THE GENERALIZATION OF NARROWBAND LOCALIZATION METHODS TO BROADBAND ENVIRONMENTS VIA PARAMETRIZATION OF THE SPATIAL CORRELATION MATRIX

Jacek Dmochowski, Jacob Benesty, and Sofiène Affes

INRS-EMT, Université du Québec
800 de la Gauchetière Ouest, H5A 1K6, Montréal, Québec
phone: +1 514-875-1266, fax: +1 514-875-0344, email: {dmochow,benesty,affes}@emt.inrs.ca

ABSTRACT

The need to localize a radiating signal source is necessitated in applications ranging from distant talker speech pick-up to automatic video camera steering. The majority of literature detailing source localization is presented in the narrowband signal context; however, acoustic signals are naturally very broadband. As a result, classical narrowband techniques do not apply in the strict sense, and as a result, the broadband localization literature is yet unrefined. This paper details the manner in which classical narrowband localization methods may be generalized to broadband signal environments – specifically, a nonlinear parametrization of the spatial correlation matrix allows one to apply steered beamforming, minimum variance, subspace, and linear predictive spectral estimation methods to the broadband localization problem.

1. INTRODUCTION

A signal whose energy is confined to a single temporal frequency is termed "narrowband." A more sophisticated definition of the term is provided in [1], but since acoustic signals are "broadband" by any definition, such definitions need not be considered for this paper.

In general, the need to localize the source of a radiating signal is vast, since the location of the signal source often conveys useful information. Examples are radar, sonar, wireless communications, and room acoustics; classical signal processing techniques stem from the first two areas. Since these are primarily narrowband signal environments, it is not surprising that the literature is focused on narrowband models.

However, the room acoustics problem does not fit into this narrowband framework, despite previous attempts to fit broadband signals into narrowband frameworks through subband techniques and subsequent frequency-domain processing. The speech bandwidth is simply too wide to make this a feasible solution.

Nevertheless, the narrowband signal framework is quite instructive and elegant, and thus this paper provides a way to extend this framework to the broadband case through a nonlinear parametrization of the spatial correlation matrix.

2. SIGNAL MODEL

Assume an array of M microphone elements, distributed in some fashion in three-dimensional space, whose outputs are denoted by $x_m(t)$, $m = 0, 1, ..., M-1$, where t denotes time. The spherical coordinate system is used, where the range is denoted by r, elevation by ϕ, and azimuth by θ. For convenience, denote $\mathbf{r} = (r, \phi, \theta)$.

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

$$x_m(t) = \alpha_m(\mathbf{r}_\mathrm{s})\, s\left[t - f_{0,m}(\mathbf{r}_\mathrm{s})\right] + v_m(t), \tag{1}$$

where x_m is the received microphone output (microphone 0 serves as the reference), 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:

$$f_{i,j}(\mathbf{r}_\mathrm{s}) = \frac{1}{c}\left[d_{\mathrm{s},j}(\mathbf{r}_\mathrm{s}) - d_{\mathrm{s},i}(\mathbf{r}_\mathrm{s})\right], \tag{2}$$

where c is the speed of sound and $d_{\mathrm{s},i}$ is the distance between the sound source and microphone i.

The received microphone signals are sampled, and the forthcoming signal processing is performed on discrete signals.

3. NARROWBAND LOCALIZATION AND THE SPATIAL CORRELATION MATRIX

In narrowband signal applications, a common space-time statistic is that of the spatial correlation matrix [2], which is given by

$$\mathbf{R} = E\{\mathbf{x}(t)\mathbf{x}^H(t)\}, \tag{3}$$

where $E\{\bullet\}$ denotes mathematical expectation and

$$\mathbf{x}(t) = \begin{bmatrix} x_0(t) & x_1(t) & \cdots & x_{M-1}(t) \end{bmatrix}^T, \tag{4}$$

the superscript ^{H} denotes conjugate transpose, as complex signals are commonly used in narrowband applications, and ^{T} denotes the transpose of a matrix or vector. To steer these array outputs to a particular location, one applies a complex weight to each sensor output, whose phase performs the steering, and then sums the sensor outputs to form the output beam. Now if the input signal is no longer narrowband, each frequency requires its own complex weight to appropriately phase-shift the signal at that frequency. In the context of broadband spatial spectral estimation, the spatial correlation matrix may be computed at each temporal frequency, and the resulting spatial spectrum is now a function of the temporal frequency. For broadband applications, these narrowband estimates may be assimilated into a time-domain statistic, a procedure termed "focusing," which is described in [3]. The resulting structure is termed a "focused covariance matrix."

