A Generalized Steered Response Power Method for Computationally Viable Source Localization

Jacek P. Dmochowski, Jacob Benesty, *Senior Member, IEEE*, and Sofiène Affes, *Senior Member, IEEE*

***Abstract*—The process of locating an acoustic source given measurements of the sound field at multiple microphones is of significant interest as both a classical array signal processing problem, and more recently, as a solution to the problems of automatic camera steering, teleconferencing, hands-free processing, and others. Despite the proven efficacy of steered-beamformer approaches to localization in harsh conditions, their practical application to real-time settings is hindered by undesirably high computational demands. This paper presents a computationally viable implementation of the steered response power (SRP) source localization method. The conventional approach is generalized by introducing an inverse mapping that maps relative delays to sets of candidate locations. Instead of traversing the three-dimensional location space, the one-dimensional relative delay space is traversed; at each lag, all locations which are inverse mapped by that delay are updated. This means that the computation of the SRP map is no longer performed sequentially in space. Most importantly, by subsetting the space of relative delays to only those that achieve a high level of cross-correlation, the required number of algorithm updates is drastically reduced without compromising localization accuracy. The generalization is scalable in the sense that the level of subsetting is an algorithm parameter. It is shown that this generalization may be viewed as a spatial decomposition of the SRP energy map into weighted basis functions—in this context, it becomes evident that the full SRP search considers all basis functions (even the ones with very low weighting). On the other hand, it is shown that by only including a few basis functions per microphone pair, the SRP map is quite accurately represented. As a result, in a real environment, the proposed generalization achieves virtually the same anomaly rate as the full SRP search while only performing 10% the amount of algorithm updates as the full search.**

***Index Terms*—Microphone arrays, source localization, spectral estimation, steered response power (SRP).**

I. Introduction

MANY applications, ranging from teleconferencing systems to artificial perception require the localization of one or more acoustic sources. Arrays of microphones serve as the spatial aperture needed to process the auditory scene and yield source location estimates. The processes of microphone array beamforming [1] and acoustic source localization are somewhat intertwined: in order to realize the signal enhancement offered by beamformers, the steering of these beamformers needs to closely correspond to the source location(s). Conversely, a robust method for localizing an acoustic source is that of maximizing the power of a steered beam across the location space.

Even though the idea of using a steered beamformer to generate location estimates originated in the 1970s [2]–[6], it was not until the early 1990s that the idea behind the steered response power (SRP) method emerged. Omologo and Svaizer [7], [8] introduced the so-called "global coherence field" (GCF), a representation that maps a hypothesized source location to the level of cross-correlation experienced at the relative delays which correspond to that location. The GCF technique varies from the SRP method [9] virtually only by name.

In the present day, steered-beamformer approaches to source localization are quickly becoming the preferred technique employed for localization applications. There are several reasons for this. The steered-conventional beamformer approach may easily be formulated in the context of multiple cross-correlation functions—this formulation is precisely the SRP method. In this context, it becomes apparent that maximizing the power of a steered conventional beamformer across the location space results in the simple yet effective combining of the various microphone pair cross-correlation measurements. In fact, it is shown in [10] that in moderately and highly reverberant environments, the SRP method surprisingly outperforms high-resolution minimum variance [11] and linear predictive spectral estimators. SRP methods utilize all cross-correlation measurements (i.e., not just the peak value for each microphone pair) in generating location estimates, in what some term the "principle of least commitment" [12]. In addition, the SRP technique is compatible with the generalized cross-correlation approach (GCC) of [13], in that prefiltering of the cross-spectra may be performed in the cross-correlation computation before performing the search process inherent to the SRP technique. For example, when the phase transform (PHAT) prefiltering is applied before computing the cross-correlations, the resulting algorithm is termed "SRP-PHAT" in [14]. Note that even though the GCF method previously proposed in [8] employs the PHAT function, the term "SRP-PHAT" is more commonly used. SRP-PHAT represents one of, if not, the most widely studied and implemented modern localization algorithms (see [15] and [16], for example). The only difference between SRP and SRP-PHAT is that the signals are prewhitened in the latter; therefore, in this paper, the term "SRP" is used interchangeably with "SRP-PHAT." In general, SRP algorithms are superior in combating the adverse effects of background noise and reverberation compared to approaches that are based on intersecting or least-squares fitting the time-differences-of-arrival (TDOA)—a performance comparison between the two categories is given in [10].

There remain two significant problems with SRP-based acoustic source localizers. Most importantly, the computational requirements of the technique are large and make real-time implementation difficult. Recently, there have been two attempts

Manuscript received January 5, 2007; revised July 3, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hiroshi Sawada.

The authors are with the INRS-EMT, University of Quebec, Montreal, QC H5A 1K6, Canada.

Digital Object Identifier 10.1109/TASL.2007.906694

1558-7916/$25.00 © 2007 IEEE

to reduce the computational requirements of the intrinsic SRP search process. In [17], a technique based on a hierarchical search procedure that eliminates candidate locations as the search continues is proposed. The search begins with a coarse grid and becomes finer and finer with time; with each search, the location space is pruned. Unfortunately, it is shown in [18] that the hierarchical localization method exhibits an undesirably high sensitivity to reverberation at reverberation times above 300 ms. In other words, the technique reduces the computational load, but trades off the original robustness benefit of the SRP-PHAT method. In [19], a second proposed solution to the computational load problem is presented: a spherical intersection of significant relative delays is performed first to generate an initial, smaller set of candidate locations. The SRP search is then performed on this initial subset, thus reducing the computational time of the overall search process. This proposed solution has the drawback of requiring the source to be in the near-field, which is a major limitation given that microphone arrays are often targeted as solutions to far-field (for example, distant talker hands-free) applications. To that end, there exists a need for a *robust, far-field applicable, and computationally viable* acoustic source localizer.

Note that one of the positive attributes of the SRP-PHAT algorithm is that it defers the final decision (i.e., choosing the estimated location) until all the cross-correlation estimates have been taken into account. In other words, a "hard-decision" is made as late as possible—after all computations. In the hierarchical localization technique, hard-decisions are made at the end of every subsearch: as soon as the map is pruned, locations that are pruned off cannot be chosen as the final estimate. Similarly, in the spherical intersection/SRP hybrid approach, only those locations which are yielded by the intersection procedure may be chosen as the final estimate; again, hard decisions are made early in the process.

Most importantly, as will be shown in this paper, these prior approaches do not explicitly address the fact that in the traditional SRP search process, many needless computations are performed, in that many of the cross-correlation measurements do not contribute significantly to the final energy map. Additionally, even these improved search methods involve an iterative traversal of the multidimensional location space. The search is sequential in space—that is, the energy emanating from location 1 is computed first, followed by location 2, etc. As a result, the entire location space must be entirely computed before we may make a decision as to the optimal location estimate.

The second problem faced by SRP methods stems from the directional nature of human sound sources. When a speaker is oriented in such a way that he/she is facing a reflective barrier, the direct-path signal component may actually arrive at the array with a lesser power than a first-order reflection, even though the distance traveled by the reflection is larger. As a result, the direction of the reflection will exhibit the greatest steered power, leading to a serious localization error. This problem is not tackled in this paper but reserved for a future publication.

This paper deals with the computational issue of the SRP technique for acoustic source localization. The classical SRP algorithm is first reformulated in terms of the inverse mapping that relates relative delays to source locations. This reformulation allows for a deeper insight into the computations involved in the SRP algorithm. Moreover, it allows for the transformation of the iterative procedure from the multidimensional location space to the one-dimensional relative delay space. As a result, the computation of the SRP energy map is no longer performed sequentially in space. By restricting the traversed relative delays to only those that are deemed to potentially inverse map to the source location, the computational requirements of the proposed search are drastically reduced. The restriction of the relative delays is scalable, thus leading to a generalized SRP method—the conventional SRP method is a particular case of this generalized paradigm. From simulation and experimental results, it is shown that even with the drastic reduction in computational load, the performance of the proposed algorithm is virtually identical to that of the conventional SRP method. Additionally, a rigorous definition of the computational load of the proposed and conventional SRP approaches is presented, with results using circular and spherical array geometries given.

II. Signal Model and Notation

A. Signal Model

Assume an array of M microphone elements, distributed in some fashion in three-dimensional space, whose outputs are denoted by $x_m(t), m = 0, 1, \ldots, M-1$, where t denotes time. The spherical coordinate system is used, where the range is denoted by r, elevation by ϕ, and azimuth by θ. In localization applications, it is more convenient to work in spherical coordinates, as certain assumptions allow us to reduce the dimensionality of the location space.

Consider a signal source located at $(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})$. Propagation of the signal to microphone m is modeled as

$$x_m(t) = \alpha_m(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s}) s[t - f_{0,m}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})] + v_m(t) \tag{1}$$

where x_m is the received microphone output (microphone 0 serves as the reference), t represents time, s is the desired signal, $v_m(t)$ is the additive noise at microphone m which includes any background or sensor noise, as well as reverberation, α_m models attenuation of the desired signal at microphone m due to propagation effects, and the function $f_{i,j}$ relates the source location to the relative delay between microphones i and j

$$f_{i,j}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s}) = \frac{1}{c}[d_{\mathrm{s},j}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s}) - d_{\mathrm{s},i}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})] \tag{2}$$

where c is the speed of sound and $d_{\mathrm{s},i}$ is the distance between the sound source and microphone i, as shown by (3) at the bottom of the page, and (r_i, ϕ_i, θ_i) are the spherical co-ordinates of microphone i. Since only the phase $f_{i,j}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})$ conveys reliable

$$d_{\mathrm{s},i}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s}) = \sqrt{(r_\mathrm{s} \sin\phi_\mathrm{s} \cos\theta_\mathrm{s} - r_i \sin\phi_i \cos\theta_i)^2 + (r_\mathrm{s} \sin\phi_\mathrm{s} \sin\theta_\mathrm{s} - r_i \sin\phi_i \sin\theta_i)^2 + (r_\mathrm{s} \cos\phi_\mathrm{s} - r_i \cos\phi_i)^2} \tag{3}$$

location information, throughout the rest of this paper, the signal attenuation term $\alpha_m(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})$ is ignored.

When the source is located in the far-field, the incoming wave front may be assumed to be planar, thus making $f_{i,j}$ independent of the range

$$f_{i,j}(r_\mathrm{s}, \phi_\mathrm{s}, \theta_\mathrm{s})\,|_\mathrm{farfield} = f_{i,j}(\phi_\mathrm{s}, \theta_\mathrm{s}) \approx \frac{1}{c}\boldsymbol{\zeta}^T_{\phi_\mathrm{s},\theta_\mathrm{s}}(\mathbf{p}_j - \mathbf{p}_i) \tag{4}$$

where

$$\boldsymbol{\zeta}_{\phi_\mathrm{s},\theta_\mathrm{s}} = [\sin\phi_\mathrm{s}\cos\theta_\mathrm{s} \quad \sin\phi_\mathrm{s}\sin\theta_\mathrm{s} \quad \cos\phi_\mathrm{s}]^T \tag{5}$$

is the unit direction vector of the source signal, and

$$\begin{aligned}\mathbf{p}_i &= [p_i^x \quad p_i^y \quad p_i^z]^T \\ &= [r_i\sin\phi_i\cos\theta_i \quad r_i\sin\phi_i\sin\theta_i \quad r_i\cos\phi_i]^T\end{aligned} \tag{6}$$

is the location vector of microphone i in Cartesian coordinates where p_i^x, p_i^y, and p_i^z are the x-, y-, and z-components of the location vector, respectively.

