A MODEL FOR AUDITORY LATERALIZATION
IN NON-STATIONARY MULTI-SOURCE ENVIRONMENTS

by

ZHENGJIN SHU

M.A.Sc., Beijing Institute of Technology, 1987
B.A.Sc., Beijing Institute of Technology, 1984

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE STUDIES
(Department of Electrical Engineering)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1995

© Zhengjin Shu, 1995
In presenting this thesis in partial fulfilment of the requirements for an advanced
degree at the University of British Columbia, I agree that the Library shall make it
freely available for reference and study. I further agree that permission for extensive
copying of this thesis for scholarly purposes may be granted by the head of my
department or by his or her representatives. It is understood that copying or
publication of this thesis for financial gain shall not be allowed without my written
permission.

(Signature)

Department of Electrical Engineering

The University of British Columbia
Vancouver, Canada

Date October 17, 1995
The problem addressed in this thesis is the determination of the positions of sound sources in non-stationary multi-source environments. This problem is approached by developing models that mimic the processing of sounds by the auditory system. It is well known that in the localization process the auditory system utilizes interaural intensity and time differences (IID and ITD) and the interaural envelope delay (IED). However, the way such cues are estimated and organized by the auditory system in non-stationary multi-source situations is not known.

It is argued in this thesis that the auditory localization process can be divided into three serial processing stages: decomposition, localization, and integration (DLI). Specifically, the signals detected by the two ears are first decomposed into their spectro-temporal distributions as represented in the neural activities of the auditory nerve fibers. Short-time spatial attributes, in terms of the localization cues, are then determined from energy concentrations in these distributions. A spatial scene of acoustic events is finally built by integrating the energy concentrations according to their spatial attributes.

A unique DLI model is proposed in which short-time cue estimation is realized by a process of pattern recognition and comparison using neural networks, and the spatial scene is represented by short-time cue distributions. Three implementations of the DLI model, which model separate auditory pathways responsible for three different types of cue sensitivity (IID, ITD, and IED) observed in the auditory system, are developed, and their performance in estimating short-time cue distributions are studied by computer simulation.

It is shown that there are unexplored patterns in the neural signals carried by the auditory nerve fibers that are important for auditory localization. These patterns are shown to contain good indications of interaural differences, and can be used to obtain robust short-time cue estimates by training neural networks that have relatively
simple structures. Furthermore, while such networks can be trained using the simplest types of inputs, they show the ability to generalize and perform well with more complex stimuli. It is demonstrated that the model works well in noise and in non-stationary multi-source situations.

The same model structure can be trained to estimate different localization cues, suggesting that the underlying structure of the different pathways responsible for different types of cue sensitivities in the auditory system may not necessarily be different. The receptive connection patterns of the hidden neurons in the model indicate that the spectro-temporal response properties of binaural neurons in the auditory system may play an important role in auditory localization, and that excitatory and inhibitory inputs to the binaural neurons play equally important roles in localization.
# TABLE OF CONTENTS

Abstract ..................................................................................................................................... ii
Table of Contents ...................................................................................................................... iv
List of Tables ............................................................................................................................... vii
List of Figures ............................................................................................................................. xii
List of Abbreviations ................................................................................................................. xviii
Acknowledgments ..................................................................................................................... xix

**INTRODUCTION** .................................................................................................................... 1

## 1 AUDITORY LOCALIZATION ................................................................................................. 5
  §1.1 Introduction ....................................................................................................................... 5
  §1.2 Approaches Used in the Study of Auditory Localization .............................................. 6
  §1.3 A Brief Review of the Structure of the Auditory System ............................................ 7
  §1.4 Binaural Hearing ............................................................................................................. 9
  §1.5 Interaural Intensity Differences (IIDs) ........................................................................ 12
  §1.6 Interaural Time Differences (ITDs) ............................................................................. 18
  §1.7 Free Field Auditory Localization ................................................................................. 25
  §1.8 Interaural Envelope Delay ......................................................................................... 28
  §1.9 Models of Auditory Localization ................................................................................ 29
  §1.10 Cross-correlation Based Models ............................................................................... 32
  §1.11 Models of cue sensitive neurons .............................................................................. 42

## 2 MODELING AUDITORY LOCALIZATION IN COMPLEX ACOUSTIC ENVIRONMENTS ......................................................................................................................... 49

## 3 A DLI MODEL FOR AUDITORY LOCALIZATION ................................................................ 58
  §3.1 A DLI Model .................................................................................................................. 58
  §3.2 Three Parallel Implementations of the DLI Model ...................................................... 58
C. Resolution Power for the Separation of Two Sound Sources .......... 104
D. Uncorrelated On-Off Modulations of Different Sound Sources ....... 104

4.3.6 Summary Table of Tests for the Evaluation of the Networks .......... 105

5 TEST RESULTS OF ITD ESTIMATION ......................................................... 106
§5.1 Pure-Tone Stimuli ................................................................. 106
§5.2 Two-Tone Complex Stimuli .................................................. 109
Analysis of the Input Patterns of the ITD Network for Two-Tone Stimuli .......... 111
§5.3 Pink Noise Stimuli .............................................................. 114
§5.4 Two Similar Sources from Different Directions ......................... 116

6 TEST RESULTS OF IID ESTIMATION .................................................... 122
§6.1 Pure-Tone Stimuli ................................................................. 122
§6.2 Two-Tone Complex Stimuli .................................................. 124
§6.3 Pink Noise Stimuli .............................................................. 127
§6.4 Two Similar Sources from Different Directions ......................... 129

7 TEST RESULTS OF IED ESTIMATION ..................................................... 133
§7.1 AM Stimuli ........................................................................... 133
§7.2 Stimuli Modulated by Two-Tone Complex Signals ..................... 135
§7.3 Two AM Sources from Different Directions ............................. 137

CONCLUSIONS ............................................................................. 142

REFERENCES ............................................................................. 152
LIST OF TABLES

Table 4.1 Networks (ITD, IID, IED) to be tested for the listed stimulus type and parameter combinations. ................................................................. 105

Table 5.1 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 250 μs ITD and a 20 dB SNR. ........................................................................................................ 107

Table 5.2 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 250 μs ITD and a 0 dB SNR. ................................................................. 107

Table 5.3 ITD estimates and the errors of the mean position estimates for pure-tone stimuli with different ITDs and a 20 dB SNR. ......................... 110

Table 5.4 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 250 μs ITD and a 20 dB SNR. ................................................................. 110

Table 5.5 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 250 μs ITD and a 0 dB SNR. ................................................................. 110

Table 5.6 ITD estimates and the errors of the mean position estimates for two-tone complex stimuli with a 250 μs ITD and a 20 dB SNR. FD: frequency difference between the two components in the stimuli ................................................................. 111

Table 5.7 ITD estimates and the errors of the mean position estimates for pink noise stimuli with different bandwidths (BW). The SNR used in the tests was 20 dB, and the ITD of the stimuli was 250 μs. ................. 115

Table 5.8 ITD estimates and the error of the mean position estimate for pink noise stimuli with a 0 dB SNR and a 250 μs ITD. The bandwidth of the stimuli was 50 Hz. ................................................................. 116
Table 5.9 ITD estimates and the errors of the mean position estimates for pink noise stimuli with different center frequencies (CF). The SNR used in the tests was 20 dB, and the ITD of the stimuli was 250 μs.

Table 5.10 ITD estimates and the position errors for two-source cases with different on-off modulation periods (MP). The SNR was 20 dB, and the ITDs of the two sources were -250 μs and 250 μs, respectively.

Table 5.11 ITD estimates and the position errors for a two-source case with a 0 dB SNR. The ITDs of the two sources were -250 μs and 250 μs, respectively.

Table 5.12 ITD estimates and the position errors for two-source cases with different spatial separations (SS, in terms of the two ITDs). The SNR was 20 dB.

Table 5.13 ITD estimates and the position errors for a two-source case with uncorrelated on-off modulations. The SNR was 20 dB.

Table 6.1 IID estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 3 dB IID and a 20 dB SNR.

Table 6.2 IID estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 3 dB IID and a 5 dB SNR.

Table 6.3 IID estimates and the errors of the mean position estimates for pure-tone stimuli with different IIDs and a 20 dB SNR.

Table 6.4 IID estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 3 dB IID and a 20 dB SNR.
Table 6.5  IID estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 3 dB IID and a 5 dB SNR. ................................................................. 126

Table 6.6  IID estimates and the errors of the mean position estimates for two-tone complex stimuli with a 3 dB IID and a 20 dB SNR. FD: frequency difference between the two components in the stimuli. .......... 126

Table 6.7  IID estimates and the errors of the mean position estimates for pink noise stimuli with different bandwidths (BW). The SNR used in the tests was 20 dB, and the IID of the stimuli was 3 dB. ..................... 127

Table 6.8  IID estimates (mean ± standard deviation) and the error of the mean position estimate for pink noise stimuli with a 5 dB SNR and a 3 dB IID. The bandwidth of the stimuli was 500 Hz. ........................................ 128

Table 6.9  IID estimates and the errors of the mean position estimates for pink noise stimuli with different center frequencies (CF). The SNR used in the tests was 20 dB, and the IID of the stimuli was 3 dB. .......... 129

Table 6.10  IID estimates and the position errors for two-source cases with different on-off modulation periods (MP). The SNR was 20 dB, and the IIDs of the two sources were -3 dB and 3 dB, respectively. ................................................................. 130

Table 6.11  IID estimates and the position errors for a two-source case with a 5 dB SNR. The IIDs of the two sources were -3 dB and 3 dB, respectively. ................................................................. 131

Table 6.12  IID estimates and the position errors for two-source cases with different spatial separation (SS, in terms of the two IIDs). The SNR was 20 dB. ................................................................. 131

Table 6.13  ITD estimates and the position errors for a two-source case with uncorrelated on-off modulations. The SNR was 20 dB. ................. 132
Table 7.1  IED estimates (mean ± standard deviation) and the error of the 
mean position estimate for AM stimuli with a 0.83 ms IED and a 
20 dB SNR. ........................................................................................................ 134

Table 7.2  IED estimates (mean ± standard deviation) and the error of the 
mean position estimate for AM stimuli with a 0.83 ms IED and a 
5 dB SNR. ........................................................................................................ 134

Table 7.3  IED estimates and the errors of the mean position estimates for AM 
stimuli with different IEDs and a 20 dB SNR. ..................................................... 134

Table 7.4  IED estimates (mean ± standard deviation) and the error of the 
mean position estimate for stimuli with two-tone complex 
modulation waveforms. The IED to be estimated was 0.83 ms. 
The SNR in the test was 20 dB. ................................................................. 136

Table 7.5  IED estimates (mean ± standard deviation) and the error of the 
mean position estimate for stimuli with two-tone complex 
modulation waveforms. The IED to be estimated was 0.83 ms. 
The SNR in the test was 5 dB. ................................................................. 136

Table 7.6  IED estimates and the errors of the mean position estimates for 
stimuli with two-tone complex modulation waveforms. The IED 
to be estimated was 0.83 ms. The SNR in the test was 20 dB. 
FD: frequency difference between the two components in the 
modulation waveforms.............................................................................. 137

Table 7.7  IED estimates and the position errors for two-source cases with 
different on-off modulation periods (OOP). The SNR was 20 
dB, and the IEDs of the two sources were ±0.83 ms, 
respectively. .................................................................................................... 138
Table 7.8  IED estimates and the position errors for a two-source cases with a
17 ms on-off modulation periods (OOP). The SNR was 5 dB, and the IEDs of the two sources were ±0.83 ms, respectively.

Table 7.9  IED estimates and the position errors for two-source cases with
different spatial separations (SS, in terms of the two IEDs). The
SNR was 20 dB.

Table 7.10  IED estimates and the position errors for a two-source case with
random on-off modulations. The SNR was 20 dB, and the IEDs
of the two sources were ±0.83 ms, respectively.
LIST OF FIGURES

Fig. 1.1 Three major parts of the auditory system. .................................................. 8
Fig. 1.2 Four parts of the auditory periphery. ........................................................... 8
Fig. 1.3 The main ascending auditory pathways of the brainstem. For abbrevia-
tions see the main text. (After Pickles, 1988, Page 170.) ............................... 9
Fig. 1.4 A coordinate system for sound localization................................................. 10
Fig. 1.5 Interaural amplitude ratio $\alpha$ for tones as a function of the azimuth
angle $\theta$ and the frequency $f$ (Shaw, 1974a, b). (After Durlach and
Colburn, 1978.) ........................................................................................................ 11
Fig. 1.6 Interaural time difference $\tau$ for tones and clicks as a function of the
azimuth angle $\theta$ (Shaw, 1974a, b). The solid lines represent data
obtained by Firestone (1930), Mills (1958), Nordlund (1962), and
Feddersen et al. (1957). The lower dashed curve is derived from the
formula $\tau = r/c(\theta + \sin(\theta))$ and the upper dashed curve from the
formula $\tau = r/c(3\sin(\theta))$, where $r = 8.75$ cm and $c = 344$ m/s
(Woodworth, 1938; Rayleigh, 1945). (After Durlach and Colburn,
1978.) ......................................................................................................................... 12
Fig. 1.7 Lateral displacement (calibrated as follows: 0 corresponds to $\theta = 0^\circ$, -5 to
$\theta = -90^\circ$, and 5 to $\theta = +90^\circ$) of the auditory event as a function of the
interaural intensity difference (dB). (After Blauert, 1983, Page 158.) .......... 14
Fig. 1.8 The perceived location of the pure tone sound image as a function of the
interaural intensity difference. The perceived location is normalized
across the studies with 0% corresponding to $\theta = 0^\circ$ and 100% to $\theta = \pm 90^\circ$.
(After Yost and Hafter, 1987.) .................................................................................. 15
Fig. 1.9 The IID threshold for perceived change as a function of frequency for
tones. The parameters (0, 9, and 15 dB) are the IIDs of the referent
tones which serve to mark positions in lateral space. (After Yost and Hafter, 1987.)

Fig. 1.10 Effects of varying IIDs on the discharge rate of a neuron in the cat dorsal nucleus of the lateral lemniscus. (After Kuwada and Yin, 1987.)

Fig. 1.11 Responses of a superior colliculus neuron of cat to changes in IID. (After Kuwada and Yin, 1987.)

Fig. 1.12 Binaural coherence curves for tones of different frequencies. P denotes the percentage of judgments “to the left,” and positive τ corresponds to the left ear leading. (After Durlach and Colburn, 1978.)

Fig. 1.13 The perceived location of the pure tone sound image as a function of the interaural time difference (in terms of degrees of interaural phase difference). The perceived location is normalized with 0% corresponding to θ = 0° and 100% to θ = ±90°. (After Yost and Hafter, 1987.)

Fig. 1.14 The perceived locations (calibrated as follows: 0 corresponds to θ = 0°, 5 to θ = 90°, and -5 to θ = -90°) of impulse sound images as a function of the interaural time difference. (After Blauert, 1983, Page 144.)

Fig. 1.15 The ITD (in terms of degrees of interaural phase difference) threshold for perceived change as a function of frequency for tones. The parameters are the ITDs (again, in terms of degrees of interaural phase difference) of the referent tones which serve to mark positions in lateral space. (After Yost and Hafter, 1987.)

Fig. 1.16 Responses (A and B) of two inferior colliculus neurons to changes in ITD, and waveforms (C, D, and E) of three types of binaural stimuli used in the stimulation of these two neurons. (After Kuwada and Yin, 1983.)
Fig. 1.17 Angle of just noticeable difference (JND) \((\Delta \theta)_0\) for tone bursts as a function of the tone frequency \(f\) and the angle \(\theta\) between the sound source and the median plane. (After Durlach and Colburn, 1978.)

Fig. 1.18 Comparison made by Mills (1960, 1972) of interaural phase JND and the interaural amplitude JND for ITD=0 and IID=0 with the changes in ITD \(\Delta \phi\) and in IID \(\Delta \alpha\) that occur when an actual source is moved a just noticeable angle from the median plane. (After Durlach and Colburn, 1978.)

Fig. 1.19 The coincidence network proposed by Jeffress in 1948 for the localization of low-frequency tones. (After Colburn and Durlach, 1978.)

Fig. 1.20 A block diagram of the cross-correlation model of Sayers and Cherry (1957). The term \(e^{-k|t|}\) in the figure corresponds to the first weighting factor, and the term \(A_L\) or \(A_R\) to the second weighting factor discussed in the main text. The term \(A_v(t)\) refers to temporal average. (After Colburn and Durlach, 1978.)

Fig. 1.21 A block diagram of the auditory nerve based model by Colburn (1973, 1977). (After Colburn and Durlach, 1978.)

Fig. 1.22 Schematic diagram of the sterausis binaural network suggested by Shamma (1989). \(c_{ij}\) denotes the output of the \((i,j)\)-th neuron in the matrix, which receives inputs from the ipsilateral fiber \(X_i\) and the contralateral fiber \(Y_j\).

Fig. 1.23 Location of the peaks of the multi-channel cross-correlation function for broad-band noise with an ITD of 1500 µs. The vertical axis indicates the center frequency of the different band-pass filter channels, while the horizontal axis indicates the argument (ITD) of the cross-correlation function. (After Stern et al., 1988.)
Fig. 1.24 Schematic diagram of the model by Sujaku et al. (1981) for ITD sensitive inferior colliculus neurons. ................................................................. 44

Fig. 1.25 Schematic diagram of the connections from AVCN and MNTB to the LSO column in the model by Reed and Blum (1990) ........................................... 47

Fig. 2.1 Simulated responses of a cochlear model (Shamma et al., 1989) to a 600 Hz sinusoidal stimulus. (After Shamma et al., 1989.) ........................................... 53

Fig. 3.1 A schematic diagram of a DLI model for auditory localization in complex acoustic environments. ................................................................. 59

Fig. 3.2 Schematic diagrams of three implementations of the generic DLI model (shown in Fig. 3.1): (a) ITD estimation; (b) IID estimation; and (c) IED estimation. ST-IID: short-time IID. ST-ITD: short-time ITD. ST-IED: short-time IED. ................................................................. 61

Fig. 3.3 Example of neural activities of a group of nine (vertical axis) modeled auditory nerve fibers in responding to a pure tone stimuli of 900 Hz. .......... 70

Fig. 3.4 Frequency spectrum of the signal in Eq. 3.2. ........................................... 72

Fig. 3.5 High modulation rates. Example of neural activities of a group of auditory nerve fibers when the stimulus has relatively high modulation rates. Each square represents the activation of a particular auditory nerve fiber at a particular time slice. The white area in each square is proportional to the activation level. AN: auditory nerve. ....................... 73

Fig. 3.6 Intermediate modulation rates. Example of neural activities of a group of auditory nerve fibers when the stimulus has intermediate modulation rates. Each square represents the activation of a particular auditory nerve fiber at a particular time slice. The white area in each square is proportional to the activation level. AN: auditory nerve. ....................... 74

Fig. 3.7 Low modulation rates. Example of neural activities of a group of auditory nerve fibers when the stimulus has relatively low modulation
rates. Each square represents the activation of a particular auditory nerve fiber at a particular time slice. The white area in each square is proportional to the activation level. AN: auditory nerve. ........................................ 75

Fig. 3.8 A schematic diagram of the structure of the networks for short-time cue estimations. The squares represent neurons in the network. The size of the white area in a square represents the amplitude of the neuron's activity, which is normalized to be from zero to one. The adjacent layers in the network are fully connected in a feed-forward fashion. See text for details on the connection between layers. ........................................ 76

Fig. 3.9 An example of the tuning curve of an output layer neuron. The arrowhead line at ITD = 0.1 ms represents the activity of the neuron for this particular ITD value. ........................................ 79

Fig. 3.10 An example of the coding of cue values by an array of output layer neurons. The arrowhead lines indicate the activities of the individual neurons. The mean of the Gaussian envelope is the encoded cue value. ...... 80

Fig. 3.11 An example of the ITD distribution function. The sound stimuli used in the simulation were 1 kHz pure-tones on which white noise was added with a 20 dB signal-noise ratio. The ITD in the stimuli was 250 µs. The peak around 250 µs indicates the correct estimate of the true ITD value. ...... 83

Fig. 4.1 A model of the neural signals in auditory nerve fibers (de Boer and de Jongh, 1978). ........................................ 86

Fig. 4.2 A schematic diagram of the auditory periphery model by Carney (1993). See text for descriptions of the different elements in the model. (Adopted from Carney, 1993.) ........................................ 87

Fig. 4.3 Parameters associated with an ITD distribution peak. ........................................ 99

Fig. 4.4 Examples of the two square-wave signals, $S_1(t)$ and $S_2(t)$, which are orthogonal to each other. ........................................ 103

xvi
Fig. 5.1 ITD estimates shown as a function of the true ITDs. .................. 109

Fig. 5.2 Example input activation patterns of the ITD network for a two-tone complex stimulus whose two components have a 250 Hz difference. ........ 113

Fig. 5.3 An example ITD distribution function for a two-source case with a 25 ms on-off modulation period. .................................................. 118

Fig. 5.4 The ITD distribution histogram for the uncorrelated two-source case. The parameters that describe the two target sources are shown in Table 5.13 ................................................................. 121

Fig. 6.1 IID estimates shown as a function of the true IIIDs. .................. 125

Fig. 6.2 An example IID distribution function for a two-source case with a 35 ms on-off modulation period. .................................................. 130

Fig. 6.3 The IID distribution function for the uncorrelated two-source case. The parameters that describe the two target sources are shown in Table 6.13 .... 132

Fig. 7.1 Estimated IEDs shown as a function of the true IEDs ................ 135

Fig. 7.2 The resulting IED distribution function for the test with a 17 ms on-off modulation period. .................................................. 139

Fig. 7.3 The IED distribution function for the two-source case with random on-off modulations. Parameters associated with the two lateral peaks are shown in Table 7.10 .................................................. 141
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>Amplitude Modulation</td>
</tr>
<tr>
<td>AN</td>
<td>Auditory Nerve</td>
</tr>
<tr>
<td>AVCN</td>
<td>Anteroventral Cochlear Nuclei</td>
</tr>
<tr>
<td>BF</td>
<td>Binaural Facilitation</td>
</tr>
<tr>
<td>BW</td>
<td>Bandwidth</td>
</tr>
<tr>
<td>CF</td>
<td>Characteristic Frequency</td>
</tr>
<tr>
<td>DC</td>
<td>Direct Current</td>
</tr>
<tr>
<td>DCN</td>
<td>Dorsal Cochlear Nuclei</td>
</tr>
<tr>
<td>DLI</td>
<td>Decomposition-Localization-Integration</td>
</tr>
<tr>
<td>EE</td>
<td>Excitatory-Excitatory</td>
</tr>
<tr>
<td>EI</td>
<td>Excitatory-Inhibitory</td>
</tr>
<tr>
<td>FD</td>
<td>Frequency Difference</td>
</tr>
<tr>
<td>IC</td>
<td>Inferior colliculus</td>
</tr>
<tr>
<td>IED</td>
<td>Interaural Envelope Delay</td>
</tr>
<tr>
<td>IID</td>
<td>Interaural Intensity Differences</td>
</tr>
<tr>
<td>ITD</td>
<td>Interaural Time Differences</td>
</tr>
<tr>
<td>JND</td>
<td>Just Noticeable Difference</td>
</tr>
<tr>
<td>LL</td>
<td>Lateral Lemniscus</td>
</tr>
<tr>
<td>LSO</td>
<td>Lateral Superior Olive</td>
</tr>
<tr>
<td>MAA</td>
<td>Minimum Audible Angle</td>
</tr>
<tr>
<td>MGB</td>
<td>Medial Geniculate Body</td>
</tr>
<tr>
<td>ML</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td>MNTB</td>
<td>Medial Nucleus of the Trapezoid Body</td>
</tr>
<tr>
<td>MP</td>
<td>Modulation Period</td>
</tr>
<tr>
<td>MSO</td>
<td>Medial Superior Olive</td>
</tr>
<tr>
<td>MTB</td>
<td>Medial Trapezoid Body</td>
</tr>
<tr>
<td>OOP</td>
<td>On-Off Modulation Period</td>
</tr>
<tr>
<td>PVCN</td>
<td>Posteroventral Cochlear Nuclei</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal Noise Ratio</td>
</tr>
<tr>
<td>SOC</td>
<td>Superior Olivary Complex</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound Pressure Level</td>
</tr>
<tr>
<td>SS</td>
<td>Spatial Separation</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

I would like to express my sincere gratitude and thanks to my supervisor in the Department of Electrical Engineering, Dr. Charles A. Laszlo; without his constant encouragement and invaluable guidance, this thesis would not have been possible. I am also deeply indebted to my supervisor in the Department of Ophthalmology, Dr. Max S. Cynader, for his invaluable support, inspiration, and guidance. I have learned a great deal from both of them during the course of this research.

Special thanks go to Dr. Nicholas V. Swindale, who has given me many suggestions over the past few years, and has commented on the draft of this thesis. Thanks also go to Dr. Robert M. Douglas for his help with computer programming and answers to questions about psychology, and to Dr. Pierre Zakarauskas for many inspiring discussions.

I would also like to take this opportunity to thank Doris Metcalf, graduate secretary in the Department of Electrical Engineering, and Barry Gibbs, secretary in the Ophthalmology Research Laboratory, for their help over the years of my study at UBC. Thanks also go to other staff members of the Departments of Electrical Engineering and Ophthalmology whose help did not go unnoticed.

To my parents
Introduction

The ability to localize sound sources enables us humans to obtain important information about what and where an event happens. It also contributes to our ability to understand speech. We could be very confused and sometimes in danger if we could not determine where a sound comes from. There is no doubt that auditory localization plays an important role in the survival of a large number of biological species. A question for us is how the auditory system solves the problem of sound localization.

The problem of localizing sound sources also arises in many engineering systems, and applications include underwater object detection, medical diagnosis, surveillance systems, robot sensing, and hearing aid devices. Passive sonar applications are particularly relevant to sound localization. Although much research has been conducted to improve the performance of such engineering systems, the physiological ear is still superior to artificial devices under similar conditions. In fact, the powerful ability of the auditory system to localize sound sources has stimulated curiosity among researchers in various disciplines for more than a century. In particular, physiologists and psychologists have been focusing their inquiries on why and how the auditory system works, while engineers and computer scientists have been interested in the duplication of auditory functions and in methods to accomplish the localization task. It is likely, however, that an integrated understanding of the problem of auditory localization would benefit from mutual interactions among research in different areas. Such integrated understanding is necessary to meet the challenge of developing artificial devices that emulate auditory localization with only two sensors.
Auditory localization occurs in both the horizontal and vertical planes (Fig. 1.4, Page 10). Horizontal localization is dominant, and under some conditions it is called lateralization. In this thesis, the term "localization" refers to horizontal localization, unless it is clear from the context that it is used in its more general meaning.

Previous studies have identified certain essential aspects of auditory localization. Cues used by the auditory system to localize sound sources, and possible neural mechanisms for the using of these cues have been studied and documented for decades. Specifically, three binaural cues, the interaural time (ITD) and intensity (IID) differences, and the interaural envelope delay (IED) of the sound signals detected by the two ears, play important roles in the sound localization process. Theoretical studies have been carried out to investigate how these cues may be processed in the brain in order to obtain estimates of the directions of sound sources. In addition, models have been created to integrate such hypotheses with experimental data. These efforts give rise to interesting ideas for the design of artificial devices that mimic the auditory localization process.

Despite the previous efforts in the study of auditory localization, the understanding of how such localization is achieved in noisy, dynamic, and multi-source environments, which are typical in nature, remains a challenging problem. In fact, this problem has not been specifically addressed with enough emphasis by published models, which are designed to fit data from experiments conducted in static acoustic environments with high signal-noise ratios (SNRs). However, the auditory system can localize sound sources with excellent performance under noisy and time-varying conditions. Investigations in this direction may provide an opportunity to discover new mechanisms of or insights into auditory localization that may not be so evident under more simplified and idealized assumptions about the acoustic environment. Another reason for the consideration of more complex and realistic acoustic environments is that any practical localization device must work in natural
(noisy) situations. Thus, it is the objective of this work to study mechanisms for auditory localization in noisy and time-varying acoustic environments.

