Tracking Multiple Simultaneous Speakers with Probabilistic Data Association Filters

Tobias Gehrig and John McDonough

Institut für Theoretische Informatik
Universität Karlsruhe

May 1, 2006

Time Delay of Arrival

- Consider the i-th pair of microphones in an array, with sensor positions $\mathbf{m}_{i1}$ and $\mathbf{m}_{i2}$.
- The *time delay of arrival* (TDOA) between the pair of microphones is defined as

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

where $\mathbf{x}$ is the position of the source and s is the speed of sound.

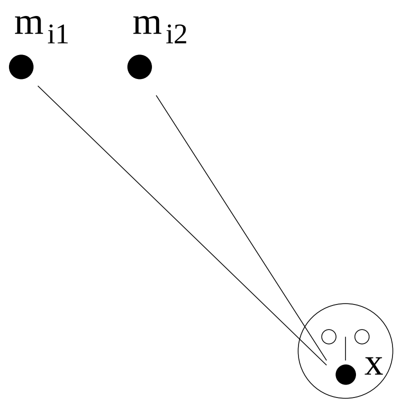


Figure 1: Time delay of arrival.

Source Localization

- Source localization based on a maximum likelihood (ML) criterion minimizes the error function

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

  where $\hat{\tau}_i$ is the observed TDOA for the i-th microphone pair, σ_i^2 is the error covariance associated with this observation, and N is the number of unique microphone pairs.

- TDOAs can be estimated with a variety of well-known techniques such as the *generalized cross correlation* (GCC),

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

- The TDOA estimate is then given by $\hat{\tau}_i = \max_\tau R_{12}(\tau)$.

Source Localization

- The nonlinear least squares criterion (1) can be linearized about the current position estimate as

$$\mathbf{\Sigma} = \text{diag}\left[\sigma_0^2 \quad \sigma_1^2 \quad \cdots \quad \sigma_{N-1}^2\right] \tag{3}$$

- The linerized least squares metric then becomes

$$\epsilon(\mathbf{x};t) = \left[\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}\right]^T \mathbf{\Sigma}^{-1} \left[\bar{\boldsymbol{\tau}}(t) - \mathbf{C}(t)\mathbf{x}\right] \tag{4}$$

where $\bar{\boldsymbol{\tau}}(t)$ and $\mathbf{C}(t)$ are defined in Klee (2005).

- The last equation is very amenable to implementation as a Kalman filter.

Extended Kalman Filter (EKF)

- Let $\mathbf{x}(t)$ denote the current state of a Kalman filter and $\mathbf{y}(t)$ the current observation.
- The operation of the Kalman filter is governed by the *process* and an *observation* equations,

$$\mathbf{x}(t+1) = \mathbf{F}(t+1,t)\,\mathbf{x}(t) + \boldsymbol{\nu}_1(t) \tag{5}$$

$$\mathbf{y}(t) = \mathbf{C}(t,\mathbf{x}(t)) + \boldsymbol{\nu}_2(t) \tag{6}$$

- $\mathbf{F}(t+1,t)$ is a known *transition matrix*.
- $\mathbf{C}(t,\mathbf{x}(t))$ is a known nonlinear, time varying *observation functional*.
- The *process* $\boldsymbol{\nu}_1(t)$ and *observation noise* $\boldsymbol{\nu}_2(t)$ terms are zero mean, white Gaussian random vector processes with covariance matrices $\mathbf{Q}_i(t)$ for $i = 1, 2$.

Innovations

- Define two estimates of the current state:
  - $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$ denotes 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 $t-1$.
  - $\hat{\mathbf{x}}(t|\mathcal{Y}_t)$ denotes the *filtered state estimate* based on all observations $\mathcal{Y}_t = \{\mathbf{y}(i)\}_{i=0}^{t}$ including the current one.
- The *predicted observation* is then given by

$$\hat{\mathbf{y}}(t|\mathcal{Y}_{t-1}) = \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \tag{7}$$

- The *innovation* is the difference

$$\boldsymbol{\alpha}(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t|\mathcal{Y}_{t-1}) \tag{8}$$

  between actual and predicted observations.
- The EKF requires the linearization of $\mathbf{C}(t, \mathbf{x}(t))$ about the predicted state estimate $\hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})$, which we will denote as $\mathbf{C}(t)$.

