Auditory System For a Mobile Robot

PhD Thesis

Jean-Marc Valin

Department of Electrical Engineering and Computer Engineering
Université de Sherbrooke, Québec, Canada
Jean-Marc.Valin@USherbrooke.ca

Motivations

- Robots need information about their environment in order to be intelligent
- Artificial vision has been popular for a long time, but artificial audition is new
- Robust audition is essential for human-robot interaction (*cocktail party effect*)

Approaches To Artificial Audition

- Single microphone
  - Human-robot interaction
  - Unreliable
- Two microphones (binaural audition)
  - Imitate human auditory system
  - Limited localisation and separation
- **Microphone array audition**
  - More information available
  - Simpler processing

UNIVERSITÉ DE SHERBROOKE
IMSI INSTITUT DES MATÉRIAUX ET SYSTÈMES INTELLIGENTS INTELLIGENT MATERIALS AND SYSTEMS INSTITUTE
LABORIUS

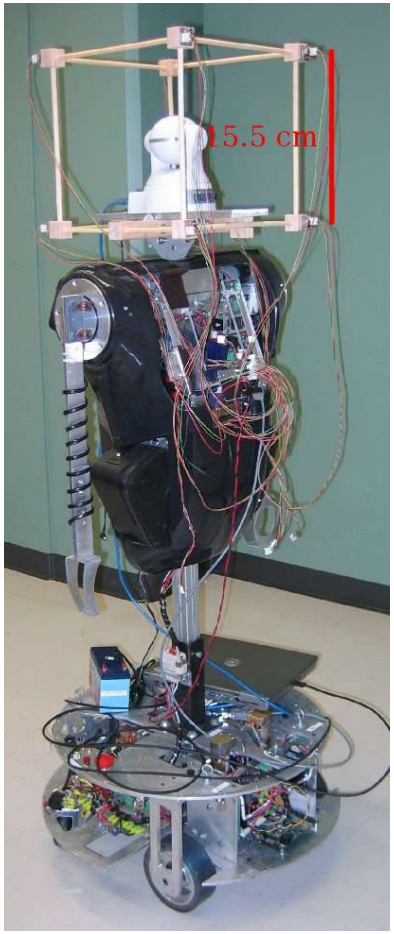
Objectives

- Localise and track simultaneous moving sound sources
- Separate sound sources
- Perform automatic speech recognition
- Remain within robotics constraints
  - complexity, algorithmic delay
  - robustness to noise and reverberation
  - weight/space/adaptability
  - moving sources, moving robot

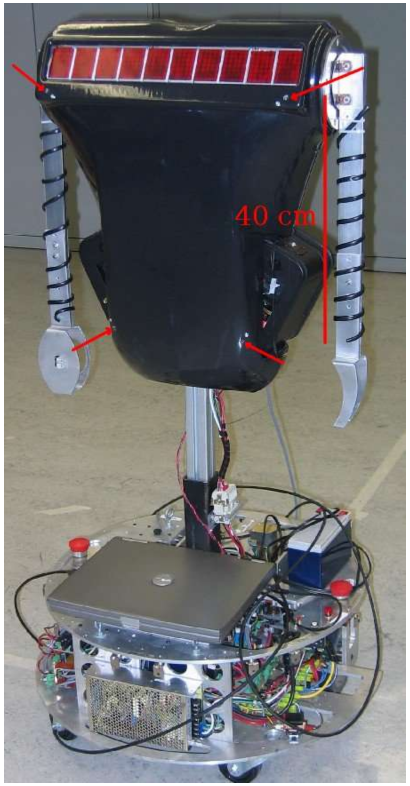
Experimental Setup

- Eight microphones on the Spartacus robot
- Two configurations
- Noisy conditions
- Two environments
- Reverberation time
  - Lab (E1) 350 ms
  - Hall (E2) 1 s

cube (C1)



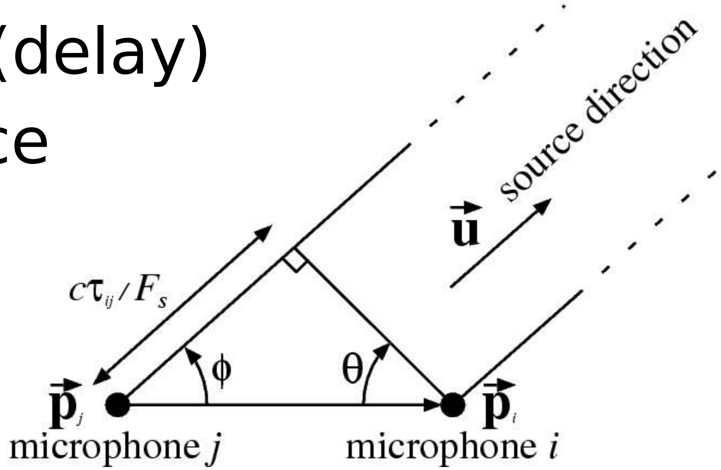
shell(C2)



