8 Robust Localization in Reverberant Rooms

Joseph H. DiBiase[1], Harvey F. Silverman[1], and Michael S. Brandstein[2]

[1] Brown University, Providence RI, USA
[2] Harvard University, Cambridge MA, USA

Abstract. Talker localization with microphone arrays has received significant attention lately as a means for the automated tracking of individuals in an enclosure and as a necessary component of any general purpose speech capture system. Several algorithmic approaches are available for speech source localization with multi-channel data. This chapter summarizes the current field and comments on the general merits and shortcomings of each genre. A new localization method is then presented in detail. By utilizing key features of existing methods, this new algorithm is shown to be significantly more robust to acoustical conditions, particularly reverberation effects, than the traditional localization techniques in use today.

8.1 Introduction

The primary goal of a speech localization system is accuracy. In general, estimate precision is dependent upon a number of factors. Major issues include (1) the quantity and quality of microphones employed, (2) microphone placement relative to each other and the speech sources to be analyzed, (3) the ambient noise and reverberation levels, and (4) the number of active sources and their spectral content. The performance of localization techniques generally improves with the number of microphones in the array, particularly when adverse acoustic effects are present. This has spawned the research and construction of large array systems (e.g. 512 elements) [1]. However, when acoustic conditions are favorable and the microphones are positioned judiciously, source localization can be performed adequately using a modest number (e.g. 4 elements) of microphones. Performance is clearly affected by the array geometry. The optimal design of the array based on localization criteria is typically dependent on the room layout, speaking scenarios, and the acoustic conditions [2]. In practice, many of these design considerations are very dependent on the specific application conditions, the hardware available, and non-scientific cost criteria. In an effort to make its applicability as general as possible, this chapter will focus primarily on speech localization effectiveness as a function of the acoustic degradations present, namely background noise and reverberations, rather than attempt to address more specific environmental scenarios.

In addition to high accuracy, these location estimates must be updated frequently in order to be useful in practical tracking and beamforming appli-

M. Brandstein et al. (eds.), *Microphone Arrays*

cations. 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 [3] and to allow for effective channel equalization [4]. A practical 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 and possess a short processing latency to make it practical for real-time systems.

These factors place tight constraints on the microphone data requirements. While the computation time required by the algorithm largely determines the latency of the locator, it is the data requirements that define theoretical limits. The work in [5], for example, focuses on reducing the size of the data segments necessary for accurate source localization in realistic room environments.

The goal of this chapter is to detail the issues associated with the problem of speech source localization in reverberant and noisy rooms and to present an effective methodology for its solution. While the focus will be the single-source scenario, the techniques described, in many cases, are applicable to situations where several individuals are conversing. The more general problem of simultaneous, multi-talker localization is addressed further in Chapter 9. The following section contains a summary of the existing genres for speech source localization using microphone arrays and highlights their relative merits. It is followed in Section 8.3 by the development of a speech source localization algorithm designed specifically for reverberant enclosures which combines two of these general approaches. Section 8.4 then offers some experimental results and conclusions.

8.2 Source Localization Strategies

Existing source localization procedures may be loosely divided into three general categories: those based upon maximizing the steered response power (SRP) of a beamformer, techniques adopting high-resolution spectral estimation concepts, and approaches employing 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 microphones.

8.2.1 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 [6–8]. Theoretical and practical variance bounds obtained via focalization are detailed in [6,7,9] and the steered-beamformer approach has been extended to the case of multiple-signal sources in [10].

The simplest type of steered response is obtained using the output of a delay-and-sum beamformer. This is what is most often referred to as a conventional beamformer. Delay-and-sum beamformers apply time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. More sophisticated beamformers apply filters to the array signals as well as this time alignment. The derivation of the filters in these filter-and-sum beamformers is what distinguishes one method from another.

Beamforming has been used extensively in speech-array applications for voice capture. However, due to the efficiency and satisfactory performance of other methods, it has rarely been applied to the talker localization problem. 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 [10]. 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 [11] 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.

Furthermore, the steered response of a conventional beamformer is highly dependent on the spectral content of the source signal. Many optimal derivations are based on *a priori* knowledge of the spectral content of the background noise, as well as the source signal [7,8]. In the presence of significant reverberation, the noise and source signals are highly correlated, making ac-

curate estimation of the noise infeasible. Furthermore, in nearly all array-applications, little or nothing is known about the source signal. Hence, such optimal estimators are not very practical in realistic speech-array environments.

The practical shortcomings of applying correlation-based localization estimation techniques without a great deal of intelligent pruning is typified by the system produced in [12]. 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 multi-rate 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.

8.2.2 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 [13,14]. 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.

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 [15], the AR model and certain eigenanalysis approaches are limited to the far-field, uniform linear array situation.

With regard to 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 sit-

uation 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 [16–18]. Either of these routes extends the computational requirements considerably.

These algorithms tend to be significantly less robust to source and sensor modeling errors than conventional beamforming methods [19,20]. 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 the reverberation conditions of primary concern here, 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[21]. Primarily for these reasons, localization methods based upon these high-resolution strategies will not considered further in this work. However, this should not exclude their judicious use in other speech localization contexts, particularly multi-source scenarios.

8.2.3 TDOA-Based Locators

With this third localization strategy, a two-step procedure is adopted. Time delay estimation (TDE) of the speech signals relative to pairs of spatially separated microphones is performed. This data along with knowledge of the microphone positions are then 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, [22–28] are examples. 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. Primarily because of their computational practicality and reasonable performance under amicable conditions, the bulk of passive talker localization systems in use today are TDOA-based.

Accurate and robust TDE is the key to the effectiveness of localizers within this genre. The two major sources of signal degradation which complicate this estimation problem are background noise and channel multi-path due to room reverberations. The noise-alone case has been addressed at length and is well understood. Assuming uncorrelated, stationary Gaussian signal and noise sources with known statistics and no multi-path, the ML time-delay estimate is derived from a SNR-weighted version of the Generalized Cross-Correlation (GCC) function [29]. An ML-type weighting appropriate for non-stationary speech sources was presented in [30] and applied successfully to

speech source localization in low-multipath environments [31]. However, once room reverberations rise above minimal levels, these methods begin to exhibit dramatic performance degradations and become unreliable [32,33]. A basic approach to dealing with multi-path channel distortions in this context has been to make the GCC function more robust by deemphasizing the frequency-dependent weightings. The Phase Transform (PHAT) [29] is one extreme of this procedure which has received considerable attention recently as the basis of speech source localization systems [34–36]. By placing equal emphasis on each component of the cross-spectrum phase, the resulting peak in the GCC-PHAT function corresponds to the dominant delay in the reverberated signal. While effective at reducing some of the degradations due to multi-path, the Phase Transform accentuates components of the spectrum with poor SNR and has the potential to provide poor results, particularly under low reverberation, high noise conditions.

