

# Kalman Filters for Time Delay of Arrival-Based Source Localization

Ulrich Klee, Tobias Gehrig, John McDonough\*

*Institut für Theoretische Informatik  
Universität Karlsruhe  
Am Fasanengarten 5  
76131 Karlsruhe, Germany*

---

## 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.

*Key words:* microphone arrays, source localization, Kalman filters

---

---

\*

*Email address:* [jmcd@ira.uka.de](mailto:jmcd@ira.uka.de) (John McDonough).

*URL:* <http://isl.ira.uka.de/~jmcd> (John McDonough).

<sup>1</sup> This work was sponsored by the European Union under the integrated project CHIL, *Computers in the Human Interaction Loop*, contract number 506909.

## 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 of the Kalman filter. It is worth noting that similar approaches have been proposed by Gannot *et al* [8] for an acoustic source localizer, as well as by Duraiswami *et al* for a combined audio-video source localization algorithm based on a particle filter [9].

We are indebted to a reviewer who called our attention to other recent work in which particle filters were applied to the acoustic source localization problem [10,11]. As explained in the tutorial by Arulampalam *et al* [12], particle filters represent a generalization of Kalman filters that can handle non-linear and non-Gaussian state estimation problems. This is certainly a desirable characteristic, and makes particle filters of interest for future study. It remains to be seen, however, whether particle filters will prove better suited for acoustic source localization than the extended Kalman filters considered here. To wit, Arulampalam *et al* [12] discuss several problems that can arise with the use of particle filters, namely, *degeneracy* and *sample impoverishment*. While solutions for circumventing these problems have appeared in the literature, the application of a particle filter to a tracking problem clearly requires a certain amount of engineering to obtain a working system, much as with our approach based on the Kalman filter. Moreover, it is not clear that the assumptions inherent in the Kalman filter, namely linearity and Gaussianity, make it unsuitable for the speaker tracking problem: Hahn and Tretter [13] show that the observation noise encountered in time delay of arrival estimation is in fact Gaussian, as required by a Kalman filter. Moreover, as shown here, the nonlinearity seen in the acoustic source localization problem is relatively mild and can be adequately handled by performing several local iterations for each time step as explained in [14]. Such theoretical considerations notwithstanding, the question of whether Kalman or particle filters are better suited for speaker tracking will only be answered by empirical studies. We believe that such studies should be conducted on real, rather than simulated, data such as we have used for the experiments discussed in Section 5, as only results obtained on real data will be truly compelling. We hope to make such empirical comparisons the topic of a future publication.

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 the standard and extended Kalman filters. It also presents a less well-known variant known as the iterated extended Kalman filter. Section 4 first discusses two possible models for speaker motion, then discusses how the preceding development 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. Section 6 presents our conclusions and plans for future work. Appendix A presents a numerically stable implementation of the Kalman filtering algorithms discussed in this work that is based on the Cholesky decomposition.

## 2 Source Localization

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

$$T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{\|\mathbf{x} - \mathbf{m}_{i1}\| - \|\mathbf{x} - \mathbf{m}_{i2}\|}{s} \quad (2.1)$$

where  $s$  is the speed of sound. Denoting

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

allows (2.1) to be rewritten as

$$T_i(\mathbf{x}) = T(\mathbf{m}_{i1}, \mathbf{m}_{i2}, \mathbf{x}) = \frac{1}{s}(d_{i1} - d_{i2}) \quad (2.2)$$

where

$$\begin{aligned} d_{ij} &= \sqrt{(x - m_{ij,x})^2 + (y - m_{ij,y})^2 + (z - m_{ij,z})^2} \\ &= \|\mathbf{x} - \mathbf{m}_{ij}\| \end{aligned} \quad (2.3)$$

is the distance from the source to microphone  $\mathbf{m}_{ij}$ . Source localization based on a maximum likelihood (ML) criterion [2] proceeds by minimizing the error

function

$$\epsilon(\mathbf{x}) = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} [\hat{\tau}_i - T_i(\mathbf{x})]^2 \quad (2.4)$$

where  $\hat{\tau}_i$  is the observed TDOA for the  $i$ -th microphone pair and  $\sigma_i^2$  is the error covariance associated with this observation. The TDOAs can be estimated with a variety of well-known techniques [1,15]. Perhaps the most popular method involves the *phase transform* (PHAT), a variant of the generalized cross correlation (GCC), which can be expressed as

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})}{|X_1(e^{j\omega\tau})X_2^*(e^{j\omega\tau})|} e^{j\omega\tau} d\omega \quad (2.5)$$

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].

Solving for that  $\mathbf{x}$  minimizing (2.4) would be eminently straightforward were it not for the fact that (2.2) is nonlinear in  $\mathbf{x} = (x, y, z)$ . In the coming development, we will find it useful to have a linear approximation. Hence, we take a partial derivative with respect to  $x$  on both sides of (2.2) 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

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

Although (2.4) implies we should find that  $\mathbf{x}$  which minimizes the instantaneous error criterion, we would be better advised to attempt to minimize such an error criterion over a *series* of time instants. In so doing, we exploit the fact that the speaker's position cannot change instantaneously; thus, both the present  $\hat{\tau}_i(t)$  and past TDOA estimates  $\{\hat{\tau}_i(n)\}_{n=0}^{t-1}$  are potentially useful in estimating a speaker's current position  $\mathbf{x}(t)$ . Hence, let us approximate  $T_i(\mathbf{x})$  with a first order Taylor series expansion about the last position estimate  $\hat{\mathbf{x}}(t-1)$  by writing

$$T_i(\mathbf{x}) \approx T_i(\hat{\mathbf{x}}(t-1)) + \mathbf{c}_i^T(t) [\mathbf{x} - \hat{\mathbf{x}}(t-1)] \quad (2.6)$$

where we have defined the row vector

$$\mathbf{c}_i^T(t) = [\nabla_{\mathbf{x}} T_i(\mathbf{x})]_{\mathbf{x}=\hat{\mathbf{x}}(t-1)}^T = \frac{1}{s} \cdot \left[ \frac{\mathbf{x} - \mathbf{m}_{i1}}{d_{i1}} - \frac{\mathbf{x} - \mathbf{m}_{i2}}{d_{i2}} \right]_{\mathbf{x}=\hat{\mathbf{x}}(t-1)}^T \quad (2.7)$$

Substituting the linearization (2.6) into (2.4) provides

$$\begin{aligned}\epsilon(\mathbf{x}; t) &\approx \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left\{ \hat{\tau}_i(t) - T_i(\hat{\mathbf{x}}(t-1)) - \mathbf{c}_i^T(t) [\mathbf{x} - \hat{\mathbf{x}}(t-1)] \right\}^2 \\ &= \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} [\bar{\tau}_i(t) - \mathbf{c}_i^T(t)\mathbf{x}]^2\end{aligned}\quad (2.8)$$

where

$$\bar{\tau}_i(t) = \hat{\tau}_i(t) - T_i(\hat{\mathbf{x}}(t-1)) + \mathbf{c}_i^T(t)\hat{\mathbf{x}}(t-1) \quad (2.9)$$

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

$$\bar{\boldsymbol{\tau}}(t) = \begin{bmatrix} \bar{\tau}_0(t) \\ \bar{\tau}_1(t) \\ \vdots \\ \bar{\tau}_{N-1}(t) \end{bmatrix} \quad \hat{\boldsymbol{\tau}}(t) = \begin{bmatrix} \hat{\tau}_0(t) \\ \hat{\tau}_1(t) \\ \vdots \\ \hat{\tau}_{N-1}(t) \end{bmatrix} \quad \mathbf{T}(\hat{\mathbf{x}}(t)) = \begin{bmatrix} T_0(\hat{\mathbf{x}}(t)) \\ T_1(\hat{\mathbf{x}}(t)) \\ \vdots \\ T_{N-1}(\hat{\mathbf{x}}(t)) \end{bmatrix}$$

and

$$\mathbf{C}(t) = \begin{bmatrix} \mathbf{c}_0^T(t) \\ \mathbf{c}_1^T(t) \\ \vdots \\ \mathbf{c}_{N-1}^T(t) \end{bmatrix} \quad (2.10)$$

so that (2.9) can be expressed in matrix form as

$$\bar{\boldsymbol{\tau}}(t) = \hat{\boldsymbol{\tau}}(t) - [\mathbf{T}(\hat{\mathbf{x}}(t-1)) - \mathbf{C}(t)\hat{\mathbf{x}}(t-1)] \quad (2.11)$$

Similarly, defining

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_0^2 & & & \\ & \sigma_1^2 & & \\ & & \ddots & \\ & & & \sigma_{N-1}^2 \end{bmatrix} \quad (2.12)$$

enables (2.8) to be expressed as

$$\epsilon(\mathbf{x}; t) = [\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}]^T \boldsymbol{\Sigma}^{-1} [\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}] \quad (2.13)$$

In past work, the criterion in (2.4) was minimized for each time instant  $t$ , typically with a closed-form approximation [6,7,5,4,16]. Thereafter, some authors have proposed using a Kalman filter to smooth the position estimates over time [17]. In this work, we propose to incorporate the smoothing stage

directly into the estimation. This is accomplished as follows: First we note that (2.13) 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 similar approaches have been proposed by Gannot *et al* [8] for an acoustic source localizer, as well as by Duraiswami *et al* for a combined audio-video source localization algorithm based on a particle filter [9].

