A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays

by

Joseph Hector DiBiase

B.S., Trinity College, 1991

Sc.M., Brown University, 1993

Thesis

Submitted in partial fulfillment of the requirements for

the Degree of Doctor of Philosophy

in the Division of Engineering at Brown University

Providence, Rhode Island

May 2000

The Vita of Joseph Hector DiBiase

Joseph was born December 26, 1969 in Providence, Rhode Island. He grew up in Cranston, Rhode Island until he and his family moved to Jamestown, Rhode Island in 1979. He attended Trinity College in Hartford, Connecticut from 1987 until 1991, graduating with a Bachelor of Science degree in Electrical Engineering with departmental honors. He went on to Brown University in Providence, Rhode Island to study signal processing and began research on microphone arrays. He received a Master of Science degree in Electrical Engineering in 1993 and continued to pursue his work towards a Doctor of Philosophy degree. While a student at Brown, he held several appointments as a research assistant. He also held several appointments as a teaching assistant for various electrical engineering courses.

Acknowledgements

I wish to thank my advisor, Professor Harvey Silverman, for his constant support and feedback. Thanks to the other members of the microphone array group for their efforts to keep the project going: John Adcock, Michael Brandstein, Stu Kirtman and Paul Meuse. I thank all my fellow graduate students, my friends, for all those special lunches, inside and outside: Michael Blane, Aaron Smith, Mike Wazlowski and the array group. Thanks to my readers, Professor Michael Brandstein and Professor David Cooper for their corrections to my thesis and their attentiveness during my defense. Thanks to Professor Bill Patterson for answering a wide range of questions for me. Thanks to the Department of Engineering for the research and teaching assistant appointments that were granted to me as I worked towards my degree. Thanks to LEMS for providing the laboratory facilities for my research and to Arpie Kaloustian for all his technical assistance on the array systems and other equipment in the lab. I would also like to thank my parents, Peter and Linda, for their investment in my education and their assurance that this was a worthwhile goal. And to my wife, Furhana, for providing support when I needed it the most and for helping me laugh during this long and challenging process.

Contents

List of Tables

List of Figures

1 Introduction

A combination of microphone arrays and sophisticated signal processing has been applied to the remote acquisition of high-quality speech audio. These applications all exploit the spatial filtering ability of an array, which allows the speech signal from one talker to be enhanced as the signals from other talkers and unwanted sources are suppressed. This process in generally referred to as *beamforming*. While some array-systems are designed to focus on sounds emanating from a preset location or direction, most employ adaptive algorithms that track the positions of one or more talkers and adjust the array's focus accordingly. This "electronically steerable" feature eliminates the need for manually operated equipment, such as shotgun or boom-mounted microphones. Furthermore, an array-system has the potential to replace the use of hand-held or head-mounted microphones in some applications.

Microphone arrays have been implemented in many applications, including teleconferencing [25][35][60][61][96], speech recognition [2][21][22][40][55][56][79], talker characterization [91] and voice capture in reverberant environments [34][39][57][98]. Some novel and interesting array designs have been studied, including a small spherical array [31] and one employing superdirectivity [24]. Both theoretical and practical aspects of array-systems are being actively researched, as reported by the participants of three special microphone array workshops [36][37][38]. Some of this work has been based on simulations using mathematical models (such as [3]) of the acoustic environment [57][82], and other work relies on pre-recorded array data of actual talkers. Still other work focuses on the design and construction of hardware [63][86], as well as the implementation of real-time software [29][76].

With the emergence of powerful and inexpensive DSP microprocessors, microphone array-systems have been introduced as commercially viable products. Examples of this are the teleconferencing products by *PictureTel* and *Ploycom*. Both companies have applied microphone-array technology to quality voice-capture products designed for use in small-room environments. There are also products by these companies that automatically steer a robotic camera and frame active talkers. The camera-steering array-system by *PictureTel* uses the location estimates produced by a 4-element array [96].

Most of these applications require accurate passive localization techniques that produce estimates at a high rate with minimal latency. When tracking multiple, moving talkers [92], there must be many

reliable location estimates produced per second. If a beamformer is to be used to focus on these talkers, then their motion must be negligible for the duration of each data segment used to compute an estimate. Furthermore, the update rate must be high enough to avoid the undesirable effects of misaiming. These effects include high-frequency rolloff in the beamformer output [5], and a general attenuation of the target source signal. Furthermore, the latency due to the accumulation of long data segments for processing before beamforming may result in unacceptable delays between the production of the speech by the talker and the output of that speech through the beamformer. For real-time applications, such long delays can be quite disruptive. These factors place tight constraints on the microphone data requirements. While the computation time required by the algorithm largely determines the latency of the locator, it is the data requirements that define theoretical limits. Hence, this thesis focuses on reducing the size of the data segments necessary for accurate source localization in realistic room environments.

The performance of voice-capture techniques generally improves with the number of microphones in the array, and this has spawned the research and construction of medium [29] and large array systems [86]. When acoustic conditions are favorable, source localization can be performed using a modest number of microphones. For example, the automatic voice-steering camera by *PictureTel* includes only four microphones. Hence, in this regard, large arrays composed of tens or hundreds of microphones are redundant. By integrating the data from a multitude of microphones, the redundancy of a large array can be exploited to improve localization in the presence of adverse acoustic effects such as reverberation and background noise.

For various reasons, including the reduction of computational costs, many source-localization algorithms break the array into pairs of microphones (See Section 1.1 for references of work in this area.). Pairwise time-delay estimation (TDE) is used to determine the time difference of arrival (TDOA) of speech sounds between the microphones comprising each pair. The redundancy of a multitude of TDOA estimates has been exploited by statistically averaging in some way to give an estimate of the talker's location [12]. However, pairwise techniques suffer considerably from acoustic reverberation. The performance of pairwise techniques improves with the amount of data used, but the desire for high update rates and low latency places strict limits on this.

When a system has a multitude of microphones, far more than a sufficient number for source-localization, they should be used in a manner that will make the algorithm robust to reverberation. The application of error-prone pairwise TDE does not seem to be the best way to achieve this. An alternative approach is one where a beamformer is used to search over a predefined spatial region looking for a peak (or peaks) in the power of its output signal [59]. While this is computationally more intensive than pairwise methods, it inherently combines the signals from multiple microphones rather than reducing the data from each pair to a single time-delay parameter. This approach is able to compensate for the short duration of each data segment used for localization by integrating the data from many, or all, of the microphones prior to parameter estimation. An additional advantage that beam-steering techniques have over TDE-based techniques is the ability to localize multiple simultaneous talkers. In such a scenario, the power of a steered beamformer should peak multiple times, and each peak should correspond to the location of an active talker. Although these techniques have not been a popular choice for speech-array applications, a new steered-beamformer method is proposed in this thesis, which combines the best features of the beamformer with those of a popular pairwise technique. It will be demonstrated that this new steered-beamformer produces highly reliable location estimates, in rooms with reverberation times of 200 and 400 milliseconds, using 25-millisecond data segments.

1.1 Methods for Pairwise Time-Delay Estimation

Many passive talker localization techniques rely on pairwise time delay estimation (TDE) [11][13][64]. These techniques use the time difference of arrival (TDOA) of speech sounds between two spatially separated microphones to parameterize the source location [7][8][12][20][93]. The best results are obtained when the microphone pairs are strategically positioned to give optimal spatial accuracy [9][10]. Pairwise TDE has been applied to automatic camera steering for videoconferencing [96]. For this application of TDE, update rates of 200-300ms are acceptable. With such long data segments, reliable estimates are produced, even in adverse acoustic conditions. However, applications such as adaptive beamforming and the tracking of multiple talkers [92] require a much higher estimate rate; positional estimates must be updated as quickly as 20-30ms. When the data segments become this short, acoustic

reverberation has a severe impact on the performance of pairwise delay estimators. Many techniques have been proposed to improve their performance in reverberant environments.

The most common pairwise TDE method is generalized cross-correlation (GCC) [64]. The type of filtering, or weighting, used with GCC is crucial to performance. Maximum likelihood (ML) weightings are theoretically optimal when there is single-path propagation in the presence of uncorrelated noise, but their performance degrades significantly with increasing reverberation [19]. The *phase transform* (PHAT) weighting is more robust against reverberation than ML, even though it is sub-optimal under ideal conditions. Also known as the *cross-power spectrum phase* (CSP), GCC-PHAT has been shown to perform well in a realistic environment [72].

Other approaches, such as cepstral prefiltering [88], attempt to deconvolve the effects of reverberation prior to applying GCC. However, deconvolution requires long data segments since the duration of a typical small-room impulse response is 200-400ms. It is also very sensitive to the high variability and non-stationarity of speech signals. In fact, the experiments performed in [88] avoided the use of speech as input altogether. Instead, colored Gaussian noise was used as the source signal.

While identification of room impulse responses is impossible when the source signal is unknown, the method proposed in [54], which is based on eigenvalue decomposition, efficiently detects the direct paths of two impulse responses. This method is effective with speech as input, but requires 250ms of microphone data to converge.

Reverberation effects can also be overcome to some degree by classifying TDEs acquired over time and associating them with the direction of arrival (DOA) of the sound waves [90]. This approach, however, is not suitable for short-time TDE. Under extreme acoustic conditions, a large percentage of the TDEs are anomalous, and it takes a considerable period (1-2 seconds in [90]) to acquire enough estimates for a statistically meaningful classification.

A short-time TDE method, which is more complex than GCC, is presented in [14]. It involves the minimization of a weighted least-squares function of the phase data. It was shown to outperform both GCC-ML and GCC-PHAT in reverberant conditions. This improvement comes at a cost; since the phase data is discontinuous, a complicated searching algorithm must be applied to the minimization. The marginal improvement over GCC-PHAT may not justify this added cost in computational complexity.

Among the methods described here, those that rely on long data segments generally outperform those that do not. [81] is another example of a GCC method that performs adequately only when the data segments are sufficiently long in duration. Cross-correlation techniques are known to improve with increasing data lengths. Hence, it is not surprising that GCC-based TDE methods also improve with more data. Those that are not GCC-based generally require larger amounts of data to be effective as well. However, the dynamic environments of many speech array applications require high update rates, which limits the duration of the data segments.

1.2 Methods for Steered-Beamformer Localization

Methods that rely on the array's ability to *focus* on signals originating from a particular location or direction in space are generally referred to as *beamformers* [59]. When the location of the source is not known, a beamformer can be used to scan, or *steer*, over a predefined spatial region by adjusting its steering parameters. The output of a beamformer, when used in this way, is known as the *steered response*. The steered response power (SRP) may peak under a variety of circumstances, but with favorable conditions, it is maximized when the point (or direction) of focus matches the location of the source.

Beamforming has been used extensively in speech-array applications for voice capture [37][38][23][41][97]. However, due to the efficiency and satisfactory performance of pairwise correlation methods, it has rarely been applied to the talker localization problem. Furthermore, the steered response of a conventional beamformer is highly dependent on the spectral content of the source signal. Many optimal derivations are based on *a priori* knowledge of the spectral content of the background noise, as well as the source signal [17][45]. In the presence of significant reverberation, the noise and source signals are highly correlated, and this makes accurate estimation of the noise nearly impossible. Furthermore, in nearly all array-applications, little or nothing is known about the source signal. Hence, such optimal estimators are not very practical in realist speech-array environments.

The simplest type of steered response is obtained using the output of a *delay-and-sum* beamformer. This is what is most often referred to as a *conventional* beamformer [59]. Delay-and-sum beamformers apply time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. Once these signals are time-aligned, they are summed together to

form a single output signal. More sophisticated beamformers apply filters to the array signals as well as this time alignment. The derivation of the filters in these filter-and-sum beamformers is what distinguishes one method from the other.

Many optimal steered-beamformer techniques have been derived for stationary, narrow-band signals. These include minimum variance beamforming [58][26][66], linear prediction [58] and generalize sidelobe cancellers [44][16]. These methods can be extended to the wideband case and are appropriate for speech signals when applied over short, stationary segments. However, the beamformer filters for all of these methods are defined in terms of the *spatial correlation matrix*. When this matrix is unknown, it must be estimated using the observed data. Such estimation, especially in adverse acoustic conditions, may require long segments of stationary data. For the dynamic conditions of speech-array applications, long interval for which the source is both spatial and temporally stationary are rarely encountered. Hence, such methods are difficult to apply to the localization of speech sources.

In this thesis, filters for a steered-beamformer are derived, which incorporate the features of a popular pairwise technique known as the *phase transform* (PHAT). The phase transform is a sub-optimal method, although it has been shown to perform well in reverberant environments. In this thesis, it will be demonstrated that this new steered-beamformer produces highly reliable location estimates, in rooms with reverberation times of 200 and 400 milliseconds, using 25-millisecond data segments. It is compared to the conventional form of steered-beamformer localization and to a pairwise technique based on the phase transform. Using unique microphone array data sets, recorded in realistic environments, the new technique is demonstrated to be more robust to reverberation than the other two methods.

1.3 This Thesis

This thesis attempts to show that pairwise localization techniques yield inadequate performance in some realistic small-room environments. Unique array data sets were collected using specially designed microphone array-systems. Through the used of this data, various localization methods were analyzed and compared. These methods are based on both the generalized cross-correlation (GCC) and the steered response power (SRP). The GCC techniques studied include the phase transform, which has been dubbed "GCC-PHAT". The beam-steering methods are based on the conventional steered response power (SRP)

and a new filter-and-sum technique dubbed "SRP-PHAT". The goals of this work can be summarized as follows:

- To show that mild reverberation can severely impact the performance of short-time GCC-based localization techniques
- To show that microphone redundancy, which exists in many array systems, can be exploited to reduce the data requirements for accurate talker localization in reverberant environments
- To examine the performance of steered-response localization techniques when applied to realistic speech-array data sets
- To propose a new steered-beamformer localization method, SRP-PHAT, that is more accurate than both the conventional method and the popular pairwise method, GCC-PHAT.

These goals were addressed over the course of six chapters. Each chapter builds on the material presented in the chapters before it. The material is organized as summarized by the following paragraphs.

In Chapter 2, a microphone signal model is derived, which is then used to describe how propagating sound waves interact with an array of microphones. This basic introduction to commonly exploited acoustic laws justifies much of the theoretical aspects of the source localization techniques explored in this thesis. The work in the following chapters builds on these acoustic models.

Chapter 3 describes the primary set of array data, which was used in the source localization experiments of Chapters 5 and 7. This data set was recorded using a custom-built microphone-array system called the *Brown Megamike II*. The Megamike microphone-array configuration, the collection procedure and the basic block-processing scheme that has been applied to the data are described. Some preliminary measurements of this data set are presented, including microphone signal-to-noise ratios (SNRs), room impulse responses and room reverberation times.

Chapter 4 introduces GCC and GCC-PHAT. It also describes the implementation of these methods in Chapters 5 and 7. Chapter 5 includes a series of three experiments designed to establish baseline performance of GCC in a mildly reverberant, high-SNR environment. The experiments in this chapter illustrate the shortcomings of GCC-based localization methods and shed some light on possible ways to improve performance. This chapter also introduces some basic performance measures that are applied to all the experiments in this thesis.

Chapter 6 introduces the steered response and how it can be used for source localization. It proposes new filters for the filter-and-sum beamformer. The new technique, SRP-PHAT, exploits microphone redundancy by combining the microphone signals, rather than combining a multitude of TDOA estimates, to enhance the accuracy of location estimation. The performance of SRP-PHAT was compared to SRP and GCC-PHAT in three experiments, which are described in Chapter 7. The experiments in this chapter include data from both small aperture and large aperture arrays.

In all the experiments in this thesis, 25-milliseconds data blocks were used to emphasize the importance of fast and accurate source localization. It will be shown by the results of the experiments that SRP-PHAT outperforms SRP and GCC-PHAT using real microphone array data that was collected in rooms having 200 and 400 millisecond reverberation times. These results, as well as future work, are summarized in Chapter 8.

2 Sound Wave Propagation

Throughout this thesis, it will be assumed that sound waves propagate as predicted by the linear wave equation [62]. With this assumption, the acoustic paths between sound sources and microphones can be modeled as linear systems [99]. This is clearly advantageous to the analysis and modeling of the signals produced by the microphones. In order for such linear models to be valid, the propagation medium must have certain properties. These properties will be defined as part of the *simple acoustic conditions* on which much of this thesis is based. These conditions are realistic in small-room speech-array environments and are regularly exploited by array-processing techniques [59]. Once defined, the simple acoustic conditions are applied to the derivation of a microphone signal model in this chapter, which is then used to describe how propagating sound waves interact with an array of microphones.

2.1 Simple Acoustic Conditions

Simple acoustic conditions will be used to describe properties of the acoustic source, such as a loudspeaker or human head, as well as the acoustic medium, which is air. These conditions are defined as follows:

1. *The source emits spherical sound waves.* It will be assumed that the size and shape of the radiator does not significantly impact sound-wave propagation. More realistic radiation patterns of the human head were reported in [68] and [69], however, incorporation of such complex models is beyond the scope of this thesis.
2. *The Doppler effect is negligible.* The source may be in motion, but its speed is obviously not comparable to the speed of sound. Hence, there is no meaningful Doppler shift in frequency.
3. *The medium is homogeneous.* Homogeneity dictates that the propagation speed of sound is constant everywhere, at all times[1], and is equal to the known value, c. In other words, the medium in non-refractive.

[1] This is for all times during the course of a single experiment and does not preclude adjustment of the speed of sound from one experiment to the next based on environmental changes such as temperature. Temperature accounts for many first order differences since its square root is inversely proportional to the speed of sound.

4. *The medium is nondispersive.* Dispersion causes propagation speed to vary with frequency, which is not consistent with the properties of a linear system. Hence, for the linear system analogy to hold, the effects of dispersion must be negligible.
5. *The medium is lossless.* A lossless medium does not absorb energy from propagating waves. Attenuation is determined strictly by the spherical shape of the waves and is independent of frequency.

2.2 Direct Path Propagation

Under simple acoustic conditions, a wave field at a spatially fixed microphone is linearly related to an assumed single, fixed source signal, $s(t)$, which created the wave field. This is true for propagation in free space as well as inside an acoustic enclosure (such as a room). In free space, sound waves propagate without interference by objects such as walls, furniture and other people. Such a free-space model is not very realistic in small-room, speech-array environments. However, it accurately describes the direct-path propagation from source to microphone, even in the presence of reverberation. The linearity of the medium allows the microphone signal to be modeled as the superposition of a direct-path component plus the sound waves that are reflected by the surfaces of the room. Signal processing algorithms rely on separating the direct-path component from reverberation and noise since it parameterizes the location of the talker.

Direct-path propagation is easily derived from the wave equation [62]. The wave-field at distance r from the sound source $s(t)$ can be expressed as follows:

$$f_{direct}(r,t) = \frac{a}{r} s(t - r/c)$$

This expression shows the wave field as a scaled and time-delayed version of the original source signal. The attenuation factor is inversely proportional to the distance from the source, and the time delay is equal to the ratio of this distance to the speed of sound, c. The constant a depends on the medium and the system of units used.

In the three-dimensional Cartesian coordinate system illustrated by Figure 2.1, the distance between the source and the microphone indexed by m is defined as:

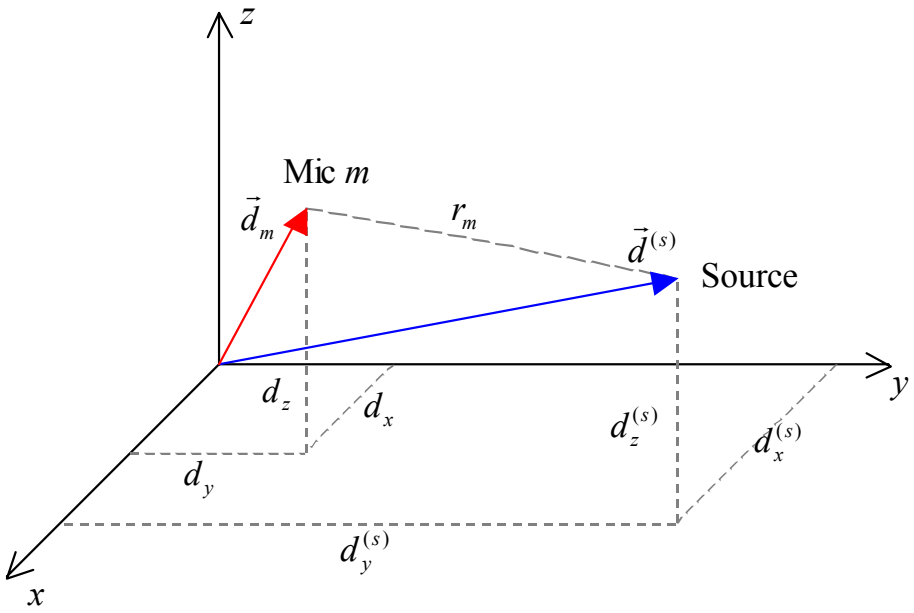

Figure 2.1 A source and microphone located in a Cartesian coordinate system.

$$r_m \equiv \left| \vec{d}_m - \vec{d}^{(s)} \right|$$

$\vec{d}_m$ and $\vec{d}^{(s)}$ are 3-element vectors, which are defined by the Cartesian coordinates of microphone m and the source, respectively:

$$\vec{d}_m \equiv \begin{bmatrix} d_x \\ d_y \\ d_z \end{bmatrix} \qquad \vec{d}^{(s)} \equiv \begin{bmatrix} d_x^{(s)} \\ d_y^{(s)} \\ d_z^{(s)} \end{bmatrix}$$

The propagation time from the source to microphone m can be defined as follows:

$$\tau_m \equiv \frac{r_m}{c}$$

Hence, the wave-field at location $\vec{d}_m$, which is produced by a single source located at $\vec{d}^{(s)}$, can also be expressed as follows:

$$f_{direct}(\vec{d}_m, \vec{d}^{(s)}, t) = \frac{a}{r_m} s(t - \tau_m) \tag{2.1}$$

2.3 Multi-Path Propagation and the Room Impulse Response

In the presence of sound-reflecting surfaces, the sound waves produced by a single source propagate along multiple acoustic paths. This gives rise to the familiar effects of reverberation; sounds reflect off objects and produce echoes. The walls of most rooms are reflective enough to create significant reverberation. While it is not always noticeable to the occupants, even mild reverberation can severely impact the performance of speech-array systems. Hence, multi-path propagation must be incorporated into the signal-processing model.

The walls of a room delineate an acoustic enclosure. Acoustic enclosures have been modeled extensively as linear systems [99], and the same techniques have been applied to room acoustics [65][71][47][94]. It has been shown that the wave field at a particular location inside a reverberant room mat also be considered to be linearly related to the source signal, $s(t)$. This relationship can be expressed in terms of the convolution of $s(t)$ with a *room impulse response* as follows:

$$f(\vec{d}_m, \vec{d}^{(s)}, t) = s(t) * h(\vec{d}_m, \vec{d}^{(s)}, t) \tag{2.2}$$

The impulse response, $h(\vec{d}_m, \vec{d}^{(s)}, t)$, characterizes all acoustic paths from the source to the location $\vec{d}_m$, including the direct path. It is a function of $\vec{d}_m$ as well as the source location, $\vec{d}^{(s)}$, and is highly dependent on these parameters.

In general, $h(\vec{d}_m, \vec{d}^{(s)}, t)$ varies with environmental changes, such as temperature and humidity. It also varies with the movement of furniture and people inside the room. While it has been shown that such variations are significant [51], it is reasonable to assume that these factors remain constant over short periods. Hence, a room impulse response may be considered time-invariant for short periods when the source and microphone are spatially fixed.

2.4 A Hybrid Multi-Path Model

The room impulse response model of Equation 2.2 does not require all simple acoustic conditions to hold. Note that, in general, frequency-dependent attenuation in the medium and phase distortion due to a non-point source radiator can be modeled by the impulse response. Hence, not only does it characterize all

reflected paths, it also models the direct-path component more realistically than Equation 2.1. However, in practical situations, the room impulse response is unknown, and there isn't enough information to estimate it. A more useful model can be expressed as a hybrid of the room impulse response model and the direct-path model of Equation 2.1. If all simple acoustic conditions apply to the direct-path sound, but not necessarily to the reflected sound, the wave field at location $\vec{d}_m$ can be approximated by the sum of the direct-path component plus the reflected sound as follows:

$$f(\vec{d}_m, \vec{d}^{(s)}, t) \approx \frac{a}{r_m} s(t - \tau_m) + s(t) * u(\vec{d}_m, \vec{d}^{(s)}, t) \tag{2.3}$$

In this equation, the reflected sound has been expressed as a filtered version of the original source signal, $s(t)$. The impulse response $u(\vec{d}_m, \vec{d}^{(s)}, t)$ characterizes all acoustic paths ***except*** the direct path; it characterizes the reverberation. Like the direct-path component, the reverberation component is a function of both the sample-location of the wave field, $\vec{d}_m$, as well as the location of the source, $\vec{d}^{(s)}$. This model can be viewed as the room impulse response model of Equation 2.2 with the following approximation for $h(\vec{d}_m, \vec{d}^{(s)}, t)$:

$$h(\vec{d}_m, \vec{d}^{(s)}, t) \approx \frac{a}{r_m} \delta(t - \tau_m) + u(\vec{d}_m, \vec{d}^{(s)}, t)$$

By separating the direct-path component from the reverberation, the hybrid model of Equation 2.3 is expressed explicitly in terms of the parameter of interest, namely the time delay, τ_m. Furthermore, it models the reflected sound as realistically as the room impulse response model of Equation 2.2. However, complete knowledge of $u(\vec{d}_m, \vec{d}^{(s)}, t)$ is generally not necessary for this hybrid model to be useful. Partial knowledge of $u(\vec{d}_m, \vec{d}^{(s)}, t)$, such as its duration and strongest peaks may yield improvements in the methods used to estimate τ_m.

2.5 Microphone Signal Model

It will be assumed that the signal produced by a microphone fixed at location $\vec{d}_m$ is the superposition of two components: a filtered version of the single-source wave field at $\vec{d}_m$ plus noise. The index of the microphone at this location is *m*, and its signal can be expressed as follows:

$$x_m(t) = f(\vec{d}_m, \vec{d}^{(s)}, t) * \gamma_m(\vec{d}^{(s)}, t) + v_m(t) \tag{2.4}$$

$\gamma_m(\vec{d}^{(s)}, t)$ is a linear filter that characterizes the frequency and phase responses of the m-th microphone channel. These responses include electrical, mechanical and acoustical properties of the microphone system. In general, the microphone's directivity pattern makes its response a function of its orientation as well as its location in space. For a microphone with a fixed location and orientation, which are implied by that microphone's index, m, $\gamma_m(\vec{d}^{(s)}, t)$ is generally a function of the source location, $\vec{d}^{(s)}$. $v_m(t)$ is the noise present in the m-th channel, which accounts for any nonlinear effects in the system. This noise term may also include any propagating isotropic noise that could be produced by fans, or other mechanical equipment inside the room. Such propagating noise is usually considerably more significant than the channel noise and tends to dominate this additive term. Generally, $v_m(t)$ is assumed to be uncorrelated with $s(t)$.

By combining Equations 2.2 and 2.4, the microphone signal can be expressed in terms of the room impulse response for a fixed source located at $\vec{d}^{(s)}$:

$$x_m(t) = s(t) * h(\vec{d}_m, \vec{d}^{(s)}, t) * \gamma_m(\vec{d}^{(s)}, t) + v_m(t) \tag{2.5}$$

From Equation 2.5, it can be seen that the impulse response from the source-output to the microphone-output is the convolution of two terms: $h(\vec{d}_m, \vec{d}^{(s)}, t)$ and $\gamma_m(\vec{d}^{(s)}, t)$. Denoting this convolution by $\widetilde{h}_m(\vec{d}^{(s)}, t)$, the microphone signal can be expressed more compactly as follows:

$$x_m(t) = s(t) * \widetilde{h}_m(\vec{d}^{(s)}, t) + v_m(t) \tag{2.6}$$

Because $\gamma_m(\vec{d}^{(s)},t)$ is not necessarily invertible or known, $h(\vec{d}_m,\vec{d}^{(s)},t)$ is not necessarily recoverable from $\widetilde{h}_m(\vec{d}^{(s)},t)$. The source-output to microphone-output is more easily measured than the room impulse response. As a result, Equation 2.6 is more useful in practice than Equation 2.5.

Using the hybrid wave-field model of Equation 2.3, $\widetilde{h}_m(\vec{d}^{(s)},t)$ can be approximated by:

$$\widetilde{h}_m(\vec{d}^{(s)},t) \approx \frac{a}{r_m}\delta(t-\tau_m) * \gamma_m(\vec{d}^{(s)},t) + u(\vec{d}_m,\vec{d}^{(s)},t) * \gamma_m(\vec{d}^{(s)},t) \tag{2.7}$$

Recall that $u(\vec{d}_m,\vec{d}^{(s)},t)$ characterizes the reverberation at microphone m, which is produced by a single source fixed at location $\vec{d}_m$. This quantity is more simply denoted by $u_m(\vec{d}^{(s)},t)$. Substituting this notation into Equation 2.7 and combining this equation with Equation 2.6 yields the following approximation for the signal produced by microphone m:

$$x_m(t) \approx \frac{1}{r_m}s(t-\tau_m) * \gamma_m(\vec{d}^{(s)},t) + s(t) * u_m(\vec{d}^{(s)},t) * \gamma_m(\vec{d}^{(s)},t) + v_m(t)$$

Note that the constant a has been absorbed by the channel filter, $\gamma_m(\vec{d}^{(s)},t)$.

For simplicity, let $\breve{v}_m(t)$ define a new noise term, which is the sum of the reverberation noise plus the original noise:

$$\breve{v}_m(t) = s(t) * u_m(\vec{d}^{(s)},t) * \gamma_m(\vec{d}^{(s)},t) + v_m(t)$$

When it is convenient to do so, the microphone signals can be expressed as the sum of this noise term plus the direct-path signal:

$$x_m(t) = \frac{1}{r_m}s(t-\tau_m) * \gamma_m(\vec{d}^{(s)},t) + \breve{v}_m(t) \tag{2.8}$$

In this form, a delayed and scaled version of the source signal, $s(t)$, is shown explicitly. Most all localization techniques depend on the direct-path component to parameterize source locations. Hence, it is useful to show that this component exists in the microphone signal, despite the fact that the noise term may include strong noise and reverberation.

2.6 Direction of Propagation and Arrival

The direct-path component of Equation 2.8 represents the propagation of sound waves along a straight line from the source to microphones m. The direction these waves travel as they impinge on microphone m is known as the *direction of propagation*. Similarly, the opposite direction defines the *direction of arrival*, or *DOA*, which is equivalent to the bearing[2] of the waves as they approach the microphone. Both terms will be used to describe the way sound waves interact with an array of microphones.

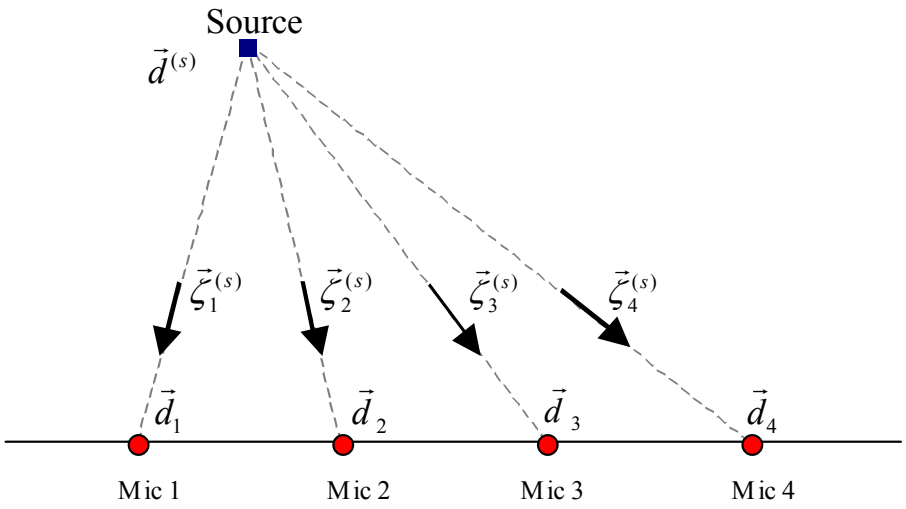

Figure 2.2 Propagation vectors.

2.6.1 Direction of Propagation

In general, an array is composed of M microphones, and each microphone is positioned at a unique spatial location. Hence, the direct-path sound waves propagate along M *bearing lines*, from the source to each microphone, simultaneously. The orientations of these lines in the global coordinate system define the *propagation directions* of the wave fronts at each microphone. The propagation vectors for a four-element,

[2] *Bearing* is used to describe a reading from a compass, on a ship, for example.

linear array are illustrated in Figure 2.2. The directions of propagation have been defined in terms of the source location, $\vec{d}^{(s)}$, and microphone locations, $\vec{d}_1, \vec{d}_2, \ldots, \vec{d}_M$, by the following unit vectors:

$$\vec{\zeta}_m^{(s)} \equiv \frac{\vec{d}_m - \vec{d}^{(s)}}{\left|\vec{d}_m - \vec{d}^{(s)}\right|} \qquad \text{for } m = 1 \ldots M$$

2.6.2 Near Field versus Far Field

When all propagation directions are approximately equal, the source is said to be in the array's *far field* [59]. This occurs when the distance from the source to the array is much larger than the array's dimensions, or *aperture size*. Under far-field conditions, the waves impinging on the array "appear" planar since the curvature of the propagating spherical wave is small with respect to the aperture size. When the distance from the source to the array is comparable to the array's aperture size, the source is said to be in *near field*. When the source is in the near field, the curvature of the wave fronts is significant compared to the aperture size. Figure 2.2 is an example of near-field conditions.

