Accelerated Speech Source Localization via a Hierarchical Search of Steered Response Power

Dmitry N. Zotkin and Ramani Duraiswami, *Member, IEEE*

***Abstract*—Accurate and fast localization of multiple speech sound sources is a problem that is of significant interest in applications such as conferencing systems. Recently, approaches that are based on search for local peaks of the steered response power are becoming popular, despite their known computational expense. Based on the observation that the wavelengths of the sound from a speech source are comparable to the dimensions of the space being searched and that the source is broadband, we have developed an efficient search algorithm. Significant speedups are achieved by using coarse-to-fine strategies in both space and frequency. We present applications of the search algorithm to speed up simple delay-and-sum beamforming and steered response power phase-transform weighted (SRP-PHAT) source localization algorithms. A systematic series of comparisons with previous algorithms are made that show that the technique is much faster, robust, and accurate. The performance of the algorithm can be further improved by using constraints from computer vision.**

***Index Terms*—Array signal processing, multimedia communication, multimedia applications, position measurement, speech enhancement, transducer arrays, user interfaces.**

I. Introduction

THE INVERSE problem of localizing a source by using signal measurements at an array of sensors is an almost classical problem in signal processing. Along with the associated problem of beamforming, it has attracted the attention of many researchers. Our interest in this problem is in the context of localizing and possibly beamforming multiple sources of speech in a conferencing environment. As noted by Brandstein [2], many of the classical beamforming algorithms were developed for applications in sonar or radar, rather than this particular problem, and consequently can perform poorly in the highly reverberant environments encountered in teleconferencing.

A recent book provides a very comprehensive introduction to the state-of-the-art in this field [3]. Generally, there are three classes of source localization algorithms [4]:

1) using steered beamformer energy response;
2) using high-resolution spectral estimation;
3) using time differences of arrival (TDOA).

Some algorithms combine features from more than one class. We will focus on algorithms that fall in the first of these three class. These algorithms, although capable of performing well, are usually slow because they involve a search for a peak of the response power. We show how the search can be performed efficiently in the case of speech sources in known rooms.

Inverse problems often exhibit ill-posed behavior in the sense of Hadamard [5], and their results are sensitive to noise in the data. In a reverberant environment, in addition, the data appears to be created by either the valid source position, or by any of a number of image sources induced by the scattering walls and surfaces. Thus, an additional element of ill-posedness is introduced in the problem, with the solution becoming multivalued. Many algorithms are posed in the context of statistical signal processing and do not treat this feature of the problem explicitly. A key to the solution of inverse problems is improved modeling that includes all available *a priori* information in the formulation. There has been some recent work in developing improved algorithms for this problem that use *a priori* information about the problem; for example, Brandstein presented algorithms for time delay estimation [2] and beamforming [6] that exploited knowledge about the pitch characteristics of speech. Our motivation here is similar, though we use different information.

A strategy that is often applied to resolve inverse problems is the iterative generation of a sequence of forward problems that might have created the data, with the best forward problem being taken to be the one that minimizes an objective function. In the selection of the candidate forward problems, we can easily satisfy any *a priori* constraints. In the present case, the forward problem that is generating the data is that there are speech source(s) in a room bounded by walls and other boundaries.

Application of this strategy to the present problem results essentially in the class of algorithm that we mentioned above—algorithms that steer the beamformer in various directions and search for peaks in the output signal, also called SRP (steered response power) algorithms. The simplest (delay-and-sum) beamformer computes the propagation delays from the source position to each microphone and compensates for these delays in order to coherently sum the signals arising from the source position. More sophisticated beamformers filter the microphone signals in addition to delaying them. In any SRP algorithm the evaluation of an objective function has to be repeatedly performed. This is usually the bottleneck in the performance of the algorithm.

As an illustration, one can build an energy map—the visual representation of variations in beamformer output energy versus the coordinates of the point that the beamformer is steered to (examples are shown later). The source manifests itself as a peak in the energy map. The map depends on the array geometry and on the spectral content of the signal. The width of the peak in the energy map is generally smaller for higher-frequency sources. Increased microphone separation also decreases the width of

Manuscript received September 1, 2003; revised April 4, 2004. This work was supported by the National Science Foundation under Award 9987944 and Award 0086075. This paper was originally presented at Proceeding IEEE ICASSP, Salt Lake City, UT, 2001 [1]. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. M. Mohan Sondhi.

The authors are with the Perceptual Interfaces and Reality Laboratory, UMIACS, University of Maryland, College Park, MD 20742 USA (e-mail: dz@umiacs.umd.edu, ramani@umiacs.umd.edu).

Digital Object Identifier 10.1109/TSA.2004.832990

the energy peak, which allows for higher localization accuracy and for better separation of sources (although increase of microphone separation is limited by the appearance of spatial aliasing for high-frequency signals if the microphones are placed too far apart). Search for peaks in energy map is an obvious source localization algorithm. However, in real applications the cost of computing the whole map at a fine enough resolution would be prohibitive, and some fast peak localization algorithm must be used. Traditional gradient descent [7] can be applied if the search space is expected to contain one peak. If this is not true and the search space is multimodal, gradient descent is likely to find a local maximum. Many trials of gradient descent with random starting points can be performed to improve the chances of finding the global maximum. Another fast search algorithm is stochastic region contraction, which is a general search algorithm for locating the global maximum of a multimodal function of many variables when the function satisfies certain conditions. It was successfully applied to microphone arrays, both in target localization [8] and in optimization of microphone placement [9], and significantly reduces the number of objective function evaluations compared to repeated gradient descent with random starting points. Use of sequential Monte Carlo methods [10], also known as particle filters, was also proposed for localization and tracking with microphone arrays (e.g., in [11], [12]). These algorithms are efficient because only a limited number of objective function evaluations are performed in the vicinity of the tracked position from previous frames.

This paper suggests a fast multilevel search strategy for an energy map using a coarse-to-fine paradigm in both the spatial and the frequency domains, which we will call doubly hierarchical beamforming (DHBF). This strategy works because it uses the characteristics of the signal producing the objective function; for example, speech sound has wavelengths that are comparable to the dimension of the space that is being searched in this application (teleconferencing). The search algorithm we develop can be applied to any underlying energy function (we use two versions of SRP beamforming, simple delay-and-sum and phase-transform weighted, for illustration). It is not limited by the number of sources (though it does require them to be spatially separated to a certain degree and have similar power) or background noise structure, and has a predictable cost in terms of the number of objective function evaluations. It is particularly suitable for implementation in environments where a prior knowledge of the spatial domain is available (e.g., the set of source locations being searched over can be restricted to an area bounded by room walls, or the source is one of several objects detected via, say, computer vision, etc.). We show in the paper that the DHBF localization algorithm is comparable in localization performance to other SRP based algorithms mentioned above and that the number of objective function evaluations is significantly reduced compared to other search strategies, resulting in much greater efficiency.