We argue that in order for the auditory system to localize sound sources under such conditions, it is essential for the system to estimate the location cues quickly and robustly. Thus, our approach is to develop and study models that estimate the well known binaural cues in such fashion in noise and in short time duration. This modeling strategy is also crucial in developing models that can localize sound sources in multi-source environments.

Localizing multiple sound sources with only two ears is a difficult problem. However, the auditory system can localize multiple sound sources in rather complicated situations. One possible reason that the auditory system can do this may be due to the ability of the system to decompose the acoustic input signals into different frequency bands. By focusing on different frequency bands while performing the localization task, sources with different frequency contents can be separately localized. When two sound sources have similar frequency contents, however, the problem of separating the directions of the two sources remains unsolved.

In order for the auditory system to solve the problem is such situations, the localization cues need to be estimated quickly and robustly in a changing acoustic environment. We note that, in a natural acoustic environment, there are many transient effects in the sense that sounds come and go, and the energies emitted from the sound sources change from time to time. Thus, we can model the acoustic environment by assuming that multiple sounds are organized in such a way that, in most small time intervals, there always is one sound that has significantly higher intensity than the other sounds in the environment. This loudest sound can then be treated as the signal in the corresponding time windows, and all other sounds as noise.
In this way it is possible to pick up multiple sound sources that emit over the entire observation period but have changing relative intensity or energy.

Our specific objectives in this thesis include the development of models that can estimate the three different binaural cues quickly and robustly, and the development of methods of integrating these estimates over time to obtain a spatial representation of the acoustic environment.

We first review, in Chapter 1, the literature related to auditory localization. We then, in Chapter 2, discuss our modeling strategies and the motivations of our model. In Chapter 3, we describe, in detail, our model for the estimation of the binaural cues in noisy, dynamic, and multi-source environments. In Chapter 4, we describe the methods used in the simulation of the model described in Chapter 3. In the following chapters (5, 6, and 7) we present estimation results of the individual cues (ITDs, IIDs, and IEDs) respectively. We finally conclude the thesis with a summary and discussion of the significance of our model, and suggestions for future work.
1 Auditory Localization

§1.1 Introduction

In everyday life, our ability to tell where a sound is coming from is so natural that we often take it for granted without paying much attention to the power of auditory localization. Scientific inquiry into auditory localization, however, dates back to the early nineteenth century. Around as early as 1800, an Italian physicist named Giovanni B. Venturi carried out some free-field tests using brief flute tones on normal and unilaterally deaf listeners. He found that his deaf listeners could not localize the sound correctly, and hypothesized that the difference in sound intensity between the two ears was the basis of localization (Carterette, 1978).

The nineteen century psychologists wondered why nonspatial tones could be spatially perceived. Rayleigh in 1877 showed that, for high-frequency tones, the interaural ratio of the sound intensities at the two ears is crucial for localization. This observation was further supported by evidence from some experiments by Thompson in 1879. These observations led to an intensity theory of localization (Carterette, 1978). Research done in the twentieth century has confirmed that, while the intensity cue is an important one, the intensity theory is not the whole story.

Early this century, using dichotic\(^1\) stimuli, researchers (e.g. Rayleigh, 1907; Wilson and Myers, 1908) demonstrated that low-frequency tones could be localized on the side of leading phase. In 1908, Mallock suggested that localization could be based on interaural time differences.

In 1907, Rayleigh formulated, for the first time, the duplex theory of localization, which states that the localization of low-frequency sounds depends on interaural time differences.

---

\(^1\) Stimuli presented through headphones are called dichotic stimuli if the two signals at the two ears are not identical.
differences (ITDs) while that of high-frequency sounds depends on interaural intensity differences (IIDs).

During the past several decades, a large number of studies have been carried out to further investigate the mechanisms involved in auditory localization. These studies give us a much more complicated picture regarding how auditory localization is achieved. While the ideas of ITDs and IIDs being important localization cues remain true, the duplex theory of localization has proved to be only partially correct. In this chapter, we shall give a detailed review of the more recent studies on auditory localization.

§1.2 Approaches Used in the Study of Auditory Localization

As described in the last section, research on auditory localization has a rather long history. Various approaches have been used to understand the possible mechanisms underlying sound localization by different biological species. This diversity of approaches is a reflection on the complexity of the problem.

At first, localization has been studied using psychophysical methods. Based on the measurements of various thresholds of sensation generated by different acoustic stimuli, functional mechanisms could be hypothesized. Indeed, research done using such methods has the longest history (over a century) producing a major part of our understanding of auditory localization.

Second, physical acoustics has been used to study the physical aspects of sound stimuli (frequency, intensity, etc.) that make localization possible. The relationship between these aspects and the direction of a sound source has been investigated using both theoretical analysis and direct measurements. These studies have established the physical foundation of directional hearing.

Another important aspect of auditory localization research is the physiology of auditory localization. By recording the electrical activities of neurons in different
sites along the auditory pathways, this approach has been a powerful tool in revealing the neural mechanisms of auditory localization.

Finally, the theoretical approach has been used along with experimental studies. Although the majority of research has been experimental, theoretical analysis has played an important role in interpreting the data, in hypothesizing the possible mechanisms of the auditory system, and in integrating our understanding of the problem. Moreover, the results of this aspect of the research, i.e. the models, are most useful in engineering applications

§1.3 A Brief Review of the Structure of the Auditory System

The auditory system can roughly be divided into three parts: the auditory periphery, the brainstem auditory nuclei, and the auditory cortex. The primary information flow in the auditory system is from the periphery through the brainstem auditory nuclei to the auditory cortex, as shown in Fig. 1.1. The auditory periphery is stimulated by the sound coming into the ear. The sound pressure signal is transformed in the auditory periphery into neural activities of the auditory nerve fibers, which link the auditory periphery to the brainstem auditory nuclei. The neural activities of the auditory nerve fibers are relayed through a number of brainstem auditory nuclei, where certain types of processing occur, to the auditory cortex. The pathways from the periphery through the brainstem to the cortex are called the ascending pathways, as opposed to the various feedback pathways (centrifugal pathways) from the cortex to the brainstem and to the periphery.

The auditory periphery itself can in turn be divided into four parts: the outer ear, the middle ear, the cochlea, and the auditory nerve, as shown in the schematic diagram in Fig. 1.2. The middle ear couples sound energy in the outer ear to the cochlea. It can be viewed as a mechanical transformer which matches the impedance of the air in the outer ear with the much higher impedance of the fluids inside the
cochlea. The cochlea transforms mechanical vibration into neural signals carried by the auditory nerve fibers. This transformation has been the subject of both experimental and modeling research for many decades (Allen, 1985).

![Diagram of the auditory system]

**Fig. 1.1** Three major parts of the auditory system.

![Diagram of the auditory periphery]

**Fig. 1.2** Four parts of the auditory periphery.

The neural signals carried by the auditory nerve fibers are processed and relayed to the auditory cortex through a number of "stations" in the brainstem, where neural information converges and diverges. These "stations" are referred to as nuclei. **Fig. 1.3** shows a schematic diagram of the main ascending auditory pathways of the brainstem (Pickles, 1988).

The neural signals in the auditory nerve are first fed into the cochlear nucleus (CN) which consists of three parts, the dorsal (DCN), posteroventral (PVCN) and anteroventral (AVCN) cochlear nuclei. One part of the pathways from the cochlear nucleus leads to the superior olivary complex (SOC), while another part bypasses this
complex and leads to the lateral lemniscus (LL) and its nucleus. The superior olivary complex consists of three parts, the lateral superior olive (LSO), the medial superior olive (MSO), and the medial trapezoid body (MTB). Among them, two parts, the LSO and the MSO, receive information from both left and right cochlea nuclei. Signals from LSO and MSO are then fed into the lateral lemniscus, which in turn sends signals to the inferior colliculus (IC). Signals from IC are sent to the medial geniculate body (MGB), which in turn sends signals to the auditory cortex.

Fig. 1.3 The main ascending auditory pathways of the brainstem. For abbreviations see the main text. (After Pickles, 1988, Page 170.)

§1.4 Binaural Hearing

Fig. 1.4 shows a coordinate system for the study of sound localization. The stimulus signals detected by the two ears when a sound source is emitting energy from a specific direction \((\theta, \phi)\) generally contain subtle differences. As mentioned in Section
1.1, the importance of such differences has long been recognized by early researchers (Rayleigh, 1907). The interaural differences caused by a single source have been studied by a number of people (e.g. Shaw, 1974a, b; Steinberg and Snow, 1934; Kuhn, 1979). Fig. 1.5 and Fig. 1.6 (After Durlach and Coiburn, 1978) show summaries made by Shaw (1974a,b) of data obtained by several researchers (Firestone, 1930; Mills, 1958; Nordlund, 1962; Feddersen et al., 1957; Woodworth, 1938; and Rayleigh, 1945).

Fig. 1.4 A coordinate system for sound localization.

Despite these differences between the two signals at the two ears, we usually perceive only one sound image coming from a direction in the vicinity of the source.
We say that a sound image is fused if the image is spatially compact and unitary. If the image is perceived to be outside the head, we say that the image is externalized.

![Interaural amplitude ratio](image)

*Fig. 1.5 Interaural amplitude ratio $\alpha$ for tones as a function of the azimuth angle $\theta$ and the frequency $f$ (Shaw, 1974a, b). (After Durlach and Colburn, 1978.)*

In order to study the different effects of different parameters in the interaural differences, researchers often use headphones to deliver stimuli to the listener’s two ears with well defined interaural differences. Headphone stimulation often produces a sound image inside the head. When the parameters in the interaural differences are changed, the image can usually be moved around inside one’s head, and such movement is along the left-right axis. Thus, the task of judging the lateral positions of such sound images is referred to as *lateralization*. By controlling the exact interaural differences in the two signals at the listener’s two ears, the effects of the interaural differences on the lateralization of the sound image can be studied. We shall refer to binaural stimuli as diotic if there is no interaural difference between the
two signals at the two ears. When there are interaural differences between the two signals, we refer to the stimuli as dichotic.

![Graph showing interaural time difference τ for tones and clicks as a function of the azimuth angle θ.](image)

Fig. 1.6 Interaural time difference τ for tones and clicks as a function of the azimuth angle θ (Shaw, 1974a, b). The solid lines represent data obtained by Firestone (1930), Mills (1958), Nordlund (1962), and Feddersen et al. (1957). The lower dashed curve is derived from the formula \( τ = r/c(θ + \sin(θ)) \) and the upper dashed curve from the formula \( τ = r/c(3\sin(θ)) \), where \( r = 8.75 \text{ cm} \) and \( c = 344 \text{ m/s} \) (Woodworth, 1938; Rayleigh, 1945). (After Durlach and Colburn, 1978.)

§1.5 Interaural Intensity Differences (IIDs)

Over the past several decades the effects of the IID on the perceived lateral position of the sound image have been studied by many researchers. One question concerning such effects is: what value of the IID can cause complete lateralization of the
perceived sound image. Pinheiro and Tobin reported in 1969 that the IID required for a noise burst to be perceived as from the side of the head was roughly 10 dB, independent of the level and duration of the burst. Similar results were reported by Bekesy in 1959 for a click train of 100 pulses per second. Flanagan et al. (1964) found that the IID required for a click to be perceived from the side of the head is less than 10 dB. Others (e.g. Guttman, 1962a; Moushegian and Jeffress, 1959; Sayers, 1964; Whitworth and Jeffress, 1961) reported that the value of IID for complete lateralization of clicks and tones was much greater than 10 dB.

Another issue is the different effects of the IID under different conditions. Bekesy in 1960 reported that the lateralization effect of the IID was approximately the same for tones above 3 kHz, for noise, and for clicks. Moushegian and Jeffress (1964) and Feddersen et al. (1957) found that the lateralization of the sound image for a certain IID value was frequency dependent for tones.

Some later studies tried to quantify the effects of the IID on lateralization for a range of IID values, giving a function of the IID over the perceived lateralization of the sound image. Fig. 1.7 shows data from Blauert, 1983, where the perceived lateral positions of both broad-band noise and 600 Hz tones are plotted as functions of the IID. Until full lateralization is attained, the function is more or less linear. Fig. 1.8 shows a comparison made by Yost and Hafter (1987) of data obtained in two studies (Watson and Mittler, 1965; and Yost, 1981). Yost and Hafter (1987) have concluded that the relationship between the lateral position and the IID is approximately the same over a considerable range of frequency, overall level, and duration.

Sensitivity of the auditory system to changes in the interaural intensity difference has also been studied by a number of researchers (Mill, 1960; Rowland and Tobias, 1967; Herskowitz and Durlach, 1969; Grantham, 1984). The smallest change in the IID that leads to a change in the perceived lateral position of the sound image was measured in these studies. Fig. 1.9 shows a comparison made by Yost and Hafter
(1987) of studies by Mills (1960), Yost and Dye (1987). Different curves in Fig. 1.9 correspond to different IID distances around which the perceived smallest changes in IID were measured. Thus, the different curves demonstrate the sensitivity of the auditory system to changes in IID at different positions of the sound images. Zero dB IID corresponds to a sound image in the center of the head; 15 dB IID corresponds to a sound image fully lateralized to one side of the head; nine dB IID corresponds to a sound image somewhere between the positions of the images produced by a 0 dB IID and a 15 dB IID.

![Diagram](image)

Fig. 1.7 Lateral displacement (calibrated as follows: 0 corresponds to $\theta = 0^\circ$, -5 to $\theta = -90^\circ$, and 5 to $\theta = +90^\circ$) of the auditory event as a function of the interaural intensity difference (dB). (After Blauert, 1983, Page 158.)

The upward shifting of the curves from the 0 dB case to 15 dB case indicates that the sensitivity to IID changes is highest when the sound image is perceived to be in the center of the head. The sensitivity becomes poorer when a sound image is perceived to be more lateral. Also note that the auditory system is less sensitive to changes in the IID for tones at frequencies around 1 kHz.
The effects of varying the IID on the responses of auditory neurons have also been studied for several decades. These studies have been aiming at elucidating the functional roles of binaural neurons and the neural mechanisms of the IID sensitivity observed in psychophysical experiments. Neurons sensitive to the IID have been found in various sites along the auditory pathways, including the superior olivary complex (Boudreau and Tsuchitani, 1968; Goldberg and Brown, 1968; Caird and Klinke, 1983), dorsal nucleus of the lateral lemniscus (Brugge et al., 1970), inferior colliculus (Ross et al., 1966; Geisler et al., 1969; Stillman, 1972), superior colliculus (Schechter et al., 1981; Hirsh et al., 1985; Yin et al., 1985), medial geniculate body (Aitkin and Webster, 1972), and auditory cortex (Brugge et al., 1969; Brugge and Merzenich, 1973; Imig and Adrian, 1977; Phillips and Irvine, 1981). The characteristic frequencies (CFs) of IID sensitive neurons fall mostly in the higher region in the audible frequency range. This is consistent with the psychophysical observation that high-frequency tones can be lateralized if there is an IID in the stimuli.

![Fig. 1.8 The perceived location of the pure tone sound image as a function of the interaural intensity difference. The perceived location is normalized across the studies with 0% corresponding to $\theta = 0^\circ$ and 100% to $\theta = 90^\circ$. (After Yost and Hafter, 1987.)](image)
Two major classes of IID sensitive neurons have been found. Neurons in the first class are excited by ipsilateral simulation and inhibited by contralateral stimulation. These neurons are referred to as "EI" cells to reflect the separate excitatory-inhibitory influences from the two ears. Fig. 1.10 (After Kuwada and Yin, 1987) shows the IID response characteristics of such a neuron found in the nucleus of the lateral lemniscus (Brugge et al., 1970). In the experiment where the data shown in Fig. 1.10 were obtained, the stimulus intensity at the contralateral ear was held constant while the intensity at the ipsilateral ear was varied. The vertical axis in Fig. 1.10 shows the percentage of the number of spikes with respect to the number of spikes generated with contralateral stimulation alone. The stimulus used was a 6.4 kHz tone burst with a 200 ms duration.

Fig. 1.9 The IID threshold for perceived change as a function of frequency for tones. The parameters (0, 9, and 15 dB) are the IIDσs of the referent tones which serve to mark positions in lateral space. (After Yost and Hafter, 1987.)
The second class of IID sensitive neurons receive excitatory inputs from both ears. Thus, they are referred to as "EE" cells to reflecting the excitatory influence from both ears. The neurons do not respond to monaural stimulation to either ear, but they do respond to binaural stimulation (Kitzes et al., 1980; Wise and Irvine, 1983; Yin et al., 1985). This phenomenon was referred to by Yin and Kuwada (1984) as binaural facilitation (BF). The relationship between the spike count and the IID for the neurons in the second class, different from that shown in Fig. 1.10, exhibits non-monotonic characteristics. Fig. 1.11 (After Kuwada and Yin, 1987) shows such a relationship for a neuron in the superior colliculus of the cat (Yin et al., 1985). The mean intensity of the stimuli at the two ears is 60 dB SPL. The stimulus frequency is 19 kHz.

Fig. 1.10 Effects of varying IIDs on the discharge rate of a neuron in the cat dorsal nucleus of the lateral lemniscus. (After Kuwada and Yin, 1987.)
§1.6 Interaural Time Differences (ITDs)

The effects of the ITD on lateralization have been studied extensively over the past several decades. Different types of stimuli have been used in these studies. For sinusoidal stimuli, the perceived lateral position of the sound image is a periodic function of the ITD as the ITD is varied (Durlach and Colburn, 1978). The period of this function is equal to the period T of the sinusoidal stimuli. If the frequency of the stimuli is substantially above 1500 Hz, the perceived lateral position of the sound image no longer follows the change of the ITD in the stimuli. For low frequency tones, as the ITD is increased from 0 to half the period of the tone, T/2, the perceived sound image moves from the center of the head towards the lead ear. When the ITD is in the vicinity of T/2, the sound image is no longer perceived as compact as it is when ITD is 0, and it is difficult to tell where the image is. As the ITD is further increased from T/2 to T, the image can again be perceived clearly, and the position of

Fig. 1.11. Responses of a superior colliculus neuron of cat to changes in IID. (After Kuwada and Yin, 1987.)
the image returns to the center of the head from the lag ear. The maximum lateralization is reached at ITDs in the vicinity of T/2. Thus, it depends on the frequency of the sinusoidal stimulus. When T/2 is larger than 700 μs, the maximum lateralization corresponds to full lateralization to the side.

The periodic function between perceived position of the sound image and the ITD is best illustrated via the binaural coherence curve (Durlach and Colburn, 1978). Fig. 1.12 shows some examples of such curves which are derived from an experiment by Sayers and Cherry in 1975. In this experiment, the listener was asked to judge whether the image was to the left or right of the center of the head. The percentage of "left" response with respect to the total number of stimulus presentation trials is plotted as a function of the ITD values tested. The positive ITDs in the plots correspond to the left ear leading. As we can see from the binaural coherence curves.
in Fig. 1.12, lateral perception of the sound image is a periodic function of the ITD, and the period of this function varies with the frequency of the sinusoidal stimuli.

The effects of the ITD on lateralization have also been studied using a method where the listener is asked to assign a position to the sound image on a linear scale (Sayers, 1964; Watson and Mittler, 1965; Yost, 1981). Fig. 1.13 from Yost and Hafter (1987) shows the results obtained by Yost, 1981. This plot illustrates an approximately linear relationship between the perceived lateral position and the ITD. This relationship remains similar for tones with frequencies up to 1200 Hz. In the plot in Fig. 1.13 the ITD is measured in terms of interaural phase difference in the range from 0 to π. This is because the relationship depicted in Fig. 1.13 is periodical with a period equal to the period of the stimuli, as demonstrated in Fig. 1.12.

![Graph](image-url)

**Fig. 1.13** The perceived location of the pure tone sound image as a function of the interaural time difference (in terms of degrees of interaural phase difference). The perceived location is normalized with 0% corresponding to θ = 0° and 100% to θ = ±90°. (After Yost and Hafter, 1987.)
The effects of the ITD on the lateralization of the sound image have also been studied with broad-band noise stimuli (e.g. Blodgett et al., 1956). Similar results to that for pure tones have been obtained except that the perception of the “side” of the sound image does not vary periodically with the ITD. Fig. 1.14 shows that for a broad band click stimulus similar relationship to that shown in Fig. 1.13 exists between the perceived location of the sound image and the ITD was observed (Blauert, 1983). Again, unlike the case for sinusoidal stimuli, this relationship is not periodical. The perception of a single image breaks down when the ITD is substantially larger than 1 ms (Durlach and Colburn, 1978).

![Fig. 1.14](image)

**Fig. 1.14** The perceived locations (calibrated as follows: 0 corresponds to \( \theta = 0^\circ \), 5 to \( \theta = 90^\circ \), and -5 to \( \theta = -90^\circ \)) of impulse sound images as a function of the interaural time difference. (After Blauert, 1983, Page 144.)

The dependence of the lateralization on the ITD for more complex stimuli have also been studied. Sayers (1964), Sayers and Cherry (1957), and Toole and Sayers (1965a) studied multiple-tone complexes. Sayers and Toole (1964) and Toole and Sayers (1965a,b) studied click trains. Cherry (1961), Cherry and Sayers (1956),
Cherry and Taylor (1954) and Sayers and Cherry (1957) used speech signals. For speech, the lateralization judgments and coherence curve are similar to that for broad band noise. For multiple-tone complexes and periodic click trains, multiple images may occur. Sophisticated listeners can often identify a variety of images and trajectories as the ITD is varied (Durlach and Colburn, 1978).

![Image of graph showing ITD threshold for perceived change as a function of frequency for tones.

Fig. 1.15 The ITD (in terms of degrees of interaural phase difference) threshold for perceived change as a function of frequency for tones. The parameters are the ITDs (again, in terms of degrees of interaural phase difference) of the referent tones which serve to mark positions in lateral space. (After Yost and Hafter, 1987.)

The sensitivity of the auditory system to small changes in the ITD for tones have been studied by Klumpp and Eady (1956), Zwislocki and Feldman (1956), Hershkowitz and Durlach (1969a), Domnitz (1973), Yost (1974), and Mills (1960). Fig. 1.15 shows data from a number of studies summarized by Yost and Hafter (1987). The data show that the binaural system is highly sensitive to changes in the ITD. The parameters in Fig. 1.15 are the ITDs around which the sensitivity of the auditory system to changes in ITD were measured. Thus, different parameters in Fig.
1.15 indicate different sensitivities of the auditory system when the sound image is perceived at different lateral positions. It is evident in Fig. 1.15 that the auditory system is more sensitive when the sound image is perceived to be in the center of the head than when the sound images is perceived to the side. Another observation is that the ITD threshold is approximately 2 degrees (in terms of interaural phase difference) at low frequencies, and the threshold increases as the frequency is increased. Above 1200 Hz, the binaural system is insensitive to changes in the ITD (Yost and Hafter, 1987).

The sensitivity of neurons from different sites along the auditory pathways to the ITD had also been studied by a number of researchers (Goldberg and Brown, 1969; Crow et al., 1978; Brugge et al., 1969; Brugge et al., 1970; Aitkin and Webster, 1972; Chan and Yin, 1984; Yin and Chan, 1990; Melssen et al., 1990; Kuwada and Yin, 1983; and Yin and Kuwada, 1983). Fig. 1.16 shows data from experiments by Yin and Kuwada (1983), demonstrating responses of two low-frequency neurons in the inferior colliculus of the cat to different ITD values in pure tone stimuli (after Kuwada and Yin, 1983). In their experiments, Kuwada and Yin (1983) used sinusoidal stimuli. Three types of delays were used which were depicted in Fig. 1.16 (C, D, and E). The stimuli shown in Fig. 1.16C are labeled as "phase compensated", reflecting the fact that, although there is an on-set delay between the two stimuli at the two ears, there is no phase difference in the two tonal stimuli. The stimuli shown in Fig. 1.16D are labeled as "delay curve", reflecting the fact that the entire waveform of the stimulus to one ear is delayed with respect to the stimulus to the other ear. The stimuli shown in Fig. 1.16E are labeled as "phase curve", reflecting the fact that there is only a phase difference between the two stimuli, and that there is no-set delay between the two stimuli. Fig. 1.16A shows the responses of a inferior neuron to the two types of stimuli shown in Fig. 1.16C and D respectively. The neuron shows selective response to different delays in the stimuli shown in Fig. 1.16D, but not to
the different on-set delays in the stimuli shown in Fig. 1.16C. Fig. 1.16B shows responses of another inferior colliculus neuron to the two types of stimuli shown in Fig. 1.16D and E respectively. The neuron shows similar sensitivities to the phase differences in the two types of stimuli shown in Fig. 1.16D and E. The data in Fig. 1.16A and B indicate that, for tonal stimuli, the neurons are sensitive to the phase differences only.

Fig. 1.16 Responses (A and B) of two inferior colliculus neurons to changes in ITD, and waveforms (C, D, and E) of three types of binaural stimuli used in the stimulation of these two neurons. (After Kuwada and Yin, 1983.)

As consistent with the psychophysical observations for low-frequency sinusoids, neural sensitivity to the ITD decreases beyond certain frequencies. For the cat, this sensitivity is restricted to frequencies below 3 kHz, which is about the upper-limit for phase-locking (i.e. firing to a particular phase in the periods of the sinusoidal
waveform) of the cat's auditory nerve fibers (Kuwada and Yin, 1987). This observation suggests that the neural mechanism for ITD sensitivity is closely related to the phase-locking of the auditory nerve fibers.

§1.7 Free Field Auditory Localization

While dichotic stimulation is useful in the study of the effects of individual interaural parameters on the sensitivity of the auditory system, free field stimulation is important to study the overall performance of the auditory system in localization. In 1936, Stevens and Newman measured a listener's ability to localize pure tone stimuli in free field environments. A major finding of their study is that a listener's ability to localize pure tones in the frequency range below 1 kHz or above 4 kHz is better than the listener's ability to localize pure tones in the range between 1 kHz and 4 kHz. They concluded that in the mid-frequency region, the listener could not effectively use either ITD or the IID cue because neither was particularly salient in this region. Another influential free field experiment was done by Mills in 1958, when he measured the smallest change that a listener could reliably report in the azimuth angle of the sound source. This smallest change in the azimuth angle is referred to as the minimum audible angle (MAA) in azimuth (Mills, 1958).

Fig. 1.17 shows the results of Mill's experiment (1958), where the MAA is plotted as a function of the frequency of the tone at different azimuth positions. These data show that the MAA in azimuth can be as small as 1 degree. The different curves in Fig. 1.17 correspond to different angles of the source with respect to the media plane at which the MAA is measured. As we can see from Fig. 1.17, the spatial resolution of the auditory system is frequency dependent. In the middle frequency range of 1-4 kHz, the auditory system shows poorer spatial sensitivity, which is consistent with the observation by Stevens and Newman (1936) cited above. Also evident from Fig. 1.17 is that as the source moves away from the median plane,
the MAA increases, and the highest spatial resolution is observed when the sound is in front of the listener.

![Fig. 1.17](image)

Fig. 1.17 Angle of just noticeable difference (JND) ($\Delta \theta$) for tone bursts as a function of the tone frequency $f$ and the angle $\theta$ between the sound source and the median plane. (After Durlach and Colburn, 1978.)

