MASTER'S THESIS

Biological versus Subspace Methods in Sound Localization

by Saurabh Dadu
Advisor: Prof. P. S. Krishnaprasad

CDCSS MS 2001-1
(ISR MS 2001-3)

The Center for Dynamics and Control of Smart Structures (CDCSS) is a joint Harvard University, Boston University, University of Maryland center, supported by the Army Research Office under the ODDR&E MURI97 Program Grant No. DAAG55-97-1-0114 (through Harvard University). This document is a technical report in the CDCSS series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CDCSS/cdcss.html

ABSTRACT

Title of Thesis: BIOLOGICAL VERSUS SUBSPACE METHODS IN SOUND LOCALIZATION

Saurabh Dadu, Master of Science, 2001

Thesis directed by: Professor P. S. Krishnaprasad
Department of Electrical and Computer Engineering

Sound localization is determining the location of sound sources using the measurements of the signals received by an array of sensors. Humans and animals possess the natural ability of localizing sound. Researchers have tried to model nature's way of solving this problem and have come up with different methods based on various neuro-physiological studies. Such methods are called biological methods. On the other hand, there is another community of researchers who has looked at this problem from pure signal processing point of view. Among the more popular methods for solving this problem using signal processing techniques are the subspace methods. In this thesis, a comparative study is done between biological methods and subspace methods. Further, an attempt has been made to incorporate the notion of head-related transfer function in the modeling of subspace methods. The implementation of a biological localization algorithm on a DSP board is also presented.

BIOLOGICAL VERSUS SUBSPACE METHODS IN SOUND LOCALIZATION

by

Saurabh Dadu

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2001

Advisory Committee:

Professor P. S. Krishnaprasad, Chairman
Professor Shihab A. Shamma
Professor K. J. R. Liu

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I wish to express my gratitude for Prof. P. S. Krishnaprasad for his guidance, inspiration and patience over the course of this project work. My interaction with him has helped me develop scientific thinking and an objective approach to engineering problems. Special thanks to Cliff Knoll for his efforts in helping me with the DSP implementation. The tools written by him made my life considerably easier. Thanks go to Prof. K. J. R. Liu and Dr. Ram Venkataraman for their useful suggestions. I express my appreciation towards Dr. Didier Depireux, Dr. Elena Grassi, and Mr. Dan Rapczynski for helping me understand the biological methods for sound localization. I would also like to thank my friends Yu Mao, Fumin Zhang and Sean Andersson for their support and help in different phases of the implementation.

The research was supported in part by the National Science Foundation Learning and Intelligent Systems Iniative Grant CMS9720334, by the Office of Naval Research under the ODDR&E MURI97 Program Grant No. N000149710501EE to the Center for Auditory and Acoustics Research and by the Army Research Office under the ODDRE&E MURI97 Program Grant No. DAAG55-97-1-0114 to the Center for Dynamics and Control of Smart Structures (through Harvard University).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The sound localization problem is to estimate the direction of sound sources using measurements of the signals received by an array of microphones. Sound localization can be useful in many applications such as robotic hearing, human-machine interface, electronic surveillance and military applications.

1.1 Sound localization in Biology

Above applications apart, sound localization is an important part of our lives. For many species such as barn owl, it is a matter of survival. The natural capabilities of human and animals to localize sound has intrigued researchers for many years. Numerous studies have attempted to determine the processes and mechanisms used by humans or animals to achieve spatial hearing.

One of the first steps in understanding nature's way of solving this problem is to understand how information is processed in the ear. A number of models for the ear have been suggested by the researchers [2, 3, 4]. These studies suggest that the cochlea effectively extracts the spectral information from the sound wave

impinging on the ear drums and converts it into the electrical signals. The cochlear output is in the form of electrical signals at different neuron points along the basilar memebrane of cochlea. The electrical signals then travel up to the brain for further processing.

Many researchers have come up with different models of processing of electrical signals in the brain for sound localization to support the experimental data from various neurophysiological studies. All these different models agree on the fundamental view that the direction of the sound is determined by two important binaural cues - the interaural time difference and the interaural level difference. These binaural cues arise from the differences in the sound waveforms entering the two ears. The interaural time difference is the temporal difference in the waveforms due to the delay in reaching the ear farther away from the sound source. The interaural level difference is the difference in the intensity of the sound reaching the two ears. In general, the ear which is farther away from the source will receive a fainter sound than the ear which is relatively closer to the source due to the attenuation effect of the head and surroundings. The phenomena of time delay and the intensity difference can be integrated into the notion of interaural transfer function which represents the transfer function between the two ears.

It is generally accepted that cross-correlation based computational models for binaural processing provide excellent qualitative and quantitative accounts of experimental studies. These models can be broadly classified into two kinds, namely, the temporal-correlation models and the spatial-correlation models. In the temporal-correlation models [11, 7], the cochlear outputs from the two ears are cross-correlated at various time delays. In the implementation of such a model, the cochlear outputs are passed through delay lines. The cochlear outputs from one

ear are continuously compared with the delayed cochlear outputs of the other ear. In the spatial-correlation models [1], the instantaneous cochlear outputs obtained from one ear are compared with the shifted image of the instantaneous outputs obtained from the other ear. Thus, the spatial correlation models eliminate the need of the delay line required to save the past cochlear outputs.

The output patterns obtained from the cross-correlation operations reflect the binaural information which can be refined further and interpreted to determine the direction of the source.

1.2 Sound localization: a signal processing view

A different community of researchers from the classical signal processing area has also been involved in solving the localization problem from a different perspective. In the signal processing community, the more commonly used term for this problem is *direction-of-arrival (DOA) estimation*. Earlier work in the field of DOA esimation has focused on narrow-band signals. It was shown that under the narrowband assumptions, the DOA problem is equivalent to a spectral analysis problem. Thus, the classical Fourier-based methods like periodograms can be used to solve it under some conditions [15]. In 1979, Schmidt [5] proposed a new algorithm, MUSIC (MUltiple SIgnal Classification), which introduced a new paradigm for solving the problem. Roy and Kailath [6] showed that the computations and the memory required by MUSIC can be reduced significantly by requiring that the sensors occur in matched pairs. The algorithm is known as ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques). The common element in MUSIC and ESPRIT is the concept of signal subspace which exploits the underlying data structure in the data model for binaural processing.

The subspace algorithms also assume that the signals are narrow-band and therefore cannot be applied directly to the sound localization problem due to the wide-band nature of the sound. The problem in wideband direction-of-arrival (DOA) estimation arises from the fact that the signal subspace is different for different frequencies. Many researchers have approached this problem by resolving the sensors outputs into discrete narrowband frequency bins and then independently applying one of the narrowband subspace techniques to each frequency bin. The estimates obtained from processing of each of the frequency bins are then averaged in some sense to obtain the final estimate of the DOAs. A brief survey of the efforts made in wideband DOA problem is given below:

In [16], a global search similar to that of spectral-MUSIC [8] is performed on the individual bins to estimate the null spectra for the narrowband components. The null spectral plots for each frequency bin are then arithmetically or geometrically averaged and the directions of arrival of the sources are determined from the peaks in the pseudo-spectrum plot. Su and Morf [17] employed a different approach in which the sensor output is modeled as multidimensional AR or ARMA process, i.e, having rational spectrum. They generalize the notions of signal subspace and array manifolds to rational vector space and develop rational signal subspace theory based on these concepts. The theory is applied to derive the unit circle eigendecomposition rational subspace (UCERSS) algorithm for source location. In UCERSS, the frequency domain representation of wideband signals is not explicitly used in the sense that the sensor outputs are not narrowband filtered to estimate correlation matrices for each frequency bins. Rather, the correlation matrix is first estimated and then transformed into the frequency domain using one of the multidimensional rational spectrum modeling schemes. The narrowband

signal subspace processing is then applied to discrete points on unit circle in the frequency domain. The individually obtained estimates are then combined in a similar fashion as [16].

Su and Morf [18] proposed another solution based on the rational signal subspace model known as modal decomposition signal subspace (MDSS) algorithm. It uses the fact that the output of the array at the system poles is characterized by the emitters sharing that pole. The column space of the residue matrices at the system poles spans the signal subspace corresponding to the emitters sharing that pole. By decomposing the emitter signals in this manner, more sources can be resolved than the number of sensors in the array. The number of sources that can be resolved at a pole is limited by the number of sensors. Otterston and Kailath [19] applied the ideas in modal decomposition signal subspace algorithm to ESPRIT that retained the basic advantages of ESPRIT as compared to MUSIC, namely the reduced number of computations and that the knowledge of array characteristics is not required.

An alternative representation of wideband signals was proposed in [20] based on a low-rank charecterization of the signal in a higher dimensional space but it requires large number of computations.

In 1983, Wang and Kaveh [21] demonstrated that it is possible to have a low-rank model of the system. They proved that there exist linear transformations that map the estimated subspace for one frequency to a focussing frequency. The linear transformations are known as focusing matrices. The sensor outputs are resolved in narrow frequency bands and their subspace estimate is mapped to a single focusing frequency by muliplying them by corresponding focusing matrices giving a low-rank model of covariance matrix. A narrowband DOA estimation scheme

can then be applied to this covariance matrix. This technique is called coherent signal subspace processing. The computation of linear transformations require premilinary knowledge of angles which can be obtained using low resolution (and hence computationally inexpensive) methods such as periodogram or conventional beamformer. In [22], it is shown that unitary focusing matrices result in improved performance. [22] and [23] describe methods to compute unitary focusing matrices.

Doron *et al.*[24] discovered a separable representation of the array manifold such that array characteristics (such as array geometry) and the frequency of the source signals can be separated from the angles-of-arrival. This made it possible to find transformations that do not require premilinary estimates of the angles. This method is termed as Array Manifold Interpolation (AMI). The separable representation in AMI is obtained by using infinite series expansion of plane waves in polar coordinates. The finite series approximation, in general, requires a large number of sensors. For the special case of a uniform circular array, termed the Circular Manifold Interpolation (CMI), the AMI method can be implemented efficiently using the FFT algorithm.

One of the primary objectives of this thesis is to compare the biological models and subspace models. Among the biological algorithms, we considered two algorithms covering both the temporal-correlation and the spatial-correlation based techniques. Among the subspace algorithms, we have considered both the MUSIC and ESPRIT-based methods. The finer details of the computations involved in the subspace models, however, differ from the models described above. An attempt has been made to incorporate the concept of interaural transfer function which is integral to the biological models.

The thesis is organized as follows. In Chapter 2, the hardware and software

setup for the implementation of sound localization has been described. Chapter 3 deals with the biological models in sound localization describing the fundamental concepts of head-related transfer function and the interaural transfer function. Lyon's cochlear model and the function of different blocks in the model are discussed. The output of the cochlear model is applied to two localization systems, one based on the temporal correlation methods and the other based on the spatial methods (stereausis). The various computations involved and the implementation on DSP hardware are described in detail. Chapter 4 focuses on the subspace methods for sound localization. The data model for subspace algorithms is developed that describes the relationship between the output of the sensors and the signals emitted by the sources and their dependence on various parameters such as the response characterstics of the HRTF and sensors, and the location of the sources with respect to the sensor array. The MUSIC and ESPRIT algorithms are derived and methods to estimate the direction of sound are developed. Finally, in Chapter 5, the results and the performance of all the four methods are presented and discussed.

Chapter 2

Implementation Setup

This chapter describes the set up for the real-time implementation of sound localization alogrithms.

2.1 Hardware Components

Figure 2.1 shows the major components in the physical set up of our system. The microphones mounted on the dummy head of a robot collect the sound signals. These signals are sent to the PC on which the Coreco board is mounted using a wireless LAN setup in the laboratory. The server program running on the PC receives the audio signals and passes them onto the Coreco Python/C67 board for processing and computation of the direction of the sound source. We will now describe each of the components in greater detail.

2.1.1 The Robot

The robot (Figure 2.2) used in the project is a Super Scout II, manufacured by Nomadic Technologies. It has an onboard computer powered by Pentium 233

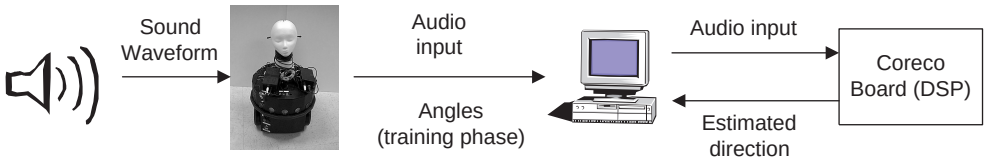

Figure 2.1: Dataflow between major hardware components.

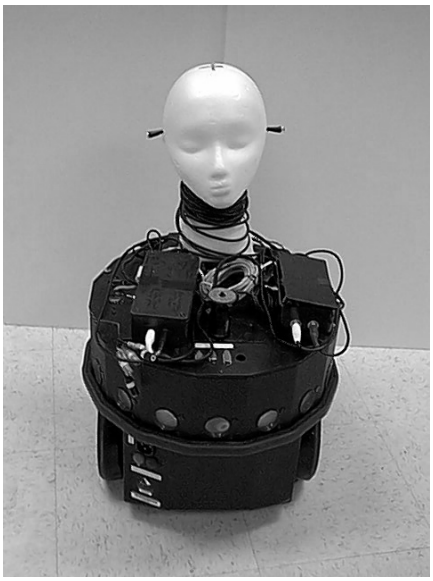

Figure 2.2: Super Scout II Robot

MHz processor which runs RedHat Linux. A dummy head made of Styrofoam was mounted on top of the robot. There were two microphones placed in the head at approximately the same locations as the human ears. The microphones were connected to a sound card in the computer system through analog amplifiers and filters. The filters were used to band-limit the input signals to 18.5 kHz. This simulated the hearing range of the humans. Secondly, these filters acted as anti-aliasing filters for discrete-time sampling by the sound card. The sound card digitized the audio signals at the rate of 40 kHz. The discrete-time samples of the

audio signals received from the two microphones were multiplexed together in one stream. The stream was then sent to the PC using a TCP/IP connection over a wireless LAN.

The reasons for using a robotic system for acquiring the sound data are two-folds:

1. The robot provided a mobile platform which was required for the online training of the sound localization system as explained later in Chapter 3.

2. Such a system can be used for further research in problems like human-machine interface, obstacle avoidance and so on.

2.1.2 The Coreco Board

The Coreco Python/C67 board was the core component of the whole system on which the sound localization algorithm was implemented. It is a multi-DSP board based on Texas Instruments' TMS320C6701 DSP chips. The configuration of our board consisted of four DSP chips connected via dedicated communication link with a peak bandwidth of 400 MB/s. The board provides up to 6400 MIPS of processing power making it suitable for intense number-crunching required by signal processing algorithms. The system is designed for multiprocessing applications. In our implementation of sound localization system, we used all the four DSPs.

2.1.3 The Host Computer

A Pentium PC running Windows NT was used as host to the Coreco board. It was connected to the Coreco board using the PCI Bus. The host computer not only provided an interface to the Coreco board but also acted as a communication link

between the robot and the Coreco board. The Code Composer Studio provided the platform for the software development and debugging environment.

2.2 Software Components

Figure 2.3 shows the key software components that were developed for the project and the flow of information between them. The *NETSRV* program[1] running on NT machine is the central link for exchange of information between the user, the robot and the Coreco board.

The console window provided by *NETSRV* is used to interact with the user for loading the sound localization programs onto the DSP chip, uploading of certain parameters required by the algorithms and setting up the socket connections between the data acquisition program and the control program running on the robot computer.

The *data acquisition program* on the robot digitizes the audio signals received from the microphones, opens a TCP/IP socket and waits for the connection to be set up. Once the connection is completed by the *NETSRV* program, it continuously sends out the audio data in blocks of size 1024 samples.

The *robot control program* is used only in the learning phase. It controls the movement of the robot and is discussed in greater detail in the next chapter.

The communication between the Coreco board and the *NETSRV* program was realized using the *application programmer's interface (API)* software modules provided with the Coreco system.

[1]The *NETSRV* program was written by Cliff Knoll of Neural Systems Lab

2.2.1 C60 Host API

The Host API follows a shared memory model whereas the DSP's memory is visible from the host application. Most of the operations are carried out by directly mapping the DSP's memory onto the host. These APIs offer a basic set of functionalities for communication between the C6701 and the host program *NETSRV*. More complex functionalities were built using the simple functions.

2.2.2 C60 Native API

C6701 offers low level APIs that are called C60 Native API. These APIs are used for message passing between the host and the DSP, memory allocation, interrupts, timer, buffers and direct memory access (DMA) management. These APIs have been used extensively by the *NETSRV* program as well as the sound localization program.

