the Mahalanobis distance:

$$d' = \sqrt{\alpha (m_1 - m_2)^T K^{-1} (m_1 - m_2)} \tag{4.7}$$

[1]As we define it, localization blur is to the JND of source position as the perceptual noise variance is to the JND in a psychophysical experiment.

$$= \sqrt{\alpha \sum_{k=1}^{n} \frac{(m_{1,k} - m_{2,k})^2}{\sigma_k^2}} \quad (4.8)$$

$$= \sqrt{\sum_{k=1}^{n} \frac{(m_{1,k} - m_{2,k})^2}{\sigma_k^2/\alpha}} \quad (4.9)$$

Thus for $\alpha < 1$, the perceptual distance is reduced by an equivalent increase of the variance estimates employed in the model. The introduction of the α factor may also be viewed as an exponentiation of the likelihood function.

Chapter 5

Comparisons with Human Performance

Several experiments have been completed to test the performance of the model. In this chapter, we report the results of an experiment designed to test the ability of the model to judge the position of a single broad-band sound source in free-field. We also present a brief analysis of the localization blur exhibited by the model, based on the broad-band localization experiment and on "perceptual distance" maps. Finally, we report the results of an experiment in which stimuli used by Zurek to quantify the precedence effect ([39]) were presented to the model.

All of the results in this chapter were calculated with the same variance "parameters" in the position estimator. We assume that the standard deviation of the IID noise is 0.65 dB and that the standard deviations of the IPD and IED noises are 0.05 ms and 0.1 ms respectively. These values were chosen as a compromise between the values implied by the Hershkowitz and Durlach JND measurements ([12]) and empirical measurement-variance estimates from the output of the interaural difference estimators with white-noise input stimuli.[1]

5.1 Localization of broad-band noise signals

Humans are quite good at localizing broad-band noise. The constant energy fluctuations in the noise signal act as continual onsets, allowing localization to be robust in the presence of reverberation. Since the interaural difference templates used in the model are based upon a broad-band input spectrum, noise signals are a logical stimulus set for testing the performance of the model. With broad-band noise, all of the cochlear filter channels will be stimulated to some degree, eliminating the need for energy thresholds or other channel-dependent weighting. This simplifies the implementation of the model

[1] Inspection of the model's output revealed that the measurement variance for the IID was approximately equal to the variance implied by the JND data, and that the IPD and IED measurement variances were larger than those implied by the ITD JND data, with especially large measurement variance in the IED. This additional variance is attributable to the measurement technique, and might be reduced by increasing the "width" of the window used in the cross-correlation.

and the analysis of its output.

The model was "trained" by calculating interaural difference templates at 177 of the KEMAR measurement positions in the right hemisphere. The training positions were chosen so as to leave 165 "test" positions interlaced between them. As described in Section 3.3.4, each interaural difference template function was approximated by 174 weighted "spherical Gaussian" kernels in the right hemisphere, with weights chosen to minimize the mean-squared error at the training positions. Templates for the left hemisphere were determined by symmetry.

Test signals were generated by convolving Gaussian pseudo-random noise with the head-related impulse responses at each KEMAR measurement position, except for positions on the median plane, for which the interaural differences are uniformly zero by construction. Each noise burst was approximately 400 ms in duration at a 44.1 kHz sampling rate. The noise was full-bandwidth (i.e., it contained energy up to at least 20 kHz) prior to convolution with the HRIRs.

Each test signal was passed through the model's front end, yielding a list of detected onsets in each filter-bank channel and a list of associated interaural difference measurements. Each list of onsets was divided into sub-lists of onsets occuring in each of four 100 ms time slices of the input signal. From each set of onsets, a spatial likelihood map was evaluated at the training positions, and a hill-climbing algorithm was used to find the global maximum of the map, which is the maximum-likelihood position estimate.

Figure 5-1 is a scatter-plot of perceived source position versus target position. Results for training and test positions are shown in Figures 5-1a and 5-1b respectively. The positions are reported in a spherical coordinate system with three polar axes: the "Left-Right" axis measures the angle that a given position makes with the median plane, the "Up-Down" axis measures the angle that a given position makes with the horizontal plane, and the "Front-Back" axis measures the angle that a given position makes with the coronal plane. The polar axis for the "Up-Down" and "Front-Back" angles is the interaural axis, and the polar axis for "Left-Right" angles is a vertical line passing through the center of the head. In Figures 5-1–5-3, all angles have been quantized to 2.5° "bins."

As can be seen in Figure 5-1a, the performance of the model at "training" positions is nearly perfect, with an average great-circle angular error of 0.66°. In the horizontal plane, the average absolute azimuthal error is 0.16° and the average absolute elevational error is 0.22°. This performance is qualitatively different from human performance in a number of respects. In humans, vertical (elevational) blur is generally larger than horizontal (azimuthal) blur, and horizontal blur generally increases with increasing azimuth. In contrast, there are no obvious trends in the localization blur implied by the model's output in this experiment. This is most likely due to the fact that no perceptual noise was explicitly added to the interaural difference measurements in the model. The JNDs measured by Hershkowitz and Durlach ([12]) increased with increasing interaural differences (a fact which might help explain the increased horizontal blur exhibited by humans at larger azimuths), but this increase is not reflected in the current model. The lack of gross errors in the model's output at training positions may be attributed to the correct extraction of interaural difference templates and reasonable extraction of interaural cues from the eardrum signals.

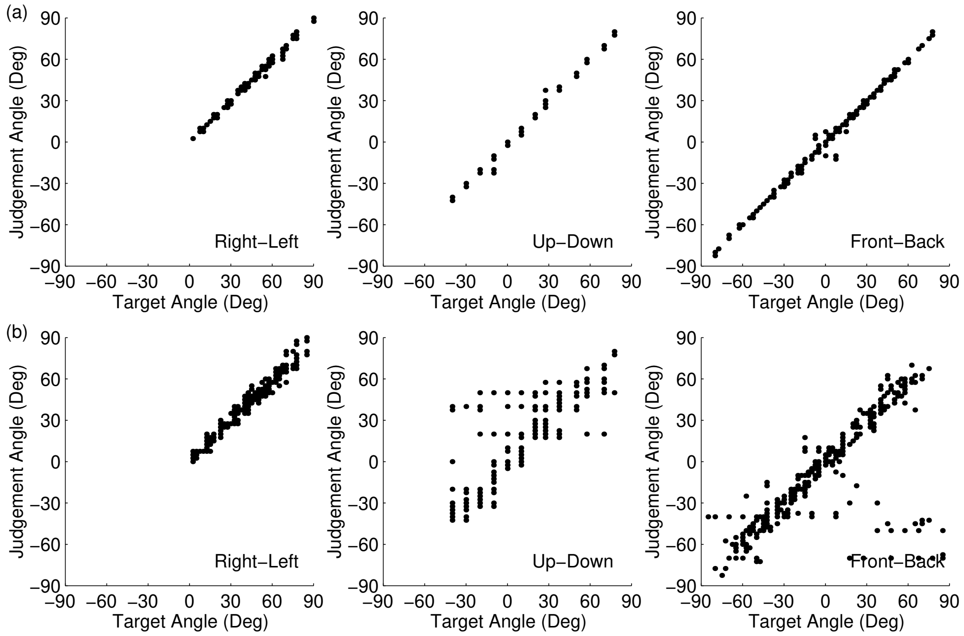

Figure 5-1: Scatter plot of perceived position versus target position. A description of the coordinate system may be found in the text. A line of unit slope passing through $(0, 0)$ represents the ideal output of the model. (a) Training positions, and (b) Test positions.

The model's performance at test positions is more comparable to human performance in terms of localization blur. It is clear from Figure 5-1b that the vertical component of the blur is much larger than the horizontal component. The mean absolute horizontal and vertical blurs are $4.15°$ and $8.21°$ respectively. It is dangerous, however, to draw strong conclusions based upon the error data in this experiment. The performance at the training positions suggests that the decreased performance at test positions is due to errors in the interpolation process. This is a *generalization* problem, which may simply be due to under-sampling of positions around the KEMAR dummy-head. It is quite likely that with a more densely sampled training set, the model would perform much better than humans over all positions more than a few degrees away from the median plane.