A comparison (Fig. 1.18) made by Mills (1960, 1972) of the sensitivities of the auditory system to the ITD, IID, and spatial location shows that the curve for the just noticeable difference in ITD as a function of the stimulus frequency (PHASE JND in Fig. 1.18) and that for the actual phase change occurs when the sound source is moved a just noticeable angle coincide in the frequency region below 1500 Hz. Similarly, the curve for the just noticeable difference in the IID (AMPLITUDE JND in Fig. 1.18) and that for the actual amplitude change occurs when the sound source is moved a just noticeable angle coincide with each other in the frequency region between 1500 and 6000 Hz. These findings suggest that the localization of tones is
determined by the ITD sensitivity of the auditory system in the frequency region below 1500 Hz, and by the IID sensitivity in the region of 1500-6000 Hz.

Fig. 1.18 Comparison made by Mills (1960, 1972) of interaural phase JND and the interaural amplitude JND for ITD=0 and IID=0 with the changes in ITD Δϕ and in IID Δα that occur when an actual source is moved a just noticeable angle from the median plane. (After Durlach and Colburn, 1978.)

Free field stimulation is also important in the study of the localization cues other than the ITD and IID, especially in the study on how localization in elevation is made possible. Many studies investigated the fundamental significance of the filtering effects of the outer ear (including the head, shoulder, pinna, and ear canal) which imposes systematic direction-dependent changes on the spectrum of the incoming sound (Blauert, 1983; Butler et al., 1990; Middlebrooks, 1992; Wenzel et al., 1993; Zakarauskas and Cynader, 1993; Musicant and Butler, 1982, 1984a, 1984b; Musicant et al., 1990; Neti et al., 1992; Wightman et al., 1987). A notable effect of the filtering
characteristics of the outer ear is that notches and peaks are introduced to an originally flat sound spectrum. Such spectral notches and peaks prove to be perceptually important (Shu et al., 1993), and may be used by the auditory system to localize sound sources in both azimuth and elevation (Neti et al., 1992).

Free field stimuli have also been used in physiological studies of auditory localization. A notable finding is the evidence for the existence of a neural map of spatial location in the central auditory system of the barn owl (Knudsen and Konishi, 1978; Moiseff and Konishi, 1981). Maps of auditory space have been found in the deep and intermediate layers of the superior colliculus of the guinea pig (Palmer and King, 1982; King and Palmer, 1983) and in the auditory system of the cat (Middlebrooks and Knudsen, 1984). Other free-field studies of neural sensitivities include that by Rajan et al. (1990a,b) who studied the azimuthal sensitivity of the neurons in the primary auditory cortex of the cat.

§1.8 Interaural Envelope Delay

As reviewed in Section 1.6, the auditory system is insensitive to the ITDs in high-frequency sinusoidal stimuli. Also, as shown in Section 1.7, auditory localization of tones is determined by the IID sensitivity of the auditory system in the frequency region of 1500-6000 Hz. Moreover, these observations have been interpreted to be evidence in supporting the duplex theory of Rayleigh (1907). However, experiments with more complex stimuli (such as the amplitude-modulated high-frequency tones) reveal that the interaural time difference can be important for the localization of high-frequency stimuli under certain conditions. Specifically, the lateralization of high-frequency signals is influenced by the interaural time delay if the high-frequency signals have low-frequency envelopes.

Envelope delay sensitivity has been demonstrated for high-frequency noise and clicks, as well as amplitude-modulated high-frequency tones by a number of
researchers (Bekesy, 1960; David et al., 1958, 1959; Harris, 1960; Henning, 1974a; Klumpp and Eady, 1956; Leakey et al., 1958; Yost et al., 1971). Experiments by Leakey et al. (1958) and David et al. (1959) demonstrated that the lateralization of such stimuli depended principally on the interaural relationship of the envelopes, not the microstructures of the stimuli. This observation was further confirmed by some more recent studies by Henning (1980), Henning and Ashton (1981), McFadden and Pasanen (1976), and Nuetzel and Hafter (1976, 1981). Furthermore, according to Henning (1974a,b), the sensitivity to ITD for a sinusoidally amplitude-modulated high-frequency tone can be as high as that to ITD for a pure tone with a frequency equal to the modulation frequency of the high-frequency tone. Results from experiments by Henning (1974a,b), Licklider and Webster (1950), and Yost et al. (1971) strongly suggest that the timing information carried solely by the auditory nerve fibers with high characteristic frequencies can be used to discriminate the ITD in high-frequency signals. Indeed, experiments by Yin et al. (1984) who used amplitude-modulated tones provided physiological evidence for the above cited psychophysical observations. They found that some neurons with high characteristic frequencies in the inferior colliculus were sensitive to the interaural envelope delay. Moller (1974) found that high-frequency cochlear nucleus neurons could phase-lock to the envelopes of amplitude-modulated signals with high-frequency carriers. Kuwada and Yin (1987) have suggested that high-frequency neurons can be sensitive to interaural envelope delay in a manner similar to low-frequency neurons being sensitive to the ITD in the microstructures of low frequency stimuli.

§1.9 Models of Auditory Localization

As reviewed in the previous sections, a large amount of experimental work has been devoted to the study of auditory localization for more than a century. In an effort to
explain the experimental data in terms of models or theories, modeling studies of auditory localization have paralleled the experimental work for many decades.

Some early models based their ideas on a model proposed by Bekesy in 1930 (Bekesy, 1960; Colburn and Durlach, 1978). In his model, Bekesy assumed a population of neurons that were innervated by fibers from both ears. Neurons in the population can become “tuned” to one of two excited states according to the source of the excitation. Specifically, a neuron is tuned left if its excitation comes from fibers originating from the left ear, or is tuned right if the excitation comes from fibers originating from the right ear. The lateralization of a stimuli is then determined by a comparison of the number of neurons tuned left with that of neurons tuned right.

In 1958, Matzker proposed a model which was essentially consistent with Bekesay’s ideas, but was more plausible with respect to anatomical and physiological observations. The population of Bekesy’s tunable cells is replaced with two symmetric auditory pathways, each of which is excited by stimulation of one ear. The lateralization of a stimuli is determined by a comparison between activity levels at some relatively central nuclei which receive information from the two pathways. Matzker (1958) also assumed that there were contralateral inhibitory pathways which block the excitatory pathways for a few milliseconds. No detailed assumption about the structure of the pathways was given in the model. Thus, the model describes only qualitatively how sound localization may be accomplished by the auditory system.

Bekesy’s ideas were further elaborated by van Bergeijk in 1962. In his model, van Bergeijk assumed that the binaural interaction occurs at a pair of relatively peripheral nuclei. Lateralization of a stimuli is assumed to be determined by a comparison of the number of neural firings in the left nucleus with that in the right nucleus. In this model, the ipsilateral inputs to each nucleus are assumed to be inhibitory and the contralateral inputs excitatory. The image of the sound stimuli is
assumed to be lateralized to the side opposite to the nucleus with the greater number of firings (Colburn and Durlach, 1978).

Fig. 1.19 The coincidence network proposed by Jeffress in 1948 for the localization of low-frequency tones. (After Colburn and Durlach, 1978.)

Another group of models (e.g. Licklider, 1959; Sayers and Cherry, 1957; Colburn, 1973, 1977) have their root in the ideas suggested by Jeffress in 1948. In his model, Jeffress described a neural network for the estimation of interaural time differences. This network is often referred to as the coincidence network, and this model as the coincidence model. Fig. 1.19 shows a schematic diagram of the coincidence network. Binaural neurons depicted in the diagram receive delayed neural signals via auditory fibers from both ears. The response of a binaural neuron is assumed to be maximum when the two excitations from the two ears coincide in time. Thus, a particular binaural neuron is tuned to respond maximally to a particular ITD depending on the delays the nerve fibers introduce before the excitation signals reach the binaural neuron. Thus, by systematically varying the delays in the nerve fibers leading to
different binaural neurons, a neural map of different ITD values can be implemented in the network.

§1.10 Cross-correlation Based Models

Jeffress’ model (1948) discussed in the last section is mathematically the same as cross-correlating the two signals detected by the two ears (Colburn and Durlach, 1978). Many other models may also be viewed to be variations of a cross-correlation mechanism (Yost and Hafter, 1987; Colburn and Durlach, 1978). The basic idea of these models is that the cross-correlation function, defined as

\[ R(t, \tau) = \int_{-\infty}^{t} x_R(t) x_L(t - \tau) dt \] (1.1)

has a maximum point at \( \tau = \tau^* \), if \( x_R(t) \) and \( x_L(t) \) are identical except for a delay \( \tau^* \), where \( x_R(t) \) and \( x_L(t) \) are input signals to the right and left ears, respectively.

The first quantitative model based on the idea of the cross-correlation mechanism was proposed by Sayers and Cherry in 1957 to model the binaural phenomena of fusion and lateralization. Fig. 1.20 from Colburn and Durlach (1978) shows a block diagram of this model. In this model, the interaural cross-correlation function in Eq. 1.1 is modified by two weighting factors. One factor limits the influence of larger interaural delays and the other reflects the influence of interaural intensity differences. The output of the model is a judgment of whether the sound image is lateralized to the left or to the right. This arrangement is convenient because the model was used to fit to the data from some psychophysical experiments where listeners were asked to make such judgments.

A multiple-channel version of this model was given by Sayers (1964) and Toole and Sayers (1965b). In this later version, the spectra of the input signals are first obtained via two banks of band-pass filters. Pairs of outputs from two band-pass
filters (one from each ear) with the same center frequency are used as the inputs to the cross-correlation device depicted in Fig. 1.20.

The idea of interaural cross-correlation has also been used in models for data from binaural signal detection experiments (Jeffress et al. 1952; Robinson and Jeffress, 1963; Dolan and Robinson, 1967; McFadden, 1968; Levitt and Lundry, 1966b; Osman, 1971). In such models, a decision variable is usually devised with the aim that the decision made by the model match that made by the listeners in the binaural detection experiments.

![Fig. 1.20](image-url) A block diagram of the cross-correlation model of Sayers and Cherry (1957). The term $e^{-\omega t}$ in the figure corresponds to the first weighting factor, and the term $A_L$ or $A_R$ to the second weighting factor discussed in the main text. The term $Av(t)$ refers to temporal average. (After Colburn and Durlach, 1978.)

In a series of papers, Colburn and his coworker (Colburn, 1973, 1977; Colburn and Latimer, 1978) developed a model for binaural interaction based on explicit quantitative description of physiological observations of auditory nerve fiber firing patterns. The description of the auditory nerve activity in the model is based on the
physiological observations of the auditory nerve of the cat made by Kiang and his associates (Kiang 1968; Kiang et al., 1965), and on the mathematical modeling work of Siebert (1965, 1968, 1970). The activity of the fibers is described in terms of statistically independent, non-homogeneous Poisson random processes. For a 500-Hz tonal stimulus described as

\[ s(t) = A \cos[2\pi 500(t - \tau)] \]  

the rate function \( r_m(t) \), which was used to describe the instantaneous probability of firing of the synchronized fiber, was assumed to be of the form

\[ r_m(t) = A \exp(\sqrt{20} \cos[2\pi 500(t - \tau)]) \]  

Fig. 1.21 shows a general block diagram used by Colburn and Durlach (1978) to present the basic ideas of Colburn’s model (Colburn, 1973, 1977). In the model, it is assumed that an ideal decision maker uses an overall decision variable that is a combination of three decision variables: a purely monaural variable for each ear and a binaural variable that carries the interaural timing information. Central to this model is a binaural displayer (Fig. 1.21b) which can be considered as a quantification and elaboration of Jeffress’ coincidence network (1948). The outputs of the coincidence counter in Fig. 1.21b, \( L_m(f_m, \tau_m) \), \( m = 1, \cdots, M \), are approximated by the following integral

\[ E[L_m(f_m, \tau_m)] = T_w \int_0^{T_s} r_{m_L}(t - \tau_m) r_{m_R}(t) \, dt \]  

for \( m = 1, \cdots, M \), where \( T_w \) is the time window for coincidence of each fiber pair, \( T_s \) is the duration of the stimulus, \( r_{m_L}(t) \) and \( r_{m_R}(t) \) are the rate functions describing the instantaneous firing probability for the corresponding left and right fibers, respectively. The integral in Eq. (1.4) is an estimate of the cross-correlation function of the instantaneous firing rates of the corresponding fiber pairs.

Many of the models that use the idea of interaural cross-correlation are essentially equivalent in terms of their operations on the acoustic input waveforms (Colburn and Durlach, 1978). They differ mainly in the detailed implementation of the interaural
correlation and in the detailed construction of the decision variables used for the prediction of psychophysical observations. These models are mainly designed to fit data from a large number of psychophysical experiments which measure the sensitivities of the auditory system to such parameters of the binaural stimuli as the ITD, IID, and IED. The models do not offer new ideas as to how the ITD cue is processed other than the coincidence mechanism which was first proposed in 1948 by Jeffress.

Jeffress' coincidence network makes use of neural delay lines that have systematic delay values. Shamma et al. (1989) proposed an alternative network, which was referred to as the sterausis binaural network, for ITD estimation without the use of neural delay lines. They argued that it was not necessary to explicitly make
use of neural delay lines in order to estimate the ITD. Specifically, they noted that the traveling wave of the motion of the basilar membrane provided timing information that was sufficient for the estimation of the ITD. The phase of the waveform describing the motion of the basilar membrane at a particular point on the basilar membrane is delayed with respect to the phase of the waveform at another point on the basilar membrane that is nearer to the apex of the basilar membrane. Thus, the correlation of the signals originating from different points on the basilar membrane can be used as the mechanism for ITD estimation. Fig. 1.22 shows a schematic diagram of this model. A matrix of binaural neurons are assumed to receive excitatory inputs from both ears through the auditory nerve fibers. Each neuron in the matrix receives neural signals from two, and only two, fibers (each from one ear), generating a measure of the correlation between the instantaneous activities of the two fibers. Due to the phase difference of the neural signals originating from fibers with different characteristic frequencies, a binaural neuron in the matrix shown in Fig. 1.22 responds maximally when the phase difference between the neural signals at the input of this neuron match the interaural time delay between the stimuli at the two ear. The idea of cross-correlating signals from two auditory nerve fibers with different characteristic frequencies in order to estimate the ITD has also been suggested earlier by Schroeder (1977) and Loeb et al. (1983).

In terms of operations or of processing of the neural signals in the auditory nerve fibers, the model of Shamma et al. (1989) is the same as that of Jeffress (1948). Binaural neurons in both models generate a measure of the cross-correlation of the two input signals. The uniqueness, however, of the sterausis model lies in the fact that the cross-correlation measure of the sterausis neuron also reflects the IID information in addition to the ITD information. In Jeffress model, a separate mechanism for the IID sensitivity has to be assumed. A weakness of the sterausis model, however, is the difficulty in evaluating the model quantitatively. It is difficult
to calibrate the ITD and the IID sensitivities of the network from the activity pattern of the binaural neuron matrix shown in Fig. 1.22.

In 1994 Bonham further extended the ideas of Shamma et al. (1989) in his Ph.D. thesis by allowing multiple fibers with different characteristic frequencies from both basilar membranes to innervate a single binaural neuron. Bonham (1994) showed that a weighted sum of the neural signals from multiple frequency bands from one ear could give rise to a waveform that was similar to the neural signals from single
auditory nerve fibers but had a different phase delay. Different sets of weights, which can be obtained using a Hebbian type learning rule (the connection weight between two neurons is increased if they fire together), result in different phase delays in the combined signals. Such combined signals from both ears can thus be correlated to generate ITD estimates. This model still obtains the ITD sensitivity by means of cross-correlating two temporal waveforms. The only difference of this model from the sterausis model of Shamma et al. (1989) and the coincidence model of Jeffress (1948) lies in the origin of the temporal signals to be cross-correlated: in Jeffress' model the signals are delayed versions of single auditory nerve fibers; in the sterausis model the signals come directly from the auditory nerve fibers without any delay; and in Bonham's model the signals are weighted sums of neural signals from different auditory nerve fibers.

When applying the cross-correlation based models to complex stimuli, a common approach is to apply the cross-correlation mechanism in different frequency bands, and the results of individual bands are combined across the relevant frequency range (Sayers 1964; Toole and Sayers, 1965b). In the more recent cross-correlation based models cited above (such as those of Colburn, 1973, 1977, Shamma et al., 1989, and Bonham, 1994), this multiple frequency channel scheme is inherent in the sense that the cross-correlation is carried out explicitly over neural signals of auditory nerve fibers. In these models, however, the emphasis is given to how the ITD and IID can be measured for individual frequency channels. In 1988, Stern at al. proposed a model that addresses explicitly how the individual frequency channels could be combined and weighted in order to model lateralization data for band-pass stimuli. They proposed a weighting strategy to predict the lateralization of low-frequency band-pass stimuli based on the outputs of multiple cross-correlation channels, which we refer to as the multi-channel cross-correlation function. Fig. 1.23 shows an example of the multi-channel cross-correlation function. The solid lines in the Fig.
1.23 indicate the locations of the peaks of the cross-correlation function across frequency space. The predicted lateralization of the stimuli in terms of the interaural delay is a weighted average of the τ (ITD) coordinates of the solid lines. The weighting is done in such a way that if a solid line is more straight and closer to \( \tau = 0 \), the corresponding \( \tau \) values are given greater weights.

![Diagram of cross-correlation function](image)

Fig. 1.23 Location of the peaks of the multi-channel cross-correlation function for broad-band noise with an ITD of 1500 \( \mu s \). The vertical axis indicates the center frequency of the different band-pass filter channels, while the horizontal axis indicates the argument (ITD) of the cross-correlation function. (After Stern et al., 1988.)

The idea of weighting of the multi-channel cross-correlation function has also been used by Shackleton et al. (1992). The weighting is carried out on the entire cross-correlation function with an implicit assumption that there is only one sound source in the acoustic environment. Thus, the model cannot deal with multi-source environments.

The idea of cross-correlation function weighting has been taken a step further by Arad et al. (1994) to address the problem of multiple sound source localization.
Unlike the model by Stern et al. (1988), the weighting is done over selected frequency channels, and short-time cross-correlation functions are used. An overall short-time cross-correlation function is obtained by integrating, across frequency, the short-time cross-correlation functions of the selected channels. A series of such overall short-time correlation function are obtained over a certain listening period. This series of correlation functions are then either averaged over time, or used to form a two-dimensional image where one dimension corresponds to the ITD and the other dimension is time. The values of the short-time correlation function are translated into gray-scale values in the image. Traces of peaks in the image along the temporal dimension indicate the presence of the sound sources. The ITD values corresponding such traces of peaks in the image are the ITD estimates of the corresponding sources. Edge detection algorithms were used to locate such traces of peaks in the image.

In this model, multiple sources with fluctuating intensities are localized by means of estimating short-time cross-correlation functions. Aral et al. (1994) tested their model in both the one- and two-speaker cases. When there are multiple sound sources in the acoustic environment, short-time cross-correlation functions give rise to noisy and faulty peaks. Aral et al. (1994) reported that, in order to obtain reliable estimates of the cross-correlation functions, several (4-12) short-time (3-12 ms) correlation function estimates have to be averaged. Furthermore, in the two-speaker case, biased ITD estimates were obtained.

The idea of short-time cross-correlation is also used by Lyon (1983) in a model for binaural localization and source separation. In this model, short-time ITD estimates are first obtained via short-time cross-correlation. Frequency- and time-dependent weights are then chosen to weight the outputs of the left and right cochlear models according to the short-time ITD estimates. The final output of Lyon's model (1983) is a spectro-temporal representation of the sound stimuli based on the weighted outputs of the cochlear models. Such spectro-temporal representations are
used to visualize sound stimuli. The model can display separate images for different sound sources, but the separation is limited to left from right or vice versa.

Another direction for the extension of the basic cross-correlation model (Sayers and Cherry, 1957) is the introduction of a lateral inhibition mechanism and monaural channels into the model. Lindemann (1986a, b) proposed a model in which correlation at one delay value inhibits correlation at other delay values. This produces the effect of sharpened cross-correlation functions. The model also provides a mechanism for the precedence effect. (The precedent effect is observed when two binaural sounds are presented with a brief time interval between the two sounds. The two sounds may be perceived as a single auditory event, and when this happens, the perceived lateralization of the single auditory event is determined by the directional cues carried by the first sound.) The centroid of the laterally inhibited cross-correlation function is taken to be the predicted lateral position of the sound image. An interesting aspect of Lindemann’s model is that the lateral inhibition mechanism allows the IID information to be processed as well as the ITD information because the level of inhibition is assumed to be intensity dependent.

An interesting extension to Lindemann’s (1986a, b) model was made by Gaik (1993) who imposed attenuation factors on the inhibited cross-correlation function defined in Lindemann’s model. The attenuation factors are used to insure that, for natural ITD-IID combinations, the intensities of the signals being correlated are equal when maximum correlation is reached. Natural ITD-IID combinations are the corresponding ITD and IID value pairs produced by free-field impulsive sound sources. Thus, the modified (with both attenuation and lateral inhibition) cross-correlation function has a single peak for natural ITD-IID combinations. Multiple peaks may occur for unnatural combinations, signaling the presence of multiple sound sources that can be identified by separate ITD and IID cues.
§1.11 Models of cue sensitive neurons

Modeling studies have been closely coupled with experimental studies of neural sensitivities of the auditory system (Goldberg and Brown, 1969; Yin and Chan, 1990). The joint effort of both theoretical modeling and experimental recording of neural activities in responding to binaural localization cues has been aimed at illuminating the neural mechanism for auditory localization.

As reviewed in Section 1.3, the superior olivary complex is the first site in the auditory system that exhibits binaural interaction. In fact, it has been identified to be the primary site of binaural processing (Kuwada and Yin, 1987; Goldberg and Brown, 1969; Yin and Chan, 1990), and most models of cue sensitive neurons are built with neurons in this complex in mind. There are two nuclei in the superior olivary complex that are thought to be involved in sound localization. In the lateral superior olive, high frequency (>3 kHz) neurons are sensitive to IIDs (Boudreau and Tsuchitani, 1968; Caird and Klucke, 1983). In the medial superior olive, on the other hand, the majority of the neurons have low characteristic frequencies (<3 kHz), and these neurons are sensitive to ITDs (Goldberg and Brown, 1969; Yin and Chan, 1990; Guinan et al., 1972).

The hypothetical coincidence detector neurons in Jeffress' ITD network (1948) have been taken as the basic model for the ITD sensitive neurons in the MSO. Goldberg and Brown (1969) and Yin and Chan (1990) showed that some neurons in the MSO exhibited properties predicted by the coincidence detector in Jeffress’ model (1948). Most other models are elaborations and refinements of the Jeffress model.

In 1990, Colburn et al. presented such a refined coincidence model of observed MSO responses. This model is based on an internal variable that is identified as the model membrane potential. There are two inputs to the model, one from each ear. The inputs are modeled as filtered sequences of pulses distributed as Poisson processes. The membrane potential is increased a certain amount whenever there is
an input pulse from either input. The modeled neuron outputs a firing pulse whenever
the membrane potential reaches a threshold. The behavior of the model was
compared with physiological data obtained by Goldberg and Brown (1969) and Yin
and Chan (1990). Good agreement between the data and the model was reported by
Colburn et al. (1990).

The model of Colburn et al. (1990) was further refined by Han and Colburn
(1993), who replaced the previous more functional model of the membrane potential
with a model neuron that has four conductance channels. One of the four channels
responds to excitatory inputs, another responds to inhibitory inputs, a third represents
the delayed potassium channel that opens in response to output action potentials, and
the last channel represents the constant, residual conductance of the cell membrane
(MacGregor, 1987). The model showed similar behavior to the previous model of
Colburn et al. (1990).

The response properties of MSO neurons have also been modeled using the
maximum likelihood (ML) estimation technique. By modeling the inputs of the MSO
neurons as Poisson processes, Dabak and Johnson (1992) investigated what a ML
estimator would do to obtain an ITD estimate, and then compared the input-output
characteristics of the ML estimator with that of the MSO neurons. They found that an
implementation of the ML estimator has a similar structure to that of the Jeffress’
coincidence detector model (1948), but the ML estimator differs from Jeffress’ model
in that the ML estimator processes its inputs in a much complicated fashion. Detailed
implementation of this more complicated processing is not given in their paper.
Although the models just described above do not contain any inhibitory input, evidence of inhibitory inputs to the MSO neurons has been found by Grothe and Sanes (1993), Adams and Mugnaini (1990), Schwartz (1992), and Cant and Hyson (1992). Few studies, however, have explored the roles of inhibition. Inhibition was seen in a model by Sujaku et al. (1981), which describes the ITD sensitivity of neurons in the inferior colliculus. The structure of the model is shown in Fig. 1.24. There are two inputs to the model, one from each ear, which make excitatory synaptic connections onto the modeled binaural cell. In this respect, the model is similar to a model of the type described above, i.e. a coincidence detector model. The uniqueness of the model, however, is the addition of the collateral presynaptic inhibition from each input. In other words, the input from one side decrease the effect of the input from the other side on the model neuron. D\textsubscript{1} through D\textsubscript{4} in Fig 1.24 are four independent delays introduced in the four branches of the inputs. The model was shown to be able to simulate the observed neural responses in the inferior colliculus of the cat. The model was also shown to be able to simulate neurons sensitive to the direction of ITD changes. Another model of ITD sensitive IC neurons that
incorporated both excitatory and inhibitory inputs is proposed by Colburn and Ibrahim (1993).

As mentioned in the beginning of this section, high-frequency neurons in the LSO are sensitive to the IIDs (Boudreau and Tsuchitani, 1968; Caird and Klucke, 1983). These neurons are excited by the ipsilateral input and inhibited by the contralateral input. The thresholds and tuning of the ipsilateral excitatory effect and that of the contralateral inhibitory effect are often comparable. This excitatory-inhibitory (EI) characteristics provides a neural mechanism for the IID sensitivity (Caird and Klucke, 1983; Pickles, 1988). Guinan et al. (1972) suggested a model for the IID sensitive LSO neurons. This model was later studied by Colburn and Moss (1981). The model neuron receives two inputs, one from each ear, and the neuron has a membrane potential which is characterized as the internal variable of the model. The input from each ear is a pulse train modeled as a Poisson process. An input pulse from the ipsilateral side causes a depolarization pulse to be added to the membrane potential while an input pulse from the contralateral side causes a hyperpolarization pulse to be added to the membrane potential. The model neuron fires an output whenever the depolarization exceeds a threshold. Colburn and Moss (1981) were able to demonstrate that the overall pattern of the response of the modeled LSO neuron was similar to those measured in the LSO by Boudreau and Tsuchitani (1968). The model neuron showed sensitivity to both the IID and the overall intensity of the stimuli.

Johnson et al. (1990) developed a maximum likelihood (ML) estimator which was used to model the response behavior of the IID sensitive LSO neurons. The ML estimator takes as its inputs the input signals to the LSO neurons. Such input signals are modeled as Poisson processes. The intensity functions of these Poisson processes are formulated to be functions of three variables: one is the lateral angle $\theta$ of the perceived sound image of the stimuli, and the other two are the intensities of the sound stimuli at the two ears. The output of the ML estimator is an estimate of the
lateral angle $\theta$. Johnson et al. (1990) found that the ML estimate of $\theta$ was based on estimation of the IID. In other word, the IID sensitive neurons can be thought to be part of an implementation of the ML estimator.

While some researchers have been concerned, as in the models just reviewed above, about the response properties of the individual LSO neurons, some others have been concerned about the organization of the LSO neurons that have functional roles in sound localization. Reed and Blum (1990) have presented a structural model for the encoding of the azimuth angle by a hypothetical column of neurons in the LSO. The focus of the model is on the connections between the LSO neurons in the column and their inputs from the anteroventral cochlear nucleus (AVCN) and the medial nucleus of the trapezoid body (MNTB). Fig. 1.25 shows a schematic diagram of these connections. A basic assumption made in the model is that the thresholds of the AVCN neurons increase monotonically from the AVCN neurons connecting to one end of the LSO column to those connecting to the other end of the column. The thresholds of the MNTB neurons (which make inhibitory connections with the LSO column), however, decrease monotonically in the same direction along the LSO column. This arrangement results in the coding of the IIDs by the position of the neuron in the LSO column whose firing first goes to zero. The encoding of the IIDs in the model was demonstrated to be independent of the absolute sound level, and to vary linearly with the IID.

