Kalman Filters for Time Delay of Arrival-Based Source Localization

Ulrich Klee, Tobias Gehrig, John McDonough

Institut für Logik, Komplexität, und Deduktionssysteme
Universität Karlsruhe
Am Fasanengarten 5
D-76131 Karlsruhe, Germany
{klee,tgehrig,jmcd}@ira.uka.de

Abstract

In this work, we propose an algorithm for acoustic source localization based on time delay of arrival (TDOA) estimation. In earlier work by other authors, an initial closed-form approximation was first used to estimate the true position of the speaker followed by a Kalman filtering stage to smooth the time series of estimates. In the proposed algorithm, this closed-form approximation is eliminated by employing a Kalman filter to directly update the speaker position estimate based on the observed TDOAs. In particular, the TDOAs comprise the observation associated with an extended Kalman filter whose state corresponds to the speaker position. We tested our algorithm on a data set consisting of seminars held by actual speakers. Our experiments revealed that the proposed algorithm provides source localization accuracy superior to the standard spherical and linear intersection techniques. Moreover, the proposed algorithm, although relying on an iterative optimization scheme, proved efficient enough for real-time operation.

1. Introduction

Most practical acoustic source localization schemes are based on *time delay of arrival estimation* (TDOA) for the following reasons: Such systems are conceptually simple. They are reasonably effective in moderately reverberant environments. Moreover, their low computational complexity makes them well-suited to real-time implementation with several sensors.

Time delay of arrival-based source localization is based on a two-step procedure:

1. The TDOA between all pairs of microphones is estimated, typically by finding the peak in a cross correlation or *generalized cross correlation* function [1].
2. For a given source location, the squared-error is calculated between the estimated TDOAs and those determined from the source location. The estimated source location then corresponds to that position which minimizes this squared error.

If the TDOA estimates are assumed to have a Gaussian-distributed error term, it can be shown that the least squares metric used in Step 2 provides the maximum likelihood (ML) estimate of the speaker location [2]. Unfortunately this least squares criterion results in a nonlinear optimization problem that can have several local minima. Several authors have proposed solving this optimization problem with standard gradient-based iterative techniques. While such techniques typically yield accurate location estimates, they are typically computationally intensive and thus ill-suited for real-time implementation [3, 4].

For any pair of microphones, the surface on which the TDOA is constant is a hyperboloid of two sheets. A second class of algorithms seeks to exploit this fact by grouping all microphones into pairs, estimating the TDOA of each pair, then finding the point where all associated hyperboloids most nearly intersect. Several closed-form position estimates based on this approach have appeared in the literature; see Chan and Ho [5] and the literature review found there. Unfortunately, the point of intersection of two hyperboloids can change significantly based on a slight change in the eccentricity of one of the hyperboloids. Hence, a third class of algorithms was developed wherein the position estimate is obtained from the intersection of several spheres. The first algorithm in this class was proposed by Schau and Robinson [6], and later came to be known as *spherical intersection*. Perhaps the best known algorithm from this class is the *spherical interpolation* method of Smith and Abel [7]. Both methods provide closed-form estimates suitable for real-time implementation.

Brandstein *et al* [4] proposed yet another closed-form approximation known as *linear intersection*. Their algorithm proceeds by first calculating a bearing line to the source for each pair of sensors. Thereafter, the point of nearest approach is calculated for each pair of bearing lines, yielding a potential source location. The final position estimate is obtained from a weighted average of these potential source locations.

In the algorithm proposed here, the closed-form approximations used in prior approaches is eliminated by employing an extended Kalman filter to directly update the speaker position estimate based on the observed TDOAs. In particular, the TDOAs comprise the observation associated with an extended Kalman filter whose state corresponds to the speaker position. Hence, the new position estimate comes directly from the update formulae associated with the Kalman filter. Similar approaches have been proposed in the past by Dvorkin and Gannot [8] for acoustic source localization, and by Duraiswami *et al* for a combined audio-video localization system based on a particle filter [9].