The received microphone signals are sampled, and the forthcoming signal processing is performed on discrete signals, which are used throughout the rest of the paper

$$x_m(n) = x_m(nT_\mathrm{sf}), \quad m = 0, 1, \ldots, M-1 \tag{7}$$

where $T_\mathrm{sf} = f_\mathrm{sf}^{-1}$ is the sampling period and f_sf is the sampling frequency. The processing is performed once per frame, with each frame consisting of N_f samples.

B. Notation

We now describe the notation pertaining to the search process inherent in the SRP technique which is outlined in the next section. L denotes the location space, and in the most general case

$$\begin{aligned}L = L_\rho \times L_\varphi \times L_\vartheta = &\left\{\rho_0, \rho_1, \ldots, \rho_{N_\rho}\right\} \\ &\times \left\{\varphi_0, \varphi_1, \ldots, \varphi_{N_\varphi}\right\} \times \left\{\vartheta_0, \vartheta_1, \ldots, \vartheta_{N_\vartheta}\right\}\end{aligned} \tag{8}$$

where $L_\rho = \{\rho_0, \rho_1, \ldots, \rho_{N_\rho}\}$ are the ranges of the candidate locations with cardinality $|L_\rho| = N_\rho, L_\varphi = \{\varphi_0, \varphi_1, \ldots, \varphi_{N_\varphi}\}$ are the elevation coordinates of the candidate locations with cardinality $|L_\varphi| = N_\varphi, L_\vartheta = \{\vartheta_0, \vartheta_1, \ldots, \vartheta_{N_\vartheta}\}$ are the azimuthal coordinates of the candidate locations with cardinality $|L_\vartheta| = N_\vartheta$, and × denotes the Cartesian vector product. The number of potential locations is given by the cardinality of L

$$|L| = N_\rho N_\varphi N_\vartheta. \tag{9}$$

The operator $|\cdot|$ denotes the cardinality operation when the argument is a set; when the argument is a scalar, $|\cdot|$ denotes the magnitude operation. In the far-field case, the location search is two-dimensional (azimuth and elevation)

$$L\,|_\mathrm{farfield} = \left\{\varphi_0, \varphi_1, \ldots, \varphi_{N_\varphi}\right\} \times \left\{\vartheta_0, \vartheta_1, \ldots, \vartheta_{N_\vartheta}\right\} \tag{10}$$

and

$$|L|_\mathrm{farfield} = N_\varphi N_\vartheta. \tag{11}$$

For a resolution of 1 degree in both elevation and azimuth, $|L| = 180 \cdot 360 = 64\,800$.

In addition, P denotes the set of all unique (order-independent) pairs of microphones, indexed by a couplet (i, j) which refers to the microphone pair formed by microphones i and j. For $M = 4$ microphones, $|P| = 6$ and

$$P = \{(0,1), (0,2), (0,3), (1,2), (1,3), (2,3)\}. \tag{12}$$

In general, the cardinality of P is

$$|P| = \binom{M}{2}. \tag{13}$$

For each microphone pairing, the set of physically realizable relative delays is given by

$$\begin{aligned}D_{i,j} = \{&-\tau_\mathrm{max}^{i,j}, -\tau_\mathrm{max}^{i,j} + 1, \ldots, -1, 0, 1, \ldots, \\ &\tau_\mathrm{max}^{i,j} - 1, \tau_\mathrm{max}^{i,j}\}\end{aligned} \tag{14}$$

where

$$\tau_\mathrm{max}^{i,j} = \mathrm{round}\left(\frac{f_\mathrm{sf}}{c}\|\mathbf{p}_i - \mathbf{p}_j\|\right) \tag{15}$$

is the maximum physically realizable relative delay between microphones i and j and $\mathrm{round}(\cdot)$ denotes the rounding operation.

III. Steered Beamformer Source Localization

The idea behind localizing an acoustic source using a steered beamformer is based on the assumption that the location of the signal source radiates more energy than all other locations. Assuming an unobstructed and omnidirectional source, this is always the case. In a reverberant environment with a directional source, this assumption may not always hold; nevertheless, the analysis of this fact is complicated and deferred for a future publication.

The output of a delay-and-sum beamformer steered to a location $(\rho, \varphi, \vartheta)$ is given by

$$z_{\rho,\varphi,\vartheta}(n) = \sum_{m=0}^{M-1} w_m x_m[n + f_{0,m}(\rho, \varphi, \vartheta)]. \tag{16}$$

The delays $f_{0,m}(\rho, \varphi, \vartheta)$ steer the beamformer to the desired direction of arrival (DOA), while the beamformer weights w_m help shape the beam accordingly. Assuming that $w_m = 1, m = 0, 1, \ldots, M-1$, the output power of the beamformer is then given by

$$E\left\{z^2_{\rho,\varphi,\vartheta}(n)\right\} = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} R_{x_i,x_j}[f_{i,j}(\rho, \varphi, \vartheta)] \tag{17}$$

where $E\{\,\cdot\,\}$ denotes mathematical expectation and

$$R_{x_i,x_j}(\tau) = E\{x_i(n)x_j(n+\tau)\} \tag{18}$$

is the cross-correlation function for two jointly wide-sense stationary real random processes and $f_{i,j} = f_{0,j} - f_{0,i}$. From (17), it is seen that the output power of a delay-and-sum beamformer is equal to the summation across all microphone pairs of pairwise cross-correlation functions. Thus, the estimate of the source location (assuming a single source) is given by

$$\begin{aligned}(\hat{r}_{\rm s}, \hat{\phi}_{\rm s}, \hat{\theta}_{\rm s}) &= \arg \max_{(\rho,\varphi,\vartheta)} E\{z^2_{\rho,\varphi,\vartheta}(n)\} \\ &= \arg \max_{(\rho,\varphi,\vartheta)} \sum_{i=0}^{M-1}\sum_{j=0}^{M-1} R_{x_i,x_j}[f_{i,j}(\rho,\varphi,\vartheta)]. \end{aligned} \tag{19}$$

IV. SRP and SRP-PHAT Algorithms

We now delve into describing the SRP algorithm, which implements the optimization given by (19). The traditional SRP technique consists of two distinct phases: the computation of cross-correlation functions and the search process which attempts to locate the source.

A. Computation of Cross Correlations

In the first phase, the cross-correlation functions $R_{x_i,x_j}(\tau)$ are computed for all unique microphone pairs $(i,j) \in P$ and the set of all physically realizable time lags (relative delays) pertaining to each microphone pairing $\tau \in D_{i,j}$.

The computation of the cross-correlation functions is typically performed in the frequency-domain via the inverse fast Fourier transform (IFFT)

$$\begin{aligned} R_{x_i,x_j}(\tau) &= \sum_{k=0}^{N_{\rm f}-1} G_{x_i,x_j}(k) e^{j2\pi \frac{k}{N_{\rm f}}\tau} \\ &= \sum_{k=0}^{N_{\rm f}-1} X_i(k)X_j^*(k) e^{j2\pi \frac{k}{N_{\rm f}}\tau} \end{aligned} \tag{20}$$

where $G_{x_i,x_j}(k) = X_i(k)X_j^*(k)$ is the cross-spectrum between channels i and j, k is the discrete frequency index, and $X_i(k)$ is the fast Fourier transform (FFT) of $x_i(n)$.

By computing the cross-correlation function in the frequency-domain, a prefiltering operation may be implemented according to the GCC method [13]

$$R^{\rm g}_{x_i,x_j}(\tau) = \sum_{k=0}^{N_{\rm f}-1} \psi_{\rm g}(k) G_{x_i,x_j}(k) e^{j2\pi \frac{k}{N_{\rm f}}\tau} \tag{21}$$

where $\psi_{\rm g}(k)$ is the prefilter and $R^{\rm g}_{x_i,x_j}$ is termed the "generalized cross-correlation function." Many choices exist for the nature of the filtering provided by $\psi_{\rm g}(k)$—a popular choice is that of the PHAT weighting function, given by $\psi_{\rm PHAT}(k) = (1/|G_{x_i,x_j}(k)|)$. The resulting cross-correlation estimate is given by

$$R^{\rm PHAT}_{x_i,x_j}(\tau) = \sum_{k=0}^{N_{\rm f}-1} \frac{G_{x_i,x_j}(k)}{|G_{x_i,x_j}(k)|} e^{j2\pi \frac{k}{N_{\rm f}}\tau}. \tag{22}$$

The PHAT weighting function whitens the signal in the sense that it removes the magnitude spectrum from the computation of the cross-correlation: the filtered cross-spectrum has a flat magnitude spectrum and the same phase spectrum as the unfiltered cross correlation. When the cross correlations involved in the steered-beamformer localization are computed using the GCC-PHAT method, the resulting algorithm is termed "SRP-PHAT."

TABLE I
Conventional Search Algorithm

```
initialization:
    for all $(\rho,\varphi,\vartheta) \in L$ , $S^{\rm SRP}(\rho,\varphi,\vartheta) := 0$

search:
    for all $(\rho,\varphi,\vartheta) \in L$
        for all $(i,j) \in P$
            look up $\tau^{i,j}_{\rho,\varphi,\vartheta} = \mathrm{round}\left[f_{i,j}(\rho,\varphi,\vartheta)\right]$
            update $S^{\rm SRP}(\rho,\varphi,\vartheta) := S^{\rm SRP}(\rho,\varphi,\vartheta) + R^{\rm g}_{x_i,x_j}\left(\tau^{i,j}_{\rho,\varphi,\vartheta}\right)$

    $\left(\hat{r}_s, \hat{\phi}_s, \hat{\theta}_s\right) := \arg\max_{(\rho,\varphi,\vartheta)\in L} S^{\rm SRP}(\rho,\varphi,\vartheta)$
```

There are a total of $\sum_{(i,j)\in P} |D_{i,j}|$ cross correlations to compute. No matter what weighting function (if any) is used to compute the cross correlations, the results of the cross correlations are stored in some form of look-up table. Note that the computational requirements of this first phase of the SRP approach are no more than those of simple TDOA-based triangulators.

B. SRP Search

The conventional SRP search process is outlined using the pseudocode in Table I. The location space is traversed iteratively, element-by-element. For each location $(\rho,\varphi,\vartheta) \in L$ and microphone pair $(i,j) \in P$, a lookup procedure translates (ρ,φ,ϑ) to a relative delay

$$\tau^{i,j}_{\rho,\varphi,\vartheta} = \mathrm{round}[f_{i,j}(\rho,\varphi,\vartheta)]. \tag{23}$$

The value of $\tau^{i,j}_{\rho,\varphi,\vartheta}$ corresponds to the discrete relative delay experienced between microphones i and j if the source is located at (ρ,φ,ϑ).

The steered response power at location (ρ,φ,ϑ) is then computed as

$$\begin{aligned} S^{\rm SRP}(\rho,\varphi,\vartheta) &= \sum_{(i,j)\in P} R^{\rm g}_{x_i,x_j}\left(\tau^{i,j}_{\rho,\varphi,\vartheta}\right) \\ &= \sum_{i=0}^{M-1}\sum_{j=i+1}^{M-1} R^{\rm g}_{x_i,x_j}\left(\tau^{i,j}_{\rho,\varphi,\vartheta}\right). \end{aligned} \tag{24}$$

The summation over only unique microphone pairs P removes the redundant microphone pairs and autocorrelation terms ("DC" components) from the computation of the steered power.

