SPEAKER LOCALIZATION EXPLOITING SPATIAL-TEMPORAL INFORMATION

Tsvi Gregory Dvorkind and *Sharon Gannot*
Faculty of Electrical Engineering, Technion, Technion City, 32000 Haifa, Israel
e-mail: dvorkind@tx.technion.ac.il, gannot@siglab.technion.ac.il

ABSTRACT

Determining the spatial position of a speaker finds a growing interest in video conference scenario where automated camera steering and tracking are required. Speaker localization can be achieved with a dual step approach. In the preliminary stage microphone array is used to extract the *time difference of arrival* (TDOA) of the speech signal. These readings are then used by the second stage for the actual localization. Since speaker trajectory must be smooth, estimates of close speaker positions might be used to improve the current position estimate. However, many methods, although exploiting the spatial information obtained by different microphone pairs, do not exploit this temporal information. In this contribution we present two localization schemes, which exploit the temporal information. The first is the well known *extended Kalman filter* (EKF). The second is a recursive form of a Gauss method, which we denote *Recursive Gauss* (RG). Experimental study supports the potential of the proposed methods.

1 Introduction and Problem Formulation

In this work we address approaches for determining speaker position which are comprised of two stages. In the first stage, the TDOA is estimated using spatially separated microphone pairs (e.g. [1],[2] and [3]). In the second stage, these readings are used for the actual localization (e.g. [4], [5] and [6]). These methods exploit the spatial information obtained by different microphone pairs, but do not exploit the temporal information available from adjoint speaker position estimates. This information is relevant for the current position estimate, due to the speaker smooth trajectory.

Consider an array of $M+1$ microphones, placed at the Cartesian coordinates $\underline{m}_i \triangleq [x_i, y_i, z_i]^T$; $i = 0, \ldots, M$ where $\underline{m}_0 = [0,0,0]^T$ is the reference microphone, placed at the axes origin and $(\cdot)^T$ stands for the transpose operation. Define the source coordinate at time instance t by $\underline{s}(t) \triangleq [x_s(t), y_s(t), z_s(t)]^T$. Each of the M microphones, combined with the reference microphone, is used at time instance t to produce a TDOA measurement $\tau_i(t)$; $i = 1 \ldots M$. Denote the i-th **range difference** measurement by $r_i(t) = c\tau_i(t)$, where c is the sound propagation speed (approximately 340[m/s] in air). The non-linear equations for estimating the source location parameters $\underline{s}(t)$ are (see for example [6]):

$$\mathbf{A}(t)\underline{f}(\underline{s}(t)) \approx \underline{d}(t) \tag{1}$$

where $\underline{f}^T(\underline{s}(t)) \triangleq \left[\underline{s}^T(t), \|\underline{s}(t)\|\right]$ and

$$\mathbf{A}(t) \triangleq \begin{bmatrix} \underline{m}_1^T, r_1(t) \\ \vdots \\ \underline{m}_M^T, r_M(t) \end{bmatrix}, \quad \underline{d}(t) \triangleq \frac{1}{2}\begin{bmatrix} \|\underline{m}_1\|^2 - r_1^2(t) \\ \vdots \\ \|\underline{m}_M\|^2 - r_M^2(t) \end{bmatrix}.$$

Note, that only approximate equality holds in (1), since the range difference measurements are noisy.

The organization of the rest of the paper is as follows. In Section 2 we derive Gauss and recursive Gauss (RG) solutions for the localization problem. A Baysian approach, namely the extended Kalman filter (EKF), is presented in Section 3. The equivalence of RG and EKF approaches is discussed in Section 4. A test case is presented in Section 5.

2 Gauss and Recursive Gauss Algorithms

Huang *et al.* [6] addressed the non-linear set in (1) and solved it by using Lagrange multiplier. Since a polynomial of degree six is involved in the proposed method (denoted *linear-correction least-squares* (LCLS)), no closed-form solution exists. Thus, the iterative secant method is used for the root search.