Other approaches for TDE of talkers in adverse environments are available. A procedure which utilizes a speech specific criterion in the design of the GCC weighting function is presented in [37]. Cepstral prefiltering [38] has been used to deconvolve the effects of reverberation prior to applying GCC. However, deconvolution requires long data segments since the duration of a typical small-room impulse response is 200-400 ms. It is also very sensitive to the high variability and non-stationarity of speech signals. In fact, the experiments performed in [38] avoided the use of speech as input altogether. Instead, colored Gaussian noise was used as the source signal. While identification of room impulse responses is extremely problematic when the source signal is unknown, the method proposed in [24], which is based on eigenvalue decomposition, efficiently detects the direct paths of the two impulse responses. This method is effective with speech as input, but requires 250 ms of microphone data to converge. A short-time TDE method, which is more complex than GCC, is presented in [33]. It involves the minimization of a weighted least-squares function of the phase data. It was shown to outperform both GCC-ML and GCC-PHAT in reverberant conditions. However, this improvement comes at the cost of a complicated searching algorithm. The marginal improvement over GCC-PHAT may not justify this added cost in computational complexity. Reverberation effects can also be overcome to some degree by classifying TDE's acquired over time and associating them with the direction of arrival (DOA) of the sound waves [39]. This approach, however, is not suitable for short-time TDE. Under extreme acoustic conditions, a large percentage of the TDE's are anomalous, and it takes a considerable period (1-2 s in [39]) to acquire enough estimates for a statistically meaningful classification.

Among the methods summarized above, those that rely on long data segments generally outperform those that do not. This result may be attributed to the ensemble averaging performed under these conditions to improve the quality of the underlying signal statistics. However, the dynamic environ-

ments of many speech array applications require high update rates, which limit the duration of the data segments used for analysis. For example, the automatic camera steering video-conferencing system detailed in [34] utilizes a TDOA-based method with GCC-PHAT TDE applied at update rates of 200-300 ms. With such long data segments, reliable estimates are produced, even in moderately adverse acoustic conditions. However, applications such as adaptive beamforming and the tracking of multiple talkers using a TDOA-based localizer require an appreciably higher estimate rate; source positions must be acquired from independent data segments as short as 20-30 ms. Over such limited durations, the lack of ensemble averaging has a severe impact on the performance of the TDE.

Given a set of TDOA figures with known error statistics, the second step of obtaining the ML location estimate necessitates solving a set of nonlinear equations. The calculation of this result is considerably less computationally expensive than that required for estimators belonging to the two previously discussed genres. There is an extensive class of sub-optimal, closed-form location estimators. 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. [22,25–28,40,41] are typical of these methods. 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.

TDOA-based locators do present several disadvantages when used as the basis of a general localization scheme. Their primary limitation is the inability to accommodate multi-source scenarios. These algorithms assume a single-source model. While TDOA-based methods with short analysis intervals may be used to track several individuals in a conversational situation [31,42], the presence of multiple simultaneous talkers, excessive ambient noise, or moderate to high reverberation levels in the acoustic field typically results in poor TDOA figures and subsequently, 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. These shortcomings may be overcome to some degree through judicious use of appropriate detection methods at each stage in the process [31].

While practical, the application of TDOA-based localization procedures is of limited utility in realistic, acoustic environments. Steered-Beamformer strategies are computationally more intensive, but tend to possess a robustness advantage and require a shorter analysis interval. The two-stage process requiring time-delay estimation prior to the actual location evaluation is suboptimal. The intermediate signal parameterization accomplished by the TDOA estimation 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.

With these relative advantages and shortcomings in mind, a new localization method, which combines the best features of the steered-beamformer with those of the Phase Transform weighting of the GCC, was introduced in [5]. The goal was to exploit the inherent robustness and short-time analysis characteristics of the steered response power approach with the insensitivity to signal conditions afforded by the Phase Transform. This new algorithm, termed SRP-PHAT, will be detailed in the following section and will be shown to produce highly reliable location estimates in rooms with reverberation times up to 200 ms, using independent 25 ms data segments.

8.3 A Robust Localization Algorithm

Before describing the SRP-PHAT algorithm, it will be necessary to develop further a number of topics addressed in the prior section. Specifically, the following subsections will provide details of the impulse response model, the GCC and its PHAT implementation, ML TDOA-based localization, and the computation of the SRP. These items will then be tied together in the final subsection to motivate and define the SRP-PHAT algorithm.

8.3.1 The Impulse Response Model

It will be assumed that sound waves propagate as predicted by the linear wave equation [43]. With this assumption, the acoustic paths between sound sources and microphones can be modeled as linear systems [44]. This is clearly advantageous to the analysis and modeling of the signals produced by the microphones of an array. Such linear models are valid under the realistic conditions encountered in small-room speech-array environments and are regularly exploited by array-processing techniques [13].

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

The wave field at a particular location inside a reverberant room may be considered to be linearly related to the source signal, $s(t)$. Let the 3-element vectors, $\boldsymbol{p}_n$ and $\boldsymbol{q}_s$, define the Cartesian coordinates of the n^{th} microphone

and the source, respectively. The received signal at the n^{th} microphone may now be expressed as

$$x_n(t) = s(t) \star h_n(\boldsymbol{q}_s, t) + v_n(t) \tag{8.1}$$

The overall impulse response, $h_n(\boldsymbol{q}_s, t)$, is the result of cascading two filters: the room impulse response and the microphone channel response. The former characterizes all acoustic paths between the source and microphone locations, including the direct path. It is a function of $\boldsymbol{p}_n$ as well as the source location, $\boldsymbol{q}_s$, and is highly dependent on these parameters. In general, the room impulse response is affected by environmental conditions, such as temperature and humidity. It also varies with the movement of furniture and individuals inside the room. While such variations are significant, it is reasonable to assume that these factors remain constant over short periods. Hence, a room impulse response may be considered time-invariant for short periods when the source and microphone are spatially fixed. The microphone channel response accounts for the electrical, mechanical and acoustical properties of the microphone system. In general, the microphone's directivity pattern makes its response a function of its orientation as well as its spatial placement relative to the source. The additive term, $v_n(t)$, is the result of channel noise in the microphone system and any propagating ambient noise such as that due to fans or other mechanical equipment. The propagating noise is usually more significant than the channel noise and tends to dominate this term. Generally, $v_n(t)$ is assumed to be uncorrelated with $s(t)$.