The implication for source localization techniques is that far-field arrays cannot resolve the source's distance, or *range*; wave curvature parameterizes range. However, the inability of an array to resolve range is often exploitable. A far field array can be used to estimate the direction of propagation while ignoring the source's range. This often simplifies the computational complexity of the localization algorithm. Hence, range resolution can be traded for decreased computational cost.

2.6.3 Direction of Arrival (DOA)

The *direction of arrival* (DOA) of sound waves at an array is simply defined by the vector that points in the direction opposite the direction of propagation; it points towards the source. When the source is in the near field, there is a unique DOA at each microphone location, just as there is a unique direction of propagation. When the source is in the far field, the wave fronts at the array appear planar, and all DOAs at the array are the same. While a DOA can be defined at each microphone, or at any point on the array, it is most commonly defined relative to the array origin. The origin of the array can be arbitrarily chosen and is not necessarily the same as the global origin. Usually, it is chosen to be the center of the array. Once the array

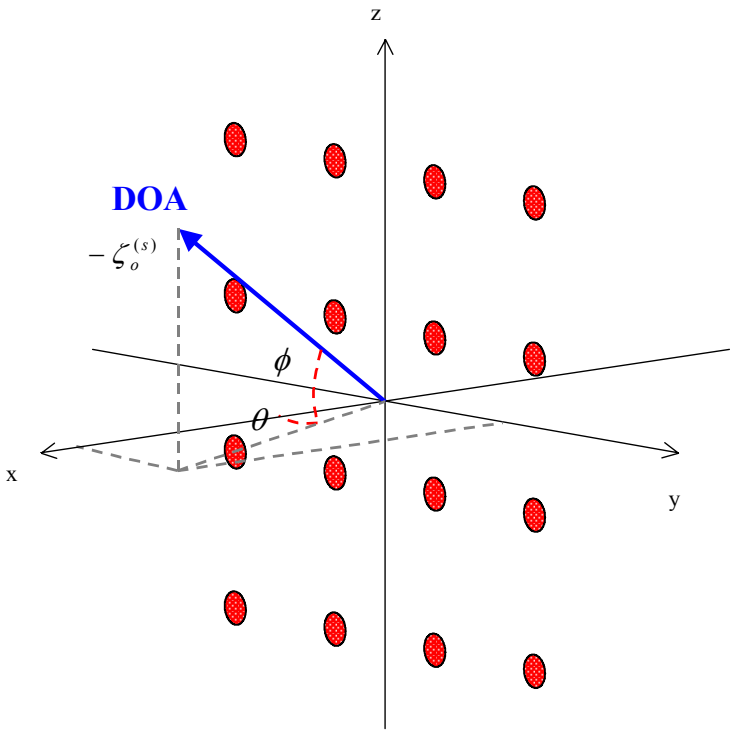

Figure 2.3 Definition of DOA in terms of azimuth, θ, elevation, ϕ, and propagation vector, $\vec{\zeta}_o^{(s)}$.

origin is established, the DOA to this origin depends only on the location of the source and is the same for both near and far fields (given that the array's global location and orientation remain fixed).

As is the propagation vector, the DOA vector is parallel to the bearing line, which passes through both the array origin and the source location. Denoting the propagation vector by $\vec{\zeta}_o^{(s)}$, the DOA vector is simply equal to $-\vec{\zeta}_o^{(s)}$. This vector can be defined in terms of the locations of the source and array origin in the global coordinate system:

$$\vec{\zeta}_o^{(s)} \equiv \frac{\vec{d}_o - \vec{d}^{(s)}}{\left|\vec{d}_o - \vec{d}^{(s)}\right|}$$

The direction of propagation, and hence the DOA, can be defined in terms of the array's local coordinate system. Figure 2.3 shows an example of how the local coordinate system might be defined for a planar microphone array. While the DOA vector, $-\vec{\zeta}_o^{(s)}$, has three elements, its orientation only depends

on two angles: *azimuth*, θ, and *elevation*, ϕ. Following a standard convention, azimuth has been defined as the angle between the projection of the DOA vector onto the local *xy*-plane and the local *x*-axis, and elevation has been defined as the angle between the DOA vector and the local *xy*-plane. The propagation vector can be defined in terms of these angles:

$$\vec{\zeta}_o^{(s)} \equiv \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix} \tag{2.9}$$

Hence, the DOA vector is also defined by the angles, θ and ϕ.

3 Microphone Array Data: Acquisition and Processing

This chapter describes the system and procedure for collection of the primary array data set, which was used in the source localization experiments presented later in this thesis. This data set was collected in a 7 by 4 by 3-meter conference room, and it is appropriately referred to as the *conference-room data set*. It was recorded using the *Brown Megamike II*. The Megamike microphone-array configuration, the collection procedure and the basic block-processing scheme that has been applied to the data are described. Some preliminary measurements of the conference-room data set are presented, including microphone signal-to-noise ratios (SNRs), room impulse responses and room reverberation times.

Figure 3.1 A picture of the Brown Megamike II, its host PC and robotic video camera with monitor.

3.1 The Brown Megamike II

The *Brown Megamike II* is a 16-channel microphone-array system that was designed to interface with a personal computer. The Megamike hardware was spawned from the *Huge Microphone Array* (HMA) project, which incorporates the same circuit board design into its 512-microphone system [86][84][85]. Real-time software has been developed for the Megamike giving it the capability to dynamically locate, track and focus the array on talkers in a conference room environment. It also has a link via its PC host to a robotic video camera, which is automatically steered, to view and follow talkers. These features were demonstrated at the *Third Biennial Roundtable on Microphone Array Technology* at Brown University [38] where it was used to automatically record the proceedings on videotape using its robotic camera while simultaneously acquiring the audio through its microphone array. A general overview of its real-time algorithms was presented there by this author, the software designer [29]. The talker-localization algorithms are based on the *linear intersection* method [7], which has been patented by Brown as part of its microphone array technology [6]. The real-time voice-capture algorithms are based on a patented *adaptive beamforming* technique [15]. The Megamike's real-time features have also been used in speech recognition experiments [55][56]. A photograph of the Megamike, its PC host and robotic camera is given in Figure 3.1.

In addition to the real-time Megamike software, the author also designed and implemented an elaborate multi-channel digital recording application. This recorder employs the Megamike as a server, which executes commands given by a *Windows95*-based application that runs on the host PC. The recorder's application window is shown in Figure 3.2. This application allows the user to give commands to the Megamike as well as view the recorded microphone signals. Commands are given using the menu on the top of the screen. The duration of the recordings is adjustable, from 1 to 15 seconds. The number of channels from which to record is also adjustable. In the example recording of Figure 3.2, 16 channels of microphone data, labeled **Mic 1** through **Mic 16**, were recorded for a 4-second duration. As the text bar just below the application menu shows, a **New Recording** was made of **80000 Samples** at a **20.00kHz Sampling Rate**.

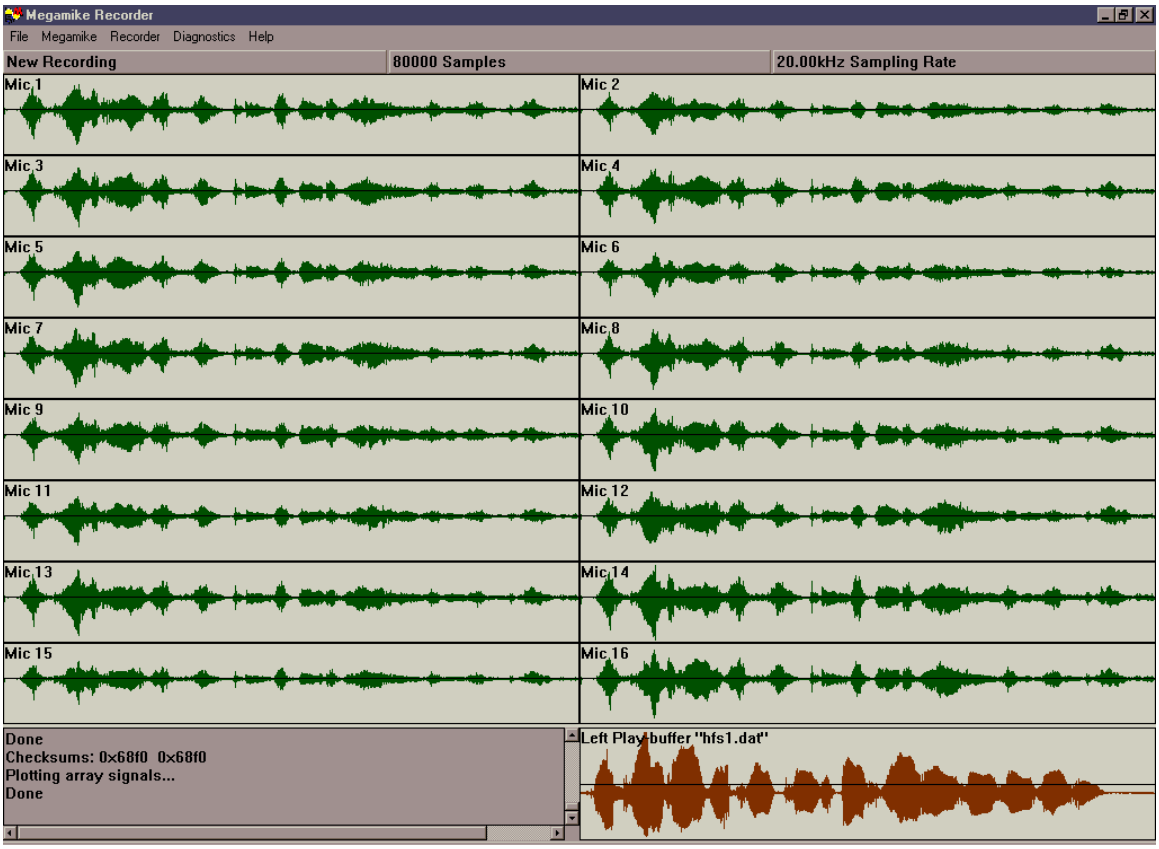

Figure 3.2 The *Megamike Recorder's* application window.

Once the microphone data is captured, the new recording can be stored in a file with a simple format known as a *Megamike Array File* (MAF). This file has an ASCII header that allows it to be browsed by any text editor on any computer platform. Within the MAF header are fields necessary to interpret the binary microphone data that follows. Next to each field-value is a text description of what the field means. For example, the fifth line of a MAF header looks like this:

```
0010 (16)..................................... Number_of_Channels
```

This line gives the number of channels in hexadecimal format followed by the base-10 equivalent in parentheses, a spacer consisting of periods, and the field description. A user who is not familiar with the MAF file format can simply view the header and use standard file I/O functions (such as *fscanf* in C) to load the array data into his program. While a simple program can be written in any language to load an

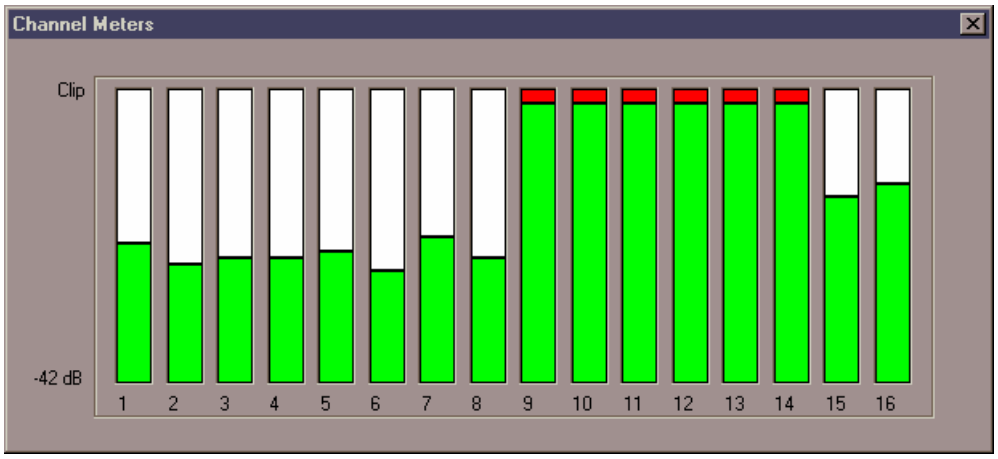

Figure 3.3 The Megamike's channel meters. Channels 9 through 14 are clipping.

MAF file, it is most commonly loaded into *MATLAB* using a script provided by the author. MAF files can also be loaded back into the recorder application for quick viewing or playback through the Meagmike's loudspeaker.

Also shown by Figure 3.2, is a window in the lower-right corner of the screen, which is named, **Left Play-buffer: "hfs1.dat"**. The plot in this window represents a signal that has been loaded from the file named "hfs1.dat" into the Megamike's *play-buffer*. By selecting *Record-Play* from the *Recorder* menu, the signal in the play-buffer is played through a loudspeaker (shown in lower left of Figure 3.1) as the signals from the microphone array are simultaneously recorded. There are two play buffers, right and left, allowing up to 2 signals to be loaded and played during a recording. Currently, only one play-buffer output is connected to an amplifier, which powers the loudspeaker. The volume of the amplifier is easily adjusted using the Megamike's *channels meters*, which are shown in Figure 3.3. The meters register the power, in dB, of the signal from each microphone channel. When the meters are activated, the play-buffer signal is played through the loudspeaker in a continuous loop allowing the volume to be adjusted while the user watches the meters. If any of the meters reaches it's maximum, which is marked by a red bar at the top of the meter, then the corresponding analog-to-digital converters (ADC) is clipping and the volume of the source must be reduced. These play-record and volume-adjustment features allow known source signals to be acquired using the full dynamic range of the array's A/Ds.

Other features of the *Megamike Recorder* include an adjustable countdown before recording, microphone scanning to listen for problematic microphone channels, an external trigger that is asserted at

Figure 3.4 A picture of the conference room.

the start of each recording, and PC-Megamike interface diagnostics. As shown in Figure 3.2, a status window, which is located in the lower-left corner of the screen, displays messages in response to the commands given to the Megamike.

3.2 The Conference-Room data set

Recordings were made in a small conference room at Brown University, which is shown in Figure 3.4, using a 15-element microphone array and the *Megamike Recorder*. The *Record-Play* feature was used to play prerecorded speech through a loudspeaker while simultaneously recording the signals from the array. The use of the loudspeaker was preferable to an actual talker since the loudspeaker could be precisely located and would be fixed over the duration of the recordings. The prerecorded speech was taken from the LEMS speech-recognizer database [80] [42] [52] [67] [43], which is composed of digital recordings from a close-talking microphone worn by talkers uttering alpha-digits. Three array recordings, ranging in duration

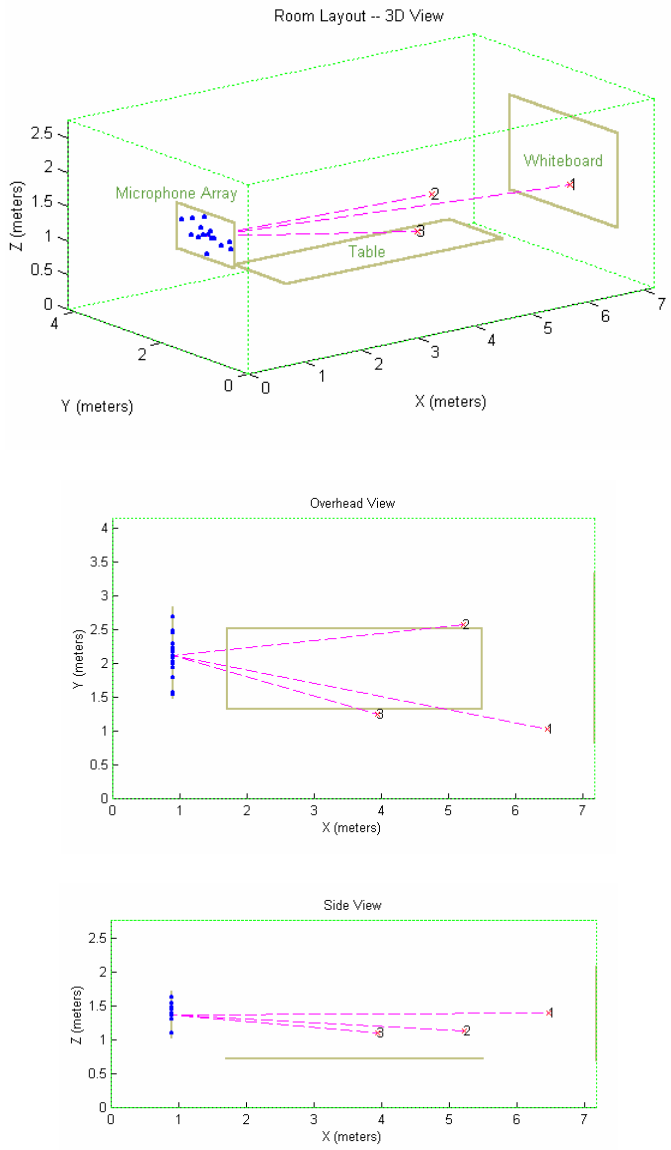

Figure 3.5 Conference room layout.

from 4 to 5 seconds, were made using the speech from this database. For each recording, there was a unique loudspeaker location, and the prerecorded speech was from a unique, male talker uttering a unique string of alpha-digits. The different source locations have been dubbed *source 1*, *source 2* and *source 3*. The conference room set-up is illustrated by Figure 3.5. Source 1 was the farthest from the array and was positioned to simulate a person who was standing in front of a white-board, as if she were giving a presentation. Source 2 and source 3 were positioned to simulate talkers sitting at a 3.8m by 1.2m conference table, which was located approximately in the center of the room.

The microphone array was composed of fifteen omni-directional electret condenser microphones [30], which were randomly distributed on a plane within a 1.34-by-0.67-meter rectangle. This is illustrated by Figure 3.6. The microphones were attached to a rectangular sheet of acoustic foam, which was supported by an aluminum frame. This frame was mounted on a tripod that was placed 0.9 meters away from the wall. The acoustic foam damped some of the reflections from this wall and isolated the microphones from vibrations traveling along the mountings. The long dimension of the array was parallel to the y-axis of the room coordinate system, as shown in Figure 3.5. The center of the

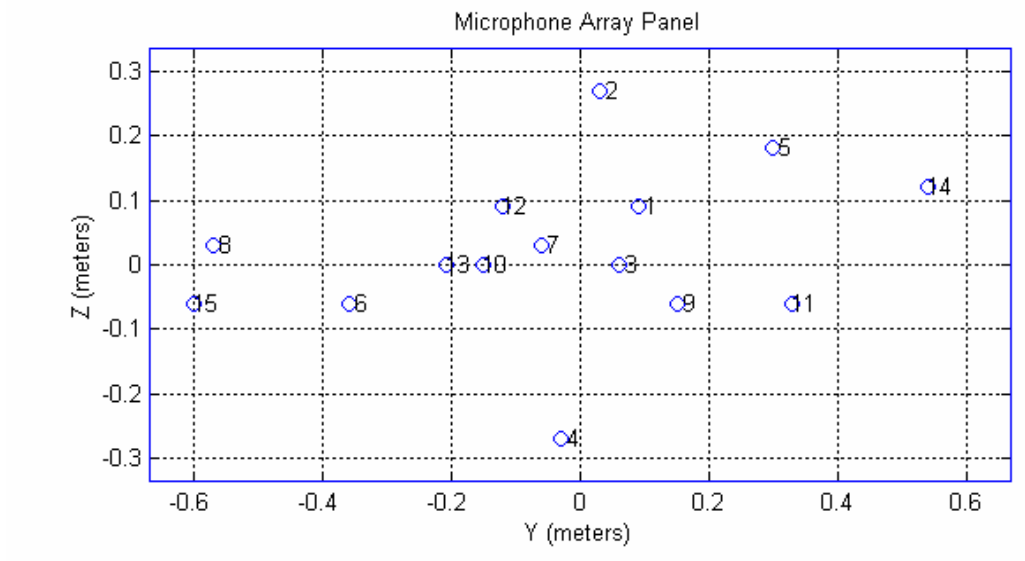

Figure 3.6 The planar, 15-element conference-room microphone array.

array, corresponding to the origin of the local coordinate system shown in Figure 3.6, was located at (0.90, 2.15, 1.37) meters in the room coordinate system.

During all three recordings, the loudspeaker was facing the microphones, and the volume had been adjusted to maximize the amplitude of the microphone signals without causing any clipping or noticeable distortion. This adjustment was made by adjusting the loudspeaker amplifier while watching the Megamike's channel meters. The channel meters clearly indicated when the A/Ds reached their limit. Using this feature in conjunction with simply listening for audible distortion from the loudspeaker ensured that the recordings were made with the highest possible signal-to-noise ratios (SNRs).

Immediately after each speech recording was made, another recording was made with the loudspeaker left undisturbed. During the second recording, a digitally generated, Gaussian noise signal was played through the loudspeaker. Again, the volume of the loudspeaker amplifier had been adjusted to maximize the amplitude of the microphone signals without causing any clipping or noticeable distortion. Finally, a single recording was made with no loudspeaker source. This source-free recording was used to measure the background noise present in the conference room.

The sixteenth channel of the Megamike was connected to a high-quality microphone that was mounted approximately 7 centimeters in front of, and facing, the loudspeaker. The signal from this microphone was recorded simultaneously with the array signals, and it provided a true reference of the

sound produced by the loudspeaker. Its gain was adjusted, independent of the array, using a preamplifier and in-line attenuators so that all sixteen channels were roughly at the same level (near maximum range). With its gain maximized, and its close proximity to the loudspeaker, this microphone essentially received a reverberation-free signal. Reference signals were included in all the array recordings made in the conference room.

While the sampling rate of the Megamike is 20kHz, the recordings were re-sampled at 16kHz using an appropriate multi-rate filtering scheme [73]. Since the prerecorded speech signals that were played through the loudspeaker were originally sampled at 16kHz, none of the speech-content was lost during the down-sampling procedure. Furthermore, the down sampling reduced the amount of digital data, which in turn reduced the amount of storage and computation needed to process the data. It was also more

Table 3.1 Source locations in the room coordinate system and DOAs relative to the array center.

	Location (Meters)			**DOA** (Degrees)	
	x	y	z	θ	ϕ
Source 1	6.47	1.03	1.40	-11.09	0.33
Source 2	5.22	2.57	1.13	5.95	-3.11
Source 3	3.94	1.25	1.10	-15.92	-4.89

convenient to have the same sampling rate for the prerecorded speech, the array recordings, and the sound card of the computer used to process the signals.

Table 3.1 lists the three source locations in relation to the room's global coordinate system. Also listed in this table are the DOAs at the array's origin, which correspond to the three source locations. These DOAs correspond to the azimuth angle, θ, and elevation angle, ϕ, as defined in Figure 2.3 for a planar array. According to this convention, sound waves traveling towards the array, on a path perpendicular to its axis, have azimuth and elevation that equal zero degrees.

3.3 Signal-to-Noise Power

The only obvious sources of noise in the conference room were the fans inside the *Megamike* and its host PC. In order to quantify this observation, the signal-to-noise ratios (SNRs) for each microphone channel were measured for the three Gaussian noise sources. The source-free recording was used to estimate the power of the background noise, which was presumed to be the same during all the recordings. Since the volumes of the source signals were set to overpower this noise, it is valid to assume that the background noise added a negligible component to the recordings of the Gaussian source. Hence, the powers of the microphone signals were used to compute the "signal" part of the SNRs.

The microphone signals from the *Megamike* are band-limited, sampled and quantized versions of the true analog signals, $x_1(t)\ldots x_M(t)$. Let these discrete-time microphone signals be denoted $x_1[n]\ldots x_M[n]$. Denoting the m-th microphone signal from the l-th Gaussian noise recording by $x_m^{(s_l)}[n]$ and from the source-free recording by $x_m^{(v)}[n]$, the SNR of the l-th source at the m-th microphone was computed using the following formula:

$$SNR_m^{(s_l)} = 10\log_{10}\left\{\frac{\sum_{n=1}^{N}\left(x_m^{(s_l)}[n]\right)^2}{\sum_{n=1}^{N}\left(x_m^{(v)}[n]\right)^2}\right\} \tag{3.1}$$

where N is the length (number of samples) of both recordings. A zero-phase, FIR high-pass filter was applied to each microphone signal prior to this computation, which removed the DC component and its skewing effects on the SNR.

Equation 3.1 can be expressed in terms of the source signal, $s_l[n]$, and the noise signal, $v_m[n]$, using the discrete-time form of Equation 2.6. With this room impulse response model, the m-th microphone signal from the N-point recording of source l can be expressed as:

$$x_m^{(s_l)}[n] = s_l[n] * \widetilde{h}_m[\vec{d}^{\,(s_l)}, n] + v_m[n] \qquad \text{for } n = 1\ldots N$$

The impulse response $\tilde{h}_m[\vec{d}^{(s_l)}, n]$ represents the acoustic system from source l, located at $\vec{d}^{(s_l)}$, to the m-th microphone in the array. In the conference-room data set, there are 15 microphones in the array and three unique source locations. Hence, $m = 1 \ldots 15$ and $l = 1 \ldots 3$. During the recording of the background noise, there was no source signal, resulting in $s[n] = 0$ and $x_m[n] = v_m[n]$. During the recording of the Gaussian source, the noise $v_m[n]$ was negligible. Using this model, the SNR, as computed by Equation 3.1, can be expressed as follows:

$$SNR_m^{(s_l)} = 10 \log_{10} \left\{ \frac{\sum_{n=1}^{N} \left(s_l[n] * \tilde{h}_m[\vec{d}^{(s_l)}, n] \right)^2}{\sum_{n=1}^{N} \left(v_m[n] \right)^2} \right\} \tag{3.2}$$

The "signal" power in Equation 3.2 (numerator) is the sum of all the power generated by the source, including the reverberation that is implicit in the convolution with the room impulse response. While it

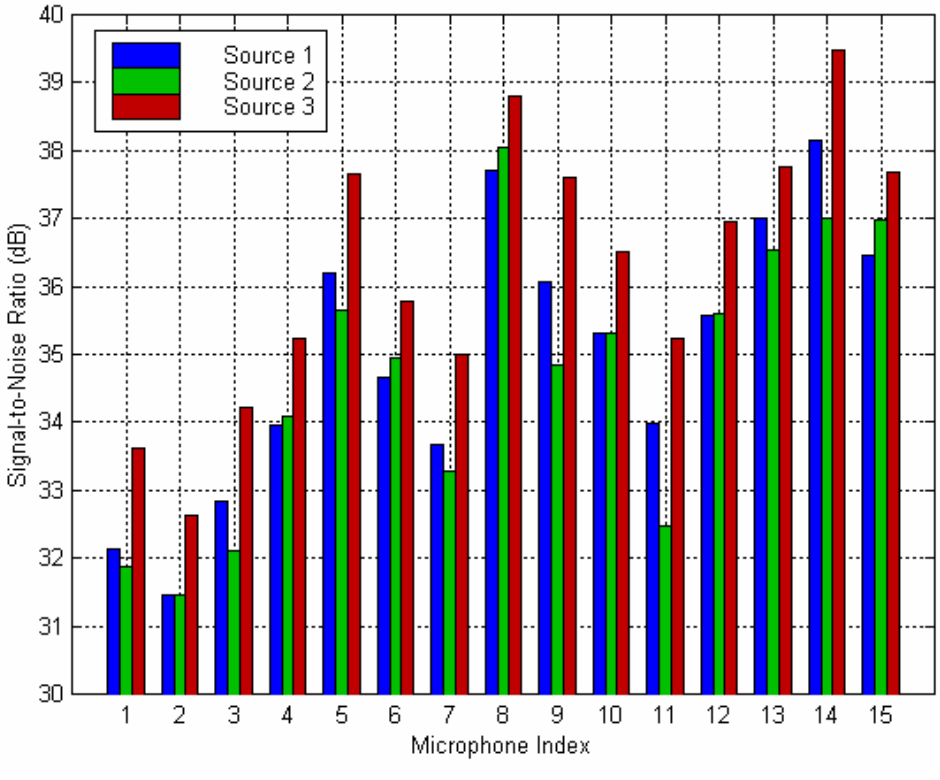

Figure 3.7 Estimated SNRs of all 15 microphone channels for each Gaussian source.

may be more accurate to use only the direct-path component of the source to compute the "signal" power, Equation 3.2 is effective in expressing the power of the source in relation to the power of the background noise. Furthermore, there is no simple way to measure the direct-path sound exclusively.

The estimated SNRs of all 15 microphone-channels and for each source location are plotted in Figure 3.7. Notice that all SNRs are generally very high (above 31 dB). As expected, source 3 has the highest SNRs, since its location was the closest to the microphone array. As the bar graph shows, there is some variation among channels. It is likely that this effect is due to variations in the system's hardware, as well as differences in the reverberation patterns for each microphone and source. Nonetheless, all microphones signals in the conference room dataset have negligible contributions from the background noise. Any effects that significantly distort the microphone signals must come from the acoustic path from source to receiver, which makes this dataset ideal for studying the sole effects of room reverberation on location estimation.

3.4 Processing the Microphone Signals in Blocks

These discrete-time microphone signals have been denoted $x_1[n] \ldots x_M[n]$. With most techniques, source-localization begins by segmenting these signals into blocks and applying the discrete Fourier transform to each block. Each block of data is windowed with a tapered window before the DFT is applied to improve the spectral representation of the signal. Consecutive DFT blocks usually overlap in the time-domain to allow the data that align with the edges of one block, and are suppressed by the tapered window, to be centered in the next, giving all data an equal weight in the analysis. Source-localization algorithms operate on the DFTs of each data block. Since each block advances in time, the algorithms are able to track moving and multiple talkers. The rate at which location estimates are produced depends on the advance of the data blocks, and the latency of each estimate depends on the block-size. Therefore, the responsiveness of the estimator to dynamic conditions is also related to these block parameters. Estimators become more responsive with a decrease in block size and an increase in the block advance rate. However, accuracy tends to increases with block-size, as does computation. Hence, there are always tradeoffs among computational demands, accuracy and responsiveness.

The discrete-time microphone signals, $x_1[n]\ldots x_M[n]$, are segmented into blocks of length L samples, and a window of length L is applied to each block:

$$x_{m,b}[n] = w[n]x[bA+n] \qquad \text{for } m = 1\ldots M, n = 0\ldots L-1$$

$x_{m,b}[n]$ is the windowed data of the m-th microphone channel and the b-th block. A is a constant, positive integer that defines the block advance. The blocks overlap when $A<L$, and A is typically set to $L/2$. A common choice for $w[n]$ is a Hanning window, which has a DFT with a mainlobe twice as wide and 25dB lower sidelobes than a rectangular window. The use of any tapered window, such as the Hanning window, eliminates many of the effects caused by the discontinuities at the ends of the window and generally is considered to improve spectral estimation, although the increased mainlobe width is a penalty.

A K-point discrete Fourier transform (DFT) is applied to each windowed block of microphone data. The DFT of the m-th microphone signal and the b-th block is defined as follows:

$$X_{m,b}[k] \equiv \sum_{n=0}^{L-1} x_{m,b}[n]e^{-jk\frac{2\pi}{K}n} \qquad \text{for } 0 \le k \le K\text{-}1, m = 1\ldots M$$

Note that the DFT length is K, and $K \ge L$. K may be greater than L if the data block needs to be zero-padded before applying a radix-2 fast Fourier transform algorithm (FFT), for example. This zero padding may also be necessary to account for the circular shifting properties of the DFTs [73]. Since the DFTs are updated for each data block, and successive data blocks advance in time, $X_{m,b}[k]$ is a time-dependent spectral estimate of the m-th microphone signal, with assumed stationarity over each interval indexed by b. By operating on $X_{m,b}[k]$, an algorithm can produce a new location estimate with each data block allowing the estimates to reflect the motion of the active talker or a switch to a different talker.