Yin et al. (1985) found that some neurons in the cat superior colliculus were sensitive to the IIDs but did not show the excitatory-inhibitory characteristics common to the IID sensitive LSO neurons. In fact, those superior colliculus neurons receive excitatory inputs from both ears, as does the ITD sensitive neurons observed in both the medial superior olive and the superior colliculus (Goldberg and Brown, 1969; Yin and Chan, 1986; Kuwada et al. 1984). Yin et al. (1985) tested the sensitivity of these IID sensitive neurons to the ITDs, and found that the neurons are
also sensitive to the ITDs. Moreover, the neurons' IID sensitivity functions are similar in shape to their corresponding ITD sensitivity functions. These observations led Yin et al. (1985) to suggest a temporal mechanism for the IID sensitivity of these neurons.

![Diagram of connections from AVCN and MNTB to the LSO column in the model by Reed and Blum (1990).]

Fig. 1.25 Schematic diagram of the connections from AVCN and MNTB to the LSO column in the model by Reed and Blum (1990).

Such a mechanism has been proposed by Jeffress in 1984 in his temporal coincidence model (see Fig. 1.19 on Page 31). Jeffress hypothesized that the neural mechanisms for the IID sensitivity also involve temporal coincidence since the neural response latency may change as a function of the stimulus intensity. This hypothesis has been referred to as the latency hypothesis (Kuwada and Yin, 1987). Electrophysiological recordings made by Yin et al. (1985) seem to have found evidence that supports such a model. In fact, a computer simulation carried out by Yin et al. (1985) showed that such a model could be used to describe the responses of some neurons in the superior colliculus. The sensitivity functions of the modeled
neuron to both the ITD and IID resemble in general shape to that observed in the experiment by Yin et al. (1985). The model, however, does not account for the observed changes in the discharge rate in responding to the changes in the overall stimulus intensity (Kuwada and Yin, 1987).
2 Modeling Auditory Localization in Complex Acoustic Environments

In Chapter 1, we reviewed both experimental and theoretical studies of auditory localization. Theses studies have shown that the auditory system is very sensitive to the spatial location as well as a number of parameters (ITD, IID, and IED) of a sound stimulus. Specifically, the smallest change in azimuth that the auditory system can detect for a pure tone source can be as small as 1/360 of the entire azimuthal range (Mills, 1958). Also, a listener can notice a 0.5 dB change in the IID under certain conditions (see Fig. 1.9 on Page 16). Perhaps the most impressive sensitivity of the auditory system is that to the ITDs. The owl can detect ITDs as short as 10 μs (Konishi 1993). The highest sensitivity to the ITD reported in human psychophysical experiments is also on the order of 10 μs (Yost and Hafter, 1987), which is very short considering the fact that a neural impulse persists for as long as 1000 μs.

As reviewed in Chapter 1, much effort has been directed to the study of the mechanisms for the sensitivities of the auditory system to localization cues. Many models have been proposed. A noticeable characteristics of these is that most models are designed to fit to the data observed in psychophysical and physiological experiments where the sensitivity measurements were made under well controlled conditions (Sayers and Cherry, 1957; Colburn, 1973, 1977; Stern et al., 1988; Gaik, 1993; Colburn et al., 1990; Han and Colburn, 1993; Sujaku et al., 1981; Colburn and Moss, 1981; Yin et al., 1985). This limits the validity of the models in more complex situations, which are typical of real acoustic environments.

Many quantitative models are centered around proposed mechanisms for the ITD sensitivity of the auditory system. Furthermore, many of the models are based on the coincidence detection mechanism first proposed by Jeffress in 1948. This mechanism has been implemented, in a number of models, in terms of the estimation of a variety
of interaural cross-correlation functions (Sayers and Cherry, 1957; Sayers, 1964; Colburn, 1973, 1977; Shamma et al., 1989; Stern et al., 1988; Lindemann, 1986a, b; Gaik, 1993; Arad et al., 1994). In order to adapt these models to deal with the situations common in real acoustic environments (which are noisy, dynamic and have multiple sound sources) short-time interaural cross-correlation functions have been used (Arad et al., 1994). However, as Arad et al. (1994) have shown, short-time cross-correlation function estimates are noisy and have faulty peaks when the time window is relatively small, and when there are two or more sources emitting sounds at the same time. Thus, previous models have difficulties when applied to sound localization tasks in complex situations which are more likely to occur in a real environment.

Natural acoustic environments are typically noisy, dynamic, and have multiple simultaneous sources. Nevertheless, the auditory system can localize sound sources in natural environments with good accuracy and reliability. Investigation into how auditory localization is achieved in such situations may provide an opportunity to discover new mechanisms or insights into auditory localization that may not be so evident under simplified and idealized assumptions about the acoustic environment. Another reason for the consideration of a more complex and realistic acoustic environment is that any practical localization device must work in such situations. Furthermore, previous models could not be adapted in a straightforward fashion for applications in the development of localization devices that would work effectively in real environments. Thus, the problem we wish to consider is how auditory localization is achieved in noisy, dynamic, and multi-source environments.

We argue that in order for the auditory system to solve the problem of sound localization in natural environments, the localization cues must be estimated robustly and quickly. Since the natural environment is usually noisy, the estimation of the localization cues must be resistant to the effect of noise. Another expectation is that
multiple sources will be present simultaneously in the majority of real-life situations. In such situations, the auditory system is demonstrated to be able to concentrate on one source to the exclusion of others (Durlach and Colburn, 1978). Robust short-time cue estimation may provide a possible mechanism for such an ability.

First, we note that, in natural environments, sound sources are often dynamic in the sense that there are many transient effects, and the energies of the sounds emitted by different sources change from time to time. Thus, it is essential for the auditory system to estimate the location cues robustly in the relatively short time periods or windows during which some sound source emit higher energy than other sources. In this way, it is possible for the auditory system to pick up multiple sound sources that emit sound energies over the whole observation period, but have changing relative intensities. To model such a process, we propose a “decomposition-localization-integration” (DLI) scheme. In this scheme, the mixture of signals detected by the two ears are first decomposed into their spectro-temporal distributions. Spatial attributes (in terms of the ITD, IID, and IED cues) are then determined robustly over small spectro-temporal windows from energy concentrations in the spectro-temporal distributions. Finally, a spatial scene of the sound sources in the environment is built by integrating the short-time energy concentrations according to their spatial attributes. This integration process is needed because a sound source may spread over the entire observation period while the spatial attributes of the energy concentrations are estimated over small temporal windows.

There is physiological evidence that supports the above modeling concept. We note that all the subsequent processes for binaural as well as monaural information processing are based on the neural signals from the auditory nerve fibers (see Figs. 1.1 and 1.2), which in turn “map” in some form the parallel analysis of the sound stimulus encoded in the motion of the basilar membrane (Allen, 1985). Such parallel analysis is in fact a spectro-temporal decomposition of the stimulus signal (Shamma
Furthermore, Takahashi et al. (1993) conducted experiments to examine the ability of neurons in the inferior colliculus of the barn owl to respond to multiple sound sources. Using two noise bursts which were time reverses of each other, they found that the owl was able to "tell" that there were two sound sources, even though the spectra of the two noise bursts were the same, and the two noise bursts were simultaneously presented to the owl. This observation suggests that the space map of the inferior colliculus relies on differences between noise bursts that exist over brief time spans, and that it builds a neural image of the acoustic environment from *multiple samples gathered over time*. It further suggests that the auditory system may estimate the location cues in short time windows, which corresponds to the "localization" process in our proposed DLI scheme. Thus, it is plausible that the auditory system detects sound sources in short time windows and accumulates short time estimates to get a complete image of the acoustic environment. The accumulation of short-time cue estimates corresponds to the "integration" process in the DLI scheme.

The question for us now is how can the auditory system estimate the location cues quickly and robustly? Most interestingly, how can the relatively slow nervous system measure the very small ITDs in a short-time window? Our approach to the problem is to look at what we know about the system, and see how we can find a process that will be able to evaluate and to measure the relatively small time differences. It is not obvious by just observing the neurons and the nuclei in the auditory pathways that the relatively slow system can actually measure very small time differences. Is there a process in the system, we wonder, that can indeed produce clear and robust time difference measures, or something that is on the order that the brain can deal with, yet it is an indicator that the small time difference exists? If we could find such a process, we can devise an algorithm that can be used in a machine, and give an idea
for the physiologists to explore, to see if indeed there are indications that the nervous system also does something similar.

If we take a closer look at the spectro-temporal decomposition of the stimulus signal as represented by the parallel neural signals coming from adjacent points on the basilar membrane, we can notice the two-dimensional pattern characteristics of the neural signals. This is best demonstrated in a waterfall diagram of the responses of a cochlear model (Shamma et al., 1989) to a 600 Hz tone, as shown in Fig. 2.1. In this diagram, the vertical axis represents the basal-to-apical (bottom-to-top in the diagram) line-ups of locations along the basilar membrane, and is labeled by the characteristic frequency corresponding to each location. The horizontal axis is time. In the auditory system, the spectro-temporal response patterns shown in Fig. 2.1 are directly related to the probability of firing of the auditory nerve fibers.

Fig. 2.1 Simulated responses of a cochlear model (Shamma et al., 1989) to a 600 Hz sinusoidal stimulus. (After Shamma et al., 1989.)
In Fig. 2.1, we can clearly see the structured pattern features (e.g. the curved line patterns of the synchronized maximum response points across frequency) in the two-dimensional spectro-temporal image representing the neural signals of the auditory nerve fibers. It is a promising avenue to explore whether these patterns hold some indicators that can be used to evaluate the very small time differences between the two ears. We argue that these patterns, readily available in the auditory system, indicate to us that, indeed, it is possible to use the relatively long time scale events, i.e. the patterns shown in Fig. 2.1, as references for the measurement of very small interaural time differences. In other words, patterns from the two ears can be compared, and pattern comparison yields measures of the very small time differences. We do not know whether the neural system actually does that, but it is possible because the patterns are readily available in the auditory system, and the neural system is good at comparing patterns. This is a question for the physiologists to decide.

The fundamental contribution of this idea is that there are unexplored patterns that are generated by the system which may embed codes for the small interaural time differences. These patterns may be generated for the very purpose of enabling the relatively slow neural elements to evaluate small time differences. This can be viewed as a coding process where the patterns are used to carry the message pertaining to time.

Thus, a key task for us is to model the pattern recognition and comparison process for the purpose of estimating short-time cues. The many techniques for solving pattern recognition and comparison problems are based on either statistical or neural network concepts (Fukunaga, 1990; Pao, 1989; Nigrin, 1993). To choose the most appropriate approach for use in our model, certain aspects of our modeling process should be considered. Firstly, our objective is to develop a model that is based on a solid foundation of knowledge about the way the auditory system achieves
localization. The auditory system, like any other part of the nervous system, involves parallel processing. Parallel processing is also important for practical applications as it has the potential to work in real time when implemented using current VLSI technology (Mead et al., 1991). Secondly, it is important to be able to relate the processes involved in the model to the processes of the auditory system. Thirdly, as we argued earlier in this chapter, estimates of different localization cues may be obtained from similar pattern recognition and comparison processes. Thus, flexible processing structures are more desirable than ad hoc algorithms specifically designed for the estimation of different cues. Fourthly, the ability to learn and to continue learning after initial training is important if the model is to be used for practical applications. Statistical techniques for pattern recognition are mostly not adaptive, but typically process all training data simultaneously before being used with new data (Lippmann, 1987). Moreover, as the acoustic environments under consideration are non-stationary, noisy and have multiple sources, assumptions about the types of acoustic stimuli and their statistical distributions should be kept to a minimum. Thus, non-parametric pattern recognition techniques are preferable as they require weaker assumptions concerning the shapes of the underlying distributions of patterns than the traditional statistical pattern recognition techniques (Lippmann, 1987). Finally, robustness and fault tolerance in the model are also important for practical applications.

Artificial neural networks have proven to be powerful tools for solving pattern recognition and comparison problems (Nigrin, 1993; Pao, 1989). More importantly, they provide an integrated pattern recognition framework which potentially has all the desirable features discussed above. Therefore, artificial neural networks were chosen to model the pattern recognition and comparison process in our proposed DLI scheme.
It should be noted that a number of recent studies (Anderson et al., 1994; Backman and Karjalainen, 1993; Lim and Duda, 1994; Neti et al., 1992; Palmieri et al., 1991; Gelfand et al., 1988; and Wang and Denbigh, 1993) have used neural networks to model auditory localization. However, these studies explored different aspects of auditory localization from the one addressed in this thesis.

Anderson et al. (1994) proposed a hierarchical neural network structure for sound localization based on cross-correlations of the left and right sensory signals and on the IID. The IID were obtained from the FFTs of the two sensory signals. Backman and Karjalainen (1993) proposed another neural network model which also used cross-correlations and FFT based IID estimates. In the model by Palmieri et al. (1991), the ITDs and IID are given pre-determined values rather than estimated ones, and neural networks were trained to perform the mapping from the ITDs and IID to the corresponding spatial locations. Although neural networks were used in these models, they were used to simulate different processes from what we are proposing to model. While we are proposing to use neural networks to obtain robust estimates of localization cues (e.g. ITDs and IID), the above models use neural networks to map estimated localization cues to spatial positions. Furthermore, our approach has the potential of extracting localization patterns that are more complex and more powerful than the traditional second-order statistics. Neural networks were also used by Neti et al. (1992) and Wang and Denbigh (1993) to model auditory localization, but these two studies focused on monaural localization cues rather than the binaural cues concerned in our work. Gelfand et al. (1988) proposed an artificial neural map based model for sensory fusion. An IID-frequency map and an ITD-frequency map were used in their model, but it is not clear how these two maps were obtained from sensory signals. No result was given for these two maps. The focus of the model was on sensory fusion rather than sound localization. Lim and Duda (1994) proposed a model for sound localization based on the output of a cochlear model. They used
auto- and cross-correlation based ITD and IID estimates to study how these cues vary with azimuth and elevation, and how well the azimuth and elevation of a single sound source can be estimated from the ITD and IID cues. It should be noted that all the above cited models examined only single-source situations. In contrast, our modeling effort is devoted to localization in multiple source environments.
3 A DLI Model for Auditory Localization

§3.1 A DLI Model

In the last chapter, considering the characteristics of real acoustic environments, we proposed a modeling scheme in which the process of localizing multiple sound sources in a noisy non-stationary environment was divided into three stages in series: decomposition, localization, and integration (DLI). Furthermore, as discussed in the last chapter, the localization process in the DLI scheme, which corresponds to the estimation of short-time location cues, can be realized by looking for and comparing patterns in the spectro-temporal distributions of the sound stimuli. Here we present a DLI model, shown in Fig. 3.1, that uses the above described idea of pattern recognition and comparison. The pattern recognition and comparison process is to be implemented with neural networks which possess the demonstrated ability to learn from examples to perform such tasks (Rumelhart et al., 1986).

§3.2 Three Parallel Implementations of the DLI Model

As reviewed in Chapter 1, three binaural differences (ITDs, IIDs, and IEDs) play important roles in auditory localization. They all have profound effects on the perceived lateral position of the sound image when the ears are stimulated by dichotic stimuli, but experimental evidence indicates that separate pathways in the auditory systems are responsible for the observed effects of different cues (Yost and Hafter, 1987; Kuwada and Yin, 1987; Konishi, 1993; Durlach and Colburn, 1978).

Specifically, as it has long been recognized in the duplex theory (Rayleigh, 1907), the localization of tones is determined by the sensitivity of the auditory system to different cues (ITDs and IID5s) in different frequency regions (Mills, 1960, 1972). The ITD sensitivity of the auditory system is limited to a frequency range bounded by
a relatively low upper frequency (1200 Hz for human) while the IID sensitivity extents beyond 10 kHz (Mills, 1960, Yost and Dye, 1987). Furthermore, physiological and anatomical experiments suggest that the processing of the IID and ITD cues occurs in different auditory nuclei (Boudreau and Tsuchitani, 1968; Caird and Klucke, 1983; Goldberg and Brown, 1969; Yin and Chan, 1990; Guinan et al., 1972).

![Diagram](image)

Fig. 3.1 A schematic diagram of a DLI model for auditory localization in complex acoustic environments.

Although the auditory sensitivity to the ITDs is limited to low frequencies for tones, the auditory system is sensitive to the envelope delays for complex high-frequency stimuli, such as click trains and amplitude modulated tones (Leakey et al., 1958; David et al., 1958, 1959; Yost et al., 1971; McFadden and Pasanen, 1976; Henning, 1980). Also, some auditory neurons with high characteristic frequencies are sensitive to the IEDs (Yin et al., 1984). Thus, it seems that the processing of ITDs
and IEDs also occurs in different neural circuitry originating form different parts of the basilar membrane.

Although, as discussed above, different binaural cues are processed in separate pathways in the auditory system, the ideas behind the generic DLI model, shown in Fig. 3.1, remain valid for all three binaural cues for the following reasons. First, the arguments made in the last chapter for the DLI modeling scheme can be referred to different binaural cues. Second, although the idea of pattern recognition and comparison was introduced in the last chapter with the measurement of small time differences in mind, a different aspect of the interaural differences, namely the IID, may also be measured by way of such a pattern comparison process. Third, the structured pattern feature observed in Fig. 2.1 on Page 53 is a result of the phase-locking property of the neural activity of the auditory nerve fibers (Kuwada and Yin, 1987; Pickles, 1988) to the fine timing structure of the stimulus waveform. Although such phase-locking to the fine timing structure disappears for high-frequency fibers, there is evidence that high-frequency fibers are able to phase-lock to the envelopes of amplitude-modulated signals with high-frequency carriers (Moller, 1974). Moller (1974) found that high-frequency cochlear nucleus neurons, which are directly innervated by high-frequency auditory nerve fibers, exhibit phase-locking property to the envelopes of amplitude-modulated stimulus signals. Thus, the patterns in the neural activities of the high-frequency auditory nerve fibers, which reflect the timing structures of the envelopes of the sound stimuli, can still be explored and compared in order to measure the IEDs. In conclusion, different implementations of the generic DLI model shown in Fig. 3.1 can be developed to model auditory localization based on different binaural cues. Fig. 3.2a through 3.2c show three such implementations corresponding to the three binaural cues (ITD, IID, and IED) respectively.
Fig. 3.2 Schematic diagrams of three implementations of the generic DLI model (shown in Fig. 3.1): (a) ITD estimation; (b) IID estimation; and (c) IED estimation. ST-IID: short-time IID. ST-ITD: short-time ITD. ST-IED: short-time IED.
Chapter 3  A DLI Model for Auditory Localization

ITD

ST-ITD Integration

Number of ST-ITDs

ITD

$S_L(t)$

Left Auditory
Peripheral Model

ITD Network

Right Auditory
Peripheral Model

$S_R(t)$

(a)
Chapter 3  A DLI Model for Auditory Localization  63

ST-IID Integration

IID Network

Left Auditory Periphery Model

Right Auditory Periphery Model

(b)

Number of ST-IIDs

IID

SL(t)

SR(t)
Chapter 3  A DLI Model for Auditory Localization

(c)
Corresponding to the different processing stages shown in the generic DLI model (Fig. 3.1), each of the three parallel models shown in Fig. 3.2 consists of three processing stages in series. Specifically, in each of the three diagrams shown in Fig. 3.2, the left and the right stimuli are first processed by the models of the left and right peripheral auditory systems, and are transformed into two sets of parallel signals representing the firing probabilities of parallel auditory nerve fibers. These two sets of parallel signals are then processed by a parallel set of neural networks (shown in each diagram in Fig. 3.2 are one of such networks in the set) whose tasks are to look for patterns in the outputs of the peripheral auditory models, and to abstract the corresponding binaural localization cues (ITD in Fig. 3.2a, IID in Fig. 3.2b, IED in Fig. 3.2c) by comparing the patterns between the two sides.

By placing different networks on different parts of the model basilar membrane in the auditory periphery, patterns in different frequency regions can be compared. Thus, to cover the entire frequency range, it is necessary to construct a parallel set of such networks in each of the three implementations shown in Fig. 3.2.

The outputs of these networks are series of short-time cue estimates. These short-time estimates are fed into an information integration mechanism, which corresponds to the integration process shown in each of the diagrams in Fig 3.2, producing histogram type distribution functions of the short-time estimates. Peaks in these distribution functions indicates the most probable cue estimates in the acoustic environment over the integration time period. In the following sections we shall discuss the detailed implementation of the three serial processing stages we have just described.

Although the three diagrams shown in Fig. 3.2 have similar structures, the tasks for the neural networks in different diagrams are different. They are required to estimate different binaural cues, namely the ITDs, IID, and IEDs, and they operate in parallel, corresponding to different neural circuitry in the auditory system that process
different cues. A question arises as to the interpretation of the outputs of the three implementations which are distribution functions of different short-time cues. As all three types of binaural cues (ITDs, IID, and IED) indicate spatial locations of sound sources in the acoustic environment, the three types of distribution functions provide a complementary description of the overall picture of the acoustic environment. A peak in any distribution function at the outputs of the three diagrams in Fig. 3.2 indicate a presence of a sound source whose location is determined by the cue value corresponding to the peak. If consistent ITD, IID, and IED are received by the two ears, the overall picture of the acoustic environment in terms of spatial distribution of the sound sources can always be obtained by combining the indication of sound sources from all three type of distribution functions through an “or” operation.

§3.3 Modeling the Auditory Periphery

As shown in both Figs. 3.1 and 3.2, the first process that needs to be modeled is the auditory periphery, which has been under study for more than a century (Helmholtz, 1862; Bekesy, 1960; Siebert, 1968; Allen, 1985). A rich body of knowledge of the physiology and anatomy as well as the signal processing characteristics of the auditory periphery has been gathered through the joint effort of researchers in various related fields. Although there are still some important questions about the system that have not been fully answered, a clear picture of the basic elements and their functions in the system has been obtained. In particular, the peripheral processing of sound stimuli by the auditory system can be roughly divided into three stages: (a) the acoustical and mechanical filtering effect of the outer and middle ears; (b) mechanical filtering due to the motion of the basilar membrane in the cochlea; and (c) transduction of the mechanical movement of the basilar membrane to neural activity in the auditory nerve fibers, through the activation of the hair cells of the cochlea. Many theoretical models of these processing stages have been proposed over the past
several decades (Siebert, 1968; Allen, 1985; Carney, 1993). Although different approaches have been used in different models, they give us a similar picture. Roughly speaking, the effect of the basilar membrane motion is similar to that of a parallel set of band-pass filters that give rise to a frequency-map representation of the sound stimuli. Such a representation is then sent to the higher processing centers of the auditory system in the form of neural activities through the auditory nerve fibers. The transduction from the mechanical movement of the basilar membrane to these neural activities is carried out by the hair cells in the cochlea, which have been modeled computationally as a rectifier plus a nonlinear transformation of the rectified signals (Allen, 1985; Carney, 1993).

We do not attempt to develop new models for the auditory periphery since it has been studied extensively, and many models exist. Rather, we selected the model reported by Carney (1993) to be used in our work. This is an integrated assembly of models incorporating the different processing stages of the peripheral auditory system. The overall model describes a process that transforms a stimulus signal into the firing rates of model auditory nerve fibers. One of the reasons for the choice of this model is that it provides us with accurate simulations of the temporal response properties of single auditory-nerve fibers for both single tones and complex sound stimuli. Since the temporal response characteristics of the auditory nerve fibers are the foundation for the ITD sensitivity observed in the auditory system, these response properties are especially important for the estimation of ITDs. Another reason why we select this model is the computational simplicity of the simulation of the model. This is important from the practical point of view: if our model of auditory localization were applied to practical applications, we would like to have a system that could work in real-time. Thus, the amount of computation becomes an important factor, and should be minimized. We leave the detailed description and computer simulation of this model to the next chapter.
§3.4 Measuring Short-Time Interaural Differences via Neural Networks

A central process in our model is that undertaken by the neural networks, shown in the three diagrams in Fig. 3.2, whose tasks are to estimate short-time interaural differences (ITDs, IID, and IED) by looking for and comparing patterns between the two sides in parallel neural activities of auditory nerve fibers. These tasks are learned by the networks through training processes. In this section, we describe the detailed implementations and training of such networks.

3.4.1 Spectro-Temporal Patterns at the Inputs of the Networks

Physiological studies show that auditory nerve fibers with lower (below 3 kHz for the cat) characteristic frequencies exhibit phase-locking of their neural activities to the stimulus waveform (Kuwada and Yin, 1987; Pickles, 1988). As discussed in Chapter 1, this phase-locking property of the low-frequency nerve fibers is the foundation of the ITD sensitivity of the auditory system. The phase shift between the phase-locked neural activities from the two ears is used as a measure of the time delay of the stimuli detected by the two ears. Such phase shift is estimated, in most previous models of the ITD sensitivity, by using the cross-correlation function of the neural activities of two corresponding single auditory nerve fibers, each from one ear (Sayers and Cherry, 1957; Sayers, 1964; Colburn, 1973, 1977; Shamma et al., 1989; Stern et al., 1988; Lindemann, 1986a, b; Gaik, 1993; Arad et al., 1994).

Another characteristics of the firing of the auditory nerve fibers is that their mean firing rates change as monotonic functions of the intensities of the sound stimuli (Pickles, 1988). Thus, the neural activation level of the auditory nerve fibers provide a reference for the measurements of the IID. Indeed, as reviewed in Chapter 1, the so called EI cells in the auditory pathways, who receive excitatory input from the ipsilateral ear and inhibitory input from the contralateral ear, are thought to be responsible for such measurements of the IID. Moreover, some previous models for
the IID sensitivity (Colburn and Moss, 1981; Johnson et al., 1990; Reed and Blum, 1990) provide detailed mechanisms for the comparison of the activation levels of single auditory nerve fibers between the two ears in order to obtain estimates of the IIDs.

A uniqueness of our model (see Fig. 3.2) lies in that the location cues are estimated by pattern comparison between group activities of sets of auditory nerve fibers, rather than by cross-correlating the two temporal waveforms (for the ITDs) or comparing the neural activation levels (for the IIDs) of two single auditory nerve fibers. Due to the phase-locking activity of individual nerve fibers, the spectro-temporal patterns are also phase-locked to the stimulus waveform for low-frequency fibers. For instance, from Fig. 3.3, which shows the activation of a group of auditory nerve fibers with adjacent characteristic frequencies responding to a 900 Hz pure-tone, one can appreciate the clear temporal alignment of the neural signals of the auditory nerve fibers. The phase information is represented in the line patterns of the neural activation across frequency. These patterns provide excellent references for small time difference measurements. Furthermore, the activation levels of these patterns provide references for the measurement of IIDs. Both the temporal shift and the activation level difference of these spectro-temporal patterns can be measured between the neural activities of two sets (one from each ear) of auditory nerve fibers, which yield short-time ITD and IID estimates respectively. These short-time estimates can be repeatedly generated in an on-going process as the ears continue to listen to the sound stimuli. Thus, in our proposed model (Fig. 3.2), the neural networks view an integrated group of auditory nerve fibers as a whole, as both the phase and intensity information is clearly encoded in the group activity patterns of the corresponding auditory fibers.

An advantage of this approach is that, in order to produce estimates of the location cues, a binaural neuron only needs to observe the input signals for a period of time
that is in the order of the reciprocal of the neuron's characteristic frequency. This provides the foundation for cue estimation over short time spans, and for the integration of the short-time cue estimates to obtain an overall picture of the acoustic environment. Also, it provides an opportunity to handle fast time-varying situations. Another advantage is that the two-dimensional spectro-temporal patterns give a relatively robust representation of both the phase and intensity information of the stimulus waveform. If some single fibers' signals contain higher noise than the signals of the rest of the fibers do, the image pattern will still be recognizable and allow pattern comparison.