2.1 Gauss Solution

The solution of (1) presented by [6] involves iterations. We suggest to mitigate the non-linearity by an alternative method, i.e the Gauss method. Note, that Eq. (1) becomes a (non-linear) *Least Squares* (LS) problem if the number of microphone pairs fulfills $M > 3$, i.e there are more equations than unknowns. The resulting non-linear LS problem can be solved by applying the Gauss method. Define $\underline{s}^{(l)}(t)$, the estimate of $\underline{s}(t)$ at the l-th iteration. Define also $\underline{h}_t(\underline{s}^{(l)}(t)) \triangleq \mathbf{A}(t)\underline{f}(\underline{s}^{(l)}(t))$, and the associated gradient matrix by $\mathbf{H}_t(\underline{s}^{(l)}(t)) = \nabla_{\underline{s}}\underline{h}_t(\underline{s}^{(l)}(t))$. By applying first-order approximation to $\underline{h}_t(\underline{s}(t))$, the Gauss iterations take the well known form

$$\underline{s}^{(l+1)}(t) = \underline{s}^{(l)}(t) + \left(\mathbf{H}_t(\underline{s}^{(l)}(t))^\dagger \mathbf{H}_t(\underline{s}^{(l)}(t))\right)^{-1} \mathbf{H}_t(\underline{s}^{(l)}(t))^\dagger \left(\underline{d}(t) - \underline{h}_t(\underline{s}^{(l)}(t))\right).$$

This solution exploits only the spatial information obtained by the separated microphone pairs at a specific time instance, but does not consider the temporal information.

2.2 Recursive Gauss (RG) Procedure

To obtain a recursive Gauss procedure we begin by evaluating Eq. (1) at all measurements from $t = 1$ to $t = N$:

$$\mathbf{A}(t=1)\underline{f}(\underline{s}(1)) \approx \underline{d}(1), \ \ldots, \ \mathbf{A}(t=N)\underline{f}(\underline{s}(N)) \approx \underline{d}(N). \tag{2}$$

Note that this is still a nonlinear equation set in the unknown positions $\underline{s}(t)$; $t = 1, \ldots, N$, due to the nonlinearity introduced by $\underline{f}$. By assuming that $\underline{s}(t)$ is slowly varying with time, a recursive solution can be derived.

The proposed method, which is also presented in [2] and for completeness reviewed here, starts by resolving the nonlinearities using first-order approximation (as with the original Gauss method), and then continues by deriving a recursion (applying further approximation). This solution will be referred to as *Recursive Gauss* (RG).

Consider a nonlinear equation set for the unknown $p \times 1$ parameter vector $\underline{\theta} \in \mathcal{C}^p$:

$$\underline{h}_{1:N}(\underline{\theta}) = \underline{d}_{1:N}$$

where $\underline{h}_{1:N}^T(\underline{\theta}) \triangleq \left[\underline{h}_1^T(\underline{\theta}) \cdots \underline{h}_N^T(\underline{\theta})\right]$ and $\underline{d}_{1:N}^T \triangleq \left[\underline{d}_1^T \cdots \underline{d}_N^T\right]$. $\underline{h}_t$ and $\underline{d}_t$ are K nonlinear equations and K measurements, available at time instance t, respectively. Applying first-order approximation around an initial guess $\underline{\theta}^{(0)}$ (as with the Gauss method) we obtain:

$$\underline{h}_{1:N}(\underline{\theta}^{(0)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(0)}) \left(\underline{\theta} - \underline{\theta}^{(0)}\right) \approx \underline{d}_{1:N} \tag{3}$$

where $\mathbf{H}_{1:N}$ is the $NK \times p$ gradient matrix:

$$\mathbf{H}_{1:N}^T(\underline{\theta}) \triangleq \left[\mathbf{H}_1^T(\underline{\theta}) \cdots \mathbf{H}_N^T(\underline{\theta})\right]$$

where $\mathbf{H}_t(\underline{\theta}) = \nabla_{\underline{\theta}} \underline{h}_t(\underline{\theta})$ is the $K \times p$ gradient matrix of $\underline{h}_t(\underline{\theta})$. According to the Gauss method, the iterative LS solution to the linearized set (3) is:

$$\underline{\theta}^{(l+1)} = \arg\min_{\underline{\theta}} \left\| \underline{d}_{1:N} - \left(\underline{h}_{1:N}(\underline{\theta}^{(l)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(l)}) \left(\underline{\theta} - \underline{\theta}^{(l)}\right)\right) \right\|$$

where the superscript denotes the iteration number. Consider the next measurements $\underline{h}_{N+1}(\underline{\theta}) = \underline{d}_{N+1}$ available at time instance $N+1$. In order to estimate $\underline{\theta}$ we will use all the available measurements simultaneously. Though we could evaluate all $(N+1)K$ equations at the current estimate $\underline{\theta}^{(l+1)}$, we will do so **only** for the new equations. Namely, instead of minimizing in the LS sense the following residual norm