There exist many narrowband spatial spectral estimators, but all are rooted in the spatial correlation matrix: the Bartlett and Capon [4] estimators apply a fixed and adaptive linear weighting, respectively. Linear predictive [4] and subspace (eigenanalysis) estimators [5] provide higher resolution estimates that aid in multiple-source scenarios. A good overview of the classical narrowband methods is found in [2].

The wideband nature of speech renders focusing multiple covariance matrices over the entire frequency range grossly impractical for real-time processing. The following section describes how the general concepts of narrowband localization may be easily extended into broadband setting by performing a simple, nonlinear parametrization of the spatial correlation matrix.

4. PARAMETERIZED SPATIAL CORRELATION MATRIX

In the narrowband case, complex weights achieve the desired effect of aligning the signal – in the broadband case, the signal may be aligned by applying a time delay (a nonlinear operation) to each received signal. To that end, the parameterized spatial correlation matrix is defined by:

$$\mathbf{R}_\mathbf{r} = E\left\{\mathbf{x}_\mathbf{r}(t)\,\mathbf{x}_\mathbf{r}^T(t)\right\}, \tag{5}$$

where

$$\mathbf{x}_\mathbf{r}(t) = \left[\begin{array}{cccc} x_0[t] & x_1\left[t + f_{0,1}(\mathbf{r})\right] & \cdots & x_0\left[t + f_{0,M-1}(\mathbf{r})\right]\end{array}\right]^T.$$

Before proceeding, assume that the desired signal s and additive noise v are mutually uncorrelated random processes, and also that the attenuation terms are independent of position: $\alpha_m(\mathbf{r}_\mathrm{s}) = \alpha_m,\ \forall \mathbf{r}_s$. In that case, substituting (1) into (5) results in the following structure for the parameterized spatial correlation matrix:

$$[\mathbf{R}_\mathbf{r}]_{i,j} = \alpha_i\alpha_j R_{\mathrm{s,s}}\left[f_{i,j}(\mathbf{r}) - f_{i,j}(\mathbf{r}_\mathrm{s})\right] + R_{v_i,v_j}\left[f_{i,j}(\mathbf{r})\right], \tag{6}$$

where

$$R_{x,y}(\tau) = E\left\{x(t)\,y(t+\tau)\right\} \tag{7}$$

is the cross-correlation function for two jointly wide-sense stationary random processes.

From (6), notice that the parametrization of the spatial correlation matrix spatially decorrelates the noise term $R_{v_i,v_j}\left[f_{i,j}(\mathbf{r})\right]$; if the additive noise is temporally white, and assuming that the lag $f_{i,j}(\mathbf{r}) \neq 0$, the noise component of the parameterized spatial correlation matrix is simply:

$$\mathbf{R}_\mathbf{r}\,|_{\text{noise}} = \sigma_v^2\mathbf{I}, \tag{8}$$

where $\sigma_v^2 = R_{v_0,v_0}(0) = \sigma_{v_0}^2 = \cdots = R_{v_{M-1},v_{M-1}}(0) = \sigma_{v_{M-1}}^2$. Therefore, it is reasonable to assume that the noise component of the parameterized spatial correlation matrix is diagonal.

Secondly, when $\mathbf{r} = \mathbf{r}_\mathrm{s}$, the signal component of the parameterized spatial correlation matrix takes the form of the following rank-one matrix:

$$\mathbf{R}_\mathbf{r}\,|_{\text{signal}} = \sigma_s^2\alpha\alpha^T, \tag{9}$$

where $\sigma_s^2 = R_{\mathrm{s,s}}(0)$ and

$$\alpha = \left[\begin{array}{cccc}\alpha_0 & \alpha_1 & \cdots & \alpha_{M-1}\end{array}\right]^T. \tag{10}$$