The APIs pass on the input audio data to the *sound localization firmware* running on TI TMS320C6701 DSP which computes the estimated direction of the sound source and relays it back to the PC. The implementation of the sound localization firmware is described in Chapter 3.

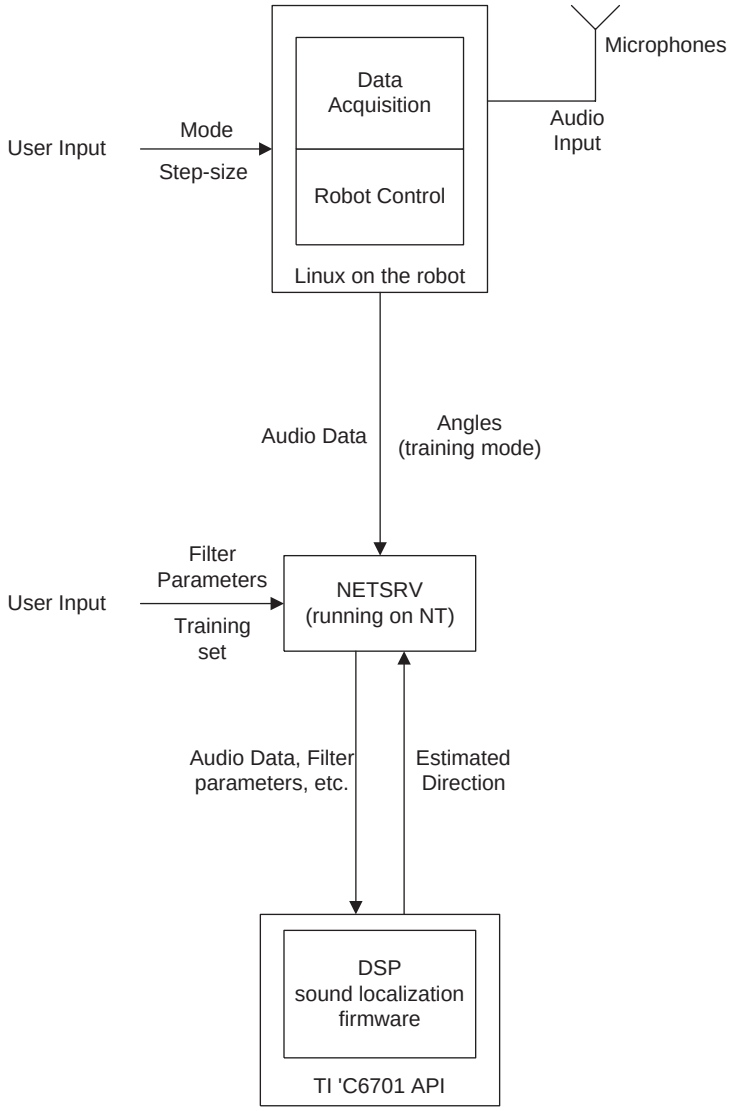

Figure 2.3: A schematic diagram of the interaction among the software blocks.

Chapter 3

Biological Methods for Sound Localization

The algorithms presented in this chapter were inspired by how humans and animals localize sounds. The algorithms use the cochlear model to separate the spectral information in the sound wave.

3.1 The Cochlear Model

The cochlear model is an attempt to model the mammalian cochlea based on neurophysiological studies. This model describes the propagation of sound in the inner ear and the conversion of the acoustical energy into neural representations. Sound that enters the outer and the middle ear is passed through the oval window into the cochlea. Once in the cochlear duct, the pressure wave propagates down the basilar membrane. The stiffness of the basilar membrane varies smoothly over its length. Thus a point in the basilar membrane is most resonant to a particular frequency in the pressure wave. The vibrations at different points in the membrane are sensed

by the hair cells which convert the mechanical signals into electrical signals. These electrical signals are then communicated to higher levels in the brain.

Since each point in the basilar membrane responds best to one frequency, it effectively decomposes the acoustical energy into different frequency bands. The cochlea near its base (where the sound enters) is most sensitive to high frequency sounds and as the wave travels down the cochlea, it becomes more sensitive to lower frequencies.

This frequency dependent response of cochlea can be best modeled as continuous differential equations. However for implementation purpose, it is normally modeled in discrete sections as a bank of bandpass filters, called cochlear filters. These filters separate the input to the ear in different frequency bands or channels. The output of each cochlear filter is passed through non-linear structures such as half-wave rectifier (HWR) and automatic gain controller (AGC) to simulate the response of actual human cochlea. Figure 3.1 shows the schematic diagram of the cochlear model. The output of the cochlear model is a set of N signals, where N is the number of cochlear filters.

3.1.1 The Cochlea Filter Bank

The cochlear filters can be emulated by the gammatone filters. In our experiments, we used a bank of $N = 129$ cochlear filters with characteristic frequencies spanning the whole audio spectrum. The frequency response of some of the cochlear filters are shown in the Figure 3.2. As we see from the figure, the filters lying in the same neighborhood have large overlap which introduces correlation across the frequency channels. This correlation is exploited by the spatial-correlation-based stereausis algorithm for binaural processing.

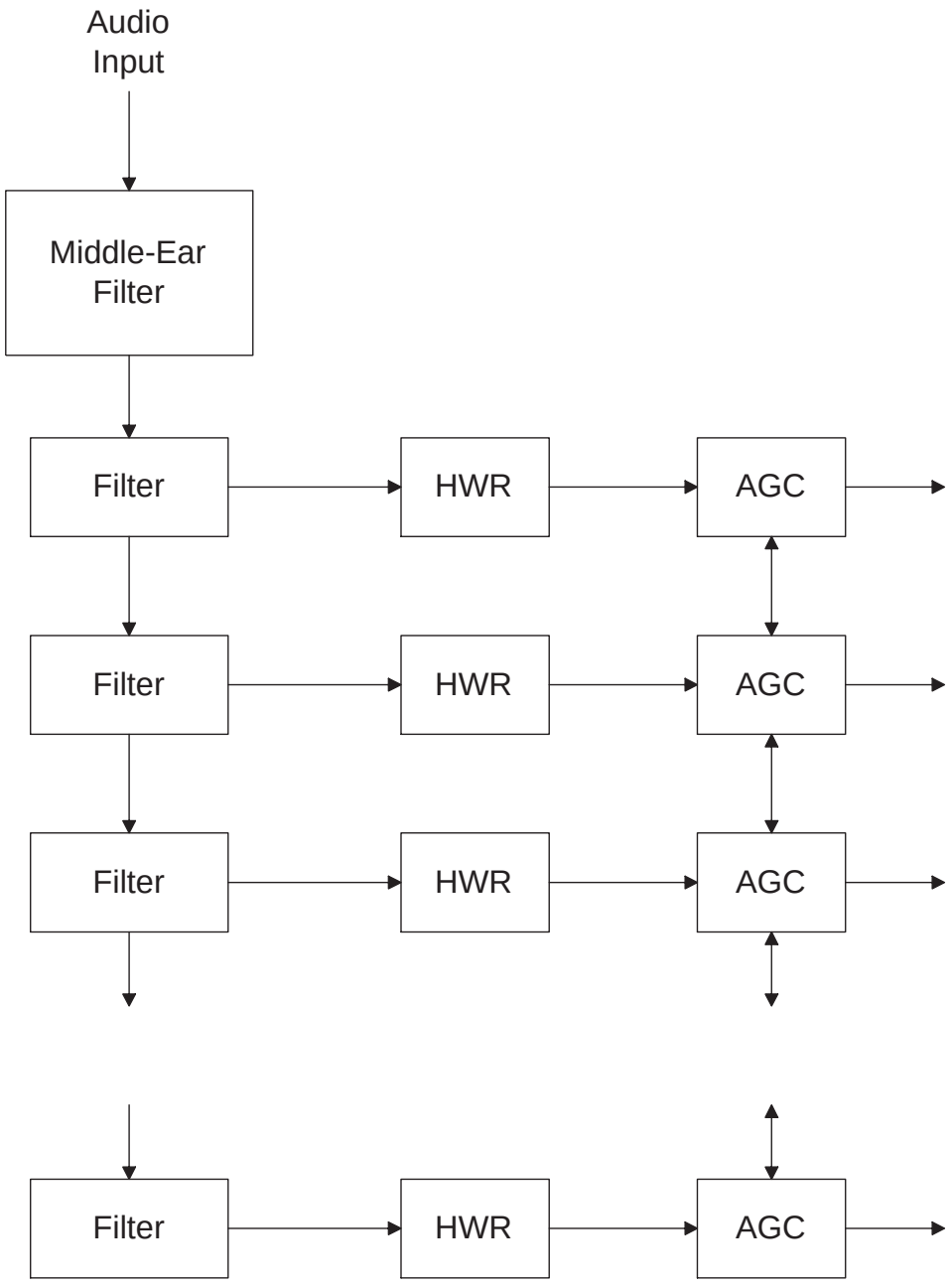

Figure 3.1: Lyon's cochlear model.

3.1.2 Automatic Gain Controller

The cochlear filters are followed by simple half-wave rectifiers. The output of a half-wave rectifier models the non-linearity of the hair cells, providing a non-negative output representing neural responses.

Automatic gain controllers are used to capture other non-linearities of the ear such as saturation and masking. A four-stage automatic gain controller (AGC) was used. The signals of each channel coming out of the HWR stages, pass through these four AGC stages. The gain of each stage depends on a time constant. The different time constants simulate the different adaptive times of our auditory system; the first AGC stage has the biggest time constant so that it reacts to the input signal more slowly, while the following stages have decreasing time constants. The AGC stages of each channel are coupled to the corresponding AGC stages of the adjacent channels. Thus a channel can affect the output of all the channels in the filter bank although the effect will decay exponentially with distance. Such a coupling, in effect, produces masking effects in the cochlear output. The outputs of the last stage approximately represent the neural firing rates produced by the transformation of various parts of the cochlea due to the sound pressure waves entering the inner ear.

Figure 3.3 shows the implementation of the AGC [25]. The objective of AGC is to attenuate the input signal so that on average it remains below a target value. The loop filter is a simple low pass filter with a feedback gain of $(1 - e)/3$. The time constant is related to the parameter e by the following equation

$$\text{time constant} \quad = \quad 1 - \exp\{-F_s/e\}$$

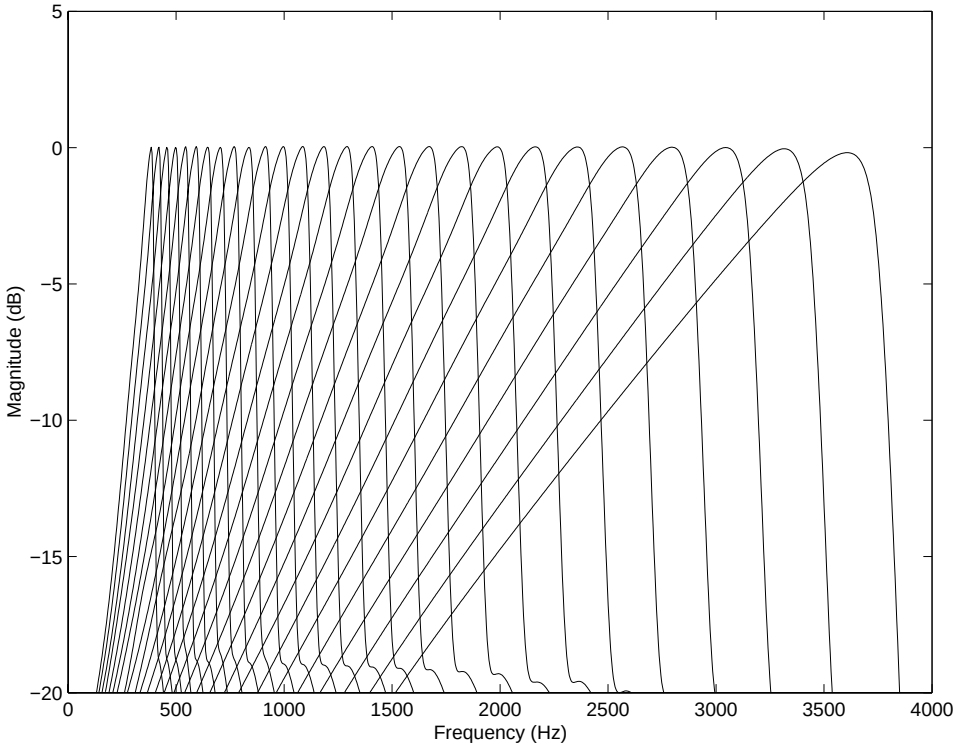

Figure 3.2: Magnitude response of the cochlear filters.

where F_s is the sampling frequency. A longer time constant means that the response of AGC to the input is slower.

The states of the two adjacent cochlear frequency channels are combined with the current channel and averaged. The target parameter is used to scale the input to the loop filter. In long run, as shown below, the state will track the value of the output of the AGC divided by the value of target.

Assuming the state values of the adjacent and the current channels are equal, the state equation can be written as

$$\text{state}(n) \quad = \quad \frac{ey}{\text{target}} + 3 \cdot \text{state}(n-1)\frac{1-e}{3} \tag{3.1}$$

In long term, for constant value of y, the state(n) is given by

$$\lim_{n\to\infty} \text{state} \quad \rightarrow \quad \frac{y}{\text{target}} \tag{3.2}$$

The output of AGC, in long term, is then given by

$$y \quad \rightarrow \quad \frac{\text{target}\, x}{\text{target} + x} \tag{3.3}$$

The values of time constant and target parameters used in the implementation were:

AGC stage	Time constant	target
First AGC	640 ms	0.0032
Second AGC	160 ms	0.0016
Third AGC	40 ms	0.0008
Fourth AGC	10 ms	0.0004

Table 3.1: Paramaters of Automatic Gain Controller

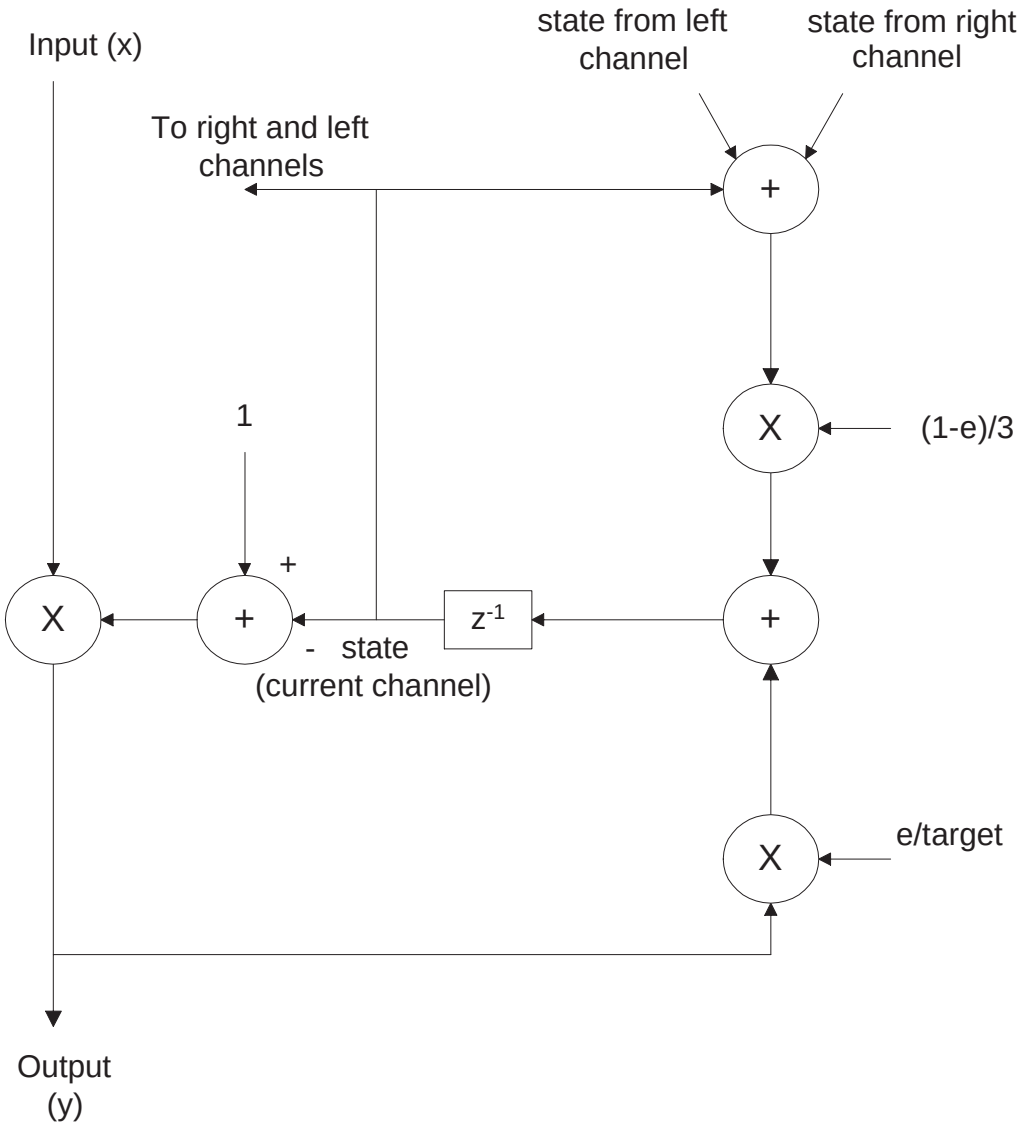

Figure 3.3: A schematic diagram of the automatic gain controller (adapted from Slaney [25]).

