Tracking and Far-Field Speech Recognition for Multiple Simultaneous Speakers

Tobias Gehrig and John McDonough

Institut für Theoretische Informatik
Universität Karlsruhe
Am Fasanengarten 5, 76131 Karlsruhe, Germany
email: {tgehrig, jmcd}@ira.uka.de

Abstract. In prior work, we developed a speaker tracking system based on an extended Kalman filter using time delays of arrival (TDOAs) as acoustic features. While this system functioned well, its utility was limited to scenarios in which a single speaker was to be tracked. In this work, we remove this restriction by generalizing the IEKF, first to a probabilistic data association filter, which incorporates a clutter model for rejection of spurious acoustic events, and then to a joint probabilistic data association filter (JPDAF), which maintains a separate state vector for each active speaker. In a set of experiments conducted on seminar and meeting data, the JPDAF speaker tracking system reduced the multiple object tracking errror from 20.7% to 14.3% with respect to the IEKF system. In a set of automatic speech recognition experiments conducted on the output of a 64 channel microphone array which was beamformed using automatic speaker position estimates, applying the JPDAF tracking system reduced word error rate from 67.3% to 66.0%. Moreover, the word error rate on the beamformed output was 13.0% absolute lower than on a single channel of the array.

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* such as the phase transform (PHAT) [1].

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

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. In prior work [3], we employed an extended Kalman filter to directly update the speaker position estimate based on the observed TDOAs. In particular, the TDOAs comprised the observation associated with an extended Kalman filter whose state corresponded to the speaker position. Hence, the new position estimate came directly from the update formulae associated with the Kalman filter. We tested our algorithm on seminar data involving actual human subjects, and found that our algorithm provided localization performance superior to the standard techniques such as [4].

In other work [5], we enhanced our audio localizer with video information. We proposed an algorithm to incorporate detected face positions in different camera views into the Kalman filter without doing any triangulation. Our algorithm differed from that proposed by Strobel *et al* [6] in that no explicit position estimates were made by the individual sensors. Rather, as in the work of Welch and Bishop [7], the observations of the individual sensors were used to *incrementally* update the state of a Kalman filter. This combined approach yielded a robust source localizer that functioned reliably both for segments wherein the speaker was silent, which would have been detrimental for an audio only tracker, and wherein many faces appear, which would have confused a video only tracker. Our experiments with actual seminar data revealed that the audio-video localizer functioned better than a localizer based solely on audio or solely on video features.

Although the systems described in our prior work functioned well, their utility was limited to scenarios wherein a single subject was to be tracked. In this work, we seek to remove this limitation and develop a system that can track several simultaneous speakers, such as might be required for meeting and small conference scenarios. Our approach is based on two generalizations of the IEKF, namely, the probabistic data association filter (PDAF) and the join probabistic data association filter (JPDAF). Such data association filters have been used extensively in the computer vision field [8], but have seen less widespread use in the field of acoustic person localization and tracking [9]. Compared with the IEKF, these generalizations provide the following advantages:

1. In the PDAF, a "clutter model" is used to model random events, such as door slams, footfalls, etc., that are not associated with any speaker, but can cause spurious peaks in the GCC of a microphone pair, and thus lead to poor tracking performance. Observations assigned with high probability to the clutter model do not affect the estimated position of the active target.

2. In the JPDAF, a unique PDAF is maintained for each active speaker and the peaks in the GCC are probabilistically associated with each of the currently active targets. This association is done jointly for all targets. Moreover, the feasible associations are defined such that a given GCC peak is associated with exactly one active speaker or the clutter model, and a target may be associated with at most one peak for a given microphone pair [10].

Through these extensions, the JPDAF is able to track multiple, simultaneous speakers, which is not possible with the simple IEKF. As we show here, this capacity for tracking multiple active speakers is the primary reason why the JPDAF system provides tracking performance superior to that achieved with the IEKF. It is worth noting that similar work in speaker segmentation based on the output of a source localizer was attempted in [11], but without exploiting the full rigor of the Kalman and data association filters.

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 provides a brief exposition of the extended Kalman, as well as it variants, the IEKF, the PDAF and JPDAF. Section 4 presents the results of our initial experiments comparing the tracking performance of the IEKF and JPDAF. The results of a set of far-field automatic speech recognition experiments based on delay-and-sum beamforming using automatic position estimates from the IEKF and JPDAF are also presented in Section 4.

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

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

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

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

where

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

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

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

where we have defined the row vector

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

Equations (4–5) are the desired linearization.