TABLE II
PROPOSED SEARCH ALGORITHM

```
initialization:
    for all $(\rho,\varphi,\vartheta) \in L$ , $S^{\mathrm{SRP}}(\rho,\varphi,\vartheta) := 0$

search:
    for all $(i,j) \in P$
        for all $\tau \in C_{i,j} \subseteq D_{i,j}$
            look up $f_{i,j}^{-1}(\tau)$
            for all $(\rho,\varphi,\vartheta) \in f_{i,j}^{-1}(\tau)$
                update $S^{\mathrm{SRP}}(\rho,\varphi,\vartheta) := S^{\mathrm{SRP}}(\rho,\varphi,\vartheta) + R^{\mathrm{g}}_{x_i x_j}(\tau)$

$\left(\hat{r}_{\mathrm{s}}, \hat{\phi}_{\mathrm{s}}, \hat{\theta}_{\mathrm{s}}\right) := \arg\max_{(\rho,\varphi,\vartheta)} S^{\mathrm{SRP}}(\rho,\varphi,\vartheta)$
```

After the last location is traversed, the SRP spatial spectrum or "energy map" is fully known. The candidate location with the highest steered power is designated as the location estimate

$$\begin{aligned}(\hat{r}_s,\hat{\phi}_s,\hat{\theta}_s) &= \arg\max_{(\rho,\varphi,\vartheta)\in L} S^{\mathrm{SRP}}(\rho,\varphi,\vartheta)\\ &= \arg\max_{(\rho,\varphi,\vartheta)\in L} \sum_{i=0}^{M-1}\sum_{j=i+1}^{M-1} R^{\mathrm{g}}_{x_i,x_j}\left(\tau^{i,j}_{\rho,\varphi,\vartheta}\right).\end{aligned} \tag{25}$$

V. Proposed Generalization of SRP Technique

This section presents the proposed generalization of the conventional SRP method, and shows how this generalization, among other things, facilitates a significant reduction in computational load.

The generalization affects only the search portion of the SRP approach—the cross-correlation functions are computed as usual. At the heart of the generalization is the inverse mapping that maps relative delays to locations. We define this mapping by

$$f_{i,j}^{-1}(\tau) = \{(\rho,\varphi,\vartheta) \in L | f_{i,j}(\rho,\varphi,\vartheta) = \tau\}. \tag{26}$$

The inverse mapping $f_{i,j}^{-1}$ maps a single relative delay (integer) to a discrete set of candidate locations. Since the inverse map is based purely on array geometry, it may be computed offline *a priori* and stored in memory. Note that the memory requirements of this inverse look-up table are identical to those of the conventional forward look-up table that maps locations to relative delays. There is one caveat: since each relative delay maps to a variable amount of locations, an irregular data structure should be used. Alternatively, a table of indices may be used in conjunction with the inverse look-up table to store the number of locations associated with each relative delay.

The proposed search is outlined using pseudocode in Table II. Throughout the proposed search, instead of traversing the three-dimensional location space, the one-dimensional relative delay space is traversed. As each delay (lag) is traversed, all locations which are inverse mapped by that delay are "simultaneously" updated. This means that *the computation of the SRP energy map is no longer performed sequentially in space*. In other words, as the various relative delays are traversed, the energy map is being built-up at the corresponding inverse-mapped candidate locations. The more relative delays and microphones that we traverse, the more accurate the map.

A. Subsetting the Relative Delay Space

How does this generalization reduce computational load? The key variable in the proposed implementation is $C_{i,j}$, a subset of $D_{i,j}$, which is the set of relative delays that is traversed in the proposed search process for microphone pair (i,j). In the proposed method

$$C_{i,j} \subsetneq D_{i,j}. \tag{27}$$

Instead, the set of traversed delays is restricted to a proper subset of all physically realizable relative delays. This proper subset includes the lags that produce high levels of cross correlation.

The traversal of the relative delay space is restricted to a subset that includes the lag that produces the peak in the cross correlation for each microphone pair. Denote this optimal lag by

$$\hat{\tau}^{i,j} = \arg\max_{\tau} R^{\mathrm{g}}_{x_i,x_j}(\tau). \tag{28}$$

The subset of traversed relative delays is then given by

$$\begin{aligned}C_{i,j} = \{&\hat{\tau}^{i,j} - p, \hat{\tau}^{i,j} - p + 1, \ldots, \hat{\tau}^{i,j} - 1, \hat{\tau}^{i,j}, \hat{\tau}^{i,j}\\ &+ 1, \ldots, \hat{\tau}^{i,j} + p - 1, \hat{\tau}^{i,j} + p\} \cap D_{i,j}\end{aligned} \tag{29}$$

which is a set of relative delays centered about $\hat{\tau}^{i,j}$. The parameter p determines how many adjacent lags are involved in the search process. The $\cap$ denotes intersection and the intersection with $D_{i,j}$ must be included to account for cases where $\hat{\tau}^{i,j}$ occurs near the edges of $D_{i,j}$. In those cases, $|C_{i,j}| \leq 2p+1$. The parameter p is crucial in that it determines both the reduction in computational load as well as the resulting source localization accuracy. This will be clear from the upcoming results.

In other words, the traversal of relative delays is only over those delays that are judged to be of significant value, or those that are deemed to potentially inverse map to a set that includes the true location. Why should we consider all relative delays when we know that only some of them may possibly correspond to the actual source location? When $R^{\mathrm{g}}_{x_i,x_j}(\tau)$ is very small or negative, we can be confident that it does not inverse map to the source. Therefore, the updating of the locations which such a delay inverse maps to is omitted. Surely it is wasteful to search over the entire space of relative delays, and this fact is the basis for the reduction in computational load offered by the generalization. By restricting the traversal of the relative delay space, we are not wasting computational time updating locations far away from the peak of the energy map. This will be illustrated in the results section.

Notice that when $C_{i,j} = D_{i,j}$, the proposed search is the conventional (full) SRP search, just performed in different order. When $C_{i,j} = \{\hat{\tau}^{i,j}\}, \forall (i,j) \in P$, the proposed search involves only those locations which are inverse mapped by the optimal (peak) cross-correlation lag of at least one microphone pair. The search is scalable in the sense of the cardinality of $C_{i,j}$.

B. Handling Multiple Sources and Large Microphone Spacing

The subsetting of the relative delay space should be matched to the characteristics of the acoustic environment; for the nominal scenario of an array which obeys the spatial aliasing criteria (i.e., properly spaced microphones) and a single active speaker, the subsetting described in the previous subsection is quite appropriate (this will be shown in the forthcoming simulation and experimental results). This section describes how the proposed algorithm described in the previous subsection may be extended to handle environments with multiple active sound sources, or arrays with large intermicrophone spacings which do not obey spatial aliasing requirements.

It is expected that the presence of multiple sources or large intermicrophone spacing will lead to multiple, widely spaced peaks in the cross-correlation functions. Therefore, iterating across adjacent lags of only the highest peak seems inappropriate. To that end, the traversal of adjacent lags is performed for the significant peaks in the cross-correlation function. Assuming that we have chosen to select N_{p} active sources or peaks using a gradient search or similar procedure, we denote the set of cross-correlation peaks in microphone pair (i,j) by

$$\kappa^{i,j} = \left\{\hat{\tau}_1^{i,j}, \hat{\tau}_2^{i,j}, \ldots, \hat{\tau}_{N_{\mathrm{p}}}^{i,j}\right\}. \tag{30}$$

The subset of traversed relative delays then follows as the union

$$\begin{aligned} C_{i,j} = \Big(&\left\{\hat{\tau}_1^{i,j} - p, \ldots, \hat{\tau}_1^{i,j}, \ldots, \hat{\tau}_1^{i,j} + p\right\} \\ &\cup \left\{\hat{\tau}_2^{i,j} - p, \ldots, \hat{\tau}_2^{i,j}, \ldots, \hat{\tau}_2^{i,j} + p\right\} \\ &\cup \ldots \cup \left\{\hat{\tau}_{N_{\mathrm{p}}}^{i,j} - p, \ldots, \hat{\tau}_{N_{\mathrm{p}}}^{i,j}, \ldots, \hat{\tau}_{N_{\mathrm{p}}}^{i,j} + p\right\}\Big) \\ &\cap D_{i,j}. \end{aligned} \tag{31}$$

Note that whether employing the subsetting of (29) or (31), the relative delays that are omitted from the search procedure correspond to relatively low values of cross correlation, and thus their omission does not significantly distort the SRP energy map.

VI. Computational Load Comparison

A. Computation of Cross-Correlation Functions

Before providing a comparison in complexity between the conventional and proposed searches, it is worthwhile to consider the computational load of the first stage of the SRP algorithm: the cross-correlation function computation. This load is common to the conventional and proposed SRP algorithms.

For all M channels, $X_i(k)$ needs to be computed for N_{f} values of k. The cross-spectra are then generated by point-wise complex multiplication of the FFTs: in order to compute the cross-spectrum of every unique microphone pair, $|P|N_{\mathrm{f}}$ complex multiplications are required.

Assuming that the PHAT weighting function is desired, each cross-spectral sample needs to be divided by its magnitude. Thus, a total of $|P|N_{\mathrm{f}}$ magnitude operations and $|P|N_{\mathrm{f}}$ real divisions are required. The computation of the generalized cross correlation then follows: for each microphone pair $(i,j) \in P$, the $R^{\mathrm{g}}_{x_i,x_j}(\tau)$ needs to be computed for $|D_{i,j}|$ values of τ.

TABLE III
Complexity of GCC-PHAT Computation

Operation	Number of Operations
compute $X_i(k)$	MN_{f}
complex multiplication	$\lvert P\rvert\, N_{\mathrm{f}}$
complex magnitude	$\lvert P\rvert\, N_{\mathrm{f}}$
division	$\lvert P\rvert\, N_{\mathrm{f}}$
compute $R^{\mathrm{g}}_{x_i,x_j}(\tau)$	$\sum_{(i,j)\in P} \lvert D_{i,j}\rvert$

TABLE IV
Complexity of Proposed and Conventional SRP Searches

Search Algorithm	N_{lookups}	N_{updates}
Conventional	$\lvert L\rvert\, \lvert P\rvert$	$\lvert L\rvert\, \lvert P\rvert$
Proposed	$\sum_{(i,j)\in P} \lvert C_{i,j}\rvert$	$\sum_{(i,j)\in P} \sum_{\tau\in C_{i,j}} \lvert f_{i,j}^{-1}(\tau)\rvert$

The computational load of the calculation of the GCC functions is summarized in Table III.

B. Search Process

From the given pseudocode, the conventional SRP search consists of $|L||P|$ look-up operations, and $|L||P|$ additions (updates).

Referring to the pseudocode of the proposed generalization, the proposed SRP search consists of $\sum_{(i,j)\in P} |C_{i,j}|$ look-up operations, and $\sum_{(i,j)\in P} \sum_{\tau\in C_{i,j}} |f_{i,j}^{-1}(\tau)|$ additions. The look-up operation involves inverse mapping a relative delay to an inverse set of locations. Note that in the proposed search, in order to form the set $C_{i,j}$ for each microphone pair, the relative delay which corresponds to the peak of that microphone pair's cross-correlation function needs to be found via an arg max operation.

The complexity of the proposed and conventional searches are summarized in Table IV, where N_{lookups} and N_{updates} denote the number of required lookups and updates, respectively. Numerical examples using spherical and circular array geometries are given in Sections VIII and IX, respectively.

