Abstract of "A Framework for Speech Source Localization Using Sensor Arrays," by Michael Shapiro Brandstein, Ph.D., Brown University, May 1995

Electronically steerable arrays of microphones have a variety of uses in speech data acquisition systems. Applications include teleconferencing, speech recognition and speaker identification, sound capture in adverse environments, and biomedical devices for the hearing impaired. An array of microphones has a number of advantages over a single-microphone system. It may be electronically aimed to provide a high-quality signal from a desired source location while simultaneously attenuating interfering talkers and ambient noise, does not necessitate local placement of transducers or encumber the talker with a hand-held or head-mounted microphone, and does not require physical movement to alter its direction of reception. Additionally, it has capabilities that a single microphone does not; namely automatic detection, localization, and tracking of active talkers in its receptive area. A fundamental requirement of sensor array systems is the ability to locate and track a speech source. An accurate fix on the primary talker, as well as knowledge of any interfering talkers or coherent noise sources, is necessary to effectively steer the array. Source location data may also be used for purposes other than beamforming; e.g. aiming a camera in a video-conferencing system. In addition to high accuracy, the location estimator must be capable of a high update rate as well as being computationally non-demanding in order to be useful for real-time tracking and beamforming applications.

This thesis addresses the specific application of source localization algorithms for estimating the position of speech sources in a real room environment given limited computational resources. The theoretical foundations of a speech source localization system are

presented. This includes the development of a source-sensor geometry for talkers and sensors in the near-field environment, the evaluation of several error criteria available to the problem, and the detailing of source detection and estimate-error prediction methods. Several practical algorithms necessary for real-time implementation are then developed, specifically the derivation and evaluation of an appropriate time-delay estimator and a novel closed-form locator. Finally, results obtained from several real systems are presented to illustrate the effectiveness of the proposed source localization techniques as well as to confirm the practicality of the theoretical models.

A Framework for Speech Source Localization Using Sensor Arrays

by

Michael Shapiro Brandstein

Sc.B., Brown University, 1988
S.M.E.E, Massachusetts Institute of Technology, 1990

Thesis

Submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
in the Division of Engineering at Brown University

May 1995

This dissertation by Michael Shapiro Brandstein is accepted in its present form by
the Division of Engineering as satisfying the
dissertation requirement for the degree of
Doctor of Philosophy

Date ....
Harvey F. Silverman

Recommended to the Graduate Council

Date ....
James L. Flanagan

Date ....
David B. Cooper

Approved by the Graduate Council

Date ....

The Vita of Michael Shapiro Brandstein

Michael was born January 11, 1967 in Willimantic, Connecticut and grew up in Woodbridge, Virginia. He attended Brown University from 1984 until 1988, graduating with a Bachelor of Science degree in Electrical Engineering, *magna cum laude*, with honors. He received an NSF fellowship for graduate study in 1988 and enrolled in the Electrical Engineering and Computer Science Department at Massachusetts Institute of Technology. As a Research Assistant for the Digital Signal Processing Group and later, the Advanced Television Research Project, he studied signal processing, doing research in speech and audio coding. He earned a Master's degree in Electrical Engineering in 1990, with a thesis on low-bit rate speech vocoders. During the 1991-92 school year, Michael taught mathematics as a faculty member of the Iolani School in Honolulu, Hawaii. He returned to Brown in the fall of 1992 to complete his graduate work, pursuing research on microphone arrays. His most recent work has focused on the application of microphone arrays for high-quality speech acquisition and efficient source localization. He is the author of two patents in this field.

Since 1980, Michael has spent his summers with the Johns Hopkins University Center for Talented Youth, first as a student and then as a member of the teaching staff. He has taught mathematics courses since 1985 as an instructor, and has been the Mathematics and Computer Science Staff Coordinator for the Lancaster, Pennsylvania site since 1989. Currently, he teaches a digital design course for the program. When not pursuing his research, Michael enjoys ultimate frisbee, sumo wrestling, and playing with dogs. He intends to continue spending his summers at camp for as long as they will have him.

Acknowledgments

I wish to express my sincere gratitude to my advisor, Professor Harvey Silverman, for his constant support, mentoring, and feedback as well as for originating the facilities and resources which made the microphone array project possible. This work could not have been accomplished without the efforts of the other members of the microphone array group: John Adcock, who provided persistent and helpful critical analysis; Joe DiBiase, who always managed to get things working for real; and Paul Meuse, who offered extensive pointers on everything. My warmest appreciation to my fellow graduate students, particularly Mike Wazlowski, Michael Blane, Aaron Smith, Yoshi Gotoh, and Daniel Mashao, for making the lab such an enjoyable environment. Thanks to my readers, Professor James Flanagan and Professor David Cooper, for their constructive comments, to laboratory technician Arpie Kaloustian for constructing equipment on a moment's notice, and to Gary Elko of Bell Laboratories for preparing the loudspeaker apparatus. This work was performed at the Laboratory for Engineering Man/Machine Systems (LEMS) of Brown University and partially funded by NSF Grants MIP-9120843 and MIP-9314625. During my tenure at Brown I have received generous financial support from a National Science Foundation Fellowship and a Perloff Graduate Research Fund Fellowship.

I would also like to acknowledge the people that matter the most: Mom and Dad, who have lovingly and generously provided me with the best things in life and are still responsible (with limited legal liability) for all my accomplishments; my sisters, Hali and Marcy, who patiently tolerated and supported me these many years; and finally, Karen, my moiety, proofreader, and the only person who has ever provided me with a worthwhile reason to graduate.

Contents

List of Tables

List of Figures

Chapter 1

Background, Motivation, and Scope

1.1 Sensor Arrays for Speech-Related Applications

A steerable array of microphones has the potential to replace the traditional head-mounted or desk-stand microphone as the input transducer system for acquiring speech data in many applications. An array of microphones has a number of advantages over a single-microphone system. First, it may be electronically aimed to provide a high-quality signal from a desired source location while it simultaneously attenuates interfering talkers and ambient noise. In this regard, an array has the potential to outperform a single, well-aimed, highly-directional microphone. Second, an array system does not necessitate local placement of transducers, will not encumber the talker with a hand-held or head-mounted microphone, and does not require physical movement to alter its direction of reception. These features make it advantageous in settings involving multiple or moving sources. Finally, it has potential capabilities that a single microphone does not; namely automatic detection, location, and

tracking of active talkers in its receptive area. Existing array systems have been used in a number of applications. These include teleconferencing [1, 2, 3, 4], speech recognition [5, 6, 7, 8], speaker identification [9], speech acquisition in an automobile environment [10, 11], sound capture in reverberant enclosures [12, 13, 14], large-room recording-conferencing [15], acoustic surveillance [16, 17], and hearing aid devices [18]. These systems also have the potential to be beneficial in several other environments, the performing arts and sporting communities, for instance.

An essential requirement of these sensor array systems is the ability to locate and track a speech source. For audio-based applications, an accurate fix on the primary talker, as well as knowledge of any interfering talkers or coherent noise sources, is necessary to effectively steer the array, enhancing a given source while simultaneously attenuating those deemed undesirable. Location data may be used as a guide for discriminating individual speakers in a multi-source scenario. With this information available, it would then be possible to automatically focus upon and follow a given source on an extended basis. Of particular interest lately, is the application of the speaker location estimates for aiming a camera or series of cameras in a video-conferencing system. In this regard, the automated localization information eliminates the need for a human or number of human camera operators.

In addition to high accuracy, these delay estimates must be updated frequently in order to be useful in practical tracking and beamforming applications. Consider the problem of beamforming to a moving speech source. It has been shown that for sources in close proximity to the microphones, the array aiming location must be accurate to within a few centimeters to prevent high-frequency rolloff in the received signal [19]. An effective beamformer must therefore be capable of including a continuous and accurate location procedure within its algorithm. This requirement necessitates the use of a location estimator

capable of fine resolution at a high update rate. Additionally, any such estimator would have to be computationally non-demanding to make it practical for real-time systems.

1.2 Source Localization Strategies

Existing source localization procedures may be loosely divided into three general categories: those based upon maximizing the output power of a steered-beamformer, techniques adopting high-resolution spectral estimation concepts, and approaches employing only time-difference of arrival (TDOA) information. These broad classifications are delineated by their application environment and method of estimation. The first refers to any situation where the location estimate is derived directly from a filtered, weighted, and summed version of the signal data received at the sensors. The second will be used to term any localization scheme relying upon an application of the signal correlation matrix. The last category includes procedures which calculate source locations from a set of delay estimates measured across various combinations of sensors.

1.2.1 Steered-Beamformer-Based Locators

The first categorization applies to passive arrays for which the system input is an acoustic signal produced by the source. The optimal Maximum Likelihood (ML) location estimator in this situation amounts to a focused beamformer which steers the array to various locations and searches for a peak in output power. Termed 'focalization', derivations of the optimality of the procedure and variations thereof are presented in [20, 21, 22]. Theoretical and practical variance bounds obtained via focalization are detailed in [20, 21, 23] and the steered-beamformer approach was been extended to the case of multiple-signal sources in [24]. The optimality of each of these procedures is dependent upon *a priori* knowledge of

the spectral content of both the primary signal and background noise. However, in practice this information is rarely available. The physical realization of the ML estimator requires the solution of a nonlinear optimization problem. The use of standard iterative optimization methods, such as steepest descent and Newton-Raphson, for this process was addressed by [24]. A shortcoming of each of these approaches is that the objective function to be minimized does not have a strong global peak and frequently contains several local maxima. As a result, this genre of efficient search methods is often inaccurate and extremely sensitive to the initial search location. In [25] an optimization method appropriate for a multimodal objective function, Stochastic Region Contraction (SRC), was applied specifically to the talker localization problem. While improving the robustness of the location estimate, the resulting search method involved an order of magnitude more evaluations of the objective function in comparison to the less robust search techniques. Overall, the computational requirements of the focalization-based ML estimator, namely the complexity of the objective function itself as well as the relative inefficiency of an appropriate optimization procedure, prohibit its use in the majority of practical, real-time source locators.

The practical shortcomings of applying correlation-based localization estimation techniques without a great deal of intelligent pruning is typified by the system produced in [26]. In this work a sub-optimal version of the ML steered-beamformer estimator was adapted for the talker-location problem. A source localization algorithm based on multirate interpolation of the sum of cross-correlations of many microphone pairs was implemented in conjunction with a real-time beamformer. However, because of the computational requirements of the procedure, it was not possible to obtain the accuracy and update rate required for effective beamforming in real-time given the hardware available.

1.2.2 High-Resolution Spectral-Estimation-Based Locators

This second categorization of location estimation techniques includes the modern beamforming methods adapted from the field of high-resolution spectral analysis: autoregressive (AR) modeling, minimum variance (MV) spectral estimation, and the variety of eigenanalysis-based techniques (of which the popular MUSIC algorithm is an example). Detailed summaries of these approaches may be found in [27, 28]. While these approaches have successfully found their way into a variety of array processing applications, they all possess certain restrictions that have been found to limit their effectiveness with the speech-source localization problem addressed here.

Each of these high-resolution processes is based upon the spatiospectral correlation matrix derived from the signals received at the sensors. When exact knowledge of this matrix is unknown (which is most always the case), it must be estimated from the observed data. This is done via ensemble averaging of the signals over an interval in which the sources and noise are assumed to be statistically stationary and their estimation parameters (location in this case) are assumed to be fixed. For speech sources, fulfilling these conditions while allowing sufficient averaging can be very problematic in practice. These algorithms tend to be significantly less robust to source and sensor modeling errors than conventional beamforming methods [29, 30]. The incorporated models typically assume ideal source radiators, uniform sensor channel characteristics, and exact knowledge of the sensor positions. Such conditions are impossible to obtain in real-world environments. While the sensitivity of these high-resolution methods to the modeling assumptions may be reduced, it is at the cost of performance. Additionally, signal coherence, such as that created by a multipath condition, is detrimental to algorithmic performance, particularly that of the eigenanalysis approaches. This situation may be improved via signal processing resources, but again at

the cost of decreased resolution[31]. With regard to the localization problem at hand, these methods were developed in the context of far-field plane waves projecting onto a linear array. While the MV and MUSIC algorithms have been shown to be extendible to the case of general array geometries and near-field sources [32], the AR model and certain eigenanalysis approaches are limited to the far-field, uniform linear array situation. Finally, there arises the issue of computational expense. A search of the location space is required in each of these scenarios. While the computational complexity at each iteration is not as demanding as the case of the steered-beamformer, the objective space typically consists of sharp peaks. This property precludes the use of iteratively efficient optimization methods. The situation is compounded if a more complex source model is adopted (incorporating source orientation or head radiator effects, for instance) in an effort to improve algorithm performance. Additionally, it should be noted that these high-resolution methods are all designed for narrowband signals. They can be extended to wideband signals, including speech, either through simple serial application of the narrowband methods or more sophisticated generalizations of these approaches, such as [33, 34, 35]. Either of these routes extends the computational requirements considerably.

1.2.3 TDOA-Based Locators

With this third localization strategy, the measure in question is not the acoustic data received by the sensors, but rather a set of relative delay estimates derived from the time signals. This approach to finding a source location has been adopted for a variety of applications where a single source may be assumed to be present in the operating environment. These applications range from navigational systems [36, 37] where the TDOA information is calculated from clocking signals transmitted from various known transmitter positions to

sonar devices [38] in which the time delays must be estimated from underwater acoustic signals detected by passive hydrophones. For the locators in this class, the TDOA and sensor position data are used to generate hyperbolic curves which are then intersected in some optimal sense to arrive at a source location estimate. A number of variations on this principle have been developed [39, 40, 41, 42, 43, 44, 45, 46]. They differ considerably in the method of derivation, the extent of their applicability (2-D vs. 3-D, near source vs. distant source, etc.), and their means of solution.

Given solely a set of TDOA figures with known error statistics, obtaining the ML location estimate necessitates solving a set of nonlinear equations. The calculation of this result can be quite cumbersome and computationally expensive, though considerably less so in either of these respects than estimators belonging to the two previously discussed genres. An exact solution is given in [47] for the situation where the number of TDOA estimates is equal to the number of spatial dimensions. However, this solution does not accommodate extra sensor measurements. Iterative methods which start with an initial guess and successively approximate the optimal solution via a local linear least-square (LLS) estimate at each step in the procedure exist [48, 49, 40]. These methods require an LLS matrix calculation at each iteration, are not guaranteed to converge in many instances, and tend to be sensitive to the choice of an initial guess. Finally, there is an extensive class of suboptimal, closed-form location estimators [39, 41, 42, 43, 44, 45, 46, 50, 51, 52, 53, 54, 55] designed to approximate the exact solution to the nonlinear problem. These techniques are computationally undemanding and, in many cases, suffer little detriment in performance relative to their more compute-intensive counterparts.

Regardless of the solution method employed, this third class of location estimation techniques possesses a significant computational advantage over the steered-beamformer or

high-resolution spectral-estimation based approaches. However, TDOA-based locators do present several disadvantages when used as the basis of a general localization scheme. For the case of acoustic sources where a time signal is available, this two-stage process requiring time-delay estimation prior to the actual location evaluation is suboptimal. The intermediate signal parameterization accomplished by the TDOA procedure represents a significant data reduction at the expense of a decrease in theoretical localization performance. However, in real situations the performance advantage inherent in the optimal steered-beamformer estimator is lessened because of incomplete knowledge of the signal and noise spectral content as well as unrealistic stationarity assumptions. In practice, the computational savings afforded by these less intensive procedures can far outweigh the moderate decline in precision. The primary limitation of delay-based locators is their inability to accommodate multi-source scenarios. These algorithms assume a single-source model. The presence of several simultaneous radiators and/or coherent noise sources in the sensor field typically results in ill-defined TDOA figures and unreliable location fixes. A TDOA-based locator operating in such an environment would require a means for evaluating the validity and accuracy of the delay and location estimates.

1.3 Elements of the Speech-Source Localization Problem

This thesis addresses the specific application of source localization algorithms for estimating the position of one or more speech sources in a real-room environment. It is assumed that limited degree of computational resources are available and that the quantity and placement of the sensors are constrained.

A speech source, whether associated with a human talker or mechanical transducer, does not represent an ideal, spherical radiator. In the case of a room-size, near-field environment,

any realistic source possesses a clear degree of directionality and spatial attenuation. This implies that a sensor which the talker is facing will tend to receive a stronger signal than those off to the side or physically behind the source. Similarly, remote sensors will be exposed to a relatively attenuated signal by virtue of the additional propagation distance. Other more subtle factors, such as the room acoustics, non-uniformity of the sensor channels, features of the talker's head and body, as well as the actual content of the speech can introduce deviations from the ideal radiator case and pose serious difficulties to accurately modeling the speech sources.

The computational liabilities and the inability to realistically model the speech sources under a wide variety of conditions prevent the use of either of the first two genres of source locators discussed for this scenario. The approach taken throughout this work will be to employ a two-stage localization procedure; delay estimation followed by a location evaluation. Although Chapter 8 presents a delay estimator specifically intended for this speech source environment, the majority of this thesis will focus on the latter process assuming that the TDOA figures are already available. Studying the problem from this perspective has several clear advantages over the stated alternatives. It is computationally non-intensive and may be parallelized in a straightforward manner. By not being overly dependent upon specific modeling conditions, it is robust and applicable to a range of situations. Furthermore, as will be demonstrated, the shortcomings associated with these techniques, most notably the difficulties with multiple coherent sources, may be overcome in practice through judicious use of appropriate detection methods at each stage in the process.

Each of the localization methods to be presented are based upon a specific source-sensor geometry; the basic unit of which consists of a pair of closely-spaced sensors and a

single delay estimate associated with the potential source. Delay estimates are evaluated exclusively with respect to the particular sensor pair. There is no attempt made to define TDOA values relative to a single reference sensor or an absolute scale. This philosophy is motivated by several arguments. Primarily, in a near-field source environment such as this, source directionality can create significant signal dissimilarities at spatially distant sensors. In the interest of obtaining accurate and reliable TDOA estimates, the individual sensors in each pair must be kept close together. Additionally, as will be shown in the following chapters, the precision of the location estimate is dependent upon the placement of the sensors relative to the actual source location. In general, this may necessitate placing sensors in a wide variety of positions throughout the enclosure. Given only a fixed number of available sensors and the requirement of spatially local sensor pairs, it is not prudent and frequently not possible, to evaluate all the TDOA figures relative to a single sensor location. The sensor-pair geometry advocated in Chapter 2 addresses the problem of source localization given these autonomous sensor pair-TDOA units.

1.4 Scope of This Work

The topics in this thesis have been grouped into two distinct parts. The first set of chapters are devoted to presenting the theoretical foundations of a speech source localization system. These methods take as their input a set of TDOA values and their associated variance figures as estimated across various combinations of sensor pairs. In Chapter 2 a source-sensor geometry appropriate for talkers and sensors in the near-field environment is detailed. Chapter 3 offers and evaluates several error criteria available to the problem. Chapters 4 and 5 provide methods for detecting the presence of a single source and evaluating the error region associated with a given location estimate, respectively. The second set of chapters

gives a performance analysis of these techniques in a real environment as well as illuminating several practical algorithms necessary for a real-time development. Chapter 6 contains some discussion of the computational aspects of these techniques. A novel closed-form locator is the subject of Chapter 7, while Chapter 8 contains the derivation and evaluation of a time-delay estimator intended specifically for the speech source environment. Chapter 9 is the culmination of this thesis, bringing together its individual facets within the context of several experiments incorporating physical systems. Results are presented to illustrate the effectiveness of the proposed source localization techniques as well as confirming the practicality of the theoretical models. Finally, Chapter 10 contains some conclusions and topics for further study.

Part I

Theory

Chapter 2

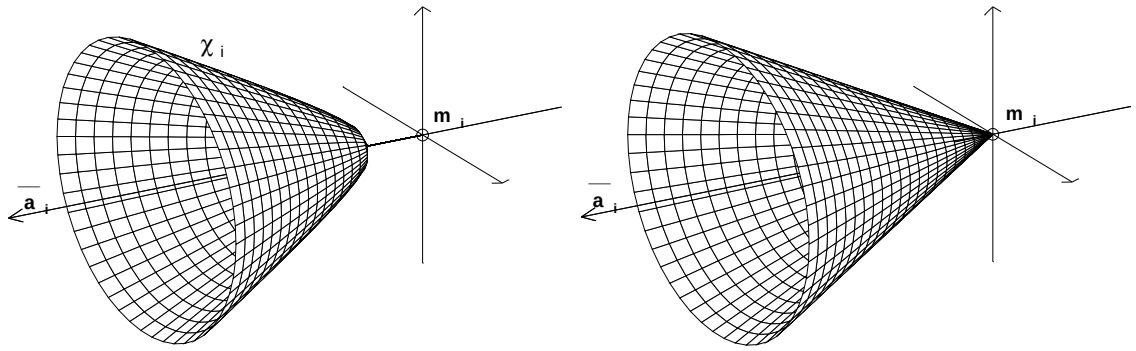
Source-Sensor Geometry

Consider the i^{th} pair of sensors, m_{i1} and m_{i2}, with spatial coordinates (x, y, z) denoted by the vectors, $\mathbf{m}_{i1}, \mathbf{m}_{i2} \in \mathcal{R}^3$, respectively. The unit vector through $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$ is denoted by $\overline{\mathbf{a}_i}$ [1] and $\mathbf{m}_i$ will be used to designate the midpoint of the sensors:

$$\begin{aligned} \overline{\mathbf{a}_i} &= \frac{\mathbf{m}_{i1} - \mathbf{m}_{i2}}{|\mathbf{m}_{i1} - \mathbf{m}_{i2}|} \\ \mathbf{m}_i &= \frac{\mathbf{m}_{i1} + \mathbf{m}_{i2}}{2} \end{aligned} \tag{2.1}$$

where $|\cdot|$ is the Euclidean distance measure. In general, the pressure waves of a signal source radiating in this region will require a specific period of time to propagate to each sensor. Given that the radiator may be modeled as a point source and the medium is uniformly ideal, these propagation times are directly related to the source's distance from the specific sensor. The constant of proportionality being the speed of propagation in the medium, c. (In air the speed of sound is $c \approx 342\frac{\mathrm{m}}{\mathrm{s}}$.) In practice, the absolute propagation times are

[1] In what follows, the notational convention adopted will be to designate vectors of ordered triplets with boldface, lowercase characters while the vectors corresponding to directed lines will be denoted by boldface, lowercase characters with an overline. In each case, the standard vector operations will apply and the difference in application will be clear from the context.



(a) (b)

Figure 2.1: Locus of potential source locations with a fixed delay k for the i^{th} sensor-TDOA combination ($\chi_i = \chi(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \tau_i = k)$) (a) and those for the cone approximation to χ_i (b).

unavailable and only the time difference of arrival (TDOA) relative to the i^{th} sensor pair may be measured.

Given a signal source with known spatial location $\mathbf{s} \in \mathcal{R}^3$, the true TDOA relative to the i^{th} sensor pair will be denoted by $T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})$, and is calculated from the expression:

$$T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s}) = \frac{|\mathbf{s} - \mathbf{m}_{i2}| - |\mathbf{s} - \mathbf{m}_{i1}|}{c} \tag{2.2}$$

The estimate of this true TDOA, the result of a time-delay estimation procedure involving the signals received at sensors m_{i1} and m_{i2}, will be given by τ_i. In practice, the TDOA estimate is a corrupted version of the true TDOA and in general, $\tau_i \neq T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})$.

For the parametric-based localization scheme addressed here, the problem is one of estimating the source location given only the TDOA estimate information. With a single sensor-pair, TDOA-estimate combination the locus of potential source locations is defined to be all spatial locations $\mathbf{s}$ for which the equation $\tau_i = T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})$ is satisfied and will be designated by $\chi(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \tau_i)$, or simply abbreviated by χ_i. The physical constraints of this problem demand that $|c \cdot \tau_i| \leq |\mathbf{m}_{i2} - \mathbf{m}_{i1}|$ and correspondingly the set χ_i is a

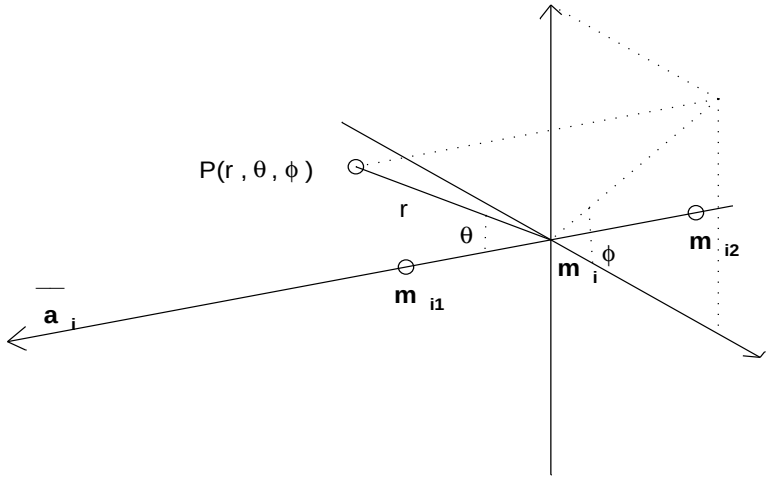


Figure 2.2: Spherical coordinate system defined relative to the sensor pair m_{i1}, m_{i2}.

continuum. In 3-space a plot of χ_i generates one-half of a hyperboloid of two sheets. This hyperboloid is centered about $\mathbf{m}_i$ and has $\mathbf{a}_i$ as its axis of symmetry. This situation is depicted in Figure 2.1a. In general, knowledge of a single sensor-pair, TDOA-estimate combination does not specify a unique source location, it only restricts the potential location to a hyperboloid in 3-space.

In an effort to analyze the nature of χ_i in more detail, spherical coordinate system is established with origin $\mathbf{m}_i$ and $\overline{\mathbf{a}_i}$ as one axis. See Figure 2.2. As a consequence of the symmetry in the following analysis, the remaining axes need not be specified, apart from their orthogonality to $\overline{\mathbf{a}_i}$ and each other. With this system, any point $\mathbf{p} \in R^3$ may be uniquely specified by $P(r, \theta, \phi)$, where r is the range of the point (i.e. its distance from the origin), θ is the angle formed by the base vector to the point and the $\overline{\mathbf{a}_i}$ axis, and ϕ is the angle formed by the projection of the base vector into the plane normal to $\overline{\mathbf{a}_i}$ relative to one of the unspecified axes. The formations to be discussed here all possess a symmetry about the $\overline{\mathbf{a}_i}$ axis and will thus be independent of the coordinate ϕ.

In terms of this spherical coordinate system, the half-hyperboloid of locus points $\mathbf{p} =$

$P(r, \theta, \phi) \in \chi_i$ must satisfy the relation:

$$\frac{\cos^2\theta}{(c \cdot \tau_i)^2} - \frac{\sin^2\theta}{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2 - (c \cdot \tau_i)^2} = \frac{1}{4r^2} \tag{2.3}$$

Note that this equation is independent of the sign of τ_i. It corresponds to a full hyperboloid of two-sheets. To be thorough, the sign of τ_i must be retained to specify the appropriate half of the hyperboloid.

For large r, the hyperboloid asymptotically converges to the cone expressed by:

$$\frac{\cos^2\theta}{(c \cdot \tau_i)^2} - \frac{\sin^2\theta}{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2 - (c \cdot \tau_i)^2} = 0 \tag{2.4}$$

or equivalently:

$$\theta = \cos^{-1}\left(\frac{c \cdot \tau_i}{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|}\right) = \theta_i \tag{2.5}$$

With this coordinate system, a cone with its vertex at the origin is expressed by the simple equation: $\theta =$ constant. For the case of the i^{th} sensor-pair, TDOA-estimate combination, this constant is the *arccosine* of the ratio of the scaled TDOA to the total sensor separation. It may be desirable to express the locus of possible source locations, χ_i, in terms of this single parameter, i.e. approximating the hyperboloid by its corresponding cone. This situation is illustrated in Figures 2.1a and 2.1b. In making such an approximation, the actual locus points (those on the hyperboloid) are displaced in position through the mapping from hyperboloid to cone. Intuitively, these distortions are most extreme for locus points close to the sensors and decrease dramatically for those locations at a greater range. To verify this intuition, consider a constant-ϕ cross-section of a cone-hyperboloid pair as depicted in Figure 2.3.

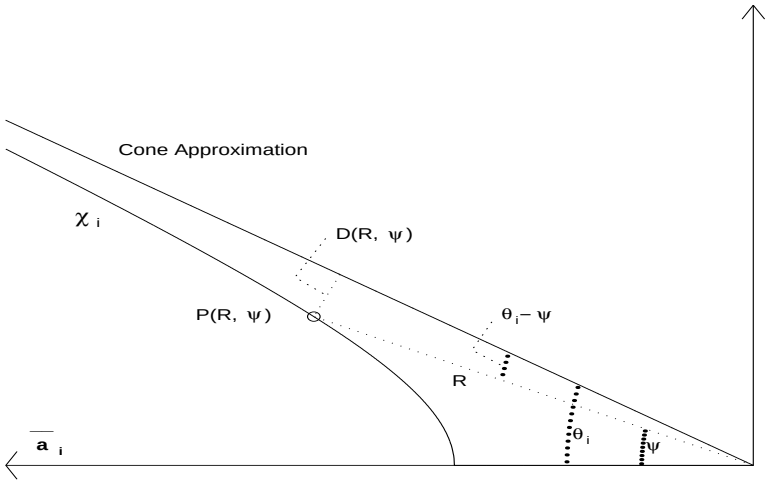


Figure 2.3: Constant-ϕ Cross-section of cone-hyperboloid pair showing the angle distortion, $\theta_i - \psi$, and total distance distortion, $D(R, \psi)$, associated with approximating a point $\mathbf{p} \in \chi_i$ by the the corresponding cone.

Here $\mathbf{p} \in \chi_i$ has coordinates $(r, \theta) = (R, \psi)$ and is a solution of Equation 2.3. The corresponding cone is given by $\theta = \theta_i$. Expressions for the angle distortion, $\theta_i - \psi$, and the total distance distortion, $D(R, \psi) = R \sin(\theta_i - \psi)$, are developed by combining Equations 2.3 and 2.5 to obtain:

$$\frac{\cos^2 \psi}{\cos^2 \theta_i} - \frac{\sin^2 \psi}{\sin^2 \theta_i} = \frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2}{4R^2}$$

which after some work, simplifies to:

$$\sin(\theta_i - \psi) = \frac{\sin^2(2\theta_i) \cdot |\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2}{\sin(\theta_i + \psi) \cdot 16R^2} \tag{2.6}$$

Equation 2.6 relates the angle distortion to the locus point's range and angle. An analysis of this relation, reveals that the angle difference, $\theta_i - \psi$, has minima at $\psi = 0, \frac{\pi}{2}, \pi$ (the end-fire and broadside conditions) and is maximized for $\psi \approx \theta_i \approx \frac{\pi}{4}$. The worst case

angle and total distance distortions are then well-approximated by the expressions:

$$\max\{\theta_i - \psi\} \approx \frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2}{16R^2} \tag{2.7}$$

$$\max\{D(R,\psi)\} \approx \frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2}{16R} \tag{2.8}$$

The maximum angle distortion varies linearly with the square of the sensor separation to source range ratio. Hence, given only an anticipated minimum source range, the worst case errors associated with the cone approximation may be calculated. In practice, these errors are quite small in comparison to the contributions of *noise* associated with the other system parameters. Therefore, a cone approximation to the hyperboloid χ_i is not unreasonable in the majority of situations. Each sensor-TDOA combination may be associated with a single parameter θ_i as given by Equation 2.5 which specifies the angle of the cone relative to the sensor pair axis. For a given source and the i^{th} pair of sensors, the parameter θ_i will be referred to as the i^{th} direction-of-arrival (DOA).

Chapter 3

Localization Error Criteria

Given a set of N sensor-TDOA combinations and their respective loci of potential source locations, χ_i, the problem remains as how to best estimate the true source location, $\mathbf{s}$. Ideally, $\mathbf{s}$ will be an element of the intersection of all the potential source loci ($\mathbf{s} \in \bigcap_{i=1}^{N} \chi_i$). (Note that depending on the number sensor pairs and the choice in their placement, this loci intersection may consist of multiple elements, even under ideal circumstances.) In practice, however, for more than two pairs of sensors this intersection is, in general, the empty set. This disparity is due in part to imprecision in the knowledge of system parameters (TDOA estimate and sensor location measurement errors) and in part to unrealistic modeling assumptions (point source radiator, ideal medium, ideal sensor characteristics, etc.).

With no ideal solution available, we must resort to estimating the source location as the point in $\mathcal{R}^3$ which best *fits* the sensor-TDOA data or more specifically, minimizes an error criterion that is a function of the given data and a hypothesized source location. Limiting our scope to the L_2 (sum-of-squares) norm, three non-linear least squares (LS) error criteria appear applicable to this situation. The first is motivated from a Maximum-Likelihood standpoint and the remaining two are heuristically derived from estimate-dependent dis-

tance measures. As a preliminary to defining these criteria, the variance associated with a DOA estimate is explored.

3.1 DOA Variance

The DOA associated with a pair of sensors, m_{i1} and m_{i2}, and an estimated TDOA for a source, τ_i, is given by Equation 2.5. τ_i is a single realization of a random variable $\mathcal{T}_i$ corresponding to the true TDOA, $T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})$, corrupted by a random noise process, which will be assumed to be additive and zero-mean. In the absence of any other information, the true TDOA is approximated by τ_i and the variance of the r.v., $var\{\mathcal{T}_i\}$, assumed to be available as a byproduct from the delay estimation procedure, is generally a function of the signal-to-noise ratio at the sensor pair.

An exact formulation of the statistics for θ_i requires knowledge of the probability distribution function of $\mathcal{T}_i$. In practice, this is not available. However, if it is assumed that the pdf of $\mathcal{T}_i$ is concentrated near its mean, the moments of θ_i may be approximated in terms of the moments of τ_i [56]. Specifically,

$$var\{\theta_i\} \approx \frac{c^2 \cdot var\{\mathcal{T}_i\}}{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2 \cdot \sin^2(\theta_i)} \tag{3.1}$$

The variance of the DOA is therefore dependent upon the estimated DOA with the minimum occurring in the broadside source case ($\theta_i = \frac{\pi}{2}$) and peaks for the endfire conditions ($\theta_i = 0, \pi$). The above approximation is most appropriate for broadside angles and small TDOA estimation variances. Intuitively, θ_i is least sensitive to the precision of the TDOA estimation procedure for source locations directly in front of the sensor pair.

3.2 The J_{TDOA} LS Error Criterion

The first LS criterion to be considered is a weighted error based upon difference between the TDOA estimates and the ideal TDOA associated with the hypothesized source location, $\mathbf{s}$:

$$J_{TDOA}(\mathbf{s}) = \sum_{i=1}^{N} \epsilon_{itdoa} \cdot [\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})]^2 \tag{3.2}$$

where ϵ_{itdoa} is a weighting figure associated with the i^{th} time delay estimate.

Equation 3.2 is motivated from a probabilistic standpoint. If the time-delay estimates at each sensor pair, τ_i, are assumed to be independently corrupted by zero-mean additive white Gaussian noise with known variance, $var\{\mathcal{T}_i\}$, i.e $\mathcal{T}_i$ is a normally distributed random variable given by:

$$\mathcal{T}_i \sim \mathcal{N}(T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s}), var\{\mathcal{T}_i\})$$

The likelihood function associated with a set of TDOA estimates, $\tau_1, \tau_2, \ldots, \tau_N$, and an hypothesized source location is given by:

$$p(\tau_1, \tau_2, \ldots, \tau_N; \mathbf{s}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\, var\{\mathcal{T}_i\}}} \exp\left(\frac{-(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s}))^2}{2var\{\mathcal{T}_i\}}\right)$$

and the corresponding log-likelihood function is:

$$\ln(p(\tau_1, \tau_2, \ldots, \tau_N; \mathbf{s})) = -\left(\sum_{i=1}^{N} \ln(\sqrt{2\pi\, var\{\mathcal{T}_i\}}) + \frac{(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s}))^2}{2var\{\mathcal{T}_i\}}\right)$$

The Maximum Likelihood (ML) location estimate, $\hat{\mathbf{s}}_{ML}$, is the position which maximizes

$\ln(p(\tau_1, \tau_2, \ldots, \tau_N; \mathbf{s}))$ or equivalently minimizes:

$$\sum_{i=1}^{N} \frac{(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s}))^2}{var\{\mathcal{T}_i\}}$$

This expression is identical to the $J_{TDOA}(\mathbf{s})$ LS error criterion in (3.2) with the weighting figures, ϵ_{itdoa}, replaced to the reciprocal of the TDOA estimate variances. For this reason, the weights are set to

$$\epsilon_{itdoa} = 1/var\{\mathcal{T}_i\} \tag{3.3}$$

and therefore, in the case of time-delay estimates corrupted by additive white Gaussian noise, the minimization of J_{TDOA} yields the ML estimate. i.e.

$$\hat{\mathbf{s}}_{TDOA} = \hat{\mathbf{s}}_{ML} = \arg\min_{\mathbf{s}} J_{TDOA}(\mathbf{s}) \tag{3.4}$$

3.3 The J_{DOA} LS Error Criterion

The second two LS criteria are based upon minimization of $\mathcal{R}^3$ distance measures, rather than the maximization of TDOA-related likelihood function. The first is a weighted error utilizing the differences between the DOA estimates and the true DOA of an hypothesized source location, $\mathbf{s}$, relative to each sensor pair. The true DOA is denoted by $\Theta(\{\mathbf{m}_1, \mathbf{m}_2\}, \mathbf{s})$ and the least-squares error criteria is defined to be:

$$J_{DOA}(\mathbf{s}) = \sum_{i=1}^{N} \epsilon_{idoa} \cdot (d\theta_i)^2 = \sum_{i=1}^{N} \epsilon_{idoa} \cdot [\theta_i - \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})]^2 \tag{3.5}$$

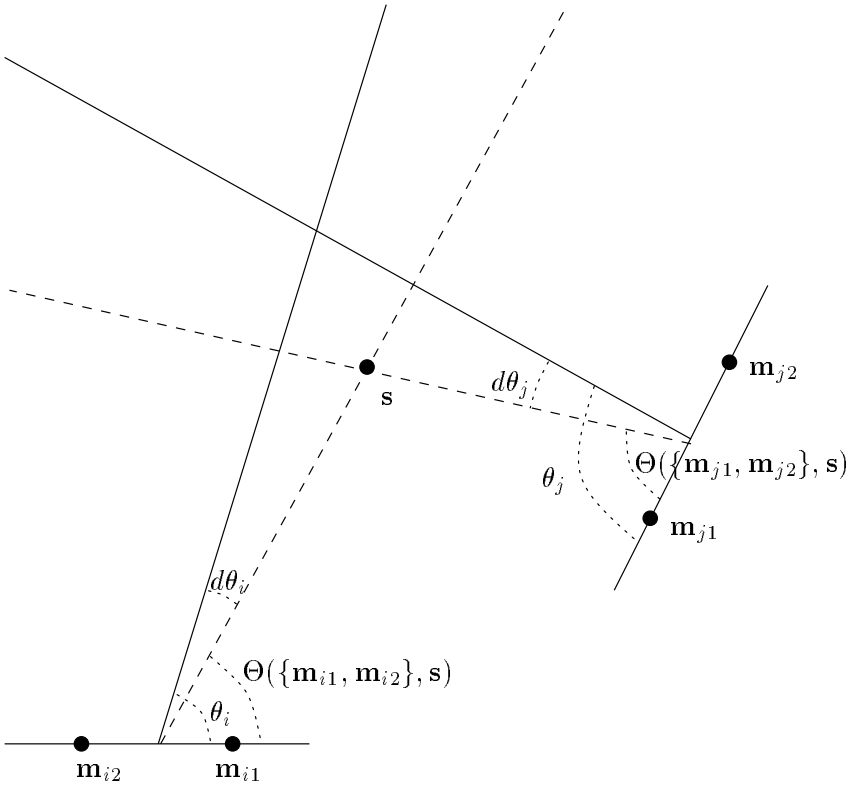


Figure 3.1: Illustration of J_{DOA} LS Error Criterion for two sensor-pair, TDOA-estimate combinations.

where ϵ_{idoa} is a weighting figure associated with the i^{th} DOA estimate. Figure 3.1 illustrates the parameters involved in evaluating J_{DOA} for two pairs of sensors. The dashed lines represent the true DOA's for a source at location $\mathbf{s}$ relative to each sensor pair while the solid lines show the estimated DOA's as determined by (2.5). In each case, the solid line illustrates the intersection of a DOA cone and the plane formed by the hypothesized source and sensor pair. The differences between the estimated and hypothesized angles are labeled with $d\theta$.

The weighting coefficients, ϵ_{idoa} in (3.5), are selected to be the reciprocal of the respective DOA estimate variances (3.1) and may be expressed as

$$\epsilon_{idoa} = 1/var\{\theta_i\} = \frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2 \cdot \sin^2(\theta_i)}{c^2 \cdot var\{\mathcal{T}_i\}} = \frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|^2}{c^2 \cdot var\{\mathcal{T}_i\}} - \frac{\tau_i^2}{var\{\mathcal{T}_i\}} \tag{3.6}$$

The source location estimate found via minimization of the J_{DOA} error is given by

$$\hat{\mathbf{s}}_{DOA} = \arg\min_{\mathbf{s}} J_{DOA}(\mathbf{s}) \tag{3.7}$$

Given TDOA estimates corrupted by additive white Gaussian noise, $\hat{\mathbf{s}}_{DOA}$ does not possess the Maximum-Likelihood property as does the estimator $\hat{\mathbf{s}}_{TDOA}$. However, the J_{DOA} error criterion does have several properties that make it preferable in specific situations. These stem from its use of a distance measure in $\mathcal{R}^3$ and the emphasis provided via its weighting coefficients. Specifically, the J_{DOA} coefficients given by (3.6) place more value on the sensor pairs with large sensor separation and/or small TDOA estimates (corresponding to broadside sources). As (3.1) suggests, these DOA's are proportionately less susceptible to noise in the TDOA estimates from which they are derived. Favoring specific DOA's based upon sensor placement allows the J_{DOA} error criteria to utilize knowledge of the array geometry in addition to the delay-estimate information when evaluating the plausibility of an hypothesized source point. The net effect is to provide the estimator with greater robustness in unfavorable conditions. As will be shown in the analysis to follow, $\hat{\mathbf{s}}_{DOA}$ possesses a performance advantage in situations where the source is off-angle to the array and the TDOA estimates are poor.

3.4 The J_D LS Error Criterion

Finally, we may also consider an error criterion based upon the distance from the hypothesized source to the individual loci of potential source locations:

$$J_D(\mathbf{s}) = \sum_{i=1}^{N} \epsilon_{id} \cdot [\mathrm{D}(\chi_i, \mathbf{s})]^2 \tag{3.8}$$

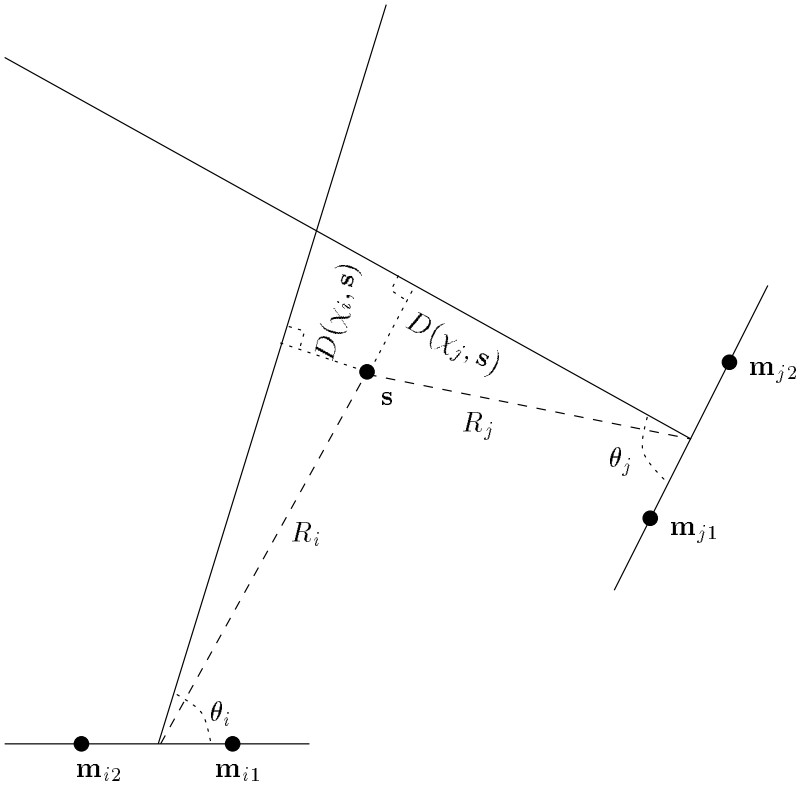


Figure 3.2: Illustration of J_D LS Error Criterion for two sensor-pair, TDOA-estimate combinations.

where $\mathrm{D}(\chi_i, \mathbf{s})$ represents the minimum distance from $\mathbf{s}$ to the locus χ_i. In practice, $\mathrm{D}(\chi_i, \mathbf{s})$ will be calculated by the orthogonal distance from $\mathbf{s}$ to the appropriate cone approximation to χ_i. The J_D LS Error criterion for two sensor-pairs is illustrated in Figure 3.2. Here again, the solid lines represent the estimated DOA's as determined from (2.5). The dotted lines show $\mathrm{D}(\chi_i, \mathbf{s})$ and $\mathrm{D}(\chi_j, \mathbf{s})$, the orthogonal distances from the hypothesized source location to the cone approximations of χ_i and χ_j, respectively. Finally, the dashed lines depict the ranges, R_i and R_j, of the hypothesized source to the midpoint of the sensor pairs.