As a further test, the model's output was recalculated under two different conditions. In the first condition, only IID information was used to make position estimates. Figure 5-2a consists of scatter plots of the resulting output at training positions. In the second condition, only ITD cues were used. The result is shown in Figure 5-2b. Similar plots, for the test positions, are found in Figure 5-3. Estimates of localization blur for all conditions are given in Table 5.1.

It is clear from the results in Table 5.1 and Figures 5-1–5-3 that the IID cues account for the performance of the current model, particularly in elevation estimation. Interestingly, it is clear that the IID information, when integrated across frequency, is sufficient to resolve elevational ambiguity on a cone of confusion with resolution at

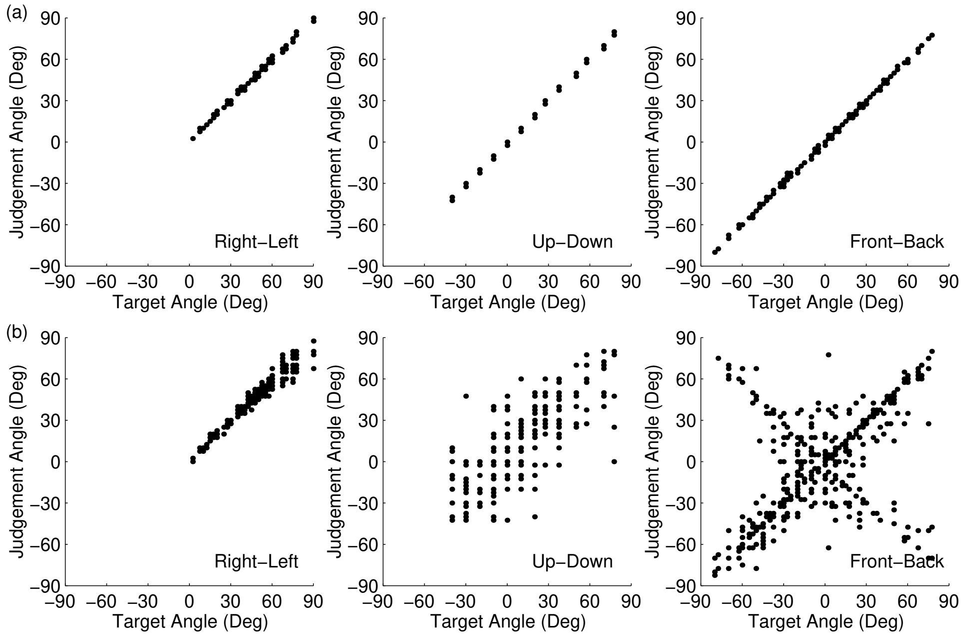

Figure 5-2: Localization data at training positions based upon: (a) IID only, and (b) ITD only.

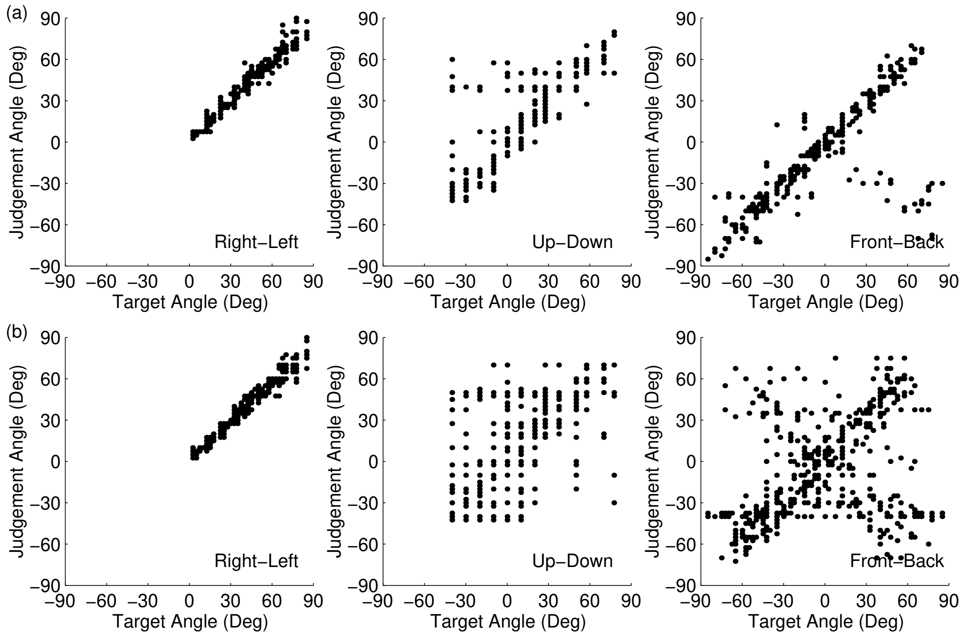

Figure 5-3: Localization performance at test positions based upon: (a) IID only, and (b) ITD only.

Position	Cues	$E[\lvert\theta - \hat{\theta}\rvert]$	σ_θ	$E[\lvert\phi - \hat{\phi}\rvert]$	σ_ϕ
Train	IID, ITD	0.16	0.30	0.22	0.40
Test	IID, ITD	4.15	4.46	8.21	13.36
Train	IID	0.15	0.31	0.28	0.56
Test	IID	3.97	4.44	9.90	16.93
Train	ITD	2.04	3.43	12.54	17.60
Test	ITD	3.94	3.83	22.64	26.54

Table 5.1: Estimates of localization blur for the experimental conditions described in 5.1. $E[|\theta - \hat{\theta}|]$ is the measured average absolute azimuth error, σ_θ is the standard deviation of the azimuthal measurement error, $E[|\phi - \hat{\phi}|]$ is the measured average absolute elevation error, and σ_ϕ is the standard deviation of the elevational measurement error. All measurements are averaged across positions in the horizontal plane.

least equal to that of humans.

Further, it is interesting to note the number of "front/back" confusions present in the ITD-only data. The confusions are represented by points close to a line of slope -1 passing through $(0, 0)$ in the "Front/Back" panels. This result is due to the strong front/back symmetry in the ITD template data.

5.2 Analysis of localization blur

As described in Section 4.4, it is possible to measure the "perceptual distance" between two stimuli by defining an appropriate metric. In this section, we make use of the expression for d' in Equation 4.9. We use the same variance estimates as were used in the noise-burst experiment.

In general, the localization blur exhibited by the model will depend upon the nature of the stimulus presented. For simplicity, we might employ impulsive stimuli, which humans are quite good at localizing. A very brief impulse will contain energy in all of the cochlear filter-bank channels, and we may assume that only one "onset" is detected in each channel. In order to avoid errors due to spherical interpolation and interaural difference estimation techniques, we employ the interpolated interaural difference templates directly as input to the model. With this formulation, the perceptual, or Mahalanobis, distance between two positions is given by:

$$d' = \sqrt{\sum_{k=1}^{N} \left[\frac{(P_k^I - T_k^I)^2}{\sigma_{I,k}^2} + \frac{(P_k^P - T_k^P)^2}{\sigma_{P,k}^2} + \frac{(P_k^E - T_k^E)^2}{\sigma_{E,k}^2} \right]}, \tag{5.1}$$

where P_k^I, P_k^P, and P_k^E are the IID, IPD and IED templates in channel k at position P, T_k^I, T_k^P, and T_k^E are the corresponding templates for position T, and $\sigma_{I,k}^2$, $\sigma_{P,k}^2$, and $\sigma_{E,k}^2$ are the corresponding variance estimates.