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

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

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, 12]. Perhaps the most popular method involves *phase transform* (PHAT), a variant of the *generalized cross correlation* (GCC) which can be expressed as

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

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

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

where

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

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

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

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

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

Similarly, defining

$$\boldsymbol{\Sigma} = \operatorname{diag}\left[\sigma_0^2 \; \sigma_1^2 \cdots \; \sigma_{N-1}^2\right] \tag{12}$$

enables (8) to be expressed as

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

While (13) is sufficient to estimate the position of a speaker at any given time instant, it takes no account of past observations, which may also be useful for determining the speaker's current position. This can be achieved, however, by defining a model of the speaker's dynamics, and applying an extended Kalman filter to this nonlinear regression problem.

3 Kalman Filters

Here we briefly review the extended Kalman filter (EKF) and its variations, the PDAF and JPDAF.

3.1 Extended Kalman Filter

Let $\mathbf{x}(t)$ denote the current state of a Kalman filter and $\mathbf{y}(t)$ the current observation. As $\mathbf{x}(t)$ cannot be observed directly, it must be inferred from the time 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,

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

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

where $\mathbf{F}(t+1,t)$ is a known *transition matrix*. The term $\mathbf{C}(t,\mathbf{x}(t))$ is the known *observation functional*, which can represent any arbitrary, nonlinear, time varying mapping from $\mathbf{x}(t)$ to $\mathbf{y}(t)$. In (14–15) 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 $\mathbf{Q}_i(t)$ for $i = 1, 2$.

In the sequel, it will prove useful to define two estimates of the current state: 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 $\mathcal{Y}_t = \{\mathbf{y}(i)\}_{i=0}^{t}$ including the current one. The *predicted observation* is then given by

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

which follows readily from (15). By definition, the *innovation* is the difference

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t|\mathcal{Y}_{t-1}) \tag{17}$$

between actual and predicted observations. Generalizing the classical Kalman filter to the EKF entails linearizing $\mathbf{C}(t,\mathbf{x}(t))$ about the predicted state estimate $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$. Let us denote this linearization as $\mathbf{C}(t)$.

The correlation matrix $\mathbf{R}(t) = \mathcal{E}\left\{\boldsymbol{\alpha}(t)\boldsymbol{\alpha}^T(t)\right\}$ of the innovations sequence can be calculated from [13, §10.3]

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{K}(t,t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \tag{18}$$

where $\mathbf{K}(t,t-1) = \mathcal{E}\left\{\boldsymbol{\epsilon}(t,t-1)\boldsymbol{\epsilon}^T(t,t-1)\right\}$ 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* for the EKF is defined as [13, §10.10]

$$\begin{aligned}\mathbf{G}_F(t) &= \mathbf{F}^{-1}(t+1,t)\,\mathcal{E}\left\{\mathbf{x}(t+1)\boldsymbol{\alpha}^T(t)\right\}\mathbf{R}^{-1}(t)\\ &= \mathbf{K}(t,t-1)\,\mathbf{C}^T(t)\,\mathbf{R}^{-1}(t)\end{aligned}$$

To calculate $\mathbf{G}_F(t)$, we must know $\mathbf{K}(t,t-1)$ in advance. The latter is available from the *Riccati equation*, which can be stated as [13, §10.4]

$$\mathbf{K}(t+1,t) = \mathbf{F}(t+1,t)\mathbf{K}(t)\mathbf{F}^T(t+1,t) + \mathbf{Q}_1(t) \tag{19}$$

$$\mathbf{K}(t) = \left[\mathbf{I} - \mathbf{F}(t,t+1)\mathbf{G}(t)\mathbf{C}(t)\right]\mathbf{K}(t,t-1) \tag{20}$$

where $\mathbf{K}(t) = \mathcal{E}\left\{\boldsymbol{\epsilon}(t)\boldsymbol{\epsilon}^T(t)\right\}$ is the correlation matrix of the *filtered state error*, $\boldsymbol{\epsilon}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t|\mathcal{Y}_t)$.

An update of the state estimate proceeds in two steps: First, the predicted state estimate $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) = \mathbf{F}(t,t-1)\hat{\mathbf{x}}(t-1|\mathcal{Y}_{t-1})$ is formed and used to calculate the innovation $\boldsymbol{\alpha}$(t) as in (16–17), as well as the linearized observation functional $\mathbf{C}(t)$. Then the correction based on the current observation is applied to obtain the filtered state estimate according to [13, §10.4]

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t)\boldsymbol{\alpha}(t) \tag{21}$$

For the sake of simplicity of exposition, we shall base our development on the EKF in the sequel; details of IEKF can be found in [3].

