Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 59625, Pages 1–17
DOI 10.1155/ASP/2006/59625

Microphone Array Speaker Localizers Using Spatial-Temporal Information

Sharon Gannot[1] and Tsvi Gregory Dvorkind[2]

[1] *School of Engineering, Bar-Ilan University, Ramat-Gan 52900, Israel*
[2] *Department of Electrical Engineering, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel*

Received 20 January 2005; Revised 17 May 2005; Accepted 22 August 2005

A dual-step approach for speaker localization based on a microphone array is addressed in this paper. In the first stage, which is not the main concern of this paper, the time difference between arrivals of the speech signal at each pair of microphones is estimated. These readings are combined in the second stage to obtain the source location. In this paper, we focus on the second stage of the localization task. In this contribution, we propose to exploit the speaker's smooth trajectory for improving the current position estimate. Three localization schemes, which use the temporal information, are presented. The first is a recursive form of the Gauss method. The other two are extensions of the Kalman filter to the nonlinear problem at hand, namely, the *extended Kalman filter* and the *unscented Kalman filter*. These methods are compared with other algorithms, which do not make use of the temporal information. An extensive experimental study demonstrates the advantage of using the spatial-temporal methods. To gain some insight on the obtainable performance of the localization algorithm, an approximate analytical evaluation, verified by an experimental study, is conducted. This study shows that in common TDOA-based localization scenarios—where the microphone array has small interelement spread relative to the source position—the elevation and azimuth angles can be accurately estimated, whereas the Cartesian coordinates as well as the range are poorly estimated.

1. INTRODUCTION AND PROBLEM FORMULATION

Determining the spatial position of a speaker finds a growing interest in video conference scenarios where automated camera steering and tracking are required. Acoustic source localization might also be used as a preprocessor stage for speech enhancement algorithms, which are based on microphone array beamformers.

Usually, methods for speaker localization are comprised of two stages. In the first stage, which is not the main concern of this paper, microphone array is used for extracting the time difference between arrivals of the speech signal at each pair of microphones. These readings are then processed by the second stage to obtain the source position. This paper focus is on the second algorithmic stage of the two-step approaches.

In the first algorithmic stage, the time difference of arrival (TDOA) is estimated using spatially separated microphone pairs. The classical method for performing this task is the generalized cross-correlation (GCC) algorithm [1]. Many improvements of this method for the reverberant case exist. Brandstein and Silverman used a robust estimate of the cross-power spectral density phase [2]. A cepstrum-based prefilter applied to the received signals prior to the application of the cross-correlation is proven by Stéphene and Champagne to be beneficial [3]. Benesty [4] and Doclo and Moonen [5] are using subspace tracking methods for performing the designated task. Recently, Dvorkind and Gannot [6–8] proposed a method for TDOA estimation, based on the nonstationarity of the speech signal, which was proven to be superior to the other methods in tracking scenarios.

During the second algorithmic stage, the noisy TDOA readings are combined to produce the source location estimate. The locus of speaker positions associated with a given microphone pair, from which we have extracted a TDOA measurement, forms one half of a hyperboloid of two sheets. By intersecting hyperboloid surfaces, one can estimate the speaker position [9]. However, this formulation is hard to compute in 3-dimensional space and tends to be noise sensitive (since small measurement errors can divert the intersection curve significantly). Another approach is useful in far-field applications, where the hyperboloid is approximated by a cone (centered at the midpoint of the microphone pair). By intersecting the bearing lines associated with such cones, location estimate can be derived by properly weighting the potential source locations according to the likelihood of the measurement. Brandstein et al. denote this method by *linear intersection* estimate [10].

By manipulating the measurement model, as will be shown in the sequel, the hyperbolic equations can be recast into a spherical form. The obtained equation set is shown to be nonlinear. Since the number of equations increases with the number of microphones, the noisy case can be solved by applying the (nonlinear) least squares (LS) approach.

The nonlinear LS problem yields a cumbersome expression. This difficulty might be alleviated in several ways. Three methods provide a closed-form solution, which differ in the way they mitigate the nonlinearity. The *spherical intersection* (SX) method was proposed by Schau and Robinson [11]. The spherical interpolation (SI) was proposed by Smith and Abel [12], while Huang et al. proposed the *one-step least squares* (OSLS) method [13]. Dealing with the differences between these methods is beyond the scope of this short survey.

Recently, Huang et al. [14] addressed the same nonlinear equation set and solved it by using Lagrange multiplier. Since a polynomial of degree six is involved in the proposed method, no closed-form solution exists. Thus, the iterative secant method [15] was used for the root search. The two-step approach is referred to as linear correction least squares (LCLS) approach. We will elaborate more on this method while formulating the problem.

Direct maximum likelihood-based algorithms are widely used in the localization task. Maximum likelihood (ML) processors require a priori knowledge of the joint probability density function of the errors in the TDOAs, and need search-based algorithms for determining the maximizer. Yao et al. [16] proposed a frequency-domain, one-step, approximate ML estimator for extracting both the source location and the received signal spectrum. They also proposed an iterative method for dealing with multiple source scenarios. Chen et al. further developed this concept and presented the Cramér-Rao lower bound (CRLB) for the localization problem in [17]. When the microphones locations are not known exactly, a two-stage estimation procedure is proposed, where iterations are performed between the ML estimation stage and a calibration stage. In the ML context, Segal et al. work should be mentioned, in which the estimate-maximize (EM) procedure is applied (in the frequency domain) for estimating both the position of several sources and their respective parameters [18]. Birchfield and Gillmor [19] utilized Bayes rule to obtain an ML estimator for the source location. In a simplified, reverberant-free room, the proposed method is shown to be more robust against additive noise than the conventional beamformer. Chen et al. [17] proposed the use of two beamformers with several look directions for extracting several candidate azimuth angles. A majority-based rule is then used for estimating the azimuth angle of the source.

All the prementioned methods exploit the spatial information obtained by different microphone pairs, but do not exploit the temporal information available from adjoint speaker position estimates. The speaker smooth trajectory can be used to obtain a more robust localization estimate. Bayesian estimation procedures were previously proposed by Ward et al. [20] and Vermaak and Blake [21]. In the former, a particle filter is used in conjunction with a beamformer to estimate the speaker position in a one-stage procedure. In the latter, the reverberation model is considered through a bimodal distribution of the noisy measurement around the true TDOA. Utilizing this distribution and giving a first-order Markov process model for the speaker trajectory, a particle filter is derived and applied to the problem at hand. Lehmann and Williamson [22] also used the particle filter. However they incorporate the importance sampling (IS) concept, in which particles are generated in each time step, based on the previous time step and the current measurement. The importance function is implemented based on a delay-and-sum beamforming results. Bechler et al. [23] proposed the use of a two-stage algorithm. In the first, the TDOA readings are used by the OSLS method [13] to obtain an initial estimate of the speaker position. These estimates are spatially smoothed by using three parallel linear Kalman filters. Each of the filters is using a different state transition model, namely, static, constant velocity, and constant acceleration. The three Kalman filters are weighted according to their a posteriori probability given the measurements. Klee and McDonough [24] showed by simulation results that the intermediate stage, in which source is localized by the SX method before applying the Kalman filter, deteriorates the overall performance. They proposed instead to apply the *iterated extended Kalman filter* directly on the TDOA readings.

In [25] we introduced two methods for exploiting the speaker's smooth trajectory for improving the tracking ability of source localizers, namely, a recursive Gauss (RG) method and the extended Kalman filter (EKF). These methods were compared with several nontemporal methods. In [26] the use of the unscented Kalman filter (UKF) for the problem at hand was proposed. The current contribution, which is an extension of the ideas presented in both [7, 26], includes a more detailed exposition of the ideas and a comprehensive comparative experimental study.

We turn now to an exact formulation of the localization problem. Consider an $M + 1$ microphones array as depicted in Figure 1. The microphones are placed at the Cartesian

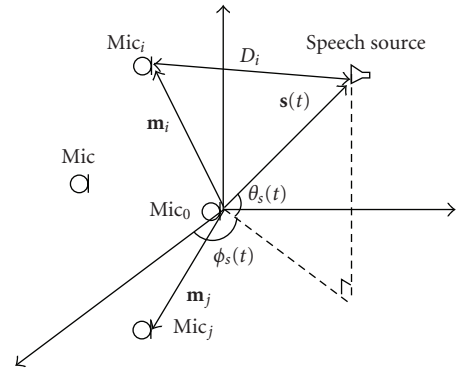

Figure 1: Microphone array. Speaker location at time instant t is $\mathbf{s}(t)$ with azimuth angle $\phi_s(t)$ and elevation angle $\theta_s(t)$. Microphone position notated by $\mathbf{m}_i$; $i = 0, \ldots, M$.

coordinates $\mathbf{m}_i \triangleq [x_i \;\; y_i \;\; z_i]^T$; $i = 0,\ldots,M$. To simplify the exposition, the location of a reference microphone $\mathbf{m}_0$ is set as the axes origin $\mathbf{m}_0 = [0 \;\; 0 \;\; 0]^T$. $(\cdot)^T$ stands for the transpose operation. Define the source coordinates at time instant t by $\mathbf{s}(t) \triangleq [x_s(t) \;\; y_s(t) \;\; z_s(t)]^T$. Each of the M microphones, combined with the reference microphone, is used at time instant t to extract a TDOA measurement $\tau_i(t)$; $i = 1,\ldots,M$ [8]. Denote the ith range difference measurement by $r_i(t) = c\tau_i(t)$, where c is the sound propagation speed (approximately 340 m/s in air). It can be easily verified from simple geometrical considerations (see Figure 1) that this range difference is related to the source and the microphone location by the nonlinear equation

$$r_i(t) = \left|\left|\mathbf{s}(t) - \mathbf{m}_i\right|\right| - \left|\left|\mathbf{s}(t)\right|\right|, \quad i = 1,\ldots,M, \tag{1}$$

where the fact that the reference microphone is positioned at the origin was used.

Usually, only an estimate of the real TDOA is available. Thus, concatenating M estimates of the quantity in (1), a nonlinear measurement model is obtained:

$$\hat{\mathbf{r}}(t) = \begin{bmatrix} \left|\left|\mathbf{s}(t) - \mathbf{m}_1\right|\right| - \left|\left|\mathbf{s}(t)\right|\right| \\ \vdots \\ \left|\left|\mathbf{s}(t) - \mathbf{m}_M\right|\right| - \left|\left|\mathbf{s}(t)\right|\right| \end{bmatrix} + \mathbf{v}(t) \triangleq \mathbf{h}(\mathbf{s}(t)) + \mathbf{v}(t). \tag{2}$$

Here, $\mathbf{v}^T(t) = [v_1(t) \;\; v_2(t) \;\; \cdots \;\; v_M(t)]$ is a vector of measurement errors, depicting the nonperfect estimate of the range differences. The goal of the localization task is to extract the speaker's trajectory $\mathbf{s}(t)$ from the measurements vector $\hat{\mathbf{r}}(t)$. Any estimation procedure (e.g., [1, 4, 5] or [8]) could be used for the TDOA estimation. The methods introduced in this contribution, constituting the second stage of the localization procedure, are independent of the choice of the first stage.

