August 5, 2005

To the Graduate School:

This thesis entitled "Acoustic Localization by Interaural Level Difference" and written by Rajitha Gangishetty is presented to the Graduate School of Clemson University. I recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science with a major in Electrical Engineering.

Dr. Stanley Birchfield, Thesis Advisor

We have reviewed this thesis
and recommend its acceptance:

Dr. John Gowdy

Dr. Hiren Maharaj

Accepted for the Graduate School:

ACOUSTIC LOCALIZATION BY INTERAURAL LEVEL DIFFERENCE

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Electrical Engineering

---

by
Rajitha Gangishetty
August 2005

Advisor: Dr. Stanley Birchfield

ABSTRACT

Interaural level difference (ILD), an important cue for acoustic localization is one of the phenomena used by the human auditory system to locate a sound source. It refers to the amplitude difference in the signals that reach the two ears from a sound source. Although its behavior has been studied extensively in natural systems, it remains an untapped resource for computer-based ones.

In this thesis, the possibility of using ILD for acoustic localization is investigated by deriving constraints on the location of a sound source given the relative energy level of the signals received by two microphones. This localization is realized by using the Inverse Square Law, which states that energy of a unit area emanating from a point source is inversely proportional to the square of the distance from the source.

An algorithm is presented for computing the sound source location by combining likelihood functions for multiple microphone pairs, one for each pair. This computation is done by using a probabilistic sampling method in which a number of candidate locations in space are selected and for each, the likelihood that the sound source is present there is computed. The total likelihood is the sum of the likelihoods for each microphone pair. Experimental results show that accurate acoustic localization can be achieved using ILD alone even under high reverberant conditions. But when reverberation exists along with extremely high noise conditions, of magnitudes almost equivalent to the signal level, ILD fails to localize the sound source accurately. Preliminary results indicate that a small improvement is obtained using a Hilbert Envelope approach.

DEDICATION

This thesis is dedicated to my mother and father who have always been supportive in my endeavors.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Birchfield, for his invaluable guidance and for providing me with the opportunity and the resources to pursue this thesis topic. I would also like to thank Dr. Gowdy and Dr. Maharaj, my committee members, for their technical guidance during my course of study at Clemson. I also would like to thank Mrs. Barbara Ramirez for reviewing my thesis.

Above all, I would like to thank my parents and my husband for their constant support, encouragement and love, without which I would never have been able to accomplish this work and reach this far.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

List of Figures (Continued)

Figure Page

List of Figures (Continued)

Figure Page

CHAPTER 1

INTRODUCTION

Sound is produced by a rapid variation in the average density or pressure of air molecules above and below the current atmospheric pressure. We perceive sound as these pressure fluctuations cause our eardrums to vibrate. When discussing sound, these fluctuations in pressure are referred to as sound waves.

Characteristics of Sound Waves

Sound waves are often characterized by four basic qualities: frequency, amplitude, wavelength and phase. Frequency is the number of cycles per unit of time. For convenience, it is most often measured in cycles per second, also referred to as Hertz (Hz). The range of human hearing is approximately 20 Hz to 20 kHz. Figure 1.1 shows a sinewave, the one to the left with a frequency of four cycles per second and the one to the right with a frequency of 8 cycles per second.

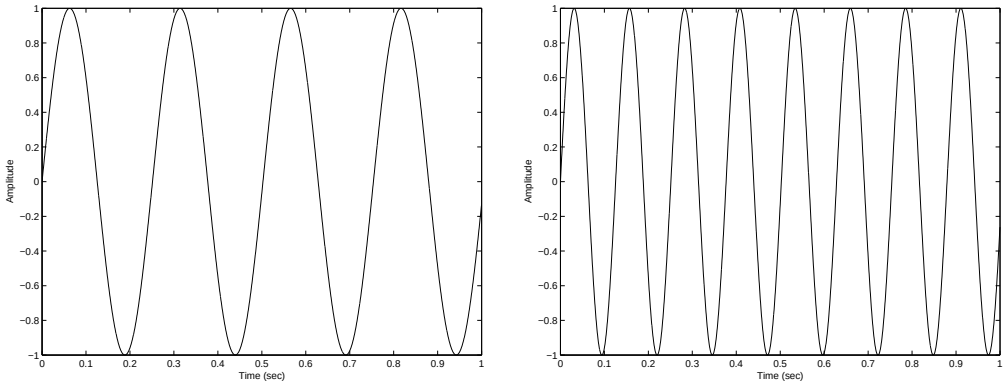

Figure 1.1 Two sine waves with different frequencies

Amplitude measures the amount of positive or negative change, in atmospheric pressure. It is measured in the amount of force applied over an area, the most common unit of measurement for acoustic waves being Newtons per square meter (N/m^2). Amplitude, also referred to as intensity of sound, is directly related to acoustic energy whose measurement is Newton per meter (N/m). A high energy wave is characterized by a high amplitude,

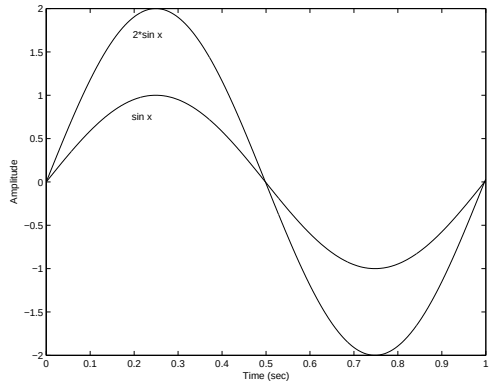

Figure 1.2 Two sine waves with different amplitudes

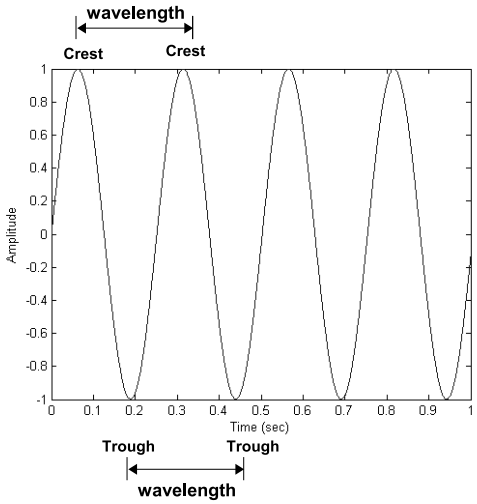

Figure 1.3 Wavelength for a sine wave

whereas a low energy wave is characterized by a low amplitude. Figure 1.2 shows two sine waves, the smaller one with an amplitude of one and the larger with an amplitude of two.

The wavelength, the distance from crest to crest or, equivalently, from trough to trough of a wave, is inversely proportional to frequency. Higher frequencies have shorter wavelengths while lower frequencies have longer ones. Figure 1.3 depicts the definition of wavelength pictorially. The last characteristic of sound, phase, denotes the particular point in the cycle of a waveform, measured as an angle in degrees. It is normally not an audible characteristic of a single wave, but can be when very low-frequency waves are used as controls in synthesis. It is a very important factor in the interaction of one wave with

another, both acoustically and electronically. Figure 1.4 shows two sine waves with a phase difference of 90 degrees.

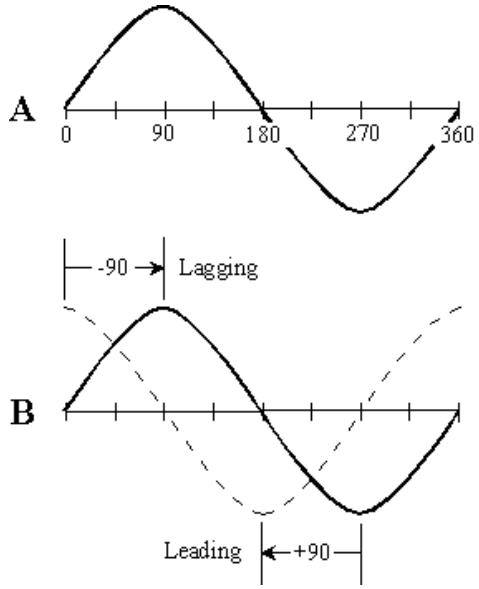

Figure 1.4 Two sine waves with a phase difference of 90 degrees

The Nature of Speech

As one of the most natural forms of communication between humans, speech is a subject which has attracted much research, especially over the past twenty years. The structure of speech, its production and perception mechanisms have long occupied linguists, psychologists and physiologists, with scientists and engineers endeavoring to construct machines to synthesize, recognize, and localize human speech. Recently, this goal has begun to be realized, though the systems that have been built are still a long way from being able to emulate human performance, because the problem is very difficult and the precise way in which human speech is produced requires further study. The following description, borrowed from [24] and from various sources on the internet, provides a brief introduction to speech sounds.

Speech sounds can be divided into three broad classes: voiced, unvoiced, and plosives, depending on the mode of excitation. Voiced sounds, the sounds made in the pronunciation

of *aah* or *oh*, for example, are produced when the vocal cords are tensed together and vibrate as the air pressure builds up, forcing the glottis to open and then subside as the air passes through it. The vibration of the cords produces an airflow waveform which is approximately triangular in shape. Being periodic, or at least quasi-periodic, this waveform has a frequency spectrum of rich harmonics at multiples of the fundamental frequency of vibration, or pitch frequency, and decaying at a rate of approximately 12dB/octave. The vocal tract acts as a resonant cavity amplifying some of these harmonics and attenuating others to produce voiced sounds. The range of pitch for an adult male is from approximately 50Hz to 250Hz, with an average value of approximately 120Hz. For an adult female the upper limit of the range is much higher, perhaps as high as 500Hz, the lower range being 50Hz.

Unlike for voiced sounds, in the production of unvoiced sounds, the vocal cords do not vibrate. The two basic types of unvoiced sounds are fricative sounds and aspirated sounds. For fricative sounds, for example *s* or *sh*, a point of constriction is created in the vocal tract and as air is forced past it, turbulence occurs, causing a random noise excitation. Since the points of constriction tend to occur near the front of the mouth, the resonances of the vocal tract have little effect on characterizing the fricative sound being produced. In aspirated sounds, for example the *h* of *hello*, the turbulent airflow occurs at the glottis because the vocal cords are held significantly apart. As a result, the resonances of the vocal tract modulate the spectrum of the random noise, and the effect clearly heard in whispered speech.

For plosive sounds, for example the *puh* at the beginning of the word *pin* or the *duh* at the beginning of *din*, the vocal tract is closed at some point; the air pressure is allowed to build up and then is suddenly released, providing a transient excitation of the vocal tract. This transient excitation occurs with or without vocal cord vibration to produce voiced (such as din) or unvoiced (such as pin) plosive sounds.

Source-filter Model of Speech Production

One of the earliest models depicting the production of speech was designed by F. J. Owens. A very simple model of the vocal tract is a uniform tube or pipe of length L, with a

sound source at one end (the vocal cords) and open at the other (the lips) as seen in Figure 1.5 below:

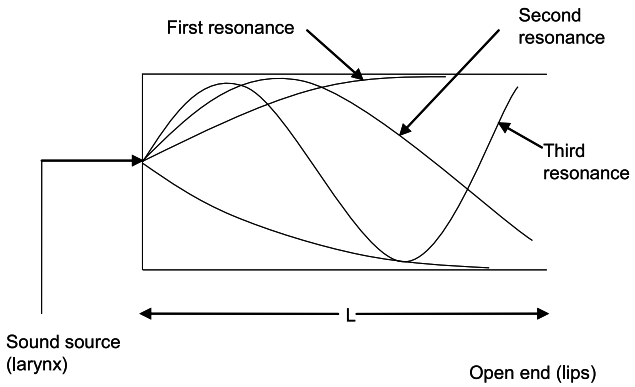

Figure 1.5 Uniform tube (pipe) model of vocal tract

Such a pipe has odd frequency resonances of f_0, $3f_0$, $5f_0$,...etc, where $f_0 = c/4L$, with c being the velocity of sound in air. In a typical vocal tract, assuming length L=17cm and c=340m/s, which usually changes with temperature and humidity, the resonant frequency values are 500Hz, 1000Hz, 1500Hz...etc. which are referred to as formants. Since, the vocal tract can take many different shapes which give rise to different resonant or formant frequency values and hence different sounds, the formant frequencies are constantly changing in continuous speech.

The preceding discussion leads to the idea of viewing the speech production processes in terms of a source-filter model (Figure 1.6) in which a signal from a sound source, either periodic pulses or random noise, is filtered by a time-varying filter with resonant properties similar to the vocal tract. Thus, the frequency spectrum of the speech signal can be obtained by multiplying the source spectrum by the frequency characteristics of the filter as illustrated in Figure 1.6 for both voiced and unvoiced speech with the gain controls A_V and A_N determining the intensity of the voiced and unvoiced excitations, respectively.

Although the vocal tract has an infinite number of resonances or formants it is only necessary to consider the first three or four, covering the range of 100Hz to approximately 3.5kHz, since the amplitudes of the higher formants in the speech signal have a high frequency roll-off of approximately -12dB/octave and thus are negligible. For an unvoiced

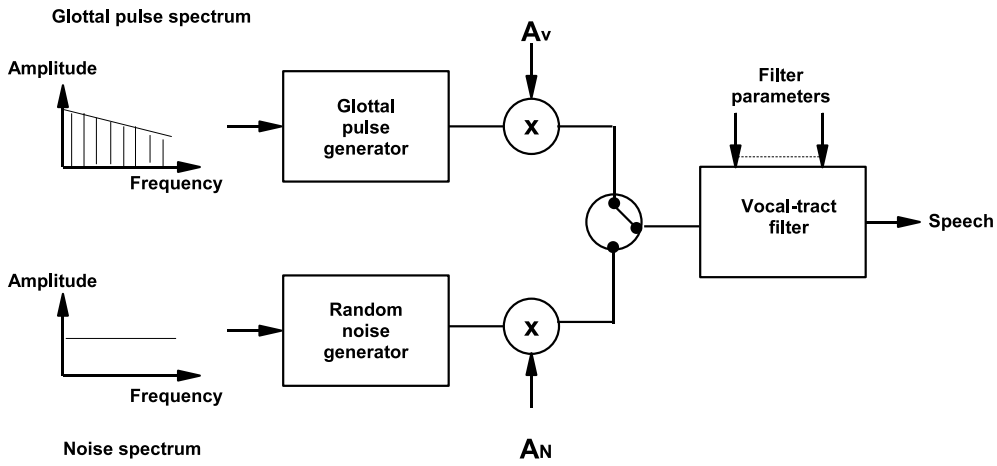

Figure 1.6 Source-Filter Model of Speech Production

source, the spectrum of which is relatively broad and flat, the same number of formants is appropriate although for proper modeling of unvoiced speech it is often necessary to extend the frequency range of interest to approximately 7 or 8 kHz. In addition to these transmission characteristics, the filter in the source-system model seen in Figure 1.6 models the effects of radiation from the mouth approximately as a first order high pass characteristic, increasing at a rate of 6dB/octave in the range 0-3kHz.

However, this source-filter model is an over-simplification of the speech production process. Fricative sounds produced when turbulent air flow occurs at a point of constriction in the vocal tract are not filtered by the resonances of the vocal tract to the same extent as voiced and aspirated sounds are. Consequently the source-filter model is not a very accurate representation for these sounds. In addition, the source-filter model assumes that the source is linearly separable from the filter with no interaction between them. This assumption is not strictly true since the vibration of the vocal cords is affected by the sound pressure inside the vocal tract and there is a coupling between the vocal tract and the lungs when the glottis is open, thereby modifying the filter characteristics every cycle of the excitation. However, these secondary factors are ignored very often, and the source-filter model is quite adequate.