Sound Source Localisation

Approaches to Sound Source Localisation

- Binaural
  - Interaural phase difference (delay)
  - Interaural intensity difference
- Microphone array
  - Estimation through TDOAs
  - Subspace methods (MUSIC)
  - **Direct search (steered beamformer)**
- Post-processing
  - Kalman filtering
  - **Particle filtering**



Steered Beamformer

- Delay-and-sum beamformer

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

- Maximise 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

- Normal cross-correlation peaks are very wide
- PHAse Transform (PHAT) has narrow peaks
- Apply weighting

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

  - Weight according to noise and reverberation

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

  - Models the precedence effect
    - Sensitivity is decreased after a loud sound

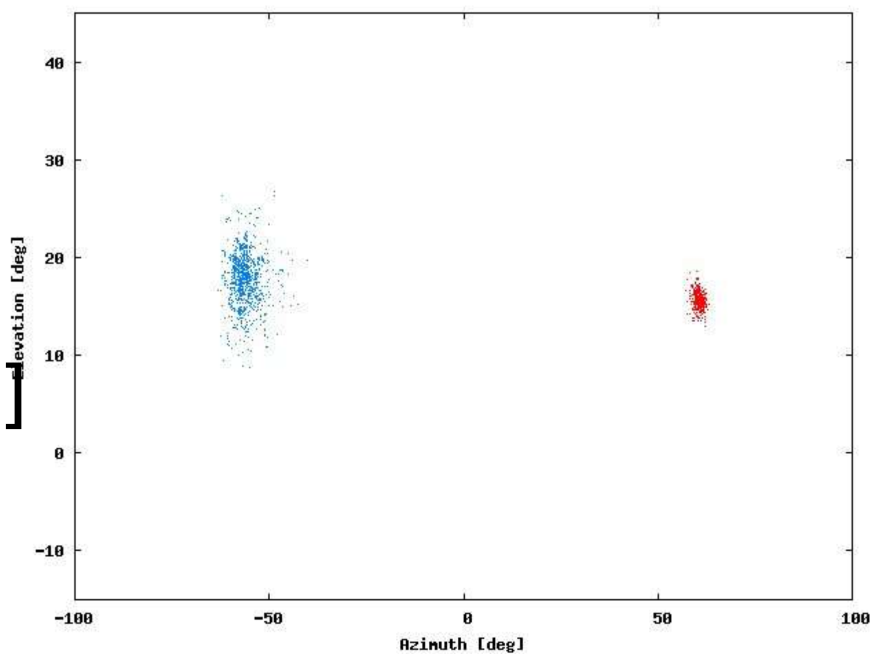
Direction Search

- Finding directions with highest energy
- Fixed number of sources *Q=4*
- Lookup-and-sum algorithm
- 25 times less complex

for $q = 0$ to $Q-1$ **do**
  for all grid index k **do**
    $E_k \leftarrow 0$
    for all microphone pair ij **do**
      $\tau \leftarrow lookup(k, ij)$
      $E_k \leftarrow E_k + R_{ij}^{(e)}(\tau)$
  $D_q \leftarrow \operatorname{argmax}_k (E_k)$
  for all microphone pair ij **do**
    $\tau \leftarrow lookup(D_q, ij)$
    $R_{ij}^{(e)}(\tau) = 0$

Post-Processing: Particle Filtering

- Need to track sources over time
- Steered beamformer output is noisy
- Representing pdf as particles
- One set of (1000) particles per source
- State=[position, speed]