II. Background Information

We summarize here the *a priori* problem information and other background material known about the problem, and present some preliminary conclusions that can be drawn from them and used to determine the coarse-to-fine strategy to speed up the steered response power search.

Spatial Extent: The source occupies a region whose spatial extent is limited, so that in the search we can refine the search to a relatively small region. The environment is usually a workplace, a conference room or rarely an auditorium. In addition, sources are typically separated by distances of ~ 1 m, and will definitely be at least 0.3 m apart (because speaking humans may be expected to be separated by this distance).

Nature of the Speech Signal: Although the frequency range of human hearing extends from 20 Hz to 20 kHz, the sound produced by the human vocal tract has significantly less range, extending from about 100 Hz to 6 kHz [13]. The spectral structure of the most energetic part of speech, the voiced phonemes (which include vowels and some consonants), consists mainly of a combination of integer multiples of a fundamental frequency f_0 that lies between 80–200 Hz for males and 150–350 Hz for females. The voiced sounds constitute the low-frequency part of the speech spectrum. In addition, stop consonants and fricative consonants contribute significant energy around 3–5 kHz.

f	λ	**Feature**	**Remarks**
20 Hz	17 m	Auditoriums	lower hearing limit
100 Hz	3.4 m	Conference rooms	speech lowest frequency
200 Hz	1.7 m	Rooms, human height	peak energy for speech
6 kHz	5.5 cm	Pinna dimensions	speech highest frequency
20 kHz	1.7 cm	Concha size	upper hearing limit

Relationship Between Frequency and Wavelength: The elementary equation $f\lambda = c$ indicates the relationship between frequency and wavelength. The wavelengths of audible sound are comparable to the dimensions of the space we live in and to the dimensions of our anatomical features. Humans use the spectral cues resulting from complex scattering of sound waves by objects with size comparable to the wavelength to determine the size of the environment and perform source localization [14]. Our goal is to exploit the relationship between speech frequency content and interesting spatial dimensions (presented in the table above) to develop fast search algorithms for locating sound sources.

Delay and Sum Beamforming Localization: Assume that the acoustic source that produces the signal $y(t)$ is located at point p and K receivers (microphones) are located at points $q_1, \ldots, q_K$. The signal $s_m(t)$ received at the mth microphone is given by

$$s_m(t) = y(t) \star h_m(q_m, p, t) + z_m(t) \tag{1}$$

where $h_m(q_m, p, t)$ is the room impulse response (RIR) function for the given positions of the source and the mth microphone, star denotes convolution, and $z_m(t)$ is the combination of the channel noise and any environmental noise that is assumed to be independent at all microphones and not correlated with $y(t)$. As suggested in [4], we decompose the RIR into a direct arrival component (which simply consists of a single peak with amplitude r_m^{-1} at time $\tau_m = r_m/c$ where $r_m = \|p - q_m\|$ and c is the sound speed) and the rest of the RIR that we denote as $h_m^*(q_m, p, t)$, in which case (1) becomes

$$s_m(t) = r_m^{-1} y(t - \tau_m) + y(t) \star h_m^*(q_m, p, t) + z_m(t). \quad (2)$$

The received signals thus contain delayed versions of the source signal plus its convolution with the rest of the RIR. We denote the time difference of arrival of a signal between receivers m and n as $\tau_{mn} = \tau_n - \tau_m$. The set of delays τ_{mn} can be associated with the location of the source. Another set of TDOAs $\hat{\tau}_{mn}$ is associated with the beamformer steering process. Given $\hat{\tau}_{mn}$, one can compute the output $s_B(t)$ of the delay-and-sum beamformer as

$$s_B(t) = \frac{1}{K} \sum_{m=1}^{K} s_m(t + \hat{\tau}_{mn}) \quad (3)$$

where n is a reference microphone, which can be chosen to be the microphone closest to the position determined by the set of $\hat{\tau}_{mn}$ so that all $\hat{\tau}_{mn}$ are negative and the beamformer is causal. To steer the beamformer, we select $\hat{\tau}_{mn}$ corresponding to different positions in space, and if $\hat{\tau}_{mn} = \tau_{mn}$, then the contributions from the source add coherently in the beamformed signal, resulting in unity gain for the source, whereas the signals of other (not steered to) sources and noise add incoherently, and their power will decrease on average by a factor of K^{-1}. Although this analysis is idealized because it assumes IID noise and neglects reverberation, it still provides useful insights into beamformer-based localization.

Equation (3) is in the time domain, but can also be expressed in the frequency domain. We recall that if a function $s(t)$ has Fourier transform $S(\omega)$, then time shifting of s by t_0 modifies its Fourier transform as $s(t-t_0) \Leftrightarrow S(\omega)e^{-j\omega t_0}$, where $j = \sqrt{-1}$. In the Fourier domain, (3) becomes

$$S_B(\omega) = \frac{1}{K} \sum_{m=1}^{K} S_m(\omega) e^{jw\hat{\tau}_{mn}} \quad (4)$$

and the output power can be expressed (up to a scale factor) as

$$P_B(\omega) = \sum_{m=1}^{K} \sum_{n=1}^{K} S_m(\omega) S_n^*(\omega) e^{jw\hat{\tau}_{mn}}. \quad (5)$$

A simple localization strategy can then be suggested by searching through the space of $\hat{\tau}_{mn}$ for an energy peak in the output $P_B(\omega)$. The search is usually performed in some organized fashion through the possible locations of sound source in three-dimensional (3-D) space, generating sets of TDOA corresponding to the spatial locations by trivial geometric computations. We describe below existing ways to improve robustness and speed of beamformer-based localization and suggest a novel fast search strategy based on hierarchical subdivision of space and frequency.