Acoustic Localization: The Problem and Solution

If the study of sound production is one half, then the study of hearing these sounds is the other. In simple terms the ability of being able to locate a sound source when a signal reaches the ears is called *acoustic localization*. According to Jens Blauert, "Acoustic localization is the law or rule by which the location of an auditory event (e.g., its direction or distance) is related to a specific attribute or attributes of a sound event, or of another event that is in some way correlated with the auditory event" ([16], pg. 37). The source of sound can be localized in the three spatial dimensions: the horizontal plane, the vertical plane and in distance. Sound localization is, therefore, the result of the human or computer auditory system's ability to process the physical parameters of sounds that correlate with the spatial location of the their sources. Figure 1.7 depicts this method of localization, showing a sound source, a microphone array (which could be either compact or distributed) and the measurements in the three spatial dimensions.

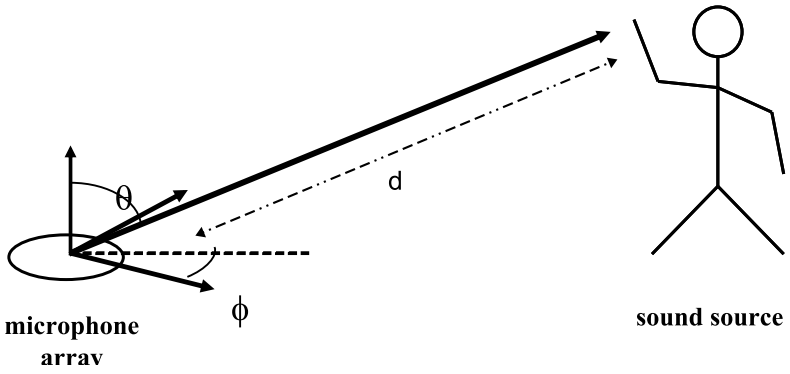

Figure 1.7 The problem of Localization

The human auditory system uses three audible cues for this purpose. The first is the interaural time difference (ITD) also called interautal phase difference (IPD), which refers to the difference in time it takes a sound to reach one ear compared to the other. Sounds located directly in front of or behind a listener will reach both ears simultaneously. If the angle of the source is moved until the difference is greater than 20 microseconds, a difference in location can be perceived. As a source moves more directly to one side of the head or the other, the ability to discriminate its location using the ITD method diminishes.

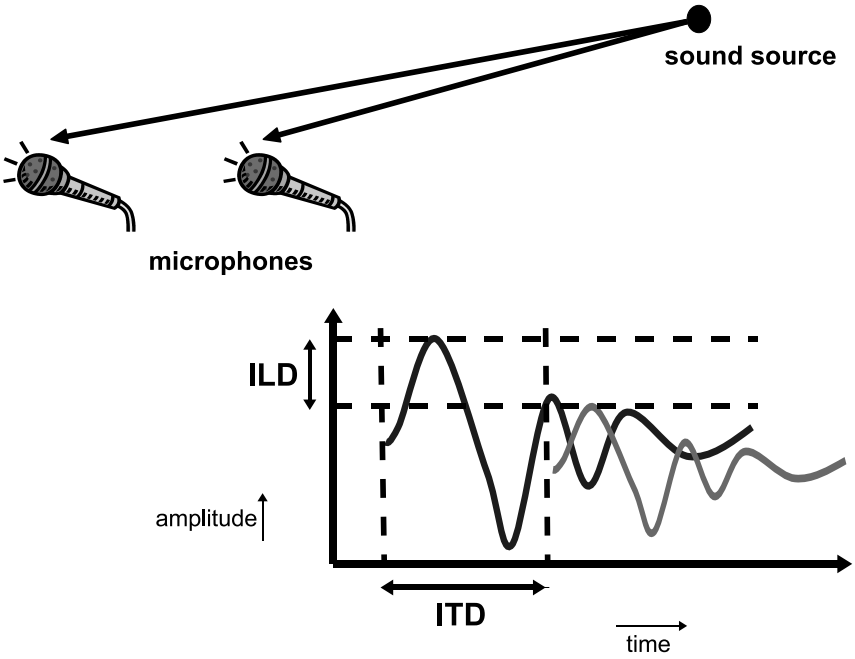

Figure 1.8 ILD and ITD

A second mechanism, called the interaural level difference (ILD), is based on the fact that the signals received by microphones not only differ in their relative time shift but also in their intensity level, with the microphone closer to the sound source receiving a higher-intensity signal than that received by a farther microphone. ILD forms the basis of the "intensity-difference theory" of directional hearing, which is the oldest theory of directional hearing going back more than 100 years [16]. It uses the difference in amplitude caused by the head physically masking sounds coming from one side or the other. Level differences between the two ear inputs in a free sound field depend significantly on frequency, a restriction that must be considered in generalizing experimental results. Even when input signals with level differences independent of frequency are used, the excitation is nonlinear, taking into consideration that with weak signals, components of one ear input signal might lie below the threshold of audibility, while the same components for the other ear are still perceptible. Because lower frequencies with longer wavelengths refract more easily around objects, this mechanism is more effective for higher frequencies. According

to Kietz (1953) [16], the auditory event for signals of any frequency moves “completely to one side” when the interaural level difference is 15-20 dB. ITD, on the other hand, is more effective for lower frequencies. ILD and ITD together form the basis for duplex theory of localization [16]. Both these phenomena are shown in Figure 1.8, with he horizontal axis representing time and the vertical axis representing amplitude. This figure shows a sound source, two microphones, and the amplitude and time differences between the signals arriving at the two microphones.

A third mechanism includes the shape of the pinna (outer ear flap) filtering frequencies depending on their angle of incidence, including the ability to place sounds in the vertical plane. A blindfolded person cannot estimate the height of the sound source accurately when the ear flaps are folded. This mechanism helps to resolve this ambiguity. All of these mechanisms are ineffective below approximately 270 Hz.

Previous Work

In recent times, there has been significant research done to use the above mentioned cues to perform localization in computer-based systems. Broadly, three approaches have been developed to solve the localization problem using computers. Of these, two of the more common methods for determining the location of a sound source are beamforming [4, 5, 6], and time-delay estimation (TDE) [1, 2, 3], in addition to accumulated correlation [8, 9]. In beamforming the original signal is reconstructed at a hypothesized location by shifting the signals from the microphones and totaling them. The energy of this reconstructed signal provides the likelihood that the sound source is present at a hypothesized location. Although accurate, it is computationally expensive because this likelihood has to be computed for all possible locations.

Time-delay estimation, also known as time difference of arrival (TDOA), is a two-step process. In the first step, the signals from each of the microphone pairs are correlated, the peak of each correlation being used to obtain the estimate of the time delay for each microphone pair. These estimates together are then used to determine the location of the sound source. The main advantage of the method is its speed, while its disadvantage is poor performance in highly reverberant environments.

Accumulated correlation [8, 9] combines the advantages of the previous two methods. Like TDE, the algorithm first computes the cross-correlation between the signals from the pairs of microphones. Instead of taking the peak of each correlation vector independently, however, all correlation vectors are mapped to a common system before finding the peak of the accumulated function. By accumulating all the available information before making a decision, the algorithm is able to provide an optimal solution, following the principle of least commitment. In this respect, it is similar to the beamforming, though it is far more computationally efficient.

Techniques for computer-based acoustic localization have therefore, to date, relied exclusively upon ITD. For example, the methods of time-delay estimation (TDE) [1, 2, 3], beamforming [4, 5, 6], hemisphere sampling [7], and accumulated correlation [8, 9] are different ways of utilizing the relative shifts in the signals received by microphones to determine the location of the sound source. A significant amount of research has also been conducted to discover prefilters to make such computations robust to noise [11, 12, 13, 14].

Motivation for this Thesis

As opposed to the earlier methods, ILD, has received little or no attention in the signal processing community. Although it is now known that ILD is not the only cue for acoustic localization, extensive psychoacoustic and psychophysical experiments have shown it to be an important cue used by the human localization system [16, 17]. Despite its importance in nature, including the localization systems of birds such as owls [18], no technique utilizing ILD has yet been proposed for computer-based systems.

In this thesis, a preliminary investigation into the possibility of using ILD in computer-based systems for acoustic localization is presented. A model is derived for computing the likelihood that a sound is placed in a particular location using only the relative energies received by microphones without any information as to their relative phase. From this formulation an algorithm is proposed to compute the sound source location using multiple microphones. Microphone-arrays are preferred here to single pair microphone systems because of their advantages over the latter. Although humans can perform localization with two microphones for computer-based systems, it is easier with multiple pairs. Micro-

phone array systems can be used to determine the positions of active talkers and can be electronically steered to provide spatially selective speech acquisition. Since it is steered electronically, a microphone array's directivity pattern can be updated rapidly to follow a moving talker or to switch between several alternating or simultaneous speakers. These features make microphone arrays an attractive alternative to single microphone systems for hands-free speech acquisition, especially those involving multiple or moving sources. The ability of microphone-array systems to determine sound source location makes them attractive for use for multimedia teleconferencing where the location of the talker can be used not only for steering the directivity of the microphone-array but also for pointing cameras or determining binaural cues for stereo imaging [21]. The algorithm developed here is experimentally tested to demonstrate its ability to localize accurately a sound source even in reverberant and noisy environments and to highlight several issues regarding ILD.

Issues

Though experimental results show that a sound source can be localized accurately using ILD alone, sometimes under highly reverberant conditions this localization is not accurate. The sound waves reaching the listener's ear directly from the source are collectively referred to as the direct sound. These waves reach the listeners ears first in most acoustic environments. In addition, the listener also hears reflected sounds, the first of such waves being called early reflections. Since they travel a longer path, the amount of time it takes the first reflected sounds to reach our ears give us clues as to the size and nature of the listening environment. Because the reflected sound may continue to bounce off many surfaces, a continuous stream of sound fuses into a single entity, which continues after the original sound ceases. This stream of continuing sound is called reverberation. The rate of build-up of this echo density is proportional to the square root of the volume of the room.

The energy of these reverberated signals depends on the position of the listener in the room as well as on the position of the sound source relative to the listener. In normal rooms, if the sound source is more than approximately three feet from the listener, the "critical distance" [15] for an ordinary microphone, the total energy of reverberation exceeds the energy of direct sound. At approximately thirty feet, the combined energy of echoes from

various directions becomes a hundred times the energy of the desired signal [15]. This reverberation in the signals reaching a listener's ears, in addition to other background noise, influences auditory localization performance.

The Objective and Overview of the Thesis

This research derives an ILD algorithm to localize accurately a sound source, the location of which is considered to be unknown, within a closed room, in the presence of noise and reverberation.

This work is comprised of the following three parts:

1. Formulating the ILD algorithm using the concept of the inverse square law.
2. Performing ILD localization with a sound source and four microphones in a closed room.
3. Making the algorithm robust to noise and reverberation.

The next chapter discusses the formulation of the ILD algorithm, describing in detail the behaviour of sound signals and how they affect the localization of the sound source. Chapter 3 develops the localization method, including the combined likelihood and nonlinear processing approaches. Chapter 4 provides the simulation results obtained for different specifications such as the size of the room, noise, reverberation and the distance between the source and the microphone. Finally, Chapter 5 explores the implications of this work, including suggestions for future study.

CHAPTER 2

ILD FORMULATION

To formulate the ILD algorithm [10], it is assumed that there are N microphones and a source signal $s(t)$ propagating through a generic free space with noise. According to the inverse-square-law, the signal received by the ith microphone can be modeled as

$$x_i(t) = s(t)/d_i + \xi_i(t),$$

where d_i is the distance from the source to the ith microphone and $\xi_i(t)$ is additive white Gaussian noise. To focus on the ILD cue, this formula ignores the relative time shift between the signals that is important for ITD.

Assuming that the sound source is audible and in a fixed location during the time interval $[0, W]$, where W is the window size, the energy received by the ith microphone can be obtained by integrating the square of the signal over this time interval:

$$\begin{aligned} E_i &= \int_0^W x_i^2(t)\,dt = \int_0^W [s^2(t)/d_i^2 + \xi_i^2(t)]\,dt \\ &= \frac{1}{d_i^2}\int_0^W s^2(t)\,dt + \int_0^W \xi_i^2(t)\,dt, \end{aligned}$$

because the integration of the cross-term is zero if $\xi_i(t)$ is uncorrelated and zero-mean. From this equation the name of the inverse-square-law is apparent: the received energy is inversely proportional to the square of the distance to the source.

Given two microphones, this equation leads to a simple relationship between the energies and distances:

$$E_1 d_1^2 = E_2 d_2^2 + \eta, \tag{2.1}$$

where $\eta = \int_0^W [\xi_1^2(t) - \xi_2^2(t)]\,dt$ is a zero-mean random variable if the variance of $\xi_i(t)$ is constant.

When $\eta = 0$ Equation 2.1 can be expressed in terms of the energy ratio $\Delta_E = E_1/E_2$, which is eqivalent to saying $\Delta_E = d_2^2/d_1^2$. Since the numerator and denominator are independent of each other, by assuming that E_1 and E_2 have a normal distribution, d_1^2 and d_2^2

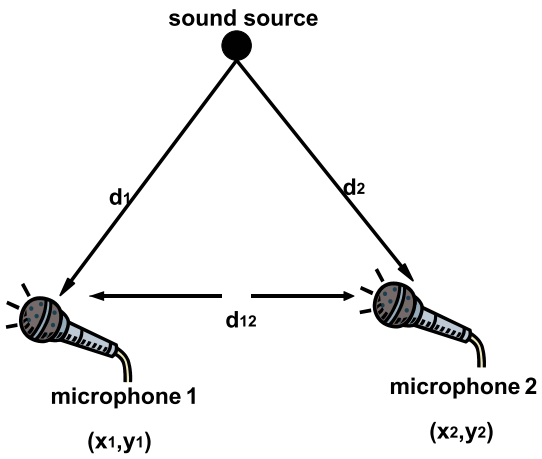

Figure 2.1 Inverse Square Law

can be considered to be Chi-Square distributions. Therefore Δ_E would be a comparision of two Chi-square distributions which can be done using the F-distribution. The F-distribution is a ratio of two Chi-square distributions. It is a non-negative, non-symmetric distribution with two degrees of freedom, one for the numerator and the other for the denominator.

In this thesis though, homegenous coordinates are used due its ease of formulation and implementation. The coordinates of the ith microphone are represented as (x_i, y_i), and the coordinates of the sound source as (x, y). To simplify the analysis, it is assumed to be a planar world throughout. Then $d_i^2 = (x - x_i)^2 + (y - y_i)^2$. Substituting this expression into Equation 2.1 yields, after algebraic manipulation, the following quadratic equation in x and y:

$$\begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} c_e & 0 & -c_x \\ 0 & c_e & -c_y \\ -c_x & -c_y & c \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \eta, \tag{2.2}$$

where

$$\begin{aligned} c_e &= E_1 - E_2 \\ c_x &= E_1 x_1 - E_2 x_2 \\ c_y &= E_1 y_1 - E_2 y_2 \\ c &= E_1(x_1^2 + y_1^2) - E_2(x_2^2 + y_2^2). \end{aligned}$$

When $\eta = 0$, this equation describes the locus of points where a source emitting a sound

will cause the two microphones to receive signals with energies of E_1 and E_2, respectively. This equation holds regardless of the overall energy of the original signal, as seen when the entire equation is divided by E_2 to obtain an equivalent expression only in terms of the energy ratio $\Delta_E = E_1/E_2$.