Particle Filtering Steps

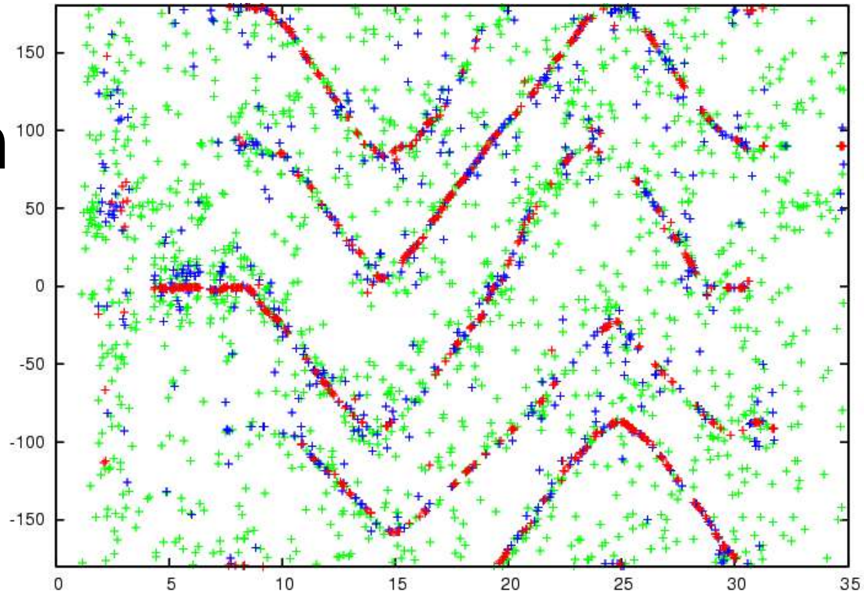
1) Prediction

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

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

2) Instantaneous probabilities estimation

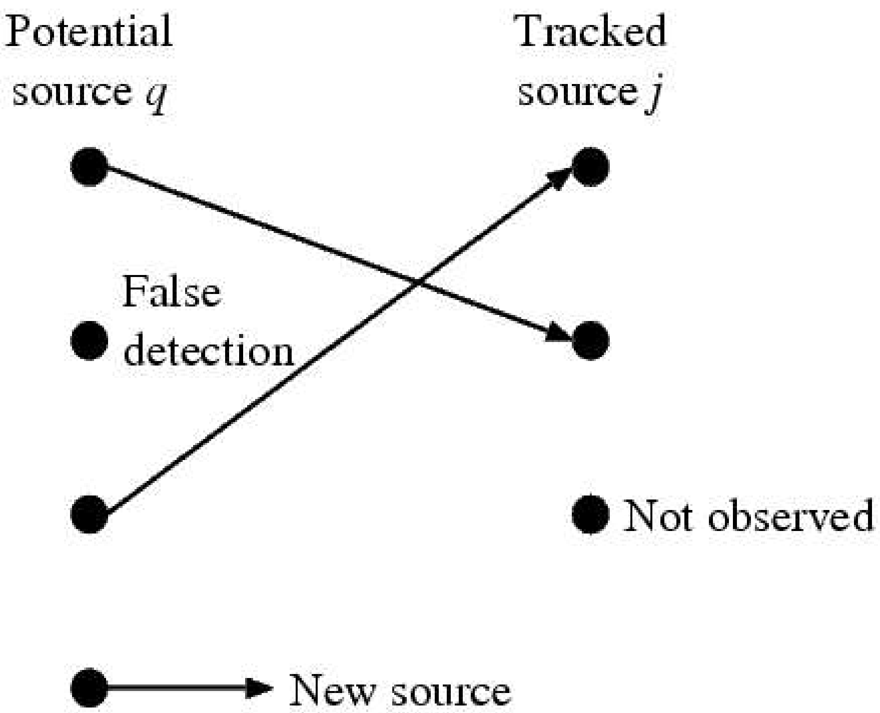
- As a function of steered beamformer energy



Particle Filtering Steps (cont.)

3) Source-observation assignment

- Need to know which observation is related to which tracked source
- Compute
  - $P_q^{(t)}(H_0)$: Probability that *q* is a false alarm
  - $P_{q,j}^{(t)}$: Probability that *q* is source *j*
  - $P_q^{(t)}(H_2)$: Probability that *q* is a new source



Particle Filtering Steps (cont.)

4) Particle weights update

$$w_{j,i}^{(t)} = p\left(\mathbf{x}_{j,i}^{(t)} \middle| \mathbf{O}^{(t)}\right)$$

- Merging past and present information
- Taking into account source-observation assignment

5) Addition or removal of sources

6) Estimation of source positions

- Weighted mean of the particle positions

7) Resampling

Localisation Results (E1)

Detection accuracy over distance

Distance	Correct (%)		Reflection (%)		Other error (%)	
	C1	C2	C1	C2	C1	C2
1 m	100	94.2	0.0	7.3	0.0	1.3
3 m	99.4	80.6	0.0	21.0	0.3	0.1
5 m	98.3	89.4	0.0	0.0	0.0	1.1
7 m	100	85.0	0.6	1.1	0.6	1.1