SRP-PHAT Localization: In localization algorithms that rely purely on TDOAs to localize the sound source, the TDOAs are usually obtained via a generalized cross-correlation (GCC) between signals s_m and s_n acquired at the mth and nth sensors respectively [15]. Denote the GCC of $s_n(t)$ and $s_m(t)$ by $r_{mn}(\tau)$ and its Fourier transform by $R_{mn}(\omega)$. Then

$$R_{mn}(\omega) = W_{mn}(\omega) S_m(\omega) S_n^*(\omega) \quad (6)$$

where $W_{mn}(\omega)$ is a weighting function. Ideally, $r_{mn}(\tau)$ (computed as the inverse Fourier transform of $R_{mn}(\omega)$) will have a peak at the true TDOA between sensors m and n (τ_{mn}). In practice, many factors such as noise, finite sampling rate, interfering sources and reverberation might affect the position of the peak. The phase transform (PHAT) weighting function was introduced in [15]

$$W_{mn}(\omega) = |S_m(\omega) S_n^*(\omega)|^{-1}. \quad (7)$$

The PHAT weighting places equal importance on each frequency by dividing the spectrum by its magnitude. It was later shown [4], [16]–[18] that it is more robust and reliable in realistic reverberant conditions than other weighting functions designed to be statistically optimal under specific nonreverberant noise conditions. The SRP-PHAT algorithm [4], [19] applies the PHAT weighting in the context of the filter-and-sum beamformer. The power in its output $P_B(\omega)$ is given by

$$P_B(\omega) = \sum_{m=1}^{K} \sum_{n=1}^{K} W_{mn}(\omega) S_m(\omega) S_n^*(\omega) e^{j\omega\hat{\tau}_{mn}} \quad (8)$$

which generalizes (5) by including the nonnegative weighting function. In SRP-PHAT the weighting function (7) is used to increase robustness in reverberant environments. As with simple delay-and-sum beamforming, the total power in the output of the beamformer $P_B = \int P_B(\omega) d\omega$ depends on how the beamformer is steered. Given the geometry of the microphone array, the steering to a position p is done by computing the set of TDOA τ_{mn} that would have been observed for a source at that position. Although acoustical sources in space produce peaks in the beamformer's output, the search for these peaks can be complicated by multiple local maxima in the search space in case of SRP-PHAT algorithm. A sophisticated search mechanism is required to ensure successful localization of the global maximum.

Time delay Imprecision: Given four or more receivers, every point in physical space (x, y, z) can be mapped to a point in delay space $(\tau_{12}, \tau_{13}, \tau_{14}, \ldots, \tau_{1K})$. The inverse map (from delays to source location) is nonlinear and ill-posed as discussed earlier. Here we consider the effect of errors in time-delays caused by an incorrect hypothesis of the source location on the computed steered response power. If the error in phase is small, then the coherence in the signals being added won't be completely destroyed, though incoherent components will also get an increase in their energy.

Consider the simplest beamformer consisting of two microphones separated by distance $2d$. If the beamformer main lobe is aligned perfectly on the source, the phase difference between signals at two microphones (after application of appropriate time delays) is zero and the beamformer gain is equal to 1 for the source whereas the gain for the incoherent noise can be expected to be equal to $\sqrt{2}/2$ and the power of

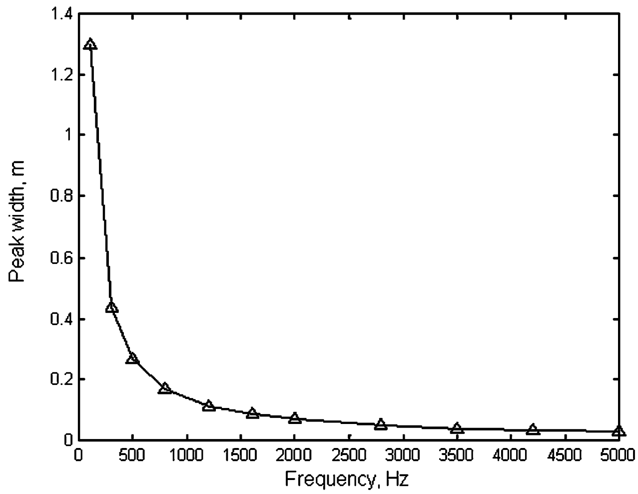

Fig. 1. Beamformer peak width as a function of frequency.

the incoherent noise is halved. Simulation shows that as the misalignment of the beamformer (and consequently the phase difference between two signals) increases, the output power of the source decreases and when the phase difference is equal to approximately $2\pi/5$ the power of the source drops close to the expected power of incoherent noise. Thus, it can be reasonably expected that when this simplest beamformer is misaligned by the distance that results in phase difference of less than $2\pi/5$ between channels, the signal of the desired source will be amplified to a certain degree compared to the incoherent part. It can also be shown that if a source located at a distance l from the array of diameter d is shifted by a distance b (which is the same as having the array misaligned by the same distance), the phase difference introduced by the shift is limited by $4\pi db/l\lambda$ (the phase difference is maximal when the source is on the axis of the array and the shift is parallel to the array plane), where λ is the source wavelength. Thus, it is conservative in far-field $(l \gg d)$ source case to estimate that an error in the source position of less than $\lambda/5$ will still result in a coherent gain in the beamformed signal. This is confirmed by tests with actual speech signals to our array. We refer to this result as the *imprecision heuristic*.

Another way to look at this result is to plot the spatial width of the peak in the energy map as a function of the source frequency (Fig. 1). From simulations performed by mixing actual room recordings of speech we see that there is an inverse relationship between the peak width in the energy map and sound wavelength. The peak in the energy map has an FWHM (full width at half maximum) of approximately $2\lambda/5$ for our array configuration, consistent with the heuristic. For the frequency of 150–160 Hz ($\sim f_0$) we get $2\lambda/5 \approx 0.8$ m.

III. Hierarchical Localization Algorithm

Let there be K microphones and let us consider beamforming a data frame sampled at a frequency higher than the Nyquist rate for the highest frequency in the signal. For now, we restrict our consideration to the two-dimensional case and work in the direction-of-arrival (DOA) space, ignoring range. We divide the space of search parameters (azimuth φ and elevation θ) into quadrants of size that is consistent with a $\lambda/5$ imprecision heuristic for the lowest signal frequency. We then evaluate the objective function (energy in the beamformed signal) at the centers of these quadrants, creating a coarse-level energy map, and search for local maxima in it. Each local maximum is tagged for further consideration. All tagged quadrants are further subdivided in the next pass of the algorithm.

There are two issues that must be fixed with this approach. First, at the coarsest level we must restrict the beamforming to frequencies that are guaranteed to see an improvement in their power despite inaccurate steering (to the quadrant center instead of the true source position). In addition, performing beamforming using either (4) or (8) on the full signals for even the relatively few coarse quadrants would be uneconomical. A first possibility to fix this problem is the decimation of the signal in the time domain. However, one quickly realizes that it can lead to significant aliasing. An approach that achieves both goals is lowpass beamforming in the frequency domain, which can be done quickly [20] simply by computing beamformed signal power only in those frequency bins where it is necessary.