Figure 8.1 illustrates a close-up view of the response that was measured in a typical conference room. The direct-path component and some of the strong reflected components are highlighted in this plot. The peaks corresponding to the reflected sound waves are comparable in size to the direct-path peak. These peaks, which occur within 20 ms of the direct-path, are responsible for many of the erroneous results produced by short-time TDE's, which operate on data blocks as small as 25 ms. The large secondary peaks in the room response are highly correlated with the false peaks in the GCC function [5].

The purpose of TDE is to evaluate the temporal disparity between the direct-path components in the two received microphone signals. To this end, it will be useful to rewrite the impulse response specifically in terms of its direct-path component. Equation 8.1 is modified to:

$$x_n(t) = \frac{1}{r_n} s(t - \tau_n) \star g_n(\boldsymbol{q}_s, t) + v_n(t) \tag{8.2}$$

where r_n is the source-microphone separation distance, τ_n is the direct path time delay, and $g_n(\boldsymbol{q}_s, t)$ is the modified impulse response which encompasses the original response minus the direct path component. The microphone signal model is now expressed explicitly in terms of the parameter of interest, namely the time delay, τ_n.

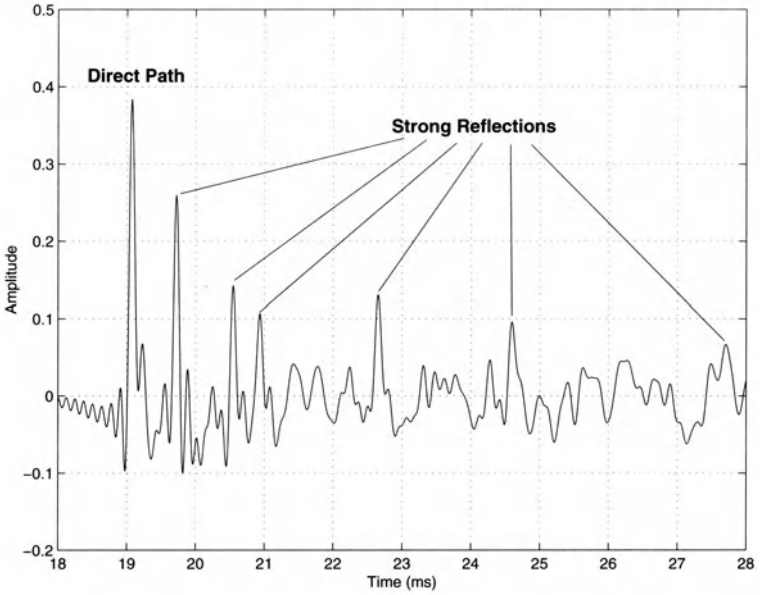

Fig. 8.1. A close-up of a 10-millisecond segment of a room impulse response measured in a typical conference room. The direct-path component and some strong reflected components are highlighted.

8.3.2 The GCC and PHAT Weighting Function

For a pair of microphones, $n = 1, 2$, their associated TDOA, τ_{12}, is defined as

$$\tau_{12} \equiv \tau_2 - \tau_1. \tag{8.3}$$

Applying this definition to their associated received microphone signal models yields

$$\begin{aligned} x_1(t) &= \frac{1}{r_1} s(t - \tau_1) \star g_1(\boldsymbol{q}_s, t) + v_1(t) \\ x_2(t) &= \frac{1}{r_2} s(t - \tau_1 - \tau_{12}) \star g_2(\boldsymbol{q}_s, t) + v_2(t). \end{aligned} \tag{8.4}$$

If the modified impulse responses for the microphone pair are similar, then (8.4) shows that a scaled version of $s(t - \tau_1)$ is present in the signal from microphone 1 and a time-shifted (and scaled) version of $s(t - \tau_1)$ is present in the signal from microphone 2. The cross-correlation of the two signals should show a peak at the time lag where the shifted versions of $s(t)$ align, corresponding to the TDOA, τ_{12}. The cross correlation of signals and is defined as:

$$c_{12}(\tau) = \int_{-\infty}^{+\infty} x_1(t) x_2(t + \tau) dt \tag{8.5}$$

The GCC function, $R_{12}(\tau)$, is defined as the cross correlation of two filtered versions of $x_1(t)$ and $x_2(t)$ [29]. With the Fourier transforms of these filters denoted by $G_1(\omega)$ and $G_2(\omega)$, respectively, the GCC function can be expressed in terms of the Fourier transforms of the microphone signals

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\big(G_1(\omega)X_1(\omega)\big)\big(G_2(\omega)X_2(\omega)\big)^* e^{j\omega\tau}d\omega \tag{8.6}$$

Rearranging the order of the signals and filters and defining the frequency dependent weighting function, $\Psi_{12} \equiv G_1(\omega)G_2(\omega)^*$, the GCC function can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Psi_{12}(\omega)X_1(\omega)X_2(\omega)^* e^{j\omega\tau}d\omega \tag{8.7}$$

Ideally, $R_{12}(\tau)$ will exhibit an explicit global maximum at the lag value which corresponds to the relative delay. The TDOA estimate is calculated from

$$\hat{\tau}_{12} = \underset{\tau\in D}{\operatorname{argmax}}\, R_{12}(\tau). \tag{8.8}$$

The range of potential TDOA values is restricted to a finite interval, D, which is determined by the physical separation between the microphones. In general, $R_{12}(\tau)$ will have multiple local maxima which may obscure the true TDOA peak and subsequently, produce an incorrect estimate. The amplitudes and corresponding time lags of these erroneous maxima depend on a number of factors, typically ambient noise levels and reverberation conditions.

The goal of the weighting function, Ψ_{12}, is to emphasize the GCC value at the true TDOA value over the undesired local extrema. A number of such functions have been investigated. As previously stated, for realsitic acoustical conditions the PHAT weighting [29] defined by

$$\Psi_{12}(\omega) \equiv \frac{1}{|X_1(\omega)X_2^*(\omega)|} \tag{8.9}$$

has been found to perform considerably better than its counterparts designed to be statistically optimal under specific non-reverberant, noise conditions. The PHAT weighting whitens the microphone signals to equally emphasize all frequencies. The utility of this strategy and its extension to steered-beamforming form the basis of the SRP-PHAT algorithm that follows.

