Tracking and Beamforming for Multiple Simultaneous Speakers with Probabilistic Data Association Filters

Tobias Gehrig, Ulrich Klee, John McDonough, Shajith Ikbal, Matthias Wölfel, Christian Fügen

Institut für Theoretische Informatik
Universität Karlsruhe
Am Fasanengarten 5, D-76131 Karlsruhe, Germany
email: {tgehrig, klee, jmcd, shajith, wolfel, fuegen}@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 tracking system reduced the multiobject tracking error 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 theJPDAF 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.

Index Terms: acoustic source localization, Kalman filter, person tracking, far-field speech recognition, microphone arrays

1. Introduction

In prior work on acoustic speaker tracking [1], we used an extended Kalman filter to directly update the speaker position estimate based on a set of observed time delays of arrival (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 pair. Let $\hat{\mathbf{x}}(\mathbf{t})$ denote the position of the speaker in $\mathbb{R}^3$. Then the TDOA between the pair of microphones is

$$T_i(x) = T(m_{i1}, m_{i2}, x) = \frac{1}{s} (d_{i1} - d_{i2})$$

(1)

where $d_{ij} = ||x - m_{ij}||$ is the distance from the source to microphone $m_{ij}$ and $s$ is the speed of sound.

Source localization based on a maximum likelihood (ML) criterion [6] proceeds by minimizing the error function

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

(2)

where $\hat{\tau}_i$ is the observed TDOA for the $i$-th microphone pair, $\sigma_i^2$ is the error covariance associated with this observation, and $N$ is the number of unique microphone pairs. The TDOAs can be estimated...
with a variety of well-known techniques [7, 8]. Perhaps the most 
popular method involves phase transform (PHAT), a variant of the 
generalized cross correlation (GCC) which can be expressed as
\[
R_{12}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X_1(e^{j\omega t})X_2^*(e^{-j\omega t}) e^{j\omega \tau} d\omega
\]  
(3)

The TDOA estimate is then given by \( \hat{\tau}, = \max, R_{12}(\tau) \).

In earlier work [1], we showed how the nonlinear least squares 
criterion (2) can be linearized about the current position estimate. 
In particular, we defined
\[
\Sigma = \text{diag}[\sigma_0^2 \; \sigma_1^2 \; \cdots \; \sigma_{n-1}^2]
\]
and derived the linearized criterion
\[
e(\chi; t) = [\tau(t) - C(t)\chi]^T \Sigma^{-1} [\hat{\tau}(t) - C(t)\chi]
\]
where \( \hat{\tau}(t) \) and \( C(t) \) are defined in [1].

While (5) is sufficient to estimate the position of a speaker at 
y 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 \( \chi(t) \) denote the current state of a Kalman filter and \( y(t) \) the 
current observation. As \( \chi(t) \) cannot be observed directly, it must 
be inferred from the time series \( \{y(t)\} \); 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,
\[
\chi(t + 1) = F(t + 1, t)\chi(t) + \nu_1(t)
\]
\[
y(t) = C(t, \chi(t)) + \nu_2(t)
\]
(6) (7)

where \( F(t + 1, t) \) is a known transition matrix. The term 
\( C(t, \chi(t)) \) is the known observation functional, which can 
represent any arbitrary, nonlinear, time varying mapping from \( \chi(t) \) 
to \( y(t) \). In (6–7) the process and observation noise terms are 
denoted by \( \nu_1(t) \) and \( \nu_2(t) \) respectively. These noise terms are by 
assumption zero mean, white Gaussian random vector processes 
with covariance matrices \( Q_i(t) \) for \( i = 1, 2 \).

In the sequel, it will prove useful to define two estimates of the 
current state: Let \( \hat{\chi}(t|Y_{t-1}) \) denote the predicted state estimate 
of \( \chi(t) \) obtained from all observations \( Y_{t-1} = \{y(i)\}_{i=0}^{t-1} \) 
up to time \( t - 1 \). The filtered state estimate \( \hat{\chi}(t|Y_{t}) \), on the other hand, is 
based on all observations \( Y_{t} = \{y(i)\}_{i=0}^{t} \), including the current 
one. The predicted observation is then given by
\[
\hat{y}(t|Y_{t-1}) = C(t, \chi(t|Y_{t-1})))
\]
(8)

which follows readily from (7). By definition, the innovation is the difference
\[
\alpha(t) = y(t) - \hat{y}(t|Y_{t-1})
\]
(9)
between actual and predicted observations. Generalizing the classical 
Kalman filter to the EKF entails linearizing \( C(t, \chi(t)) \) about 
the predicted state estimate \( \hat{\chi}(t|Y_{t-1}) \). Let us denote this linearization as \( \mathcal{C}(t) \).

The correlation matrix \( \mathbf{R}(t) = \mathcal{E}[(\alpha(t)\alpha^T(t))] \) of the innova-
tions sequence can be calculated from [9, §10.3]
\[
\mathbf{R}(t) = C(t)K(t, t-1)C^T(t) + Q_2(t)
\]
(10)

where \( K(t, t-1) = \mathcal{E}[(\epsilon(t, t-1)^T\epsilon(t, t-1))] \) is the correlation matrix of the predicted state error, \( \epsilon(t, t-1) = \chi(t) - \hat{\chi}(t|Y_{t-1}) \).