In this approach we compute the FFTs of each of the received signals at K channels. We decimate the signal in the frequency domain (performing a lowpass operation) with a cutoff frequency determined by the quadrant size at the coarsest level and compute the power at the centers of the quadrants. Let there be k sources, leading to k tagged quadrants. These are further recursively divided, using the quadtree data structure described below, and the power is computed at child nodes with the cutoff frequency twice as high at each step (because the length of the quadrant side gets divided by two at each level of quadtree subdivision). The child node with the highest energy level is selected as the most probable source position, and the subdivision is repeated recursively until desired node level and desired precision is achieved.

A. Quadtrees and Octrees

To perform a hierarchical space division we use quadtrees. A quadtree is a data structure for hierarchical representation and processing of two-dimensional (2-D) spatial data, which is typically organized as follows. The root of a quadtree is associated with a 2-D region bounded by a parallelogram (often a square). This parent region is subdivided into four similar equally sized quadrants, each carrying more specific information about its portion of space. Each child quadrant is, in turn, recursively partitioned into four children. The process of subdivision can continue infinitely, but in practice it has to stop (e.g., when further subdivisions do not significantly help the search). The hierarchical nature of quadtrees allows one to efficiently represent and search data distributed in 2-D space, which is possible because the size of a quadtree representing a 2-D region is $O(s)$ where s is the perimeter of the object. Algorithms that execute on quadtrees rather than on pixel arrays have running times proportional to the number of blocks in the quadtree.

Similarly, the octree is the data structure representing 3-D volumetric data, with eight children per node and the number of nodes proportional to the area of the object's surface. Use of quadtrees and octrees leads to the dimension reduction effect;

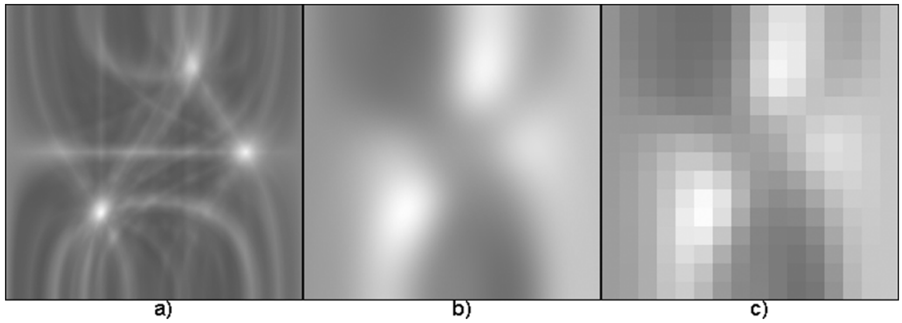

Fig. 2. (a) Sample energy map (energy versus azimuth and elevation) obtained using SRP-PHAT algorithm with three sources, resolution 512 × 512. (b) Same map for the signals lowpassed at 2 kHz. (c) Same map for the signals lowpassed at 2 kHz, resolution 16 × 16.

an octree algorithm applied to a 3-D problem is analogous to an array-based algorithm in 2-D [21].

B. Implementation

The algorithm for localization of multiple sound sources using the proposed doubly-hierarchical search of steered response power starts by forming a quadtree of level l and corresponding coarse $2^l \times 2^l$ search grid. When it is known that the sources are located in front of the microphone array (for example, when the array is on a wall) the search area can be defined as a square in the DOA space $\varphi \in [-\pi/2, \pi/2]$, $\theta \in [-\pi/2, \pi/2]$. The search space is divided by the coarse grid into 2^{2l} quadrants with the quadrant size chosen sufficiently small so that it is unlikely that two sound sources share the same quadrant and that local maxima of the energy can be computed. The latter condition is ensured by the $\lambda/5$ heuristic. Note that the quadrants in the DOA space are not rectangular in the world coordinate frame, and the cutoff frequency and the initial grid size have to be selected by remapping the DOA grid onto a (x, y, z) grid using the maximum possible source depth (determined by known room dimensions). The nodes that correspond to the local maxima in the constructed energy map are selected for further processing. Every node is recursively searched by partitioning it into four children, and the child with the maximum energy level is selected for subdivision at the next level. The recursion terminates when the quadtree branch reaches a certain depth, corresponding to a fixed minimal quadrant size. The procedure is repeated for about ten levels of quadtree subdivision, which yields a quadrant of size of about 1 cm (when recomputed to spatial units) in our implementation. The center of the maximal energy quadrant at the deepest level is output as the source position.

A problem that may arise is that the actual peak may lie at the boundary of a quadrant in the DOA space resulting in possible mislocalization of the energy maximum during the coarsest stage of algorithm. To avoid this, we perform a simple check at the last step. If at the end of the recursive search within a quadrant the peak is localized at the boundary P of the original coarse quadrant V, a search is also performed in the neighboring coarse quadrant V' that shares P with V. In practice this rarely happens.

The algorithm as described above is developed for 2-D sound source localization. Sometimes it is desirable to perform full 3-D localization with one or more arrays [22], and our search algorithm can be adapted for this situation. In this case, the search algorithm is executed in a 3-D space using an octree based peak search instead of a quadtree, which directly determines the Cartesian coordinates of the sound source(s) relative to the arrays.

IV. Performance Evaluation of the Algorithm

We compare the developed search algorithm with other steered response power algorithms. The algorithms we test are full search [4], repeated gradient descent [7], stochastic region contraction [9] and DHBF. We will denote them with one-letter designations (for the plots) as follows:

- (F) Full search for the SRP-PHAT energy maximum over the 1024-by-1024 grid of all possible DOA's in the $Z > 0$ hemisphere.
- (G) Repeated gradient descent with $N = 1000$ different starting points (we found this number is the minimum sufficient for repeatable and robust localization of the global maximum).
- (S) Stochastic region contraction with the parameters suggested in [9]$(K_1 = 100, K_2 = 400, \hat{K} = 8)$.
- (D) The proposed DHBF algorithm with 16 × 16 initial grid size, 1 kHz initial high frequency cutoff, and maximum quadtree subdivision level of 10.