VII. Spatial Decomposition Formulation

Motivated by the proposed generalization, it is possible to expand the expression for the steered energy of a given location $(\rho, \varphi, \vartheta)$

$$\begin{aligned} S(\rho,\varphi,\vartheta) &= \sum_{(i,j)\in P} R^{\mathrm{g}}_{x_i,x_j}[f_{i,j}(\rho,\varphi,\vartheta)] \\ &= \sum_{(i,j)\in P} \sum_{\tau\in D_{i,j}} R^{\mathrm{g}}_{x_i,x_j}(\tau) \\ &\quad\times \sum_{(r,\phi,\theta)\in f_{i,j}^{-1}(\tau)} \delta[(\rho,\varphi,\vartheta) - (r,\phi,\theta)] \end{aligned} \tag{32}$$

where the basis functions are identified as $\sum_{(r,\phi,\theta)\in f_{i,j}^{-1}(\tau)} \delta [(\rho,\varphi,\vartheta) - (r,\phi,\theta)]$—a function whose value is 1 at locations belonging to the set $f_{i,j}^{-1}(\tau)$ and 0 elsewhere—and the weighting coefficients of the basis functions are $R^{(\mathrm{g})}_{x_i,x_j}(\tau)$. Since the second summation in (32) is over $D_{i,j}$, this refers to

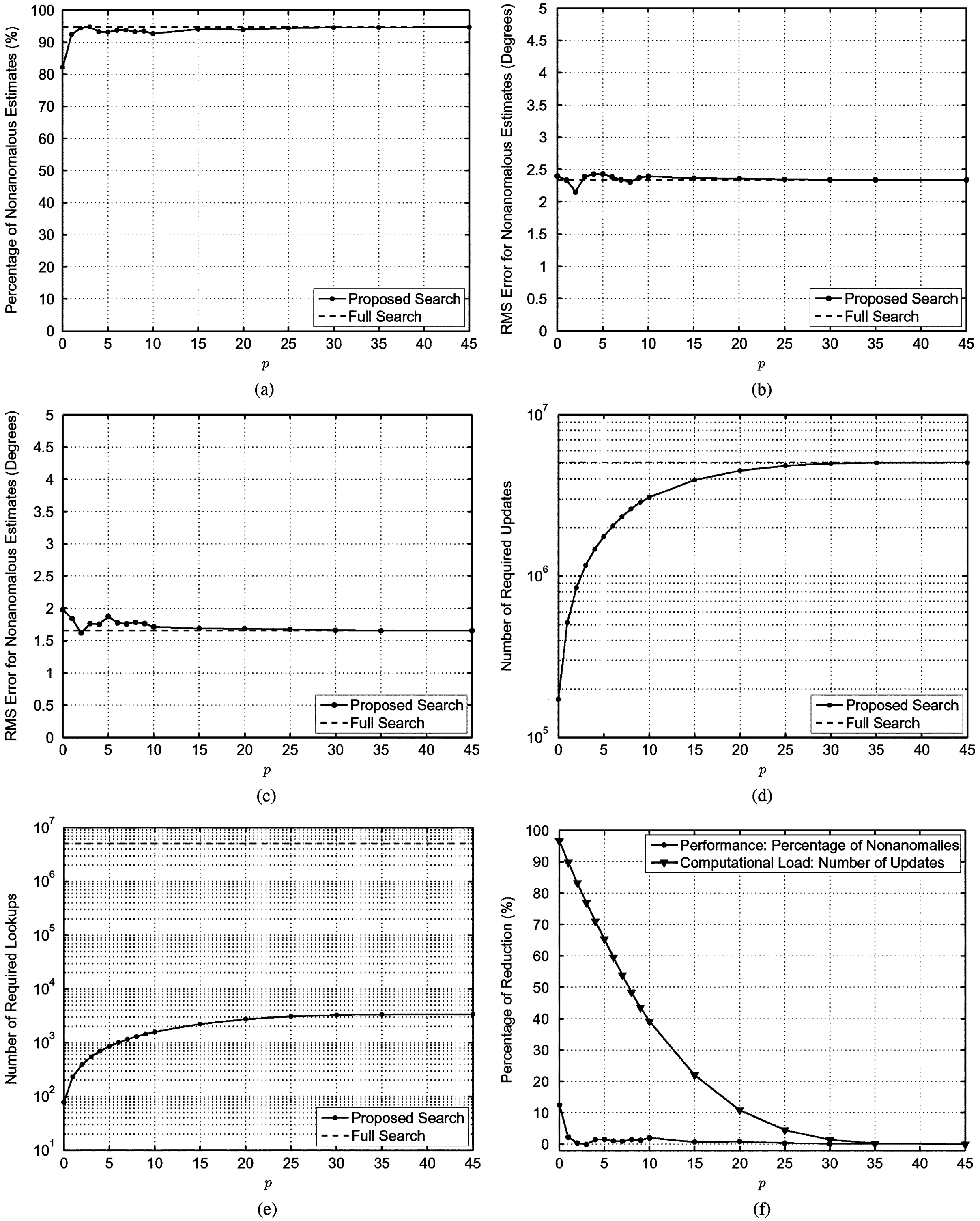


Fig. 1. Performance and computational load of proposed search as a function of p with SNR $= 20$ dB. (a) Percentage of nonanomalous estimates. (b) RMS error in azimuth for nonanomalous estimates. (c) RMS error in elevation for nonanomalous estimates. (d) Number of required updates. (e) Number of required lookups. (f) Corresponding reductions in performance and computational load.

the full SRP map, with all relative delays used. The proposed generalization is indicated by simply switching $D_{i,j}$ to $C_{i,j}$

$$S'(\rho,\varphi,\vartheta) = \sum_{(i,j)\in P}\sum_{\tau\in C_{i,j}} R^{\mathrm{g}}_{x_i,x_j}(\tau) \times \sum_{(r,\phi,\theta)\in f_{i,j}^{-1}(\tau)} \delta[(\rho,\varphi,\vartheta)-(r,\phi,\theta)]. \quad (33)$$

Each basis function is identified by both the microphone pair and lag which defines the corresponding inverse set $f_{i,j}^{-1}(\tau)$; therefore, the basis functions are unique to a given array geometry. Each geometry has $\sum_{(i,j)\in P}|D_{i,j}|$ basis functions; defining $\tau_{\max} = \max\{\tau_{\max}^{0,1}, \tau_{\max}^{0,2}, \ldots, \tau_{\max}^{M-2,M-1}\}$, the number of basis functions is upper-bounded by $|P|(2\tau_{\max}+1)$. In (33), the summation over $C_{i,j}$ means that in the decomposed representation

TABLE V
PERFORMANCE AND COMPUTATIONAL LOAD OF PROPOSED AND CONVENTIONAL SRP SEARCHES WITH SNR = 20 dB

p	$\%_{\mathrm{nonanomalous}}$	$e_{\theta,\mathrm{RMS}}$	$e_{\phi,\mathrm{RMS}}$	N_{updates}	N_{lookups}	$\%_{\mathrm{perf.\ reduction}}$	$\%_{\mathrm{comp.\ reduction}}$
0	82.25	2.34	1.65	172,400	78	12.47	96.59
1	92.47	2.40	1.98	514,990	234	2.25	89.81
2	94.38	2.34	1.84	849,470	390	0.34	83.19
3	94.83	2.15	1.62	1,165,400	545	-0.11	76.94
4	93.26	2.39	1.76	1,464,500	699	1.46	71.03
5	93.15	2.43	1.75	1,755,100	852	1.57	65.28
6	93.71	2.43	1.88	2,045,000	1003	1.01	59.54
7	93.82	2.38	1.77	2,331,700	1153	0.90	53.87
8	93.26	2.34	1.76	2,610,600	1300	1.46	48.35
9	93.48	2.31	1.78	2,855,200	1443	1.24	43.51
10	92.70	2.37	1.76	3,076,700	1584	2.02	39.13
15	94.04	2.39	1.71	3,940,600	2227	0.67	22.04
20	93.93	2.36	1.69	4,507,100	2734	0.79	10.83
25	94.38	2.36	1.68	4,825,000	3072	0.34	4.54
30	94.61	2.35	1.67	4,982,500	3259	0.11	1.42
35	94.61	2.34	1.66	5,041,500	3337	0.11	0.26
43	94.72	2.34	1.65	5,054,400	3354	0	0
full	94.72	2.34	1.65	5,054,400	5,054,400	0	0

of $S(\rho,\varphi,\vartheta)$, the basis functions with low weighting $R^{\mathrm{g}}_{x_i,x_j}(\tau)$ are omitted. By selectively neglecting many basis functions, the SRP energy map is represented accurately with a lesser amount of coefficients—this is another way of viewing the computational reduction of the proposed generalization. Notice that the selection of which basis functions to include in the representation is *adaptive* or data-dependent, in that the results of the cross correlations themselves determine this selection.

VIII. SIMULATION EVALUATION

A. Simulation Environment

The conventional and generalized SRP algorithms are evaluated in a computer simulation. An equiangled open spherical array [20] of $M = 13$ omnidirectional microphones and a radius of 7.62 cm is employed as the spatial aperture. The equiangled scheme samples the azimuth and elevation dimensions uniformly with N_{s} samples in each dimension. The set of azimuth samples is formed as

$$S_\vartheta = \left[0, \frac{2\pi}{N_{\mathrm{s}}}, \ldots, \frac{2\pi(N_{\mathrm{s}}-1)}{N_{\mathrm{s}}}\right]. \tag{34}$$

Similarly, the set of elevation samples is given by

$$S_\varphi = \left[0, \frac{\pi}{N_{\mathrm{s}}}, \ldots, \frac{\pi(N_{\mathrm{s}}-1)}{N_{\mathrm{s}}}\right]. \tag{35}$$

The locations of the microphones then follow from the set $S_\vartheta \times S_\varphi$. For each element in the set, the location of the microphone is easily obtained by converting from spherical to Cartesian coordinates, with each microphone position having the same radial component. In the simulations, $N_{\mathrm{s}} = 4$, leading to 16 samples; however, four of the samples are common (refer to the same location), meaning that there are only $M = 13$ unique spatial samples or microphone positions.

With 13 microphones, $|P| = 78$. Given the array geometry, $\tau_{\max} = 21$; for simplicity, this range of physically realizable delays is applied to all microphone pairs: $|D_{i,j}| := 43, \forall (i,j) \in P$. The proposed search is run for $p \in \{0,1,2,\ldots,9,10,15,20,\ldots,35,43\}$.

A reverberant acoustic environment is simulated using the image model method [21]. The simulated room is rectangular with plane reflective boundaries (walls, ceiling, and floor). Each boundary is characterized by a frequency-independent uniform reflection coefficient which does not vary with the angle of incidence of the source signal. The room dimensions in centimeters are (304.8, 457.2, 381.0). The array is located in the center of the room: the center of the sphere is at (152.4, 228.6, 101.6). The speaker is situated at (254, 406.4, 203.2). The correct azimuth angle of arrival is 60°; the correct elevation is 64°. The distance from the center of the array to the source is 228.6 cm. The source is (correctly) assumed to be in the far-field, reducing the dimensionality of the location space to 2.

Two levels of signal-to-noise ratio (SNR) are simulated: 0 dB and 20 dB. The additive noise is spatially uncorrelated across the array and temporally uncorrelated with itself and the desired signal. For the computation of the SNR, the signal component includes reverberation. The reverberation time is measured using the reverse-time integrated impulse response method of [22]. The frequency-independent reflection coefficients of the walls and ceiling are adjusted to achieve the desired level of reverberation: a 60-dB reverberation decay time (T_{60}) of 600 ms. The desired source signal is convolved with the synthetic impulse responses, and appropriately scaled Gaussian noise is then added at the microphones to achieve the required SNR.