In our implementation, this expression can be rewritten:

$$d' = \sqrt{\sum_{i=1}^{24} \frac{(P_i^I - T_i^I)^2}{\sigma_I^2} + \sum_{j=1}^{14} \frac{(P_j^P - T_j^P)^2}{\sigma_P^2} + \sum_{k=12}^{24} \frac{(P_k^E - T_k^E)^2}{\sigma_E^2}}. \tag{5.2}$$

Here, $\sigma_{I,k}^2$, $\sigma_{P,k}^2$, and $\sigma_{E,k}^2$ are independent of frequency, and there are 24 filter-bank channels. The IPD is evaluated only in channels 1–14 (center frequencies below 1.6 kHz), and the IED is evaluated only in channels 12–24 (center frequencies above 1.0 kHz).

Some examples of perceptual-distance maps are given in Figures 5-4–5-8. In the figures, seven contours are drawn, at $d' = 2$, 4, 8, 16, 32, 64, and 128. There are several interesting features to note in the contour maps. In general, the ridges in the surfaces tend to be nearly circular in the latitude/longitude coordinate system and are of nearly constant azimuth in the "interaural" coordinate system. This result corresponds well with the notion of "cones of confusion" as discussed with regard to the spherical head model. It is worth noting that as the "zero-distance" point moves closer to the median plane, the localization blur becomes smaller in azimuth, but much larger in elevation. In Figures 5-4–5-8, the contours for small values of d' are difficult to see, but the $d' = 8$ and $d' = 16$ contours, for example, are sufficiently visible in the graphs that the positional variation of the localization blur may be seen with the naked eye.

To measure the localization blur more quantitatively as a function of position, the perceptual-distance map was evaluated for each of the training and test positions in the horizontal plane. At each position, the horizontal and vertical "widths" of the surface were quantified by measuring the great-circle angles in each direction at which the perceptual distance crossed the $d' = 2$, 4, 8, and 16 contours. The search area was limited to 40° total arc along each axis to reduce the computational load. Figures 5-9a and 5-9b show the horizontal and vertical blur respectively as a function of azimuth in the horizontal plane, with $d' = 4$. Corresponding curves for the other values of d' were not qualitatively different; thus, they have not been included.

5.3 Tests of the precedence effect

To test the operation of the precedence effect model, we employed stimuli used by Zurek to quantify the precedence effect in humans [39]. In the first experiment, two short noise bursts are presented in succession, the first diotic and the second containing a pure interaural delay or intensity difference. As the delay between the two bursts varied, Zurek measured IID and ITD JNDs in the trailing burst. In the second experiment, a 50 ms diotic noise burst is presented. At some time within the burst, a 5 ms portion of the signal is replaced with a new noise burst containing either a pure ITD or a pure IID. IID and ITD JNDs were measured as a function of sub-burst delay with respect to the beginning of the 50 ms burst.

These stimuli are not similar to any that occur in "natural" listening environments; in a given stimulus, only IIDs *or* ITDs are present. The model must therefore contend with "unnatural" combinations of interaural differences. Rakerd and Hartmann suggest that, in such a situation, humans "decide" which set of cues (ITD or IID) is more

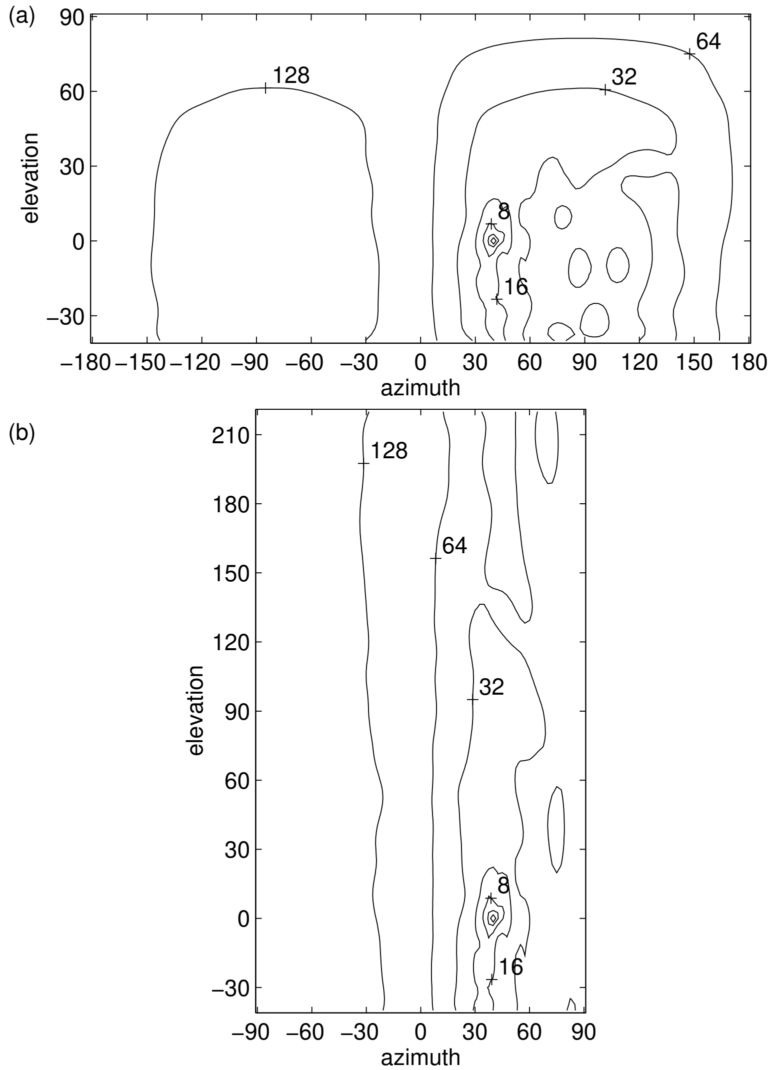

Figure 5-4: Equal perceptual-distance contour map based on interaural difference templates and the variance estimates used in Section 5.1. The zero-distance point in this example is at ($\theta = 40°, \phi = 0°$). (a) The latitude/longitude coordinate system, and (b) the "interaural" coordinate system.

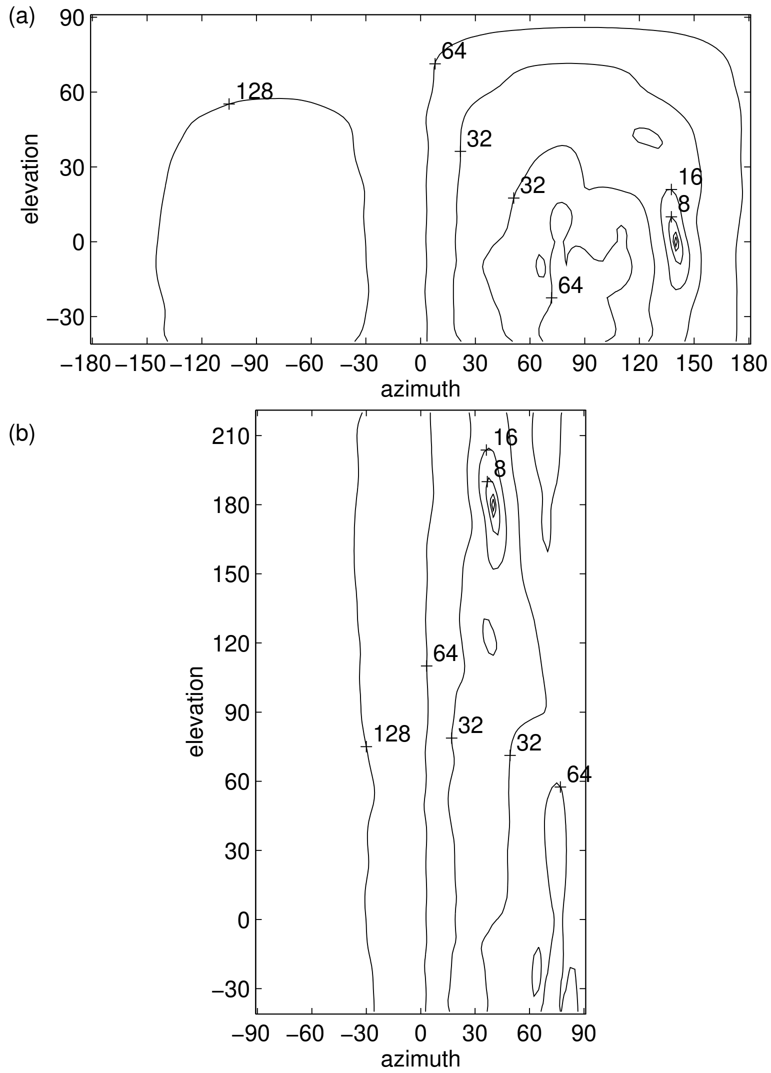

Figure 5-5: As in Figure 5-4, with the zero-distance point at $(\theta = 140°, \phi = 0°)$ in the latitude/longitude coordinate system.