To see more clearly how this approach can be implemented, we review the Kalman filter and several variations thereof in Section 3.

### 3 Kalman Filtering

To set the stage for the development to follow, this section summarizes the Kalman filter based on the Riccati equation, as well as the extended Kalman filter.

#### 3.1 Riccati-Based Kalman Filter

Our purpose here is to present, without proof, the principal quantities and equations required to implement a Kalman filter based on the Riccati equation. Let  $\mathbf{x}(t)$  denote the current state of a Kalman filter and  $\mathbf{y}(t)$  the current observation. Normally,  $\mathbf{x}(t)$  cannot be observed directly and thus must be inferred from the times series  $\{\mathbf{y}(t)\}_t$ ; this is the primary function of the Kalman filter. The operation of the Kalman filter is governed by a *state space model* consisting of a *process* and an *observation* equation, respectively,

$$\mathbf{x}(t+1) = \mathbf{F}(t+1, t) \mathbf{x}(t) + \boldsymbol{\nu}_1(t) \quad (3.1)$$

$$\mathbf{y}(t) = \mathbf{C}(t) \mathbf{x}(t) + \boldsymbol{\nu}_2(t) \quad (3.2)$$

where  $\mathbf{F}(t+1, t)$  and  $\mathbf{C}(t)$  are the known *transition* and *observation* matrices. By definition, the transition matrix satisfies

$$\mathbf{F}(t+1, t) \mathbf{F}(t, t+1) = \mathbf{F}(t, t+1) \mathbf{F}(t+1, t) = \mathbf{I} \quad (3.3)$$

In (3.1–3.2) the *process* and *observation noise* terms are denoted by  $\boldsymbol{\nu}_1(t)$  and  $\boldsymbol{\nu}_2(t)$  respectively. These noise terms are by assumption zero mean, white

Gaussian random vector processes with covariance matrices defined by

$$\mathcal{E}\{\boldsymbol{\nu}_i(t)\boldsymbol{\nu}_i^T(k)\} = \begin{cases} \mathbf{Q}_i(t) & \text{for } t = k \\ \mathbf{0} & \text{otherwise} \end{cases}$$

for  $i = 1, 2$ . Moreover,  $\boldsymbol{\nu}_1(t)$  and  $\boldsymbol{\nu}_2(k)$  are statistically independent such that  $\mathcal{E}\{\boldsymbol{\nu}_1(t)\boldsymbol{\nu}_2^T(k)\} = \mathbf{0}$  for all  $t$  and  $k$ .

In the sequel, it will prove useful to define two estimates of the current state  $\mathbf{x}(t)$ : Let  $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$  denote the *predicted state estimate* of  $\mathbf{x}(t)$  obtained from all observations  $\mathcal{Y}_{t-1} = \{\mathbf{y}(i)\}_{i=0}^{t-1}$  up to time  $t - 1$ . The *filtered state estimate*  $\hat{\mathbf{x}}(t|\mathcal{Y}_t)$ , on the other hand, is based on all observations  $\mathcal{Y}_t = \{\mathbf{y}(i)\}_{i=0}^t$  up to time  $t$ . The *predicted observation* is then given by

$$\hat{\mathbf{y}}(t|\mathcal{Y}_{t-1}) = \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$$

as can be readily derived from (3.1). By definition, the *innovation* is the difference

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) \quad (3.4)$$

between actual and predicted observations. Exploiting the statistical independence of  $\boldsymbol{\nu}_1(t)$  and  $\boldsymbol{\nu}_2(t)$ , the correlation matrix of the innovations sequence can be expressed as

$$\begin{aligned} \mathbf{R}(t) &= \mathcal{E}\{\boldsymbol{\alpha}(t)\boldsymbol{\alpha}^T(t)\} \\ &= \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \end{aligned} \quad (3.5)$$

where

$$\mathbf{K}(t, t-1) = \mathcal{E}\{\boldsymbol{\epsilon}(t, t-1)\boldsymbol{\epsilon}^T(t, t-1)\}$$

is the correlation matrix of the *predicted state error*,

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

The *Kalman gain* is defined as

$$\mathbf{G}(t) = \mathcal{E}\{\mathbf{x}(t+1)\boldsymbol{\alpha}^T(t)\}\mathbf{R}^{-1}(t)$$

This definition can be readily shown to be equivalent to

$$\mathbf{G}(t) = \mathbf{F}(t+1, t)\mathbf{K}(t, t-1)\mathbf{C}^T(t)\mathbf{R}^{-1}(t) \quad (3.6)$$

To calculate  $\mathbf{G}(t)$ , we must know  $\mathbf{K}(t, t-1)$  in advance. The latter is available from the *Riccati equation*, which can be stated as

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \quad (3.7)$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{F}(t, t+1)\mathbf{G}(t)\mathbf{C}(t)]\mathbf{K}(t, t-1) \quad (3.8)$$

*Input vector process:*  $\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(t)$

*Known parameters:*

- state transition matrix:  $\mathbf{F}(t+1, t)$
- measurement matrix:  $\mathbf{C}(t)$
- covariance matrix of process noise:  $\mathbf{Q}_1(t)$
- covariance matrix of measurement noise:  $\mathbf{Q}_2(t)$
- initial diagonal loading:  $\sigma_D^2$

*Initial conditions:*

$$\begin{aligned}\hat{\mathbf{x}}(1|\mathcal{Y}_0) &= \mathbf{x}_0 \\ \mathbf{K}(1, 0) &= \frac{1}{\sigma_D^2} \mathbf{I}\end{aligned}$$

*Computation:*  $t = 1, 2, 3, \dots$

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \quad (3.10)$$

$$\mathbf{G}(t) = \mathbf{F}(t+1, t)\mathbf{K}(t, t-1)\mathbf{C}^T(t)\mathbf{R}^{-1}(t) \quad (3.11)$$

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) \quad (3.12)$$

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}(t)\boldsymbol{\alpha}(t) \quad (3.13)$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{F}(t, t+1)\mathbf{G}(t)\mathbf{C}(t)] \mathbf{K}(t, t-1) \quad (3.14)$$

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \quad (3.15)$$

Table 1

Calculations for Kalman filter based on the Riccati equation.

where

$$\mathbf{K}(t) = \mathcal{E} \left\{ \boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}^T(t) \right\}$$

is the correlation matrix of the *filtered state error*,

$$\boldsymbol{\epsilon}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t|\mathcal{Y}_t)$$

Finally, the filtered state estimate can be updated based on the Kalman gain  $\mathbf{G}(t)$  and innovation  $\boldsymbol{\alpha}(t)$  according to

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}(t)\boldsymbol{\alpha}(t) \quad (3.9)$$

These relations are summarized in Table 1.