3.2 Interaural Transfer Function

Before we define the interaural transfer function, we will describe the important concepts that characterize the interaural transfer function. They are as follows.

3.2.1 Binaural Cues

The differences in the sound waves impinging on the two ears are known as binaural cues. Such differences are essential to locate sound sources in space. Interaural time difference (ITD) and the interaural level difference (ILD) are recognized as the two most important binaural cues for localization.

Interaural Time Difference

The interaural time difference is the time delay between the signals reaching the two ears that arises because the separation of the ears introduces a path length. The time delay depends on the separation distance between the ears, the angle of arrival of the sound wave, and its speed of propagation. It is generally difficult to measure the time delay, per se. So, usually, the phase difference in the two signals is used as a measure of ITD. For this reason, ITD is also know as interaural phase difference.

Interaural Level Difference

The interaural level difference is the difference in the intensity of the signals reaching the ears. Sound waves that come from different directions in the space are diffracted and scattered by the head, shoulders, torso, etc. This causes differences in the wave appearing on the ear drums and is the basis of ILD.

It may seem that the ITD information should be sufficient for sound localization but that is not so. ILD complements the ITD information in many situations. One of them is in the case of higher frequencies (> 1500 Hz) when the phase information becomes ambiguous. This can be explained using the Nyquist sampling theorem and its equivalence in the spatial sampling case. Indeed, at a particular instant of time, the sound waveform is sampled by microphones at two points in space. The distance between the microphones is analogous to a time-sampling period. If this distance is greater than half the wavelength of the signal, then the phase information cannot be determined with certainity.

The ILD information may also be useful in solving the *cone of confusion* problem. The cone of confusion is the set of all directions for which the time delay is same. If the ILD information is different for each of the directions on a cone, it can be used to discriminate and locate the direction of the source.

3.2.2 Head-Related Transfer Function

The transformation of the sound wave from the source to the ear is normally described by a transfer function called the head-related transfer function (HRTF). The HRTF is a function of the frequency of the signal and the location of the source with respect to the head. The location of the source can be specified by its range, azimuth angle and the elevation angle. In this thesis, we shall be concerned with the estimation of azimuth angle only. The elevation angle will be asssumed to be zero at all times. However, the extension to 2-D case of azimuth and elevation estimation is straightforward in most cases. Further, the source will be assumed to be in the far field; thus the dependence of HRTF on the range will be ignored. To this end, consider a sound source located at azimuth angle θ with respect to

the head. Let $S(\omega)$ be the Fourier transform of the source signal, $H_X(\omega, \theta)$ and $H_Y(\omega, \theta)$ be the HRTF's of left-ear and right-ear respectively, then the Fourier transform of the signals received at the two ears can be given by

$$X(\omega, \theta) = H_X(\omega, \theta)S(\omega) \tag{3.4}$$

$$Y(\omega, \theta) = H_Y(\omega, \theta)S(\omega) \tag{3.5}$$

Next, we define,

$$F(\omega, \theta) = \frac{H_Y(\omega, \theta)}{H_X(\omega, \theta)} \tag{3.6}$$

$F(\omega, \theta)$ is known as the interaural transfer function (ITF). The interaural transfer function captures the important binaural cues. The interaural time difference is captured in the phase information of the ITF. More specifically, the derivative of $\arg(F(\omega, \theta))$ with respect to ω gives the ITD. Note that introduction of frequency-dependent HRTF results in dependence of the interaural time (phase) difference on frequency. On the other hand, the magnitude of $F(\omega, \theta)$ provides a measure of frequency-dependent ILD.

The ITF can be estimated by taking the ratio of Fourier transforms of the signals received at the left-ear and the right-ear.

$$F(\omega, \theta) = \frac{Y(\omega, \theta)}{X(\omega, \theta)} \tag{3.7}$$

It is important to note that $F(\omega, \theta)$ is independent of the source spectrum and thus can be used to find the location of any wideband source. This observation can be utilized for finding a simple way of solving the sound localization problem using *a priori* information. Suppose the actual interaural transfer function of the head, $F(\omega, \theta)$, is known *a priori*. This information may be obtained from a training process. Later, in order to estimate the direction of an unknown source signal, one

can estimate the interaural transfer function of the head from the received signals using (3.7) and compare it with the known $F(\omega, \theta)$. The value of θ for which the estimated interaural transfer function is 'closest' to the actual interaural transfer function gives the direction of the source.

3.3 The Localization System

The method of sound localization described in this section was proposed by Lim and Duda [7]. Figure 3.4 shows the schematic diagram of the localization system. The input source signal received at the ears are processed through a cochlear model. The output of the cochlear model is used to obtain the binaural cues, namely the ITD and the ILD.

A common way of calculating the ITD cue is to first crosscorrelate the cochlear outputs of corresponding left-ear and right-ear channels for different time-lags and finding the time-lag for maximum crosscorrelation. A single interaural time difference is arrived at by averaging the ITD values obtained for each channel. In mathematical notation, if ITD_i represents the ITD for the frequency channel i, then

$$ITD_i = \arg \max_{l \in [-l_{max}, l_{max}]} \sum_k x_i(k) y_i(k-l) \tag{3.8}$$

$$ITD = \frac{1}{N} \sum_i ITD_i \tag{3.9}$$

where, $x_i(k)$, $y_i(k) \in \mathcal{R}$ are, respectively, discrete-time left-ear and righ-ear cochlear outputs for channel i. The notation l is the time-lag for crosscorrelation, and l_{max} is the maximum time-lag possible between the signals received at the two ears. Maximum time-lag $l_{max} = (\Delta/c)F_s$ depends on the distance between the ears Δ, the sampling frequency F_s, and the propagation speed of the sound c.

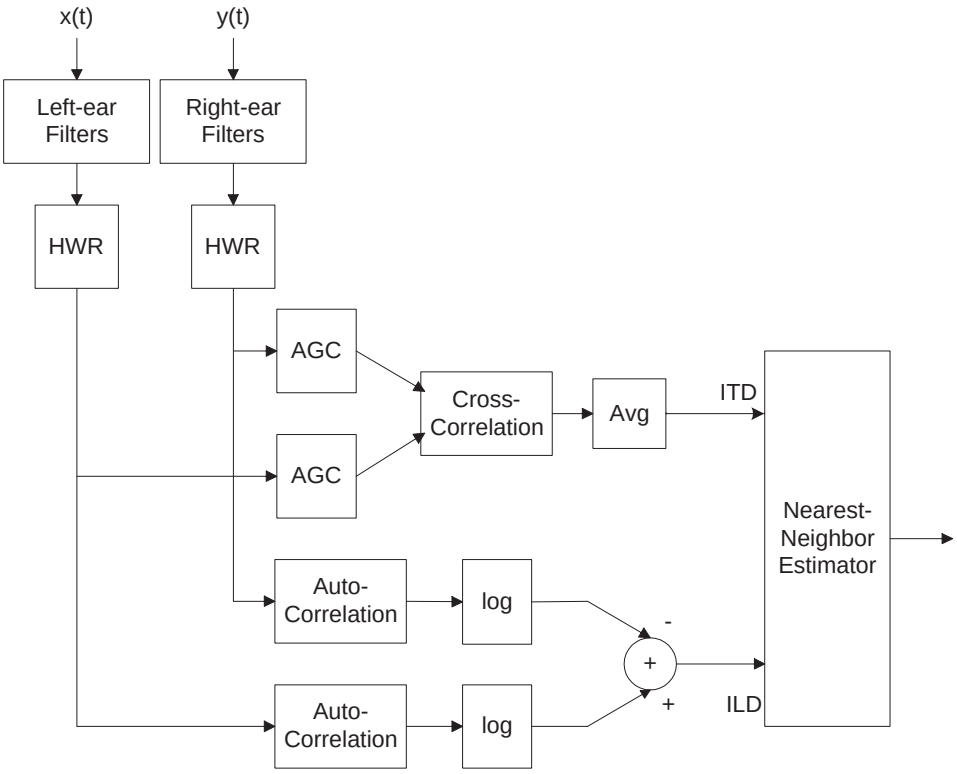

Figure 3.4: The Localization System (taken from Lim and Duda [7]).

The Jeffress network [14] shown in Figure 3.5 provides an efficient way to compute temporal correlations at different time-lags.

To compute the ILD, the AGC is disabled in order to preserve information regarding the amplitude level differences in the signals. The ILD spectrum is obtained by calculating the logarithm of the ratio of the signal energies for the corresponding channels in the left-ear and the right-ear. The signal energy in each channel is estimated by computing the zero-lag autocorrelation of the channel output.

$$ILD_i = 10(\log_{10}(\sum_k x_i(k)x_i(k)) - \log_{10}(\sum_k y_i(k)y_i(k))) \quad (3.10)$$

The vector $[ITD, ILD_1, \cdots, ILD_N]$ contains information regarding the interaural transfer function and will be known as ITF vector.

The ITF vector is an approximation to the interaural transfer function $F(\omega, \theta)$ as described in the previous section. Firstly, it assumes that the phase in $F(\omega, \theta)$ does not depend on the frequency, and uses a single value of ITD to represent the phase information. Secondly, the magnitude of $F(\omega, \theta)$ is not computed using Discrete-Fourier Transforms (DFTs) but from the frequency channels in the cochlear model.

3.3.1 Learning of Interaural Transfer Function

The training of the system to learn the interaural transfer function is the first step towards estimating the direction of an unknown sound source. It requires a controlled environment to reduce the errors due to random noise. A white noise sound source is used for the training purpose. In order to compute the ITF vector for angle θ, the source is placed at that angle. The measurements obtained from the microphones are used in equations (3.8), (3.9) and (3.10) to compute the

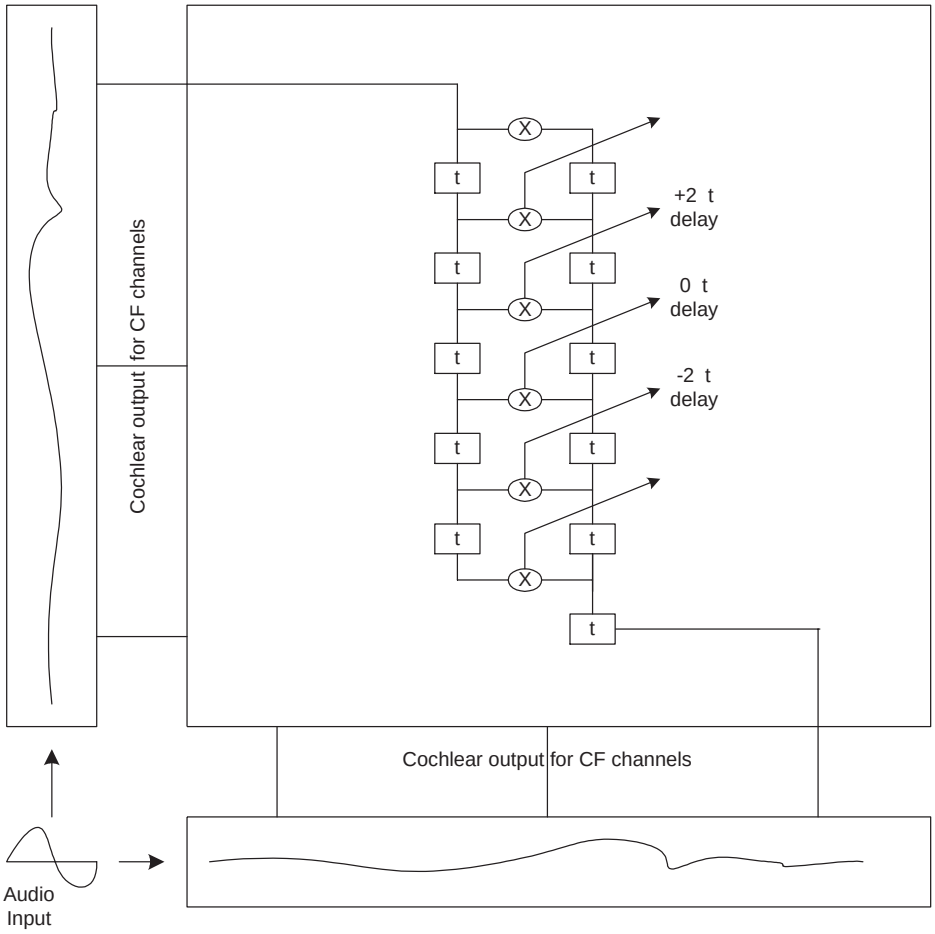

Figure 3.5: The Jeffress network.

ITF vector at angle θ. Since the vector incorporates the information regarding all the frequency channels, it is denoted by $F(\theta)$, with slight abuse of notation, to represent the interaural transfer function of the system at angle θ. The set $\{F(\theta), \theta \in \Theta\}$, represents the learning or the training set of interaural transfer function, where Θ denote the set of all angles for which the system is trained.

3.3.2 Estimation of direction

To estimate[1] the direction, $\widehat{\theta}$, of an unknown source, the waveforms received at the two ears are processed through the cochlear model and the interaural transfer function represented by the ITF vector $\widehat{F}$ is estimated. The vector $\widehat{F}$ may be different from the vectors in $\{F(\theta), \theta \in \Theta\}$ because of the random noise and the variation of the location of the source from the angles in Θ. The maximum likelihood (ML) approach is followed to estimate $\widehat{\theta}$. Under standard assumptions of independence, additive Gaussian noise and arbitarily large training set, the ML method says that the best estimate is given by the following expression

$$\widehat{\theta} \;=\; \arg\min_{\theta}(\|\widehat{F} - F(\theta)\|^2), \quad \theta \in \Theta \tag{3.11}$$

The authors of [7] also call it as the nearest-neighbor estimator as it involves finding the a vector in $\{F(\theta), \theta \in \Theta\}$ which is closest to $\widehat{F}$ in the sense of Euclidean distance.

[1]Throughout this thesis, the notation $\widehat{\ }$ is used to denote the corresponding quantities in the estimation phase to be differentiated from the theoretical values and the quantities in the training phase.

3.4 Implementation in Real-Time

The algorithm was implemented on a Coreco board in C. A cochlear model with 129 channels was used. The implementation of 129 channels required a large number of computations and memory. So, three DSPs were used with 43 channels implemented on each of them. The fourth DSP was used for implementing the ML nearest-neighbor estimator. The system can be run in two modes, the estimation mode and the learning mode.

- **Estimation Mode:** In the estimation mode, the audio signals emmitted by a source is received by the microphones and sent to the Coreco board which computes the vector $\widehat{F}$. The estimator implemented on the fourth DSP picks up this vector, compares it with the training data and gives out the angle corresponding to the closest match as the estimated direction of the source.

- **Learning/training mode:** The user can switch the system from the estimation mode to learning mode from the robot console. The whole process of learning interaural transfer function is automated; except that the system assumes that a broadband sound source is present at azimuth angle equal to zero. As soon as the user specifies switching to learning mode, the old data buffers in the Coreco system are flushed, and the program switches to learning mode. On the robot side, the robot console program instructs the robot to rotate in incremental steps (the step size in degrees can be specified by the user). The white noise sound data emitted from a speaker is recorded at different angles and is sent to Coreco board for further processing. At the same time, the robot also sends out the angle that the robot is making with the sound source (computed using the step size). This angle is needed so that

it can be attached as a tag to the corresponding ITF vector in the training set. Proper care is taken to keep coordination between the robot and the Coreco board during the transfer of the sound data from the robot and the computation of ITF information at the Coreco board in order to ensure that the sound data used for the computation matches with the angle associated with it.

The Coreco board presented severe constraints on the availability of memory; both in terms of the program size as well as the data size. In fact, it was so severe that it was not possible to hard code the coefficients of the cochlear filter in the program. This made the program so large that it would not fit into the program memory. To get around this problem, 'Read file' utility provided by *NETSRV* program was utilized. A MATLAB file was written that generated three binary files for each of the three DSPs on which the cochlear filters were implemented. The files consisted of a header followed by the filter coefficients. The information in the header was used to dynamically allocate memory where all the filter coefficients were stored. Pointers to special data structures were utilized to retrieve the proper coefficients of a cochlear filter.