The balance of this work is organized as follows. In Section 2, we review the process of source localization based on time-delay of arrival estimation. In particular, we formulate source localization as a problem in nonlinear least squares estimation, then develop an appropriate linearized model. Section 3 summarizes a less well-known variant of the Kalman filter, known as the iterated extended Kalman filter. Section 4 presents a simple model for speaker motion, then discusses how the development in the preceding sections can be combined to develop an acoustic localization algorithm capable of tracking a moving speaker. Section 5 presents the results of our initial experiments comparing the proposed algorithm to the standard techniques.

2. Source Localization

Consider the i-th pair of microphones, and let $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$ respectively be the positions of the first and second microphones in the pair. Let $\mathbf{x}$ denote the position of the speaker in $\mathbf{R}^3$. Then the *time delay of arrival* (TDOA) between the two microphones

of the pair can be expressed as

$$T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_{i1}\| - \|\mathbf{x} - \mathbf{m}_{i2}\|}{s} \quad (1)$$

where s is the speed of sound. Denoting

$$\mathbf{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \qquad \mathbf{m}_{ij} = \begin{bmatrix} m_{ij,x} \\ m_{ij,y} \\ m_{ij,z} \end{bmatrix}$$

allows (1) to be rewritten as

$$T_i(\mathbf{x}) = T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{1}{s}(d_{i1} - d_{i2}) \quad (2)$$

where

$$\begin{aligned} d_{ij} &= \sqrt{(x - m_{ij,x})^2 + (y - m_{ij,y})^2 + (z - m_{ij,z})^2} \\ &= \|\mathbf{x} - \mathbf{m}_{ij}\| \end{aligned} \quad (3)$$

is the distance from the source to microphone $\mathbf{m}_{ij}$. Equation (2) is clearly nonlinear in $\mathbf{x} = (x, y, z)$. In the coming development, we will find it useful to have a linear approximation. Hence, we can take a partial derivative with respect to x on both sides of (2) and write

$$\frac{\partial T_i(\mathbf{x})}{\partial x} = \frac{1}{s} \cdot \left[\frac{x - m_{i1,x}}{d_{i1}} - \frac{x - m_{i2,x}}{d_{i2}} \right]$$

Taking partial derivatives with respect to y and z similarly, we find

$$\nabla_{\mathbf{x}} T_i(\mathbf{x}) = \frac{1}{s} \cdot \left[\frac{\mathbf{x} - \mathbf{m}_{i1}}{d_{i1}} - \frac{\mathbf{x} - \mathbf{m}_{i2}}{d_{i2}} \right]$$

We can approximate $T_i(\mathbf{x})$ with a first order Taylor series expansion about the last position estimate $\hat{\mathbf{x}}(t-1)$ as

$$\begin{aligned} T_i(\mathbf{x}) &\approx T_i(\hat{\mathbf{x}}(t-1)) + \nabla_{\mathbf{x}} T_i(\mathbf{x})(\mathbf{x} - \hat{\mathbf{x}}(t-1)) \\ &= T_i(\hat{\mathbf{x}}(t-1)) + \mathbf{c}_i(t)(\mathbf{x} - \hat{\mathbf{x}}(t-1)) \end{aligned} \quad (4)$$

where we have defined the row vector

$$\mathbf{c}_i(t) = [\nabla_{\mathbf{x}} T_i(\mathbf{x})]^T = \frac{1}{s} \cdot \left[\frac{\mathbf{x} - \mathbf{m}_{i1}}{d_{i1}} - \frac{\mathbf{x} - \mathbf{m}_{i2}}{d_{i2}} \right]^T \quad (5)$$

Equation (4) is the desired linearization.

Source localization based on a maximum likelihood (ML) criterion [2] proceeds by minimizing the error function

$$\epsilon(\mathbf{x}) = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[\hat{\tau}_i - T_i(\mathbf{x}) \right]^2 \quad (6)$$

where $\hat{\tau}_i$ is the observed TDOA for the i-th microphone pair and σ_i^2 is the error covariance associated with this observation. The TDOAs can be estimated with a variety of well-known techniques [1, 10]. Perhaps the most popular method involves the generalized cross correlation (GCC), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau}) X_2^*(e^{j\omega\tau})|} \, e^{j\omega\tau} \, d\omega \quad (7)$$

For reasons of computational efficiency, $R_{12}(\tau)$ is typically calculated with an inverse FFT. Thereafter, an interpolation is performed to overcome the granularity in the estimate corresponding to the sampling interval [1].