3.5 Speech/Silence Detection: Block SNR and the SNR Mask

As shown by the SNR measurements in Section 3.3, the conference-room data set was recorded in a low-noise environment. However, unlike the Gaussian sources, which produce a high SNR for the duration of each array recording, the SNR of the speech recordings fluctuate considerably. When speech recordings

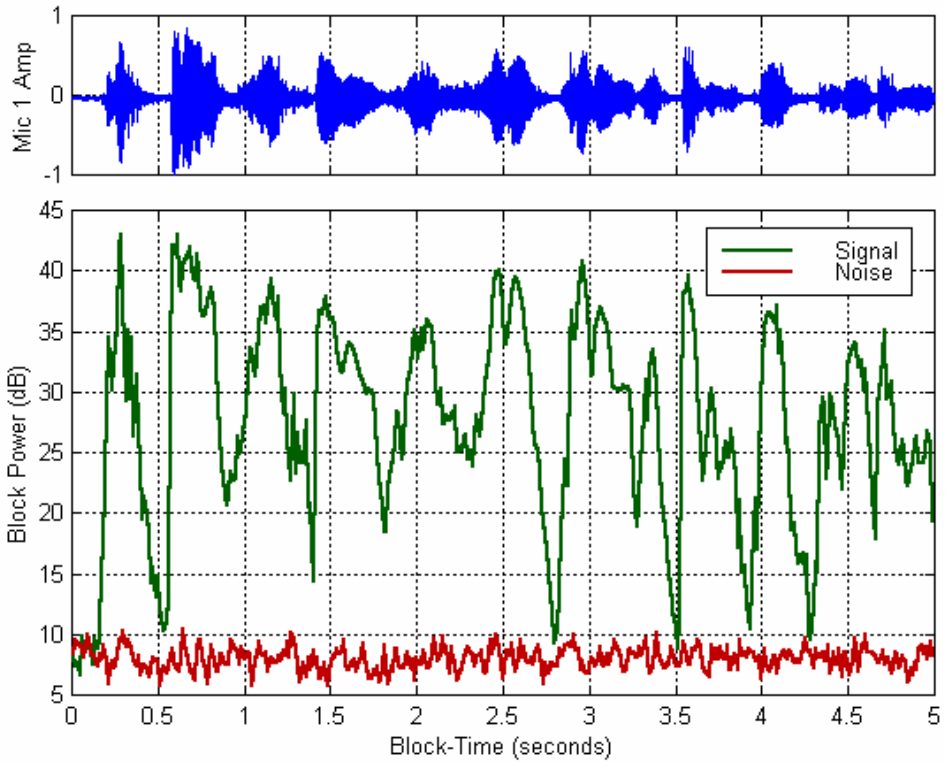

Figure 3.8 Block powers of the speech signal and background noise at microphone 1 (bottom). Top shows the amplitude of the speech signal at microphone 1, which corresponds to the power below.

are segmented according to the block-processing scheme described in Section 3.4, the background noise becomes significant in the blocks where there are pauses or very low-power speech. It is advantageous to discard, or *masked out*, these blocks since they typically result in poor location estimates.

In order to derive such SNR *masks*, the SNRs of the speech-array recordings were computed on a block-by-block basis. The blocks were 25 milliseconds long, and the advance was 12.5 milliseconds. Figure 3.8 shows the block power of speech source 2 and microphone 1 as well as the block power of the background noise recording. Notice how the speech power fluctuates from block to block. At certain points, the block power falls to a level near that of the background.

Although the speech power varies from block to block, Figure 3.8 shows that the noise power varies very little. With the noise power essentially constant across all the blocks, the plot of the speech power in microphone 1 is nearly equivalent to the SNR, within a constant offset in dB. Therefore, the

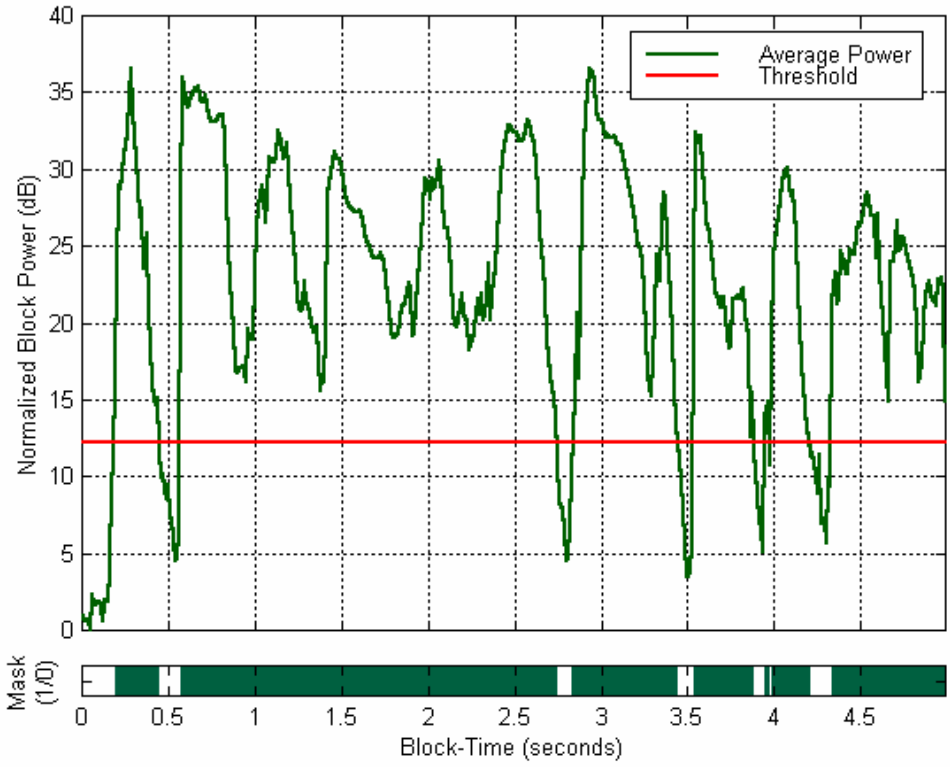

Figure 3.9 The top plot is the block power averaged over microphone during the speech recording of source 2. The bottom plot shows the SNR mask derived from the average block power and the SNR threshold, which is marked by the horizontal line in the top plot.

microphone power alone is sufficient for defining the SNR mask. By setting a threshold on the block power of the microphone signal, the low-SNR blocks can be masked. This same approach can be applied to the average block power of all the microphones in the array to ensure that all the array signals have sufficient SNR. Figure 3.9 shows an SNR mask that was derived using a threshold of 12 dB. The top plot shows the average block power across all the microphone signals during the recording of speech source 2. This plot was normalized (offset in dB) so that silence corresponds to 0 dB, and the SNR threshold, which has been set to 0.33 of the maximum block power in this recording, was equal to 12 dB. Any block with average power below 12 dB was masked, and masked blocks are indicated by the white portions of the lower plot. This procedure will be used in the experiments that follow later in this thesis.

3.6 Measuring Room Impulse Responses

Perceptually, the conference room seemed to be an acoustically "dead" room with negligible reverberation. This observation was quantified using the Gaussian noise recordings, including the signals from the reference microphone mounted in front of the loudspeaker, to estimate the room impulse responses and reverberation times. The Gaussian test signals were similar to those used in other impulse response measuring experiments [77][95].

Recall the room impulse response model of Equation 2.6:

$$x_m(t) = s(t) * \widetilde{h}_m(\vec{d}^{(s)}, t) + v_m(t)$$

$\widetilde{h}_m(\vec{d}^{(s)}, t)$ is the source-output to microphone-output impulse response, which is the convolution of the room impulse response with the microphone-channel impulse response. $\widetilde{h}_m(\vec{d}^{(s)}, t)$ can be measured when $s(t)$ is known and is sufficiently white over the duration of an array recording. The reference microphone provided the known source signal, $s(t)$, while its Gaussian properties yielded sufficient wide-band power over the duration of the recording. Since there were no anechoic measurements of the microphone system, there was no way to recover the actual room impulse response from $\widetilde{h}_m(\vec{d}^{(s)}, t)$. However, if the microphone system is linear and behaves similar to a bandpass filter, then $\widetilde{h}_m(\vec{d}^{(s)}, t)$ is a band-limited approximation to the true room impulse response, $h_m(\vec{d}^{(s)}, t)$.

3.6.1 Least-Squares Fit to Input-Output Data

A room impulse response estimation procedure was developed based on the discrete-time form of Equation 2.6:

$$x_m[n] = s[n] * \widetilde{h}_m[\vec{d}^{(s)}, n] + v_m[n] \tag{3.3}$$

As discussed in Chapter 2, the impulse responses are highly dependent on $\vec{d}_m$ and $\vec{d}^{(s)}$, as well as the environmental factors. However, it will be assumed that the environment was constant over the duration of each Gaussian array recording, which is 5 seconds long, and that the room impulse responses remained

fixed as well for this duration. Under these conditions, the explicit dependence on the source location, $\vec{d}^{(s)}$, can be dropped, and Equation 3.3 can be re-written as:

$$x_m[n] = s[n] * \tilde{h}_m[n] + v_m[n]$$

Considering only one microphone signal at a time, the microphone index, m, can also be dropped from the notation:

$$x[n] = s[n] * \tilde{h}[n] + v[n] \tag{3.4}$$

Room impulse responses have infinite duration in nature. However, their power becomes negligible in comparison to the power of the direct-path sound in finite time (See Section 3.6.3), and they can be modeled accurately by finite-duration sequences. If the duration of the truncated, discrete-time room impulse response is I points, then the convolution, represented by "*" in Equation 3.4 can be expanded as follows:

$$x[n] = \sum_{i=0}^{I-1} \tilde{h}[i] s[n-i] * + v[n] \tag{3.5}$$

Hence, the goal is to estimate $\tilde{h}[n]$ for $0 \le n \le I-1$ using the data from the N-point Gaussian noise recordings where $N>>I$. This was done using the block-processing scheme described in Section 3.4 in conjunction with a DFT-based method that minimized the sum of squares error between the DFT of the modeled microphone signals and the DFT of the observed microphone signals.

According to the block-processing scheme described in Section 3.4, $X_{m,b}[k]$ denotes the DFT of the m-th microphone signal over the windowed block indexed by b, where the duration of each block is L points. For a single microphone, this quantity is more simply denoted by $X_b[k]$. Let N_b denote the total number of blocks in the N-point recording. Applying this block-processing scheme to the model of Equation 3.4, the DFT of block b can be expressed as:

$$X_b[k] = S_b[k]\tilde{H}[k] + \mathrm{V_b}[k] \qquad \text{for } 0 \le k \le K\text{-}1,\ 0 \le \mathrm{b} \le N_b \tag{3.6}$$

$S_b[k]$ and $V_b[k]$ are the block DFTs of the source and noise signals, respectively. Note that the DFT of the room impulse response, $\widetilde{H}[k]$, remains fixed over all blocks since it was assumed fixed over the duration of the recording. Hence, each block of the microphone signal is the convolution of the impulse response with a unique block of the source signal, plus uncorrelated noise. An estimate of $\widetilde{H}[k]$ can be obtained by "fitting" the input-output data from all blocks. To do this, it is more convenient to express Equation 3.6 in vector notation as follows:

$$\underbrace{\mathbf{X}[k]}_{\begin{bmatrix} X_0[k] \\ X_1[k] \\ \vdots \\ X_{N_b-1}[k] \end{bmatrix}} = \underbrace{\mathbf{S}[k]}_{\begin{bmatrix} S_0[k] \\ S_1[k] \\ \vdots \\ S_{N_b-1}[k] \end{bmatrix}} \widetilde{H}[k] + \underbrace{\mathbf{V}[k]}_{\begin{bmatrix} V_0[k] \\ V_1[k] \\ \vdots \\ V_{N_b-1}[k] \end{bmatrix}} \tag{3.7}$$

The data is "fitted" by minimizing the (square root of the) error sum of squares for each value of k, which has been defined as follows:

$$E[k] \equiv \left\| \mathbf{X}[k] - \hat{\widetilde{H}}[k]\mathbf{S}[k] \right\|$$

where $\|\cdot\|$ denotes the vector norm. The DFT of the impulse response estimate, $\hat{\widetilde{H}}[k]$, that minimizes this error is given by [49][89]:

$$\hat{\widetilde{H}}[k] = \frac{\mathbf{S}'[k]\mathbf{X}[k]}{\mathbf{S}'[k]\mathbf{S}[k]} \tag{3.8}$$

$\mathbf{S}'[k]$ denotes the conjugate transpose of $\mathbf{S}[k]$. The quantity $\mathbf{S}'[k]\mathbf{X}[k]$ is the time-averaged *cross-spectral density*, or *cross spectrum* of $\mathbf{S}'[k]$ and $\mathbf{X}[k]$, and $\mathbf{S}'[k]\mathbf{S}[k]$ the time-averaged *power spectral density*, or *power spectrum* of $\mathbf{S}[k]$. Cross-spectra and power-spectra are quantities that arise frequently in statistical signal processing [48], and this is a common means of estimating a system's response based on measured input-output data.

Equation 3.8 uses time-segmented blocks of a linear system's input and output sequences to estimate the DFT of an I-point impulse response. Since the convolution of two, equal-length sequences (input and output) produces a sequence that is twice as long, the block size must be at least twice the duration of the impulse response. This places a constraint on the block size parameter, L, which must be at

least twice the length of the impulse response. Although the duration of $\tilde{h}[n]$ was not known ahead of time, by trial and error, these parameters were eventually set such that $L \geq 2I$. The power in $\tilde{h}[n]$ for $n > I$ was considered negligible.

The time-domain impulse responses can be obtained by applying the inverse DFT to Equation 3.8:

$$\hat{\tilde{h}}[n] = \frac{1}{K}\sum_{k=0}^{K-1} \hat{\tilde{H}}[k] e^{-jnk\frac{2\pi}{K}} \qquad \text{for } 0 \leq n \leq I \tag{3.9}$$

Notice the range for n in Equation 3.9. Recall that $\tilde{h}[n]$ is an I-point sequence, and the inverse DFT, in general, yields K points. The DFT size, K, and the block size, L, have been chosen so that $2I \leq L \leq K$. Hence, only the first I points of $\hat{\tilde{h}}[n]$ in Equation 3.9 need be saved since $\tilde{h}[n]$ is assumed to be zero for $n > I$. Furthermore, because of the circular shifting property of DFTs [73], the last $K/2$ points of the inverse DFT correspond to those for $n < 0$, and these points must be discarded. By choosing K as it has been, the DFT size was large enough to discard these points and still having more than I points left in the first half of the inverse DFT.

3.6.2 Application to the Conference-Room Data Set

Some of the room impulse responses were estimated from the conference-room data set using Equations 3.6 through 3.9. A block size of 16000 points was used, which resulted in 8000-point impulse responses, or 500 milliseconds at a 16kHz-sampling rate. The block advance parameter was set to 500, and a rectangular window was applied to each data block. With these parameters, the 5-second Gaussian noise recordings produced 289 blocks. That is, $N_b = 289$.

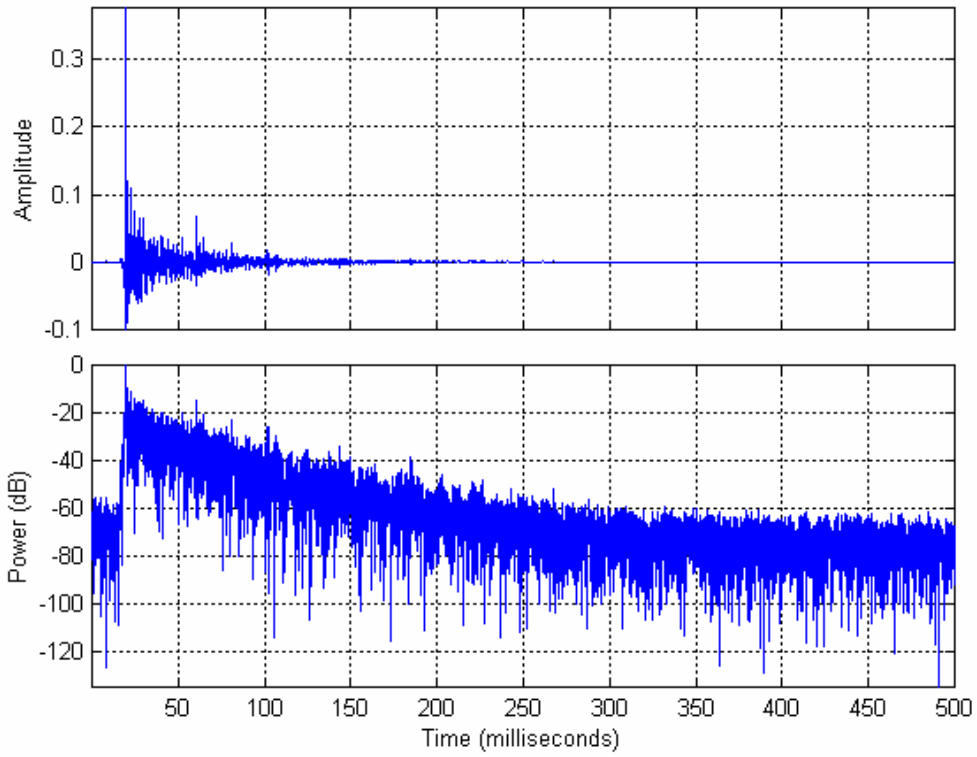

Figure 3.10 Room impulse response of microphone 1 and source 1. The top plot shows the amplitude of the impulse response, and the bottom plot shows its power in dB.

Figure 3.10 shows the room impulse response of microphone 1 and Gaussian source 1. The top plot shows the amplitude of the impulse response, and the bottom shows its power in dB. The maximum power, which occurs near 20 milliseconds, corresponds to the direct-path sound waves and has been normalized to 0 dB. Notice that the power falls off almost linearly from the maximum, to where it "flattens" out, just below -60 dB. This "flattening" indicates that the noise floor is just about 65 dB down from the power of the direct-path sound. Hence, the block size was large enough for estimating the impulse response within the limits of the noise in the data. Perhaps, with a longer recording and more data blocks, the noise floor could be lowered. However, this data is sufficient for measuring the *reverberation time*, T_{60}, which corresponds to the point where the impulse response power falls below -60dB. From the lower plot of Figure 3.10, the reverberation time for microphone 1 and source 1 appears to be about 200 milliseconds. Reverberation curves are given, for all three sources, in Section 3.6.3.

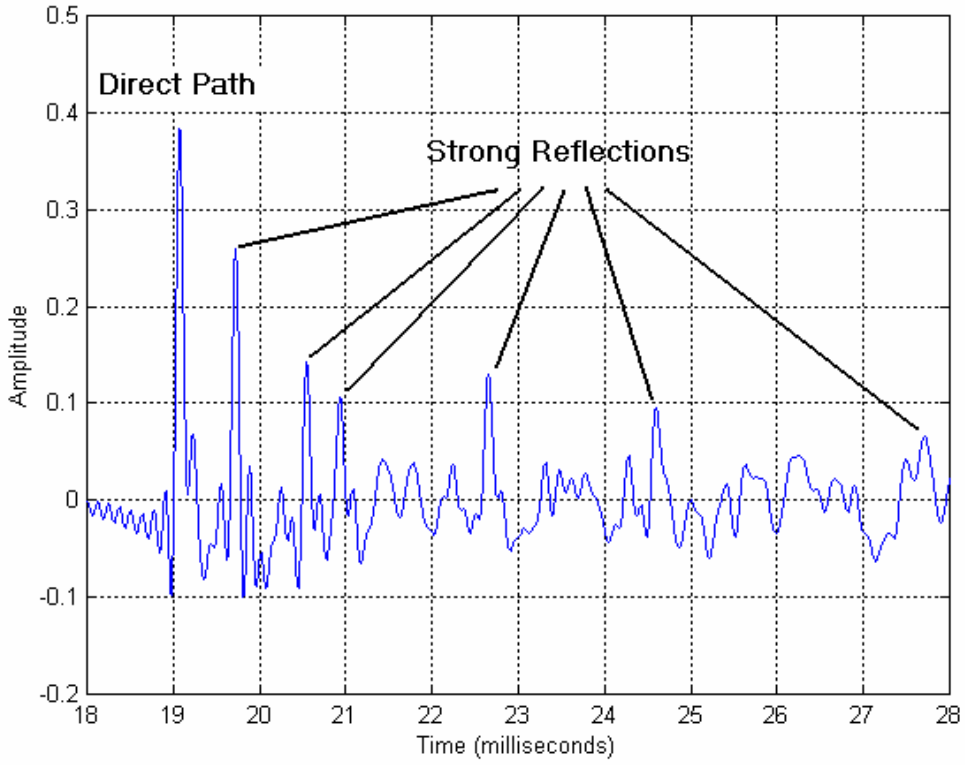

Figure 3.11 A close-up of a 10-millisecond segment of the room impulse response from source 1 to microphone. The direct-path component and some strong reflected components are highlighted.

Figure 3.11 is a close-up view of the impulse response of Figure 3.10. It has been interpolated to show more detail. The direct-path component and some of the strong reflected components are highlighted in this plot. The peaks corresponding to the reflected sound waves are comparable in size to the direct-path peak. These peaks, which occur within 20 milliseconds of the direct-path, are responsible for many of the erroneous locations produced by short-time estimators, which operate on blocks as small as 25 milliseconds. For example, source localization techniques that employ the *generalized cross-correlation* (GCC) function are severely impacted by reverberation, as shown in [19]. The large secondary peaks in the room impulse responses are directly correlated with the false peaks in the GCC function. These effects, along with results from GCC-based experiments with the conference-room data set, will be discussed in more detail later in this thesis.

3.6.3 The Conference Room Reverberation Time

The impulse responses from each Gaussian source to microphone 1 were computed. The smoothed powers of these impulse responses are plotted in Figure 3.12. The direct-path peaks in all three plots have been aligned and shifted to zero on the time axis. The amplitudes were normalized so that the maximum power of each plot equals zero dB. With these adjustments, the reverberation times can be easily determined. Notice that the decay of each power curve is approximately the same, which implies that the room reverberation time is the same for any source and microphone locations. The point where these curves fall below -60dB, which corresponds to T_{60}, is about 200 milliseconds. This qualifies as a mildly reverberant room. However, the near-end peaks in the impulse response combined with a 200-millisecond reverberation time have a surprising effect on source localization. This will be examined further later in this thesis.

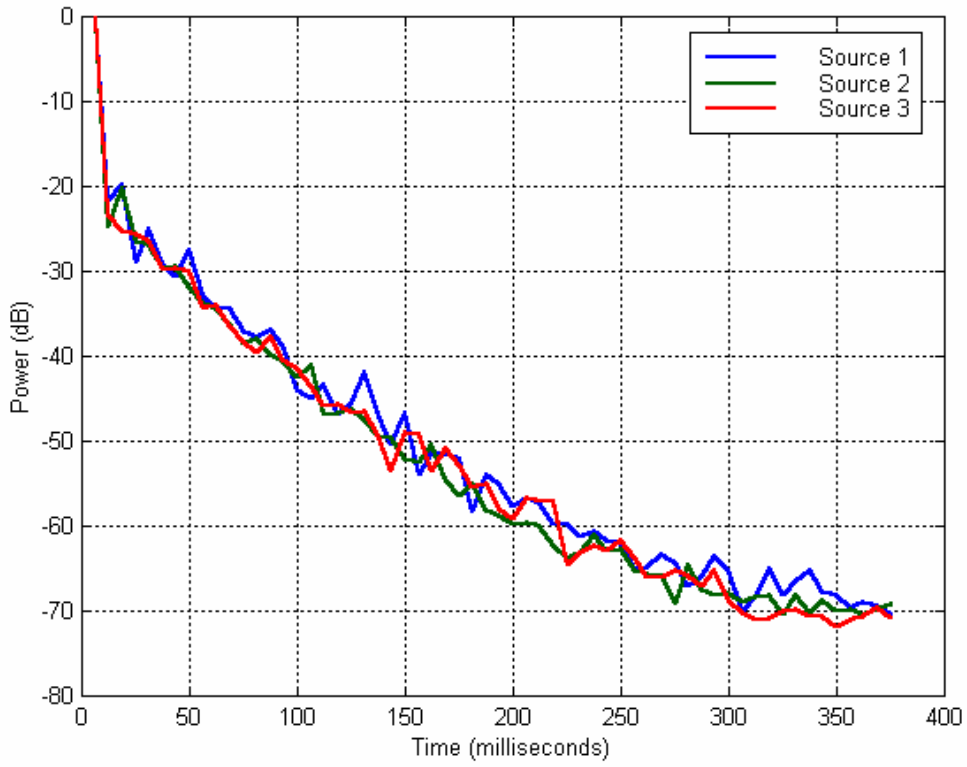

Figure 3.12 The smoothed powers of the impulse responses from sources 1, 2 and 3 to microphone 1. The reverberation time, T_{60}, is 200 milliseconds.

4 Generalized Cross Correlation (GCC)

Generalized cross correlation (GCC) has been used successfully to determine the time difference of arrival (TDOA) of propagating waves between two microphones. TDOA estimates from multiple microphone pairs can be used to parameterize the location of a sound source. An example of this is depicted in Figure 4.1. The distance from the source to each microphone has been denoted by $r_m^{(s)}$ for $m = 1 \ldots 4$. As

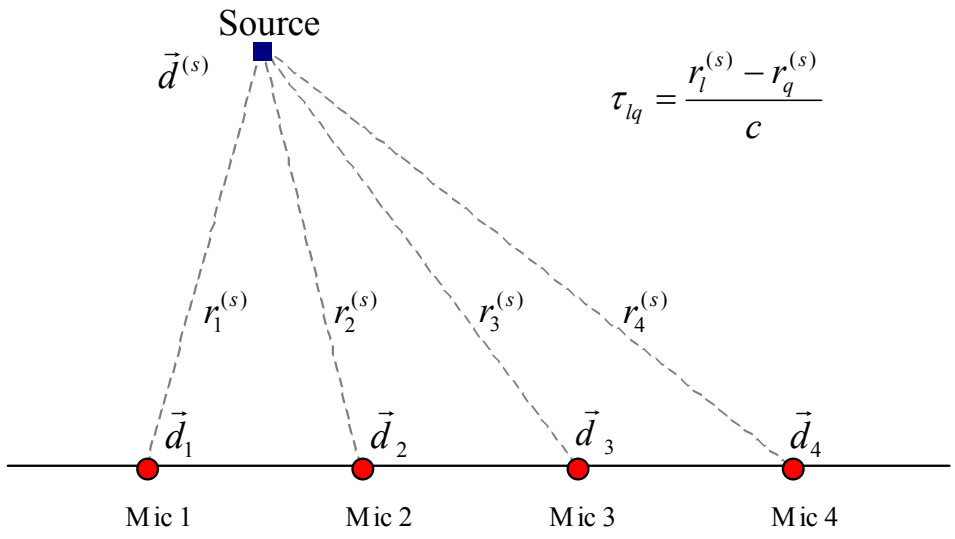

Figure 4.1 An example of how TDOAs parameterize source location.

defined in Section 2.6, under simple acoustic conditions, the relationship between these distances and the propagation delays was given as:

$$\tau_m = \frac{r_m^{(s)}}{c} \tag{4.1}$$

where c equals the speed of sound. The TDOA for the pair consisting of microphones l and q has been denoted by τ_{lq}, and it is defined simply as the difference between the propagation delays as follows:

$$\tau_{lq} \equiv \tau_l - \tau_q \tag{4.2}$$

By substituting Equation 4.1 into Equation 4.2, the TDOAs can be expressed, as shown in Figure 4.1, in terms of the distances from the source to each microphone as follows:

$$\tau_{lq} = \frac{r_l^{(s)} - r_q^{(s)}}{c}$$

This equation can be re-written to express the distances in terms of the TDOAs:

$$r_l^{(s)} - r_q^{(s)} = c\,\tau_{lq}$$

Hence, the TDOAs parameterize the source's location, and by various techniques ([8][93] are two examples), the location can be derived from a multitude of TDOA estimates.

The performance of GCC-based localization techniques will be studied in the following chapters using the conference-room data set. Since GCC-based methods are so widely used in speech-array applications, it is important to quantify its performance in what might be considered high-SNR and mildly reverberant conditions. This chapter introduces GCC and the implementation of it in the following experiments.

4.1 GCC Defined

Recall the microphone signal model of Equation 2.8:

$$x_m(t) = \frac{1}{r_m} s(t - \tau_m) * \gamma_m(\vec{d}^{(s)}, t) + \breve{v}_m(t)$$

For a pair of microphones, $m = 1,2$, and the TDOA from microphone 1 to microphone 2 is defined as follows:

$$\tau_{12} \equiv \tau_2 - \tau_1$$

This definition implies that $\tau_2 = \tau_1 + \tau_{12}$, and by substituting this into the signal equation for microphone 2, the TDOA can be explicitly expressed in the microphone signal equations as follows:

$$\begin{aligned} x_1(t) &= \frac{1}{r_1} s(t - \tau_1) * \gamma_1(\vec{d}^{(s)}, t) + \breve{v}_1(t) \\ x_2(t) &= \frac{1}{r_2} s(t - \tau_1 - \tau_{12}) * \gamma_2(\vec{d}^{(s)}, t) + \breve{v}_2(t) \end{aligned} \qquad (4.3)\text{-}(4.4)$$

If the channel impulse responses, $\gamma_1(\vec{d}^{(s)},t)$ and $\gamma_2(\vec{d}^{(s)},t)$, are similar for both microphone systems, then Equations 4.3 and 4.4 show that a scaled version of $s(t-\tau_1)$ is present in the signal from microphone 1 and a time-shifted (and scaled) version of $s(t-\tau_1)$ is present in the signal from microphone 2. The *cross correlation* of the two signals should show a peak at the time lag where the shifted versions of $s(t)$ align, corresponding to the TDOA, τ_{12}. The cross correlation of signals $x_1(t)$ and $x_2(t)$ is defined as [48][50]:

$$c_{12}(\tau) \equiv \int_{-\infty}^{+\infty} x_1(t)x_2(t+\tau)dt \tag{4.5}$$

The Fourier transform of the cross correlation function is known as the *cross spectral density*, or *cross spectrum*, and is given by the following:

$$C_{12}(\omega) = \int_{-\infty}^{\infty} c_{12}(\tau)e^{-j\omega\tau}d\tau \tag{4.6}$$

By substituting Equation 4.5 into Equation 4.6, and applying the convolution property of Fourier transforms [73], the cross-spectral density can be expressed in terms of the Fourier transforms of $x_1(t)$ and $x_2(t)$:

$$C_{12}(\omega) = X_1(\omega)X_2'(\omega) \tag{4.7}$$

$X_1(\omega)$ is the Fourier transform of $x_1(t)$ and $X_2'(\omega)$ is the complex conjugate of the Fourier transform of $x_2(t)$. The inverse Fourier transform of Equation 4.7 gives the cross correlation function in terms $X_1(\omega)$ and $X_2'(\omega)$:

$$c_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X_1(\omega)X_2'(\omega)e^{j\omega\tau}d\omega \tag{4.8}$$

The *generalized cross correlation (GCC) function* [64], $R_{12}(\tau)$, is the cross correlation of two filtered versions of $x_1(t)$ and $x_2(t)$. With the Fourier transforms of these filters denoted by $G_1(\omega)$ and $G_2(\omega)$, the GCC function can be expressed as:

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(G_1(\omega)X_1(\omega)\right)\left(G_2(\omega)X_2(\omega)\right)' e^{j\omega\tau} d\omega$$

Rearranging the order of the signals and filters gives:

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} G_1(\omega)G_2'(\omega)X_1(\omega)X_2'(\omega)e^{j\omega\tau} d\omega$$

By defining the frequency dependent weighting function, $\Psi_{12}(\omega) \equiv G_1(\omega)G_2'(\omega)$, the GCC function can be defined as:

$$R_{12}(\tau) \equiv \frac{1}{2\pi}\int_{-\infty}^{\infty} \Psi_{12}(\omega)X_1(\omega)X_2'(\omega)e^{j\omega\tau} d\omega \tag{4.9}$$

Ideally, with an appropriate weighting function, $R_{12}(\tau)$ should exhibit a peak, over a restricted range[3] of τ (i.e. $\tau \in D$), which corresponds to the TDOA between microphone 1 and 2. The TDOA estimate is the time lag that maximizes $R_{12}(\tau)$:

$$\hat{\tau}_{12} = \arg\max_{\tau \in D} R_{12}(\tau)$$

Note that finding $\hat{\tau}_{12}$ requires a simple, one-dimensional search. In general, Equation 4.9 has multiple local maxima. The amplitudes and corresponding time lags of these maxima depend on a number of factors. These factors include the separation distance of the microphones, the nature of the source signal and noise signals, and the choice of the weighting function $\Psi_{12}(\omega)$.

4.1.1 Maximum Likelihood (ML) Weighting Function

When there is no multipath propagation (no reverberation), and the source and noise terms are uncorrelated Gaussian signals, the *maximum likelihood* (ML) weighting yields an estimator that is asymptotically unbiased and efficient. The ML weighting function is given in terms of the power spectral densities of the source signal, $S(\omega)$, and noise signals, $V_1(\omega)$ and $V_2(\omega)$, [64]:

---

[3] This finite range is determined by the distance between the microphones divided by the speed of sound.

$$\Psi_{12}^{ML}(\omega) \equiv \frac{|S(\omega)|}{|V_1(\omega)||V_2(\omega)|}\left\{1+\frac{|S(\omega)|}{|V_1(\omega)|}+\frac{|S(\omega)|}{|V_2(\omega)|}\right\}^{-1}$$

The idealized conditions for which this weighting is optimal are rarely encountered in practice. It has been shown that even mild reverberation greatly degrades performance of the ML estimator [19]. The coherence function, on which the ML estimator is based, is estimated in practice using a temporal averaging technique, such as the one described in [18]. However, this process can be problematic for non-stationary source signals, such as long segments of speech. An approximation to this weighting, which has been shown to work well with speech signals, operates on a single, short segment of speech, and can be given in terms of the magnitude spectra of the microphone signals and noise signals [11]:

$$\hat{\Psi}_{12}^{ML} \equiv \frac{|X_1(\omega)||X_2(\omega)|}{|V_1(\omega)|^2|X_2(\omega)|^2+|V_2(\omega)|^2|X_1(\omega)|^2}$$

Here $X_1(\omega)$ and $X_2(\omega)$ are the received microphone spectra, and $V_1(\omega)$ and $V_2(\omega)$ represent the additive noise components that are assumed to be estimated over silence regions.