The orthogonal distance, $\mathrm{D}(\chi_i, \mathbf{s})$ may be calculated from the existing parameters by:

$$\mathrm{D}(\chi_i, \mathbf{s}) = R_i \cdot \sin\left(\theta_i - \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})\right) \tag{3.9}$$

The weighting coefficients, ϵ_{id}, will be calculated using (3.6) in the same fashion as those

for the J_{DOA}, and the J_D-based source location estimate is defined to be

$$\hat{\mathbf{s}}_D = \arg\min_{\mathbf{s}} J_D(\mathbf{s}) \tag{3.10}$$

The J_{DOA} and J_D error criteria are similar in that they both evaluate a distance measure in $\mathcal{R}^3$. However, the J_D criterion, by virtue of the R_i term in (3.9), has a strong tendency to bias the J_D-based estimator $\hat{\mathbf{s}}_D$ to the benefit of hypothesized source locations with small ranges. The effect is to dramatically pull the estimate towards the sensor array. The J_{DOA} error criteria possesses no such dependency on source range and does not exhibit this trend in practice.

A practical consideration that must be addressed is the computational procedure required for the evaluation of these three estimators. Since each of the error criteria that has been presented is a nonlinear function of $\mathbf{s}$, the solutions of (3.4), (3.7), and (3.10) require some form of a numerical search (see Chapter 6 for details); search methods have the potential to be computationally burdensome and problematic due to local minima in the error space.

3.5 An Analysis of the Least-Squares Error Criteria

The properties of these three error criteria were evaluated through a series of Monte Carlo simulations. In each case, a ten-element, bi-linear sensor array as depicted in Figure 3.3 was employed. Sensor spacings were set at 0.5m and the eight pairings of diagonally adjacent sensors (i.e. sensors 1 and 4, 2 and 3, 3 and 6, 4 and 5, etc.) were selected as the sensor pairs used for TDOA calculations. This choice of array geometry and sensor pairings is somewhat arbitrary. The use of a bi-linear array in this case was motivated from its potential use as

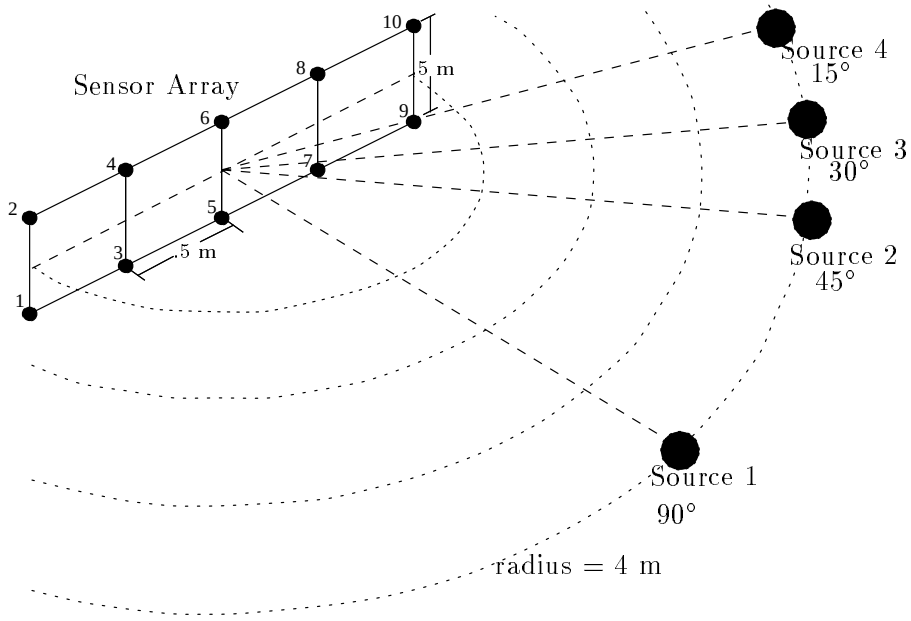


Figure 3.3: Illustration of the experimental set-up used to evaluate the LS location estimators: A 10-sensor planar array with 0.5m spacings and four sources at 90°, 45°, 30°, and 15° with a common range of 4m.

part of a portable teleconferencing unit. The array may be easily deployed at the site and offers reasonable coverage of a typical conference-room table. In general, the design of array geometry for the purposes of source localization and/or sound capture is dependent upon the room environment. The results to be presented here would presumably scale appropriately for different array dimensions and configurations.

The first simulation compared the three LS error-based location estimators using four sources with a common range of 4m and varying bearing angles (90°, 45°, 30°, and 15°) relative to the array center. Figure 3.3 shows this experimental-setup. The true TDOA values for each sensor pair were calculated and then corrupted by additive white Gaussian noise of various power levels. For those instances where the corrupted TDOA value exceeded the maximum time-delay possible for a given sensor pair separation distance (i.e. when $|c \cdot \tau_i| > |\mathbf{m}_{i2} - \mathbf{m}_{i1}|$), the TDOA value in question was set equal to the maximum possible TDOA for that sensor pair.

For each set of corrupted TDOA figures, three location estimates were computed via minimization of the appropriate error criterion. The estimates in (3.4), (3.7), and (3.10) were computed via a search method[1] with the initial guess set equal to the true location. Clearly this is not a practical algorithm since it requires prior knowledge of the actual source location, but for the purposes of comparing the statistical properties of these three estimators it is a computationally reasonable alternative to a more comprehensive search. 100 trials were performed at each of 11 noise levels ranging from a standard deviation the equivalent of 10^{-3}m to 10^{-1}m when scaled by the propagation speed of sound in air ($c \approx 342\frac{\mathrm{m}}{\mathrm{s}}$). The sample means and standard deviations were calculated from the source location DOA and range estimates generated by the three error criteria at each noise condition.

For each estimator, at a constant noise level, the location-estimation accuracy was greatest for the 90° broadside source and progressively declined as the source was moved further toward the endfire condition. All of the estimators exhibited some degree of bias. This bias generally grew as the variance of the additive noise was increased and as the source was moved away from the broadside location. This situation was most extreme for the J_D-based estimator which displayed significantly greater bias in both range and DOA estimation when compared to its J_{TDOA} and J_{DOA}-based counterparts. The reason behind this behavior was alluded to in the previous section and was attributed to the range term in (3.9).

Given this estimator bias, it is more appropriate to consider the root-mean-square error (RMSE) of each estimator rather than the estimators' variance or bias alone. The RMSE is defined by:

$$RMSE[\hat{x}] = \sqrt{E[(\hat{x} - x)^2]}$$

[1]Chapter 6 addresses a number of issues relating to the specific application of nonlinear optimization procedures for the evaluation of these location estimates.

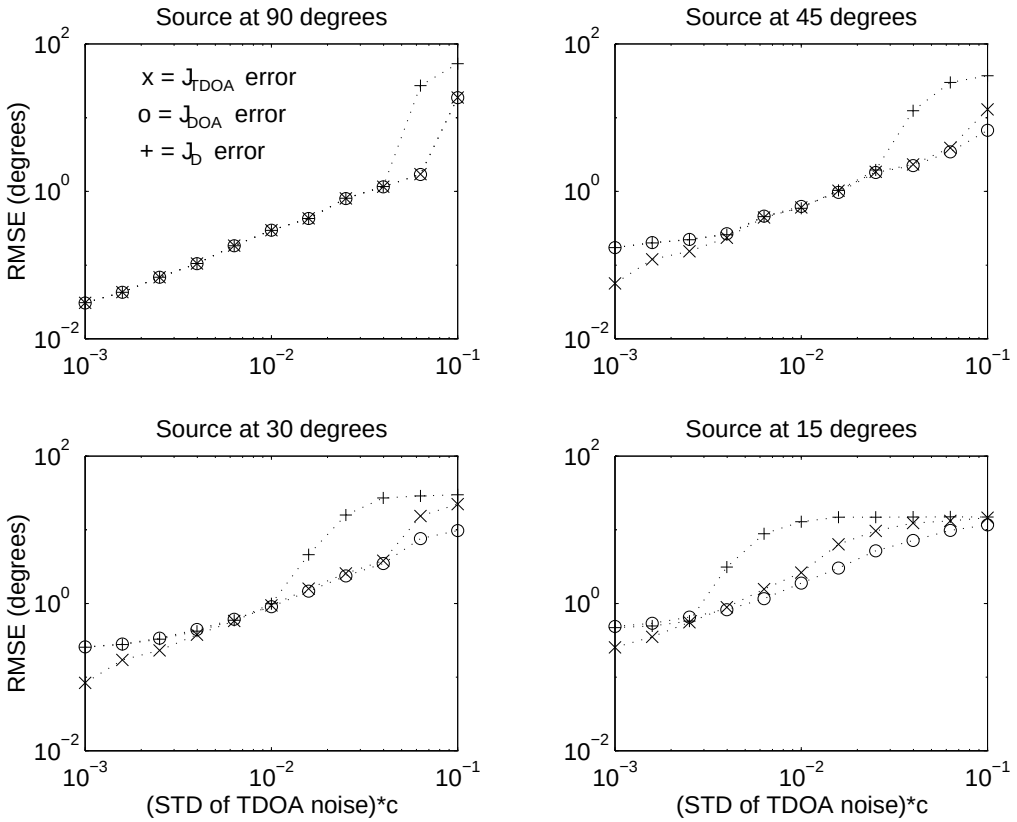


Figure 3.4: Source DOA Estimate RMSE for sources at 90°, 45°, 30°, and 15° relative to the array and common range of 4m. For each plot the x-axis represents the standard deviation of the white Gaussian noise added to the true TDOA's scaled by c. and the y-axis represents the RMSE of the DOA estimate.

where $\hat{x}$ is the estimate of the true value x. In practice, the expectation operator is replaced by the ensemble average. The RMSE can be shown to be equivalent to:

$$RMSE[\hat{x}] = \sqrt{bias\{\hat{x}\}^2 + var\{\hat{x}\}}$$

and thus the RMSE incorporates the tradeoff between bias and various into a single statistic.

Figure 3.4 displays the RMSE results of these Monte Carlo simulations for the source DOA estimates produced by each of the three error criteria. The four graphs correspond to the distinct source locations and in each case the horizontal axis plots the standard

deviation of the added white Gaussian noise scaled by the propagation speed of sound in air. For the broadside source at 90° there is very little to distinguish the performance of these three estimators in the low to moderate noise conditions. However, at the two most extreme noise levels the J_D-based estimator exhibits a marked increase in RMSE value. This distinct 'knee' in the J_D performance line is apparent at all four positions and occurs at progressively smaller noise levels as the source's angle of arrival is decreased. In general, the J_D estimate is by a considerable degree the least robust of the three to the additive noise and DOA conditions. The J_{TDOA} and J_{DOA} estimators display a specific trend as well. At low noise levels, the J_{TDOA}-based estimate, which is the ML estimate in this case, possesses a distinct performance advantage over the J_{DOA} estimate. However, with the higher noise levels this situation is reversed and the J_{DOA} is superior. The performance crossover point occurs at lower noise levels the more endfire the source is positioned.

The preceeding results presented the RMSE values of the source DOA estimates. A similar analysis with the range estimates, does not reveal as distinct differences between the respective estimators. While the J_D estimate possesses an extreme bias towards the array origin, this is countered by a small range variance. Conversely, the remaining two maintain little bias, but do have a significantly greater range estimate variance. The net effect is to produce roughly equivalent range RMSE values for all three estimators.

Based upon these results a second simulation was performed, this time fixing the source DOA and noise level to 15° and .01m, respectively, and allowing the source range to vary from 2 to 10m. The DOA estimate RMSE results are displayed in Figure 3.5. At a roughly constant 2° RMSE, the J_{DOA}-based estimator offers consistently better performance than its counterparts. The J_{TDOA}-based estimator does slightly worse, particularly at close range, while the J_D-based estimator quickly climbs to a peak RMSE value of 14° for this

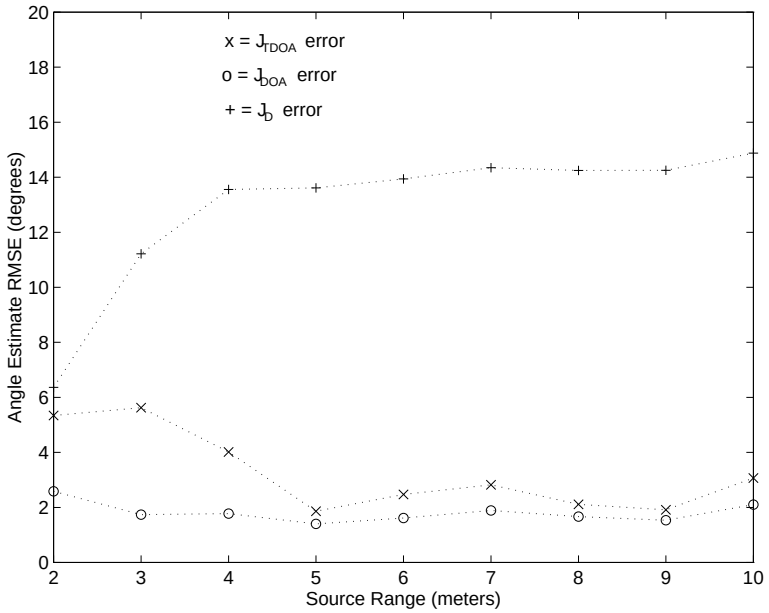


Figure 3.5: Source DOA Estimate RMSE for sources at 15° relative to the array and ranges varying from 2 to 10m. Standard deviation of TDOA estimate noise scaled by c is fixed at .01m for each trial. The x-axis represents the source range and y-axis represents the RMSE of the DOA estimate.

15° source. The J_D error appears to be quite sensitive to the true source range, the J_{TDOA} less so in this respect, and the J_{DOA} error very little at all. The independence of source DOA estimate precision from the source range is a desirable estimator property, particularly in applications where only the source's bearing is of interest (pointing some cameras, for instance).

To summarize the results of these simulations: For broadside sources and clean TDOA estimates, the location estimate $\hat{s}_{TDOA}$, which is based entirely on a least-squares error criterion employing the time-delay estimates alone, proved advantageous. However, under less favorable noise conditions and with sources located off-angle to the sensor array, the DOA-based location estimator, $\hat{s}_{DOA}$, appears to offer preferable performance. None of the data available advocated the use of the distance-based estimator, $\hat{s}_D$, and it will not be considered further.

Chapter 4

Detection of Sources

The source-sensor-TDOA model which has been employed for the source localization problem may also be incorporated with statistical hypothesis-testing procedures to produce a means for detecting the presence of a signal source. Given TDOA estimates which are assumed to be samples from mutually uncorrelated, Gaussian random variables, the TDOA-based LS error criterion, J_{TDOA}, is shown to be the basis for a probabilistically optimal detection process. The specifications of the decision rules are dependent upon the source/non-source model adopted, three of which will be considered here. The first scenario includes specific models for the TDOA estimates under both the source and non-source conditions. The resulting detection rule corresponds to a binary hypothesis test. In the second scenario, no assumptions are made with regard to the nature of the TDOA estimates during periods when no signal source is present. Instead of attributing the observations to a particular hypothesis, the consistency of the source model and the data is evaluated. For the final scenario, no statistical modeling assumptions are adopted and an empirical detection test is presented.

4.1 Source/Non-Source Modeling

The purpose of this chapter is to provide a method of identifying when a signal source is present in the radiation field of the sensors. When appropriate statistical models are available, this is accomplished using a source/non-source decision process. In this context the term "source" will refer to a single, radiating source which has presumably been effectively located via one of the estimation procedures developed in the preceding chapter. The label "non-source" will be applied to any location estimate for which the "source" condition is not satisfied, most notably incorrect location estimates or those estimates produced during periods of no source activity. The development of a practical source/non-source model requires knowledge of the system application environment and the performance specifications of the delay estimator responsible for generating the TDOA estimates and variance figures.

With regard to environment, issues that must be addressed are the number of potential sources and the nature of the background signal during non-source periods. In the simplest case, the source/non-source model may be be reduced to well-defined single-source and silence hypotheses. With more complex situations, the non-source condition may include instances of radically errant location estimates, multiple-simultaneous signal sources, and silence periods with incomplete or unknown statistics.

The source localization and detection procedures depend principally upon the TDOA information. In many instances, the time-varying nature of the source necessitates the use of a delay estimator responsive to short-term signal characteristics. For a delay estimate to be accurate and meaningful, the analysis window associated with a single delay estimate must be small enough to assure that the signal is statistically stationary throughout the analysis time interval. The appropriate time-interval limit is dependent upon the nature of the signal source. In the case of speech, this time frame is on the order of 20ms to 30ms. An additional

factor is that frames of speech will be interspersed with periods of silence. Ideally, the delay estimator will be capable of producing independent estimates on a frame-by-frame basis. For those frames containing speech, the TDOA value is a reflection of the source DOA and the variance is a function of the signal content and the SNR conditions. During periods of silence, the TDOA statistics are not clearly defined. Presumably, in the absence of a source, the background noise at the sensors is uncorrelated and the resulting TDOA estimates will possess a zero mean. If the estimate variance is calculated on a short-term basis as well, the variance figure reported with the TDOA estimate will correspond to the silence condition. However, if the variance value is evaluated over several analysis frames, some of which may include periods of source activity, the variance term is not appropriate for modeling the silence regions.

Two source/non-source models will be addressed here. The first assumes that a very simple scenario is appropriate for the application. Only one signal source is operating at a time and the delay estimator is capable of reporting accurate variance figures during both source and non-source periods. This model, referred to as the "binary source/silence model", is investigated in scenario #1. The second model is more general, assuming that TDOA statistics are valid only during "source" periods. The "non-source" condition is left unspecified. The subject of scenario #2, this situation is termed the "source-only model". Scenario #3 involves an alternative, non-statistical approach. In the absence of any clear statistical models, the physical clustering of the estimated DOA bearings relative to the estimated source is adopted as a detection measure.

4.2 Scenario # 1: Binary Source/Silence Model

4.2.1 Binary Hypothesis Testing

For a set of observational data and a number of probabilistic models which may have produced the data, statistical hypothesis testing provides a systematic means for identifying the appropriate model and quantifying the confidence of this choice [57]. In the event that only two specific, statistically well-defined hypotheses are considered:

$$(H_0) \text{ "non-source"}$$

$$(H_1) \text{ "source"}$$

the hypothesis selection may be accomplished using a binary hypothesis test. In the absence of any prior detection probabilities or costs associated with misclassified decisions, the decision rule will be derived from the Neyman-Pearson criterion. The objective behind this approach is to select an appropriate false-alarm probability (P_F) and then determine a decision strategy that obtains this value while simultaneously maximizing the probability of detection (P_D). In this context, P_F associated with a decision rule is defined as the probability that the source is identified as present when it is not (H_1 selected when H_0 is true), and P_D is the corresponding probability that the source is identified as present when it is (H_1 selected when H_1 is true). A decision rule which satisfies these restrictions is referred to as the "most powerful test" of the hypothesis H_1 with respect to the alternative H_0. Let $p_0(\mathbf{y})$ and $p_1(\mathbf{y})$ be the joint probability density functions for an observation set $\mathbf{y}$ under hypotheses H_0 and H_1, respectively. The likelihood-ratio, $\Lambda(\mathbf{y})$, appropriate for this binary decision is defined by:

$$\Lambda(\mathbf{y}) = \frac{p_1(\mathbf{y})}{p_0(\mathbf{y})}$$

A theorem attributed to Neyman and Pearson [58] shows that the most powerful test with the false-alarm constraint ($P_F = \alpha$) is found from the likelihood-ratio test:

$$\text{if} \quad \begin{cases} \Lambda(\mathbf{y}) \geq \lambda & \text{accept} \quad H_1 \\ \Lambda(\mathbf{y}) < \lambda & \text{accept} \quad H_0 \end{cases}$$

$$\text{where} \qquad Pr(\Lambda(\mathbf{y}) \geq \lambda \mid H_0) = \int_{\lambda}^{\infty} P_0(\Lambda) d\Lambda = \alpha$$

$P_0(\Lambda)$ is the probability density function of the random variable $\Lambda(\mathbf{y})$ under hypothesis H_0. The maximum probability of detection obtained by this optimal test is:

$$P_D = Pr(\Lambda(\mathbf{y}) \geq \lambda \mid H_1) = \int_{\lambda}^{\infty} P_1(\Lambda) d\Lambda$$

where $P_1(\Lambda)$ is the probability density function of $\Lambda(\mathbf{y})$ under H_1.

The above test is frequently expressed in terms of a monotonic function $G = G(\Lambda)$ of the likelihood ratio. Assuming, without loss of generality, this function to be increasing, the likelihood test is rewritten as:

$$\text{if} \quad \begin{cases} G(\Lambda) \geq G_0 & \text{accept} \quad H_1 \\ G(\Lambda) < G_0 & \text{accept} \quad H_0 \end{cases}$$

$$\text{where} \qquad Pr(G(\Lambda) \geq G_0 \mid H_0) = \int_{G_0}^{\infty} P_0(G) dG = \alpha$$

and the corresponding P_D is found from:

$$P_D = Pr(G(\Lambda) \geq G_0 \mid H_1) = \int_{G_0}^{\infty} P_1(G) dG$$

Here $P_0(G)$ and $P_1(G)$ are the pdf's of $G(\Lambda)$ under each of the respective hypotheses.

4.2.2 Binary Source Detection Test

The source/silence model assumes that for the set of N sensor-pairs, the TDOA estimates $(\tau_1, \tau_2, \ldots, \tau_N)$ and their associated variances $(var\{\mathcal{T}_1\}, var\{\mathcal{T}_2\}, \ldots, var\{\mathcal{T}_N\})$ are available, and that the source location estimate, $\hat{\mathbf{s}}$, has been evaluated. With regard to the earlier discussion, the TDOA estimates are further assumed to be observations from independent, Gaussian processes with the hypothesis-dependent parameters:

$$(H_0): \quad \text{“silence”} \quad \mathcal{T}_i \sim \mathcal{N}(0, var\{\mathcal{T}_i\})$$

$$(H_1): \quad \text{“source”} \quad \mathcal{T}_i \sim \mathcal{N}(T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}), var\{\mathcal{T}_i\})$$

In each case the reported variance figure is assumed to be consistent with either the “source” or “silence” conditions. The hypotheses are distinguished only by their differing mean values.

The likelihood-ratio for these binary hypotheses is given by:

$$\begin{aligned}\Lambda(\tau_1, \tau_2, \ldots, \tau_N) = \frac{p_1(\tau_1, \tau_2, \ldots, \tau_N)}{p_0(\tau_1, \tau_2, \ldots, \tau_N)} &= \frac{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi var\{\mathcal{T}_i\}}} \exp\left(\frac{-(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}))^2}{2\, var\{\mathcal{T}_i\}}\right)}{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi var\{\mathcal{T}_i\}}} \exp\left(\frac{-\tau_i^2}{2\, var\{\mathcal{T}_i\}}\right)} \\ &= \frac{\prod_{i=1}^{N} \exp\left(\frac{-(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}))^2}{2\, var\{\mathcal{T}_i\}}\right)}{\prod_{i=1}^{N} \exp\left(\frac{-\tau_i^2}{2\, var\{\mathcal{T}_i\}}\right)}\end{aligned}$$

Defining the function $G(\Lambda) \equiv -2\ln(\Lambda)$, the expression is reduced to the sufficient statistic:

$$G(\tau_1, \tau_2, \ldots, \tau_N) = \sum_{i=1}^{N} \frac{(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}))^2}{var\{\mathcal{T}_i\}} - \sum_{i=1}^{N} \frac{\tau_i^2}{var\{\mathcal{T}_i\}}$$

$$= \quad J_{TDOA}(\hat{\mathbf{s}}) - \sum_{i=1}^{N} \frac{\tau_i^2}{var\{\mathcal{T}_i\}} \tag{4.1}$$

and the optimal likelihood test becomes:

$$\text{if} \quad \begin{cases} G(\tau_1, \tau_2, \ldots, \tau_N) \leq G_0 & \text{accept} \quad H_1 \\ G(\tau_1, \tau_2, \ldots, \tau_N) > G_0 & \text{accept} \quad H_0 \end{cases}$$

$$\text{where} \qquad Pr(G(\Lambda) \leq G_0 \mid H_0) = \int_{-\infty}^{G_0} P_0(G) dG = \alpha \tag{4.2}$$

Note that the inequalities have been reversed as a result using a monotonically decreasing transformation function $G(\Lambda)$. The probability of detection given the false-alarm constraint is:

$$P_D = Pr(G(\Lambda) \leq G_0 \mid H_1) = \int_{-\infty}^{G_0} P_1(G) dG \tag{4.3}$$

In order to determine the test threshold, G_0, and the resulting detection probability, the pdf of the test statistic $G(\tau_1, \tau_2, \ldots, \tau_N)$ must be obtained. Simplifying (4.1) yields:

$$G(\tau_1, \tau_2, \ldots, \tau_N) = \sum_{i=1}^{N} \frac{T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})^2 - 2\tau_i \; T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})}{var\{\mathcal{T}_i\}}$$

The statistic is a linear combination of Gaussian variables and therefore a Gaussian variable itself under each of the hypotheses. Specifically,

$$(H_0): \quad \text{“silence”} \quad G \sim \mathcal{N}\Bigg(\overbrace{\sum_{i=1}^{N} \frac{T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})^2}{var\{\mathcal{T}_i\}}}^{+m}, \; \overbrace{4 \sum_{i=1}^{N} \frac{T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})^2}{var\{\mathcal{T}_i\}}}^{s^2} \Bigg)$$

$$(H_1): \quad \text{“source”} \quad G \sim \mathcal{N}\Bigg(\underbrace{-\sum_{i=1}^{N} \frac{T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})^2}{var\{\mathcal{T}_i\}}}_{-m}, \; \underbrace{4\sum_{i=1}^{N} \frac{T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})^2}{var\{\mathcal{T}_i\}}}_{s^2} \Bigg) \tag{4.4}$$

Evaluating (4.2) and (4.3) requires the calculation of the area under a Gaussian probability density function. While no closed-form solution to this problem exists, values of the cumulative unit normal distribution function are available via numerical integration methods. This function is defined by:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-x^2}{2}\right) dx$$

and $\Phi(z)$ may be interpreted as the probability that a unit-mean, unit variance Gaussian variable is less than z. In terms of this notation, solving (4.2) for the detection threshold yields:

$$\begin{aligned} \int_{-\infty}^{G_0} P_0(G)dG = \Phi\left(\frac{G_0 - m}{s}\right) &= \alpha \\ \Longrightarrow \qquad G_0 &= s \cdot \Phi^{-1}(\alpha) + m \end{aligned}$$

where $\Phi^{-1}(x)$ denotes the inverse function of $\Phi(z)$. The detection probability (4.3) is then:

$$\begin{aligned} P_D &= \int_{-\infty}^{G_0} P_1(G)dG \\ &= \Phi\left(\frac{G_0 + m}{s}\right) = \Phi\left(\Phi^{-1}(\alpha) + \frac{2m}{s}\right) \\ &= \Phi\left(\Phi^{-1}(\alpha) + \sqrt{m}\right) \end{aligned} \tag{4.5}$$

Figure 4.1 illustrates the relationship between the G statistic's probability distributions under each of the hypotheses and the calculation of the detection and false-alarm probabili-

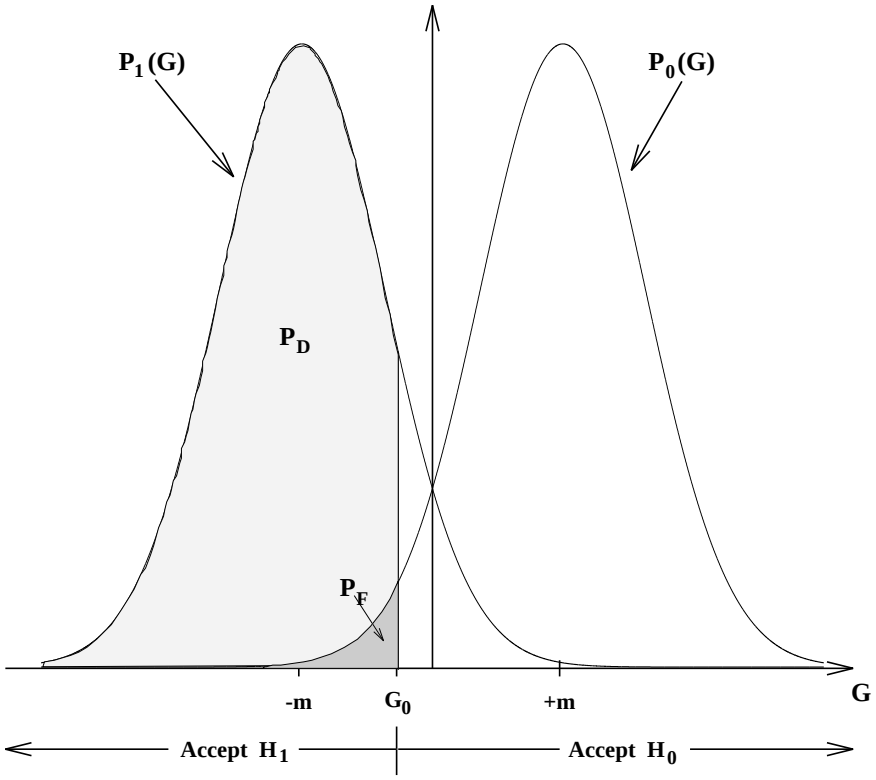


Figure 4.1: Binary Hypothesis Test: The probability distribution functions for the statistic G under each hypothesis are illustrated. The detection (P_D) and false-alarm (P_F) probabilities are indicated by the shaded regions under each curve and to the left of the decision threshold (G_0).

ties. P_D and P_F are indicated by the area under the curves of $P_1(G)$ and $P_0(G)$, respectively, for values of G less than the decision threshold G_0. Note that both probabilities exhibit monotonic growth(decay) as G_0 is increased(decreased).

The false-alarm and detection probabilities are functions of the mean (m) and variance (s^2) of the statistic G as defined in (4.4). These figures are directly related to the source location estimate and the TDOA variances. As (4.5) indicates, for a fixed false-alarm rate, the probability of detection improves as m increases. In general, this implies that a signal source at a location possessing small TDOA values with respect to the sensor pairs is more difficult to distinguish from silence than the same source found at a location with larger TDOA values. For a linear array, this means that end-fire sources have more favorable

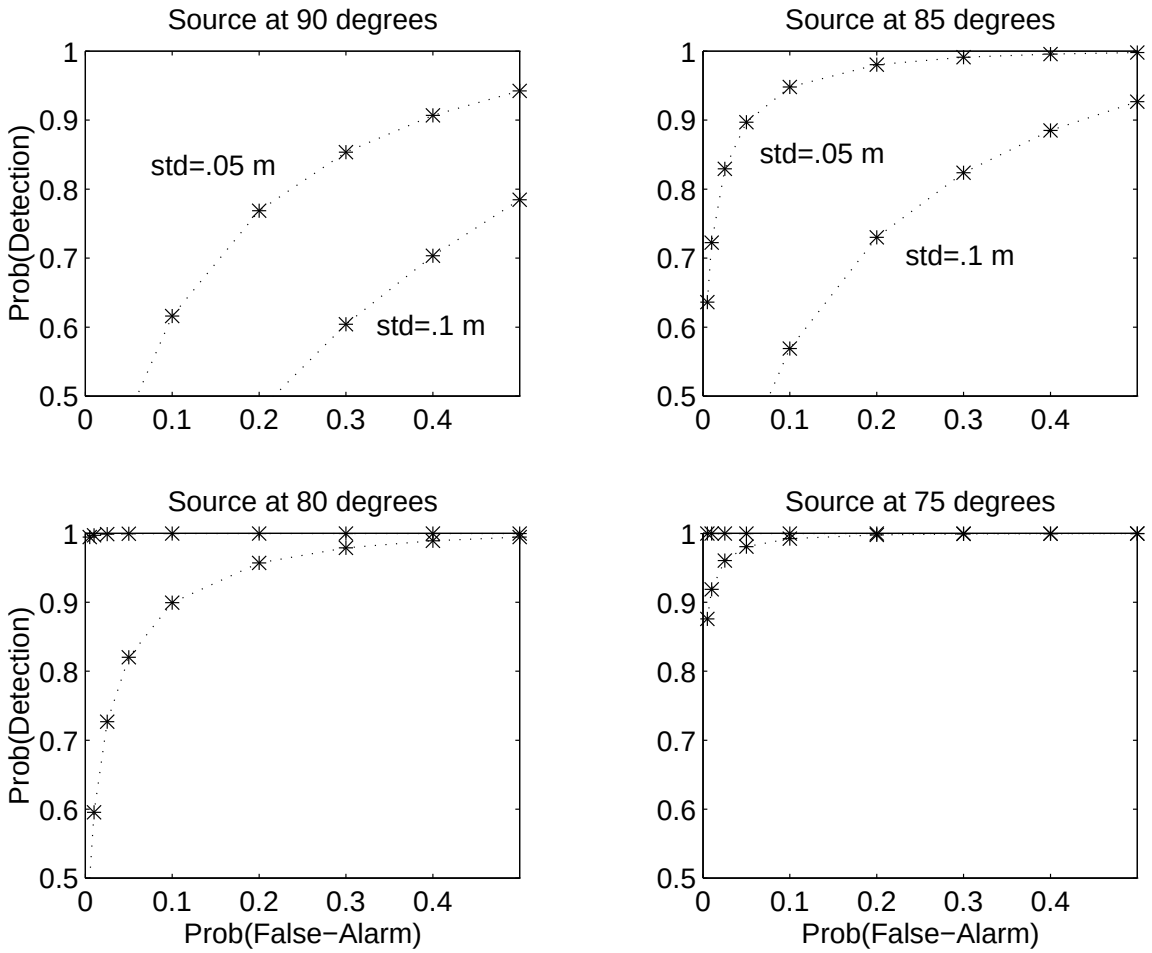


Figure 4.2: Binary Hypothesis Test: Plots of Probability of Detection versus Probability of False-Alarm for two sets of TDOA estimate variances. The four source locations are each at a range of 10m from the 10-element bilinear array with bearing angles varying from 90° (broadside) to 75°.

detection statistics than broadside ones.

To exhibit this phenomena, the ten-element bi-linear depicted in Figure 3.3 and used for the experiments in Section 3.5 was reemployed for simulations involving four different source locations. The sources were placed at a range of 10m relative to the midpoint of the array and at the same height as the array mid-line. The location bearing angles were begun at broadside (90°) and varied in 5° decrements. Sensor-pair selection was done the same as Section 3.5. Figure 4.2 plots the receiver operating characteristic (ROC), P_D versus P_F, for each location and two sets of TDOA estimate variance levels. The source at 90°

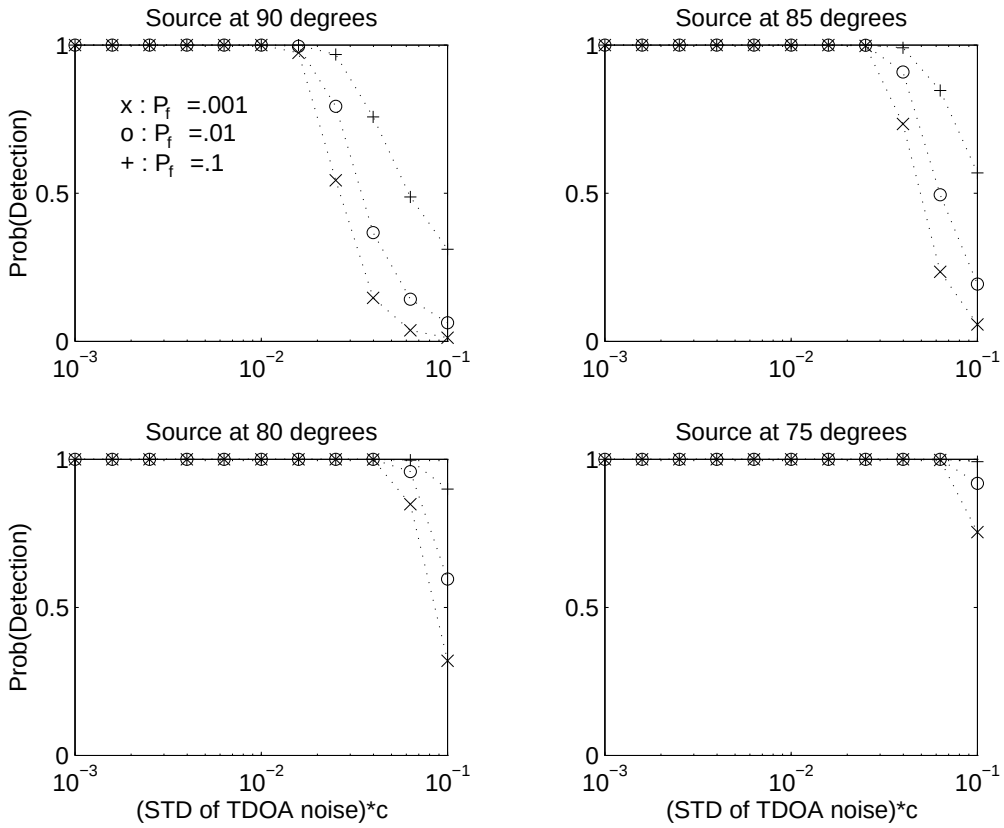


Figure 4.3: Binary Hypothesis Test: Plots of Probability of Detection versus standard deviation of TDOA estimates scaled by c (in meters) for three false-alarm probabilities: P_F=.001, .01, and .1. The four source locations are each at a range of 10m from the 10-element bilinear array with varying angles.

presents the worst-case location at this range. In this situation there is a relatively low detection probability associated with each P_F value at these variance levels. As the source is moved from broadside, the P_D values quickly grow until the ROC curve is nearly pinned at $P_D \approx 1.0$ for the 75° location. Note that the source range (10m) and TDOA estimate variances are quite large in comparison to those used in Section 3.5. These conditions were selected in an effort to highlight the location dependence inherent in this binary hypothesis test. With more moderate variance levels and a smaller range, the ROC curves would all appear nearly flat at $P_D \approx 1.0$ (including the worst-case broadside source) and thus unenlightening.

The properties of this binary hypothesis test were further analyzed through a second set of simulations. This time the probability of detection for the four source locations was evaluated as a function of the TDOA estimate variance while the false-alarm rate was held constant. Figure 4.3 displays the plots of P_D versus the TDOA variance for P_F values of .001, .01, and .1. Each of these curves is flat at $P_D \approx 1.0$ for small TDOA variances and possesses a distinct knee as the variance is increased. The robustness of the test improves as the source is moved from broadside, but even with the worst-case source at this relatively large range, the detector performance does not degrade until the TDOA variance has increased to over 10^{-2}m.

Given a situation in which the source/silence model assumed for this scenario is appropriate, the binary hypothesis test presented provides an effective means for assessing the validity of a source location estimate. Further, the significance of this detection decision may be evaluated from the sensor geometry and the location estimate information. For most practical applications, this confidence level is quite high.

4.3 Scenario # 2: Source-Only Model

4.3.1 Model Consistency Testing

In the previous scenario, the "silence" hypothesis modeled the non-source frames as periods of no source activity with known statistics. As discussed earlier, the delay estimate variance figures may not be valid for the silence frames and the binary hypothesis test presented there would be non-applicable. Furthermore, with many situations, the source-silence dichotomy is inappropriate. An uncorrelated, zero-mean background noise may not be the only alternative to a single source. Consider the case of several simultaneous sources. A delay

estimator which does not distinguish this situation will tend to produce a single TDOA estimate that is a weighted mixture of the individual delay figures. The TDOA estimates in this case would not correspond to either of the hypotheses of the previous scenario[1]. For these reasons, it is desirable to have a hypothesis test which attempts to validate the consistency of the "source" hypothesis given the TDOA information without pronouncing a more favorable description. This is the goal of model consistency testing.

Given a set of observational data, $\mathbf{y}$, and a statistical model believed to be responsible for generating this data, two hypotheses are defined:

(H_0) "observations are not consistent with model"

(H_1) "observations are consistent with model"

With hypotheses specified in this manner, identifying the "most powerful test" of H_1 with respect to H_0 is not possible. In the absence of any knowledge of the alternative hypotheses, the approach taken is to select the smallest acceptance region satisfying a fixed probability of detection constraint, $P_D = \beta$. The acceptance region represents the set of observations $\mathbf{y}$ that is maximally consistent with the H_1 hypothesis subject to the $P_D = \beta$ restriction while attempting to minimize the false-alarm probability associated with the unknown alternatives. Denoting the acceptance region by $\mathcal{R}_1$, the decision rule may be expressed as:

$$\text{if} \quad \begin{cases} \mathbf{y} \notin \mathcal{R}_1 & \text{accept} \quad H_0 \\ \mathbf{y} \in \mathcal{R}_1 & \text{accept} \quad H_1 \end{cases}$$

[1]The localization of multiple-sources will be addressed in Chapter 9 in the context of competing speech sources.

where $\mathcal{R}_1$ is found from the solution of the constrained minimization problem:

$$\min_{\mathcal{R}_1} \int_{\mathcal{R}_1} d\mathbf{y} \qquad \text{subject to} \qquad Pr(\mathbf{y} \in \mathcal{R}_1 \mid H_1) = \int_{\mathcal{R}_1} p_1(\mathbf{y}) d\mathbf{y} = \beta \tag{4.6}$$

4.3.2 Source Consistency Test

The "source-only model" may be expressed as:

$$(H_0): \quad \text{"source not valid"} \quad \mathcal{T}_i \not\sim \mathcal{N}(T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}), var\{\mathcal{T}_i\})$$

$$(H_1): \quad \text{"source valid"} \quad \mathcal{T}_i \sim \mathcal{N}(T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}), var\{\mathcal{T}_i\})$$

In the event of a single source, with estimated location $\hat{\mathbf{s}}$, the TDOA estimates are again assumed to be observations from uncorrelated, Gaussian random variables with known mean and variance. However, no assumptions are made concerning the nature of the alternative hypothesis.