Homogeneous coordinates are used in Equation 2.2 to show all possible cases using a single expression. One such case occurs when the received energies are not identical, i.e., $E_1 \neq E_2$; then the equation can be written in a more familiar form

$$\left(x - \frac{c_x}{c_e}\right)^2 + \left(y - \frac{c_y}{c_e}\right)^2 = \frac{E_1 E_2 d_{12}^2}{c_e^2} + \eta',$$

where $d_{12} = (x_1 - x_2)^2 + (y_1 - y_2)^2$ is the squared distance between the two microphones, and $\eta' = \eta/c_e$. According to this expression, the sound source is constrained to lie on a circle centered at $(c_x/c_e, c_y/c_e)$ with a radius of $d_{12}\sqrt{E_1 E_2}/c_e$, ignoring noise. In 3D, the circle becomes a sphere.

Another case arises when $E_1 = E_2$; then the equation reduces to

$$2c_x x + 2c_y y = c + \eta,$$

which is the equation of the line passing halfway between the microphones and perpendicular to the line joining them (i.e., the perpendicular bisector). In 3D, the line becomes a plane.

This line for $E_1 = E_2$ and the circles for $E_1 \neq E_2$ are evident in the isocontours of the quadratic equation displayed in Figure 2.2. The shape of these isocontours correspond qualitatively with those measured in the ILD localization system of owls [17].

In this figure, the sound source lies on a circle (sphere) unless the two energies are equal, in which case it lies on a line (plane, the *mid-sagittal plane*) between the microphones. Here microphones 1 and 2 are located at $(-0.5, 0)$ and $(0.5, 0)$, respectively.

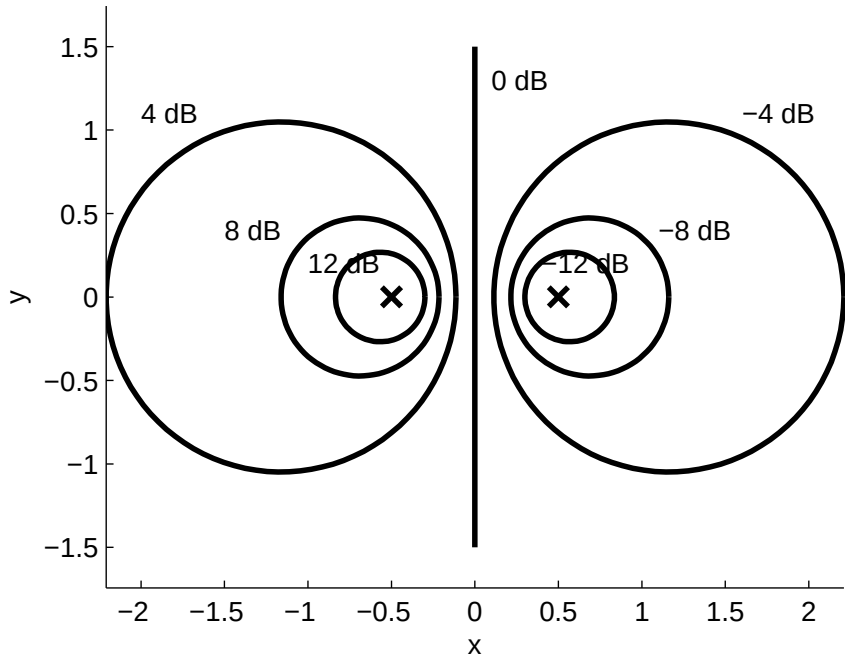

Figure 2.2 Isocontours of Equation 2.2 for different values of $10 \log \Delta_E$.

CHAPTER 3
ILD LOCALIZATION

Sounds that lie in the mid-sagittal plane form a *cone of confusion* where all sounds located on the cone produce the same interaural differences. ILD cues vary with both "cone of confusion" [17] and the relative distance from the source to the left and right ears. At low frequencies, only the relative distance from the source to the ears contributes to the overall ILD. For all frequencies, the spatial information in the ILD that is independent of ITD depends only on the distances from source to two ears and is constant on a sphere symmetrical about the interaural axis. Since, with only two microphones ILD is not able to pinpoint the sound source location, the source is constrained to lie on a curve (or surface in 3D), similar to the "cone of confusion". Ignoring noise, all sources emanating from a point on this curve yield an identical interaural level difference. For each sound source location, there exists a cone of confusion describing the location of other sound sources that produce the same interaural differences. Therefore, sound localization mistakes often occur along this cone of confusion. For example, in the mid-sagittal plane, listeners often confuse sounds from directly in front with those from directly behind (*front-back confusions*) and vice versa (*back-front confusions*) thereby causing localization errors in the vertical plane. One way to overcome this confusion is to use spectral cues derived from the Head-Related Transfer Function (HRTFs) in addition to the interaural differences. The HRTF describes how the torso and and head change the amplitudes and phases of a sound as it travels from a source toward the outer ear. At high frequencies the HRTFs are also affected by pinnae.

Another criteria to consider is the smallest angular separation between two sound sources that a receiver could just detect which is called the "minimal audible angle" (MAA) [16]. Experiments show that when a sound is in front of a listener, a change in location can be better detected than when it is to one side.

Combined Likelihood Approach

The approach proposed here to solve these ambiguities is to employ multiple microphone pairs, each determining a different curve in the environment so that the intersection

of these curves yields the sound source location. However, instead of computing this intersection directly using a closed form or a least squares solution, probabilistic sampling is used. That is, a number of candidate locations in the space is selected and, for each of these locations, the likelihood that the sound source is located there is computed. This total likelihood is computed as the sum of the likelihoods using each microphone pair. Assuming that the microphone pairs yield independent measurements, this technique is equivalent to computing the joint probability by multiplying the individual probabilities using the sum of log likelihoods. This simple approach to sensor fusion has been used successfully in ITD acoustic localization [8].

The final issue here that remains to be solved is to compute the likelihood at an arbitrary candidate location given a curve (circle or line) for a microphone pair. This problem is solved by calculating the expected value for Δ_E for any given candidate location $(\tilde{x}, \tilde{y})$, as the ratio of the squares of the distances to the two microphones:

$$\tilde{\Delta}_E = \frac{(\tilde{x} - x_2)^2 + (\tilde{y} - y_2)^2}{(\tilde{x} - x_1)^2 + (\tilde{y} - y_1)^2}.$$

This result is obtained by substituting $(\tilde{x}, \tilde{y})$ for (x, y) in Equation 2.2, setting $\eta = 0$, and solving for Δ_E. Using this expression, $\tilde{\Delta}_E$ for all the candidate locations is computed once off-line. Then, at run time, the likelihood that the sound source is at a candidate location is computed by treating $10 \log \Delta_E$ as a Gaussian random variable with a mean of $10 \log \tilde{\Delta}_E$ and a variance of σ_e^2. This approach is able to localize the sound source accurately in most cases but fails sometimes in highly reverberant environments.

Hilbert Envelope Approach

To improve these results, we tried a Hilbert envelope approach. If a signal were a perfect impulse, reverberation would have no serious effect, the echoes contributing an aggregate of impulses, none exceeding the desired signal. Therefore, it is better to exploit the impulse nature of voiced speech [15]. In the range above 1KHz speech energy rises sharply at the start of a pitch period and decays by approximately 20dB before the next pitch pulse. These pitch pulses are distinctly evident in direct sound. Reverberation, however, the composite of many small signals arriving at different times, has a much lower peak factor. Energy

peaks which are usually dominated by direct sound even at five or more times the critical distance are substantially larger than average energy.

Fischell and Cocker proposed in [15] the use of a non-linear processing method that responds primarily to energy peaks. The Hilbert Envelope [15] is generated by using the Hilbert Transformer as shown in Figure 3.1; an all-pass filter circuit produces two signals with equal amplitude but 90 degrees out of phase. These are then squared, the squares summed and the square root of the result is determined.

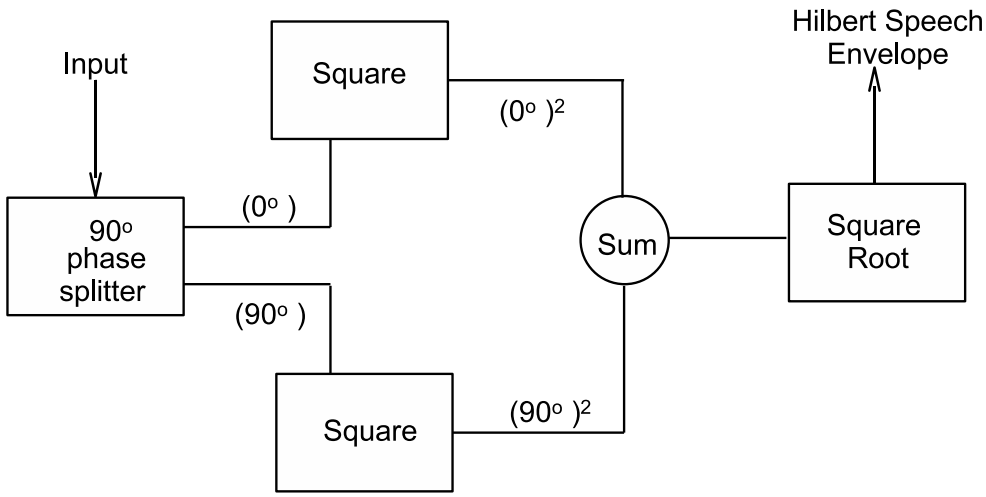

Figure 3.1 Block Diagram for Hilbert Envelope Generation

Hilbert Transform

The derivation of Hilbert tranform [23] relations is based on the notion of causality or one-sidedness. Since the relations are between real and imaginary parts of a complex sequence, one-sidedness is applied to the Fourier transform of the sequence. Since the Fourier transform is periodic it cannot be specified as zero for $\omega < 0$, where ω is the angular frequency. Therefore, sequences for which the Fourier transform is zero in the second half of each period i.e. the z-transform is zero on the bottom half of the unit circle $(-\pi \leq \omega < 0)$, are considered. If $x[n]$ is the sequence, and $X(e^{j\omega})$ is the Fourier transform, then

$$X(e^{j\omega}) = 0, \qquad -\pi \leq \omega < 0$$

The sequence $x[n]$ corresponding to $X(e^{j\omega})$ must be complex, since if $x[n]$ were real, $X(e^{j\omega})$ would be conjugate symmetric, i.e., $X(e^{j\omega}) = X^*(e^{j\omega})$. Therefore, $x[n]$ is of the form

$$x[n] = x_r[n] + jx_i[n]$$

Here $x_r[n]$ and $x_i[n]$ are real sequences. If $X_r(e^{j\omega})$ and $X_i(e^{j\omega})$ denotes the Fourier transforms of the real sequences $x_r[n]$ and $x_i[n]$ respectively, then

$$X(e^{j\omega}) = X_r(e^{j\omega}) + jX_i(e^{j\omega})$$

meaning

$$X_r(e^{j\omega}) = \frac{1}{2}[X(e^{j\omega}) + X^*(e^{-j\omega})]$$
$$jX_i(e^{j\omega}) = \frac{1}{2}[X(e^{j\omega}) - X^*(e^{-j\omega})]$$

$X_r(e^{j\omega})$ and $X_i(e^{j\omega})$ are complex valued functions in general. $X_r(e^{j\omega})$ is conjugate symmetric, i.e., $X_r(e^{j\omega}) = X_r^*(e^{j\omega})$, and $jX_i(e^{j\omega})$ is conjugate antisymmetric, i.e., $jX_i(e^{j\omega}) = -jX_i^*(e^{j\omega})$.

From these equations,

$$X(e^{j\omega}) = \begin{cases} 2X_r(e^{j\omega}) & , \quad 0 \leq \omega < \pi \\ 0 & , \quad -\pi \leq \omega < 0 \end{cases}$$

and

$$X(e^{j\omega}) = \begin{cases} 2jX_i(e^{j\omega}) & , \quad 0 \leq \omega < \pi \\ 0 & , \quad -\pi \leq \omega < 0 \end{cases}$$

$X_r(e^{j\omega})$ and $X_i(e^{j\omega})$ can be related directly by

$$X_i(e^{j\omega}) = \begin{cases} -jX_r(e^{j\omega}) & , \quad 0 \leq \omega < \pi \\ jX_r(e^{j\omega}) & , \quad -\pi \leq \omega < 0 \end{cases}$$

or

$$X_i(e^{j\omega}) = H(e^{j\omega})X_r(e^{j\omega}) \tag{3.1}$$

where

$$H(e^{j\omega}) = \begin{cases} -j & , \quad 0 < \omega < \pi \\ j & , \quad -\pi < \omega < 0 \end{cases}$$

This frequency response has unity magnitude, a phase angle of $-\pi/2$ for $0 < \omega < \pi$, and a phase angle of $+\pi/2$ for $-\pi < \omega < 0$. Such a system is called an ideal 90-degree phase shifter or a Hilbert transformer.

From Equation 3.1,

$$X_r(e^{j\omega}) = \frac{1}{H(e^{j\omega})}X_i(e^{j\omega}) = -H(e^{j\omega})X_i(e^{j\omega})$$

Thus, $-x_r[n]$ can also be obtained from $x_i[n]$ using a 90-degree phase shifter.

The impulse response $h[n]$ of a 90-degree phase shifter, corresponding to the frequency response $H(e^{j\omega})$ can be represented by

$$h[n] = \frac{1}{2}\int_{-\pi}^{0} je^{j\omega n}d\omega - \frac{1}{2\pi}\int_{0}^{\pi} je^{j\omega n}d\omega,$$

or

$$h[n] = \begin{cases} \frac{2}{\pi}\frac{sin^2(\pi n/2)}{n} & , \quad n \neq 0 \\ 0 & , \quad n = 0 \end{cases}$$

Therefore,

$$x_i[n] = \sum_{m=-\infty}^{\infty} h[n-m]x_r[m]$$

$$x_r[n] = -\sum_{m=-\infty}^{\infty} h[n-m]x_i[m]$$

Figure 3.2 shows how a discerte-time Hilbert transformer system can be used to form a complex analytic signal, which is simply a pair of real signals.

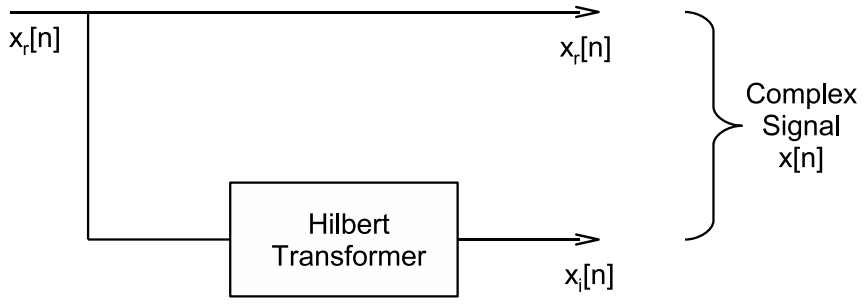

Figure 3.2 Block Diagram representation of the creation of a complex sequence having a one sided Fourier transform

The Design of a Hilbert Transformer

A Hilbert Transformer can be designed using a Kaiser window defined as

$$w[n] = \begin{cases} \frac{I_o\beta(1-[(n-\alpha)/\alpha]^2)^1/2}{I_o(\beta)} & , \quad 0 \leq n \leq M \\ 0 & , \quad otherwise \end{cases}$$

where $\alpha = M/2$ and $I_o(.)$ represent the zeroth-order modified Bessel function of the first kind. In contrast to other windows, the Kaiser window has two parameters: the length (M+1) and a shape parameter β. By varying (M+1) and β, the window length and shape can be adjusted to trade side-lobe amplitude for main lobe width.