Due to the high order of the cochlear filters, circular buffers were used. In general, implementation of digital filters requires shift registers to realize the delay lines. The disadvantage of shift registers is that every time a new sample comes in, the data in the shift register needs to be shifted to accomodate the new sample. This process of shifting the registers reduces the efficiency. In circular buffers, on the other hand, the new data simply overwrites the oldest data. The TI 'C6701 DSP processor provides hardware support for the circular buffers. To utilize this facility, the filters were implemented in Assembly language which resulted in faster

execution of the filtering operation.

3.5 The Stereausis Algorithm

The stereausis model was proposed by Shamma *et. al* [1]. It is a two-dimensional model and measures the binaural cues by detecting the instantaneous disparities in the cochlear responses along the basilar membrane in the two ears.

At any instant of time, the outputs of a cochlea can be viewed as a snapshot of a traveling wave in the basilar membrane. The stereausis model utilizes the disparities in the two traveling waves of the ears to compute the binaural cues. For instance, the interaural delay in the sound signals received by the two ears will produce traveling waves such that one is phase-shifted in one ear relative to the other (see Figure 3.6(b)). In other words, the snapshots of the waves traveling along the basilar membrane will appear shifted in space. Similarily, the interaural level difference due to the HRTFs will produce relative disparities in the amplitudes of the traveling waves. Such differences are known as *spatial disparities* in the stereausis model. Figure 3.5 shows the stereausis network that is used to measure binaural differences from the spatial disparities.

The channel outputs of the cochlear model is fed into the network which produces a 2-D image or a matrix. Both the axes of the image represent the characterstic frequencies (CF) of the channels. The elements of the image, c_{ij}, are computed by cross-correlating the outputs of the ith frequency channel of the left ear with that of the jth frequency channel of the right ear. If $C(\cdot, \cdot)$ represents the cross-correlation function, then

$$c_{ij} = C(x_i, y_j) \tag{3.12}$$

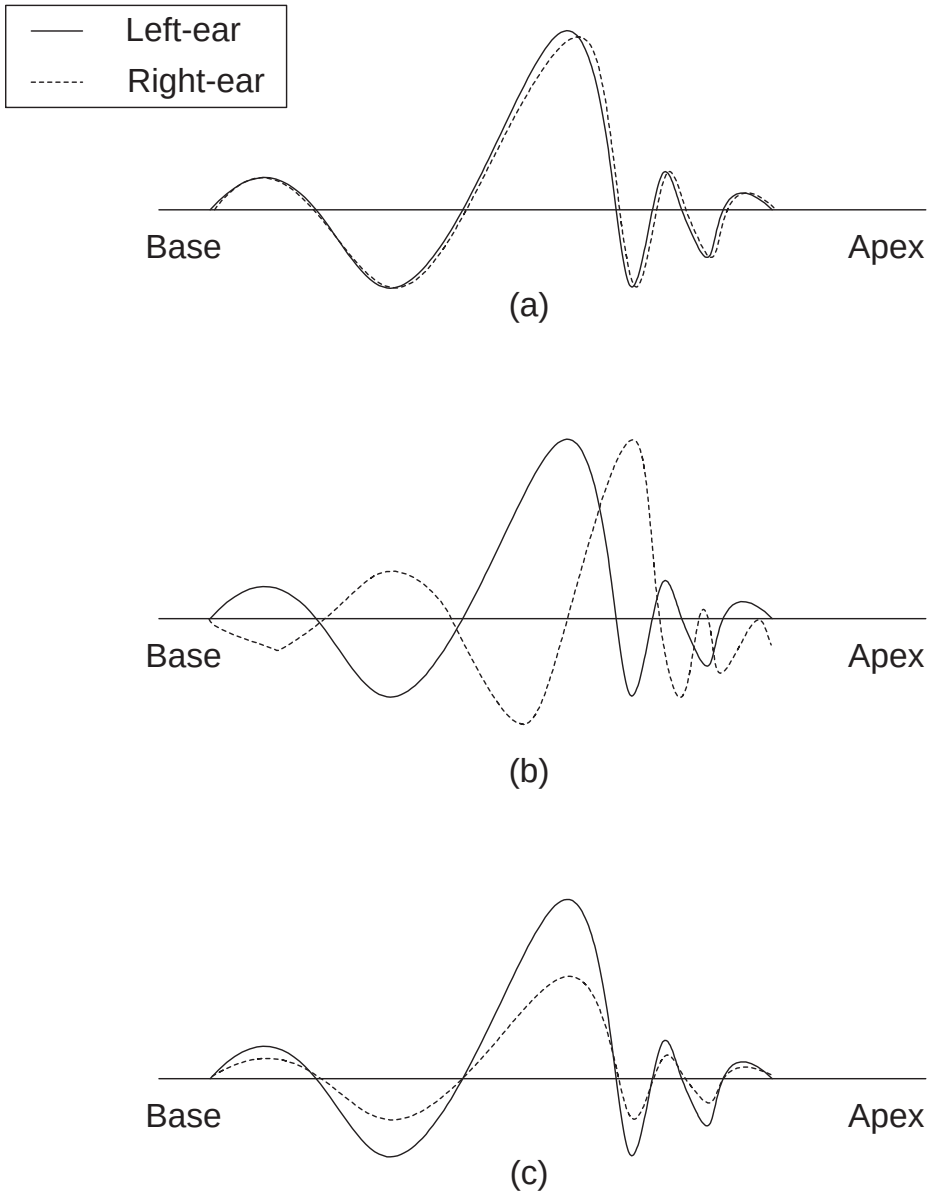

Figure 3.6: Traveling waves in the basilar membrane of the cochlea (adapted from Shamma *et. al* [1]).

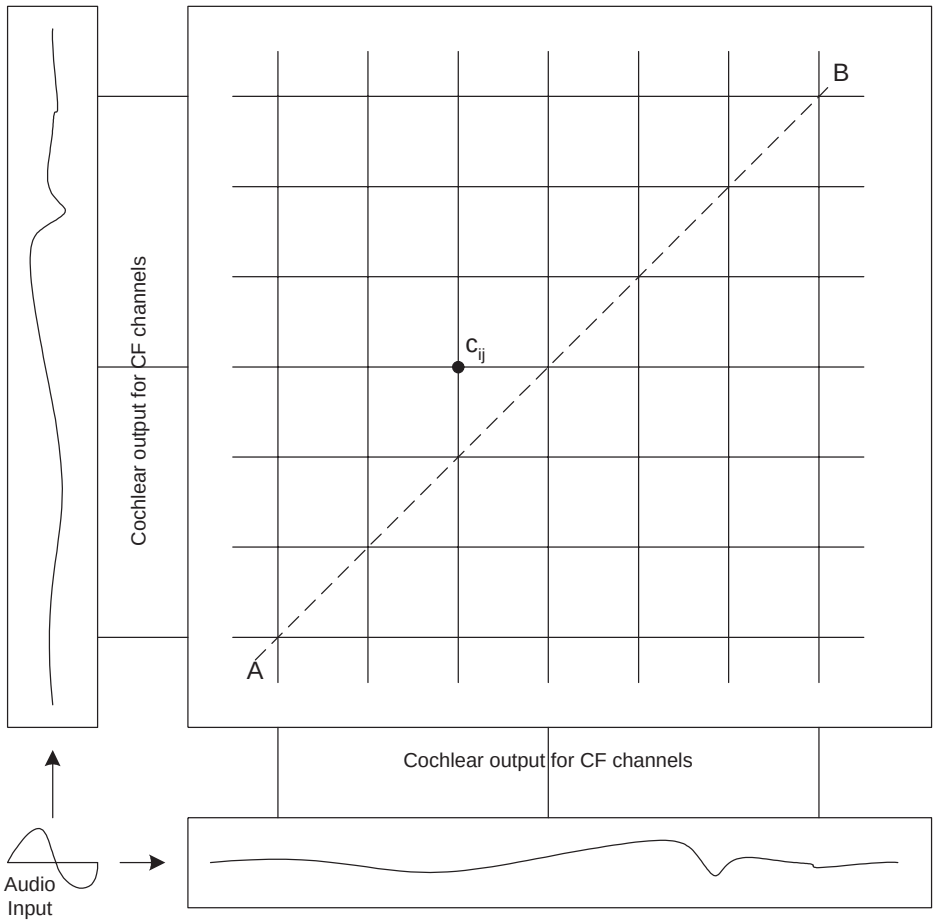

Figure 3.7: The stereausis representation (adapted from Shamma *et. al* [1]).

where x_i and y_j represents the instantaneous outputs of ith and jth channels of the left-ear and the right-ear respectively. The stereausis images are computed over a period of time and then averaged. Figure 3.8 shows typical stereausis images for an input signal which is a mixture of 600, 800, 1000 and 1200 Hz tones. The valleys and ridges in the image represent a measure of the output activity with the regions of ridges (darker regions) depicting high correlation activity and vice versa.

There are two important axes of information in the stereausis image

- The Disparity or Lateralization Axis: The axis normal to the diagonal AB in Figure 3.8 is called the disparity or the laterization axis (represented by CD). The correlation activity along the phase disparity axis shows the disparity between the left and right channel signals.

  The stereausis network systematically correlates the cochlear responses x_i at a given CF location i in one ear with outputs y_j from CF ($j = i$) and off-CF ($j \neq i$) cochelar locations of the other ear. Since the off-CF cochlear outputs represent the delayed versions of the responses at CF, the elements along a diagonal parallel to AB represent the correlation between the cochlear output of one ear and the spatially shifted cochlear output of the other ear. In other words, the diagonals represent the correlation of the two cochlear images at different horizontal spatial shifts.

  The distance of the correlation activity from AB signifies the amount of spatial disparity between the left-ear and the right-ear signals. Since this spatial disparity can be interpreted as the temporal disparity too, the disparity axis is important for ITD cues. Figure 3.8(b) shows the stereausis image for an off-centered signal input. Comparing it with Figure 3.8(a), we see that the

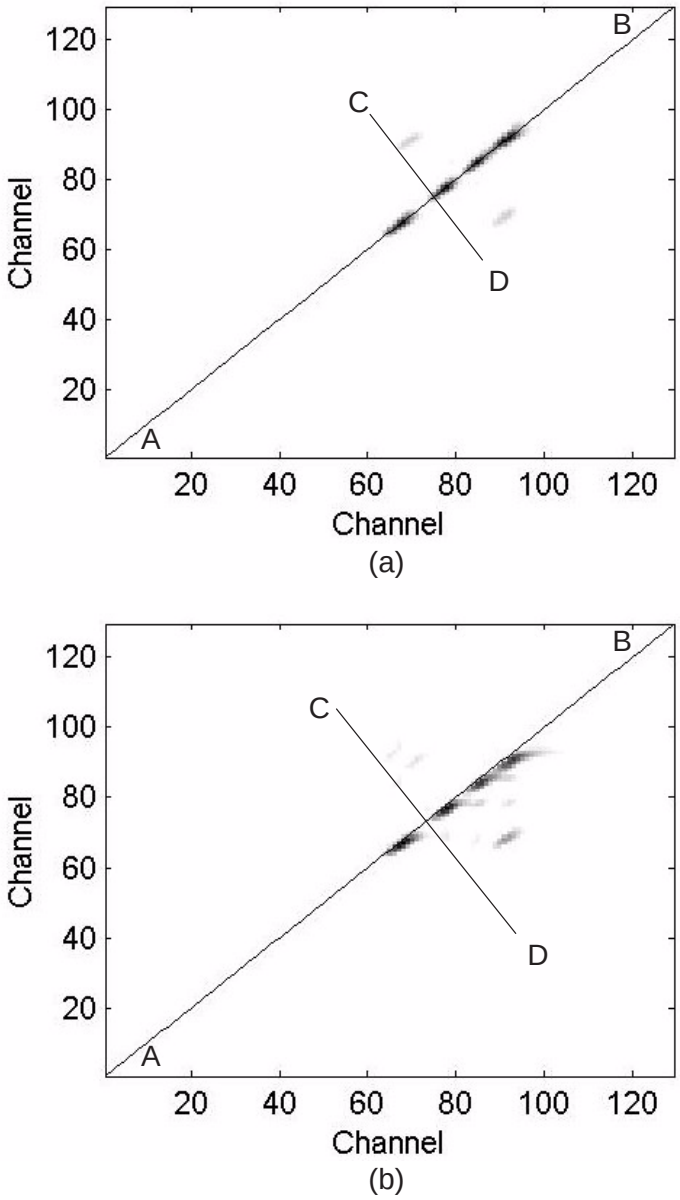

Figure 3.8: Stereausis images (a) source at 0^0 (b) source at 22.5^o.

time delay results in the shifting of the ridges along the disparity axis.

- The Spectral (CF) Axis: The spectral axis is the axis parallel to the diagonal. This axis provides information on the spectral content of the signal. In Figure 3.8, we see high activity in the channels corresponding to the frequencies in the mixture of tones.

It can be shown, under some assumptions, that the diagonals close to the center approximate the temporal correlation methods for ITD information. Indeed, consider the correlation function as follows

$$c_{ij} = C(x_i, y_j) = \sum_k x_i(k) y_j(k) \tag{3.13}$$

Assume that a narrow-band signal of low frequency ω impinges on the ear drums. Ignoring the nonlinear structures of HWR and AGC, the output signals for the i-th channel of the left-ear and the j-th channel of the right-ear can be expressed as

$$x_i(k) = A_i(\omega) \sin(\omega k T + \alpha_i(\omega)) \tag{3.14}$$

$$y_j(k) = A_j(\omega) \sin(\omega k T + \alpha_j(\omega)) \tag{3.15}$$

where $A_i(\omega)$ and $A_j(\omega)$ are the amplitude responses and $\alpha_i(\omega)$ and $\alpha_j(\omega)$ represent the phase transformations due to cochlear filters at locations i and j, respectively. The sets $\{x_i(k), i = 1, \cdots, N\}$ and $\{y_j(k), j = 1, \cdots, N\}$ represent the snapshots of the traveling waves at time k. Further, assume that the channels i and j have close characteristic frequencies. Then, from the shape of the cochlear filters (Figure 3.2) it can be seen that for low frequencies, the magnitude responses of the filters close together in space are nearly same[2]. Thus, we can assume that

[2]It is interesting to note that while the stereausis algorithm utilizes the shape and the overlap

$A_i(\omega) \approx A_j(\omega)$. Next, defining $\alpha_j(\omega) = \alpha_i(\omega) - \delta\alpha(\omega)$, we get

$$y_j(k) \approx A_i(\omega)\sin((\omega kT + \alpha_i(\omega) - \delta\alpha(\omega)) \tag{3.16}$$

Again, assuming that the velocity of the traveling waves is constant over the small distance between the i-th and j-th locations on basilar membrane, the phase difference $\delta\alpha(\omega)$ can be expresssed as

$$\delta\alpha(\omega) = \omega\tau_s \tag{3.17}$$

where τ_s is the time taken by the traveling to travel between the two locations, i and j. Thus, $y_j(k)$ can be re-written as

$$y_j(k) \approx A_i(\omega)\sin((\omega kT + \alpha_i(\omega) - \omega\tau_s) \tag{3.18}$$

$$= y_i(kT - \tau_s) \tag{3.19}$$

Thus $y_j(k)$ is a delayed version of $y_i(k)$. The spatial correlation as defined above is therefore equivalent to the temporal correlation

$$c_{ij} = \sum_k x_i(k)y_i(k - \tau_s) \tag{3.20}$$

The above analysis shows that as long as the two channels are not far enough, the elements of the image represent a measure of temporal correlation that can be used to measure the interaural time difference. In other words, the elements that are close to the center diagonal AB are important for the detection of ITD cue.

A simple method similar to the temporal-correlation method was used to measure the ITD from the spatially correlated outputs from the stereausis network. The elements along a diagonal were summed together. The sum represented the

of the cochlear filters, the algorithm by Lim and Duda [7] ignores the overlap and treats the filters as approximations to DFT.

correlation between spatially shifted cochlear outputs. The sums for different diagonals represented the correlation at different spatial shifts. They were then plotted along the disparity axis and after a post processing to suppress dual peaks in the plot, the peak was searched. The distance of the peak from the diagonal was used as a measure of ITD.