8.3.3 ML TDOA-Based Source Localization

Consider the i^{th} pair of microphones with spatial coordinates denoted by the 3-element vectors, $\boldsymbol{p}_{i1}$ and $\boldsymbol{p}_{i2}$, respectively. For a signal source with known

spatial location, $\boldsymbol{q}_s$, the true TDOA relative to the i^{th} sensor pair will be denoted by $T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$, and is calculated from the expression

$$T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s) = \frac{|\boldsymbol{q}_s - \boldsymbol{p}_{i2}| - |\boldsymbol{q}_s - \boldsymbol{p}_{i1}|}{c} \tag{8.10}$$

where c is the speed of sound in air. The estimate of this true TDOA, the result of a TDE procedure involving the signals received at the two microphones, will be given by $\hat{\tau}_i$. In practice, the TDOA estimate is a corrupted version of the true TDOA and in general, $\hat{\tau}_i \neq T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}_s)$.

For a single microphone pair and its TDOA estimate, the locus of potential source locations in 3-space which satisfy (8.10) corresponds to one-half of a hyperboloid of two sheets. This hyperboloid is centered about the midpoint of the microphones and has $\boldsymbol{p}_{i2} - \boldsymbol{p}_{i1}$ as its axis of symmetry.

For sources with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry. The corresponding estimated direction angle, $\hat{\theta}_i$, for the microphone pair is given by:

$$\hat{\theta}_i = \cos^{-1}\left(\frac{c \cdot \hat{\tau}_i}{|\mathbf{m}_{i1} - \mathbf{m}_{i2}|}\right) \tag{8.11}$$

In this manner each microphone pair and TDOA estimate combination may be associated with a single parameter which specifies the angle of the cone relative to the sensor pair axis. For a given source and TDOA estimate, $\hat{\theta}_i$ is referred to as the DOA relative to the i^{th} pair of microphones.

Given a set of M TDOA estimates derived from the signals received at multiple pairs of microphones, the problem remains as how to best estimate the true source location, $\boldsymbol{q}_s$. Ideally, the estimate will be an element of the intersection of all the potential source loci. 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, the source location must be estimated as the point in space 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. If the time-delay estimates at each microphone pair are assumed to be independently corrupted by zero-mean additive white Gaussian noise of equal variance then the ML location estimate can be shown to be the position which minimizes the least squares error criterion

$$E(\boldsymbol{q}) = \sum_{i=1}^{M} (\hat{\tau}_i - T(\{\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}\}, \boldsymbol{q}))^2. \tag{8.12}$$

The location estimate is then found from

$$\hat{\boldsymbol{q}}_s = \underset{\boldsymbol{q}}{\operatorname{argmin}}\, E(\boldsymbol{q}). \tag{8.13}$$

The criterion in (8.12) will be referred to as the LS-TDOA error. As stated earlier, the evaluation of $\hat{\boldsymbol{q}}_s$ in this manner involves the optimization of a non-linear function and necessitates the use of search methods. Closed-form approximations to this method were given earlier.

8.3.4 SRP-Based Source Localization

The microphone signal model in (8.2) shows that for an array of N microphones in the reception region of a source, a delayed, filtered, and noise corrupted version of the source signal, $s(t)$, is present in each of the received microphone signals. The delay-and-sum beamformer time aligns and sums together the $x_n(t)$, in an effort to preserve unmodified the signal from a given spatial location while attenuating to some degree the noise and convolutional components. It is defined as simply as

$$y(t, \boldsymbol{q}_s) = \sum_{n=1}^{N} x_n(t + \Delta_n) \tag{8.14}$$

where Δ_n are the *steering delays* appropriate for focusing the array to the source spatial location, $\boldsymbol{q}_s$, and compensating for the direct path propagation delay associated with the desired signal at each microphone. In practice, the delays relative to a reference microphone are used instead of the absolute delays. This makes all shifting operations causal, which is a requirement of any practical system, and implies that $y(t, \boldsymbol{q}_s)$ will contain an overall delayed version of the desired signal which in practice is not detrimental. The use of a single reference microphone means that the steering delays may be determined directly from the TDOA's (estimated or theoretical) between each microphone and the reference. This implies that knowledge of the TDOA's alone is sufficient for steering the beamformer without an explicit source location.

In the most ideal case with no additive noise and channel effects, the output of the deal-and-sum beamformer represents a scaled and potentially delayed version of the desired signal. For the limited case of additive, uncorrelated, and uniform variance noise and equal source-microphone distances this simple beamformer is optimal. These are certainly very restrictive conditions. In practice, convolutional channel effects are nontrivial and the additive noise is more complicated. The degree to which these noise and reverberation components of the microphone signals are suppressed by the delay-and-sum beamformer is frequently minimal and difficult to analyze. Other methods have been developed to extend the delay-and-sum concept to the more general filter-and-sum approach, which applies adaptive filtering to the microphone

signals before they are time-aligned and summed. Again, these methods tend to not be robust to non-theoretical conditions, particularly with regard to the channel effects.

The output of an N-element, filter-and-sum beamformer can be defined in the frequency domain as

$$Y(\omega, \boldsymbol{q}) = \sum_{n=1}^{N} G_n(\omega) X_n(\omega) e^{j\omega \Delta_n} \tag{8.15}$$

where $X_n(\omega)$ and $G_n(\omega)$ are the Fourier Transforms of the n^{th} microphone signal and its associated filter, respectively. The microphone signals are phase-aligned by the steering delays appropriate for the source location, $\boldsymbol{q}$. This is equivalent to the time-domain beamformer version. The addition of microphone and frequency-dependent filtering allows for some means to compensate for the environmental and channel effects. Choosing the appropriate filters depends on a number of factors, including the nature of the source signal and the type of noise and reverberations present. As will be seen, the strategy used by the PHAT of weighing each frequency component equally will prove advantageous for practical situations where the ideal filters are unobtainable.