![Fig. 3.3 Example of neural activities of a group of nine (vertical axis) modeled auditory nerve fibers in responding to a pure tone stimuli of 900 Hz.](image)

Neural networks trained to estimate ITDs will not work for high-frequency stimuli because auditory nerve fibers with high characteristic frequencies cannot phase-lock to the fine timing structure of the stimulus waveform (Pickles, 1988; Kuwada and Yin, 1987). However, as reviewed in Chapter 1, Moller's experiment (1974) suggests that auditory nerve fibers with high characteristic frequencies show phase-locking to the envelopes of amplitude-modulated signals with high-frequency carriers. Thus, for complex high-frequency stimuli, the group activities of high-frequency auditory nerve
fibers show spectro-temporal patterns that reflect the temporal structures of the envelopes of the stimulus signals. Such patterns can then be compared by neural networks (see Fig. 3.2c) to measure the temporal delays of envelopes (i.e. the IED) between the two ears. A major difference between the ITD estimation and the IED estimation is that for the ITD case, the patterns in the input to the neural networks reflect the fine timing structures of the sound stimuli whereas for the IED case the patterns reflect the timing structures of the envelopes of the sound stimuli. Thus, a similar structured network may be used to estimate IEDs. A unique feature of the envelope spectro-temporal patterns of the stimuli is that the patterns will not show up if the stimuli are not amplitude-modulated with a modulation rate falling in an appropriate range as defined below.

Consider an amplitude-modulated sound stimulus signal of the following form

$$S(t) = A \sin(\omega_m t + \varphi_m) \sin(\omega_c t + \varphi_c)$$

(3.1)

where $\omega_m$ is the modulation frequency, $\omega_c >> \omega_m$ is the carrier frequency, $\varphi_m$ and $\varphi_c$ are the phases of the modulation and carrier signals, respectively. Eq. 3.1 can be rewritten as

$$S(t) = \frac{A}{2} [\cos(\omega_1 t + \varphi_1) - \cos(\omega_2 t + \varphi_2)]$$

(3.2)

where $\omega_1 = \omega_c - \omega_m$, $\omega_2 = \omega_c + \omega_m$, $\varphi_1 = \varphi_c - \varphi_m$, and $\varphi_1 = \varphi_c + \varphi_m$. Thus, the Fourier spectrum of $S(t)$ consists of two adjacent sinusoids. The distance in frequency space between the two sinusoids is twice as much as the modulation frequency. The midpoint between the two sinusoids on the frequency axis is the carrier frequency, as shown in Fig. 3.4.

As reviewed in Section 3.3, the effect of the basilar membrane motion is similar to that of a parallel set of band-pass filters. If the two sinusoids in Fig. 3.4 are separated at such a distance that at most only one sinusoid falls within the pass band of a particular filter (i.e. the modulation rate is too high), the output of that filter will
at most contain one major sinusoid. Thus, the envelope of this output will be largely flat with a major DC component. Fig. 3.5 shows such a case where a window of 13 time slices of the neural activities of a set of nine auditory nerve fibers are presented in a matrix of small squares. The larger the white area in each square in the figure, the higher the activation of the corresponding fiber at the corresponding time slice.

![Fig. 3.4 Frequency spectrum of the signal in Eq. 3.2.](image)

When the two sinusoids in Fig. 3.4 are sufficiently close to each other so that both of them fall within the pass bands of a set of auditory nerve fibers, the activation of these auditory nerve fibers will show patterns reflecting the amplitude modulation of the stimuli. Fig. 3.6 shows such a case.

If the two sinusoids in Fig. 3.4 are so close to each other that the modulation rate becomes very slow, the activation of the auditory nerve fibers will not show significant temporal variation within a relatively small temporal window. Fig. 3.7 shows such a case.

In summary, there are upper and lower limits of the modulation rate, within which the windowed activities of the auditory nerve fibers show spectro-temporal patterns that reflect the envelope of the stimulus signal.
3.4.2 Structure of the Networks

It is possible that networks with different structures may be trained to perform the same task. Our objective here is to find a network structure that can be trained to perform the tasks required, rather than to investigate the difference of performance of different network structures that may be used for the task. We have chosen the back-propagation network (Rumelhart et al., 1986) for the task. This type of network has been shown to be able to learn, from training examples, complex input-output transformations (Maren et al., 1990). More importantly, multi-layer back-propagation networks are able to form arbitrarily complex decision regions for the purpose of pattern classification and recognition (Lippmann, 1987). This ability is directly related to the non-linearity of the individual neurons and to the fact that the networks have multiple layers. In our simulation study, a three-layer structure was used. As
will be evident from the results to be presented in Chapters 5, 6, and 7, the same network structure can be trained to perform the three different tasks required in the three parallel implementations (see Fig. 3.2) of the generic DLI model shown in Fig. 3.1. Fig. 3.8 shows a schematic diagram of the structure.

![Diagram of network structure](image)

**Fig. 3.6 Intermediate modulation rates.** Example of neural activities of a group of auditory nerve fibers when the stimulus has intermediate modulation rates. Each square represents the activation of a particular auditory nerve fiber at a particular time slice. The white area in each square is proportional to the activation level. AN: auditory nerve.

The three layers of the network shown in Fig. 3.8 are called the input, hidden, and output layers, respectively. The input layer is fully connected with the hidden layer, and the hidden layer with the output layer. Both of these connections are feed-forward which means that each neuron in the receiving layer receives weighted input from every neuron in the sending layer, and that the activation of a receiving neuron is governed by the following equation:
where \( s_i \) is the output of the \( i \)-th neuron in the receiving layer, \( u_j \)'s are the outputs of the neurons in the sending layer, \( b \) is a bias constant, \( w_{ij} \)'s are the connection weights between the two layers, \( N \) is the number of neurons in the sending layer, and the function \( f(\cdot) \) is defined as a sigmoid nonlinearity

\[
f(x) = \frac{1}{1 + e^{-x}}.
\]

The connection weights and the bias constant are to be determined in a training process as will be described in the next subsection.

Fig. 3.7 Low modulation rates. Example of neural activities of a group of auditory nerve fibers when the stimulus has relatively low modulation rates. Each square represents the activation of a particular auditory nerve fiber at a particular time slice. The white area in each square is proportional to the activation level. AN: auditory nerve.
The input layer of the network consists of two parts, with each corresponding to one of the two ears. The two parts have the same structure in the form of a rectangular matrix, as shown in the input layer in Fig. 3.8. The activity in the matrix of neurons represents the neural activation pattern of a group of auditory nerve fibers over a short-time interval. Each column of the matrix corresponds to one auditory nerve fiber in the group, and the rows correspond to the neural firing rates of the auditory nerve fibers at successive discrete time instants in the time interval shown. The sampling rate is minimized to achieve simulation efficiency without causing any aliasing problem. This is possible because the neural activities of the auditory nerve fibers correspond to outputs of the band-pass filters used to model the mechanical filtering effect of the basilar membrane.
In particular, the activation of the neurons in the input matrix can be viewed as a small patch of the overall neural spectro-temporal representation of the sound stimulus. An alternative view is that the matrix of the input neurons is a small spectro-temporal window that can be used to scan the spectro-temporal input along the time axis. To determine the size of this spectro-temporal window, or the number of columns and rows in the input matrix, several considerations are relevant. First, the activation patterns in the window must be representative of the patterns exhibited in the overall spectro-temporal representation produced by the auditory peripheral model that conveys the timing and intensity information. Accordingly, the window size should not be too small to accommodate such patterns. On the other hand, in order to follow the possibly time-varying cues in non-stationary situations, the temporal dimension of the window should be as small as possible. Also, in order to differentiate the possibly multiple sound sources having different frequency contents, the spectral dimension of the window should also be small. Finally, a small window is also desirable from the point of view of computational convenience in the training and simulation of the network. For the simulation presented in Chapters 5, 6, and 7, a 13 by 9 input matrix is used. This corresponds to nine adjacent auditory nerve fiber models sampled at thirteen consecutive time instants. For low-frequency fibers, the sampling interval is determined by the upper limit of the frequency range covered by the relevant auditory nerve fibers. For high-frequency fibers, which cannot phase-lock to the fine temporal waveform of the sound stimulus, the sampling interval is determined by the upper limit of the relevant modulation frequency range described in the last section. The characteristic frequencies of the nine auditory nerve fibers at the input to the ITD network are chosen to be frequency values uniformly sampled in the equivalent-rectangular-bandwidth-rate (ERB-rate) scale (Moore and Glasberg, 1983). A sampling interval of 0.2 in the ERB-rate scale is used. This interval corresponds to a frequency interval of about 100 Hz at 1 kHz. In Fig. 3.8, where an example of the
activation of the input matrices of neurons stimulated by a 900 Hz tone are shown, a 13 by 9 window contains a complete period of the periodical patterns (lines) in the neural spectro-temporal representation of the sound stimulus (see also Fig. 3.3).

In contrast to the input layer, the hidden layer is an array of neurons that do not have any particular order. Each hidden neuron receives a weighted sum of the inputs from both matrices of neurons in the input layer. Different neurons usually have different weighting vectors, abstracting different information in the input activation patterns. These weighting factors are determined by a training process which is designed to force the network to learn its tasks, as will be described in the next subsection. The number of neurons in the hidden layer is usually determined empirically depending on the specific tasks of the network. A general guideline is to choose the smallest possible number of hidden neurons that are sufficient for the network to learn the specific tasks. This not only minimizes the computing resources required for the simulation, but also prevents the network from directly memorizing the training examples instead of learning generalization for use on other stimuli.

The task of the network is to estimate the cues from the information available in the windowed neural activities of the model auditory nerve fibers, as represented by the two input matrices of neurons corresponding to the two ears. This task is realized by encoding the location cue values in the collective activities of the output layer neurons of the network. A convenient and biologically plausible way to encode the cues is to define tuning curves for the output layer neurons in responding to different cue values (Lehky and Sejnowski, 1990). Specifically, the activity or response of an output layer neuron is defined as a non-monotonic function with a single peak at a particular cue value. For computational convenience, we choose a Gaussian function for the shape of the tuning curves. Fig. 3.9 shows such a tuning curve for the ITD cue which specify the response activity of a particular output layer neuron as a Gaussian function of the relevant ITDs.
Fig. 3.9 An example of the tuning curve of an output layer neuron. The arrowhead line at ITD = 0.1 ms represents the activity of the neuron for this particular ITD value.

There are two problems associated with the use of non-monotonic tuning curves (Lehky and Sejnowski, 1990). A first problem is that if the tuning curves are narrowly shaped and each neuron is used to encode only one particular cue value, a large number of neurons are needed to obtain relatively high resolution. Another problem is that a neuron may give the same response to different cue values when the tuning curves overlap. One way of resolving this problem is to encode a range of cue values by the collective responses of a group of neurons rather than by single-neuron responses to each cue value within the range. Fig. 3.10 shows an example of the activity pattern of an ordered array of output neurons corresponding to a specific cue value. In this particular example, these responses are determined by their Gaussian shaped tuning curves which are defined as

\[
R_i(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu_i)^2}{2\sigma^2}}
\]

(3.9)

where \( \theta \) corresponds to the encoded cue value, \( i \) is the index of the output neurons.
The maximum-response cue values (which correspond to the $\mu_i$'s in Eq. 3.9) of the neurons in the array change in an orderly fashion from the neuron at one end of the array to the one at the other end. Thus, the spatial arrangement of the neurons in the output array is just as important as it is for the neurons in the input matrix.

![Diagram](image)

Fig. 3.10 An example of the coding of cue values by an array of output layer neurons. The arrowhead lines indicate the activities of the individual neurons. The mean of the Gaussian envelope is the encoded cue value.

The choice of the number of neurons in the output layer is also an empirical one. As a general guideline, more neurons in the layer will increase the precision of the coding scheme. However, because there will be noise in the response of the individual neurons, the incremental improvement in precision will become smaller and smaller as the number of neurons increases. For the simulation presented in Chapters 5, 6, and 7, there are 20 neurons in the output layers of the corresponding networks.
3.4.3 Training of the Networks

The tasks of the networks are to estimate the location cues as encoded in their output activity patterns upon scanning a section of the neural spectro-temporal representation of the sound stimuli. The performance of these tasks are learned through training processes in which the connection weights in the network are adjusted systematically so that the outputs of the networks will approximate the desired outputs for the training examples. The desired output activity patterns are given, and they correspond to the known cue values of the training examples. After the training is completed, the networks are expected to give appropriate cue estimates for arbitrary sound stimuli.

As the training process can be formulated as an optimization problem, various optimization techniques can be used to design the training algorithm. The objective is to find a set of the connection weights in the network that minimize the difference between the desired outputs and the actual outputs of the network for all the training examples.

As the networks are trained only to produce the correct output encoding of the known cue values of the training examples, an important question arises on how well the networks will encode the cue for arbitrary sound stimuli, or, how well the network will extrapolate and interpolate in the entire space of all possible stimuli. To achieve good ability to generalize, the choice of training examples is crucial. This in turn depends on the structure of the input space. For all three types of networks shown in Fig. 3.2, there exist multiple parameter dimensions that effectively affect the input patterns of the networks. Our strategy in designing the training example sets is, thus, to uniformly sample the major relevant parameter dimensions.

In addition to the proper sampling of the stimulus parameter space for the selection of the training example sets, another aspect must also be considered. Although the shifts or the activation level differences of the spectro-temporal patterns
between the two ears are constant for a certain corresponding cue value, the absolute phase of the pattern in the spectrotemporal representation from each ear is changing continuously. Thus, the networks should be trained to measure the constant phase or activation level differences independent of the absolute phase of the spectrotemporal pattern from each ear. This is achieved by including, in the training set, examples that have different absolute phases of the same spectro-temporal patterns.

§3.5 Integrating Short-Time Cue Estimates

A simple way of integrating the short-time cue estimates (produced by the networks discussed in the last section) is to use histograms. Histograms of such short-time estimates are formed by dividing a corresponding range of cue values into a number of bins of relatively small interval. The short-time estimates are then put into the corresponding bins depending on their values. In a dynamic situation, the short-time estimates changes from one time window to another. Thus, the resulting histograms are estimates of short-time cue distributions over the corresponding listening period. Such distributions give us direct indications of the spatial distribution of sound sources in the observation time period. Peaks in such distribution functions indicate the most likely estimates of the cue values that correspond to different sound sources. Fig. 3.11 shows such a distribution function for the ITD cue. The single peak in the distribution indicates a single source localized at a position that corresponds to the ITD value at the peak.

The idea of using histograms to integrate short-time cue estimates is biologically plausible. A possible neural implementation of the process of building a histogram is to have an array of neurons representing the different bins in the histogram. The firing rate of a neuron in the array corresponds to the height of the corresponding histogram bins, and this firing rate is driven by the short-time estimates that coincide with the cue value represented by the neuron.
§3.6 Summary of the DLI Model

In summary, our model for auditory localization consists of several processing stages in series, with each stage having parallel processing pathways or modules. The stimulus signals detected by the two ears are first processed by models of the left and right peripheral auditory systems, respectively. These two models transfer their incoming sound signals into parallel sets of band-pass signals with different center frequencies, forming two neural spectro-temporal images of the sound stimuli at the two ears. The second stage consists of three parallel processing pathways, each of which is arranged to generate short-time estimates of one of the three binaural cues: ITD, IID, and IED. Within each pathway, a parallel set of neural networks are trained

Fig. 3.11 An example of the ITD distribution function. The sound stimuli used in the simulation were 1 kHz pure-tones on which white noise was added with a 20 dB signal-noise ratio. The ITD in the stimuli was 250 μs. The peak around 250 μs indicates the correct estimate of the true ITD value.
to estimate the corresponding short-time cues based on information from small sections of the entire frequency range. In effect, each neural network in the set takes, as its input, two patches of the neural spectro-temporal images of the sound stimuli at the two ears as represented in the neural activities of the model auditory nerve fibers. The ITD and IED networks are trained to measure the temporal shift of patterns in the neural images corresponding to the two ears, while the IID networks are trained to measure the activation level differences between the two neural images. The outputs of these networks are series of short-time estimates of the ITD, IID, and IED which are fed into an information integration process, giving rise to distributions of the corresponding short-time cues. Peaks of the distribution functions correspond to sound sources appearing in the corresponding listening period.
4 Simulation Methods

In the last two chapters we have introduced and described our model for auditory localization. The model is studied using computer simulations. In this chapter, we discuss the methods used in our simulations. We first discuss the detailed simulation of a model (by Carney, 1993) for the peripheral auditory system, which is used as the first module or processing stage in our DLI model shown in Fig. 3.1. Second, we discuss the simulation and training of the neural networks which are the central processes in our model. As mentioned in the last chapter, only one network in each of the three parallel implementations (Fig. 3.2) of the DLI model (Fig. 3.1) needs to be simulated to gain a good understanding of the behavior of the model. Finally, we discuss the evaluation methods used to assess the performance of the trained networks.

§4.1 Simulation of the Peripheral Auditory System

As mentioned in Section 3.3, a model by Carney (1993) is used in our simulation. The model transforms a stimulus signal into the firing rates on the model auditory nerve fibers. This transformation has previously been modeled as a cascade of a linear filter, a static nonlinearity, and a stochastic pulse generator, as shown in Fig. 4.1 (de Boer and Kuyper, 1968; de Boer and de Jongh, 1978). It has been found (Johannesma, 1972; de Boer, 1975; de Boer and Kruidenier, 1990) that a filter of the form

\[
g(t) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right) \cdot \cos(\omega_c(t - \alpha)), \quad \text{for } t \geq \alpha,
\]

\[
g(t) = 0, \quad \text{for } t < \alpha.
\]

(4.1)

can be used in the model shown in Fig. 4.1 to describe the responses of auditory nerve fibers. The parameters in Eq. 4.1 are estimated by fitting the model in Fig. 4.1 to
activities of auditory nerve fibers responding to wideband noise stimuli. The function in Eq. 4.1 provides an estimate of the linear aspects of the acoustical and mechanical filtering of the middle ear and the basilar membrane (de Boer and de Jongh, 1978).

A non-linear characteristics of basilar membrane mechanics, which is not considered in the model shown in Fig. 4.1 above, is that the bandwidths of the cochlear filters changes as a function of the intensity of the sound stimulus (Rhode, 1971; Hall, 1974; Johnstone et al., 1986). Carney (1993) incorporated this non-linearity in the model shown in Fig. 4.1 by introducing a feedback mechanism to control the bandwidth of the filter in Eq. 4.1 according to the intensity of the sound stimulus. A schematic diagram of Carney’s model is shown in Fig. 4.2. Four processing elements can be identified in the diagram: (i) a time-varying narrow-band filter that describes the linear aspects of the mechanical tuning of the basilar membrane; (ii) a feedback mechanism for the control of the bandwidth of the narrow-band filter in (i); (iii) an overall time delay consisting of the traveling wave, acoustical and synaptic delays; and (iv) a static nonlinearity that describes the conversion of mechanical movement to neural signals by the inner hair cells. In the following subsections (4.1.1 through 4.1.4), we shall give more detailed descriptions of these elements (Carney, 1993).
Fig. 4.2 A schematic diagram of the auditory periphery model by Carney (1993). See text for descriptions of the different elements in the model. (Adopted from Carney, 1993.)
4.1.1 Time-Varying Narrow-Band Filter

For the parameter ranges relevant to auditory nerve fiber responses, the function in Eq. 4.1 has a simple frequency domain approximation (Patterson et al., 1988):

\[ G(\omega) \propto \left( \frac{1}{1 + j\tau(\omega - \omega_{cp})} \right)^\gamma e^{-j\omega \alpha}. \]  (4.2)

The parameter \( \tau \) has a strong influence on the bandwidth of the filter. The time delay \( \alpha \) is included as part of the delay introduced after the narrow-band filter. The value of \( \gamma \) is set equal to 4 for fibers with relatively low characteristic frequencies (de Boer, 1975).

In our simulation, Eq. 4.2 is implemented through a cascade of \( \gamma \) digital filters of the following form:

\[ q_1(kT) = A[q_0(kT) + q_0(kT-T) + B(kT)q_1(kT-T)] \]  (4.3)

where \( A = 1/(1.5\tau_0 C + 1), \) and \( B(kT) = F(kT)C - 1, \) with \( C = 2/T. \) \( \tau_0 \) represents the time constant of the exponential damping of the function in Eq. 4.1 for a 75 dB sound-pressure level. \( F(kT) \) is the feedback signal shown in Fig. 4.2, and will be described in the next subsection.

4.1.2 Feedback Control of the Filter Bandwidth

An important aspect of the mechanical tuning of the basilar membrane is that the bandwidth of the tuning varies over time as a function of the sound-pressure level of the stimulus (Rhode, 1971; Hall, 1974; Robles et al., 1976; Patuzzi et al., 1984; Johnstone et al., 1986; Ruggero and Rich, 1991). When the amplitude of the stimulus is low, the filter is relatively sharply tuned. As the amplitude of the input signal increases, the bandwidth of the filter increases. This aspect of the nonlinearity of the basilar membrane tuning is modeled through a feedback control mechanism which
includes a saturating nonlinearity and a low-pass filter. The saturating nonlinearity is modeled using a hyperbolic function of the following form:

\[ V_{\text{sat}}(t) = \frac{V_{\text{max}}}{1 + \tanh P_0} \left[ \tanh \left( \frac{0.707 \frac{P_f(t)}{P_{\text{db}}} - P_0}{P_{\text{db}}} \right) + \tanh P_0 \right], \quad (4.4) \]

where the parameter \( V_{\text{max}} \) determines the saturation value, \( P_0 \) determines the asymmetry of the nonlinearity, and \( P_{\text{db}} \) sets the operating point of the nonlinearity.

The low-pass filter in the feedback loop is implemented in the time domain as a digital filter of the form:

\[ V_{\text{out}}(kT) = C_1 V_{\text{out}}(kT - T) + C_2 \left[ V_{\text{in}}(kT) + V_{\text{in}}(kT - T) \right], \quad (4.5) \]

where \( T \) is the sampling interval, and \( C_1 \) and \( C_2 \) are the filter coefficients determined by the cut-off frequency of the filter. The output of the low-pass filter is scaled and biased according to the following equation:

\[ F(t) = \frac{3}{2} \tau_0 - \left( \frac{V_{\text{fb}}(t)}{V_{\text{max}}} \right) \left( \frac{\tau_0}{2} \right), \quad (4.6) \]

where \( F(t) \) is the feedback signal to the time-varying narrow-band filter of the basilar membrane model, and \( \tau_0 \), again, represents the time constant of the exponential damping of the function in Eq. 4.1 for a 75 dB sound-pressure level.

### 4.1.3 Traveling Wave Delay

The responses of the auditory nerve fibers have specific latencies due to the acoustical, synaptic, and traveling wave delays. To align the latency of the output of the narrow-band filter with the measured latencies of the auditory nerve fibers, a time delay is introduced after the narrow-band filtering, as shown in Fig. 4.2. The time delay varies as a function of the characteristic frequency \( \omega_{\text{CR}} \) of the auditory nerve fiber as described by the following equation:

\[ P_{\text{bm}}(t) = P_f \left( t - A_D \exp\left( -\frac{\omega_{\text{CR}}}{A_L} \right) - \frac{2\pi}{\omega_{\text{CR}}} \right), \quad (4.7) \]
where $A_D$ and $A_L$ are obtained from experimental measurements by Carney and Yin (1988), and $x$ is the distance in mm from the apex of the basilar membrane.

### 4.1.4 Model of the Inner Hair Cell

The inner hair cell is modeled as a saturating nonlinearity, which has the same form as Eq. 4.4, followed by two low-pass filters that have the same form as Eq. 4.5. The asymmetrical saturating nonlinearity of Eq. 4.4 has previously been proposed to model the input/output characteristics of the inner hair cells by Russell and Sellick (1978). The two low-pass filters represent the electrical filtering of the inner hair cell membrane (Russell and Sellick, 1983). The cut-off frequency of the two low-pass filters is set to 1100 Hz, which is the same as that used for the low-pass filter in the feedback loop shown in Fig. 4.2.

### 4.1.5 Modeling High-Frequency Auditory Nerve Fibers

In order to model the effect that high-frequency auditory nerve fibers cannot phase-lock to the fine timing structure of the sound stimulus, we add an extra processing stage in the peripheral model described in the previous subsections. Specifically, the envelope of the output of the time-varying narrow-band filter in Fig. 4.2 is abstracted and sent to the next processing stage. This model is used for the auditory nerve fibers at the inputs to both IID and IED networks shown in Fig. 3.2.

### §4.2 Simulation of the Neural Networks

We have once tried to write our own computer programs to facilitate the simulation studies of the neural networks in our model. A critical issue in the simulation of neural networks is the implementation of the training algorithms. It is important for us to implement different training algorithms in the simulation programs. This will allow us to test with different training methods in order to chose the most appropriate
(in terms of the training speed and the ability to avoid local minima) ones. Another issue concerns the user interface. A graphical user interface is desired, which allow us to plot the changes of important parameters during training, to view the activity patterns in different layers of the network during and after training, and to view the connection patterns between layers of the network. The interface should be interactive in the sense that the user has full control of the processes taking place in the network, such as setting break points and stepping through the training process, changing parameters at any point during training, and testing a trained network by individual test stimuli.

In the early stage of our simulation study, we became aware of a neural network simulation package called Xerion (van Camp, 1993) that meets the requirements of our simulation. Xerion is a collection of neural network simulation tools developed by a research group in the Department of Computer Science at the University of Toronto. It runs on the UNIX platform, and allows easy creation of different simulators for different types of neural networks. Thus, instead of continuing with writing our own simulation programs, we have since used the back-propagation network simulator, which comes with the Xerion package, to train the networks in our model. Most of the popular training algorithms for back-propagation networks are implemented in the simulator. Moreover, the simulator has an excellent graphical user interface that allow the user to visualize and manipulate the networks in the way discussed above. It also has a UNIX-like script shell that allow automation of the training and manipulation of the networks.

4.2.1 Training Examples

To limit the number of training examples, only the simplest types are used. It is also interesting to know whether a network trained with the simplest type of stimuli can "generalize" and perform well with more complex stimuli. For the ITD and IID
networks, only pure-tone stimuli are used. For the TED networks, sinusoidally amplitude-modulated (AM) tones are used. In order to train the networks to perform their tasks (estimating short-time cues) in noisy situations, white noise is added to the training stimuli.

As mentioned in Subsection 3.4.3, the choice of training examples is to uniformly sample the relevant parameter dimensions of the stimulus space. For the ITD and IID networks (Fig. 3.2a, b), the relevant parameters of the stimuli are frequency, intensity, signal-noise ratio (SNR), ITD, and IID. For the TED networks (Fig. 3.2c), the modulation frequencies of AM stimuli are also relevant.

A training example to a network consists of two parts. One part is the input pattern to the network, and the other part is the corresponding output pattern the network is expected to produce upon being presented with the input pattern. As discussed in Subsection 3.4.2, the output patterns of the networks in our model shown in Fig. 3.2 encode the corresponding localization cues. Such encoding is defined by the tuning curves of the output layer neurons in responding to different cue values. These tuning curves have Gaussian shapes whose maximum response points vary in an orderly fashion from one end of the output layer to the other end, covering a specific cue value range. The ratio between the standard deviation of the Gaussian curves and the cue value range covered by the output layer neurons is 15. By choosing such a ratio, the activities of the output layer neurons show a Gaussian shaped pattern that occupies one-fifth of the entire output layer, and the position of this pattern shift along the output layer depending on the encoded cue value. For an ITD network, the cue value range is determined by the maximum characteristic frequency of the auditory nerve fibers at the input to the network. For IID networks, the encoded cue value range is from -12 dB to 12 dB. This value (12 dB) of IID corresponds to a sound image completely lateralized to one ear, as reviewed in
Chapter 1. For a IED network, the cue value range is determined by the upper limit of modulation frequency discussed in Subsection 3.4.1.