The signal is female English speech with silences removed. The DOA estimates are computed once per 128-ms frame. The sampling rate is chosen to be 48 kHz, resulting in frames of

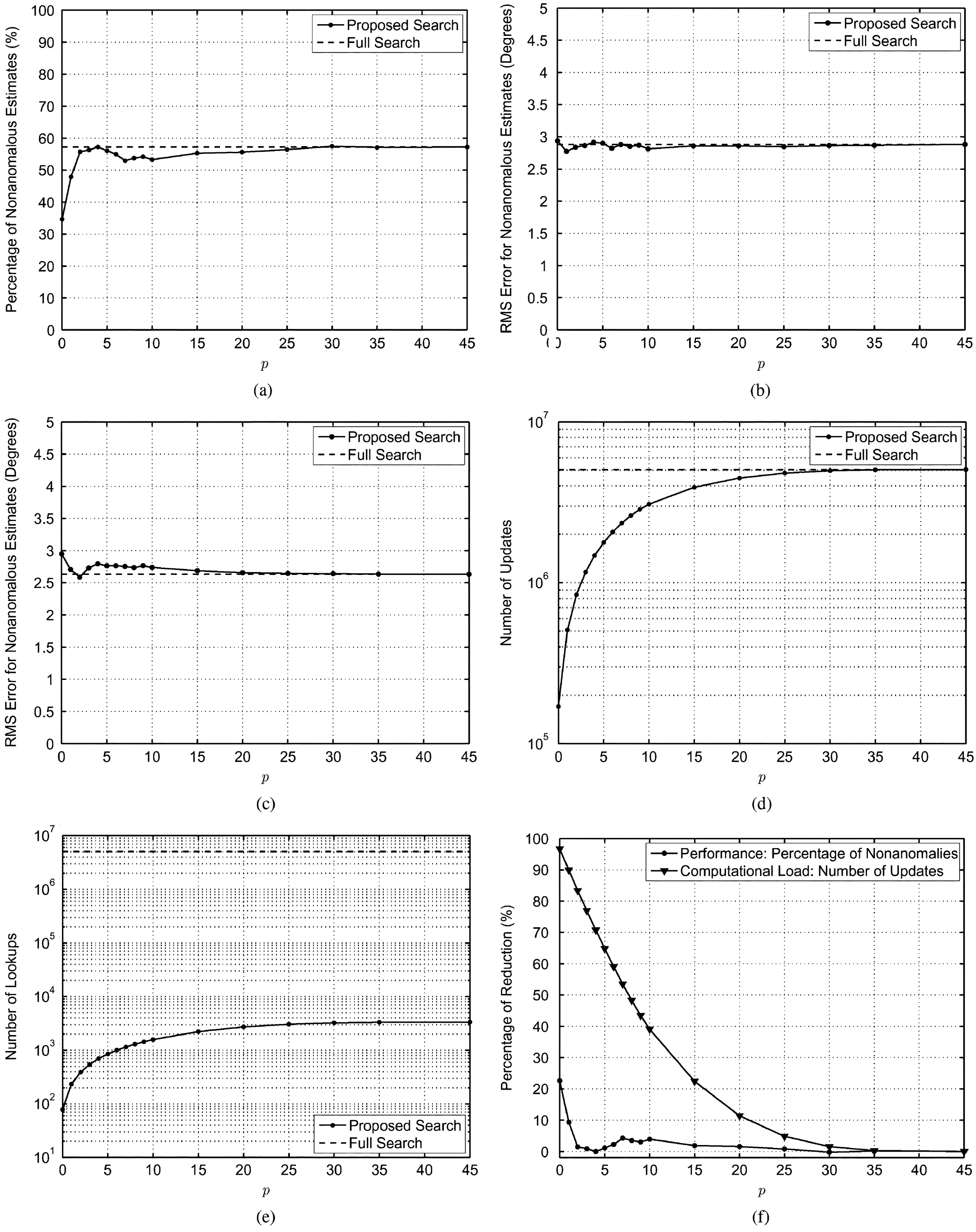


Fig. 2. Performance and computational load of proposed search as a function of p with SNR $= 0$ dB. (a) percentage of nonanomalous estimates. (b) RMS error in azimuth for nonanomalous estimates. (c) RMS error in elevation for nonanomalous estimates. (d) Number of required updates. (e) Number of required lookups. (f) Corresponding reductions in performance and computational load.

$N = 6144$ samples each. A simulation run consisting of $K = 890$ frames is performed. The location space is denoted by

$$L = L_\varphi \times L_\vartheta = \{0^\circ, 1^\circ, 2^\circ, \ldots, 178^\circ, 179^\circ\} \times\{0^\circ, 1^\circ, 2^\circ, \ldots, 358^\circ, 359^\circ\} \quad (36)$$

and thus $|L| = 64\,800$.

To achieve good angular resolution in the resulting SRP spectrum, two approaches may be taken. The cross-correlation measurements may be interpolated; alternatively, redundant microphone pairs may be utilized such that the mapping from a point in space to the corresponding set of expected relative delays is unique: even though two locations may map to the same relative delay at one microphone pair, there are other microphone

TABLE VI
PERFORMANCE AND COMPUTATIONAL LOAD OF PROPOSED AND CONVENTIONAL SRP SEARCHES WITH SNR = 0 dB

p	$\%_{\text{nonanomalous}}$	$e_{\theta,\text{RMS}}$	$e_{\phi,\text{RMS}}$	N_{updates}	N_{lookups}	$\%_{\text{perf. reduction}}$	$\%_{\text{comp. reduction}}$
0	34.61	2.94	2.95	170,301	78	22.58	96.63
1	47.87	2.78	2.71	509,247	234	9.33	89.92
2	55.73	2.84	2.59	842,301	389	1.46	83.34
3	56.29	2.86	2.73	1,165,048	544	0.90	76.95
4	57.19	2.92	2.80	1,476,499	697	0	70.79
5	56.07	2.90	2.76	1,777,897	850	1.12	64.82
6	54.94	2.82	2.76	2,069,985	1000	2.25	59.05
7	52.92	2.88	2.75	2,350,266	1149	4.27	53.50
8	53.71	2.85	2.74	2,614,568	1296	3.48	48.27
9	54.16	2.87	2.77	2,858,304	1439	3.03	43.45
10	53.26	2.81	2.74	3,080,066	1580	3.93	39.06
15	55.28	2.86	2.69	3,921,827	2221	1.91	22.41
20	55.62	2.86	2.65	4,478,208	2723	1.57	11.40
25	56.40	2.85	2.65	4,809,340	3065	0.79	4.85
30	57.42	2.86	2.64	4,975,650	3254	-0.22	1.56
35	57.08	2.87	2.63	5,040,195	3333	0.11	0.28
43	57.19	2.88	2.63	5,054,400	3354	0	0
full	57.19	2.88	2.63	5,054,400	5,054,400	0	0

pairs at which the locations map to distinct delays. The result is a unique mapping between a location and a vector of TDOAs. In the simulation evaluation, since $|P| = 78$, interpolation of the cross-correlation measurements is not necessary and is thus not performed.

B. Performance Criteria

The conventional and proposed algorithms are evaluated from a DOA estimation standpoint using the percentage of anomalies—estimates that differ from either the true azimuth or the true elevation by more than 5°, and root-mean-square (rms) error measure for the nonanomalous estimates

$$\mathrm{e}_{\theta,\text{rms}} = \sqrt{\frac{1}{|\chi_{\text{na}}|}\sum_{l\in\chi_{\text{na}}}[|\hat{\theta}_{s,l} - \theta_s|']^2} \tag{37}$$

where $\hat{\theta}_{s,l}$ is the estimate of the azimuth during frame l, χ_{na} is the set of all nonanomalous estimates, and the prime operator is included to take into account the cyclicity of the angular space

$$|\hat{\theta}_{s,l} - \theta_s|' = \begin{cases} |\hat{\theta}_{s,l} - \theta_s|, & \text{if } |\hat{\theta}_{s,l} - \theta_s| \le \pi \\ 2\pi - |(\hat{\theta}_{s,l} - \theta_s)|, & \text{if } |\hat{\theta}_{s,l} - \theta_s| > \pi. \end{cases} \tag{38}$$

For the elevation angle estimates

$$\mathrm{e}_{\phi,\text{rms}} = \sqrt{\frac{1}{|\chi_{\text{na}}|}\sum_{l\in\chi_{\text{na}}}[|\hat{\phi}_{s,l} - \phi_s|]^2} \tag{39}$$

where $\hat{\phi}_{s,l}$ is the estimate of the elevation during frame l.

C. Results

Fig. 1(a) depicts the percentage of nonanomalous estimates as a function of p for the proposed algorithm with a SNR of 20 dB. Figs. 1(b) and (c) show the rms errors for the nonanomalous estimates in the azimuth and elevation, respectively. The required number of updates and lookups of the proposed SRP search are shown in Figs. 1(d) and (e), respectively. In all of these curves, the quantities are compared to a reference: the conventional (full) SRP search. Fig. 1(f) depicts the percentage reductions in performance (i.e., percentage of nonanomalies) and computational load (i.e., the number of required updates) as a function of p. The exact numbers used to create the plots are found in Table V.

TABLE VII
COMPLEXITY OF GCC-PHAT COMPUTATION WITH SIMULATED DATA

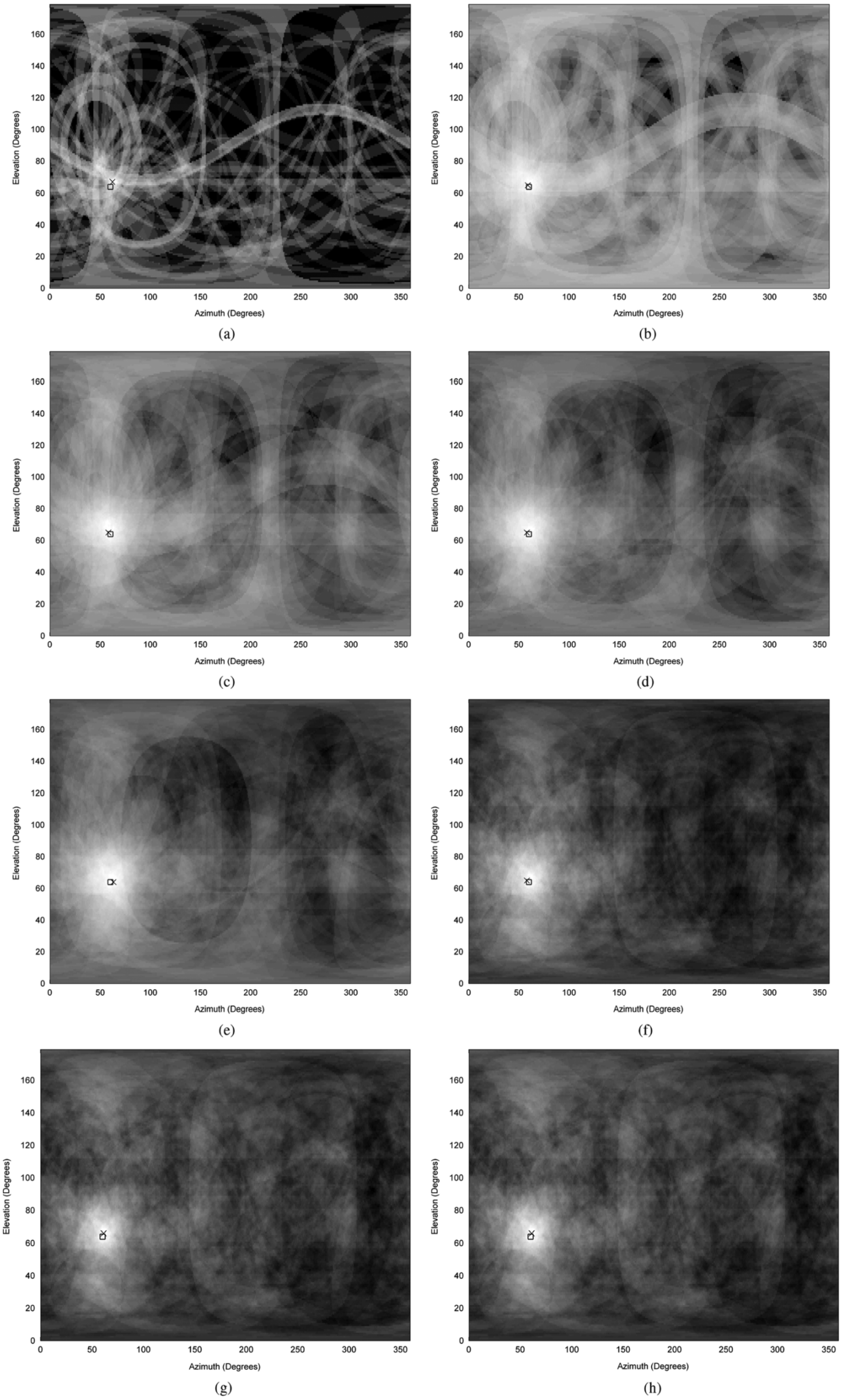
Operation	Number of Operations
compute $X_i(k)$	$(13)(6144) = 79,782$
complex multiplication	$(78)(6144) = 479,232$
complex magnitude	$(78)(6144) = 479,232$
division	$(78)(6144) = 479,232$
compute $R^{\text{g}}_{x_i,x_j}(\tau)$	$(78)(43) = 3,354$