Substituting the linearization (4) into (6) and introducing a time dependence provides

$$\epsilon(\mathbf{x}; t) \approx \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[\bar{\tau}_i(t) - \mathbf{c}_i(t)\mathbf{x} \right]^2 \quad (8)$$

where

$$\bar{\tau}_i(t) = \hat{\tau}_i(t) - T_i(\mathbf{x}(t-1)) + \mathbf{c}_i(t)\hat{\mathbf{x}}(t-1) \quad (9)$$

for $i = 0, \ldots, N-1$. Let us define

$$\bar{\boldsymbol{\tau}}(t) = \begin{bmatrix} \bar{\tau}_0(t) \\ \bar{\tau}_1(t) \\ \vdots \\ \bar{\tau}_{N-1}(t) \end{bmatrix} \qquad \hat{\boldsymbol{\tau}}(t) = \begin{bmatrix} \hat{\tau}_0(t) \\ \hat{\tau}_1(t) \\ \vdots \\ \hat{\tau}_{N-1}(t) \end{bmatrix}$$

and

$$\mathbf{T}(\hat{\mathbf{x}}(t)) = \begin{bmatrix} T_0(\hat{\mathbf{x}}(t)) \\ T_1(\hat{\mathbf{x}}(t)) \\ \vdots \\ T_{N-1}(\hat{\mathbf{x}}(t)) \end{bmatrix} \qquad \mathbf{C}(t) = \begin{bmatrix} \mathbf{c}_0(t) \\ \mathbf{c}_1(t) \\ \vdots \\ \mathbf{c}_{N-1}(t) \end{bmatrix} \quad (10)$$

so that (9) can be expressed in matrix form as

$$\bar{\boldsymbol{\tau}}(t) = \hat{\boldsymbol{\tau}}(t) - [\mathbf{T}(\mathbf{x}(t-1)) - \mathbf{C}(t)\hat{\mathbf{x}}(t-1)] \quad (11)$$

Similarly, defining

$$\boldsymbol{\Sigma} = \text{diag}\begin{bmatrix} \sigma_0^2 & \sigma_1^2 & \cdots & \sigma_{N-1}^2 \end{bmatrix} \quad (12)$$

enables (8) to be expressed as

$$\epsilon(\mathbf{x}; t) = [\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}]^T \boldsymbol{\Sigma}^{-1} [\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}] \quad (13)$$

In past work, the criterion (6) was minimized for each time instant t, typically with a closed-form approximation to the true minimum [6, 7, 5, 4]. Thereafter, some authors have proposed using a Kalman filter to smooth the position estimates over time [4, 11]. In this work, we propose to incorporate the smoothing stage directly into the estimation. This is acomplished as follows: First we note that (13) represents a *nonlinear* least squares estimation problem that has been appropriately linearized; we can associate $\hat{\boldsymbol{\tau}}(t)$ with the *observation* vector appearing in a Kalman filter such as we will encounter in Section 3. Moreover, we can define a model for the motion of the speaker, in the form typically seen in the *process equation* of a Kalman filter. Thereafter, we can apply the standard Kalman filter update formulae directly to the given recursive estimation problem without ever having recourse to a closed-form approximation for the speaker position. It is worth noting that a similar approach was used by Duraiswami *et al* in developing an algorithm for combined audio-video source localization based on a particle filter [9].

To see more clearly how this approach can be implemented, we briefly review a variation of the Kalman filter in Section 3.