Instead of investigating the observation space consisting of the TDOA estimates themselves, it is advantageous to consider the one-dimensional statistic $F = F(\tau_1, \tau_2, \ldots, \tau_N)$ based upon the J_{TDOA} error criterion:

$$F(\tau_1, \tau_2, \ldots, \tau_N) = \sum_{i=1}^{N} \frac{(\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}))^2}{var\{\mathcal{T}_i\}} = J_{TDOA}(\hat{\mathbf{s}})$$

Under the H_1 hypothesis, F is a random variable possessing a chi-squared distribution with N-degrees of freedom ($F \sim \chi^2_N$). Since the chi-squared pdf is unimodal, the acceptance region $\mathcal{R}_1$ will be the closed interval $[a, b]$ and the constrained minimization problem may be solved analytically. Applying Lagrange multipliers to (4.6) indicates that the probability

density function values associated with the optimal acceptance region must satisfy [59]:

$$p_{\chi^2_N}(a) = p_{\chi^2_N}(b)$$

where $p_{\chi^2_N}(x)$ is the pdf associated with the χ^2_N distribution given by:

$$p_{\chi^2_N}(x) = \frac{1}{2^{N/2}\Gamma(N/2)} x^{(N-2)/2} e^{-x/2}$$

Finding the acceptance region endpoints then requires searching $p_{\chi^2_N}$ for equal values and expanding/contracting the endpoints until the required probability, $P_D = \beta$ is contained within the interval. Defining the cumulative N-degree of freedom chi-square distribution function as:

$$\Psi(\chi^2; N) = \int_0^{\chi^2} p_{\chi^2_N}(x)dx$$

the final decision rule may be written as:

$$\text{if} \quad \begin{cases} F < a \quad \text{or} \quad F > b & \text{“source not valid”} \\ a \leq F \leq b & \text{“source valid”} \end{cases}$$

with a and b calculated from:

$$p_{\chi^2_N}(a) = p_{\chi^2_N}(b) \qquad \text{such that} \qquad \Psi(b; N) - \Psi(a; N) = \beta \tag{4.7}$$

With no alternative hypotheses assumptions, an analysis of the test's operating characteristics is less straightforward than in the previous scenario. To illustrate some of the consistency test properties, four potential signal situations were created and simulated using the 10-element bilinear array with its eight sensor-pairs. Case A corresponds to the “source

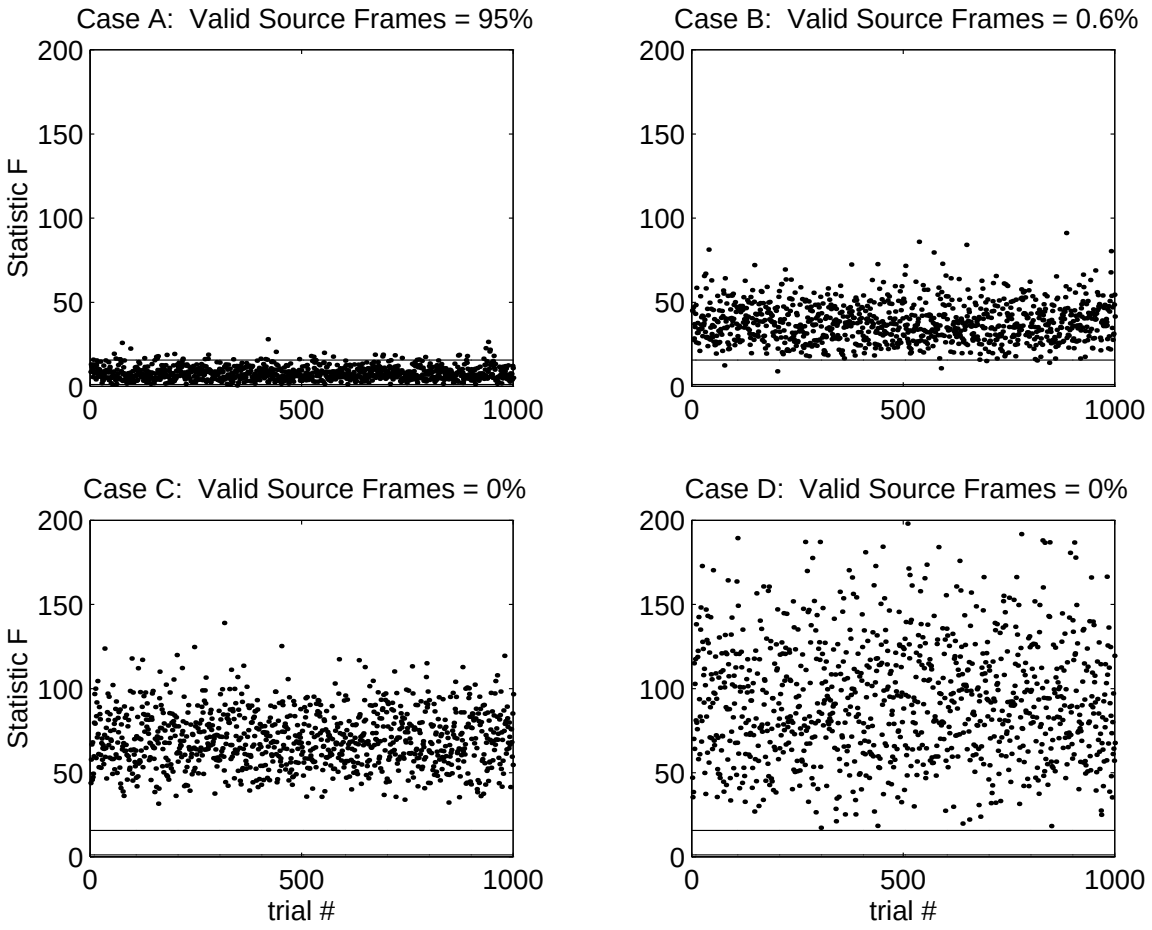


Figure 4.4: Hypothesis consistency test for four signal situations with a source estimated to be broadside to the 10-element bilinear array and at a range of 10m. Each plot presents the statistic F for a 1000 trials under each signal condition. The acceptance region $[1.2, 15.7]$ (shown as a horizontal region in each plot) was calculated from (4.7) with a detection constraint of $\beta = .95$. "Valid Source Frames" refers to the percentage of trials falling into the acceptance region. The signal situations represent: (Case A) single active source, (Case B) two active sources, (Case C) no source and correct variance figures, and (Case D) no source and underestimated variance figures.

valid" hypothesis, while Cases B-D represent various signal circumstances inconsistent with the valid source model. For each case, the estimated signal location, $\hat{\mathbf{s}}$, was assumed to be broadside to the sensor array at a range of $10m$. This is identical to the 90° source in the scenario #1 simulations. The signal situations are described below:

Case A: Valid source. True TDOA values are corrupted by samples of uncorrelated

Gaussian noise with a known standard deviation $.01m$.

Case B: Two simultaneous active sources. The sources are assumed to be symmetrically situated $2.8m$ about the the broadside location such that the corresponding TDOA estimates reflect equally the contributions of each true TDOA set and the source location is estimated to be $\hat{\mathbf{s}}$. TDOA values are corrupted by samples of uncorrelated Gaussian noise with a known standard deviation $.01m$. This is an instance of multiple sources producing TDOA estimates that are not representative of a single source location. The non-valid location estimate found through minimization of the error criteria is a reflection of the true source locations, in this case, their midpoint.

Case C: No source present and correct TDOA variance figures. TDOA values are set to zero and uncorrelated Gaussian noise with a known standard deviation $.01m$ is then added. This corresponds to the "source absent" hypothesis in the binary hypothesis test of scenario #1.

Case D: No source present and underestimated TDOA variance figures. TDOA values are again set to zero and uncorrelated Gaussian noise with a standard deviation $.02m$ is then added. However, the reported TDOA variance is $.01m$. This corresponds to the case of a delay estimator that does not accurately model the variance term during silence intervals and for which the binary hypothesis test is inappropriate.

Note that for each of the above situations, the reported variance figures and the source location estimate are identical. Under the "source valid" hypothesis, the statistic F would be distributed as $F \sim \chi^2_8$ in each case and therefore the acceptance regions will be identical.

Selecting a detection probability constraint of $\beta = .95$, the acceptance interval calculated from (4.7) is found to be $[a, b] = [1.2, 15.7]$.

The simulations consisted of 1000 trials with each signal situation. The results are shown in Figure 4.4. In each of these graphs, the statistic F value for a given trial is denoted by a dot at the appropriate height. The boundaries of the acceptance region are represented by the two horizontal lines in the lower portion of each plot. The percentage figures listed in the figure titles refer to the fraction of trials falling within the acceptance region. For the Case A simulation, the valid source frame value of 95% is consistent with the P_D constraint of $\beta = .95$. Cases B-D demonstrate the ability of the consistency test to reject non-valid source scenarios. For the two source situation presented the false-alarm rate is less than 1% and with the silence conditions, the distinction between models is large enough that no misclassifications are made. The results of Case C may be compared to those of the binary hypothesis test of scenario #1. Referring back to the 90° source in Figure 4.3, the graph indicates that with these conditions ($P_D = .95$ and TDOA noise $= 10^{-2}$m) the false-alarm rate associated with the binary decision rule is negligible ($P_F \ll .001$). While the consistency test presented here is generally more conservative due to its lack of alternate hypothesis assumptions, the results of the Case C simulation certainly agree with those predicted for the binary decision test. In this instance, little in the way of performance has been sacrificed by substituting the more general consistency test for the binary hypothesis test.

4.4 Scenario # 3: No Statistical Model

In scenario #1, simple statistical models were assumed to be available for both the source and non-source conditions. Scenario #2 involves the limitation of this knowledge to the

characterization of the TDOA estimates only during valid source periods. In this final scenario, no clear statistical models are assumed to be applicable to either the source or non-source hypotheses. The resulting detection rule is based entirely upon empirical criteria rather than a probabilistic derivation, and as such, a performance analysis is difficult to quantify independent of the particular application. This empirical test is appropriate for those cases in which the "source-only model" is unrealistic. These situations include instances in which the TDOA estimates are found to deviate significantly from a Gaussian distribution or their reported variance figures are inaccurate on an absolute scale[2].

Given a source location estimate, $\hat{\mathbf{s}}$, the empirical detection measure, E, is defined as the average of the absolute value of the differences between the estimated DOA, θ_i, and the true DOA associated with the location $\hat{\mathbf{s}}$ relative to each sensor pair, i.e.

$$E = \frac{1}{N} \sum_{i=1}^{N} |\theta_i - \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}})| \tag{4.8}$$

The physical significance of this detection measure is illustrated in Figure 4.5 for the case of 3 sensor pairs. In the interest of clarity, the DOA cones are shown as bearing lines which represent the intersection the respective cone with the plane formed by the estimated source location and the appropriate sensor-pair axis. The solid lines indicate estimated bearing lines for each sensor pair while the dashed lines denote the DOA bearings for the location estimate. The expression in (4.8) is a reflection of the degree that the estimated bearings are clustered about $\hat{\mathbf{s}}$. A tight clustering produces a small value for E and is indicative of

[2]It is important to distinguish the absolute and relative precision of the variance figures associated with the TDOA estimates. A misrepresentation of these values on an absolute scale prevents the use of the statistical models incorporated into the hypothesis-testing procedures of scenarios #1 and #2, but may not be detrimental to the location estimation itself. The LS-error criteria presented are dependent upon the ratio of these variance values relative to one another. An error in the scale of these terms will not effect the minimization process. In practice, knowledge of the relative TDOA variance may be simpler to obtain than an estimation of the absolute variance figures.

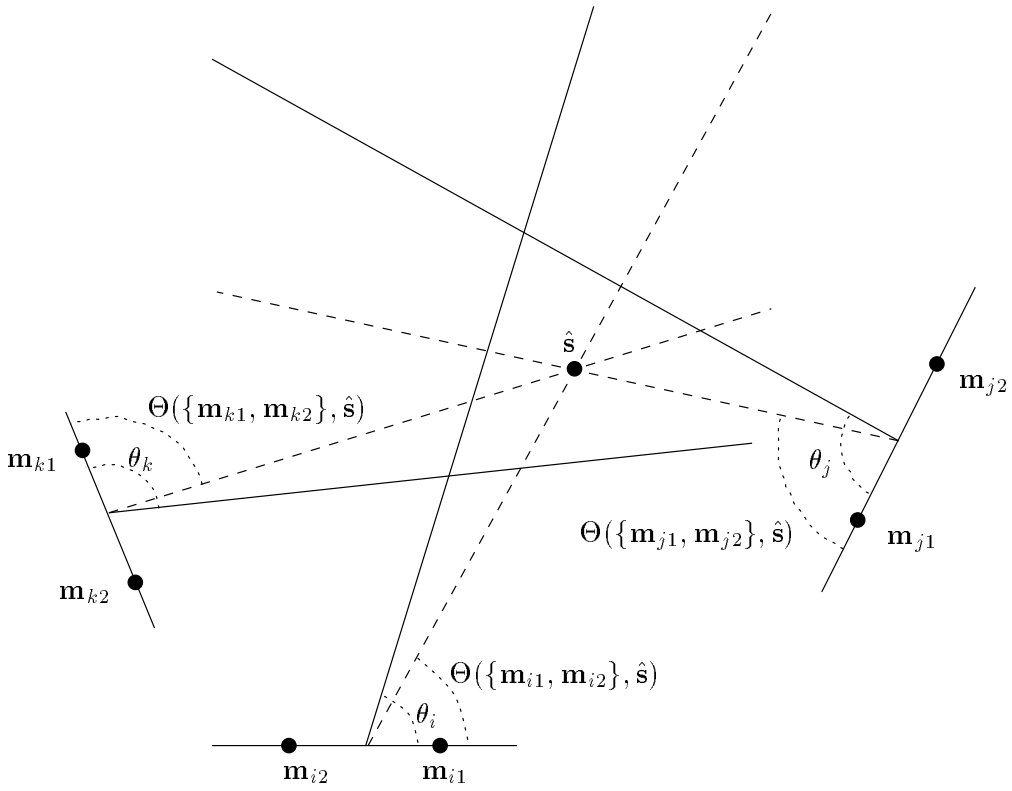


Figure 4.5: Empirical Detection Test: Illustration of the physical significance associated with the detection measure E for the case of 3 sensor pairs.

a valid source location estimate. Excessive values are typical for an inaccurate estimate or a situation where a single source is not present. In practice, a detection threshold of 1° to 2° provides an effective means of identifying valid source locations.

A detection statistic which incorporates average bearing angle deviation is preferable to tests based upon TDOA disparity or overall distance. With regard to physical significance, a bearing angle measure is advantageous to a TDOA approach. Because of the nonlinear mapping from spatial bearings to TDOA values, displacements stated in terms of time-delay figures will have varying physical interpretations depending upon source bearing. A mean distance measure is unfavorable due to its bias towards locations close to the sensors. This was the shortcoming of the J_D error criterion presented in the previous chapter. Remote sources, in general, possess a greater total distance from the DOA cones thereby making

it difficult to devise a detection threshold that is independent of source range. The use of a detection measure based upon direction of arrival alone avoids both these difficulties. It is invariant to both source bearing and range as well as possessing a physical significance suitable for a source/non-source selection.

4.5 Discussion

Three distinct source detection tests have been detailed in this chapter. Their use is dependent upon the environmental and system constraints imposed by the practical application. The first, the binary source detection test, was designed with a very specific circumstance in mind, namely those instances where the "source/silence model" is valid. When this is the case, the test is statistically optimal. The second, the source consistency test, was derived assuming a general source/non-source model. It is applicable to a wider range of situations, but may not represent the most powerful test available when specific statistical models are known. The empirical detection test presents the extreme end of the utilization-performance spectrum. While being universally applicable, it does not offer any guarantee of optimality or performance predictability. In general, the selection of a particular detection test is a function of the information available. When specific knowledge of the source/non-source statistics is known, it may be possible to generate a test which fully exploits this understanding, as was done in scenario #1. For those cases where the hypotheses are inadequately defined or unspecified altogether, the general detection test of scenario #2 and the empirical test of scenario #3 are appropriate.

Chapter 5

Estimation of Localization Error Region

Given the location estimate of a source, an assessment of the spatial region of uncertainty related to the estimate is essential before the information can be judiciously employed in a practical application. The geometric framework developed here lends itself to a straightforward analysis of the spatial covariance associated with the location estimators.

5.1 Displacement Geometry

Let $\hat{\mathbf{s}}$ be the 3-dimensional location estimate of a source with true location $\mathbf{t}$. For a pair of sensors, m_{i1} and m_{i2}, with midpoint $\mathbf{m}_i$ and unit axis $\overline{\mathbf{a}_i}$ as shown in Figure 5.1, R_i is defined as the distance from $\mathbf{t}$ to $\mathbf{m}_i$ and $\psi_i = \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{t})$ as the angle between the directed line segment $\mathbf{t} - \mathbf{m}_i$ and the sensor pair axis $\overline{\mathbf{a}_i}$. The values $\hat{R}_i$ and $\hat{\theta}_i$ are defined similarly for the location estimate $\hat{\mathbf{s}}$. The 3-dimensional Cartesian displacement vector from

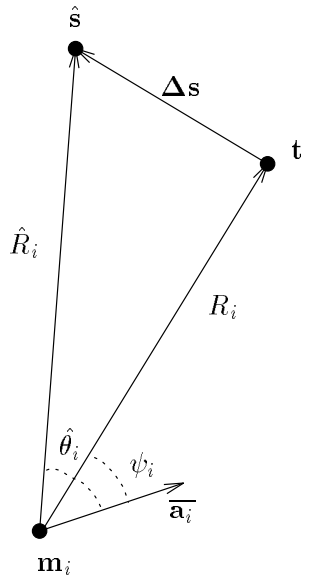


Figure 5.1: The relationship between the true source location $\mathbf{t}$ and an estimate of the location $\hat{\mathbf{s}}$ relative to the i^{th} sensor pair.

$\mathbf{t}$ to $\hat{\mathbf{s}}$ is denoted by:

$$\boldsymbol{\Delta}\mathbf{s} = \begin{bmatrix} \Delta s_x \\ \Delta s_y \\ \Delta s_z \end{bmatrix}$$

In what follows, it is assumed that the true source location is known and the goal is to develop a statistical analysis of the precision associated with the source estimate $\hat{\mathbf{s}}$.

$\hat{\theta}_i$ is related to the positional vectors via the dot product:

$$\begin{aligned} \hat{R}_i \cos\hat{\theta}_i &= (\mathbf{t} + \boldsymbol{\Delta}\mathbf{s} - \mathbf{m}_i) \cdot \overline{\mathbf{a}_i} \\ &= (\mathbf{t} - \mathbf{m}_i) \cdot \overline{\mathbf{a}_i} + \boldsymbol{\Delta}\mathbf{s} \cdot \overline{\mathbf{a}_i} \\ &= R_i \cos\psi_i + \boldsymbol{\Delta}\mathbf{s} \cdot \overline{\mathbf{a}_i} \end{aligned} \tag{5.1}$$

Following [43], $\hat{R}_i$ is approximated by its first-order Taylor series expansion about the

true source location:

$$\hat{R}_i \approx R_i + \left(\frac{\mathbf{t} - \mathbf{m}_i}{R_i}\right) \cdot \mathbf{\Delta s} \tag{5.2}$$

This linearization of the source range value requires the assumption that source location estimate be sufficiently close to the true location such that the error induced in this approximation is negligible in comparison to localization errors associated with the TDOA estimate errors. As will be shown, this assumption is quite reasonable in practice.

Substituting (5.2) into (5.1) yields:

$$R_i \cos\hat{\theta}_i + \left(\frac{\mathbf{t} - \mathbf{m}_i}{R_i}\right) \cdot \mathbf{\Delta s} \cos\hat{\theta}_i = R_i \cos\psi_i + \mathbf{\Delta s} \cdot \overline{\mathbf{a}_i}$$

or equivalently:

$$\cos\hat{\theta}_i - \cos\psi_i = \left[\frac{\overline{\mathbf{a}_i}}{R_i} - (\mathbf{t} - \mathbf{m}_i)\frac{\cos\hat{\theta}_i}{R_i^2}\right] \cdot \mathbf{\Delta s} \tag{5.3}$$

By making the assumption that $\frac{\cos\psi_i}{R_i^2} \approx \frac{\cos\hat{\theta}_i}{R_i^2}$ and applying (2.5) to the cosine terms on the left side of the equation, arriving at:

$$\left(\frac{c}{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|}\right)[T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}) - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{t})] = \left[\frac{\overline{\mathbf{a}_i}}{R_i} - (\mathbf{t} - \mathbf{m}_i)\frac{\cos\psi_i}{R_i^2}\right] \cdot \mathbf{\Delta s}$$

And finally:

$$\begin{aligned}[T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \hat{\mathbf{s}}) - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{t})] &= \underbrace{\left(\frac{|\mathbf{m}_{i2} - \mathbf{m}_{i1}|}{c}\right)\left[\frac{\overline{\mathbf{a}_i}}{R_i} - (\mathbf{t} - \mathbf{m}_i)\frac{\cos\psi_i}{R_i^2}\right]}_{\mathbf{h}_i^T} \cdot \mathbf{\Delta s} \\ &= \mathbf{h}_i^T \cdot \mathbf{\Delta s} \end{aligned} \tag{5.4}$$

where $\mathbf{h}_i^T$ is the (1×3) vector relating the difference in TDOA for the i^{th} sensor pair to the estimate displacement vector.

It will be useful to express (5.4) for the N sensor pairs via matrix notation. The $(N \times 1)$ vector of TDOA differences is denoted by $\boldsymbol{\Delta}\tau_{\hat{s}t}$ and the $(N \times 3)$ matrix composed of the $\mathbf{h}_i^T$ vectors will be given by $\mathbf{H}$, i.e.

$$\boldsymbol{\Delta}\tau_{\hat{s}t} = \begin{bmatrix} [T(\{\mathbf{m}_{11}, \mathbf{m}_{12}\}, \hat{\mathbf{s}}) - T(\{\mathbf{m}_{11}, \mathbf{m}_{12}\}, \mathbf{t})] \\ [T(\{\mathbf{m}_{21}, \mathbf{m}_{22}\}, \hat{\mathbf{s}}) - T(\{\mathbf{m}_{21}, \mathbf{m}_{22}\}, \mathbf{t})] \\ \vdots \\ [T(\{\mathbf{m}_{N1}, \mathbf{m}_{N2}\}, \hat{\mathbf{s}}) - T(\{\mathbf{m}_{N1}, \mathbf{m}_{N2}\}, \mathbf{t})] \end{bmatrix} \qquad \mathbf{H} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_2^T \\ \vdots \\ \mathbf{h}_N^T \end{bmatrix}$$

The estimate displacement is then related to the N delay-estimate, sensor-pair combinations by

$$\boldsymbol{\Delta}\tau_{\hat{\mathbf{s}}\mathbf{t}} = \mathbf{H}\boldsymbol{\Delta}\mathbf{s} \tag{5.5}$$

5.2 Source Estimate Based Upon J_{TDOA}

The case where the position estimate in question was derived from minimization of J_{TDOA} LS error criterion and correspondingly $\hat{\mathbf{s}} = \hat{\mathbf{s}}_{TDOA}$ as given by (3.4) is now examined.

The LS error (3.2) may be rewritten as:

$$\begin{aligned} J_{TDOA}(\mathbf{s}) &= \sum_{i=1}^{N} \epsilon_{itdoa} \cdot \Big[\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})\Big]^2 \\ &= \sum_{i=1}^{N} \epsilon_{itdoa} \cdot \Big[\tau_i - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{t}) \quad + \\ &\qquad T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{t}) - T(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})\Big]^2 \end{aligned} \tag{5.6}$$

Defining $\mathbf{W}_{tdoa}$ as the $(N \times N)$ diagonal matrix of weighting coefficients ϵ_{itdoa} given by

(3.3) and $\mathbf{\Delta}\tau_{\tau t}$ to be the $(N \times 1)$ vector of differences between the estimated TDOA and true TDOA,

$$\mathbf{\Delta}\tau_{\tau t} = \begin{bmatrix} \tau_1 - T(\{\mathbf{m}_{11}, \mathbf{m}_{12}\}, \mathbf{t}) \\ \tau_2 - T(\{\mathbf{m}_{21}, \mathbf{m}_{22}\}, \mathbf{t}) \\ \vdots \\ \tau_N - T(\{\mathbf{m}_{N1}, \mathbf{m}_{N2}\}, \mathbf{t}) \end{bmatrix} \qquad \mathbf{W}_{tdoa} = \begin{bmatrix} \epsilon_{1tdoa} & & & \\ & \epsilon_{2tdoa} & & \\ & & \ddots & \\ & & & \epsilon_{Ntdoa} \end{bmatrix}$$

Equation (5.6) is rewritten as:

$$J_{TDOA}(\mathbf{s}) = (\mathbf{\Delta}\tau_{\tau t} - \mathbf{\Delta}\tau_{st})^T \mathbf{W}_{tdoa} (\mathbf{\Delta}\tau_{\tau t} - \mathbf{\Delta}\tau_{st})$$

The LS criterion is assumed to be minimized when $\mathbf{s} = \hat{\mathbf{s}}$, which gives:

$$\begin{aligned} J_{TDOA}(\hat{\mathbf{s}}) = \min_{\mathbf{s}} J_{TDOA}(\mathbf{s}) &= (\mathbf{\Delta}\tau_{\tau t} - \mathbf{\Delta}\tau_{\hat{s}t})^T \mathbf{W}_{tdoa} (\mathbf{\Delta}\tau_{\tau t} - \mathbf{\Delta}\tau_{\hat{s}t}) \\ &= (\mathbf{\Delta}\tau_{\tau t} - \mathbf{H}\mathbf{\Delta}\mathbf{s}_{tdoa})^T \mathbf{W}_{tdoa} (\mathbf{\Delta}\tau_{\tau t} - \mathbf{H}\mathbf{\Delta}\mathbf{s}_{tdoa}) \end{aligned} \quad (5.7)$$

The right side of this equation is identical in form to the weighted linear least squares error and can be shown [60] to be minimized when:

$$\mathbf{\Delta}\mathbf{s}_{tdoa} = (\mathbf{H}^T \mathbf{W}_{tdoa} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{W}_{tdoa} \mathbf{\Delta}\tau_{\tau t} \quad (5.8)$$

Therefore, minimization of J_{TDOA} would result in a $\mathbf{\Delta}\mathbf{s}_{tdoa}$ as given above. Equation (5.8) relates the displacement vector associated with a location estimate $\hat{\mathbf{s}}$ to the TDOA estimates that would produce this particular estimate via minimization of the nonlinear LS error criterion J_{TDOA}.

The covariance of $\mathbf{\Delta s}_{tdoa}$ is given by:

$$cov\{\mathbf{\Delta s}_{tdoa}\} = E\left[(\mathbf{\Delta s}_{tdoa} - E(\mathbf{\Delta s}_{tdoa}))(\mathbf{\Delta s}_{tdoa} - E(\mathbf{\Delta s}_{tdoa}))^T\right]$$

The delay estimates have been assumed to be corrupted by a zero-mean, uncorrelated noise source and therefore $E(\mathbf{\Delta}\tau_{\tau t}) = \mathbf{0}$. Substituting this and (5.8) into the above yields:

$$\begin{aligned} cov\{\mathbf{\Delta s}_{tdoa}\} &= E\left[\mathbf{\Delta s}_{tdoa}\mathbf{\Delta s}_{tdoa}{}^T\right] \\ &= (\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}_{tdoa}E(\mathbf{\Delta}\tau_{\tau t}\mathbf{\Delta}\tau_{\tau t}{}^T)\mathbf{W}_{tdoa}{}^T\mathbf{H}\left((\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1}\right)^T \\ &= (\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}_{tdoa}E(\mathbf{\Delta}\tau_{\tau t}\mathbf{\Delta}\tau_{\tau t}{}^T)\mathbf{W}_{tdoa}{}^T\mathbf{H}(\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1} \end{aligned}$$

The term $E\left(\mathbf{\Delta}\tau_{\tau t}\mathbf{\Delta}\tau_{\tau t}{}^T\right)$ is equivalent to $cov\{\mathbf{\Delta}\tau_{\tau t}\}$ which is an $(N \times N)$ diagonal matrix with entries $var\{\tau_i\}$. Similarly, the weighting coefficients that comprise the diagonal elements of $\mathbf{W}_{tdoa}$ were selected to be $(1/var\{\tau_i\})$, and thus, $E(\mathbf{\Delta}\tau_{\tau t}\mathbf{\Delta}\tau_{\tau t}{}^T)\mathbf{W}_{tdoa}{}^T = \mathbf{I_N}$. The expression therefore simplifies to:

$$\begin{aligned} cov\{\mathbf{\Delta s}_{tdoa}\} &= (\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1}(\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})(\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1} \\ &= (\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H})^{-1} \end{aligned} \tag{5.9}$$

Equation (5.9) predicts the covariance of the $\hat{\mathbf{s}}_{TDOA}$ estimate given knowledge of the source and sensor locations as well the TDOA estimate variances.

5.3 Source Estimate Based Upon J_{DOA}

A similar procedure may be followed for analyzing the precision of the location estimate $\hat{\mathbf{s}}_{DOA}$, found via minimization of the J_{DOA} LS error criterion.

The LS error-criteria (3.5) is expressed as

$$\begin{aligned} J_{DOA}(\mathbf{s}) &= \sum_{i=1}^{N} \epsilon_{idoa} \cdot \Big[\theta_i - \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})\Big]^2 \\ &= \sum_{i=1}^{N} \epsilon_{idoa} \cdot \Big[\theta_i - \psi_i + \psi_i - \Theta(\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}, \mathbf{s})\Big]^2 \\ &= (\boldsymbol{\Delta\theta}_{\theta t} - \boldsymbol{\Delta\theta}_{st})^T \mathbf{W}_{doa} (\boldsymbol{\Delta\theta}_{\theta t} - \boldsymbol{\Delta\theta}_{st}) \end{aligned}$$

where $\mathbf{W}_{doa}$ is the diagonal matrix of weighting coefficients ϵ_{idoa} defined by (3.6), $\boldsymbol{\Delta\theta}_{\theta t}$ is the vector of differences between the estimated and true DOA's, and $\boldsymbol{\Delta\theta}_{st}$ is the vector of DOA differences between the hypothesized source location $\mathbf{s}$ and the true location $\mathbf{t}$,

$$\mathbf{W}_{doa} = \begin{bmatrix} \epsilon_{1doa} & & & \\ & \epsilon_{2doa} & & \\ & & \ddots & \\ & & & \epsilon_{Ndoa} \end{bmatrix}$$

$$\boldsymbol{\Delta\theta}_{\theta t} = \begin{bmatrix} \theta_1 - \psi_1 \\ \theta_2 - \psi_2 \\ \vdots \\ \theta_N - \psi_N \end{bmatrix} \qquad \boldsymbol{\Delta\theta}_{st} = \begin{bmatrix} \Theta(\{\mathbf{m}_{11}, \mathbf{m}_{12}\}, \mathbf{s}) - \psi_1 \\ \Theta(\{\mathbf{m}_{21}, \mathbf{m}_{22}\}, \mathbf{s}) - \psi_2 \\ \vdots \\ \Theta(\{\mathbf{m}_{N1}, \mathbf{m}_{N2}\}, \mathbf{s}) - \psi_N \end{bmatrix}$$

Returning now to (5.3) and applying the trigonometric identity:

$$\cos\hat{\theta}_i - \cos\psi_i = 2\sin\left(\frac{\psi_i + \hat{\theta}_i}{2}\right)\sin\left(\frac{\psi_i - \hat{\theta}_i}{2}\right)$$

which for small angle differences ($\hat{\theta}_i \approx \psi_i$) is well approximated by:

$$\cos\hat{\theta}_i - \cos\psi_i \approx -(\hat{\theta}_i - \psi_i)\sin\psi_i$$

or

$$\hat{\theta}_i - \psi_i \approx \frac{-(\cos\hat{\theta}_i - \cos\psi_i)}{\sin\psi_i}$$

Substitution of (5.3) into this expression yields:

$$\hat{\theta}_i - \psi_i \approx \underbrace{\left(\frac{-1}{\sin\psi_i}\right)\left[\frac{\overline{\mathbf{a}_i}}{R_i} - (\mathbf{t} - \mathbf{m}_i)\frac{\cos\psi_i}{R_i^2}\right]}_{\mathrm{g}_i^T} \cdot \boldsymbol{\Delta}\mathbf{s}_{doa}$$

The vector of estimate and true DOA differences, $\boldsymbol{\Delta}\boldsymbol{\theta}_{\hat{s}t}$, may then be written as

$$\boldsymbol{\Delta}\boldsymbol{\theta}_{\hat{s}t} = \mathbf{G}\boldsymbol{\Delta}\mathbf{s}_{doa} \qquad \text{where} \qquad \mathbf{G} = \begin{bmatrix} \mathbf{g}_1^T \\ \mathbf{g}_2^T \\ \vdots \\ \mathbf{g}_N^T \end{bmatrix}$$

Following an argument similar to that used for the derivation of the expressions in (5.8) and (5.9), the displacement vector associated with $\hat{\mathbf{s}}_{DOA}$ is calculated from

$$\boldsymbol{\Delta}\mathbf{s}_{doa} = (\mathbf{G}^T\mathbf{W}_{doa}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{W}_{doa}\boldsymbol{\Delta}\boldsymbol{\theta}_{\theta t} \tag{5.10}$$

and the corresponding displacement covariance is found to be

$$cov\{\boldsymbol{\Delta}\mathbf{s}_{doa}\} = (\mathbf{G}^T\mathbf{W}_{doa}\mathbf{G})^{-1} \tag{5.11}$$

A comparison of (5.9) and (5.11) and their constituent matrices reveals that $\mathbf{H}^T\mathbf{W}_{tdoa}\mathbf{H} = \mathbf{G}^T\mathbf{W}_{doa}\mathbf{G}$ and therefore the covariance predictors are equivalent. While the estimates $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ are clearly not identical, the approximations made in deriving these closed-form estimate error expressions yield indistinguishable results. In situations where the aforementioned approximations are less appropriate (severe noise conditions, extreme source location bearings, etc.) the $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimators do vary considerably in performance, as was demonstrated in the simulation results of the preceding section. Under these conditions the estimate covariance expressions will be less applicable in predicting the actual estimator error. However, as will be shown, the estimate error expressions, (5.9) and (5.11), are accurate predictors of the estimators' true performance given reasonable source position and signal quality scenarios.

5.4 Analysis of Estimate Error Predictors

To evaluate the accuracy of (5.9) and (5.11) as predictors of the estimators' true covariance, two sets of simulations were conducted with the varying parameter being the positioning of the sensor pairs.

Evaluation #1

In this first experiment, the ten-element, bi-linear sensor array shown in Figure 3.3 was reemployed. The array was situated at the center of one wall in a $6m \times 6m \times 4m$ rectangular room as depicted in Figure 5.2. Once again, the eight pairings of diagonally adjacent sensors were selected as the sensor pairs. Monte Carlo simulations consisting of 100 trials each were conducted across a grid of 36 source locations within the room. Source

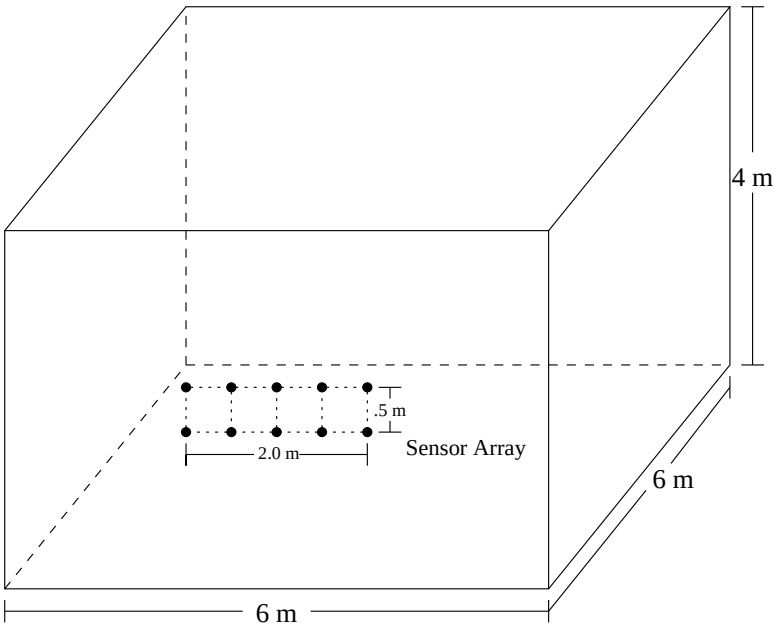


Figure 5.2: Location Error Evaluation #1: The ten-element bilinear sensor array with 0.5m spacings is centered along one wall of a $6m \times 6m \times 4m$ rectangular room.

locations were spaced a meter apart along two distinct horizontal planes. For each source location the true TDOA values for each sensor pair were calculated and then corrupted by uncorrelated additive white Gaussian noise. The corrupting noise level at each sensor pair was fixed at a moderate level, a standard deviation of 10^{-2}m when scaled by c. The LS-based estimates $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ were then calculated for each trial via a quasi-Newton algorithm constrained to search within the physical dimensions of the room.

Figure 5.3 displays the results of these simulations. Each of the plots in this figure is from the perspective of a viewer directly above the room and looking downward. The sensor array is represented by the five circles on the left vertical axis. In the top graph, the 3600 (36 locations, 100 estimates per location) $\hat{\mathbf{s}}_{TDOA}$ estimates have been plotted with dots. While a dot's position as projected onto the floor is clear from the figure, the height is ambiguous. Because of the symmetry involved in the setup of this array within the room environment, the choice of source locations at each height was limited to the half-plane

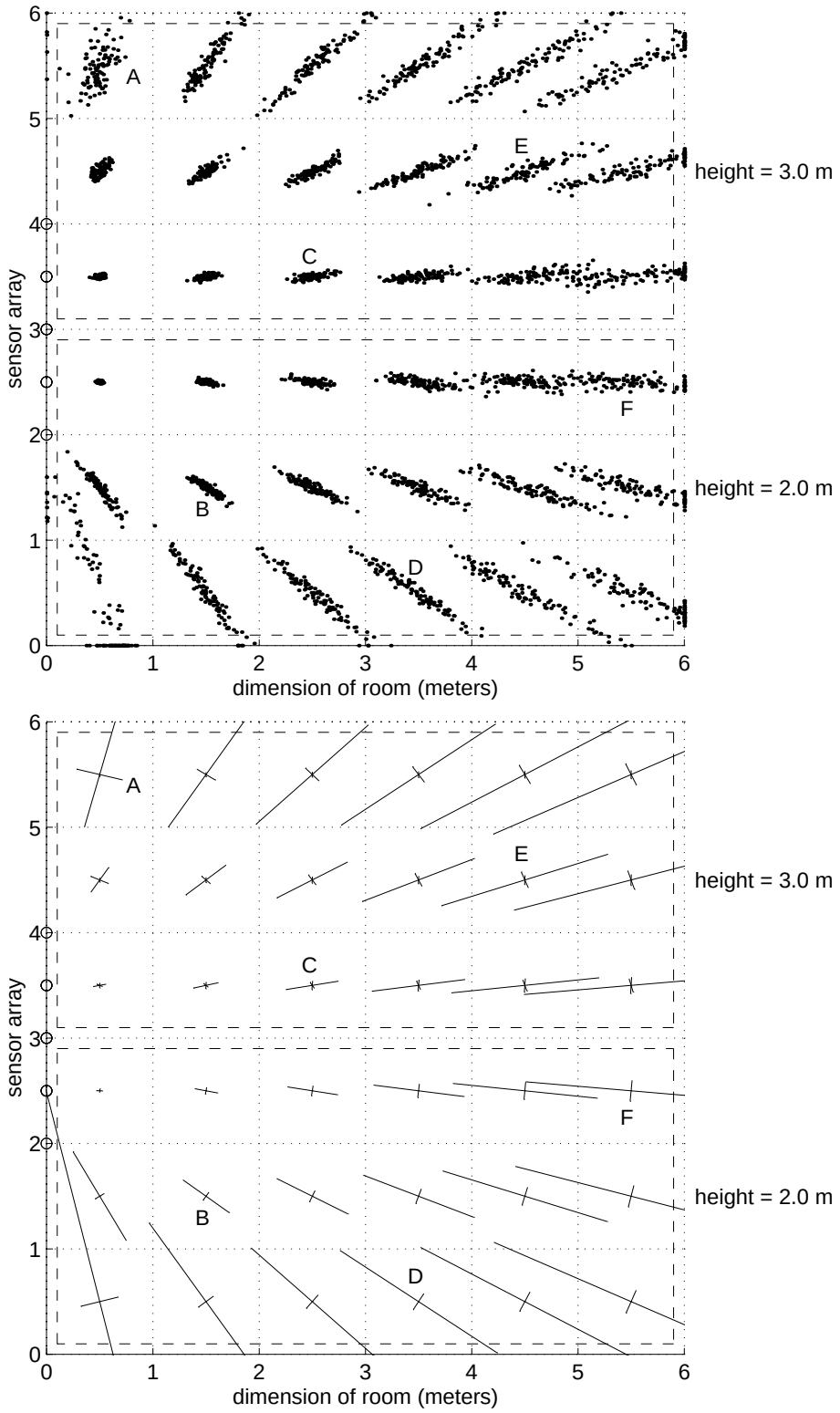


Figure 5.3: Location Error Evaluation #1: (top graph) A top-down view of the $\hat{\mathbf{s}}_{TDOA}$ source estimates for the 100 trial simulations at 36 source locations. (bottom graph) Principal component axes of the predicted estimate error covariance for each source location. Alphabetic labels refer to those source locations for which numerical data is presented in Table 5.1.

delineated by the line normal to the center of the array. Source locations on the remaining half-plane would presumably display the same properties. In each of these plots, the lower horizontal half-plane contains source locations at a height of 2m, level with the mid-line of the sensor array. The upper half-plane is at a height of 3m, a meter vertically above the array.

The bottom graph in Figure 5.3 shows the principal component vectors of the predicted covariance matrix scaled to 2.5 standard deviations. For each source location, the predicted error covariance matrix was calculated via (5.9), or equivalently by (5.11). An eigenvector-eigenvalue decomposition of the (3×3) matrix yields its principal components vectors [61]. Geometrically, if $cov\{\mathbf{\Delta s}_{doa}\}$ is positive definite with eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ for $i = 1, 2, 3$, all the (3×1) vectors $\mathbf{x}$ which satisfy:

$$(\mathbf{x} - \overline{\mathbf{x}})^T (cov\{\mathbf{\Delta s}_{doa}\})^{-1} (\mathbf{x} - \overline{\mathbf{x}}) = h^2$$

define a hyperellipsoid centered about $\overline{\mathbf{x}}$ with axes $\pm h\sqrt{\lambda_i}\mathbf{e}_i$. The eigenvalues correspond to the variance of the data set projected onto the corresponding eigenvector or principal component. Setting $h = 2.5$ in the above expression will therefore generate an hyperellipsoid with axes extending 2.5 standard deviations in either direction from the center of the conic along each of the principal component vectors. If the distribution of source estimates possesses a trivariate normal density, a given estimate would have a 0.9 probability of falling on or within such a hyperellipsoid [61]. In each case, the estimator has been assumed to be zero-biased and thus the center of the hyperellipsoid is the given source location. The lines in the bottom graph display the scaled principal component vectors which correspond to the axes of the hyperellipsoid associated with the predicted error covariance of each source

Point Label	Standard Deviations of Principal Components (cm)	pred.	$\hat{\mathbf{s}}_{\text{TDOA}}$	$\hat{\mathbf{s}}_{\text{DOA}}$
A	1^{st}	22.7	21.0	21.7
	2^{nd}	9.0	14.9	9.1
	3^{rd}	1.8	1.8	1.8
	total	24.6	25.8	23.6
B	1^{st}	10.6	10.5	10.5
	2^{nd}	1.9	1.9	1.9
	3^{rd}	1.5	1.3	1.3
	total	10.9	10.8	10.7
C	1^{st}	10.7	10.7	10.7
	2^{nd}	2.2	2.2	2.2
	3^{rd}	2.0	1.9	1.9
	total	11.1	11.1	11.1

Point Label	Standard Deviations of Principal Components (cm)	pred.	$\hat{\mathbf{s}}_{\text{TDOA}}$	$\hat{\mathbf{s}}_{\text{DOA}}$
D	1^{st}	35.4	33.0	33.0
	2^{nd}	3.7	3.2	3.2
	3^{rd}	3.0	2.8	2.8
	total	35.7	33.3	33.3
E	1^{st}	33.6	34.1	34.0
	2^{nd}	3.7	3.8	3.8
	3^{rd}	3.5	3.6	3.6
	total	33.9	34.5	34.4
F	1^{st}	39.9	35.6	35.3
	2^{nd}	4.0	4.2	4.2
	3^{rd}	3.9	3.8	3.8
	total	40.2	36.0	35.7

Table 5.1: Location Error Evaluation #1: The principal component standard deviations for the predicted error covariance and the sample covariances derived from the $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimates. Point labels refer to source locations in Figure 5.3.

location shown in the top graph.

Table 5.1 presents detailed numerical data for selected source locations in Figure 5.3. In each case, the principal component standard deviations are listed for the predicted error covariance as well as the sample covariances associated with the sets of $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimates. A fourth row gives the total standard deviation. This value is calculated as the square root of the component variance summation and is equivalent to the square root of the trace of the particular covariance matrix.

Several observations are apparent from Figure 5.3 and Table 5.1. The predicted error covariance closely models, to within a few centimeters, the true performance of both the $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimators. As the figure suggests and the table quantifies, disparities between the predicted and observed are most extreme in those cases involving relatively large error variances, where the linearity assumptions used in the derivation of the error covariance predictor are less valid. For instance, consider the source point labeled 'D', located at $(3.5m, 0.5m)$ and at a height of 2m. The total observed standard deviation is

33.3cm for each estimator while the predicted total is 35.7cm. The 2.4cm difference is largely accounted for by the first principle component. At the other extreme is the point 'B' at $(1.5m, 1.5m)$ and height 2m. The disparity between predicted and observed total standard deviation is only .1cm. For sources positioned near the boundary of the room, this disparity between observation and prediction may be due, in part, to artificially low observed covariance values brought about by the search constraint placed on the LS-error criteria minimizer. Those source locations that are estimated to be outside of the physical room are placed at the room boundary and the spread of source estimates is subsequently skewed. This effect is apparent in the statistics for point 'F' where there is a visible cluster of estimate points at the wall of the room and the observed first principal component standard deviations are sizably less than the predicted value.

As expected, the source estimation procedure is most accurate for broadside sources close to the sensor array. Estimate precision is extremely sensitive to bearing for sources near end-fire conditions and the quality of the range estimates degrades rapidly as the true source range increases. These observations are consistent with results reported for standard linear arrays, as in [38]. For a fixed (x,y) position relative to the floor, the variation in height of the half-planes had little effect on the estimators' precision. An exception to this rule being the source location labeled 'A' at location $(0.5m, 5.5m)$ and height 3m and its symmetric counterpart at $(0.5m, 0.5m)$ and height 2m. While the former is farther from the sensors than the latter, it possesses a milder bearing condition relative to the array. As the figure illustrates, this small improvement in bearing angle has a dramatic effect on the error spread of the source's location estimates in comparison to its counterpart's. Finally, with regard to the LS-error criteria, the results of this set of simulations are consistent with those of the previous section. For broadside sources with a large DOA angle, the $\hat{\mathbf{s}}_{TDOA}$ and

$\hat{\mathbf{s}}_{DOA}$ estimators perform comparably. With source locations close to the end-fire condition, the $\hat{\mathbf{s}}_{DOA}$ estimate obtains a slight performance advantage at this noise level.