The Hilbert transformer can be approximated by the Kaiser window approximation of order M (length M+1) in the form of

$$h[n] = \begin{cases} \frac{I_o\beta(1-[(n-n_d)/n_d]^2)^1/2}{I_o(\beta)}[\frac{2}{\pi}\frac{sin^2[\pi(n-n_d)/2}{n-n_d}] & , \quad 0 \leq n \leq M \\ 0 & , \quad otherwise \end{cases}$$

where $n_d = M/2$. Figure 3.3 shows the impulse response and Figure 3.4 the magnitude of frequency response for M=18 and β=2.629. Because $h[n]$ satisfies the symmetry condition $h[n] = -h[M-n]$ for $0 \leq n \leq M$, this phase is exactly 90 degrees plus a linear component corresponding to a delay of $n_d = 18/2 = 9$ samples; i.e.

$$\angle H(e^{j\omega}) = -\frac{\pi}{2} - 9\omega, \qquad 0 < \omega < \pi$$

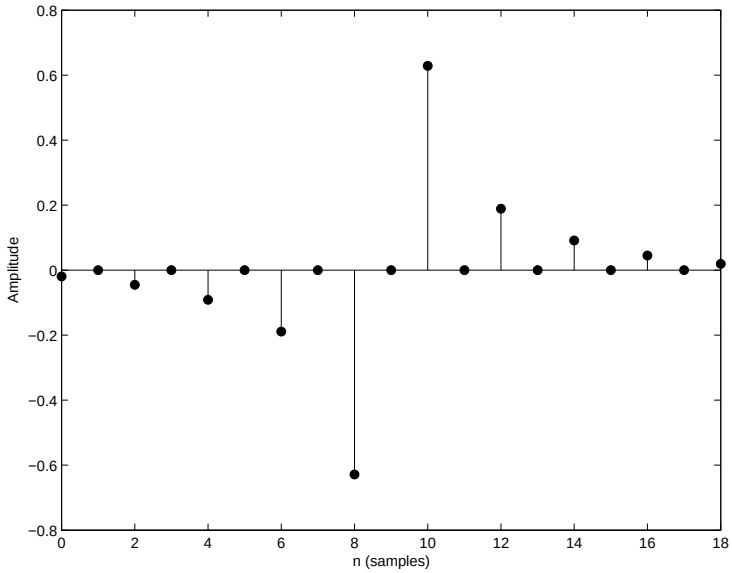

Figure 3.3 Impulse response of an FIR Hilbert Transformer designed using the Kaiser window (M=18 and β=2.629)

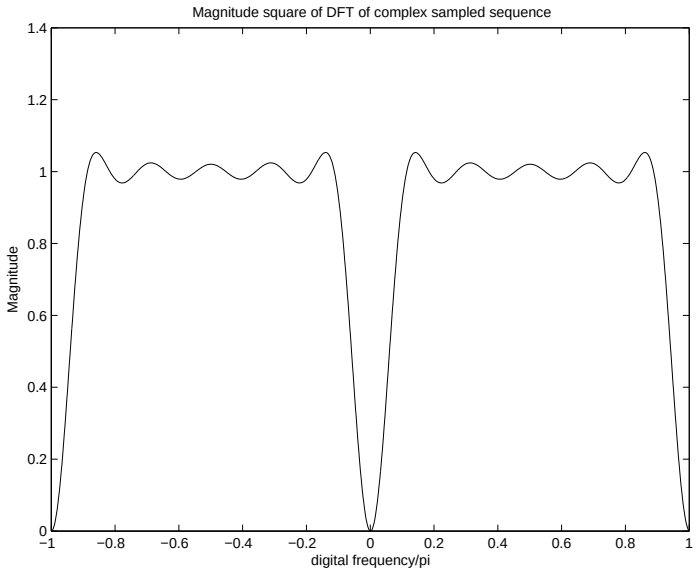

Figure 3.4 Magnitude response of an FIR Hilbert Transformer designed using the Kaiser window (M=18 and β=2.629)

CHAPTER 4

SIMULATION RESULTS

The ILD algorithm was tested in a 5 m $\times$ 5 m simulated room with four microphones arranged in a square so that opposing ones were separated by 1 m, as shown in Figure 4.1.

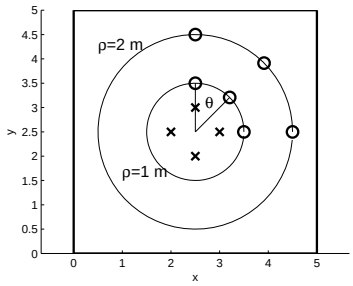

Figure 4.1 The simulated room with four microphones (x) and six sound source locations (o).

A sound file of a male voice counting from one to ten (16-bit, 44.1 kHz) was played at a predetermined location in the room and captured by the microphones, using the image method [18] with linear interpolation between samples up to sixth order reflections for the four walls. For these experiments, the entire 2.5-second utterance was treated as a single audio frame.

Figure 4.2 shows the results of Δ_E estimation using the two horizontal microphones for the six sound source locations. From top to bottom, $\theta = 0, 45, 90$ degrees; from left to right $\rho = 1, 2$ m. The solid line is ground truth.

The error $|10 \log \Delta_E - 10 \log \tilde{\Delta}_E| = 10|\log(\Delta_E/\tilde{\Delta}_E)|$ is calculated using the two horizontal microphones at different source locations and with different values of the reflection coefficient β as seen in in Figure 4.2. The accuracy of the Δ_E estimation is highly dependent on the sound source location and the amount of reverberation. At higher reverberations, or in positions where the reverberations are asymmetric, the accuracy of the estimation decreases significantly. Not only does the error increase but there also appears to be a systematic bias in the estimation. This behavior is because ILD conveys additional information about both the source distance and direction only when sources are within a

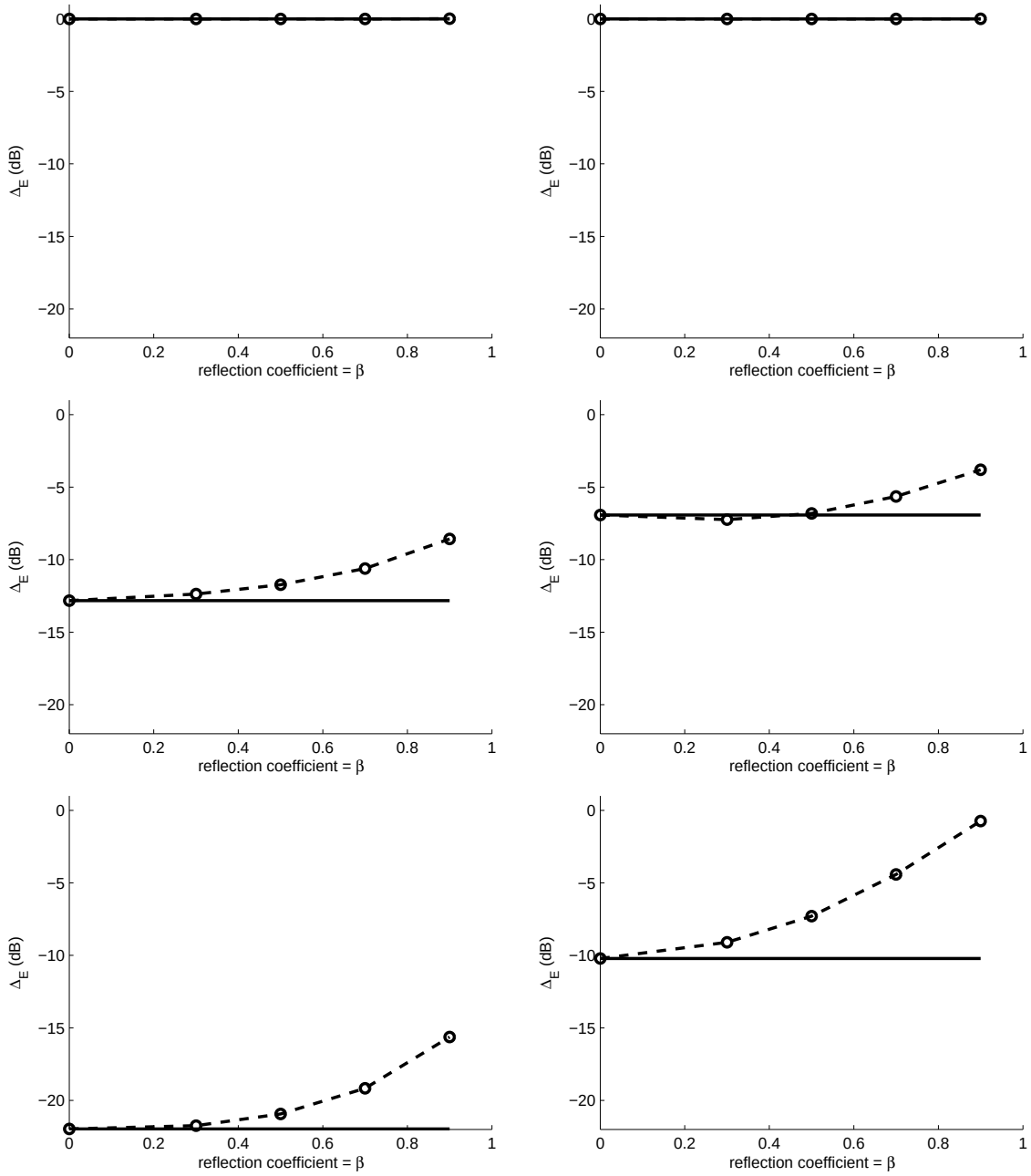

Figure 4.2 The results of Δ_E estimation using the two horizontal microphones for the six sound source locations. From top to bottom, $\theta = 0, 45, 90$ degrees; from left to right $\rho = 1, 2$ m. The solid line is ground truth.

meter of the listener [22]. In addition, reverberation also has a large impact on the perceived characteristics of a listening environment, in part because reverberation decorrelates the signals reaching the two ears. The perceived spaciousness of a room increases with reverberation time, reverberation level, and/or the amount of decorrelation between left and right ear signals. Reverberation physically distorts steady-state “directional” cues like interaural differences and spectral shape, and according to Shinn and Cunningham [2000], there is evidence that more “realistic” reverberation does interfere with directional perception.

The likelihood plots for a symmetric room (5m x 5m) with a sound source at 45 degrees and a distance of 2m, no noise and a reflection coefficient of 0 are shown in the Figure 4.3. In this figure, the likelihood function computed by the horizontal (left) and vertical (center) microphone pairs, and the contour plot for the overlaid likelihood (right) are shown at the top. The contour plots on the bottom show the two likelihood functions for the horizontal (left), and vertical (center) microphone pairs, and the contour plot for the combined likelihood (right), in addition to the microphones (x), the true sound source location (o), the peak of the combined function (*), and the computed bearing angle to the peak (solid line). The likelihood plots with a sound source at 72 degrees and a distance of 1m, for the same specifications are shown in the Figure 4.4.

If the sound is complex, such as noise, then different frequencies will be attenuated and delayed by different amounts depending on the size of the objects (such as the pinna and various parts of the pinna, the nose, and the torso) the sound encounters before reaching the ear. The amount of attenuation and delay provided by an obstacle will also depend on the direction from which the sound originates. For instance, the pinna offers more attenuation for sounds coming from behind than those coming from the front.

The ILD algorithm is able to localize the sound source accurately even with high noise and reverberation. This is indicated by Figures 4.5 and 4.6, which show the likelihood plots for a source angle of 45 degrees at 2m, with an SNR of 0dB and a reflection coefficient of 0.9 for a 5m x 5m and a 10m x 10m room respectively while Figures 4.7, 4.8 and 4.9 show the likelihood plots for the same specifications with a source angle of 0 degrees at 2m, and 90 and 0 degrees at 1m respectively.

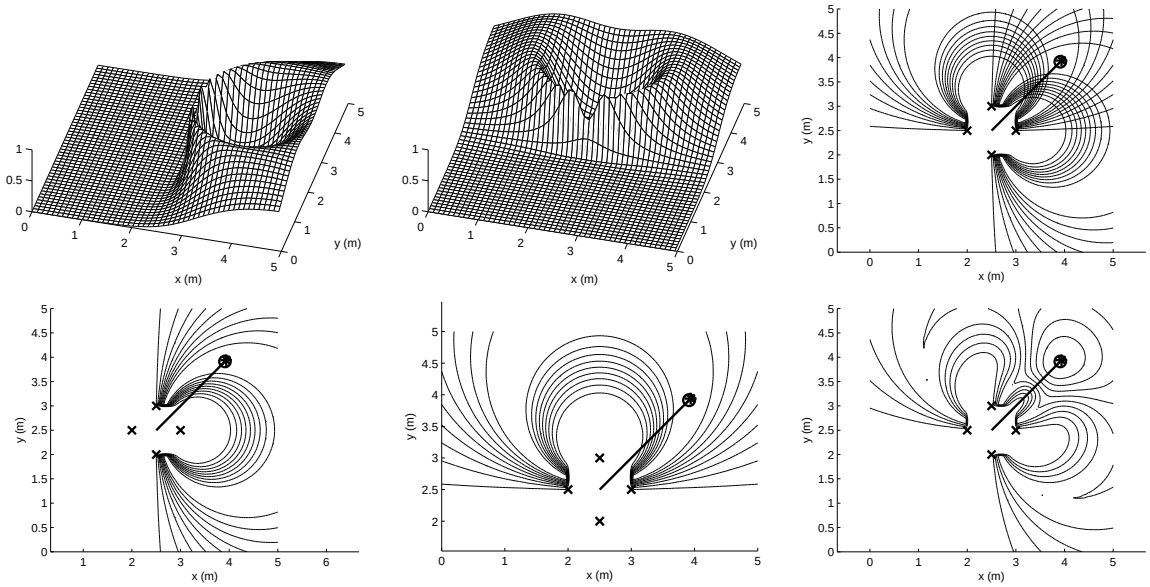

Figure 4.3 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 45 degrees and a distance of 2m, no noise, no reflection.

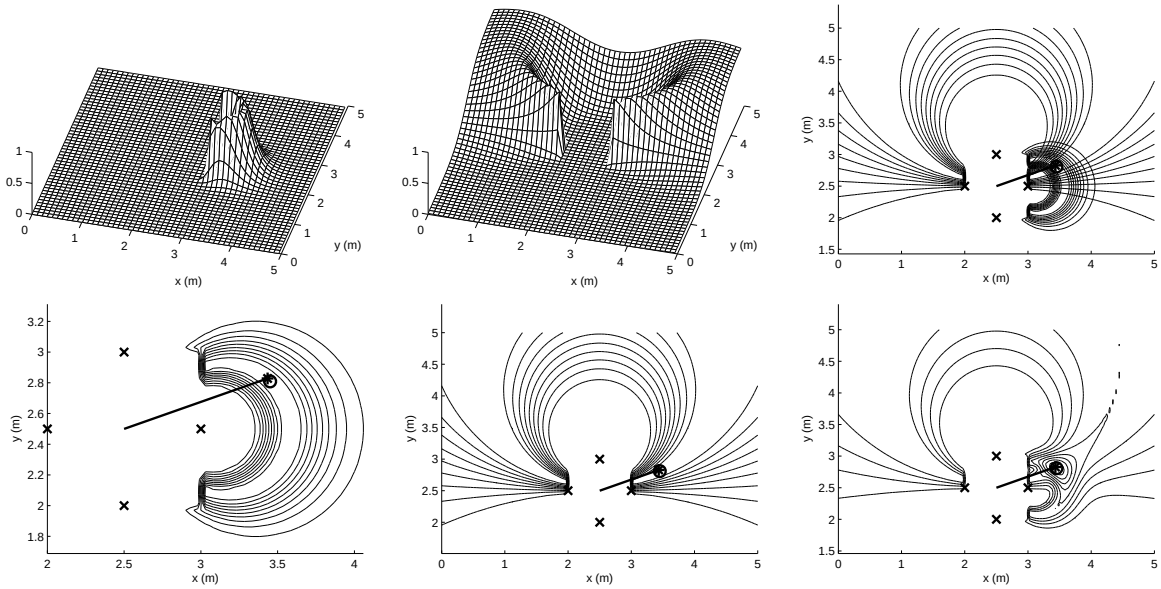

Figure 4.4 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 72 degrees and distance of 1m, no noise, no reflection.