A. Multisource Energy Map Illustration

Let us start by observing some interesting properties of the energy map derived using the SRP-PHAT energy function. In Fig. 2, energy in the beamformed signal is plotted as a pixel intensity on a two-dimensional plane with the horizontal axis being the azimuth $\varphi \in [-\pi/2, \pi/2]$ and the vertical axis being the elevation $\theta \in [-\pi/2, \pi/2]$ (so essentially a hemisphere of directions with $Z > 0$ is plotted). From here on we will use simulated results on a circular planar array of seven microphones with one microphone at the center and the other six located at the circumference of the array to match a real array we have. The radius of the array is 0.3 m. Assume that the origin of the coordinate system is placed at the center of the array with X and Y axes in the array plane and Z axis orthogonal to it, and the point with $\varphi = \theta = 0$ is located directly in front of the array on the positive Z axis. The sources in these plots are placed at $(-3.3, 1.7, 2.5)$, $(2.5, -2.7, 3.5)$, and $(0.0, 4.2, 3.0)$ meters.

In Fig. 2(a), the sources are well defined and sharp. However, the number of local maxima in the energy map is much more than the number of sources. Each bright curved line is formed

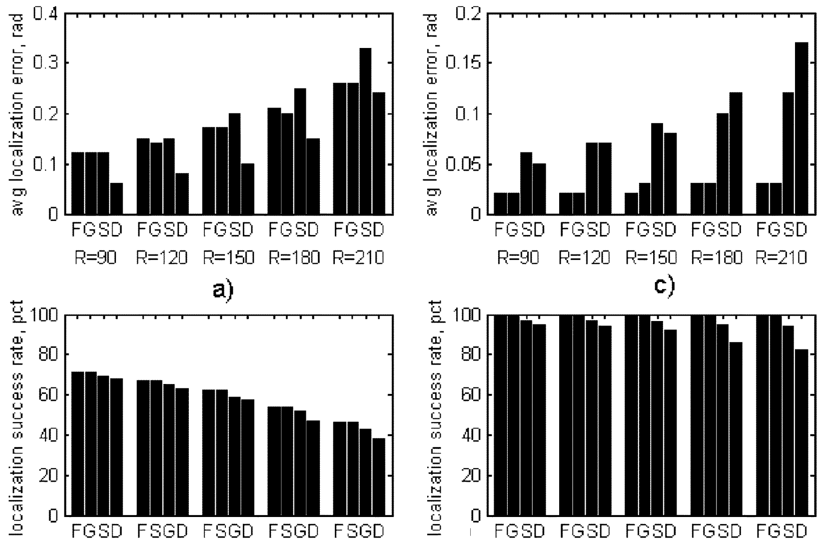

Fig. 3. (a) Average localization error for 4 algorithms. (b) Percentage of correct localization for 4 algorithms. (c) Average localization error with data frames where full search fails omitted. d) Percentage of correct localization under the same condition. R is the reverberation time in ms.

by a projection of the TDOA locus (a hyperboloid for a given microphone pair) into the (φ, θ) space. This projection can be viewed as taking an intersection of the equi-TDOA surface centered at the point between the microphones in the pair with the sphere of some radius ρ centered at the center of the array and letting $\rho \to \infty$. As $\rho \to \infty$, the intersection locus describes all DOAs that satisfy the observed TDOA. Lines corresponding to the same source intersect at the same spot, which is the visual illustration of the source localization by intersection of cones defined by each TDOA.

It is difficult to localize multiple sources in the map because of many local maxima. In contrast, in Fig. 2(b) the same map is plotted with the signal lowpassed at 2 kHz. All fine details are removed, and the map consists of three broad peaks located roughly at the source positions. This can be expected because of the relationship between peak width and frequency—the lower the frequency, the larger the spatial shift that keeps the source still in partial focus. Fine details are not present in the map because high frequencies (i.e., small wavelengths) that can contribute to fast spatial variations of the beamformer output are missing. The rationale behind the coarse map of lowpassed signals as an initial step of our localization algorithm is that if the map in Fig. 2(b) is downsampled [Fig. 2(c)], the resulting coarse map has peaks in approximately correct positions, which are used as starting points for a recursive search.

Bandlimiting at low resolution is thus crucial to avoid misleading peaks and obtain correct initial estimates of source locations. It is also important to keep the signal appropriately bandlimited (determined by the quadrant size) during a hierarchical subdivision of space. This strategy also uses the important prior information that the peaks that arise at each frequency in a spatial neighborhood are caused by the same physical source.

B. Accuracy in Reverberant Environments

First, we measured and compared performance of several algorithms under the simulated reverberation, which is perhaps more degrading to the localization performance than the noise alone. We used clean speech (utterance of ten consecutive digits "zero, one, two, three…" with short pauses between the digits) recorded by a microphone placed close to the speaker's mouth as a source signal. From a 5 s utterance 33 frames of length 50 ms with SNR greater that 12 dB were selected for processing. We simulated reverberation using a simple image model [23] in a rectangular room. In this model, a regular lattice of virtual sources representing the reflections of the acoustic source in room walls (including floor and ceiling) is created, and a room transfer function (or, alternatively, room impulse response, or RIR) can be computed [24]. Simulated microphone outputs can then be computed by convolving the source waveform with the appropriate RIR's for each microphone position. The sampling frequency was set to 40.44 kHz to match the setup used in the real experiments (described later). The simulated room had dimensions $5 \times 2.5 \times 4$ m. The origin of the coordinate system was placed at the center of the room, and the Y axis was vertical so that the coordinates of one of the room corners was (2.5,1.25,2.0). The center of the microphone array was at (0.0,0.0,-2.0) and the array was a 0.3 m radius array with six microphones equispaced on the array circumference and one microphone placed in the center. The RIR was computed up to the 24th reflection and lowpassed with cutoff frequency of 100 Hz as described in [23]. We ran several simulations with different wall reflective properties to simulate different reverberation times. For each reverberation time, we performed 64 runs of the four algorithms being tested using 64 randomly selected source positions within the room subject to constraints $X \in [-2.0, 2.0]$, $Y \in [-1.0, 1.0]$, $Z \in [-1.0, 1.0]$.

From the known source position, the correct DOA was computed and compared to the DOA estimate produced by the four search algorithms. The beamformed energy in all algorithms was computed using the SRP-PHAT technique using frequency components from 300 Hz to 11 kHz, except for the DHBF algorithm, where the high cutoff was adjusted during the search as described previously. We present experimental data in concise form here. In the top left histogram in Fig. 3, we show the average localization error for the four algorithms for five different reverberation times (90, 120, 150, 180, and 210 ms, respectively). In the bottom left histogram, percentages of successful localizations are shown (where "successful localization" is defined as localization within ten degrees of the true DOA). In the right part of Fig. 3, we show the same plots with all frames where full search fails to localize the source omitted (see discussion below), in which case full search obviously shows 100% correct localization.