We have now formulated the *one-step prediction* form of the Kalman filter, which returns the predicted state estimate  $\hat{\mathbf{x}}(t+1|\mathcal{Y}_t)$ . In Section 3.2, we will require the filtered state estimate  $\hat{\mathbf{x}}(t|\mathcal{Y}_t)$ , which can be obtained as follows. Taking an expectation conditioned on  $\mathcal{Y}_t$  on both sides of (3.1), we can write

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_t) + \hat{\boldsymbol{\nu}}_1(t|\mathcal{Y}_t) \quad (3.16)$$

As the process noise is zero mean, we have  $\hat{\boldsymbol{\nu}}_1(t|\mathcal{Y}_t) = \mathbf{0}$ , so that (3.16) becomes

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_t) \quad (3.17)$$

To obtain the desired filtered estimate, we multiply both sides of (3.17) by  $\mathbf{F}(t|t+1)$  and invoke (3.3) and write

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \mathbf{F}(t, t+1)\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) \quad (3.18)$$

### 3.2 Extended Kalman Filter (EKF)

For the sake of completeness, we provide here a brief derivation of the general extended Kalman filter (EKF). This development is based on that in Haykin [18, §10.10].

To begin, let us split the filtered state estimate update formula (3.18) into two steps. Firstly, we make a one-step prediction to update  $\hat{\mathbf{x}}(t-1|\mathcal{Y}_{t-1})$  to  $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ , which is achieved by (3.17). Secondly, we update  $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$  to  $\hat{\mathbf{x}}(t|\mathcal{Y}_t)$ , which is achieved by substituting (3.9) into (3.18) and using (3.3) to simplify:

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

where the *filtered Kalman gain*  $\mathbf{G}_F(t)$  is defined as

$$\mathbf{G}_F(t) = \mathbf{F}(t, t+1)\mathbf{G}(t) \quad (3.19)$$

The complete filtering algorithm is then

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \quad (3.20)$$

$$\mathbf{G}_F(t) = \mathbf{K}(t, t-1)\mathbf{C}^T(t)\mathbf{R}^{-1}(t) \quad (3.21)$$

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) \quad (3.22)$$

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t)\boldsymbol{\alpha}(t) \quad (3.23)$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{G}_F(t)\mathbf{C}(t)]\mathbf{K}(t, t-1) \quad (3.24)$$

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \quad (3.25)$$

and

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_t) \quad (3.26)$$

To formulate the extended Kalman filter, we first posit a less restrictive state-space model, namely

$$\mathbf{x}(t+1) = \mathbf{F}(t+1, t)\mathbf{x}(t) + \boldsymbol{\nu}_1(t) \quad (3.27)$$

$$\mathbf{y}(t) = \mathbf{C}(t, \mathbf{x}(t)) + \boldsymbol{\nu}_2(t) \quad (3.28)$$

where the observation *functional*<sup>2</sup>  $\mathbf{C}(t, \mathbf{x}(t))$  is in general nonlinear and time-varying. The main idea behind the EKF is then to linearize this functional about the most recent state estimate  $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ . The corresponding linearization can be written as

$$\mathbf{C}(t) = \left. \frac{\partial \mathbf{C}(t, \mathbf{x})}{\partial \mathbf{x}} \right|_{\mathbf{x} = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})} \quad (3.29)$$

In the above, entry  $(i, j)$  of  $\mathbf{C}(t, \mathbf{x})$  is the partial derivative of the  $i$ -th component of  $\mathbf{C}(t, \mathbf{x})$  with respect to the  $j$ -th component of  $\mathbf{x}$ .

Based on (3.29), we can express the first order Taylor series of  $\mathbf{C}(t, \mathbf{x}(t))$  as

$$\mathbf{C}(t, \mathbf{x}(t)) \approx \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) + \mathbf{C}(t) [\mathbf{x}(t) - \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})]$$

Using this linearization, the nonlinear state-space equations (3.27–3.28) can be written as

$$\mathbf{x}(t+1) = \mathbf{F}(t+1, t)\mathbf{x}(t) + \boldsymbol{\nu}_1(t) \quad (3.30)$$

$$\bar{\mathbf{y}}(t) \approx \mathbf{C}(t)\mathbf{x}(t) + \boldsymbol{\nu}_2(t) \quad (3.31)$$

where we have defined

$$\bar{\mathbf{y}}(t) = \mathbf{y}(t) - [\mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) - \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})] \quad (3.32)$$

As everything on the right hand side of (3.32) is known at time  $t$ ,  $\bar{\mathbf{y}}(t)$  can be regarded as an observation.

The extended Kalman filter is obtained by applying the computations in (3.26–3.24) to the linearized model in (3.30–3.31), whereupon we find

$$\begin{aligned} \hat{\mathbf{x}}(t+1|\mathcal{Y}_t) &= \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_t) \\ \hat{\mathbf{x}}(t|\mathcal{Y}_t) &= \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t) \boldsymbol{\alpha}(t) \end{aligned} \quad (3.33)$$

$$\begin{aligned} \boldsymbol{\alpha}(t) &= \bar{\mathbf{y}}(t) - \mathbf{C}(t)\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) \\ &= \mathbf{y}(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \end{aligned} \quad (3.34)$$

These computations are summarized in Table 2.

---

<sup>2</sup> Most authors formulate the extended Kalman filter with a nonlinear process functional  $\mathbf{F}(t, \mathbf{x}(t))$  in addition to the observation functional  $\mathbf{C}(t, \mathbf{x}(t))$ ; see, for example, Haykin [18, §10.10]. This more general formulation is not required here.

*Input vector process:*  $\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(t)$

*Known parameters:*

- state transition matrix:  $\mathbf{F}(t+1, t)$
- nonlinear measurement functional:  $\mathbf{C}(t, \mathbf{x}(t))$
- covariance matrix of process noise:  $\mathbf{Q}_1(t)$
- covariance matrix of measurement noise:  $\mathbf{Q}_2(t)$
- initial diagonal loading:  $\sigma_D^2$

*Initial conditions:*

$$\begin{aligned}\hat{\mathbf{x}}(1|\mathcal{Y}_0) &= \mathbf{x}_0 \\ \mathbf{K}(1, 0) &= \frac{1}{\sigma_D^2} \mathbf{I}\end{aligned}$$

*Computation:*  $t = 1, 2, 3, \dots$

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \quad (3.35)$$

$$\mathbf{G}_F(t) = \mathbf{K}(t, t-1)\mathbf{C}^T(t)\mathbf{R}^{-1}(t) \quad (3.36)$$

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (3.37)$$

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t)\boldsymbol{\alpha}(t) \quad (3.38)$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{G}_F(t)\mathbf{C}(t)]\mathbf{K}(t, t-1) \quad (3.39)$$

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \quad (3.40)$$

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t)\hat{\mathbf{x}}(t|\mathcal{Y}_t) \quad (3.41)$$

*Note:* The linearized matrix  $\mathbf{C}(t)$  is computed from the nonlinear functional  $\mathbf{C}(t, \mathbf{x}(t))$  as in (3.29).

Table 2

Calculations for extended Kalman filter.

### 3.3 Iterated Extended Kalman Filter (IEKF)

We now consider a further refinement of the extended Kalman filter. Repeating (3.35–3.38) of Table 2, we can write

$$\mathbf{R}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \quad (3.42)$$

$$\mathbf{G}_F(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{K}(t, t-1)\mathbf{C}^T(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\mathbf{R}^{-1}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (3.43)$$

$$\boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{y}(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (3.44)$$

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (3.45)$$

where we have explicitly indicated the dependence of the relevant quantities on  $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ . Jazwinski [14, §8.3] describes an *iterated extended Kalman filter*

(IEKF), in which (3.42–3.45) are replaced with the *local iteration*,

$$\mathbf{R}(t, \boldsymbol{\eta}_i) = \mathbf{C}(\boldsymbol{\eta}_i)\mathbf{K}(t, t-1)\mathbf{C}^T(\boldsymbol{\eta}_i) + \mathbf{Q}_2(t) \quad (3.46)$$

$$\mathbf{G}_F(t, \boldsymbol{\eta}_i) = \mathbf{K}(t, t-1)\mathbf{C}^T(\boldsymbol{\eta}_i)\mathbf{R}^{-1}(t, \boldsymbol{\eta}_i) \quad (3.47)$$

$$\boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) = \mathbf{y}(t) - \mathbf{C}(t, \boldsymbol{\eta}_i) \quad (3.48)$$

$$\boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) = \boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) - \mathbf{C}(\boldsymbol{\eta}_i) [\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) - \boldsymbol{\eta}_i] \quad (3.49)$$

$$\boldsymbol{\eta}_{i+1} = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t, \boldsymbol{\eta}_i)\boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) \quad (3.50)$$

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

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

Note that  $\boldsymbol{\eta}_2 = \hat{\mathbf{x}}(t|\mathcal{Y})$  as defined in (3.45). Hence, if the local iteration is run only once, the IEKF reduces to the EKF. Normally (3.46–3.50) 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

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \boldsymbol{\eta}_f$$

and this value is used to update  $\mathbf{K}(t)$  and  $\mathbf{K}(t+1, t)$ . Jazwinski [14, §8.3] reports that the IEKF provides faster convergence in the presence of significant nonlinearities in the observation equation, especially when the initial state estimate  $\boldsymbol{\eta}_1 = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$  is far from the optimal value. The calculations for the iterated extended Kalman filter are summarized in Table 3.