Kalman Gain

- The correlation matrix of the innovations sequence

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

can be calculated from

$$\mathbf{R}(t) = \mathbf{C}(t)\mathbf{K}(t, t-1)\mathbf{C}^T(t) + \mathbf{Q}_2(t) \tag{9}$$

- Here

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

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

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

Riccati Equation

- To calculate $\mathbf{G}_F(t)$, we must know $\mathbf{K}(t, t-1)$ in advance.
- $\mathbf{K}(t, t-1)$ can be calculated from the *Riccati equation*,

$$\mathbf{K}(t+1, t) = \mathbf{F}(t+1, t)\mathbf{K}(t)\mathbf{F}^T(t+1, t) + \mathbf{Q}_1(t) \tag{10}$$

$$\mathbf{K}(t) = [\mathbf{I} - \mathbf{F}(t, t+1)\mathbf{G}(t)\mathbf{C}(t)]\,\mathbf{K}(t, t-1) \tag{11}$$

- Here

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

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

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

EKF Schematic

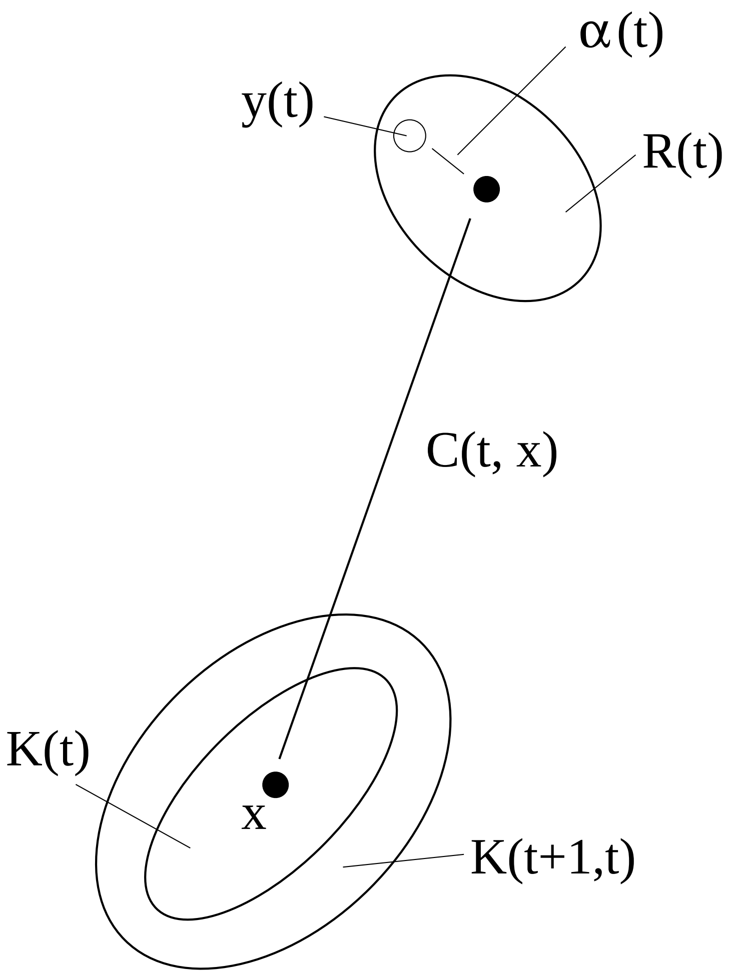


Figure 2: Schematic of the extended Kalman filter.

State Update

An update of the state estimate proceeds in two steps:

1. 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 $\boldsymbol{\alpha}(t)$, $\mathbf{C}(t)$ and $\mathbf{G}_F(t)$.

2. Then the correction based on the current observation is applied to obtain the filtered state estimate,

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t)\boldsymbol{\alpha}(t) \tag{12}$$

Probabilistic Data Association Filter

- The probabilistic data association Filter (PDAF) augments the target pdf with a *clutter* model.
- Define the *association events*

$$\theta_i(t) = \{\mathbf{y}_i(t) \text{ is the target observation at time } t\} \quad (13)$$

$$\theta_0(t) = \{\text{all observations are clutter}\} \quad (14)$$

and the posterior probability of each association event $\beta_i(t) = P\{\theta_i(t)|\mathcal{Y}_t\}$

- The conditional innovation is then

$$\boldsymbol{\alpha}_i(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \mathbf{y}_i(t) - \mathbf{C}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (15)$$

- The combined update is

$$\hat{\mathbf{x}}(t|\mathcal{Y}_t) = \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}) + \mathbf{G}_F(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\, \boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) \quad (16)$$

where the *combined innovation* is

$$\boldsymbol{\alpha}(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1})) = \sum_{i=1}^{m_t} \boldsymbol{\alpha}_i(t, \hat{\mathbf{x}}(t|\mathcal{Y}_{t-1}))\, \beta_i(t) \quad (17)$$