Evaluation #2

In previous set of simulations, those source locations that were estimated with the highest degree of precision possessed two key features: they were broadside (or nearly broadside) to the sensor array and they were not particularly distant from the sensors. Short of placing physical obstacles at those positions deemed undesirable, the span of potential source locations within a room cannot be dictated. However, there may be a great deal of liberty granted in the placement of the sensors. The results of the Evaluation #1 motivated the choice of the array configuration illustrated by Figure 5.4 in which a $0.5m \times 0.5m$ square array has been centered along each wall of the $6m \times 6m \times 4m$ rectangular room. This sensor arrangement provides for an improved coverage of the room environment. The vast majority of potential source locations are at a broadside angle to and in the proximity of at least one sensor pair. The diagonal combinations within each sub-array were selected as the sensor pairs, yielding the same number of TDOA estimates (eight) and the same sensor spacings ($.5\sqrt{2}$m) as used in the previous experiment. Monte Carlo simulations were conducted in an identical manner to those of Evaluation #1. However, now that the array configuration possesses the added plane of symmetry relative to the rectangular room, the grid of source locations was selected from four parallel quarter-planes at heights ranging from the midpoint of the room, 2m, to a half-meter short of the ceiling, 3.5m.

The results of this experiment are presented in Figure 5.5 and Table 5.2. Once again, expressions (5.9) and (5.11) accurately predict the results of the Monte Carlo simulations. Discrepancies between observed and predicted values continue to be greatest in the large

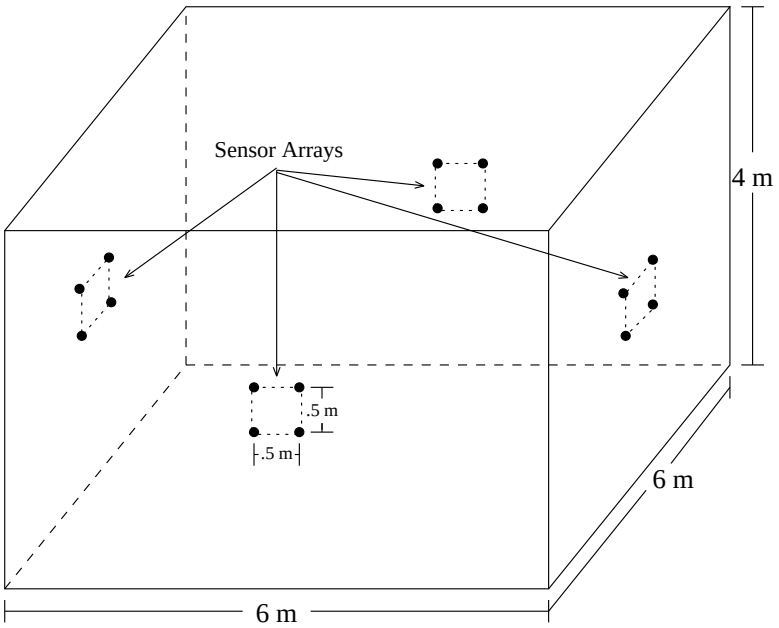


Figure 5.4: Location Error Evaluation #2: Four sets of $0.5m \times 0.5m$ square sensor arrays are positioned at the center of each wall in $6m \times 6m \times 4m$ rectangular room.

variance cases. The $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimators perform comparably under these circumstances as well with the $\hat{\mathbf{s}}_{DOA}$ estimate being mildly preferable for the extreme source location conditions. Source height has little effect on the overall estimation precision, except under those circumstances where altering source height significantly alters a source's bearing angle relative to a sensor pair. In short, the trends from Evaluation # 1 remain apparent with this alternative sensor arrangement. However, the overall source localization error has been reduced significantly as a result of the more judicious placement of sensors. For nearly all 36 source locations, the total error standard deviation has declined markedly and the error hyperellipsoids are considerably less eccentric than those generated via the bilinear array of the previous experiment.

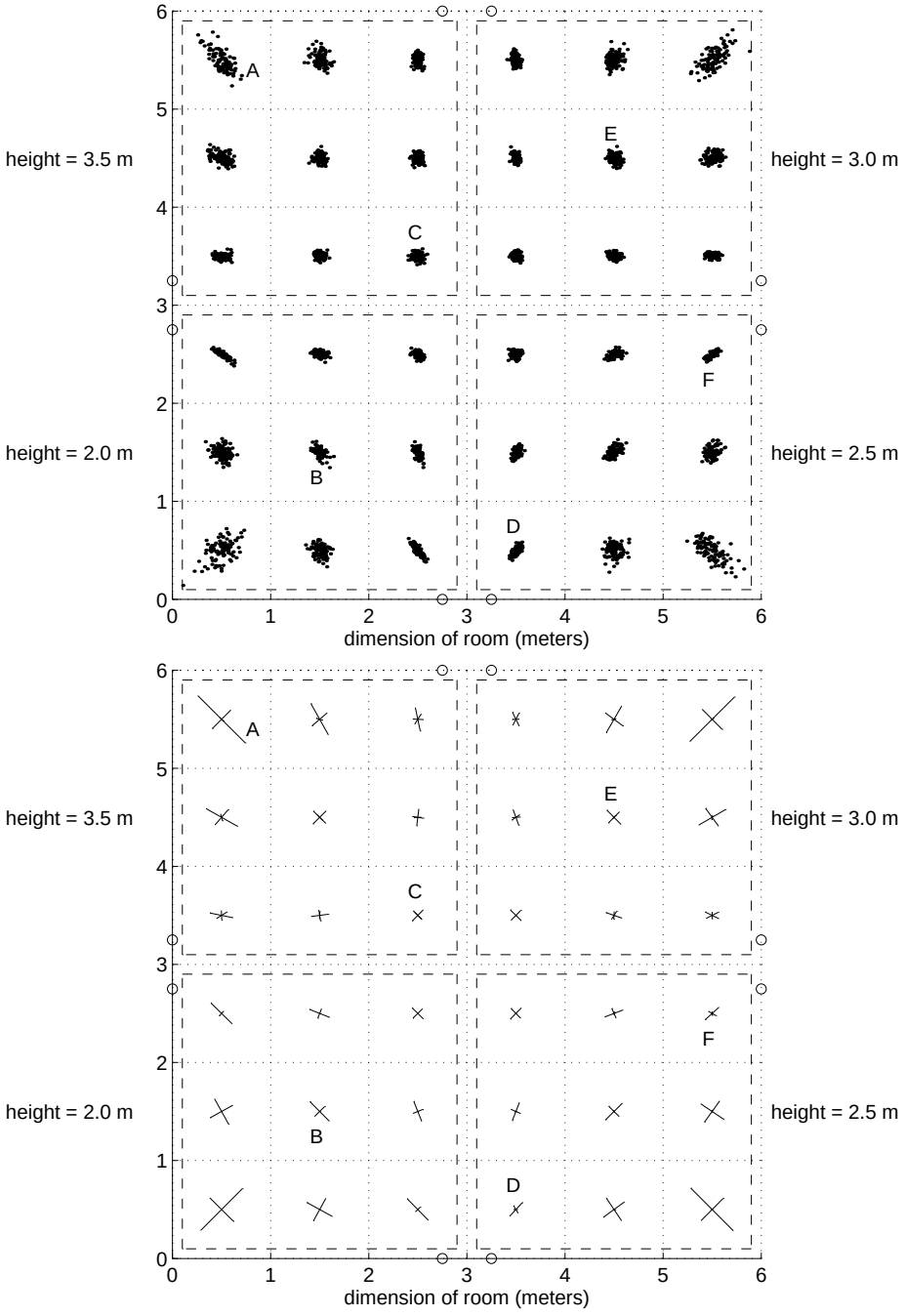


Figure 5.5: Location Error Evaluation # 2: (top graph) A top-down view of the $\hat{\mathbf{s}}_{TDOA}$ source estimates for the 100 trial simulations at 36 source locations. (bottom graph) Principal component axes of the predicted estimate error covariance for each source location.

Point Label	Standard Deviations of Principal Components (cm)	pred.	$\hat{\mathbf{s}}_{\mathrm{TDOA}}$	$\hat{\mathbf{s}}_{\mathrm{DOA}}$
A	1^{st}	14.7	13.5	13.3
	2^{nd}	5.4	4.9	4.9
	3^{rd}	2.8	2.8	2.8
	total	15.9	14.6	14.4
B	1^{st}	6.6	5.1	5.0
	2^{nd}	2.9	2.9	2.9
	3^{rd}	1.9	2.0	2.0
	total	6.6	6.2	6.1
C	1^{st}	3.2	3.8	3.8
	2^{nd}	3.1	3.3	3.2
	3^{rd}	2.9	2.7	2.7
	total	5.3	5.7	5.7

Point Label	Standard Deviations of Principal Components (cm)	pred.	$\hat{\mathbf{s}}_{\mathrm{TDOA}}$	$\hat{\mathbf{s}}_{\mathrm{DOA}}$
D	1^{st}	4.5	4.9	4.7
	2^{nd}	2.0	2.0	2.1
	3^{rd}	1.2	1.2	1.2
	total	5.0	5.4	5.3
E	1^{st}	4.1	4.2	4.2
	2^{nd}	3.5	3.8	3.8
	3^{rd}	2.3	2.3	2.2
	total	5.8	6.1	6.0
F	1^{st}	4.5	4.7	4.6
	2^{nd}	2.0	1.8	1.8
	3^{rd}	1.2	1.5	1.4
	total	5.0	5.2	5.1

Table 5.2: Location Error Evaluation #2: The principal component standard deviations for the predicted error covariance and the sample covariances derived from the $\hat{\mathbf{s}}_{TDOA}$ and $\hat{\mathbf{s}}_{DOA}$ estimates. Point labels refer to source locations in Figure 5.5.

5.5 Discussion

As Evaluation #2 illustrates, the placement of sensors within a room can dramatically impact the quality of the source location estimates. Sensor positioning is usually subject to a number of restrictions. These may be due to the physical or aesthetic constraints of the environment. They may also be due in large part to the requirements of the time-delay estimation procedure. It has been assumed throughout this discussion that the source localization process is independent of the TDOA estimation, requiring that only the parameters of sensor pair locations, TDOA estimates, and the estimate variances be passed from the latter to the former. However, the precision of time-delay estimators is highly dependent upon the coherence or similarity of the signals received at the two sensors. It is therefore essential to the quality of the TDOA estimates that the separation of the individual sensors within each sensor pair be small enough to prevent significant disparities in the received signal quality or content across the sensor pair. This qualification makes certain placement

scenarios, that are seemingly advantageous from a purely localization standpoint, ineffective due to the detrimental effects on the quality of the time-delay estimates. The increased TDOA variances essentially overwhelm the advantages of a broadened baseline. In practice, the selection of sensor separation distances requires a knowledge of the environmentally-dependent performance characteristics associated with the time-delay estimation procedure employed. The simulations presented here have used a sensor separation distance of $.5\sqrt{2}$m along with a TDOA noise standard deviation of .01m. This has proven to be a realistic and appropriate combination of these parameters for this room setting.

The considerations expounded upon in the preceding paragraph apply only to the separation of individual sensors within a sensor pair, not to the overall placement of the sensor pairs themselves. The choice of sensor pair numbers and positions ultimately depends upon minimizing some form of a precision-based cost function that is constrained by the requirements of the physical environment and the intra-pair separation distances. The details and method of minimizing such a cost function will vary dramatically from one application to another. Some work in this area related specifically to speech source acquisition has been reported in [5, 1, 62, 63]. In many scenarios, prior information concerning the potential locations of signal sources or a set of spatial regions from which it is desirable to obtain 'good' location estimates may be specified and the complexity of the cost function will be greatly reduced. Regardless of the specifics, at the core of this procedure there must be a means of evaluating estimation accuracy given source and sensors locations. It is here that the expressions for predicting error covariance find application.

Another application of the error covariance predictors is as the basis of a scheme for distinguishing sources in a multi-source tracking system. Consider a situation where a number of source location estimates $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$ have been evaluated over a period of

time. It is desired to know whether or not these estimates are associated with a single, non-moving source or several such sources. A null hypothesis testing approach may be adopted to determine if the observed location estimates are consistent with a single source hypothesis [57]. Let $\overline{\mathbf{s}}$ be the sample mean of the estimate points and the assumed true location of the hypothesized single source. Using (5.9) or (5.11), the predicted error covariance for the location $\overline{\mathbf{s}}$ is evaluated and denoted the matrix by $\mathbf{C}$. The hypothesized error region is assumed to be normally distributed and accordingly, if the single source model were valid, the location estimate samples would be derived from a p-dimensional normal distribution with mean $\overline{\mathbf{s}}$ and covariance $\mathbf{C}$. The scalar statistic

$$S = \sum_{i=1}^{N} (\mathbf{s}_i - \overline{\mathbf{s}})^T \mathbf{C}^{-1} (\mathbf{s}_i - \overline{\mathbf{s}})$$

would possess a chi-squared distribution with $p(N-1)$-degrees of freedom ($S \sim \chi^2_{p(N-1)}$). Letting P_F be the desired false-alarm probability, the acceptance region that is maximally consistent with the hypothesized model while satisfying the false-alarm criterion is given by the interval:

$$a \leq S \leq b \qquad \text{such that} \qquad p_{\chi^2_{p(N-1)}}(a) = p_{\chi^2_{p(N-1)}}(b)$$
$$\text{and} \quad \int_a^b p_{\chi^2_{p(N-1)}}(x)dx = 1 - P_F$$

where $p_{\chi^2_{p(N-1)}}(x)$ is the probability density function associated with the $\chi^2_{p(N-1)}$ distribution. If S is not in this interval, the single source model is rejected in favor of a multiple-source or moving-source scenario.

Part II

Practice

Chapter 6

Practical and Computational Considerations

This chapter is intended to illustrate some of the practical issues involved in calculating the location estimates detailed in Chapter 3. These estimation procedures involve nonlinear error-criterion, J_{TDOA} or J_{DOA}, the minimization of which cannot, in general, be performed analytically. Nonlinear function optimization typically entails some form of iterative search in the function parameter space. While this process may be facilitated through an efficient selection of candidate points, these techniques are computationally burdensome and subject to a host of practical considerations. In general, there is a fundamental trade-off between algorithm efficiency and robustness. The most efficient means are sensitive to discontinuities in the objective function as well as to the choice of the initial search point. When applied to an overly-complicated function, such methods will frequently obtain local minima or pursue an undesirable tangent. Meanwhile, more robust approaches involving a number of starting points or grid searches, will tend to produce more reliable results under these circumstances but at the cost of time and resources. The choice of an optimization method

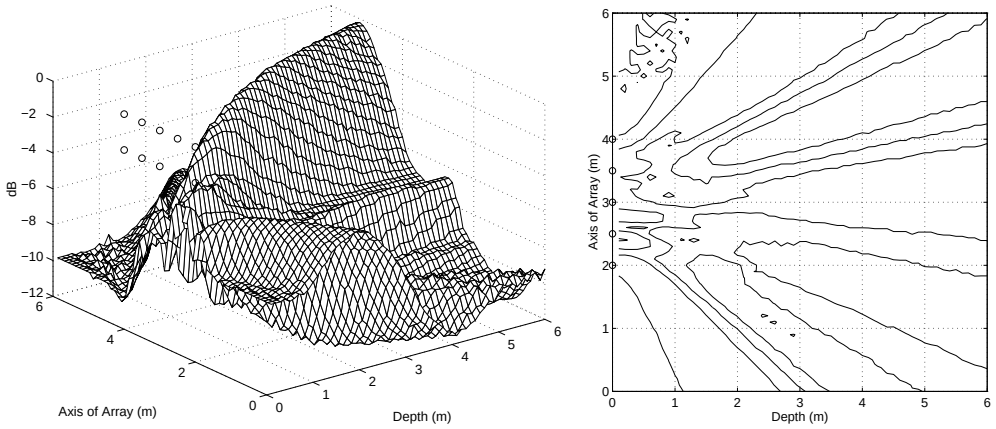


Figure 6.1: Illustration of the Error Criterion Associated with a Steered-Beamformer-Based Locator: A 3-dimensional mesh plot (left) and a contour image (right) of the beamformer output power generated by a set of candidate source locations. The source locations represent a horizontal rectangular grid at a height of 2m inside the enclosure with 10cm spacings between test points.

ideally depends upon the nature of the function to be minimized and necessitates an element of 'conventional wisdom' on the part of the user.

As will be shown in this chapter, the J_{TDOA} and J_{DOA} error criteria exhibit the continuity and unimodal properties appropriate for efficient optimization. This situation is contrasted by the objective function associated with the 'focalization' procedure alluded to in Chapter 1. This presentation is then followed by a comparison of several appropriate nonlinear optimization routines applied specifically to this localization problem.

6.1 Characterization of Error Criterion

Figure 6.1 illustrates the nature of the localization criterion associated with the steered-beamformer-based genre of locators discussed in Chapter 1. The displayed plot was produced using the ten-element bilinear array and $6m \times 6m \times 4m$ enclosure of Figure 5.2 and an ideal source located 20° off broadside at a range of 4.25m and height of 2m (identical to the height of the array midline). For this example, the sensor recordings were generated from

a high-quality speech segment by artificially applying delays appropriate for a point source at the designated location. No additional modeling was performed. The resulting set of sensor signals corresponds to a highly ideal situation, flawlessly delayed versions of identical, noiseless source signals as well as complete knowledge of the signal spectral content.

Following [21], the ML source location estimate is found by focusing the array at a number of specific locations and searching for the point that maximizes the beamformer output power. In this instance, a single 512-point window of a 20kHz discrete signal containing the vowel portion of the word 'They' was used. Figure 6.1 displays the output power in dB for a horizontal rectangular grid of test points separated by 10cm intervals. The level of the analysis plane was 2 m, the same height as the actual source location. The left-hand plot shows a 3-dimensional mesh of the power figures plotted as a function of location within the search plane. The bilinear array sensor positions, denoted by 'o', have been included to establish room orientation. Their vertical placement is arbitrary on this scale. The right-hand plot illustrates this same data via a contour image.

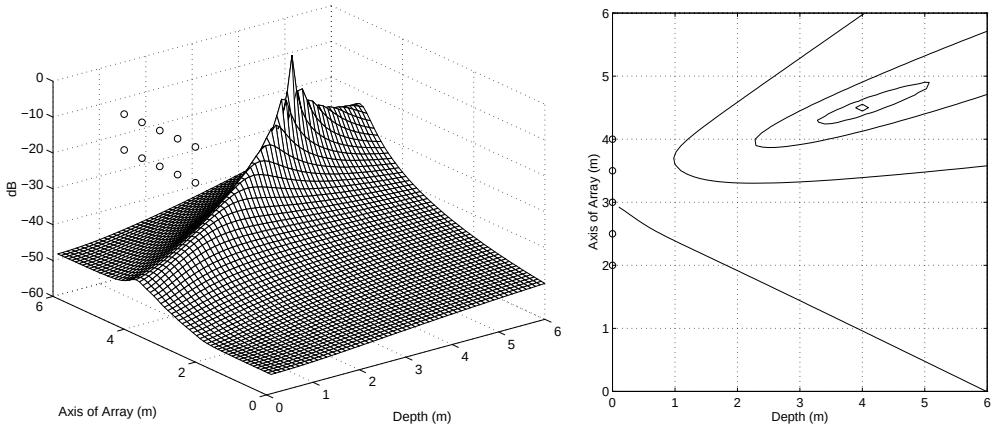
In Chapter 1, the steered-beamformer-based class of locators was dismissed as the basis for a practical speech localization system. These plots highlight some of the undesirable features associated with the genre's error criterion, making the locators computationally infeasible. As is apparent from the figure, the error function contains several local maxima and does not possess a strong peak at the true source location, even under these ideal circumstances. Reliable estimation of the global maximum in this situation would necessitate the use of a sophisticated and laborious search procedure, typically demanding an order of magnitude more function evaluations than would be required by more efficient optimization procedures. Several of these more practical iterative optimization methods will be successfully employed with TDOA-based locators in the following section but are ineffective

for this application. Furthermore, the work involved in computing the beamformer output power for a single candidate location is substantially greater than that for the J_{TDOA} or J_{DOA} error criteria. The increase in workload depends on the quantity and arrangement of the sensors. For this simulation, each calculation of the steered-beamformer error criterion represents a factor of 10 to 20 more operations than a single evaluation of the TDOA-based criteria[1].

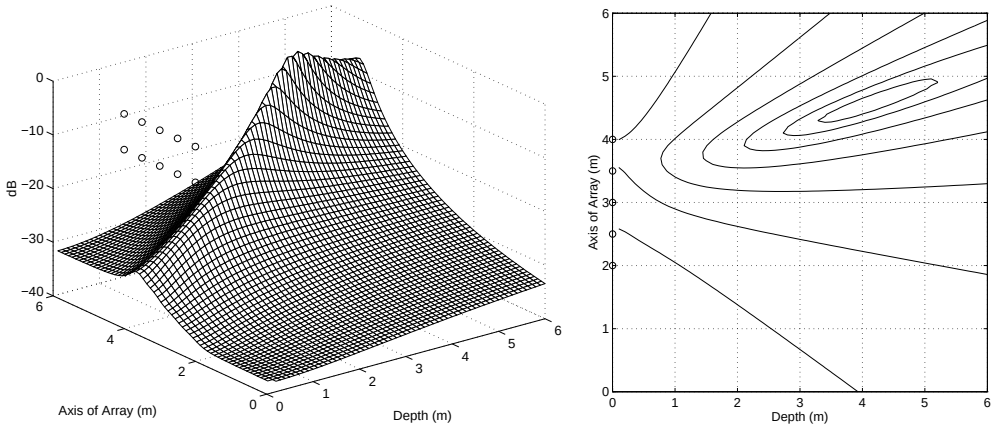
Figures 6.2 and 6.3 illustrate the results of a similar analysis conducted with the LS error criteria, J_{TDOA} and J_{DOA}, respectively. This second set of simulations involved the same source location and set of candidate test points as the steered-beamformer error criterion situation. Using the ten-element bilinear array, the sensor pairs were selected as the eight doublets of diagonally adjacent sensors. In each case, the true TDOA values for the prescribed source location were calculated relative to the individual sensor pairs and then corrupted by additive Gaussian noise. Three TDOA noise conditions were evaluated; these represented standard deviations of .001m, .01m, and .1m (when scaled by the speed of sound in air). These two localization methods require finding the global minimum of their respective error criteria. To aid in their visualization and comparison to the earlier simulation, the plots in Figures 6.2 and 6.3 display the reciprocal of the J_{TDOA} and J_{DOA} functions (in dB) and thus the location estimate would now correspond to the coordinates producing the global maximum.

The multimodal nature of the steered-beamformer error criterion with its numerous peaks and valleys is starkly contrasted by the set of functions plotted in Figures 6.2 and 6.3. The practical utility of these two TDOA-based error criteria is clearly demonstrated by

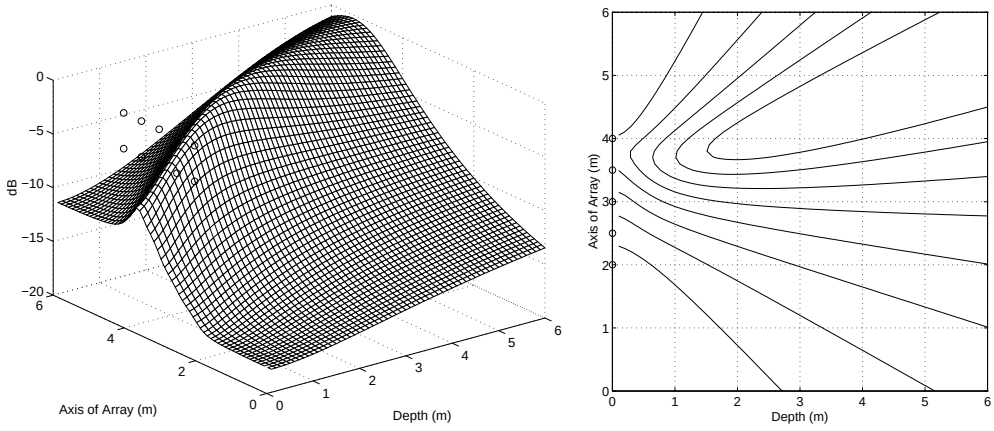
[1]Granted, the TDOA-based schemes require TDOA estimation prior to the localization stage. However, the individual TDOA's need only be evaluated once per frame and, given the independence of the sensor-pairs, this pre-processing may be easily parallelized.



(a) TDOA Standard Deviation: .001m

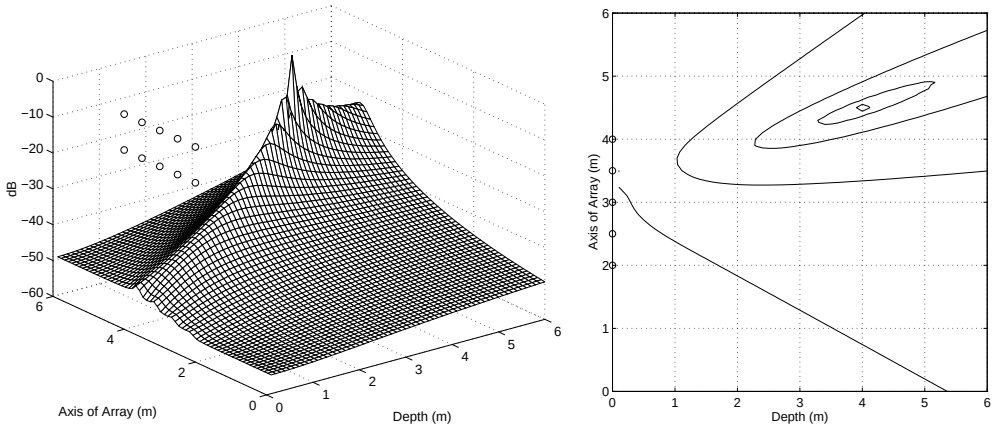


(b) TDOA Standard Deviation: .01m

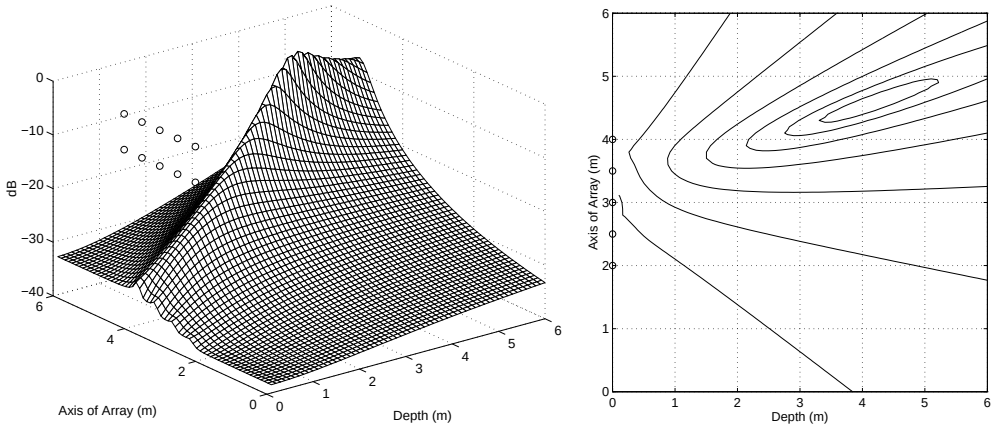


(c) TDOA Standard Deviation: .1m

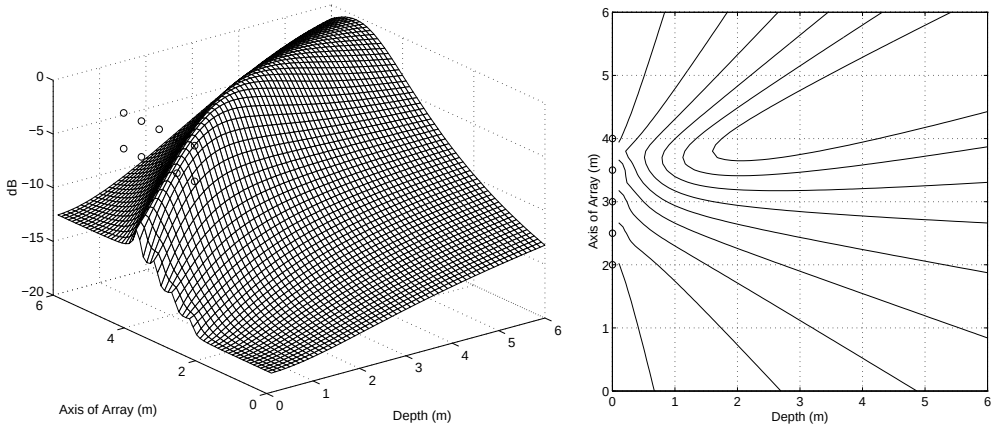
Figure 6.2: Illustration of the reciprocal J_{TDOA} Error Function Space: Evaluation of the error criterion over a horizontal, 10cm interval grid of candidate source points at a height of 2m. Each row of plots represents a different TDOA precision condition.



(a) TDOA Standard Deviation: .001m



(b) TDOA Standard Deviation: .01m



(c) TDOA Standard Deviation: .1m

Figure 6.3: Illustration of the reciprocal J_{DOA} Error Function Space: Evaluation of the error criterion over a horizontal, 10cm interval grid of candidate source points at a height of 2m. Each row of plots represents a different TDOA precision condition.

this second second simulations. For each combination of error criterion and TDOA noise condition, the resulting function is smoothly varying with a single peak; ideally suited for optimization by computationally efficient means. The dynamic range of the functions as well as the resolution of the maximum point depends greatly upon the precision of the TDOA estimates incorporated. With the .001m TDOA noise cases, a very distinct peak is evident at the actual source location. As the noise levels are increased the acme is progressively less distinguished from the alternative locations. For this source location there is very little variation between the two LS criteria themselves over the range of noise conditions. This observation is consistent with the results of the experiments in Section 3.5 where the J_{TDOA}- and J_{DOA}-based locators were found to exhibit nearly identical performance with near-broadside sources.

6.2 Comparison of Nonlinear Optimization Routines

Several Monte Carlo simulations were performed to evaluate the relative effectiveness of four representative nonlinear optimization routines. Each of these methods involves a sequential search starting from a single initial guess location and may be classified as belonging at the computationally efficient end of the efficiency vs. robustness spectrum. In the interest of standardization, the simulations were conducted using generic applications of routines included in the widely-available MATLAB software package[2]. The routines are:

Quasi-Newton: The Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [64, 65, 66, 67], a gradient-descent approach using a mixed quadratic and cubic line search. MATLAB function: **fminu**.

[2]MATLAB is a trademark of the Math Works, Inc. of Natick, MA. The basic software, as well as the Optimization Toolbox employed here, are furnished under licensing agreement.

Simplex Method: The Nelder and Mead Downhill Simplex algorithm [68], a slower, but more robust geometrical approach not requiring derivative estimates. MATLAB function: **fmins**.

SQP: Sequential Quadratic Programming (SQP) [69], an adaptation of linear programming techniques to the constrained nonlinear problem. MATLAB function: **constr**

NLLS: The Levenberg-Marquardt Algorithm [70] adapted from the nonlinear least-square (NLLS) modeling problem. MATLAB function: **leastsq**

An excellent summary of the each of these methods as well as a discussion of their relative merits and applicability is available in [71] while [72] presents their MATLAB implementation details. For the simulations that follow, the worst-case estimate precision was fixed at 1mm. The remaining search parameters were set to the default values.

The parameters for these simulations were again modeled after those of Evaluation #1 in Section5.4. The ten-element bilinear array depicted in Figure 5.2 was employed with the sensor pairs selected as the eight doublets of diagonally adjacent sensors. 100 source locations were randomly selected within the $6m \times 6m \times 4m$ rectangular enclosure, and the true TDOA values were calculated and then corrupted by additive Gaussian noise with a standard deviation of .01m. Location estimation was performed with the 100 individual TDOA sets using each combination of the error criterion and optimization routine. The entire process was performed twice, first using the center of the room as the initial guess location (full search) and then employing an initial search guess that was a random perturbation of the actual location (abbreviated search). The full search reflects the case when no prior knowledge of the source location is available. The abbreviated search is motivated by a scenario in which an approximate location has been provided, either via previous estimates

J_{TDOA}-Based Locator w/ Full Search												
	Quasi-Newton			Simplex Method			SQP			NLLS		
	min	median	max	min	median	max	min	median	max	min	median	max
Error (cm)	1.0	12.4	239.8	1.0	16.5	540.2	1.0	12.4	101.8	1.0	12.3	129.1
Iterations	31	60	332	180	253	423	41	63	170	23	41	91
FLOPS	2e+04	3e+04	2e+05	8e+04	1e+05	2e+05	3e+04	4e+04	1e+05	1e+04	2e+04	5e+04

J_{TDOA}-Based Locator w/ Abbreviated Search												
	Quasi-Newton			Simplex Method			SQP			NLLS		
	min	median	max	min	median	max	min	median	max	min	median	max
Error (cm)	1.0	12.3	129.1	1.0	12.3	129.1	1.0	12.3	80.3	1.0	12.3	129.0
Iterations	31	37	56	71	116	186	38	53	77	22	29	51
FLOPS	2e+04	2e+04	3e+04	3e+04	5e+04	8e+04	2e+04	3e+04	5e+04	1e+04	2e+04	3e+04

J_{DOA}-Based Locator w/ Full Search												
	Quasi-Newton			Simplex Method			SQP			NLLS		
	min	median	max	min	median	max	min	median	max	min	median	max
Error (cm)	0.7	12.5	692.8	0.7	19.1	10164.7	0.7	12.5	113.9	0.7	13.7	226.2
Iterations	30	46	96	164	251	779	25	59	84	28	47	95
FLOPS	2e+04	3e+04	6e+04	1e+05	2e+05	5e+05	2e+04	5e+04	7e+04	2e+04	3e+04	7e+04

J_{DOA}-Based Locator w/ Abbreviated Search												
	Quasi-Newton			Simplex Method			SQP			NLLS		
	min	median	max	min	median	max	min	median	max	min	median	max
Error (cm)	0.7	12.0	130.4	0.8	12.4	130.4	0.7	12.2	80.3	0.7	12.1	130.4
Iterations	17	36	80	53	111	240	23	40	78	23	35	83
FLOPS	1e+04	2e+04	5e+04	3e+04	7e+04	1e+05	2e+04	3e+04	6e+04	2e+04	3e+04	6e+04

Table 6.1: Comparison of Nonlinear Optimization Routines: Minimum, median, and maximum values out of the 100 trials for the four optimization routines (Quasi-Newton, Simplex Method, SQP, and NLLS) reported for three categories: error (the Euclidean distance between the estimated and actual locations), the number of algorithm iterations, and the total number of floating point operations (FLOPS) utilized as given by the MATLAB **flops** function. The top two charts represent the results obtained for the J_{TDOA}-based estimator while the bottom two give those of the J_{DOA}-based estimator. Correspondingly, the top table in each pair lists the full-search condition; the lower tables display the results of the abbreviated search.

or through a preliminary closed-form estimation procedure (see Chapter 7). In the case of an abbreviated search, the initial search location was computed by adding .25m standard

deviation Gaussian noise to each of the 3 dimensional values of the true location. Table 6.1 presents the collective results of these simulations. For each of the four optimization routines (Quasi-Newton, Simplex Method, SQP, and NLLS) the minimum, median, and maximum values out of the 100 trials are reported for three categories: error (the Euclidean distance between the estimated and actual locations), the number of algorithm iterations, and the total number of floating point operations (FLOPS) utilized as given by the MATLAB **flops** function. The top two charts represent the results obtained for the J_{TDOA}-based estimator while the bottom two give those of the J_{DOA}-based estimator. Correspondingly, the top table in each pair lists the full-search condition; the lower tables display the results of the abbreviated search.

First, with regard to estimate error relative to the LS error criterion employed, these results are consistent with the the simulations of Section 3.5. The J_{TDOA} and J_{DOA} criteria yield comparable estimator performance at this moderate noise condition. With the exception of the NLLS routine, the J_{DOA} searches exhibit a tendency to converge to their final estimate in fewer iterations. However, this does not result in reduced computational load on the part of these methods. Each evaluation of the J_{DOA} criterion requires the calculation of direction-of-arrival data in addition to a weighted sum. The result is more operations per iteration in comparison to the J_{TDOA} measure, offsetting the gains of fewer algorithm steps. Consequently, the FLOPS figures associated with the J_{DOA} are generally higher than those given for their J_{TDOA}-based counterparts.

Several observations may be made regarding the performance of the individual optimization routines. First, the Simplex Method is clearly the least desirable, possessing by far the worst precision, most iterations, highest FLOPS counts, and surprisingly, the least robustness of the four routines tested. The distinctions between the remaining three algorithms

are less apparent. The Quasi-Newton, SQP, and NNLS methods offer nearly equivalent error results under the respective testing conditions. The SQP method does present some evidence of increased robustness, consistently obtaining the least maximum error scores. In terms of iterations and FLOPS values, there is a distinct trend which is dependent upon the LS error criterion employed. For the J_{TDOA}-based estimates, the NLLS technique achieves the highest computational efficiency. With the J_{DOA}-based locators, the Quasi-Newton technique has a small computing advantage. Not surprisingly, each of these four optimization methods benefited significantly from the inclusion of a prescient initial search location. The results of the abbreviated searches exhibit dramatically improved figures for each of the performance statistics evaluated relative to their full-search equivalents. The reduction in iteration and total operation counts are to expected because of the reduced search distance; the improvements in the error figures are indicative of the overall robustness of the methods to the nature of the objective function utilized and their dependence upon initial guess selection. With the general search, the Quasi-Newton and Simplex Methods exhibit the highest tendency to obtain local minima or pursue an incorrect tangent. Correspondingly, they appear to benefit the most from an informed initial guess.

These simulations were intended to illustrate the applicability of a class of optimization techniques to this problem. As for the selection of a single routine, none of these approaches appears to offer a distinct advantage under all the testing conditions. Clearly, in terms of performance criteria alone, there is no reason to utilize the Simplex Method. If the selection process is guided strictly by computational efficiency, either the Levenberg-Marquardt (NLLS) or the Quasi-Newton BFGS algorithms may be preferable. The Sequential Quadratic Programming (SQP) approach offers a modicum of added robustness without a sizable increase in computational demands, and for these reasons, was the optimization

function used in the experiments performed for this work. Additional considerations may be relevant to a specific application. For instance, the size of the software encoding or the ease of translating an algorithm to a specific real-time hardware platform, may be cause to favor one optimization routine over another of comparable performance.

6.3 Discussion

In the first section of this chapter, the general nature of the steered-beamformer objective function and TDOA-based localization criteria, J_{TDOA} and J_{DOA}, were characterized. The latter were found to be amenable to minimization by efficient nonlinear iterative optimization procedures. The second section evaluated several such algorithms, demonstrating their overall effectiveness as a means for obtaining reliable and accurate TDOA-based location estimates as well as comparing their respective performance attributes. Each of the routines evaluated was a 'canned' package and entailed no specific tailoring for this task. In an effort to further aid search robustness and efficiency, it may be possible to incorporate specific knowledge of an error criterion's nature into the optimization process. However, given the success and practicality of these existing methods, no attempts along those lines have been made to date.

Chapter 7

A Closed-Form Source Localization Algorithm

Each of the LS criterion-based estimators detailed in Chapter 3 involves the minimization of an error measure that is a nonlinear function of the potential source location. As a result, these estimators require a numerical search of a potential location space (a subset of $\mathcal{R}^3$). While the utility of these objective spaces for minimization by efficient search algorithms that rapidly converge to the desired location estimate has been demonstrated, there may be applications where a full-search is not feasible due to limited computational resources. This is particularly true for real-time situations requiring a high update rate and/or many sensors. These circumstances necessitate the development of closed-form location estimators that, while providing sub-optimal localization data, are computationally inexpensive. For those circumstances where the optimal estimate is required the closed-form solution may be used as an intermediate solution, providing the initial starting point for a less burdensome, partial search.

A closed-form solution to the source localization problem, termed the *linear intersection*

(LI) method, is presented in this chapter. The algorithm is derived in the context of the sensor-pair geometry developed in this work and shown to provide results on par with the search-based estimators as well as performance superior to that of a quality, previously-reported algorithm.

7.1 Closed-Form Location Estimation

In Section 3.2, it was shown that for the case of time-delay estimates corrupted by additive white Gaussian noise, the J_{TDOA}-based estimator, $\hat{\mathbf{s}}_{TDOA}$, produced the Maximum Likelihood location estimate. Under certain conditions, the J_{DOA}-based estimate, $\hat{\mathbf{s}}_{TDOA}$, was found to yield performance results superior to those of its TDOA-based counterpart. Each of these estimates is found via minimization of the sum-squared error of the differences between the observed TDOA or DOA and those of the hypothesized source, $\mathbf{s}$. Because these error criteria are non-linear functions of $\mathbf{s}$, neither estimate possesses a closed-form solution. For the location estimator presented here and the many others found in the literature the requirement of a closed-form solution necessitates the development of alternative error criteria. These alternative error criteria take several forms and vary in the degree to which they approximate the LS error criteria and perform in comparison to the search-based estimators. A discussion of several of these closed-form estimators as well as a relative performance evaluation is presented in [41]. Smith and Abel found their proposed estimation procedure, an approach which linearizes the TDOA differences and obtains an estimate through a linear least-squares matrix solution, to exhibit an RMS error superior to that of the estimators presented in [46] and [39]. Their estimation procedure is termed the *spherical interpolation (SI)* method and will be employed in Section 7.3 as a benchmark for evaluating the proposed closed-form algorithm.

Existing closed-form solutions to this type of localization problem have been developed with different situations in mind. Radar, sonar, and global positioning are the most common examples. These applications differ from the speech source localization problem addressed here in several respects. Primarily, the TDOA estimates for these other scenarios are evaluated relative to an absolute time-scale or a single reference sensor. The sensor-pair geometry detailed in this work requires only that TDOA estimates be found between isolated doublets of sensors. This generalization has been imposed out of necessity. Given a general placement of sensors within an environment and realistic speech sources possessing non-ideal radiation patterns, there is no assurance that the received signal coherence across the span of the sensors will be appropriate to allow precise TDOA estimation relative to a single reference sensor. In Section 5.5, it was argued that such conditions restrict the intra-pair separation distance in practice. The closed-form location estimator presented in the following section is derived specifically from the context of the sensor-pair geometry and is designed to closely approximate the ML estimator.