### 3.4 Numerical Stability

All variants of the Kalman filter discussed in Sections 3.1–3.3 are based on the Riccati equation (3.7–3.8). Unfortunately, the Riccati equation possesses poor numerical stability properties [18, §11] as can be seen from the following: Substituting (3.8) into (3.7) and making use of (3.3) provides

$$\begin{aligned} \mathbf{K}(t+1, t) &= \mathbf{F}(t+1, t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t) \\ &\quad - \mathbf{G}(t)\mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \end{aligned} \quad (3.59)$$

Manipulating (3.6), we can write

$$\mathbf{R}(t)\mathbf{G}^T(t) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t) \quad (3.60)$$

Then upon substituting (3.60) for the matrix product  $\mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t)$  appearing in (3.59), we find

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t) - \mathbf{G}(t)\mathbf{R}(t)\mathbf{G}^T(t) + \mathbf{Q}_1(t) \quad (3.61)$$

*Input vector process:*  $\mathbf{y}(1), \mathbf{y}(2), \dots, \mathbf{y}(t)$

*Known parameters:*

- state transition matrix:  $\mathbf{F}(t+1, t)$
- nonlinear measurement functional:  $\mathbf{C}(t, \mathbf{x}(t))$
- covariance matrix of process noise:  $\mathbf{Q}_1(t)$
- covariance matrix of measurement noise:  $\mathbf{Q}_2(t)$
- initial diagonal loading:  $\sigma_D^2$

*Initial conditions:*

$$\begin{aligned}\hat{\mathbf{x}}(1|\mathcal{Y}_0) &= \mathbf{x}_0 \\ \mathbf{K}(1, 0) &= \frac{1}{\sigma_D^2} \mathbf{I}\end{aligned}$$

*Computation:*  $t = 1, 2, 3, \dots$

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

*Iterate:*  $i = 1, 2, 3, \dots, f-1$

$$\mathbf{R}(t, \boldsymbol{\eta}_i) = \mathbf{C}(\boldsymbol{\eta}_i) \mathbf{K}(t, t-1) \mathbf{C}^T(\boldsymbol{\eta}_i) + \mathbf{Q}_2(t) \quad (3.51)$$

$$\mathbf{G}_F(t, \boldsymbol{\eta}_i) = \mathbf{K}(t, t-1) \mathbf{C}^T(\boldsymbol{\eta}_i) \mathbf{R}^{-1}(t, \boldsymbol{\eta}_i) \quad (3.52)$$

$$\boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) = \mathbf{y}(t) - \mathbf{C}(t, \boldsymbol{\eta}_i) \quad (3.53)$$

$$\boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) = \boldsymbol{\alpha}(t, \boldsymbol{\eta}_i) - \mathbf{C}(\boldsymbol{\eta}_i) [\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) - \boldsymbol{\eta}_i] \quad (3.54)$$

$$\boldsymbol{\eta}_{i+1} = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t, \boldsymbol{\eta}_i) \boldsymbol{\zeta}(t, \boldsymbol{\eta}_i) \quad (3.55)$$

*Calculate:*

$$\begin{aligned}\hat{\mathbf{x}}(t|\mathcal{Y}_t) &= \boldsymbol{\eta}_f \\ \mathbf{K}(t) &= [\mathbf{I} - \mathbf{G}_F(t, \hat{\mathbf{x}}(t|\mathcal{Y}_t)) \mathbf{C}(\hat{\mathbf{x}}(t|\mathcal{Y}_t))] \mathbf{K}(t, t-1)\end{aligned} \quad (3.56)$$

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t) \mathbf{K}(t) \mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \quad (3.57)$$

$$\hat{\mathbf{x}}(t+1|\mathcal{Y}_t) = \mathbf{F}(t+1, t) \hat{\mathbf{x}}(t|\mathcal{Y}_t) \quad (3.58)$$

*Note:* The local iteration over  $i$  continues until there is no significant difference between  $\boldsymbol{\eta}_i$  and  $\boldsymbol{\eta}_{i+1}$ .

Table 3

Calculations for the iterated extended Kalman filter.

Equation (3.61) illustrates the problem inherent in the Riccati equation: As  $\mathbf{K}(t+1, t)$  is the covariance matrix of the predicted state error  $\boldsymbol{\epsilon}(t+1, t)$ , it must be positive definite. Similarly,  $\mathbf{R}(t)$  is the covariance matrix of the innovation  $\boldsymbol{\alpha}(t)$  and must also be positive definite. Moreover, if  $\mathbf{F}(t+1, t)$  and  $\mathbf{G}(t)$  are full rank, then the terms  $\mathbf{F}(t+1, t) \mathbf{K}(t, t-1) \mathbf{F}^T(t+1, t)$  and  $\mathbf{G}(t) \mathbf{R}(t) \mathbf{G}^T(t)$  are positive definite as well. Therefore, Equation (3.61) implies

that a positive definite matrix  $\mathbf{K}(t+1, t)$  must be calculated as the *difference* of the positive definite matrix  $\mathbf{F}(t+1, t)\mathbf{K}(t, t-1)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t)$  and positive definite matrix  $\mathbf{G}(t)\mathbf{R}(t)\mathbf{G}^T(t)$ . Due to finite precision errors, the resulting matrix  $\mathbf{K}(t+1, t)$  can become indefinite after a sufficient number of iterations, at which point the Kalman filter exhibits a behavior known as *explosive divergence* [18, §11].

As discussed in Appendix A, a more stable implementation of the Kalman filter can be developed based on the *Cholesky decomposition* or *square-root* of  $\mathbf{K}(t+1, t)$ , which is by definition that unique lower triangular matrix  $\mathbf{K}^{1/2}(t+1, t)$  achieving

$$\mathbf{K}(t+1, t) \triangleq \mathbf{K}^{1/2}(t+1, t) \mathbf{K}^{T/2}(t+1, t)$$

The Cholesky decomposition of a matrix exists if and *only* if the matrix is symmetric and positive definite [19, §4.2.3]. The basic idea behind the square-root implementation of the Kalman filter is to update  $\mathbf{K}^{1/2}(t+1, t)$  instead of  $\mathbf{K}(t+1, t)$  directly. By updating or *propagating*  $\mathbf{K}^{1/2}(t+1, t)$  forward in time, it can be assured that  $\mathbf{K}(t+1, t)$  remains positive definite. Thereby a numerically stable algorithm is obtained *regardless* of the precision of the machine on which it runs. Appendix A presents a procedure whereby  $\mathbf{K}^{1/2}(t+1, t)$  can be efficiently propagated in time using a series of *Givens rotations* [19, §5.1.8].

In the acoustic source localization experiments conducted thus far, the numerical stability has proven adequate even using the Kalman filter based directly on the Riccati equation. Instabilities can arise, however, when the audio features are supplemented with video information as in Gehrig *et al* [20]. Hence, we have included Appendix A for the sake of completeness.

#### 4 Speaker Tracking with the Kalman Filter

In this section, we discuss the specifics of how the linearized least squares position estimation criterion (2.13) can be recursively minimized with the iterated extended Kalman filter presented in Section 3.3. We begin by associating the “linearized” TDOA estimate  $\bar{\boldsymbol{\tau}}(t)$  in (2.11) with the modified observation  $\bar{\mathbf{y}}(t)$  appearing in (3.32). Moreover, we recognize that the linearized observation functional  $\mathbf{C}(t)$  in (3.29) required for the Kalman filter is given by (2.7) and (2.10) for our acoustic localization problem. Furthermore, we can equate the TDOA error covariance matrix  $\boldsymbol{\Sigma}$  in (2.12) 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 supplement 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.