To our surprise, in initial simulations we found that the DHBF algorithm, in addition to its speed, was superior to all other methods in terms of average localization error, even to the full search. This means that under simulated reverberation there exist some cases when the highest energy peak does not coincide with the true DOA (and thus, full search, which presumably should be the most robust method, fails). This happens in about 30% of all cases with the smallest reverberation time (90 ms) and in about 55% of all cases with the largest one (210 ms). Examination of the energy maps in such cases shows that generally the highest (false) peak is located close to the $\varphi = \theta = 0$ DOA, in which case all search methods except DHBF are distracted by it. On the other hand, the lowpassed energy map that is the starting point for the DHBF algorithm does not contain the false peak, and thus the DHBF search is initialized in the vicinity of the source and stays in the neighborhood of the initialization point during the refinement stage. This results in a DOA estimation that is close to the correct DOA.

The appearance of a false peak located around the room center is due to the several factors, in particular to the symmetric arrangement of the microphones in the array and to the effect of the windowing operation, which causes a bias in Fourier transform coefficient phases (moreover, remember that SRP-PHAT operates essentially only on the phases of the coefficients and thus is particularly sensitive to the bias). It is also enhanced by the idealized room reverberation model with a regular lattice of virtual sources. In real conditions, slight array and room asymmetries and objects present in the room are likely to diminish this effect. Indeed, in real experiments we did observe some data frames with such behavior (in which all methods but DHBF find a false peak located around the room center) but in reduced proportion compared to the simulations. We also performed preliminary simulated experiments using a planar microphone array consisting of randomly placed microphones and a smoother windowing operation, and both of these somewhat decrease the frequency of the false peak appearance but do not completely eliminate such cases. This topic is a subject of our future research.

In summary, we see that the DHBF localization performance is comparable to the performance of other search algorithms in case of simulated reverberation. As the reverberation time increases, the DHBF performance somewhat worsens if the energy map has a pronounced peak at the true DOA. However, this is partially offset by good DHBF performance if there is a false energy peak near the center of the room. Moreover, the algorithm that is closest to the DHBF algorithm in terms of speed, stochastic region contraction, also performs worse than the repeated gradient descent and the full search.

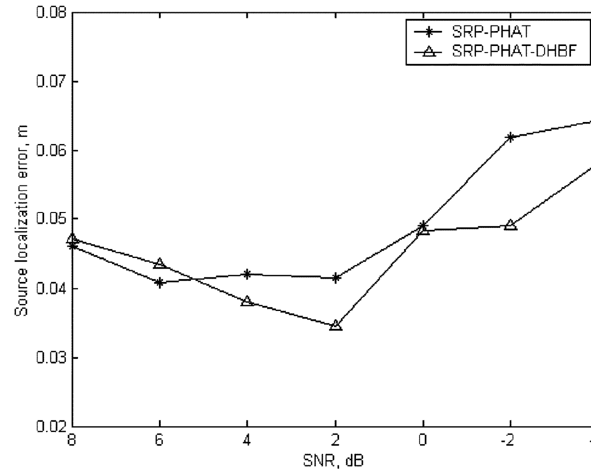

Fig. 4. Localization error versus SNR for SRP-PHAT and SRP-PHAT-DHBF.

C. Accuracy in Noisy Environments

We performed testing of the localization accuracy of the SRP-PHAT algorithm with and without DHBF search technique under different noise conditions. We synthesized an acoustic scene in which one source was located at a known position and contaminated the signal in each channel with white noise to achieve different SNR's. In Fig. 4, we show the average localization error in 20 trials with random source positions for the original SRP-PHAT algorithm and its acceleration using our search algorithm (SRP-PHAT-DHBF). For the SRP-PHAT algorithm, we created a spatial energy map with the same resolution as for the SRP-PHAT-DHBF algorithm at the deepest level (ten) of spatial subdivision, so that the map size was 1024×1024 pixels and the resolution was 0.176 degrees. A data frame length of 2048 points (93 ms at 22.05 kHz) was used. The difference in accuracy between two algorithms is not very significant. In fact, SRP-PHAT-DHBF is even more accurate, perhaps by avoiding false peaks generated by SRP-PHAT at a fine resolution with noisy data. Other experiments conducted show that both SRP-PHAT-DHBF and SRP-PHAT exhibit the same noise robustness level in comparison with TDOA-based localizers.

In addition, the frequency hierarchy allows for selection of multiple sources on the coarsest energy map as long as they are separated by certain minimal angular spacing that depends on array geometry and cutoff frequency. We performed tests to determine the minimum resolvable multiple source spacing. We placed two sources at random points (φ_1, θ_1) and (φ_2, θ_2) on the $Z > 0$ hemisphere $\varphi \in [-\pi/2, \pi/2]$, $\theta \in [-\pi/2, \pi/2]$ of a fixed radius $\rho = 3.5$ m and ran the SRP-PHAT-DHBF algorithm to find out the probability of finding both sources as a function of an angle between their DOA's. The resulting histogram is presented in Fig. 5 and shows that the angular spacing corresponding to 50% chance of correct source separation is about 40 degrees.

D. Accuracy in Real Conditions

Finally, we performed a test of the algorithms in real reverberant and noisy conditions. We used a system consisting of a 0.3 m radius seven-channel microphone array, a data

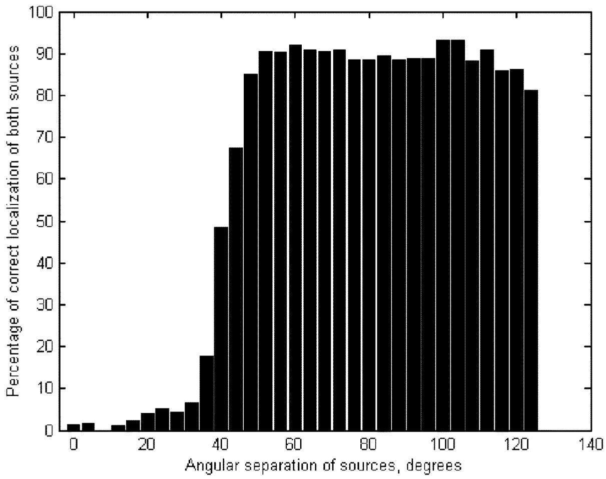

Fig. 5. Histogram of the percentage of correct localization of two simultaneous acoustic sources versus their angular separation for the SRP-PHAT-DHBF algorithm.

acquisition board and a PC for data collection. The experiments were conducted in a large $5.6 \times 2.7 \times 4.7$ m (width $\times$ height $\times$ depth) office room. The microphone array configuration was the same as in the simulations above. We used Panasonic WM-61A omnidirectional speech-band button microphones with a custom preamplifier on AD797 chips connected to the PowerDAQ PD-MF-16-333/12L data acquisition board capturing seven channels at 40.44 kHz sampling frequency and 12 bit resolution. The microphones were mounted on a large sheet of thick foam rubber to dampen wall reflections near the array, and the array was placed on the longest room wall a third of the wall length from the corner. The measured room reverberation time was about 350 ms, and the main interference source was the noise of computer equipment fans [25].