If the parametrization is not matched to the location of the signal (i.e., $f_{i,j}(\mathbf{r}) \neq f_{i,j}(\mathbf{r}_s),\ \forall i,j$), the signal component is no longer rank-one. Assuming that the desired signal is a white process, the signal component takes the form of

$$\mathbf{R}_\mathbf{r}\,|_{\text{signal}} = \sigma_s^2\,\mathrm{diag}\left(\alpha_0^2, \alpha_1^2, \cdots, \alpha_{M-1}^2\right), \tag{11}$$

where $\mathrm{diag}(\bullet)$ is a diagonal matrix with its entries denoted by the arguments. Putting all of this together, we arrive at

$$\mathbf{R}_\mathbf{r} = \begin{cases} \sigma_s^2\alpha\alpha^T + \sigma_v^2\mathbf{I}, & \text{if } \mathbf{r} = \mathbf{r}_s \\ \sigma_s^2\,\mathrm{diag}\left(\alpha_0^2, \alpha_1^2, \cdots, \alpha_{M-1}^2\right) + \sigma_v^2\mathbf{I}, & \text{otherwise} \end{cases}. \tag{12}$$

5. BROADBAND LOCALIZATION METHODS

5.1 Steered Conventional Beamforming and the Steered Response Power Method

A delay-and-sum beamformer (DSB) is a simple fixed beamformer which attempts to time-align the received signals in such a way that the signal arriving from a certain location is emphasized. Using the model of Section 2, the output of a DSB steered to a location $\mathbf{r}$ is given as

$$z_\mathbf{r}(n) = \sum_{m=0}^{M-1} w_{\mathbf{r},m}x_m\left[n + f_{0,m}(\mathbf{r})\right]. \tag{13}$$

The delays $f_{0,m}(\mathbf{r})$ steer the beamformer to the desired location, while the beamformer weights $w_{\mathbf{r},m}$ help shape the beam accordingly. The weights here have been made dependent on the steered location $\mathbf{r}$ for a reason that will become apparent in future subsections.

The estimate of the spatial spectral power at location $\mathbf{r}$ is given by the power of the beamformer output when steered to azimuth $\mathbf{r}$. The steered-beamformer spectral estimate is given by

$$S^{\mathrm{DSB}}(\mathbf{r}) = E\left\{z_\mathbf{r}^2(n)\right\}. \tag{14}$$

Substitution of (13) into (14) leads to

$$S^{\mathrm{DSB}}(\mathbf{r}) = \mathbf{w}_\mathbf{r}^T\mathbf{R}_\mathbf{r}\mathbf{w}_\mathbf{r}, \tag{15}$$

where

$$\mathbf{w}_\mathbf{r} = \left[\begin{array}{cccc} w_{\mathbf{r},0} & w_{\mathbf{r},1} & \cdots & w_{\mathbf{r},L}\end{array}\right]^T. \tag{16}$$

The location estimate is thus given by

$$\hat{\mathbf{r}}_\mathrm{s} = \arg\max_\mathbf{r} \mathbf{w}_\mathbf{r}^T\mathbf{R}_\mathbf{r}\mathbf{w}_\mathbf{r}. \tag{17}$$

The well-known steered response power (SRP) algorithm [6] follows directly from a special case of (17), where $\mathbf{w}_\mathbf{r} = \mathbf{1}$ for all $\mathbf{r}$, and $\mathbf{1}$ is a vector of ones:

$$\hat{\mathbf{r}}_{\mathrm{s,SRP}} = \arg\max_\mathbf{r} \mathbf{1}^T\mathbf{R}_\mathbf{r}\mathbf{1}. \tag{18}$$

For this special case of fixed unit weights, this means that the maximization of the power of a steered DSB is equivalent to the maximization of the sum of the entries of $\mathbf{R}_\mathbf{r}$.

It is interesting to note that in the narrowband case, the spatial correlation matrix is location-independent and the weights perform the delay operation. In the broadband case, the spatial correlation matrix is parameterized by the location and thus the delay operations are embedded in the matrix. Nevertheless, the weights add more design degrees of freedom in the estimation procedure.

It is somewhat surprising that even though the SRP algorithm has attracted significant attention recently (see [6], [7], and [8]), the weighting $\mathbf{w}_\mathbf{r} = \mathbf{1}$ is used exclusively in the literature. This weighting is fixed with respect to both the data and the steered location. Notice that from (15), this is an effectively "narrowband" weight selection, in that the pre-aligning of the microphones requires only the selection of a single weight per channel. Note, however, that this weight selection must be performed for all locations $\mathbf{r}$. To that end, the following section presents one such adaptive weighting scheme, proposed (in a somewhat different context to be explained in the next section) by Krolik and Swingler [9].