The beamformer may be used as a means for source localization by steering the array to specific spatial points of interest in some fashion and evaluating the output signal, typically its power. When the focus corresponds to the location of the sound source, the SRP should reach a global maximum. In practice, peaks are produced at a number of incorrect locations as well. These may be due to strong reflective sources or merely a byproduct of the array geometry and signal conditions. In some cases, these extraneous maxima in the SRP space may obscure the true location and in any case, complicate the search for the global peak. The SRP for a potential source location can be expressed as the output power of a filter-and-sum beamformer by

$$P(\boldsymbol{q}) = \int_{-\infty}^{+\infty} |Y(\omega)|^2 d\omega \tag{8.16}$$

and location estimate is found from

$$\hat{\boldsymbol{q}}_s = \underset{\boldsymbol{q}}{\operatorname{argmax}}\, P(\boldsymbol{q}). \tag{8.17}$$

8.3.5 The SRP-PHAT Algorithm

Given this background, the SRP-PHAT algorithm may now be defined. With respect to GCC-based TDE, the PHAT weighting has been found to provide an enhanced robustness in low to moderate reverberation conditions. While improving the quality of the underlying delay estimates, it is still not sufficient to render TDOA-based localization effective under more adverse conditions.

The delay-and-sum SRP approach requires shorter analysis intervals and exhibits an elevated insensitivity to environmental conditions, though again, not to a degree that allows for their use under excessive multi-path. The filter-and-sum version of the SRP adds flexibility but the design of the filters is typically geared towards optimizing SNR in noise-only conditions and is excessively dependent on knowledge of the signal and channel content. Originally introduced in [5], the goal of the SRP-PHAT algorithm is to combine the advantages of the steered beamformer for source localization with the signal and condition independent robustness offered by the PHAT weighting.

The SRP of the filter-and-sum beamformer can be expressed as

$$P(\boldsymbol{q}) = \sum_{l=1}^{N}\sum_{k=1}^{N}\int_{-\infty}^{\infty}\Psi_{lk}(\omega)X_l(\omega)X_k^*(\omega)e^{j\omega(\Delta_k-\Delta_l)}d\omega \tag{8.18}$$

where $\Psi_{lk}(\omega) = G_l(\omega)G_k^*(\omega)$ is analogous to the two-channel GCC weighting term in (8.7). The corresponding multi-channel version of the PHAT weighting is given by

$$\Psi_{lk}(\omega) = \frac{1}{|X_l(\omega)X_k^*(\omega)|} \tag{8.19}$$

which in the context of the filter-and-sum beamformer (8.15) is equivalent to the use of the individual channel filters

$$G_n(\omega) = \frac{1}{|X_n(\omega)|}. \tag{8.20}$$

These are the desired SRP-PHAT filters. They may be implemented from the frequency-domain expression above. Alternatively, it may be shown that (8.18) is equivalent to the sum of the GCC's of all possible N-choose-2 microphone pairings. This means that the SRP of a 2-element array is equivalent to the GCC of those two microphones. Hence, as the number of microphones is increased, SRP naturally extends the GCC method from a pairwise to a multi-microphone technique. Denoting $R_{lk}(\tau)$ as the PHAT-weighted GCC of the l^{th} and k^{th} microphone signals, a time-domain version of SRP-PHAT functional can now be expressed as

$$P(\boldsymbol{q}) = 2\pi\sum_{l=1}^{N}\sum_{k=1}^{N}R_{lk}(\Delta_k - \Delta_l). \tag{8.21}$$

This is the sum of all possible pairwise GCC permutations which are time-shifted by the differences in the steering delays. Included in this summation is the sum of the N autocorrelations, which is the GCC evaluated at a lag of zero. These terms contribute only a DC offset to the steered response power since they are independent of the steering delays.

Given either method of computation, SRP-PHAT localization is performed in a manner similar to the standard SRP-based approaches. Namely,

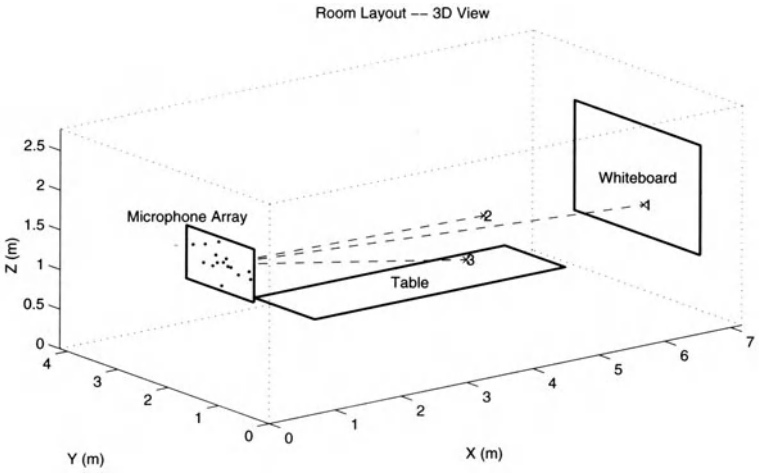

Fig. 8.2. Conference room layout.

$P(\boldsymbol{q})$ is maximized over a region of potential source locations. As will be shown in the next section, relative to the search space indicative of the standard SRP approach, the SRP-PHAT functional significantly deemphasizes extraneous peaks and dramatically sharpens the resolution of the true peak. These desirable features result in a decreased sensitivity to noise and reverberations and more precise location estimates than the existing localization methods offer. Additionally, this is achieved using a very short analysis interval.

8.4 Experimental Comparison

While more extensive results are available in [5], an experiment is offered here to evaluate and compare the relative characteristics and performance of three different source locators: SRP, SRP-PHAT and ML-TDOA. Five second recordings were made for three source locations in a 7 by 4 by 3 m conference room at Brown University using a 15-element microphone array. Figure 8.2 illustrates the room layout. Pre-recorded speech, which was acquired using a close-talking microphone, was played through a loudspeaker while simultaneously recording the signals from the array. The use of the loudspeaker was preferable to an actual talker since the loudspeaker could be precisely located and would be fixed over the duration of the recordings. The talkers were males uttering a unique string of alpha-digits. Source 1 was most distant from the array and was positioned at standing height in front of a white-board. The other two sources were positioned at a seated level

around a conference table, which was located approximately in the center of the room.

The microphone array was composed of eight omni-directional electret condenser microphones, which were randomly distributed on a plane within a .33 by 0.36 m rectangle. The microphones were attached to a rectangular sheet of acoustic foam, which was supported by an aluminum frame. This frame was mounted on a tripod that was placed parallel to the back wall at a distance of 0.9 m. The acoustic foam damps some of the multi-path reflections from this wall and isolates the microphones from vibrations traveling along the mountings.