Fig. 2 depicts the corresponding curves for an SNR of 0 dB, while the related numeric values are shown in Table VI. Table VII shows the computational load incurred by the calculation of the GCC-PHAT functions, which precedes both the conventional and proposed searches.

The standard SRP search requires

$$N_{\text{updates}} = |L||P| = N_{\text{lookups}} = 5\,054\,400 \tag{40}$$

lookups and updates each, and yields a $\%_{\text{nonanomalous}} = 94.72\%$ rate of nonanomalous estimates with a SNR of 20 dB. With an SNR of 0 dB, the nonanomalous rate falls to $\%_{\text{nonanomalous}} = 57.19\%$.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Steered response power maps for proposed generalization. (a) $p = 0$. (b) $p = 1$. (c) $p = 2$. (d) $p = 3$. (e) $p = 4$. (f) $p = 15$. (g) $p = 43$. (h) Conventional (full) search.

Referring first to the SNR $= 20$ dB case, at $p = 0$ (inverse mapping only the peak relative delay for each microphone pair), a 12.47% reduction in the number of nonanomalous estimates ensues—a significant, but surprisingly low reduction, considering that this scenario represents a reduction of 96.59% in the number of required updates. It is evident that even with just one relative delay per microphone pair, the inverse mapping technique is substantially efficacious. Moreover, at $p = 1$, or using three relative delays per microphone pair, the rate of nonanomalous frames is 92.47%, which corresponds to a 2.25% decrease in performance and an 89.81% decrease in the required number of update operations. At $p \geq 2$, the rate of nonanomalous estimates hovers around that of the full search, while the reduction in computational load decreases commensurately with p. It is remarkable that at $p = 3$, the proposed search achieves optimal (i.e., that of the full search)

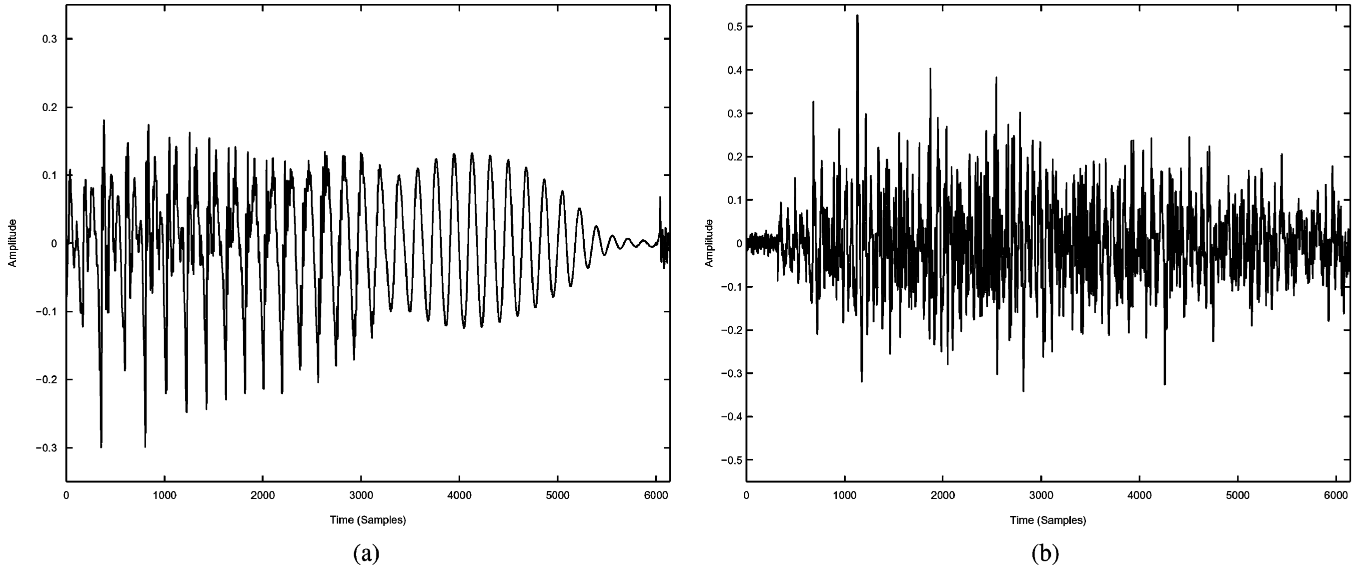


Fig. 4. (a) Clean speech signal and (b) received microphone output for sample frame.

localization accuracy while requiring only about a fifth the amount of updates.

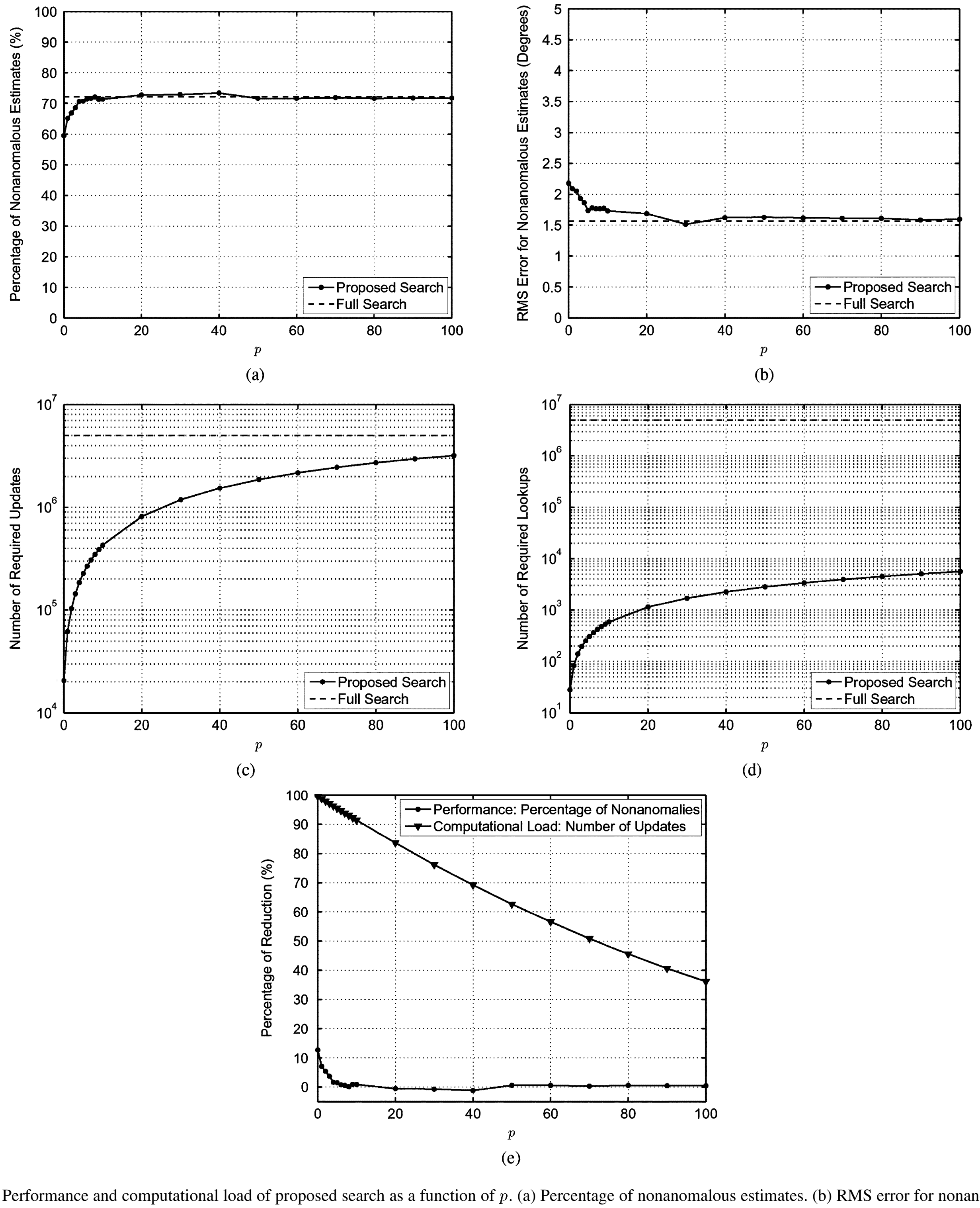
The slight fluctuations in localization accuracy at $p \geq 2$ occur as a result of including additional relative delays in the search process: a high cross-correlation level corresponding to a strong reflection may be added, hence boosting the probability of an anomalous estimate. Conversely, adding an adjacent relative delay which inverse maps to the source location or one near it reduces the probability of an anomaly. As a result, the performance of the proposed generalization fluctuates stochastically about a monotonically increasing curve which tends to the nonanomalous estimate rate achieved by the full SRP search. It should be noted that even in a heavily reverberant environment as modeled in the simulations, the proposed algorithm's performance converges to that of the full search at low values of p. Nevertheless, in practice, it may be wise to choose a slightly larger value of p.

In the $\text{SNR} = 0$ dB scenario, the positive impact of the proposed SRP search is very similar: at $p = 0$ and $p = 1$, there is a significant reduction in localization accuracy. However, the performance of the proposed search converges to that of the full search at $p = 4$, which corresponds to a 70.79% reduction in the required number of updates. This supports the notion that the positive impact of the proposed search does not depend on the level of noise in the environment: whether the conventional localization performance is quite accurate or not, the same accuracy may be achieved with the reduced complexity search procedure.