The ILD cue depends on the kind of the correlation function $C(\cdot,\cdot)$ used for forming the spatial image. We now show that the multiplicative correlation function used above does not provide good ILD cues. Assume that we have a constant ILD, so that the cochlear output for the right-ear is just a scaled value of the left-ear output, i.e,

$$y_i(k) = ax_i(k) \tag{3.21}$$

$$y_j(k) = ax_j(k) \tag{3.22}$$

where a is the scalar factor representing ILD. Then for the multiplicative correlation function, c_{ij} and c_{ji} are given by

$$c_{ij} = \sum_k ax_i(k)x_j(k) \tag{3.23}$$

$$c_{ji} = \sum_k ax_i(k)x_j(k) \tag{3.24}$$

As is obvious from the above equations, $c_{ij} = c_{ji}$. Therefore, such a correlation function will not provide any asymmetry around the diagonal AB which can be used to detect the ILD cue. Different correlation functions such as addition $C(x_i, y_j) = \sum_k (x_i(k) + y_j(k))$ can be used instead [1].

In this thesis, for the purpose of measuring the ILD we follow the same methodology as in Section 3.3. A vector consisting of ILD values is formed by taking the ratio of signal energies in each of the channels.

$$ILD_i = 10(\log_{10}(\sum_k x_i(k)x_i(k)) - \log_{10}(\sum_k y_i(k)y_i(k))) \tag{3.25}$$

The ITD and the ILDs are used to form the ITF vector. The rest of the procedure for training and estimating the direction remains the same as in the temporal correlation case[3].

[3]In this thesis, we have used a simple method for binaural processing. The stereausis image is highly informative and a much more sophisticated processing can be used to extract the sound localization information. Please refer to [12, 13] for more details.

Chapter 4

Subspace Methods

4.1 Introduction

The algorithms described in the previous chapter are based on matching of measured interaural transfer function with a known set of interaural transfer functions. These algorithms inspired us to explore statistical signal processing tools to compute the interaural transfer functions and follow the same procedure of exhaustive search for finding the closest match. The statistical methods provide effective techniques to tackle measurement noise inevitable in all practical systems.

A simple way of measuring the interaural transfer function is to compute the short-time DFT coefficients of the signals received at the two sensors and take their ratios which will give the short-time DFT coefficients for the interaural transfer. One can then average these coefficients over a period of time to get the statistical mean. This technique is akin to averaged periodogram methods which have been shown to perform poorly in comparison to parameteric methods [8]. Thus, we have used parameteric subspace methods, specifically MUSIC and ESPRIT, as the main tools for the spectral estimation of the interaural transfer function.

To get started, we will first derive the data model for the subspace methods. After that, the genesis of subspace methods, MUSIC will be described and the concept of signal subspace will be explained. Later, we will talk about ESPRIT and the methods to tackle the problem at hand.

4.2 The Data Model

The popular DOA methods including the subspace methods assume that the impinging signals are narrowband. However this assumption is not true in the case of sound signals. For the ease of presentation, the narrowband model is first derived and then extended to form the wideband model.

4.2.1 The Narrowband Model

A number of assumptions are made to simplify the derivation of the model equation. Some of these assumptions are given below.

The transmission medium is assumed to be homogenous and isotropic. The sources are assumed to be in the far field of the array. Hence the signals received by the sensors are plane waves. The signals are assumed to be sample functions of narrowband stationary processes with center frequencies, ω_i, for the i-th signal. Thus, the i-th signal $s_i(t) \in \mathcal{C}$ can be written as

$$s_i(t) \quad = \quad u_i(t)e^{j(\omega_i t + v_i(t))} \tag{4.1}$$

where $u_i(t)$ and $v_i(t)$ are "slowly varying" functions of time such that for small propagation delays τ_i, the following conditions are true

$$\begin{aligned} u_i(t-\tau_i) &\approx u_i(t) \\ v_i(t-\tau_i) &\approx v_i(t) \end{aligned}$$

Then,

$$\begin{aligned} s_i(t-\tau_i) &= u_i(t-\tau_i)e^{j(\omega_i(t-\tau_i)+v_i(t-\tau_i))} \\ &\approx u_i(t)e^{j(\omega_i t+v_i(t))}e^{-j\omega_i\tau_i} \end{aligned}$$

which can be written as

$$s_i(t-\tau_i) \approx s_i(t)e^{-j\omega_i\tau_i} \tag{4.2}$$

Thus, the effect of small time delays for narrowband signals is simply a phase-shift. Regarding the sensor array, we assume that the sensor elements and the head-related response can be modeled as linear time-invariant systems having linear transfer functions. *It is important to mention that these transfer functions have spatial characterstics which means that the transfer functions may be different for signals arriving from different directions.*

With above assumptions in place, let us consider L narrowband signals with *known* center frequencies $\{\omega_i\}_{i=1}^{L}$ impinging on an array of M sensors from directions $\{\theta_i\}_{i=1}^{L}$. Since the signal i is sampled both in time and space (by spatially distributed sensors), it is important to specify it in both parameters. So let us denote $s_i(t)$ as the value of i-th signal waveform at a reference point in space, at time t. The reference point is normally one of the sensors in the array.

Let τ_{ki} be relative delay of the i-th waveform in reaching sensor k. Then, using the superposition principle, the output of sensor k can be written as

$$z_k(t) = \sum_{i=1}^{L} h_{ki}(t) * s_i(t-\tau_{ki}) + e_k(t) \tag{4.3}$$

where $h_{ki}(t)$ is the combined impulse response of the sensor k and the head (HRTF) to signal i. The impulse response depends on the direction-of-arrival of the signal, and hence the subscript i in the impulse response. $e_k(t)$ is the additive

measurement noise. Assuming that the propagation delay, τ_{ki}, is small over the extent of the array such that (4.2) holds, the above equation can be re-written as

$$z_k(t) = \sum_{i=1}^{L} h_{ki}(t) * s_i(t)e^{-j\omega_i\tau_{ki}} + e_k(t) \quad (4.4)$$

This equation can be simplified further using the narrowband assumption. Since the spectrum of $s_i(t)$ is centered around ω_i and falls off rapidly for increasing $|\omega-\omega_i|$, the convolution with $h_{ki}(t)$ can be replaced by multiplication with complex gain $H_{ki}(\omega_i)$ at frequency ω_i. Thus, (4.4) becomes

$$z_k(t) = \sum_{i=1}^{L} H_{ki}(\omega_i)s_i(t)e^{-j\omega_i\tau_{ki}} + e_k(t) \quad (4.5)$$

Next, we introduce a few vector notations to facilitate writing the ouput equations for all M sensors in a compact form.

1. The output vector

$$z(t) = [z_1(t), \ldots, z_M(t)]^T \quad (4.6)$$

2. The signal vector

$$s(t) = [s_1(t), \ldots, s_L(t)]^T \quad (4.7)$$

3. The additive measurement noise vector

$$e(t) = [e_1(t), \ldots, e_M(t)]^T \quad (4.8)$$

4. The array steering column vector

$$a(\theta_i) = [H_{1i}(\omega_i)e^{-j\omega_i\tau_{1i}}, \ldots, H_{Mi}(\omega_i)e^{-j\omega_i\tau_{Mi}}]^T, \quad i = 1, \ldots, L \quad (4.9)$$

The center frequencies, ω_i, are assumed to be known. The gain at microphone k to signal i at frequency ω_i, $H_{ki}(\omega_i)$ depends only on the angle-of-arrival θ_i

of the i-th signal. Further, if the array geometry is assumed to be known, the propagation delays, τ_{ki}, depend only on the θ_i. Thus, the array steering vector is a function of θ_i only.

5. The array steering matrix

$$A(\theta) = [a(\theta_1), \ldots, a(\theta_L)] \tag{4.10}$$

Using the above vector notations, we can combine the output equations for all M sensors and write the output data model for narrowband signals as

$$z(t) = A(\theta)s(t) + e(t) \tag{4.11}$$

4.2.2 The Wide-band Model

We now assume that the source signals impinging on the array are wide-band. Using the same notation as in the case of narrow-band sources, the signal received at the k-th sensor can be expressed as

$$z_k(t) = \sum_{i=1}^{L} h_{ki}(t) * s_i(t - \tau_{ki}) + e_k(t) \tag{4.12}$$

Unlike the narrow-band case, it is more convenient to represent the model in frequency domain. Assume that the source signals and the received signals have a Fourier series representation, then the above relation can be expressed as

$$Z_k(\omega, \theta) = \sum_{i=1}^{L} H_{ki}(\omega) e^{-j\omega\tau_{ki}} S_i(\omega) + E_k(\omega) \tag{4.13}$$

$$= \sum_{i=1}^{L} \widetilde{H}_{ki}(\omega) S_i(\omega) + E_k(\omega) \tag{4.14}$$

where, $\widetilde{H}_{ki}(\omega) = H_{ki}(\omega)e^{-j\omega\tau_{ki}}$. In matrix notation, the above equation becomes

$$Z(\omega, \theta) = A(\omega, \theta)S(\omega) + E(\omega) \tag{4.15}$$

where,

$$Z(\omega,\theta) = [Z_1(\omega,\theta),\ldots,Z_M(\omega,\theta)]^T \tag{4.16}$$

$$S(\omega) = [S_1(\omega),\ldots,S_L(\omega)]^T \tag{4.17}$$

$$E(\omega) = [E_1(\omega),\ldots,E_M(\omega)]^T \tag{4.18}$$

and the matrix $A(\omega,\theta)$ is given by

$$A(\omega,\theta) = \begin{bmatrix} H_{11}(\omega) & \cdots & H_{1L}(\omega) \\ H_{21}(\omega)e^{-j\omega\tau_{21}} & \cdots & H_{2L}(\omega)e^{-j\omega\tau_{2L}} \\ \vdots & \ddots & \vdots \\ H_{M1}(\omega)e^{-j\omega\tau_{M1}} & \cdots & H_{ML}(\omega)e^{-j\omega\tau_{ML}} \end{bmatrix} \tag{4.19}$$

Observe that each column of frequency-domain array steering matrix $A(\omega,\theta)$ is associated with a different source. The subspace spanned by array steering matrix is called signal subspace. A quick look at (4.15) shows that in the absence of noise term $E(\omega)$, the output vector belongs to the subspace of matrix $A(\omega,\theta)$. This concept is elaborated upon in the next section. Further, note that the columns of $A(\omega,\theta)$ span different spaces at different frequencies even if the sensors and HRTF have flat, omni-directional frequency response. This property makes it extremely difficult to combine the subspaces of different frequencies for coherent DOA estimation.

4.3 MUSIC

MUSIC [5] algorithm was proposed by R. O. Schmidt. It is derived using the correlation structure of the output data. Consider the wide-band model (4.15), reproduced here for convenience.

$$Z(\omega,\theta) = A(\omega,\theta)S(\omega) + E(\omega) \tag{4.20}$$

Assuming that the vector $S(\omega)$ consists of complex envelope of $L(< M)$ uncorrelated zero-mean source signals $S_i(\omega)$ at frequency ω and the frequency-domain additive measurement noise vector $E(\omega)$ is white with zero mean and variance σ^2, we can express the correlation matrix of $Z(\omega, \theta)$ as follows

$$R(\omega, \theta) \quad = \quad A(\omega, \theta) R_s(\omega) A^H(\omega, \theta) + \sigma^2 I \tag{4.21}$$

where I is the M-by-M identity matrix, $R_s(\omega)$ is the correlation matrix of signal vector $S(\omega)$, and the superscript H denotes transpose and complex conjugation. Since the signals $S_i(\omega)$ are uncorrelated, $R_s(\omega)$ is a diagonal matrix

$$R_s(\omega) \quad = \quad \text{diag}\{P_1(\omega), \ldots, P_L(\omega)\} \tag{4.22}$$

where $P_i(\omega) = E[|S_i(\omega)|^2]$, $i = 1, \ldots, L$ is the spectral power density of the i-th signal.

To this end, let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$ denote the eigenvalues of the correlation matrix $R(\omega)$, and $\nu_1 \geq \nu_2 \geq \ldots \geq \nu_M$ denote the eigenvalues of the matrix $A(\omega, \theta) R_s(\omega) A^H(\omega, \theta)$, then from (4.21), we get

$$\lambda_i \quad = \quad \nu_i + \sigma^2, \qquad\qquad i = 1, 2, \ldots, M \tag{4.23}$$

Now, since $R_s(\omega)$ is a diagonal matrix with rank L, the smallest $(M - L)$ eigenvalues of $A(\omega, \theta) R_s(\omega) A^H(\omega, \theta)$ are zero, i.e, $\nu_{L+1} = \ldots = \nu_M = 0$. Rewrite (4.21) in the following form

$$R(\omega, \theta) - \sigma^2 I \quad = \quad A(\omega, \theta) R_s(\omega) A^H(\omega, \theta) \tag{4.24}$$

Then, if $\{q_l(\omega, \theta), l = 1, \cdots, M\}$ denote the eigenvectors of $R(\omega, \theta)$, we get

$$(R(\omega, \theta) - \sigma^2 I) q_l(\omega, \theta) \quad = \quad A(\omega, \theta) R_s(\omega) A^H(\omega, \theta) q_l(\omega, \theta) \tag{4.25}$$

Equivalently,

$$\lambda_l q_l(\omega,\theta) - \sigma^2 q_l(\omega,\theta) \;=\; A(\omega,\theta)R_s(\omega)A^H(\omega,\theta)q_l(\omega,\theta) \tag{4.26}$$

From the above discussion, it follows that the eigenvectors of $R(\omega,\theta)$ associated with the smallest $(M-L)$ eigenvalues satisfy the following relationship

$$A(\omega,\theta)R_s(\omega)A^H(\omega,\theta)q_l(\omega,\theta) \;=\; 0, \qquad l = L+1,\ldots,M \tag{4.27}$$

Since the matrix $R_s(\omega)$ is a real-valued diagonal matrix of full rank, it follows from the above equation that

$$A^H(\omega,\theta)q_l(\omega,\theta) \;=\; 0, \qquad l = L+1,\ldots,M \tag{4.28}$$

or equivalently from the definition of $A^H(\omega,\theta)$,

$$a^H(\omega,\theta_i)q_l(\omega,\theta) \;=\; 0, \qquad l = L+1,\ldots,M, \quad i = 1,\ldots,L \tag{4.29}$$

If we define $Q_N(\omega,\theta)$ and $Q_S(\omega,\theta)$ as follows

$$Q_N(\omega,\theta) \;=\; [q_{L+1}(\omega,\theta),\ldots,q_M(\omega,\theta)], \tag{4.30}$$

$$Q_S(\omega,\theta) \;=\; [q_1(\omega,\theta),\ldots,q_L(\omega,\theta)] \tag{4.31}$$

Then from (4.29), we get

$$Q_N^H(\omega,\theta)a(\omega,\theta_i) \;=\; 0, \qquad i = 1,\ldots,L \tag{4.32}$$

From the above equation, we make the following key observation. The DOAs, θ_i, are the roots of the following equation

$$a^H(\omega,\theta_i)Q_N(\omega,\theta)Q_N^H(\omega,\theta)a(\omega,\theta_i) \;=\; 0, \qquad i = 1,\ldots,L \tag{4.33}$$

4.3.1 Signal and Noise Subspaces

From the relation (4.32), we note that the steering vector $a(\omega, \theta_i)$ belongs to the null space of $Q_N(\omega, \theta)$ which is denoted by $a(\omega, \theta_i) \in \mathcal{N}(Q_N(\omega, \theta))$. Also, since the correlation matrix $R(\omega, \theta)$ is hermitian, its eigenvectors are mutually orthogonal and hence,

$$Q_N^H(\omega, \theta) Q_S(\omega, \theta) \;=\; 0 \tag{4.34}$$

Thus if we denote the range space of $Q_S(\omega, \theta)$ by $\mathcal{R}(Q_S(\omega, \theta))$,

$$\mathcal{R}(Q_S(\omega, \theta)) \;=\; \mathcal{N}(Q_N(\omega, \theta)) \tag{4.35}$$

and hence,

$$a(\omega, \theta_i) \;\in\; \mathcal{R}(Q_S(\omega, \theta)) \tag{4.36}$$

The range space, $\mathcal{R}(Q_S(\omega, \theta))$, spanned by the first L eigenvectors (associated with the L largest eigenvalues) of the output correlation matrix $R(\omega, \theta)$ is called the *signal subspace* and the space, $\mathcal{R}(Q_N(\omega, \theta))$, spanned by the last $(M - L)$ eigenvectors is called the *noise subspace*.

4.3.2 Direction-of-Arrival Estimation

The equation (4.33) provides a straightforward way of estimating the directions. Suppose $a(\omega, \theta)$ is known for the complete range of ω and θ and the noise subspace, $Q_N(\omega, \cdot)$, obtained from the data received from the microphones. Then, an exhaustive search is done to find L array steering vectors which are most orthogonal to the noise subspace, i.e, the L vectors which provide the least values to the expression $a^H(\omega, \theta_i) Q_N(\omega, \cdot) Q_N^H(\omega, \cdot) a(\omega, \theta_i)$. The angles corresponding to the L most orthogonal array steering vectors are the estimated directions of the sources.