The loudspeaker faced the array and the volume level was adjusted at each location to maximize SNR conditions. SNR levels at each microphone averaged about 25 dB for the three source locations. Source 3, with its location the closest to the microphone array, had SNRs as high as 36 dB. With such high SNRs, all microphones signals in the conference room dataset have minimal contributions from the background noise, which was primarily produced by the fans inside the computer equipment.

The measured reverberation time of the room was determined to be 200 ms. This qualifies as a mildly reverberant room. However, the near-end peaks in the impulse responses (as in Figure 8.1) combined with a 200 ms reverberation time do, in fact, have a significant impact on localization. This will be demonstrated by the following performance comparisons.

Given the size of the array aperture relative to the source ranges, all three talkers can be considered to lie in the far field of array. Under such conditions, range estimates are ambiguous, and only the azimuth and elevation angles can be estimated reliably. Accordingly, this experiment will focus on DOA measures as opposed to 3-D Cartesian coordinates. Results obtained with more extensive arrays and near-field sources are available in [5].

The recorded data was segmented into 25 ms frames using a half-overlapping Hanning window. SNR-based speech detection was performed for each frame. All frames where any of the eight microphone channels had SNR within 12dB of the background noise were eliminated. Out of the 399 frames per recording, 313, 340, and 297 were retained for sources 1,2, and 3, respectively. The DOA's of the sources were estimated by minimization of the LS-TDOA error and maximization of SRP and SRP-PHAT evaluated over azimuth and elevation relative to the array's origin. The frequency range used to compute both the steered responses and the GCC's was 300 Hz to 8 kHz. These functions were computed over a range of $-60°$ to $+60°$ for both azimuth and elevation with a $0.1°$ resolution.

By taking all possible combinations, 28 microphone pairs were formed using the 8-element array. Hence, for each data frame, 28 TDOA estimates were made for each of the three speech recordings using GCC-PHAT. Figure 8.3 illustrates the LS-TDOA error as a function of azimuth and elevation for a segment of nine successive frames recorded for source 1. The white point in

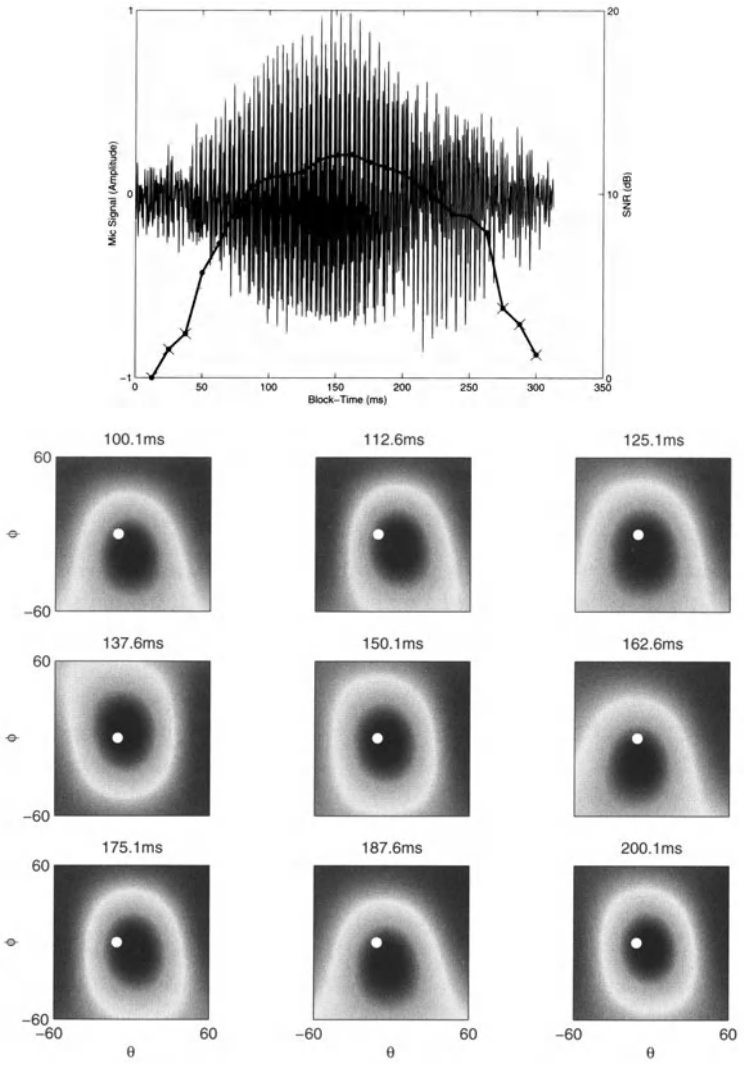

Fig. 8.3. Speech segment (top) with nine frames of the LS-TDOA error surfaces.

each contour plot marks the true DOA. The dark area in the center of the images represents the minima of the LS-TDOA error. At the top of this figure is a plot of the amplitude of the corresponding speech segment, which is the letter "R", spoken as in "Are we there yet?" Superimposed on this speech signal is a curve representing the average power of the signals from the array, with the scale of its vertical axis labeled on the right side of the graph. Each point along this power curve corresponds to the average frame SNR. The three frames at the beginning and end of this speech segment

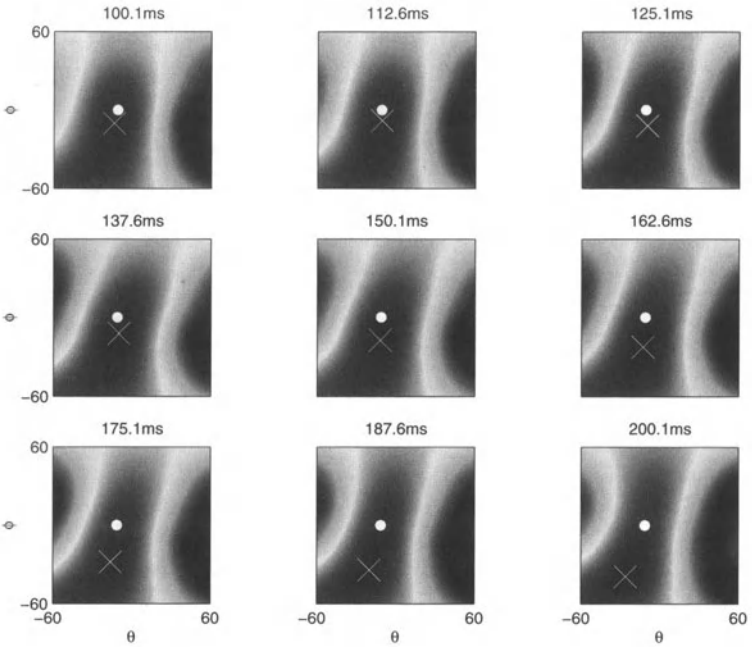

Fig. 8.4. Delay-and-sum beamformer SRP over nine, 25 ms frames.

lacked sufficient SNR to included in the analysis. These plots show that the LS-TDOA error is generally a smooth surface with a global minimum over the angular range of $\pm 60°$. However, from frame to frame the minima vary from the true source location. This inaccuracy is caused by erroneous TDOA estimates. Note also that because of the smooth nature of the error space, the resolution of the DOA estimates is considerably limited.