To construct a speaker tracking system, we need only associate the observation $\mathbf{y}(t)$ with the TDOA estimate $\boldsymbol{\tau}(t)$ for the audio features. The observation functional $\mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))$ required for the Kalman filter is formulated as a mapping from the speaker position $\mathbf{x}$ to a vector of time delays $\boldsymbol{\tau}(t)$, as in (2). The TDOA error covariance matrix $\boldsymbol{\Sigma}$ in (12) can be associated 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 finding his position at a single time instant.

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 set $\mathbf{Q}_1(t) = \sigma^2 T^2 \mathbf{I}$ where T is the time since the last state update. Note that T can vary given that a speaker does not always speak, and no update is possible when the speaker is silent.

3.2 Probabilistic Data Association Filter

The PDAF is a generalization of the Kalman filter wherein the Gaussian probability density function (pdf) associated with the location of the speaker or *target* is supplemented with a pdf for random false alarms or *clutter* [10, §6.4]. Through the inclusion of the clutter model, the PDAF is able to make use of several observations $\{\mathbf{y}_i(t)\}_{i=1}^{m_t}$ for each time instant, where m_t is the total number of observations for time t. Each observation can then be attributed either to the target itself, or to the background model. Let us define the *association events*

$$\theta_i(t) = \{\mathbf{y}_i(t) \text{ is the target observation at time } t\} \tag{22}$$

$$\theta_0(t) = \{\text{all observations are clutter}\} \tag{23}$$

and the posterior probability of each event

$$\beta_i(t) = P\{\theta_i(t)|\mathcal{Y}_t\} \tag{24}$$

As the events $\{\theta_i(t)\}_{i=0}^{m_t}$ are exhaustive and mutually exclusive, we have $\sum_{i=0}^{m_t} \beta_i(t) = 1$. Moreover, invoking the total probability theorem, the filtered state estimate can be expressed as

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \sum_{i=0}^{m_t} \hat{\mathbf{x}}_i(t|\mathcal{Y}_t)\beta_i(t)$$

where $\hat{\mathbf{x}}_i(t|\mathcal{Y}_t) = E\{\mathbf{x}(t)|\theta_i(t), \mathcal{Y}_t\}$ is the updated state estimate conditioned on $\theta_i(t)$. It can be readily shown that this state estimate can be calculated as

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

where

$$\boldsymbol{\alpha}_i(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{y}_i(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \tag{25}$$

is the innovation for observation $\mathbf{y}_i(t)$. The combined update is

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\, \boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \tag{26}$$

where the *combined innovation* is

$$\boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \sum_{i=1}^{m_t} \boldsymbol{\alpha}_i(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\, \beta_i(t) \tag{27}$$

The Riccati equation (19–20) must be suitably modified to account for the additional uncertainty associated with the multiple innovations $\{\boldsymbol{\alpha}_i(t)\}$, as well as the possibility of the null event $\theta_0(t)$; see Bar-Shalom and Fortmann [10, §6.4] for details.

3.3 Joint Probabilistic Data Association Filter

The JPDAF is an extension of the PDAF to the case of multiple targets. Consider the set $\mathbf{Y}(t) = \{\mathbf{y}_i(t)\}_{i=1}^{m_t}$ of all observations occuring at time instant t and let $\mathcal{Y}_{t-1} = \{\mathbf{Y}(i)\}_{i=0}^{t-1}$ denote the set of all past observations. The first step in the JPDA algorithm is the evaluation of the conditional probabilities of the *joint association events*

$$\boldsymbol{\theta} = \bigcap_{i=1}^{m_t} \theta_{ik_i}$$

where the atomic events are defined as

$$\theta_{ik} = \{\text{observation } i \text{ originated from target } k\}$$

for all $i = 1, \ldots, m_t$; $t = 0, 1, \ldots, T$. Here, k_i denotes the index of the target to which the i-th observation is associated in the event currently under consideration. A *feasible event* is defined as an event wherein

1. An observation has exactly one source, which can be the clutter model;
2. No more than one observation can originate from any target.