5.2 Minimum Variance

The minimum variance approach to spatial spectral estimation involves selecting weights that pass a signal [i.e., a broadband plane wave $s(t)$] propagating from location $\mathbf{r}$ with unity gain, while minimizing the total output power, given by $\mathbf{w}_\mathbf{r}^T\mathbf{R}_\mathbf{r}\mathbf{w}_\mathbf{r}$. The application of the minimum variance method to broadband spatial spectral estimation is given in [9].

The unity gain constraint proposed by [9] is

$$\mathbf{w}_\mathbf{r}^T\mathbf{1} = 1. \tag{19}$$

and the $\mathbf{1}$ vector follows from the fact that the signal is already time-aligned across the array before minimum variance processing. It is as if the signal is coming from the broadside of a linear array.

Using the method of Lagrange multipliers in conjunction with the cost function $\mathbf{w}_\mathbf{r}^T\mathbf{R}_\mathbf{r}\mathbf{w}_\mathbf{r}$, the minimum variance weights become

$$\mathbf{w}_{\mathbf{r},\mathrm{mv}} = \frac{\mathbf{R}_\mathbf{r}^{-1}\mathbf{1}}{\mathbf{1}^T\mathbf{R}_\mathbf{r}^{-1}\mathbf{1}}. \tag{20}$$

The resulting minimum variance spatial spectral estimate is found by substituting the weights of (20) into the cost function:

$$S^{\mathrm{mv}}(\mathbf{r}) = \mathbf{w}_{\mathbf{r},\mathrm{mv}}^T \mathbf{R}_\mathbf{r} \mathbf{w}_{\mathbf{r},\mathrm{mv}} = \left(\mathbf{1}^T \mathbf{R}_\mathbf{r}^{-1} \mathbf{1}\right)^{-1}. \tag{21}$$

The broadband minimum variance DOA estimator is thus given by:

$$\hat{\mathbf{r}}_{\mathrm{s,mv}} = \arg\max_{\mathbf{r}} \left(\mathbf{1}^T \mathbf{R}_\mathbf{r}^{-1} \mathbf{1}\right)^{-1}. \tag{22}$$

The interesting part of Krolik and Swingler's proposal is in what they term the "steered covariance matrix" – although it might seem at first glance that this matrix is equivalent to the parameterized spatial correlation matrix, Krolik and Swingler propose estimating this steered covariance matrix in the *frequency-domain* using subsequent combining of the frequency bands. In this paper, it is clearly understood that the parameterized spatial correlation matrix is to be computed using simple cross-correlation in the time domain.

5.3 Subspace Approach

A subspace approach to broadband spatial spectral estimation was first presented in [10] and [11]. Referring to (12), consider only the signal component of $\mathbf{R}_\mathbf{r}$. It may be easily shown that this matrix has one non-zero eigenvector, that eigenvector being α, with the corresponding eigenvalue being $\sigma_s^2 \|\alpha\|^2$. The vector of attenuation constants α is generally unknown; however, from the above discussion, it is apparent that the vector may be estimated from the eigenanalysis of $\mathbf{R}_\mathbf{r}$.

To that end, consider another adaptive weight selection method, which follows from the ideas of narrowband beamforming [2]. This weight selection attempts to non-trivially maximize the output energy of the steered-beamformer for a given location $\mathbf{r}$:

$$\mathbf{e}_{\max,\mathbf{r}} = \arg\max_{\mathbf{w}_\mathbf{r}} \mathbf{w}_\mathbf{r}^T \mathbf{R}_\mathbf{r} \mathbf{w}_\mathbf{r} \tag{23}$$

subject to

$$\mathbf{w}_\mathbf{r}^T \mathbf{w}_\mathbf{r} = 1. \tag{24}$$

It is well-known that the solution to the above constrained optimization is the vector that maximizes the Rayleigh quotient [12], $\frac{\mathbf{w}_\mathbf{r}^T \mathbf{R}_\mathbf{r} \mathbf{w}_\mathbf{r}}{\mathbf{w}_\mathbf{r}^T \mathbf{w}_\mathbf{r}}$, which is in turn given by the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_\mathbf{r}$. The resulting spatial spectral estimate is given by:

$$S^{\mathrm{EIG}}(\mathbf{r}) = \mathbf{e}_{\max,\mathbf{r}}^T \mathbf{R}_\mathbf{r} \mathbf{e}_{\max,\mathbf{r}} = \lambda_{\max,\mathbf{r}}, \tag{25}$$

where $\lambda_{\max,\mathbf{r}}$ is the maximum eigenvalue of $\mathbf{R}_\mathbf{r}$ and $\mathbf{e}_{\max,\mathbf{r}}$ is the corresponding (principal) eigenvector. The localization involves searching for the $\mathbf{r}$ that produces the largest maximum eigenvalue of $\mathbf{R}_\mathbf{r}$:

$$\hat{\theta}_{\mathrm{EIG}} = \arg\max_{\mathbf{r}} \lambda_{\max,\mathbf{r}}. \tag{26}$$

In addition to producing another spatial spectrum estimate, the above eigenanalysis allows one to estimate α:

$$\hat{\alpha} = \mathbf{e}_{\max,\hat{\theta}_{\mathrm{EIG}}}. \tag{27}$$

5.4 Linear Predictive Methods

Linear spatial prediction involves predicting the output of one microphone with a linear combination of the remaining microphones. In a broadband setting, this requires the pre-aligning of the microphones with respect to a certain location. Interestingly, the resulting broadband linear spatial predictive model is intimately related to the parameterized spatial correlation matrix. This concept was first presented in [13] and [14] in the context of linear array time delay estimation; the idea was generalized to arbitrary array geometries, transforming the problem from time delay estimation to localization in [10].

Using a spatial autoregressive (AR) model, the linear predictive framework is given by

$$x_0(t) = \sum_{l=1}^{L} a_{\mathbf{r},l}\, x_l\left[t + f_l(\mathbf{r})\right] + e(t), \tag{28}$$

where $e(t)$ may be interpreted as either the spatially white noise that drives the AR model, or the prediction error. For each $\mathbf{r}$ in the azimuth space, one finds the weight vector

$$\mathbf{a}_\mathbf{r} = \begin{bmatrix} -1 & a_{\mathbf{r},1} & \cdots & a_{\mathbf{r},L} \end{bmatrix}^T \tag{29}$$

which minimizes the criterion

$$J_\mathbf{r} = E\left\{ \left| -\mathbf{a}_\mathbf{r}^T \mathbf{x}_\mathbf{r}(t) \right|^2 \right\} \tag{30}$$

subject to the constraint

$$\delta^T \mathbf{a}_\mathbf{r} = -1, \tag{31}$$

where

$$\delta = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}^T. \tag{32}$$

Using the method of Lagrange multipliers, the optimal predictive weights are given by:

$$\mathbf{a}_{\mathbf{r},\mathrm{opt}} = -\frac{\mathbf{R}_\mathbf{r}^{-1}\delta}{\delta^T \mathbf{R}_\mathbf{r}^{-1}\delta}, \tag{33}$$

and the resulting minimum mean-squared error (MMSE) is

$$J_{\mathbf{r},\min} = \frac{1}{\delta^T \mathbf{R}_\mathbf{r}^{-1}\delta}. \tag{34}$$

Note that both the optimal predictive coefficients and the MMSE are a function of the steered location $\mathbf{r}$.

In [13] and [14], it is shown that

$$0 \le \det\left(\widetilde{\mathbf{R}}_\mathbf{r}\right) \le \frac{J_{\mathbf{r},\min}}{E\left\{x_0^2(t)\right\}} \le 1, \tag{35}$$

where the following factorization is used [13], [14]:

$$\mathbf{R}_\mathbf{r} \quad = \quad \mathbf{D}\widetilde{\mathbf{R}}_\mathbf{r}\mathbf{D}, \tag{36}$$

where

$$\mathbf{D} = \begin{bmatrix} \sqrt{E\{x_0^2(t)\}} & 0 & \cdots & 0 \\ 0 & \sqrt{E\{x_1^2(t)\}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sqrt{E\{x_{M-1}^2(t)\}} \end{bmatrix} \tag{37}$$

is a diagonal matrix,

$$\widetilde{\mathbf{R}}_\mathbf{r} = \begin{bmatrix} 1 & \rho_{\mathbf{r},0,1} & \cdots & \rho_{\mathbf{r},0,M-1} \\ \rho_{\mathbf{r},0,1} & 1 & \cdots & \rho_{\mathbf{r},1,M-1} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{\mathbf{r},0,M-1} & \cdots & \rho_{\mathbf{r},M-2,M-1} & 1 \end{bmatrix} \tag{38}$$

is a symmetric matrix and

$$\rho_{\mathbf{r},k,l} = \frac{E\{x_k[t+f_k(\mathbf{r})]\, x_l[t+f_l(\mathbf{r})]\}}{\sqrt{E\{x_k^2(t)\}E\{x_l^2(t)\}}}, \quad k,l = 0,1,...,M-1, \tag{39}$$

is the cross-correlation coefficient between $x_k\left[t+f_k(\mathbf{r})\right]$ and $x_l\left[t+f_l(\mathbf{r})\right]$.