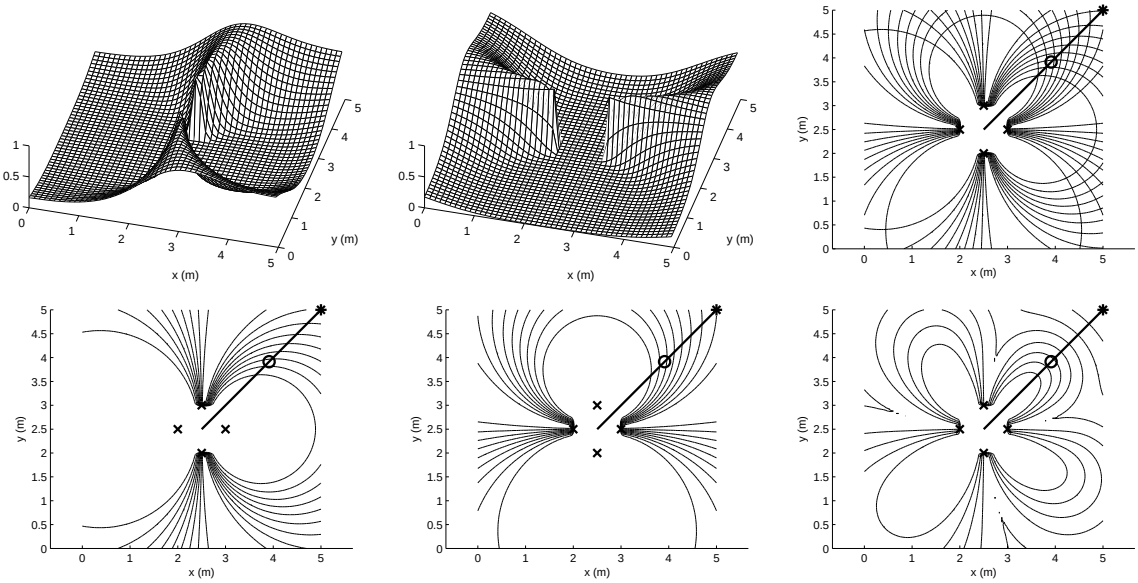

Figure 4.5 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 45 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

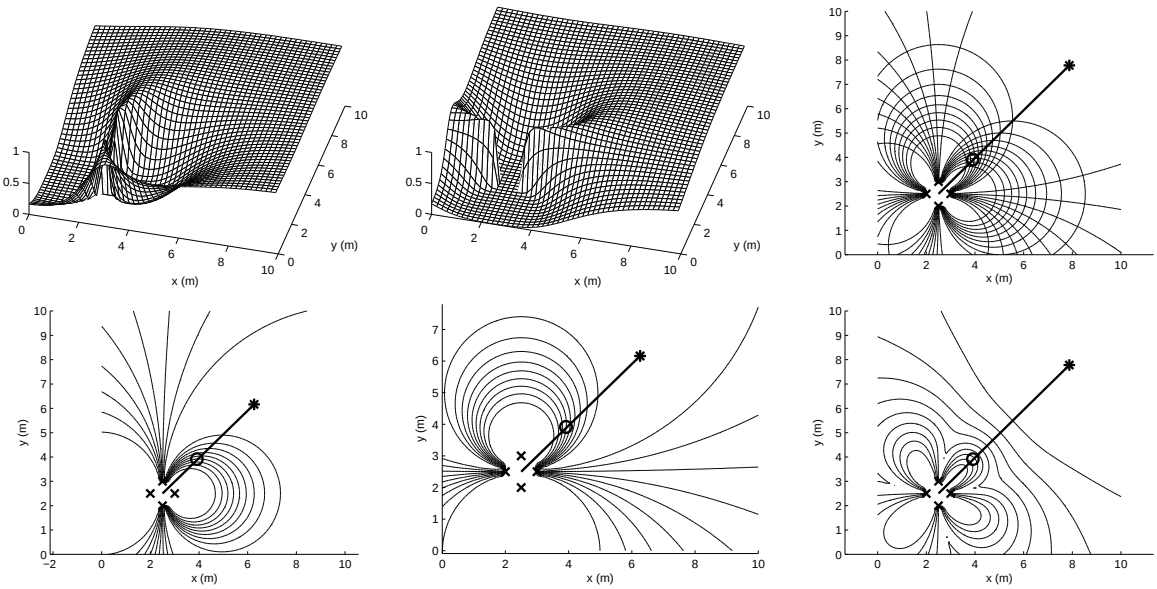

Figure 4.6 Likelihood plots for a symmetric room (10m x 10m) with a sound source at 45 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

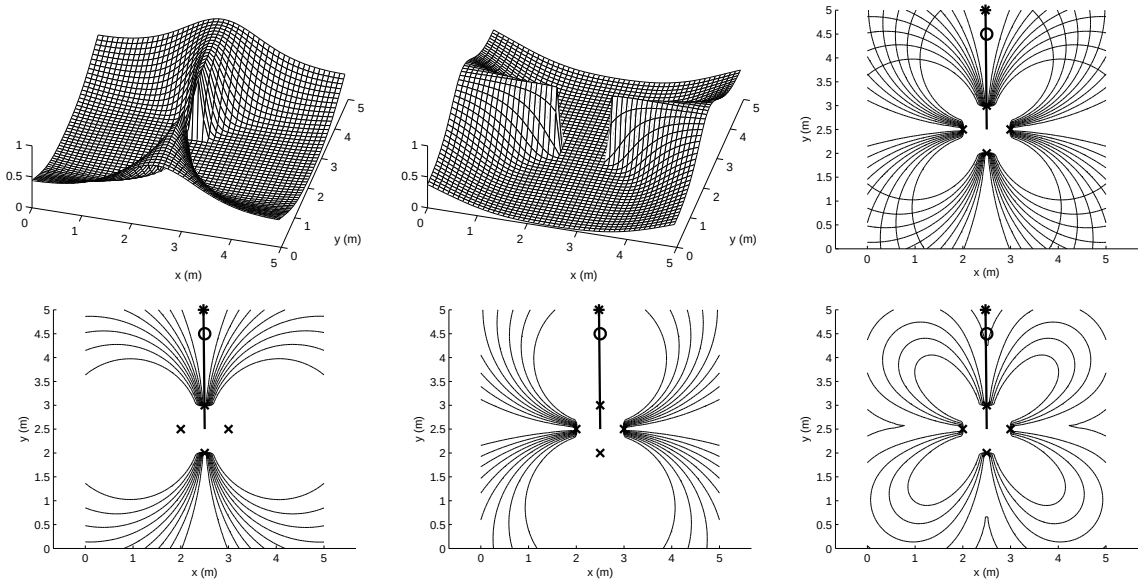

Figure 4.7 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 0 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

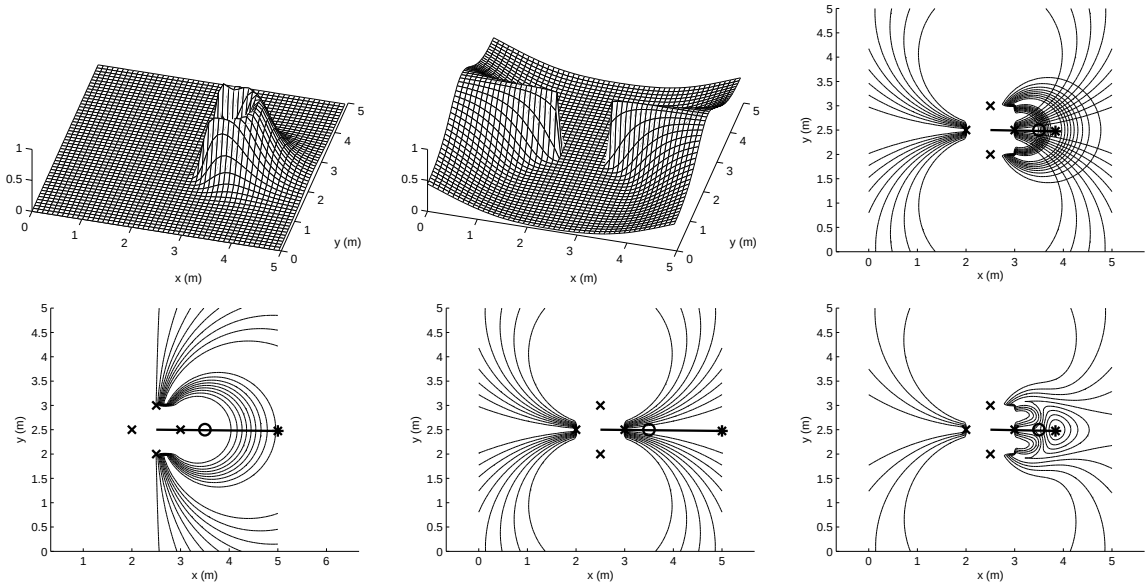

Figure 4.8 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 90 degrees and a distance of 1m, an SNR of 0dB and a reflection coefficient of 0.9.

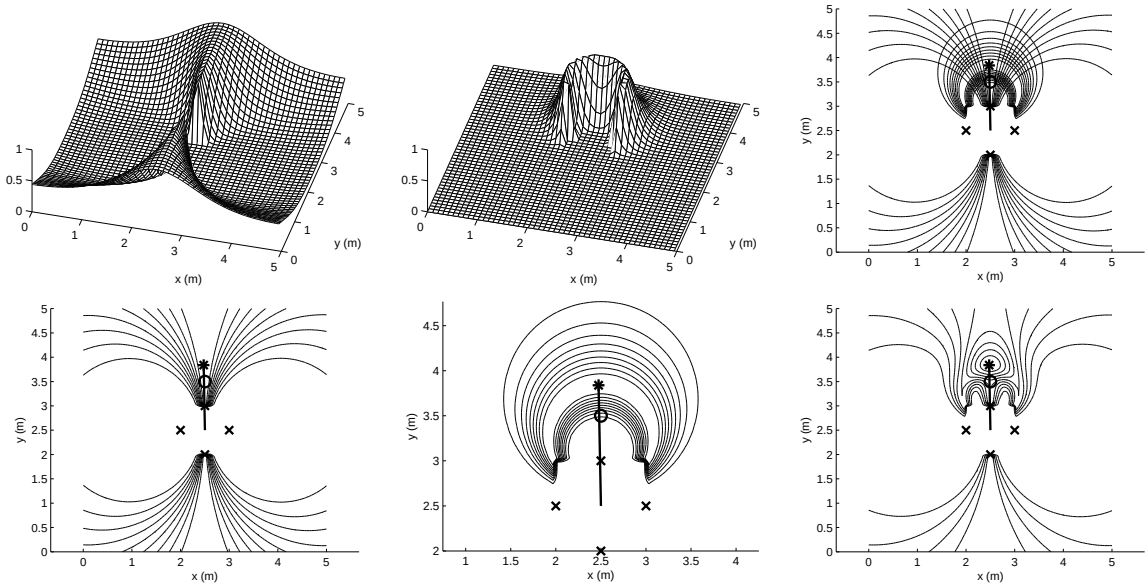

Figure 4.9 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 0 degrees and a distance of 1m an SNR of 0dB and a reflection coefficient of 0.9.

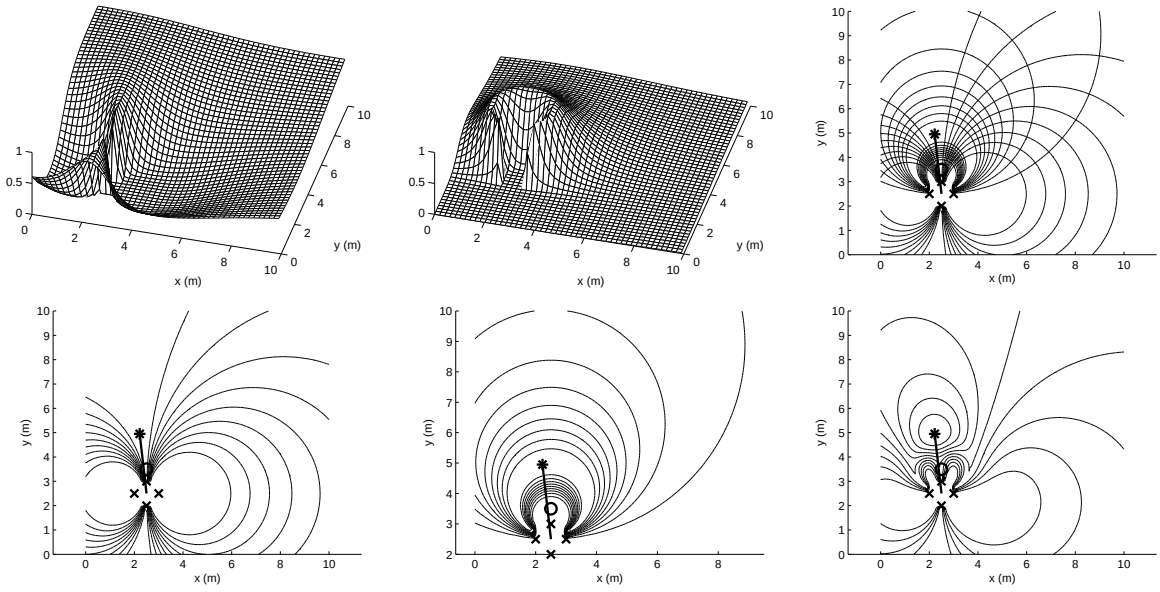

Figure 4.10 Likelihood plots for a symmetric room (10m x 10m) with a sound source at 0 degrees and a distance of 1m, an SNR of 0dB and a reflection coefficient of 0.9.