When reverberation is present in the noise terms, the basic assumption that the noise and source signals are uncorrelated is violated. While ML-type weightings are widely used in speech-array applications, they are inadequate in reverberant environments and will not be used in the experiments of this thesis.

4.1.2 The Phase Transform (PHAT) Weighting Function

Another weighting function, known as the *phase transform* (PHAT) [64], is sub-optimal under reverberation-free conditions, yet performs considerably better than ML in realistic environments. It is a popular form of GCC because of its robustness to reverberation. GCC-PHAT whitens the microphone signals, which in turn whitens the cross-spectrum, $X_1(\omega)X_2'(\omega)$. It is defined as follows:

$$\Psi_{12}(\omega) \equiv \frac{1}{|X_1(\omega)X_2'(\omega)|} \tag{4.10}$$

GCC-PHAT has been shown to be effective in real environments [72][93]. It will be studied in more detail later in this thesis.

4.1.3 Bandpass Weighting Function

The simplest weighting function is one that attenuates frequencies outside the band of interest. For speech, this band is typically 300Hz-6kHz. Hence, a weighting function may be defined as:

$$\Psi_{12}(\omega) \equiv \begin{cases} 1 & 2\pi \cdot 300 \le \omega \le 2\pi \cdot 6000 \\ 0 & \text{Otherwise} \end{cases}$$

It is advantageous to suppress frequency components below 300Hz since much of the power in this range is from background noise, which is generated by air conditioning and heating units, for example. Furthermore, the long wavelengths of low-frequency propagating waves are not of much use to a small-aperture array; it is difficult to determine their direction of propagation. A bandpass weight is often used in conjunction with ML or PHAT, emphasizing only the frequency band where most of the speech energy lies.

4.2 Implementation of GCC

GCC is most commonly implemented using the block-processing scheme described in Section 3.4. The array signals are segmented into small blocks under the assumption that the location of the source is stationary for the duration of each block. An expression for the DFT-based generalized cross correlation of block b can be defined by substituting the block DFTs for the Fourier transforms in Equation 4.9. Let microphones l and q define pair $\{l, q\}$, whose signals are used to compute the DFT-based GCC function from block b, which will be denoted by $\widetilde{R}_{lq,b}(\tau)$. With the block DFTs denoted by $X_{l,b}[k]$ and $X_{q,b}[k]$, this GCC function can be evaluated for any value of the continuous, free variable, τ, as follows:

$$\widetilde{R}_{lq,b}(\tau) \equiv \frac{1}{K}\sum_{k=0}^{K-1} \Psi_{lq}[k] X_{l,b}[k] X'_{q,b}[k] e^{jk\frac{2\pi}{K}\tau} \tag{4.11}$$

$\Psi_{lq}[k]$ is a discrete version of the frequency weighting function $\Psi_{lq}(\omega)$. Identifying that $X_{l,b}[k]X'_{q,b}[k]$ in Equation 4.11 as the DFT-based version of the cross spectrum given by Equation 4.7, which will be denoted by $C_{lq,b}[k]$, $\widetilde{R}_{lq,b}(\tau)$ can also be expressed by:

$$\widetilde{R}_{lq,b}(\tau) \equiv \frac{1}{K}\sum_{k=0}^{K-1}\Psi_{lq}[k]C_{lq,b}[k]e^{jk\frac{2\pi}{K}\tau} \tag{4.12}$$

The separation distance of the microphones physically limits the range of valid time delays. Consider the "end-fire" case where the source is in line with the two microphones indexed by l and q. This case yields the largest TDOA (absolute value) possible, which is equal to d/c where d is the separation distance between these microphones, and c is the speed of sound. Therefore, the range of possible TDOAs is $-d/c$ to $+d/c$. While τ is a continuous variable, Equations 4.11 and 4.12 are sampled in practice, using a suitable step-size, over the range of possible TDOAs.

When the source is in the far field, the time delay parameter, τ, in Equation 4.12 can be expressed in terms of the angle of arrival, θ, as follows:

$$\tau = \frac{d}{c}\sin\theta$$

θ is measured from the perpendicular bisector of the line segment connecting microphones l and q. In the end-fire case, this angle equals either $-\pi/2$ or $+\pi/2$ (-90 or +90 degrees). Hence, Equation 4.12 can be re-expressed in terms of this angle as follows:

$$\widetilde{R}_{lq,b}(\theta) = \frac{1}{K}\sum_{k=0}^{K-1}\Psi_{lq}[k]C_{lq,b}[k]e^{-jk\frac{2\pi}{K}\frac{d}{c}\sin\theta}, \qquad -\tfrac{\pi}{2} \le \theta \le \tfrac{\pi}{2} \tag{4.13}$$

While θ is a continuous variable, the above equation is sampled in practice, using a suitable step-size, over the range $-\frac{\pi}{2} \le \theta \le -\frac{\pi}{2}$.

It has been assumed that the talker's location remains fixed for the duration of each data block. This assumption is valid when the block size is as small as 25 milliseconds. However, at the possible cost of reduced responsiveness to moving talkers, the cross-spectrum estimates from several successive blocks can be averaged to increase GCC's performance. Such averaging suppresses the noise and reverberation that is uncorrelated from block to block, and gives an improved estimate of the true cross spectrum [48]. The cross-spectrum between microphone signals l and q, which is computed for the b-th analysis block, and averages data from blocks $b \ldots b+I-1$, is defined as follows:

$$C_{lq,b}[k] \equiv \frac{1}{I} \sum_{i=b}^{b+I-1} X_{l,i}[k] X'_{q,i}[k] \tag{4.14}$$

Hence, $C_{lq,b}[k]$ is obtained by averaging $X_{l,b}[k]X'_{q,b}[k]$ over I blocks. The performance of cross-correlation techniques generally improves with longer data segments. Therefore, there is a temptation to make I as long as possible. However, there is always a tradeoff between responsiveness and robustness. If there is too much averaging, DOA estimates cannot keep up with moving talkers. If there isn't enough averaging, room reverberation and noise severely impact accuracy.

Implementation of the phase-transform using the DFT-based cross-correlation given by Equation 4.12 leads to the following discrete version of the weighting function:

$$\Psi_{lq,b}[k] \equiv \frac{1}{\left| C_{lq,b}[k] \right|} \tag{4.15}$$

Substitution of this equation into Equation 4.12 gives the DFT-based phase transform function, which uses data from blocks $b \ldots b + I - 1$, beginning with the block indexed by b:

$$\widetilde{R}_{lq,b}(\tau) = \frac{1}{K} \sum_{k=0}^{K-1} \frac{C_{lq,b}[k]}{\left| C_{lq,b}[k] \right|} e^{jk\frac{2\pi}{K}\tau} \tag{4.16}$$

4.3 RMS TDOA Error for an Array

A useful accuracy measure of a collection of TDOA estimates can be defined using the root mean square (RMS) error. The *RMS TDOA error* will be defined as the RMS of the errors in the individual, pairwise TDOA estimates. The individual TDOA errors are the differences between the estimated and known TDOAs. Denoting the known TDOA for microphones l and q by τ_{lq}^0 and the estimated TDOA by $\hat{\tau}_{lq}$, the TDOA error for pair $\{l, q\}$ is given by:

$$\hat{\tau}_{lq} - \tau_{lq}^0$$

The known TDOAs can be calculated using the known positions of the source and microphones. With the location of the source denoted by $\vec{d}^{(s)}$, and the locations of microphone l and q denoted by $\vec{d}_l$ and $\vec{d}_q$, respectively, the true TDOAs, under simple acoustic conditions (see Chapter 2), can be computed by:

$$\tau_{lq}^{0} = \frac{\left|\vec{d}_l - \vec{d}^{(s)}\right| - \left|\vec{d}_q - \vec{d}^{(s)}\right|}{c}$$

where c is the speed of sound. The true delays can also be derived from the measured source-to-microphone-output impulse responses.

For an array of M microphones, there are a total of $\frac{M(M-1)}{2}$ possible pairwise combinations (i.e. there are "*M choose 2*" 2-combinations of an M-set). Any subset of these pairings can be used for TDOA estimation, resulting in a multitude of TDOA estimates. By taking the root mean square of the individual TDOA errors, a single RMS error characterizes the accuracy of all TDOA estimates taken from the same array.

It is convenient to define the RMS error in terms of TDOA vectors. At most, such a vector has $\frac{M(M-1)}{2}$ elements, one for every possible pairwise TDOA. Hence, a complete TDOA vector is given by:

$$\boldsymbol{\tau} \equiv \left[\tau_{12} \quad \tau_{13} \quad \ldots \quad \tau_{1M} \quad \tau_{23} \quad \tau_{24} \quad \ldots \quad \tau_{2M} \quad \ldots \quad \tau_{(M-1)M}\right]'$$

Denoting the vector of known TDOAs by $\boldsymbol{\tau}^0$ and the vector of estimated TDOAs by $\hat{\boldsymbol{\tau}}$, the RMS error can be defined as follows:

$$E_{RMS}(\hat{\boldsymbol{\tau}}) \equiv sqrt\left\{\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^0\right)'\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^0\right)\right\} \tag{4.17}$$

4.4 Source Localization by Minimization of the RMS TDOA Error

The source can be localized by minimizing the RMS TDOA error, as defined by Equation 4.17. By searching over a pre-defined set of spatial coordinates, the RMS error can be computed for each candidate point. Instead of using $\boldsymbol{\tau}^0$, which is the actual TDOA vector, the TDOAs corresponding to each candidate point are used to compute this error. Denoting the candidate point by $\vec{d}$, the corresponding TDOA vector, $\boldsymbol{\tau}$, can be constructed from the following elements:

$$\tau_{lq} = \frac{\left|\vec{d}_l - \vec{d}\right| - \left|\vec{d}_q - \vec{d}\right|}{c} \qquad \text{for } l, q \in \{1 \ldots M, l \neq q\}$$

For a fixed set of microphone locations, $\vec{d}_1 \ldots \vec{d}_M$, and a given TDOA-estimate vector, $\hat{\boldsymbol{\tau}}$, the RMS error is a function of the candidate location, $\vec{d}$, and it can be defined as follows:

$$E(\vec{d}) \equiv sqrt\left\{\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\vec{d}, \vec{d}_1 \ldots \vec{d}_M)\right)' \left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\vec{d}, \vec{d}_1 \ldots \vec{d}_M)\right)\right\} \tag{4.18}$$

An estimate of the source's location is given by the candidate location that minimizes $E(\vec{d})$:

$$\hat{\vec{d}} = \arg\min_{\vec{d}} E(\vec{d})$$

Since each candidate point is a 3-element vector with the Cartesian coordinates of the candidate location, Equation 4.18 is a function of three spatial variables. To minimize this error, a search must be performed over these variables, and it can be computationally intensive. This computational burden can be eased by using a *simplex search* [74], for example, which works well because $E(\vec{d})$ tends to be smooth and unimodal.

As discussed in Section 2.6, far field conditions limit the ability of an array to estimate range. When the range of the source is ambiguous, the RMS TDOA error is a function of only two spatial dimensions, azimuth and elevation. Hence, by defining the candidate delays in terms of these spatial variables, the RMS TDOA error can be minimized over direction of arrival (DOA) instead of source-location. While this obviously eases the computational load, the source cannot be completely localized in 3-D space. Generally, the DOAs from multiple far-field arrays can be used, via triangulation, to yield an estimate of the source's location.

In the far field case, the candidate TDOAs can be defined in terms of the assumed direction of arrival at the array's origin. As defined in Section 2.6, the direction of arrival is opposite the direction of propagation, $\vec{\zeta}_o^{(s)}$. Denoting the assumed propagation vector by $\vec{\zeta}_o$, the candidate TDOAs can be computed by:

$$\tau_{lq} = \frac{-\vec{\zeta}_o \cdot \left(\vec{d}_l - \vec{d}_q\right)}{c} \qquad \text{for } l, q \in \{1 \ldots M, l \neq q\}$$

The vector that defines the assumed direction of arrival, which is $-\vec{\zeta}_o$, is also known as the *look vector*, since this is the direction the array is "looking" to find the source. Like the propagation vector of Equation 2.9, the look vector can be defined in terms of the azimuth and elevation angles, θ and ϕ, as follows:

$$-\vec{\zeta}_o \equiv \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix}$$

These angles, θ and ϕ, define the assumed direction of arrival, or look direction, relative to the array's local origin. Now, the RMS TDOA error for a far-field source can be expressed as a function of azimuth and elevation as follows:

$$E_{FAR}(\theta,\phi) \equiv sqrt\left\{\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\theta,\phi,\vec{d}_1 \ldots \vec{d}_M)\right)'\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}(\theta,\phi,\vec{d}_1 \ldots \vec{d}_M)\right)\right\} \tag{4.19}$$

To find the DOA of the source, this error must be minimized over two spatial dimensions:

$$(\hat{\theta},\hat{\phi}) = \arg\min_{\theta,\phi} E_{FAR}(\theta,\phi)$$

The lower and upper bounds on these angles correspond to -90 degrees and +90 degrees, respectively. However, this range can be reduced according to the geometric constraints of the specific application. For example, the array may be positioned in such a way that valid talkers can only be located in a field that corresponds to -60 to 60 degrees.

5 Experimental Performance Evaluations of GCC

This chapter includes a series of three experiments designed to establish baseline performance of GCC TDOA estimation in a mildly reverberant, high-SNR environment. Subsets of the conference-room data set, described in Section 3.2, were used for this series. As reported in Chapter 3, this data set had an average SNR of 35dB and a reverberation time of about 200 milliseconds. Hence, these experiments truly examine the effects of reverberation on GCC algorithms' performance (hereafter shorted to "GCC's performance") using real data from a realistic environment. The goals of the experiments in this chapter are summarized as follows:

- Show that GCC's performance, in a mildly reverberant environment, is poor when the data blocks are as short as 25 milliseconds.
- Show a connection between anomalous TDOA estimates and the secondary peaks of the cross-correlation of the source-to-microphone room impulse responses.
- Define the TDOA *error rate*, which will be used to evaluate performance, in these and the following experiments.
- Show that GCC-PHAT is an appropriate weighting function for speech sources.
- Show that GCC's performance improves with the amount of data used to compute the cross-correlation function.

The first experiment examines the performance of GCC using a single pair of microphones. This simple scenario allows a clear introduction to GCC and the way reverberation impacts its performance. The second experiment employed a 3-element array, which has been dubbed the "triad array". This experiment introduces the use of the RMS TDOA error, defined in Section 4.3. It also introduces the TDOA *error rate*. The third and final experiment in this series, uses the data from an 8-element array to estimate the DOAs of the speech sources by minimization of the RMS TDOA error as described in Section 4.4. In part of this experiment, cross-correlations are averaged, over multiple blocks, and the performance of GCC-PHAT is reported for the various block-averaging lengths.

5.1 GCC Experiment #1: TDOA Estimation with a Single Pair of Microphones

Two microphones were selected from the conference-room array panel. As illustrated by Figure 5.1, microphones 2 and 9 comprised this pair. The orientation of the pair was approximately 18 degrees from vertical, and the separation distance of the microphones was 36 centimeters. The Gaussian noise recordings were used in conjunction with GCC to estimate the single TDOA between microphones 2 and 9.

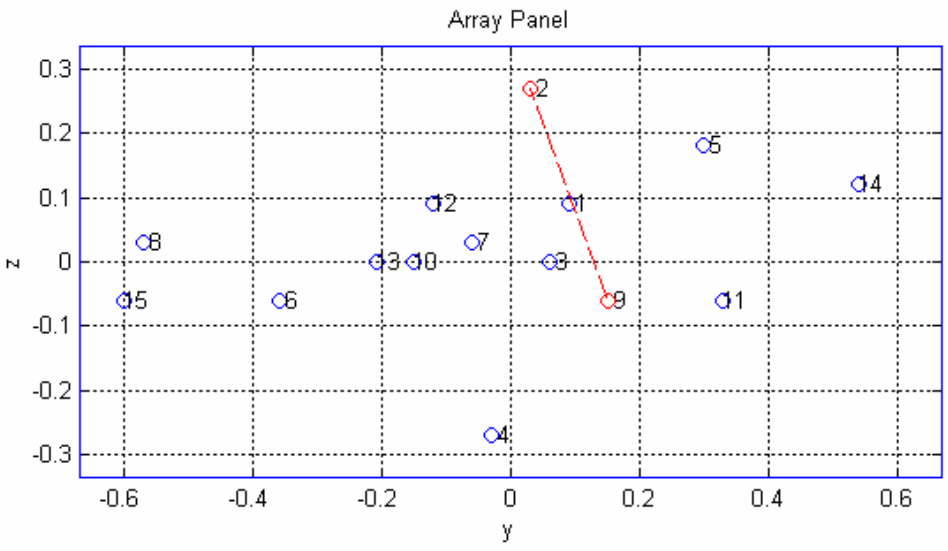

Figure 5.1 Microphones 2 and 9 comprise a pair with a 36cm separation distance.

With the range of the sources from the array (greater than 3 meters) being much larger than the microphone separation distance, it can be assumed that the wave fronts impinging on this pair were planar (far field conditions). Hence, the DOA, α, is related to the TDOA, τ_{29}, by the following:

$$\alpha = \sin^{-1}\left(\frac{c\tau_{29}}{d}\right)$$

where c is the speed of sound, 342 meters per second, and d is the separation distance, 0.36 meters. At a sampling rate of 16kHz, α is related to the TDOA in sample units as shown by the graph of Figure 5.2. This graph shows that the maximum possible TDOA for this separation distance is 16.4 samples. Hence,

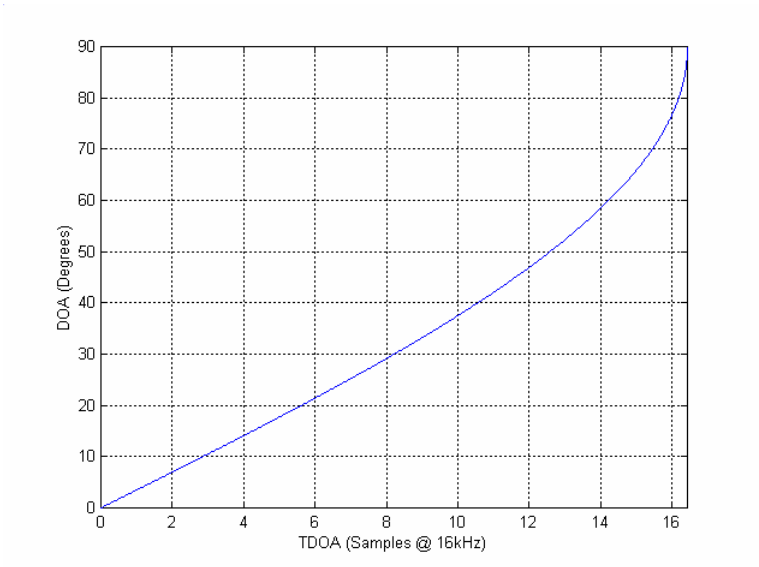

Figure 5.2 Plane wave DOA-TDOA relationship for a microphone pair with a 36cm separation.

the valid range of TDOAs, which corresponds to a DOA range of $-90 \le \alpha \le 90$ degrees, is ± 16.4 samples. The true TDOA for this pair and source 1 was –0.9 samples, which corresponds to approximately –4 degrees. The TDOAs for sources 2 and 3 were 2.0 samples (8 degrees) and 0.1 samples (0.5 degrees), respectively.

5.1.1 TDOA Estimation

The signals from microphones 2 and 9 were segmented into 25-millisecond blocks using the procedure described in Section 3.4. A Hanning window was applied to each block before the DFTs were taken. With a block advance of 12.5 milliseconds, the 5-second recordings yielded 399 blocks. The TDOA was estimated by finding the time delay that maximized the cross-correlation function from each data block. Hence, there were 399 TDOA estimates made over the duration of each 5-second recording (one estimate per block).

Equation 4.11 was used to compute the GCC function for each data block. Each GCC response was computed over the range of possible TDOAs (-16.4 to +16.4 samples) with a step-size of 0.1 samples. According to the plot of Figure 5.2, this step size corresponds to less than 0.5 degrees in look direction. A discrete high-pass weighting function was used, which is defined by:

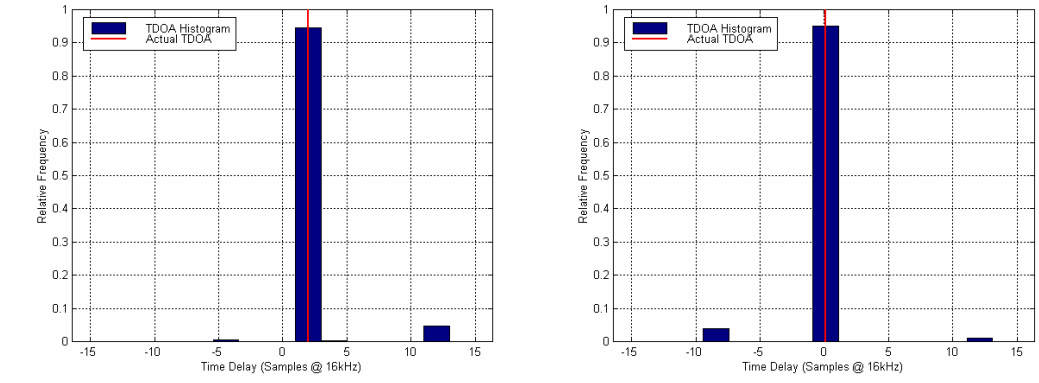

Figure 5.3 Histograms of the TDOA estimates of source 2 (left) and source 3 (right).

$$\Psi^{HP}[k] \equiv \begin{cases} 1 & \frac{300}{f_s}K \le k \le \frac{8000}{f_s}K \\ 0 & \text{otherwise} \end{cases}$$

K is the DFT length and f_s is the sampling rate. With the exception of the frequencies below 300Hz, which tend to include much of the background noise, all GCC frequency components were weighted equally by this high-pass weighting function. With a Gaussian source signal and high SNR-conditions, the use of such a weighting function is justified; a nearly white source and no noise should not require pre-filtering. Furthermore, any significant erroneous behavior by the TDOA estimator would have to be a result of multipath propagation. Since the conference room is perceived to be acoustically "dead", one would expect there to be predominately single-path propagation. Hence, pairwise GCC *should* perform well using a uniform weighting function.

5.1.2 Experimental Results and Discussion

Histograms of the pairwise TDOA estimates for Gaussian sources 2 and 3 are shown by the bar graphs of Figure 5.3. The vertical axes of these histograms range from 0 to 1, where 1 corresponds to the total number of estimates and values within this range correspond to some fraction of the total (similar to a probability). The actual TDOAs are shown by the vertical line in each graph, which correspond to 2.0 and 0.1 samples, respectively. As theses histograms show, over 90 percent of the estimates fell within two

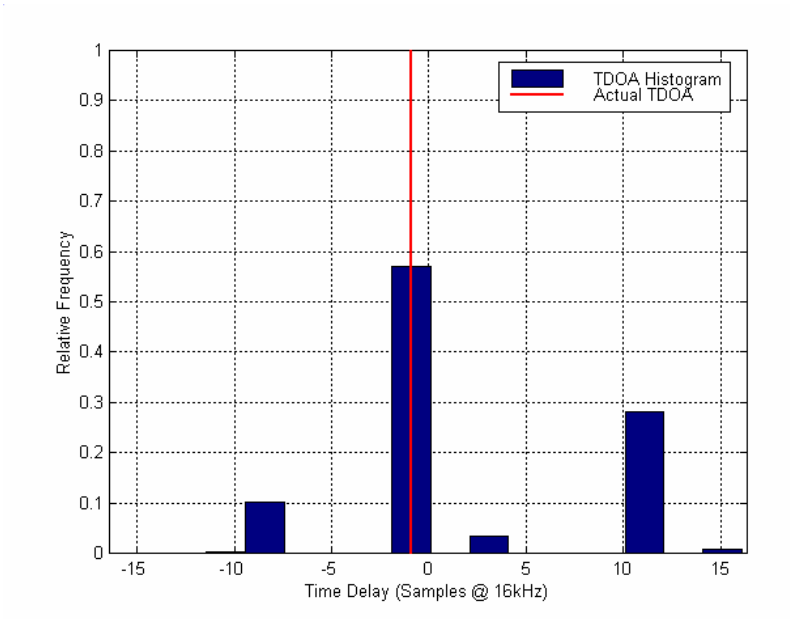

Figure 5.4 Histogram of TDOA estimates from source 1.

samples of the actual TDOAs. Hence, for these two source locations, which were between 3 and 4.5 meters from the array, GCC performed as expected under nearly ideal SNR and source-signal conditions.

A histogram of the pairwise TDOA estimates from source 1 is shown by the bar graph of Figure 5.4. The actual TDOA is shown by the vertical line, which corresponds to –0.9 samples. Less than 60 percent of the estimates fell within two samples of this actual value. Considering the nearly ideal source-signal and low-noise conditions, GCC performed surprising poorly in this experiment. This result supports the notion that sound wave propagation in a room, which is perceived to be acoustically "dead", does not necessarily follow the single-path propagation model. If this is true, then the erroneous TDOA estimates should correspond to contributions from the room impulse responses. This connection was investigated using the impulse response measurements of Section 3.6.

Recall the microphone signal model of Equation 2.6:

$$x_m(t) = s(t) * \tilde{h}_m(\vec{d}^{(s)}, t) + v_m(t)$$

This can be expressed in the temporal frequency domain as follows:

$$X_m(\omega) = S(\omega)\tilde{H}_m(\vec{d}^{(s)}, \omega) + V_m(\omega) \tag{5.1}$$

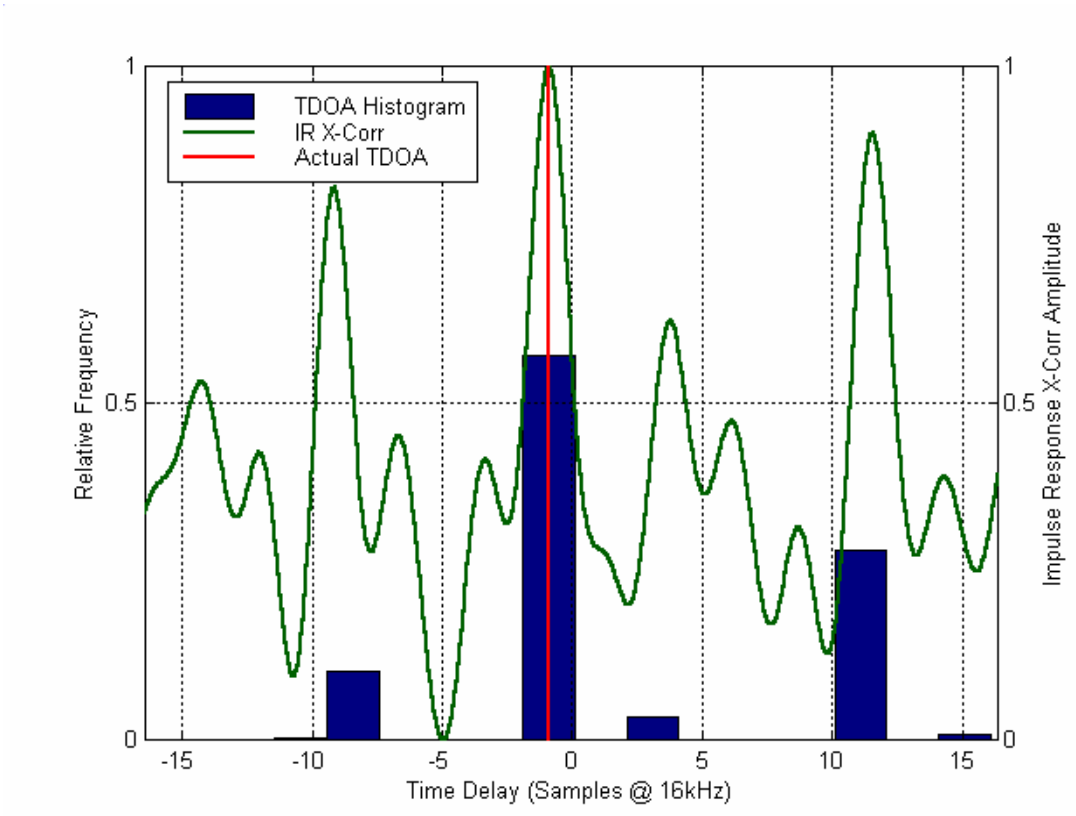

Figure 5.5 Histogram of TDOA estimates with cross-correlation ("X-Corr") of measured room impulse responses ("IR").

With SNRs between 31dB and 38dB in the conference-room data set, the contribution from $V_m(\omega)$ was negligible. Setting this term to zero in Equation 5.1, the cross-correlation of microphones signals 2 and 9 can be expressed using Equation 4.8 as follows:

$$c_{29}(\tau) = \frac{1}{2\pi}\int |S(\omega)|^2 \widetilde{H}_2(\omega)\widetilde{H}_9'(\omega)e^{-j\omega\tau}d\omega \tag{5.2}$$

The power spectra of the Gaussian source signal, $|S(\omega)|$, was nearly constant over frequency. Hence, with $|S(\omega)| \to S$, Equation 5.2 can be expressed as follows:

$$c_{29}(\tau) = S^2 \cdot \widetilde{h}_2(\tau) * \widetilde{h}_9(-\tau) \tag{5.3}$$

where "*" denotes convolution.

Using the discrete form of Equation 5.3, the cross-correlation of the measured impulse responses, $\hat{\tilde{h}}_2[n]$ and $\hat{\tilde{h}}_9[n]$ was computed for Gaussian source 1 and up-sampled to 10-times the sampling rate, giving it the same resolution as the block GCCs. Figure 5.5 shows the normalized cross-correlation, labeled "IR X-Corr", plotted as a function of delay (time lag). Also shown in this figure is the TDOA histogram from Figure 5.4. It is apparent in Figure 5.5 that the erroneous TDOA estimates in the histogram occur where there are large secondary peaks in the cross-correlation function. Even though the height of these secondary peaks are less than the direct-path peak, the far-end reverberation appears as noise in the short 25-millisecond blocks and corrupts the true amplitudes of the peaks in the GCC function from each block. Since these secondary peaks are nearly as tall as the direct-path peak, the reverberation noise produces many anomalies that correspond to picking these false peaks instead of the direct-path peak. These anomalies tend to increases with the separation distance between the microphones; more secondary peaks appear in the range of valid TDOA and cause more erroneous TDOA estimates.

Figure 5.6 shows the GCC response for each data block of Gaussian source 1 over a 2-second interval of the recording. The vertical axis of this 2-D image plot is time-delay parameter, τ, and the

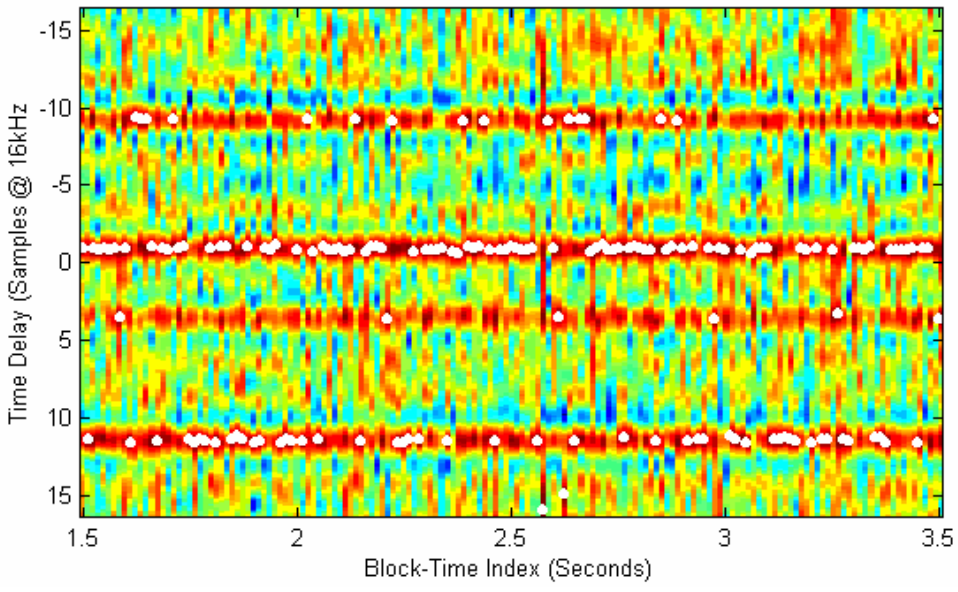

Figure 5.6 Normalized GCC responses over time (each block) for Gaussian source 1.

horizontal axis is the time index of each block. The color of the image represents the amplitude of the responses (the darker/redder colors represent higher amplitudes). Superimposed on the image plot are white points, which highlight the maximum, over τ, of each response (time slice). The delay values, on the vertical axis, that correspond to these maxima are the TDOA estimates. The amplitude of each response has been normalized so that the maximum value from each response equals one. While this normalization does not affect the TDOA estimation, it does make the responses clearer when plotted in this manner.