Essentially, minimizing the prediction error corresponds to minimizing $\det\left(\widetilde{\mathbf{R}}_{\mathbf{r}}\right)$, and thus the following spatial spectrum is proposed:

$$S^{\mathrm{MCCC}}(\mathbf{r}) = \rho_{\mathbf{r}}^2 = 1-\det\left(\widetilde{\mathbf{R}}_{\mathbf{r}}\right) = 1-\det\left(\mathbf{D}^{-1}\mathbf{R}_{\mathbf{r}}\mathbf{D}^{-1}\right), \qquad (40)$$

where $\rho_{\mathbf{r}}$ is the multichannel cross-correlation coefficient (MCCC).

The location estimation easily follows as

$$\hat{\theta}_{\mathrm{MCCC}} = \arg\max_{\mathbf{r}} \rho_{\mathbf{r}}^2. \qquad (41)$$

6. SIMULATION EVALUATION

The proposed estimators are evaluated in a computer simulation using the image method model of [15]. A 10-microphone uniform circular array of 6.9 cm is simulated. The simulated room is rectangular with plane reflective boundaries (walls, ceiling and floor). The room dimensions in centimeters are (304.8, 457.2, 381). The centre of the array sits at (152.4, 228.6, 101.6). The speaker is located at (152.4, 406.4, 101.6). The reverberation times are measured using the method of [16]: three reverberation levels ranging from $T_{60}=0$ ms to $T_{60}=600$ ms are considered. The source signal is convolved using the synthetic impulse responses. Appropriately-scaled temporally white Gaussian noise is then added at the microphones to achieve an SNR of 0 dB. Two signal types are examined: white Gaussian noise and female English speech. The DOA estimates are computed once per 128 ms frame over a two-minute signal. The sampling rate is 48 kHz. Due to the planar array geometry and far-field source, the location space is limited to the set of azimuth angles in the range $0-360$ degrees, with a resolution of 1 degree.

The algorithms are evaluated in terms of the percentage of anomalous estimates – those that vary from the true azimuth by more than 5 degrees, and by the root-mean-square (RMS) error for the nonanomalous estimates. For comparison, the proposed estimators are compared to a standard two-step algorithm [10] that consists of TDOA measurements in the first stage and maximum likelihood least-squares mapping of relative delays to source location in the second stage; this algorithm is indicated by TDOA in the tables.

Following the publication of [10], a bug was found in the program used to evaluate the algorithms. The bug has been fixed and the results presented herein reflect that correction.

Tables 1 and 2 display the results. In the white signal case, all estimators based on the parameterized spatial correlation matrix yield no anomalies even in the heavily reverberant case. The TDOA-based algorithm yields a low (i.e., 2 %) anomaly rate in the heavily reverberant environment. However, the simulations using a speech signal yield much higher anomaly rates; in all speech simulations, the proposed estimators yield significantly more accurate localization than the TDOA-based method. In an anechoic environment, the SRP and MVDR methods prove to be most effective. However, the MCCC method shows the greatest robustness to the effects of reverberation, as it significantly outperforms all other estimators in the the moderately and heavily reverberant case.

7. EVALUATION WITH REAL RECORDINGS

The estimators are also evaluated with data obtained using the IDIAP Research Institute's Smart Meeting Room – please refer to [17] for details. The array used is a planar, uniform circular array with $M=8$ omnidirectional microphones and a radius of 10 cm. The planar geometry of the array, coupled with its small radius, means that in a far-field setting, only the azimuth angle of arrival can be reasonably extracted. Thus, the evaluation focuses on locating the source in the azimuth plane only.