Following the derivation presented in [11–14], a practical approach for solving the nonlinear problem can be derived. Defining the distance between the speaker and the ith microphone as $D_i(t) \triangleq \|\mathbf{s}(t) - \mathbf{m}_i\|$ (see Figure 1), we get

$$D_i^2(t) = \left|\left|\mathbf{s}(t) - \mathbf{m}_i\right|\right|^2 = \left|\left|\mathbf{s}(t)\right|\right|^2 - 2\mathbf{m}_i^T\mathbf{s}(t) + \left|\left|\mathbf{m}_i\right|\right|^2. \tag{3}$$

However, using (1), the estimated distance is given by

$$\hat{D}_i(t) = \hat{r}_i(t) + \left|\left|\mathbf{s}(t)\right|\right|, \quad i = 1,\ldots,M. \tag{4}$$

An estimator of the speaker location is derived by minimizing the error between the estimated and the true squared distance:

$$\begin{aligned} \epsilon_i(t) &\triangleq \frac{1}{2}\left(\hat{D}_i^2(t) - D_i^2(t)\right) \\ &= \mathbf{m}_i^T\mathbf{s}(t) + \hat{r}_i(t)\left|\left|\mathbf{s}(t)\right|\right| \\ &\quad - \frac{1}{2}\left(\left|\left|\mathbf{m}_i\right|\right|^2 - \hat{r}_i^2(t)\right), \quad i = 1,\ldots,M. \end{aligned} \tag{5}$$

Concatenating the equations in (5), we have

$$\boldsymbol{\epsilon}(t) = A(t)\mathbf{g}(\mathbf{s}(t)) - \mathbf{b}(t), \tag{6}$$

where

$$A(t) \triangleq \begin{bmatrix} x_1 & y_1 & z_1 & \hat{r}_1(t) \\ x_2 & y_2 & z_2 & \hat{r}_2(t) \\ & & \vdots & \\ x_M & y_M & z_M & \hat{r}_M(t) \end{bmatrix},$$

$$\mathbf{b}(t) \triangleq \frac{1}{2}\begin{bmatrix} \left|\left|\mathbf{m}_1\right|\right|^2 - \hat{r}_1^2(t) \\ \left|\left|\mathbf{m}_2\right|\right|^2 - \hat{r}_2^2(t) \\ \vdots \\ \left|\left|\mathbf{m}_M\right|\right|^2 - \hat{r}_M^2(t) \end{bmatrix}, \tag{7}$$

$$\mathbf{g}(\mathbf{s}(t)) \triangleq \begin{bmatrix} x_s(t) \\ y_s(t) \\ z_s(t) \\ \left|\left|\mathbf{s}(t)\right|\right| \end{bmatrix}, \qquad \boldsymbol{\epsilon}(t) \triangleq \begin{bmatrix} \epsilon_1(t) \\ \epsilon_2(t) \\ \vdots \\ \epsilon_M(t) \end{bmatrix}.$$

The estimation problem is thus converted into a minimization problem of the quantity $\boldsymbol{\epsilon}^T(t)\boldsymbol{\epsilon}(t)$ with respect to the nonlinear functional $\mathbf{g}(\mathbf{s}(t))$. Since the fourth component of the vector $\mathbf{g}(\mathbf{s}(t))$ is related to the first three, the minimization problem becomes a constrained LS problem.

In [14] this problem was solved by using the *Lagrange multipliers* technique yielding

$$\hat{\mathbf{g}}(\mathbf{s}(t)) = \left(A^T(t)A(t) + \lambda\Sigma\right)^{-1}A^T(t)\mathbf{b}(t), \tag{8}$$

where $\Sigma \triangleq \mathrm{diag}[1 \;\; 1 \;\; 1 \;\; -1]$[1] and λ is the Lagrange multiplier, imposing the (quadratic) constraint on $\mathbf{g}(\mathbf{s}(t))$ structure. It can be shown that λ is obtained by finding the roots of a polynomial of degree six. Due to the complexity of the polynomial equation, numerical methods for root finding should be used. Therefore it is proposed in [14] to first solve the unconstrained LS problem and then use a linear correction

[1] We denote by $\mathrm{diag}(m_1, m_2, \ldots)$ a diagonal matrix with $m_1, m_2, \ldots$ on its main diagonal.

in the second phase. The method was hence denoted by the LCLS approach. We note that this approach lacks the temporal information as it makes no use of the fact that an estimate of $\mathbf{s}(t)$ should be spatially close to the estimate obtained during the previous time instant.

The organization of the rest of the paper is as follows. In Section 2 we derive a solution to the nonlinear problem using Gauss iterations. We proceed by approximating this batch solution by a recursive version applicable for tracking scenarios. The obtained RG solution constitutes our first spatial-temporal solution to the localization problem. Other spatial-temporal solutions can be derived by introducing a Bayesian framework for the problem at hand. The first solution, discussed in Section 3, is the well-known EKF, commonly applied to nonlinear optimal filtering problems. Less known nonlinear extension of the Kalman filter is introduced in Section 4, where the recently proposed UKF is applied to the speaker tracking problem. The CRLB on the position estimate is calculated in Section 5 for the simple unimodal noise model. In a typical TDOA-based localization scenario, the microphone array has small interelement spread relative to the source position. An approximate calculation shows that while the Cartesian coordinate estimation bound might become extremely high, the polar coordinates estimation bound is relatively small. We conclude this work in Section 6 by presenting an extensive simulation study for several test scenarios, showing the advantage of the spatial-temporal methods over the spatial-only methods.

2. GAUSS AND RECURSIVE GAUSS ALGORITHMS

The solution to the nonlinear problem in (6), presented by [14], involves several iterations for finding the Lagrange multiplier, due to the resulting sixth-order polynomial equation. We suggest an alternative method to mitigate the nonlinearity by using the Gauss method.

2.1. Gauss solution

Starting again from (6) we can state the nonlinear weighted LS (WLS) problem

$$\min_{\mathbf{s}(t)} \left[\mathbf{b}(t) - A(t)\mathbf{g}(\mathbf{s}(t))\right]^T W \left[\mathbf{b}(t) - A(t)\mathbf{g}(\mathbf{s}(t))\right] \tag{9}$$

with an arbitrary weighting matrix W. Note that (9) becomes a (nonlinear) LS problem if the number of microphone pairs fulfills $M > 3$, that is, if there are more equations than unknowns. This nonlinear set can be solved by applying the Gauss method rather than following [14]. The Gauss method, which is an iterative procedure for solving the nonlinear LS problem, is presented in Appendix A. Define $\mathbf{f}(\hat{\mathbf{s}}^{(l)}(t)) \triangleq A(t)\mathbf{g}(\hat{\mathbf{s}}^{(l)}(t))$ and the associated gradient matrix $F(\hat{\mathbf{s}}^{(l)}(t)) \triangleq \nabla_{\mathbf{s}(t)}\mathbf{f}(\hat{\mathbf{s}}^{(l)}(t))$ calculated at the current iteration (l). Gauss iterations for obtaining $\mathbf{s}(t)$ take the well-known form (see Appendix A):

$$\begin{aligned}\hat{\mathbf{s}}^{(l+1)}(t) = \hat{\mathbf{s}}^{(l)}(t) &+ \left[F^T\left(\hat{\mathbf{s}}^{(l)}(t)\right) W F\left(\hat{\mathbf{s}}^{(l)}(t)\right)\right]^{-1} \\ &\times F^T\left(\hat{\mathbf{s}}^{(l)}(t)\right) W\left[\mathbf{b}(t) - \mathbf{f}\left(\hat{\mathbf{s}}^{(l)}(t)\right)\right].\end{aligned} \tag{10}$$

This solution, as the solution in [14], only exploits the spatial information obtained by the separated microphone pairs at a specific time instant, but does not consider the temporal information.

2.2. RG procedure

Exploiting the temporal information embedded in the tracking problem necessitates the derivation of a recursive version of the Gauss method. We begin by concatenating (6) at all available measurements at time instances $1 \le \tau \le t$:

$$\begin{aligned}\boldsymbol{\epsilon}(1) &= A(1)\mathbf{g}(\mathbf{s}(1)) - \mathbf{b}(1) = \mathbf{f}(\mathbf{s}(1)) - \mathbf{b}(1), \\ \boldsymbol{\epsilon}(2) &= A(2)\mathbf{g}(\mathbf{s}(2)) - \mathbf{b}(2) = \mathbf{f}(\mathbf{s}(2)) - \mathbf{b}(2), \\ &\vdots \\ \boldsymbol{\epsilon}(t) &= A(t)\mathbf{g}(\mathbf{s}(t)) - \mathbf{b}(t) = \mathbf{f}(\mathbf{s}(t)) - \mathbf{b}(t).\end{aligned} \tag{11}$$

Note that each of the equations is referring to a distinct unknown source location $\mathbf{s}(\tau)$; $\tau = 1, \ldots, t$, and can be independently solved by using the iterative Gauss method of Section 2.1. However, since we assume that the source position $\mathbf{s}(t)$ is slowly varying with time, a more efficient, recursive solution can be derived. Linearizing each of the equations in (11) around $\mathbf{s}^*(\tau)$, as in Appendix A, one obtains

$$\begin{aligned}\boldsymbol{\epsilon}(1) &\simeq \mathbf{b}(1) - \mathbf{f}(\mathbf{s}^*(1)) - F(\mathbf{s}^*(1))(\mathbf{s}(1) - \hat{\mathbf{s}}^*(1)), \\ \boldsymbol{\epsilon}(2) &\simeq \mathbf{b}(2) - \mathbf{f}(\mathbf{s}^*(2)) - F(\mathbf{s}^*(2))(\mathbf{s}(2) - \mathbf{s}^*(2)), \\ &\vdots \\ \boldsymbol{\epsilon}(t) &\simeq \mathbf{b}(t) - \mathbf{f}(\hat{\mathbf{s}}^*(t)) - F(\mathbf{s}^*(t))(\mathbf{s}(t) - \hat{\mathbf{s}}^*(t)).\end{aligned} \tag{12}$$

Assuming slow movement of the speaker, an initial guess for the speaker location at each time instant τ can be taken from its estimated location at the previous time instant. Namely, the recursion $\mathbf{s}^*(\tau) = \hat{\mathbf{s}}(\tau - 1)$ can be used. As no significant movement of the speaker is expected from one time instant to another, only one more Gauss iteration suffices for obtaining a new estimate. By this *stochastic approximation*, we obtain a fast adaptation procedure but yet taking into account past measurements for stabilizing the estimate.