$$\min_{\underline{\theta}} \left\| \underline{d}_{1:N+1} - \left(\underline{h}_{1:N+1}(\underline{\theta}^{(l+1)}) + \mathbf{H}_{1:N+1}(\underline{\theta}^{(l+1)}) \left(\underline{\theta} - \underline{\theta}^{(l+1)}\right)\right) \right\|$$

we will minimize a modified LS problem

$$\min_{\underline{\theta}} \left\| \begin{array}{l} \underline{d}_{1:N} - \\ \underline{d}_{N+1} - \\ \left(\underline{h}_{1:N}(\underline{\theta}^{(l)}) + \mathbf{H}_{1:N}(\underline{\theta}^{(l)}) \left(\underline{\theta} - \underline{\theta}^{(l)}\right)\right) \\ \left(\underline{h}_{N+1}(\underline{\theta}^{(l+1)}) + \mathbf{H}_{N+1}(\underline{\theta}^{(l+1)}) \left(\underline{\theta} - \underline{\theta}^{(l+1)}\right)\right) \end{array} \right\|.$$

The reason for this approximation is to keep past solutions intact, thus enabling a recursive solution to be derived. Now, using *stochastic approximation*, i.e. replacing the iteration index by the time index, a sequential algorithm is obtained. To summarize the procedure, an estimate for $\underline{\theta}$ at the current time instance t (denoted by $\hat{\underline{\theta}}(t)$) is obtained by solving the following LS problem sequentially using the *recursive LS* (RLS) procedure:

$$\hat{\underline{\theta}}(t) = \arg\min_{\underline{\theta}} \left\| \begin{bmatrix} \mathbf{H}_1(\hat{\underline{\theta}}(0)) \\ \vdots \\ \mathbf{H}_t(\hat{\underline{\theta}}(t-1)) \end{bmatrix} \underline{\theta} - \underline{y}_{1:t} \right\| \tag{4}$$

where

$$\underline{y}_{1:t} = \begin{bmatrix} \underline{y}_1 \\ \vdots \\ \underline{y}_t \end{bmatrix} \triangleq \begin{bmatrix} \underline{d}_1 - \underline{h}_1(\hat{\underline{\theta}}(0)) + \mathbf{H}_1(\hat{\underline{\theta}}(0))\hat{\underline{\theta}}(0) \\ \vdots \\ \underline{d}_t - \underline{h}_t(\hat{\underline{\theta}}(t-1)) + \mathbf{H}_t(\hat{\underline{\theta}}(t-1))\hat{\underline{\theta}}(t-1) \end{bmatrix}$$

with $\hat{\underline{\theta}}(0)$ the initial estimate for the parameter set. We note that in many practical situations the parameter set $\underline{\theta}$ might slowly vary with time. In these cases a common practise is to apply the RLS algorithm with a diagonal weight matrix that uses a forgetting factor $0 < \alpha \leq 1$ to weight past equations.

Another practical issue concerns the computational burden. At each time instance new K equations become available, resulting a $K \times K$ matrix inversion at each RLS iteration. However, by properly varying the forgetting factor α the computational complexity can be further reduced. This procedure is beyond the scope of this paper.

2.3 Recursive Gauss (RG) Application

Denote the parameter set by $\underline{\theta} = \underline{s}$. The RG procedure takes the following form. Let,

$$\mathbf{H}_t(\underline{s}) = \begin{bmatrix} \underline{m}_1^T + \frac{r_1(t)}{\|\underline{s}\|}\underline{s}^T \\ \vdots \\ \underline{m}_M^T + \frac{r_M(t)}{\|\underline{s}\|}\underline{s}^T \end{bmatrix}, \; \underline{y}_t = \underline{d}_t = \frac{1}{2} \begin{bmatrix} \|\underline{m}_1\|^2 - r_1^2(t) \\ \vdots \\ \|\underline{m}_M\|^2 - r_M^2(t) \end{bmatrix}.$$

Then $\hat{\underline{s}}(t)$ is evaluated by solving (4) with RLS and a forgetting factor $\alpha < 1$.

