Robust 3D Localization and Tracking of Sound Sources Using Beamforming and Particle Filtering

Jean-Marc Valin, François Michaud, Jean Rouat

17/5/2006

CeNTIE is supported by the Australian Government through the Advanced Networks Program (ANP) of the Department of Communications, Information Technology and the Arts and the CSIRO ICT Centre

Context

- Application: tracking speakers in a video-conferencing environment with a microphone array
- Camera not located near the microphones (parallax problem)
  - Distance estimation is required
- Tracking of multiple sources in 3 dimensions in a noisy, reverberant environment

Microphone Array Sound Source Localization and Tracking

- Spatial cues
  - Intensity cues
  - **Phase (delay) cues**
- Microphone array techniques
  - TDOA estimation followed by location estimation
  - Subspace methods (MUSIC, ESPRIT, ...)
  - **Direct search (steered beamformer)**
- Tracking algorithms
  - Kalman filtering
  - **Particle filtering (sequential Monte Carlo estimation)**

Steered Beamformer

- Delay-and-sum beamformer

$$y(n_t) = \sum_{n=0}^{N-1} x_n \left(n_t - \tau_n\right)$$

- Maximize output energy

$$E = \sum_{n_t=0}^{L-1} \left[y(n_t)\right]^2$$

- Frequency-domain computation

$$E = K + 2 \sum_{m_1=0}^{M-1} \sum_{m_2=0}^{m_1-1} R_{x_{m_1},x_{m_2}} \left(\tau_{m_1} - \tau_{m_2}\right)$$

$$R_{ij}(\tau) \approx \sum_{k=0}^{L-1} X_i(k) X_j(k)^* e^{\jmath 2\pi k\tau/L}$$

Spectral Weighting

- Standard cross-correlation has wide peaks
- PHAse Transform (PHAT) is sensitive to noise
- Introducing Reliability-Weighted PHAT (RWPHAT)
  - Apply weighting

$$R_{i,j}^{RWPHAT}(\tau) = \sum_{k=0}^{L-1} \frac{\zeta_i(k) X_i(k) \zeta_j(k) X_j(k)^*}{|X_i(k)|\,|X_j(k)|} e^{\jmath 2\pi k\tau / L}$$

  - Weight based on noise and reverberation

$$\zeta_i(k) = \frac{\text{signal}}{\text{signal} + \text{noise} + \text{reverberation}}$$

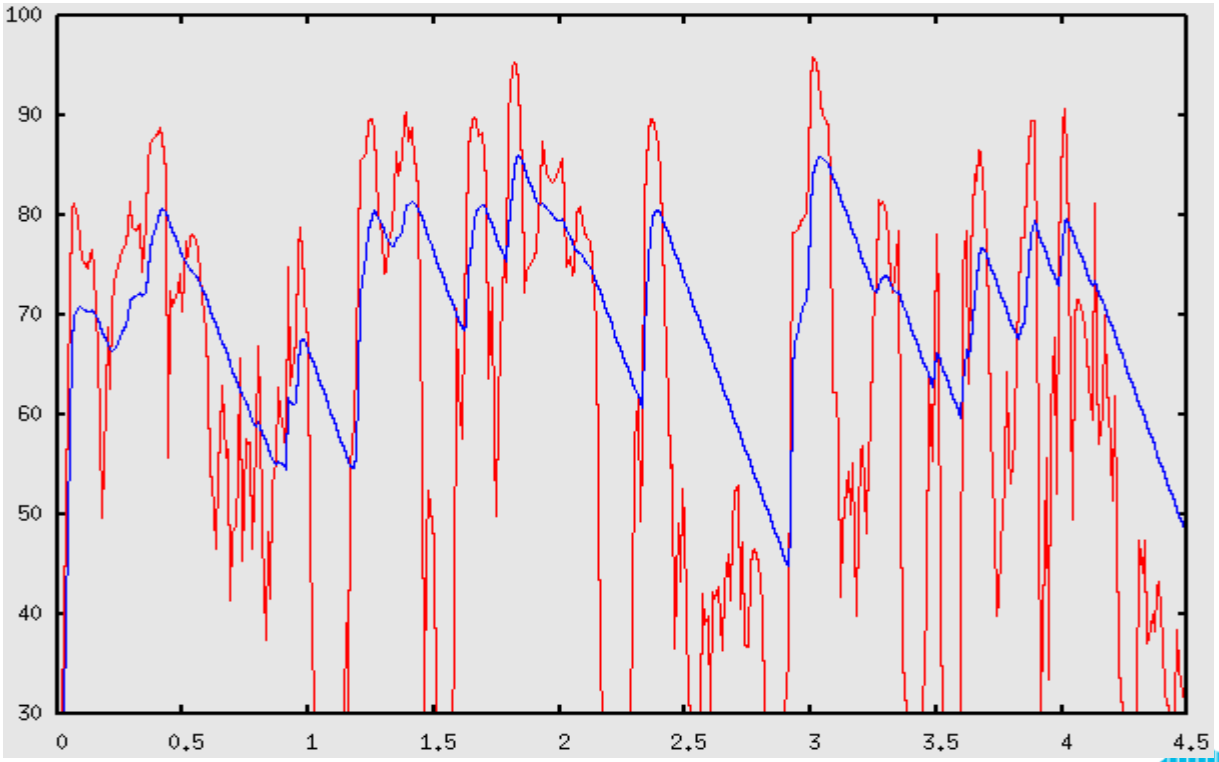
  - Discards unreliable frequency bands
  - Models precedence effect