7.2 The Linear Intersection Algorithm

As presented in Chapter 2, given a specific sensor pair $\{\mathbf{m}_{i1}, \mathbf{m}_{i2}\}$ and their associated TDOA estimate, τ_i, the locus of potential source locations in 3-space forms one-half of a hyperboloid of two sheets. This hyperboloid is centered about the midpoint of $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$ and has the directed line segment $\overline{\mathbf{m}_{i1}\mathbf{m}_{i2}}$ as its axis of symmetry. For sources with a large source-range to sensor-separation ratio, the hyperboloid may be well-approximated by the cone with vertex at the sensors' midpoint, having $\overline{\mathbf{m}_{i1}\mathbf{m}_{i2}}$ as a symmetry axis, and a constant direction angle relative to this axis. The direction angle, θ_i, for a sensor-pair,

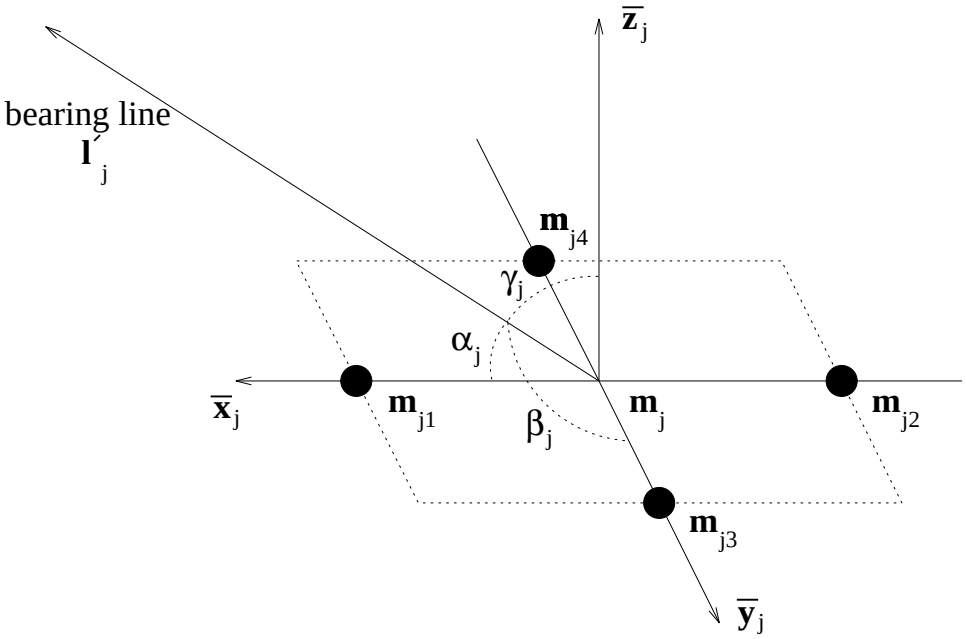


Figure 7.1: Quadruple sensor arrangement and local Cartesian coordinate system

TDOA-estimate combination is given by:

$$\theta_i = \cos^{-1}\left(\frac{c \cdot \tau_i}{|\mathbf{m}_{i1} - \mathbf{m}_{i2}|}\right) \tag{7.1}$$

Now consider two pairs of sensors $\{\mathbf{m}_{j1}, \mathbf{m}_{j2}\}$ and $\{\mathbf{m}_{j3}, \mathbf{m}_{j4}\}$, where j is used to index the sets of sensor quadruples, along with their associated TDOA estimates, τ_{j12} and τ_{j34}, respectively. The sensors' placement positions are constrained to lie on the midpoints of a rectangle and as a result $\overline{\mathbf{m}_{j1}\mathbf{m}_{j2}}$ and $\overline{\mathbf{m}_{j3}\mathbf{m}_{j4}}$ are orthogonal and mutually bisecting. A local Cartesian coordinate system is established with unit vectors defined as $\overline{\mathbf{x}_j} = \frac{\overline{\mathbf{m}_{j1}\mathbf{m}_{j2}}}{|\mathbf{m}_{j1}-\mathbf{m}_{j2}|}$, $\overline{\mathbf{y}_j} = \frac{\overline{\mathbf{m}_{j3}\mathbf{m}_{j4}}}{|\mathbf{m}_{j3}-\mathbf{m}_{j4}|}$, and $\overline{\mathbf{z}_j} = \overline{\mathbf{x}_j} \times \overline{\mathbf{y}_j}$ with the origin at the common midpoint of the two pairs, denoted by $\mathbf{m}_j$. This geometry is depicted in Figure 7.1. The first sensor-pair TDOA-estimate approximately determines a cone with constant direction angle, α_j, relative to the $\overline{\mathbf{x}_j}$ axis, as given by (7.1). The second specifies a cone with constant direction angle, β_j, relative to the $\overline{\mathbf{y}_j}$ axis. Each has a vertex at the local origin. If the potential source location

is restricted to the positive-z half-space, the locus of potential source points common to these two cones is the bearing line, $\mathbf{l'}_j$, in 3-space. The remaining direction angle, γ_j, may be calculated from the identity

$$\cos^2 \alpha_j + \cos^2 \beta_j + \cos^2 \gamma_j = 1$$

with $0 \leq \gamma_j \leq \frac{\pi}{2}$ and the line may be expressed in terms of the local coordinate system by the parametric equation

$$\mathbf{l'}_j = \begin{bmatrix} x_j \\ y_j \\ z_j \end{bmatrix} = r_j \begin{bmatrix} \cos \alpha_j \\ \cos \beta_j \\ \cos \gamma_j \end{bmatrix} = r_j \mathbf{a'}_j$$

where r_j is the range of a point on the line from the local origin at $\mathbf{m}_j$ and $\mathbf{a'}_j$ is the vector of direction cosines. The line $\mathbf{l'}_j$ may then be expressed in terms of the global Cartesian coordinate system via the appropriate translation and rotation. Namely,

$$\mathbf{l}_j = r_j \mathbf{R}_j \mathbf{a'}_j + \mathbf{m}_j$$

where $\mathbf{R}_j$ is the 3×3 rotation matrix from the j^{th} local coordinate system to the global coordinate system. Alternatively, if $\mathbf{a}_j$ represents the rotated direction cosine vector then

$$\mathbf{l}_j = r_j \mathbf{a}_j + \mathbf{m}_j$$

Given M sets of sensor quadruples and their corresponding bearing lines

$$\mathbf{l}_j = r_j \mathbf{a}_j + \mathbf{m}_j \qquad \text{for } j = 1, \ldots, M$$

the problem of estimating a specific source location remains. The approach taken here will be to calculate a number of potential source locations from the points of closest intersection for all pairs of bearing lines and use a weighted average of these locations to generate a final source-location estimate. Figure 7.2 illustrates the process used to determine the points of closest intersection. Specifically, given two bearing lines

$$\begin{aligned} \mathbf{l}_j &= r_j \mathbf{a}_j + \mathbf{m}_j \\ \mathbf{l}_k &= r_k \mathbf{a}_k + \mathbf{m}_k \end{aligned} \tag{7.2}$$

the shortest distance between the lines is measured along a line parallel to their common normal and is given by [73]:

$$d_{jk} = \frac{|(\mathbf{a}_j \times \mathbf{a}_k) \cdot (\mathbf{m}_j - \mathbf{m}_k)|}{|\mathbf{a}_j \times \mathbf{a}_k|}$$

Accordingly, the point on $\mathbf{l}_j$ with closest intersection to $\mathbf{l}_k$ (denoted by $\mathbf{t}_{jk}$) and the point on $\mathbf{l}_k$ with closest intersection to $\mathbf{l}_j$ (denoted by $\mathbf{t}_{kj}$) may be found by first solving for the local ranges, r_j and r_k, and substituting these values into (7.2). The local ranges are found via solution of the overconstrained matrix equation:

$$\left[\begin{array}{c|c} \mathbf{a}_j & -\mathbf{a}_k \end{array} \right] \left[\begin{array}{c} r_j \\ \\ r_k \end{array} \right] = \left[\begin{array}{c} \mathbf{m}_k - \mathbf{m}_j + d_{jk} \cdot (\mathbf{a}_j \times \mathbf{a}_k) \end{array} \right]$$

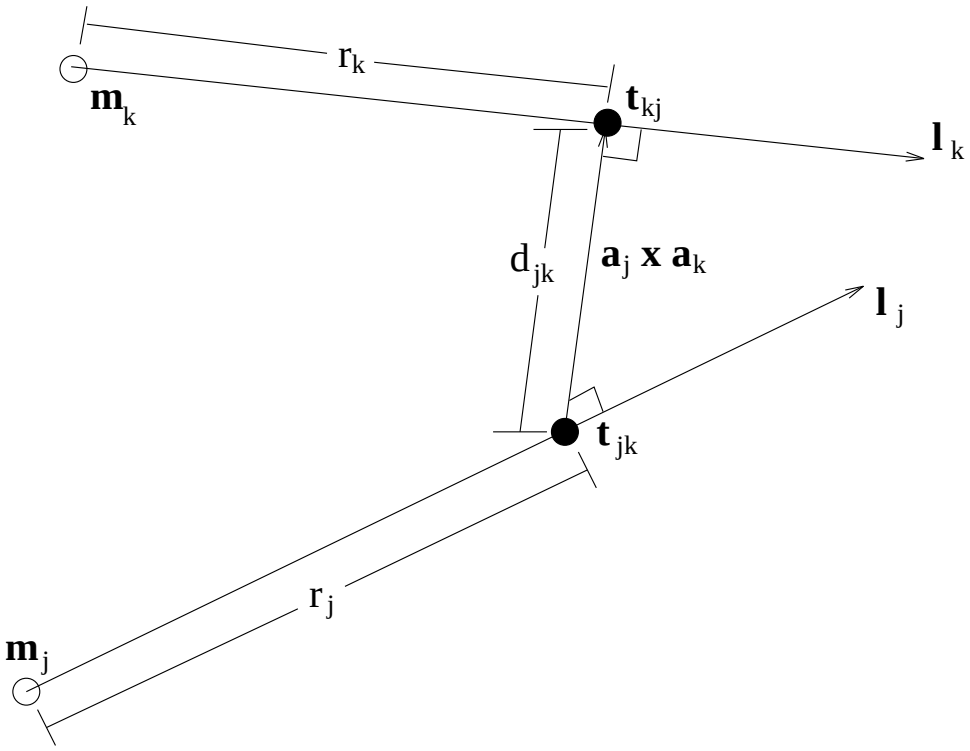


Figure 7.2: Illustration of the process used to calculate the points of closest intersection, $\mathbf{t}_{jk}$ and $\mathbf{t}_{kj}$, for a pair of bearing lines, $\mathbf{l}_j$ and $\mathbf{l}_k$, in 3-space.

Each of the potential source locations is weighted based upon its probability conditioned on the observed set of $2M$ sensor-pair, TDOA-estimate combinations. The TDOA estimates are assumed to be normal distributions with mean given by the estimate itself. The weight associated with the potential source location, $\mathbf{t}_{jk}$, is calculated from:

$$\begin{aligned} W_{jk} &= \prod_{l=1}^{M} P(\mathcal{T}(\{\mathbf{m}_{l1}, \mathbf{m}_{l2}\}, \mathbf{t}_{jk}), \tau_{l12}, \sigma_{l12}^2) \\ &\quad \cdot P(\mathcal{T}(\{\mathbf{m}_{l3}, \mathbf{m}_{l4}\}, \mathbf{t}_{jk}), \tau_{l34}, \sigma_{l34}^2) \end{aligned} \tag{7.3}$$

where $P(x, m, \sigma^2)$ is the value of a Gaussian probability distribution function with mean m

and variance σ^2 evaluated at x, i.e.

$$P(x, m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp(\frac{-(x-m)^2}{2\sigma^2})$$

The final location estimate, which will be referred as the *linear intersection estimate* ($\hat{\mathbf{s}}_{LI}$), is then calculated as the weighted average of the potential source locations:

$$\hat{\mathbf{s}}_{LI} = \frac{\sum_{j=1}^{M} \sum_{\substack{k=1, \\ k \neq j}}^{M} W_{jk} \mathbf{t}_{jk}}{\sum_{j=1}^{M} \sum_{\substack{k=1, \\ k \neq j}}^{M} W_{jk}} \tag{7.4}$$

Evaluated in this manner, $\hat{\mathbf{s}}_{LI}$ represents the expected value of a partially known random variable. The points of closest intersection, $\mathbf{t}_{jk}$, are assumed to be points of high probability clustered about the peak of a symmetrical probability distribution. In this sense, the LI algorithm attempts to model the ML estimate which searches for the maximum in the joint probability distribution of the TDOA estimate set.

Figure 7.3 depicts the LI localization method for the sensor arrangement shown in Figure 5.4 of Section 5.4. Note that each of the four quadruple units satisfies the LI sensor positional constraint when sensor-pairs are selected from the diagonal elements at the vertices of each square. The top graph in the figure displays the bearing lines projecting from the quadruple units for a simulated source at location $(2m, 4m, 3m)$. To generate this situation, the true TDOA values have been corrupted by additive noise with a standard deviation of .01m. The points of closest intersection, $\mathbf{t}_{jk}$, and the final LI estimate are shown along with their projections onto the xy-, xz-, and yz-planes. The bottom graph in Figure 7.3 presents an enlarged overhead view of the intersection region alone. The individual $\mathbf{t}_{jk}$ locations are now visible and denoted by 'x' and the corresponding normal

vectors are plotted as dashed lines. For these 4 bearing lines, there are a total of $\binom{4}{2} = 6$ normal vectors and 12 $\mathbf{t}_{jk}$ locations. The final locations estimate, $\hat{\mathbf{s}}_{LI}$, is indicated by the '*' near the center of the graph, surrounded by the intermediate points of closest intersection. In this example, the final estimate was found to deviate from the actual location by less than 3cm in any dimension.

7.3 Closed-Form Estimator Comparison

As a means of evaluating the LI location estimator, the statistical characteristics of the LI and spherical interpolation localization methods were compared through a series of Monte Carlo simulations modeled after those conducted in [41]. The experimental set-up, a nine-sensor orthogonal array with half-meter spacings and a source located at a range of 5m with equal direction angles, is depicted in Figure 7.4. The true TDOA values (2.2) were corrupted by additive white Gaussian noise. 100 trials were performed at each of 11 noise levels ranging from a standard deviation the equivalent of 10^{-3}m to 10^{-1}m when scaled by the propagation speed of sound in air. The LI method partitioned the array into 4 square sensor quadruples and required the evaluation of 8 TDOA estimates, one for each diagonal sensor-pair. The SI method required that all the TDOA values be relative to a reference sensor. The sensor at the origin was chosen for this purpose and the TDOA for each of the remaining 8 sensors relative to the reference were calculated. In addition to calculating the LI and SI estimates, the ML estimate, $\hat{\mathbf{s}}_{TDOA}$, was computed via a search method with the initial guess set equal to the true location.

Figures 7.5 and 7.6 summarize the results of these simulations. The top plot in Figure 7.5 presents the sample bias for the estimated source bearing and range for each of the estimation methods as a function of the level of noise added to the true TDOA values.

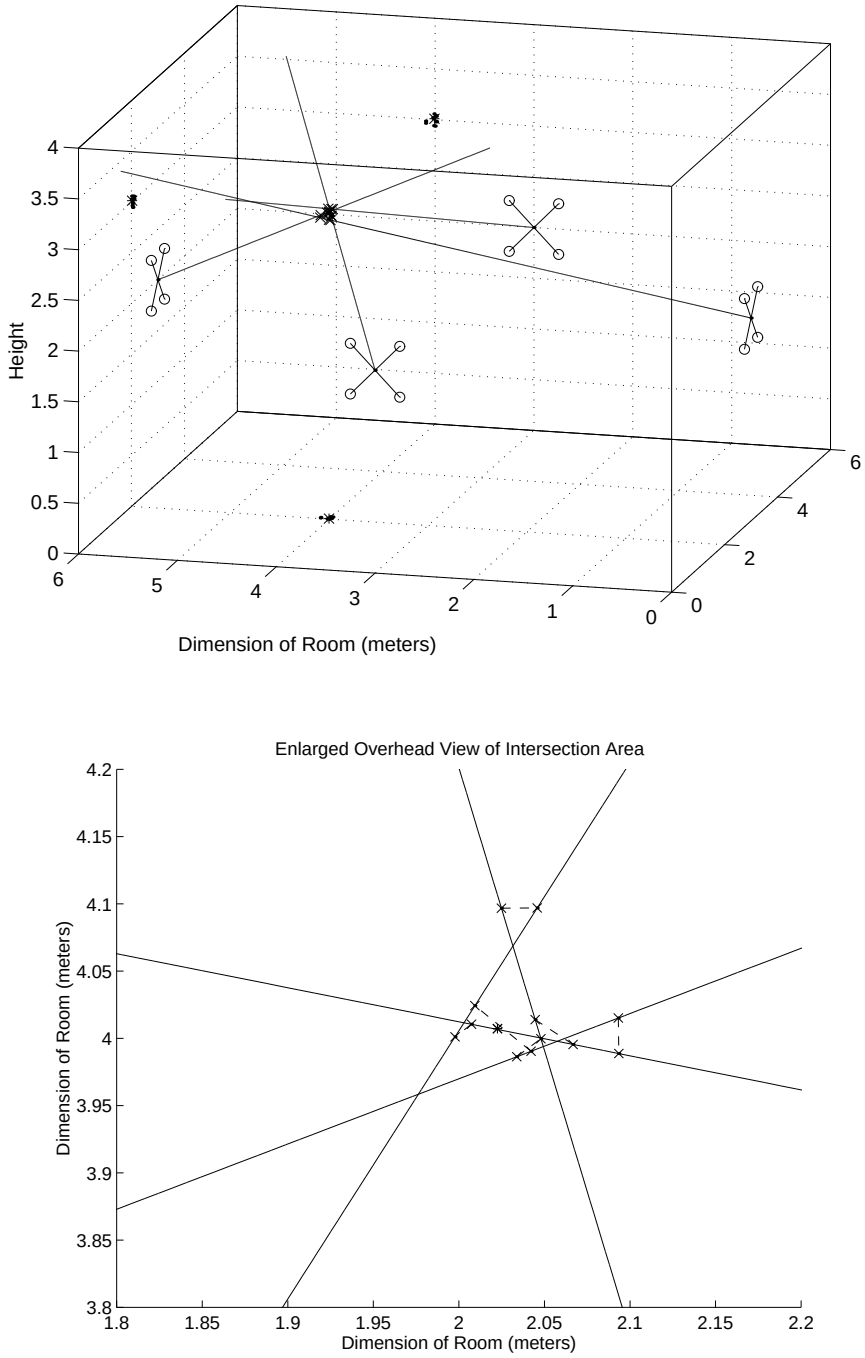


Figure 7.3: Illustration of Linear Intersection Algorithm. The top graph shows the bearing lines projecting from the quadruple units for a simulated source at location $(2, 4, 3)$ along with the $\hat{\mathbf{s}}_{LI}$ and $\mathbf{t}_{jk}$ locations and their projections onto the xy-, xz-, and yz-planes. The bottom graph is an enlarged overhead view of the intersection region alone. The individual $\mathbf{t}_{jk}$ locations are now visible and denoted by 'x's, their corresponding normal vectors by dashed lines, and the final LI estimate by a '*'.

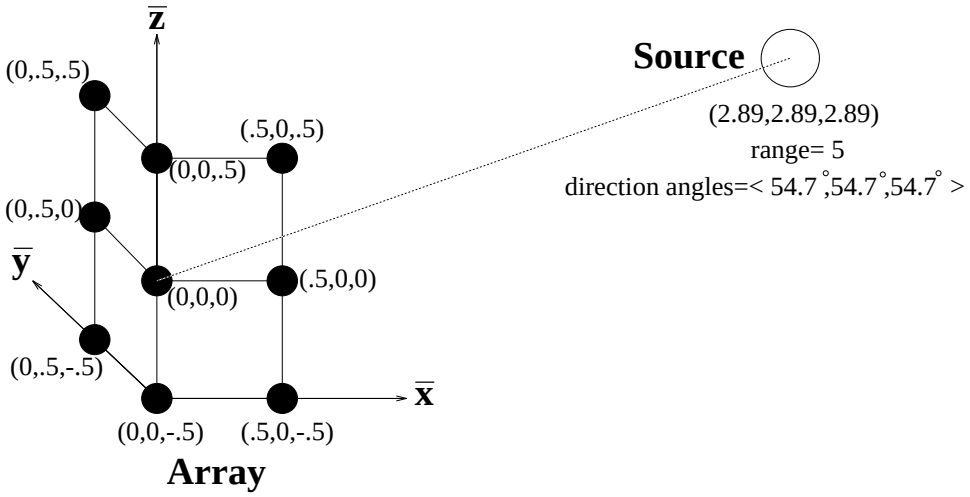


Figure 7.4: Closed-Form Estimator Comparison: The nine-element orthogonal array used for the comparison simulations. All distances are in meters.

While each of the methods exhibits some degree of bias in the noisier trials, the situation is most extreme for the SI method. This tendency for the SI method to consistently bias its estimates towards the origin was noted by the authors of [41]. The SI algorithm may be shown to approximately minimize an error criterion similar to that of the discarded J_D LS error detailed in Section 3.8. Indeed, the simulation results found here for the SI estimate are very similar to those of the $\hat{\mathbf{s}}_D$ estimator used in conjunction with the analysis in Section 3.5. The LI method performs comparably to the ML estimate for all but the most extreme noise conditions. The bottom plot Figure 7.5 shows the sample standard deviations. For the standard deviation of the bearing estimates, a trend similar to the bearing bias is observed. The SI method's performance decays rapidly for noise levels above 10^{-2}m. However, in terms of the range, each of the closed form estimators displays a smaller variance than the ML estimator at the higher noise conditions. This is a consequence of the estimator biases observed previously. Finally, In Figure 7.6 the root-mean-square errors (RMSE) are illustrated. Once again, the LI method closely tracks the ML estimator in all but the most extreme condition while the SI method exhibits a marked performance

decrease in both bearing and range for moderate and large noise levels.

Simulations performed over a broad range of source positions exhibit trends similar to those in Figures 7.5 and 7.6 The LI estimator is consistently less sensitive to noise conditions and possesses a significantly smaller bias in both its range and bearing estimates when compared to the SI estimator.

7.4 Discussion

A closed-form method for the localization of source positions given only TDOA information has been presented. It was shown to be a robust and accurate estimator, closely modeling the ML estimator, and clearly outperforming a representative algorithm.

From an implementation standpoint, the constraint that the array be composed of rectangular 4-element sub-arrays is not problematic for typical room-oriented microphone-array applications. It is an advantage of the LI method that localization in 3-space can be performed with a 2-dimensional array. The SI method as well many similar approaches requires that the matrix of sensor locations be full-rank. This necessitates the use of a 3-dimensional sensor for localization in 3-space. Also, since the LI method does not require the estimation of delays between sensors other than those in the local sub-array, the sub-arrays can be placed far apart and delay-estimation processing can be performed locally. The SI algorithm evaluates all TDOA estimates relative to a single reference sensor.

In Chapter 9 The *linear intersection* method will be used in conjunction with several real microphone-array systems. It will be shown to be an effective source localization procedure when used alone or as a means of providing initial search conditions to the more computationally demanding search-based algorithms.

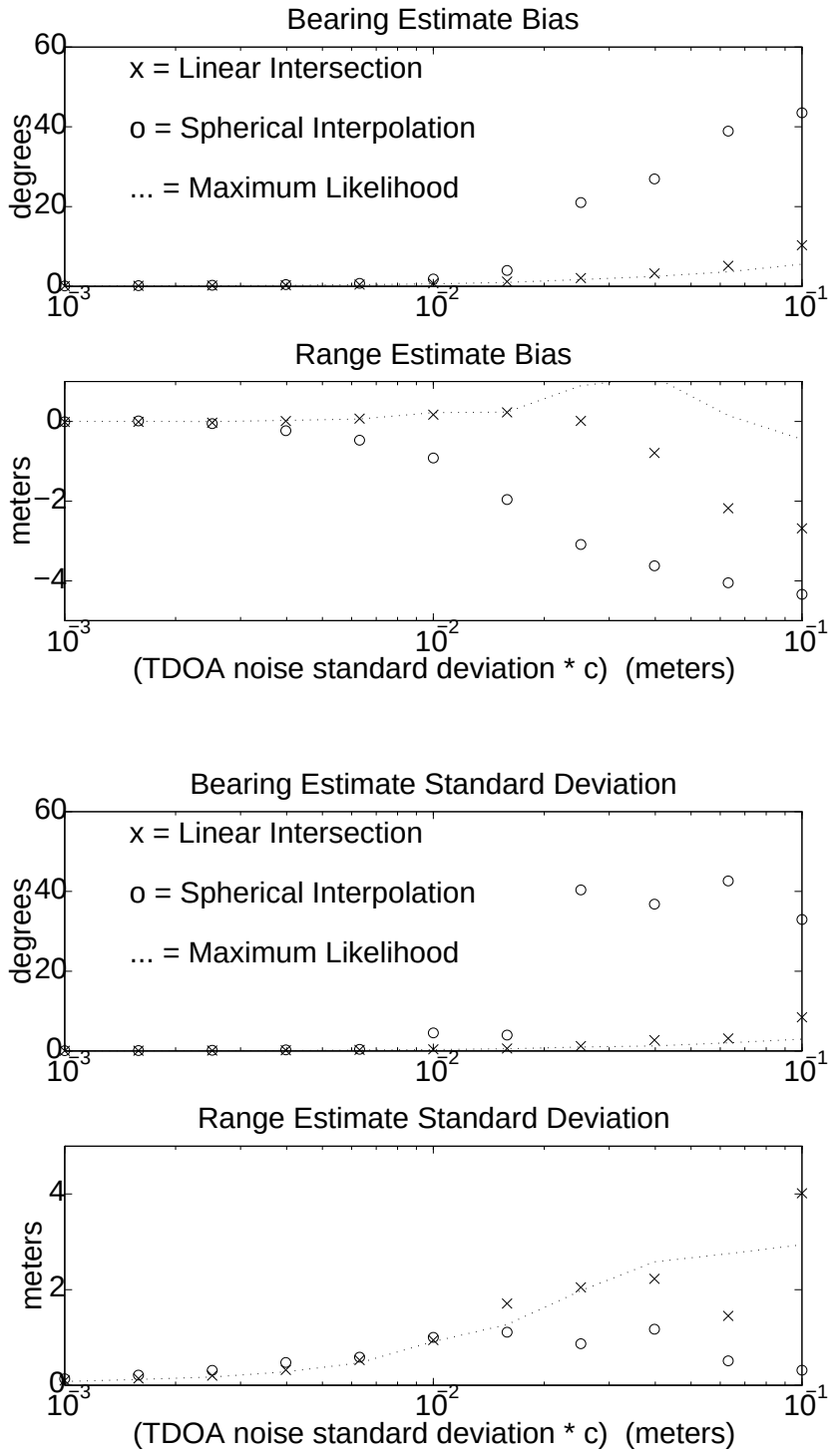


Figure 7.5: Closed-Form Estimator Comparison: Sample bias and standard deviation plots for the three estimation procedures, LI, SI, and ML, as a function of the level of noise added to the true TDOA values.

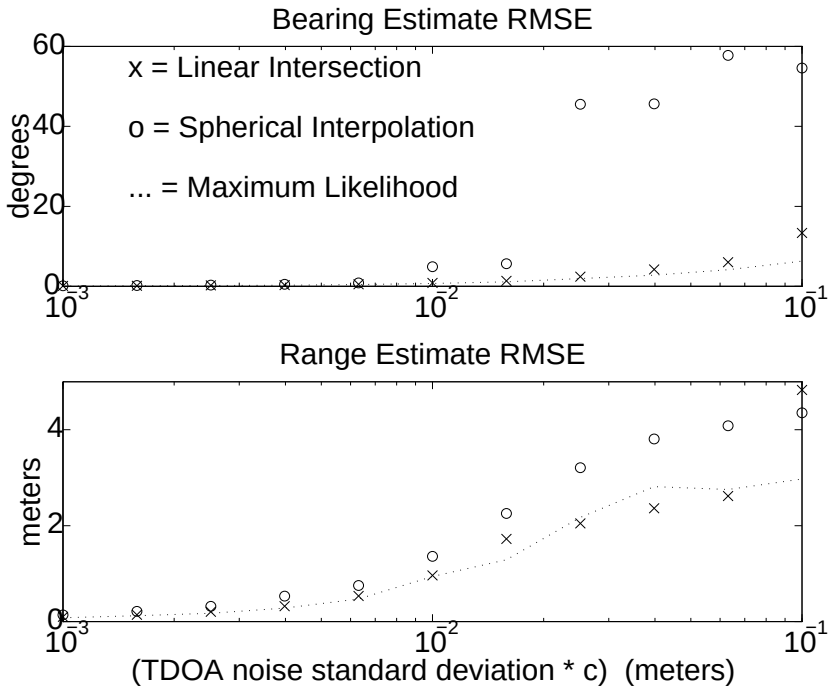


Figure 7.6: Closed-Form Estimator Comparison: Root mean square error plots for the three estimation procedures, LI, SI, and ML, as a function of the level of noise added to the true TDOA values.

Chapter 8

A Practical TDOA Estimator for Speech Sources

A fundamental requirement of the source-localization procedures presented in the preceding chapters is the ability to determine the relative time delay between signal arrivals at distinct sensor locations. The precision and robustness of these TDOA estimates are crucial factors in the quality of an associated source-location scheme. In addition to high accuracy, these delay estimates must be updated frequently in order to be useful in practical tracking and beamforming applications. Furthermore, any such estimator would also have to be computationally non-demanding to make it practical for real-time systems.

In general, correlation strategies have been used for estimating the time delay between signals received at two spatially distinct sensors. Specifically, the cross-correlation function of the two signals is computed, filtered in some "optimal" sense, and the maximum is found with a peak detector [74, 75, 76]. While the filtering criteria and the methods used for peak detection vary considerably, these techniques are all based on maximizing the cross-correlation function. Estimate resolution is limited by the sampling period unless some kind

of interpolation method is employed. These methods range from upsampling the signal to parabolic fitting of the cross-correlation function [74]; for each there is a general trade-off between the increased accuracy achieved and the computational expense incurred by the procedure. This genre of delay estimation has been applied to the same problem addressed in this work, source localization in the radiation field of a microphone sensor array.

In this chapter a frequency-domain TDOA estimator appropriate for a speech source application is described. It is designed to provide high resolution estimates in a single-source environment, to have minimal computational requirements, and to be capable of providing independent delay estimates many times (≈ 70) a second. The TDOA estimator is evaluated and shown to be extremely accurate under a wide range of signal conditions.

8.1 Mathematical Development

Consider two microphone receivers in the acoustic-field of a single speech source. Assuming that microphone placement is such that relative signal attenuation between the microphones due to propagation distance and source size and orientation are negligible, the sampled received signals, $r_1(l)$ and $r_2(l)$, may be expressed as:

$$\begin{aligned} r_1(l) &= s(l) + n_1(l) \\ r_2(l) &= s(l-\tau) + n_2(l) \end{aligned} \tag{8.1}$$

where l is the discrete-time index, $n_1(l)$ and $n_2(l)$ are background noise sources with known statistical characteristics and assumed to be uncorrelated to $s(l)$ and each other, and τ is the TDOA in sample units of the source wavefront between the receivers.

The problem here is to estimate τ from finite-duration sequences of the processes $r_1(l)$

and $r_2(l)$. In typical situations this delay will vary significantly with time, due to the physical movement, (e.g. body and head motion) of the audio source. Measurement consistency is affected by the time-varying nature of the source signal. For instance, a typical speech source may only be considered statistically stationary over a short time frame ($\approx$ 30 ms) and will have periods of signal production interspersed with durations of silence. For these reasons it is advantageous to estimate τ periodically using a small analysis window and to avoid inter-frame averaging in the signal analysis. In what follows, a restriction is imposed that the proposed TDOA estimator must compute an independent estimate of τ from a single 20-30 ms frame of data.

The DFT coefficients of the N-point, windowed received signals in (8.1) and their cross-spectrum are given by

$$\begin{aligned} R_1(k) &= W(k) * (S(k) + N_1(k)) \\ R_2(k) &= W(k) * (S(k)e^{-j\omega_k\tau} + N_2(k)) \\ G_{R_1R_2}(k) &= R_1(k)R_2(k)' \end{aligned}$$

where $W(k)$ is the N-point DFT of the analysis window, $k = 0, 1, \ldots, \frac{N}{2}$, $\omega_k = \frac{2\pi k}{N}$, and $*$ and $'$ denote the convolution and complex conjugate operators, respectively. The TDOA τ now appears as part of the complex phase term and as such, is not restricted to integer values. The phase of the cross-spectrum, may be expressed as

$$\theta_k = arg(G_{R_1R_2}(k)) = \omega_k\tau + \epsilon_k. \tag{8.2}$$

Here ϵ_k, the phase deviation, is a random variable that summarizes the contributions of the noise terms and analysis window to the overall phase term at each discrete frequency. Given

that ϵ_k is zero-mean for all k (demonstrated in Table 8.1), the expected value of the phase term, θ_k, is directly proportional to the discrete radian frequency, ω_k, with the constant of proportionality being the signal delay, τ. *i.e.*

$$E(\theta_k) = \omega_k \tau$$

In this sense τ may be interpreted as the slope of the line that "fits" the series of phase terms. Assuming that the ϵ_k terms are uncorrelated (In the case of Gaussian noise sources, this assumption is valid for the wideband speech signals and observation intervals considered here. See [77].), the best linear unbiased estimator of τ is given by the expression [78]:

$$\hat{\tau} = \frac{\sum_{k=1}^{N-1} W_k \omega_k \theta_k}{\sum_{k=1}^{N-1} W_k \omega_k^2} \tag{8.3}$$

where W_k are weighting coefficients equivalent to the reciprocal of the phase deviation variance, i.e.

$$W_k = \frac{1}{var(\epsilon_k)}$$

The variance associated with the estimate $\hat{\tau}$ is calculated from:

$$var(\hat{\tau}) = \frac{1}{\sum_{k=1}^{N-1} W_k \omega_k^2} \tag{8.4}$$

The above analytical expression for calculating $\hat{\tau}$ has several advantages over its time-domain counterpart. It is computationally simple, does not necessitate the use of search methods, and, as will be shown, is capable of intra-sample precision. In addition, if the ϵ_k terms are Gaussian, $\hat{\tau}$ can be shown to be the minimum variance unbiased(MVU) estimator of τ as well [78].

8.1.1 Calculation of Estimator Parameters

In practice, the variance terms required for (8.3) and (8.4) are unavailable *a priori* and must be evaluated directly from the data. A means for computing these variances using the magnitude-squared coherence of the spectra derived from overlapping windowed segments is given in [79]. The conditions required for ϵ_k to be Gaussian and thus $\hat{\tau}$ to be the MVU estimator are stated in [80]. A limitation of the method for this application is the long-term averaging required for the coherence estimate. For instance, with a 20 kHz sampling rate and half-overlapping 25.6ms Hanning windows, the estimate is shown to be equivalent to the maximum-likelihood estimate when averaged over two seconds of data. This analysis interval vastly exceeds the independent analysis constraint and is clearly inappropriate for speech signals. A sub-optimal variation of this technique which restricts coherence-estimate analysis to a single 20-30ms time interval will be considered in Section 8.2.

Given these analysis restrictions, an appropriate alternative is to estimate the error variance independently for each data frame using the following approximation.

$$\hat{var}(\epsilon_k) = \frac{\lambda_{1k}}{\mid R_1(k) \mid^2} + \frac{\lambda_{2k}}{\mid R_2(k) \mid^2} \tag{8.5}$$

Where the λ_{1k} and λ_{2k} coefficients are derived from the frequency-dependent background noise power at each receiver as follows:

$$\lambda_{1k} = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} \mid M_{1j}(k) \mid^2$$

$$\lambda_{2k} = \lim_{J\to\infty} \frac{1}{J} \sum_{j=1}^{J} \mid M_{2j}(k) \mid^2$$

with $M_{1j}(k)$ and $M_{2j}(k)$ being the DFT coefficients of individual windowed frames of the

S/N (dB)	experimental mean(ϵ_k) $(radians)$	experimental var(ϵ_k) $(radians)^2$	estimated var(ϵ_k) $(radians)^2$
36	.0016	.09	.05
18	-.0052	.52	.36
12	-.0031	.84	.71
0	-.0235	1.77	1.96

Table 8.1: Experimental mean, variance and estimated variance of ϵ_k as a function of S/N ratio

background noise sources $n_1(l)$ and $n_2(l)$. The variance estimate may be interpreted as the sum of the approximate inverse S/N ratios at each receiver. Equation (8.5) was derived assuming relatively large S/N ratios and that the $M_{1j}(k)$ and $M_{2j}(k)$ terms have uniformly random phases. With the phase deviation variance approximated in this manner, the weighting coefficients used in (8.3) and (8.4) are calculated from:

$$\hat{W}_k = \frac{1}{\hat{var}(\epsilon_k)} = \frac{\mid R_1(k) \mid^2 \mid R_2(k) \mid^2}{\lambda_{1k} \mid R_2(k) \mid^2 + \lambda_{2k} \mid R_1(k) \mid^2} \tag{8.6}$$

Data showing the validity of the assumptions made on the ϵ_k random variables and the accuracy of (8.5) in estimating the true error variance are presented in Table 8.1. The results here have been generated with a Gaussian white random source (variance σ_s^2) delayed 1 sample relative to the receivers and then corrupted by uncorrelated additive white Gaussian noise sources (variance σ_n^2). The signals were sampled at 20kHz, segmented into 200 25.6ms (512-point) Hanning windows, zero-padded, a 1024-point DFT implemented to generate the spectral coefficients, and the ϵ_k terms were then calculated via (8.2). The choice of this window type and DFT length is based upon an analysis presented in [81]. The 2 left-hand columns in the table report the sample mean and variance of the ϵ_k terms for each of the S/N conditions. The right-hand column lists the predicted variance of ϵ_k calculated

from (8.5). As the table illustrates, the zero-mean assumption for ϵ_k is appropriate for the entire range of signal conditions. Furthermore, the error variance estimated using (8.5) accurately models the experimental variance.

8.1.2 Application Considerations

A practical issue that must be considered when applying the proposed estimator is that of phase continuity. The cross-spectrum phase θ_k as evaluated by (8.2) is modulo 2π whereas the delay estimator (8.3) requires a phase angle that varies in a continuous linear fashion with the radian frequency. This situation necessitates the use of a "phase-unwrapping" algorithm to remove the 2π discontinuities from the initial θ_k before evaluating $\hat{\tau}$. Several algorithms for this purpose are available from cepstral processing applications, [82] is typical. An alternative solution to the phase discontinuity problem is given in [81]. The "phase-unwrapping" technique used in the following experiments is along the lines of [82] but less general since the phase difference function is assumed to be linear.

Consider a frame of modulo 2π cross-spectrum phase terms, θ_k, and their associated weighting coefficients, $\hat{W}_k$. A reordering of the frequency components with respect to the weighting terms (high to low) is performed. As the linear fit in Equation (8.3) is performed, progressively summing over the the ordered set of frequency components, an intermediate slope estimate may be used to predict the value of the next phase term. The measured value of the phase angle is unwrapped around this predicted value (by adding an integer multiple of 2π to minimize their difference) before it is included in the linear-fit sums. A second pass is performed to correct values of θ_k that may been improperly unwrapped due to the variation of the slope estimate over the course of the linear-fit/phase-unwrapping process. Because the unwrapped value of the initial phase term in the series is undetermined, this

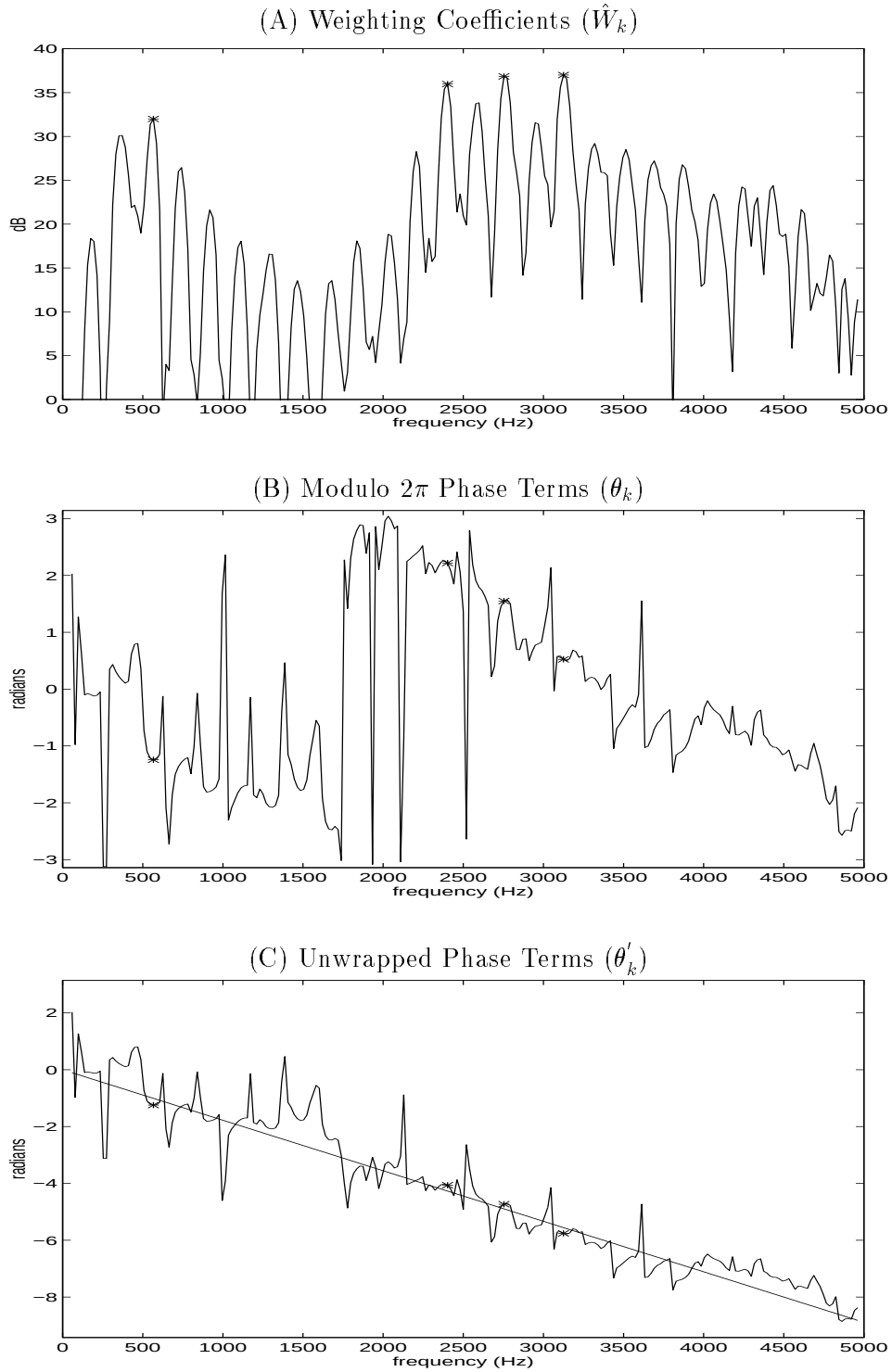


Figure 8.1: Illustration of the Linear-Fit/Phase-Unwrapping Process: Plot (A) graphs the weighting coefficients as a function of frequency. Plot (B) displays the modulo 2π phase terms. Plot (C) shows the phase values unwrapped around the line with a slope corresponding to the final TDOA estimate. Four highly weighted frequency bins have been indicated by a '*' to illustrate their contribution to the line-fitting procedure.

process must be repeated several times with different potential unwrapped versions of this starting phase. The case that provides the best linear-fit to the now unwrapped phases, $\theta_k^{'}$, provides the final TDOA estimate.

Figure 8.1 illustrates the results of this linear-fit/phase-unwrapping process. Plot (A) graphs the weighting coefficients, $\hat{W}_k$, estimated from a 25.6ms segment of voiced speech over a frequency range of 100Hz to 5kHz. Four highly weighted frequency bins have been indicated by a '*'. Plot (B) displays the measured modulo 2π phase terms, θ_k. Note that while these phase terms exhibit a linear trend, there is an apparent discontinuity between 1500Hz and 2000Hz. Plot (C) shows the unwrapped phases and the line with a slope corresponding to the final TDOA estimate. The line provides a close fit to the to the highly weighted phases as indicated by the proximity of the '*' symbols to the line. Sizable phase deviations from the estimated line generally occur only for those frequency bins with minimal weighting coefficients.

Another practical issue that must be addressed is microphone placement. In a near-field setting, such as a room, excessive separation between the microphone receivers may result in significant deviations from the the source-model assumptions which have the potential of seriously degrading the quality of the delay estimate. Deviations from the model may be due to non-uniform radiation of the source, or unequal signal attenuation and filtering due to the acoustics of the room. Also the short window length employed in the signal processing imposes a practical limit on the maximum TDOA and therefore the sensor separation. If the relative delay between two channels is an appreciable fraction of the window length one can no longer be confident that there is good correlation between the segments of the source signal captured by the two sensors. A feedback mechanism to realign the time sequences by repeating the windowing process with skewed time indexes is not computationally feasible

on a frame-by-frame basis. One means of overcoming both these problems is to limit the microphone separation distance. This minimizes near-field radiation effects and restricts the range of potential signal TDOA's. The upper bound for microphone separation distances is dictated by the physical environment and location of signal sources.

8.2 TDOA Estimator Comparison

8.2.1 Experiment # 1

Two computer simulations were performed to evaluate the accuracy of the TDOA estimator described in the previous section. In the first, a single phoneme (the /e/ in 'ketchup') of 20kHz sampled speech was bandlimited to the range 100Hz to 5kHz and isolated with a 25.6ms Hanning window. Relative delays of 1, 5, and 10 samples were introduced to simulate a second sensor and uncorrelated white Gaussian noise added to both channels. The noise variance was adjusted to create the appropriate S/N ratio. Here S/N ratio is defined as:

$$S/NdB = 10\log_{10}\left(\frac{\sum_l [w(l)s(l)]^2}{\sum_l [w(l)n(l)]^2}\right)$$

for a finite length window $w(l)$. A 1024 point FFT was computed for each signal. The delay estimate, $\hat{\tau}$, and the predicted variance of $\hat{\tau}$ were then calculated from (8.3) and (8.4), respectively, using the weighting coefficients given by (8.6). Table 8.2 lists each estimate's sample mean and standard deviation determined from 100 trials at each S/N condition. The values in parentheses represent the averages of the predicted standard deviation. For comparison purposes, the least-squares (LS) delay estimate given in [80] was also calculated. The LS estimator involves a weighted line fit of the phase data but differs from the proposed delay estimator in a key respect; the weighting coefficients and phase

sample delay	S/N (dB)	Proposed Estimator		LS Estimator	
		mean	std. dev. (predicted)	mean	std. dev.
1	36	1.00	.017 (.012)	1.00	.014
	24	1.00	.036 (.024)	1.00	.030
	12	1.01	.070 (.048)	1.00	.077
	0	1.00	.162 (.091)	.99	.233
5	36	5.00	.020 (.012)	4.97	.021
	24	5.00	.032 (.024)	4.98	.036
	12	5.01	.074 (.048)	4.99	.077
	0	5.00	.177 (.091)	4.97	.206
10	36	10.00	.017 (.012)	9.93	.028
	24	10.00	.036 (.024)	9.93	.041
	12	10.00	.067 (.048)	9.97	.087
	0	10.03	.182 (.091)	9.99	.265

Table 8.2: Results of TDOA Estimator Experiment #1 (Single phoneme): Sample Mean and Standard Deviation of the Proposed TDOA Estimator and the LS TDOA Estimator for varying S/N ratios and sample delays. All values are in terms of samples at 20kHz.

information required for (8.3) and (8.4) are derived via the multiple-window magnitude-squared coherence estimation procedure referred to in the preceding section. While the merits of this approach are apparent for long-term statistically and physically stationary signal sources, the expectation is that given the time constraint of the analysis interval, the LS estimator will be at a disadvantage. A number of scenarios were considered for the partitioning of the 25.6ms speech segment required for the coherence estimation. The most favorable, which was used to generate the LS estimator results reported in Table 8.2, incorporated 7 half-overlapping 6.4ms subwindows.