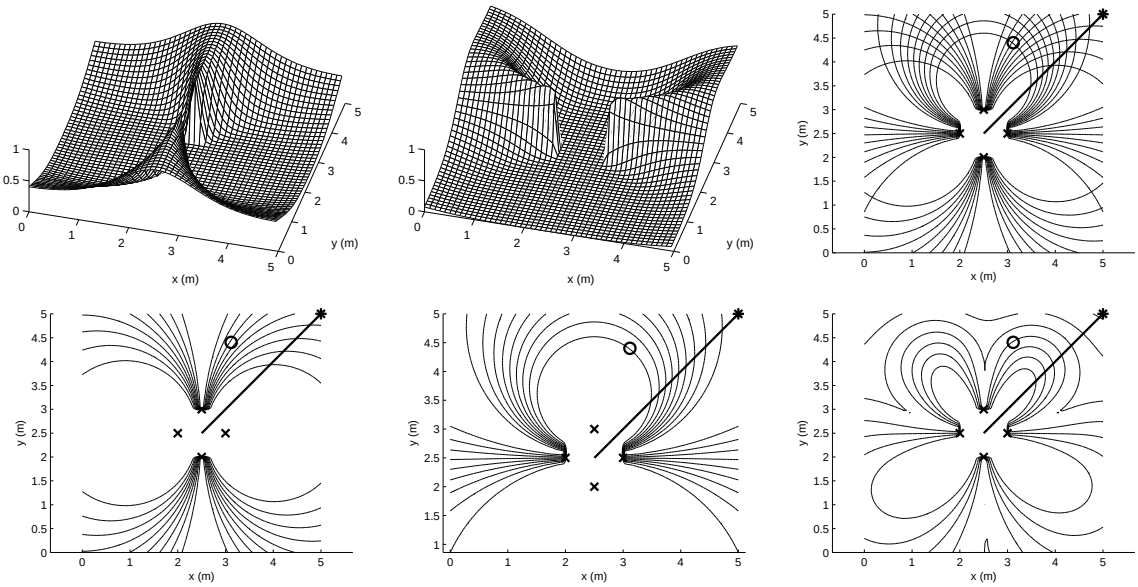

Figure 4.11 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 18 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

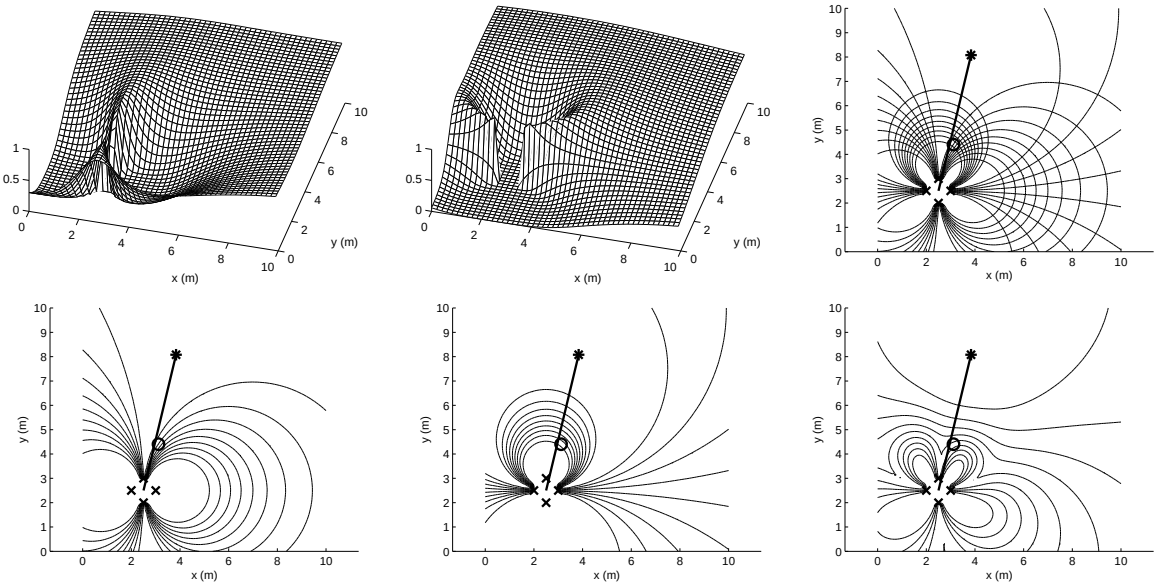

Figure 4.12 Likelihood plots for a symmetric room (10m x 10m) with a sound source at 18 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

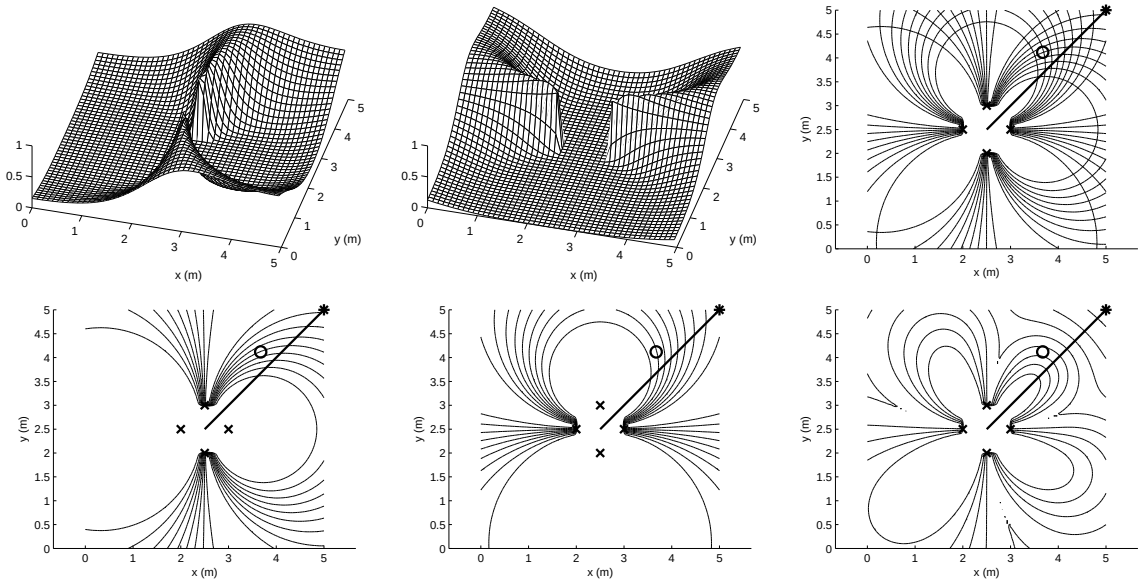

Figure 4.13 Likelihood plots for a symmetric room (5m x 5m) with a sound source at 36 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

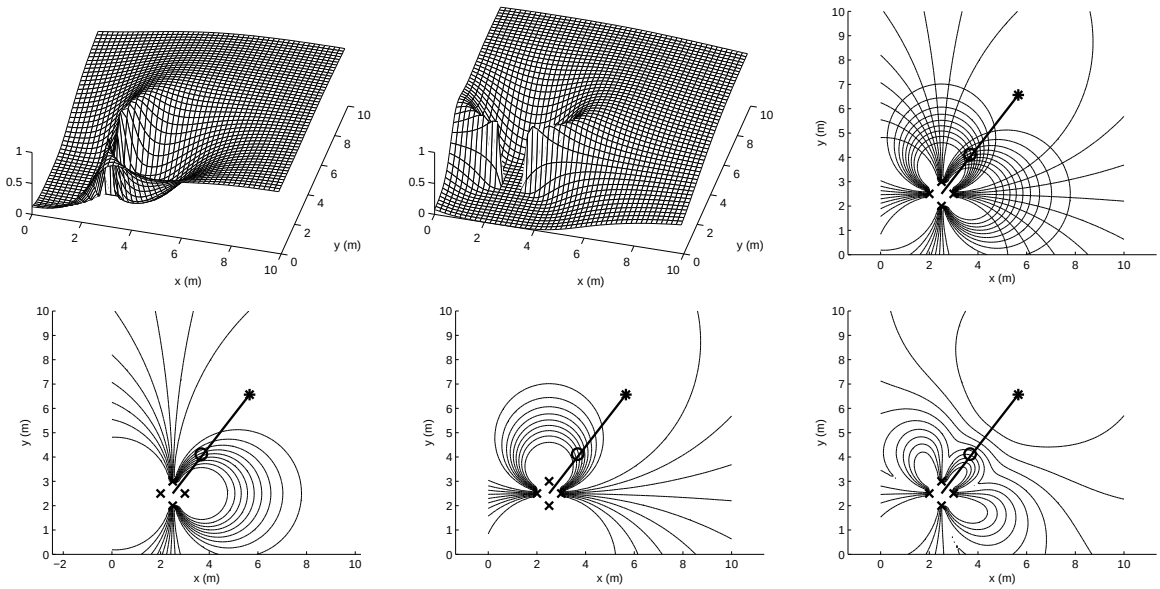

Figure 4.14 Likelihood plots for a symmetric room (10m x 10m) with a sound source at 36 degrees and a distance of 2m, an SNR of 0dB and a reflection coefficient of 0.9.

However, there are cases like the ones shown in Figures 4.10, 4.11, 4.12, 4.13 and 4.14 in which the localization is not very accurate, i.e. there exists an angle error between the actual source location and the predicted source location. These cases have angle errors of 27.0 and 9.0 degrees, when the source is at 18 and 36 degrees respectively for the 5m x 5m room, and 6.5, 4.5 and 1.8 degrees, when the source is at 0, 18 and 36 degrees respectively for the 10m x 10m room. In the 0 degree case the source distance is 1m while it is 2m for the 18 and 36 degree cases.

Table 4.1 Angle error in degrees for the 5m x 5m room when the source is at a distance of 1m.

	5m x 5m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	0.5	1.7	0.9	0.0
β=0.7, SNR=10dB	1.1	0.9	0.2	0.0
β=0.7, SNR=20dB	1.2	1.1	1.3	0.0
β=0.8, SNR= 0dB	0.5	1.7	2.4	0.7
β=0.8, SNR=10dB	1.0	4.2	2.0	0.0
β=0.8, SNR=20dB	1.2	2.5	2.0	0.0
β=0.9, SNR= 0dB	0.5	3.8	1.9	0.0
β=0.9, SNR=10dB	0.9	5.1	4.4	0.0
β=0.9, SNR=20dB	1.1	5.8	4.7	0.0

Table 4.2 Angle error in degrees for the 10m x 10m room when the source is at a distance of 1m.

	10m x 10m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	4.1	2.3	1.9	1.7
β=0.7, SNR=10dB	3.8	1.9	2.1	0.0
β=0.7, SNR=20dB	4.2	1.4	1.4	0.0
β=0.8, SNR= 0dB	4.1	2.3	0.5	0.0
β=0.8, SNR=10dB	3.8	2.7	2.1	0.0
β=0.8, SNR=20dB	4.2	0.4	1.4	0.0
β=0.9, SNR= 0dB	6.5	1.5	0.5	0.0
β=0.9, SNR=10dB	3.5	2.7	2.0	0.0
β=0.9, SNR=20dB	4.2	0.4	1.4	0.0

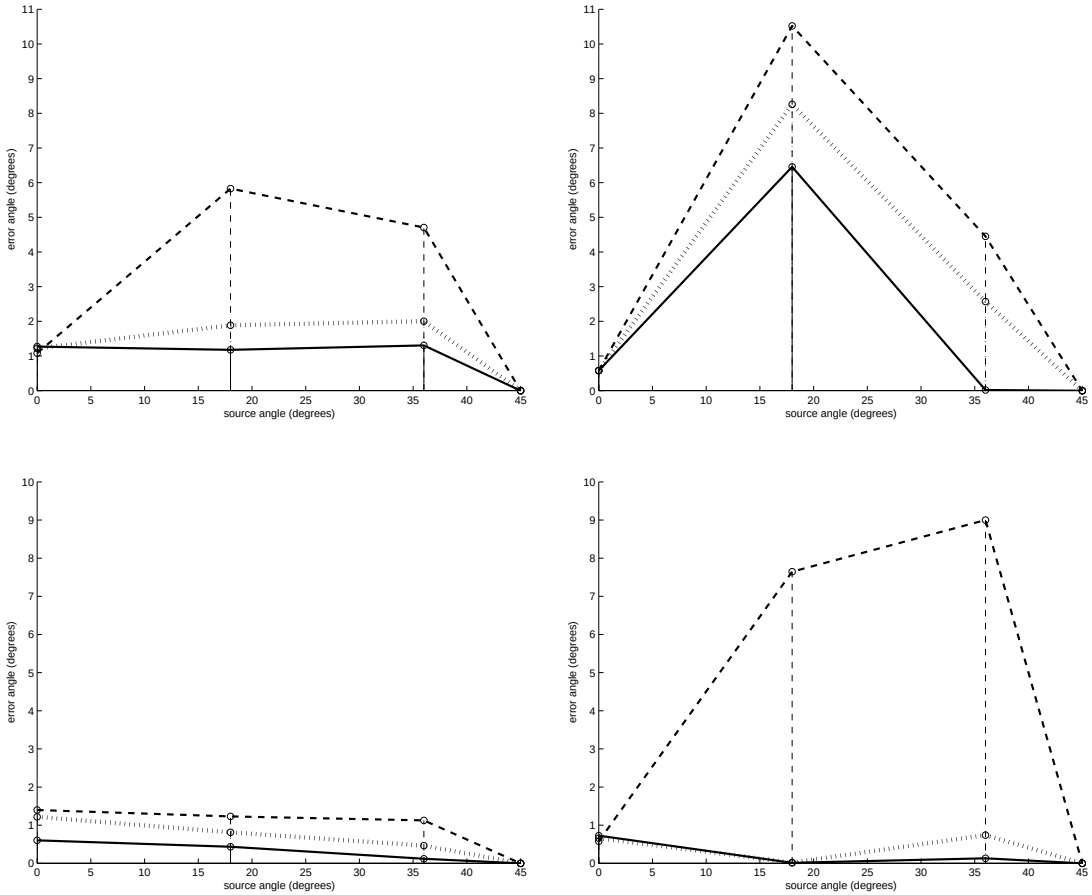

Figure 4.15 Angle errors when either noise or reverberation are present in a 5m x 5m room. The angle error plots for reverberation of 0.7 (solid line), 0.8 (dotted), 0.9 (dashed) for 1m (top left) and 2m (top right) rooms and the plots for SNR of 20 (solid line), 10 (dotted), and 0dB (dashed) for 1m (bottom left)and 2m (bottom right) rooms are shown.

Figure 4.15 shows the angle error plots for cases where only reverberation or noise are present in a 5m x 5m room. The figure indicates that the angle errors are caused due to both reverberation and noise and increase as the reverberation and the amount of noise increases.

Table 4.3 Angle error in degrees for the 5m x 5m room when the source is at a distance of 2m.

	5m x 5m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	45.0	26.4	9.0	0.0
β=0.7, SNR=10dB	43.8	6.0	0.0	0.0
β=0.7, SNR=20dB	0.5	6.7	0.9	0.0
β=0.8, SNR= 0dB	45.0	27.0	9.0	0.0
β=0.8, SNR=10dB	45.0	7.1	2.9	0.0
β=0.8, SNR=20dB	45.0	8.2	1.8	0.0
β=0.9, SNR= 0dB	0.5	27.0	9.0	0.0
β=0.9, SNR=10dB	0.5	10.5	5.9	0.0
β=0.9, SNR=20dB	0.5	10.5	4.4	0.0

Table 4.4 Angle error in degrees using the Hilbert Envelope (HILBERT function in Matlab) for the 5m x 5m room when the source is at a distance of 2m.

	5m x 5m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	45.0	3.5	3.9	0.0
β=0.7, SNR=10dB	0.5	6.0	0.5	0.0
β=0.7, SNR=20dB	0.5	6.6	0.9	0.0
β=0.8, SNR= 0dB	45.0	11.0	8.4	0.0
β=0.8, SNR=10dB	45.0	7.1	2.2	0.0
β=0.8, SNR=20dB	45.0	8.2	1.8	0.0
β=0.9, SNR= 0dB	0.5	21.9	9.0	0.0
β=0.9, SNR=10dB	0.5	10.5	4.6	0.0
β=0.9, SNR=20dB	0.5	10.5	4.4	0.0

Tables 4.1 and 4.2 show the angle error values when the sound source is at a distance of 1m in a 5m x 5m room and a 10m x 10m room for sound source angles of 0, 18, 36 and 45 degrees, with different permutations of, a reflection coefficient of 0.7, 0.8 and 0.9, and SNR of 0, 10 and 20 dB. Tables 4.3 and 4.6 show the angle error values for the same

Table 4.5 Angle error in degrees using the Hilbert Envelope (Kaiser Hilbert Transformer) for the 5m x 5m room when the source is at a distance of 2m.