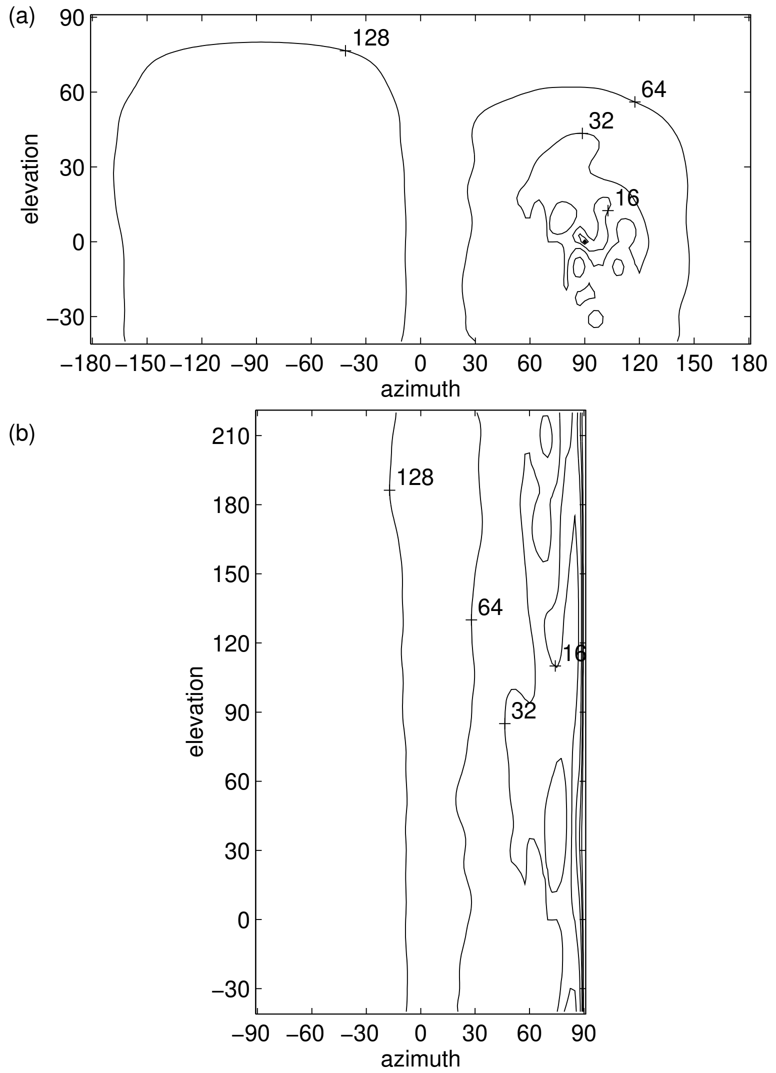

Figure 5-6: As in Figure 5-4, with the zero-distance point at ($\theta = 90°, \phi = 0°$) in the latitude/longitude coordinate system.

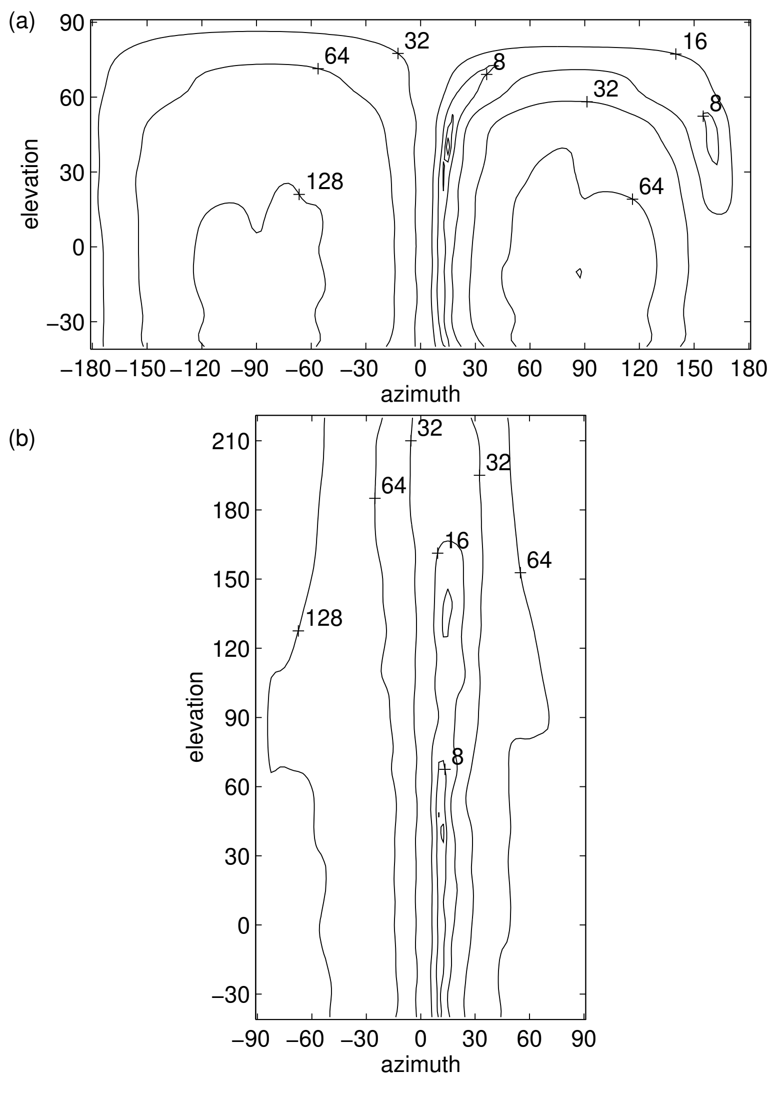

Figure 5-7: As in Figure 5-4, with the zero-distance point at ($\theta = 15°, \phi = 40°$) in the latitude/longitude coordinate system.