Reverberation Estimation

- Exponential decay model

$$\lambda_i^n(k) = \gamma\lambda_i^{n-1}(k) + (1-\gamma)\delta^{-1}\left|\zeta_i^n(k)X_i^{n-1}(k)\right|^2$$

- Example: 500 Hz frequency bin



CSIRO

Search

- Only N(N-1) lookup-and-sum operations per location
- Assumes fixed number of sources
- Coarse (41x41x5) – fine (201x210x25) grid search

for $q = 1$ to assumed number of sources **do**
  for all grid index k **do**
    $E_k \leftarrow \sum_{i,j} R_{i,j}^{RWPHAT}(lookup(k,i,j))$
  end for
  $D_q \leftarrow \text{argmax}_k\ (E_k)$
  for all microphone pair i, j **do**
    $R_{i,j}^{RWPHAT}(lookup(D_q,i,j)) \leftarrow 0$
  end for
end for

Tracking With Particle Filtering

- Integrate beamformer observations in time
- State = [location, velocity]
- PDF represented as a set of particles
  - 1000 particles per tracked source
  - Sequential Importance Resampling
- Why not Kalman filtering?
  - Multi-modal distributions
    - Multiple observations
    - False detections in steered beamformer
  - Flexibility of predictor in particle filter

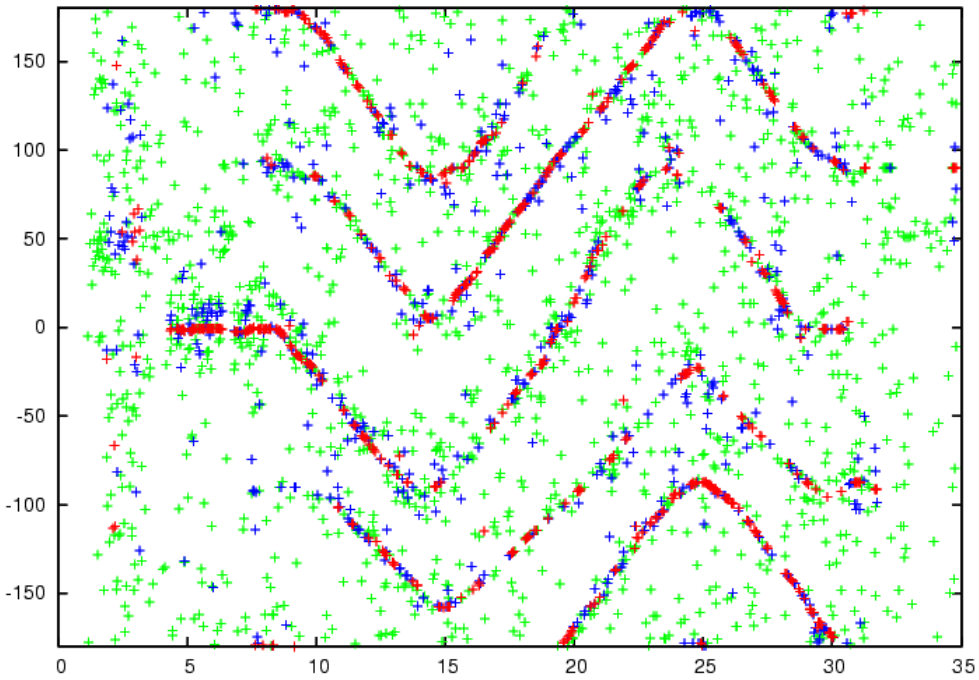
Particle Filtering Steps

1) Prediction
   - Position and velocity
   - Excitation-damping model
   - Random excitation