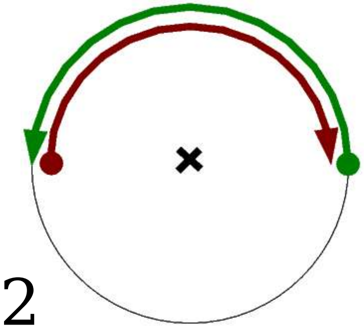
Localisation accuracy

Localisation error	C1 (deg)	C2 (deg)
Azimuth	1.10	1.44
Elevation	0.89	1.41

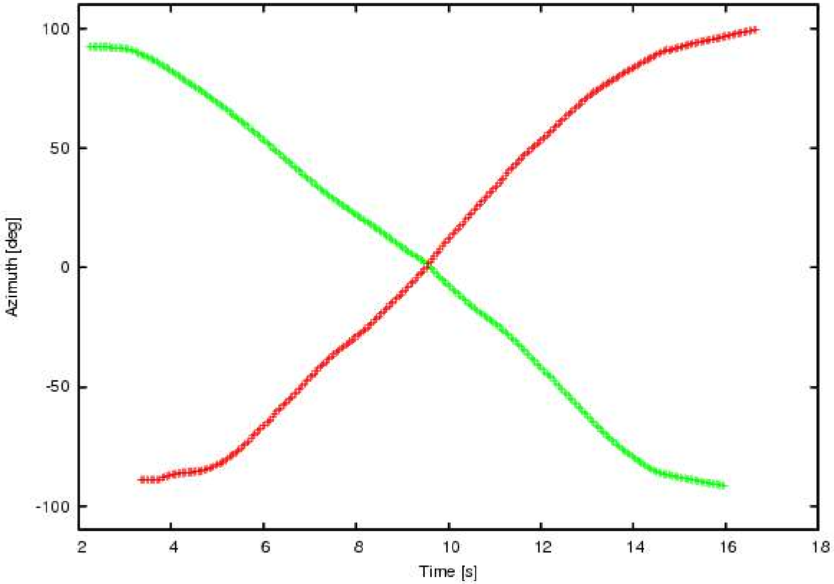
Tracking Results

Two sources crossing with C2

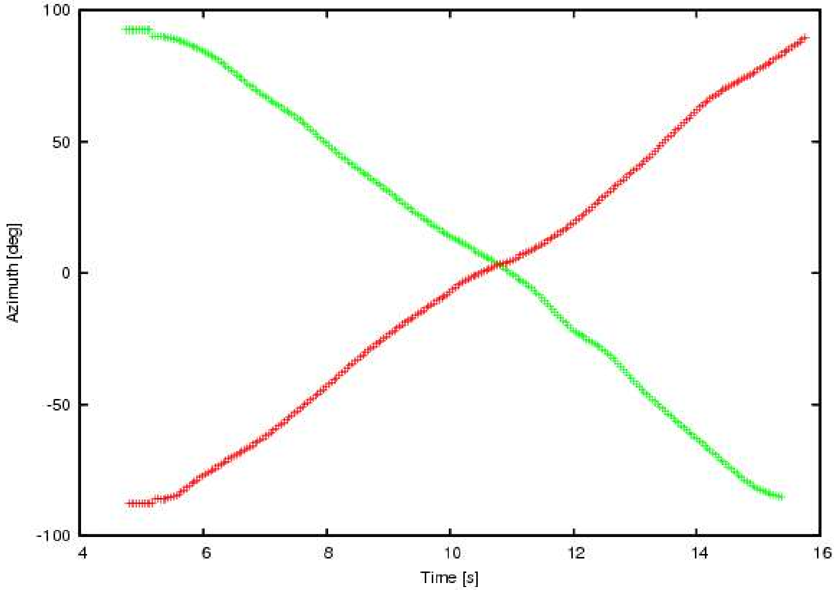
- Video



E1

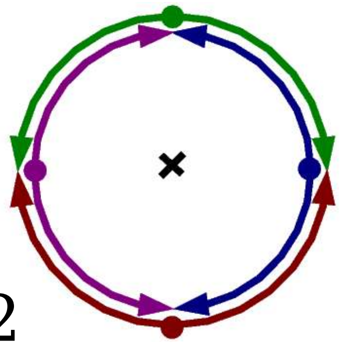


E2

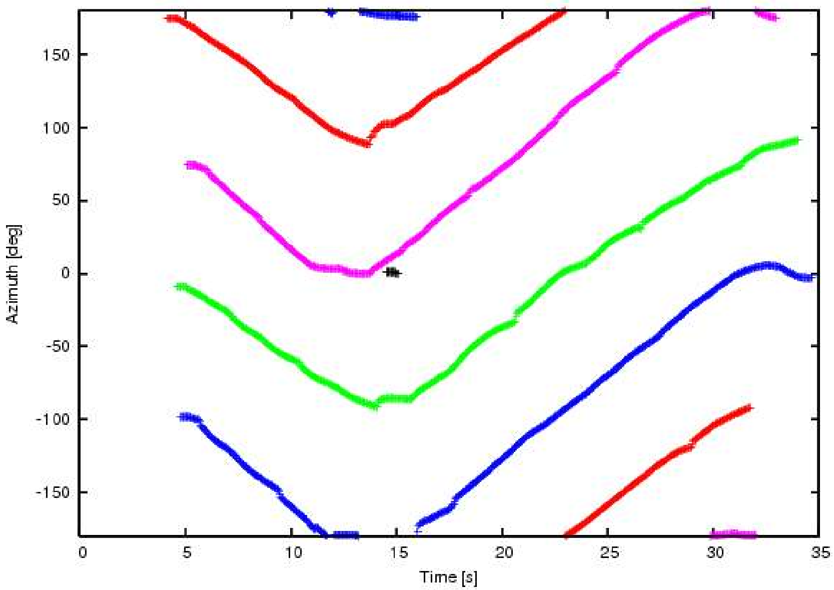


Tracking Results (cont.)