PDAF Schematic

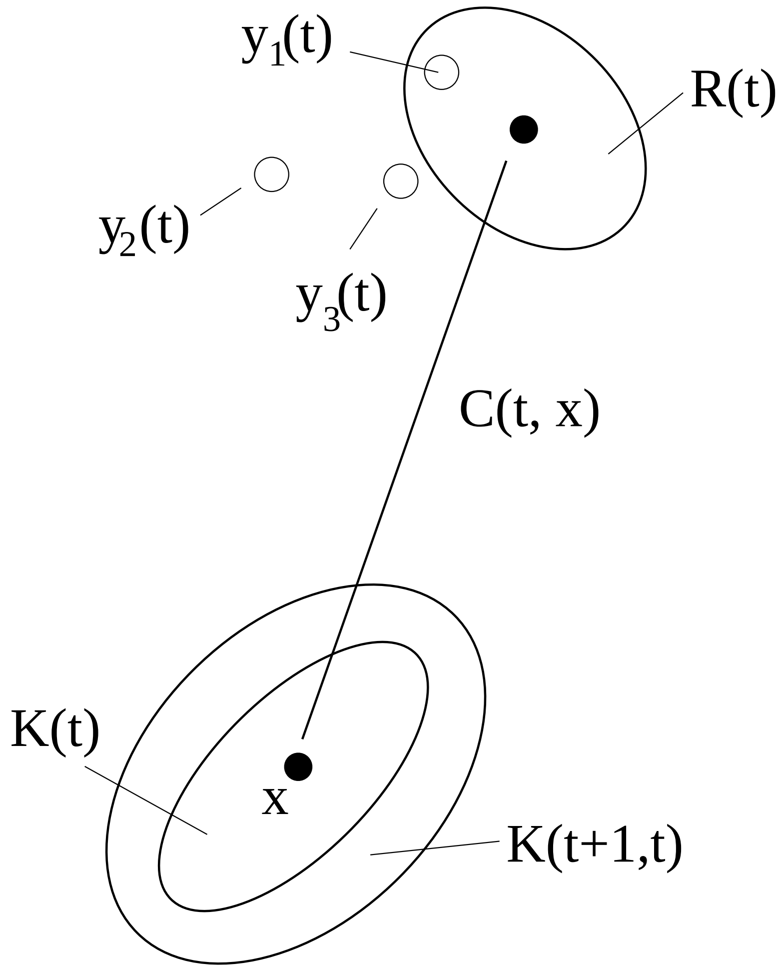


Figure 3: Schematic of the probabilistic data association filter.

Joint Probabilistic Data Association Filter

- The JPDAF algorithm defines 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