The determination of array steering vectors, $a(\omega, \theta)$, however requires precise knowledge of HRTFs and the sensor responses and, thus, is a long, time-consuming process. It is preferable to use (4.34) instead of (4.33) since the signal subspace can be determined using a similar system set-up as needed for estimation. We now describe a practical implementation of localization system based on MUSIC.

Training Phase

As in the biological algorithms, we go through a training process to determine the training set of signal subspace $Q_S(\omega, \theta)$ for different values of $\theta \in \Theta$. Due to practical limitations, the training set can only be recorded for discrete frequencies $\omega = \omega_n$. Hence, we denote the training set consisting of signal subspaces as $\{Q_S(\omega_n, \theta), \omega_n \in B, \theta \in \Theta\}$, where B is the set of discrete frequencies covering the bandwidth of the signals.

A broadband white noise is transmitted and the data received by the microphones is recorded at different angles of arrival of the white noise. For each angle θ, the data in time series is converted to frequency domain by taking the DFT for frequencies ω_n. The correlation matrix $R(\omega_n, \theta)$ is formed. The eigenvector corresponding to the largest eigenvalue of $R(\omega_n, \theta)$ is the estimated signal subspace $Q_S(\omega_n, \theta)$ at angle θ. Such a set of vectors representing signal subspaces form the training data and is later used during the estimation of the directions of unknown sources.

Estimation Phase

The important steps in the estimation phase of MUSIC algorithm are summarized as follows:

1. Convert the time series of output data obtained from the sensors into frequency domain, $\widehat{Z}(\omega_n, \widehat{\theta})$, $\omega_n \in B$. The notation $\widehat{\theta} = [\widehat{\theta}_1, \cdots, \widehat{\theta}_L]$ is the set of unknown directions.

2. Estimate the frequency domain correlation matrix

$$\widehat{R}(\omega_n, \widehat{\theta}) \;=\; \widehat{Z}^H(\omega_n, \widehat{\theta})\widehat{Z}(\omega_n, \widehat{\theta})$$

3. Compute the eigenvectors of $\widehat{R}(\omega_n, \widehat{\theta})$ using eigendecomposition methods.

4. Estimate the matrix $\widehat{Q}_N(\omega_n, \widehat{\theta})$ from the $(M - L)$ eigenvectors associated with the smallest $(M - L)$ eigenvalues[1].

5. Assuming that the training process is already completed, determine the angles-of-arrivals, $\widehat{\theta}_i$, by searching for the subspace vectors which are orthogonal to the noise subspace. In spectral MUSIC, the search is performed by plotting the following expression, known as pseudo-spectrum, over the domain of θ.

$$J(\omega_n, \theta) \;=\; \frac{1}{Q_S^H(\omega_n, \theta)\widehat{Q}_N(\omega_n, \widehat{\theta})\widehat{Q}_N^H(\omega_n, \widehat{\theta})Q_S(\omega_n, \theta)} \qquad (4.37)$$

   Then it follows from (4.34), that the angles, $\widehat{\theta}_i$, can be estimated as the peaks in the plot of $J(\omega_n, \theta)$.

6. Since each frequency ω_n will provide a pseudo-spectrum $J(\omega_n, \theta)$ and its own set of directions, a scheme is needed to combine the results of all the

[1]In the above analysis we assumed that L is known. But there are many practical situations in which the number of sources is unknown. In that case, we need to estimate L. One way to do this is to calculate the eigenvalues of $\widehat{R}(\omega, \widehat{\theta})$ and taking the number of eigenvalues greater than a pre-defined threshold as the number of sources.

frequencies. The approach of "averaging" the pseudo-spectrums is followed in order to take into account all the frequencies. The final estimates of directions are obtained from the peaks of the average:

$$J(\theta) = \frac{1}{\sum_{\omega_n} Q_S^H(\omega_n,\theta)\widehat{Q}_N(\omega_n,\widehat{\theta})\widehat{Q}_N^H(\omega_n,\widehat{\theta})Q_S(\omega_n,\theta)} \tag{4.38}$$

4.4 ESPRIT

The ESPRIT [6] algorithm is another popular subspace alogrithm for direction-of-arrival problem. It was introduced by Roy and Kailath and is similar to MUSIC in that it exploits the underlying data model. The ESPRIT algorithm can achieve significant computational and storage cost advantages over MUSIC by requiring that the sensors occur in matched pairs with identical displacement vectors. However, in our case, the sensors see different HRTFs and the transfer characteristics of the two sensors are not same. A search procedure as in the previous methods is employed to overcome this problem which reduces the computational advantages of ESPRIT over MUSIC. Still, the processing in ESPRIT is different from MUSIC. Moreover, it is closer in spirit to the biological algorithms and incorporates the concept of interaural transfer function naturally.

Consider an array of M sensors in which the sensors can be grouped in doublets such that the displacement between the sensor elements in a doublet is constant both in magnitude and direction for all the doublets. The exact location of the doublet pairs in the space is not important.

It shall be convenient to visualize the array as being comprised of two subarrays $\mathcal{Z}_X$ and $\mathcal{Z}_Y$, identical in geometry and response characteristics but translated in space by a fixed displacement vector. Let Δ be the displacement vector with

magnitude d. Under this special array geometry, we redefine our data model (4.11) as follows.

Assume that the array is illuminated by L wide-band plane waveforms. Let $H_{ki}(\omega)$ be the complex response of the first sensor in the kth doublet and the corresponding HRTF to the i-th wavefront incident from the direction θ_i measured with respect to the normal to the displacement vector Δ. Then, proceeding as in previous section, the Fourier transform of the output signal received at the first sensor of the kth doublet can be expressed as

$$X_k(\omega,\theta) \;=\; \sum_{i=1}^{L} H_{ki}(\omega)e^{-j\omega\tau_{ki}}S_i(\omega) + E_{x_k}(\omega) \tag{4.39}$$

where τ_{ki} is the propagation delay for waveform i from the reference point to the first element in the k-th doublet and $E_{x_k}(\omega)$ represents the additive measurement noise. Since the second sensor element in the doublet is displaced further by the distance d from the first element, the signal received by the second sensor will be delayed further by time $d\sin\theta_i/c$, where c is the speed of propagation of sound. The received signal at the second sensor is given by

$$Y_k(\omega,\theta) \;=\; \sum_{i=1}^{L} H_{ki}(\omega)e^{-j\omega\tau_{ki}}\widetilde{F}(\omega,\theta_i)e^{-j\omega d\sin\theta_i/c}S_i(\omega) + E_{y_k}(\omega) \tag{4.40}$$

where $F(\omega,\theta_i) = \widetilde{F}(\omega,\theta_i)e^{-j\omega d\sin\theta_i/c}$ represents the "interaural transfer function" between the two sensors in the kth doublet. The additive measurement noise corresponding to $E_{x_k}(\omega)$, $E_{y_k}(\omega)$ are uncorrelated with signals and are assumed to be stationary zero-mean spatially white random processes with a *known* covariance.

Using the matrix notation to express equations (4.39) and (4.40) for $k = 1,\ldots,M/2$, the output of the array can be expressed as

$$X(\omega,\theta) \;=\; A(\omega,\theta)S(\omega) + E_x(\omega) \tag{4.41}$$

$$Y(\omega,\theta) \;=\; A(\omega,\theta)\Phi(\omega,\theta)S(\omega) + E_y(\omega) \tag{4.42}$$

where,

$X(\omega,\theta) \in \mathcal{C}^{M/2}$ is the output vector of first sensors at frequency ω

$Y(\omega,\theta) \in \mathcal{C}^{M/2}$ is the output vector of second sensors in each doublet

$A(\omega,\theta) \in \mathcal{C}^{M/2\times L}$ is an unknown matrix of array steering vectors

$E_x(\omega), E_y(\omega) \in \mathcal{C}^{M/2}$ are the measurement noise vectors

$\Phi(\omega,\theta) = \text{diag}\ \{|\widetilde{F}(\omega,\theta_1)|e^{-j\phi_1(\omega)}, \ldots, |\widetilde{F}(\omega,\theta_L)|e^{-j\phi_L(\omega)}\}$,

$\Phi(\omega,\theta)$ is the matrix of interaural transfer functions with

$$\phi_i(\omega,\theta_i) = \omega\frac{d\sin\theta_i}{c} + \arg(\widetilde{F}(\omega,\theta_i)) \tag{4.43}$$

Next, combining equations (4.41) and (4.42), we define the data model for the problem as

$$Z(\omega,\theta) = \begin{bmatrix} X(\omega,\theta) \\ Y(\omega,\theta) \end{bmatrix} = \bar{A}(\omega,\theta)S(\omega) + E(\omega) \tag{4.44}$$

$$\bar{A}(\omega,\theta) = \begin{bmatrix} A(\omega,\theta) \\ A(\omega,\theta)\Phi(\omega,\theta) \end{bmatrix}, \quad E(\omega) = \begin{bmatrix} E_x(\omega) \\ E_y(\omega) \end{bmatrix} \tag{4.45}$$

The objective is to estimate the number of signals L and the directions-of-arrival θ_i. For this it is sufficient to estimate the matrix $\Phi(\omega,\theta)$. It is the structure of the matrix $\bar{A}(\omega,\theta)$ that is exploited to obtain $\Phi(\omega,\theta)$ *without having to know* the exact HRTF and the sensor transfer functions in $A(\omega,\theta)$.

Proceeding as in the case of MUSIC algorithm, we compute the correlation matrix $R(\omega,\theta)$ of total array output vector $Z(\omega,\theta)$. The L eigenvectors of $R(\omega,\theta)$, denoted by $\{q_l(\omega,\theta),\ l = 1,\ldots,L\}$ corresponding to L largest eigenvalues are used to obtain the signal subspace $\mathcal{R}(Q_S(\omega,\theta))$ at frequency ω, where $Q_S(\omega,\theta) = [q_1(\omega,\theta),\ldots,q_L(\omega,\theta)]$. Since $\mathcal{R}(Q_S(\omega,\theta)) = \mathcal{R}(\bar{A}(\omega,\theta))$, there must exist a unique

nonsingular matrix T such that

$$Q_S(\omega,\theta) \;=\; \bar{A}(\omega,\theta)T \;=\; \begin{bmatrix} A(\omega,\theta)T \\ A(\omega,\theta)\Phi(\omega,\theta)T \end{bmatrix} \tag{4.46}$$

Partitioning $Q_S(\omega,\theta)$ into $Q_X \in \mathcal{C}^{M/2\times L}$ and $Q_Y \in \mathcal{C}^{M/2\times L}$, we get

$$\begin{bmatrix} Q_X \\ Q_Y \end{bmatrix} \;=\; \begin{bmatrix} A(\omega,\theta)T \\ A(\omega,\theta)\Phi(\omega,\theta)T \end{bmatrix} \tag{4.47}$$

It follows from the above equation that

$$\mathcal{R}(Q_X) = \mathcal{R}(Q_Y) = \mathcal{R}(A(\omega,\theta)) \tag{4.48}$$

Next, define

$$Q_{XY} \;=\; [Q_X|Q_Y] \tag{4.49}$$

Since Q_X and Q_Y span the same column space, the rank of Q_{XY} is L. Since $Q_{XY} \in \mathcal{C}^{M/2\times 2L}$, the null space of Q_{XY} has dimension L. Let $G \in \mathcal{C}^{2L\times L}$ of rank L span the null space of Q_{XY}, denoted by $\mathcal{N}(Q_{XY})$, then

$$[Q_X|Q_Y]G \;=\; 0 \tag{4.50}$$

Partitioning $G^H = [G_X^H|G_Y^H]$, where $G_X, G_Y \in \mathcal{C}^{L\times L}$, we get

$$Q_X G_X + Q_Y G_Y \;=\; 0 \tag{4.51}$$

$$\Rightarrow A(\omega,\theta)TG_X + A(\omega,\theta)\Phi(\omega,\theta)TG_Y \;=\; 0 \tag{4.52}$$

Since G is rank L, the matrices G_X and G_Y are nonsingular. Define

$$\Psi \;=\; -G_X G_Y^{-1} \tag{4.53}$$

Then (4.52) can be expressed as

$$-A(\omega,\theta)T\Psi + A(\omega,\theta)\Phi(\omega,\theta)T \;=\; 0 \tag{4.54}$$

Rearranging, we get

$$A(\omega,\theta)\Phi(\omega,\theta) = A(\omega,\theta)T\Psi T^{-1} \tag{4.55}$$

Finally, assuming $A(\omega,\theta)$ to be full rank, we get

$$\Phi(\omega,\theta) = T\Psi T^{-1} \tag{4.56}$$

Thus the eigenvalues of Ψ are the same as the diagonal elements of $\Phi(\omega,\theta)$.

4.4.1 Direction-of-Arrival Estimation

Given the diagonal elements of $\Phi(\omega,\theta)$, the interaural trasfer functions $F(\omega,\theta_i)$ can be determined from (4.43). However, the term $\widetilde{F}(\omega,\theta_i)$ in the interaural transfer function depends on the response characteristics of the sensors and the surroundings and is generally unknown; thus making it impossible to directly find the directions, θ_i. We, therefore, follow the same two-phase scheme as in the the case of earlier methods.

Training Phase

In the first phase, the system undergoes a training process that provides a set of interaural transfer functions, $\{F(\theta), \theta \in \Theta\}$. The notation $F(\theta)$ represents the vector $\{F(\omega_n,\theta), \omega_n \in B\}$.

A broadband white noise is transmitted and the data received by the microphones is recorded at different angles of arrival of the white noise. For each angle θ, the data in time series is converted to frequency domain by taking the DFT for frequencies ω_n. For each frequency ω_n, the correlation matrix $R(\omega_n,\theta)$ is formed. The eigenvector corresponding to the largest eigenvalue of $R(\omega_n,\theta)$ gives estimated signal subspace $Q_S(\omega_n,\theta)$ at angle θ. Then, the equations (4.49), (4.50), (4.53) and

(4.56) are used to compute $F(\omega_n, \theta)$. Repeating the process for all $\omega_n \in B$ and $\theta \in \Theta$ provides the training set.

Estimation Phase

The computational steps in the estimation phase of ESPRIT are summarized as follows:

1. Convert the time series of output data obtained from the sensors into frequency domain, $\widehat{Z}(\omega_n, \widehat{\theta})$, $\omega_n \in B$. The notation $\widehat{\theta} = [\widehat{\theta}_1, \cdots, \widehat{\theta}_L]$ is the set of unknown directions.

2. Estimate the frequency-domain correlation matrix for the frequency bin ω_n, $\widehat{R}(\omega_n, \widehat{\theta}) = \widehat{Z}^H(\omega_n, \widehat{\theta})\widehat{Z}(\omega_n, \widehat{\theta})$.

3. Compute the eigenvectors of $\widehat{R}(\omega_n, \widehat{\theta})$ using eigendecomposition methods.

4. Estimate the matrix $\widehat{Q}_S(\omega_n, \widehat{\theta})$ from the L eigenvectors associated with the largest L eigenvalues.

5. Partition $\widehat{Q}_S(\omega_n, \widehat{\theta}) = [\widehat{Q}_X^H | \widehat{Q}_Y^H]^H$ and form matrix $\widehat{Q}_{XY}$.

6. Compute the null space of $\widehat{Q}_{XY}$, given by $\widehat{G}$.

7. Partition $\widehat{G}^H = [\widehat{G}_X^H | \widehat{G}_Y^H]^H$ and compute $\widehat{\Psi} = -\widehat{G}_X \widehat{G}_Y^{-1}$.

8. Compute the L eigenvalues of matrix $\widehat{\Psi}$.

9. Case $L = 1$: In the case of single source, there will be just one eigenvalue of $\widehat{\Psi}$. Repeating the steps (2-8) for all frequency bins, $\omega_n \in B$, a vector of eigenvalues is formed, represented by $\widehat{F}$. Assuming that the training set $\{F(\theta), \theta \in \Theta\}$ is available, the nearest-neighbor approach as in Chapter 3 is followed to estimate the location of the source.

10. Case $L > 1$: When $L > 1$, the picture becomes much more complicated. In the case of more than one sources, the ESPRIT algorithm will give a set of L eigenvalues, $F(\omega_n, \widehat{\theta}_i), \{i = 1, \cdots, L\}$, for frequency ω_n. Repeating the steps (2-8) for all frequencies $\in B$, similar sets of eigenvalues are obtained. The process of obtaining the eigenvalues for one frequency bin is independent of that for another frequency bin. It is, thus, not clear how to associate these eigenvalues with the sources. In order to solve this problem, the nearest neighbor estimator is extended and a global search on all the possible associations of the eigenvalues with the sources is performed to find the vectors that minimize the distance measure between the training set of interaural transfer functions and the estimated interaural transfer functions from the received data.