3. Iterated Extended Kalman Filter

Let $\mathbf{x}(t)$ denote the *state* of a Kalman filter at time t, and let $\mathbf{y}(t)$ denote the associated observation. Moreover, define a transition matrix $\mathbf{F}(t+1, t)$ which specifies how the state evolves in time, and functional $\mathbf{C}(t, \mathbf{x}(t))$ which specifies how the state is related to the current observation. The Kalman filter is then described by the *process* and *observation* equations:

$$\mathbf{x}(t+1) = \mathbf{F}(t+1, t)\,\mathbf{x}(t) + \boldsymbol{\nu}_1(t) \quad (14)$$

$$\mathbf{y}(t) = \mathbf{C}(t, \mathbf{x}(t)) + \boldsymbol{\nu}_2(t) \quad (15)$$

where $\boldsymbol{\nu}_1(t)$ and $\boldsymbol{\nu}_2(t)$ are the *process* and *observation noise* respectively. By assumption, $\boldsymbol{\nu}_1(t)$ and $\boldsymbol{\nu}_2(t)$ are zero mean with covariances matrices $\mathbf{Q}_1(t)$ and $\mathbf{Q}_2(t)$.

Let $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ denote the predicted estimated state of a Kalman filter using all the observations $\mathcal{Y}_{t-1}$ up to time $t-1$. The *innovation* is defined as the difference between the observation $\mathbf{y}(t)$ and the *prediction* $\mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))$ at time t:

$$\boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{y}(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \tag{16}$$

Using the innovation and the *Kalman gain* $\mathbf{G}_f(t, \mathbf{x}(t|\mathcal{Y}_{t-1}))$, the state estimate can be updated according to [12, §10]

$$\mathbf{x}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_f(t, \mathbf{x}(t|\mathcal{Y}_{t-1}))\, \boldsymbol{\alpha}(t, \mathbf{x}(t|\mathcal{Y}_{t-1})) \tag{17}$$

The details of the recursive calculation of $\mathbf{G}_f(t, \mathbf{x}(t|\mathcal{Y}_{t-1}))$ can be found in Haykin [12, §10].

Jazwinski [13, §8.3] describes an *iterated extended Kalman filter* (IEKF), in which (16–17) are replaced with the *local iteration*,

$$\boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) = \mathbf{y}(t) - \mathbf{C}(t, \boldsymbol{\eta}_i) \tag{18}$$

$$\boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) = \boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) - \mathbf{C}(\boldsymbol{\eta}_i)\,[\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) - \boldsymbol{\eta}_i] \tag{19}$$

$$\boldsymbol{\eta}_{i+1} = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_f(t, \boldsymbol{\eta}_i)\boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) \tag{20}$$

where $\mathbf{C}(\boldsymbol{\eta}_i)$ is the linearization of $\mathbf{C}(t, \boldsymbol{\eta}_i)$ about $\boldsymbol{\eta}_i$. The local iteration is initialized by setting

$$\boldsymbol{\eta}_1 = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) = \mathbf{F}(t, \hat{\mathbf{x}}(t-1|\mathcal{Y}_{t-1}))$$

Note that $\boldsymbol{\eta}_2 = \hat{\mathbf{x}}(t|\mathcal{Y})$ as defined in (17). Hence, if the local iteration is run only once, the IEKF reduces to an extended Kalman filter. Normally (18–19) are repeated, however, until there are no substantial changes between $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_{i+1}$. Both $\mathbf{G}_f(t, \boldsymbol{\eta}_i)$ and $\mathbf{C}(\boldsymbol{\eta}_i)$ are updated for each local iteration. After the last iteration, we set $\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \boldsymbol{\eta}_f$. Jazwinski [13, §8.3] reports that the IEKF provides faster convergence in the presence of significant nonlinearities in the observation equation, especially when the initial state $\mathbf{x}(1|\mathcal{Y}_0)$ is far from the optimal value.

4. Speaker Tracking with the Kalman Filter

In this section, we discuss the specifics of how the linearized least squares position estimation criterion (13) can be recursively minimized with the iterated extended Kalman filter presented in the prior section. We begin by associating the TDOA estimate $\boldsymbol{\tau}(t)$ with the observation $\mathbf{y}(t)$. Moreover, we recognize that the linearized observation functional $\mathbf{C}(t)$ required for the Kalman filter is given by (5) and (10) for our acoustic localization problem. Furthermore, we can equate the TDOA error covariance matrix $\boldsymbol{\Sigma}$ in (12) with the observation noise covariance $\mathbf{Q}_2(t)$. Hence, we have all relations needed on the observation side of the Kalman filter. We need only supplment these with an appropriate model of the speaker's dynamics to develop an algorithm capable of tracking a moving speaker, as opposed to merely finding his position at a single time instant. This is our next task.