The room dimensions are 8.2-by-3.6-by-2.4 m. The array rests on a centrally located table with dimensions 4.8-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 generalized cross-correlation (GCC) method of [18] using the phase transform (PHAT) is employed to compute the cross-correlation measurements. 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. The frame length is 1024 samples or 64 ms. 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 [17]) are taken into account; there are 1426 such frames.

Table 3: Performance of estimators in a real environment: percentage of anomalies (%) and RMS error (degrees).

	TDOA	SRP	MVDR	MAX-EIG	MCCC
Anomaly rate	29.10	35.27	36.25	35.20	30.43
RMS error	1.84	1.66	1.55	1.75	1.84

The results, listed in Table 3. The TDOA and MCCC techniques provide the lowest anomaly rates. The former algorithm's good performance is somewhat of a surprise. Notice that the SRP-PHAT method, generally considered to be the most robust localization algorithm to date, is outperformed by the proposed MCCC estimator in conjunction with the PHAT preprocessing. Further research is required to thoroughly evaluate the robustness of the proposed estimators in comparison to conventional two-stage techniques in real environments.

8. CONCLUSION

This paper has described how a nonlinear parametrization of the spatial correlation matrix allows for the extension of classical narrowband spectral estimation methods to the broadband source localization problem. In general, the linear predictive MCCC estimator shows the highest level of robustness to the effects of noise and reverberation. Nevertheless, acoustic signal location anomaly rates are still higher than would be desired in many applications, and the problem continues to be somewhat open; with the parameterized spatial correlation matrix serving as the framework for broadband localization, further research into how to exploit the location information contained in this matrix is likely to yield even more robust methods.

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] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Mag.*, pp. 67–94, July 1996.

[3] H. Wang and M. Kaveh, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wideband sources," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 823–831, Aug. 1985.

[4] D.H. Johnson, "The application of spectral estimation methods to bearing estimation problems," *Proc. IEEE*, vol. 70, pp. 1018–1028, Sept. 1982.

[5] R.O. Schmidt, "Multiple emitter location and signal parameter estimation", *IEEE Trans. Antennas Propag.*, vol. AP-34, no.3, pp. 276 – 280, March 1986.

[6] 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.), pp. 157–180, Springer-Verlag, Berlin, 2001.

[7] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 499–508, Sept. 2004.

[8] A. Johansson and S. Nordholm, "Robust acoustic direction of arrival estimation using Root-SRP-PHAT, a realtime implementation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2005, vol. 4, pp. 933–936.

Table 1: Percentage of anomalies (%) in a simulated environment.

Parameters	**TDOA**	**SRP**	**MVDR**	**MAX-EIG**	**MCCC**
white source, $T_{60} = 0$ ms	0	0	0	0	0
white source, $T_{60} = 300$ ms	0	0	0	0	0
white source, $T_{60} = 600$ ms	2.35	0	0	0	0
speech source, $T_{60} = 0$ ms	18.16	16.45	16.03	18.80	19.87
speech source, $T_{60} = 300$ ms	56.52	44.02	44.44	44.02	33.55
speech source, $T_{60} = 600$ ms	73.50	57.27	52.35	58.97	44.87

Table 2: RMS error (degrees) for nonanomalous estimates in a simulated environment.

Parameters	**TDOA**	**SRP**	**MVDR**	**MAX-EIG**	**MCCC**
white source, $T_{60} = 0$ ms	1.00	1.00	1.00	1.00	1.00
white source, $T_{60} = 300$ ms	1.00	1.00	1.00	1.00	1.00
white source, $T_{60} = 600$ ms	1.48	1.04	1.02	1.13	1.00
speech source, $T_{60} = 0$ ms	1.60	1.87	1.92	1.81	1.77
speech source, $T_{60} = 300$ ms	2.24	2.34	2.42	2.36	2.88
speech source, $T_{60} = 600$ ms	2.44	2.52	2.27	2.42	3.08

[9] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1481–1494, Oct. 1989.

[10] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech and Language Proceessing*, vol. 15, pp. 1327 – 1339, May. 2007.

[11] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using eigenanalysis of the parameterized spatial correlation matrix," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, 2007.

[12] D. E. Dudgeon and D. H. Johnson, *Array Signal Processing*, Prentice-Hall, NJ, 1993.

[13] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 549–557, Nov. 2003.

[14] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross-correlation," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 509–519, Sept. 2004.

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

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

[17] 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.

[18] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 320–327, Aug. 1976.