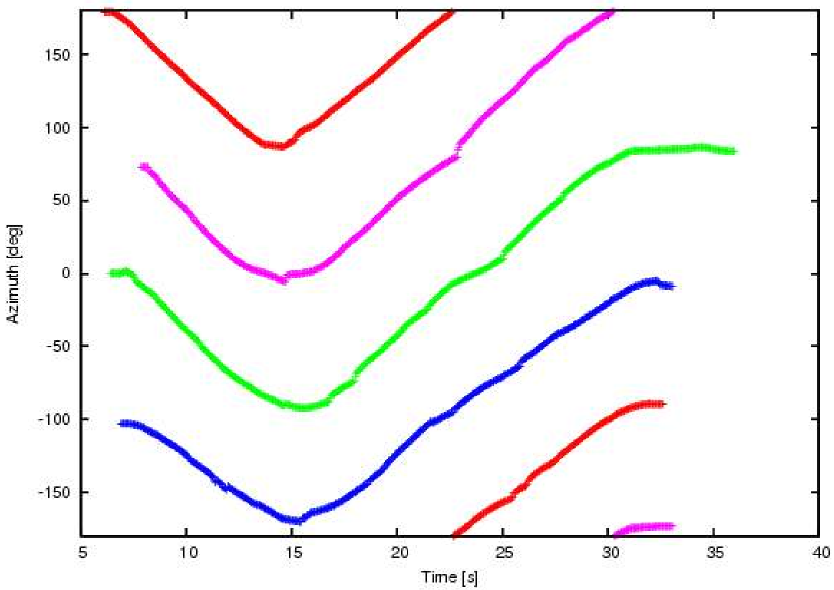
Four moving sources with C2



E1



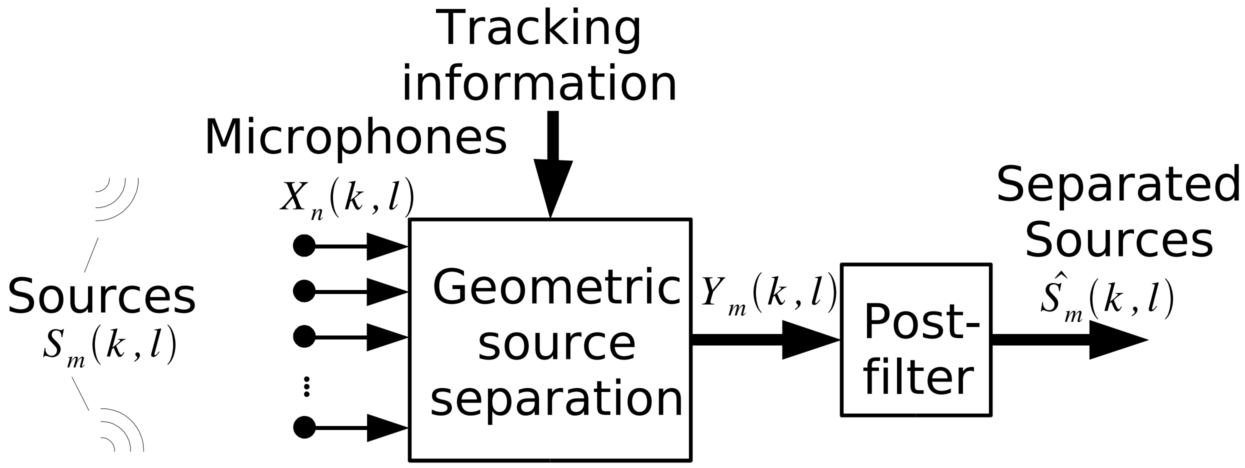
E2



Sound Source Separation
&
Speech Recognition

Overview of Sound Source Separation

- Frequency domain processing
  - Simple, low complexity
- Linear source separation
- Non-linear post-filter



Geometric Source Separation

- Frequency domain:

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) + \mathbf{n}(k)$$

- Constrained optimization $\mathbf{y}(k) = \mathbf{W}(k)\mathbf{x}(k)$
  - Minimize correlation of the outputs:

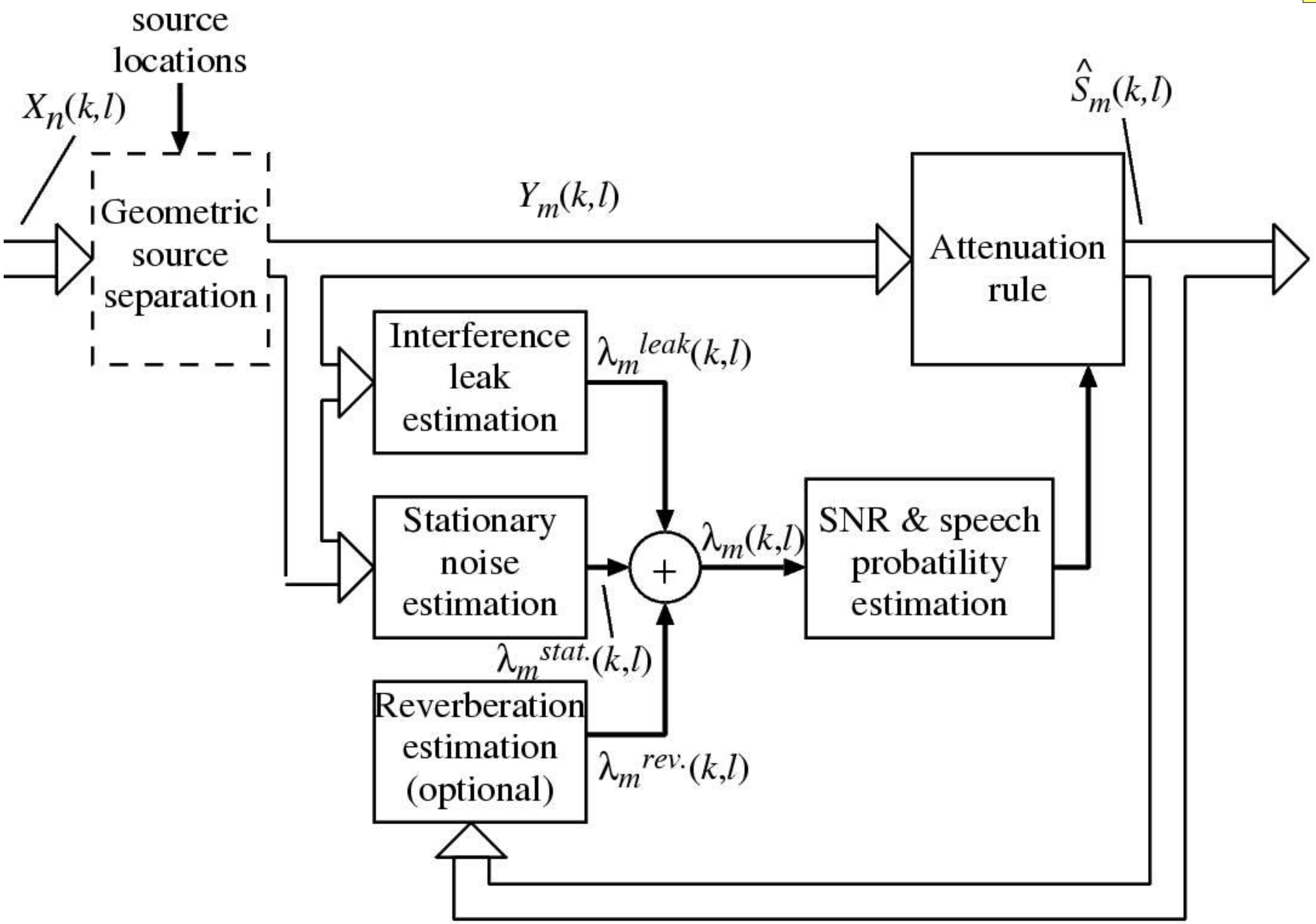
$$J_1(\mathbf{W}(k)) \quad = \quad \left\|\mathbf{R}_{\mathbf{yy}}(k) - \mathrm{diag}\left[\mathbf{R}_{\mathbf{yy}}(k)\right]\right\|^2$$

  - Subject to geometric constraint:

$$J_2(\mathbf{W}(k)) \quad = \quad \left\|\mathbf{W}(k)\mathbf{A}(k) - \mathbf{I}\right\|^2$$

- Modifications to original GSS algorithm
  - Instantaneous computation of correlations
  - Regularisation