One of the four strong red horizontal bands in Figure 5.6 corresponds to the true TDOA of -0.9 samples. The others correspond to secondary peaks in the cross-correlation of the room impulse responses, as shown by Figure 5.5. The secondary bands also correspond to the erroneous TDOA estimates, which is shown by the histogram of Figure 5.5. The white points in the responses of Figure 5.6 show that these erroneous peaks (among others) were picked quite frequently over the course of the Gaussian noise recording. Even though the Gaussian signal had nearly constant power across all blocks, the small fluctuations in the signal were enough to excite the secondary peaks and sometimes make them taller than the main peak. The far-end reverberation also contributed to this effect. Knowing that the recordings had high SNRs, 35dB on average, uncorrelated background noise could not have played a role here.

5.2 GCC Experiment #2: RMS TDOA Error with a Triad Array

A subset of the conference-room data set was used to form a 3-element array, which has been dubbed the "triad array". The microphones were chosen so that they lie on the vertices of an equilateral triangle with 36-cm sides. This is shown in Figure 5.7. The TDOAs were estimated for each of the three possible microphone pairings using the procedure from the previous experiment, described in Section 5.1.1. With the separation distances between the microphones equal, the range of possible TDOAs was the same for all three pairs. This range was the same as the previous experiment, ± 16.4 samples, since pair {2,9} is

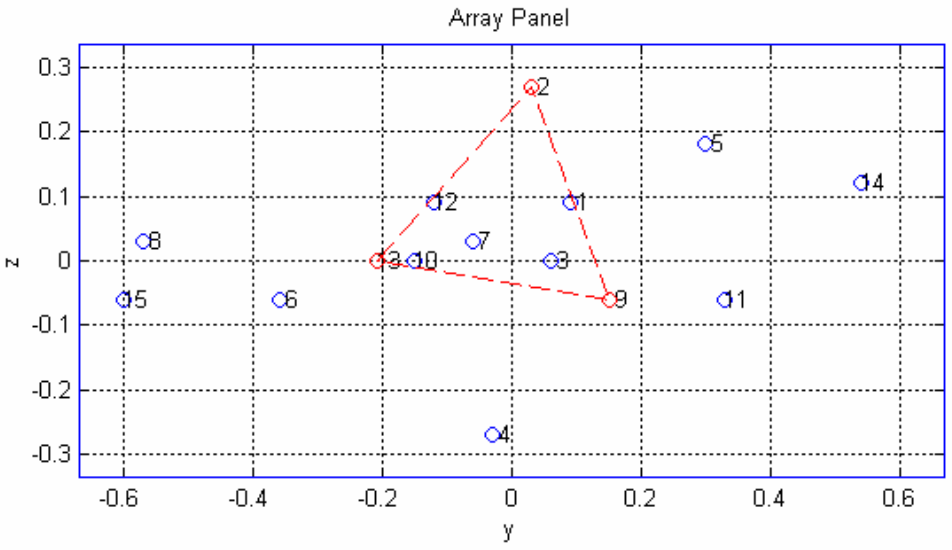

Figure 5.7 Microphones 2, 9 and 13 form a triad with a 36cm separation distance.

included in the triad array. Hence, the DOA-TDOA plot of Figure 5.2 applies to each pair in this experiment as well. This setup gave three pairs with three unique spatial orientations and comparable TDOA ranges.

5.2.1 RMS TDOA Errors

The root mean square (RMS) TDOA error of an array was defined in Section 4.3. This error was computed for each Gaussian source, and it included the TDOA estimates from all three microphone-pairs. Figure 5.8 shows the histograms of the RMS TDOA error. The data from sources 1 and 2, which were more distant

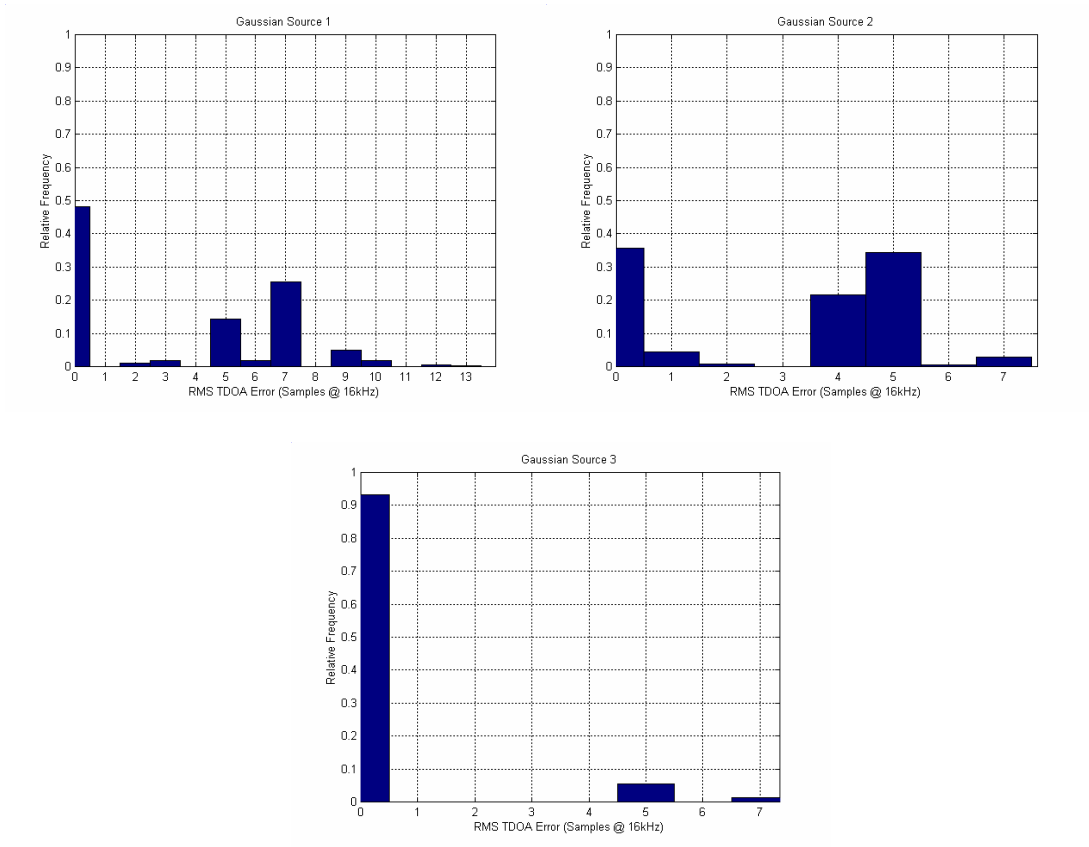

Figure 5.8 RMS TDOA error histograms for three Gaussian sources and the triad array.

than source 3, produced poor results. Only about 50 percent and 40 percent of the source-1 and source-2 estimates, respectively had error less than 3 samples. Source 3, which was about 3 meters from the array, gave a much better performance than the other two sources, which were both more than four meters away. This is consistent with the hypothesis that signal-to-reverberation ratios decrease with increasing source-to-microphone distance. As the signal-to-reverberation decreases, reflected sound waves become comparable to the direct-path sound in strength. When this occurs, the single-path propagation model is no longer valid, and estimators such as GCC, which are based on this model, exhibit poor performance.

Notice how the performance of the triad array, evaluated using the RMS error, was considerably worse than that of one of its pairs, {2,9}, as reported in Section 5.1, for the source-2 recording. Over 90 percent of the estimates produced by this pair (on its own) were within two samples of the true TDOA, as compared to 40 percent for the triad array. At least one of the other two pairs brought the average performance down to an unacceptable level even though pair {2,9} produced highly accurate estimates.

This shows that spatial orientation matters; all three pairs had the same separation distance and nearly the same locations (they even shared microphones), yet not all the pairs could accurately resolve the relative strengths of the direct-path sound and reflected sound. This is a major shortcoming of pairwise techniques; they can only resolve DOA in the spatial dimension that coincides with the pair's orientation.

5.2.2 RMS TDOA Error Rates with Gaussian Sources

The *error rate* will be used extensively in the remainder of this thesis to evaluate performance of talker-localization techniques. In general, the error rate is the percentage of estimates with error greater than or equal to some threshold, plotted as function of that threshold. The error can be measured in whatever way is appropriate, but it must be non-negative. One such error is the RMS TDOA error, which was used in Section 5.2.1 to evaluate the performance of the triad array.

Figure 5.9 shows the same histogram of the RMS TDOA error for Gaussian source 1 that was first shown in Figure 5.8. Superimposed on this histogram is the error rate, with the scale of its vertical axis

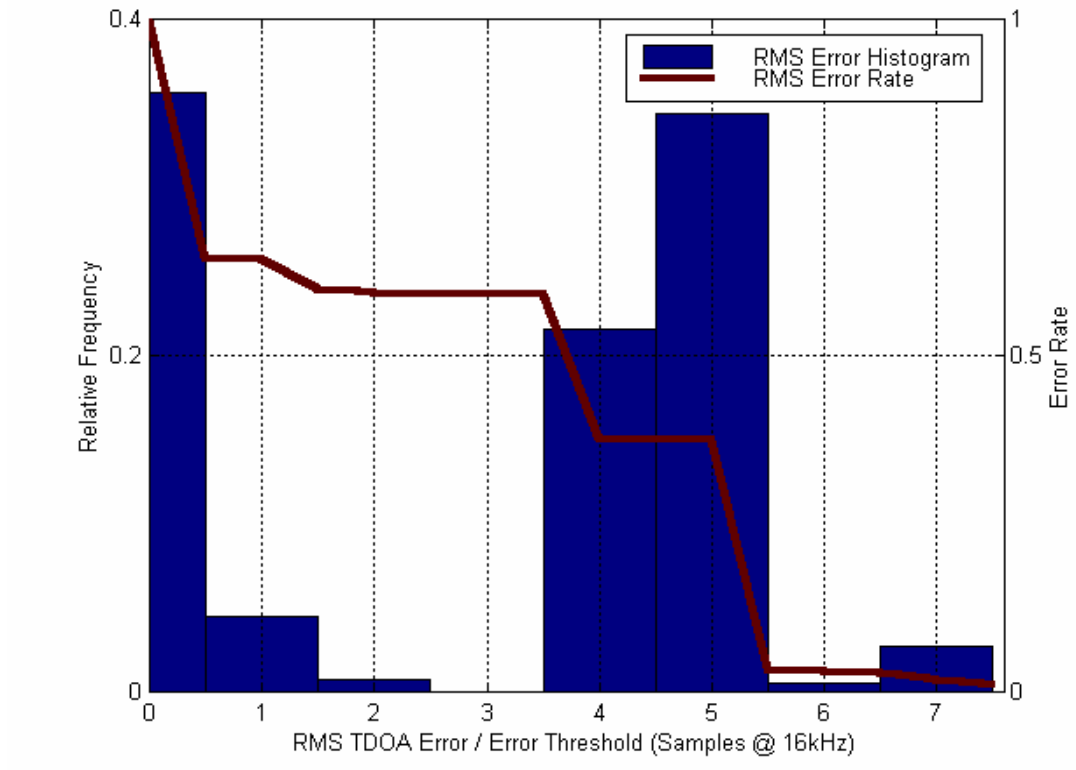

Figure 5.9 Histogram and error rate of RMS TDOA error for Gaussian source 1.

labeled on the right side of the graph. This example clearly shows the connection between the two plots. Notice that the changes in the error rate correspond to the distribution of the RMS error; the error rate is equivalent to one minus the sum of the histogram over the error axis (a cumulative distribution). Recall, from Figures 5.5 and 5.6, that the erroneous TDOA estimates tend to cluster near the delays that correspond to secondary peaks in the cross-correlations. This causes the error rate to sometimes have nearly discrete steps at the corresponding error thresholds. This effect is less pronounced with a large number of GCC pairs; as the number of pairs increase, the RMS TDOA error becomes more uniformly distributed since each microphone pair has a slightly different cross-correlation with its secondary peaks corresponding to different delays. The RMS error rate of Figure 5.9 is relatively smooth, but does reflect the large clusters of error near 0, 5 and 7 samples.

The error rates for all three Gaussian sources are plotted in Figure 5.10. These three error-rate plots report the accuracy of the TDOA estimates in a way that is easier to compare than the error

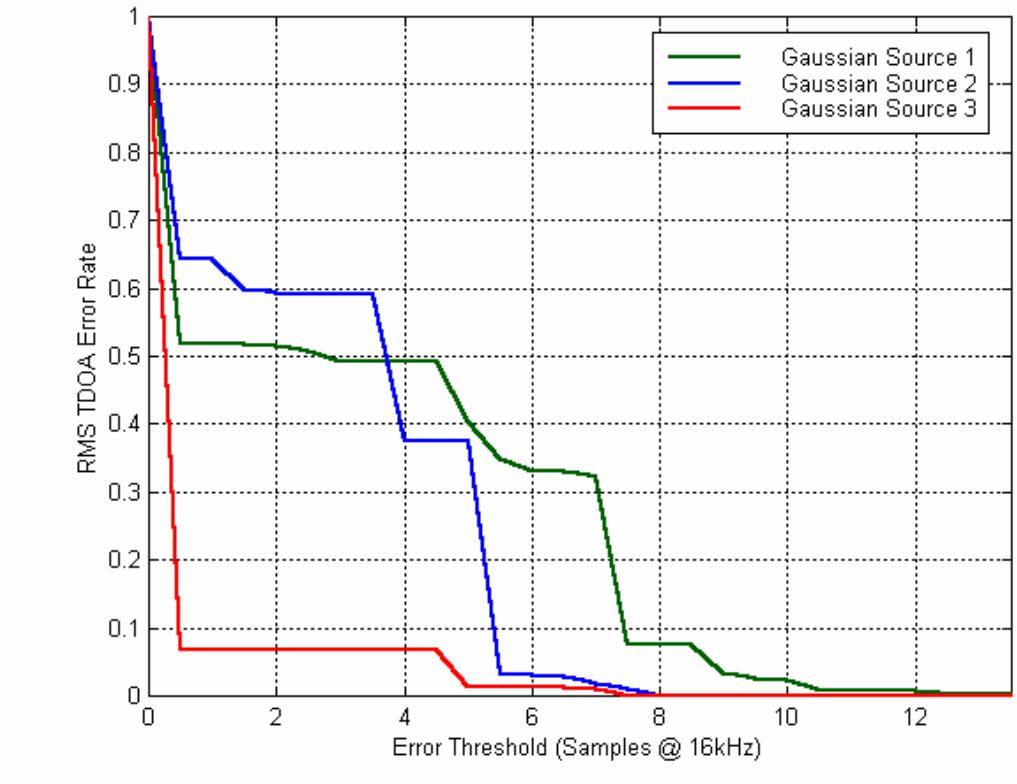

Figure 5.10 RMS TDOA error rates for Gaussian sources 1, 2 and 3.

histograms of Figure 5.8, while containing nearly as much information. They are far more informative than statistics such as standard deviation and bias, which are used frequently to evaluate performance. From Figure 5.10, it is easy to see that the data from the closest source, source 3, produced many more accurate estimates than the other two. Less than 10 percent of the estimates from source 3 had error greater than or equal to 0.5 samples (or 90 percent had error less than 0.5 samples) while about 50 percent of the others' estimates had error greater than 4 samples.

5.2.3 RMS TDOA Error Rates with Speech Sources

The experiment was run again using the array recordings of speech and the same triad array. As described in Section 3.2, there are three speech recordings in the conference-room data set, and the loudspeaker location during each was the same as it was during the Gaussian noise recordings. TDOA estimation was again performed using the same parameters: a Hanning window, 25-millisecond blocks, a 12.5-millisecond block advance, and a 0.1-sample TDOA resolution. The procedure was the same as the one applied to the Gaussian noise recordings, with one additional step. An SNR mask was used to discard any TDOAs that were produced by low-SNR speech blocks. The mask was derived using a 0.33 threshold and applied as explained in Section 3.5. Out of 399 blocks per recording, the mask passed 313 from source 1, 340 from source 2 and 297 from source 3. These speech recordings were processed using two GCC weighting functions. The first was the high-pass weighting function from Section 5.1.1, which was also applied to the Gaussian noise recordings. The second was a combination of this high-pass weighting function and the *phase transform* (PHAT) weighing function, which was presented in Section 4.1.2.

RMS error rates were computed using the TDOA estimates produced by the speech recordings. These are shown in Figure 5.11. Also shown in this figure for comparison sake are the error rates for the corresponding Gaussian sources from Figure 5.10. Hence, for each source location, there are three error rates. The first two were derived from GCC with the high-pass weighing function, one for the Gaussian signals (labeled "Gaussian / GCC") and one for the speech signals (labeled "Speech / GCC"). The third error rate in each plot was from GCC with the combined GCC-PHAT and high-pass weighing function for the speech signals (labeled "Speech / GCC-PHAT").

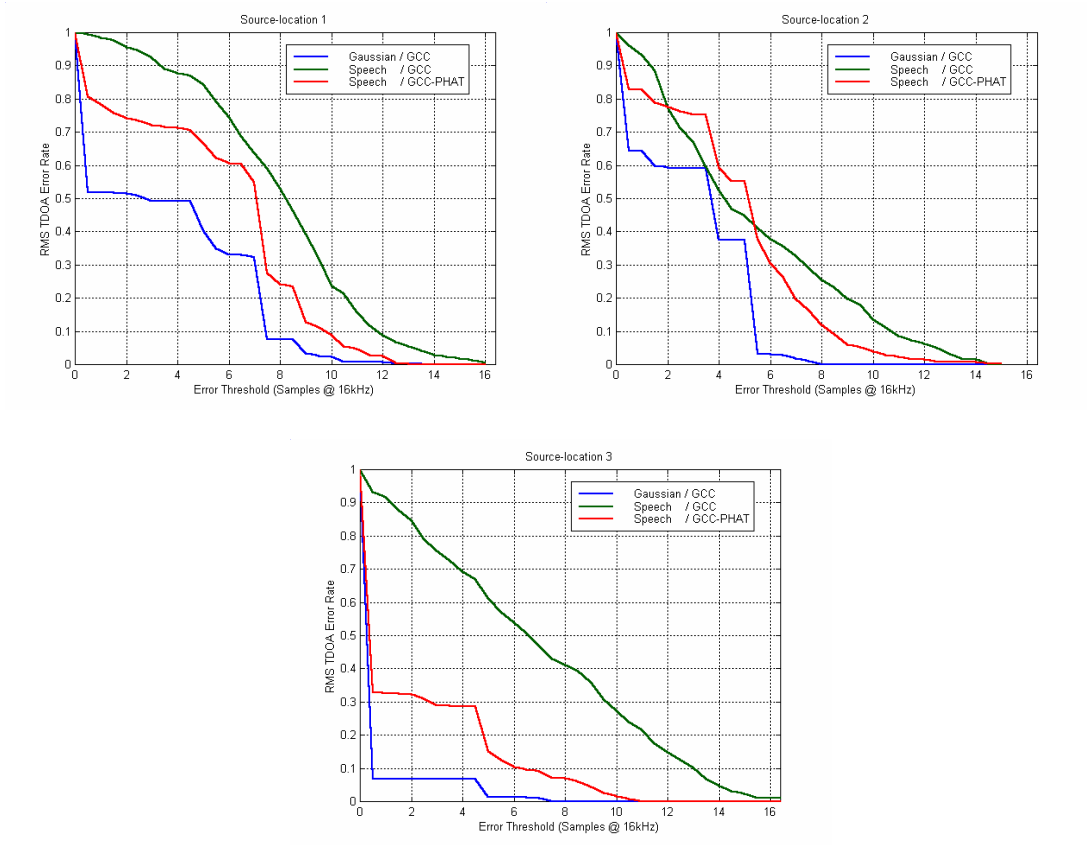

Figure 5.11 Error rates for the three source locations and two source signals at each location: Gaussian and speech. The speech sources were processed using both GCC and GCC-PHAT with SNR masks.

In Figure 5.11, the Gaussian-source error rates serve as the baseline performance of GCC with the triad array in the conference room environment. These signals have properties that are nearly ideal and give GCC, in some sense, the "best you can do" results. The error rates reflect this, showing that the Gaussian sources consistently produce estimates that are more accurate than the speech sources. In general, the PHAT weighting function puts GCC performance somewhere between that of the Gaussian sources and the speech sources without PHAT. Without the PHAT weighting, the speech sources produce unacceptable results, and this is true for all the source locations. Unfortunately, for source locations 1 and 2, which are about 4.5 and 5.5 meters from the array, respectively, the baseline performance from the Gaussian signals is very poor, and there's little GCC-PHAT can do to boost the performance from the speech recordings. However, it is clear that for the closest source, source 3, which is about 3 meters from the array, GCC-PHAT brings the speech performance close to the Gaussian performance, which is quite good.

These results indicate GCC produces acceptable results only when the source is close to the array (3 meters in this case). They also show that GCC-PHAT works quite well with speech sources, and it produces estimates that are nearly as accurate as the estimates produced by the baseline, Gaussian signals. Hence, GCC-PHAT will be further studied in this thesis.

5.3 GCC Experiment #3: DOA Estimation with an 8-Element Array

Eight microphones were selected from the conference-room array panel to form the sub-array shown in Figure 5.12. These microphones lie within a 33 by 36 centimeter rectangle, resulting in an aperture size that is much smaller than the distance from the array to the nearest source. Hence, is can be assumed that all three sources in the conference-room data set lie in the far field of this sub-array. Under such conditions, range estimates are ambiguous, and only the azimuth and elevation angles can be estimated reliably. As presented in Section 4.4, when far-field conditions hold, the DOA of the source can be

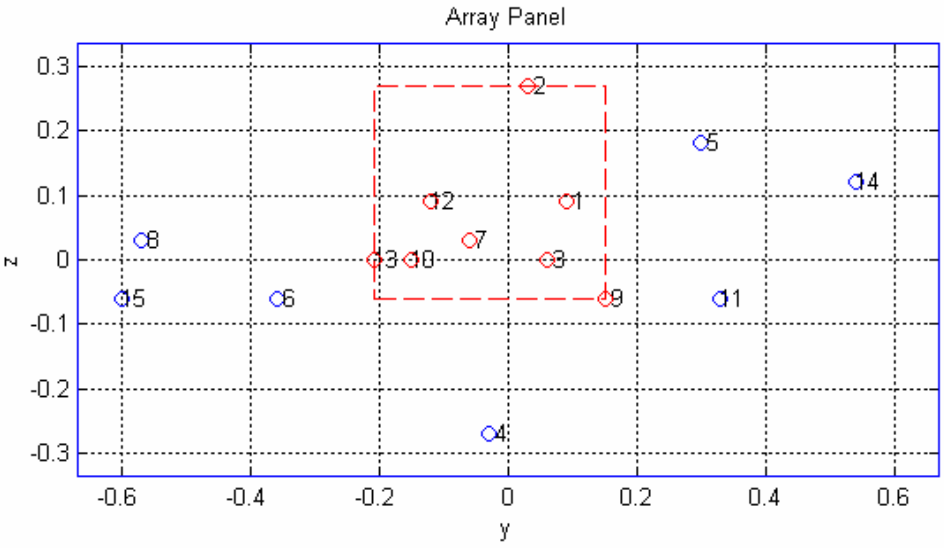

Figure 5.12 An 8-element, 33 by 36 centimeter sub-array.

estimated by minimization of the RMS TDOA error evaluated over azimuth and elevation relative to the array's origin. This technique was used in conjunction with the GCC-PHAT TDOA-estimation procedure from the previous experiment (See Section 5.2.3) to estimate the DOAs of the three speech sources.

5.3.1 DOA Estimation by Minimization of the RMS TDOA Errors

By taking all possible combinations, 28 microphone pairs were formed using the 8-element array. Hence, for each data block, 28 TDOA estimates were made for each of the three speech recordings using GCC-PHAT. An SNR mask was applied to the data using the same technique described in Section 3.5. The

SNR threshold was again set to 0.33, and out of 399 blocks per recording, the mask passed 313 from source 1, 340 from source 2 and 297 from source 3.

With the diversity in the spatial orientations of the microphone pairs, the collection of 28 TDOAs represented an over-determined system of parameters to solve for the DOA of the source in both azimuth and elevation. Hence, for each data block that was not rejected by the SNR mask, the RMS TDOA error of Equation 4.19, $E_{FAR}(\theta,\phi)$, was computed using all 28 TDOA estimates over a predefined range of azimuth, θ, and elevation, φ. While these angles range from minus ninety to ninety degrees in general, Equation 4.19 was computed over the more limited range of ± 60 degrees based on the valid talker-locations. This covers the region around the conference table, including the three loudspeaker locations used to make the recordings (see Figure 3.5). The RMS TDOA error was computed and minimized over this region using a grid of 0.1 degrees. For each DOA estimate produced in this way, an RMS DOA error was computed, which is simply defined as follows:

$$E_{DOA}(\hat{\theta},\hat{\phi}) = \sqrt{\left(\hat{\theta}-\theta^{(s)}\right)^2 + \left(\hat{\phi}-\phi^{(s)}\right)^2} \tag{5.4}$$

where $\theta^{(s)}$ and $\phi^{(s)}$ are the actual DOA angles from the array to the source.

Figure 5.13 shows a plot of the error rates, computed over all the block-DOA estimates for each source, of the minimized RMS TDOA error (from Section 4.3) and the RMS DOA error, $E_{DOA}(\hat{\theta},\hat{\phi})$. With the exception of source 3, which was closest to the array, the DOA error rates were very high. Over

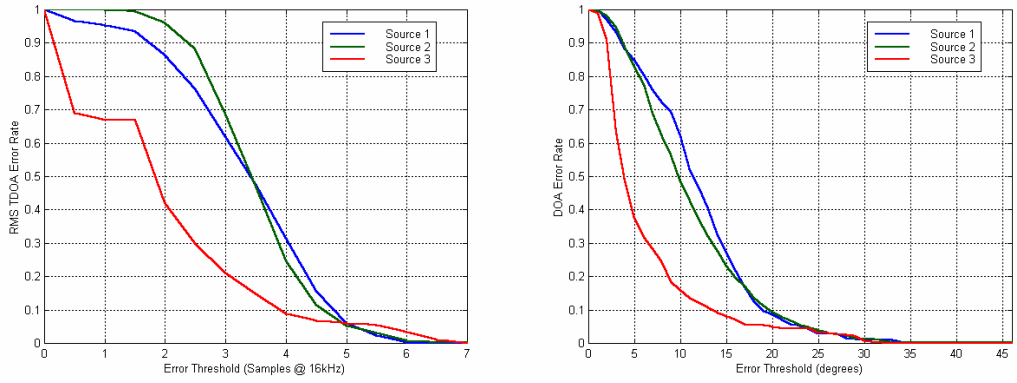

Figure 5.13 RMS error rates; TDOA error rate (left) and DOA error rate (right).

80 percent of the DOA estimates from sources 1 and 2 had errors greater than 5 degrees. At a distance of 6 meters, this angular error corresponds to about a 1-meter uncertainty in location. About 50 to 60 percent of these had errors greater than 10 degrees, which corresponds to about a 2-meter location uncertainty. Source 3 yielded a better performance; only about 40 percent had error greater than 5 degrees, and less than 20 percent were greater than 10 degrees.

The RMS TDOA error in Figure 5.13 (the left-hand plot) reflects the performance in DOA estimation; the higher errors in the TDOA estimates result in higher errors in the corresponding DOA estimates. Not surprisingly, the DOA estimates can only be as good as the TDOA estimates, which were better, on average, for this 8-element array than the triad array of Section 5.2.3. The increase in the number of microphones from the triad array to this 8-element array greatly increased the number TDOA pairs, from 3 to 28, and the average performance of the larger group was better than the smaller group. Hence, there are obviously some advantages to be gained by increasing the number of microphones used to estimate DOA. The question, which will be addressed in the following chapters, is "Can an increase in the number of microphones be exploited in a better way?" It seems that increasing the number of TDOA pairs simply increases the number of erroneous TDOA estimates and the averaging that occurs during the DOA fitting simply smoothes away some of the outliers. While pairwise techniques, such as GCC-PHAT, are attractive because of their computational simplicity and autonomy, there may be a significant increase in performance realized by combining the data from multiple microphones earlier in the DOA estimation process, and any consequential increase in computational costs may be justified.

5.3.2 Visualizing the RMS TDOA Error

To get an intuition for the RMS TDOA error, $E_{FAR}(\theta,\phi)$, it was computed over a short segment of speech from the recording of source 1 and plotted for nine successive, half-overlapping, 25-milliseconds blocks. This is illustrated by the series of images plotted in Figure 5.14. The white point in each image marks the true DOA. The dark (dark blue) color in the images represents the minima of the RMS error. At the top of this figure is a plot of the amplitude of the corresponding speech segment. Superimposed on this speech signal is a curve representing the average power of the signals from the array, with the scale of its vertical axis labeled on the right side of the graph. Each point along this power curve corresponds to the

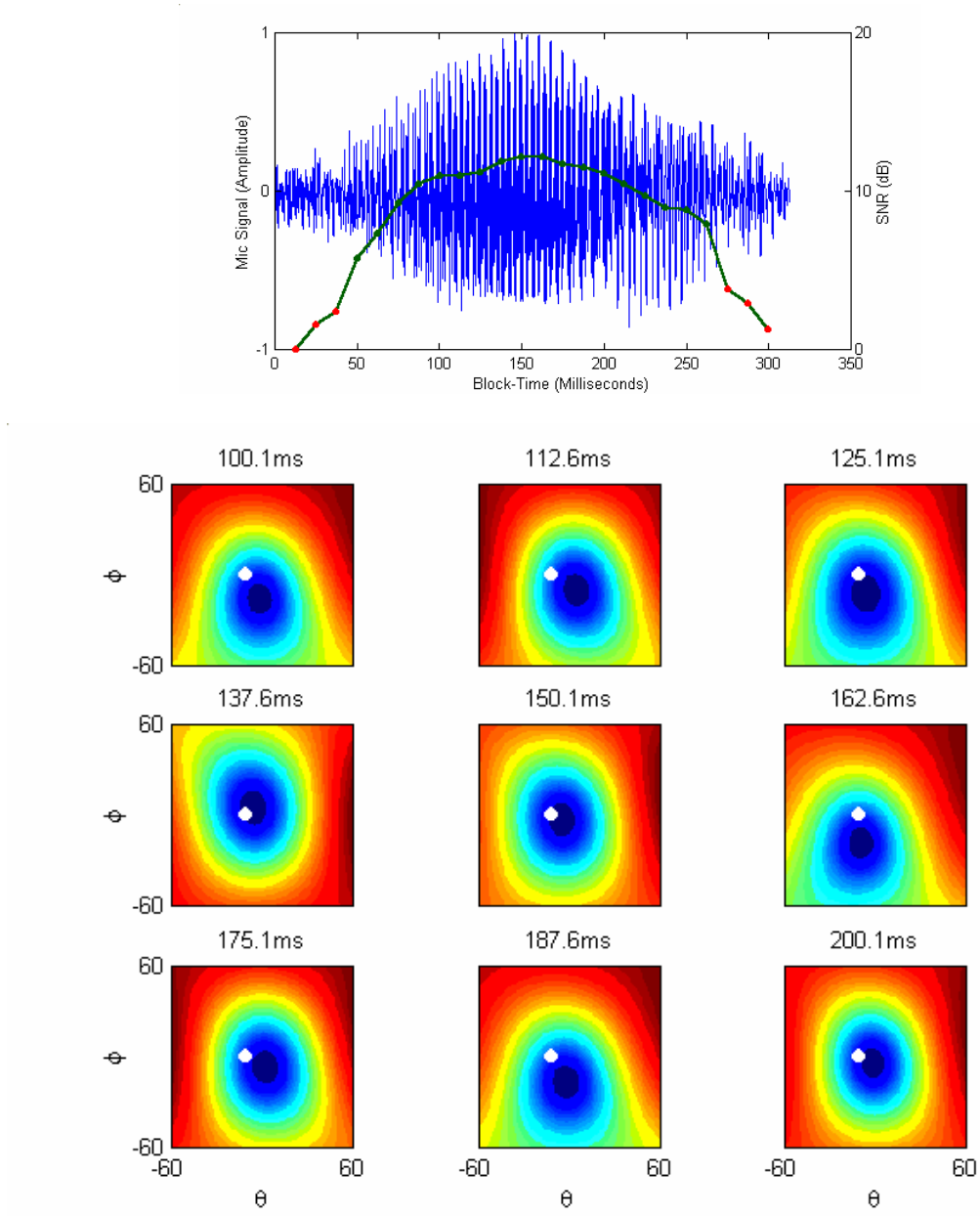

Figure 5.14 Speech segment (top) with nine frames of TDOA error surfaces (bottom).

average block SNR as described in Section 3.5. The three blocks at the beginning and the three blocks at the end of this speech segment (highlight by red points) were masked out during the TDOA estimation procedure. However, only the middle nine blocks were used to produce the images plots below.

As shown by the images of Figure 5.14, $E_{FAR}(\theta,\phi)$ is generally a smooth surface with a global minimum over the angular range of ± 60 degrees. However, the minima seem to "wobble" around the

actual source-location from one frame to the next. This wobbling effect is caused by erroneous TDOA estimates, and according to the results of Sections 5.1 and 5.2, these erroneous TDOA estimates are primarily caused by the mild reverberation (200 millisecond reverberation time) in the conference room.