4.5 Tracking of moving source

Continuous tracking is required in applications in which the source is moving. In such a case, the localization system needs to continuously update the direction of the source. In subspace methods, the intermediate step for estimating the direction is the computation of signal subspace. As the direction of the source changes, so does the signal subspace. An approach to update the signal subspace is to apply a forgetting factor $0 < \beta < 1$ that damps out the effects of the older data and gives higher weightage to more recent data.

$$\widehat{R}_t(\omega_n) \quad = \quad \beta \widehat{R}_{t-1}(\omega_n) + (1-\beta)\widehat{Z}_t^H(\omega_n, \widehat{\theta}_t)\widehat{Z}_t(\omega_n, \widehat{\theta}_t) \tag{4.57}$$

where $\widehat{Z}_t(\omega_n, \widehat{\theta}_t)$ is the short-time DFT vector computed from the latest data obtained from the microphones, and t is the running time index. The updated

matrix $\widehat{R}_t(\omega_n)$ is used to compute the new signal subspace using eigendecomposition methods. This method, however, does not make use of the results of the previous subspace computations for $\widehat{R}_{t-1}(\omega_n)$ and is highly expensive in terms of the computational cost.

A better approach is to use the data matrix for subspace computation instead of the correlation matrix. The data matrix is formed by computing $\widehat{Z}_t(\omega_n, \widehat{\theta}_t)$ at discrete instants of time t and stacking them in a form of matrix. In order to take into account the forgetting factor, the old data matrix is multiplied by forgetting factor before appending a column of newly arrived data.

$$\widehat{D}_t(\omega_n) \;=\; \left[\beta\widehat{D}_{t-1}(\omega_n)\vdots(1-\beta)\widehat{Z}_t(\omega_n,\widehat{\theta}_t)\right] \tag{4.58}$$

It can be easily seen that $\widehat{R}_t(\omega_n) = \widehat{D}_t^H(\omega_n)\widehat{D}_t(\omega_n)$.

We then follow the URV Decomposition method introduced by Stewart [10]. He showed that there exists a matrix decomposition, called the URV decomposition, of $\widehat{D}_t(\omega_n)$ which is of the form

$$\widehat{D}_t^H(\omega_n) \;=\; U_t^H\Lambda_t V_t \tag{4.59}$$

where $U_t \in \mathcal{C}^{t\times M}$ is the left orthogonal matrix, $\Lambda_t \in \mathcal{C}^{M\times M}$ is a right triangular matrix and $V_t \in \mathcal{C}^{M\times M}$ is the right orthogonal matrix. The matrix Λ_t can be written as

$$\Lambda_t \;=\; \begin{bmatrix} \Sigma & \Gamma_1 \\ 0 & \Gamma_2 \end{bmatrix} \tag{4.60}$$

and satisfies the following properties.

1. Σ and Γ_2 are upper triangular,

2. smaller singular value of $\Sigma \approx \sqrt{\lambda_L}$,

3. $\sqrt{\|\Gamma_1\|^2 + \|\Gamma_2\|^2} \approx \sqrt{\lambda_{L+1} + \cdots + \lambda_M}$.

where $\lambda_1 \geq \cdots \lambda_L > \lambda_{L+1} \geq \cdots \lambda_M$ represent the eigenvalues of $\widehat{D}_t(\omega_n)$. Under these conditions, it can be easily proved that the subspace spanned by V_t is equal to the space spanned by the eigenvectors of $\widehat{R}_t(\omega_n)$ [10]. Moreover, the subspace spanned by the first L columns of V_t is approximately equal to the subspace spanned by the L eigenvectors of the correlation matrix $\widehat{R}_t(\omega_n)$ and therefore represent the signal subspace.

Everytime a new row of data $\widehat{Z}^H_{t+1}(\omega_n, \widehat{\theta}_{t+1})$ is added to the data matrix, the matrices U_t, Λ_t and V_t need to be updated. The updating of the signal subspace does not require the knowledge of the matrix U_t. So in the computations of the signal subspace, U_t is completely ignored. This results in huge savings in the computational cost and the storage requirements. The updating of Λ_t and V_t with the arrival of a new row is an $O(n^2)$ process. It is to be compared with the eigenvalue decomposition of the correlation matrix and the singular value decomposition method which are of the order of $O(n^3)$ process. Though, these methods are generally more accurate than the URV decomposition method, Liu *et. al* [9] showed that the results are comparable. For more details on the computations involved in URV decomposition, please refer to [10] and [9].

Chapter 5

Results and Discussion

The experimental studies were conducted to evaluate the perfomance of the localization algorithms described in the chapters. The following two set ups were considered.

1. The effects of the head and the outer ears were simulated by using experimental HRTF measurements from KEMAR manikin [26]. A wide-band noise source was used as the sound source which was convoluted with the KEMAR HRTFs to simulate the output data of the sensors. Since the KEMAR HRTF measurements were done in a controlled anechoic environment, the simulations in this case follow rather ideal conditions.

2. The data was collected from the Scout robot (Figure 2.2) at the sampling rate of 40 kHz. A wide band noise was generated and reproduced from a speaker kept at around 3 meters from the robot. The measurements in this case were done in a highly-reverberant room environment.

5.1 Experimental results for KEMAR

A wide-band source was simulated at angle 20^o. The sensor measurement errors were introduced by zero mean normal additive noise with signal-to-noise ratio (SNR) of 20 dB. In the training mode, however, the sensor errors were assumed to be zero. The HRTF provided data at step-size of 5^o. Therefore, the training set consisted of 72 ITF vectors.

For temporal-correlation based (Lim-Duda), spatial-correlation based (stereausis) and ESPRIT methods, we considered three cases where ITD only, ILD only and both ITD and ILD cues were respectively used for direction estimation. The histograms of the estimated directions were obtained. The results are shown in Figure 5.1 thru Figure 5.4.

5.2 Experimental results for Scout robot

A wide-band signal was produced from the speaker kept at direction 18^o from the normal of the line connecting the two microphones on the robot dummy head. The data collected from the microphones was used for estimating the directions. For the training mode, the same environment was used and the data was collected by rotating the robot in step-size of 3^o at angles $\{0, 3, 6, \cdots, 357\}$. The histogram of the estimated direction are shown in Figure 5.5 thru Figure 5.8.

As earlier, for temporal-correlation based (Lim-Duda), spatial-correlation based (stereausis) and ESPRIT methods, we considered three cases where ITD only, ILD only and both ITD and ILD cues were respectively used for direction estimation.

5.3 Discussion

The results indicate that the estimates of the subspace based methods are unbiased compared to the biological methods. The subspace methods also provided higher percentage of accurate estimates and therefore lower variance. In the case of KEMAR, experiments were performed with decreasing values of SNR. It was found that the degradation of performance for the biological methods was higher than that of subspace methods. In fact, at SNR of 6 dB, the biological methods failed to localize the sources, whereas the subspace methods provided reasonably accurate results.

Among the biological algorithms, it seems that the temporal-correlation methods are better suited than spatial-correlation based methods. However, it is possible that the simplified method of extracting ITD used in this thesis failed to capture the necessary information embedded in the stereausis image.

Among the subspace methods, ESPRIT estimates showed higher variance and lesser percentage accuracy than those of MUSIC. This points to the fact that the projection method in MUSIC performed better than the least-mean-square method used in ESPRIT for comparing the proximity of the estimation data from the training set.

It is interesting to note that the cone of confusion effect (the reason for peaks around 160^o) is very visible in the KEMAR case, even when ILD only and no ITD is used for estimation. This is due to the similarity in the attenuation characteristics of the front part and the rear part of the KEMAR dummy head. In the case of Scout robot, there was no such symmetry. Therefore, although the peaks due to cone effect were present in the ITD only case, the ILD only case didn't show considerable effect. This helped in achieving better localization when both ITD

and ILD were used.

The experiments show that localization accuracy is much better in the KEMAR case than the Scout robot measurements. This can be easily explained by the fact that in the latter case the environment is highly echoic. The echoes can be viewed as virtual sources. Such an environment presents multiple weak sources which violates the assumptions made in the algorithms[1]. Furthermore, these virtual sources are correlated which results in further degradation in the performance. Nevertheless, the results show that good localization can be achieved, especially by subspace methods, even in an extremely reverberant environment by using training/estimation approach.

[1]The subspace algorithms can handle multiple sources. However, in our experiments, the number of sensors, $M = 2$. Therefore, the maximum number of sources that subspace algorithm can localize is one.