Consider the simplest model of speaker dynamics, wherein the speaker is "stationary" inasmuch as he moves only under the influence of the process noise $\boldsymbol{\nu}_1(t)$. The transition matrix is then $\mathbf{F}(t+1|t) = \mathbf{I}$. Assuming the process noise components in the three directions are statistically independent, we can write

$$\mathbf{Q}_1(t) = \sigma^2\, T^2 \mathbf{I} \tag{21}$$

where T is the time since the last state update. Although the audio sampling is synchronous for all sensors, it cannot be assumed that the speaker constantly speaks, nor that all microphones receive the direct signal from the speaker's mouth; i.e., the speaker sometimes turns so that he is no longer facing the microphone array. As only the direct signal is useful for localization [14], the TDOA estimates returned by those sensors receiving only the indirect signal reflected from the walls should not be used for position updates. This is most easily done by setting a threshold on the GCC (7), and using for source localization only those microphone pairs returning a peak in the GCC above the threshold [14]. This implies that no update at all is made if the speaker is not speaking. Alternatively, a partial update can be made based only on the prediction

$$\hat{\mathbf{X}}(t+1|\mathcal{Y}_t) = \mathsf{F}(t, \hat{\mathbf{X}}(t|\mathcal{Y}_{t-1})$$

The nonlinear functional $\mathbf{C}(t, \mathbf{x}(t))$ corresponds to the TDOA model

$$\mathbf{T}(t, \mathbf{x}(t)) = \begin{bmatrix} T_0(\mathbf{x}(t)) \\ T_1(\mathbf{x}((t)) \\ \vdots \\ T_{N-1}(\mathbf{x}(t)) \end{bmatrix}$$

where the individual components $T_i(\mathbf{x}(t))$ are given by (2–3). The linearized functional $\mathbf{C}(\mathbf{x}(t))$ is given by (5) and (10). As explained in [12], a numerically stable version of the Kalman filter based on the Cholesky decomposition can be readily developed.

Although the IEKF with the local iteration (18–20) was used for the experiments reported in Section 5, the localization system ran in less than real time on a Pentium Xeon processor with a clock speed of 3.0 GHz. This is so because during normal operation very few local iterations are required before the estimate converges. The local iteration compensates for the difference between the original nonlinear least squares estimation criterion (6) and the linearized criterion (8). The difference between the two is only significant during the starting phase, when the estimated position is far from the true speaker location; once the speaker position has been acquired to a reasonable accuracy, the linearized model (8) matches the original (6) quite well. The use of such a linearized model can be equated with the *Gauss-Newton method*, wherein higher order terms in the series expansion of the criterion function are neglected. The connection between the Kalman filter and the Gauss-Newton method is well-known, as is the fact that the convergence rate of the latter is superlinear if the error $\hat{\tau}_i - T_i(\mathbf{x})$ is small near the optimal solution $\mathbf{x} = \mathbf{x}^*$. Further details can be found in Bertsekas [15, §1.5].

5. Experiments

The test set used to evaluate the algorithms proposed here contains approximately three hours of audio and video data recorded during seven seminars by students and faculty at the University of Karlsruhe (UKA) in Karlsruhe, Germany. Prior to the start of the seminars, four video cameras in the corners of the room had been calibrated with the technique of Zhang [16]. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.7 second. Using these hand-marked labels, the true position of the speaker's head in three dimensions was calculated using the technique described in [17]. These "ground truth" speaker positions are accurate to within 10 cm.

As the seminars took place in an open lab area used both by seminar participants as well as students and staff engaged in other activities, the recordings are optimally-suited for evaluating acoustic source localization and other technologies in a realistic, natural setting. In addition to speech from the seminar speaker, the farfield recordings contain noise from fans, computers, and doors, in addition to cross-talk from other people

Algorithm	RMS Error	
	Azimuth (deg)	Depth (cm)
SX	24.6	148
SX + Kalman filter	20.4	145
LI	17.6	234
LI + Kalman filter	13.3	207
IEKF	11.4	119

Table 1: Experimental results of source localization- and tracking algorithms

Algorithm	RMS Error	
	Azimuth (deg)	Depth (cm)
IEKF	11.4	119
IEKF with adaptive threshold	8.64	65

Table 2: IEKF with and without adaptive threshold

present in the room. For these initial experiments, the seminars were recorded with an sixteen-element, linear array with an inter-element spacing of 4 cm.