2) Instantaneous probability estimation
   - Based on steered beamformer alone
   - Function of beamformer energy

$$\dot{\mathbf{x}}_{j,i}^{(t)} = a\dot{\mathbf{x}}_{j,i}^{(t-1)} + bF_{\mathbf{x}}$$

$$\mathbf{x}_{j,i}^{(t)} = \mathbf{x}_{j,i}^{(t-1)} + \Delta T\dot{\mathbf{x}}_{j,i}^{(t)}$$

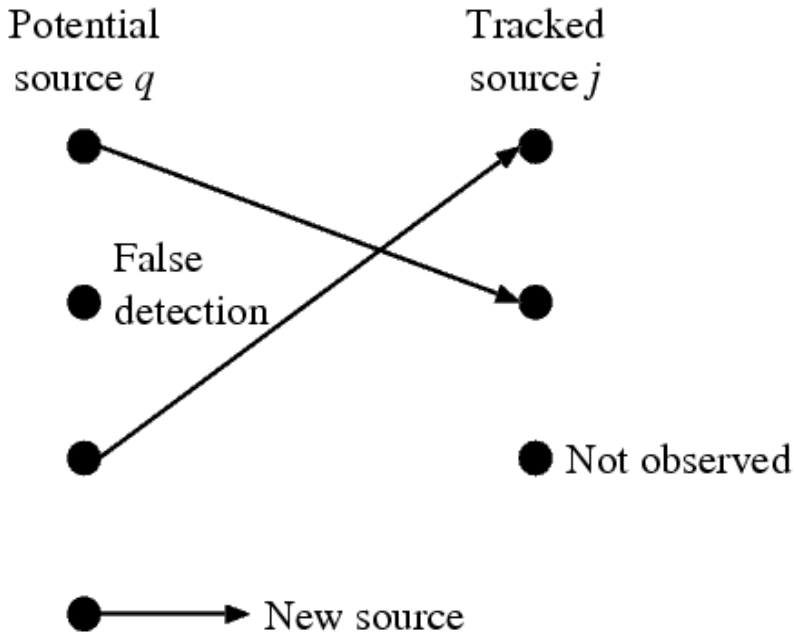


CSIRO

Particle Filtering Steps (cont.)

3) Source-observation assignment

- Match beamformer observations to tracked sources
- Compute:
  - Probability of false alarm
  - Probability of new source
  - Probability for each tracked source



4) Update particle weights

- Applying Bayes' rule
- Merging past and present information
- Taking into account source-observation assignment

Particle Filtering Steps (cont.)

5) Addition or removal of sources

6) Location estimation

- Weighted mean of particle positions

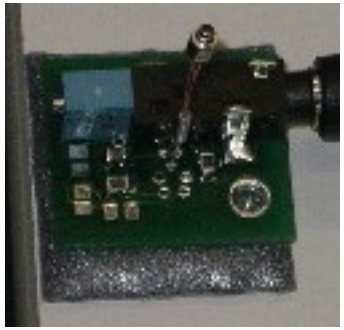
7) Resampling

- Eliminate particles with low probability
- Increase number of particles in regions of high probability
- Performed only when necessary

- Example (animation)