The computational load incurred by both conventional and proposed SRP algorithms due to the calculation of the GCC-PHAT functions for the various microphone pairs is found in Table VII. For every frame, a total of 79 782 computations of $X_i(k)$ are required. By direct computation, each of these operations requires $N_{\text{f}} = 6144$ multiplications and $N_{\text{f}} - 1 = 6143$ additions. However, it is well-known that there exist efficient methods for computing FFTs that reduce the order of the entire FFT operation (i.e., compute $X_i(k)$ for $k = 1, 2, \ldots, N_{\text{f}}$) to $O(N_{\text{f}} \log_2 N_{\text{f}})$, where $O(\cdot)$ denotes "order of." As a result, the computation of the $M = 13$ FFTs is actually in the order of $13 \cdot 6144 \log_2 6144 = 1\,005\,186$. Therefore, it is understood from Table VII that it is the iterative nature of the SRP search procedure, and not so much the computation of the cross-correlation functions, which poses the greatest problem to the real-time viability of the SRP algorithm. Note also that it is difficult to envision a localization scheme that does not require the computation of the cross-correlation functions: this cost is unavoidable. On the other hand, the computational cost of the conventional SRP search is quite reducible as evidenced by the simulation results.

To gain deeper insight into the proposed generalization and to understand how it is possible to make such drastic cuts in computational load while maintaining optimal performance, Fig. 3 displays the SRP maps produced by the proposed search at various values of p for a sample frame ($\text{SNR} = 20$ dB), whose corresponding clean and received speech signals are shown in Fig. 4(a) and (b), respectively. In these maps, energy is plotted as a function of the azimuth and elevation. White shades denote high levels of energy, the square denotes the actual source location, and the cross denotes the location chosen by the SRP search.

From Fig. 3(a), the SRP map produced by the proposed algorithm with $p = 0$ yields a map with much dark space—this follows from the fact when inverse mapping only a single relative delay per microphone pair, many locations are not updated by a single relative delay. The light-shaded "ovals" are precisely the basis functions $\sum_{(r,\phi,\theta)\in f_{i,j}^{-1}(\tau)} \delta([\rho, \varphi, \vartheta] - [r, \phi, \theta])$, corresponding to the inverse sets in spherical coordinates (cones in Cartesian coordinates). The SRP map is simply a superposition of a finite number of weighted spatial basis functions. The weighting (i.e., shade) corresponds to the level of the cross-correlation experienced at the lag whose inverse mapping yields the basis function. In Fig. 3(a), the number of basis functions is $|P| = 78$. Despite this low number, the resulting SRP map shows a clear energy peak in the vicinity of the actual



(a)

(b)

(c)

(d)

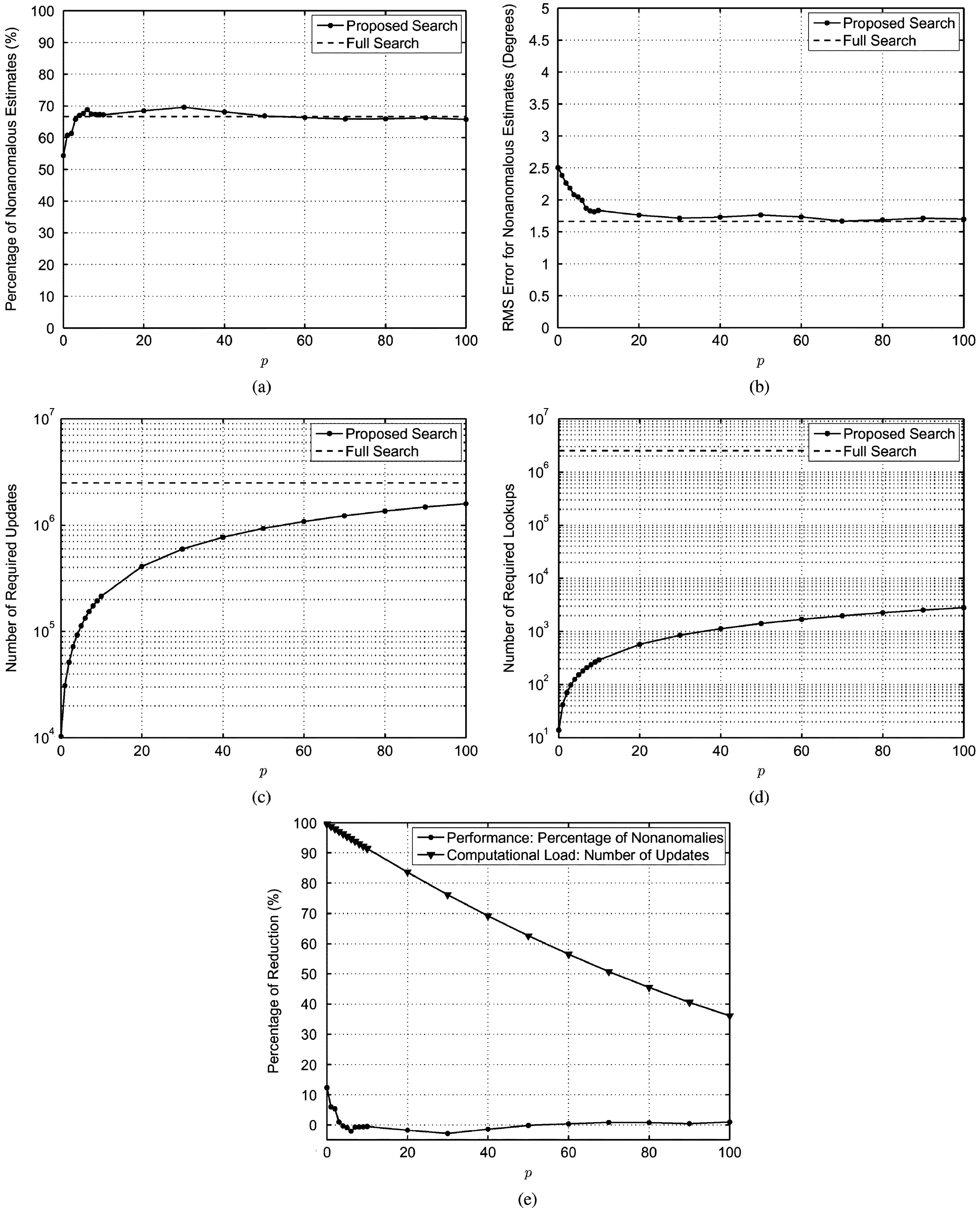
(e)

Fig. 5. Performance and computational load of proposed search as a function of p. (a) Percentage of nonanomalous estimates. (b) RMS error for nonanomalous estimates. (c) Number of required updates. (d) Number of required lookups. (e) Corresponding reductions in performance and computational load.

source location, and the estimate is nonanomalous. Moreover, it is clear from Figs. 3(b) through (e) that the SRP map is quite accurately represented by only a few basis functions per microphone pair (i.e., $1 \leq p \leq 4$). The detail added by including many more basis functions corresponds to the locations far away from the source: see the map corresponding to $p = 15$, for example. The locations in the vicinity of the source are updated frequently by the first few heavily weighted basis functions of each microphone pair. At $p = 43$ (the value of p which ensures that all relative delays are included in the search), the SRP energy map is identical to that produced by the conventional search.

IX. Evaluation With Real Recordings

The proposed and conventional SRP-PHAT algorithms are also evaluated with data obtained using the IDIAP Research



(a)

(b)

(c)

(d)

(e)

Fig. 6. Performance and computational load of proposed search as a function of p using only 14 microphone pairs. (a) Percentage of nonanomalous estimates. (b) RMS error for nonanomalous estimates. (c) Number of required updates. (d) Number of required lookups. (e) Corresponding reductions in performance and computational load.

Institute's Smart Meeting Room—please refer to [23] for details. The array used is a planar, uniform circular array with $M = 8$ omnidirectional microphones and a radius of 10 cm. Since the array is planar with a small radius, the array is not able to provide precise elevation discrimination. Therefore, the evaluation focuses on the localization accuracy of only the azimuth angle of arrival. The IDIAP Institute provides a public MATLAB implementation of SRP-PHAT; this implementation is employed in the experimental evaluation. In order to assess the impact of the proposed generalization of SRP, the program is modified to reflect the proposed search procedure.

TABLE VIII
PERFORMANCE AND COMPUTATIONAL LOAD OF PROPOSED AND CONVENTIONAL SRP SEARCHES

p	$\%_{\text{nonanomalous}}$	$e_{\theta,\text{RMS}}$	N_{updates}	N_{lookups}	$\%_{\text{perf. reduction}}$	$\%_{\text{comp. reduction}}$
0	59.54	2.18	20,643	28	12.69	99.59
1	65.15	2.09	61,914	84	7.08	98.76
2	66.90	2.05	103,125	140	5.33	97.93
3	68.58	1.93	144,255	196	3.65	97.11
4	70.62	1.86	185,292	252	1.61	96.29
5	70.76	1.74	226,207	308	1.47	95.46
6	71.46	1.78	267,006	364	0.77	94.65
7	71.67	1.77	307,659	420	0.56	93.83
8	72.16	1.77	348,147	476	0.07	93.02
9	71.32	1.77	388,443	532	0.91	92.21
10	71.39	1.73	428,499	588	0.84	91.41
20	72.79	1.69	816,607	1148	-0.56	83.63
30	72.93	1.51	1,187,821	1708	-0.70	76.19
40	73.42	1.62	1,536,124	2268	-1.19	69.20
50	71.67	1.63	1,862,421	2828	0.56	62.66
60	71.67	1.62	2,165,097	3388	0.56	56.59
70	71.95	1.61	2,450,833	3948	0.28	50.86
80	71.67	1.61	2,715,927	4508	0.56	45.55
90	71.81	1.58	2,962,396	5068	0.42	40.61
100	71.81	1.59	3,184,031	5628	0.42	36.16
full	72.23	1.56	4,987,892	4,987,892	0	0

The room dimensions are 8.2 m by 3.6 m by 2.4 m. The array rests on a centrally located table with dimensions 4.8 m by 1.2 m. Throughout the recording process, the speaker moves to 16 locations in an L-shaped corner area of the room and utters a sequence of digits, followed by "this is position 1 (i.e.,)." The microphones are sampled at 16 kHz. Since this sampling rate is lower than that required for fine location resolution, the GCC measurements are interpolated by a factor of 20 before running the searches. Prior to the upsampling, the range of considered relative delays is set to $|D_{i,j}| := 35, \forall(i,j) \in P$. However, with the upsampled rate, $|D_{i,j}| := 700, \forall(i,j) \in P$. The forward and inverse look-up tables are formed according to this upsampled rate. Since some of the source locations are in the near-field, the search is performed across the three Cartesian dimensions—assuming a two-dimensional DOA space would imply a plane-wave model and lead to a modeling error in the ensuing SRP search. A total of $|L| = 178\,139$ locations are included in the search grid. Since the number of unique microphone pairs is $|P| = 28$, the total number of lookups and updates required is $N_{\text{lookups}} = N_{\text{updates}} = |L||P| = 4\,987\,892$. From this, the computational expense of the SRP-PHAT technique is evident: a grid of 178 139 locations operating in conjunction with a 700-by-28 element look-up table. Once the grid element with the highest steered energy is found, the azimuth angle of arrival (relative to the array center) is computed from trivial geometric calculations. The frame length is 1024 samples or 64 ms. The proposed search is run for $p \in \{0, 1, 2, \ldots, 9, 10, 20, 30, \ldots, 90, 100\}$. The location estimates are computed for all 3498 frames—however, in the performance evaluation, only the frames during which the speaker is active (for details, please see [23]) are taken into account; there are 1426 such frames.