	5m x 5m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	45.0	9.2	5.3	0.0
β=0.7, SNR=10dB	0.5	4.9	0.5	0.0
β=0.7, SNR=20dB	0.5	4.3	0.9	0.0
β=0.8, SNR= 0dB	45.0	16.0	9.0	0.0
β=0.8, SNR=10dB	45.0	4.9	2.2	0.0
β=0.8, SNR=20dB	45.0	6.0	1.8	0.0
β=0.9, SNR= 0dB	0.5	24.5	9.0	0.0
β=0.9, SNR=10dB	0.5	7.1	3.9	0.0
β=0.9, SNR=20dB	0.5	7.1	3.1	0.0

Table 4.6 Angle error in degrees for the 10m x 10m room when the source is at a distance of 2m.

	10m x 10m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	19.5	4.6	1.1	0.9
β=0.7, SNR=10dB	19.1	3.5	1.4	0.0
β=0.7, SNR=20dB	19.2	4.7	0.6	0.0
β=0.8, SNR= 0dB	23.2	3.7	1.5	0.0
β=0.8, SNR=10dB	24.1	4.6	0.8	0.0
β=0.8, SNR=20dB	23.9	3.5	0.0	0.0
β=0.9, SNR= 0dB	30.3	4.5	1.8	0.0
β=0.9, SNR=10dB	30.6	2.2	0.1	0.0
β=0.9, SNR=20dB	29.3	3.1	0.8	0.0

specifications when the sound source is at a distance of 2m. From the likelihood contour plots and the tables it becomes obvious that the localization errors are rather small when the sound source is at a distance of 1m when compared to a distance of 2m. As discussed earlier, this is because ILD is able to provide accurate information about the direction and distance only upto a distance of 1m. This implies that sound localization depends upon the relative distance between the source and the microphones. Comparing the tables for the 5m x 5m room and 10m x 10m room it can also be seen that localization depends upon the size of the room because the angle errors for the 5m x 5m room are different from the errors for

Table 4.7 Angle error in degrees using the Hilbert Envelope (HILBERT function in Matlab) for the 10m x 10m room when the source is at a distance of 2m.

	10m x 10m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	17.9	4.0	1.0	0.0
β=0.7, SNR=10dB	18.4	2.9	1.0	0.0
β=0.7, SNR=20dB	18.8	3.5	0.5	0.0
β=0.8, SNR= 0dB	21.2	2.8	1.1	0.0
β=0.8, SNR=10dB	21.0	4.0	0.4	0.0
β=0.8, SNR=20dB	23.3	3.2	1.0	0.0
β=0.9, SNR= 0dB	30.2	3.3	0.3	0.0
β=0.9, SNR=10dB	30.0	1.6	0.0	0.0
β=0.9, SNR=20dB	27.3	1.6	0.6	0.0

Table 4.8 Angle error in degrees using the Hilbert Envelope (Kaiser Hilbert Transformer) for the 10m x 10m room when the source is at a distance of 2m.

	10m x 10m room			
Angle in degrees	0	18	36	45
β=0.7, SNR= 0dB	16.9	2.8	1.1	0.0
β=0.7, SNR=10dB	15.5	2.9	1.1	0.0
β=0.7, SNR=20dB	16.1	2.3	0.6	0.0
β=0.8, SNR= 0dB	20.2	3.6	1.7	0.0
β=0.8, SNR=10dB	21.1	3.0	1.4	0.0
β=0.8, SNR=20dB	19.8	2.9	0.0	0.0
β=0.9, SNR= 0dB	26.1	2.3	0.9	0.0
β=0.9, SNR=10dB	25.9	2.6	0.3	0.0
β=0.9, SNR=20dB	26.7	3.1	0.8	0.0

the 10m x 10m room for the same specifications.

These angle errors have been reduced to a certain extent by using the Hilbert envelope approach. The Tables 4.4, 4.5, 4.7 and 4.8 show the comparison of the angle error values in a 5m x 5m room and a 10m x 10m room for the same sound source locations as above at a distance of 2m. The tables 4.4 and 4.7 show the angle errors using the Hilbert function in Matlab, while the tables 4.5 and 4.8 show the angle errors using the Kaiser Hilbert Transformer. The same are also depicted in the angle error plots in Figures 4.16 and 4.17 in a 5m x 5m room and a 10m x 10m room. The x-axis represents the sound source angle and

the y-axis represents the error angle in degrees. The error without the use of the Hilbert Envelope approach is represented by a solid line, while the one with the use of the Hilbert function in Matlab is represented by a dotted line and that with the use of the Kaiser Hilbert Transformer is represented by a dashed line.

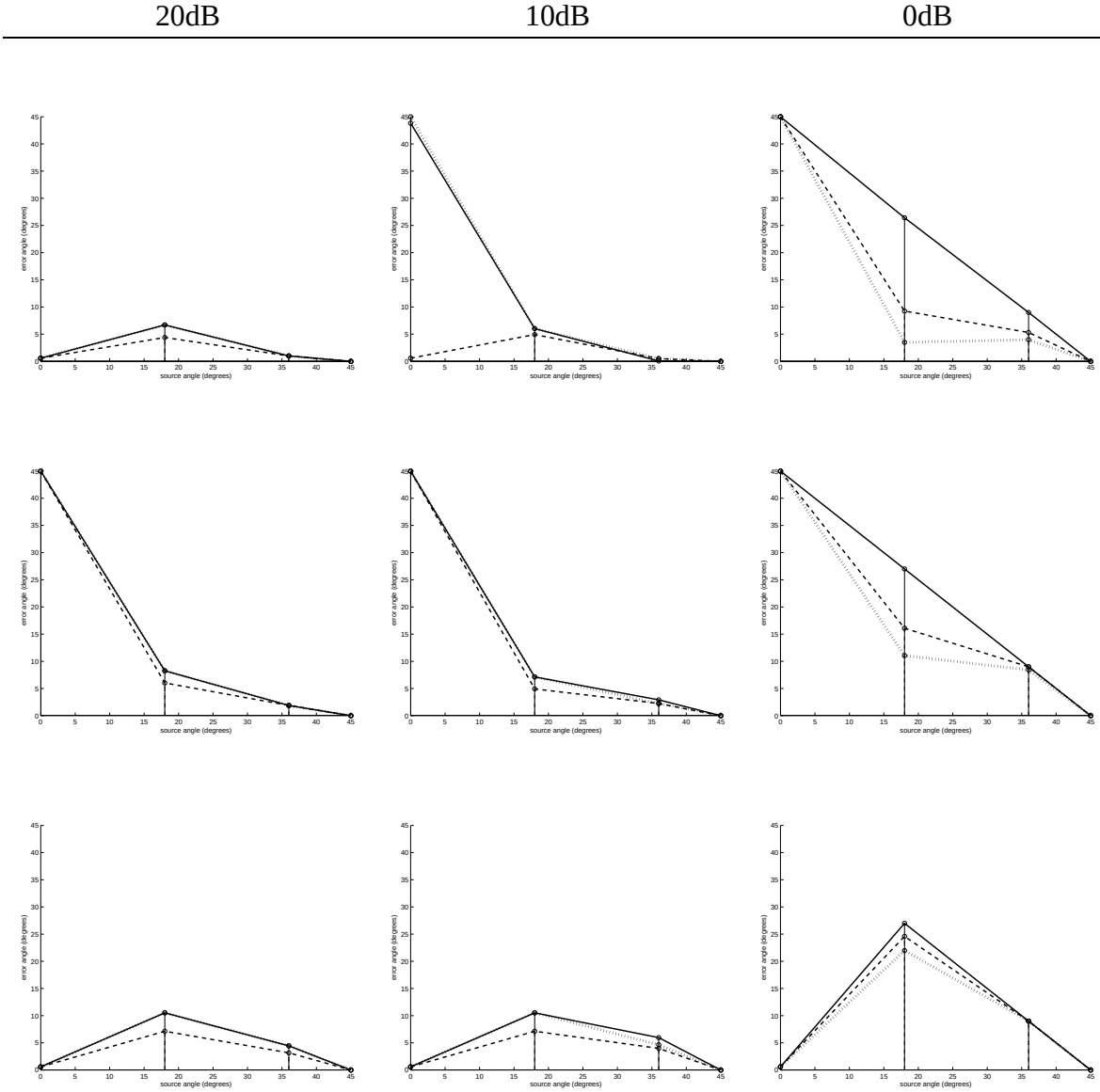

Figure 4.16 Angle error for a reverberation of 0.7 (top), 0.8 (center) and 0.9(bottom) and an SNR of 20, 10 and 0dB in a 5m x 5m room. The solid line indicates the angle error without the use of Hilbert Envelope approach, dotted indicates the error using the Hilbert function in Matlab and dashed the error using the Kaiser Hilbert Tranformer.

20dB	10dB	0dB

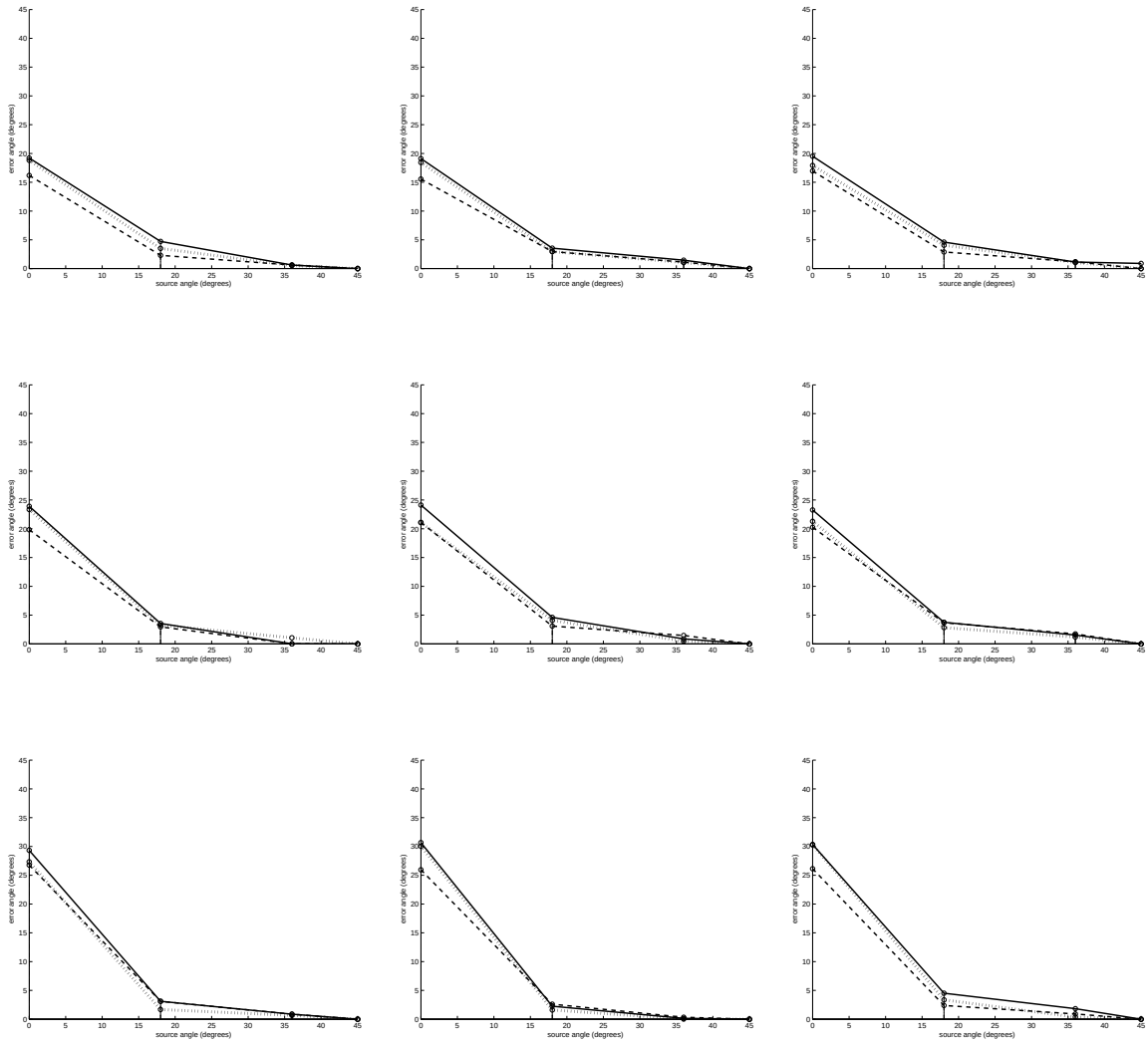

Figure 4.17 Angle error for a reverberation of 0.7 (top), 0.8 (center) and 0.9 (bottom) and an SNR of 20, 10 and 0dB in a 10m x 10m room. The solid line indicates the angle error without the use of Hilbert Envelope approach, dotted indicates the error using the Hilbert function in Matlab and dashed the error using the Kaiser Hilbert Tranformer.

It is seen that, in almost all the cases, the errors are less when the Hilbert Envelope approach is used, the lowest errors alternating between the regular Hilbert and the Kaiser Hilbert Transormer. But this still doesn't solve the problem, because the Hilbert Envelope method does not eliminate the error. From the results it seems to reduce the error on an average by about 25%.

Another idea to solve this would be to divide the entire signal into a number of frames and look at the angle error in each frame. In this case the signal is divided into 50 frames. The sampling rate is 44100Hz. Each frame has 4096 samples with an overlap of 2000 samples, i.e., about a 50% overlap. The length of each frame is 4096/44100 = 92.8ms. A couple of cases with large angle errors are presented here to see how this works.

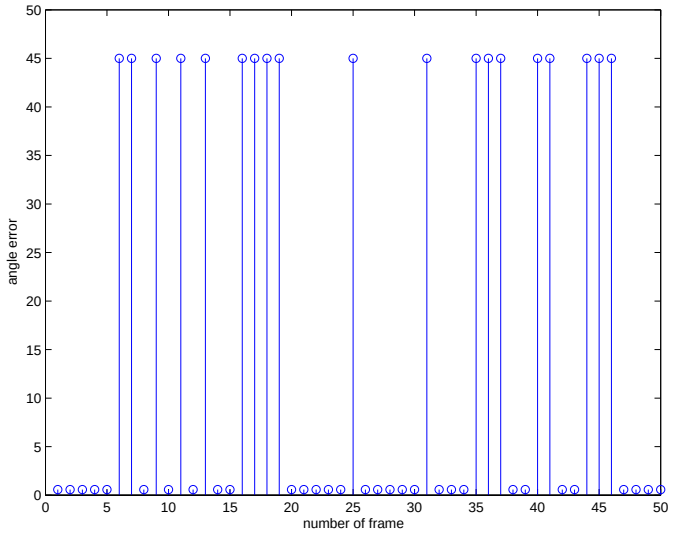

Figure 4.18 Angle error in each frame when the sound source is at 0 degrees and a distance of 2m, with an SNR of 0dB and a reflection coefficient of 0.8.