Experimental Setup

- Circular array of 8 microphones
  - 60 cm diameter
- ~ 7dB SNR

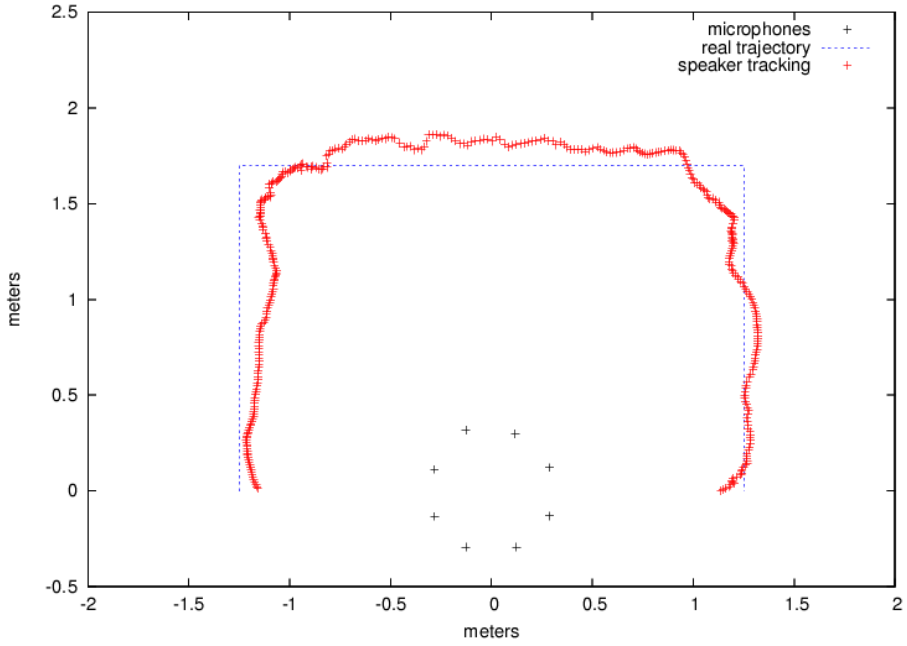


Localization Results

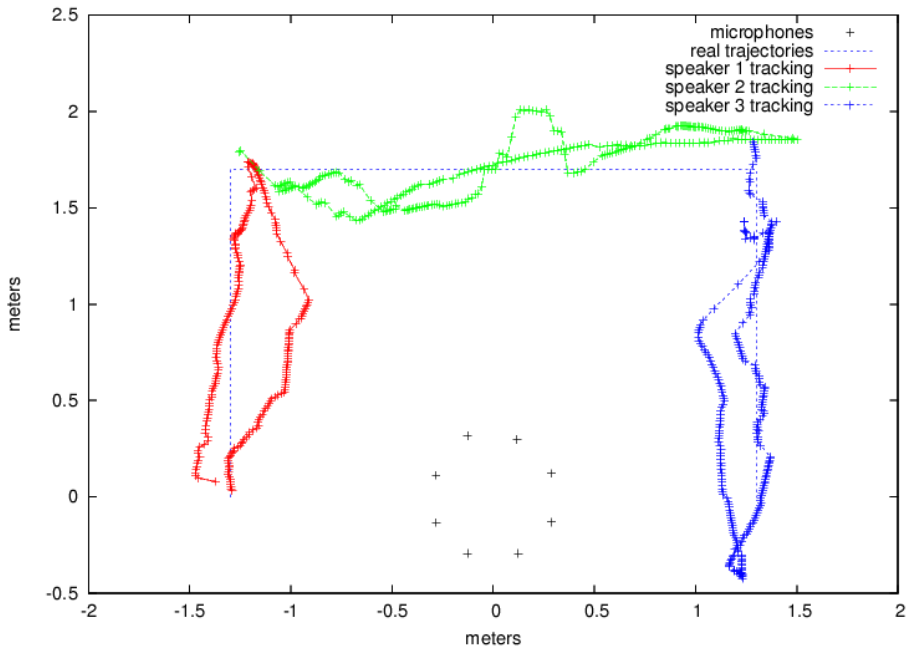
- One stationary source
  - < 1 degree angular resolution
  - 10 % accuracy on distance
- Multiple moving sources
  - Impossible to measure angular accuracy
  - ~10% accuracy on distance

Tracking Results

1 moving speaker



3 moving speakers



CSIRO

Conclusion

- Two-step approach
  - Steered beamformer
  - Particle filtering
- Accurate localization and tracking
  - < 1 degree angular error
  - ~10 % distance error
  - Tracking up to 3 speakers
- Future work
  - Improve distance accuracy
  - Handling of uncertainty on new sources
  - Merge visual and audio information

Questions?