Then, a recursive speaker location estimate is obtained by solving the linearized WLS problem:

$$\hat{\mathbf{s}}(t) = \arg\min_{\mathbf{s}(t)} \left\| \begin{bmatrix} F(\hat{\mathbf{s}}(0)) \\ \vdots \\ F(\hat{\mathbf{s}}(t-1)) \end{bmatrix} \mathbf{s}(t) - \begin{bmatrix} \mathbf{b}(1) - \mathbf{f}(\hat{\mathbf{s}}(0)) + F(\hat{\mathbf{s}}(0))\hat{\mathbf{s}}(0) \\ \vdots \\ \mathbf{b}(t) - \mathbf{f}(\hat{\mathbf{s}}(t-1)) + F(\hat{\mathbf{s}}(t-1))\hat{\mathbf{s}}(t-1) \end{bmatrix} \right\|_W^2 \tag{13}$$

with $\hat{\mathbf{s}}(0)$ being the initial estimate for the parameter set. Recalling that $\mathbf{f}(\mathbf{s}(t)) = A(t)\mathbf{g}(\mathbf{s}(t))$ and using the definitions of $A(t)$ and $\mathbf{g}(\mathbf{s}(t))$, we calculate the derivative matrix to be

$$\begin{aligned} F(\hat{\mathbf{s}}(\tau)) &= \nabla_{\mathbf{s}(\tau)} \mathbf{f}(\hat{\mathbf{s}}(\tau)) \\ &= \begin{bmatrix} \mathbf{m}_1^T + \hat{r}_1(\tau) \dfrac{\hat{\mathbf{s}}^T(\tau)}{||\hat{\mathbf{s}}(\tau)||} \\ \mathbf{m}_2^T + \hat{r}_2(\tau) \dfrac{\hat{\mathbf{s}}^T(\tau)}{||\hat{\mathbf{s}}(\tau)||} \\ \vdots \\ \mathbf{m}_M^T + \hat{r}_M(\tau) \dfrac{\hat{\mathbf{s}}^T(\tau)}{||\hat{\mathbf{s}}(\tau)||} \end{bmatrix}, \quad \tau = 0, 2, \ldots, t-1. \end{aligned} \tag{14}$$

For solving this WLS problem recursively, we further choose the weighting matrix to be[2]

$$\begin{aligned} W = \text{blkdiag}\{&\text{diag}(\alpha^t, \ldots, \alpha^t); \text{diag}(\alpha^{t-1}, \ldots, \alpha^{t-1}); \ldots; \\ &\text{diag}(\alpha, \ldots, \alpha); \text{diag}(1, \ldots, 1)\}, \end{aligned} \tag{15}$$

with parameter $0 < \alpha \leq 1$. Note that an equal weight is given to all measurement in each time instant, hence all microphone readings have the same weight, while past measurements are reweighted by a factor of α, hence exponentially discarding the history. By using this weighting matrix, a *recursive least squares* (RLS) [27] algorithm is easily derived.

Another practical issue concerns the computational burden. At each time instant new M equations become available (relating to the number of microphones M), resulting in an $M \times M$ matrix inversion at each RLS iteration. However, by properly varying the forgetting factor within the well-known RLS algorithm, the computational complexity can be further reduced. This procedure is described in Appendix B.

3. THE EXTENDED KALMAN FILTER

The source location problem can be stated in the Bayesian framework as well. In this framework a dynamic model for the source trajectory should be given. As the actual track is unknown, a simplified random walk model is used instead.

$$\mathbf{s}(t+1) = \Phi \mathbf{s}(t) + \mathbf{w}(t), \tag{16}$$

$\mathbf{w}(t)$ is the coordinate-wise temporally white driving noise with covariance matrix $Q(t)$, Φ is a transition matrix assumed to be close to the identity matrix. A nonlinear measurement model was given in (2). Note that in this framework we are using the original hyperbolic model without using the spherical exposition. The measurement model is repeated here for the clarity of the exposition:

$$\mathbf{r}(t) = \begin{bmatrix} ||\mathbf{s}(t) - \mathbf{m}_1|| - ||\mathbf{s}(t)|| \\ \vdots \\ ||\mathbf{s}(t) - \mathbf{m}_M|| - ||\mathbf{s}(t)|| \end{bmatrix} + \mathbf{v}(t) \triangleq \mathbf{h}(\mathbf{s}(t)) + \mathbf{v}(t), \tag{17}$$

where $v(t)$ is a temporally white measurement noise signal with covariance matrix $R(t)$. Note that we are treating here $\mathbf{r}(t)$ as a measured process rather than estimates of the true range difference. For that sake we have omitted the estimation notation from the equation.

Equations (16) and (2) constitute the state-space model of the problem at hand. Since this model is nonlinear (due to the measurement equation), the classical Kalman filter cannot be used for estimating the state vector. Hence, nonlinear extensions thereof are called upon. Therefore, we propose to use the EKF. This procedure only gives a suboptimal solution to the problem at hand. We note that the usage of similar EKF formulation was also suggested in [28] where the localization problem was addressed in the context of multipath problems in wireless communication.

We give here, for the completeness of the exposition, the calculations involved in the EKF aiming to solve the localization problem. The EKF is essentially a Kalman filter in which the nonlinearity is mitigated by linearizing the transition and measurement matrices in each time instant (a complete derivation of the EKF can be found in many textbooks, e.g., [27]). Note that, in our case, (16) is already linear. However the measurement model in (2) still needs to be linearized.

Assume that an estimate $\hat{\mathbf{s}}(t-1 \mid t-1)$ of the speaker location at time instant $t-1$ is known, as well as its corresponding error-covariance matrix, $P(t-1 \mid t-1)$. Then, recalling that the transition matrix is linear, the EKF recursion takes the following form.

(i) Propagation equations:

$$\begin{aligned} \hat{\mathbf{s}}(t \mid t-1) &= \Phi \hat{\mathbf{s}}(t-1 \mid t-1), \\ P(t \mid t-1) &= \Phi P(t-1 \mid t-1) \Phi^T + Q(t). \end{aligned} \tag{18}$$

[2] We denote by $\text{blkdiag}(M_1, M_2, \ldots)$ a block-diagonal matrix with the matrices $M_1, M_2, \ldots$ on its main diagonal.

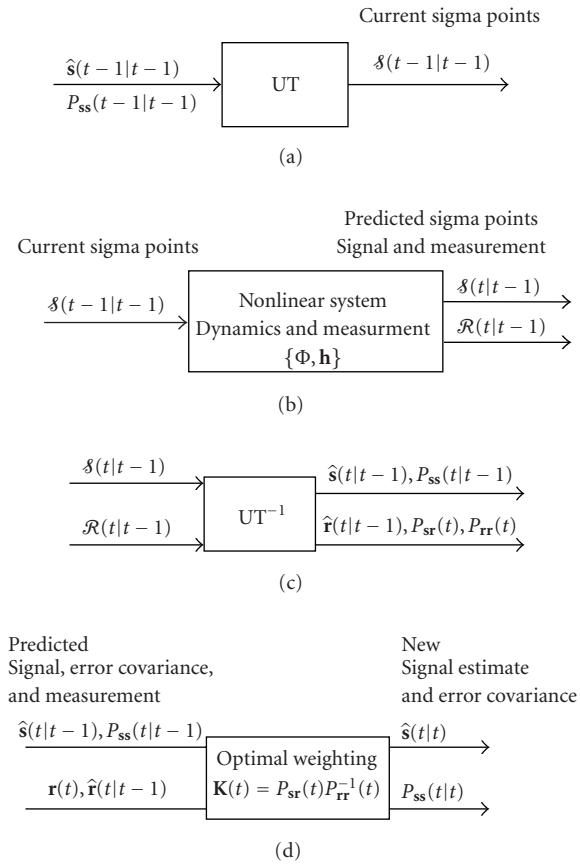

Figure 2: UKF: (a) UT, (b) propagation equations, (c) inverse UT, and (d) update equations.

(ii) Update equations:

$$\hat{\mathbf{s}}(t \mid t) = \hat{\mathbf{s}}(t \mid t-1) + K(t)\big(\mathbf{r}(t) - \mathbf{h}\big(\hat{\mathbf{s}}(t \mid t-1)\big)\big),$$

$$H(t) \triangleq \nabla_{\mathbf{s}(t)}\mathbf{h}\big(\hat{\mathbf{s}}(t \mid t-1)\big) = \begin{bmatrix} \left(\dfrac{\hat{\mathbf{s}}(t \mid t-1) - \mathbf{m}_1}{\|\hat{\mathbf{s}}(t \mid t-1) - \mathbf{m}_1\|} - \dfrac{\hat{\mathbf{s}}(t \mid t-1)}{\|\hat{\mathbf{s}}(t \mid t-1)\|}\right)^T \\ \vdots \\ \left(\dfrac{\hat{\mathbf{s}}(t \mid t-1) - \mathbf{m}_M}{\|\hat{\mathbf{s}}(t \mid t-1) - \mathbf{m}_M\|} - \dfrac{\hat{\mathbf{s}}(t \mid t-1)}{\|\hat{\mathbf{s}}(t \mid t-1)\|}\right)^T \end{bmatrix},$$

$$P(t \mid t) = \big(I - K(t)H(t)\big)P(t \mid t-1). \tag{19}$$

(iii) Kalman gain:

$$K(t) = P(t \mid t-1)H^T(t)\big(H(t)P(t \mid t-1)H^T(t) + R(t)\big)^{-1} \tag{20}$$

with the initialization $\hat{\mathbf{s}}(0 \mid -1)$ and its respective covariance $P(0 \mid -1)$.

4. THE UNSCENTED KALMAN FILTER

The EKF is not the only possible procedure for mitigating the nonlinearity in recursive optimal estimation. Julier and Uhlmann [29] proposed to use the UKF rather than the EKF for nonlinear recursive estimation problems and showed that an improved performance may be obtained.

Figure 2 summarizes the steps involved in the UKF. The method consists of calculating the mean and covariance of a state vector, undergoing a known nonlinear transform by using the unscented transform (UT). For details on the UT, the reader is referred to Appendix C.

Denote by $\hat{\mathbf{s}}(t-1 \mid t-1)$ the current source position estimate and by $P_{\mathbf{ss}}(t-1 \mid t-1)$ its respective covariance. The method is comprised of four stages. In stage (a), $\hat{\mathbf{s}}(t-1 \mid t-1)$ is split into σ-points $\mathcal{S}(t-1 \mid t-1)$ approximating the probability density function of the state vector (see [29]). By using this method, the mean and covariance propagate through the nonlinearities better than in the EKF method. However, no claims of optimality hold. Then, in stage (b), each of the σ-points is undergoing the known nonlinearity yielding the σ-points of the *predicted* state vector, $\mathcal{S}(t \mid t-1)$. The σ-points of the predicted noisy measurement, $\mathcal{R}(t \mid t-1)$, are calculated as well. In step (c), the σ-points are collected together yielding the predicted values $\hat{\mathbf{s}}(t \mid t-1)$ and $\hat{\mathbf{r}}(t \mid t-1)$. This concludes the propagation stage of the UKF. In step (d), similar to the conventional filter, the Kalman gain is calculated by $\mathbf{K}(t) = P_{\mathbf{sr}}(t)P_{\mathbf{rr}}^{-1}(t)$. Note that the covariance matrices estimates are obtained by the UT. Finally, the update stage is implemented by properly weighting the predicted values and the current measurement yielding the new source location estimate $\hat{\mathbf{s}}(t \mid t)$ and its respective covariance $P_{\mathbf{ss}}(t \mid t)$.