In the acoustic person tracking application where the observations are peaks in the cross correlation function for pairs of microphones, the second point must be interpreted as referring to the observations for any given pair of microphones. Applying Bayes' rule, the conditional probability of $\boldsymbol{\theta}(t)$ can be expressed as

$$P\{\boldsymbol{\theta}(t)|\mathcal{Y}_t\} = \frac{P\{\mathbf{Y}(t)|\boldsymbol{\theta}(t), \mathcal{Y}_{t-1}\} P\{\boldsymbol{\theta}(t)\}}{P\{\mathbf{Y}(t)|\mathcal{Y}_{t-1}\}} \tag{28}$$

where the marginal probability $P\{\mathbf{Y}(t)|\mathcal{Y}_{t-1}\}$ is computed by summing the joint probability in the numerator of (28) over all possible $\boldsymbol{\theta}(t)$. The conditional probability of $\mathbf{Y}(t)$ required in (28) can be calculated from

$$P\{\mathbf{Y}(t)|\boldsymbol{\theta}(t), \mathcal{Y}_{t-1}\} = \prod_{i=1}^{m_t} p(\mathbf{y}_i(t)|\theta_{ik_i}(t), \mathcal{Y}_{t-1}) \tag{29}$$

The individual probabilities on the right side of (29) can be easily evaluated given the fundamental assumption of the JPDAF, namely,

$$\mathbf{y}_i(t) \backsim \mathcal{N}(\hat{\mathbf{y}}_{k_i}(t|\mathcal{Y}_{t-1}), \mathbf{R}_{k_i}(t))$$

where $\hat{\mathbf{y}}_{k_i}(t|\mathcal{Y}_{t-1})$ and $\mathbf{R}_{k_i}(t)$ are, respectively, the predicted observation (16) and innovation covariance matrix (18) for target k_i. The prior probability $P\{\boldsymbol{\theta}(t)\}$ in (28) can be readily evaluated through combinatorial arguments [10, §9.3]. Once the posterior probabilities of the joint events $\{\boldsymbol{\theta}(t)\}$ have been evaluated for all targets together, the state update for each target can be made separately according to (25–27).

Once the posterior probabilities of the joint events $\{\boldsymbol{\theta}(t)\}$ have been evaluated for all targets together, the state update for each target can be made separately according to (25–27). For any given target, it is only necessary to marginalize out the effect of all other targets to obtain the required posterior probabilities $\{\beta_i(t)\}$.

As the JPDAF can track multiple targets, it was necessary to formulate rules for deciding when a new target should be created, when two targets should be merged and when a target should be deleted. A new target was always created as soon as a measurement could not be associated with any existing target. But if the time to initialize the filter exceeded a time threshold, the newly created target was immediately deleted. The initialization time of the filter is defined as the time required until the variance of each dimension of $\boldsymbol{\epsilon}(t, t-1)$ in (**??**) fell below a given threshold. Normally this initialization time is relatively short for a target that emits sufficient measurements and long for spurious noises. To merge two or more targets, a list was maintained with the timestamp when the two targets became closer than a given distance. If, after some allowed interval of overlap, the two targets did not move apart, then the target with the larger $|K(t, t-1)|$ was deleted. In all cases, targets were deleted if their position estimate had not been updated for a given length of time. To detect the active sound source, we simply used the target with the smallest error covariance matrix, since an active sound source should emit enough measurements so that the covariance decreases and others that are inactive should increase at the same time.

4 Experiments

The test set used to evaluate the algorithms proposed here contains approximately three hours of audio and video data recorded during 18 seminars held by students and faculty at University of Karlsruhe (UKA) in Karlsruhe, Germany. An additional hour of test data was recorded at Athens Information Technology in Athens, Greece, IBM at Yorktown Heights, New York, USA, Instituto Trentino di Cultura in Trento, Italy, and Universitat Politecnica de Catalunya in Barcelona, Spain. These recordings were made in connection with the European Union integrated project CHIL, *Computers in the Human Interaction Loop*. In the sequel, we describe out speaker tracking and STT experiments.

4.1 Speaker Tracking Experiments

Prior to the start of the recordings, four video cameras in the corners of the room had been calibrated with the technique of Zhang [14]. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every second. Using these hand-marked labels, the true position of the speaker's head in the three dimensions was calculated using the technique described in [14]. These "ground truth" speaker positions are accurate to within 10 cm. For the speaker tracking experiments described here, the seminars were recorded with several four-element T-shaped arrays. A precise description of the sensor and room configuration at UKA is provided in [3].