Figures 8.4 and 8.5 illustrate the error spaces of the SRP and SRP-PHAT as evaluated for the same nine 25 ms frames of speech. Relative to the prior figure the contour images are now inverted in darkness to emphasize the maxima. The plots of the delay-and-sum beamformer SRP in Figure 8.4 bear a noticeable similarity in general shape to their LS-TDOA counterparts. The maximum value in each SRP image, marked by an X, occurs at points distant from the actual DOA, indicated by a white dot. The main beam of the delay-and-sum beamformer is broad and fluctuates considerably over the duration of the speech segment. As a result, many inaccurate location estimates are produced by this method. In contrast to the LS-TDOA and SRP cases, the peaks of SRP-PHAT plots in Figure 8.5 match the actual DOA almost exactly. The main beam of the PHAT beamformer is sharp and consistent over each frame. This produces contour images which appear quite different from the LS-TDOA and SRP versions. The PHAT filters, when applied to the filter-and-sum beamformer, yield an error space that is superior to that of

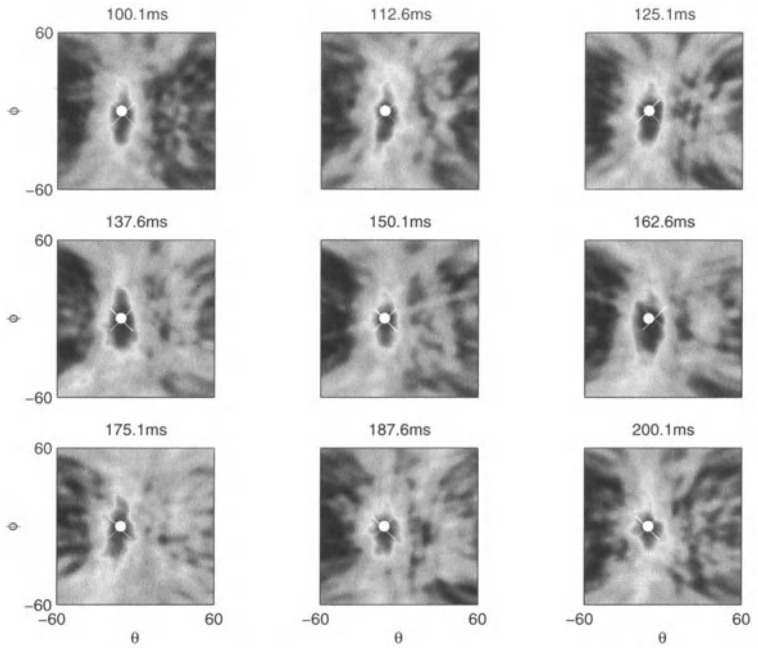

Fig. 8.5. SRP-PHAT response over nine, 25 ms frames.

the delay-and-sum beamformer or the TDOA-based criterion. This qualitative observation will now be corroborated through a numerical performance comparison.

For the DOA estimates produced for each of the three source locations, an RMS DOA error was computed from

$$E_{\text{RMS}}(\hat{\theta}, \hat{\phi}) = \sqrt{(\hat{\theta} - \theta)^2 + (\hat{\phi} - \phi)^2} \tag{8.22}$$

where ϕ and θ are the true azimuth and elevation angles and $\hat{\phi}$ and $\hat{\theta}$ are their estimated counterparts. Figure 8.6 illustrates the results. These plots show the fraction of DOA estimates in each case which exceed a given RMS error threshold. Using this metric, the SRP-PHAT consistently outperforms the other two methods for each of the source locations. The ML-TDOA exhibits definite advantages over the SRP. While the SRP-PHAT's results are nearly identical for all the source locations, including the most distant source 1, the ML-TDOA locator is highly dependent on source location. For example, 60% percent of the estimates from source 1 had error greater than 10° while 50% percent from source 2 and 15% percent from source 3 had error greater 10°. In contrast, nearly all the estimates produced by SRP-PHAT had error less than 10°. About 90% of the estimates from sources 2 and 3, and 80% from source 1 had errors less than 4°.

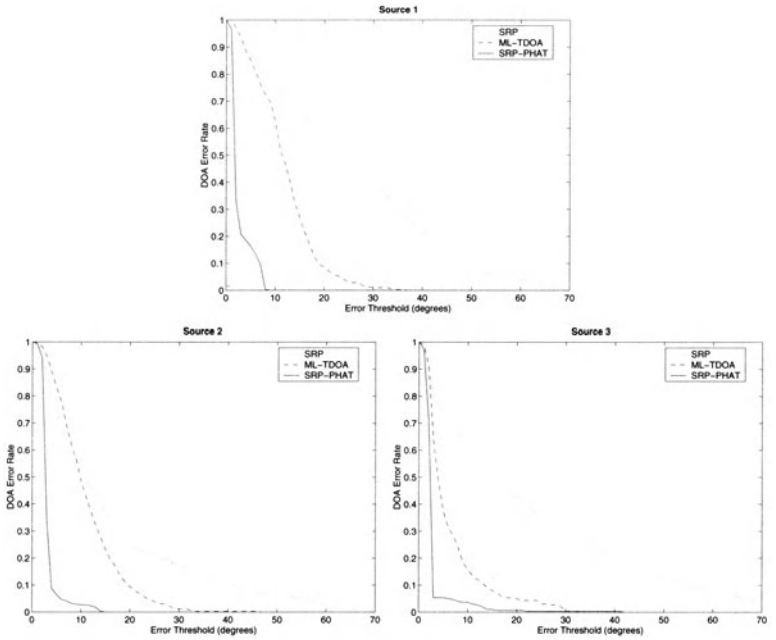

Fig. 8.6. Localizer DOA error rates for three different sources.