Multi-Source Post-Filter



Interference Estimation

- Source separation leaks
  - Incomplete adaptation
  - Inaccuracy in localization
  - Reverberation/diffraction
  - Imperfect microphones
- Estimation from other separated sources

$$\lambda_m^{leak}(k,\ell) = \eta \sum_{i=0,i\neq m}^{M-1} Z_i(k,\ell)$$

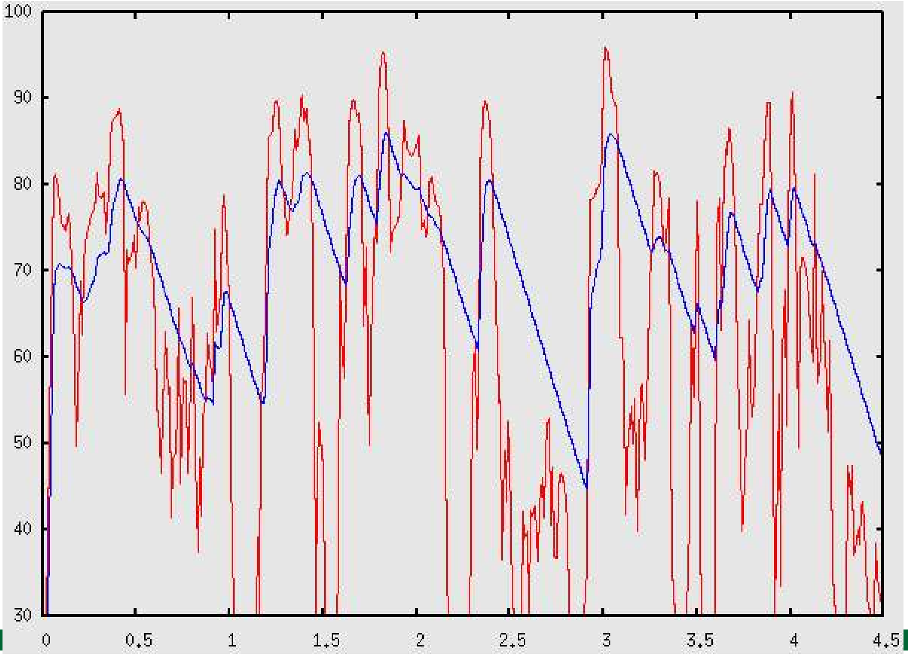
$$Z_m(k,\ell) = \alpha_s Z_m(k,\ell-1) + (1-\alpha_s)\left|Y_m(k,\ell)\right|^2$$

Reverberation Estimation

- Exponential decay model

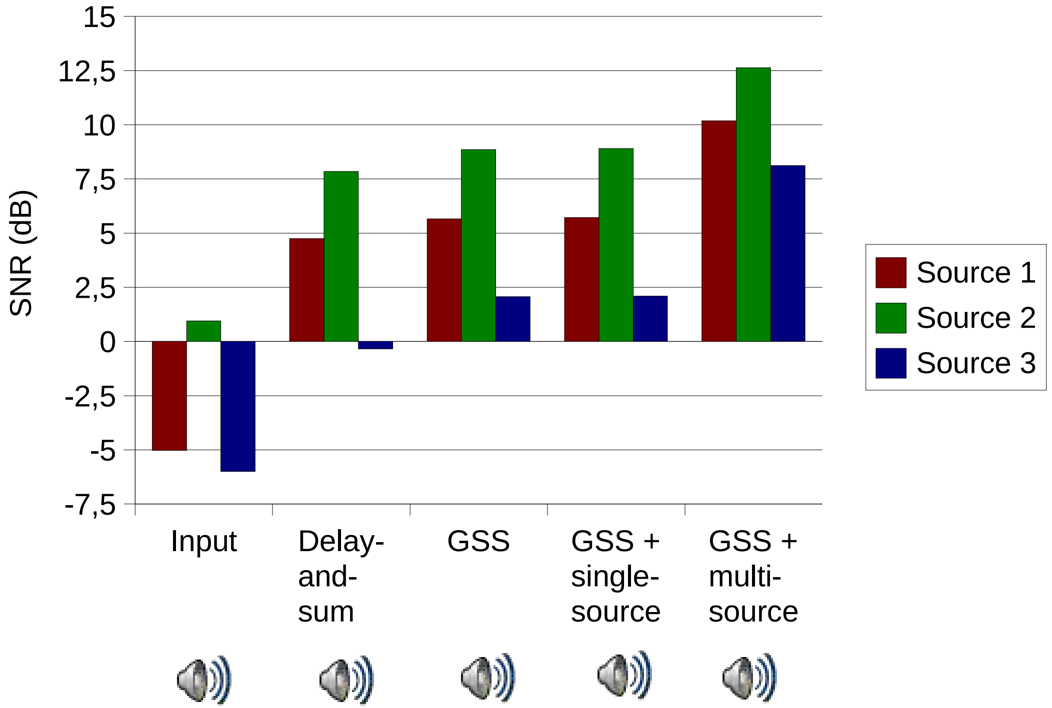
$$\lambda_i^{rev}(k,\ell) = \gamma \lambda_i^{rev}(k,\ell-1) + \frac{(1-\gamma)}{\delta}\left|\hat{S}_i(k,\ell-1)\right|^2$$