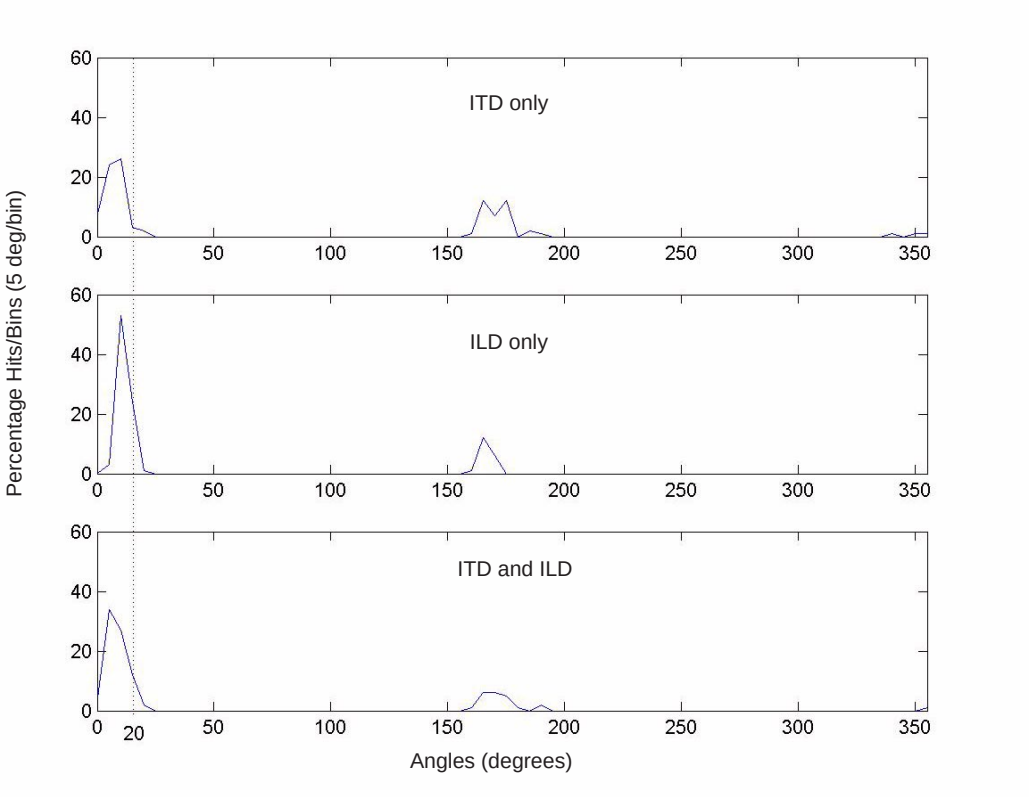

Figure 5.1: Histograms of temporal-correlation method for KEMAR

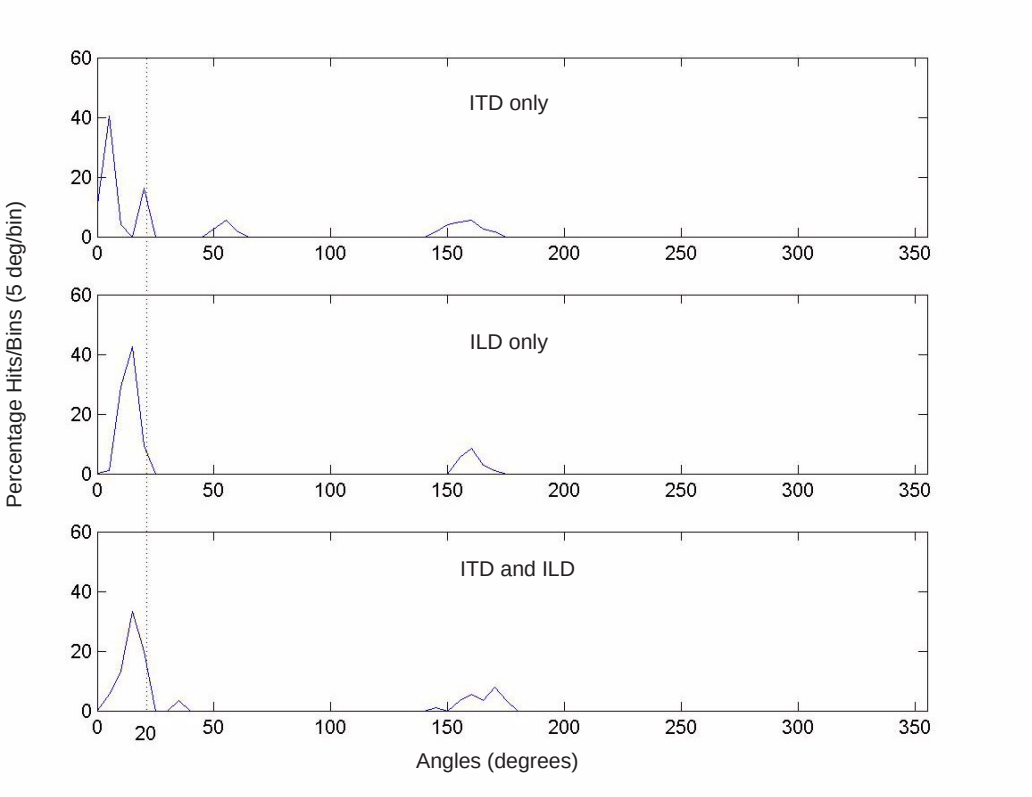

Figure 5.2: Histograms of spatial-correlation method for KEMAR

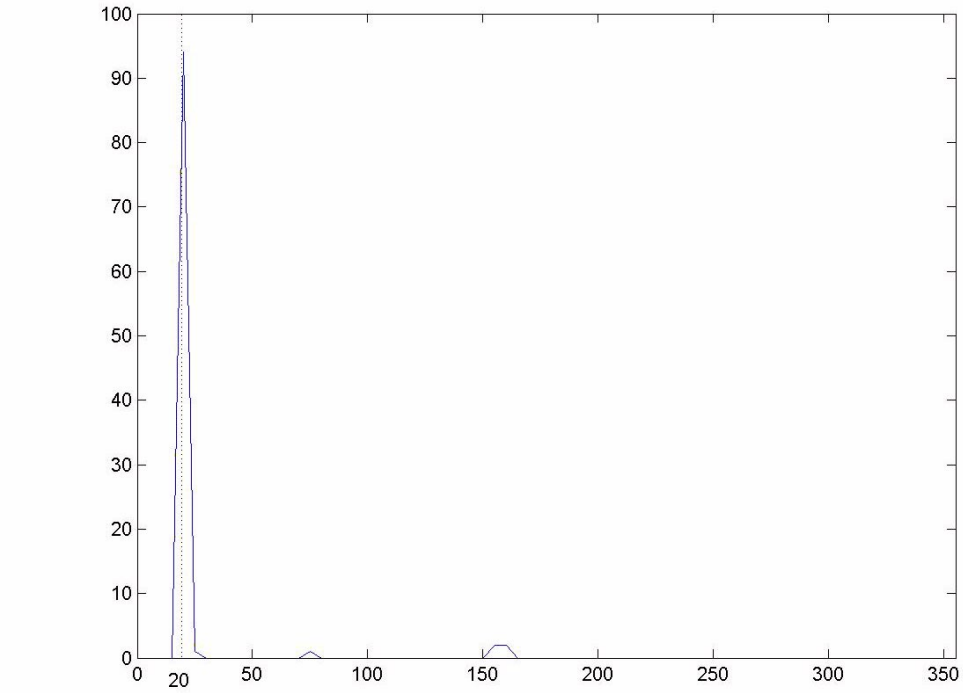

Figure 5.3: Histograms of MUSIC method for KEMAR

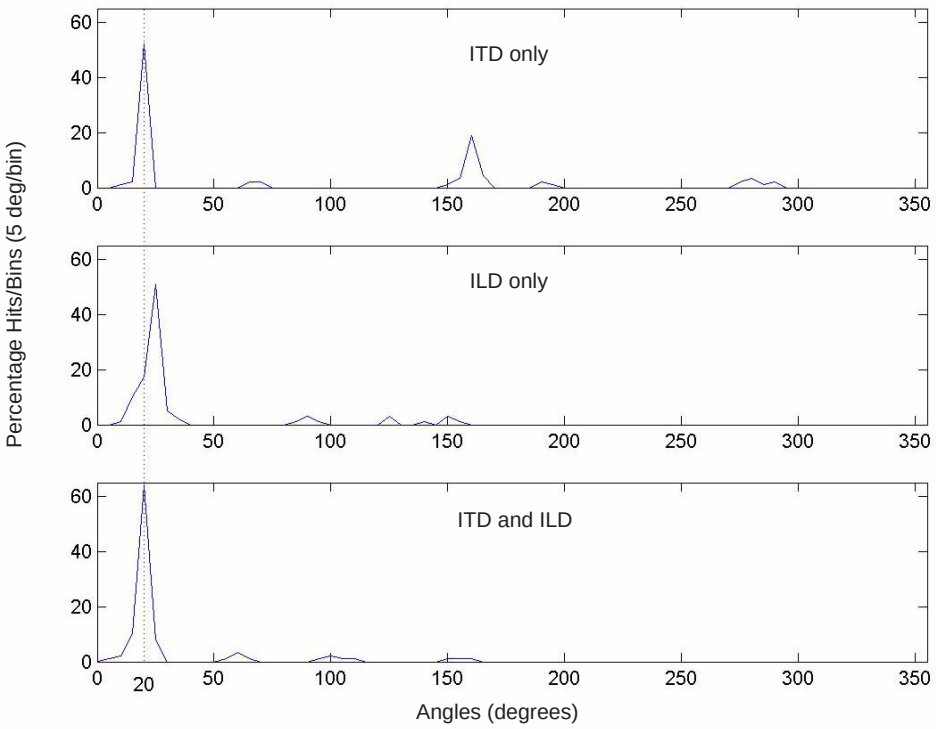

Figure 5.4: Histograms of ESPRIT method for KEMAR

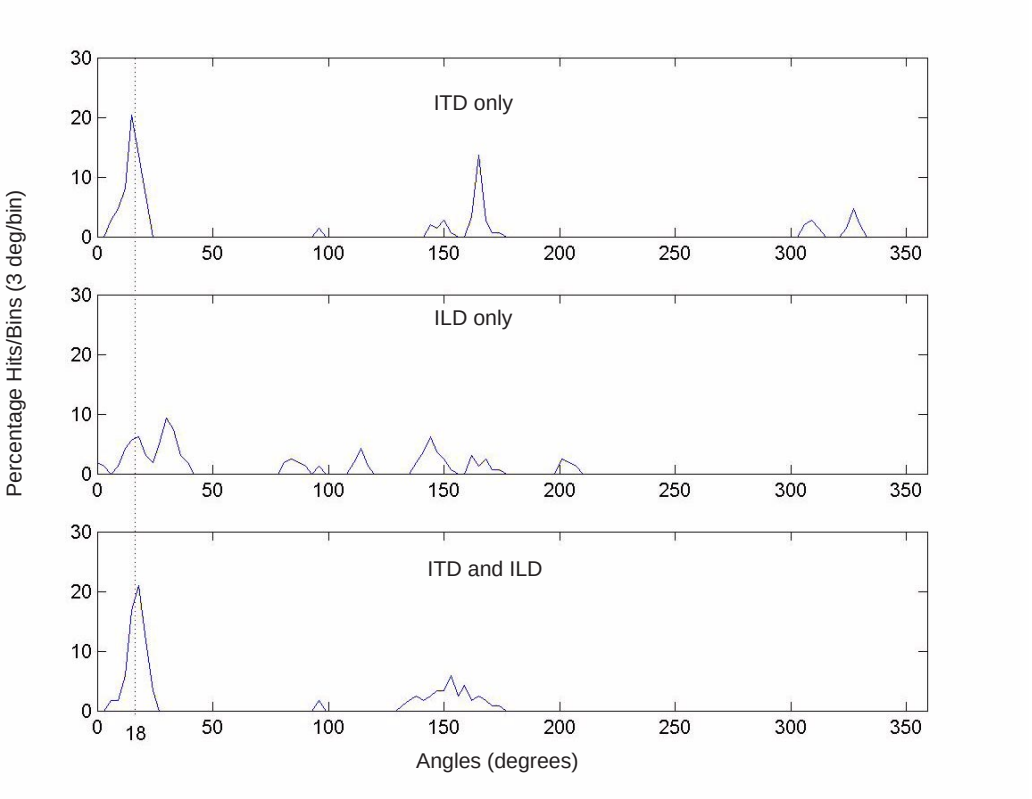

Figure 5.5: Histograms of temporal-correlation method for Scout robot

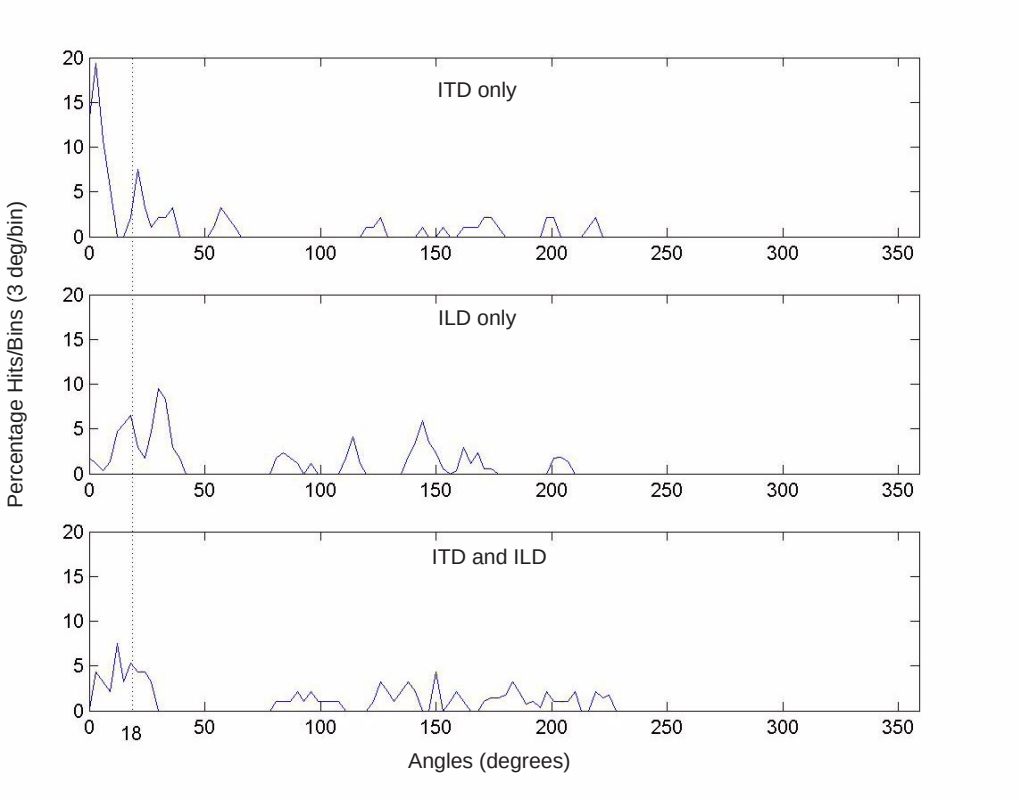

Figure 5.6: Histograms of spatial-correlation method for Scout robot

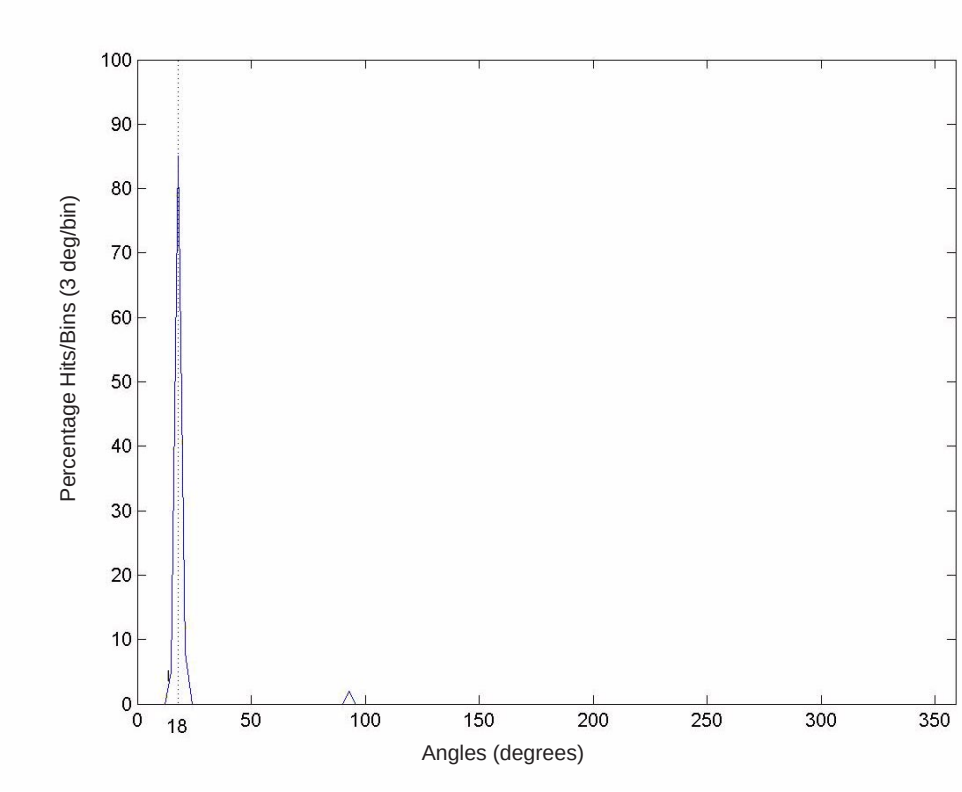

Figure 5.7: Histograms of MUSIC method for Scout robot

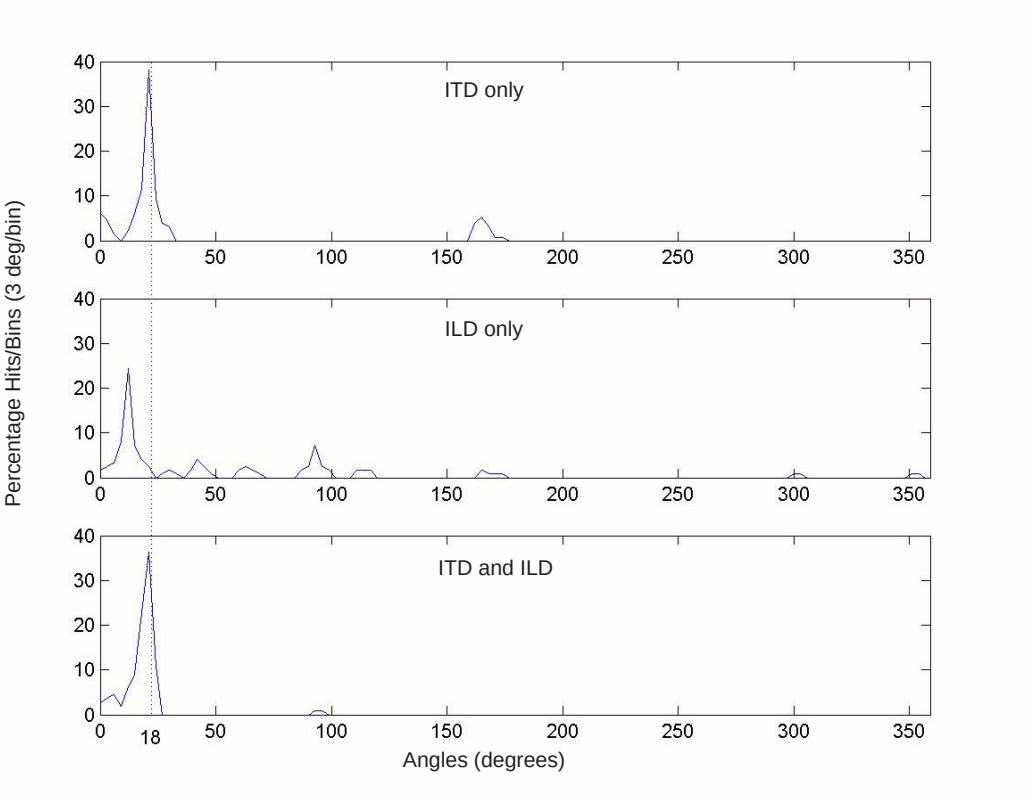

Figure 5.8: Histograms of ESPRIT method for Scout robot

BIBLIOGRAPHY

[1] S. Shamma, N. Shen, P. Gopalaswamy, *Stereausis: binaural processing without neural delays*, J. Acoustic Soc. Am., Vol. 86, pp. 989-1006, Sept. 1989.

[2] R. F. Lyon, *A computational model of filtering,detection and compression in the cochlea*, Proc. of the IEEE Int. Conf. Acoust., Speech, Signal Processing, Paris, France, May 1982.

[3] J. O. Pickels, *An Introduction to the Physiology of Hearing*, Academic Press, London, 1982.

[4] M. R. Schroeder, *An integrable model for the basilar membrane*, JASA 53, pp. 429-434, 1973.

[5] R. O. Schmidt, *A signal subspace approach to multiple emitter location and spectral estimation*, Ph.D. dissertation, Stanford Univ, CA, 1981.

[6] R. Roy and T. Kailath, *ESPRIT-Estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol.37 no. 7, July 1989.

[7] C. Lim and R. O. Duda, *Estimating the Azimuth and Elevation of a Sound Source from the Output of a Cochlear Model*, Proc. 28th Asilomar Conf. on Signals, Systems and Computers (Asimolar, CA, 1994).

[8] P. Stoica and R. L. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, 1997.

[9] K. J. Ray Liu, D. P. O'Leary, G. W. Stewart and Yuan-Jye J. Wu, *URV ESPRIT for Tracking Time-Varying Signals*, IEEE Trans. Signal Processing, vol. 42(12), Dec. 1994.

[10] G. W. Stewart, *An updating algorithm for subspace tracking*, IEEE Trans. Signal Processing, vol. 40, pp. 1535-1541, 1992.

[11] J. Licklider, *A duplex theory of pitch perception*, Experientia 7, pp. 128-134, 1951.

[12] E. Grassi, D. Euston, T. Takahashi, S. Shamma, *Model for Sound Localization in the Barn Owl*, Society for Neuroscience Annual Meeting, New Orleans, Oct. 2000.

[13] D. J. Rapczynski, *A Robotic Implementation of a Two-Dimensional Human Sound Localization Model Employing Neural Networks*, Scholarly Paper, University of Maryland, College Park, Dec. 2000.

[14] L. Jeffress, *A place theory of sound localization*, J. Comp. Physiol. Psych. 61, pp. 468-486, 1948.

[15] S. Haykin, *Adaptive Filter Theory* Prentice Hall, NJ, second ed., 1991.

[16] M. Wax, T. J. Shan, and T. Kailath, *Spatio-temporal spectral analysis by eigenstructure methods*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-32, pp. 817-827, Aug. 1984.

[17] G. Su and M. Morf, *Signal subspace approach for multiple wide-band emitter location*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 31, pp. 1502-1522, Dec. 1983.

[18] G. Su and M. Morf, *Modal decomposition signal subspace algorithms*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 34, pp. 585-602, June 1986.

[19] B. Ottersten and T. Kailath, *Direction-of-arrival estimation for wide-band signals using the ESPRIT algorithm*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 38, pp. 317-327, Feb. 1990.

[20] K. M. Buckley and L. J. Griffiths, *Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, pp. 953-964, July 1988.

[21] H. Wang and M. Kaveh, *Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. ASSP-33, pp. 823-831, Aug. 1985.

[22] H. Hung and M. Kaveh, *Focussing matrices for coherent signal-subspace processing*, IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, pp. 1272-1281, Aug. 1988.

[23] M. A. Doron and A. J. Weiss, *On focusing matrices for wide-band array processing*, IEEE Trans. Signal Processing, vol. 40, pp. 1295-1302, June 1992.

[24] E. Doron and M. A. Doron, *Coherent wideband array processing*, Proc ICASSP, 1992.

[25] M. Slaney, *Lyon's cochlear model*, Apple Technical Report No. 13, Advanced Technology Group, Apple Computer Inc., Cupertino, CA (1998).

[26] W. G. Gardner and K. D. Martin, *HRTF measurements of a KEMAR dummy-head microphone*, MIT Media Lab Perceptual Computing Technical Report #280, 1994. *http://sound.media.mit.edu/KEMAR.html*.