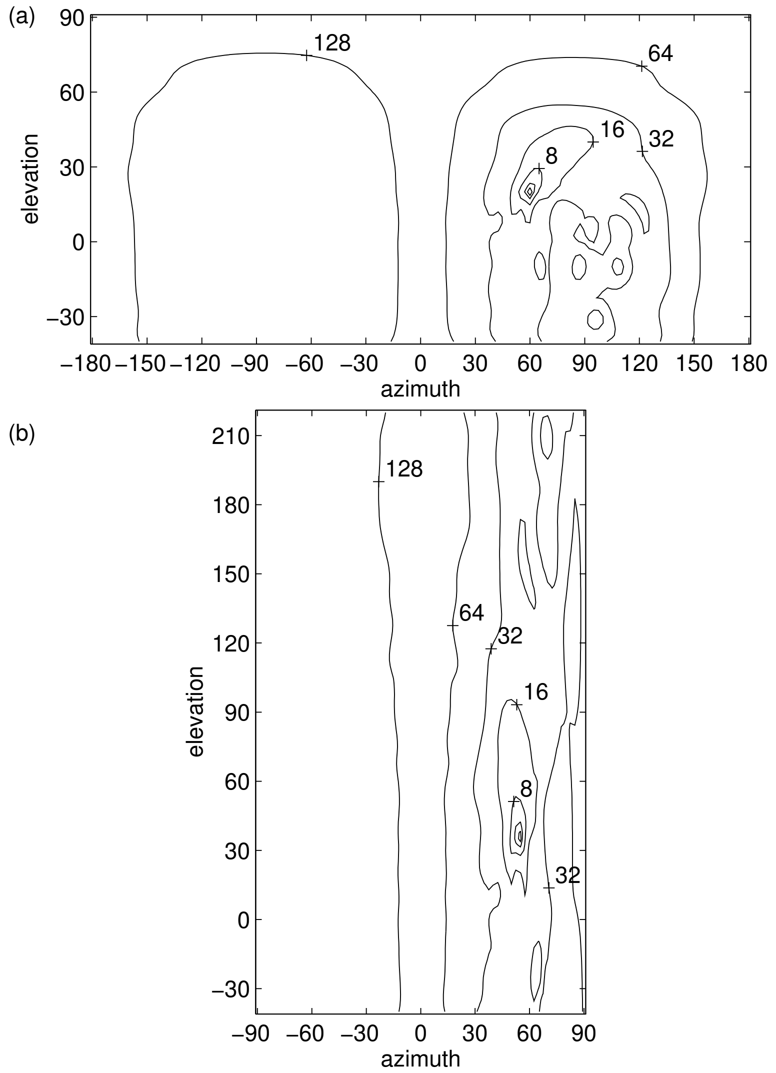

Figure 5-8: As in Figure 5-4, with the zero-distance point at ($\theta = 40°, \phi = 20°$) in the latitude/longitude coordinate system.

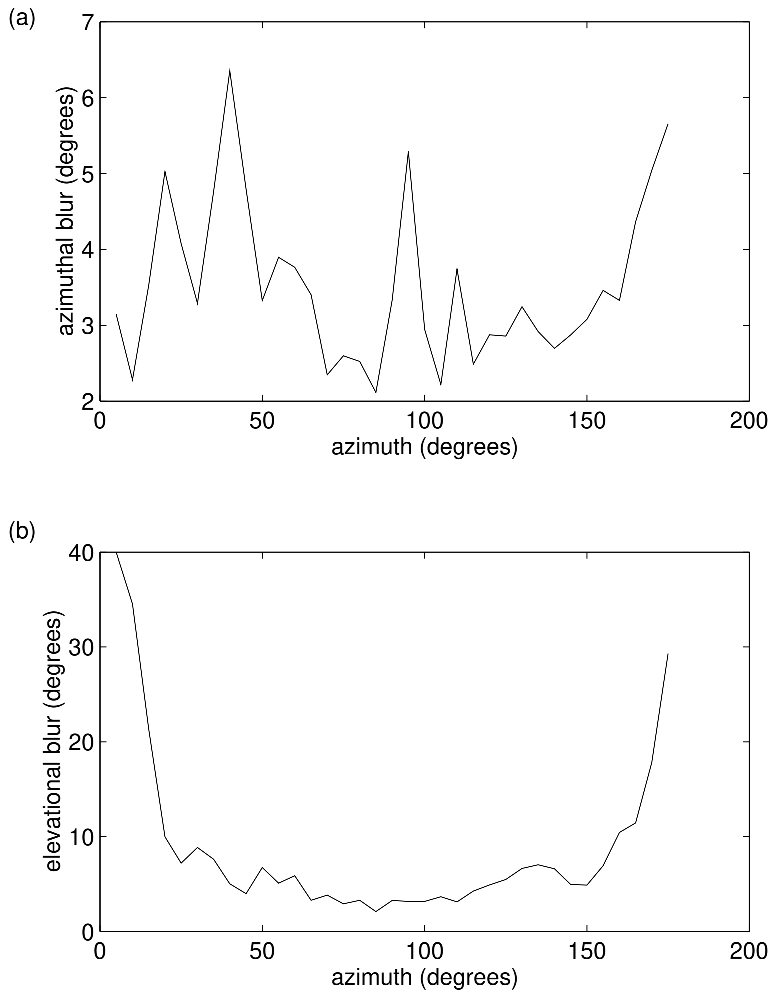

Figure 5-9: Localization blur as a function of azimuth in the horizontal plane (longitude/latitude coordinate system), as estimated from "perceptual-distance" maps. (a) azimuthal blur, and (b) elevational blur. The sharp rise in blur for sources near the median plane results from the fact that all interaural difference cues are very small for positions near the median plane.

reliable, and discount the other set [23].

The current model has no provision for determining which cues are more reliable. In order to facilitate comparisons with human responses, it was decided to modify the model slightly. In stimuli used to estimate the ITD JND, the IID is uniformly zero; thus, the model was programmed to dismiss the contradictory IID cues. Likewise, the ITD cues are dismissed in stimuli used to measure the IID JND. With this modification of the model, we turn to a detailed description of the experiments.

5.3.1 Experiment 1: Two noise bursts

The stimuli for the first precedence effect experiment consisted of two rectangularly gated 5 ms noise bursts, with a brief inter-burst delay. The first burst was diotic in all presentations, while the inter-burst delay and interaural characteristics of the trailing burst were varied from trial to trial. On a given trial, the trailing burst contained either a pure interaural delay or a pure intensity difference. An in-depth description of the stimuli may be found in [39].

In total, stimuli were presented with 11 different inter-burst delays, ranging from 0 to 20 ms. In Zurek's original experiment, JNDs were measured for IID and ITD at each inter-burst delay. In the present experiment, a number of different interaural difference values were presented to the model at each inter-burst delay. The ITD was probed in 25 μs increments from 25 μs to 400 μs, and the IID was probed in 1 dB increments from 1 dB to 15 dB. There were three repetitions of each condition.

Interpreting the output of the current model for the stimuli in this experiment was problematic for several reasons. There is no mechanism in the current model that decides which interaural cues are weighted most significantly in determining the perceived position of a stimulus. As already mentioned, it was decided that the model should dismiss IID cues in stimuli used to measure the ITD JND (and vice versa), in order to alleviate this problem. In the "interaural" coordinate system, the spatial likelihood maps arising from the stimuli in this experiment were generally well-described as ridges of nearly constant azimuth. To facilitate comparisons with Zurek's data, the likelihood map was numerically integrated over elevation to yield a unidimensional lateralization axis. The peak of the resulting "lateralization map" was recorded for each stimulus. There is no *a priori* reason to suspect that this approach will yield a good lateralization estimate, but it seemed a reasonable first step.

The lateralization output varied greatly over the three repetitions of each fixed condition. Figure 5-10 shows the results of lateralization measurements for various ITDs and IIDs at a 1 ms inter-burst delay. The cause of the large inter-trial variation is not apparent. A careful examination of individual interaural difference estimates for the stimuli in this experiment is warranted, but has not been completed due to time constraints. Though empirical evidence is currently absent, it does appear that the onset sampling algorithm employed in the current model is too rigid. The suppression curve in Figure 3-7a suggests a slow return of sensitivity. Zurek did not present quantitative measurements of the suppression curve *per se*, but it appears that the return of sensitivity in the current model may be too abrupt. In the current experiment, with stimuli containing short inter-burst delays (i.e., < 5 ms), information from the trailing sub-burst might be almost completely suppressed. Information from the trailing sub-burst in stimuli with larger inter-burst delays, however, would be weighed heavily. Clearly, the onset sampling strategy must be reconsidered.