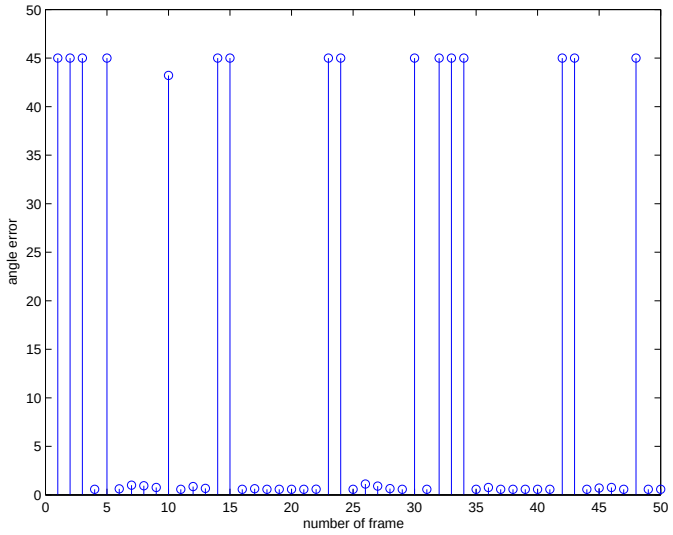

Figure 4.19 Angle error in each frame when the sound source is at 0 degrees and a distance of 1m, an SNR of 0dB and a reflection coefficient of 0.8.

Figure 4.18 shows the angle error in each frame when the sound source is at 0 degrees at a distance of 2m, with an SNR of 0dB and a reflection coefficient of 0.9. The mean angle error for this case is 18.1796 degrees, the standard deviation being 21.9349 degrees. Figure 4.19 show the angle error vs frame for the same specifications when the source is 1m away. The mean angle error in this case is 14.8477 degrees and the standard deviation is 20.8472 degrees. This does not do a good job for the 1m case, but it is expected because there would a few frames in the signal which would be blank and upon adding noise, these frames that have only noise would lead to large angle errors.

Figures 4.20 and 4.21 show the errors vs frame for the sound source at 18 degrees, when the distances are 2m and 1m, with an SNR of 0dB and a reflection coefficient of 0.9. The mean angle error in the former case is 15.9143 degrees, the standard deviation being 11.0557 degrees, while for the latter the mean error is 7.4964 degrees, the standard deviation being 5.8302 degrees. It can be seen that the angle error reduces by approximately 40% when the sound source is 2m away thereby improving the localization performance.

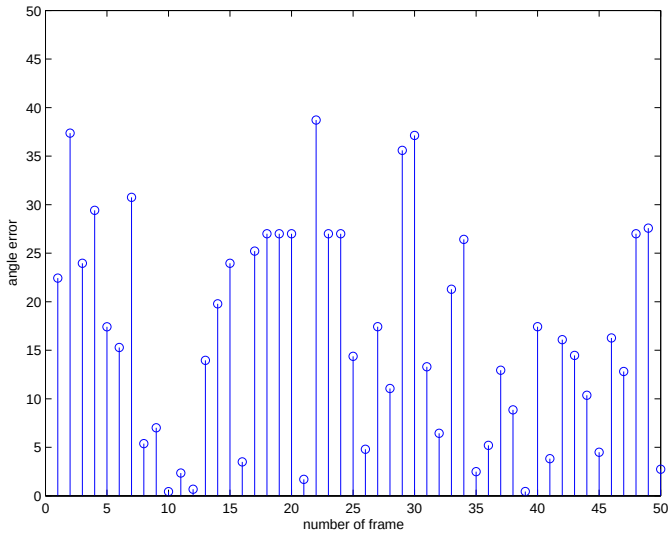

Figure 4.20 Angle error in each frame when the sound source is at 18 degrees and a distance of 2m, with an SNR of 0dB and a reflection coefficient of 0.8.

From the localization results obtained so far, it is seen that the algorithm computes the bearing angle to the sound source with 0.0 degree error, with the exception of extremely high noise and reverberation conditions. A number of experiments have been conducted

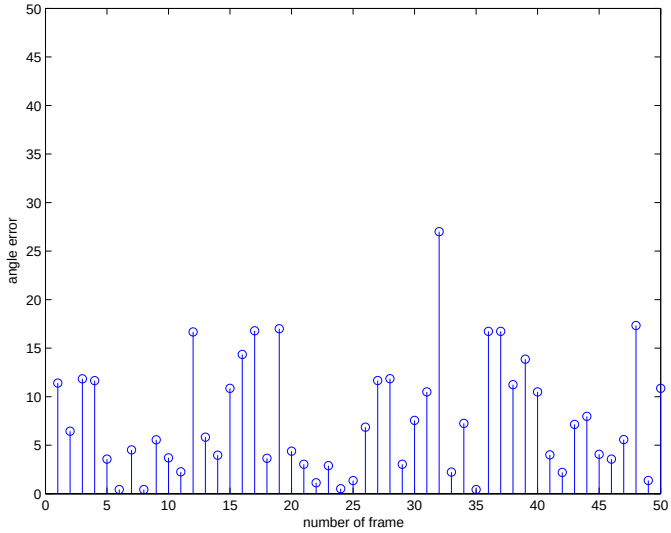

Figure 4.21 Angle error in each frame when the sound source is at 0 degrees and a distance of 1m, an SNR of 0dB and a reflection coefficient of 0.8.

by placing the sound source at some of the more challenging positions and setting reverberation to maximum values ($\theta = 0, 18, 36, 45, 72, 90$, $\rho = 1, 2$, $\beta = 0.7, 0.8, 0.9$). The variance is found from the plot of the Δ_E estimation. Under high reverberation and noise conditions, the localization and bearing angle depends not only on the size of the room but also on the source location and the relative positions of the microphones. However, the algorithm exhibits a bias toward locations far from the microphones and hence is unable to estimate distance accurately.

CHAPTER 5
CONCLUSION

Interaural level difference (ILD) is an important cue for acoustic localization in natural systems. The possibility of using ILD in computer-based systems has been investigated. Equations were derived that constrain the location of a sound source based upon received energy levels of two microphones, and an algorithm for computing the location using multiple microphone pairs has been proposed. The localization was performed using a combined likelihood approach, the results of which have been further improved by using a nonlinear processing approach known as the Hilbert envelope. Experiments in reverberant environments demonstrate the algorithm's ability to yield accurate results for several configurations even in a noisy and reverberant environment, thus validating the utility of the cue.

However, there are questions that need to be answered like the bias toward distant locations and sensitivity to reverberation, which can be cosidered as topics for further investigation. Also, experiments need to be performed in real environments to study the performance of ILD. The effect of occlusion on the level differences can be studied by placing an object inside the microphone array, or by placing the microphones on either side of the head. ILD can also be combined with ITD for performing Acoustic Localization in order to obtain more robust results.

APPENDICES

Appendix A

Allen and Berkley Algorithm

In this algorithm, a talker in a room is modelled as a point source in a rectangular cavity in the algorithm. A single frequency point source of acceleration in free space emits a pressure wave of the form

$$P(t; X, X') = \frac{exp[i\omega(R/c - t)]}{4\pi R},$$

where,

$$\begin{aligned}
P &= pressure, \\
\omega &= 2\pi f, \\
f &= frequency, \\
t &= time, \\
R &= |X - X'|, \\
X &= vector\ \ talker\ \ location\ \ (x, y, z), \\
X' &= vector\ \ microphone\ \ location\ \ (x', y', z'), \\
i &= \sqrt{-1}, \\
c &= speed\ \ of\ \ sound,
\end{aligned}$$

Let $x = [x, y, z]$ be a vector pointing to the source, and let $x' = [x', y', z']$ be a vector pointing to the microphone. Using the image method, each cluster contains eight points at $v_p = x^T(2p - 1)$, where $p = [p_x, p_y, p_z]$ contains three values each of which can be zero or one. The pointer to the origin of a cluster is $v_r = 2r^T l$, where $r = [r_x, r_y, r_z]$ and $l = [l_x, l_y, l_z]$ are the room dimensions. The distance from source to microphone is then given by $v_{p,r} = ||x' + v_p + v_r||$.

The pressure wave is given by

$$p(t; X, X') = \sum_{p=0}^{1} \sum_{r=-\infty}^{\infty} \frac{Q_{p,r}\delta[t - (v_{p,r}/c)]}{4\pi v_{p,r}} \tag{A.1}$$

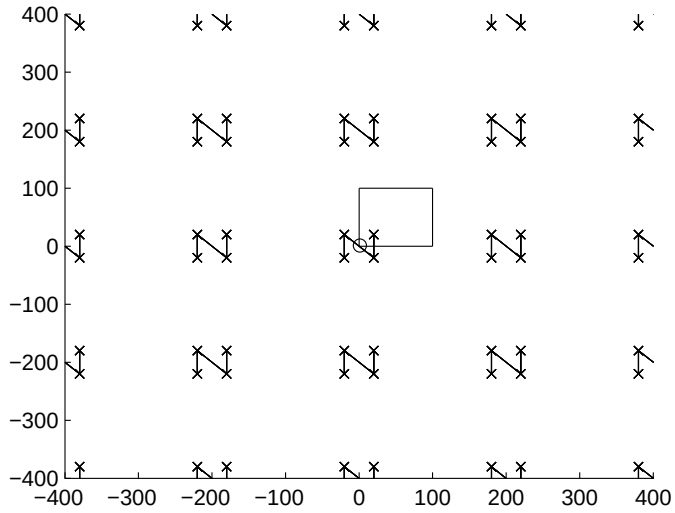

Figure A.1 A slice through the image space showing the spatial arrangement of the images of the source.

An interpretation of Equation A.1 is given in Figure A.1. The solid box represents the original room. The actual image space is three Dimensional.

With rigid(lossless) walls, $Q_{p,r} = Q_o$. With nonrigid walls, however,

$Q_{p,r} = Q_o \prod_{\in x,y,z} \beta_{i1}^{|r_i - p_i|} \beta_{i2}^{|r_i|}$

If we consider a one Dimensional case with two walls β_{x1} and β_{x2}, the signals reaching the microphone are

$$\begin{aligned}
p(t; X, X') &= \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x1} \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x2} \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x1}\beta_{x2} \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x2}\beta_{x1} \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x1}^2\beta_{x2} \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \beta_{x1}\beta_{x2}^2 \times \frac{Q_{p,r}\delta[t-(v_{p,r}/c)]}{4\pi v_{p,r}} \\
&+ \cdots
\end{aligned}$$

Appendix B

Derivation of Equation of Locus of ILD (Equation 2.2)

This section provides a detailed derivation of Equation 2.2, i.e. the locus of ILD:

$$
\begin{aligned}
E_1 d_1^2 &= E_2 d_2^2 \\
0 &= E_1[(x-x_1)^2 + (y-y_1)^2] - E_2[(x-x_2)^2 + (y-y_2)^2] \\
0 &= E_1[x^2 - 2x_1x + x_1^2 + y^2 - 2y_1y + y_1^2] - E_2[x^2 - 2x_2x + x_2^2 + y^2 - 2y_2y + y_2^2] \\
0 &= c_e x^2 - 2c_x x + c_e y^2 - 2c_y y + c \\
0 &= x^2 - 2\frac{c_x}{c_e}x + \frac{c_x^2}{c_e^2} - \frac{c_x^2}{c_e^2} + y^2 - 2\frac{c_y}{c_e}y + \frac{c_y^2}{c_e^2} - \frac{c_y^2}{c_x^2} + \frac{c}{c_e} \\
\frac{E_1 E_2 {d_{12}}^2}{c_e^2} &= \left(x - \frac{c_x}{c_e}\right)^2 + \left(y - \frac{c_y}{c_e}\right)^2
\end{aligned}
$$

This is the case for $E_1 \neq E_2$.

When $E_1 = E_2$, the equation reduces to

$$2c_x x + 2c_y y = c + \eta,$$

which is the equation of the line passing halfway between the microphones and perpendicular to the line joining them (i.e., the perpendicular bisector).

The generalized equation, therefore, is

$$[\,x \quad y \quad 1\,] \begin{bmatrix} c_e & 0 & -c_x \\ 0 & c_e & -c_y \\ -c_x & -c_y & c \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \eta, \tag{B.1}$$

where

$$\begin{aligned} c_e &= E_1 - E_2 \\ c_x &= E_1 x_1 - E_2 x_2 \\ c_y &= E_1 y_1 - E_2 y_2 \\ c &= E_1(x_1^2 + y_1^2) - E_2(x_2^2 + y_2^2). \end{aligned}$$

BIBLIOGRAPHY

[1] Michael S. Brandstein and Harvey F. Silverman, "Practical Methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, no. 2, pp. 91-126, 1997.

[2] Piergiorgio Svaizer, Marco Matassoni, and Maurizio Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, 1997, vol. 1, pp. 231-234.

[3] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environemnts," *ICASSP*, 2001.

[4] J. L. Flanagan, J.D. Johnston, R. Zahn, and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *Journal of the Acoustical Society of America(JASA)*, vol. 78, no. 5, 1985.

[5] R. Duraiswami, D. Zotkin, and L. Davis, "Active speech source localization by a dual coarse-to-fine search," *ICASSP*, 2001.

[6] Darren B. Ward and Robert C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," *ICASSP*, 2002.

[7] Stanley T. Birchfield and Daniel K. Gillmor, "Acoustic source direction by hemisphere sampling," *ICASSP*, 2001.

[8] Stanley T. Birchfield and Daniel K. Gillmor, "Fast Bayesian Acoustic Localization," *ICASSP*, 2002.

[9] Stanley T. Birchfield, "A Unifying framework for Acoustic Localization," *Proceedings of the 12th European Signal Processing conference(EUSIPCO)*, 2004.

[10] Stanley T. Birchfield and Rajitha Gangishetty, "Acoustic Localization by Interaural level difference," *ICASSP*, 2005.

[11] Charles H. Knapp and G. Clifford Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320-327, Aug. 1976.

[12] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *Journal of Acoustical Society of America*, vol. 105, no. 5, pp. 2914-2919, 1999.

[13] Maurizio Omologo and Piergiorgio Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no.3, 1997.

[14] Alex Stephenne and Benoit Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," *ICASSP*, vol. 5, pp. 3055-3058, 1995.

[15] David R. Fischell and Cecil H. Coker, "A Speech Direction Finder," *ICASSP*, 1984.

[16] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, Cambridge, Massachusetts: The MIT Press, 2001.

[17] William. A. Yost, *Fundamentals of Hearing: An Introduction*, San Diego: Academic Press, 2000.

[18] M. S. Brainard and Eric I. Knudsen and Steven D. Esterly, "Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues," *Journal of Acoustical Society of America*, vol. 91, no.2, pp. 1015-1027, 1992.

[19] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

[20] Y. Huang, J.Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," *Proceedings of the IEEE, ICASSP*, pp. 937-940, 1999.

[21] M. S. Brandstein, J. E. Adcock, J. H. DiBiase and H. F. Silverman, " A Closed-Form Method for Finding Source Locations from Microphone-Array Time-Delay Estimates," *Proceedings of the IEEE, ICASSP*, pp. 3019-3022, 1995.

[22] Barbara Shinn, Cunningham, "Learning Reverberation: Considerations for Spatial Auditory Displays," *Proceedings of the International Conference on Auditory Display*, pp. 126-133, 2000.

[23] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, India: Pearson Education, Inc., 2003.

[24] F. J. Owens, *Signal Processing of Speech*, United Kingdom: Macmillan New Electronics, 2003.

[25] Mario F. Triola, *Elementary Statistics*, Massachusetts: Addison-Wesley Reading, seventh edition, 1998.