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

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

JPDAF Schematic

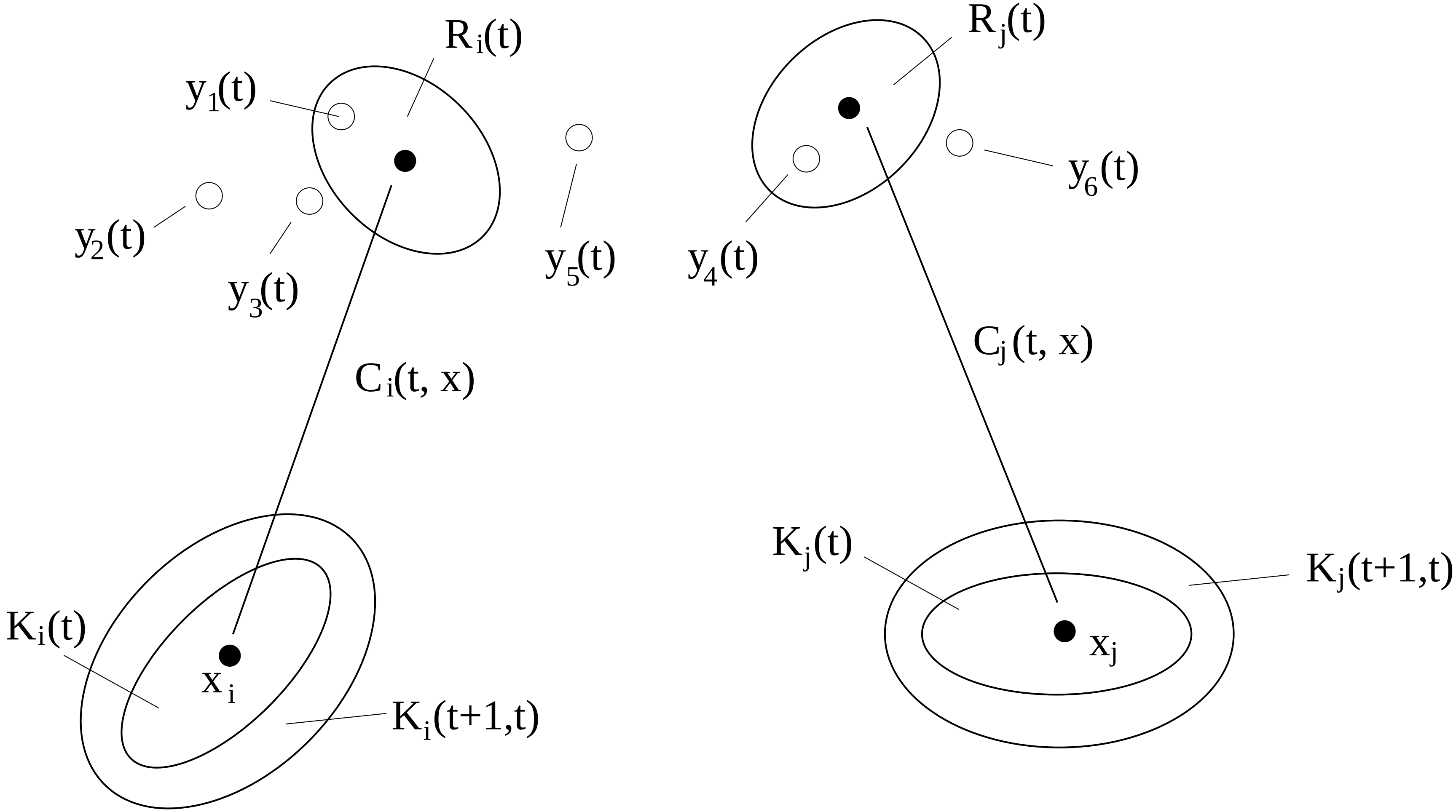


Figure 4: Schematic of the joint probabilistic data association filter.

Speaker Tracking Metrics

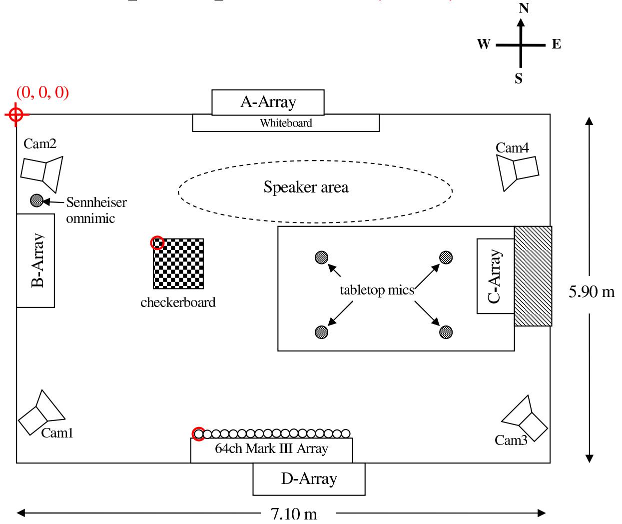
- A threshold of 50 cm between the ground truth and the estimated position was defined.
- Any instance where the error exceeded this threshold was treated as a *false positive* (FP) and was not considered when calculating the *multiple object tracking precision* (MOTP), which is defined as the average 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

$$\text{MOTE} = \frac{N_{\text{fp}} + N_{\text{m}}}{N}$$

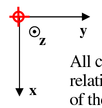
where N is the total number of ground truth positions.

Sensor Configuration at the University of Karlsruhe

ISL_SEMINAR_2004 ROOM SETUP (UPDATE)



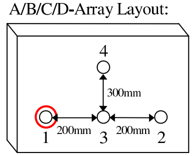
	x	y	z
Checkerboard 2004_11	2130	3260	732
Checkerboard 2004_06/07/08	2000	3110	730
Mark III	5665	2900	1710
Array A1	105	3060	2370
Array B1	2150	105	2290
Array C1	2700	6210	2190
Array D1	5795	4280	2400



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:



Speaker Tracking Results

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.

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

Table 1: Speaker tracking performance for IEKF and JPDAF systems.

Far Field Speech-to-Text Results

For the purpose of beamforming and STT experiments a 64 channel Mark III microphone array developed at the US National Institute of Standards and Technologies (NIST) was used.

	% 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

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

Conclusions and Future Work

- We have improved our single-person tracker to handle multiple simultaneous speakers through the generalization of the iterated extended Kalman filter (IKEF) to a joint probabilistic data association filter (JPDAF).
- On the 2006 CLEAR development data, this generalization reduced the multiple object tracking error (MOTE) from 20.7% to 14.3%.
- Using the new tracking system for beamforming followed by STT reduced word error rate from 67.3% to 66.0 % on the RT06 evaluation set.
- The beamformed signal of the 64 channel Mark III provided a 13.0% absolute reduction in WER with respect to a single channel of the Mark III.
- In future we will use our multiple speaker tracker in a multiple stream STT system.
- Our goal is to create an accurate, reliable system for speaker attributed STT.