The Kalman gain for the EKF is defined as [9, §10.10]
\[
G_F(t) = F^{-1}(t+1, t)\mathcal{E}\{x(t+1)\alpha^T(t)\}R^{-1}(t)
\]
\[= K(t, t-1)C^T(t)R^{-1}(t) \]
To calculate \( G_F(t) \), we must know \( K(t, t-1) \) in advance. The 
latter is available from the Riccati equation, which can be stated as [9, §10.4]
\[
K(t + 1, t) = F(t + 1, t)K(t)F^T(t + 1, t) + Q_1(t)
\]
\[= [I - F(t, t + 1)C(t)C^T(t)]K(t, t-1) \]
(11) (12)

where \( K(t) = \mathcal{E}[(\epsilon(t)\epsilon^T(t))] \) is the correlation matrix of the 
filtered state error, \( \epsilon(t) = \chi(t) - \hat{\chi}(t|Y_{t}) \). An update of the state estimate proceeds in two steps: First, the 
predicted state estimate \( \hat{\chi}(t|Y_{t-1}) = F(t, t-1)\hat{\chi}(t-1|Y_{t-1}) \) 
is formed and used to calculate the innovation \( \alpha(t) \) as in (8–9), 
as well as the linearized observation functional \( C(t) \). Then the 
correction based on the current observation is applied to obtain the 
filtered state estimate according to [9, §10.4]
\[
\hat{\chi}(t|Y_{t}) = \hat{\chi}(t|Y_{t-1}) + G_F(t)\alpha(t)
\]
(13)

For the sake of simplicity of exposition, we shall base our develop-
ment on the EKF in the sequel; details of IIEKF can be found in [1].

To construct a speaker tracking system, we need only associate 
the observation \( y(t) \) with the TDOA estimate \( \tau(t) \) for the audio 
features. The observation functional \( C(t, \hat{\chi}(t|Y_{t-1})) \) required for 
the Kalman filter is formulated as a mapping from the speaker po-


tion \( \chi \) to a vector of time delays \( \tau(t) \), as in (1). The TDOA error 
covariance matrix \( \Sigma \) in (4) can be associated with the observation 
noise covariance \( Q_2(t) \). 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 \( \nu_1(t) \). The transition matrix is 
then \( F(t + 1, t) = I \). Assuming the process noise components 
in the three directions are statistically independent, we can set 
\( Q_1(t) = \sigma^2 T^2 I \) 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 loca-
tion 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 
\( \{y(i)\}_{i=0}^{t} \) for each time instant, where \( m_i \) 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

$$\theta_i(t) = \{ x_i(t) \text{ is the target observation at time } t \} \quad (14)$$
$$\theta_0(t) = \{ \text{all observations are clutter} \} \quad (15)$$

and the posterior probability of each event

$$\beta_i(t) = P[\theta_i(t)|y_t] \quad (16)$$

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

$$\hat{x}(t|y_t) = \sum_{i=0}^{m_t} \hat{x}_i(t|y_t) \beta_i(t)$$

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

$$\hat{x}_i(t|y_t) = \hat{x}(t|y_t) + G_F(t, \hat{x}(t|y_{t-1})) \alpha_i(t, \hat{x}(t|y_{t-1}))$$

where

$$\alpha_i(t, \hat{x}(t|y_{t-1})) = y_i(t) - C(t, \hat{x}(t|y_{t-1}))$$

is the innovation for observation $y_i(t)$. The combined update is

$$\hat{x}(t|y_t) = \hat{x}(t|y_{t-1}) + G_F(t, \hat{x}(t|y_{t-1})) \alpha(t, \hat{x}(t|y_{t-1}))$$

where the combined innovation is

$$\alpha(t, \hat{x}(t|y_{t-1})) = \sum_{i=1}^{m_t} \alpha_i(t, \hat{x}(t|y_{t-1})) \beta_i(t)$$

(19)

The Riccati equation (11–12) must be suitably modified to account for the additional uncertainty associated with the multiple innovations $\{\alpha_i(t)\}$, 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 $Y(t) = \{ y_i(t) \}_{i=0}^{m_t}$ of all observations occurring at time instant $t$ and let $Y_{t-1} = \{ Y(i) \}_{i=0}^{m_t - 1}$ 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

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

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 $\theta(t)$ can be expressed as

$$P[\theta(t)|y_t] = \frac{P[Y(t)|\theta(t), y_t] \cdot P[\theta(t)]}{P(Y(t)|y_t)} \quad (20)$$

where the marginal probability $P[Y(t)|\theta(t), y_t]$ is computed by summing the joint probability in the numerator of (20) over all possible $\theta(t)$. The conditional probability of $Y(t)$ required in (20) can be calculated from

$$P[Y(t)|\theta(t), y_{t-1}] = \prod_{i=1}^{m_t} p(y_i(t)|\theta_{ik}(t), y_{t-1})$$

(21)

The individual probabilities on the right side of (21) can be easily evaluated given the fundamental assumption of the JPDAF, namely,

$$y_i(t) \sim N(y_{ik}(t), R_{ik}(t))$$

where $y_{ik}(t|y_{t-1})$ and $R_{ik}(t)$ are, respectively, the predicted observation (8) and innovation covariance matrix (10) for target $k_i$. The prior probability $P[\theta(t)]$ in (20) can be readily evaluated through combinatorial arguments [10, §9.3]. Once the posterior probabilities of the joint events $\{\theta(t)\}$ have been evaluated for all targets together, the state update for each target can be made separately according to (17–19).

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. Details of the heuristics used for this purpose can be found in [11].

### 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 Politècnica 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 [12]. 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 [12]. 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 [1].

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 than 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 $N_0$ and $N_m$, respectively, denote the total number of false positives and misses, the multiple object tracking error (MOTE) is defined as $(N_0 + N_m)/N$ 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 4.1 are the results of our experiments. These results clearly 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 National Institute of Standards and Technologies (NIST); see [1] 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. [13], 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 development 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 4.2 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.

### 5. Acknowledgements

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

### 6. References