Table 1 presents the results of a set of experiments comparing the new IEKF algorithm proposed in this work, to the the spherical intersection (SX) method of Schau and Robinson [6], as well as the linear intersection (LI) technique of Brandstein *et al* [4].

The SX method used three microphones of the array (mic-number 0, 2 and 4) with a distance of 8.1 cm between a pair to make an initial estimate of the speaker's position. The LI and IEKF techniques, on the other hand, made use of the same set of 12 microphone pairs. These pairs were formed out of the microphone array by dividing the array into two 8-channel subarrays and taking each possible pair of microphones with an inter-element distance of 8.1 cm . In all cases, the TDOAs were estimated using the generalized cross correlation [1].

The results shown in Table 1 summarize the position estimation error over the 14 segments of the CHIL seminar data. The root mean square (RMS) errors for azimuth and depth were obtained by comparing the true speaker positions obtained from the video labels with the position estimates produced by the several acoustic source localization algorithms. Position estimates from the SX and LI methods lying outside the physical borders of the room were omitted.

Without any smoothing, the source localization estimates returned by both the LI and SX methods are very inaccurate. The LI method provides particularly poor estimates in depth. Kalman filtering improved the position estimates provided by both the SX and LI methods, yet the average RMS distance from the true source location remained large. The new IEKF approach outperformed both the SX and LI methods for both azimuth and depth. We attribute this superior performance largely to the elimination of the initial closed-form estimate associated with the LI and SX methods, and its inherent inaccuracy.

The result of the IEKF could be further improved by implementing an adaptive threshold as proposed by [14]. The total gain is about 30 percent in terms of azimuth and about 40 percent in depth as shown in Table 2.

6. Acknowledgements

This work was sponsored by the European Union under the integrated project CHIL, *Computers in the Human Interaction Loop*, contract number 506909.

7. References

[1] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proc. ICASSP*, vol. II, 1994, pp. 273–6.

[2] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[3] M. S. Brandstein, "A framework for speech source localization using sensor arrays," Ph.D. dissertation, Brown University, Providence, RI, May 1995.

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

[5] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Trans. Signal Proc.*, vol. 42, no. 8, pp. 1905–15, August 1994.

[6] H. C. Schau and A. Z. Robinson, "Passive source localization employing intersecting spherical surfaces from time-of-arrival differences," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. ASSP-35, no. 8, pp. 1223–5, August 1987.

[7] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoust. Speech Signal Proc.*, vol. ASSP-35, no. 12, pp. 1661–9, December 1987.

[8] T. Dvorkin and S. Gannot, "Speaker localization exploiting spatial-temporal information," in *The International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Sep. 2003, pp. 295–298.

[9] R. Duraiswami, D. Zotkin, and L. Davis, "Multimodal 3-d tracking and event detection via the particle filter," in *Workshop on Event Detection in Video, International Conference on Computer Vision*, 2001, pp. 20–27.

[10] J. Chen, J. Benesty, and Y. A. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Proc.*, vol. 11, no. 6, pp. 549–57, November 2003.

[11] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video signal processing for object localization and tracking," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Heidelberg, Germany: Springer Verlag, 2001, ch. 10.

[12] S. Haykin, *Adaptive Filter Theory*, 4th ed. New York: Prentice Hall, 2002.

[13] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic Press, 1970.

[14] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," in *Proc. Eurospeech*, vol. II, 2003, pp. 501–4.

[15] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1995.

[16] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis Machine Intel.*, vol. 22, pp. 1330–1334, 2000.

[17] D. Focken and R. Stiefelhagen, "Towards vision-based 3-D people tracking in a smart room," in *IEEE Int. Conf. Multimodal Interfaces*, October 2002.