#### 4.1 Dynamic Model

Consider the simplest model of speaker dynamics, wherein the speaker is “stationary” inasmuch as he moves only under the influence of the process noise  $\boldsymbol{\nu}_1(t)$ . Assuming the process noise components in the three directions are statistically independent, we can write

$$\mathbf{Q}_1(t) = \sigma_P^2 T^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4.1)$$

where  $T$  is the time elapsed since the last update of the state estimate, and  $\sigma_P^2$  is the process noise power. Typically  $\sigma_P^2$  is set based on a set of empirical trials to achieve the best localization results.

#### 4.2 Position Updates

Before performing an update, it is first necessary to determine the time  $T$  that has elapsed since an observation was last received. This factor appears as a parameter of the process noise covariance matrix  $\mathbf{Q}_1(t)$  in (4.1). Although we assume 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 [21], 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 assured by setting a threshold on the PHAT (2.5), and using for source localization only those microphone pairs returning a peak in the PHAT above the threshold [21]. This implies that no update at all is made if the speaker is silent.

Given the dynamic model in Section 4.1, we now have everything required for an acoustic speaker tracking system. The position update equations are given in Table 3. The nonlinear functional  $\mathbf{C}(t, \mathbf{x}(t))$  appearing there 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 (2.2–2.3). The linearized functional  $\mathbf{C}(t) = \mathbf{C}(\mathbf{x}(t))$  is given by (2.7) and (2.10). To gain an appreciation for the severity of the nonlinearity in this particular Kalman filtering application, we plotted the actual value of  $T_i(\mathbf{x}(t))$  against the linearized version. These plots are shown in Figures 1 and 2, respectively, for deviations

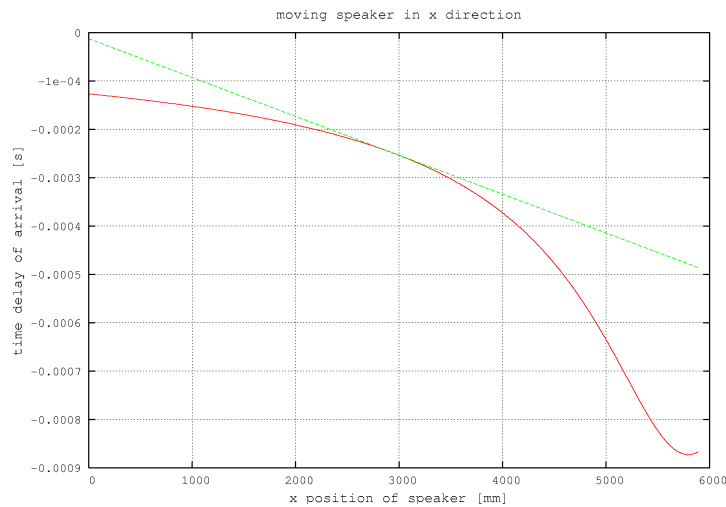


Fig. 1. Actual vs. linearized  $T_i(\mathbf{x}(t))$  for movement in the  $x$ -direction.

in the  $x$ - and  $y$ -directions from the point about which  $T_i(\mathbf{x}(t))$  was linearized. For these plots, the D-Array in Figure 5 was used and the operating point was taken as  $(x, y, z) = (2.950, 4.080, 1.700)$  m in room coordinates, which is approximately in the middle of the room. As is clear from the plots, for deviations of  $\pm 1$  m from the nominal, the linearized TDOA is within 2.33% of the true value for movement in the  $x$ -direction, and within 1.38% for movement in the  $y$ -direction.

Although the IEKF with the local iteration (3.53–3.55) 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, typically five or fewer. The local iteration compensates for the difference between the original nonlinear least squares estimation criterion (2.4) and the linearized criterion (2.8). The difference between the two is only significant during startup and when a significant amount of time has

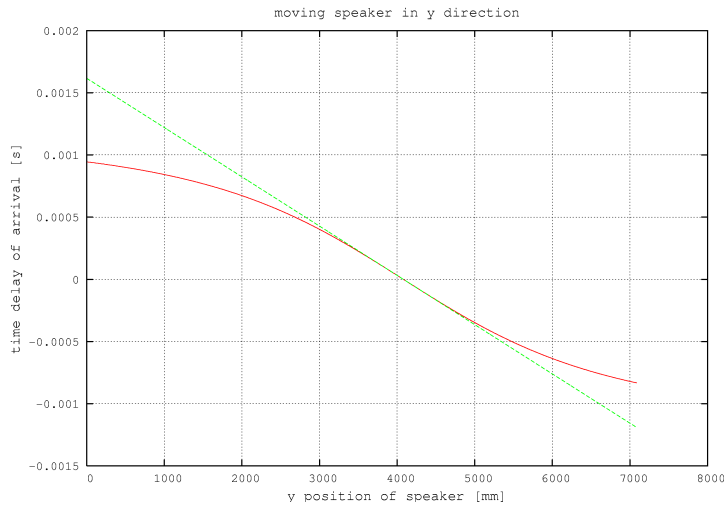


Fig. 2. Actual vs. linearized  $T_i(\mathbf{x}(t))$  for movement in the  $y$ -direction.

passed since the last update, as in such cases the initial position estimate is far from the true speaker location. Once the speaker position has been acquired to a reasonable accuracy, the linearized model (2.8) matches the original (2.4) 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 (2.6) 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 [22, §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 a series of seminars by students and faculty at the Universität Karlsruhe in Karlsruhe, Germany. The seminar room is approximately  $6 \times 7$  m with a calibrated camera in each corner. An initial set of seven seminars was recorded in the Fall of 2003. At that time, the seminar room was equipped with a single linear 16-channel microphone array with an intersensor spacing of 4.1 cm, as shown in Figure 3. The linear array was used for both beamforming and source localization experiments. In 2004, the audio sensors in the seminar room were enhanced to the configuration shown in Figure 5. The 16-channel linear array was replaced with a 64-channel Mark III microphone array developed at the US National Institute of Standards (NIST). This large array is intended primarily for beamforming, and was not used for the source localization experiments reported here. Four T-shaped arrays were mounted on the walls of seminar

ISL\_SEMINAR\_2003 ROOMSETUP

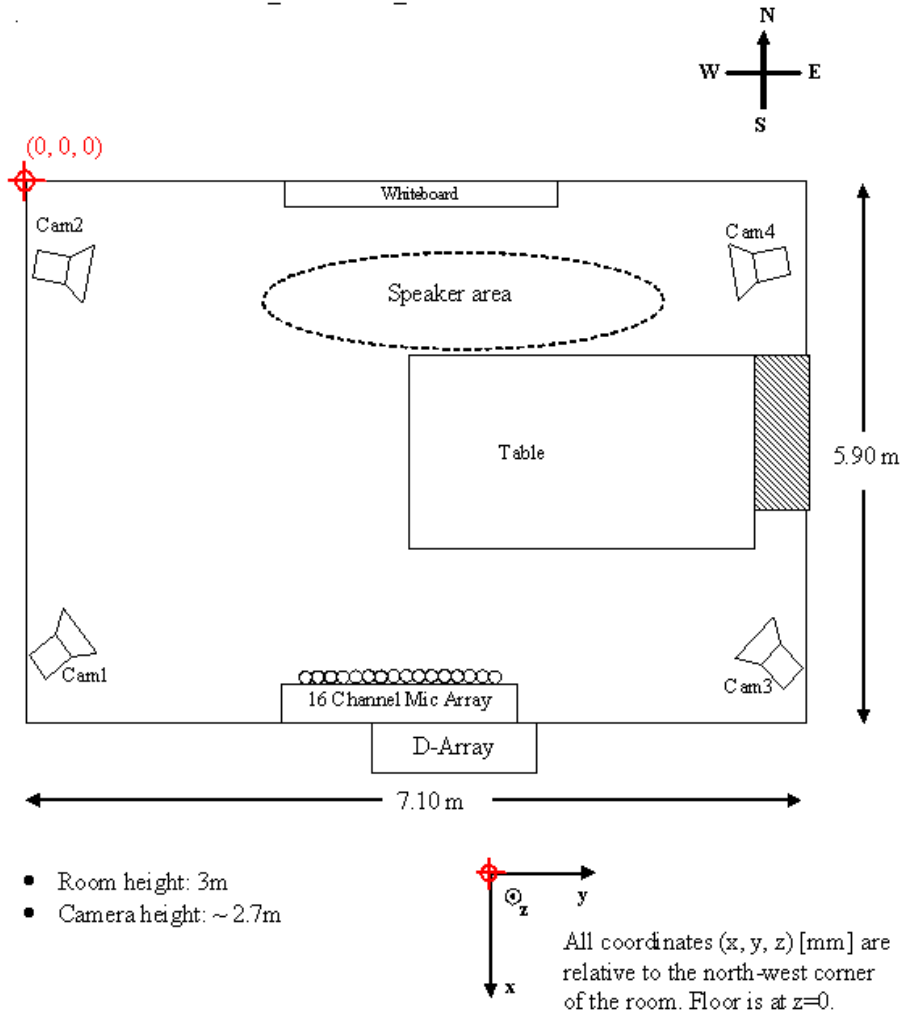


Fig. 3. The room setup for the seminar recordings 2003 at University Karlsruhe.