Similar to the EKF, (16) and (2) constitute the state and measurement equations for the UKF. As the nonlinearity is known, the UKF can be applied for solving the localization problem.

5. THE CRAMÉR-RAO LOWER BOUND

Calculating a bound for the performance of the localizer in the dynamic case is a cumbersome task. To get a rough estimate of the predicted performance, following [14], we assume a simplified model of the source locations. Specifically, we assume that the true range difference readings in the measurement equation (2) are contaminated by Gaussian distributed noise with zero-mean and covariance matrix $C_{\mathbf{v}}$. Note that the existence of directional interferences and reverberation phenomenon might cause high level of noise correlation between microphone pairs and across time. Moreover, in high noise level the TDOA estimation algorithm might produce readings related to the directional noise source, causing multimodal noise distribution. Nevertheless, for simplicity, we start by assuming (like Huang et al. [14]) that the noise is unimodal (Gaussian) distributed spatially and temporally white. Now, CRLB for unbiased estimation of the source position can be calculated.

Huang et al. [14] calculated the CRLB in Cartesian coordinates:

$$J(\mathbf{s}(t)) = G^T C_{\mathbf{v}}^{-1} G, \tag{21}$$

where

$$G = \begin{bmatrix} \left(\dfrac{\mathbf{s}(t) - \mathbf{m}_1}{||\mathbf{s}(t) - \mathbf{m}_1||} - \dfrac{\mathbf{s}(t)}{||\mathbf{s}(t)||} \right)^T \\ \vdots \\ \left(\dfrac{\mathbf{s}(t) - \mathbf{m}_M}{||\mathbf{s}(t) - \mathbf{m}_M||} - \dfrac{\mathbf{s}(t)}{||\mathbf{s}(t)||} \right)^T \end{bmatrix}. \tag{22}$$

Note that as no temporal information was used, the obtained result is time independent. When temporal information is used, the calculations become too complex to be evaluated analytically. However, we may assume that the obtainable bound should be lower.

It is interesting to evaluate the CRLB in polar coordinates. Define the transformation from the Cartesian coordinates $\mathbf{s}(t) = [x_s(t) \;\; y_s(t) \;\; z_s(t)]^T$ to the polar coordinates $\mathbf{s}_p(t) \triangleq [\phi_s(t) \;\; \theta_s(t) \;\; \rho_s(t)]^T$ as

$$\begin{aligned} \rho_s(t) &= \sqrt{x_s^2(t) + y_s^2(t) + z_s^2(t)}, \\ \phi_s(t) &= \cos^{-1}\left(\frac{x_s(t)}{\sqrt{x_s^2(t) + y_s^2(t)}} \right), \\ \theta_s(t) &= \sin^{-1}\left(\frac{z_s(t)}{\rho_s(t)} \right). \end{aligned} \tag{23}$$

The Jacobian of the transformation (in Cartesian coordinates terms) can be easily verified to be

$$P(\mathbf{s}(t)) = \begin{bmatrix} -\dfrac{y_s(t)}{x_s^2(t) + y_s^2(t)} & \dfrac{x_s(t)}{x_s^2(t) + y_s^2(t)} & 0 \\ -\dfrac{z_s(t)x_s(t)}{(x_s^2(t) + y_s^2(t) + z_s^2(t))\sqrt{x_s^2(t) + y_s^2(t)}} & -\dfrac{z_s(t)y_s(t)}{(x_s^2(t) + y_s^2(t) + z_s^2(t))\sqrt{x_s^2(t) + y_s^2(t)}} & \dfrac{\sqrt{x_s^2(t) + y_s^2(t)}}{x_s^2(t) + y_s^2(t) + z_s^2(t)} \\ \dfrac{x_s(t)}{\sqrt{x_s^2(t) + y_s^2(t) + z_s^2(t)}} & \dfrac{y_s(t)}{\sqrt{x_s^2(t) + y_s^2(t) + z_s^2(t)}} & \dfrac{z_s(t)}{\sqrt{x_s^2(t) + y_s^2(t) + z_s^2(t)}} \end{bmatrix}. \tag{24}$$

Therefore, the CRLB in polar coordinates is given by

$$J(\mathbf{s}_p(t)) = P(\mathbf{s}(t)) J(\mathbf{s}(t)) P(\mathbf{s}(t))^T. \tag{25}$$

In a typical TDOA-based localization scenarios, the microphone array has small interelement spread relative to the source position. As the microphone separation distance is relatively small, it allows for an efficient calculation of the TDOA readings. In such circumstances, as we will also demonstrate by our simulative study of Section 6, the obtainable performance in polar coordinates (concerning only the estimate of the azimuth and the elevation angles in far-field scenario) is superior to the obtainable performance in Cartesian coordinates. For that reason we will present throughout this work the results transformed into polar coordinates.

6. EXPERIMENTAL STUDY

In this section we compare the performance obtained by the various localization methods presented in this work. We start by evaluating the CRLB for a simplified unimodal scenario. This calculation leads us to a conclusion that the meaningful information lies in the azimuth and elevation angles rather than in the Cartesian coordinates or the range information. Fortunately, these angle estimates are sufficient for camera steering applications. We proceed by assessing the performance of five localization methods presented in this work. Namely, the two nontemporal methods (LCLS and Gauss iterations) and the three spatial-temporal methods (RG, EKF, and UKF). The methods are first assessed by using artificially contaminated true TDOA readings, in which the speaker is moving along a helix-shaped trajectory. We then proceed with a more realistic scenario for which the available data are estimated TDOA readings obtained from alternating speakers. The TDOA readings are extracted by a previously proposed method, which exploits speech nonstationarity [8]. It was shown that this method (notated RS1 in [8]) outperforms other state-of-the-art algorithms.

6.1. Test scenario

A set of eight microphones is placed on a sphere of radius 0.9 m around a reference microphone placed at the origin,

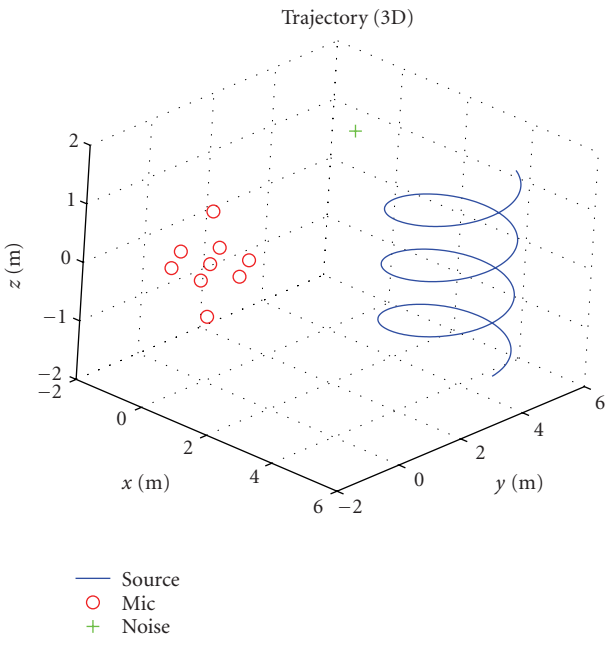

Figure 3: Speaker trajectory, noise position, and microphones positions.

$\mathbf{m}_0^T = [0 \;\; 0 \;\; 0]$, at the following positions:[3]

$$\begin{aligned}
\mathbf{m}_1^T &= \begin{bmatrix} 0.9 & 0 & 0 \end{bmatrix}, & \mathbf{m}_2^T &= \begin{bmatrix} 0.45 & 0.7794 & 0 \end{bmatrix}, \\
\mathbf{m}_3^T &= \begin{bmatrix} -0.45 & 0.7794 & 0 \end{bmatrix}, & \mathbf{m}_4^T &= \begin{bmatrix} -0.9 & 0 & 0 \end{bmatrix}, \\
\mathbf{m}_5^T &= \begin{bmatrix} -0.45 & -0.7794 & 0 \end{bmatrix}, & \mathbf{m}_6^T &= \begin{bmatrix} 0.45 & -0.7794 & 0 \end{bmatrix}, \\
\mathbf{m}_7^T &= \begin{bmatrix} 0 & 0 & 0.9 \end{bmatrix}, & \mathbf{m}_8^T &= \begin{bmatrix} 0 & 0 & -0.9 \end{bmatrix}.
\end{aligned} \tag{26}$$

The speaker trajectory is set to a helix with a radius of $R = 1.5$ m, given in Cartesian coordinates by (27) and shown in Figure 3:

$$\begin{aligned}
x_s(t) &= R\left(\cos\left(\frac{t}{R}\right) + 2.5\right), \\
y_s(t) &= R\left(\sin\left(\frac{t}{R}\right) + 2.5\right), \\
z_s(t) &= \frac{t}{10} - 1.5.
\end{aligned} \tag{27}$$

The main axis of the helix is parallel to the z-axis, 3.75 m away from the origin. The speaker completes one full circle, $2\pi R$ meters long, in $2\pi R$ seconds, hence its tangent speed is 1 m/s. The speaker speed along the z-axis is set to 1/10 m/s. The time span of the trajectory is $t \in [0, T]$ and the total duration of the movement is $T = 30$ s. The entire scenario is depicted in Figure 3.

[3] All dimensions are in meters.

6.2. The CRLB evaluation

We now calculate the CRLB for the tested scenario. We assume that the true range difference (or, equivalently, the TDOA) readings are contaminated by a unimodal Gaussian distributed noise signal, with zero mean and standard deviation (STD) of $\sigma_v = 0.2$ m in each coordinate. This STD is equivalent to 4.7 samples at a sample rate of $F_s = 8000$ Hz. Under these conditions, the CRLB is calculated for both Cartesian and polar coordinates using the derivations in Section 5. The resulting bound (in meters for the Cartesian coordinates and the range, and in degrees for the azimuth and elevation angles) is depicted in Figure 4. The CRLB naturally depends on the source position. Using (27), we give the CRLB as a function of the time instant, as it completely parameterizes the speaker's trajectory. Note that the Cartesian coordinates, as well as the range, cannot be accurately estimated in this scenario. Actually, the obtainable STD renders the estimated quantity useless. However, the azimuth and elevation angles may be estimated in high accuracy. Fortunately, for camera steering applications, estimation of the azimuth and elevation angles suffices. Note also that the presented CRLB serves as a bound to the nontemporal methods alone, since past measurements are disregarded at each time instant.

Finally, we comment that the CRLB can be dramatically reduced to an acceptable level (especially, for the Cartesian coordinates and range) if, for instance, we set the radius of the array to 5 m instead of 0.9 m. The new microphone constellation and the associated CRLB is shown in Figure 5. However, the larger dimensions of the array impose huge computational burden on the first stage of the localizer, namely, the TDOA extraction. In this work, we will concentrate on the more practical scenario, where the speaker distance from the microphones is significantly larger than the array dimensions.