While the proposed TDOA estimator and the LS estimator perform comparably for the small-delay, high-S/N conditions, the experimental results distinctly favor the proposed estimator for the S/N=0dB case and for the larger sample delays. The LS estimator exhibits a sizeable bias and increased standard deviation at delays of 5 and 10 samples. Part of this effect may be attributed to window misalignment; these delays represent a sizeable fraction of the 6.4ms subwindows employed by the LS estimator. The proposed estimator with a

single 25.6ms window does not display a marked bias or inflated standard deviation at these larger sample delays. Finally, note that the standard-deviation predictor figures from (8.4) in parentheses accurately model the measured estimator variance for all but the S/N=0dB case.

8.2.2 Experiment # 2

For the second simulation, 100 different frames of randomly-segmented speech representing a wide range of phonemes were prepared under similar conditions to those of the previous experiment. The sample means and standard deviations for the proposed TDOA estimator and the LS estimator are given in Table 8.3. In general the variances measured in this experiment are greater than those reported for the single-phoneme experiment. This is most likely due to the varying spectral content of the phonemes used in the second experiment. The /e/ phoneme used in the first experiment is strongly voiced and has very good S/N at the formant frequencies, and thus is well suited to the frequency-dependent weighting used in the delay estimator. In the second experiment the phonemes used cover a broad variety, many of which do not have spectral characteristics quite as favorable for the delay estimation algorithm.

A comparison of the TDOA estimators' relative performance in this second simulation reveals trends similar, but more pronounced, than those demonstrated in the previous experiment. The larger LS estimator bias is evident even at the high S/N conditions and the variance is larger than that of the proposed estimator's even for the low S/N conditions. The difference in the variances is particularly large for the cases with 5 and 10 sample delays and high S/N. The proposed delay estimator outperforms its counterpart for all the simulation conditions and is relatively insensitive to the differing sample delays although it

sample delay	S/N (dB)	Proposed Estimator		LS Estimator	
		mean	std. dev.	mean	std. dev.
1	36	1.01	.059	.98	.061
	24	1.00	.144	.96	.205
	12	1.00	.405	.98	.631
	0	.90	.909	.94	1.101
5	36	4.99	.057	4.86	.114
	24	5.01	.147	4.85	.243
	12	4.98	.494	4.77	.641
	0	4.83	1.208	4.74	1.254
10	36	9.99	.061	9.75	.228
	24	9.97	.135	9.70	.462
	12	10.06	.407	9.70	.731
	0	9.90	.999	9.29	1.538

Table 8.3: Results of TDOA Estimator Experiment #2 (100 frames of speech): Sample Mean and Standard Deviation of the Proposed TDOA Estimator and the LS TDOA Estimator for varying S/N ratios and sample delays. All values are in terms of samples at 20kHz.

does begin to show an estimate bias for the S/N=0dB case.

The results of these experiments clearly show that the proposed TDOA estimator has superior performance properties in comparison to the Least-Squares TDOA estimator presented and is capable of intra-sample precision. For instance, at $S/N \geq 24$dB, the standard deviation of the estimate is less than .15 samples for all the conditions tested. With regard to computational considerations, the proposed estimator again has a marked advantage over its LS counterpart. For each delay estimate generated in these simulations, the bulk of the proposed delay estimator's computational load is contained in the 2 1024-point FFT's. The remaining elements, such as the calculation of the phases and weights, the phase-unwrapping, and final delay estimation, require computation equal to approximately one-half of a 1024-point FFT. Roughly speaking, the total number of floating point operations required for a single TDOA estimate is equivalent to that of performing 2.5 1024-point

FFT's. The LS estimator used in these experiments needs approximately 3 times this number of operations. A similar correlation-based delay estimator would require a minimum of 3 FFT's to compute the cross-correlation function alone.

8.3 Source Detection with the TDOA Estimator

The TDOA estimate $\hat{\tau}$ is readily shown to be the delay value that minimizes the weighted-least-squares error:

$$\text{Error}(\tau) = \sum_{k=1}^{N-1} \hat{W}_k (\theta'_k - \omega_k \tau)^2$$

and

$$\hat{\tau} = \arg\min_{\tau} \text{Error}(\tau)$$

This LS error may be interpreted as the 'line fit' error associated with the unwrapped phase terms, θ'_k, and the line with slope τ. The minimum error, $\text{Error}(\hat{\tau})$, provides a useful statistic for evaluating the significance of the TDOA estimate $\hat{\tau}$. A relatively small error indicates that the single source model is applicable to the windowed signal frame and that $\hat{\tau}$ is a reliable measure of the true TDOA. Large values demonstrate that the estimated TDOA is not valid, either through imprecision or because of an inconsistency between the data and the single source model. This would be expected, for example, during silence intervals. The effect may also be due to the presence of simultaneous interfering sources, in which case the derivation model is inappropriate. A further possible cause could arise from severe reverberations. In acoustically live environments, the TDOA estimate possesses an increased inaccuracy similar to the effects of diminished SNR conditions [83]. Each of these situations is manifested through an enlarged LS error.

In practice, this detection statistic is calculated from a normalized version of the weighted

LS error:

$$D_{\text{error}} = A\frac{\sum_{k=1}^{N-1} \hat{W}_k(\theta_k' - \omega_k\tau)^2}{\sum_{k=1}^{N-1} \hat{W}_k} \tag{8.7}$$

The appearance of the denominator term is necessary to adjust the error to a uniform scale across a range of signal SNR situations. The constant A, discussed below, is used to regulate D_{error} relative to a precalculated non-source error. The use of line-fit error as a detection statistic is outlined in [80]. However, the statistic presented there is an unweighted version of (8.7) and the subsequent decision rule was found to be ineffective for speech signals.

During periods of silence, the received sensor signals are assumed to be uncorrelated with known spectral-density coefficients, λ_{1k} and λ_{2k} . The cross-spectrum phase, θ_k, and the corresponding phase deviation terms, ϵ_k, are uniformly distributed between $-\pi$ and π [84]. Under these conditions the second-order statistics for ϵ_k are:

$$E(\epsilon_k \mid \text{silence}) = 0 \qquad\qquad E(\epsilon_k^2 \mid \text{silence}) = \pi^2/3$$

and the predicted expectation of D_{error} simplifies to:

$$E(D_{\text{error}} \mid \text{silence}) = A(\pi^2/3) \tag{8.8}$$

However, when the TDOA estimator encounters a silence frame, the estimation procedure produces an arbitrary value of τ that minimizes Error(τ). The actual values of D_{error} produced from this process are biased below the predicted expectation $A(\pi^2/3)$. The degree of bias is dependent upon the power spectral density of the noise and the frequency components included in the TDOA estimate summation.

In practice, given an interval ($\approx 1 - 2s$) of background silence conditions, a condition-

dependent correction term may be evaluated to produce a uniform D_{error} statistic. This is done by estimating the TDOA and D_{error} values for the silence interval. The sample mean and standard deviation for D_{error} are then computed and the constant A is selected to scale the statistic mean to 1. i.e. choose A such that

$$\bar{D}_{\text{error}} = \frac{1}{L} \sum_{l=1}^{\text{L frames}} D_{\text{error}}(l) = 1$$

As an example of this procedure, two-second (40,000 samples at 20kHz sampling) uncorrelated sequences were generated from equal-power, white, Gaussian noise processes to simulate background silence conditions at a pair of sensors. TDOA and D_{error} estimates were prepared using a half-overlapping 512-point Hanning window, a 1024-point FFT, and a frequency bandlimit of 100Hz to 5kHz. The normalization constant was calculated to be $A = \frac{1.28}{\pi^2/3}$, indicating that the actual statistic mean is roughly three-fourths of that predicted by (8.8). The final second-order statistics for the scaled detection statistic were:

$$\begin{aligned} E(D_{\text{error}} \mid \text{silence}) &= 1 \\ std(D_{\text{error}} \mid \text{silence}) &= .132 \end{aligned}$$

Once the D_{error} silence statistics have been determined, a detection threshold may be calculated for a predetermined false-alarm rate, α. Using the central-limit theorem [56] to make the simplifying assumption that D_{error} is approximately Gaussian during the silence intervals, the threshold is found from the cumulative unit normal distribution function,

$\Phi(z)$, and the decision rule is stated as:

$$\text{if} \quad \begin{cases} D_{\text{error}} < D_0 & \text{source present} \\ D_{\text{error}} \geq D_0 & \text{source absent} \end{cases}$$

$$\text{where} \quad D_0 = s \cdot \Phi^{-1}(\alpha) + 1$$

Here s is the standard deviation of the scaled silence-only detection statistic. For the above example, some false-alarm rate/detection threshold pairs are:

$$\alpha = 10^{-2} \implies D_0 = .69$$

$$\alpha = 10^{-3} \implies D_0 = .59$$

$$\alpha = 10^{-4} \implies D_0 = .52$$

$$\alpha = 10^{-5} \implies D_0 = .48$$

8.4 TDOA Estimator Demonstrations

8.4.1 Single, Moving Talker

Figure 8.2 illustrates the elements of the TDOA estimation procedure for a single, moving speech source. The talker was recorded by a pair of pressure-gradient microphones, placed 20.5cm apart, and digitally sampled at 20kHz. Plot (A) in the figure represents 1.25s of speech, the utterance "One Two Three", received at one of the microphones. The background noise statistics and detection normalization constant were estimated using a 1s sampling of silence conditions. In this case, the background noise is dominated by the whirr from a computer fan in the vicinity. While not creating the ideal silence conditions, this

does provide for a realistic testing scenario. Once again, TDOA estimates were evaluated using the 512-point, half-overlapping Hanning window, a 1024-point DFT, and a frequency band range of 100Hz to 5kHz. With these parameters, 97 independent analysis windows are applied to the 1.25s speech segment. Plot (B) graphs the detection statistic, D_{error}, as a function of the sample midpoint in each analysis window. A value of 1 in this plot corresponds to a silence frame, while a 0 indicates an ideal source. The horizontal line represents the detection threshold, $D_0 = .35$, calculated from a desired false-alarm rate of .05. Note how the frames with a D_{error} value below this threshold are aligned with the periods of source activity in Plot (A), and conversely, the D_{error} values above the line are associated with speech pauses. This demonstrates the effectiveness of the detection statistic in accurately dichotomizing source/silence periods in the speech signal. Plot (C) shows the TDOA estimates for those analysis frames in which a source was detected. The vertical axis in this graph has been scaled by the speed of sound, c, so that the values represent the difference in source propagation distance. The maximum absolute value on this scale corresponds to the sensor separation distance, .205m. The TDOA estimates exhibit a continuous downward progression throughout the segment. This indicates that the talker was moving relative to the sensor-pair, crossing from one of the half-spaces delineated by the sensors' perpendicular bisecting plane, to the other. The small analysis window incorporated into the estimation procedure allows for a high TDOA update rate, providing the time-resolution necessary to effectively track moving sources. Finally, Plot (D) illustrates the predicted standard deviation associated with each TDOA estimate. These figures have been scaled by c and are presented in meters. Each prediction figure is independently calculated using (8.4) for a single analysis frame. The standard deviation is primarily a function of the signal SNR and the predicted figures trace out a curve that is roughly inversely related to the signal energy.

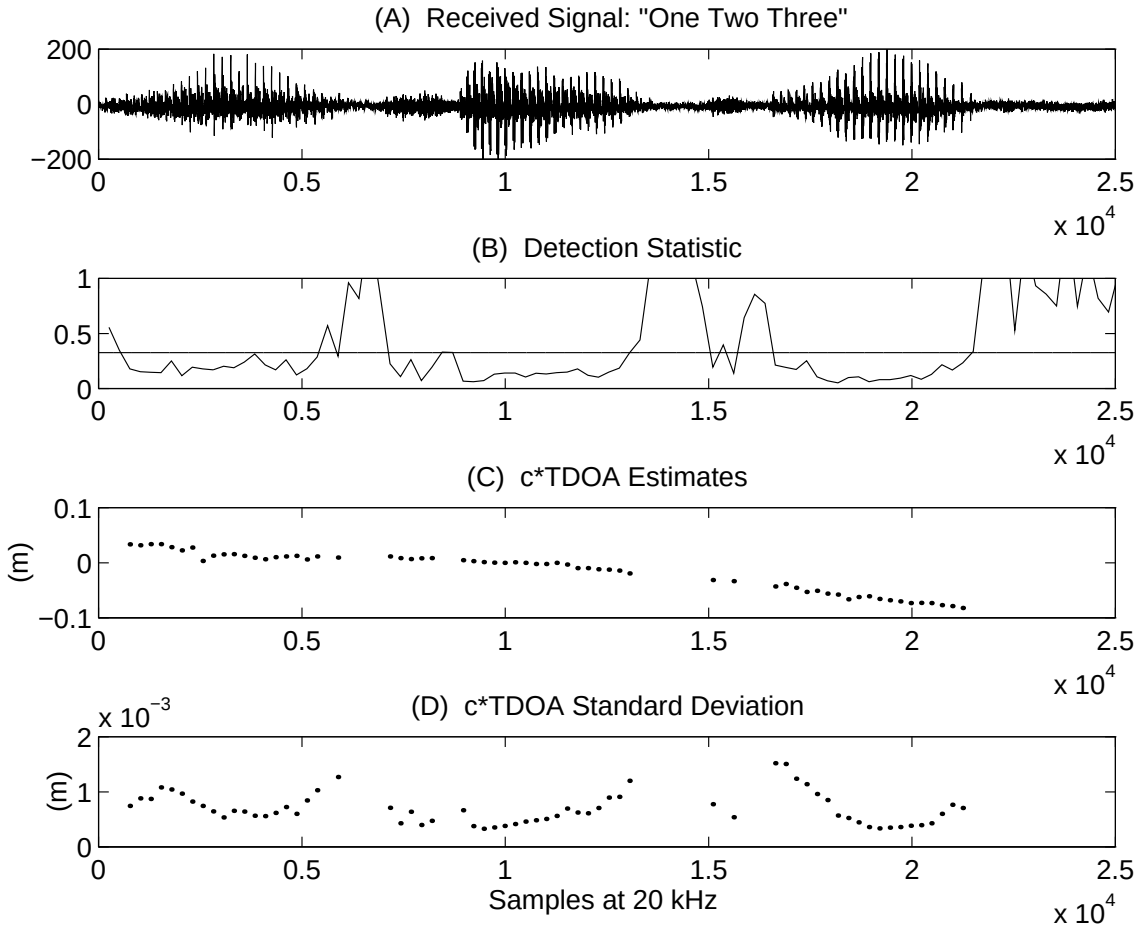


Figure 8.2: TDOA Estimation for a single moving talker: Plot (A) illustrates the received signal. Plot (B) shows the scaled detection statistic, D_{error}, relative to the $\alpha = .05$ detection threshold. Plots (C) and (D) track the TDOA estimates and their predicted standard deviations for the detected source frames. Both plots have been scaled by c to present these values in terms of meters. The x-axis for each plot represents samples at 20kHz.

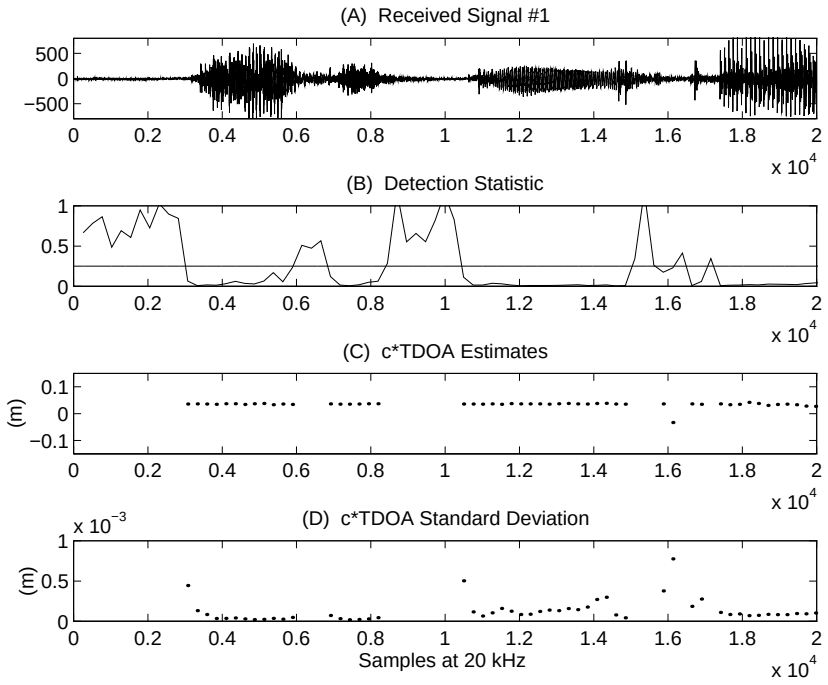
8.4.2 Multiple Talkers

In the previous demonstration, the detection statistic was incorporated into a source/silence decision rule. However, the significance of the D_{error} value may also be applied as a means for validating the consistency of the single source model that has been assumed in the derivation of the TDOA estimator. This is especially useful in situations in which several speech sources may be simultaneously active. In this context the D_{error} statistic detects when an analysis frame contains speech from a single talker versus periods of multiple

or no source activity. Typical speech patterns contain distinct intervals of active talking interspersed with silence. For single-talker speech the average silence duration is on the order of 120ms [85], equivalent to nearly 10 of the TDOA estimator's overlapped analysis windows. With conversational speech, fewer than 20% of the overall frames include more than a single active talker [86]. Under these conditions, the TDOA estimator presented here will have ample access to signal frames containing each of the active talkers speaking, essentially, in isolation, and given the effectiveness of the detection statistic, will be able to identify these valid frames.

To demonstrate the performance of the TDOA estimator in a multiple-talker scenario three recordings were done with a pair of pressure-gradient microphones, separated by 16.5cm. The first two recordings were each done with a single source speaking continuously at a distinct, fixed location. The third recording repeated the same two utterances, this time simultaneously. Each of the speech signals was pre-recorded and played out by a computer to insure synchronization from the individual to simultaneous scenarios. This represents a particularly extreme two-talker case. Since the individuals are not engaging in a conversation, but rather talking continuously and simultaneously, the signal overlap is much greater than would be expected in a typical (polite) dialogue. Indeed, while the individual recordings are quite clear, the dual recording is unintelligible. The background statistics and TDOA estimation parameters were adjusted as in the previous example.

Figures 8.3 and 8.4 display the results of the TDOA estimation for 1s segments (77 analysis frames) of each recording with the false-alarm rate set at a conservative 10^{-3} level. Each of the individual recordings in Figure 8.3 exhibits a near constant TDOA throughout the utterance. This indicates that the sources were positionally fixed and, because of the difference in $\pm$ signs, were located on opposite sides of the sensor-pair perpendicular bisect-



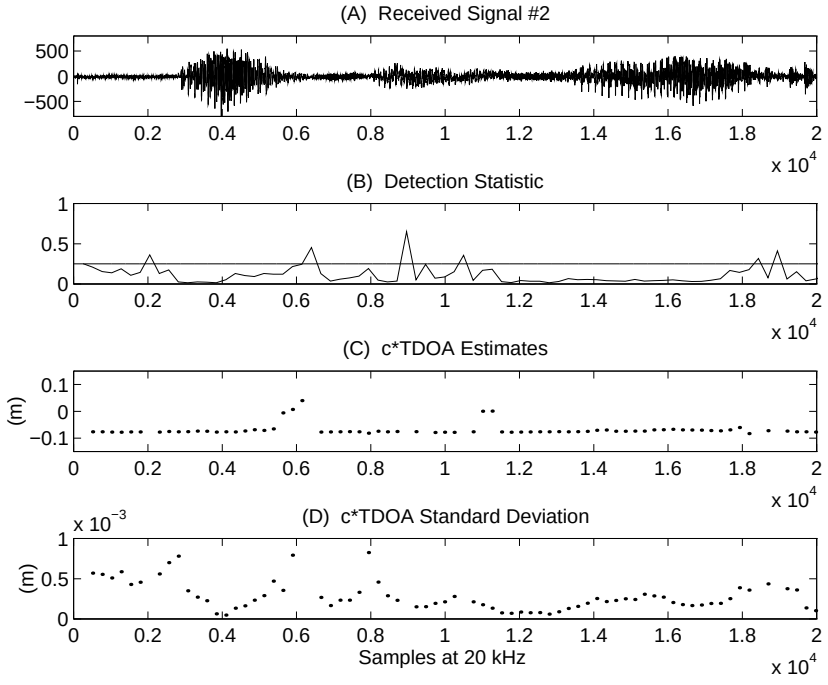


Figure 8.3: TDOA Estimation for Two Isolated Single Talkers: Plot (A) in each case illustrates the time sequence of the received signals. The (B) plots show the scaled detection statistics, D_{error}, relative to the $\alpha = 10^{-3}$ detection threshold. The (C) and (D) plots track the TDOA estimates and their predicted standard deviations for the detected source frames in each recording. Again, the x-axis for each plot represents samples at 20kHz.

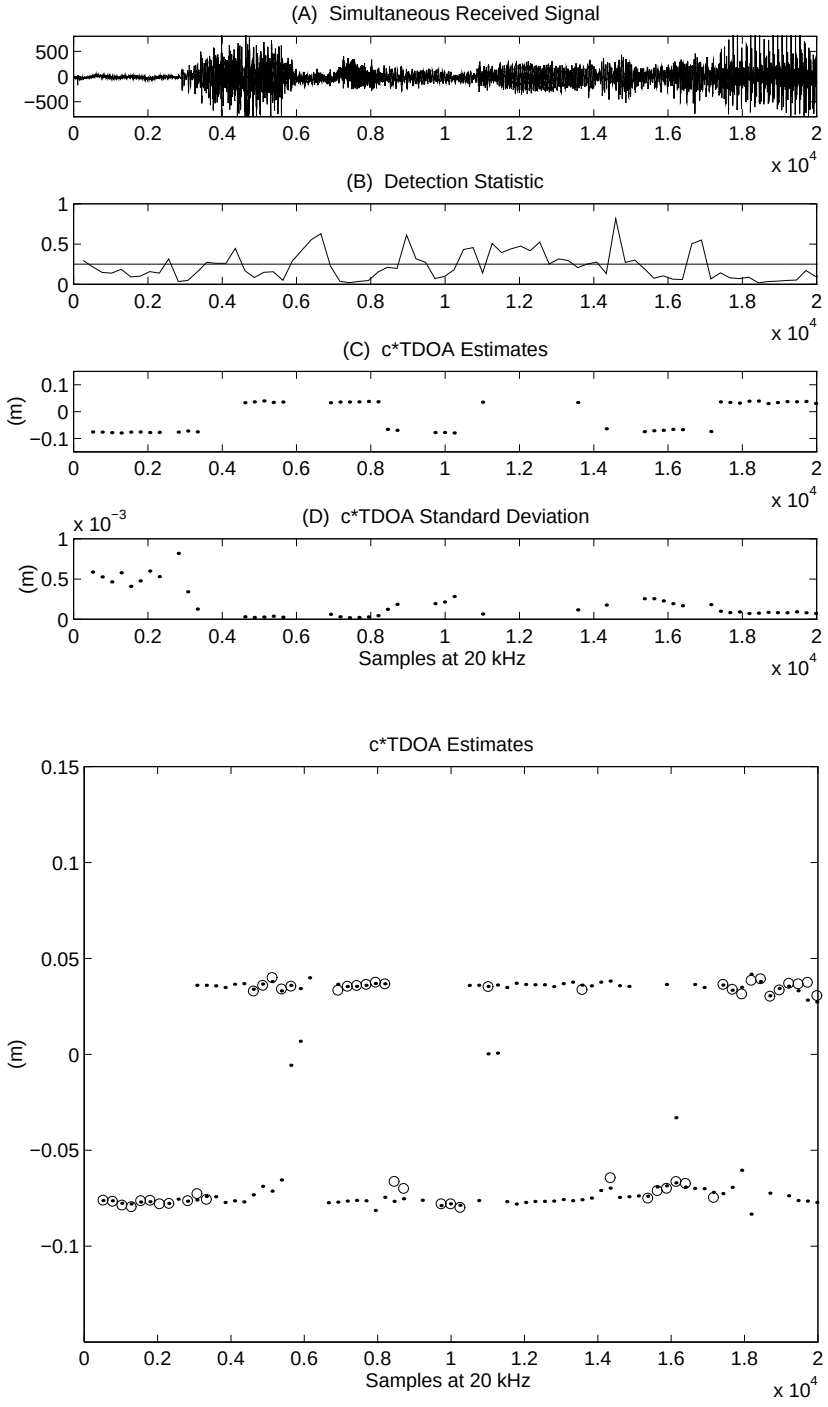


Figure 8.4: TDOA Estimation for Two Simultaneous Talkers: Plot (A) illustrates the received signal. Plot (B) shows the scaled detection statistic, D_{error}, relative to the $\alpha = 10^{-3}$ detection threshold. Plots (C) and (D) track the TDOA estimates and their predicted standard deviations for the detected source frames. The bottom plot presents an enlarged view of the TDOA estimates for all three recordings. The TDOA values for the individual recordings are denoted by '.' while those for the dual recording are marked by 'o'.

ing plane. A few errant TDOA values are evident in each Plot (C) in the figure. These are visible as deviations from the horizontal line that dominates each plot and are a result of borderline detection misclassification during speech/silence transitions in the respective signals. The recorded signal in Figure 8.4 appears to be the summation of the two signals in Figure 8.3. The corresponding TDOA plot possesses a distinct bi-linear nature, consisting of two horizontal rows at the levels found for the single source recordings. The detection statistic in this case has effectively identified those frames in which a single source contributes predominant energy and the subsequent TDOA estimate is a valid representation of that source's true TDOA. This is further verified by the bottom plot in Figure 8.4 which shows an enlarged version of the TDOA estimates graphs for all three recordings. The single source results are denoted by '.' while the simultaneous recording TDOA values are indicated by 'o'. Note how the 'o' values overlap '.' values during single source periods and are absent during silence and when both sources are of comparable strength.

8.5 Discussion

The results of these experiments illustrate the ability of the TDOA estimator to provide reliable source delay figures over a wide range of scenarios. The estimator is robust to signal/noise conditions, capable of a high update rate necessary for tracking, and is able to distinguish individual sources in a multi-party environment. Furthermore, the computational requirements of the algorithm are non-demanding, allowing for real-time hardware applications. Because of these features, the estimator presented in this chapter is an appropriate, if not ideal, means for generating the sensor-pair TDOA information required by the source localization procedures detailed in this work.

Chapter 9

Experiments with Real Systems

The source localization procedures detailed in the preceding chapters are evaluated through a series of experiments conducted with two real microphone array systems: a 10-element bilinear array mounted in a laboratory environment and a 14-element array consisting of three autonomous units placed in a conference-room setting. In each case, several recordings were obtained and processed offline.

9.1 Experimental Design

Figure 9.1 presents a flow chart of the procedures conducted in each of the localization experiments presented in this chapter. At the top level, the time signals (sampled at 20khz and digitized with 12-bit A/D converters) from the sensors were fed to TDOA estimation blocks, one per sensor pair. The TDOA estimator presented in Chapter 8 was employed for all of the experiments. The silence information and detection thresholds required by the TDOA estimators were derived by processing a two-second segment of background noise. The source-detection thresholds were calculated using a false-alarm rate of 10^{-5}. Within each TDOA estimation block, the signals were bandlimited to the range 100Hz to 5kHz and

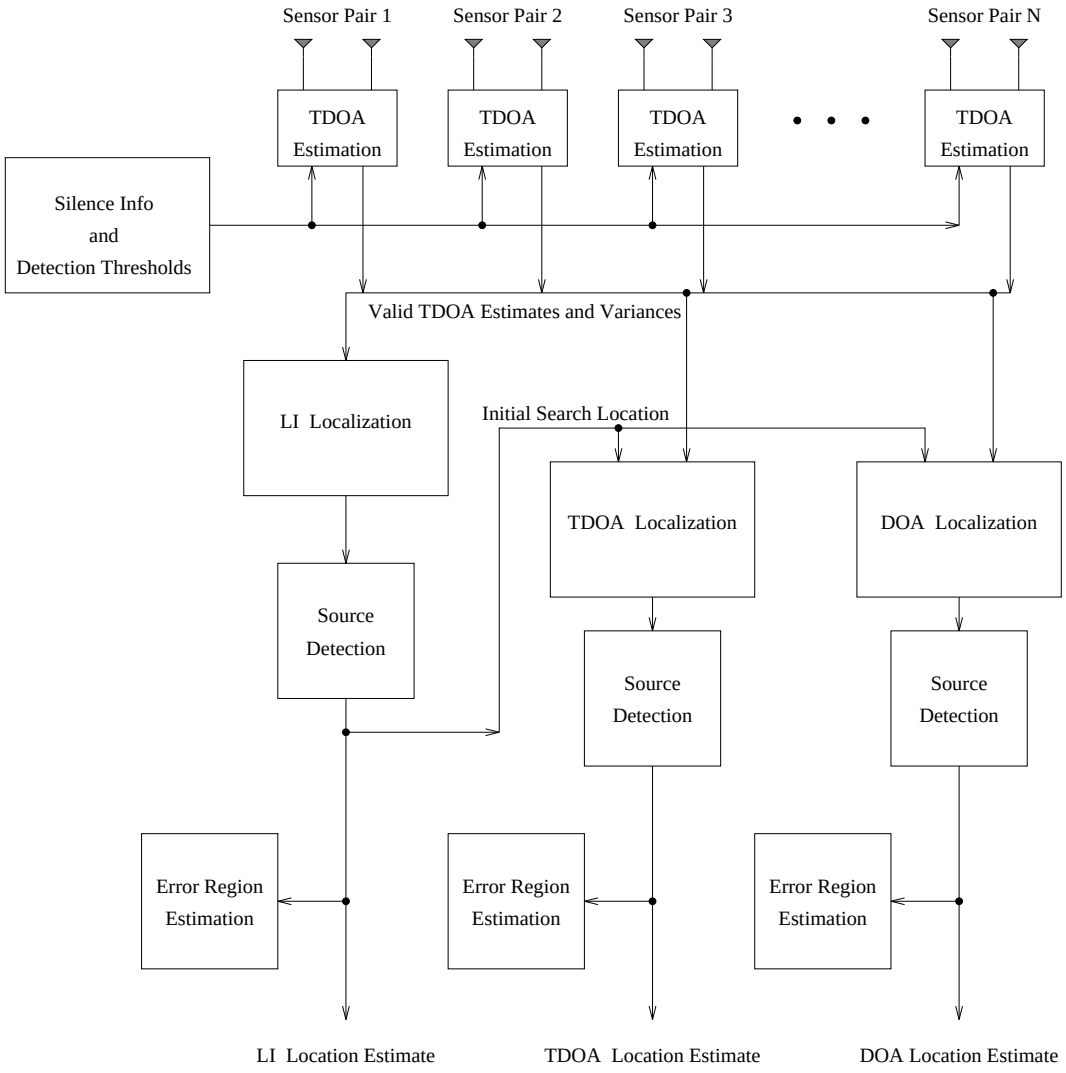


Figure 9.1: Flowchart of Localization Experiments: An outline of the localization procedures employed for each of the experiments presented in this chapter.

segmented into individual frames using a 512-point (25.6msec) half-overlapping Hanning window. A 1,024-point FFT was applied, and the TDOA estimate, a variance figure, and the detection error were calculated. Those frames possessing a detection error less than the specified threshold were declared valid, and their respective TDOA estimates and variances were made available to the localization algorithms. No TDOA information is reported for non-valid frames.

The closed-form linear intersection (LI) location estimate of Chapter 7 is evaluated first.

The LI method requires that, for each of the quadruple sensor units, both sensor-pair TDOA estimates must be available to generate a bearing line. While the LI method can generate an estimate given just two bearing lines, a minimum of three valid TDOA pairs, corresponding to three bearing lines, were required for the LI processing. This restriction was imposed to insure some redundancy in the localization algorithm and provide more reliable results. The LI algorithm is guaranteed to produce at least six points of closest intersection which were then weight-averaged to produce the final location estimate. For those frames not satisfying the three-pair valid TDOA limit, no LI estimate was evaluated.

Each reported LI estimate was then subjected to one of the source-detection tests detailed in Chapter 4. Given that the TDOA estimators evaluate a source/non-source decision based upon the pairs of time signals, this second detection test may be interpreted as a means of verifying the location estimate's significance. Either the general source consistency test or the empirical detection test is appropriate under these circumstances. With the experiments that follow, each test will be examined. The consistency test is designed with a detection rate of .99 while the empirical test employs a 1° cutoff.

The TDOA and DOA localization schemes of Chapter 3 were processed next. For those frames in which an LI estimate has been calculated and certified as valid, the LI location was used as the initial value for the search routines required in minimizing the J_{TDOA} and J_{DOA} LS-error criteria. When a valid LI estimate was not available, this initial value was established as the search region's center. The TDOA- and DOA-based localizers do not possess a TDOA valid-pair restriction as does the LI method, however they do require a minimum of three valid TDOA estimates from a set of non-collinear sensor pairs to identify a unique location in 3-space. For added reliability this limit is set at four. Any frames not possessing this minimal number of valid TDOA estimates are left unreported. The

TDOA and DOA locations that were estimated were then subjected to a source detection test similar to that performed on the LI estimate.

The LI, TDOA, and DOA location estimates declared valid by the source detection test are reported as the final location estimates. As a final analysis, the error region associated with each estimate is analyzed via the formulae derived in Chapter 5. While the derivation is expressed in terms of the J_{TDOA} and J_{DOA} error criteria, it is applied to the LI estimate as well.

9.2 A 10-Element Bilinear Array System

The 10-element bilinear array represents a scaled realization of the bilinear array first introduced in the simulations of Section 3.5. The array itself consists of pressure-gradient microphones mounted in a wire mesh at .25m intervals along two parallel rows. This structure is horizontally centered at a height of 1.58m along one wall of a $3.0m \times 3.5m$ enclosure as illustrated in Figure 9.2.

Approximately 70% of the surface area of the enclosure walls is covered with 7.5cm acoustic foam, the 3m ceiling is untreated plaster with large semi-cylindrical cavities, and the floor is light carpet over concrete. The reverberation time within the enclosure is approximately 250ms. The enclosure is a partially walled-off area contained within an acoustically-untreated workstation lab. The primary source of background noise in the recording area is computer equipment located both within the experimental enclosure and in the room surrounding the enclosure.

Sensor pairs consist of the eight diagonally adjacent microphones (i.e. sensors 1 and 4, 2 and 3, 3 and 6, 4 and 5, etc.) for a total of 8. The 4 sensor quadruples 1-2-3-4, 3-4-5-6, 5-6-7-8, and 7-8-9-10 obey the bisection-orthogonality constraint required by the

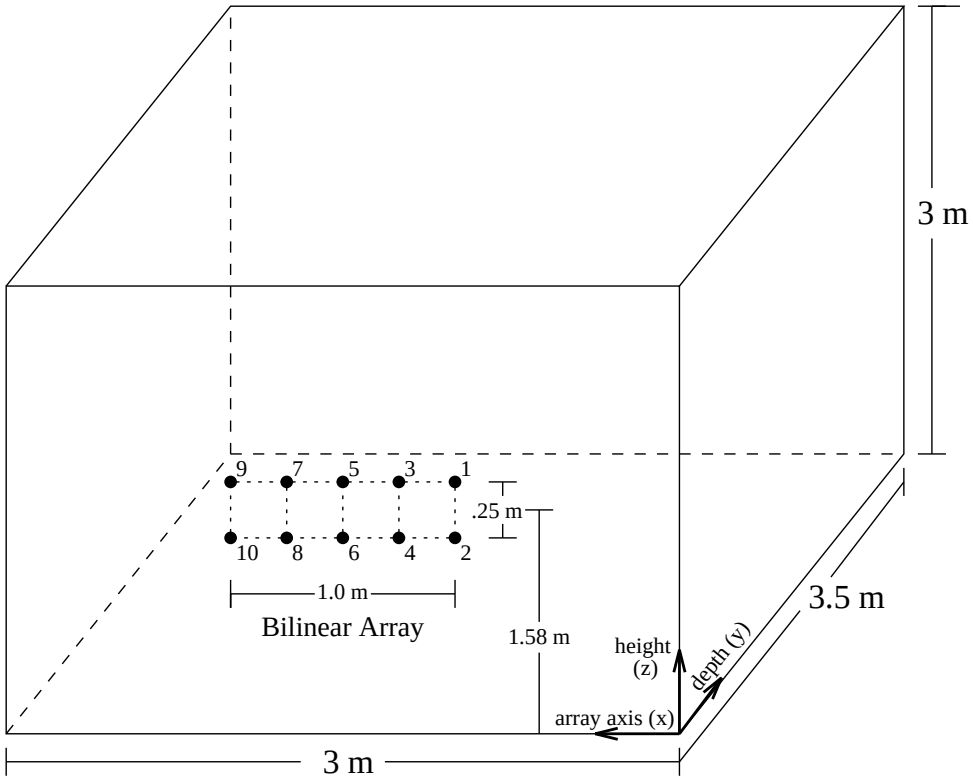


Figure 9.2: 10-Element Bilinear Array System: Illustration of the array set-up within the experimental enclosure. The array is horizontally centered on the near wall at a height of 1.58m.

LI algorithm.

9.2.1 Experiment #1: A Source Grid

This first experiment evaluated the performance of the localization schemes over a regular grid of positions within the enclosure. The experimental locations were spaced at .5m intervals along the axis of the array, x, and 1.0m intervals in the direction normal to the array, y. The symmetry of the array-enclosure setup allowed for two distinct heights. Locations on the left-side of the grid were placed at the height of the array midline, 1.58m, while the right-side was at 1.08m. These heights correspond to those of standing and sitting talkers, respectively. There were a total of 18 test locations.

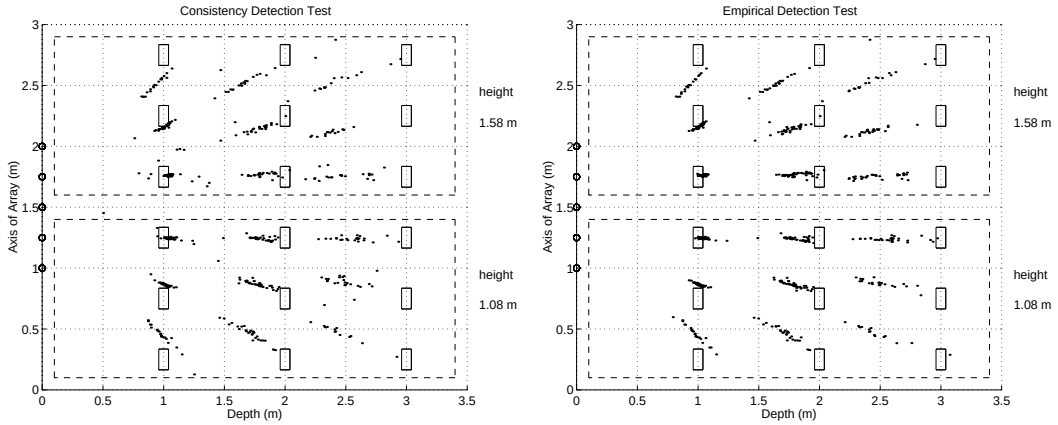
A loudspeaker was used to play back a recording of the two-second spoken phrase "h-e-

i-n-z". The transducer had a 5cm diameter cone and was contained in an acoustically and sealed enclosure 17cm on a side and 8cm deep. The front baffle of the speaker enclosure was covered with sound absorbing foam. At each location the speaker was oriented toward the center of the array and the recorded phrase was simultaneously played back and recorded by the 10 microphones. The synchronization was achieved via computer control. The peak recorded signal-to-noise ratio ranged between 5 and 30 dB, varying as a function of speaker location and orientation relative to the microphone in question.

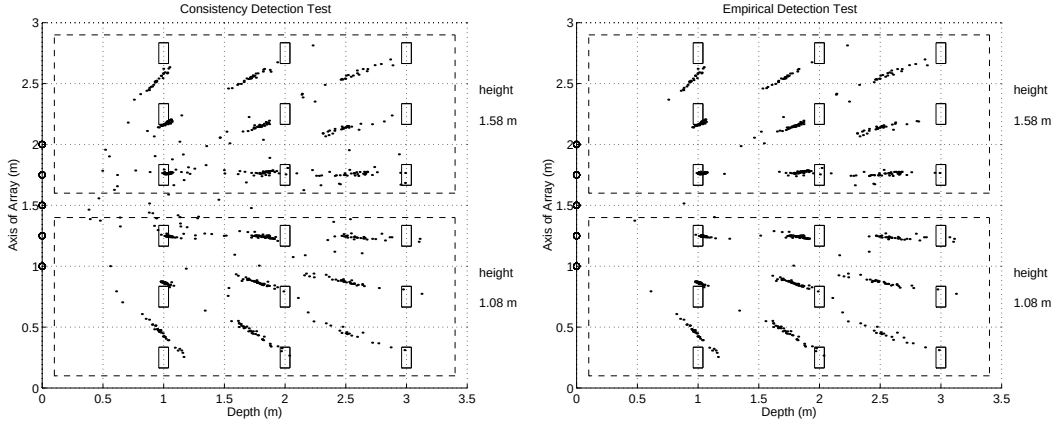
Figure 9.3 contains overhead plots of the location estimates generated by the three localization procedures and validated by each of the two source detection tests. The left-hand column graphs employed the statistical source consistency detection test, while those in the right column incorporated the empirical detection test. The location estimators are organized by row. The LI, TDOA, and DOA estimates are found in the top, middle, and bottom rows, respectively. Individual location estimates are denoted by '.', the microphones by 'o', and speaker positions as rectangular boxes. The plots represent a projection of each of these elements onto the xy-plane of the enclosure. The two distinct heights are indicated by the dashed rectangular regions.

These results are quantified in Tables 9.1 and 9.2. For each of the 18 speaker positions and three location estimators, the number of valid frames detected by each procedure is listed along with the mean location and total standard deviation of the estimated clusters. Table 9.1 presents the totals associated with the source consistency detection test and Table 9.2 contains those for its empirical counterpart. All locations are presented as ordered triplets corresponding to the coordinate system specified in Figure 9.2. The total standard deviation is calculated from the square root of the trace of the estimated covariance matrices.

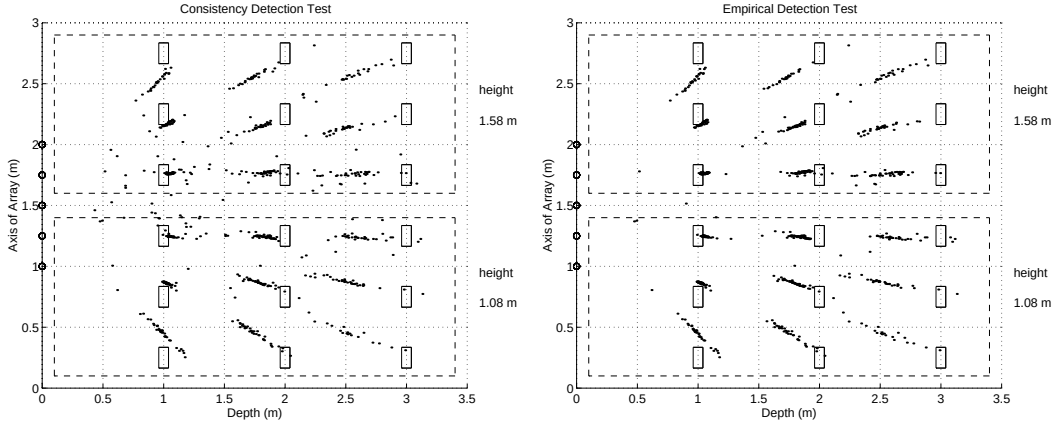
For the windowing parameters applied in these experiments, there are 77.5 independent



(a) LI Localization



(b) TDOA Localization



(c) DOA Localization

Figure 9.3: Bilinear Array Experiment #1: Clusters of valid location estimates for the three location estimators, LI, TDOA, and DOA. The plots in the left column were generated using the statistical source consistency test while those in the right column employed the empirical detection test. The actual speaker location is indicated by a box while the individual estimates are denoted by '.'. Each graph represents an overhead view of enclosure with the points projected onto the xy-plane. The heights are indicated by the dashed rectangular regions.