room. The intersensor spacing of the T-arrays was chosen as either 20 or 30 cm in order to permit more accurate source localization. The T-shape was chosen to permit three-dimensional localization, which was impossible with the initial linear configuration. In the balance of this section, we report experimental results on both sensor configurations.

Prior to the start of the seminars, the four video cameras in the corners of the room had been calibrated with the technique of Zhang [23]. 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 [24]. These “ground truth” speaker positions are accurate to within 10 cm.

Algorithm	RMS Error			
	$x$ (cm)	$y$ (cm)	Azimuth (deg)	Depth (cm)
SX	144.6	166.6	24.6	148
SX + Kalman filter	138.1	153.5	20.4	145
SI	121.4	132.0	16.5	133
SI + Kalman filter	121.2	131.7	16.4	132
LI	225.0	130.5	17.6	234
LI + Kalman filter	196.1	111.5	13.3	207
IEKF	91.7	119.7	12.9	100

Table 4

Experimental results of source localization- and tracking algorithms

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 far field recordings contain noise from fans, computers, and doors, in addition to cross-talk from other people present in the room.

### 5.1 Experiments with Linear Microphone Array

The simple dynamic model presented in Section 4.1 was used for all source localization experiments based on the IEKF reported in this section and in Section 5.2. Table 4 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], the spherical interpolation (SI) method of [25] as well as the linear intersection (LI) technique of Brandstein *et al* [4].

The SX method used three microphones of the array, numbers 0, 2 and 4, to make an initial estimate of the speaker’s position. Only three microphones can be used, as the array is not two-dimensional, unlike the array in the original work [6]. To improve estimation results, the SI method extends the ideas of the SX and enables the use of more than four microphones to take advantage of the redundancy. The array was divided into two subarrays and all microphones of these subarrays were used to estimate two positions. The final position estimate was the average of the two initial estimates. 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 eight-channel subarrays and taking each possible

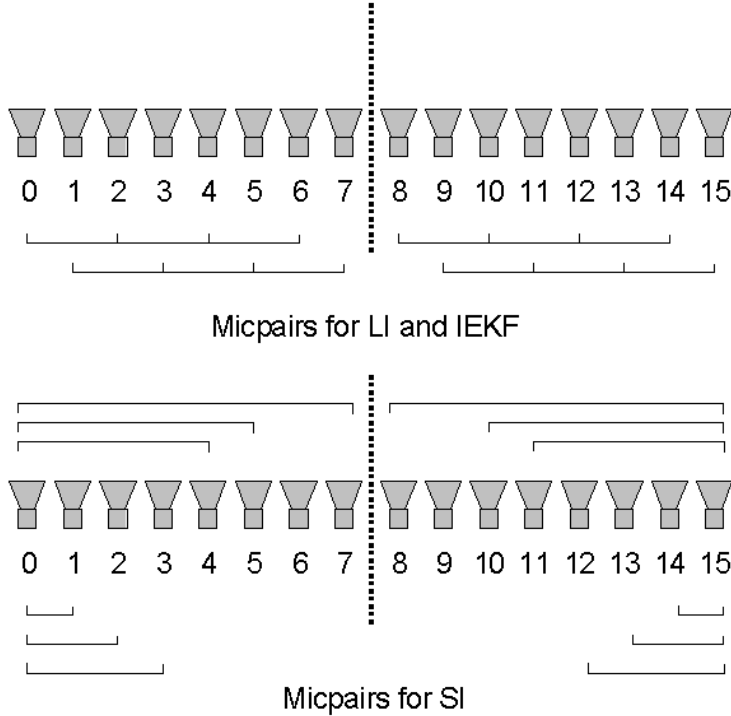


Fig. 4. Use of microphone-pairs for the different methods.

pair of microphones with an inter-element distance of 8.14 cm. In all cases studied here, the TDOAs were estimated using the PHAT (2.5). Figure 4 shows the configuration of the array and definition of the microphone pairs in detail.

The results shown in Table 4 summarize the position estimation error over the 14 segments of the seminar data. The root mean square (RMS) errors 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. The position estimates from the Kalman filter were initially produced in Cartesian coordinates then converted to azimuth and depth. Table 4 reports results in both the original Cartesian coordinates, as well as azimuth and depth, as the latter are more meaningful for the linear array considered here. Position estimates from the SX, SI 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 SI method shows significantly better performance than the SX and LI method both in total precision as well as in stability of the position estimations, as the results didn't show

Algorithm	RMS Error			
	$x$ (cm)	$y$ (cm)	Azimuth (deg)	Depth (cm)
IEKF	91.7	119.7	12.9	100
IEKF with adaptive threshold	52.2	61.9	6.97	52.9

Table 5

IEKF with and without adaptive threshold

the big variance of the first two methods. On the other hand, this improved stability reduces the improvement given by Kalman filtering, hence the filtered results don't show the big improvements noticeable for the SX and LI method. The new IEKF approach outperformed all 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, SI and SX methods, and its inherent inaccuracy.

The performance of the IEKF could be further improved by implementing an adaptive threshold on the PHAT as proposed in [21]. The total gain is about 46% relative in terms of azimuth and about 47% in depth as shown in Table 5.

## 5.2 Experiments with T-shaped Microphone Arrays

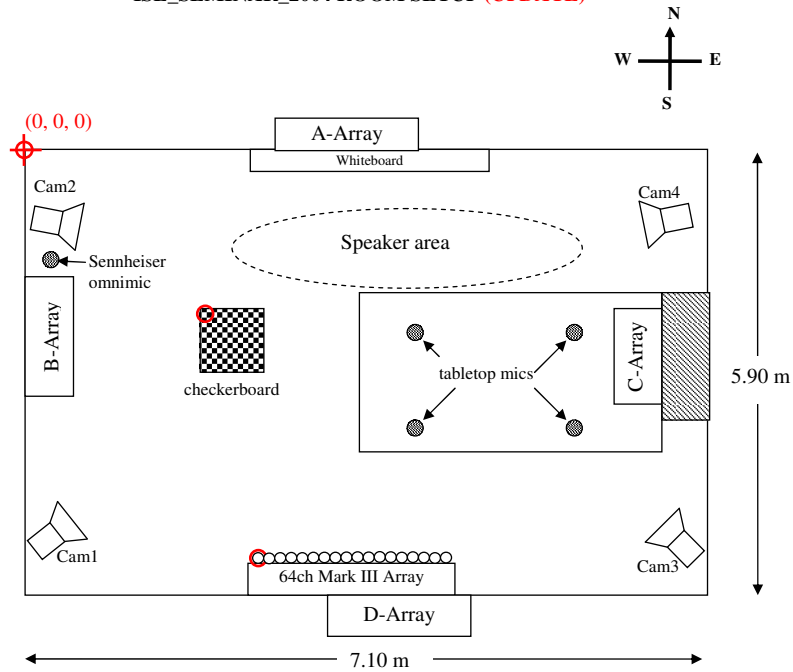
Here we report the results of a set of experiments conducted with the T-shaped arrays whose location is indicated in Figure 5. For the new sensor configuration, we report results only in Cartesian coordinates, as azimuth and depth are no longer meaningful. Shown in Table 6 are source localization results obtained with the IEKF on two data sets: the *development set* consisting of three 15-minute segments, on which the parameters of the Kalman filter were tuned, and the *evaluation set* consisting of ten 15-minute segments chosen from five seminars. These results were obtained with a fixed PHAT threshold.