3 Extended Kalman Filter (EKF)

The non-linear set in Eq. (1) can be also solved in the Bayesian framework. The optimal *minimum mean square error* (MMSE) solution becomes complicated in this nonlinear case, and sub-optimal solutions are called upon. Such a solution is the *extended Kalman filter* (EKF). As the actual movement model is not known in advance, we use a *random walk* model instead

$$\begin{cases} \underline{s}(t+1) &= \underline{s}(t) + \underline{w}(t) \\ \underline{r}(t) &= \underline{h}(\underline{s}(t)) + \underline{v}(t) \end{cases} \tag{5}$$

where $\underline{w}(t)$ is the state driving noise and $\underline{v}(t)$ is the measurement noise. $\underline{h}$ represents the nonlinear range difference measurement equations, given by:

$$\underline{h}(\underline{s}) \triangleq \begin{bmatrix} \|\underline{m}_1 - \underline{s}\| - \|\underline{s}\| \\ \vdots \\ \|\underline{m}_M - \underline{s}\| - \|\underline{s}\| \end{bmatrix}. \tag{6}$$

We note that the same approach was taken in [7], but in a different context.

4 Equivalence of RG and EKF

It is well known that the RLS algorithm can be viewed as a special case of the Kalman filter. We show now that the same equivalence holds for the recursive Gauss algorithm, derived in Section 2.2, and the extended Kalman filter. Using a diagonal weight matrix and setting the forgetting factor to α, the RG algorithm coincide with the EKF formulation for the following state-space model,

$$\begin{cases} \underline{\theta}(t+1) &= \underline{\theta}(t) \\ \underline{d}_t &= \underline{h}_t(\underline{\theta}(t)) + \underline{v}(t) \end{cases}$$

The equivalence holds when $\mathbf{R}(t) \triangleq \mathrm{Cov}(\underline{v}(t)) = \alpha \mathbf{I}$ (where $\mathbf{I}$ stands for the identity matrix) and with the initial condition

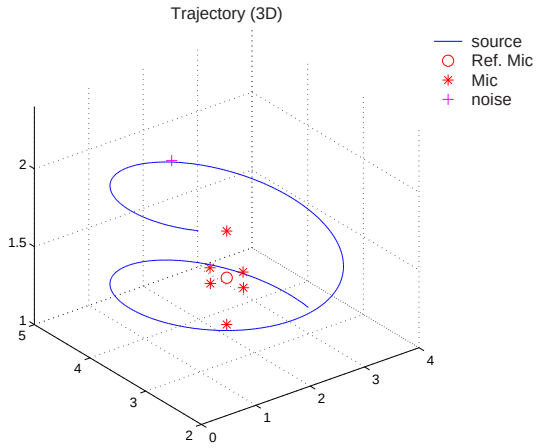

Figure 1: Speaker trajectory

$\mathbf{P}_{0|-1} \triangleq \mathrm{Cov}(\underline{\theta}(0)) = \frac{1}{\alpha}\mathbf{P}(0)$. Moreover, this formulation is exactly the same as the one relating the RLS algorithm and (linear) Kalman filter.

5 Experimental Study

In this section we perform simulative comparison of several localization methods. To gain some insight on the obtainable performance of a microphone array with a small inter-element spread relative to the source position, calculation of the CRLB for a specific scenario is performed. This calculation leads us to a conclusion that the meaningful information lies in the azimuth and elevation angles estimates. We proceed by assessing four localization methods. Two of them are non-temporal (LCLS and Gauss iterations), and the other two (RG and EKF) exploit the temporal information.

5.1 Test Scenario

A set of 6 microphone pairs is placed on a sphere of radius $0.3[m]$ around a reference microphone placed at the origin, $\underline{m}_0 = [0,0,0]^T$, at the following positions:

$$\underline{m}_1^T = [0.3,\ 0,\ 0]\,,\quad \underline{m}_2^T = [-0.3,\ 0,\ 0]\,,\ \underline{m}_3^T = [0,\ 0.3,\ 0]$$
$$\underline{m}_4^T = [0,\ -0.3,\ 0]\,,\ \underline{m}_5^T = [0,\ 0,\ 0.3]\,,\quad \underline{m}_6^T = [0,\ 0,\ -0.3]\,.$$

The speaker trajectory is set to be an helix with radius $R = 1.5$[m] around the reference microphone. The speaker movement speed is set to 0.5[m/s] and the total duration of the movement is $T = 30$[sec]. In Cartesian coordinates, the position of the speaker (in $\underline{m}_0$ coordinate system) as a function of time in the interval $t \in [0,T]$ is given by:

$$x(t) = R\cos(2\pi ft),\ y(t) = R\sin(2\pi ft),\ z(t) = \tfrac{t}{T} - 0.375$$

with $f = 0.0529[Hz]$. Along this trajectory, the overall change of the azimuth angle is within $\theta \in [0^\circ, 570^\circ]$ and of the elevation angle is within $\gamma \in [-14^\circ, 22.5^\circ]$. The entire scenario is depicted in Fig. 1.