6.3. Artificially contaminated range difference

The setup presented in Section 6.1 is evaluated by five localization methods. The true range differences are assumed to be contaminated by spatially and temporally white Gaussian noise with covariance matrix $\mathrm{Cov}\{\mathbf{v}(t)\} = \sigma_v^2 I$, $\sigma_v = 0.2$ m.

The first localization algorithm is the LCLS method, presented by Huang et al. [14]. The second is the batch Gauss method (denoted BG) with three iterations at each time instant. The third is the RG with forgetting factor $\alpha = 0.85$. We emphasize that no attempt to optimize this quantity was made. The value of $\alpha = 0.85$ was set as a compromise between fast adaptation requirements and stable estimation. The fourth is the EKF method evaluated with random-walk model having driving noise with a STD of 0.5 m along each Cartesian coordinate, that is, $Q(t) = 0.5^2 I_3$. This value was chosen to be compatible with the assumed changing rate of the speaker's position. The performance was found to be robust to a wide region of this parameter values. Exact prior knowledge of the measurement noise is not assumed as well, and the measurement covariance matrix is deliberately

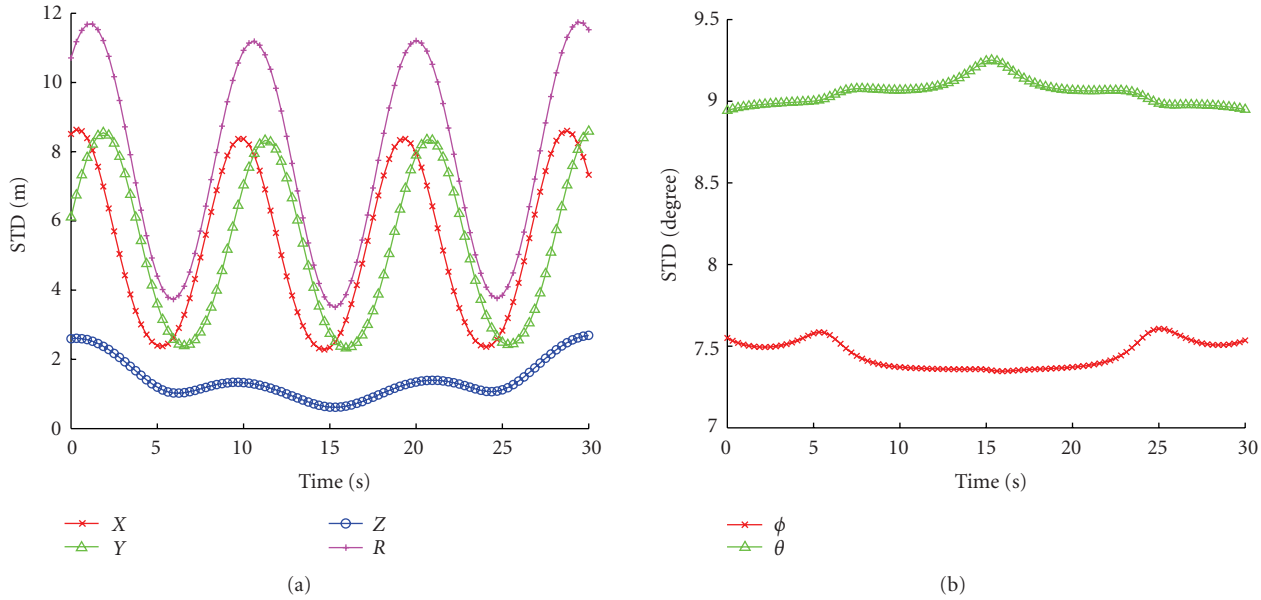

FIGURE 4: CRLB results for position estimate along the speaker trajectory for the scenario in Figure 3 with array radius set to 0.9 m. (a) Cartesian coordinates and range. (b) Azimuth (ϕ) and elevation (θ) angles.

overestimated to $R(t) = 10\sigma_v^2 I$; $\sigma_v = 0.2$ m. To allow a slight decay of past estimates, we set the transition matrix to the value $\Phi = 0.99I$. The fifth tested method is the UKF method using the same setup as the EKF. No attempt was made to adapt the parameters of the filters to a given scenario. One thousand Monte Carlo trials are performed to obtain a meaningful evaluation of the root mean square error (RMSE) of the angles estimate. The results for this setup are depicted in Figure 6.

We have also repeated this experiment with an additional point noise source which is placed at the $[0.5 \;\; 4 \;\; 1.5]^T$ coordinate (see Figure 3). By replacing 20% of the range difference readings by readings associated with the point noise location rather than the speech source position, we aim to simulate a scenario where, due to the directional interferer, the first localization stage, that is, the estimation of TDOA values, is disrupted by the point noise source.[4] Results for this scenario are depicted in Figure 7. As can be seen, for both scenarios, the LCLS method has better performance than the Gauss iterations method. However the RG which exploits the temporal information obtains better results. The EKF and the UKF methods remarkably outperform the other methods, with slight advantage to the latter. Overall, the results of the Kalman filter-based methods demonstrate acceptable performance even in these harsh conditions. By comparing Figures 6 and 7, we see that the obtainable performance in the first, anomaly-free case is better than that of the latter scenario. We also remark, that no advantage was gained by directly estimating the polar coordinates rather than transforming the estimates of Cartesian coordinates into polar coordinates.

We conclude this section by presenting in Figure 8 a typical realization for the tracking ability of both the EKF and UKF methods for the directional interference case. The small bias depicted in the figure is probably due to the fact that the Kalman-based localizers cannot track the fast maneuvering speaker in this specific setup.

6.4. Switching scenario

We proceed by testing a more realistic scenario. Consider the following simulation which is typical to a video conference scenario. Two speakers located at two different and fixed locations alternately speak. The camera should be able to maneuver from one person to the other. For this scenario, simulation is conducted with one speaker located at the polar position $[\phi = (\pi/4)\,\text{rad} \;\; \theta = (\pi/4)\,\text{rad} \;\; R = 1.5\,\text{m}]$ and the other at $[\phi = (3\pi/4)\,\text{rad} \;\; \theta = (\pi/3)\,\text{rad} \;\; R = 1.5\,\text{m}]$. A directional interference is placed at the position $[\phi = (\pi/2)\,\text{rad} \;\; \theta = (\pi/4)\,\text{rad} \;\; R = 1.0\,\text{m}]$. Six microphones were mounted at the following positions (in meters), relative to the reference microphone (which is at the axes origin):

$$\begin{aligned} \mathbf{m}_1^T &= \begin{bmatrix} 0.3 & 0 & 0 \end{bmatrix}, & \mathbf{m}_2^T &= \begin{bmatrix} -0.3 & 0 & 0 \end{bmatrix}, \\ \mathbf{m}_3^T &= \begin{bmatrix} 0 & 0.3 & 0 \end{bmatrix}, & \mathbf{m}_4^T &= \begin{bmatrix} 0 & -0.3 & 0 \end{bmatrix}, \\ \mathbf{m}_5^T &= \begin{bmatrix} 0 & 0 & 0.3 \end{bmatrix}, & \mathbf{m}_6^T &= \begin{bmatrix} 0 & 0 & -0.3 \end{bmatrix}. \end{aligned} \tag{28}$$

For this scenario, rather than adding white Gaussian noise to

[4] We note that the 80% true range difference readings are still corrupted by the white Gaussian noise, as in the previous scenario.

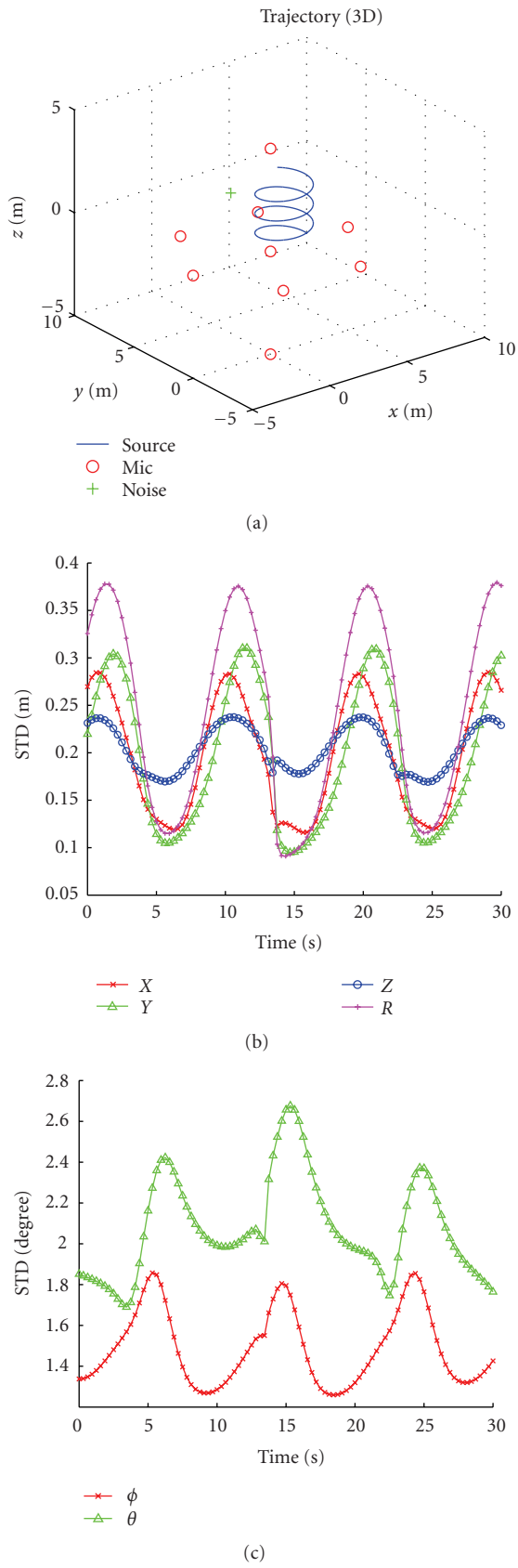

FIGURE 5: CRLB results. (a) Test scenario with array radius set to 5 m. (b) Cartesian coordinates and range. (c) Azimuth (ϕ) and elevation (θ) angles.

the true range differences, estimated TDOA values (equivalently, range differences) were used. We note that any method for TDOA extraction can be used in conjunction with our localization algorithm. However, to give specific simulations, we used TDOA readings, extracted from the noisy microphone data, by the RS1 algorithm described in [7, 8]. For that estimation stage, room reverberation (set to reverberation time of $T_r = 0.25$ s) and the directional interferer were taken into account. Room reverberation was simulated by the *image method* [30]. Mean SNR level was set to 10 dB. The same setup for the localization methods is applied here as well. Namely, the EKF and UKF localizers still use the random walk model though a better choice might have been asserted.