5.3.3 GCC Time-averaging

It has been reported that GCC-based TDOA estimators perform well with long data segments. The DOAs of the speech sources were estimated again, using the same general GCC-PHAT procedure. However, this time Equation 4.16 was used to compute the GCC function. This equation is expressed in terms of the cross-spectrum of the pairwise microphone signals, which is given by Equation 4.14:

$$C_{lq,b}[k] \equiv \frac{1}{I} \sum_{i=b}^{b+I-1} X_{l,i}[k] X'_{q,i}[k]$$

The performance of GCC was evaluated for an increasing number of blocks used in the computation of

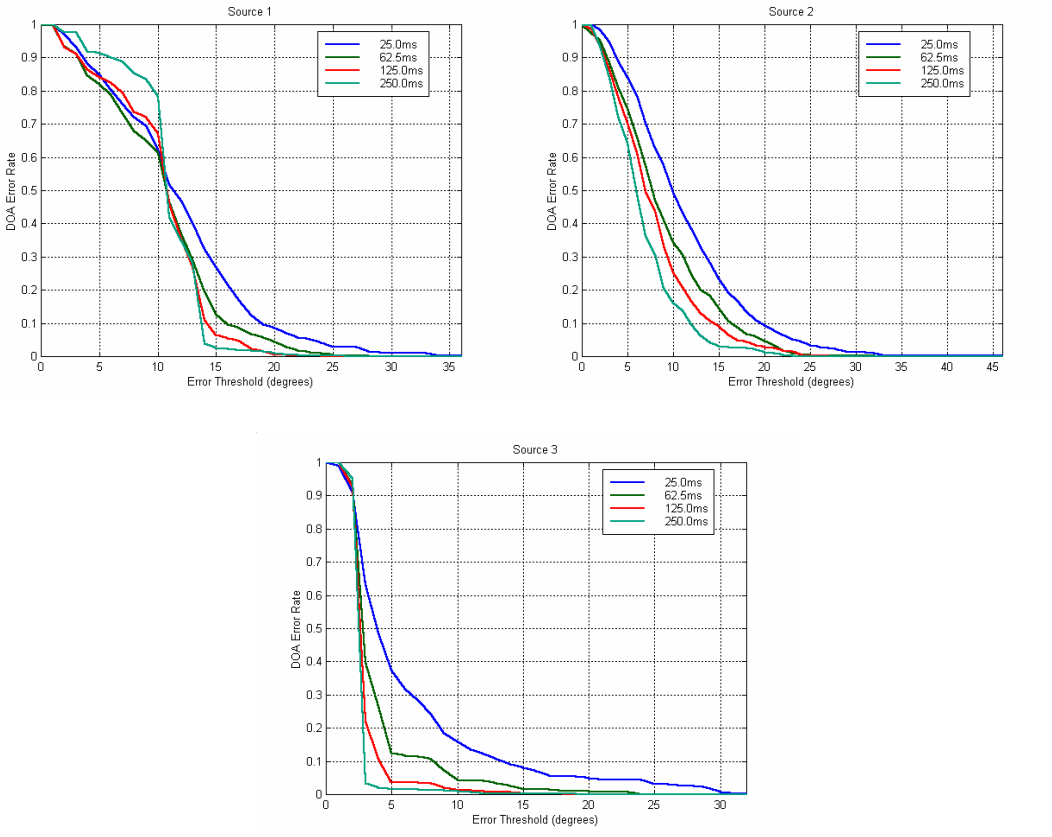

Figure 5.15 DOA error rates for various cross-spectrum accumulation times.

each cross-spectrum. That is, I in Equation 4.14 was increased and used to compute the cross spectra in Equation 4.16. This was done using I equal to 1, 4, 9 and 19. With a block size of 25 milliseconds and a 12.5-millisecond advance, this corresponds to the following block-averaging times: 25.0ms, 62.5ms, 125.0ms and 250.0ms.

As given by Equation 5.4, the RMS errors of the DOA estimates produced over the coarse of each speech recording were used to compile error rates for each source location. These are plotted in Figure 5.15. Notice how performance generally increases with the accumulation time of cross-spectra data. Hence, averaging does improve the TDOA estimation process. However, cross-spectra averaging requires long data segments to be effective. In this experiment, it took 10 times more time-data to improve the error rate for source 3, for example, from 80 percent to near 0 percent at a threshold of 3 degrees, which is about the maximum resolution for this array aperture.

With the desire to maintain the short 25-millisecond data blocks and the associated benefits, time averaging in not always a desirable means for improving performance. However, these results show that "more data is better". The underlying pairwise process incorporates the data from only two microphones. Rather than increasing the accumulation time, an increase in data may be achieved by incorporating the data from several microphones. That is, average over the spatial dimension instead of the temporal one. With the microphones sampling the wave field at different points in space, a combination of their data give a spatial accumulation rather than a temporal one. The following chapters explore this possibility.

6 The Steered Response Power (SRP)

Array signal processing techniques rely on the ability to *focus* on signals originating from a particular location or direction in space. Most of these techniques employ some type of *beamforming*, which generally includes any algorithm that exploits an array's sound-capture ability [59]. Beamforming, in the conventional sense, can be defined by a *filter-and-sum* process, which applies some temporal filters to the microphone signals before summing them to produce a single, focused signal. These filters are often adapted during the beamforming process to enhance the desired source signal while attenuating others. The simplest filters execute time shifts that have been matched to the source signal's propagation delays. This method is referred to as *delay-and-sum* beamforming; it delays the microphone signals so that all versions of the source signal are time-aligned before they are summed. The filters of more sophisticated filter-and-sum techniques usually apply this time alignment as well as other signal-enhancing processes.

Beamforming techniques have been applied to both source-signal capture and source localization. If the location of the source is known (and perhaps something about the nature of the source signal is known as well), then a beamformer can be focused on the source, and its output becomes an enhanced version (in some sense) of the inputs from the microphones. If the location of the source is not known, then a beamformer can be used to scan, or *steer*, over a predefined spatial region by adjusting its steering delays (and possibly its filters). The output of a beamformer, when used in this way, is known as the *steered response*. The steered response power (SRP) may peak under a variety of circumstances, but with favorable conditions, it is maximized when the steering delays match the propagation delays. By predicting the properties of the propagating waves, these steering delays can be mapped to a location, which should coincide with the location of the source.

Beamforming has been used extensively in speech-array applications for voice capture [38][37][23][41][97]. For this application, the filters applied by the filter-and-sum technique must not only suppress the background noise and contributions from unwanted sources, they must also do this in way that does not significantly distort the desired signal. However, when beamforming techniques are applied to source-localization, these filters need only boost the power of the desired source signal in the beamformer's output when the array is focused on it. This important distinction is exploited in this chapter where a new

type of filter is proposed for localization. These filters are derived from the phase transform (PHAT), which applies a magnitude-normalizing weighting function to the cross-spectrum of two microphone signals (see Section 4.1.2). This procedure produces a function that is useful for TDOA estimation but is obviously a distortion of the input (and source) signals. In the same way, beamformer filters can be designed to produce a steered response that is useful for source localization but not for voice-capture.

The phase transform was studied in previous chapters, where it was demonstrated that it is a suitable choice for TDOA estimation using speech sources. It was also shown to have limitations in reverberant environments, and it was hypothesized that incorporation of multiple microphone signals may improve performance of this commonly used pairwise technique. This chapter proposes the application of filters that makes the steered response power (SRP) equivalent to the sum of all possible pairwise phase transforms. The new technique, which has been dubbed "SRP-PHAT", exploits microphone redundancy by combining the microphone signals, rather than combining a multitude of TDOA estimates, to enhance the accuracy of location estimation. In the following chapter, an experiment with the conference-room data set of Section 3.2 demonstrates that this approach yields substantial improvements in performance over the RMS TDOA-error minimization technique, which was described in Section 4.4 and first employed in the experiments of Chapter 5.

6.1 Beamforming

Recall the microphone signal model of Equation 2.8:

$$x_m(t) = \frac{1}{r_m} s(t - \tau_m) * \gamma_m(\vec{d}^{(s)}, t) + \breve{v}_m(t)$$

This equations clearly shows that for an array of M microphones, a delayed and filtered version of the source signal, $s(t)$, exists in each microphone signal. By time-aligning the delayed versions of $s(t)$, the resulting signals can be summed together so that all copies add constructively while the uncorrelated noise signals present in $\breve{v}_m(t)$ cancel. Hence, the *delay-and-sum beamformer* can be defined as follows:

$$y(t, \Delta_1 \ldots \Delta_M) \equiv \sum_{m=1}^{M} x_m(t - \Delta_m) \tag{6.1}$$

$\Delta_1 \ldots \Delta_M$ are the M *steering delays*, which are chosen to focus or *steer* the array to the source's spatial location or direction. The copies of $s(t)$ in the microphone signals can be time-aligned by setting the steering delays equal to the negative values of the propagation delays plus some constant delay, τ_0:

$$\Delta_m = \tau_0 - \tau_m \qquad \text{for } m = 1 \ldots M \tag{6.2}$$

τ_0 defines the *phase center* of the array, and is usually set to the largest propagation delay making all the steering delays greater than or equal to zero. This makes all shifting operations causal, which is a requirement of any practical system. This also makes the steering delays relative to one microphone, and therefore they are equivalent to the TDOAs between each microphone and the reference microphone. This implies that knowledge of the TDOAs is sufficient for steering the beamformer.

The delay-and-sum beamformer output, as defined by Equation 6.1, can now be expressed in terms of the microphone signal model of Equation 2.8 and the steering delays of Equation 6.2:

$$y(t, \Delta_1 \ldots \Delta_M) = s(t - \tau_0) * \sum_{m=1}^{M} \frac{1}{r_m} \gamma_m(\vec{d}^{(s)}, t - \tau_0 + \tau_m) + \sum_{m=1}^{M} \frac{1}{r_m} \breve{v}_m(t - \tau_0 + \tau_m) \tag{6.3}$$

If the responses of the microphone channels, $\gamma_m(\vec{d}^{(s)}, t)$, are similar and approximate a bandpass filter, and the distances from the source to each microphone are nearly the same, then the output of the beamformer, as given by Equation 6.3, contains a band-limited version of $s(t)$ that has amplitude M times larger than any one microphone signal. However, the degree to which the noise signals are suppressed (or amplified) depends on the nature of the noise. The model of Equation 2.8, as discussed in Section 2.5, allows $\breve{v}_m(t)$ to include reverberation as well as background noise. Since reverberation consists of multipath versions of the source signal, $\breve{v}_m(t)$ can be highly correlated with the $s(t)$. Hence, delay-and-sum beamforming is not always effective at suppressing this noise. Other adaptive beamforming methods extend the delay-and-sum concept to the more general *filter-and-sum* approach, which applies adaptive filtering to the microphone signals before, and possible after, they pass through the delay-and-sum beamformer.

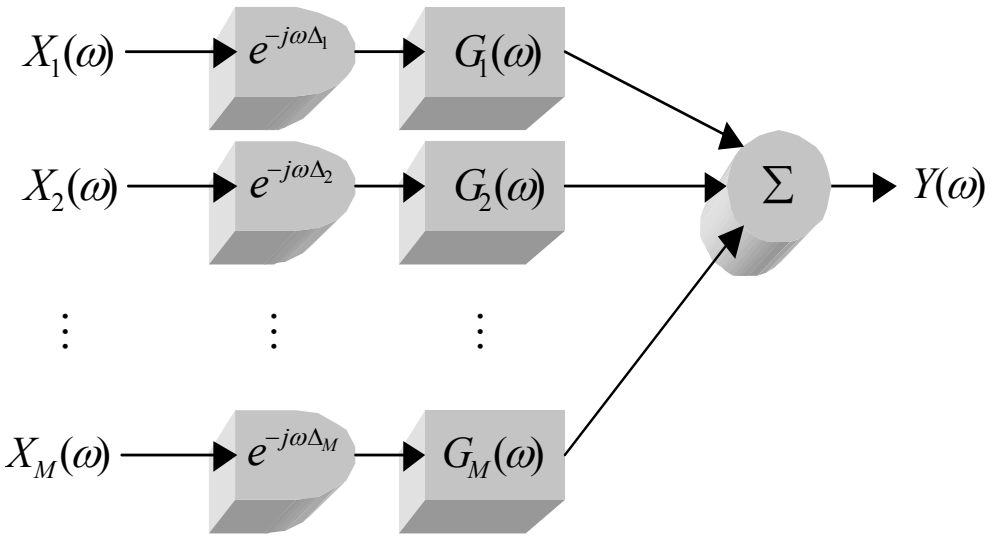

Figure 6.1 The structure of a filter-and-sum beamformer.

The output of an *M*-element, filter-and-sum beamformer can be defined in the frequency domain as follows:

$$Y(\omega, \Delta_1 \ldots \Delta_M) \equiv \sum_{m=1}^{M} G_m(\omega) X_m(\omega) e^{-j\omega\Delta_m} \tag{6.4}$$

In the above equation, $X_1(\omega) \ldots X_M(\omega)$ are the Fourier Transforms of the microphone signals, and $G_1(\omega) \ldots G_M(\omega)$ are the Fourier transforms of some temporal filters. This process is illustrated by the diagram of Figure 6.1. The microphone signals are delayed by the steering delays, as they would when passing though a simple delay-and-sum beamformer. However, the filter-and-sum beamformer applies additional temporal filters to the microphone signal before summing the result to form the output. Choosing the appropriate filters depends on a number of factors, including the nature of the source signal and the type of noise present.

6.2 The Steered Response

The *steered response* is generally a function of *M* steering delays, $\Delta_1 \ldots \Delta_M$. The steering delays are used to aim a beamformer, which acoustically focuses the array to a particular position or direction in space.

The steered response is obtained by sweeping the focus of the beamformer. When the focus corresponds to the location of a sound source, the power of the steered response reaches a maximum, although a variety of circumstances may causes it to peak when focused on other locations as well. The steered response power (SRP) can be expressed as the output power of a filter-and-sum beamformer and is defined as follows[4]:

$$P(\Delta_1 \ldots \Delta_M) \equiv \int_{-\infty}^{+\infty} Y(\omega, \Delta_1 \ldots \Delta_M) Y'(\omega, \Delta_1 \ldots \Delta_M) d\omega \tag{6.5}$$

$Y(\omega, \Delta_1 \ldots \Delta_M)$ is the output of the filter-and-sum beamformer, as defined by Equation 6.4, and $Y'(\omega, \Delta_1 \ldots \Delta_M)$ its complex conjugate. The steering delays, $\hat{\Delta}_1 \ldots \hat{\Delta}_M$, that maximize Equation 6.5 correspond to the TDOA estimates among microphones. This is similar to the behavior of the generalized cross-correlation for two microphones (m=2); it peaks when the time delay, τ, corresponds to the TDOA of the sound waves between two microphones. The TDOA estimate between the l-th and q-th microphone signals is the difference between the l-th and q-th steering delays from the set that maximizes the steered response power:

$$\hat{\tau}_{lq} \equiv \hat{\Delta}_l - \hat{\Delta}_q$$

Recall, from Section 4.4 and the experiment of Section 5.3, that minimization of the RMS TDOA error over a predefined set of spatial points leads to the localization of the source. This was achieved by applying the simple acoustic conditions from Section 2.1 to the propagation of the sound waves from the source to each microphone. Accordingly, the candidate TDOA vectors were computed using the assumed propagation delays. These propagation delays are equivalent to the steering delays used by the filter-and-sum beamformer of Equation 6.4. Hence, using a similar technique, the steered response power can be maximized over a predefined set of spatial points from which the steering delays can be computed and used to focus the beamformer. The corresponding spatial point where the beamformer is focused when its output power peaks (global maximum), gives the estimate of the source's location. Hence, denoting $\vec{d}$ as

[4] Technically speaking, the power of the beamformer output is, in some sense, proportional to this integral. There is a dependence on the filters $G_1(\omega) \ldots G_M(\omega)$.

the candidate location, the steered response power (SRP), which is a function of $\vec{d}$, is equivalent to Equation 6.5 evaluated as follows:

$$P(\vec{d}) = P(\Delta_1 \ldots \Delta_M) \quad \text{for } \Delta_m = \tau_0 - \frac{\left|\vec{d}_m - \vec{d}\right|}{c} \tag{6.6}$$

As presented in Section 4.4 for GCC, under far-field conditions, the propagation delays can be expressed in terms of the assumed direction of propagation, $\vec{\zeta}_o$, as follows:

$$\Delta_m = \frac{-\vec{\zeta}_o \cdot \vec{d}_m}{c} \tag{6.7}$$

The look vector, which points in the direction opposite the direction of propagation can be defined in terms of the azimuth and elevation angles, θ and ϕ, as follows (See Figure 2.3):

$$-\vec{\zeta}_o \equiv \begin{bmatrix} \cos\phi \sin\theta \\ \cos\phi \cos\theta \\ \sin\phi \end{bmatrix} \tag{6.8}$$

These angles, θ and ϕ, define the assumed direction of arrival, or look direction, relative to the array's local origin, o. By evaluating Equation 6.5 using the steering delays of Equation 6.7, the far-field steered response power can be defined in terms of θ and ϕ as follows:

$$P_{FAR}(\theta,\phi) \equiv P(\Delta_1 \ldots \Delta_M) \quad \text{for } \Delta_m = \frac{-\vec{\zeta}_o(\theta,\phi) \cdot \vec{d}_m}{c} \tag{6.9}$$

6.3 SRP in Terms of GCC

The steered response power (SRP) inherently averages the data from multiple microphones. This section shows that the SRP of an *M*-element array is actually equivalent to the sum of the generalized cross-correlations (GCCs) of all possible *M-choose-2*, i.e. $\binom{M}{2}$, microphone pairings. This means that the SRP of a 2-element array is equivalent to the GCC of those two microphones. Hence, as the number of microphones is increased, SRP naturally extends the GCC method from a pairwise to a multi-microphone technique.

By combining Equation 6.4 and 6.5, the steered response power of the filter-and-sum beamformer can be expressed as follows:

$$P(\Delta_1 \ldots \Delta_M) = \int_{-\infty}^{\infty} \left(\sum_{l=1}^{M} G_l(\omega) X_l(\omega) e^{-j\omega\Delta_l} \right) \left(\sum_{q=1}^{M} G'_q(\omega) X'_q(\omega) e^{j\omega\Delta_q} \right) d\omega$$

Expanding the multiplication of the summation terms yields:

$$P(\Delta_1 \ldots \Delta_M) = \int_{-\infty}^{\infty} \sum_{l=1}^{M} \sum_{q=1}^{M} \left(G_l(\omega) X_l(\omega) e^{-j\omega\Delta_l} \right) \left(G'_q(\omega) X'_q(\omega) e^{j\omega\Delta_q} \right) d\omega$$

Rearranging the terms inside the double summation gives:

$$P(\Delta_1 \ldots \Delta_M) = \int_{-\infty}^{\infty} \sum_{l=1}^{M} \sum_{q=1}^{M} G_l(\omega) G'_q(\omega)\, X_l(\omega) X'_q(\omega)\, e^{j\omega(\Delta_q - \Delta_l)}\, d\omega$$

The integral converges since the microphone signals and the filters have finite energy (in practice), and therefore it can be interchanged with the summations:

$$P(\Delta_1 \ldots \Delta_M) = \sum_{l=1}^{M} \sum_{q=1}^{M} \int_{-\infty}^{\infty} G_l(\omega) G'_q(\omega)\, X_l(\omega) X'_q(\omega)\, e^{j\omega(\Delta_q - \Delta_l)}\, d\omega$$

Defining the weighting function as follows:

$$\Psi_{lq}(\omega) \equiv G_l(\omega) G'_q(\omega) \tag{6.10}$$

By replacing the filters by this weighing function, the steered response power becomes:

$$P(\Delta_1 \ldots \Delta_M) = \sum_{l=1}^{M} \sum_{q=1}^{M} \int_{-\infty}^{\infty} \Psi_{lq}(\omega)\, X_l(\omega) X'_q(\omega)\, e^{j\omega(\Delta_q - \Delta_l)}\, d\omega$$

The generalized cross-correlation of two microphone signals, indexed by m=1,2, has been defined by Equation 4.9 as follows:

$$R_{12}(\tau) \equiv \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{12}(\omega) X_1(\omega) X'_2(\omega) e^{j\omega\tau} d\omega$$

Applying this to the pair composed of the microphones indexed by l and q, and denoting the time lag for this pair as Δ_{lq}, the generalized cross-correlation of the l-th and q-th microphone signals can be expressed as follows:

$$R_{lq}(\Delta_{lq}) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Psi_{lq}(\omega)X_l(\omega)X'_q(\omega)e^{j\omega\Delta_{lq}}d\omega$$

With $\Delta_{lq} \equiv \Delta_l - \Delta_q$, the steered response power can now be expressed as a function of the generalized cross-correlations:

$$P(\Delta_1 \ldots \Delta_M) = 2\pi\sum_{l=1}^{M}\sum_{q=1}^{M}R_{lq}(\Delta_q - \Delta_l) \tag{6.11}$$

This is the sum of all possible pairwise GCC crossings, which are time-shifted by the differences in the steering delays. Included in this summation is the sum of the *M* autocorrelations, which is the generalized cross-correlation evaluated at $\Delta_{lq} = 0$. This contributes only a DC offset to the steered response power since it is independent of the steering delays:

$$\sum_{m=1}^{M}R_{mm}(0) = \frac{1}{2\pi}\sum_{m=1}^{M}\int_{-\infty}^{\infty}\left|G_m(\omega)\right|^2\left|X_m(\omega)\right|^2 d\omega$$

Equation 6.11 also includes both permutations of each crossing. However, because the associated difference in steering delays is opposite for each permutation, summing a GCC combination plus its "time-flipped" permutation is equivalent to scaling one permutation by two:

$$R_{lq}(\Delta_q - \Delta_l) = R_{ql}(\Delta_l - \Delta_q) \quad \Rightarrow \quad R_{lq}(\Delta_q - \Delta_l) + R_{ql}(\Delta_l - \Delta_q) = 2R_{lq}(\Delta_q - \Delta_l)$$

Therefore, Equation 6.11 really is the summation of all possible GCC combinations, within a scale factor and constant offset. Hence, it has been shown that SRP is equivalent to summing all possible GCC combinations.

6.4 Combining the Phase Transform and Steered Response Power: SRP-PHAT

By the experiments in Chapter 5, it was demonstrated that the phase transform was an effective weighting for GCC when applied to speech signals. However, the same experiments underscored the shortcomings of pairwise GCC. In Section 6.3, the relationship between the steered response power (SRP) and pairwise GCC was presented. Using this relationship and the appropriate filters, the sum of all possible pairwise GCC-PHAT combinations can be formed using the filter-and-sum structure, illustrated in Figure 6.1.

These filters lead to the combination of the phase transform and the steered response power, which has been dubbed "SRP-PHAT".

Recall that the generalized cross-correlation (GCC) of the l-th and q-th microphone signals has been expressed as follows:

$$R_{lq}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Psi_{lq}(\omega)X_l(\omega)X'_q(\omega)e^{j\omega\tau}d\omega$$

From Section 4.1.2, the phase transform weighting function has been defined by Equation 4.10 as:

$$\Psi_{lq}(\omega) \equiv \frac{1}{\left|X_l(\omega)X'_q(\omega)\right|}$$

In Section 6.3, it was shown that the steered response power is equivalent to the sum of all possible GCC pairings in an *M*-element array. If the filters of the filter-and-sum beamformer used to compute the steered response are chosen appropriately, then the steered response power becomes the sum of all possible GCC-PHAT pairings. Recall, from Equation 6.10, that the relationship between these filters and the GCC weighting function is:

$$\Psi_{lq}(\omega) = G_l(\omega)G'_q(\omega)$$

Combining this equation with the definition of the phase transform weighting of Equation 4.10 yields the following:

$$G_l(\omega)G'_q(\omega) = \frac{1}{\left|X_l(\omega)X'_q(\omega)\right|} \tag{6.12}$$

Noting that $\frac{1}{|X_l(\omega)||X'_q(\omega)|} = \frac{1}{|X_l(\omega)|}\frac{1}{|X'_q(\omega)|}$ and $\frac{1}{\left|X'_q(\omega)\right|} = \frac{1}{\left|X_q(\omega)\right|}$, Equation 6.12 can be expressed as:

$$G_l(\omega)G'_q(\omega) = \frac{1}{\left|X_l(\omega)\right|}\frac{1}{\left|X_q(\omega)\right|}$$

This equation holds for the following choice of filters:

$$G_l(\omega) \equiv \frac{1}{\left|X_l(\omega)\right|} \qquad G_q(\omega) \equiv \frac{1}{\left|X_q(\omega)\right|}$$

These are the desired SRP-PHAT filters. They can be defined for all M microphones of the array as follows:

$$G_m(\omega) \equiv \frac{1}{|X_m(\omega)|} \qquad \text{for } m = 1 \ldots M \tag{6.13}$$

Just as with the phase transform, these filters whiten the microphone signals. This whitening technique effectively sharpens the peaks in the phase transform, and therefore should have the same effect on the steered response power. Unlike the typical, narrow-band signals found in the radar and sonar applications where SRP is widely used, the spectral content of speech signals fluctuates and is unknown. By whitening the microphone signals, SRP can be used just as effectively in speech-array applications.

6.5 Implementation of SRP

As presented in Section 4.2 for GCC, SRP can be implemented using a block-processing scheme that employs short-time DFTs as estimates of the microphone signals' spectra. Using the method described in Section 3.4, the array signals are segmented in small blocks and the steered response is computed for each block. The block DFTs have been denoted by $X_{m,b}[k]$ where m is microphone index and b is the block index. Equation 6.4 defines the steered response in terms of the continuous temporal frequency variable, ω, and the continuous steering delays, $\Delta_1 \ldots \Delta_M$. Replacing the Fourier transforms in this equation with their respective block DFTs, the steered response of block b can be defined as follows:

$$\widetilde{Y}_b[k, \Delta_1 \ldots \Delta_M) \equiv \sum_{m=1}^{M} G_{m,b}[k] X_{m,b}[k] e^{-j\omega \Delta_m} \tag{6.14}$$

$G_{m,b}[k]$ is the DFT of a discrete-time filter for microphone m, which is updated at every block, b, in general. $\widetilde{Y}_b[k, \Delta_1 \ldots \Delta_M)$ is a function of discrete temporal frequency, indexed by k, and a continuous set of m steering delays. By taking the summation over K discrete frequencies, the steered response power is obtained:

$$\widetilde{P}_b(\Delta_1 \ldots \Delta_M) \equiv \sum_{k=1}^{K} \widetilde{Y}_b[k, \Delta_1 \ldots \Delta_M) \widetilde{Y}_b'[k, \Delta_1 \ldots \Delta_M) \tag{6.15}$$

Although the steering delays are continuous, Equation 6.15 is sampled in practice, based on a predefined set of spatial locations (or directions).

Implementation of SRP-PHAT requires a discrete version of the filters defined by Equation 6.13. These discrete filters, which depend on the microphone signals at each block, are defined as follows:

$$G_{m,b}[k] \equiv \frac{1}{\left|X_{m,b}[k]\right|} \qquad \text{for } m = 1 \ldots M$$

By substituting these filters into Equation 6.14, the PHAT steered response can be expressed as follows:

$$\widetilde{Y}_b^{PHAT}[k, \Delta_1 \ldots \Delta_M) \equiv \sum_{m=1}^{M} \frac{X_{m,b}[k]}{\left|X_{m,b}[k]\right|} e^{-j\omega \Delta_m}$$

By applying Equation 6.15, the PHAT steered response power, or SRP-PHAT, can be obtained:

$$\widetilde{P}_b^{PHAT}(\Delta_1 \ldots \Delta_M) \equiv \sum_{k=1}^{K} \widetilde{Y}_b^{PHAT}[k, \Delta_1 \ldots \Delta_M) \widetilde{Y}_b'^{PHAT}[k, \Delta_1 \ldots \Delta_M)$$

6.6 Time Averaging versus Spatial Averaging

Recall, from Section 6.3, that the steered response power is the sum of all possible pairwise GCC crossings, which includes all possible GCC combinations, as expressed by Equation 6.11:

$$P(\Delta_1 \ldots \Delta_M) = 2\pi \sum_{l=1}^{M} \sum_{q=1}^{M} R_{lq}(\Delta_q - \Delta_l)$$

From this, it is easy to see that SRP-PHAT relies on diversity among microphone signals and averages cross-correlations to improve performance under adverse conditions. This is equivalent to spatial averaging since each microphone samples the sound field at a different point in space. Recall, from Section 4.2, that an alternate implementation of GCC-PHAT averages cross-spectra from successive data blocks in time, which can then be used to compute the generalized cross-correlation. Hence, GCC-PHAT also averages multiple cross-correlations, but relies on temporal diversity in the microphone signals and averages over time to improve performance.

There are tradeoffs between these two methods of averaging. SRP-PHAT requires more microphones than GCC-PHAT, while GCC-PHAT requires more time data. When emphasis is placed on producing accurate DOA estimates every 20 to 30 milliseconds, as is necessary to track the dynamic conditions in speech-array applications, the amount of time averaging performed during each analysis must be minimized. With such a constraint on the amount of time data available for each analysis, GCC-PHAT

is unable to perform adequately in reverberant environments, as the experiments of Chapter 5 demonstrate. The cost of adding microphones to the location-estimator system seems to be a practical means of improving performance while keeping the analysis period short. In many systems, such as the *Megamike* and the *HMA* (See Section 3.1), microphones are abundant, and instead of employing them to compute a multitude of generalized cross-correlations, it may be more effective to use them to compute one, or a few, steered responses. This will be examined, through some experiments, in the following chapter.

7 Experimental Performance Comparisons of SRP, SRP-PHAT and GCC-PHAT

A series of experiments were performed to evaluate and compare the performance of three different source locators: SRP, SRP-PHAT and GCC-PHAT. As described in Chapter 6, SRP employs the steered response of a delay-and-sum beamformer to localize the source, while SRP-PHAT uses a filter-and-sum beamformer with the phase transform (PHAT) filters introduced in Section 6.4. GCC-PHAT was used in the experiment of Section 5.3 to localize the source by minimization of the RMS TDOA. Using these three techniques and three different microphone arrays, performance and accuracy has been evaluated using error-rate plots of either DOA estimates or 3D Cartesian estimates of the source location. In all these experiments, 25-millisecond data blocks were used to emphasize the importance of fast and accurate source localization.

The first two experiments in this chapter used the conference-room data set, which was recorded in a high-SNR and mildly reverberant environment (See Chapter 3). The arrays used in these experiments were planar: the 8-element array from Section 5.3, which was a subset of the conference room array, and a 15-element array, which was composed of all 15 microphones connected to the Megamike during the conference-room recordings. The sources in these experiments were in the far field, and therefore, DOA estimation was performed. The third experiment used recordings made by the *Huge Microphone Array* (HMA) [83][84][85][86][87]. These recordings were taken from 128 microphones spread over a large aperture, which encompassed the talker(s) on three sides. The environment where the HMA lives is considerably more noisy and reverberant than the conference room where the data set of Section 3.2 was recorded. The HMA recordings were used to examine the performance of SRP, SRP-PHAT and GCC-PHAT when applied to a large aperture array. With such a large aperture, it was also possible to demonstrate the inherent ability of beam-steering techniques, such as SRP and SRP-PHAT, to localize multiple, simultaneously active talkers.

7.1 Experiment #1: DOA Estimation with an 8-Element Array

The same 8-element sub-array from the conference-room data set used in the GCC experiment of Section 5.3 was used in this experiment. As illustrated by Figure 7.1 (taken from Figure 5.12), the microphones are randomly positioned within a 33 by 36 centimeter rectangle. It has been assumed, as it was in Section 5.3, that all three sources lie in the far field of this sub-array. Using the steered response power (SRP), which was described in Chapter 6, the DOAs of these three speech sources were estimated. The performance of SRP was compared to that of the GCC DOA estimation procedure used in Section 5.3.

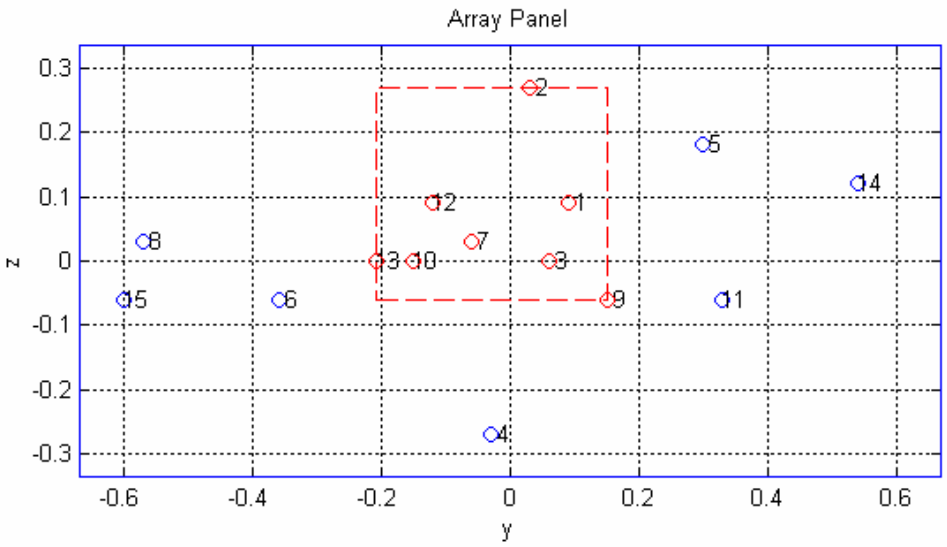

Figure 7.1 The 15-element array used to take the conference-room data set. The highlighted microphones form an 8-element array, which was used in the last GCC DOA experiment.