### 4.2.2 Specifications of Three Networks to be Tested

As reviewed in Chapter 1, the ITD sensitivity of the auditory system is limited to frequencies below 1200 Hz (Yost and Hafer, 1987), whereas the IED sensitivity is only observed in high-frequency stimuli (Bekesy, 1960; Henning, 1980; Yin et al., 1984). For the IID sensitivity, although it is observed in the entire frequency range (see Fig. 1.9 on Page 16), free-field stimuli can only induce significant IIDs in the high frequency range (see Fig. 1.5). Thus, different location cues are estimated not only in separated pathways in the auditory system, as discussed in Section 3.2, but also in different frequency regions. Simulation of the ITD networks will be limited to the lower frequency region (below 1200 Hz), whereas that of the IID and IED networks limited to the higher frequency region.

From the model for the peripheral auditory processing, it follows (see Eq. 4.1) that if we normalize the frequency tuning curves that correspond to different points along the basilar membrane according to their corresponding characteristic frequencies, they have similar shapes. This suggests that, within the same frequency region in which different networks of the same type are constructed, the input patterns to these networks have similar characteristics. As these networks rely on such input patterns to abstract the interaural differences, they are expected to have similar behaviors. Thus, only one network of each of the three types (ITD, IID, and IED; see Fig. 3.2) will be studied, and the choice of the frequency ranges covered by these networks is somewhat arbitrary.

For the chosen ITD network (to be presented in Chapter 5), nine fibers from each of the two ears are used at the input of the network. The characteristic frequencies of the fibers cover the frequency range of 800-1000 Hz. This range is adjacent to the
upper limit of the frequency region of ITD sensitivity. Thirteen consecutive samples
(at a sampling interval of 200 μs) of each signal from these fibers are sent to the
network. The ITD range covered by the output layer neurons is from -505 μs to 505
μs. There are 20 neurons in the output layer that respond maximally at ITD values of
μ_i = -505 + 50.5(i - 1) μs, i = 1,⋯,20, respectively. These μ_i 's are uniform samples
over the range from -505 μs to 505 μs.

For the chosen IID network (to be presented in Chapter 6), another set of nine
auditory nerve fibers from each ear is used at the input of the network. The model of
the auditory periphery used in our simulation is derived largely from the observed
behavior of the auditory periphery at low frequencies (below 5000 Hz), as most
auditory periphery models are (Carney, 1993). Thus, we have chosen an intermediate
frequency range within the region of 1200-5000 Hz for the simulation of the IID
network. Specifically, the characteristic frequencies of the IID network cover the
range of 2500-3500 Hz. The activities of these auditory nerve fibers, which reflect
the timing structure of the envelope of the stimulus signals, are sampled at 500 μs
intervals. Thirteen consecutive samples from each fiber are sent to the network.
There are also twenty neurons in the output layer of the network which respond
maximally at IID values of μ_i = -12 + 1.26(i - 1) dB, i = 1,⋯,20, respectively. These
μ_i 's are uniform samples over the range of -12 dB to 12 dB.

The same fibers as those in the IID network described above are used at the input
to the trained IED network to be presented in Chapter 7. However, the IED network
is trained differently: rather than measuring the activation level differences between
the two ears (as is the IID network trained to do), the IED network is trained to
measure the temporal delay between the envelopes of the stimulus signals at the two
ears. The effective modulation frequency (Subsection 3.4.1), at which the activation
patterns of the relevant auditory nerve fibers reflect the temporal structure of the
modulation, ranges from 154 to 305 Hz. The same number of output layer neurons as
that used in the IID network are used in the IED network, but they are trained to encode a range of IEDs rather than IIDs. The IED range encoded by the IED network is from -1.64 to 1.64 ms, which is determined by the effective modulation frequency range. The \( i \)-th neuron in the output layer responds maximally at an IED value of \( \mu_i = -1.64 + 0.164(i - 1) \) ms. These \( \mu_i \)'s are uniform samples over the range of -1.64 to 1.64 ms.

### 4.2.3 The Training Process

The training example set for a network consists of a series of input-output pairs that the network is to be trained to associate. The training process consists of a series of iterations. In each iteration, the input patterns of the training examples are presented to the network, and the actual outputs of the network are compared to the corresponding output patterns in the examples which are the desired outputs of the network. The weights of the network are adjusted to minimize the mean squared error between the actual and the desired outputs. The training process is terminated when a local minimum is reached. The process is repeated several times using different random initial weights of the network. The final weights that correspond to the smallest local minima are taken to be the final weights of the network.

The minimization problem in the training process is solved using the conjugate-gradient method well known in optimization theory (Hestenes, 1980). This method is suitable for effective handling of large-scale problems with hundreds of variables (Fletcher, 1975; Powell, 1977; Hestenes, 1980), which is the case in our simulation. A typical minimization process consists of a series of iterations of local search in the weight space. In each iteration, the search starts with the weights resulting from last iteration, and a search direction is chosen. The error function that is to be minimized is then evaluated in the neighborhood of the starting point and along the chosen direction. The point along the direction that gives rise to the smallest error is chosen.
to be the search result of this iteration. This process is called line search. In the conjugate-gradient method, the search directions of successive iterations are chosen to be conjugate to each other (Hestenes, 1980). It is well known in optimization theory that if the error function is quadratic with respect to the weights, the exact minimum of the error function can be found in a limited number of iterations if the search directions are conjugate to each other (Fletcher, 1975; Powell, 1977; Hestenes, 1980). As any error function can be approximated by a quadratic function in the vicinity of a starting point, conjugate direction search will lead to a point in the weight space where the error function has the smallest value in the vicinity of the starting point. When such a point is reached, the search will be repeated with another set of conjugate directions in the vicinity of the new starting point. The search ends where a local minimum of the error function is reached. The error functions of the networks in our model are far more complex than a simple quadratic function due to the non-linearity (see Eq. 3.8) of the transformation from a neurons’ inputs to its output. It is difficult to predict how many iterations are needed before a local minimum is reached. Furthermore, as a line search is conducted in each iteration to find the minimum point along the search direction of the iteration, it is also difficult to estimate how long each iteration will take. Our experience with the training of the networks presented in the thesis is that the search of a local minimum needs a number of iterations that is on the order of 1000, and takes about two weeks on the Sun workstations (SparcStation 2) available.

Several factors contribute to the lengthy training of the networks in our model. A first factor is the size of the networks. There are 234 neurons in the input layer, and 20 in each of the hidden and output layers of a network. A network, thus, contains 4700 connection weights between the input and the hidden layers, and 420 weights between the hidden and output layers. A second factor is the size of the training example sets. The training examples are saved in computer files in the ASCII format.
Such an example file typically contains about 20,000,000 bytes of data. It takes the network simulator several hours to read all these data into computer memory before the actual training process starts. The limited computer resource available for our research makes it necessary to terminate a training process before it is finished, and restart the process from the termination point in a later time. This considerably increases the overall time required for the training because the entire training example set needs to be re-loaded into computer memory before the training process can be restarted. The third factor concerns the demand of computer memory of the network simulator. The memory requirement of our simulation exceeds the available memory in a workstation. Thus, the computer must constantly swap data between its memory and hard disk, slowing down the computer operations in the training process.

Because of the enormous computational resource requirement, our simulations are limited to one narrow band of frequencies in the audible spectrum for each of the three networks. Thus, the problem of across-frequency integration (Stem et al., 1988) is not addressed in this work.

4.2.4 Short-Time Cue Estimates

A trained network produces short-time cue estimates upon being presented with stimulus signals. As discussed in Subsection 3.4.2, such short-time cue estimates are encoded in the activation patterns of the output layer neurons (see Fig. 3.10). To decode these activation patterns to obtain the corresponding short-time cue estimates we need to fit Gaussian curves to the patterns. The means of the fitted Gaussian functions give the short-time cue estimates.

§4.3 Evaluation Methods of Trained Networks

After a network is trained, some important issues arise concerning the performance of the trained network when it carries out its tasks. Since the network is only trained on
limited number of training examples of the simplest type, one such issue is its
generalization ability. Another issue concerns the robustness of the network in noisy
environments. Finally, how well the network works in multi-source situations needs
to be tested.

4.3.1 Parametric Description of Cue Estimates

As discussed in Subsection 4.2.3, the trained networks give out short-time cue
estimates. The actual cue estimate for a relatively long listening period is an
integration of these short-time cue estimates generated through an integration process,
which follows the network processing stage, as shown in Fig. 3.2. The result of such
an integration process is a cue distribution function, an example of which is shown in
Fig. 4.3. Roughly speaking, peaks in such distribution functions indicate the
estimated cue values. More specifically, however, certain characteristics of the
“hump” surrounding a peak should be considered when evaluating the cue estimates.
Relevant parameters are the position and width of the hump. The position of a hump
is defined as the centroid of the area under the hump. The width of the hump is
defined as the difference between the upper and lower 6 dB cut-off positions. For the
example shown in Fig. 4.3, the width is defined to be

\[ d = \Delta t_{\text{upper}} - \Delta t_{\text{lower}} \]  \hspace{1cm} (4.8)

where \( \Delta t_{\text{upper}} \) is the ITD position below which the height of the distribution function
in the vicinity of the peak is at least 6 dB lower than the height of the peak, and \( \Delta t_{\text{lower}} \)
is the ITD position above which the height of the distribution function in the vicinity
of the peak is at least 6 dB lower than the height of the peak.

The two parameters (position and width), as defined above, are measures of how
well the cue estimates reflect the true cue values. For instance, the position gives an
estimate of the true cue value, and the width gives us some ideas about the resolution
of the estimate. Thus, the performance of the trained networks will be evaluated in terms of these two parameters.

Fig. 4.3 Parameters associated with an ITD distribution peak.

4.3.2 Statistical Evaluation of the Trained Networks

Since white noise is added to the stimuli, the performance of the networks is stochastic in nature, i.e. there will be variation in the resulting cue estimates for stimuli with similar characteristics. Also, since the networks are only trained to learn a limited number of training examples, there will be variation in the cue estimates for different stimuli that have the same cue value. Furthermore, there are an unlimited number of possible stimuli, and we cannot test all of them. Thus, the performance of the trained networks should be evaluated from a statistical point of view.
Specifically, in a typical test presented in the following chapters, a random sample of 120 stimuli is first selected from the entire assembly of all possible stimuli that are relevant to the test and have the same specific cue value. The network under examination is then tested with all 120 stimuli in the random sample, and the sample mean ($\bar{X}$) and standard deviation ($S$) of the cue estimates resulting from the testing stimuli are calculated. According to the central limit theorem of statistics (Romano, 1977), this sample mean ($\bar{X}$) has an approximate normal distribution. The 95% confidence interval for the true mean of the cue estimates is between about $\bar{X} - 0.18S$ and $\bar{X} + 0.18S$, where $\bar{X}$ and $S$ are the sample mean and standard deviation, respectively.

4.3.3 Generalization Ability of the Networks

As mentioned in Subsection 4.2.1, only a limited number of the simplest type stimuli are used as the training examples. An important evaluation of the trained networks is, then, to test how well the networks work with stimuli not seen in the training example sets. A first test for this purpose is to examine the performance of the networks for stimuli that are of the same type as those used in training, but have different stimulus parameters. Tests with other representative types of sound stimuli are also necessary. For the ITD and IID networks, a second test is to examine the networks with stimuli that contain two frequency components. The purpose of this test is to assess the ability of the networks to "generalize", and how well they will perform with stimuli that have discrete frequency spectra. For the IED network, sinusoidal signals modulated with two-tone-complex waveforms are used for the corresponding test. A third test for the ITD and IID networks is to examine the networks with stimuli that have continuous frequency spectra. Band-pass filtered noise (pink noise) samples are used in this test. Since a broad-band stimulus is subject to band-pass filtering by the
cochlea, the test results with pink noise stimuli provide us with some concrete ideas about how well the networks work for more general stimuli than pink noise.

4.3.4 Robustness Test of the Networks

Robustness of the cue estimates generated from the trained networks is an important aspect of performance because one of our goal is to develop models that can localize sound sources in noisy situations. Thus, in all the tests discussed in the last subsection, both low and high signal-noise ratios (SNRs) are used in order to obtain an idea of how the performance of the networks changes from high SNRs to low SNRs.

4.3.5 Testing in Multi-Source Situations

In the tests described in the last two subsections, we are concerned with how accurate and robust is the cue estimation made by the trained networks. Thus, only one target sound source are used in those tests. As discussed in Chapter 2, an important task for our model is to localize multiple sound sources. In this section, we discuss the testing methods used to assess the ability of the networks to localize multiple sound sources.

A. A Model for Multi-Source Environments

It is an important characteristics of natural acoustic environments that different sources appear at different times and have time-varying acoustic powers. The concept of the DLI modeling scheme (see Fig. 3.1) is proposed to take advantage of these characteristics in order to resolve multiple sound sources that have similar frequency contents.

To emphasize the above temporal characteristics of an acoustic environment, we model the environment to be having \( n \) independent sound sources and add white noise to the signals reaching the two ears. The target sources may have the same or similar
frequency contents, but have time varying intensities. In mathematical terms, the two signals detected by the two ears have the following general form:

\[ s_L(t) = S_1(t)u_1(t) + S_2(t)u_2(t) + \cdots + S_n(t)u_n(t) + w_L(t) \]  
\[ s_R(t) = S_1(t - \tau_1)B_1u_1(t - \tau_1) + S_2(t - \tau_2)B_2u_2(t - \tau_2) + \cdots + S_n(t - \tau_n)B_nu_n(t - \tau_n) + w_R(t) \]

where \( S_i(t), i = 1, \cdots, n \), describe the time-varying modulation of the \( n \) sources \( u_i(t), i = 1, \cdots, n \). \( B_i, i = 1, \cdots, n \), reflect the IIDs, and \( \tau_i, i = 1, \cdots, n \), are the ITDs. \( w_L(t) \) and \( w_R(t) \) are independent white noise samples.

**B. Orthogonal On-Off Modulations of Different Sound Sources**

To test the trained networks for the separation of the directions of multiple sound sources, we first consider an idealized situation where there are two target sound sources from two different directions. The intensities of the two sounds are on-off modulated with their modulation functions being square-waves that are orthogonal to each other. In this case, the sound signals detected by the two ears can be described as follows:

\[ s_L(t) = S_1(t)u_1(t) + S_2(t)u_2(t) + w_L(t) \]  
\[ s_R(t) = S_1(t - \tau_1)B_1u_1(t - \tau_1) + S_2(t - \tau_2)B_2u_2(t - \tau_2) + w_R(t) \]

where \( u_1(t) \) and \( u_2(t) \) are the sound waveforms of the two target sources, \( \tau_1 \) and \( \tau_2 \) are two time delays, \( B_1 \) and \( B_2 \) are two interaural amplitude ratios corresponding to the two sources, and \( S_1(t) \) and \( S_2(t) \) are two orthogonal square-wave signals that satisfy the following equation:

\[ S_2(t) = S_1(t - p/2) \]

where \( p \) is the period of the square-wave signals. Fig. 4.4 shows examples of the two square-wave signals in Eq. 4.13.
In the actual simulation, we drop the time delay terms in the on-off modulation signals. By doing so, the only cue that make the separation of the two sources possible is the differential on-off modulations of the two sources. The two signals at the two ears then become:

\[
\begin{align*}
    s_L(t) &= S_1(t)u_1(t) + S_2(t)u_2(t) + w_L(t) \\
    s_R(t) &= S_1(t)B_1u_1(t - \tau_1) + S_2(t)B_2u_2(t - \tau_2) + w_R(t)
\end{align*}
\]  

(4.14)  

(4.15)

If, in one extreme case of Eq. 4.14 and Eq. 4.15, the rate of the on-off modulation is so high that the two signals from the different directions are mixed in most of the short-time windows in which the short-time cues are estimated, the two sources are not separable by the networks. If, in another extreme case of Eq. 4.14 and Eq. 4.15, the rate of the modulation is so slow that in the entire listening period there is only one source is turned on, then the situation degenerate to a single target source case.
Thus, an important parameter that will affect the results of the networks is the on-off period of the modulation functions.

C. Resolution Power for the Separation of Two Sound Sources

Another important aspect of the performance of the networks is their resolution power in separating the directions of two closely placed targets. Although the width parameter corresponding to a cue distribution peak gives us a measure of the resolution for single target cases, it is not clear how close two target sources may be placed before the distribution peaks that correspond to the two targets merge into one. This is tested in situations where two orthogonal on-off modulated tonal sources are separated by different spatial distances.

D. Uncorrelated On-Off Modulations of Different Sound Sources

In the test described in B earlier this subsection, the two target sources are completely temporally segregated by means of orthogonal on-off modulations. However, in more realistic situations, different sound sources are not correlated to each other. To model a more general situation, the networks are tested with two independently on-off modulated target sources. More specifically, the on-off modulation intervals of the two sources have random lengths. The random on-off modulation of the two targets are not correlated with each other. In this case, at any small interval of the listening period, there are four possible combination of the two target signals: (i) target A is on, but target B is off; (ii) target A is off, but target B is on; (iii) both targets are on; and finally (iv) both targets are off. Thus, in effect, the two targets are only partly segregated. This test is designed to see if the networks are still able to pick up the two spatially separated targets in the first two (i and ii) combinations.
### 4.3.6 Summary Table of Tests for the Evaluation of the Networks

Table 4.1 summarizes the planned tests described in this section.

**Table 4.1** Networks (ITD, IID, IED) to be tested for the listed stimulus type and parameter combinations.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Low SNR (0 or 5 dB)</th>
<th>High SNR (20 dB)</th>
<th>Frequency difference</th>
<th>Center frequency</th>
<th>Bandwidth</th>
<th>Spatial separation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure tone</td>
<td>ITD, IID</td>
<td>ITD, IID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-tone complex</td>
<td>ITD, IID</td>
<td>ITD, IID</td>
<td>ITD, IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pink noise</td>
<td>ITD, IID</td>
<td>ITD, IID</td>
<td></td>
<td>ITD, IID</td>
<td>ITD, IID</td>
<td></td>
</tr>
<tr>
<td>AM stimuli</td>
<td>IED</td>
<td>IED</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-tone complex as modulation signal</td>
<td>IED</td>
<td>IED</td>
<td>IED</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthogonal two-source case</td>
<td>ITD, IID, IED</td>
<td>ITD, IID, IED</td>
<td></td>
<td>ITD, IID</td>
<td>IED</td>
<td></td>
</tr>
<tr>
<td>Random modulation two-source case</td>
<td>ITD, IID, IED</td>
<td>ITD, IID, IED</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5 Test Results of ITD Estimation

As mentioned in Subsection 4.2.1, three networks (ITD, IID, and IED), each in one of the three implementations (Fig. 3.2) of the DLI model shown in Fig. 3.1, were trained and evaluated following the simulation methods described in the last chapter. In this chapter, we present the test results of the trained ITD network. Results for the trained IID and IED networks will be presented in Chapters 6 and 7, respectively.

§5.1 Pure-Tone Stimuli

Following the evaluation strategy described in Section 4.3, the ITD network was first tested with pure-tone stimuli having fixed values for the ITD and signal-noise ratio (SNR). A random sample of 120 pure-tone stimuli that have a certain ITD value was selected to test the estimation accuracy for that particular ITD value. Several parameters, frequency, intensity, phase, and IID, are particularly relevant to the random sampling of stimuli. The intensities of the test stimuli were randomly selected from the range of 30-60 dB SPL, which covers a moderate intensity range. For lower intensity values, our model generated ITD distribution functions that did not show clear peaks, indicating that the model was not stimulated sufficiently to give out locations of the sound sources, if any. The IIDs of the test stimuli were randomly selected from the range of -9 to 9 dB. The phases of the pure-tones are randomly selected from the range of 0 to 2π. The frequencies were randomly selected from the range of 800-1000 Hz, which covers the characteristic frequencies of the auditory nerve fibers at the input of the ITD network. Since the neural signals of the auditory nerve fibers are modeled as outputs of a process that involves narrow band-pass filters, a sound stimulus with its frequency outside the range adjacent to the characteristic frequencies of the fibers will be much attenuated such that the ITD
network will not be stimulated sufficiently to give out valid ITD estimates. As discussed in Subsection 4.2.2, the range of ITD values encoded by the output layer neurons of the ITD network is from -505 μs to 505 μs. We used a 250 μs ITD in the test, which is an intermediate value between 0 second and the upper limit (505 μs) of the encoded ITD range. White noise was added to all test stimuli, resulting in a 20 dB SNR.

ITD estimates (in terms of the position and width defined in Subsection 4.3.1) were obtained for the above described random sample of stimuli, and the mean and standard deviation of these estimates were calculated, as shown in Table 5.1. It is clear that accurate ITD estimates were obtained.

Table 5.1 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 250 μs ITD and a 20 dB SNR.

<table>
<thead>
<tr>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.6±10.4</td>
<td>42.9±16.7</td>
<td>0.6</td>
</tr>
</tbody>
</table>

The robustness of the ITD estimates was tested by repeating the test shown in Table 5.1 using a 0 dB SNR. All other aspects of the test remain unchanged. Table 5.2 shows the results.

Table 5.2 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 250 μs ITD and a 0 dB SNR.

<table>
<thead>
<tr>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>250.4±10.6</td>
<td>86.9±35.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Data in Table 5.2 are similar to that in Table 5.1 except for a decrease (50 μs versus 100 μs) in the resolution of the estimates. This indicates that the ITD estimates are robust in the presence of noise for pure-tone stimuli.

The test shown in Table 5.1 was repeated for a small range of ITD values around 0 second. The results are tabulated in Table 5.3, where the first column shows the tested ITD values, and each row shows the characteristics of the corresponding ITD estimates. The numbers after "±" are standard deviations of the corresponding estimates.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>-300</td>
<td>-304.0 ± 35.1</td>
<td>60.0 ± 23.6</td>
<td>-4.0</td>
</tr>
<tr>
<td>-200</td>
<td>-194.2 ± 24.6</td>
<td>55.3 ± 20.0</td>
<td>5.8</td>
</tr>
<tr>
<td>-100</td>
<td>-91.6 ± 10.5</td>
<td>53.5 ± 19.5</td>
<td>8.4</td>
</tr>
<tr>
<td>0</td>
<td>12.9 ± 19.1</td>
<td>41.3 ± 20.4</td>
<td>12.9</td>
</tr>
<tr>
<td>100</td>
<td>106.8 ± 17.3</td>
<td>51.1 ± 20.9</td>
<td>6.8</td>
</tr>
<tr>
<td>200</td>
<td>206.0 ± 11.1</td>
<td>48.4 ± 18.3</td>
<td>6.0</td>
</tr>
<tr>
<td>300</td>
<td>294.1 ± 8.7</td>
<td>46.9 ± 17.7</td>
<td>-5.9</td>
</tr>
</tbody>
</table>

The estimated ITDs (in the second column of Table 5.3) are plotted as a function of the true ITDs (in the first column of Table 5.3), as shown in Fig. 5.1. These results clearly show that the trained ITD network is able to produce accurate ITD estimates with high resolutions.
Fig. 5.1 ITD estimates shown as a function of the true ITDs.

§5.2 Two-Tone Complex Stimuli

Next, the network was tested with stimuli that have two frequency components. Again, a random sample of stimuli was used. An important parameter of this type of stimuli is the frequency spacing of the two components in the stimuli. Thus, we first conducted a test using a random sample of stimuli that have a fixed frequency spacing value. Specifically, one frequency component was randomly selected from the range of 700-1100 Hz while the other was chosen to have a 100 Hz difference from the first component. The intensities of the stimuli in the sample were randomly selected from the range of 30-60 dB SPL. The IIDs of the stimuli were randomly selected from the range of -9 to 9 dB. White noise was added to the stimuli, resulting in a 20 dB SNR. The ITD of the stimuli is 250 μs. Table 5.4 shows the test results.
Table 5.4 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 250 μs ITD and a 20 dB SNR.

<table>
<thead>
<tr>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>246.3±17.3</td>
<td>45.2±15.7</td>
<td>-3.7</td>
</tr>
</tbody>
</table>

The robustness of the ITD estimation for two-tone complex stimuli was tested by repeating the test shown in Table 5.4 with a 0 dB SNR. The results are shown in Table 5.5.

Table 5.5 ITD estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 250 μs ITD and a 0 dB SNR.

<table>
<thead>
<tr>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>239.4±19.6</td>
<td>93.2±41.1</td>
<td>-10.6</td>
</tr>
</tbody>
</table>

Data shown in Tables 5.4 and 5.5 are similar to the corresponding data (shown in Tables 5.1 and 5.2) for the pure-tone case presented in the last section. This demonstrates the ability of the ITD network to "generalize" and perform well with two-tone complex stimuli. Furthermore, although there is a two fold decrease in the resolution of the ITD estimates when the SNR is reduced from 20 dB (Table 5.4) to 0 dB (Table 5.5), the otherwise comparable data between Table 5.4 and Table 5.5 indicate that the ITD estimates are robust in the presence of noise for two-tone complexes.

The effect of the frequency spacing of the two components in the stimuli on the performance of the ITD network was studied by repeating the test shown in Table 5.4 for different values of the frequency difference between the two components. The results are shown in Table 5.6, where the first column shows the tested frequency
differences, and the subsequent columns show the corresponding ITD estimates. The true ITD in the test stimuli was 250 μs.

Table 5.6 ITD estimates and the errors of the mean position estimates for two-tone complex stimuli with a 250 μs ITD and a 20 dB SNR. FD: frequency difference between the two components in the stimuli.

<table>
<thead>
<tr>
<th>FD: Δf (Hz)</th>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>247.2 ± 21.9</td>
<td>47.8 ± 24.4</td>
<td>-2.8</td>
</tr>
<tr>
<td>50</td>
<td>246.3 ± 20.6</td>
<td>45.7 ± 19.6</td>
<td>-3.7</td>
</tr>
<tr>
<td>100</td>
<td>246.3 ± 17.3</td>
<td>45.2 ± 15.7</td>
<td>-3.7</td>
</tr>
<tr>
<td>150</td>
<td>244.5 ± 15.2</td>
<td>47.0 ± 16.5</td>
<td>-5.5</td>
</tr>
<tr>
<td>200</td>
<td>238.6 ± 21.0</td>
<td>51.7 ± 22.9</td>
<td>-11.4</td>
</tr>
<tr>
<td>250</td>
<td>234.6 ± 19.4</td>
<td>57.8 ± 28.4</td>
<td>-15.4</td>
</tr>
</tbody>
</table>

Data in Table 5.6 indicate that the frequency spacing between the two components in the stimuli has little effect on the estimation results.

Analysis of the Input Patterns of the ITD Network for Two-Tone Stimuli

We have seen that although the ITD network was trained by using pure-tone stimuli only, it also works with two-tone complex stimuli. We wondered why this was the case. As the ITD is measured by comparing input patterns to the network, these patterns may hold the key to the answer. With a single tone stimulus, the activities of the auditory nerve fibers form line patterns (see Fig. 3.3) which provide references for the estimation of the ITDs. What would the input patterns look like if the stimulus contains two sinusoidal components?