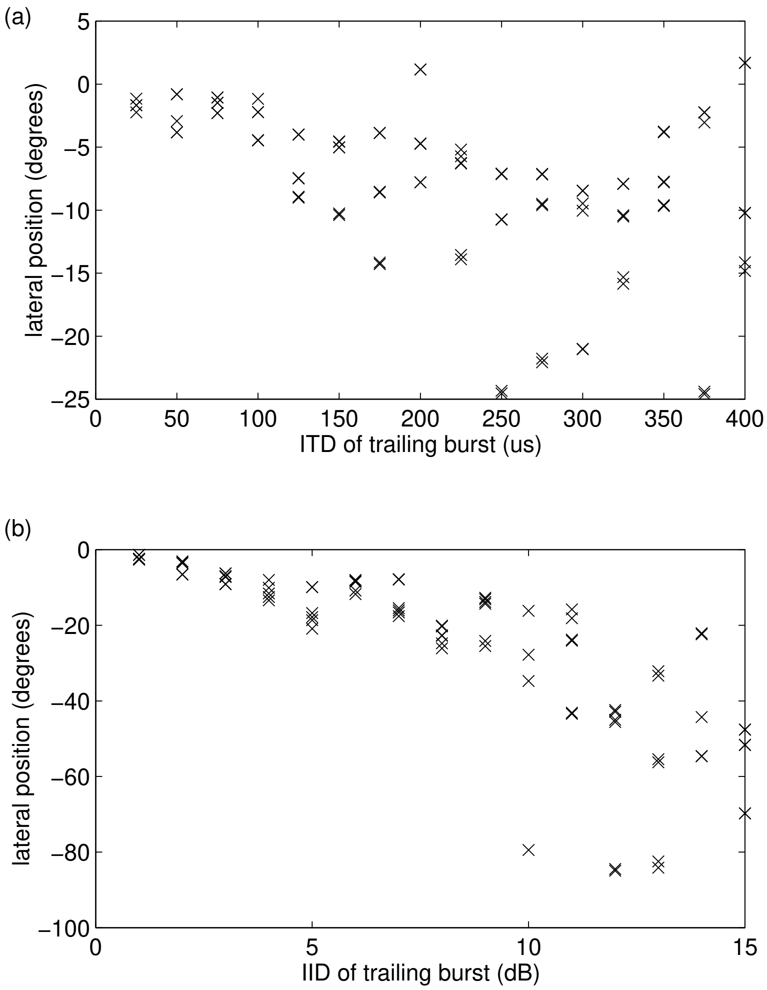

Figure 5-10: Outputs of the model for the two-burst stimuli with a 1 ms inter-burst delay and various interaural difference values. It may be noted that there are actually 6 ($\times$)s at each interaural value (though some may coincide). There were in fact two lateralization estimates performed for each of the three trials, based upon different values of the α term introduced in Section 4.4.

There does not appear to be any "clean" method of interpreting the results of this experiment. The best linear fit to the lateralization data at each inter-burst delay did not yield any information about the "suppression" of interaural cues as a function of inter-burst delay. The model will have to be more adaptable to "unnatural" stimuli and may require a revised onset/suppression algorithm before better comparisons may be made.

5.3.2 Experiment 2: Single burst with embedded sub-burst

In the second precedence effect experiment, each stimulus consisted of a 50 ms diotic noise burst with a 5 ms embedded dichotic segment. In each trial, the delay of the sub-burst relative to the beginning of the larger burst and the interaural characteristics of the sub-burst were varied. On a given trial, the rectangularly-gated sub-burst consisted of either a pure interaural delay or a pure intensity difference. In total, stimuli were presented with 14 different sub-burst delays, ranging from 0 to 45 ms. In Zurek's original experiment, JNDs were measured for IID and ITD at each sub-burst delay. In the present experiment, a number of different interaural difference values were presented to the model at each sub-burst delay. The ITD was probed in 25 μs increments from 25 μs to 325 μs, and the IID was probed in 1 dB increments from 1 dB to 15 dB. There were three repetitions of each condition. An in-depth description of the stimuli may be found in [39].

As in Experiment 1, the results of this experiment are difficult to interpret. As before, the model was programmed to dismiss IID cues in stimuli used to measure the ITD JND, and vice-versa. As before, the spatial-likelihood map was integrated over elevation to yield a lateralization estimate. Figure 5-11 shows the results of lateralization measurements for various ITDs and IIDs at a 1 ms sub-burst delay. Again, there is a great deal of variation between trials with the same parameters, and the onset sampling strategy may be the root cause. If an onset is detected shortly before the beginning of the 5 ms sub-burst, the onset sampling might "overlook" sub-burst lateralization information in that filter-bank channel. It is therefore possible that, on different trials, a variable number of channels contribute to the lateralization, thus causing the large output variation.

Further interpretation of the results will not be worthwhile until the model has been modified to treat "unnatural" stimuli in a more reasonable manner.

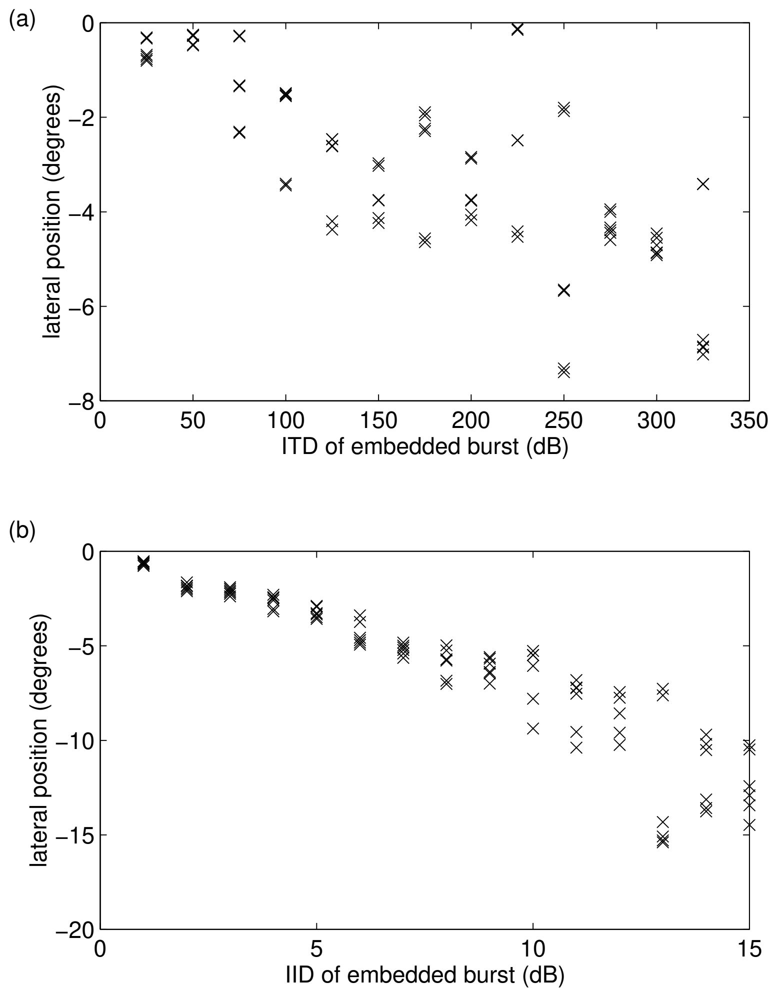

Figure 5-11: Outputs of the model for the single-burst stimuli with a 1 ms sub-burst delay and various interaural difference values. It may be noted that there are actually 6 ($\times$)s at each interaural value (though some may coincide). There were in fact two lateralization estimates performed for each of the three trials, based upon different values of the α term introduced in Section 4.4.

Chapter 6

Conclusions and Future Work

In this thesis, we have described a computational model of spatial hearing based on interaural difference cues. The current model is based on a loose interpretation of human auditory physiology, coupled with the well-known pattern-recognition technique of maximum-likelihood estimation.

As part of the construction of the model, we have described a method of extracting interaural intensity and time-delay differences from arbitrary binaural signals, and we have presented mathematical formulations for the "average" values of these signals as functions of sound source position based on HRTF measurements. Subsequent testing of the model with white-noise input signals has yielded several significant results:

- The interaural difference template extraction method yields templates which closely match actual interaural difference measurements made for many input signals.
- The IID extraction method, as implemented in the current model, appears to operate sufficiently well to achieve localization performance in both azimuth and elevation that is similar to or better than human performance.
- The IED and IPD extraction techniques do not appear to operate well-enough to correspond with human performance, as illustrated by the large measurement variance compared to the variance implied by the JNDs. It is likely that increasing the effective window width in the running cross-correlation functions will alleviate this problem to some degree. The current short window was chosen to allow the model to follow rapid changes in the interaural differences. A closer examination of this tradeoff in the running cross-correlation is certainly warranted.