Tracking performance was evaluated only on those parts of the seminars where only a single speaker was active. For these parts, it was determined whether the error between the ground truth and the estimated position is less 50 cm. Any instance where the error exceeded this threshold was treated as a *false positive* (FP) and was not considered when calculating the *multiple object tracking precision* (MOTP), which is defined as the average horizontal position error. If no estimate fell within 50 cm of the ground truth, it was treated as a *miss*. Letting N_{fp} and N_{m}, respectively, denote the total number of false positives and misses, the *multiple object tracking error* (MOTE) is defined as $(N_{\text{fp}} + N_{\text{m}})/N$ where N is the total number of ground truth positions. We evaluated performance separately for the portion of the seminar during which only the lecturer spoke, and that during which the lecturer interacted with the audience. Shown in Table 1 are the results of our experiments. These results clearly

Table 1. Single and multi-speaker tracking performance of the IEKF and JPDAF.

Filter	Test Set	MOTP (cm)	% Miss	% FP	% MOTE
IEKF	lecture	11.4	8.32	8.30	16.6
IEKF	interactive	18.0	28.75	28.75	57.5
IEKF	complete	12.1	10.37	10.35	20.7
JPDAF	lecture	11.6	5.81	5.78	11.6
JPDAF	interactive	17.7	19.60	19.60	39.2
JPDAF	complete	12.3	7.19	7.16	14.3

show that the JPDAF provided better tracking performance for both the lecture and interactive portions of the seminar. As one might expect, the reduction in MOTE was largest for the interactive portion, where multiple speakers were often simultaneously active.

4.2 STT Experiments

For the purpose of beamforming and STT experiments, the seminars were also recorded with a 64 channel Mark III microphone array developed at the US Na-

tional Institute of Standards and Technologies (NIST); see [3] for the placement of the various sensors. The training procedures and data sets used for training the acoustic and language model components of the STT systems used for the experiments undertaken here are described in Fügen *et al* [15], as are the automatic speech segmentation and speaker clustering procedures. Our STT experiments were conducted on two sets of data: the first was the development data distributed to all participants in the NIST RT06 speech-to-text evaluation, which consisted of approximately 3.5 hours of CHIL seminar data recorded with the T-arrays as well as the Mark III. The second set was the NIST RT06 evaluation data, which was similar in nature and duration to the development data. Experiments on the evaluation data, however, were "blind" in that no transcriptions were available for system development. Rather, the raw STT hypotheses were submitted to NIST as part of the RT06 evaluation. NIST then scored the hypotheses against undisclosed references, tabulated the scores, and distributed the results to all RT06 participants.

Shown in Table 2 are the results of our STT experiments, wherein we compared word error rates (WERs) for a single channel of the Mark III, as well as the beamformed output of the entire array using automatic position estimates based on the IEKF and the JPDAF. Clearly, both beamformed outputs provided word error rates more than 12% absolute lower than the single channel. Moreover, it is apparent that beamforming based on the position estimates returned by the JPDAF provides significantly better STT performance than beamforming on the IEKF position estimates.

Table 2. STT performance for single channel and beamformed array output using IEKF and JPDAF position estimates.

	% Word Error Rate		
Test Set	Single Channel	IEKF	JPDAF
RT06 Dev	61.8	49.4	48.8
RT06 Eval	N/A	67.3	66.0

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. U. Klee, T. Gehrig, and J. McDonough, "Kalman filters for time delay of arrival-based source localization," *Journal of Advanced Signal Processing, Special Issue on Multi-Channel Speech Processing*, to appear.
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. T. Gehrig, K. Nickel, H. K. Ekenel, U. Klee, and J. McDonough, "Kalman filters for audio-video source localization," in *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 2005.
6. 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.
7. G. Welch and G. Bishop, "SCAAT: Incremental tracking with incomplete information," in *Proc. Computer Graphics and Interactive Techniques*, August 1997.
8. G. Gennari and G. D. Hager, "Probabilistic data association methods in the visual tracking of groups," in *Proc. CVPR*, 2004, pp. 1063–1069.
9. D. Bechler, "Akustische Sprecherlokalisation mit Hilfe eines Mikrofonarrays," Ph.D. dissertation, Universität Karlsruhe, Karlsruhe, Germany, 2006.
10. Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. San Diego: Academic Press, 1988.
11. J. Ajmera, G. Lathoud, and I. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP*, 2004, pp. I–605–8.
12. 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.
13. S. Haykin, *Adaptive Filter Theory*, 4th ed. New York: Prentice Hall, 2002.
14. Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Analysis Machine Intel.*, vol. 22, pp. 1330–1334, 2000.
15. C. Fügen, M. Wölfel, J. McDonough, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani, "Advances in lecture recognition: The ISL RT-06s evaluation system," in *Proc. Interspeech*, submitted for publication, 2006.