Results for Consistency Detection Test									
Speaker	LI Localization			TDOA Localization			DOA Localization		
Location	valid	mean	std	valid	mean	std	valid	mean	std
(x,y,z) (m)	(#)	(x,y,z) (m)	(cm)	(#)	(x,y,z) (m)	(cm)	(#)	(x,y,z) (m)	(cm)
(0.25,1.00,1.08)	22	(0.44,0.99,1.14)	14.2	37	(0.42,0.97,1.11)	39.9	36	(0.42,0.99,1.10)	40.0
(0.25,2.00,1.08)	28	(0.46,1.71,1.20)	13.8	39	(0.48,1.81,1.22)	41.8	39	(0.48,1.82,1.22)	42.2
(0.25,3.00,1.08)	16	(0.59,2.31,1.30)	57.7	28	(0.67,2.22,1.33)	73.9	27	(0.66,2.27,1.32)	73.0
(0.75,1.00,1.08)	32	(0.87,1.02,1.14)	10.2	43	(0.86,1.01,1.14)	11.1	42	(0.86,1.02,1.14)	9.0
(0.75,2.00,1.08)	39	(0.86,1.80,1.18)	18.8	53	(0.89,1.79,1.22)	25.6	52	(0.89,1.81,1.22)	27.6
(0.75,3.00,1.08)	21	(0.89,2.47,1.23)	12.0	38	(1.02,2.33,1.21)	61.6	37	(1.00,2.37,1.21)	56.0
(1.25,1.00,1.08)	46	(1.24,1.04,1.13)	4.8	63	(1.26,1.05,1.14)	10.3	55	(1.26,1.07,1.14)	11.0
(1.25,2.00,1.08)	34	(1.24,1.82,1.18)	15.1	52	(1.25,1.81,1.20)	15.2	52	(1.25,1.82,1.20)	15.1
(1.25,3.00,1.08)	24	(1.24,2.49,1.25)	37.3	37	(1.26,2.43,1.26)	59.2	36	(1.26,2.46,1.27)	58.3
(1.75,1.00,1.58)	48	(1.76,1.05,1.58)	9.7	57	(1.77,1.06,1.57)	15.7	56	(1.77,1.08,1.57)	16.0
(1.75,2.00,1.58)	35	(1.76,1.80,1.62)	34.4	51	(1.78,1.83,1.70)	49.7	51	(1.77,1.85,1.70)	48.3
(1.75,3.00,1.58)	19	(1.76,2.52,1.63)	17.6	36	(1.71,2.42,1.58)	55.0	36	(1.71,2.43,1.59)	53.3
(2.25,1.00,1.58)	33	(2.16,1.00,1.57)	7.5	45	(2.17,1.01,1.57)	9.4	40	(2.17,1.02,1.57)	9.5
(2.25,2.00,1.58)	24	(2.15,1.78,1.60)	12.0	36	(2.07,1.77,1.59)	40.9	36	(2.07,1.78,1.59)	40.4
(2.25,3.00,1.58)	19	(2.04,1.95,1.61)	112.8	35	(2.04,2.27,1.63)	59.8	35	(2.04,2.28,1.63)	59.1
(2.75,1.00,1.58)	20	(2.51,0.96,1.55)	16.7	30	(2.64,1.05,1.55)	68.9	29	(2.64,1.07,1.55)	70.3
(2.75,2.00,1.58)	22	(2.54,1.71,1.59)	25.4	34	(2.43,1.64,1.63)	61.1	34	(2.43,1.66,1.63)	60.4
(2.75,3.00,1.58)	16	(2.51,2.39,1.62)	33.9	28	(2.45,2.30,1.55)	75.1	27	(2.49,2.36,1.62)	53.4

Table 9.1: Bilinear Array Experiment #1: Numerical comparison of location estimates detected with the source consistency detection test. For each of the three location estimators (LI, TDOA, and DOA) plotted in Figure 9.3, the number of valid frames is given for each speaker location along with the mean location and total standard deviation of the estimated cluster.

analysis frames per second of recorded signal. The two-second recordings used here therefore contain 155 analysis frames. The number declared valid is a function of the speech signal, the recording conditions, the localization scheme, and the detection criterion. The utterances and their playback volume are identical in each case. However, the SNR of the received signals diminishes as the source-sensor distance is increased. Subsequently, valid source detection in the TDOA and location estimators becomes less frequent as the source range is enlarged. This trend is evident in Figure 9.3 and verified by Tables 9.1 and 9.2. In general, remote source positions generate a smaller number of valid estimates than positions in close proximity to the sensors. A further cause for variations in valid frame numbers is

Results for Empirical Detection Test									
Speaker	LI Localization			TDOA Localization			DOA Localization		
Location	valid	mean	std	valid	mean	std	valid	mean	std
(x,y,z) (m)	(#)	(x,y,z) (m)	(cm)	(#)	(x,y,z) (m)	(cm)	(#)	(x,y,z) (m)	(cm)
(0.25,1.00,1.08)	22	(0.46,0.97,1.15)	11.9	35	(0.45,0.98,1.15)	15.2	36	(0.45,0.99,1.15)	15.1
(0.25,2.00,1.08)	29	(0.46,1.72,1.19)	13.8	38	(0.47,1.78,1.19)	27.1	38	(0.47,1.79,1.19)	27.1
(0.25,3.00,1.08)	15	(0.46,2.48,1.22)	22.3	22	(0.49,2.56,1.26)	32.6	22	(0.49,2.56,1.26)	32.6
(0.75,1.00,1.08)	39	(0.87,1.01,1.14)	4.3	49	(0.86,1.02,1.13)	6.8	50	(0.86,1.03,1.13)	3.1
(0.75,2.00,1.08)	43	(0.87,1.77,1.19)	9.2	53	(0.86,1.79,1.18)	9.3	53	(0.86,1.80,1.18)	9.3
(0.75,3.00,1.08)	22	(0.88,2.50,1.22)	16.0	32	(0.87,2.50,1.22)	17.7	32	(0.87,2.51,1.22)	17.6
(1.25,1.00,1.08)	62	(1.24,1.04,1.13)	4.2	67	(1.25,1.04,1.14)	7.5	67	(1.25,1.06,1.14)	7.4
(1.25,2.00,1.08)	45	(1.24,1.81,1.19)	9.4	56	(1.24,1.82,1.19)	7.8	56	(1.24,1.83,1.19)	7.7
(1.25,3.00,1.08)	31	(1.24,2.58,1.23)	17.4	36	(1.24,2.63,1.22)	23.5	36	(1.24,2.63,1.22)	23.4
(1.75,1.00,1.58)	60	(1.76,1.04,1.59)	2.5	64	(1.76,1.03,1.59)	2.4	64	(1.76,1.05,1.59)	2.4
(1.75,2.00,1.58)	46	(1.77,1.82,1.61)	11.2	56	(1.77,1.85,1.61)	7.2	56	(1.77,1.86,1.61)	7.2
(1.75,3.00,1.58)	31	(1.75,2.49,1.63)	16.6	41	(1.76,2.56,1.63)	15.4	42	(1.75,2.52,1.61)	37.7
(2.25,1.00,1.58)	38	(2.17,1.01,1.57)	4.0	46	(2.18,1.02,1.57)	3.5	46	(2.17,1.03,1.57)	3.4
(2.25,2.00,1.58)	34	(2.14,1.76,1.60)	10.2	41	(2.14,1.78,1.59)	16.0	41	(2.14,1.79,1.59)	16.0
(2.25,3.00,1.58)	25	(2.12,2.42,1.62)	12.7	38	(2.07,2.35,1.65)	51.2	38	(2.07,2.36,1.65)	50.9
(2.75,1.00,1.58)	19	(2.50,0.92,1.56)	9.8	27	(2.53,0.95,1.56)	9.1	28	(2.52,0.96,1.56)	9.6
(2.75,2.00,1.58)	23	(2.53,1.67,1.59)	21.3	27	(2.57,1.79,1.60)	30.4	28	(2.54,1.75,1.61)	41.5
(2.75,3.00,1.58)	20	(2.52,2.39,1.62)	30.8	26	(2.56,2.52,1.62)	26.1	26	(2.56,2.52,1.62)	26.0

Table 9.2: Bilinear Array Experiment #1: Numerical comparison of location estimates detected with the empirical detection test. For each of the three location estimators (LI, TDOA, and DOA) plotted in Figure 9.3, the number of valid frames is given for each speaker location along with the mean location and total standard deviation of the estimated cluster.

the the localization procedure employed. The three valid TDOA-estimate-pair minimum imposed as a prerequisite for performing LI localization eliminates a number of analysis frames from contention as potential LI estimates. The TDOA restrictions for the TDOA- and DOA-based search procedures are less stringent and thus many frames not considered by the LI procedure produce acceptable TDOA and DOA location estimates. Additionally, with some frames the LI estimation is performed and the result deemed non-valid by the detection test, but the search-based locators succeed in finding a valid location. Finally, the particular detection test utilized has a primary effect on the number of valid frames detected. As Figure 9.3 illustrates, there is significant deviation in the quantity and quality

of location estimates accepted by the two detection schemes. The statistically oriented source consistency test is inclined to accept fewer location estimates than its empirical-based counterpart. Exceptions to this rule occur for the off-broadside conditions (x=.25m and x=2.75m) where the bearing angle deviation employed by the empirical detection test tends to be increasingly severe and the detection test more discriminating as the source approaches an end-fire position relative to the linear array.

Each of the localization schemes exhibits some degree of range bias in its estimates. This tendency to underestimate a source's distance from the bilinear array was apparent in the simulations conducted in Sections 3.5 and 7.3 and found to progress as the precision of the TDOA estimates decreased and the source's range was expanded its bearing moved further from broadside. Each of these effects is evident in the results of this experiment. The broadside, close-range sources display very little range bias while off-broadside, remote positions possess a significant shift towards the array center, as much as .5m in the worst cases. In addition to the range bias inherent in a source's actual location, this detrimental result is exacerbated by the lower SNR and less accurate TDOA estimates that accompany the more distant sources. It is feasible that this systematic disparity in range measurements could be calculated as a function of the estimated location and the TDOA variances. It would then be possible to correct the location estimate in those instances where the range figure was critical. However, it should also be remarked that this bias is a consequence of the array-enclosure geometry. A more general placement of sensor-pairs within the enclosure (e.g. microphone groups on several different walls) would improve this situation and increase the overall accuracy of the location estimates within the enclosure as a whole. The dramatic dependence of localization error upon sensor placement has been presented in evaluations #1 and #2 of Section 5.4.

Each of the detection tests displays a distinct behavior with regard to the accepted location estimates. While, on average, validating fewer locations, the source consistency test produces estimate clusters which are significantly larger than those created by the empirical detection test. The empirical test clearly yields superior detection performance for each of the localization procedures evaluated. The use of an error-spread detection criterion rather than a statistical test appears to offer a clear advantage in this practical scenario. The lackluster product of the source-consistency test may be attributed, in part, to the inaccuracy of the statistical assumptions made regarding the nature of the TDOA estimates.

A relative performance comparison of the localization schemes is less straightforward. Referring to the empirical detection test data of Table 9.2, of the three procedures, the closed-form LI locator has the smallest cluster total standard deviation for 11 of the 18 speaker locations. However, this is at the expense of markedly fewer valid frames. In some cases, the search-based methods may detect up to 50% more frames. Comparing the TDOA and DOA estimators alone, with their nearly identical valid frame numbers, the DOA scheme achieves a smaller cluster size in nine instances, the TDOA method is better in three, and the remaining six locations result in a tie. The DOA procedure appears to possess a very slight advantage in cluster mean location as well. Overall, the closed-form LI localization procedure demonstrates performance characteristics just mildly less desirable than those of the more costly, search-based methods. For those situations, where the additional computational expense is unwarranted, use of the LI method will not incur significantly inferior results. When one of the search-based methods is required, the DOA-based procedure appears to narrowly surpass the TDOA-based alternative.

Finally, the predicted error region associated with the location estimates is analyzed.

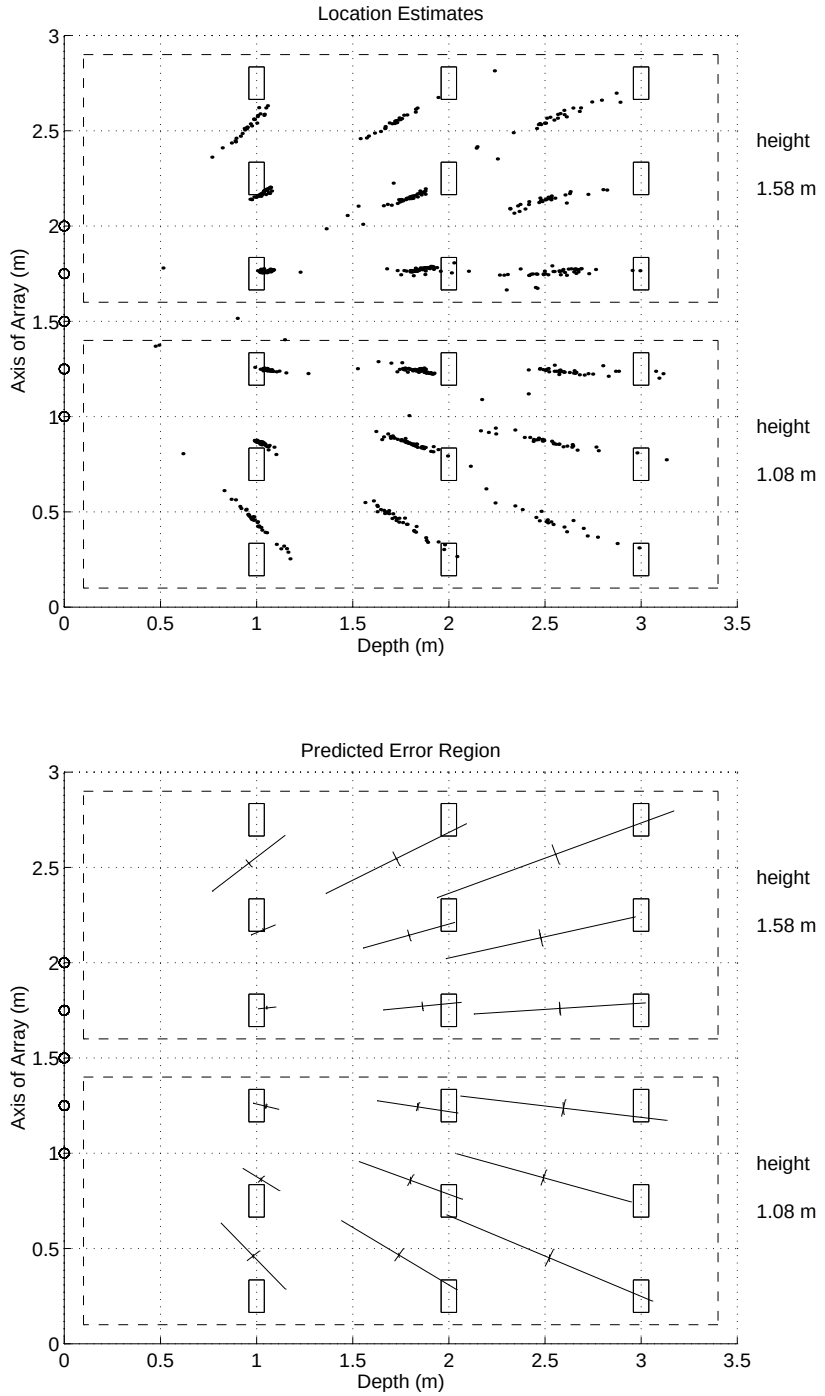
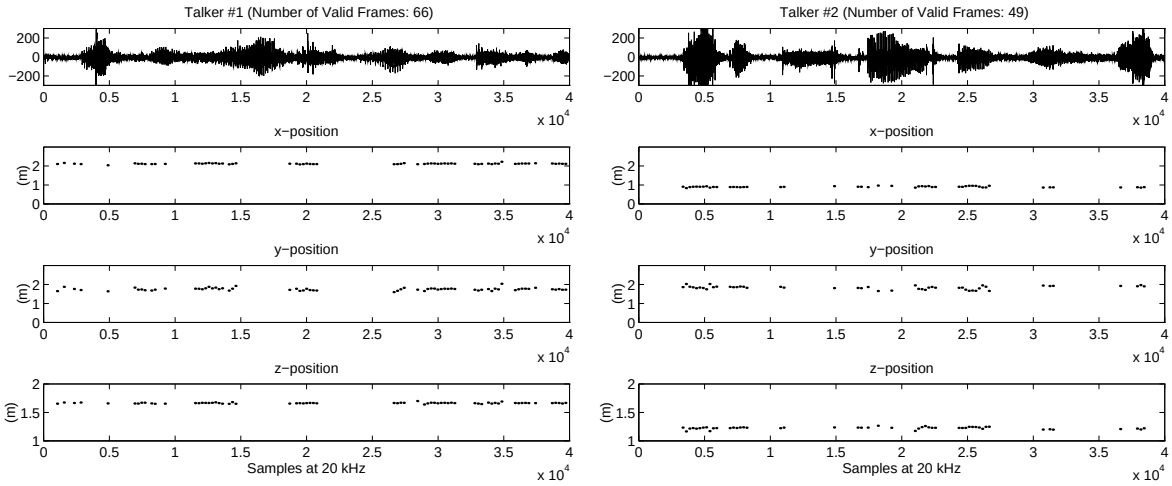


Figure 9.4: Bilinear Array Experiment #1: Experimental cluster and predicted error region. The top graph is an overhead view of the location estimates produced by the DOA localization procedure and validated using the empirical detection test. The bottom graph shows the principal component vectors of the predicted covariance matrices calculated via the median TDOA variance figures and scaled to 2.5 standard deviations.

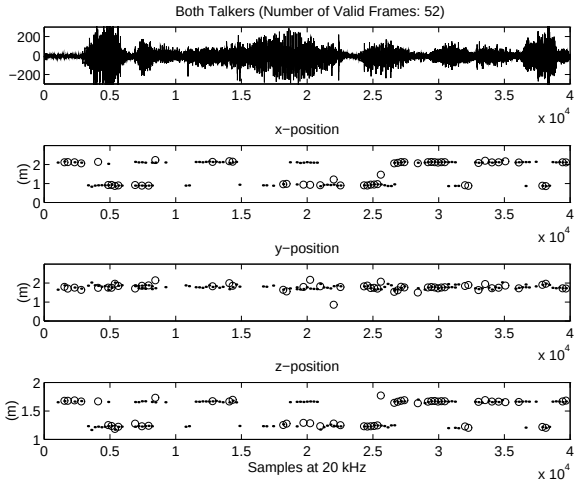
The predicted-error region is a function of the source estimate, the sensor-pair positions, and the TDOA variances. Because the TDOA variances fluctuate from frame to frame, it is difficult to compute a predicted error region indicative of the source location throughout the entire utterance. The approach taken here will be to use the median of the TDOA variances for the valid frames at a particular speaker position. Figure 9.4 shows overhead views of the experimental cluster and the predicted error regions calculated via the median TDOA variance figures. The top graph in the figure plots the location estimates produced by the DOA localization procedure and validated using the empirical detection test. The bottom graph shows the principal component vectors of the predicted covariance matrices scaled to 2.5 standard deviations. As the plots illustrate, the predicted error regions closely model the experimental clusters in both orientation and extent. The largest deviations between experiment and theory occur at the remote locations. Part of this disparity is due to the empirical detection test which presumably eliminates the more extreme points in the top graph because they fail to satisfy the bearing error threshold.

9.2.2 Experiment #2: Multi-Talkers

In Section 8.4.2 the proposed TDOA estimator was applied to the case of two simultaneous and continuous-speech sources. The detection statistic associated with the TDOA estimator was shown to effectively identify those frames in which a single source contributes predominant energy and the subsequent TDOA estimate is a valid representation of that source's true TDOA. Here the experiment presented in that section is taken a step further and the locations of the individual talkers are evaluated. Again, three recordings were taken. The first two recordings were each done with distinct fixed sources while the third repeated the same two utterances simultaneously. The content of these utterances is identical in pat-



(a) Individual Talkers

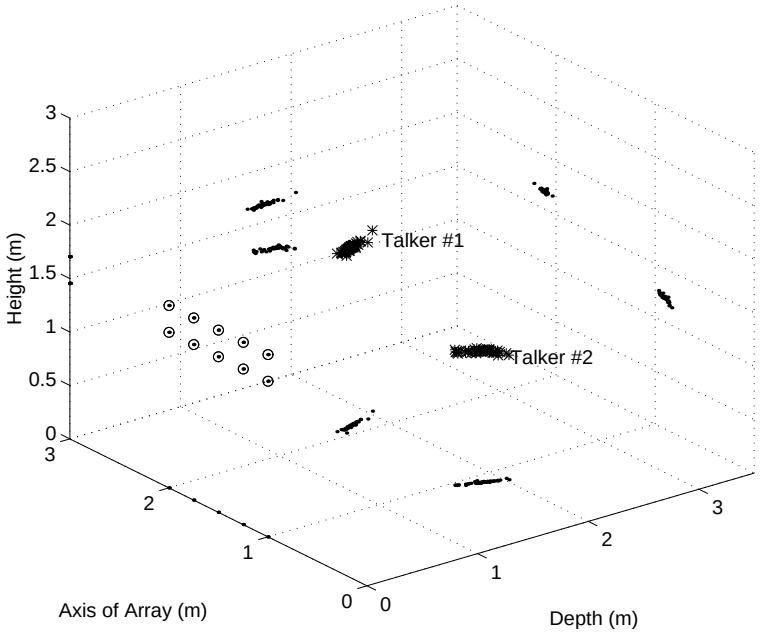


(b) Simultaneous Talkers

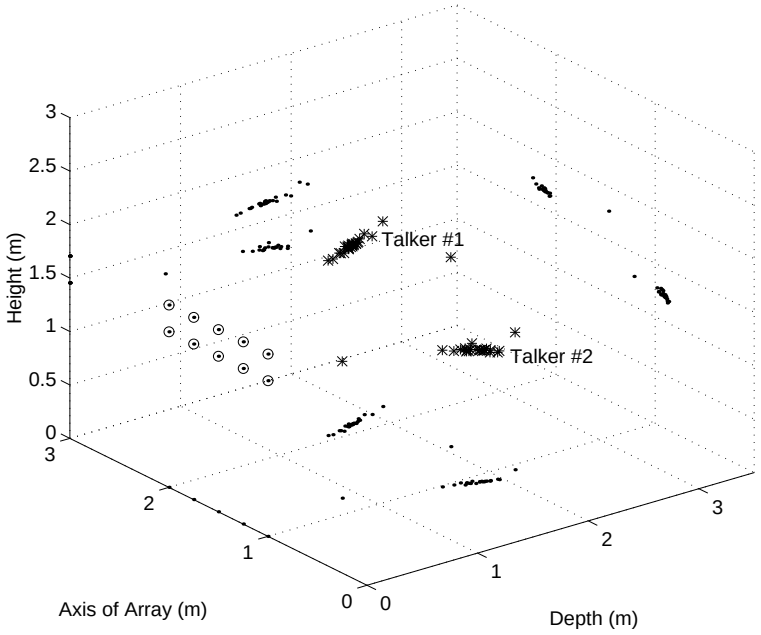
Figure 9.5: Bilinear Array Experiment #2: Multi-Talkers. The top two graphs illustrate the valid x, y, and z-positions for each of the individual recordings as functions of time. The lower graph presents the location estimates for the simultaneous recording. The single talker positions are replotted as '.' while the simultaneous location estimates are given by 'o'.

tern and speed to those employed in Section 8.4.2 and once again represents a particularly extreme two-talker case with significant periods of signal overlap.

The results of this experiment are presented in Figures 9.5 and 9.6. The first of these figures illustrates the valid x, y, and z-positions as functions of time. These estimates were generated using the DOA localization method and the empirical detection test. The top two plots represent the individual recordings and their respective location data. Note that



(a) Individual Talkers



(b) Simultaneous Talkers

Figure 9.6: Bilinear Array Experiment #2: 3-dimensional scatter plots of the position versus time data of Figure 9.5. The top graph presents the results of the two individual recordings and the lower graph shows the simultaneous situation. In each case, the 3-dimensional location is denoted by '*' and a '.' is used to show the orthogonal projection of the location onto the respective planes.

in each case the positions remain nearly constant throughout the utterances, indicating fixed sources. For each of these two-second recordings, the signals were segmented into 155 analysis frames. Talker #1 possessed 66 valid frames, and talker #2 had 49 valid frames. The lower graph in the figure illustrates the same information for the simultaneous recording. The single talker positions are replotted as '.' while the simultaneous location estimates are given by 'o'. This time 52 valid frames were detected. Once again, as in the case of the TDOA estimation, the algorithm is correctly able to discriminate periods of single-source activity from multi-source intervals. The location estimates achieved clearly demonstrates the two-party nature of the received signal. Despite the overlapping nature of the signals, a significant fraction of each of the valid individual estimates are preserved in the simultaneous recording. Figure 9.6 contains 3-dimensional scatter plots of the position versus time data of Figure 9.5. The top graph presents the results of the two individual recordings and the lower graph shows the simultaneous situation. In each case, the 3-dimensional location is denoted by '*' and a '.' is used to show the orthogonal projection of the location onto the respective planes. Two distinct sources are evident in these graphs. Each is approximately 1.75m from the array. Talker #1 is in front of the extreme left sensors at a height of 1.60m and Talker #2 is at the right edge of the array at height 1.20m.

9.2.3 Experiment #3: Moving Talkers

With a 25.6ms analysis window and the ability to generate independent location estimates on the order of 70 times per second, the localization algorithms presented here are appropriate for tracking moving speech sources. As a result of the short analysis interval, a source's change of location within the estimation period is insubstantial and has minimal impact on the precision of the calculated delays and derived location. The high update rate allows

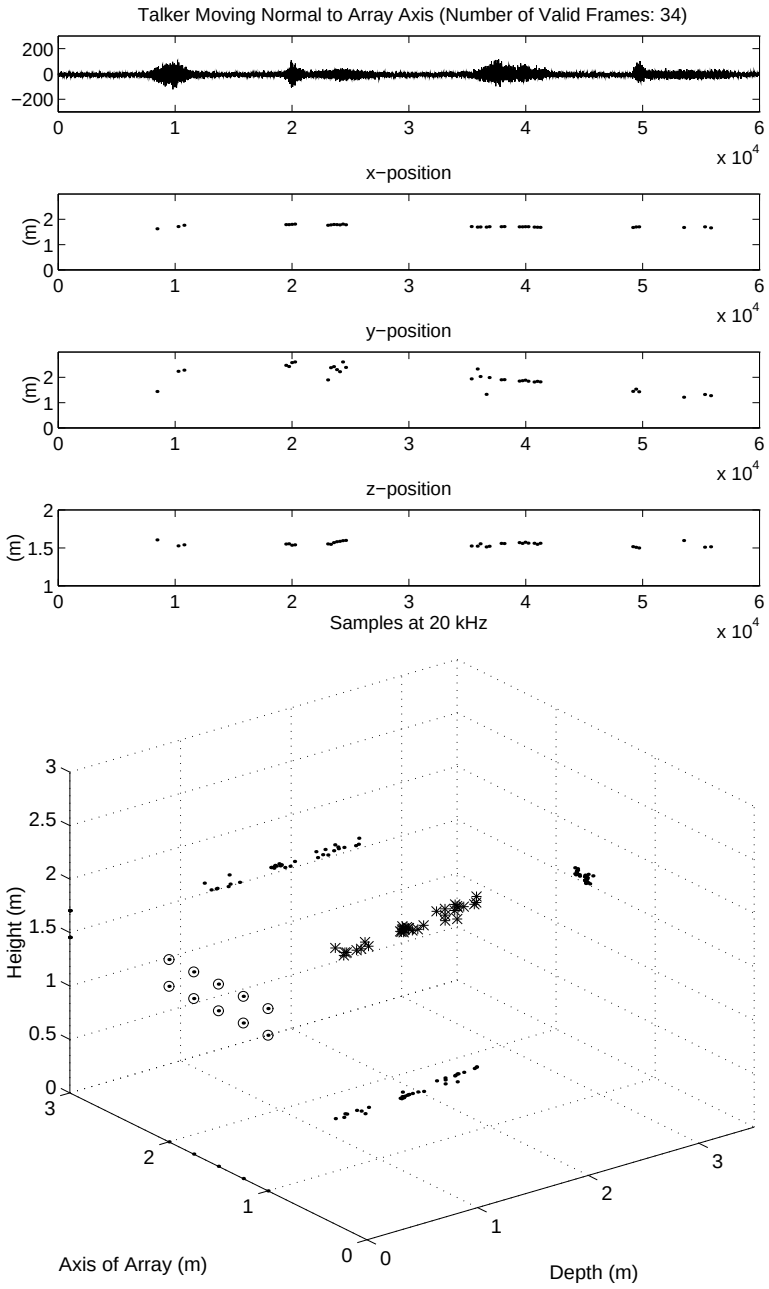


Figure 9.7: Bilinear Array Experiment #3: Talker Moving Normal to Array Axis. The top graph contains plots the signal received at a single microphone as well as the valid x-, y-, and z-positions of the moving source as a function of time. The lower graph presents the localization data in 3-dimensional scatter plot. A '*' denotes the location estimates while a '.' is used to show the orthogonal projection of the location onto the respective planes.

for near continuous localization of even the most rapidly moving sources in a typical talker scenario.

Figures 9.7 through 9.10 illustrate the ability of the location algorithm to track a

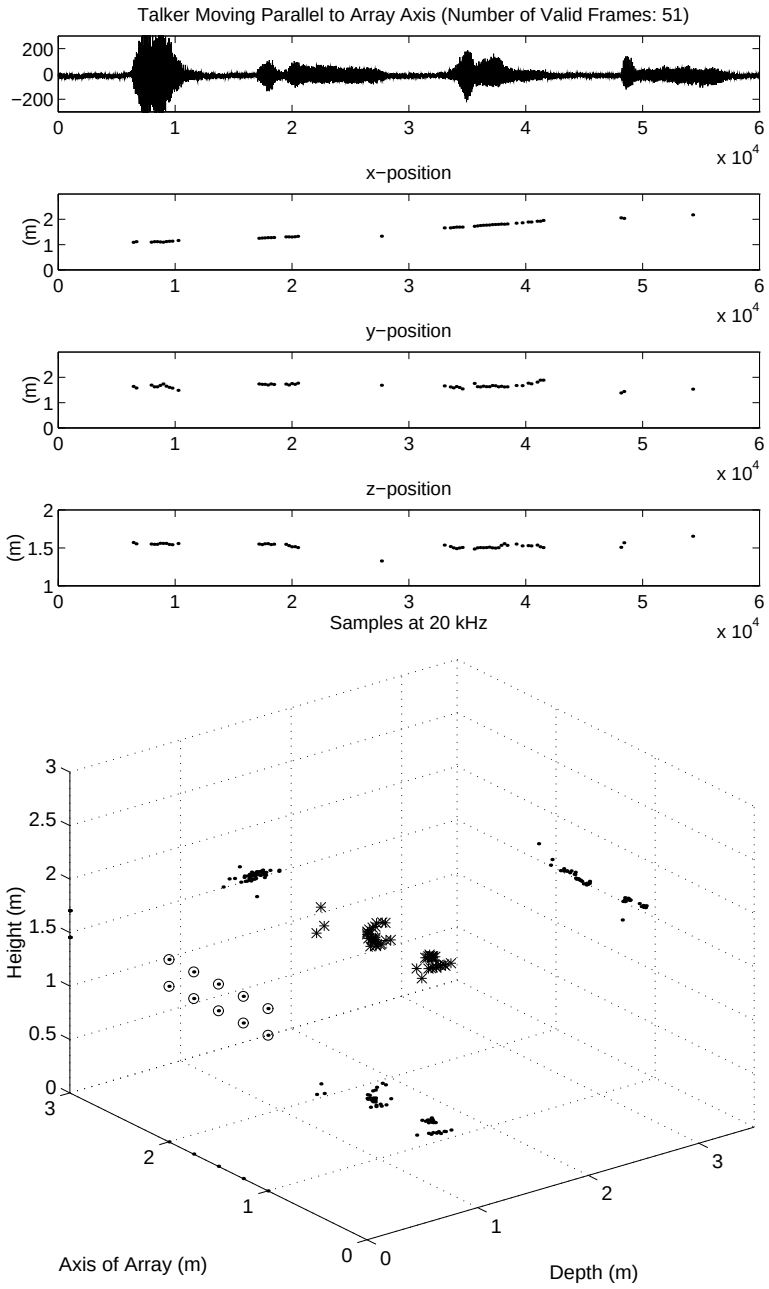


Figure 9.8: Bilinear Array Experiment #3: Talker Moving Parallel to Array Axis. The top graph contains plots the signal received at a single microphone as well as the valid x-, y-, and z-positions of the moving source as a function of time. The lower graph presents the localization data in 3-dimensional scatter plot. A '*' denotes the location estimates while a '.' is used to show the orthogonal projection of the location onto the respective planes.

moving talker. In each of these examples, a talker spoke the phrase "One Two Three Four" with varying degrees of loudness while walking in a number of directions relative to the bilinear array. Location estimation was performed for each of the three-second (60,000

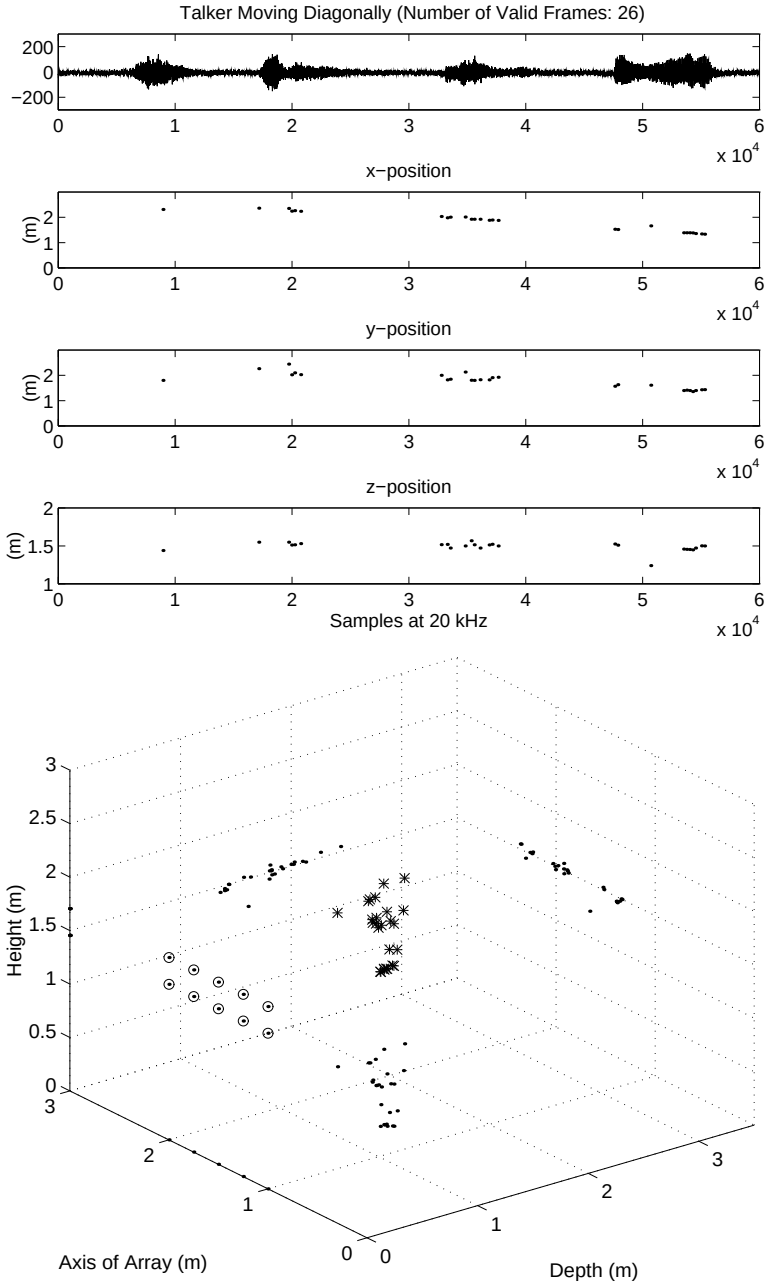


Figure 9.9: Bilinear Array Experiment #3: Talker Moving Diagonally Across Array Axis. The top graph contains plots the signal received at a single microphone as well as the valid x-, y-, and z-positions of the moving source as a function of time. The lower graph presents the localization data in 3-dimensional scatter plot. A '*' denotes the location estimates while a '.' is used to show the orthogonal projection of the location onto the respective planes.

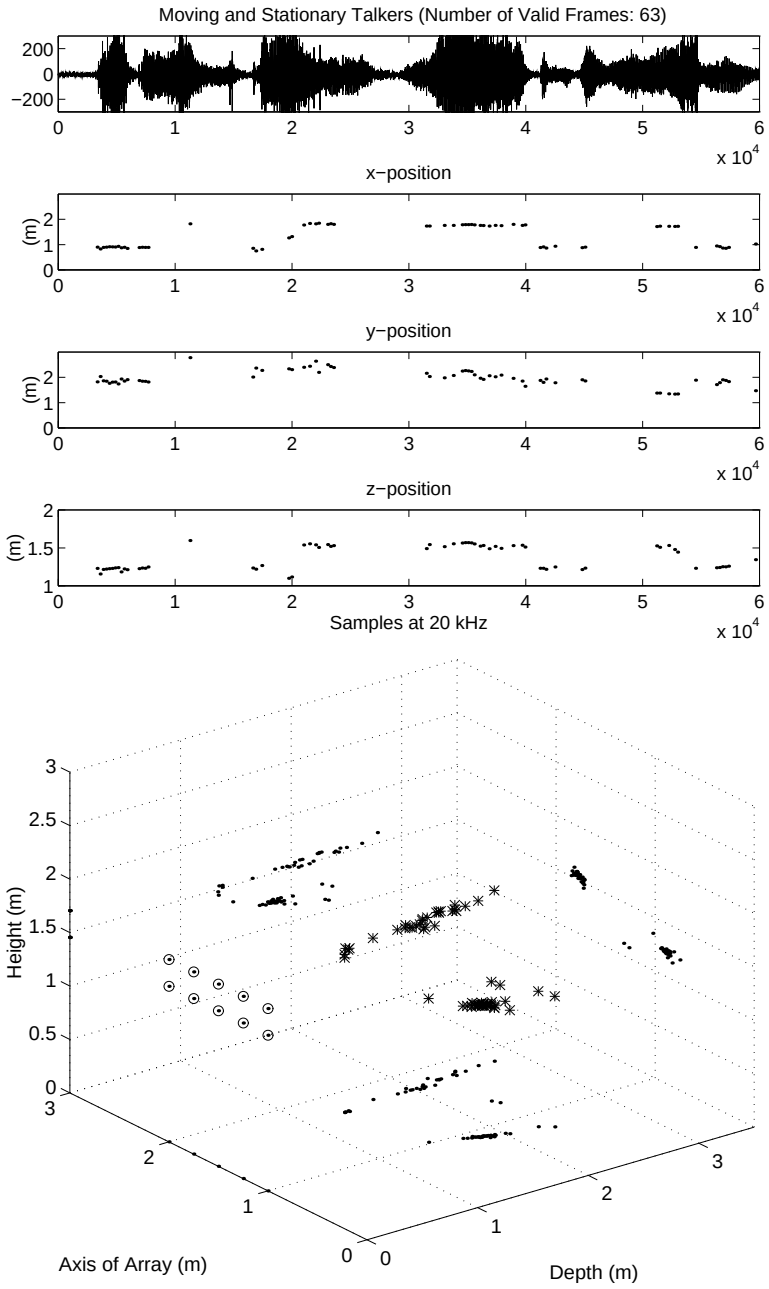


Figure 9.10: Bilinear Array Experiment #3: Moving and Fixed Talkers. The top graph contains plots the signal received at a single microphone as well as the valid x-, y-, and z-positions of the moving and fixed sources as a function of time. The lower graph presents the localization data in 3-dimensional scatter plot. A '*' denotes the location estimates while a '.' is used to show the orthogonal projection of the location onto the respective planes.

sample) recordings. The figures display the valid results obtained from the DOA-based locator with the empirical detection test. The top graph in each case shows the time signal received at a single microphone along with the x-, y-, and z-positions of the valid locations as functions of time. The lower graphs contain 3-dimensional scatter plots of this same location information. Once again, the locations themselves are plotted with '*', and a '.' is used to denote their projections onto the back planes. Figure 9.7 presents a source moving towards the sensors in a direction normal to the array axis. Note the relative consistency of the x- and z-positions as the y value steadily decreases. While there was no mechanism available to accurately determine the exact path traversed by the talker, the smooth, nearly linear path detected by the locator certainly suggests the algorithm was performing accurately. For this particular recording, the talker was speaking quite softly and as the time signal indicates, the signal to noise ratio at this microphone is relatively poor. The SNR condition results in fewer valid locations, 34 out of 233 analysis frames, and reduced estimate accuracy. The source displayed in Figure 9.8 was recorded while moving parallel to the array. A moderate speech volume produced 51 valid estimate frames which are manifested in the graph as a smooth upward transition in the x-position throughout the course of the utterance. The third case, presented in Figure 9.9, was done with the talker walking inward, diagonally across the array. For the 26 valid estimates detected, both the x- and y-positions may be seen to slowly decrease over the three-second interval.

With each recording the y-position parameter displays a noticeable deviation from the linear nature that would be expected for these source motions. The x- and z-estimates are less sensitive in this regard. This behavior is consistent with the results of experiment #1 in which the range estimate was shown to possess the bulk of the error inherent in localization with the bilinear array. For these experiments, the range component contributes primarily

to the y-position value.

The last example, Figure 9.10, repeats the scenario of a source moving normal to the array, but this time a second, fixed source has been included as well. As a whole the estimates reveal a distinct two-source situation with one source appearing to be moving in a continuous fashion while the other is relatively stationary. However, if taken in isolation, these location values may be the source of reasonable confusion. In practice, the product provided by these algorithms may be combined with single and multi-source tracking schemes to follow and discriminate individual talkers in a multi-party environment. To achieve these results many tracking techniques may be adapted from sonar and radar applications [27, 87, 88, 89]. One important distinction between this situation and the radar/sonar scenarios is the source-motion model employed. The latter may assume that tracked elements are constrained to roughly linear motion with limited acceleration potential. Furthermore, because these methods are usually active or rely on a continuous source signal, location updates are available on a regular basis. However, for a typical multi-talker situation, location information is evaluated only when a particular source is speaking. These periods may be well separated. This difficulty is compounded by the fact that talker motion is subject to a variety of discontinuities, necessitating the use of a much more general source-motion model. A general model, Kalman filter approach to tracking a single speech source was explored in [90]. The multi-source problem is significantly more complicated, and, in addition to more sophisticated tracking methods, may benefit tremendously from the incorporation of speaker identification procedures applied to the received microphone signals [91].

9.3 A Multi-Unit Conferencing Array System

The second set of experiments was performed with a multi-unit array placed within a conference room. The room is 4m $\times$ 7m with a carpeted floor and an acoustically tiled ceiling at a height of 2.75m. Four full-length windows are found along one of the side walls while the door is located opposite. Acoustic partitions have been hung in the space remaining along each of the side walls. One of the end walls consists of a white-board attached to painted concrete block. The other is untreated plaster with no obstructions. The primary feature of the room is a 4m Formica conference table. The table is slightly oblong, .95m in width at the ends and 1.20m at the center, and at a height of .7m. The enclosure is well-insulated from background noise and possesses a reverberation time of approximately 300ms.