As part of the interaural difference template extraction process, a method of "spherical interpolation" based on functional approximation has been introduced. This method has proven reasonable for interpolating the interaural difference functions, but it appears that the measurement positions used in the interpolation should be more densely spaced on the surface of the sphere than the 10° sampling that was employed. A much more dense sampling will not be possible without extensive HRTF measurements.

The current implementation of the model, in the interpreted MATLAB environment, is inadequate for exhaustive testing. On a DEC Alpha workstation, the model's front end requires approximately 20 minutes to analyze a 0.5 s segment of input signal. A careful re-implementation in C could probably reduce this time requirement by an order of magnitude, and it may be possible, on dedicated hardware, to implement the front-end in real-time. Taking great pains to improve the speed of the model is probably not a useful step at this time, however.

In describing the estimation portion of the model, we have attempted to justify a method of choosing appropriate variance estimates for the interaural differences based on human psychoacoustic data. This seems like a reasonable approach to fitting the model's performance to that of humans, but a more exhaustive set of human JND measurements will clearly be required.

Based on a set of variance estimates, we have proposed a "perceptual-distance" metric which is the Mahalanobis distance. In particular, the concept of a perceptual-distance map was introduced. Based on this paradigm, we have investigated the variation of localization blur over position. Several results are of interest:

- The elevational blur exhibited by the model increases with source proximity to the median plane. In the limit, sources on the median plane are indistinguishable in the current formulation because their interaural differences are uniformly zero by construction. It is suggested that a spectral-cue localization model like the one proposed by Zakarauskas and Cynader ([38]) will be an essential complement to the current model for purposes of explaining human localization in generality. Our belief is that the effect of the spectral-cue model will be significant only for sources in or near the median plane.

- "Cones of confusion" are present to some degree in the KEMAR HRTF data ([11]), and they are visible in the perceptual-distance maps as "circles" in the latitude/longitude coordinate system and as "lines" of constant azimuth in the "interaural" coordinate system.

- We have found no evidence in the KEMAR HRTF data to suggest that localization blur should be larger for sources *behind* a listener than for sources in front. In contrast, most investigations of human localization performance suggest that localization blur is much larger for sources at rear positions [1, 18]. Two possible explanations are: (1) the measurement technique used by researchers exaggerates localization blur for sources in the rear hemisphere, and (2) "interaural templates" used by humans may not be as good in the rear hemisphere, possibly because of the lack of a direct feedback system (the vision system in humans provides a strong feedback system for frontal sources).

Experiments designed to test the existence of a precedence effect in the model have highlighted several weaknesses in the current implementation. It is quite clear that the model must be adapted to deal with "unnatural" stimuli if its output is intended to be compared with human performance. It appears that the onset sampling strategy intended to model the precedence effect is inadequate. Localization information must be integrated more smoothly over time in order for the model's output to be consistent over multiple presentations of statistically similar input stimuli.

On the other hand, the discrete onset paradigm used in the current implementation holds great promise for extending the current model in other directions. In an acoustic environment containing multiple sources and reverberant energy, discrete onsets might be separated into groups arising from the various sources, and reverberant energy might be separated from the direct sound. Top-down approaches to solving this segregation problem have become our principal area of interest in the field of spatial hearing and auditory scene analysis. Before steps may be taken in this direction, however, it is evident that the onset-detection algorithm must be modified. One intuitively important aspect that is missing in the current implementation is a measure of the "sharpness" of an onset. Clearly, a sharp rise in energy is a more salient onset than a gradual rise.

Appendix A

Coordinate Systems

In this thesis, we refer to two head-centered spherical coordinate systems, both employing the variables azimuth (θ) and elevation (ϕ). The first is the familiar latitude/longitude system used for global positioning, which is useful for visualization since it "unwraps" to the familiar Mercator projection, with distortions occuring near the "north" and "south poles." In this coordinate system, the azimuth angle θ ranges through a full 360° interval, while the elevation angle ϕ is restricted to the range $-90° < \phi < 90°$.

The second coordinate system is similar, but the polar axis coincides with the interaural axis, and azimuth and elevation exchange roles. The azimuth angle θ is restricted to the range $-90° < \theta < 90°$, while the elevation angle ϕ ranges through a full 360° interval. This system is useful for localization since each well-known "cone of confusion" is equivalent to a single azimuth angle. Therefore, the cones of confusion encountered by the model will "unwrap" to lines of constant θ on the (θ, ϕ) plane, and distortions will occur at extreme azimuth angles (near the poles).

When viewing "unwrapped" surfaces in spherical coordinate systems, one must interpret the apparent results carefully. In an unwrapped map, distances are exaggerated near the poles of the coordinate system. This, for example, accounts for the appearance that Greenland is nearly the same size as the United States on a Mercator projection of the globe, when in fact it is much smaller in total area, but lies closer to the North pole. A comparison of positions, relative distances, and relative "sizes" for the two coordinate systems used in this thesis is shown in Figure A-2.

As the current model does not estimate the distance of sound sources, the two degrees of freedom (θ, ϕ) are sufficient to describe the range of possible source positions. The choice of coordinate system will be specified if it is not clear from the context.

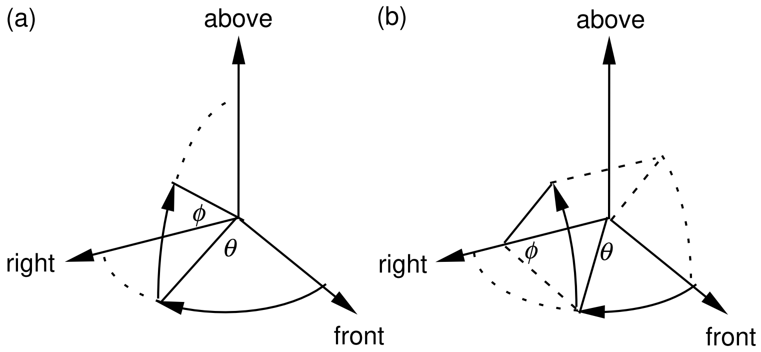

Figure A-1: The coordinate systems used in this paper: (a) latitude/longitude ($-180° < \theta < 180°, -90° < \phi < 90°$), (b) "interaural" ($-90° < \theta < 90°, -90° < \phi < 270°$).

Position Number	Latitude/Longitude θ	Latitude/Longitude ϕ	"Interaural" θ	"Interaural" ϕ
1	0	0	0	0
2	0	75	0	75
3	-155	-9	-25	190
4	-95	9	-80	120
5	70	0	70	0.0
6	120	-5	60	190
7	-47	-14	-45	-20

Table A.1: Kernel positions for the equal-distance contours shown in Figure A-2. All angles are given in degrees.