Again, the same block parameters were used: Hanning windowed, 25-millseconds blocks with a 12.5-millisecond block advance. The same SNR mask was used to reject the estimates derived from low-SNR blocks. The frequency range used to compute both the steered responses and the generalized cross-correlations was again 300Hz to 8kHz. Equations 6.14 and 6.15 were used to compute the steered responses of each block with steering delays as defined by the far-field Equations 6.8 and 6.9. These responses were computed over a range of -60 to +60 degrees for both azimuth and elevation on a grid of 0.1 degrees. These DOA search parameters were the same as those used by the GCC technique in the experiment of Section 5.3.

7.1.1 Performance Comparison

Error rates of the DOA error, $E_{DOA}(\hat{\theta},\hat{\phi})$, as defined in Section 5.3.1, were computed for the three speech sources in the conference-room data set. Each source-location was estimated using three different techniques: SRP, GCC-PHAT, and SRP-PHAT. The results are shown in Figure 7.2. Notice how SRP-PHAT consistently outperforms the other two methods. There is a tremendous improvement between the delay-and-sum steered response power (SRP) and the filter-and-sum steered response using the PHAT filters (SRP-PHAT). It is clear from the error rates that SRP-PHAT is greatly superior to GCC-PHAT in the low-noise and mildly reverberant conditions of the conference-room data set. SRP-PHAT's accuracy is nearly the same for all source locations, including the most distant, source 1. In contrast, GCC-PHAT's performance was highly dependent on source location. For example, 60 percent of the estimates from

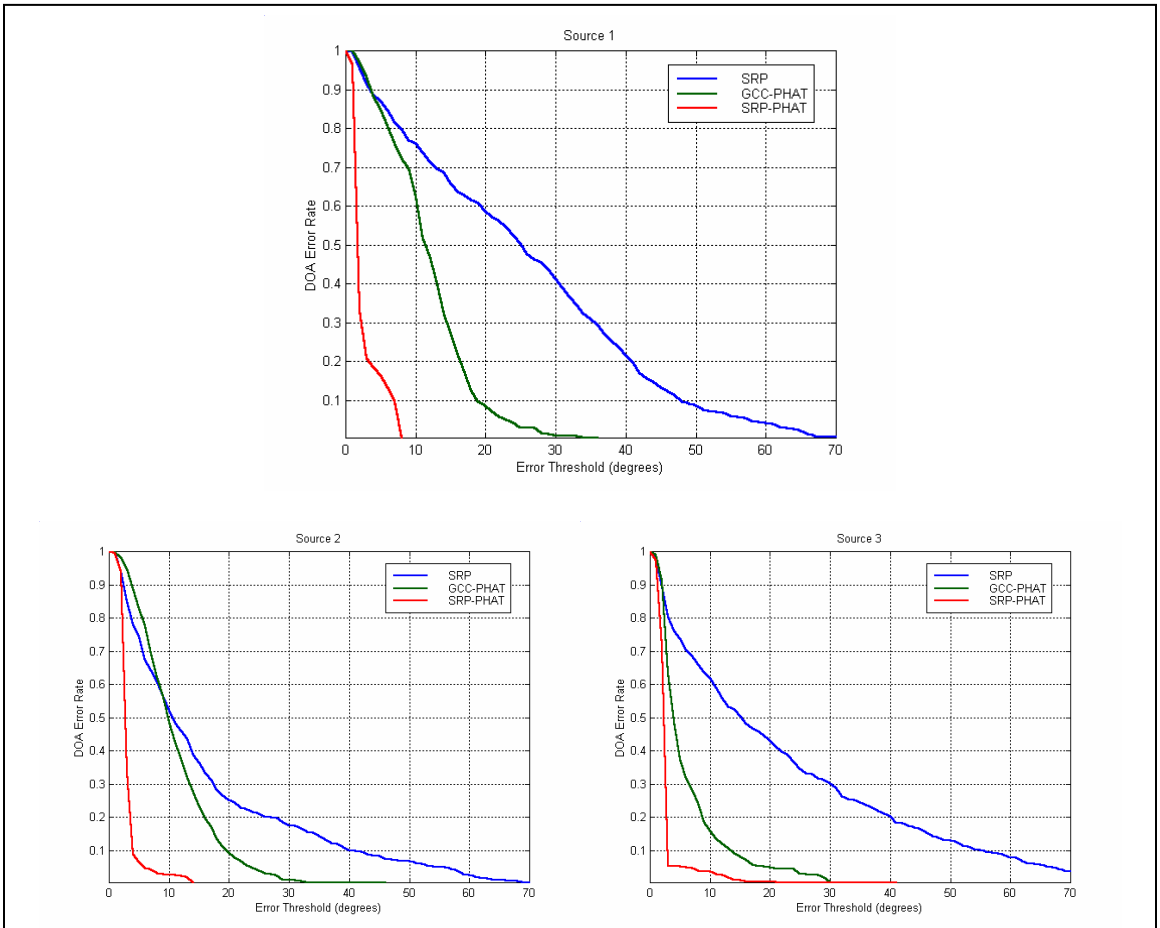

Figure 7.2 DOA error rates for three different sources.

source 1 had error greater than 10 degrees, while 50 percent from source 2 and 15 percent from source 3 had error greater than 10 degrees. Nearly 100 percent of all the estimates produced by SRP-PHAT had error less than 10 degrees. About 90 percent of the estimates from source 2 and source 3, and 80 percent from source 1, had error less than 4 degrees.

7.1.2 Visualizing the Steered Response Power

The steered responses of SRP and SRP-PHAT were computed and plotted for nine 25-millisecond blocks from a short segment of speech. The plot in Figure 7.3 is the amplitude signal from microphone 1 over the

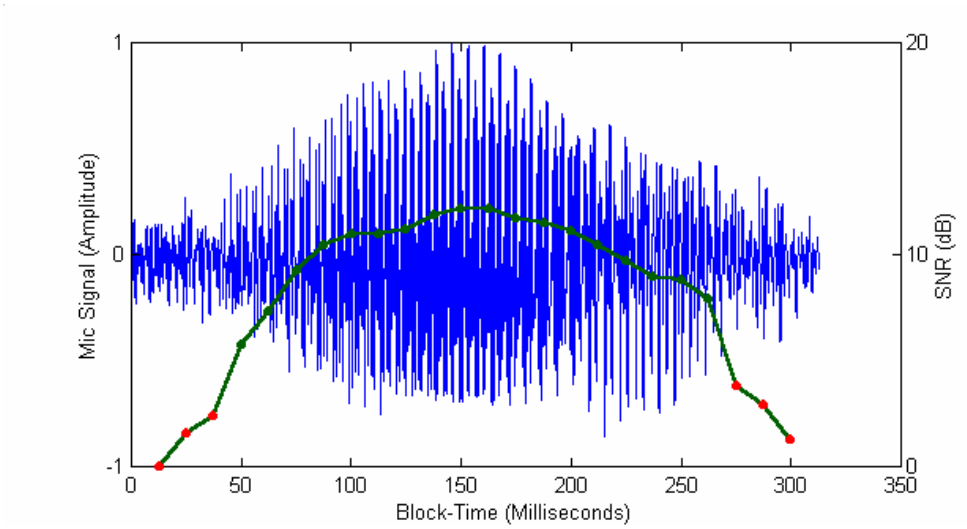

Figure 7.3 The speech segment used to compute the steered responses of Figures 7.4 and 7.5. The right-hand axis corresponds to the block-power curve plotted over the speech signal.

9-block recording interval. Also in this plot is the average array-signal power for each block, with the scale of it vertical axis labeled on the right-hand side of the plot. The points along this power curve mark the centers of the blocks. The block at the center of this segment (150.1 ms), plus the four blocks on either side of the center block, was used to produce the series of steered responses. This recording was speech from the source-location 1, and the segment in Figure 7.3 is the letter "R", spoken as in "*Are* we there yet?" This was the same segment used to visualize the RMS TDOA errors in Figure 5.14.

The steered response power for this far-field array is a function of azimuth and elevation. A series of 2D plots is shown in Figure 7.4, which represent the steered response of the delay-and-sum beamformer

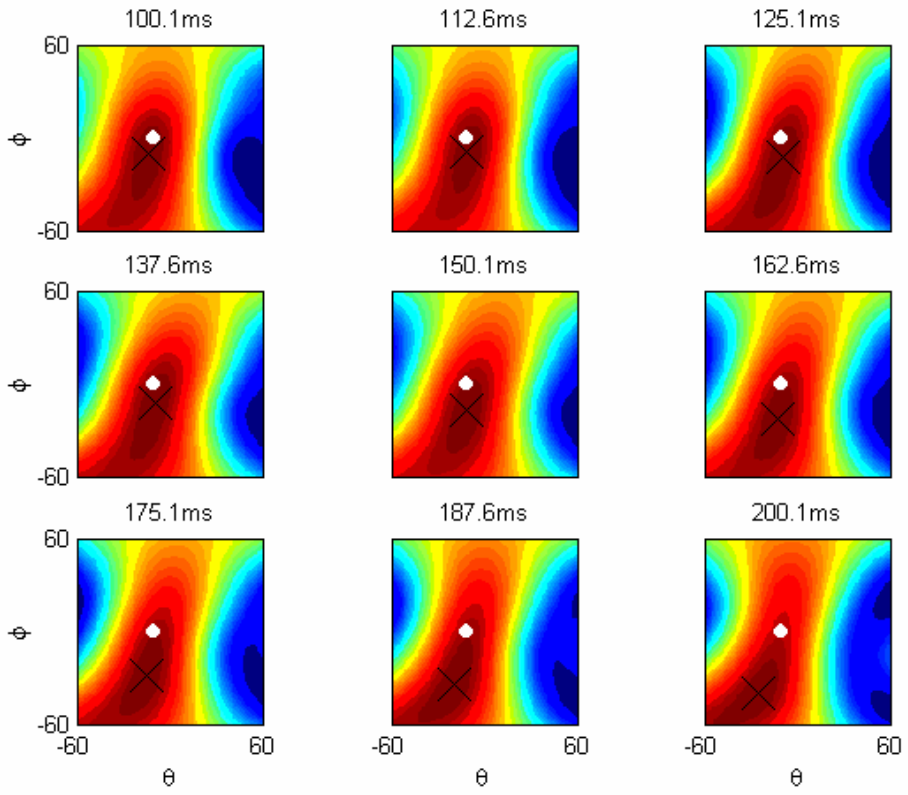

Figure 7.4 Steered responses of the delay-and-sum beamformer over nine, 25-millisecond blocks.

over 9 blocks of array data. This illustration is similar to the plots of the RMS TDOA error of Figure 5.14, which are also a function of azimuth and elevation. In this case, however, the 2D functions of the steered response power reach a maximum at the corresponding DOA, while the RMS TDOA error reaches a minimum at this point.

Notice how the maximum value in each image of Figure 7.4, which is marked an "X", occurs at points distant from the actual DOA, which is marked by a white dot. The main beam of the delay-and-sum beamformer is broad, and it fluctuates considerably over the duration of the speech segment. This accounts for the poor performance seen in the error rate plots of Figure 7.2 for SRP.

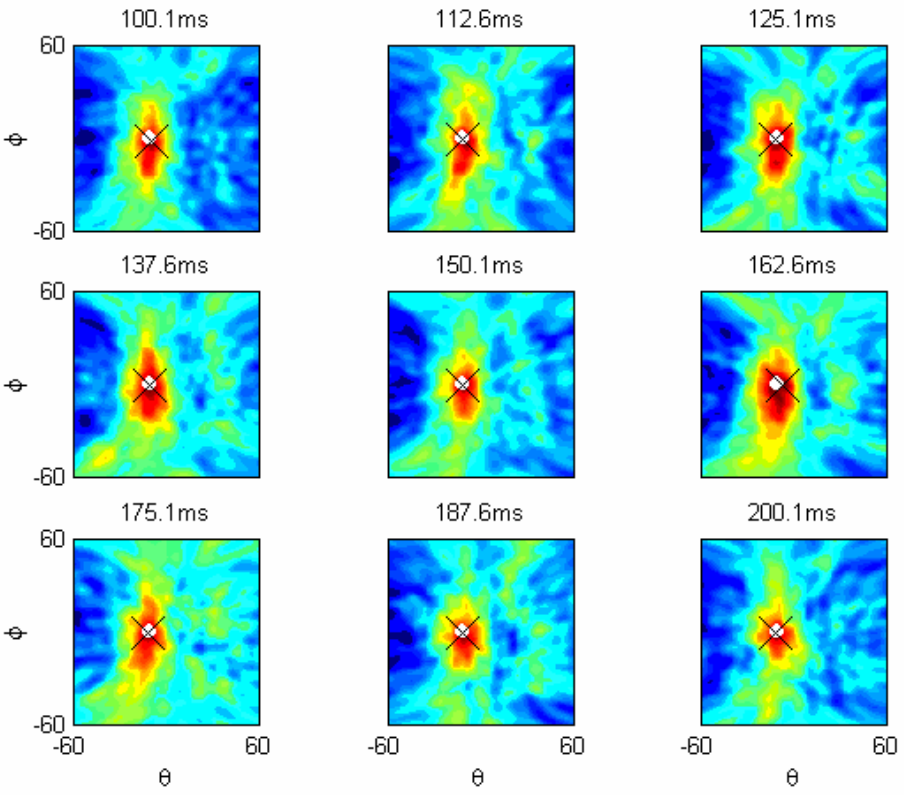

Figure 7.5 Steered responses over nine, 25-millisecond blocks using SRP-PHAT.

Another series of images plots was produced using the steered response of SRP-PHAT and the same segment of speech as Figure 7.4. The SRP-PHAT responses are shown in Figure 7.5. In contrast to the SRP responses, the peaks of SRP-PHAT are very close to the actual DOA, and the main beam of the PHAT beamformer is sharp and consistent over each block. Hence, the PHAT filters, when applied to the filter-and-sum beamformer, yield a steered response that is superior to that of the delay-and-sum beamformer, and this is reflected in the error rates of the DOA estimates shown in Figure 7.2.

7.2 Experiment #2: DOA Estimation with a 15-Element Array

The experiment of Section 7.1 was repeated, using the same parameters and procedure, with all 15 microphones of the conference room array, which is illustrated by Figure 7.1. This experiment was performed for speech source 1 only, which was about 5.5 meters from the array (See Figure 3.5). Since the aperture of this 15-element array is 120 by 60 centimeters, only source 1 could be safely categorized as a far-field source. The closer sources (2 and 3) were in the "gray" area; the array might have a large enough aperture to estimate range, but these estimates would not be very accurate. While the range of source 1 was only about four-and-half times the larger dimension of the array's aperture, it was sufficiently distant to allow the array's depth of focus to be set to infinity and still yield accurate estimates of the source's DOA.

With 15 microphones, 105 GCC pairs were formed to produce TDOA estimates of the source, which in turn were used to estimate DOA by minimization of the RMS TDOA error over azimuth and

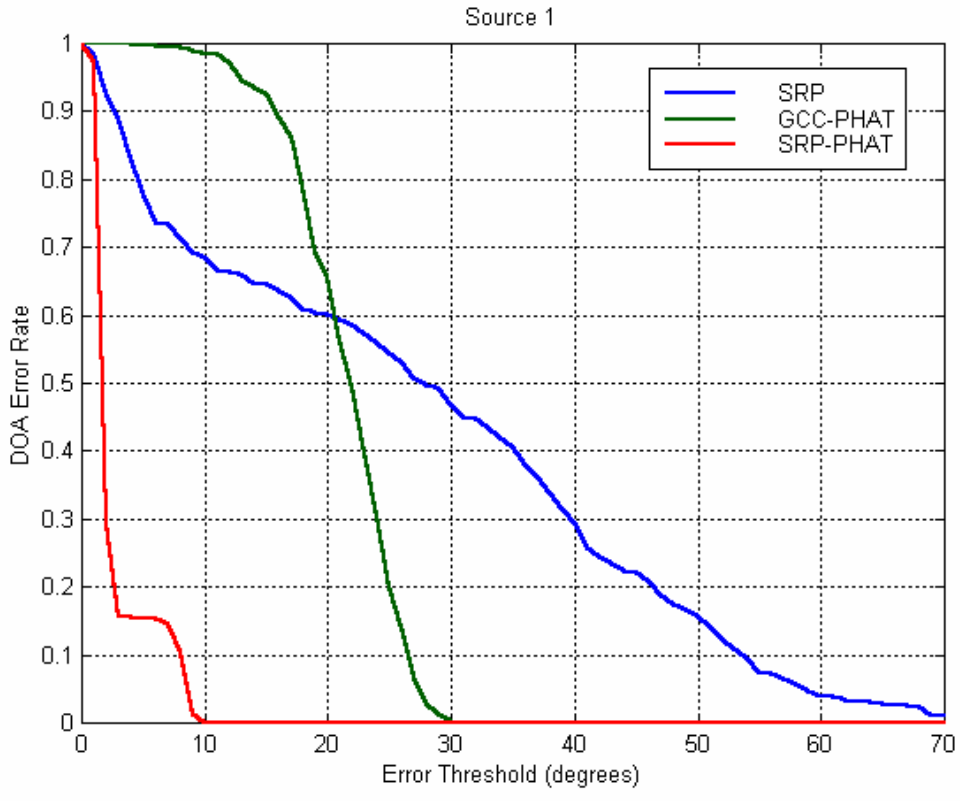

Figure 7.6 DOA error rates for the 15-element planar array and the recording of speech source 1.

elevation. Again the array recording was segmented into half-overlapping, 25-millisecond blocks, and an SNR mask was applied, which passed 313 out of 399 blocks for each technique.

7.2.1 Performance Comparison

The performance of SRP, GCC-PHAT and SRP-PHAT were compared using the error rates of the DOA error, $E_{DOA}(\hat{\theta},\hat{\phi})$, which are plotted in Figure 7.6. While the performances of SRP and SRP-PHAT were nearly the same as with the 8-element array, GCC-PHAT's performance is considerably worse. As the error rate for source 1 in Figure 7.2 shows, about 30 percent of the estimates produced by the 8-element array had error less than 10 degrees, and 10 percent had error greater than 20 degrees. The 15-element array produced nearly zero estimates with error less than 10 degrees, and about 70 percent had error greater than 20 degrees.

An explanation for this drastic decline in performance can be given based on the observation from Section 5.1, where the connection between the room impulse responses and erroneous TDOA estimates was made. The jump in aperture size from the 8-element array to the 15-element array resulted in many GCC pairs having longer separation distances, and this caused an increase in the range of valid time delays considered when maximizing the GCC functions to find the TDOAs. With this range of delays increased, more of the secondary peaks that are formed by the reflection-peaks in the room impulse responses appear in the GCC functions causing more erroneous TDOA estimates. While this effect could be minimized by judiciously choosing the GCC pairs with small separation distances, this reduces the overall resolution of the array, which is defined by its overall aperture size, to that of the longest pairwise separation distance. Despite that fact the SRP-PHAT did not seem to benefit from the increased aperture size in this experiment, it can be exploited to resolve multiple sources, which is something that the GCC-PHAT technique is unable to do.

7.3 Experiment #3: 3D Source Localization using the *Huge Microphone Array* (HMA)

The Huge Microphone Array (HMA), which is fully described in [84] and [85], can support up to 512 microphones. Its current installation includes 256 omni-directional microphones mounted on a series of 1.34 by 0.67-meter panels. The microphones on each panel have been placed at a random subset of nodes that are intersections of a 3 by 3 centimeter grid. The total array forms a "U"-shape aperture that is about 3 by 5 meters, when viewing it from the top. Thus far, the HMA has real-time processing algorithms that perform three basic functions; it can locate a talker, decide whether the estimated location is accurate, and apply a time-domain delay-and-sum beamformer. The HMA also has the ability to make 3-second recordings using all 256 microphones. This feature was employed to acquire data for the following

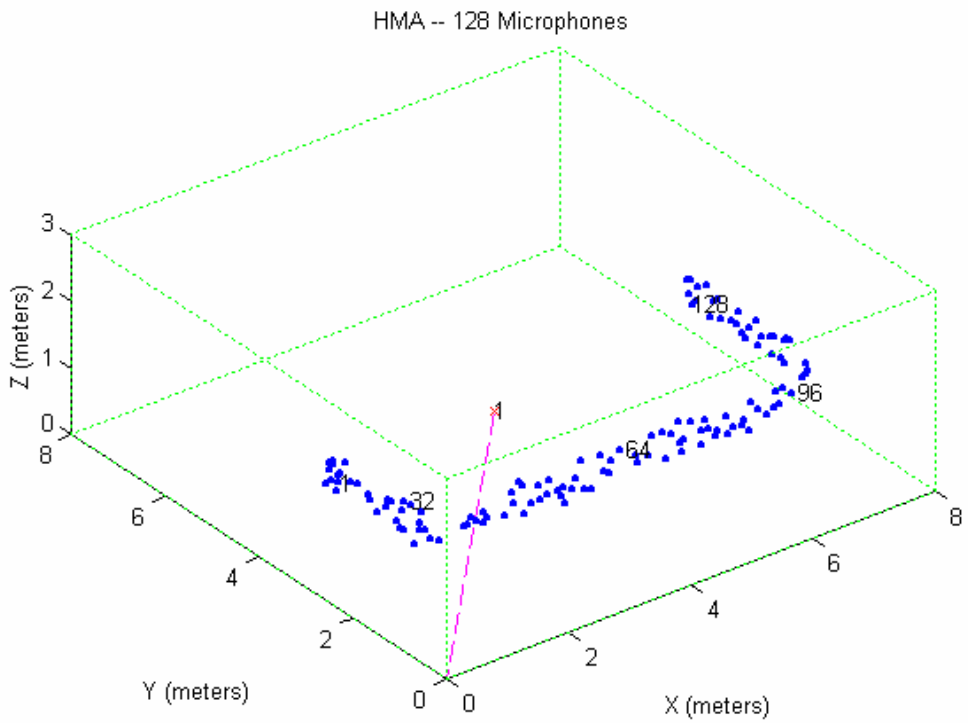

Figure 7.7 The HMA layout with 128 (of 256) microphones. Source "1" is located at (3.11,3.03,1.53) meters**.**

experiment, which evaluates the performance of SRP, SRP-PHAT and GCC-PHAT when applied to a large aperture array.

7.3.1 Data and Setup

Two 3-second recordings were made of ***real*** talkers using 128 microphones, which were randomly chosen from the aperture of the HMA. The positions of these microphones in a Cartesian coordinate system are shown in Figure 7.7. Also depicted in this figure is the room size, which is 8 by 8 by 3 meters, and the location of the single talker used for the first recording. This talker was located at (3.11,3.03,1.53) meters from the origin of the array, which was chosen to be the same as the global origin in the room coordinate system. This recording was used to compare the performance of SRP, SRP-PHAT and GCC-PHAT.

A second recording was made of three simultaneous talkers, which was used to demonstrate the resolution of SRP-PHAT as compared to SRP. Inherent in these beam-steering techniques is the ability to find multiple sources; each source corresponds to a peak in the steered response. GCC-based techniques don't typically allow for such multi-talker situations since their derivations are based on single-path propagation. In fact, the presence of multiple sources generally degrades the performance of GCC. Hence, GCC was not applied to this multi-talker recording.

Unlike the conference-room data set, which was collected by playing pre-recorded speech through a loudspeaker, the HMA recordings made for this experiment were of actual talkers. These talkers were instructed to stand as still as possible during each recording, although the exact positions of the their mouths could not be controlled. Hence, the actual locations of the talkers were known to within 10 to 20 centimeters of the true location. There may have been some movement by the talker's heads as the recordings were made, although it has been assumed that such movement is negligible and within the uncertainty of the true location.

The reverberation time of the room, T_{60}, where the HMA has been installed was measured in [87] and [78] and reported to be approximately 400 milliseconds. This value is twice that of the conference room where the data set of Section 3.2 was collected using the Megamike. The HMA environment was also noisier than the conference room since there is more noise-making hardware associated with the 512-

channel HMA than the 16-channel Megamike. Hence, the recordings used in this experiment were considerably more reverberant and noisy than those used in the previous experiments of this thesis.

7.3.2 Location Estimation

Location estimation was performed using SRP, SRP-PHAT and GCC-PHAT with the single-source recording. The large aperture of the HMA provided sufficient spatial resolution in all three Cartesian dimensions, and therefore searches were performed over three free variables. These searches, required to maximize the steered response power, $P(\vec{d})$, or minimize the GCC TDOA error, $E(\vec{d})$, were performed on a 0.1-meter grid for $1.5 \le x \le 4.5$, $1 \le y \le 4$ and $1.5 \le z \le 2.5$ meters. GCC pairs were formed using all 128 microphones by combining adjacent microphone indices. This formed 127 pairs, with separation distances that ranged from 5.9 centimeters to 97 centimeters and had a median separation of 22.2 centimeters. The GCC functions for each pair were computed with a 0.1 sample resolution.

The same block parameters were maintained from the previous experiments: Hanning windowed, 25-millisecond blocks with half-overlap. While the sampling rate of the HMA is 20kHz, the recording was

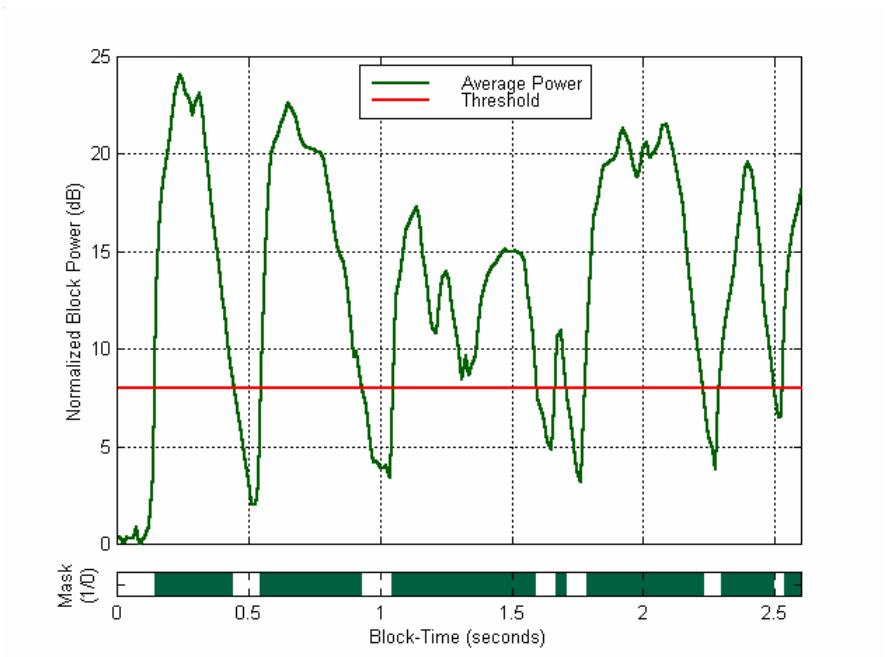

Figure 7.8 Plot of block power averaged over microphone with SNR threshold marked by the horizontal (red) line. Below this plot is the SNR mask.

re-sampled at 16kHz for the sake of consistency with the conference-room data set. Nearly the entire band from the 16kHz data was used to compute both the SRP responses and GCC functions. As in the previous experiments of this chapter, this frequency range was 300Hz to 8000kHz.

The average block power for the 128-element array was computed for this recording and used to derive an SNR mask as described in Section 3.5. The block power and SNR mask are shown in Figure 7.8. The beginning of the recording captured the background noise only, and the block power in Figure 7.8 has been offset so that the power of this background noise corresponds to 0 zero dB. From this, an SNR threshold was established at 8 dB, which is equal to 0.33 of the maximum block power (24 dB). With this threshold, the mask passed 160 out of 208 location estimates.

It has been mentioned that the SNR was lower in this recording than in the recordings of the conference-room data set. With the background-noise power approximately the same for all blocks, as it was shown for the conference-room data set in Section 3.5, the block power plot of Figure 7.8 is essentially the SNR of each block within a constant offset. Comparing this plot, to the similar plot for speech source 2 of the conference-room data set in Figure 3.9, it is apparent that the SNR of the HMA recording was, on average, about 12 dB lower. Figure 3.9 shows that the SNR in the conference room averaged about 25 dB over the duration of the recording and was as high as 37 dB for some blocks. In contrast, the average SNR of the HMA recording, as shown in Figure 7.8, was about 13 dB and only reached 24 dB at its most powerful block.

7.3.3 Experimental Results

The error of each location estimate was computed using the geometric distance from the true source location, $\vec{d}^{(s)}$, to the estimated location, $\hat{\vec{d}}$:

$$E_{LOC} \equiv \sqrt{\left(\vec{d}^{(s)} - \hat{\vec{d}}\right) \cdot \left(\vec{d}^{(s)} - \hat{\vec{d}}\right)}$$

Error rates were computed for SRP, SRP-PHAT and GCC-PHAT, and these are plotted in Figure 7.9. All of the SRP-PHAT estimates have error less than 0.1 meters (which is equal to the spacing of the search grid). About 70 percent of the SRP-estimates are equally accurate. In sharp contrast, GCC-PHAT is far less accurate, with 70 percent of the estimates having error greater than 0.5 meters. It is apparent from

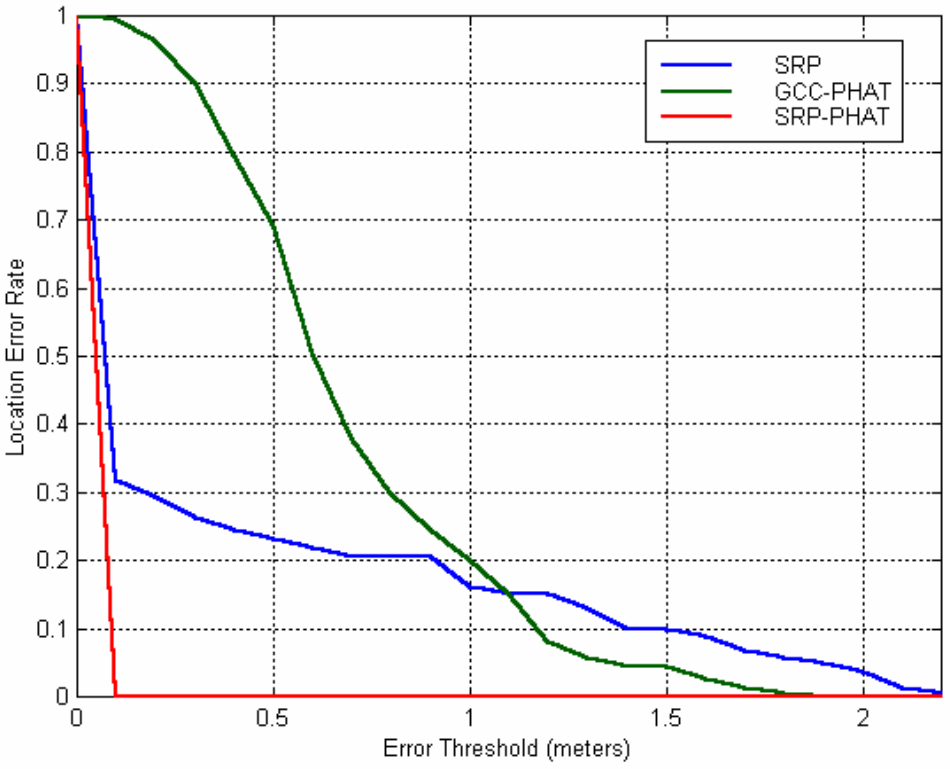

Figure 7.9 Location error rates for SRP, GCC-PHAT and SRP-PHAT using 128 microphones.

Figure 7.9 that SRP-PHAT is superior to the GCC-PHAT method, which uses the same number of microphones as the SRP-PHAT method in a pairwise fashion.