5.2 The Cramér-Rao Lower Bound

We calculate now the CRLB for the tested scenario. We assume that the true range difference (or, equivalently, the TDOA) readings are contaminated by Gaussian distributed noise with zero-mean and *standard deviation* (STD) of $\sigma = 0.0425$[m]. This STD is equivalent to 1[sample] at a sample rate of $F_s = 8000$[Hz]. 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 multi-modal noise distribution. Nevertheless, for simplicity, we assume (like Huang *et al.* [6]) that the noise is uni-modal (Gaussian) distributed and temporally white. In this setup we further assume that the noise is spatially white. Under these conditions, the CRLB is calculated for both Cartesian and polar coordinates. The resulting bound (in meters, for the Cartesian coordinates and the radius, and in degrees for the azimuth and elevation angles) is depicted in Fig. 2. Note, that the Cartesian coordinates, as well as the radius, can not be accurately estimated in this scenario. This conclusion corresponds with the results presented in [6]. However the azimuth and elevation angles might 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 non-temporal methods alone, since past measurements are disregarded at each time instance.

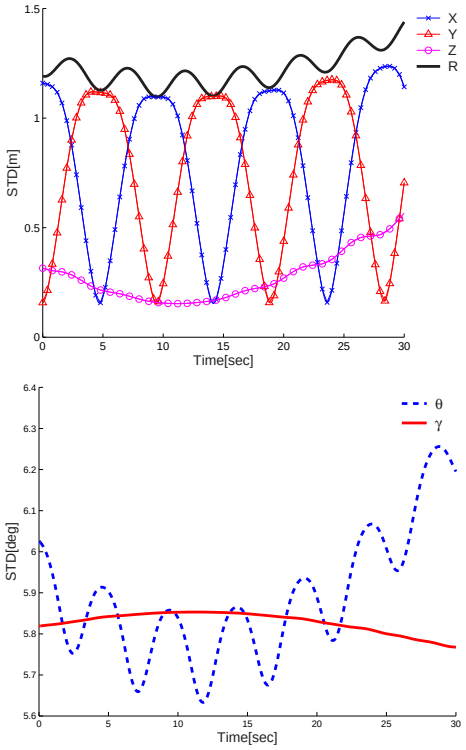

Figure 2: CRLB results. Top: Cartesian coordinates and radius. Bottom: Azimuth (θ) and elevation (γ) angles.

5.3 Simulation Results

The previously presented setup is evaluated by four localization methods. The first is *Linear Correction Least Squares* (LCLS), presented by Huang *et al.* [6]. The second is the Gauss method (denoted G) with 3 iterations at each time instance. The third is the recursive Gauss (RG) with forgetting factor $\alpha = 0.85$. The fourth is the EKF method evaluated with random-walk model and driving noise STD of 0.1[m] at each axis. The measurements covariance matrix is over-estimated to $10\sigma^2\mathbf{I}$. 1000 Monte-Carlo trials are performed. The *Root Mean Square Error* (RMSE) of the angles estimate is presented in Fig. 3. As can be seen, the Gauss iterations and the LCLS method have comparable performance. How-

ever the RG and the EKF methods remarkably outperform them. We proceed by testing a more realistic scenario, where