Figure 9 presents the azimuth angle estimates obtained by the five methods. Figure 10 presents the respective elevation angle estimates. As can be seen from the plots, the temporal methods, especially the EKF and UKF algorithms, clearly outperform the other methods. The transition instances are the main cause of errors in this scenario. While the batch methods (Gauss and LCLS) demonstrate unstable behavior in these regions, the recursive methods demonstrate smooth transition curves due to their inherent memory. Although the Kalman-based methods are not using a valid state-space model, their performance is obviously better than the nonrecursive methods. The UKF method obtains slightly better results than the EKF method in wide range of parameters' value selection. The computational burden of both methods is comparable.

7. CONCLUSIONS

We presented both nontemporal and temporal algorithms for talker localization and tracking. The nontemporal methods are commonly used in speech localization applications. Among the two batch methods, the LCLS method outperforms the Gauss method. Three temporal methods were derived. One is within a non-Bayesian framework (RG algorithm) and the other two are within the Bayesian framework, namely, the EKF and UKF algorithms. Both these Kalman filter-based methods are known to be computationally simpler than the particle filter. The UKF method marginally outperforms the EKF method for a wide range of parameters' values. Nevertheless, the imposed computational burden is almost equivalent. Evaluation of the CRLB showed that for a microphone array with a small interelement spread relative to the source position, angle estimates might be obtained reliably (as opposed to the Cartesian coordinates estimates). This justifies the use of polar coordinates rather than Cartesian coordinates in our simulations. Empirical results demonstrate the effectiveness of using the temporal information. Finally, we emphasize that only a simplified model was used in the Kalman-based methods and no attempt was made to optimize their parameters. However, we demonstrated that even with this simple model and without any optimization of the parameters, the temporal methods outperform the commonly used nontemporal methods. A more accurate model, in conjunction with the nonlinear

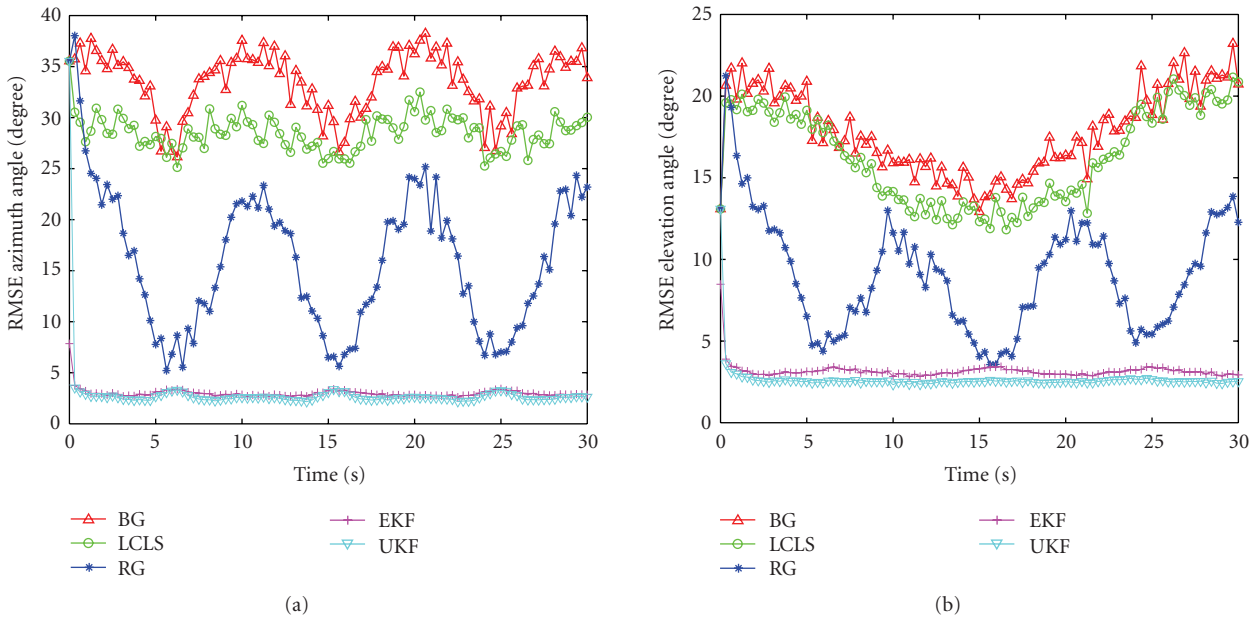

FIGURE 6: RMSE results averaging 1000 trials with white Gaussian noise. (a) Azimuth angle (ϕ). (b) Elevation angle (θ).

extensions of the Kalman filter, might be able to improve the tracking ability of the algorithms, in particular, at the abrupt changes instances.

APPENDICES

A. GAUSS METHOD

Consider the weighted nonlinear LS problem:

$$\min_{\mathbf{s}(t)} \left[\mathbf{b}(t) - \mathbf{f}(\mathbf{s}(t))\right]^T W \left[\mathbf{b}(t) - \mathbf{f}(\mathbf{s}(t))\right], \tag{A.1}$$

where W is an arbitrary weighting matrix. Expanding $\mathbf{f}(\mathbf{s}(t))$ to a Taylor series around $\mathbf{s}^*(t)$ and taking only first-order approximation, we obtain

$$\mathbf{f}(\mathbf{s}(t)) \simeq \mathbf{f}(\mathbf{s}^*(t)) + \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\left(\mathbf{s}(t) - \mathbf{s}^*(t)\right). \tag{A.2}$$

Define the error term $\boldsymbol{\epsilon}(t) \triangleq \mathbf{b}(t) - \mathbf{f}(\mathbf{s}(t))$. Then

$$\begin{aligned}\boldsymbol{\epsilon}(t) &\simeq \mathbf{b}(t) - \mathbf{f}(\mathbf{s}^*(t)) - \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\left(\mathbf{s}(t) - \mathbf{s}^*(t)\right)\\ &= \mathbf{b}(t) - \mathbf{f}(\mathbf{s}^*(t)) - \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\mathbf{s}(t)\\ &\quad + \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\mathbf{s}^*(t)\\ &= \widetilde{\mathbf{b}}(t) - \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\mathbf{s}(t),\end{aligned} \tag{A.3}$$

where $\widetilde{\mathbf{b}}(t) = \mathbf{b}(t) - \mathbf{f}(\mathbf{s}^*(t)) + \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))\mathbf{s}^*(t)$. Using the gradient matrix definition, $F(\mathbf{s}^*(t)) = \nabla_{\mathbf{s}(t)}\mathbf{f}(\mathbf{s}^*(t))$, we obtain a linearized LS problem:

$$\min_{\mathbf{s}(t)} \left[\widetilde{\mathbf{b}}(t) - F(\mathbf{s}^*(t))\mathbf{s}(t)\right]^T W \left[\widetilde{\mathbf{b}}(t) - F(\mathbf{s}^*(t))\mathbf{s}(t)\right]. \tag{A.4}$$

The LS solution is given by

$$\begin{aligned}\hat{\mathbf{s}}(t) &= \left[F^T(\mathbf{s}^*(t))\,WF(\mathbf{s}^*(t))\right]^{-1}F^T(\mathbf{s}^*(t))\\ &\quad \times W\left[\mathbf{b}(t) - \mathbf{f}(\mathbf{s}^*(t)) + F(\mathbf{s}^*(t))\mathbf{s}^*(t)\right]\\ &= \mathbf{s}^*(t) + \left[F^T(\mathbf{s}^*(t))\,WF(\mathbf{s}^*(t))\right]^{-1}F^T(\mathbf{s}^*(t))\\ &\quad \times W\left[\mathbf{b}(t) - \mathbf{f}(\mathbf{s}^*(t))\right].\end{aligned} \tag{A.5}$$

Since this solution is valid for any $\mathbf{s}^*(t)$, we can use it iteratively to obtain the Gauss method:

$$\begin{aligned}\hat{\mathbf{s}}^{(l+1)}(t) = \hat{\mathbf{s}}^{(l)}(t) &+ \left[F^T(\hat{\mathbf{s}}^{(l)}(t))\,WF(\hat{\mathbf{s}}^{(l)}(t))\right]^{-1}\\ &\times F^T(\hat{\mathbf{s}}^{(l)}(t))\,W\left[\mathbf{b}(t) - \mathbf{f}(\hat{\mathbf{s}}^{(l)}(t))\right]\end{aligned} \tag{A.6}$$

starting from an initial guess $\hat{\mathbf{s}}^{(0)}(t)$.

B. RLS FOR MULTIPLE READINGS

Assume a scenario in which for each time instant we have K scalar measurements $\mathbf{z}(\tau) \in \mathbb{R}^K$ related to an unknown $p \times 1$

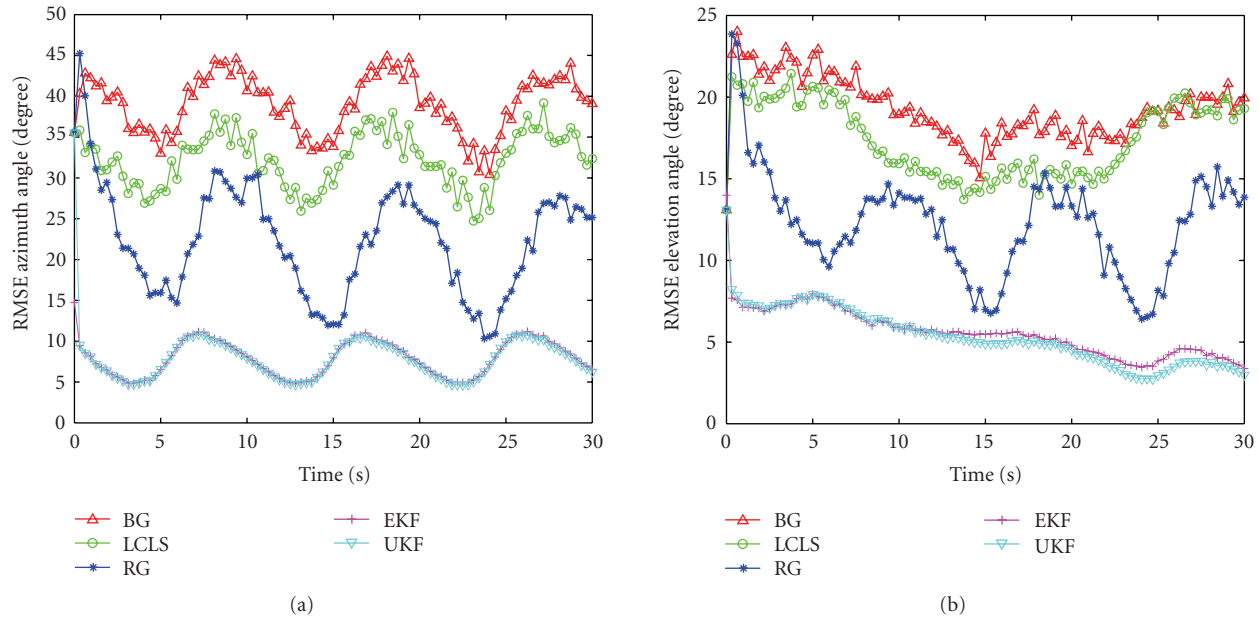

Figure 7: RMSE results averaging 1000 trial with white Gaussian noise and 20% anomaly. (a) Azimuth angle (ϕ). (b) Elevation angle (θ).