The 14 microphones were partitioned among three autonomous array units: the main array consisting of six microphones in a rectangular arrangement and two remote arrays each with four microphones forming orthogonal pairs. The main and remote array structures are displayed in Figure 9.11. The arrays were designed for their portability, effectiveness, and adaptability. In each case, the pressure-gradient microphones were mounted into rubber grommets and placed within two parallel tracks made of stretched rubber 'o' rings. This arrangement allowed for free motion of the microphone along the line of constraint while acoustically decoupling the sensor from the array structure. The rubber 'o' ring tracks were affixed to 15cm screws drilled into a panel of aluminum which was in turn, attached to a tripod. Finally, a portion of the space between the microphones and metal backplane was filled with 10cm acoustic foam designed to prevent direct back-reflections while leaving open space behind the sensors, a requirement for nominal pressure-gradient microphone operation. In the experiments that follow, the microphone spacings were set at .25m. With

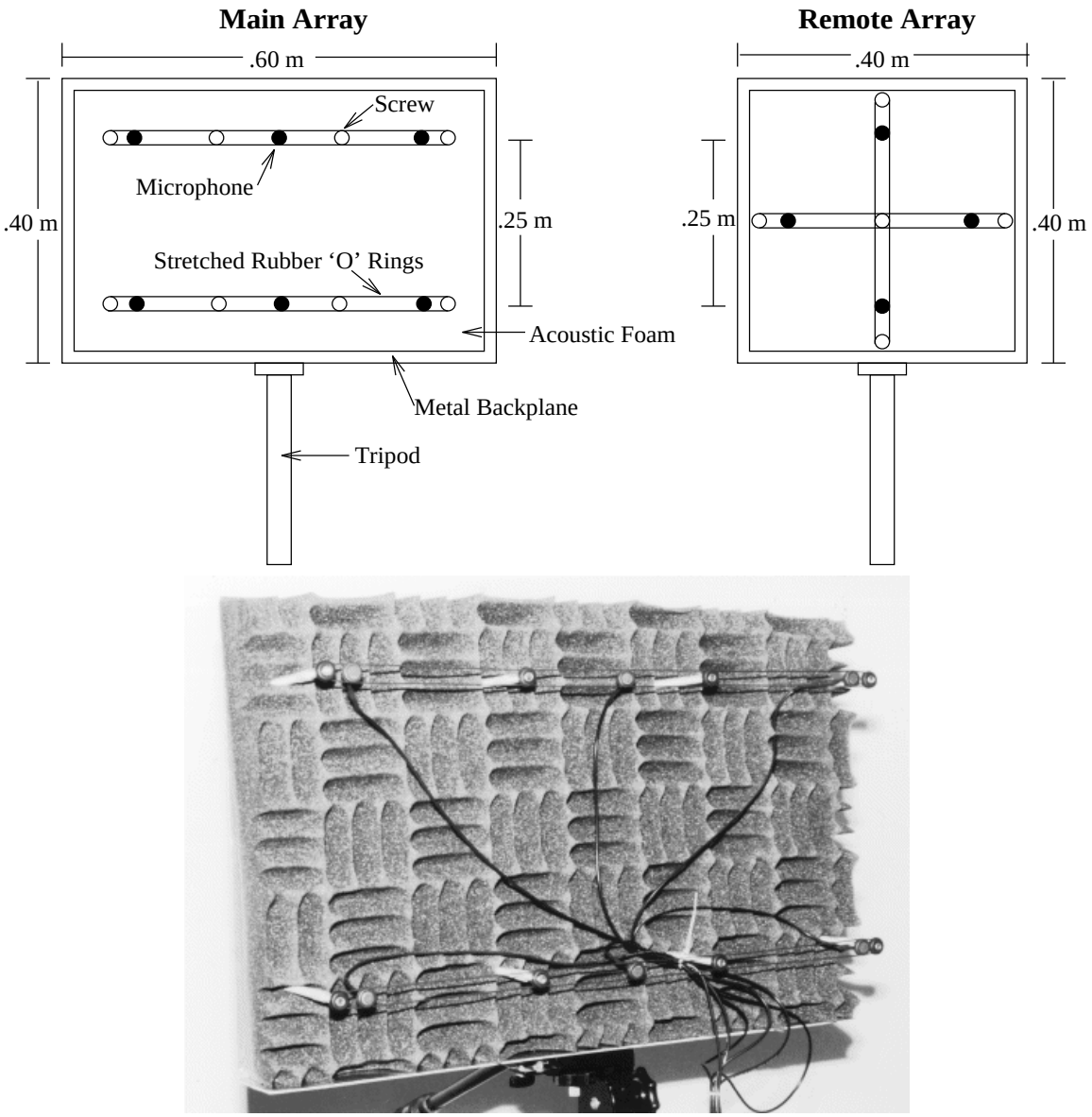


Figure 9.11: Multi-Unit Conference Array System: Diagrams of the main and remote arrays used for the experiments. The photo represents the main array.

the six-element main array, four sensor pairs were defined as the diagonal elements of the two square units, yielding a sensor-pair separation of $.25\sqrt{2}$m in each case. The remote arrays consisted of two sensor pairs apiece, each with a .25m separation.

The main array was positioned at one end of the conference table, 1m from the end wall and facing the participants. The two remote arrays were placed at the midpoint of the

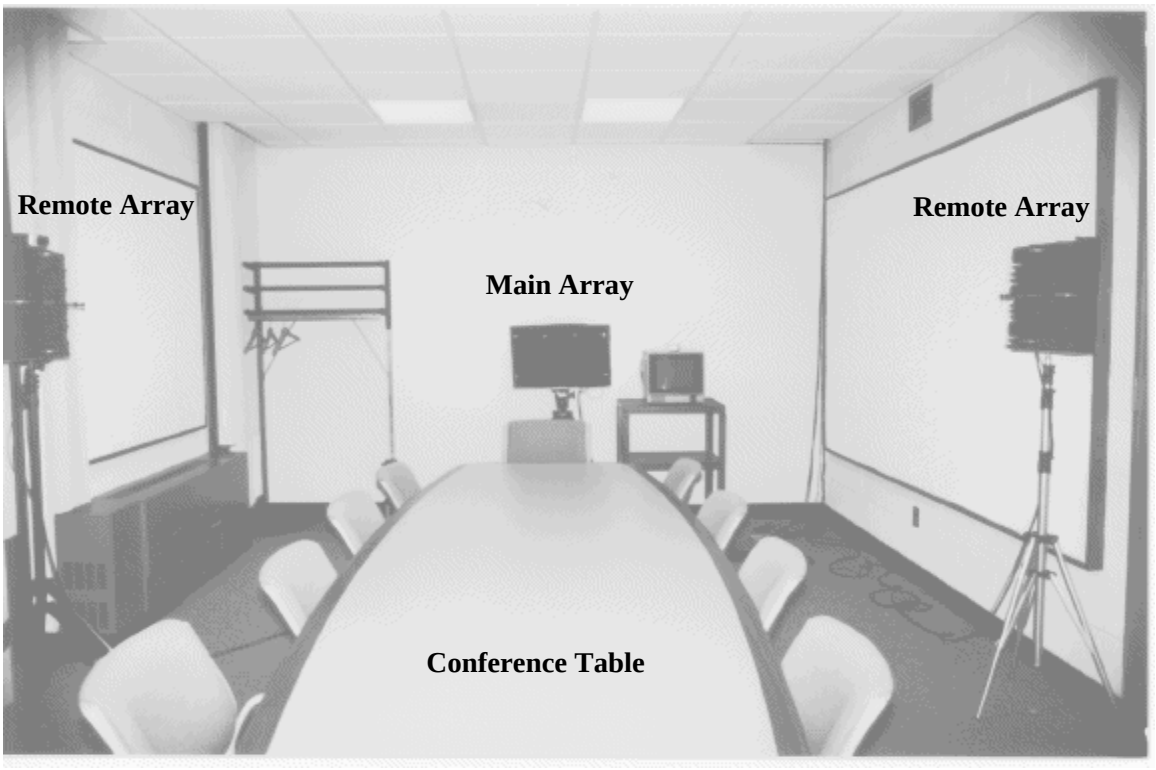


Figure 9.12: Multi-Unit Conference Array System: Photo of the conference room and the three array setup.

room, abutting the side walls, and facing the conference table. The two remote arrays were centered at a height of 1.58m, approximately standing height, while the main array was placed at a height of 1.27m to accommodate seated participants. This heuristic choice of array positions was guided by the intent to provide a general coverage of the area surrounding the conference table given the three array units and the practical restrictions of the room. All of these positioning measurements were done by hand using an ultrasonic measuring device. and are subject to limited accuracy, on the order of centimeters. Figure 9.12 presents a photo of the conference room along with the three array units.

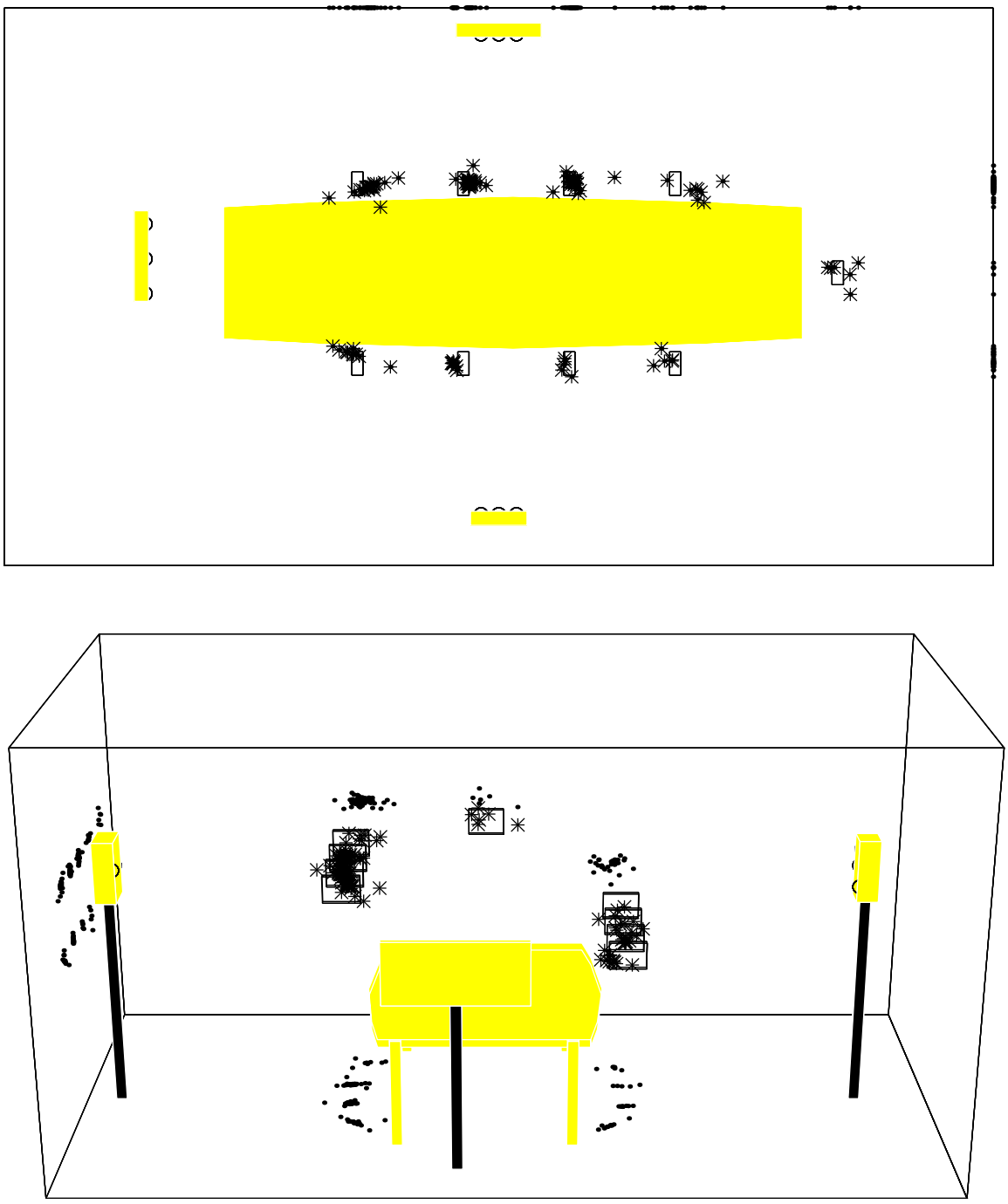


Figure 9.13: Multi-Unit Conference Array System: Experiment #1. Location estimates. generated using the DOA localization procedure and validated by the empirical detection test. The top plot presents an overhead view of the room with the table and arrays shaded and the individual speaker positions indicated by rectangular boxes. The lower plot presents the same data from the perspective of the wall behind the main array. The location estimates produced by the nine two-second recordings are plotted with a '*' symbol and their orthogonal projections are denoted by a '.' on the respective walls.

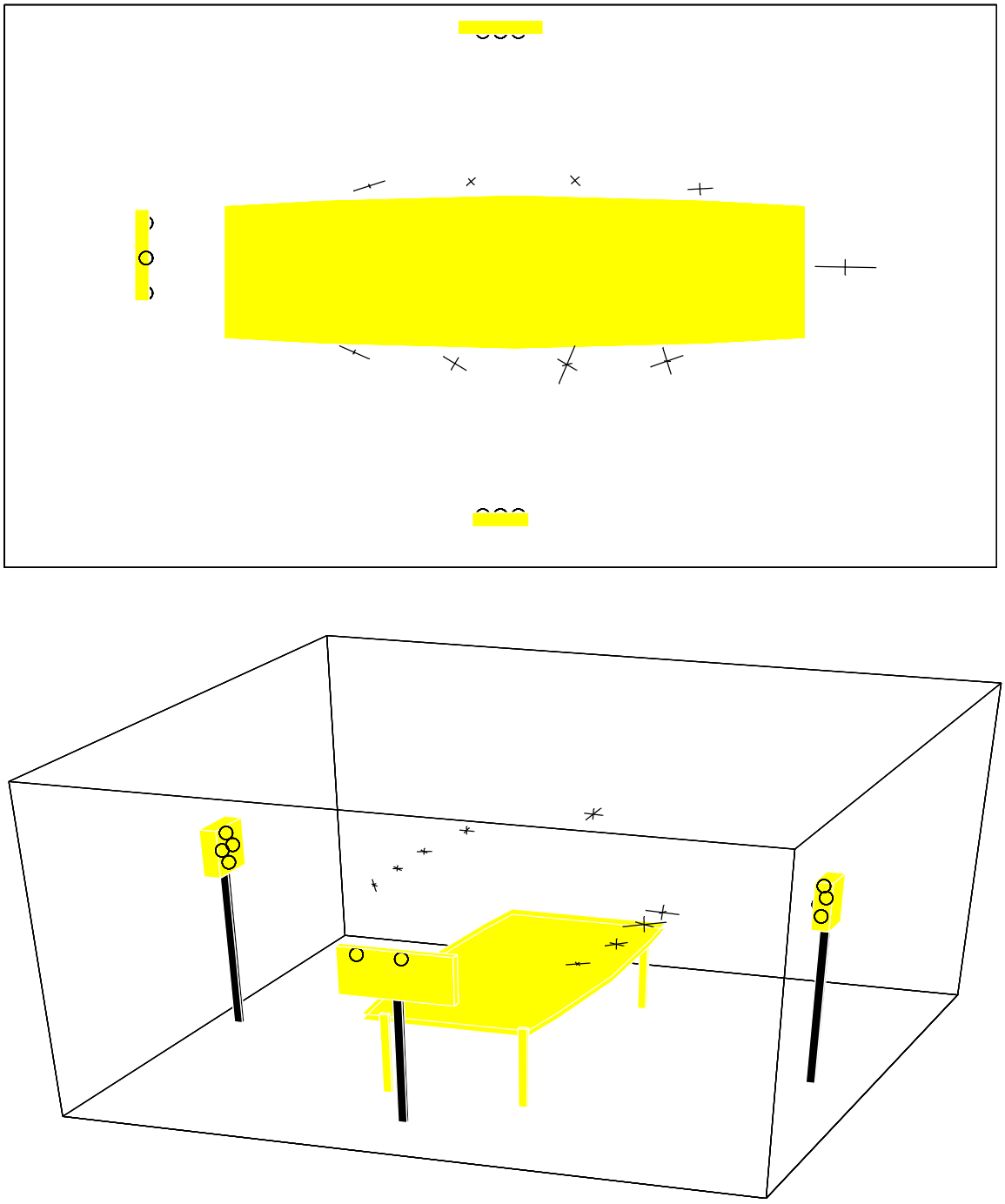


Figure 9.14: Multi-Unit Conference Array System: Experiment #1 Predicted Error Region. The principal component vectors of the predicted covariance matrices calculated via the median TDOA variance figures and scaled to 2.5 standard deviations. The top plot presents an overhead view while the lower graph provides a front-wall perspective.

9.3.1 Experiment #1: A Source Grid

The first experiment performed with the Conferencing Array System involved localization of sources at fixed positions around the conference table. The design is similar to that of the experiments presented in Section 9.2.1. A two-second phrase was simultaneously played through a speaker and recorded by the 14 microphones. The utterance itself, as well as the playback and recording levels, were identical to those of the earlier experiment. The peak SNR levels again varied from approximately 5 to 30dB, depending on the speaker's location and orientation relative to a particular microphone. The loudspeaker locations were selected in an attempt to emulate a true video-conference scenario, being placed at .75m intervals along the length of the table with one placed at the head. This resulted in nine distinct locations, four symmetrically placed on either side. In each case the loudspeaker was oriented towards the main array at the end of the table where the camera/video display would presumably be found. Because of the midline symmetry of the array/room setup, two specific source heights were adopted. For one side of the table, the loudspeaker was placed at 1.58m to simulate standing sources and with the other side, a height of 1.15m was utilized, corresponding to a seated talker.

Some partial results of this experiment are presented in Figure 9.13. These graphs were generated using the DOA location estimates validated by the empirical-detection test. The top plot is an overhead view of the room with the table and arrays shaded and the individual speaker positions indicated by rectangular boxes. The location estimates produced by the nine two-second recordings are plotted with a '*' symbol and their orthogonal projections are denoted by a '.' on the respective walls. The lower plot presents the same data from the perspective of the wall behind the main array. Height information is apparent in this second plot. The locations on the right side are seated while those on the left side and at

the head of the table are standing.

The predicted error regions associated with each of these source positions relative to the array geometry are illustrated in Figure 9.14. The principal component vectors were generated in the same fashion as those of Figure 9.4. Specifically, the TDOA variances required for the error region estimate have been calculated from the median of the valid frame variance figures at each speaker position and the principal components displayed have been scaled by 2.5 standard deviations. The top plot in Figure 9.14 represents an overhead view of the predicted error region principal components while the lower graph provides a front-wall perspective.

As an initial observation, the estimate clusters of Figure 9.13 and their predicted error regions in Figure 9.14 appear to approximately agree in orientation and extent. On the whole, the location estimates of this section possess significantly smaller and less eccentric error regions than those results obtained with a similar experiment employing the Bilinear Array (Figure 9.3). This is a byproduct of the more robust placement of microphones throughout the source region and is consistent with the simulations conducted in Section 5.4. The localization-error region clearly benefits from the more uniform sensor positioning. However, relative to the earlier experiment the quantity of valid detection estimates for this two-second utterance has been reduced dramatically. In the present case, the number of valid estimates detected by the DOA and TDOA localization schemes for the nine source locations was found to range between 5 and 36 out of a total of 155 analysis frames. With the LI localization scheme, this value peaked at 29 and was as little as 2. Referring to Table 9.2, for the experiment conducted in Section 9.2.1 equivalent valid frame numbers span from 22 to 67 frames for the TDOA/DOA schemes and 15 to 62 for the closed-form LI method. Given that the content and playback conditions of the spoken utterance were

identical in each scenario, several practical factors may be responsible for this disparity. The conferencing array system employs a significantly larger room than the bilinear array system and accordingly, worst case source to microphone distances are greater in the former case. This increased distance results in a reduced signal SNR due to propagation attenuation. Loudspeaker orientation may also be responsible in part for the reduced number of detected TDOA and location estimates. Each recording was produced with the loudspeaker situated facing the front array. As Figure 9.13 indicates, for several locations microphones on the remote arrays are situated off-angle to the direct source radiation path and may receive a distorted version of the source signal. The effects of source orientation on radiation pattern have been investigated in [92, 93]. This situation is compounded by the reception characteristics of the microphones themselves. The pressure-gradient microphones employed here have a cardioid directivity designed to attenuate off-angle sources. While this may be a useful feature for reducing the contribution of noise sources outside the desired source-field, it is detrimental to the signal quality of several sources in the conference array system.

Overall, these arguments highlight a general tradeoff between the increased location accuracy obtained via the multi-unit array at the expense of the frequency of valid estimates. The selection of an array geometry is dependent upon the application in mind. In this instance, achieving a small overall location error was the priority. With other scenarios, room restrictions may prevent an unconstrained placement of sensors. If range information is considered unimportant, the single bilinear array would be quite effective for providing bearing-only estimates.

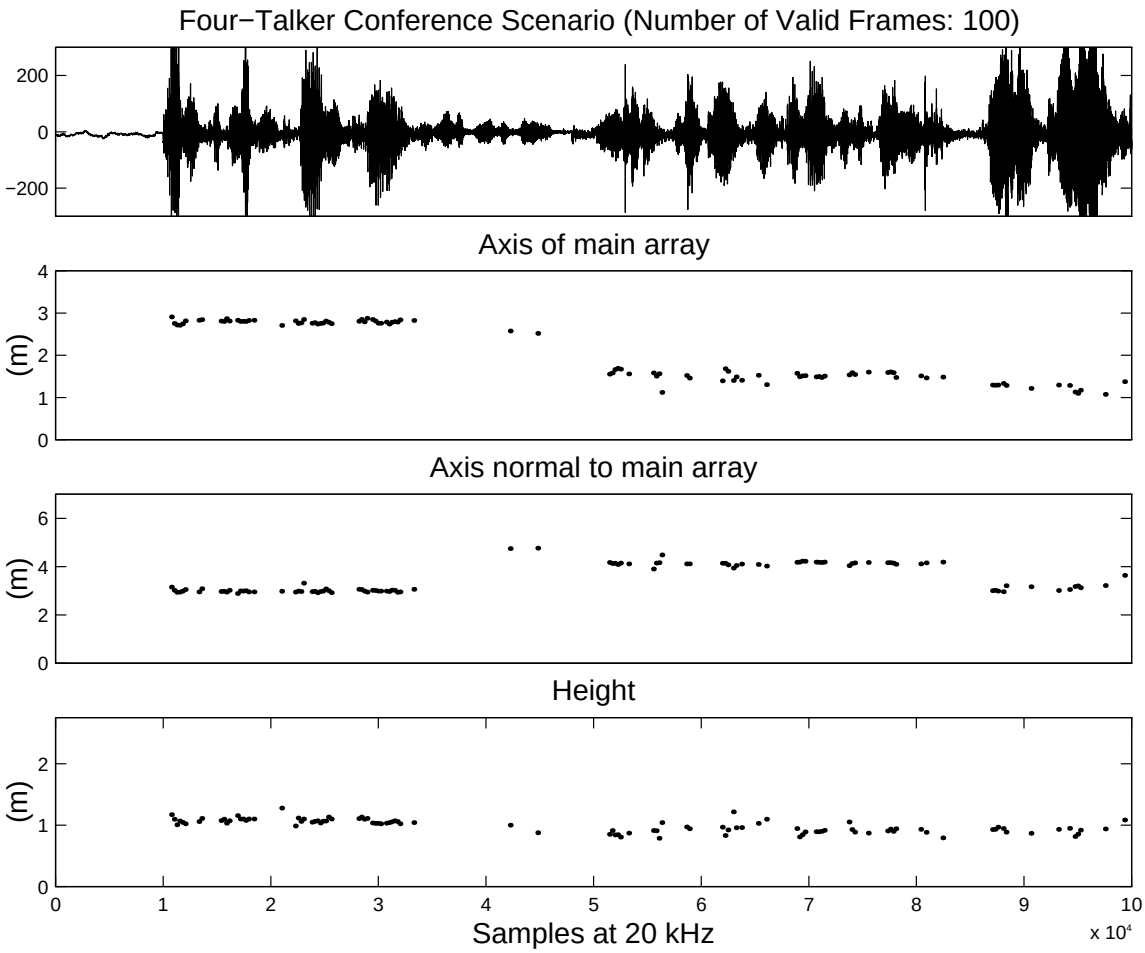


Figure 9.15: Multi-Unit Conference Array System: Experiment #2. Location estimates of the four talkers plotted as a function of time samples at 20kHz. The top graph shows the signal received at one of the microphones. The lower three plots display the valid location estimates along each of the three room axes as a function time samples.

9.3.2 Experiment #2: A Conference Scenario

As a final demonstration of the effectiveness of the localization methods presented in this work, a typical conference scenario was created using four talkers. The individuals were seated at various positions around the conference table and each asked to speak a single sentence in turn during the course of a five-second recording. Figures 9.15 and 9.16 present the localization results obtained. Figure 9.15 graphs the location information as a function of the sample time index while Figure 9.16 illustrates the same data via three-dimensional

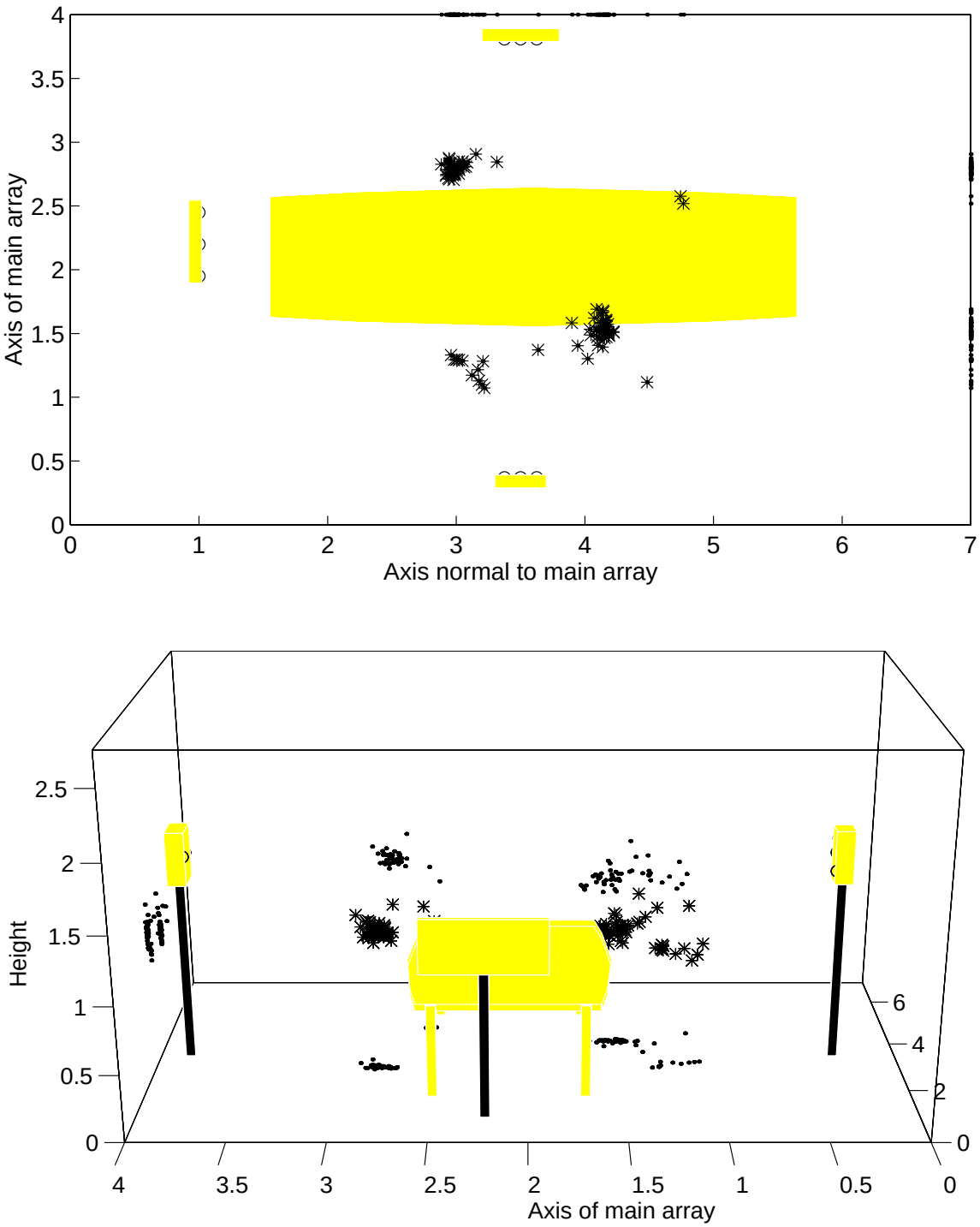


Figure 9.16: Multi-Unit Conference Array System: Experiment #2. Scatter plots of the localization data in Figure 9.15. The top plot presents an overhead view while the lower plot shows the room from the perspective of the wall behind the main array.

scatter plots. As in the previous experiment, these estimates were obtained using the combination of the DOA localization procedure and the empirical detection test.

The four individual talkers are discernible from their varying amplitude levels in the signal plotted in Figure 9.15. The first sequential pair of talkers are seated on the left side of the table and produce similar location estimates along the axis of the main array and with regard to height. Their positions in the direction normal to the array clearly vary. The situation is identical for the second talker pair. With this five-second recording 100 of the total 354 analysis frames were identified as producing valid location estimates. The second talker has a particularly short utterance (less than one second) and possesses a relatively weak signal strength. Accordingly, only two valid frames associated with this talker were determined.

From the scatter plots in Figure 9.16, four estimate clusters are apparent. The two at the far end of the table, opposite the main array, are notable in that both talkers appear to be leaning over the conference table. In one case a number of valid locations have been detected, but several outliers are associated with the cluster. These errant estimates are due in part to the error region inherent in the source's location and signal content. However, a further cause of this error may be attributed to an excessive multipath condition created by the presence of the table. In this situation, the talker is facing the main array and reading from a document placed on the table. Projecting downward in this manner, a significant secondary signal is generated off of the reflective, Formica table surface. While multipath conditions have existed in all the experiments presented in this chapter, the form and degree of the source reflection is particularly extreme in this case. This situation may be responsible for producing lower TDOA estimate precision than would be expected given the SNR conditions alone and therefore reduced localization accuracy for this source.

Despite the practical obstacles involved in translating the source localization process from a laboratory to applied environment, the results of this experiment demonstrate that the localization methods provide an effective means for placing and discriminating individual talkers in a real-world conferencing scenario.

9.4 Discussion

Several experiments have been presented in this chapter to establish the utility of the speech source localization framework detailed in this thesis. The source grid experiment of Section 9.2.1 quantified and compared the various localization procedures, two detection tests, and the error prediction method in the context of a laboratory-environment, bilinear array. Sections 9.2.2 and 9.2.3 were designed to illustrate the system's ability to distinguish and locate multiple and moving talkers, respectively. Finally, Section 9.3 presented the algorithms operating with an alternative array configuration in a real-world, conference-room setting.

Chapter 10

Conclusions and Future Work

The goal of this work has been to detail an effective system for the localization of one or more speech sources in a real-room environment. In Chapter 2 an appropriate source-sensor geometry was developed and forwarded as the basis for a series of localization error criteria, means for detecting a source presence, and techniques to evaluate the accuracy of a location estimate. Three least-squares error criteria were introduced in Chapter 3. The TDOA-based error, J_{TDOA}, was shown to yield the ML estimate under Gaussian noise conditions. However, a second criterion based upon the direction-of-arrival information, J_{DOA}, was found in simulations to produce superior performance when confronted with less favorable conditions. The third error criterion, employing a total distance measure, exhibited an extreme estimator bias and was not considered for further development (although it is brought up again in Chapter 7 in relation to the error criterion minimized by the closed-form spherical interpolation (SI) locator). Chapter 4 presented three source detection tests, two statistical and one empirical, designed for various TDOA noise scenarios. The two more general methods, the source consistency and the empirical tests, were included into the real-system experiments that followed. Chapter 5 contained an analysis of the error region

associated with a location estimate. Explicit formulae were derived relating the estimation confidence to the TDOA and sensor-pair data. These techniques were demonstrated through simulations to accurately predict the exhibited estimation spread.

The second half of this thesis focused on the development of practical algorithms to implement the source localization procedures and the evaluation of system performance in a number of real-world experiments. Chapter 6 highlighted a number of computational issues involved in obtaining the optimal nonlinear location estimates. Chapter 7 was devoted to the development of a closed-form location estimator, the Linear Intersection (LI) method, designed specifically for this application. The LI estimator was shown to generate precise results, on par with the more burdensome nonlinear estimators and superior to those of a representative algorithm. The closed-form solution may be used alone or as the initial guess for the nonlinear estimator's optimization routines. A set of time delay estimates are the basis for each of the techniques offered in this work. In Chapter 8 a practical TDOA estimator for speech sources was spotlighted. The algorithm requires minimal computational resources to produce precise TDOA figures. Through the incorporation of a short analysis window and source detection methods, it is capable of tracking moving sources as well as identifying multi-source situations. Chapter 9 brought together each of the individual aspects of this work through a series of experiments with real microphone array systems. Two distinct environments and array geometries were employed for these evaluations, a bilinear array placed along a wall of a computer laboratory and three small independent arrays set up in a conference room. The experiments performed corroborated the results of earlier simulations and demonstrated the effectiveness and applicability of the source localization techniques advanced in this thesis for real-world scenarios. With these methods, it was possible to detect and localize sources in 3-space to within centimeters precision, track

moving sources reliably, identify individual talkers in multi-source scenarios, and predict consistently the error region associated with a particular location estimate.

Along the lines of this research, there are several avenues available for future study. First, the problem of tracking and distinguishing multiple sources given the location data provided by these procedures was alluded to in the discussion within Chapter 9. Considerable work still remains in this area for adapting existing tracking methods and creating novel techniques appropriate for the specific application. Second, the issue of optimal sensor placement has not been addressed substantially here. In Chapter 5 it was suggested that the error region predictor developed there could be incorporated into some optimal means for array design. However, for the experiments and simulations performed here the choice of sensor placement was primarily *ad hoc* and guided by practical considerations and algorithmic constraints. Third, the estimation and correction for source orientation is an important aspect of the talker localization problem. In the context of this work, the orientation angle was effectively lumped into the larger parameter of signal SNR, but in many scenarios knowledge of a talker's orientation in addition to location may be vital. The research in this area, referred to in Chapter 9, has taken important steps in characterizing source radiation patterns and may eventually provide acoustic tools for assessing talker orientation as well as other parameters. Finally, there is the practical issue of sensor calibration. For the experiments presented here, the sensors were placed and measured by hand while the signal processing channels (anti-aliasing filters, A/D converter, etc.) were normalized in an informal fashion. Each of these procedures introduces a degree of imprecision into the overall results and provides for a practical inconvenience. The development of methods to automatically and accurately identify sensor positions as well as calibrate the acquisition channels, will be required to facilitate the incorporation of this technology into desirable

commercial products.

Bibliography

[1] J. L. Flanagan. Bandwidth design for speech-seeking microphone arrays. In *Proceedings of ICASSP85*, pages 732–735, Tampa, FL, March 1985.

[2] W. Kellerman. A self-steering digital microphone array. In *Proceedings of ICASSP91*, pages 3581–3584, Toronto, CA, May 1991.

[3] J. Flanagan. Beamwidth and usable bandwidth of delay-steered microphone arrays. *AT&T Technical Journal*, Vol.64:983–995, April 1985.

[4] J. Flanagan, D. Berkley, G. Elko, J. West, and M. Sondhi. Autodirective microphone systems. *Acustica*, 73:58–71, 1991.

[5] H. F. Silverman. Some analysis of microphone arrays for speech data acquisition. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-35(2):1699–1712, December 1987.

[6] C. Che, M. Rahim, and J. Flanagan. Robust speech recognition in a multimedia teleconferencing environment. *J. Acoust. Soc. Am.*, Vol.92(4, pt.2):2476(A), 1992.

[7] 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. ARPA-Software & Intelligent Systems Technology Office.

[8] 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, Sep. 1994.

[9] Q. Lin, E. Jan, and J. Flanagan. Microphone arrays and speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):622–629, October 1994.

[10] Y. Grenier. A microphone array for car environments. In *Proceedings of ICASSP92*, pages I–305 – I–309, San Francisco, CA, April 1992.

[11] S. Oh, V. Viswanathan, and P. Papamichalis. Hands-free voice communication in an automobile with a microphone array. In *Proceedings of ICASSP92*, pages I–281 – I–284, San Francisco, CA, April 1992.

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

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

[14] E. Adugna. *Speech Enhancement Using Microphone Arrays.* PhD thesis, Rutgers University, New Brunswick, NJ, October 1994.

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

[16] Maurizio Omologo and Piergiorgio Svaizer. Acoustic event localization using a crosspower-spectrum phase based technique. In *Proceedings of ICASSP94*, pages II–273– II–276. IEEE, 1994.

[17] M. Omologo and P. Svaizer. Use of the cross-power spectrum phase in acoustic event localization. Technical Report Technical Report No. 9303-13, IRST, Povo di Trento, Italy, March 1993.

[18] J. E. Greenberg and P. M. Zurek. Evaluation of an adaptive beamforming method for hearing aids. *J. Acoust. Soc. Amer.*, 91:1662–1676, March 1992.

[19] J. L. Flanagan and H. F. Silverman. Material for international workshop on microphone array systems: Theory and practice. Technical report, Division of Engineering, Brown University, Providence, RI 02912, October 1992.

[20] W. Bangs and P. Schultheis. Space-time processing for optimal parameter estimation. In J. Griffiths, P. Stocklin, and C. Van Schooneveld, editors, *Signal Processing*, pages 577–590. New York:Academic, 1973.

[21] G. Carter. Variance bounds for passively locating an acoustic source with a symmetric line array. *J. Acoust. Soc. Am.*, Vol.62(4):922–926, October 1977.

[22] 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.

[23] W. Hahn. Optimum signal processing for passive sonar range and bearing estimation. *J. Acoust. Soc. Am.*, Vol.58(1):201–207, July 1975.

[24] M. Wax and T. Kailath. Optimum localization of multiple sources by passive arrays. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-31(5):1210–1217, October 1983.

[25] V. M. Alvarado. *Talker Localization and Optimal Placement of Microphones for a Linear MIcrophone Array using Stochastic Region Contraction.* PhD thesis, Brown University, Providence, RI, May 1990.

[26] 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.

[27] D. Johnson and D. Dudgeon. *Array Signal Processing- Concepts and Techniques.* Prentice Hall, first edition, 1993.

[28] S. Haykin. *Adaptive Filter Theory.* Prentice Hall, second edition, 1991.

[29] A. Vural. Effects of pertubations on the performance of optimum/adaptive arrays. *IEEE Trans. Aerosp. Electron.*, AES-15(1):76–87, January 1979.

[30] R. Compton Jr. *Adaptive Antennas*. Prentice Hall, first edition, 1988.

[31] T. Shan, M. Wax, and T. Kailath. On spatial smoothing for direction-of-arrival estimation in coherent signals. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33(8):806–811, August 1985.

[32] R. Schmidt. *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. PhD thesis, Stanford University, Stanford, CA, 1981.

[33] J. Krolik. Focussed wide-band array processing for spatial spectral estimation. In S. Haykin, editor, *Advances in Spectrum Analysis and Array Processing*, volume 2, pages 221–261. Prentice Hall, 1991.

[34] H. Wang and M. Kaveh. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33(4):823–831, August 1985.

[35] K. Buckley and L. Griffiths. Broad-band signal-subspace spatial-spectrum (bass-ale) estimation. *IEEE Trans. Acoust., Speech, Signal Processing*, vol.36(7):953–964, July 1988.

[36] J. Van Etten. Navigational systems: Fundamentals of low and very low frequency hyperbolic techniques. *Electrical Commun.*, vol. 45(3):192–212, 1970.

[37] P. Janiczek, editor. *Global Positioning System*. Washington, D.C.:The Institute of Navigation, 1980.

[38] G. Carter. Time delay estimation for passive sonar signal processing. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(3):463–470, June 1981.

[39] R. Schmidt. A new approach to geometry of range difference location. *IEEE Trans. Aerosp. Electron.*, AES-8(6):821–835, November 1972.

[40] J. Delosme, M. Morf, and B. Friedlander. A linear equation approach to locating sources from time-difference-of-arrival measurements. In *Proceedings of ICASSP80*, pages 818–824. IEEE, 1980.

[41] J. Smith and J. Abel. Closed-form least-squares source location estimation from range-difference measurements. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-35(12):1661–1669, December 1987.

[42] J. Smith and J. Abel. The spherical interpolation method for closed-form passive source localization using range difference measurements. In *Proceedings of ICASSP87*. IEEE, 1987.

[43] H. Lee. A novel procedure for accessing the accuracy of hyperbolic multilateration systems. *IEEE Trans. Aerosp. Electron.*, AES-11(1):2–15, January 1975.

[44] N. Marchand. Error distributions of best estimate of position from multiple time difference hyperbolic networks. *IEEE Trans. Aerosp. Navigat. Electron.*, vol.11:96–100, June 1964.

[45] B. Friedlander. A passive localization algorithm and its accuracy analysis. *IEEE Jour. Oceanic Engineering*, OE-12(1):234–245, January 1987.

[46] H. Schau and A. Robinson. Passive source localization employing intersecting spherical surfaces from time-of-arrival differences. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-35(8):1223–1225, August 1987.

[47] B. Fang. Simple solutions for hyperbolic and related position fixes. *IEEE Trans. Aerosp. Electron.*, vol. 26(9):748–753, September 1990.

[48] W. Foy. Position-location solutions by taylor-series estimation. *IEEE Trans. Aerosp. Electron.*, AES-12:187–194, March 1976.

[49] D. Torrieri. Statistical theory of passive location systems. *IEEE Trans. Aerosp. Electron.*, AES-20:183–198, March 1984.

[50] Y. Chan and K. Ho. A simple and efficient estimator for hyperbolic location. *IEEE Trans. Signal Processing*, 42(8):1905–1915, August 1994.

[51] L. Krause. A direct solution to gps-type navigation equations. *IEEE Trans. Aerosp. Electron.*, AES-23(2):225–232, March 1987.

[52] S. Bancroft. An algebraic solution to the gps equations. *IEEE Trans. Aerosp. Electron.*, AES-21(7):56–59, January 1985.

[53] S. Reddi. An exact solution to range computation with time delay information for arbitrary array geometries. *IEEE Trans. Signal Processing*, 41(1):485–486, January 1993.

[54] N. Blachman. On combining target-location ellipses. *IEEE Trans. Aerosp. Electron.*, AES-25(2):284–287, March 1989.

[55] J. Roecker. On combining multidimensional target location ellipsoids. *IEEE Trans. Aerosp. Electron.*, vol. 27(1):175–177, January 1991.

[56] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, second edition, 1984.

[57] H. Van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley & Sons, first edition, 1968.

[58] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Royal Society of London*, vol.A231(9):289–337, 1933.

[59] S. Meyer. *Data Analysis for Scientists and Engineers*. John Wiley and Sons, first edition, 1975.

[60] S. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, first edition, 1993.

[61] D. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, first edition, 1982.

[62] F. Pirz. Design of a wideband, constant beamwidth, array microphone for use in the near field. *Bell System Technical Journal*, 58(8):1839–1850, October 1979.

[63] M. Goodwin and G. Elko. Constant beamwidth beamforming. In *Proceedings of ICASSP93*, pages I–169–I–172. IEEE, 1993.

[64] C. Broydon. The convergence of a class of double-rank minimization algorithms. *J. of the Inst. of Mathematics and its Applic.*, vol.6:76–90, 1970.

[65] R. Fletcher. A new approach to variable metric algorithms. *Computer Journal*, vol.13:317–322, 1970.

[66] D. Goldfarb. A family of variable metric updates derived by variational means. *Mathematics of Computing*, vol.24:23–26, 1970.

[67] D. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computing*, vol.24:647–656, 1970.

[68] J. Nelder and R. Mead. A simplex method for function minimization. *Computer Journal*, vol.7:308–313, 1965.

[69] P. Gill, W. Murray, and M. Wright. *Practical Optimization*. Academic Press, 1981.

[70] J. More. The levenberg-marquardt algorithm: Implementation and theory. In G.A. Watson, editor, *Numerical Analysis*, Lecture Notes in Mathematics, pages 105–116. Springer-Varleg, 1977.

[71] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[72] A. Grace. *Optimization Toolbox Uer's Guide*. The Math Works, Inc., Natick, MA, August 1992.

[73] E. Swokowski. *Calculus with Analytic Geometry*. Prindle, Weber, and Schmidt, second edition, 1979.

[74] J. C. Hassab. *Underwater Signal and Data Processing*. CRC Press, Inc., 1989.

[75] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-24(4):320–327, August 1976.

[76] IEEE Trans. Acoust., Speech, Signal Processing. *Special Issue on Time-Delay Estimation*, volume ASSP-29, June 1981.

[77] W. S. Hodgkiss and L. W. Nolte. Covariance between fourier coefficients representing the time waveforms observed from an array of sensors. *J. Acoust. Soc. Am.*, 59:582–590, March 1976.

[78] S. M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice Hall, Inc., 1993.

[79] 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 Audio and Electroacoustics*, AU-21(4):337–344, August 1973.

[80] Y. T. Chan, R. V. Hattin, and J. B. Plant. The least squares estimation of time delay and its use in signal detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:217–222, 1978.

[81] M. S. Brandstein and H. F. Silverman. A new time-delay estimator for finding source locations using a microphone array. LEMS Technical Report 116, LEMS, Division of Engineering, Brown University, Providence, RI 02912, March 1993.

[82] J. M. Tribolet. A new phase unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25:170–177, 1977.

[83] S. Bédard, B. Champagne, and A. Stéphenne. Effects of room reverberation on time-delay estimation performance. In *Proceedings of ICASSP94*, pages II–261– II–264. IEEE, 1994.

[84] R. Shiavi. *Introduction to Applied Statistical Signal Analysis*. Aksen Associates, first edition, 1991.

[85] J. Gruber. A comparison of measured and calculated speech temporal parameters relevant to speech activity detection. *IEEE Trans.Commun.*, Com-30(4):728–738, April 1982.

[86] P. Brady. A technique for investigating on-off patterns of speech. *Bell Sys. Tech. Jour.*, 44:1–22, January 1965.

[87] Y. Bar-Shalom and T. Fortmann. *Tracking and Data Association*. Academic Press, Inc., first edition, 1988.

[88] Y. Bar-Shalom, editor. *Multitarget-Multisensor Tracking: Advanced Applications*. Artech House, first edition, 1990.

[89] Y. Bar-Shalom. *Estimation and Tracking: Principles, Techniques, and Software*. Artech House, first edition, 1993.

[90] M. Brandstein. Model-based tracking of a single talker. unpublished work, May 1993.

[91] Q. Lin, E. Jan, and J. Flanagan. Microphone arrays and speaker identification. *IEEE Trans. Speech Audio Proc.*, 2(4):622–629, October 1994.

[92] P. Meuse and H. Silverman. Characterization of talker radiation pattern using a microphone array. In *Proceedings of ICASSP94*, pages II–257– II–260. IEEE, 1994.

[93] J. Flanagan. Analog measurements of sound radiation from the mouth. *J. Acoust. Soc. Am.*, Vol.32(12):1613–1620, December 1960.