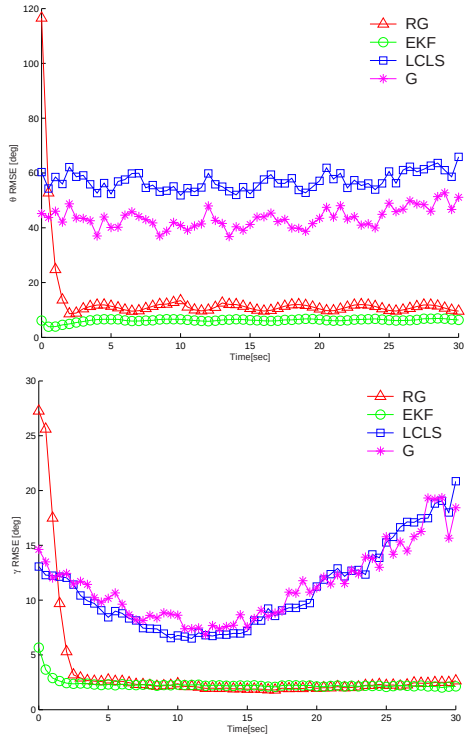

Figure 3: Uni-modal white Gaussian noise. Top: RMSE for azimuth angle (θ). Bottom: RMSE for elevation angle (γ)

the measurement noise is not uni-modal and highly spatially correlated. This is the case of directional interference which might occur in a video conference scenario. This setup is simulated by a Gaussian mixture distribution. The mixing probabilities were set to 0.9 for the correct mode (speaker) and 0.1 for the directional noise mode. For this purpose a directional noise source is placed at the polar coordinates $[\theta = \frac{3\pi}{4}, \gamma = \frac{\pi}{4}, R = 1]$, as depicted in Fig. 1. Furthermore, while in the noise mode, the measurement noise is set to be spatially correlated with correlation matrix $\mathbf{C} \triangleq \mathrm{Cov}\{\underline{v}_t\}$ where the i,j-th ($i \neq j$) element of $\mathbf{C}$ is $0.9\sigma^2$ and σ^2 along the diagonal. As can be seen from Fig. 4, the performance of all methods degrades. However, the EKF method clearly outperforms the other methods. This is despite of the fact that the EKF is not using the new noise model. Note that the use of the random walk model in the EKF formulation explains the divergence from the RG method and is more appropriate for the tracking problem.

6 Conclusions

We presented both non-temporal and temporal algorithms for talker localization and tracking. The Gauss method was shown to have comparable performance to the LCLS method. Two temporal methods were derived. One is within a non-Bayesian framework (RG algorithm) and the other is within the Baysian framework (EKF). The RG method is shown to be a degenerate case of the EKF. The empirical results demonstrate the effectiveness of the use of the temporal information.

7 *

References

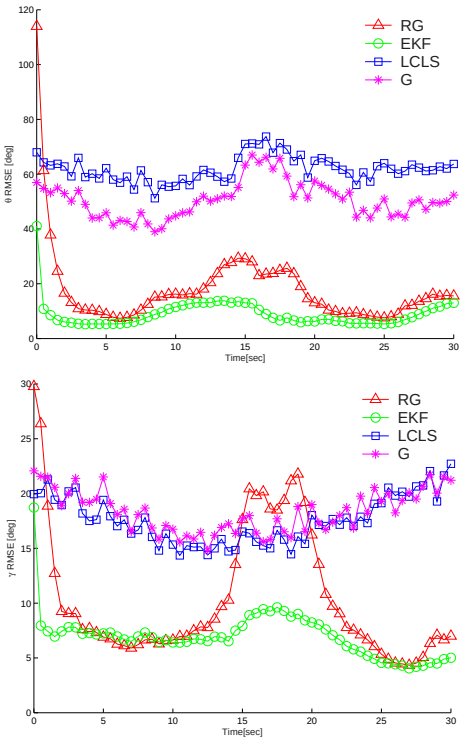

Figure 4: Bimodal noise. Top: RMSE for azimuth angle (θ). Bottom: RMSE for elevation angle (γ)

[1] T. Dvorkind and S. Gannot, "Speaker Localization in a Reverberant Environment," *IEEE proceedings, The 22nd convention of Electrical and Electronics Engineers in Israel.*, pp. 7–9, Dec. 2002.

[2] T. Dvorkind, S. Gannot, "Approaches for Time Difference of Arrival Estimation in a Noisy and Reverberant Environment," in *IWAENC Kyoto, Japan*, Sep 2003.

[3] C.H. Knapp and G.C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[4] J.O. Smith and J. S. 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, Dec. 1987.

[5] 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, Jan. 1997.

[6] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "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, Jan. 2001.

[7] D.C. Popescu and C. Rose, "Emitter Localization in a Multipath Environment Using Extended Kalman Filter," in *33rd Conf. on Information Sciences and Systems - CISS*, The Johns Hopkins University, Baltimore, Maryland, USA, March 1999, vol. 1, pp. 147–150.