TABLE IX
COMPLEXITY OF GCC-PHAT COMPUTATION WITH REAL DATA

Operation	Number of Operations
compute $X_i(k)$	$(8)(1024) = 8,192$
complex multiplication	$(28)(1024) = 28,672$
complex magnitude	$(28)(1024) = 28,672$
division	$(28)(1024) = 28,672$
compute $R^{\text{g}}_{x_i,x_j}(\tau)$	$(28)(35) = 980$

The performances of the proposed and conventional SRP-PHAT algorithms are displayed in Fig. 5. The exact numerical values used to create the plots of the figure are shown in Table VIII. Table IX lists the computational load stemming from the calculation of the GCC functions. To investigate the benefits of the proposed search algorithm in SRP implementations utilizing less microphones (or less microphone pairs), Fig. 6 shows the performances of the conventional and proposed searches when employing only 14 of the 28 unique microphone pairs, with the corresponding numeric values given in Table X.

From Fig. 5 and Table VIII, it is immediately seen that the proposed generalization produces striking results in the lower values for p; in fact, for $p = 0$ (using only 1 out of 700 relative delays per microphone pair), the percentage of anomalies only increases by 12.69%—this represents a reduction in computational load of 99.59%! As in the simulation results, the performance of the proposed scheme rapidly approaches that of the

TABLE X
PERFORMANCE AND COMPUTATIONAL LOAD OF PROPOSED AND CONVENTIONAL SRP SEARCHES USING ONLY 14 MICROPHONE PAIRS

p	$\%_{\text{nonanomalous}}$	$e_{\theta,\text{RMS}}$	N_{updates}	N_{lookups}	$\%_{\text{perf. reduction}}$	$\%_{\text{comp. reduction}}$
0	54.35	2.51	10,318	14	12.34	99.59
1	60.73	2.38	30,935	42	5.96	98.76
2	61.36	2.26	51,534	70	5.33	97.93
3	65.71	2.18	72,090	98	0.98	97.11
4	66.97	2.08	92,594	126	-0.28	96.29
5	67.60	2.05	113,048	154	-0.91	95.47
6	68.79	2.00	133,443	182	-2.10	94.65
7	67.39	1.87	153,770	210	-0.70	93.83
8	67.32	1.83	174,014	238	-0.63	93.02
9	67.32	1.82	194,160	266	-0.63	92.21
10	67.18	1.84	214,194	294	-0.49	91.41
20	68.44	1.76	408,418	574	-1.75	83.63
30	69.57	1.72	594,312	854	-2.88	76.17
40	68.09	1.73	768,809	1134	-1.40	69.17
50	66.83	1.77	9,323,397	1414	-0.14	62.62
60	66.34	1.73	1,084,039	1694	0.35	56.53
70	65.85	1.67	1,227,247	1974	0.84	50.79
80	65.92	1.69	1,360,015	2254	0.77	45.47
90	66.27	1.72	1,483,350	2534	0.42	40.52
100	65.71	1.70	1,594,266	2814	0.98	36.07
full	66.69	1.66	2,493,946	2,493,946	0	0

full search with low values of p. In fact, at $p = 8$, the number of computations is reduced by 93% without sacrificing the localization accuracy (i.e., 0.07% reduction). This again follows from the fact that a small number of basis functions per microphone pair are required to accurately represent the energy map. Even at $p = 100$, the algorithm only performs 63.84% of the total number of updates performed by the full search, even though the performance converges at $p \approx 10$. This illustrates the wastefulness of the conventional SRP search and also the problem it poses for real-time implementation.

From Fig. 6 and Table X, utilizing only half of the unique microphone pairs leads to a higher anomaly rate in the conventional (and proposed) SRP algorithms. This is because the redundancy offered by utilizing additional microphone pairs aids in combating the effects of noise and reverberation [10]. Nevertheless, the positive impact of the proposed search remains the same: the nonanomalous rate offered by the proposed SRP search converges to that of the full search at low values of p (i.e., $p = 4$), while the reduction in computational load is drastic (i.e., 96.29% at $p = 4$). Notice that the reduction in computational load is achieved by subsetting the range of relative delays for each microphone pair; thus, the complexity reduction is independent of the number of microphone pairs employed.

X. CONCLUSION

This paper has presented a generalized paradigm for steered-energy based acoustic source localization. The traditional SRP (and all of its variants, including SRP-PHAT) were reformulated in terms of an inverse function that maps relative delays to sets of candidate locations. With this reformulation, the resulting search is not performed sequentially in space.

It was shown that this reformulation corresponds to a spatial decomposition of the SRP energy map—the spatial basis functions of this decomposition are constructed using the inverse sets. The weights of each basis function correspond to the level of cross-correlation experienced at the relative delay which inverse maps to the locations comprising the basis function.

Results using both simulated and real data showed that only a few basis functions per microphone pair are required to accurately represent the SRP-PHAT energy map. As a result, a drastic reduction in the number of required computations is afforded. In fact, it was shown that it is possible to perform less than 10% of the number of computations required by the conventional full search without significantly sacrificing localization accuracy. The complexity reduction offered by the proposed SRP generalization is independent of the array geometry and the number of microphone pairs utilized.

In order to reduce the heavy computational load of the SRP approach, it is required to remove some of the updates from the search process. In this paper, the values of the cross correlations themselves are used to determine which updates to omit. The drastic reduction in the computation time afforded by the proposed generalization makes SRP-PHAT more amenable to real-time operation.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, pp. 4–24, Apr. 1988.

[2] W. Bangs and P. Schultheis, "Space-time processing for optimal parameter estimation," in *Signal Process.*, J. Griffiths, P. Stocklin, and C. V. Schooneveld, Eds. New York: Academic, 1973, pp. 577–590.

[3] G. Carter, "Variance bounds for passively locating an acoustic source with a symmetric line array," *J. Acoust. Soc. Amer.*, vol. 62, pp. 922–926, Oct. 1977.

[4] W. Hahn and S. Tretter, "Optimum processing for delay-vector estimation in passive signal arrays," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 9, pp. 608–614, Sep. 1973.

[5] W. Hahn, "Optimum signal processing for passive sonar range and bearing estimates," *J. Acoust. Soc. Amer.*, vol. 58, pp. 201–207, Jul. 1975.

[6] M. Wax and T. Kailath, "Optimum localization of multiple sources by passive arrays," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 5, pp. 1210–1217, Oct. 1983.

[7] M. Omologo and P. G. Svaizer, "Use of the cross-power-spectrum phase in acoustic event localization," ITC-IRST Tech. Rep. 9303-13, Mar. 1993.

[8] M. Omologo, P. G. Svaizer, and R. De Mori, "Acoustic transduction," in *Spoken Dialogue with Computers*, R. De Mori, Ed. San Diego, CA: Academic, 1998, pp. 23–69.

[9] J. 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.

[10] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Proceess.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[11] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 10, pp. 1481–1494, Oct. 1989.

[12] S. T. Birchfield and D. K. Gillmor, "Acoustic source direction by hemisphere sampling," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2001, vol. 5, pp. 3053–3056.

[13] 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, no. 4, pp. 320–327, Aug. 1976.

[14] J. Dibiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments," Ph.D. dissertation, Brown Univ., Providence, RI, 2000.

[15] J. M. Valin *et al.*, "Localization of simultaneous moving sound sources for mobile robot using a frequency- domain steered beamformer approach," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2004, pp. 1033–1038.

[16] J. M. Valin, F. Michaud, and J. Rouat, "Robust 3D localization and tracking of sound sources using beamforming and particle filtering," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2006, vol. 4, pp. 841–844.

[17] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 499–508, Sep. 2004.

[18] J. Peterson and C. Kyriakakis, "Analysis of fast localization algorithms for acoustical environments," in *Proc. 39th Asilomar Conf. Signals, Syst., Comput.*, 2005, pp. 1385–1389.

[19] J. Peterson and C. Kyriakakis, "Hybrid algorithm for robust, real-time source localization in reverberant environments," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2005, vol. 4, pp. 1053–1056.

[20] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 713–716, Jan. 2005.

[21] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

[22] M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Amer.*, vol. 37, pp. 409–412, 1965.

[23] G. Lathoud, J. M. Odobez, and D. Gatica-Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI'04 Workshop*, 2006, pp. 182–195.

Jacek P. Dmochowski was born in Gdansk, Poland, in December 1979. He received the B.Eng. degree (with high distinction) in communications engineering and the M.S. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2003 and 2005, respectively. He is currently pursuing the Ph.D. degree at the INRS-EMT, University of Quebec, Montreal, QC, Canada.

His research interests include microphone array beamforming and source localization, blind source separation, as well as frequency-domain uncertainty analysis.

Mr. Dmochowski is the recipient of the National Sciences and Engineering Research Council (NSERC) Post Graduate Scholarship at the Doctoral Level (2005–2007).

Jacob Benesty (M'92–SM'04) was born in 1963. He received the M.S. degree in microwaves from Pierre and Marie Curie University, Paris, France, in 1987, and the Ph.D. degree in control and signal processing from Orsay University, Paris, in April 1991.

During the Ph.D. degree (from November 1989 to April 1991), he worked on adaptive filters and fast algorithms at the Centre National d'Etudes des Telecommunications (CNET), Paris. From January 1994 to July 1995, he worked at Telecom Paris University on multichannel adaptive filters and acoustic echo cancellation. From October 1995 to May 2003, he was first a Consultant and then a Member of the Technical Staff at Bell Laboratories, Murray Hill, NJ. In May 2003, he joined the INRS-EMT, University of Quebec, Montreal, QC, Canada, as an Associate Professor. His research interests are in signal processing, acoustic signal processing, and multimedia communications. He coauthored the books *Acoustic MIMO Signal Processing* (Springer-Verlag, 2006) and *Advances in Network and Acoustic Echo Cancellation* (Springer-Verlag, 2001). He is also a coeditor/coauthor of the books *Speech Enhancement* (Spinger-Verlag, 2005), *Audio Signal Processing for Next Generation Multimedia Communication Systems* (Kluwer, 2004), *Adaptive Signal Processing: Applications to Real-World Problems* (Springer-Verlag, 2003), and *Acoustic Signal Processing for Telecommunication* (Kluwer, 2000). He was a member of the Editorial Board of the *EURASIP Journal on Applied Signal Processing*.

Dr. Benesty received the 2001 Best Paper Award from the IEEE Signal Processing Society. He was the Co-Chair of the 1999 International Workshop on Acoustic Echo and Noise Control.

Sofiène Affès (M'94–SM'04) received the Diplôme d'Ingénieur in electrical engineering and the Ph.D. degree (with honors) in signal processing, both from the École Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1992 and 1995, respectively.

He has since been with the INRS-EMT, University of Quebec, Montreal, QC, Canada, as a Research Associate from 1995 until 1997, then as an Assistant Professor until 2000. Currently, he is an Associate Professor in the Personal Communications Group. His research interests are in wireless communications, statistical signal and array processing, adaptive space–time processing and MIMO. From 1998 to 2002, he led the radio-design and signal processing activities of the Bell/Nortel/NSERC Industrial Research Chair in Personal Communications at INRS-EMT. Currently, he is actively involved in a major project in wireless of PROMPT-Québec (Partnerships for Research on Microelectronics, Photonics and Telecommunications). He currently acts as a member of Editorial Board of the *Wiley Journal on Wireless Communications and Mobile Computing*.

Prof. Affes is the corecipient of the 2002 Prize for Research Excellence of INRS and currently holds a Canada Research Chair in High-Speed Wireless Communications. He served as a General Co-Chair of the IEEE VTC'2006-Fall conference, Montreal.