The results of this limited experiment illustrate the performance advantages of the SRP-PHAT localizer relative to more traditional approaches for talker localzation with microphone arrays. Other experiments conducted under more general and adverse conditions are consistent with the results here and serve to confirm the utility of combining steered-beamforming and a uniform-magnitude spectral weighting for this purpose.

While the TDOA-based localization method performed satisfactorily for a talker relatively close to the array, it was severely impacted by even the mild reverberation levels encountered when the source was more distant. This result is due to the fact that signal-to-reverberation ratios decrease with increasing source-to-microphone distance. As the reverberation component of the received signal increases relative to the direct path component, the validity of the single-source model inherent in the TDE development is no longer valid. As a result TDOA-based schemes rapidly exhibit poor performance as the talker moves away from the microphones. The SRP-PHAT algorithm is relatively insensitive to this effect. As the results here suggest the proposed algorithm exhibits no marked performance degradation from the near to distant source conditions tested.

The SRP-PHAT algorithm is computationally more demanding than the TDOA-based localization methods. However, its significantly superior performance may easily warrant the additional processing expense. Additionally,

while not discussed here, it is possible to alter the algorithm to dramatically reduce its computational load while maintaining much of its benefit.

References

1. H. Silverman, W. Patterson, J. Flanagan, and D. Rabinkin, "A digital processing system for source location and sound capture by large microphone arrays," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 251–254, April 1997.
2. M. Brandstein, J. Adcock, and H. Silverman, "Microphone array localization error estimation with application to sensor placement," *J. Acoust. Soc. Am.*, vol. 99, no. 6, pp. 3807–3816, 1996.
3. J. Flanagan and H. Silverman, eds., *International Workshop on Microphone-Array Systems: Theory and Practice*, Brown University, Providence RI, USA, October 1992.
4. B. Radlovic, R. Williamson, and R. Kennedy, "On the poor robustness of sound equalization in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 881–884, March 1999.
5. J. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments*, PhD thesis, Brown University, Providence RI, USA, May 2000.
6. W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," in *Signal Processing* (J. Griffiths, P. Stocklin, and C. V. Schooneveld, eds.), pp. 577–590, Academic Press, 1973.
7. G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Am.*, vol. .62, pp. 922–926, October 1977.
8. W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inform Theory*, vol. IT-19, pp. 608–614, September 1973.
9. W. Hahn, "Optimum signal processing for passive sonar range and bearing estimation," *J. Acoust. Soc. Am.*, vol. 58, pp. 201–207, July 1975.
10. M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 1210–1217, October 1983.
11. 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, USA, May 1990.
12. H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone-array data," *Computer, Speech, and Language*, vol. 6, pp. 129–152, April 1992.
13. D. Johnson and D. Dudgeon, *Array Signal Processing- Concepts and Techniques*, Prentice Hall, 1993.
14. S. Haykin, *Adaptive Filter Theory*, Prentice Hall, second ed., 1991.
15. R. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*, PhD thesis, Stanford University, Stanford CA, USA, 1981.
16. J. Krolik, "Focussed wide-band array processing for spatial spectral estimation," in *Advances in Spectrum Analysis and Array Processing* (S. Haykin, ed.), vol. 2, pp. 221–261, Prentice Hall, 1991.

17. 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*, vol. ASSP-33, pp. 823–831, August 1985.
18. K. Buckley and L. Griffiths, "Broad-band signal-subspace spatial-spectrum (BASS-ALE) estimation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. .36, pp. 953–964, July 1988.
19. A. Vural, "Effects of pertubations on the performance of optimum/adaptive arrays," *IEEE Trans. Aerosp. Electron.*, vol. AES-15, pp. 76–87, January 1979.
20. R. Compton Jr., *Adaptive Antennas*, Prentice Hall, 1988.
21. T. Shan, M. Wax, and T. Kailath, "On spatial smoothing for direction-of-arrival estimation in coherent signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 806–811, August 1985.
22. M. Brandstein, J. Adcock, and H. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 45–50, January 1997.
23. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
24. Y. Huang, J. Benesty, and G. W. Elko, "Adaptive eigenvalue decomposition algorithm for realtime acoustic source localization system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 937–940, March 1999.
25. R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron.*, vol. AES-8, pp. 821–835, November 1972.
26. J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1661–1669, December 1987.
27. H. Lee, "A novel procedure for accessing the accuracy of hyperbolic multilateration systems," *IEEE Trans. Aerosp. Electron.*, vol. AES-11, pp. 2–15, January 1975.
28. N. Marchand, "Error distributions of best estimate of position from multiple time difference hyperbolic networks," *IEEE Trans. Aerosp. Navigat. Electron.*, vol. .11, pp. 96–100, June 1964.
29. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-24, pp. 320–327, August 1976.
30. M. Brandstein, J. Adcock, and H. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *Computer, Speech, and Language*, vol. 9, pp. 153–169, April 1995.
31. M. Brandstein and H. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer, Speech, and Language*, vol. 11, pp. 91–126, April 1997.
32. S. Bédard, B. Champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-94)*, Adelaide, Australia, pp. II:261–264, April 1994.
33. M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 375–378, April 1997.

34. H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 187–190, April 1997.
35. M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event localization," *IEEE Trans. Speech Audio Proc.*, vol. 5, pp. 288–292, May 1997.
36. P. Svaizer, M. Matassoni, and M. Omologo, "Acoustic source location in a three-dimensional space using crosspower spectrum phase," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 231–234, April 1997.
37. M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.
38. A. Stéphenne and B. Champagne, "Cepstral prefiltering for time delay estimation in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-95)*, Detroit MI, USA, pp. 3055–3058, May 1995.
39. N. Strobel and R. Rabenstein, "Classification of time delay estimates in reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-99)*, Phoenix AZ, USA, pp. 3081–3084, March 1999.
40. B. Friedlander, "A passive localization algorithm and its accuracy analysis," *IEEE Jour. Oceanic Engineering*, vol. OE-12, pp. 234–245, January 1987.
41. Y. Chan and K. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Processing*, vol. 42, pp. 1905–1915, August 1994.
42. D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone-array measurements," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97)*, Munich, Germany, pp. 371–374, April 1997.
43. L. Kinsler, A. Frey, A. Coppens, and J. Sanders, *Fundamentals of Acoustics*, John Wiley & Sons, third ed., 1982.
44. L. Ziomek, *Fundamentals of Acoustic Field Theory and Space-Time Signal Processing*, CRC Press, 1995.