The GCC-PHAT algorithm was run a second time using more data to compute each cross-correlation. This was done, as it was in Section 5.3.3, by accumulating cross-spectra over several consecutive blocks. Each phase transform was computed using the average cross-spectrum accumulated over 19 half-overlapping blocks, giving 250 milliseconds of data per TDOA estimate. The resulting location error rate is plotted in Figure 7.10. Also plotted for comparison are the GCC-PHAT and SRP-PHAT error rates from Figure 7.9, which were produced using single, 25-millisecond blocks per TDOA estimate. As the error rates show, the performance of GCC-PHAT improves considerably with a ten-fold increase in the cross-spectra accumulation time. About 80 percent of the location-estimates have error less than 0.4 meters. However, this performance is still not equal to that of SRP-PHAT using the smaller

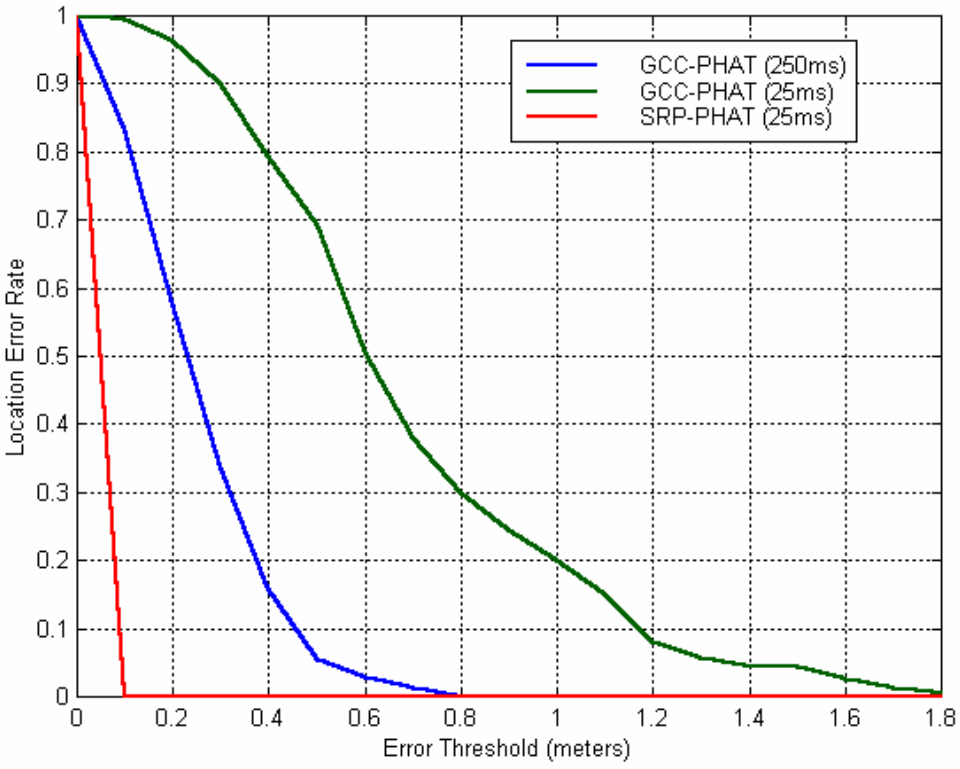

Figure 7.10 Cartesian error rates for two different cross-spectra accumulation times: 25ms and 250ms.

analysis time of 25 milliseconds per estimate, which produced estimates that were all within 0.1 meters of the actual location.

As reported in [87], the HMA's real-time locator is able to achieve accurate performance using GCC-PHAT when combining long accumulation times with peak-picking heuristics designed to reject location estimates derived from erroneous peaks in the GCC functions. The real-time system uses the same TDOA RMS error minimization technique as was used in this experiment to localize talkers. Each set of TDOAs is estimated using 204.8 milliseconds of microphone data; cross-spectra are accumulated over eight, non-overlapping 25.6ms-blocks. While the HMA currently includes 256 microphones, the locator uses only 24 of them due to the computation limits of the system. However, with such long data segments, and with the heuristics used to discard erroneous locations, the HMA'a locator is quite accurate. Unfortunately, many of the estimates are qualified as erroneous, and it is very difficult to track the dynamic conditions encountered. Hence, with sufficient computational power, which is available on the HMA, a

version of SRP-PHAT could be implemented that would give the same accuracy as the current real-time system with a decrease in estimate latency and increase in talker-tracking ability.

7.3.4 Multi-talker Resolution

Inherent in the beam-steering methods of SRP and SRP-PHAT is the ability to find multiple, active sources. The second HMA recording, which was of three people talking simultaneously, was used to

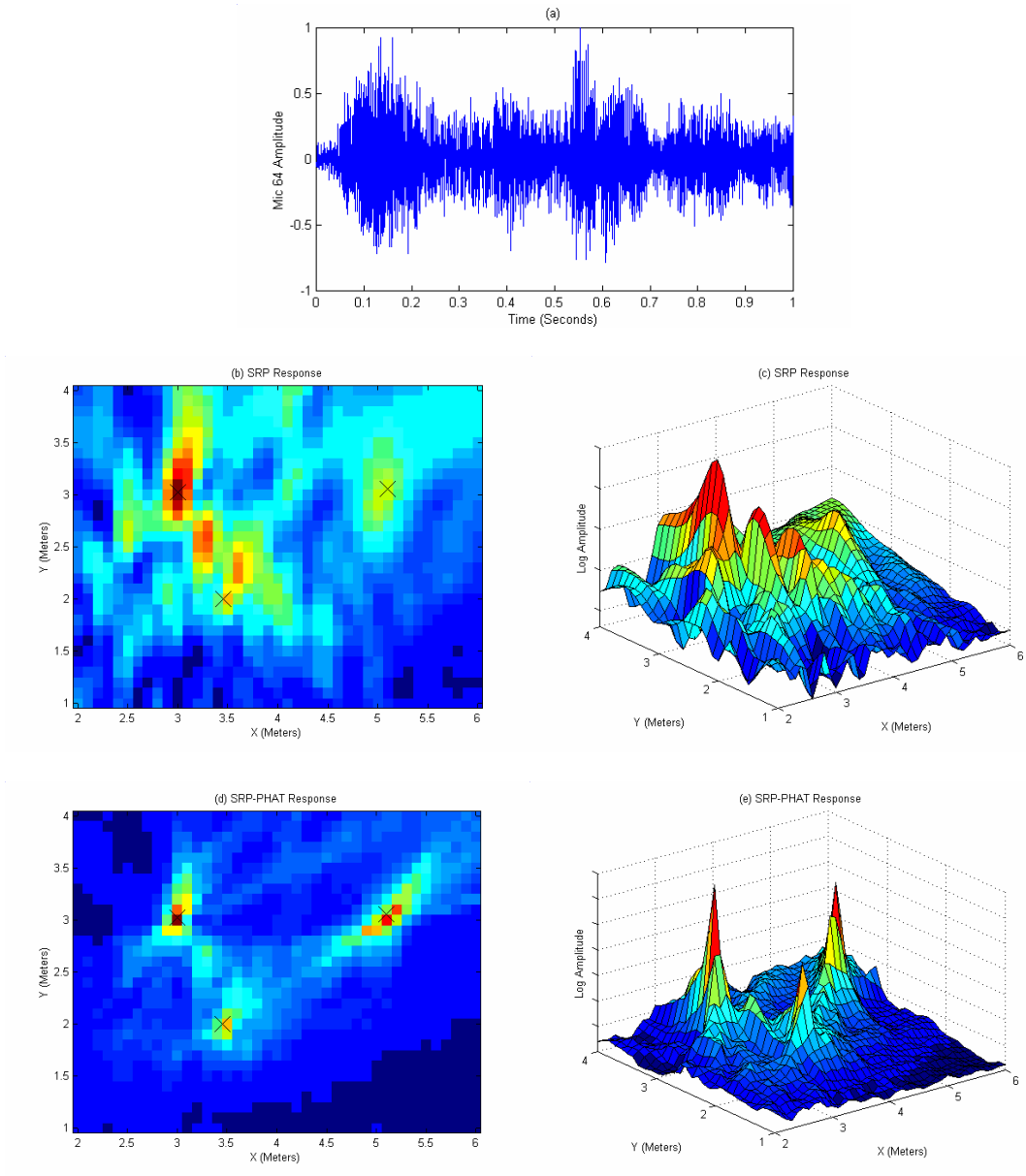

Figure 7.11 Steered responses using the 128-element array and three simultaneous talkers. The top plot (e) shows the amplitude of microphone 64 during the 1-second recording used to compute the SRP and SRP-PHAT responses below it.

demonstrate this ability, and to show that the filters used by SRP-PHAT enhance the steered responses by sharpening the peaks corresponding to the locations of the talkers. The recording was performed using the same sub-set of 128 microphones illustrated in Figure 7.7. With this large-aperture array, there is sufficient spatial resolution in the steered responses to distinguish talkers that are less than a meter apart.

The steered responses were computed using SRP and SRP-PHAT for 80 25-millisecond blocks over 1 second of the multi-talker recording. For the sake of visualizing these 3D functions, each response was summed over the z-dimension for values in the range $1.5 \leq z \leq 2.5$ meters with a 0.1-meter resolution. The resulting 2D functions are the superposition of all the responses in this range of z. The 2D functions were also averaged over all 80 blocks to give a single 2D representation of the steered response over the duration of this 1-second interval. This averaging ensured that the contributions from all three talkers appeared in the final response. While the resulting responses integrated 1 second of array data, each of the 80 block-responses accurately depicted the locations of the active talkers.

The resulting 2D responses have been plotted for $2 \leq x \leq 6$ and $1 \leq y \leq 4$ meters, with a 0.1-meter resolution, and are presented in Figure 7.11. At the top of this figure, there is plot labeled "(a)" that is the amplitude of microphone 64 during the 1-second recording segment. This plot appears unintelligible, containing the utterances from three talkers, and the reverberation caused by them. However, as the responses show, their locations can be identified using this short recording.

The image plot labeled "(b)" is the steered response power of the delay-and-sum beamformer with each source location marked by an "X". Just to the right of this plot, is a surface plot labeled "(c)" showing the relative heights of the peaks corresponding to each talker. Notice that the talker at (3.0,3.0) can been clearly identified, but the other two are amidst secondary peaks with comparable amplitudes. With this method, it would not be possible to accurately determine the locations and number of active talkers.

The steered responses of the filter-and-sum beamformer using the PHAT filters are depicted by the plots labeled "(d)" and "(e)" in Figure 7.11. Notice the three distinct peaks that accurately correspond to the source locations. Unlike the delay-and-sum peaks, the SRP-PHAT peaks are sharp and stand out from any secondary peaks. Hence, SRP-PHAT could be employed as the basis of a robust multi-talker localization algorithm taking great advantage of the short-data blocks used for the estimates.

8 Summary, Conclusions and Future Work

The results of carefully performed experiments, using real microphone-array data, show that SRP-PHAT is vastly superior to SRP and GCC-PHAT in accuracy. SRP-PHAT computes its accurate estimates using small blocks of array data. To achieve similar results, GCC-PHAT requires a significant increase in data requirements (over 10 times the data), which increases latency and decreases responsiveness. Without intelligent pruning and sensitive heuristics, GCC-PHAT did not achieve the accuracy of SRP-PHAT in adverse acoustic conditions. The GCC and GCC-PHAT pairwise methods were severely impacted by mild reverberation in experiments where the source was more than 3 meters from the array. The significant improvement of SRP-PHAT over GCC-PHAT using various microphone arrays and different acoustic environments compels further study of non-pairwise methods for talker localization.

8.1 Summary

Microphone-array data was collected in a 7 by 4 by 3-meter conference room using a custom built array-system, the *Megamike*. This data was used in a series of experiments, which were described in Chapters 5 and 7. An additional data set was collected in a 7 by 7 by 3-meter room using another custom built system, the *Huge Microphone Array*. Using subsets of the microphones from these two systems, the performance of both generalized cross-correlation and steered-response talker localization methods were studied for various array configurations. These configurations included small-aperture and medium-aperture planar arrays, and a large-aperture, 128-element, "U"-shaped array. These data sets captured the characteristics of two different environments: **1)** a high-SNR, mildly reverberant (200-millisecond reverberation time) small room, **2)** a low-SNR, highly reverberant (400-millisecond reverberation time) medium-size room.

The experiments of Chapter 5 showed that mild reverberation could severely impact the performance of GCC-based localization techniques when the block size is short (25 milliseconds). Using the measured room impulse responses, a connection was made between anomalous TDOA estimates and the secondary peaks of the cross-correlation of the room impulse responses. These anomalies were responsible for the poor performance of the GCC-PHAT-based DOA estimator. This estimator was most severely affected by reverberation when the source was placed greater than 3 meters from the 8-element, 33

by 36-centimeter planar array. However, performance did increase at closer distances and with longer ensemble averages over time for computation of the phase transforms. These factors suggested that the phase transform was well suited to talker localization over a limited spatial range when there was sufficient microphone data available. This motivated the development of a new steered-beamformer localization method in Chapter 6, SRP-PHAT, which extended the phase transform from a pairwise to a multi-microphone function. By choosing appropriate filters for a filter-and-sum beamformer, SRP-PHAT effectively incorporates short segments of data from multiple microphones using a beam-steering process that is equivalent to averaging the phase transforms of all possible pairs in the array.

The experiments in Chapter 7 evaluated the performance of SRP-PHAT, as well as that of the conventional steered-response method, SRP. These beam-steering methods were compared to similar GCC-PHAT localization methods. For both the small and large arrays and their respective environments, SRP-PHAT was more accurate and robust than SRP and GCC-PHAT when implemented using 25-millisecond data blocks. It was also demonstrated that the filters used by the PHAT beamformer improve the resolution of the steered response. This allowed a clear separation of three simultaneous talkers using the 128-element array. These experiments demonstrated that microphone redundancy could be exploited to reduce the data requirements for accurate talker localization in reverberant environments using the SRP-PHAT method.

8.2 Computational Complexity

There is an obvious price to pay for the improvement in performance of SRP-PHAT over GCC-PHAT. This price comes in the form of a significant increase in computational complexity. A quick analysis of the computation required by SRP-PHAT and GCC-PHAT can be performed using the following definitions for the numbers of each operation:

$N_l \equiv$ *# of evaluations of objective function (steered response or TDOA RMS error)*

$N_k \equiv$ *# of DFT components used in computation*

$N_m \equiv$ *# of microphones in the array*

$N_p \equiv$ *# of GCC pairs used for TDOA RMS error*

$N_\tau \equiv$ *# of points computed for each GCC function in the time-lag domain*

Using "big-O" notation [27], the number of operations required for each evaluation of the steered response power is $O(N_k N_m)$. This is performed a total of N_l times. Hence, the total computation for SRP-PHAT is $O(N_l N_k N_m)$. The computation of each generalized cross-correlation requires $O(N_\tau N_k)$. This must be performed for each microphone pair, which results in a total of $O(N_\tau N_k N_p)$ In addition to the cross-correlation functions, the TDOA RMS error must be evaluated N_l times. Hence, GCC-PHAT requires a total of $O(N_\tau N_k N_p)+O(N_l)$ operations. The ratio of the SRP-PHAT operations to the GCC-PHAT operations can now be expressed as follows:

$$\text{Compute Ratio} = \frac{\text{SRP - PHAT operations}}{\text{GCC - PHAT operations}} = \frac{O(N_l N_k N_m)}{O(N_\tau N_k N_p) + O(N_l)}$$

Dividing both the numerator and denominator by $O(N_\tau N_k N_p)$ yields:

$$\text{Compute Ratio} = \frac{O(N_l N_m) \Big/ O(N_\tau N_p)}{1 + O(N_l) \Big/ O(N_\tau N_k N_p)}$$

For time-delays ranging from -20 to +20 samples and a 0.1-sample step-size, N_τ is on the order of 400. With N_k on the order of 500 and N_p on the order of 50, $O(N_\tau N_k N_p) >> O(N_l)$ for any reasonable value of N_l. Hence, the compute ratio simplifies to:

$$\text{Compute Ratio} = \frac{O(N_l N_m)}{O(N_\tau N_p)}$$

Therefore, SRP-PHAT requires approximately $N_l N_m \Big/ N_\tau N_p$ more operations than GCC-PHAT.

Consider the far-field experiment of Section 7.1, where an 8-element planar array was used to search over a grid of DOAs ranging from -60 to +60 degrees with a 1-degree step size. In this case, N_l=121*x*121=14641. The range of valid TDOAs varied with the separation distance of each microphone pair resulting in a minimum TDOA range of 5 samples, a maximum of 31 samples, and an average of 18 samples. Using the average TDOA range, at a 0.1-sample step size, the number of points computed for

each GCC function was N_τ=180. Using the relationship just derived, the SRP-PHAT method required 23 times the computation required by GCC-PHAT in this experiment.

The HMA experiment of Section 7.3 required a much larger number of functional evaluations since the searches were performed in three dimensions using 128 microphones. In fact, computational limits somewhat dictated the range searched in the z-dimension, which was ultimately chosen to be 1 meter. The x and y-dimensions were searched over a 3-meter range, and all 3 dimensions were searched with a 0.1-meter step size. Hence, the number of evaluations of the objective functions was N_f=9000. With a mean TDOA range of 11 samples and a 0.1-sample step-size, the GCC functions were computed, on average, for N_τ=110 points. In this case, the SRP-PHAT method required 83 times the computation required by GCC-PHAT.

8.3 Future Work

With the promising results of Chapter 7, future work should focus on techniques that combine the signals, rather than the time-delay parameters, from multiple microphones. There is a clear advantage to doing this as the superior performance of SRP-PHAT shows. The obvious drawback is the increase in computational requirements over pairwise, time-delay methods. The far-field case might be an acceptable increase in computation (23 times) for some applications. While it is not the ideal solution, a large aperture-array, such as the HMA, could be broken into multiple far-field sub-arrays. Using a method such as the linear intersection method for talker localization [7], the DOAs from these sub-arrays could be combined to yield location estimates. Such an approach is still an improvement over breaking the array into microphone pairs. Using these simplifying array geometries in conjunction with more efficient searching techniques could make SRP-PHAT the basis for an efficient and practical source-localization algorithm. The significant improvement of SRP-PHAT over GCC-PHAT, which was demonstrated in the experiments of this thesis, motivates the development of a computationally efficient form of SRP-PHAT or a similar method that yields the same performance by fully exploiting all the microphones in a given array-system.

Bibliography

[1] J. E. Adcock, J. H. DiBiase, M. S. Brandstein and H. F. Silverman. Practical issues in the use of a frequency-domain delay estimator for microphone-array applications. *128th Meeting of the Acoustical Society of America*, November 1994.

[2] J. E. Adcock, Y. Gotoh, D. J. Mashao and H. F. Silverman. Microphone-array speech recognition via incremental MAP training. In *Proceeding of ICASSP96*, Atlanta, GA, May 1996.

[3] J. B. Allen and D. A. Berkely. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Amer.*, 65(4):943-949, 1979.

[4] R. Bouquin-Jeanns, A. A. Azirani and G. Faucon. Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator. *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 5, September 1997.

[5] M. S. Brandstein. Informal presentation. *International Workshop on Microphone-Array Systems: Theory and Practice*, Brown University, Providence, RI, October 1992.

[6] M. S. Brandstein, J. E. Adcock and H. F. Silverman. *Method and Apparatus for Source Location Estimation from Microphone-Array Time-Delay Estimates*, US Patent No. 5737431, issued April 7, 1998.

[7] M. S. Brandstein, J. E. Adcock and H. F. Silverman. A closed-form location estimator for use with room environment microphone arrays. *IEEE Transaction on Speech and Audio Processing*, 5(1):45-50, January 1997.

[8] M. S. Brandstein, J. E. Adcock and H. F. Silverman. A closed-form method for finding source locations from microphone-array time-delay estimates. In *Proceedings of ICASSP95*, pages 3019-3022, IEEE, May 1995.

[9] M. S. Brandstein, J. E. Adcock and H. F. Silverman. A localization-error based method for microphone-array design. In *Proceedings of ICASSP96*, IEEE, May 1996.

[10] M. S. Brandstein, J. E. Adcock, and H. F. Silverman. Microphone array localization error estimation with application to sensor placement. *J. Acoust. Soc. Amer.*, 99(6):3807-3816, June 1996.

[11] M. S. Brandstein, J. E. Adcock and H. F. Silverman. A practical time-delay estimator for localizing speech sources with a microphone array. *Computer, Speech and Language*, Volume 9, pages 153-169, September 1995.

[12] M. S. Brandstein. *A Framework for Speech Source Localization Using Sensor Arrays*. Ph. D. thesis, Brown University, Providence, RI, May 1995.

[13] M. S. Brandstein. A pitch-based approach to time-delay estimation of reverberant speech. In *Transactions of 1997 Workshop on Applications of Signals Processing to Audio and Acoustics*, New Paltz, N.Y., October 1997.

[14] M. S. Brandstein and H. F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *Proceedings of ICASSP97*, IEEE, April 1997.

[15] M. S. Brandstein and H. F. Silverman. *Method and Apparatus for Adaptive Beamforming*, US Patent No. 5581620, issued December 3, 1996.

[16] D. Burshtein and E. Weinstein. Confidence intervals for the maximum entropy spectrum. *IEEE Trans Acoustics, Speech and Signal Processing*, ASSP-35:504-510, April 1987.

[17] G. Carter. Variance bounds for passively locating an acoustic source with a symmetric line array. *J. Acoust. Soc. Am*., 64(4):922-926, October 1977.

[18] G. C. Carter, C. H. Knapp, and A. H. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. *IEEE Transactions on Audio and Electroacoustics*, AU-21(4):337-344, August 1973.

[19] B. Champagne, S. Bédard and A. Stéphenne. Performance of time-delay estimation in the presence of room reverberation. *IEEE Trans. Speech Audio Proc.*, 4(2):148-152, March 1996.

[20] Y. T. Chan and K. C. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Transactions on Signal Processing*, 42(8):1905-1915, 1994.

[21] C. Che, Q. Lin, J. Pearson, B. deVries and J. Flanagan. Microphone arrays and neural networks for robust speech recognition. In *Proceedings of the Human Language Technology Workshop*, Pages 342-347, Plainsboro, NJ, March 8-11, 1994.

[22] C. Che, M. Rahim and J. Flanagan. Robust speech recognition in a multimedia teleconferencing environment. *J. Acoust. Soc. Am*., 92(4):2476, 1992.

[23] T. Chou. Frequency-independent beamformer with low response error. In *Proceedings of ICASSP95*, IEEE, 1995.

[24] P. L. Chu. Superdirective microphone array for a set-top videoconferencing system. In *Proceedings of ICASSP97*, vol. 1, pp. 235-38, May 1997.

[25] P. L. Chu. Desktop mic array for teleconferencing. In *Proceedings of ICASSP95*, IEEE, 1995.

[26] R. T. Compton, Jr. *Adaptive Antennas*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[27] T. H. Cormen, C. E. Leiserson and R. L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, Massachusetts, 1990.

[28] L. Cremer. *Principles and Applications of Room Acoustics,* volumes 1 & 2. Applied Science Publishers, Ripple Road, Barking, Essex, England, English language edition, 1982.

[29] J. H. DiBiase. Informal presentation of the Megmaike array-system. *Third Biennial Roundtable on Microphone Array Technology*, Brown University, Providence, RI, October 1996.

[30] Electret condenser microphone cartridge, ceramic microphone receiver cartridge, dynamic microphone cartridge publication. vol. 3. Secaucus, NJ: Panasonic Electronic Components Div., Panasonic Indust. Co., 1994.

[31] G. W. Elko and A. N. Pong. A steerable and variable first-order differential microphone array. In *Proceedings of ICASSP97*, May 1997.

[32] R. Ellis and D. Gulick. *Calculus with Analytical Geometry*. 3rd edition, Harcourt Brace Jovanovich Publishers, Inc. San Diego, 1986.

[33] C. F. Eyring. Reverberation time in 'dead' rooms. *J. Acoust. Soc. Amer.*, vol. 1, pp. 217-241, 1930.

[34] J. Flanagan, H. Johnson, R. Zahn and G. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Amer.*, 78(5):1508-1518, November 1985.

[35] J. L. Flanagan, D. A. Berkley, G.W. Elko, J. E. West, and M. M. Shondhi. Autodirective microphone systems. *Acoustica*, vol. 73, pp. 58-71, 1991.

[36] J. L. Flanagan and H. F Silverman, *International Workshop on Microphone-Array Systems: Theory and Practice*, Brown University, Providence, RI, October 1992.

[37] J. L. Flanagan and H. F Silverman, *Workshop on Microphone Arrays: Theory, Design & Application*, CAIP Center, Rutgers University, October, 1994.

[38] J. L. Flanagan and H. F Silverman, *Third Biennial Roundtable on Microphone Array Technology*, Brown University, Providence, RI, October 1996.

[39] J. Flanagan, A. Surendran and E. Jan. Spatially selective sound capture for speech and audio processing. *Speech Communication*, 13(1-2):207-222,1993.

[40] D. Giuliani, M. Omologo and P. Svaizer. Talker localization and speech recognition using a microphone array and a cross-power spectrum phase analysis. In Proceedings of ICSLP, volume 3, pages 1243-1246, September 1994.

[41] M. M. Goodwin and G. W. Elko. Constant beamwidth beamforming. In *Proceedings of ICASSP93*, IEEE, 1993.

[42] Y. Gotoh. *Incremental Algorithms and MAP Estimation: Efficient HMM Learning of Speech Signals*. Doctoral Dissertation, Brown University, Providence, RI

[43] Y. Gotoh, M. M. Hochberg, D. J. Mashao, and H. F. Silverman. Incremental map estimation of HMM's for efficient training and improved performance. In *Proceedings of ICASSP95*, vol.1, pp. 457-460, IEEE, 1995.

[44] L. J. Griffiths and C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Ant. Prop*., AP-30:27-34, January 1982.

[45] W. Hahn and S. Tretter. Optimum processing for delay-vector estimation in passive signal arrays. *IEEE Trans. Inform. Theory*, IT-19(5):608-614, September 1973.

[46] D. Halliday and R. Resnick. *Fundamentals of Physics*. John Wiley & Sons, Inc., New York, 1988.

[47] Y. Haneda, S. Makino and Y. Kaneda. Common acoustical pole and zero modeling of room transfer functions. *IEEE Trans. Speech Audio Proc.*, 2(7):320-328, April 1994.

[48] M. H. Hayes. *Statistical Signal Processing and Modeling.* John Wiley and Sons, Inc., New York, 1996.

[49] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, Englewood Cliffs, New Jersey 07632, second edition, 1991.

[50] S. Haykin. *An Introduction to Analog and Digital Communications.* John Wiley and Sons, Inc., New York, 1989.

[51] T. Hikichi and F. Itakura. Time variation of room acoustic transfer functions. In *Proceedings of ICASSP93*, IEEE, 1993.

[52] M. M. Hochberg. *A Comparison of State-Duration Modeling Techniques for Connected Speech Recognition*. Doctoral Dissertation, Brown University, Providence, RI, 1993.

[53] O. Hoshuyama and A. Sugiyama. A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. In *Proceedings of ICASSP96*, IEEE, 1996.

[54] Y. Huang, J. Benesty, and G. W. Elko. Adaptive eigenvalue decomposition algorithm for real-time acoustic source localization. In *Proceedings of ICASSP99*, IEEE, 1999.

[55] T. B. Hughes, H. Kim, J. H. DiBiase, and H. F. Silverman. Performance of an HMM speech recognizer using a real-time tracking microphone array as input. *IEEE Trans. Speech Audio Proc.*, 7(3):346-349, May 1999.

[56] T. B. Hughes, H. Kim, J. H. DiBiase, and H. F. Silverman. Using a real-time, tracking microphone array as input to an HMM speech recognizer. In *Proceedings of ICASSP98*, IEEE, 1998.

[57] E. Jan, P. Svaizer and J. Flanagan. Matched-filter processing of microphone array for spatial volume selectivity. In *Proceedings of ICASSP95*, pages 1460-1463, IEEE, 1995.

[58] D. H. Johnson. The application of spectral estimation methods to bearing estimation problems. In *Proceedings of IEEE*, 70:1018-1028, September 1982.

[59] D. H. Johnson and D. E. Dudgeon. *Array Signal Processing: Concepts and Techniques*. P T R Prentice Hall, Englewood Cliffs, New Jersey 07632, 1993.

[60] W. Kellerman. A self-steering digital microphone array. In *Proceedings of ICASSP91*, pages 3581-3584, IEEE, May 1991.

[61] F. Khali, J. P. Jullien and A. Gilloire. Microphone array for sound pickup in teleconferencing systems. *J. Audio Eng. Soc*., vol. 42, no. 9, September 1994.

[62] L. E. Kinsler et al. *Fundamentals of Acoustics*. John Wiley & Sons, New York, 3rd edition, 1982.

[63] S. E. Kirtman and H. F. Silverman. A user-friendly system for microphone array research. In *Proceedings of ICASSP95*, vol. 5, pp. 3015-3018, May 1995.

[64] C. H. Knapp and G.C. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320-327, August 1976.

[65] H. Kuttruff. *Room Acoustics*. London, U.K.: Elsevier, 1991, 3rd edition.

[66] R. T. Lacoss. Data adaptive spectral analysis methods. *Geophysics*, 36:661-675, August 1971.

[67] D. J. Mashao. *Computations and Evaluations of an Optimal Feature-set for an HMM-based Recognizer*. Doctoral Dissertation, Brown University, Providence, RI, 1996.

[68] P. C. Meuse. *Characterization of Talker Radiation Pattern Using a Microphone Arrays*. Ph. D. thesis, Brown University, Providence, RI, May 2000.

[69] P. C. Meuse and H F Silverman. Characterization of Talker Radiation Pattern Using a Microphone Array. *Proceedings of ICASSP94*, May 1994.

[70] J. Mourjopoulos. On the variation and invertibility of room impulse responses. *Journal of Sound and Vibration*, 102(2):217-228, 1985.

[71] J. Mourjopoulos. Pole and zero modeling of room transfer functions. *Journal of Sound and Vibration*, 146(2):281-302, 1991.

[72] M. Omologo and P. Svaizer. Acoustic source localization in noisy and reverberant environments using CSP analysis. In *Proceedings of ICASSP96*, pages 901-904. IEEE, 1996.

[73] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, New Jersey 07632, 1989.

[74] W. H. Press, S. A. Teukolsky, W. T. Vettering and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, New York, 1992.

[75] J. G. Proakis and D. G. Manolakis. *Introduction to Digital Signal Processing*. Macmillan Publishing Company, New York, 1988.

[76] D. V. Rabinkin, R. J. Renomeron, A. Dahl, J. C. French, J. L. Flanagan and M. H. Bianchi. A DSP implementation of source location using microphone arrays. *J. Acoust. Soc. Amer.*, 99(4):2503-, April 1996.

[77] D. D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *J. Audio Eng. Soc.*, 37(6):419-444, June 1989.

[78] J. M. Sachar. *Measurement Techniques and Results for Assessing the Performance of a Large-Aperture Microphone Array in a Highly Reverberant Room*. Sc. M. thesis, Brown University, Providence, RI, May 2000.

[79] H. F. Silverman. Some analysis of microphone arrays for speech data acquisition. Trans. on *Acoustics, Speech, and Signal Processing*, 35(12):1699-1711, IEEE, December 1987.

[80] H. F. Silverman and Y. Gotoh. On the implementation and computation of training an HMM recognizer having explicit state durations and multiple-feature-set, tied-mixture output probabilities. LEMS Monograph Ser., no. 1-1, LEMS, Division of Engineering, Brown University, Providence, RI, March 1994.

[81] H. F. Silverman and S. E. Kirtman. A two-stage algorithm for determining talker location from linear microphone-array data. *Computer, Speech and Language*, 6(2):129-152, April 1992.

[82] H. F. Silverman and W. R. Patterson. Visualizing the performance of large-aperture microphone arrays. In *Proceedings of ICASP-1999*, pp. 969-972, Phoenix, AZ, March, 1999.

[83] H. F. Silverman, W. R. Patterson and J. L. Flanagan. *The Huge Microphone Array (HMA)*. LEMS Technical Report, Brown University, Providence, RI, May 1996.

[84] H. F. Silverman, W. R. Patterson and J. L. Flanagan. The Huge Microphone Array (HMA) – Part I. *IEEE Transactions on Concurrency*, 6(4):36-46, October-December, 1998.

[85] H. F. Silverman, W. R. Patterson and J. L. Flanagan. The Huge Microphone Array (HMA) – Part II. *IEEE Transactions on Concurrency*, 7(1):32-47, January-March, 1999.

[86] H. F. Silverman, W. R. Patterson, J. L. Flanagan, and D. Rabinkin. A digital processing system for source location and sound capture by large microphone arrays. In *Proceedings of ICASSP97*, Munich, Germany, pp. 251-254, 1997.

[87] H. F. Silverman, W. R. Patterson and J. M. Sachar. First measurements of a large-aperture microphone array system for remote audio acquisition. White paper, Brown University, Providence, RI, January 2000.

[88] A. Stéphenne and B. Champagne. Cepstral prefiltering for time delay estimation in reverberant environments. In *Proceedings of ICASSP*, pages 3055-3058. IEEE, 1995.

[89] G. Strang. *Linear Algebra and Its Applications.* Harcourt Brace Jovanovich, Inc., Orlando, Florida 32887, 1988.

[90] N. Strobel and N. Rabenstein. Classification of time delay estimates for robust speaker localization. *In Proceedings of ICASSP99*, IEEE, 1999.

[91] D. E. Sturim. *Talker Characterization Using Microphone-Array Measurements*. Ph. D. thesis, Brown University, Providence, RI, May, 1999.

[92] D. E. Sturim, M. S. Brandstein, and H. F. Silverman. Tracking multiple talkers using microphone-array measurements. In *Proceedings of ICASSP97*, IEEE, April 1997.

[93] P. Svaizer, M. Mattassoni and M. Omologo. Acoustic source localization in a three-dimensional space using crosspower spectrum phase. In *Proceedings of ICASSP97*, IEEE, 1997.

[94] W. Tager and Y. Mahieux. Reverberant sound field analysis using a microphone array. In *Proceedings of ICASSP97*, IEEE, 1997.

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

[96] H. Wang and P. Chu. Voice source localization for automatic camera pointing system in videoconferencing. In *Proceedings of ICASSP*, volume 1, pages 187-190. IEEE, 1997.

[97] D. B. Ward, R. A. Kennedy and R. C. Williamson. An adaptive algorithm for broadband frequency invariant beamforming. In *Proceedings of ICASSP97*, IEEE, 1997.

[98] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of ICASSP98*, IEEE, 1998.

[99] L. J. Ziomek. *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing.* CRC Press, Inc., 2000 Corporate Blvd., N. W., Boca Raton, Florida 33431, 1995.