Experiment	RMS Error (cm)				
	$x$	$y$	$z$	2D	3D
dev. set	39.0	40.4	9.9	56.3	57.1
eval. set	35.3	34.9	10.3	51.8	53.0
dev. set with SAD on CTM	36.0	35.6	9.4	50.7	51.6
eval. set with SAD on CTM	34.5	32.8	9.9	49.2	50.3

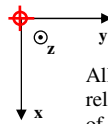
Table 6

IEKF source localization results with the T-shaped arrays, both without and with explicit speech activity detection (SAD) on the close-talking microphone (CTM).

ISL\_SEMINAR\_2004 ROOM SETUP (UPDATE)



	x	y	z
Checkerboard 2004_11	2130	3260	732
Checkerboard 2004_06/07/08	2000	3110	730
Mark III	5665	2900	1710
Array A1	105	3060	2370
Array B1	2150	105	2290
Array C1	2700	6210	2190
Array D1	5795	4280	2400



All coordinates (x, y, z) [mm] are relative to the north-west corner of the room. Floor is at z=0.

- Mark III: 64 ch, 20mm mic distance
- Checkerboard square size: 105mm. Position of the first *inner* crossing is given.
- Checkerboard for *internal* calibration: 42mm square size
- Room height: 3m
- Camera height: ~ 2.7m

A/B/C/D-Array Layout:

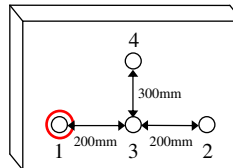


Fig. 5. Current sensor configuration in the seminar room at the Universität Karlsruhe.

Only TDOAs for pairs of microphones from the *same* T-array were estimated, as the distances between the T-arrays were too great to allow for reliable cross correlation. The TDOAs from all microphone pairs for which the PHAT was above the fixed threshold were concatenated into a *single* observation vector, which was then used to update the position estimate. As can be seen upon comparing the results in Table 6 with those in Table 4, the T-shaped arrays provide for significantly more accurate speaker localization. Moreover, the T-shape enables three-dimensional estimation. In the column of Table 6 labeled “3D,” we report the total RMS error for all dimensions; in the column labeled “2D,” the height or  $z$ -dimension is ignored.

Given that the position estimate can only be updated when the speaker is actually speaking, we also investigated the application of speech activity detection (SAD) to the source localization problem. We trained a neural net based speech activity detector on the data from the close-talking microphone worn by the speaker, and only updated for time segments declared to be speech by the detector. We still retained the threshold criterion on the PHAT of each microphone pair, to ensure the update was based the signal received directly from the speaker’s mouth. As shown in Table 6, the use of an explicit SAD module provided a marginal improvement in the performance of the source localizer. We suspect that this is so because the threshold on the PHAT already provides an effective means of speech activity detection.

We also tested the LI method on the T-shaped arrays; the SI and SX methods require the calculation of all times delays with respect to a reference microphone and, as noted previously, the distances between the T-arrays is too great to estimate TDOAs reliably. In this case, we calculated a single bearing line for each microphone array, then calculated the point of nearest intersection for each unique pair of bearing lines. The final position estimate was given by the average of the points of nearest intersection, as specified in Brandstein *et al* [4]. The LI results are shown in Table 7 for the evaluation set.

Experiment	RMS Error (cm)				
	$x$	$y$	$z$	2D	3D
LI	114.2	177.2	45.7	211.5	216.6
LI + Kalman filter	109.9	175.4	45.4	207.6	212.8

Table 7

Source localization results for the T-shaped arrays based on linear intersection.

Comparing the results of Tables 6 and 7, we see that the IEKF still clearly outperforms LI. The accuracy of the LI method improves in the  $x$ -direction when the four T-arrays are used instead of the single linear array. The accuracy in the  $y$ -direction, however, degrades due to the fact that fewer microphone pairs are available to resolve the location of the source along this dimension.

In a final set of studies, we investigated the effect of speaker movement on the number of local iterations required by the IEKF. In Figure 6, we plotted the number of local iterations vs. the time since the last update. Figure 7 shows the

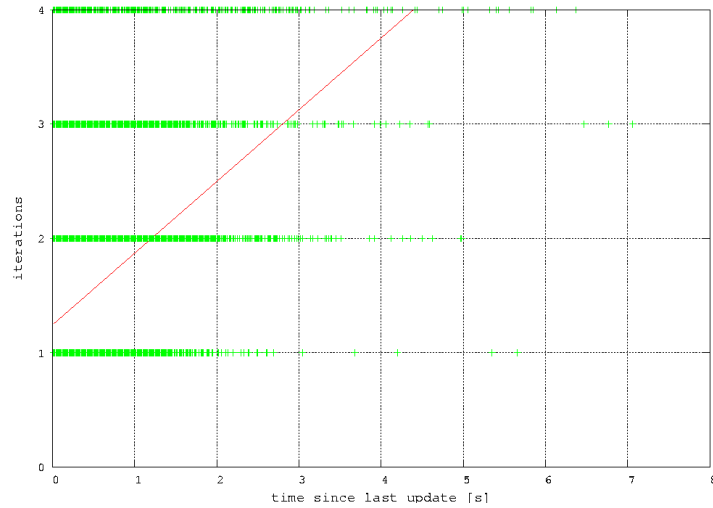


Fig. 6. Number of local iterations vs. time since last update.

number of local iterations plotted against the distance the speaker has moved since the last update. Finally, Figure 8 displays local iterations vs. speaker

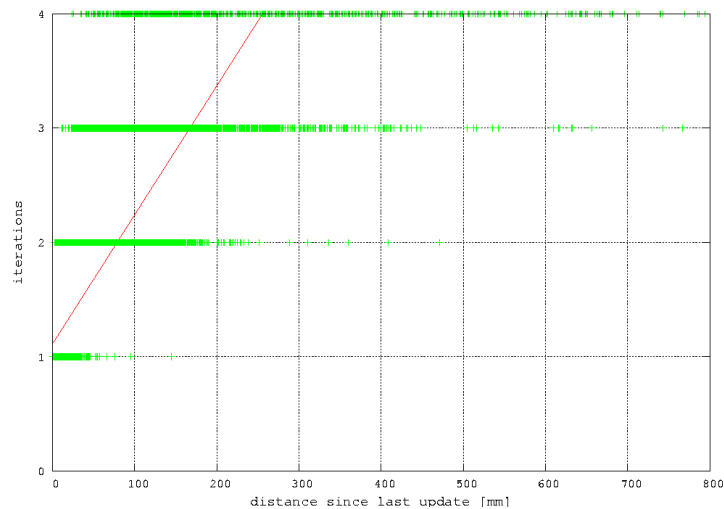


Fig. 7. Number of local iterations vs. distance moved since last update.

velocity. For each of the three cases, we calculated a regression line, which is also shown in the plot. As is clear from the figures, the average number of local iterations increases in proportion to the time since the last update, distance moved since the last update, and speaker velocity. These results correspond to our expectations, in that significant speaker movement implies that the

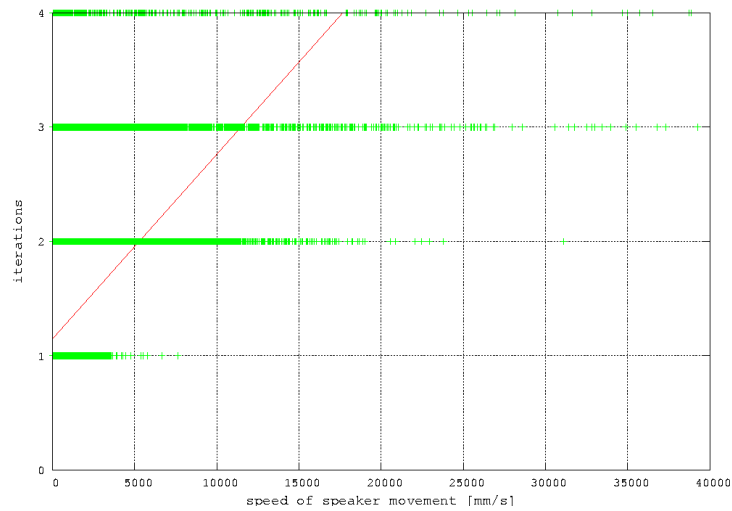


Fig. 8. Number of local iterations vs. speaker velocity.

linearized error criterion (2.8) does not initially match the true criterion (2.4). Hence, several local iterations are required for the position estimate to converge. Five or fewer local iterations were required for convergence in all cases, however, so that the system is still suitable for real time speaker tracking.