Consider a stimulus of the following form:
\[ y(t) = A \sin(\omega_1 t + \varphi_1) + B \sin(\omega_2 t + \varphi_2) \quad (5.1) \]

It can be re-written in the following forms

\[
y(t) = C \left[ \sin \theta \sin(\omega_1 t + \varphi_1) + \cos \theta \sin(\omega_2 t + \varphi_2) \right]
= \frac{C}{2} \left[ \cos(\omega_1 t + \varphi_1 - \theta) - \cos(\omega_1 t + \varphi_1 + \theta) + \right.
+ \sin(\omega_2 t + \varphi_2 + \theta) + \sin(\omega_2 t + \varphi_2 - \theta) \bigg]
\]

\[
= \frac{C}{2} \left[ \sin(\omega_1 t + \varphi_1 - \frac{\pi}{2}) - \sin(\omega_1 t + \varphi_1 + \frac{\pi}{2}) \right]
+ \left[ \sin(\omega_2 t + \varphi_2 + \frac{\pi}{2}) + \sin(\omega_2 t + \varphi_2 - \frac{\pi}{2}) \right]
\]

\[
= C \left[ \sin \left( \frac{\omega_1 + \omega_2}{2} \right) t + \frac{\varphi_1 + \varphi_2}{2} + \frac{\pi}{4} \right] \cos \left( \frac{\omega_1 - \omega_2}{2} \right) t + \frac{\varphi_1 - \varphi_2}{2} - \theta + \frac{\pi}{4} \bigg]
+ \left[ \sin \left( \frac{\omega_1 + \omega_2}{2} \right) t + \frac{\varphi_1 + \varphi_2}{2} - \frac{\pi}{4} \right] \cos \left( \frac{\omega_1 - \omega_2}{2} \right) t + \frac{\varphi_1 - \varphi_2}{2} + \theta - \frac{\pi}{4} \bigg]
\]

\[
(5.2)
\]

where \( C = \sqrt{A^2 + B^2} \), and \( \theta = \arcsin(A/C), -\pi/2 < \theta < \pi/2 \). From Eq. 5.2, it is clear that the stimulus is a linear combination of two amplitude-modulated tones which have the same carrier frequency \((\omega_1 + \omega_2)/2\) and modulating frequency \((\omega_1 - \omega_2)/2\).

In addition, the carrier tones are \( \pi/2 \) out of phase.

The stimulus can be analyzed in three cases: (i) the two frequencies \( \omega_1 \) and \( \omega_2 \) in Eq. 5.1 are very close to each other; (ii) the two frequencies are very far from each other; and (iii) the two frequencies have an intermediate distance in the frequency domain. In the first case, as \( \omega_1 \) and \( \omega_2 \) are very close, the modulating frequency in Eq. 5.2 is very small. Therefore, the amplitude-modulated tones in Eq. 5.2 can be simplified to two pure-tones of the same frequency in a short time window. Thus, over short time windows, Eq. 5.2 is reduced to

\[
y(t) = A \sin \omega t - B \cos \omega t = C \cos(\omega t + \theta) \quad (5.3)
\]

which is a single-tone stimulus.
In the second case, where $\omega_1$ and $\omega_2$ are very far from each other, the band-pass filtering effect of the basilar membrane will filter out the frequency component that is further away from the center frequency of the range covered by the ITD network. Thus, this case is also reduced to the single-tone stimulus case.

The last case, where $\omega_1$ and $\omega_2$ have an intermediate distance, is somewhat more complicated than the first two. In this case, both tones in Eq. 5.2 are amplitude-modulated. The highest modulation rate occurs when $\omega_1$ and $\omega_2$ have the largest distance, but none of them are filtered out by the model auditory nerve fibers that activate the ITD network. However, due to the fact that the ITDs are estimated in very short time windows, as long as the modulation rate is slow in the sense that the amplitude of the carrier tones does not change much in these short time windows, the two-tone case can still be reduced to the single-tone case. In fact, the activation of the auditory nerve fibers for the worst cases of two-tone stimuli still shows straight patterns, as shown in Fig. 5.2.

![Fig. 5.2 Example input activation patterns of the ITD network for a two-tone complex stimulus whose two components have a 250 Hz difference.](image)
§5.3 Pink Noise Stimuli

Following the evaluation strategy discussed in Section 4.3, the trained ITD network was also tested with pink noise stimuli. A pink noise is characterized by its bandwidth and center frequency. Two series of tests were conducted to assess the effects of these two parameters on the performance of the ITD network. In both series of tests, the bandwidths of the pink noise tested are no larger than 150 Hz. Tests with wider bandwidths are not necessary because (i) the characteristic frequencies of the input auditory nerve fibers to the ITD network cover a narrow band of frequencies, and (ii) the models of the auditory nerve fibers are band-pass filters. If the bandwidth of a stimulus is significantly wider than the frequency band covered by the ITD network, the energy outside the covered band will be filtered out by the group of auditory nerve fibers at the input to the ITD network.

In the first series of tests, the effect of the bandwidth is studied. More specifically, in each experiment in the series, a random sample of 120 pink noise stimuli were selected. The bandwidth of these stimuli was the same, but other stimulus parameters were randomly selected from uniform distributions. The ITD estimates for these stimuli were then obtained. The mean and standard deviation of these ITD estimates (in terms of the corresponding position and width parameters) were tabulated in Table 5.7, where the first column shows the tested values of the stimulus bandwidth. In the results presented in Table 5.7, the center frequencies of the stimuli were randomly selected from the range of 800-1000 Hz. To make sure that the stimuli contain energy in the frequency range covered by the ITD network, we did not use frequencies outside this range. If this is not the case, the ITD network will not be excited sufficiently by the stimuli, and will not be able to tell which directions the stimuli are coming from. The intensities of the stimuli were randomly selected from the range of 30-60 dB. The IIDs were randomly selected from the
range from -9 to 9 dB. All test stimuli have a 250 µs ITD, with white noise added, resulting in a 20 dB SNR.

Data shown in Table 5.7 are similar to that obtained in the pure-tone case (Table 5.1), demonstrating the ability of the network to “generalize” and perform well with pink noise stimuli. Furthermore, the bandwidth of the pink noise stimuli shows no effect on the performance of the network.

The robustness of the ITD estimates for pink noise stimuli was tested by repeating the tests presented in Table 5.7 using stimuli that had a 50 Hz bandwidth and on which white noise was added so that the SNR was 0 dB. The results are shown in Table 5.8. Again, except for a two fold decrease in resolution, the estimates shown in Table 5.8 are similar to that obtained in the higher (20 dB) SNR case. Thus, the ITD estimates for pink noise stimuli are also robust in the presence of white noise.

Table 5.7  ITD estimates and the errors of the mean position estimates for pink noise stimuli with different bandwidths (BW). The SNR used in the tests was 20 dB, and the ITD of the stimuli was 250 µs.

<table>
<thead>
<tr>
<th>BW: w (Hz)</th>
<th>Position: p (µs)</th>
<th>Width: d (µs)</th>
<th>Position Error (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>249.6 ± 11.8</td>
<td>51.4 ± 22.0</td>
<td>-0.4</td>
</tr>
<tr>
<td>50</td>
<td>250.7 ± 10.2</td>
<td>50.7 ± 17.7</td>
<td>0.7</td>
</tr>
<tr>
<td>100</td>
<td>251.0 ± 8.9</td>
<td>49.2 ± 15.3</td>
<td>1.0</td>
</tr>
<tr>
<td>150</td>
<td>248.3 ± 9.0</td>
<td>48.7 ± 13.3</td>
<td>-1.7</td>
</tr>
</tbody>
</table>

Parallel to the first series of experiments presented in Table 5.7, in the second series of experiments, we tested the effect of the center frequencies of pink noise stimuli on the performance of the ITD network. In different experiments in the series, a different center frequency was tested so that a function of performance versus stimulus center frequency was obtained. In each experiment in the series, a random
sample of 120 stimuli that had the same center frequency was first selected, and the ITD estimates for these stimuli were obtained. The mean and the standard deviation of these ITD estimates were then calculated. In all of the test stimuli in the random samples used in the series of experiments, the bandwidths were randomly selected from the range of 25-150 Hz. The intensities were randomly selected from the range of 30-60 dB. The IIDs were randomly selected from the range from -9 to 9 dB. White noise was added to the stimuli, resulting a 20 dB SNR. The ITD of the stimuli is 250 μs. The test results are tabulated in Table 5.9, where the first column shows the different center frequencies tested in different experiments in the series. These results show that the center frequency of a pink noise stimulus has little effect on the performance of the ITD network.

Table 5.8 ITD estimates and the error of the mean position estimate for pink noise stimuli with a 0 dB SNR and a 250 μs ITD. The bandwidth of the stimuli was 50 Hz.

<table>
<thead>
<tr>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>235.7±21.5</td>
<td>84.8±31.3</td>
<td>-14.3</td>
</tr>
</tbody>
</table>

§5.4 Two Similar Sources from Different Directions

Following the testing strategy outlined in Subsection 4.3.5, the ITD network was also tested in cases where there were two sound sources. A first test was carried out using two pure-tone sources that were periodically on-off modulated with orthogonal modulation functions (Subsection 4.3.5 B). The two pure-tones had the same 50 dB SPL intensity and the same 900 Hz frequency. The IID for both tone sources was 0 dB. The ITDs for the two sources were -250 μs and 250 μs, respectively. White noise was added to the signals from the two targets, resulting in a 20 dB SNR.
Different on-off modulation periods were tested, as shown in the first column of Table 5.10. There are two peaks in each of the ITD distribution functions generated in the tests, as shown in Fig. 5.3 where the ITD distribution function that corresponds to a 25 ms on-off modulation period is plotted. The position and width parameters associated with the two peaks are listed in the second through fifth columns of Table 5.10. These data demonstrate that correct estimates of the two ITDs (-250 and 250 μs) associated with the two pure tone sources were obtained. We found that the smallest on-off modulation period at which the ITD network could still separate the two sources was 15 ms.

<table>
<thead>
<tr>
<th>CF: f (Hz)</th>
<th>Position: p (μs)</th>
<th>Width: d (μs)</th>
<th>Position Error (μs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>239.9 ± 11.8</td>
<td>70.9 ± 28.2</td>
<td>-10.1</td>
</tr>
<tr>
<td>850</td>
<td>249.4 ± 8.9</td>
<td>52.7 ± 13.0</td>
<td>-0.6</td>
</tr>
<tr>
<td>900</td>
<td>252.2 ± 7.8</td>
<td>48.7 ± 11.5</td>
<td>2.2</td>
</tr>
<tr>
<td>950</td>
<td>251.5 ± 7.0</td>
<td>44.4 ± 10.1</td>
<td>1.5</td>
</tr>
<tr>
<td>1000</td>
<td>254.4 ± 8.6</td>
<td>43.3 ± 11.0</td>
<td>4.4</td>
</tr>
</tbody>
</table>

The robustness of ITD network to separate two pure-tone sources was tested by repeating the test shown in Table 5.10 with a 0 dB SNR. Only the smallest (15 ms) on-off modulation period was used. The results, which is similar to that shown in Table 5.10, is shown in Table 5.11. The comparable results between the two tests shown in Tables 5.10 and 5.11 demonstrate the robust ability of the network to separate two similar sources originating form different directions.
Fig. 5.3 An example ITD distribution function for a two-source case with a 25 ms on-off modulation period.

Table 5.10 ITD estimates and the position errors for two-source cases with different on-off modulation periods (MP). The SNR was 20 dB, and the ITDs of the two sources were -250 μs and 250 μs, respectively.

<table>
<thead>
<tr>
<th align="center">MP (ms)</th>
<th align="center">First Peak</th>
<th align="center"></th>
<th align="center"></th>
<th align="center">Second Peak</th>
<th align="center"></th>
<th align="center"></th>
</tr>
</thead>
<tbody>
<tr>
<td align="center"></td>
<td align="center">Position (μs)</td>
<td align="center">Width (μs)</td>
<td align="center">Position Error (μs)</td>
<td align="center">Position (μs)</td>
<td align="center">Width (μs)</td>
<td align="center">Position Error (μs)</td>
</tr>
<tr>
<td align="center">15</td>
<td align="center">-288.6</td>
<td align="center">101.0</td>
<td align="center">-38.6</td>
<td align="center">274.8</td>
<td align="center">50.5</td>
<td align="center">24.8</td>
</tr>
<tr>
<td align="center">20</td>
<td align="center">-259.8</td>
<td align="center">67.3</td>
<td align="center">-9.8</td>
<td align="center">244.7</td>
<td align="center">33.7</td>
<td align="center">-5.3</td>
</tr>
<tr>
<td align="center">25</td>
<td align="center">-236.3</td>
<td align="center">50.5</td>
<td align="center">13.7</td>
<td align="center">236.5</td>
<td align="center">33.7</td>
<td align="center">-13.5</td>
</tr>
<tr>
<td align="center">30</td>
<td align="center">-234.7</td>
<td align="center">84.2</td>
<td align="center">15.3</td>
<td align="center">236.6</td>
<td align="center">50.5</td>
<td align="center">-13.4</td>
</tr>
<tr>
<td align="center">35</td>
<td align="center">-250.0</td>
<td align="center">67.3</td>
<td align="center">0.0</td>
<td align="center">257.3</td>
<td align="center">33.7</td>
<td align="center">7.3</td>
</tr>
</tbody>
</table>
Table 5.11 ITD estimates and the position errors for a two-source case with a 0 dB SNR. The ITDs of the two sources were -250 μs and 250 μs, respectively.

<table>
<thead>
<tr>
<th>MP (ms)</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (μs)</td>
<td>Width (μs)</td>
</tr>
<tr>
<td>15</td>
<td>-289.4</td>
<td>117.8</td>
</tr>
</tbody>
</table>

The resolution power of the network to separate the directions of two sources was tested by repeating the test in Table 5.10 for different spatial separations of the two pure-tone sources. A 15 ms on-off modulation period was used in the test. Table 5.12 shows the results, where the first column lists the spatial separations (in terms of the two ITD values) of two pure-tone sources. The smallest ITD difference was found to be about 200 μs.

Table 5.12 ITD estimates and the position errors for two-source cases with different spatial separations (SS, in terms of the two ITDs). The SNR was 20 dB.

<table>
<thead>
<tr>
<th>SS (μs)</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (μs)</td>
<td>Width (μs)</td>
</tr>
<tr>
<td>(-100, 100)</td>
<td>-80.1</td>
<td>84.2</td>
</tr>
<tr>
<td>(-125, 125)</td>
<td>-118.1</td>
<td>50.5</td>
</tr>
<tr>
<td>(-188, 188)</td>
<td>-180.4</td>
<td>50.5</td>
</tr>
<tr>
<td>(-250, 250)</td>
<td>-288.6</td>
<td>101.0</td>
</tr>
</tbody>
</table>
Finally, as described in Subsection 4.3.5 D, the network was tested in a case that has two uncorrelated pure-tone targets. The two targets were also on-off modulated, but with on-off intervals of random lengths selected from the range of \([0, 35]\) ms. The two tones had the same 50 dB SPL intensity and the same 900 Hz frequency. The IID of both sources was 0 dB. The ITDs of the two sources were -250 and 250 \(\mu s\), respectively. White noise was added to the signals from the two targets, resulting in a 20 dB SNR. Table 5.13 shows the test results. The corresponding ITD distribution histogram is shown in Fig. 5.4. Two peaks in this distribution function are identifiable that correspond to the two targets. This indicates that the network is able to pick up the individual targets even when they are not completely segregated. The estimated parameters (position and width) associated with the two peaks are comparable to that for the corresponding orthogonal two-target case presented earlier this section (Table 5.10). The unequal sizes of the “humps” corresponding to the two peaks are a reflection of the unequal “show-up” times of the two targets, which in turn are results of the independent random on-off modulations of the two targets. The third (middle) peak in the distribution function shown in Fig. 5.4 corresponds to short-time ITD estimates obtained in time periods during which the two sources are both present.

<table>
<thead>
<tr>
<th>True ITDs ((\mu s))</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Width</td>
</tr>
<tr>
<td>((-250, 250))</td>
<td>-247.7</td>
<td>67.3</td>
</tr>
</tbody>
</table>
Fig. 5.4 The ITD distribution histogram for the uncorrelated two-source case. The parameters that describe the two target sources are shown in Table 5.13.
6 Test Results of IID Estimation

In this chapter, we present parallel test results of the trained IID network to that of the ITD network presented in the last chapter. Again, the evaluation methods for the IID network, which are the same as that for the ITD network, are outlined in Section 4.3, and will not be repeated here. Only aspects that are unique to the IID network will be discussed.

§6.1 Pure-Tone Stimuli

The IID network was first tested with pure-tone stimuli. A random sample of 120 pure-tone stimuli was used in the test. Parameters relevant to the random sampling of the stimuli are the frequency, intensity, phase, and ITD. The intensities of the test stimuli were randomly selected from the range of 30-60 dB SPL. The ITDs were randomly selected from the range of -1.64 to 1.64 ms. The phases of the pure-tones are randomly selected from the range of 0 to 2\pi. The frequencies were randomly selected from the range of 2500-3500 Hz, which covers the characteristic frequencies of the auditory nerve fibers at the input of the IID network. A 3 dB IID was used in the test. With 10-12 dB IIDs corresponding to sound images completely lateralized to one ear (see Figs. 1.7 and 1.8), a 3 dB IID corresponds to a sound image perceived at a position slightly to the side. White noise was added to all sample stimuli, resulting in a 20 dB SNR. The resulting IID estimates (in terms of the associated “position” and “width” parameters defined in Subsection 4.3.1) were obtained for the above described random sample of stimuli. The mean and standard deviation of these estimates are shown in Table 6.1.

The robustness of the IID estimates was tested by repeating the test shown in Table 6.1 using a 5 dB SNR. Table 6.2 shows the results.
Despite the poorer resolution of the IID estimates than that for the 20 dB SNR case (Table 6.1), the network gave out similar IID estimates to that shown in Table 6.1, indicating that the IID estimates are robust in the presence of noise for pure-tone stimuli.

Table 6.1  IID estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 3 dB IID and a 20 dB SNR.

<table>
<thead>
<tr>
<th>Position: $p$ (dB)</th>
<th>Width: $d$ (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9±0.87</td>
<td>0.8±0.27</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 6.2  IID estimates (mean ± standard deviation) and the error of the mean position estimate for pure-tone stimuli with a 3 dB IID and a 5 dB SNR.

<table>
<thead>
<tr>
<th>Position: $p$ (dB)</th>
<th>Width: $d$ (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.71±0.92</td>
<td>3.56±1.86</td>
<td>0.71</td>
</tr>
</tbody>
</table>

The test shown in Table 6.1 was repeated for nine IID values in the range of -12 to 12 dB. The results are tabulated in Table 6.3, where the first column shows the tested IID values.

The estimated IIDs (in the second column in Table 6.3) are plotted as a function of the true IIDs (in the first column in Table 6.3), as shown in Fig. 6.1. This clearly shows that the IID network is able to produce accurate IID estimates.
Table 6.3 IID estimates and the errors of the mean position estimates for pure-tone stimuli with different IID\(\Delta I\)s and a 20 dB SNR.

<table>
<thead>
<tr>
<th>IID: (\Delta I) (dB)</th>
<th>Position: (p) (dB)</th>
<th>Width: (d) (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-12</td>
<td>-11 ± 1.6</td>
<td>0.51 ± 0.21</td>
<td>1</td>
</tr>
<tr>
<td>-9</td>
<td>-9.7 ± 1.9</td>
<td>0.78 ± 0.42</td>
<td>-0.7</td>
</tr>
<tr>
<td>-6</td>
<td>-6.8 ± 1.6</td>
<td>0.75 ± 0.44</td>
<td>-0.8</td>
</tr>
<tr>
<td>-3</td>
<td>-3.6 ± 1.2</td>
<td>0.75 ± 0.33</td>
<td>-0.6</td>
</tr>
<tr>
<td>0</td>
<td>0.26 ± 1.1</td>
<td>0.66 ± 0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>3</td>
<td>3.9 ± 0.87</td>
<td>0.8 ± 0.27</td>
<td>0.9</td>
</tr>
<tr>
<td>6</td>
<td>7.5 ± 1.7</td>
<td>0.66 ± 0.22</td>
<td>1.5</td>
</tr>
<tr>
<td>9</td>
<td>10 ± 0.83</td>
<td>0.65 ± 0.36</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>11 ± 2.2</td>
<td>0.46 ± 0.25</td>
<td>-1</td>
</tr>
</tbody>
</table>

§6.2 Two-Tone Complex Stimuli

The network was then tested with stimuli that have two frequency components. We first conducted a test using a random sample of stimuli that have a fixed frequency difference between the two components in the stimuli. Specifically, the two frequency components in a test stimulus were randomly selected from the range of 2000-4000 Hz except that the difference between the two components was always 500 Hz. The intensities of the stimuli were randomly selected from the range of 30-60 dB SPL. The ITDs were randomly selected from the range of -1.64 to 1.64 ms. White noise was added to the stimuli in the sample, resulting a 20 dB SNR. The IID of the test stimuli is 3 dB. Table 6.4 shows the test results.
Chapter 6  Test Results of IID Estimation

Fig. 6.1 IID estimates shown as a function of the true IID.

Table 6.4 IID estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 3 dB IID and a 20 dB SNR.

<table>
<thead>
<tr>
<th>Position: $p$ (dB)</th>
<th>Width: $d$ (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5±0.94</td>
<td>1.35±0.81</td>
<td>0.5</td>
</tr>
</tbody>
</table>

The robustness of the IID estimates for two-tone complex stimuli was tested by repeating the test shown in Table 6.4 with a 5 dB SNR. The results are shown in Table 6.5.

Data shown in Tables 6.4 and 6.5 are similar to the corresponding data (shown in Tables 6.1 and 6.2) for the pure-tone case presented in the last section. This demonstrates the ability of the IID network to “generalize” and perform well with two-tone complex stimuli. Furthermore, although the resolution of the IID estimates...
becomes poorer (4 dB versus 1.35 dB) when the SNR is reduced from 20 dB (Table 6.4) to 5 dB (Table 6.5), the network is able to give out similar "position" estimates at the two noise levels.

Table 6.5 IID estimates (mean ± standard deviation) and the error of the mean position estimate for two-tone complex stimuli with a 3 dB IID and a 5 dB SNR.

<table>
<thead>
<tr>
<th>Position: ( p ) (dB)</th>
<th>Width: ( d ) (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9±1.3</td>
<td>4±1.7</td>
<td>0.9</td>
</tr>
</tbody>
</table>

The test shown in Table 6.4 was repeated for different frequency differences between the two components in the test stimuli. The results are shown in Table 6.6, where the first column show the tested frequency differences, and the subsequent columns show the corresponding IID estimates. The true IID in the test stimuli was 3 dB.

Table 6.6 IID estimates and the errors of the mean position estimates for two-tone complex stimuli with a 3 dB IID and a 20 dB SNR. FD: frequency difference between the two components in the stimuli.

<table>
<thead>
<tr>
<th>FD: ( \Delta f ) (Hz)</th>
<th>Position: ( p ) (dB)</th>
<th>Width: ( d ) (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3.77 ± 0.97</td>
<td>1.99 ± 1.27</td>
<td>0.77</td>
</tr>
<tr>
<td>300</td>
<td>3.4 ± 1.13</td>
<td>1.28 ± 0.73</td>
<td>0.4</td>
</tr>
<tr>
<td>500</td>
<td>3.5 ± 0.94</td>
<td>1.35 ± 0.81</td>
<td>0.5</td>
</tr>
<tr>
<td>700</td>
<td>3.77 ± 0.89</td>
<td>1.02 ± 0.65</td>
<td>0.77</td>
</tr>
<tr>
<td>900</td>
<td>3.79 ± 0.72</td>
<td>0.7 ± 0.33</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Data in Table 6.6 indicate that the frequency difference between the two components of the stimuli has little effect on the estimation results.

§6.3 Pink Noise Stimuli

The network was also tested with pink noise stimuli in two series of experiments. In the first series, the effect of the stimulus bandwidth was tested. In each test in the series, the test stimuli had the same bandwidth. The center frequencies of the stimuli were randomly selected from the range of 2500-3500 Hz. The intensities were randomly selected from the range of 30-60 dB. The ITDs were randomly selected from the range of -1.64 to 1.64 ms. All test stimuli had the same 3 dB IID, with white noise added, resulting in a 20 dB SNR. The test results are tabulated in Table 6.7, where the first column shows the values of the stimulus bandwidth used.

<table>
<thead>
<tr>
<th>BW: w (Hz)</th>
<th>Position: p (dB)</th>
<th>Width: d (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>3.9 ± 1</td>
<td>3.7 ± 1.4</td>
<td>0.9</td>
</tr>
<tr>
<td>300</td>
<td>3.3 ± 0.5</td>
<td>3 ± 0.97</td>
<td>0.3</td>
</tr>
<tr>
<td>500</td>
<td>2.9 ± 0.69</td>
<td>2.9 ± 0.92</td>
<td>-0.1</td>
</tr>
<tr>
<td>700</td>
<td>2.6 ± 0.4</td>
<td>3 ± 0.9</td>
<td>-0.4</td>
</tr>
<tr>
<td>900</td>
<td>2.5 ± 0.62</td>
<td>3.1 ± 1</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

The results shown in Table 6.7 are similar to that obtained in the pure-tone case (Table 6.1), demonstrating the ability of the network to “generalize” and perform well with pink noise stimuli. Furthermore, the bandwidth of the pink noise stimuli shows
only a small effect on the performance of the network: there is a small downward shift of the estimated IID (the second column of Table 6.7) as the stimulus bandwidth increases from 100 to 900 Hz.

The robustness of the IID estimates for pink noise stimuli was tested by repeating the tests shown in Table 6.7 using stimuli that have a 500 Hz bandwidth and a lower (5 dB) SNR. The results are shown in Table 6.8. There is a two fold decrease in the resolution of the IID estimates when the SNR is reduced from 20 dB to 5 dB SNR. Otherwise, the data in Table 6.8 are similar to the corresponding data entry in Table 6.7.

In the second series of experiments with pink noise stimuli, we tested the effect of the center frequencies of stimuli on the performance of the network. Different center frequencies were used to obtain a function of performance versus frequency. In each experiment in the series, the bandwidths of the test stimuli were randomly selected from the range of 50-1000 Hz. The intensities were randomly selected from the range of 30-60 dB. The ITDs were randomly selected from the range of -1.64 to 1.64 ms. All test stimuli had the same 3 dB IID, and on which white noise was added with a 20 dB SNR. The test results are shown in Table 6.9, where the first column shows the different center frequencies tested. The center frequency of the stimuli shows little effect on the performance of the network.

Table 6.8 IID estimates (mean ± standard deviation) and the error of the mean position estimate for pink noise stimuli with a 5 dB SNR and a 3 dB IID. The bandwidth of the stimuli was 500 Hz.

<table>
<thead>
<tr>
<th>Position: $p$ (dB)</th>
<th>Width: $d$ (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.9±1.2</td>
<td>6.1±1.6</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

Chapter 6  Test Results of IID Estimation  128
Table 6.9  IID estimates and the errors of the mean position estimates for pink noise stimuli with different center frequencies (CF). The SNR used in the tests was 20 dB, and the IID of the stimuli was 3 dB.

<table>
<thead>
<tr>
<th>CF: f (Hz)</th>
<th>Position: p (dB)</th>
<th>Width: d (dB)</th>
<th>Position Error (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2500</td>
<td>2.9 ± 0.76</td>
<td>2.6 ± 0.8</td>
<td>-0.1</td>
</tr>
<tr>
<td>2750</td>
<td>4.2 ± 0.89</td>
<td>3.1 ± 1.3</td>
<td>1.2</td>
</tr>
<tr>
<td>3000</td>
<td>2.9 ± 0.53</td>
<td>2.8 ± 0.82</td>
<td>-0.1</td>
</tr>
<tr>
<td>3250</td>
<td>2.9 ± 0.74</td>
<td>2.5 ± 0.86</td>
<td>-0.1</td>
</tr>
<tr>
<td>3500</td>
<td>2.4 ± 0.38</td>
<td>3 ± 0.94</td>
<td>-0.6</td>
</tr>
</tbody>
</table>

§6.4 Two Similar Sources from Different Directions

Finally, the network was tested in cases where there were two sound sources. In the first test in this category, two pure-tone sources (both have a 50 dB SPL intensity and a 3000 Hz frequency) were periodically on-off modulated with orthogonal modulation functions. The ITD for both tone sources was 0 second. The IIDs for the two sources were -3 dB and 3 dB, respectively. The SNR used in the test was 20 dB. Different on-off modulation periods, which are listed in the first column of Table 6.10, were tested. The test results are listed in the other columns of the table. Two peaks in the IID distribution functions generated in the tests are identifiable which correspond to the IIDs of the two sources. Fig. 6.2 shows such an IID distribution function resulted from the test with a 35 ms on-off modulation period. We found that the smallest on-off modulation period for two-source separation was about 20 ms.