In the experiment, the person stood in five positions in the room and uttered the same sentence ("Dear fellow radio listeners!…Pass pass pass pass pass…") at all positions. The positions were different in azimuth, elevation and distance, with the furthest one being about 2.5 m from the array. From the recording, we selected all frames where the SNR of the recorded speech was greater than 12 dB and ran the same four algorithms on these frames. We repeated the processing with 4 different frame lengths (25, 50, 75, and 100 ms). The total number of frames selected for the processing was 548, 276, 190 and 143 frames for frame length of 25, 50, 75, and 100 ms, respectively. We show the results in Fig. 6 using error rate versus error threshold metric (e.g., if the dotted line shows an error rate of 70% at an error threshold of 7 degrees, it means that the DOA estimate produced by DHBF algorithm is within 7 degrees of true DOA in 30% of the cases). In this metric, for a fixed error rate, the lower the error threshold, the better the algorithm performance.

It can be seen that the DHBF algorithm has a region of superior performance when the frame length is 25 ms, and the performance is comparable to other algorithms using frames of 50 ms. When the frame length is equal to 75 and 100 ms, the performance of DHBF is notably worse than that of the stochastic region contraction and of the gradient descent. The negative effect of the short data frame on the full search is caused by the windowing operation, which biases the bin phases and causes the appearance of the false energy peak (as described before, DHBF is often able to get a good DOA estimate even when the full computed energy map has a maximum in an incorrect location because the bias is more pronounced at higher frequencies and the initial coarse-grid energy map estimation done by DHBF excludes them from consideration). In real applications it is often necessary to track moving sound sources and to use short processing frames to minimize tracker latency, in which case DHBF is a viable alternative to existing fast localization algorithms.

E. Computational Speed

The main advantage of the proposed search algorithm is its speed. In [4], the full spatial map with a resolution of 0.1 degrees is computed in the DOA space for the SRP and the SRP-PHAT algorithms to compare their performance. In the proposed search algorithm, the number of objective function evaluations is much smaller than in previously proposed fast search techniques. We performed a direct comparison between the number of operations needed to localize the source using full search, repeated gradient descent, stochastic region contraction and DHBF. We explicitly counted the average number of objective function evaluations in our code while doing source localization in real environment as described above and obtained the following estimates.

- Full search: $N_f = 2^{20} \approx 10^6$ evaluations per frame processed.
- Repeated gradient descent: $N_g \approx 3.8 \cdot 10^4$ evaluations per frame processed.
- Stochastic region contraction: $N_s \approx 4.6 \cdot 10^3$ evaluations per frame processed.
- DHBF: $N_d \approx 3.7 \cdot 10^2$ evaluations per frame processed. In addition, most DHBF evaluations are *far cheaper* than the evaluations used in the other algorithms.

Indeed, for the maximum quadtree subdivision level m, $k = 3$ sources and initial (coarse) grid quadtree subdivision level l, our DHBF algorithm performs $N_d = 2^{2l} + 4k(m-l)$ energy evaluations, compared to $N_f = 2^{2m}$ evaluations done if the full spatial map at the fine resolution is computed. For $m = 10$, $l = 4, k = 3$ $N_d = 328$, which agrees with experimental results. Furthermore, most evaluations are performed with *bandpassed versions* of the signal, which additionally decreases the DHBF computational load. If we assume that the room is a cube with the side length of Z, the highest signal frequency is $f_d/2$ and frequency decimation is used according to the rules described above, then the effective number of evaluations N_{dd} can be obtained by introducing evaluation weights $\xi(\lambda)$ into the formula for N_d

$$N_{dd} = 2^{2l}\xi\left(\frac{Z}{2^l}\right) + 4k \sum_{k=l+1}^{m} \xi\left(\frac{Z}{2^k}\right) \tag{9}$$

where $\xi(\lambda)$ represents the ratio of the bandlimited signal frequency range to the full-band signal frequency range and can be directly derived from the quadrant size heuristic as $\xi(\lambda) = c/(\lambda f_d)$ when $c < \lambda f_d$ and $\xi(\lambda) = 1$ otherwise. For the case considered above, using the same values of m, k, l and letting

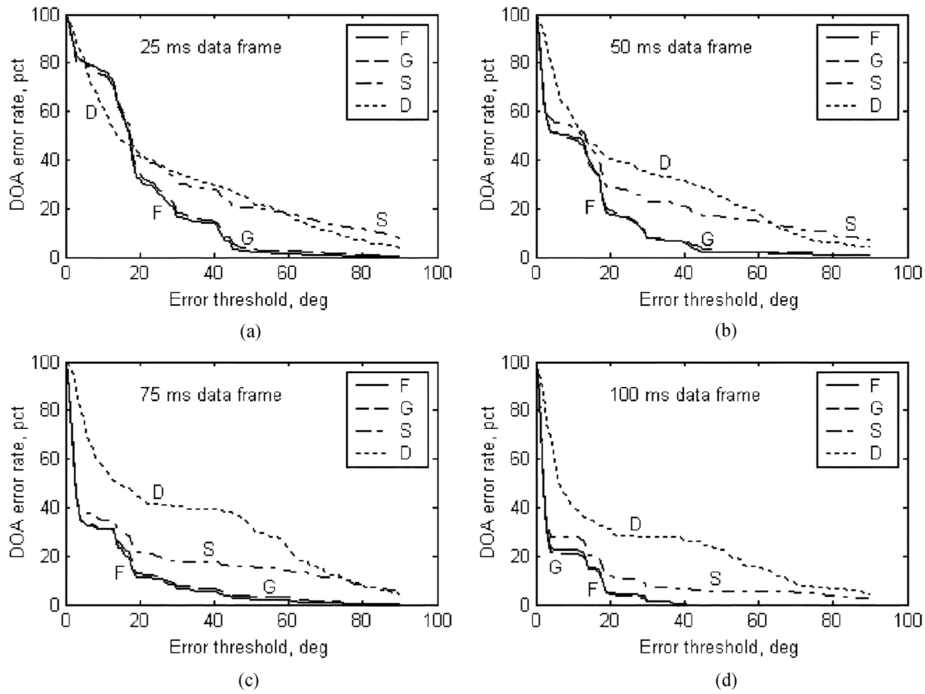

Fig. 6. Error rate (percents) versus error threshold (degrees) for varying frame length in real conditions.