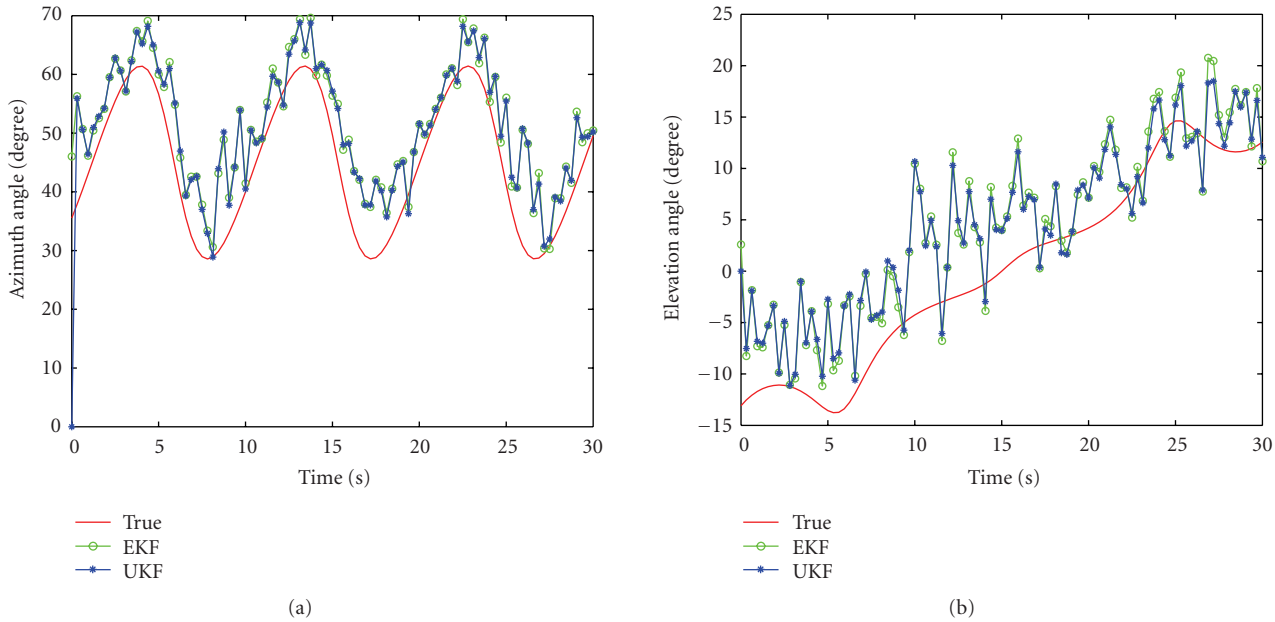

Figure 8: One realization of tracking results with white Gaussian noise and 20% anomaly for EKF and UKF. (a) Azimuth angle (ϕ). (b) Elevation angle (θ).

parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$ by a linear $K \times p$ transformation $H(\tau)$:

$$\mathbf{z}(\tau) \approx H(\tau)\boldsymbol{\theta}. \tag{B.1}$$

The approximation is due to the fact that the measurements are noisy or due to slight modelling errors. $\tau = 1, 2, \ldots, t$ time instants can be augmented to a matrix form $\mathbf{z}(1:t) \approx H(1:t)\boldsymbol{\theta}$ where

$$\mathbf{z}(1:t) \triangleq \begin{bmatrix} \mathbf{z}(1) \\ \mathbf{z}(2) \\ \vdots \\ \mathbf{z}(t) \end{bmatrix}, \qquad H(1:t) \triangleq \begin{bmatrix} H(1) \\ H(2) \\ \vdots \\ H(t) \end{bmatrix}. \tag{B.2}$$

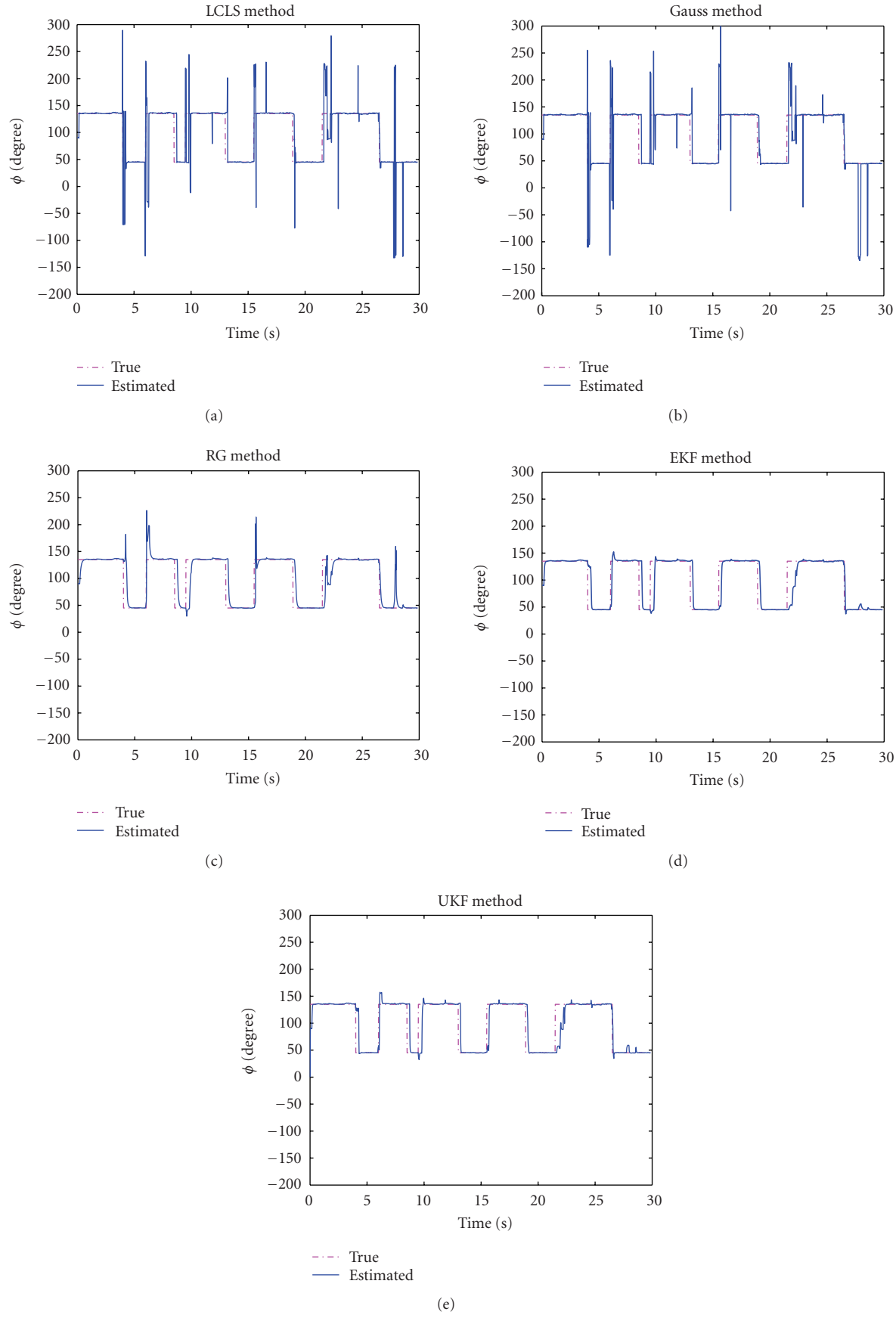

FIGURE 9: Azimuth angle ϕ estimation results. The method's name is presented in the title of each plot.

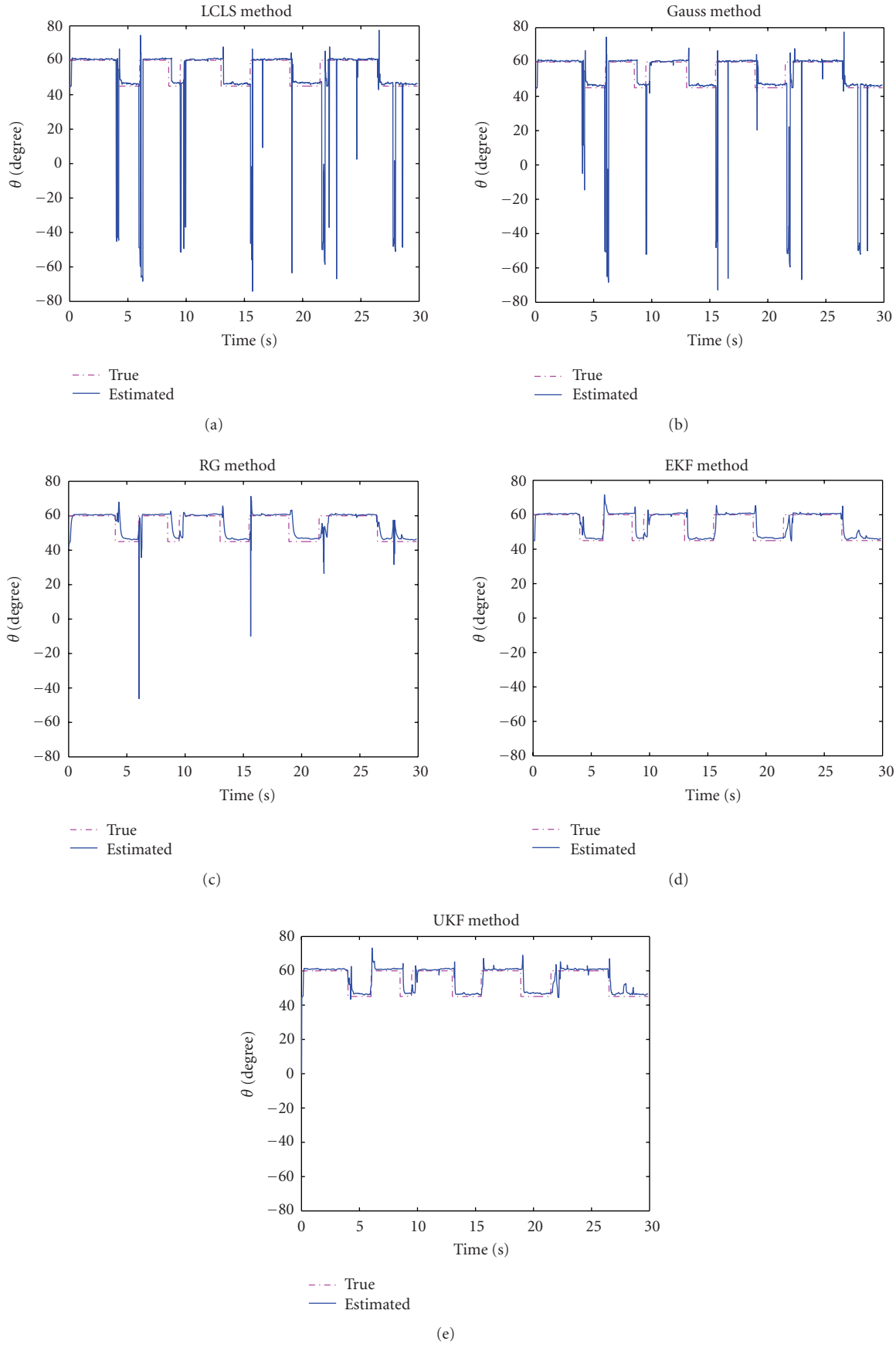

Figure 10: Elevation angle θ estimation results. The method's name is presented in the title of each plot.