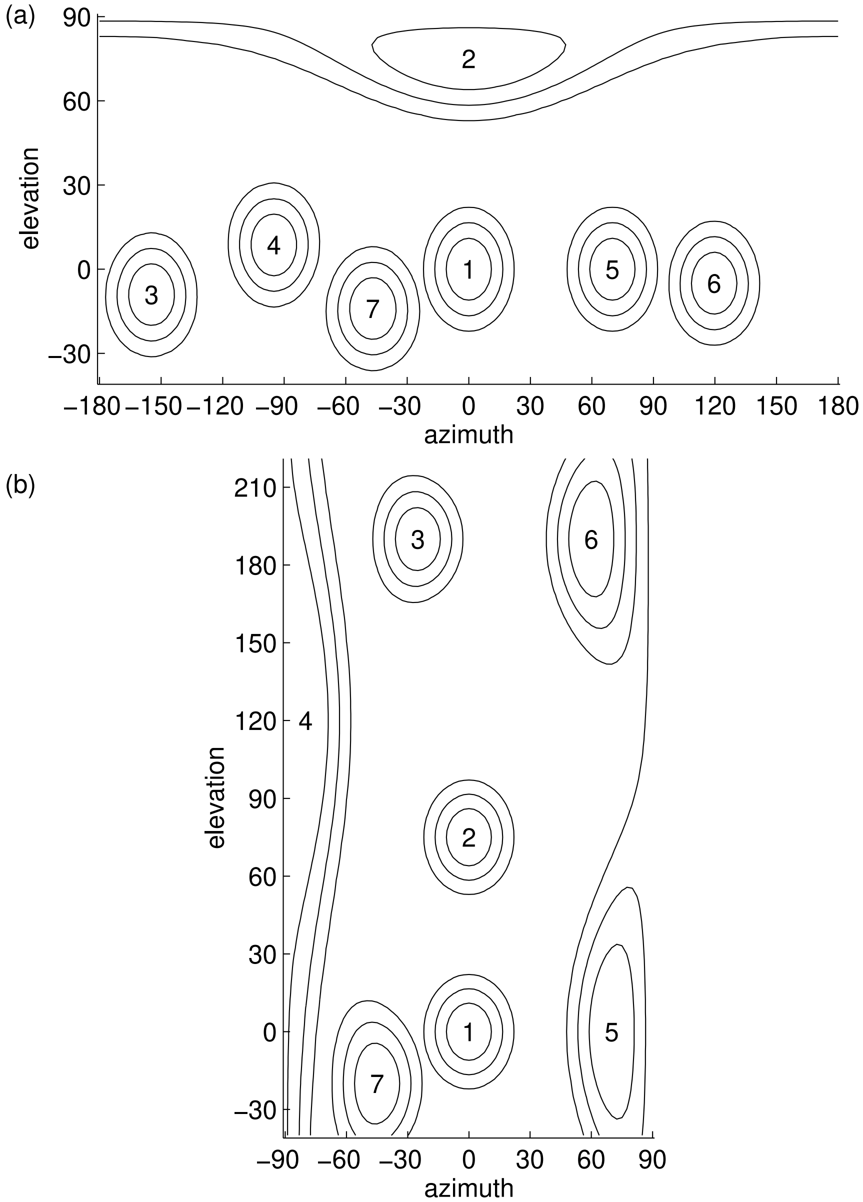

Figure A-2: Contours of equal great-circle distance in the two spherical coordinate systems employed in this thesis. A list of the plotted positions may be found in Table A.1. (a) The latitude/longitude coordinate system. Note the apparent increase in size at position two. (b) The "interaural" coordinate system. Note the "larger" contours at positions 4–6.

Bibliography

[1] J. Blauert. *Spatial Hearing.* MIT Press, Cambridge, MA, 1983.

[2] J. Blauert and W. Cobben. "Some Consideration of Binaural Cross Correlation Analysis". *Acustica*, 39(2):96–104, 1978.

[3] M. Bodden. "Modeling human sound-source localization and the cocktail-party-effect". *acta acustica*, 1:43–55, 1993.

[4] A. S. Bregman. *Auditory Scene Analysis.* MIT Press, Cambridge, MA, 1990.

[5] J. Chen, B. D. Van Veen, and K. E. Hecox. "A spatial feature extraction and regularization model for the head-related transfer function". *J. Acoust. Soc. Am.*, 97(1):439–452, 1995.

[6] H. S. Colburn and N. I. Durlach. "Models of Binaural Interaction". In *Handbook of Perception. Vol. IV*, chapter 11, pages 467–518. Academic Press, 1978.

[7] N. I. Durlach. "A Decision Model for Psychophysics". RLE Technical Report, Massachusetts Institute of Technology, 1968.

[8] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. "On the Externalization of Auditory Images". *Presence*, 1(2):251–257, Spring 1992.

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

[10] B. Gardner and K. Martin. "HRTF Measurements of a KEMAR Dummy-Head Microphone". Perceptual Computing Technical Report #280, MIT Media Lab, May 1994.

[11] W. G. Gardner and K. D. Martin. "HRTF measurements of a KEMAR". *J. Acoust. Soc. Am.*, 97(6), 1995.

[12] R. M. Hershkowitz and N. I. Durlach. "Interaural Time and Amplitude jnds for a 500-Hz Tone". *J. Acoust. Soc. Am.*, 46(6 (Part 2)):1464–1467, 1969.

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

[14] G. F. Kuhn. "Physical Acoustics and Measurements Pertaining to Directional Hearing". In W. A. Yost and G. Gourevitch, editors, *Directional Hearing*, chapter 1, pages 3–25. Springer-Verlag, New York, 1987.

[15] C. Lim and R. O. Duda. "Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model". In Preprint for the *28th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 1994.

[16] 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.

[17] E. A. Macpherson. "A Computer Model of Binaural Localization for Stereo Imaging Measurement". *J. Audio Eng. Soc.*, 39(9):604–622, 1991.

[18] J. C. Middlebrooks and D. M. Green. "Sound Localization by Human Listeners". *Annu. Rev. Psychol.*, 42:135–159, 1991.

[19] B. C. J. Moore. *An Introduction to the Psychology of Hearing.* Academic Press, Boston, MA, 1989.

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

[21] A. V. Oppenheim and A. S. Willsky. *Signals and Systems.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[22] J. O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press, Boston, MA, 1988.

[23] B. Rakerd and W. M. Hartmann. "Localization of sound in rooms, II: The effects of a single reflecting surface". *J. Acoust. Soc. Am.*, 78:524–533, 1985.

[24] L. Rayleigh. "On our perception of sound direction". *Philos. Mag.*, 13:214–232, 1907.

[25] D. D. Rife and J. Vanderkooy. "Transfer-Function Measurements using Maximum-Length Sequences". *J. Audio Eng. Soc.*, 37(6):419–444, 1989.

[26] B. M. Sayers and E. C. Cherry. "Mechanism of binaural fusion in the hearing of speech". *J. Acoust. Soc. Am.*, 29:973–987, 1957.

[27] E. A. G. Shaw. "Transformation of sound pressure level fron the free field to the eardrum in the horizontal plane". *J. Acoust. Soc. Am.*, 56(6):1848–1861, 1975.

[28] R. M. Stern and C. Trahiotis. "The Role of Consistency of Interaural Timing Over Frequency in Binaural Lateralization". In Y. Cazals, L. Demany, and K. Horner, editors, *Proceedings of the Ninth International Symposium on Auditory Physiology and Perception, Carcans, France.* Pergamon, Oxford, 1991.

[29] R. M. Stern, A. S. Zeiberg, and C. Trahiotis. "Lateralization of complex binaural stimuli: A weighted-image model". *J. Acoust. Soc. Am.*, 84(1):156–165, 1988.

[30] C. W. Therrien. *Decision Estimation and Classification.* John Wiley & Sons, New York, NY, 1992.

[31] C. Trahiotis and R. M. Stern. "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase". *J. Acoust. Soc. Am.*, 86(4):1285–1293, 1989.

[32] C. Trahiotis and R. M. Stern. "Across-frequency interaction in lateralization of complex binaural stimuli". *J. Acoust. Soc. Am.*, 96(6):3804–3806, 1994.

[33] J. Vanderkooy. "Aspects of MLS Measuring Systems". *J. Audio Eng. Soc.*, 42(4):219–231, 1994.

[34] F. L. Wightman and D. J. Kistler. "Headphone simulation of free-field listening. I: Stimulus Synthesis". *J. Acoust. Soc. Am.*, 85(2):858–867, 1989.

[35] F. L. Wightman and D. J. Kistler. "Headphone simulation of free-field listening. II: Verification". *J. Acoust. Soc. Am.*, 85(2):868–878, 1989.

[36] 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*, chapter 2, pages 26–48. Springer-Verlag, New York, 1987.

[37] R. S. Woodworth. *Experimental Psychology.* Holt, New York, 1938.

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

[39] P. M. Zurek. "The precedence effect and its possible role in the avoidance of interaural ambiguities". *J. Acoust. Soc. Am.*, 67(3):952–964, 1980.

[40] P. M. Zurek. "The Precedence Effect". In W. A. Yost and G. Gourevitch, editors, *Directional Hearing*, chapter 4, pages 85–105. Springer-Verlag, New York, 1987.

[41] P. M. Zurek. "A note on onset effects in binaural hearing". *J. Acoust. Soc. Am.*, 93(2):1200–1201, 1993.