$Z = 4$ m, $f_d/2 = 11$ kHz, $N_{dd}/N_d \approx 0.19$, signifying an additional 5-fold computational load reduction.

We have implemented our algorithm together with several other SRP-based algorithms and achieved real-time DHBF algorithm operation on a dual Pentium III 600 MHz Dell Precision 620 PC under Windows 2000 operating system with no specialized hardware.

F. Vision-Constrained Source Localization

One application area for the DHBF algorithm is source localization in multimodal user interface systems. In such system, *a priori* information from video will be available [26]. Knowing the location of the foreground objects in the image plane of the camera we can restrict the search to areas that are likely to be sources. In our implementation of this concept an active camera is used to scan the room. A background model is constructed by mosaicing several images taken at different camera orientations. The room is constantly monitored for foreground objects, and a simple background subtraction method based on pixel intensities is used to classify every pixel as foreground/background. Because the camera image is 2-D, the contour of the foreground object defines a "visual cone" in which the object lies, with the cone origin at the camera center. The cone is bounded by the room walls. The union U of the visual cones of all foreground objects is either used directly in a full 3-D search by ignoring voxels in the initial coarse octree that do not intersect U, or U is reprojected back onto the DOA space using known geometric relationships for the 2-D DOA search. For the sample videoconferencing application, we use two-camera setup. One camera collects the image from the active source and the other one constantly scans the room to dynamically provide constraining data. In this way, one can obtain an additional system latency reduction. Such uses are further described in [26].

V. Conclusions

We have presented a generic doubly hierarchical search algorithm for speech source localization using steered response power. The algorithm was designed with prior information about the speech in mind, and is able to achieve the accuracy that is comparable with other steered response power search algorithms in reduced processing time. Our experiments show that significant speedups can be achieved while keeping reasonable localization accuracy. The algorithm also has the ability to localize reasonably separated multiple active sources simultaneously. The search algorithm has applications in multiple areas including multimodal human-computer interaction, videoconferencing, and other entertainment, educational and remote collaboration applications.

Acknowledgment

The authors wish to thank the anonymous referees for their comments and suggestions.

References

[1] R. Duraiswami, D. N. Zotkin, and L. S. Davis, "Active speech source localization by a dual coarse-to-fine search," in *Proc. IEEE ICASSP 2001*, vol. 5, Salt Lake City, UT, 2001, pp. 3309–3312.

[2] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Am.*, vol. 105, no. 5, pp. 2914–2919, 1999.

[3] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001.

[4] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 157–180.

[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*. Berlin, Germany: Springer-Verlag, 1992.

[6] M. S. Brandstein and S. Griebel, "Explicit speech modeling for microphone array applications," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 131–153.
[7] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-31, pp. 1210–1218, Dec. 1983.
[8] V. M. Alvarado, "Talker localization and optimal placement of microphones for a linear microphone array using stochastic region contraction," Ph.D. dissertation, Dept. Eng., Brown Univ., Providence, RI, 1990.
[9] M. F. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction," *IEEE Trans. Signal Processing*, vol. 39, pp. 2377–2386, Nov. 1991.
[10] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. Berlin, Germany: Springer-Verlag, 2001.
[11] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE ICASSP 2001*, vol. 5, Salt Lake City, UT, 2001, pp. 3021–3024.
[12] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. IEEE ICASSP 2002*, vol. 2, Orlando, FL, 2002, pp. 1777–1780.
[13] H. Silverman, "Some analysis of microphone arrays for speech data acquisition," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP–35, pp. 1699–1712, Dec. 1987.
[14] W. M. Hartmann, "How we localize sound," *Phys. Today*, vol. 52, no. 11, pp. 24–29, 1999.
[15] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. ASSP-24, pp. 320–327, Aug. 1976.
[16] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE ICASSP*, Munich, Germany, 1997, pp. 375–378.
[17] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 288–292, 1997.
[18] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE ICASSP*, Munich, Germany, 1997, pp. 187–190.
[19] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Dept. Eng., Brown Univ., Providence, RI, 2000.
[20] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
[21] H. Samet, *Applications of Spatial Data Structures*. Boston, MA: Addison-Wesley, 1990.
[22] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput., Speech Lang.*, vol. 11, no. 2, pp. 91–126, 1997.
[23] J. B. Allen and D. A. Barkeley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 5, pp. 943–950, 1979.
[24] R. Duraiswami, N. A. Gumerov, D. N. Zotkin, and L. S. Davis, "Efficient evaluation of reverberant sound fields," in *Proc. IEEE WASPAA*, New Paltz, NY, Oct. 2001, pp. 203–206.
[25] D. N. Zotkin, R. Duraiswami, L. S. Davis, and I. Haritaoglu, "An audio-video front end for multimedia applications," in *Proc. IEEE SMC*, Nashville, TN, 2000, pp. 786–791.
[26] D. N. Zotkin, R. Duraiswami, V. Philomin, and L. S. Davis, "Smart videoconferencing," in *Proc. IEEE ICME*, vol. 3, New York, 2000, pp. 1597–1600.

Dmitry N. Zotkin was born in Moscow, Russia, in 1973. He received the B.S./M.S. degrees in applied mathematics and physics from the Moscow Institute of Physics and Technology, Dolgoprudny, Moscow, Russia, in 1996, and the M.S. and Ph.D. degrees in computer science from University of Maryland, College Park, in 1999 and 2002, respectively.

He is currently an Assistant Research Scientist at the Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies (UMIACS), University of Maryland, College Park. His current research interests are in multichannel signal processing for tracking and multimedia. He is also working in the general area of spatial audio, including virtual auditory scene synthesis, customizable virtual auditory displays, perceptual processing interfaces, and associated problems.

Ramani Duraiswami (M'99) received the B.Tech. degree in mechanical engineering from Indian Institute of Technology, Bombay, India, in 1985, and the Ph.D. in mechanical engineering and applied mathematics from The Johns Hopkins University, Baltimore, MD, in 1991.

He is a Member of the Faculty in the Department of Computer Science and in the Institute for Advanced Computer Studies (UMIACS) at the University of Maryland, College Park. He is the Director of the Perceptual Interfaces and Reality Laboratory. His research interests are broad and include spatial audio, virtual environments, microphone arrays, computer vision, statistical machine learning, fast multiple methods, and integral equations.