The WLS solution for $\boldsymbol{\theta}$ using nonnegative weight matrix $W(1:t)$ (of size $Kt \times Kt$) is given by

$$\hat{\boldsymbol{\theta}} = \left(H(1:t)^T W(1:t) H(1:t)\right)^{-1} H(1:t)^T W(1:t)\mathbf{z}(1:t). \tag{B.3}$$

Our goal is to evaluate (B.3) recursively. If the parameters slowly change, a common approach is to apply a diagonal weight matrix $W(1:t)$ with powers of a forgetting factor $0 < \alpha \leq 1$ along its diagonal. Note that for measurements associated with the same time instant, we wish to apply the same factor, since equations of the same time instant have equal importance. Such weight matrix can be represented recursively as

$$W(1:t) = \begin{bmatrix} \alpha W(1:t-1) & \mathcal{O} \\ \mathcal{O}^T & I \end{bmatrix}, \qquad W_{1:1} = I, \tag{B.4}$$

where I and $\mathcal{O}$ stand for the identity and zero matrices of sizes $K \times K$ and $(t-1)K \times K$, respectively. At first glance it seems that a recursive solution to (B.3) necessitates $(K \times K)$-matrix inversion in each RLS iteration. However, in practice, the complexity can be further reduced. This is obtained by applying the well-known RLS algorithm with a minor twist. Consider a single equation. If this equation belongs to one of the K equations constituting the current time instant (but not the first one), a forgetting factor of 1 should be used. However, if this equation is the first at the new time instant τ, a forgetting factor $\alpha \leq 1$ must be used instead. Thus, in order to derive a recursion, where the update stage considers only a *single* equation, the forgetting factor should vary. Notating the time instant by τ ($\tau = 1, 2, \ldots$) and the sequential number of the equation by $(\tau - 1)K + k$ (where $k \in \{1, \ldots, K\}$), the forgetting factor becomes

$$\text{forgetting factor} = \begin{cases} \alpha, & k = 1, \\ 1, & \text{otherwise.} \end{cases} \tag{B.5}$$

It is easily verified that a matrix inversion is not necessary in this case.

C. THE UNSCENTED TRANSFORM

Let $\mathbf{x}$ be an L-dimensional random vector with mean $\bar{\mathbf{x}}$ and covariance matrix P_{xx}. Let $\mathbf{y} = f(\mathbf{x})$ be a nonlinear transformation from the random vector $\mathbf{x}$ to another random vector $\mathbf{y}$. The first- and second-order statistics of the vector $\mathbf{y}$ should be calculated. We briefly summarize the method proposed in [29]. The mean and covariance of $\mathbf{x}$ can be presented by the $2L + 1$ σ-points

$$\begin{aligned} \mathcal{X}_0 &= \bar{\mathbf{x}}, \\ \mathcal{X}_l &= \bar{\mathbf{x}} + \left(\sqrt{(L+\lambda)P_{xx}}\right)_l, \quad l = 1, \ldots, L, \\ \mathcal{X}_{l+L} &= \bar{\mathbf{x}} - \left(\sqrt{(L+\lambda)P_{xx}}\right)_l, \quad l = 1, \ldots, L, \end{aligned} \tag{C.1}$$

where $(\sqrt{(L+\lambda)P_{xx}})_l$ is the lth row or column of the corresponding matrix square root and $\lambda = \alpha^2(L + \kappa) - L$. α determines the spread of the sigma points. $\alpha = 1$ was used throughout our simulations. κ is a secondary scaling parameter. The choice $\kappa = 3 - L$ maintains the kurtosis of a Gaussian vector. Throughout our simulations, κ is set to 0.

Define the weights

$$\begin{aligned} W_0^{(m)} &= \lambda/(L+\lambda), \\ W_0^{(c)} &= \lambda/(L+\lambda) + (1 - \alpha^2 + \beta), \\ W_l^{(m)} &= W_l^{(c)} = 1/2(L+\lambda), \quad l = 1, 2, \ldots, 2L, \end{aligned} \tag{C.2}$$

where β is used to incorporate prior knowledge of the distribution ($\beta = 2$ for Gaussian distributions). A proper choice of these parameters and its influence on the obtainable performance is still an open topic. Then the mean and covariance of the vector $\mathbf{y}$ can be calculated using the following procedure.

(1) Construct $\mathbf{x}$ σ-points: $\mathcal{X}_l$, $l = 0, \ldots, 2L$.
(2) Transform each point to the respective $\mathbf{y}$ σ-points: $\mathcal{Y}_l = f(\mathcal{X}_l)$, $l = 0, \ldots, 2L$.
(3) Use weighted averaging $\bar{\mathbf{y}} \approx \sum_{l=0}^{2L} W_l^{(m)} \mathcal{Y}_l$ to estimate $\mathbf{y}$ mean.
(4) Use weighted outer product $P_{yy} \approx \sum_{l=0}^{2L} W_l^{(c)} (\mathcal{Y}_l - \bar{\mathbf{y}})(\mathcal{Y}_l - \bar{\mathbf{y}})^T$ to estimate $\mathbf{y}$ covariance and $P_{xy} \approx \sum_{l=0}^{2L} W_l^{(c)} (\mathcal{X}_l - \bar{\mathbf{x}})(\mathcal{Y}_l - \bar{\mathbf{y}})^T$ to estimate the cross-covariance between $\mathbf{x}$ and $\mathbf{y}$.

The benefits of using the UT are presented in [29, 31].

REFERENCES

[1] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 1, pp. 375–378, Munich, Germany, April 1997.

[3] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Processing*, vol. 59, no. 3, pp. 253–266, 1997.

[4] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *The Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 384–391, 2000.

[5] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[6] T. Dvorkind and S. Gannot, "Speaker localization in a reverberant environment," in *Proceedings of the 22nd IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI '02)*, pp. 7–9, Tel-Aviv, Israel, December 2002.

[7] T. G. Dvorkind and S. Gannot, "Approaches for time difference of arrival estimation in a noisy and reverberant environment," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 215–218, Kyoto, Japan, September 2003.

[8] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2004.

[9] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing*, vol. 42, no. 8, pp. 1905–1915, 1994.

[10] M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A closed-form location estimator for use with room environment microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 45–50, 1997.

[11] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 8, pp. 1223–1225, 1987.

[12] J. Smith and J. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 12, pp. 1661–1669, 1987.

[13] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, pp. 909–912, Istanbul, Turkey, June 2000.

[14] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least squares approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1988.

[16] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood acoustic source localization: experimental results," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 3, pp. 2949–2952, Orlando, Fla, USA, May 2002.

[17] J. C. Chen, R. E. Hudson, and K. Yao, "Maximum-likelihood source localization and unknown sensor location estimation for wideband signals in the near-field ," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 1843–1854, 2002.

[18] M. Segal, E. Weinstein, and B. R. Musicus, "Estimate-maximize algorithms for multichannel time delay and signal estimation," *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 1–16, 1991.

[19] S. T. Birchfield and D. K. Gillmor, "Fast Bayesian acoustic localization," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 2, pp. 1793–1796, Orlando, Fla, USA, May 2002.

[20] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment ," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 826–836, 2003.

[21] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 5, pp. 3021–3024, Salt Lake City, Utah, USA, May 2001.

[22] E. A. Lehmann and R. C. Williamson, "Importance sampling particle filter for robust acoustic source localization and tracking in reverberant environments," in *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA '05)*, vol. C, pp. 17–18, Piscataway, NJ, USA, March 2005.

[23] D. Bechler, M. Grimm, and K. Kroschel, "Speaker tracking with a microphone array using Kalman filtering," *Advances in Radio Science*, vol. 1, pp. 113–117, 2003.

[24] U. Klee and J. McDonough, "Kalman filtering for acoustic source localization based on time delay of arrival," in *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA '05)*, vol. C, pp. 5–6, Piscataway, NJ, USA, March 2005.

[25] T. G. Dvorkind and S. Gannot, "Speaker localization exploiting spatial-temporal information," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 295–298, Kyoto, Japan, September 2003.

[26] T. G. Dvorkind and S. Gannot, "Speaker localization using the unscented Kalman filter," in *Proceedings of the Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA'05)*, vol. C, pp. 3–4, Piscataway, NJ, USA, March 2005.

[27] S. Haykin, *Adaptive Filter Theory*, Information and System Sciences, Prentice Hall, Upper Saddle River, NJ, USA, 4th edition, 2002.

[28] D. C. Popescu and C. Rose, "Emitter localization in a multipath environment using extended Kalman filter," in *Proceedings of the 33rd Conference on Information Sciences and Systems (CISS '99)*, vol. 1, pp. 147–150, Baltimore, Md, USA, March 1999.

[29] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[30] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[31] E. A. Wan and R. van der Merwe, "The unscented Kalman filter for nonlinear estimation," in *Proceedings of IEEE Symposium on Adaptive Systems for Signal Processing, Communication and Control (AS-SPCC '00)*, pp. 153–158, Lake Louise, Alberta, Canada, October 2000.

Sharon Gannot received his B.S. degree (summa cum laude) from the Technion – Israeli Institute of Technology, Israel, in 1986 and the M.S. (cum laude) and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering. Between 1986 and 1993 he was the Head of a research and development section in R&D Center of the Israeli Defense Forces. In the year 2001 he held a postdoctoral position at the Department of Electrical Engineering (SISTA) at K.U. Leuven, Belgium. Between 2002 and 2003 he held a research and teaching position at the Signal and Image Processing Laboratory (SIPL), Faculty of Electrical Engineering, Technion – Israeli Institute of Technology, Israel. Currently, he is a Lecturer in the School of Engineering, Bar-Ilan University, Israel. He is also an Associate Editor of the EURASIP Journal on Applied Signal Processing, a Guest Editor in the Speech Communication Journal, and a Reviewer of many IEEE journals. His research interests include parameter estimation, statistical signal processing, and speech processing using either single- or multimicrophone arrays.

Tsvi Gregory Dvorkind received his B.S. degree in computer engineering in 2000 and the M.S. degree in electrical engineering in 2003, both summa cum laude and both from the Technion – Israeli Institute of Technology, Israel. He is now at the Technion, pursuing his Ph.D. degree in electrical engineering. From 1998 to 2000 he worked at the Electro-Optics Research and Development Company at the Technion, and during 2000–2001 at the Jigami Corporation. Starting from 2001 he is a Research Assistant and a Project Supervisor at the Signal and Image Processing Laboratory (SIPL), Faculty of Electrical Engineering, Technion. His research interests include speech enhancement and acoustical localization, general parameter estimation problems, and sampling theory.