The above test was then repeated with a 5 dB SNR and for a 20 ms on-off modulation period. The results, as shown in Table 6.11, are similar to that for the 20 dB case shown in Table 6.10 except for a decrease (the width parameter changes from 0.8 to 2.8) in the resolution of the IID estimates.
Table 6.10  IID estimates and the position errors for two-source cases with different on-off modulation periods (MP). The SNR was 20 dB, and the IIDs of the two sources were -3 dB and 3 dB, respectively.

<table>
<thead>
<tr>
<th>MP (ms)</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (dB)</td>
<td>Width (dB)</td>
</tr>
<tr>
<td>20</td>
<td>-3.4</td>
<td>0.4</td>
</tr>
<tr>
<td>25</td>
<td>-3.4</td>
<td>0.4</td>
</tr>
<tr>
<td>30</td>
<td>-3.4</td>
<td>0.4</td>
</tr>
<tr>
<td>35</td>
<td>-3.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Fig. 6.2  An example IID distribution function for a two-source case with a 35 ms on-off modulation period.
Table 6.11 IID estimates and the position errors for a two-source case with a 5 dB SNR. The IID’s of the two sources were -3 dB and 3 dB, respectively.

<table>
<thead>
<tr>
<th>MP (ms)</th>
<th>First Peak</th>
<th></th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Width</td>
<td>Position Error</td>
</tr>
<tr>
<td>30</td>
<td>-3</td>
<td>2.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Next, the test shown in Table 6.10 was repeated again, but for different spatial separations of the two pure-tone sources. A 20 ms on-off modulation period was used in the test. Table 6.12 shows the results, where the first column lists the spatial separations (in terms of the two IID values) of the two pure-tone sources. The smallest IID difference was found to be about 3 dB.

Table 6.12 IID estimates and the position errors for two-source cases with different spatial separation (SS, in terms of the two IID’s). The SNR was 20 dB.

<table>
<thead>
<tr>
<th>SS (dB)</th>
<th>First Peak</th>
<th></th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Width</td>
<td>Position Error</td>
</tr>
<tr>
<td>(-1, 1)</td>
<td>-1.3</td>
<td>2</td>
<td>-0.3</td>
</tr>
<tr>
<td>(-1.5, 1.5)</td>
<td>-2.4</td>
<td>0.8</td>
<td>-0.9</td>
</tr>
<tr>
<td>(-2, 2)</td>
<td>-3.4</td>
<td>0.4</td>
<td>-1.4</td>
</tr>
</tbody>
</table>

The network was finally tested in a case where the two target sources were modulated with on-off intervals having random lengths selected from the range of [0, 35] ms. The two tones had the same 50 dB SPL intensity and the same 3000 Hz
frequency. The ITD of both sources was 0 second. The IIDs of the two sources were -3 dB and 3 dB, respectively. The SNR in the test was 20 dB. Table 6.13 shows the test results. The corresponding IID distribution function is shown in Fig. 6.3. Two peaks in this distribution function are identifiable that correspond to the two targets. The third (middle) peak corresponds to short-time IID estimates generated in time periods during which the two sources are both present.

Table 6.13 ITD estimates and the position errors for a two-source case with uncorrelated on-off modulations. The SNR was 20 dB.

<table>
<thead>
<tr>
<th></th>
<th>First Peak</th>
<th></th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position (dB)</td>
<td>-3.4</td>
<td>Width (dB)</td>
<td>3</td>
</tr>
<tr>
<td>Position Error (dB)</td>
<td>0.4</td>
<td>Width (dB)</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>-0.4</td>
<td>Position Error (dB)</td>
<td>0.0</td>
</tr>
</tbody>
</table>

![Fig. 6.3](image) Fig. 6.3 The IID distribution function for the uncorrelated two-source case. The parameters that describe the two target sources are shown in Table 6.13.
7 Test Results of IED Estimation

Similar to the ITD and IID networks presented in the last two chapters, the trained IED network was tested following the methods outlined in Section 4.3. The test results parallel those of the ITD and IID networks, and are presented in the following sections.

§7.1 AM Stimuli

As mentioned in Subsection 4.2.1, the IED network was trained using sinusoidally amplitude-modulated tonal stimuli, which we refer to as the AM stimuli. Thus, the first test for the IED network was carried out using a random sample of 120 AM stimuli to which white noise was added with a 20 dB SNR. The intensities of the test stimuli were randomly selected from the range of 30-60 dB SPL. The IIDs were randomly selected from the range of -12 to 12 dB. The carrier frequencies were randomly selected from the range (2500-3500 Hz) that covers the characteristic frequencies of the auditory nerve fibers at the input of the IED network. The modulation frequencies were randomly selected from the effective range (154-305 Hz) discussed in Subsections 3.4.1 and 4.2.1. As mentioned in Subsection 4.2.2, the IED values encoded by the output layer neurons range from -1.64 to 1.64 ms. We used a 0.83 ms IED in the test, which is an intermediate value between 0 second and the upper bound (1.64 ms) of the encoded IED range. The resulting IED estimates (in terms of the associated “position” and “width” parameters defined in Subsection 4.3.1) were shown in Table 7.1.

The above test was repeated using a lower (5 dB) SNR. Table 7.2 shows the results. Despite an about two-fold decrease (as compared with the data in Table 7.1) in resolution, correct IED estimates were obtained.
Table 7.1 IED estimates (mean ± standard deviation) and the error of the mean position estimate for AM stimuli with a 0.83 ms IED and a 20 dB SNR.

<table>
<thead>
<tr>
<th>Position: (p) (ms)</th>
<th>Width: (d) (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85±0.06</td>
<td>0.24±0.012</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 7.2 IED estimates (mean ± standard deviation) and the error of the mean position estimate for AM stimuli with a 0.83 ms IED and a 5 dB SNR.

<table>
<thead>
<tr>
<th>Position: (p) (ms)</th>
<th>Width: (d) (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84±0.12</td>
<td>0.54±0.34</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 7.3 IED estimates and the errors of the mean position estimates for AM stimuli with different IEDs and a 20 dB SNR.

<table>
<thead>
<tr>
<th>IED: (\Delta t) (ms)</th>
<th>Position: (p) (ms)</th>
<th>Width: (d) (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>1.2±0.08</td>
<td>0.25±0.12</td>
<td>-0.05</td>
</tr>
<tr>
<td>0.83</td>
<td>0.85±0.06</td>
<td>0.24±0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>0.42</td>
<td>0.42±0.09</td>
<td>0.24±0.11</td>
<td>0.0</td>
</tr>
<tr>
<td>0</td>
<td>0.004±0.09</td>
<td>0.19±0.06</td>
<td>0.004</td>
</tr>
<tr>
<td>-0.42</td>
<td>-0.42±0.07</td>
<td>0.2±0.07</td>
<td>0.0</td>
</tr>
<tr>
<td>-0.83</td>
<td>-0.82±0.07</td>
<td>0.27±0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>-1.25</td>
<td>-1.2±0.08</td>
<td>0.3±0.32</td>
<td>0.05</td>
</tr>
</tbody>
</table>

The test shown in Table 7.1 was also repeated for different IED values around 0 second. The results are tabulated in Table 7.3, where the first column shows the tested IED values.
The estimated IEDs (in the second column of Table 7.3) are plotted as a function of the true IEDs (in the first column of Table 7.3), as shown in Fig. 7.1. It is clear, from this plot, that the IED network is able to produce accurate IED estimates for AM stimuli.

![Fig. 7.1 Estimated IEDs shown as a function of the true IEDs.](image)

§7.2 Stimuli Modulated by Two-Tone Complex Signals

To test how well does the model work for stimuli with more complex modulation waveforms, we used modulation signals that contain two sinusoids of different frequencies while keeping the carrier signal sinusoidal. In this category, we first conducted a test using a random sample of 120 stimuli that have a fixed frequency difference between the two components in the modulation signals. Specifically, the two frequency components of the modulation signals were randomly selected from the range of 150-300 Hz except that the difference between the two components was
always 50 Hz. Other relevant parameters were randomly selected from the same corresponding ranges as that used in the tests discussed in the last section. The SNR was 20 dB. The IED to be estimated was 0.83 ms. Table 7.4 shows the test results.

<table>
<thead>
<tr>
<th>Position: p (ms)</th>
<th>Width: d (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85±0.06</td>
<td>0.25±0.11</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The network was also tested with a lower (5 dB) SNR while keeping other aspects of the test unchanged. The results are shown in Table 7.5.

<table>
<thead>
<tr>
<th>Position: p (ms)</th>
<th>Width: d (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85±0.09</td>
<td>0.49±0.21</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The results shown in Tables 7.4 and 7.5 are similar to the corresponding results (shown in Tables 7.1 and 7.2) for the AM stimuli case presented in the last section. This demonstrates the ability of the IED network to “generalize” and perform well with stimuli modulated by more complex waveforms than simple sinusoidal ones. Moreover, although the resolution of the IED estimates shows a two-fold decrease when the SNR is reduced from 20 dB (Table 7.4) to 5 dB (Table 7.5), the network is able to give out similar “position” estimates at the two noise levels.
The test shown in Table 7.4 was further repeated for different frequency differences between the two frequency components in the modulation waveforms. The results are shown in Table 7.6, where the first column shows the tested frequency differences, and the subsequent columns show the corresponding IED estimates. The true IED in the test stimuli was 0.83 ms. The SNR was 20 dB. It is evident that the frequency difference between the two components in the modulation waveforms has little effect on the estimation results.

Table 7.6 IED estimates and the errors of the mean position estimates for stimuli with two-tone complex modulation waveforms. The IED to be estimated was 0.83 ms. The SNR in the test was 20 dB. FD: frequency difference between the two components in the modulation waveforms.

<table>
<thead>
<tr>
<th>FD: Δf (Hz)</th>
<th>Position: p (ms)</th>
<th>Width: d (ms)</th>
<th>Position Error (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.87 ± 0.07</td>
<td>0.24 ± 0.11</td>
<td>0.04</td>
</tr>
<tr>
<td>50</td>
<td>0.85 ± 0.06</td>
<td>0.25 ± 0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>75</td>
<td>0.85 ± 0.08</td>
<td>0.31 ± 0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>100</td>
<td>0.85 ± 0.10</td>
<td>0.33 ± 0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>125</td>
<td>0.86 ± 0.06</td>
<td>0.39 ± 0.18</td>
<td>0.03</td>
</tr>
</tbody>
</table>

§7.3 Two AM Sources from Different Directions

Finally, the IED network was tested in situations where there were two sources emitting AM signals from different directions. In a first series of tests in this category, the two sources were periodically on-off modulated with orthogonal functions (see Subsection 4.3.5 B). The IID of both sources was 0 dB. The IEDs for the two sources were ±0.83 ms, respectively. The SNR in the tests was 20 dB. Table 7.7 shows the test results. The first column of Table 7.7 shows the different on-off
modulation periods tested. The other columns of the table show the corresponding IED estimates for the two sources. Fig. 7.2 shows the resulting IED distribution function for a specific on-off modulation period (17 ms). The two peaks in the distribution indicate the two IED estimates. We found that the smallest on-off modulation period for two-source separation by the network was about 17 ms.

Table 7.7 IED estimates and the position errors for two-source cases with different on-off modulation periods (OOP). The SNR was 20 dB, and the IEDs of the two sources were ±0.83 ms, respectively.

<table>
<thead>
<tr>
<th>OOP (ms)</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (ms)</td>
<td>Width (ms)</td>
</tr>
<tr>
<td>17</td>
<td>-0.86</td>
<td>0.16</td>
</tr>
<tr>
<td>23</td>
<td>-0.82</td>
<td>0.11</td>
</tr>
<tr>
<td>33</td>
<td>-0.85</td>
<td>0.16</td>
</tr>
<tr>
<td>67</td>
<td>-0.82</td>
<td>0.11</td>
</tr>
</tbody>
</table>

The above test was repeated with a lower (5 dB) SNR and a 17 ms on-off modulation period. The results are shown in Table 7.8, which demonstrate the network's ability to separate the directions of two AM sources in noise.

The test shown in Table 7.7 was also repeated for different spatial separations of the two AM sources. A 17 ms on-off modulation period was used in these tests. Table 7.9 shows the results, where the first column lists the spatial separations (in terms of the two IED values) of the two AM sources. The smallest IED difference between the two sources was found to be about 0.7 ms.
Table 7.8 IED estimates and the position errors for a two-source cases with a 17 ms on-off modulation periods (OOP). The SNR was 5 dB, and the IEDs of the two sources were ±0.83 ms, respectively.

<table>
<thead>
<tr>
<th>OOP (ms)</th>
<th>First Peak</th>
<th>Second Peak</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (ms)</td>
<td>Width (ms)</td>
</tr>
<tr>
<td>17</td>
<td>-0.78</td>
<td>0.33</td>
</tr>
</tbody>
</table>

A last test for the IED network was carried out in a case where the two AM sources were modulated with on-off intervals having random lengths selected from the range of [0, 33] ms. Again, the IEDs of the two sources were ±0.83 ms, respectively. The SNR in the test was 20 dB. Table 7.10 shows the test results. Fig.
7.3 shows the corresponding IED distribution function. The two lateral peaks in the distribution correspond to the two sources. The middle peak corresponds to short-time IED estimates produced by the network in time periods during which the two sources are both present. The time periods with both sources present are results of the random on-off modulation of the two sources.

Table 7.9 IED estimates and the position errors for two-source cases with different spatial separations (SS, in terms of the two IEDs). The SNR was 20 dB.

<table>
<thead>
<tr>
<th>SS (ms)</th>
<th>First Peak</th>
<th></th>
<th>Second Peak</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (ms)</td>
<td>Width (ms)</td>
<td>Position Error (ms)</td>
<td>Position (ms)</td>
</tr>
<tr>
<td>(-0.33, 0.33)</td>
<td>-0.3</td>
<td>0.16</td>
<td>0.03</td>
<td>0.28</td>
</tr>
<tr>
<td>(-0.5, 0.5)</td>
<td>-0.51</td>
<td>0.16</td>
<td>-0.01</td>
<td>0.5</td>
</tr>
<tr>
<td>(-0.67, 0.67)</td>
<td>-0.63</td>
<td>0.16</td>
<td>0.04</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 7.10 IED estimates and the position errors for a two-source case with random on-off modulations. The SNR was 20 dB, and the IEDs of the two sources were ±0.83 ms, respectively.

<table>
<thead>
<tr>
<th></th>
<th>First Peak</th>
<th></th>
<th>Second Peak</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position (ms)</td>
<td>Width (ms)</td>
<td>Position Error (ms)</td>
<td>Position (ms)</td>
</tr>
<tr>
<td>-0.81</td>
<td>0.11</td>
<td>0.02</td>
<td>0.85</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Fig. 7.3 The IED distribution function for the two-source case with random on-off modulations. Parameters associated with the two lateral peaks are shown in Table 7.10.
Conclusions

This work is concerned with the determination of the positions of sound sources in multi-source acoustic environments. Solutions to this problem have many applications in such areas as speech recognition, robot sensing, tele-operation, and man-machine interfaces. The ability to localize sounds is essential for animals and humans.

Sound localization is a non-trivial problem, especially in natural environments which are often noisy, dynamic, and have multiple simultaneous sources at different locations in space. Nevertheless, the human auditory system is able to localize with ease multiple sound sources in such complicated situations. It is a great challenge to develop machines that can mimic the processes that make auditory localization possible. This is the motivation for our work.

To approach this problem, in Chapter 2, we proposed the DLI modeling scheme. In this scheme the signals detected by the two ears are first decomposed into their spectro-temporal distributions as represented in the neural activities of the auditory nerve fibers. Spatial attributes or localization cues are then determined from energy concentrations in the distributions. A spatial scene of acoustic events is finally built by integrating the short-time energy concentrations according to their spatial attributes. While much work has been done previously on modeling how the peripheral auditory system decomposes the stimulus signals into spectro-temporal distributions, our work has focused on modeling the later two processes in the DLI scheme: brief-interval cue estimation and integration. We have proposed, in Chapter 3, a DLI model in which the task of brief-interval cue estimation is realized by a process of pattern recognition and comparison.
We have further proposed that three parallel implementations of the DLI model can be realized by placing, in each implementation, a set of neural networks along the basilar membrane. As discussed in Chapter 3, these separate sets of networks are used to model the separate processing of different binaural cues (ITD, IID, and IED) along separate pathways in the auditory system. Moreover, the ITD networks are placed in the lower frequency region (below 1200 Hz), and the IID and IED networks are placed in the higher region. Although the IID and the IED networks are placed in the same frequency region, they are trained to abstract different interaural differences (delay or intensity).

We have only presented results for one network of each type (ITD, IID, or IED). As discussed in Subsection 4.2.2, if we normalize the frequency tuning curves of the auditory periphery corresponding to different points along the basilar membrane with respect to their characteristic frequencies, they turn out to be similar in shape (see Eq. 4.2). Thus, the input patterns to different networks of the same type placed at different parts of the basilar membrane have similar characteristics. This in turn suggests that the test results of one such network are representative of those of other networks of the same type. In fact, we have trained several different ITD networks that cover different frequency ranges within the lower frequency region (from about 100-1200 Hz). Similar results were obtained for these networks. As mentioned in Subsection 4.2.3, it takes a relatively long time (weeks) to train just one network, and resource limitations have prevented us from training more complete sets of networks to cover the entire frequency region. Nevertheless, we have shown that robust short-time cue estimates can be obtained by training the relatively simple networks used in our model.

A rough estimate made from Fig. 1.14, which gives the perceived locations of impulse sound images as a function of the ITD (Blauert, 1983), suggests that a 10 μs ITD difference corresponds approximately to a one-degree difference in direction.
Corresponding to the estimated width parameter shown in Table 5.1, the ITD network is able to localize sound sources within about 4.5 degrees in direction, which is about the just noticeable angle difference measured in psychophysical experiments for sounds coming from an angle between 60 and 75 degrees from the median plane (see Fig. 1.17). In terms of interaural phase difference, the estimated width parameter shown in Table 5.1 corresponds to about 15 phase degrees, which is about the human ITD threshold for perceived change for a lateralized sound image, as shown in Fig. 1.15. As reviewed in Chapter 1, an important observation that is evident from both Fig. 1.15 and Fig. 1.17 is that the ITD threshold of perceived change is at minimum when the sound image is perceived in the center of the head (0 second ITD), and increases when the sound image becomes lateralized (larger absolute ITDs). In other words, the auditory system is more sensitive to ITD changes if the absolute ITD value is smaller. This phenomenon of differential ITD sensitivity is less evident in the test results of the ITD network, if any.

In the results of IID estimation presented in Table 6.3 the estimated values for the width parameter are generally smaller than (about half) the variation of the estimated values of the position parameter. Thus, we compare the standard deviations of the position estimates (rather than the width) with IID sensitivities measured in psychophysical experiments. The position standard deviations shown in Table 6.3 vary from about 1 to 2 dB. They are generally comparable to the measured IID threshold of perceived change for lateralized sound images, although the IID network in our model is slightly less sensitive (0.5 dB versus 1 dB) than the observed human performance for centralized sound images (see Fig. 1.9 on Page 16). As reviewed in Chapter 1, the IID threshold of perceived change is smaller (0.5 versus 1.5 dB) when the sound image is centralized than it is when the sound image is lateralized (see also Fig. 1.9 on Page 16). This phenomenon is also observable in the test results of the
IID network. From Table 6.3, the standard deviation of position varies from about 2 dB for the fully lateralized image case to about 1 dB for the centralized image case.

A characteristic of our model is that the binaural neurons in the three types of networks in the model "listen" to multiple auditory nerve fibers with different, but adjacent, characteristic frequencies. Such multiple-fiber innervation is also seen in Bonham's (1994) convergence model reviewed in Chapter 1. In Bonham's model, however, the signals carried by multiple nerve fibers from a single ear are first combined to obtain a composite signal, which is then used to estimate ITDs via a coincidence detection mechanism (Jeffress, 1948). In effect, the binaural neurons do not "see" the spectro-temporal patterns seen by neurons in our model. The fact that auditory nerve fibers with different characteristic frequencies from different ears are sent to the same binaural neurons is also seen in Shamma's (1989) stereausis model. But in Shamma's model the neurons receive only two inputs, namely a single fiber from each ear. Again, the binaural neuron does not see the spectro-temporal profile of the signals carried by the nerve fibers.

Fig. C.1 shows some examples of the connection patterns between the hidden layer neurons and the input layer ones in the trained ITD network. Two observations can be made: (i) the hidden neurons can be viewed as spectro-temporal filters; and (ii) there are inhibitory inputs from the auditory nerve fibers to the hidden layer neurons (black square shown in Fig. C.1). The first observation indicates to the physiologists that the spectro-temporal response properties of binaural neurons may be important in the study of neural mechanisms of auditory localization. The second observation provides us with a possible explanation for the usefulness of inhibitory inputs to the binaural neurons observed in the physiological experiments reviewed in Chapter 1 (Grothe and Sanes, 1993; Adams and Mugnaini, 1990; Schwartz, 1992; Cant and Hyson, 1992). The connection patterns in our model are different from many previous models that do not have inhibitory input at all (e.g. Jeffress, 1948; Colburn
et al., 1990; Dabak and Johnson, 1992). Inhibitory inputs to the binaural neurons are seen in a model by Sujaku et al. (1981), a schematic diagram of which is shown in Fig. 1.24. This model can be seen as a degenerated case of the connection patterns observed in our model.

Fig. C.1 (next page) Connection patterns between hidden layer neurons and input layer neurons in the trained ITD network. Each larger box in the figure corresponds to one hidden neuron, Inside these larger box are two matrices of connection weights (corresponding to the two matrices of input layer neurons) and a bias. The weights and bias are represented by white (positive weight value) or black (negative weight value) area in a small gray square. The sizes of the white or black areas in the small squares indicate the magnitudes of the weights.
To conclude this thesis, the following specific contributions are claimed. First, we have proposed a DLI scheme for models of auditory localization in non-stationary multi-source acoustic environments, and have accordingly developed a unique DLI model. The signals detected by the two ears are superpositions of the sound signals emitted from the spatially distributed sound sources in the acoustic environment. The ultimate task of sound localization is to decompose the sensory signals into a spatial map of the original sound sources. While it is very difficult to map the sensory signals directly to their spatial distribution, the idea behind the DLI modeling scheme is to find a "bridge", or an intermediate representation, between the sensory signals and the spatial distribution of sound sources. In order for such a bridge to be helpful in solving our problem, it must have the following property: it should consist of elements that can be mapped to no more than one source in the spatial distribution. We argue that the spectro-temporal distribution of the stimulus signal, which is manifested in the activity of parallel auditory nerve fibers, can serve as such a bridge. Small "patches" of the spectro-temporal distribution can be assumed to correspond to no more than one source in space. As it is very likely that different elements of the intermediate representation (the bridge) having the same short-time cues correspond to the same sources in space, the final spatial distribution of sound sources can be obtained by integrating, or grouping, these elements according to these cues.

Secondly, we have shown that there are unexplored patterns in the neural signals carried by parallel auditory nerve fibers that are important for sound localization. Spectro-temporal patterns are traditionally used in signal recognition tasks (Lippmann, 1989; Dror et al., 1995). Our work is the first to use such patterns for the purpose of passive sound localization.

Thirdly, we have trained a simple type of neural network to show that such patterns are good indications of interaural differences, and can be used to obtain robust short-time location cue estimates.
Fourthly, we have shown that the simplest types of stimuli, pure-tones (for ITD and IID estimation) or AM stimuli (for IED estimation), are adequate for the training of the neural networks in our model. The trained networks are shown to be able to "generalize" and perform well when the stimuli are more complex signals. Furthermore, we have shown that our model works in low SNRs and in non-stationary multi-source situations.

Finally, we have demonstrated that the same model structure can be used and trained to estimate different localization cues. This suggests that there is no real difference between the ITD and IED pathways except that the spectro-temporal representations of the sound stimuli are different; for low-frequency stimuli, the neural activity in the auditory nerve fibers reflects the fine timing structure of the stimuli, whereas for high-frequency stimuli, the neural activity reflects the envelope timing structure of the stimuli. In other words, the difference between the ITD and the IED pathway comes naturally from the results of the differential peripheral processing of low-frequency and high-frequency stimuli.

Although our model is computationally intensive and inherently parallel, as most brain models are, it is not far from practical implementation. Hardware chips (Lyon and Mead, 1988) already exist that mimic the peripheral processing of the auditory system. There are other chips (Lazzaro and Mead, 1989; Mead et al., 1991) that take the output of a cochlear chip as their input, and mimic certain types of binaural processing. Our model provide an effective new computational structure or algorithm that can be implemented in practical applications.

There are several directions in which further research may improve our model. One of the limitations of our model is that the networks in the model are trained and tested using stimuli that simulate anechoic environments. However, room reverberation and echoes may lead to biased location estimates or multiple estimates corresponding to both an original sound source and its echoes. The precedence effect
(Zurek, 1987) is thought to be related to the robust ability of the auditory system to localize in normal rooms, and the onset transients of sound stimuli play an important role in this effect. In order for our model to localize in room environments, methods that are more sophisticated than histogram accumulation should be used in the implementation of the "integration" step in the DLI modeling scheme. Specifically, a refractory mechanism that involves lateral inhibition may be developed to reduce the effect of echoes and room reverberation. In such a mechanism, the short-time cue estimates obtained immediately after the onsets will be inhibited if these estimates do not agree with the estimates obtained from the onset transients.

Another limitation of our model is that the networks in the model are trained using stimuli whose loudness levels are restricted to 30-60 dB. This is a relatively narrow range considering the large dynamic range of the sensitivity of the ear. Due to the non-linearity of the cochlear model, networks trained with stimuli in a limited range may not generalize or perform well for stimuli that have loudness levels below or above this range.

We have made an assumption about the acoustic environments which states that in most short (in the order of several milliseconds) time intervals only one sound has significant intensity. This assumption is likely to hold for such situations as a conversation between two or three people, but not for many environmental sounds or for most music. Thus, the solution provided by our model to the multi-source problem is relatively limited. To overcome this limitation, networks that "cover" frequency bands that are much wider than those used in our current simulation (or the entire audible frequency range) may be used. The task of such networks would be to recognize and then localize spectro-temporal patterns that are more complex than those limited to narrow frequency bands. Such more complex short-time spectro-temporal patterns may help distinguish simultaneous complex sounds. Another
alternative is to combine the outputs from parallel networks over different narrow frequency bands which may provide a satisfactory solution to the problem.

The training of the networks in our model takes a relatively long time. Further effort may be devoted to developing more efficient ways of training, and ways of improving the performance of the networks after the initial training. We have used simple multi-layer feed-forward neural networks for the task of pattern recognition and comparison, but have not investigated other possible network structures that may be more effective in performing similar tasks. A future project is to develop alternative network structures and to compare the performance of different network structures. Examination of these alternatives may also help to simplify or speed up the training process. Finally, it should be worthwhile to explore methods for the implementation of the present model using current integrated circuit technologies.
References


References


References