Shown in Figure 9 are several images from a seminar recording. The light



Fig. 9. Images from four calibrated video cameras taken during a seminar recording at the Universität Karlsruhe.

square marks the true speaker position obtained from the hand labeled im-

ages. The dark square is the three-dimensional position estimate after back projection to the two-dimensional image plane.

## 6 Conclusions

In this work, we have proposed an algorithm for acoustic source localization based on time delay of arrival 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 position estimates. In our proposed algorithm, the closed-form approximation is eliminated by employing a Kalman filter to directly update the speaker position estimate based on the observed TDOAs. To be more precise, 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, and recorded simultaneously with one or more microphone arrays, as well as four calibrated video cameras. From the video images, the true position of the speaker was extracted and used as “ground truth” for the automatic localization experiments. Our experiments revealed that the proposed algorithm provided source localization superior to the standard spherical and linear intersection techniques. We found that further improvements in localization accuracy could be obtained by adaptively setting a threshold on the PHAT function. Moreover, our algorithm ran in less than real time on an Intel Xeon processor with a 3.0 GHz clock speed.

In other recent work [20], we have extended our technique to include features obtained from calibrated video cameras, as in the work by Strobel *et al* [17]. This can be accomplished in a straightforward manner through the use of a video-based detector to find the speaker’s face in a video image. Thereafter, the difference between the location of the detected face in the image and its predicted position obtained by back projecting from three-dimensional room coordinates to two-dimensional image coordinates serves as the innovation vector. Hence, it is never necessary to triangulate the speaker’s position from several video images as was done in [17]. We need only follow the approach adopted in this work and perform the standard update for a Kalman filter. Such an *incremental* update procedure was investigated by Welch and Bishop [26] for a different set of sensors. Welch [27] also analyzed this incremental scheme in terms of the observability criterion typically encountered in state-space control theory. He found that although the state of the system (i.e., the speaker’s position) is *not* observable based on the information from any single sensor, the state becomes observable when the information from all sensors is combined. To achieve a stable update, it is only necessary to ensure that all sensors are sampled frequently enough.







in Computer Science in spring 2006. Parts of this work were done during a one semester scholarship at Carnegie Mellon University, Pittsburgh, USA.”

**Tobias Gehrig** was born in Karlsruhe, Germany in 1980. He finished high school in 1999 and received an award for extraordinary achievements in natural science from the alumni association. Since October, 2000 he has studied computer science at the University of Karlsruhe. He began working at the Institut für Theoretische Informatik in 2004 as a student assistant involved in data collection and source localization. His student thesis (*Studienarbeit*) was about audio-video source localization using Kalman filters.

**John McDonough** received his Bachelor of Science in 1989 and Master of Science in 1992 from Rensselaer Polytechnic Institute. From January 1993 until August 1997 he worked at the Bolt, Beranek, and Newman Corporation in Cambridge, Massachusetts primarily on large vocabulary speech recognition systems. In September 1997 he began doctoral studies at the Johns Hopkins University in Baltimore, Maryland, which he completed in April 2000. Since January of 2000 he has been employed at the Interactive Systems Laboratories at the Universität Karlsruhe (TH), Germany as a researcher and lecturer.

## References

- [1] M. Omologo, 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, Prentice-Hall, Englewood Cliffs, NJ, 1993.
- [3] M. S. Brandstein, A framework for speech source localization using sensor arrays, Ph.D. thesis, Brown University, Providence, RI (May 1995).
- [4] M. S. Brandstein, J. E. Adcock, H. F. Silverman, A closed-form location estimator for use with room environment microphone arrays, IEEE Trans. Speech Audio Proc. 5 (1) (1997) 45–50.
- [5] Y. T. Chan, K. C. Ho, A simple and efficient estimator for hyperbolic location, IEEE Trans. Signal Proc. 42 (8) (1994) 1905–15.
- [6] H. C. Schau, A. Z. Robinson, Passive source localization employing intersecting spherical surfaces from time-of-arrival differences, IEEE Trans. Acoust. Speech Signal Proc. ASSP-35 (8) (1987) 1223–5.
- [7] J. O. Smith, J. S. Abel, Closed-form least-squares source location estimation from range-difference measurements, IEEE Trans. Acoust. Speech Signal Proc. ASSP-35 (12) (1987) 1661–9.
- [8] T. Dvorkin, S. Gannot, Speaker localization exploiting spatial-temporal information, in: The International Workshop on Acoustic Echo and Noise Control (IWAENC), 2003, pp. 295–298.
- [9] R. Duraiswami, D. Zotkin, 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] D. B. Ward, A. Lehmann, R. C. Williamson, Particle filtering algorithms for tracking an acoustic source in a reverberant environment, IEEE Trans. Speech Audio Proc. 11 (2003) 826–36.
- [11] E. A. Lehmann, D. B. Ward, R. C. Williamson, Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room, in: Proc. ICASSP, Vol. V, 2003, pp. 177–80.
- [12] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, IEEE Trans. Sig. Proc. 50 (2002) 174–188.
- [13] W. R. Hahn, S. A. Tretter, Optimum processing for delay-vector estimation in passive signal arrays, IEEE Trans. Info. Theory IT-19 (1973) 608–614.
- [14] A. H. Jazwinski, Stochastic Processes and Filtering Theory, Academic Press, New York, 1970.

- [15] J. Chen, J. Benesty, Y. A. Huang, Robust time delay estimation exploiting redundancy among multiple microphones, *IEEE Trans. Speech Audio Proc.* 11 (6) (2003) 549–57.
- [16] Y. Huang, J. Benesty, G. Elko, R. M. Mersereau, Real-time passive source localization: A practical linear-correction least-squares approach, *IEEE Trans. Speech Audio Proc.* 9 (8) (2001) 943–956.
- [17] N. Strobel, S. Spors, R. Rabenstein, Joint audio-video signal processing for object localization and tracking, in: M. Brandstein, D. Ward (Eds.), *Microphone Arrays*, Springer Verlag, Heidelberg, Germany, 2001, Ch. 10.
- [18] S. Haykin, *Adaptive Filter Theory*, 4th Edition, Prentice Hall, New York, 2002.
- [19] G. H. Golub, C. F. Van Loan, *Matrix Computations*, 3rd Edition, The Johns Hopkins University Press, Baltimore, 1996.
- [20] T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, J. McDonough, Kalman filters for audio-video source localization, in: *Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, submitted for publication.
- [21] L. Armani, M. Matassoni, M. Omologo, P. Svaizer, Use of a CSP-based voice activity detector for distant-talking ASR, in: *Proc. Eurospeech*, Vol. II, 2003, pp. 501–4.
- [22] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, USA, 1995.
- [23] Z. Zhang, A flexible new technique for camera calibration, *IEEE Trans. Pattern Analysis Machine Intel.* 22 (2000) 1330–1334.
- [24] D. Focken, R. Stiefelhagen, Towards vision-based 3-D people tracking in a smart room, in: *IEEE Int. Conf. Multimodal Interfaces*, 2002.
- [25] J. S. Abel, J. O. Smith, The spherical interpolation method for closed-form passive source localization using range difference measurements, in: *Proc. ICASSP*, 1987.
- [26] G. Welch, G. Bishop, SCAAT: Incremental tracking with incomplete information, in: *Proc. Computer Graphics and Interactive Techniques*, 1997.
- [27] G. F. Welch, SCAAT: Incremental tracking with incomplete information, Ph.D. thesis, University of North Carolina, Chapel Hill, NC (1996).
- [28] A. H. Sayed, T. Kailath, A state-space approach to adaptive RLS filtering, *IEEE Signal Processing Magazine* (1994) 18–60.
- [29] J. McDonough, U. Klee, T. Gehrig, Kalman filtering for time delay of arrival-based source localization, Tech. Rep. 104, Interactive Systems Lab, Universität Karlsruhe (December 2004).