- Example: 500 Hz frequency bin



Results (SNR)

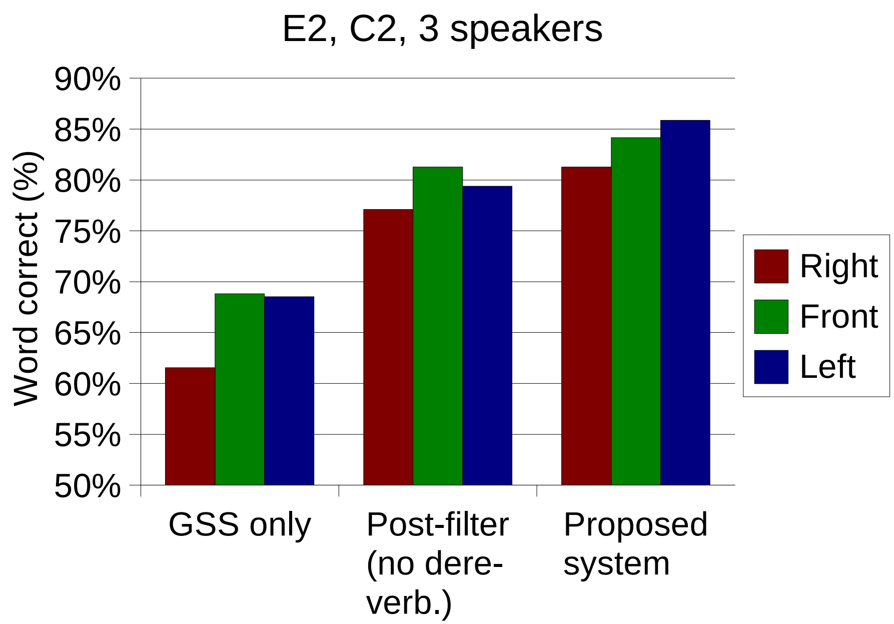
- Three speakers
- C2 (shell), E1 (lab)



Speech Recognition Accuracy (Nuance)

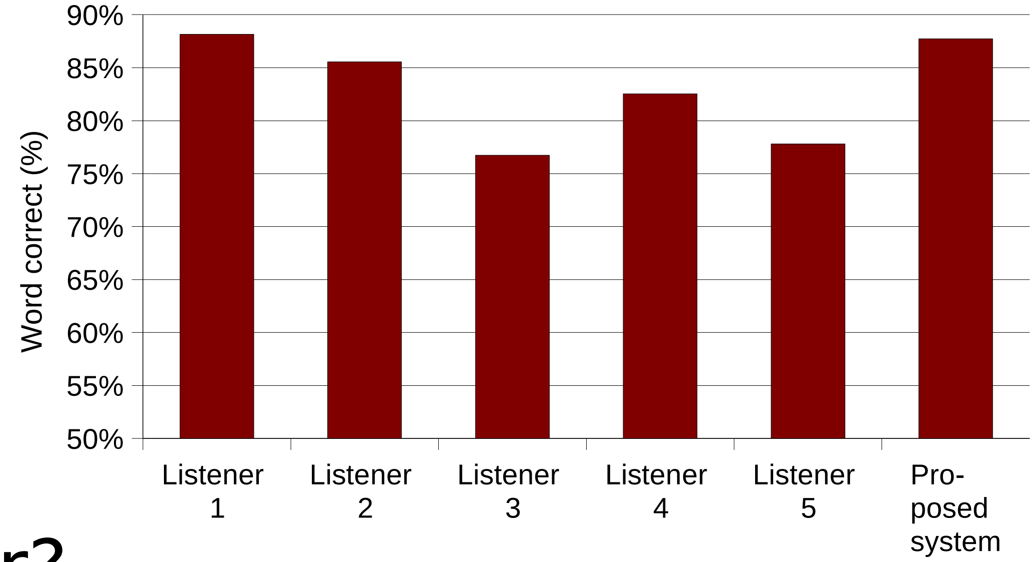
- Proposed post-filter reduces errors by 50%
- Reverberation removal helps in E2 only
- No significant difference between C1 and C2
- Digit recognition
- 3 speakers: 83%
- 2 speakers: 90%

microphone separated



Man vs. Machine

- How does a human compare?



- Is it fair?
  - Yes and no!

Real-Time Application

- Video from AAAI conference

Speech Recognition With Missing Feature Theory

- Speech is transformed into features (~12)
- Not all features are reliable
- MFT = ignore unreliable features
  - Compute missing feature mask
  - Use the mask to compute probabilities

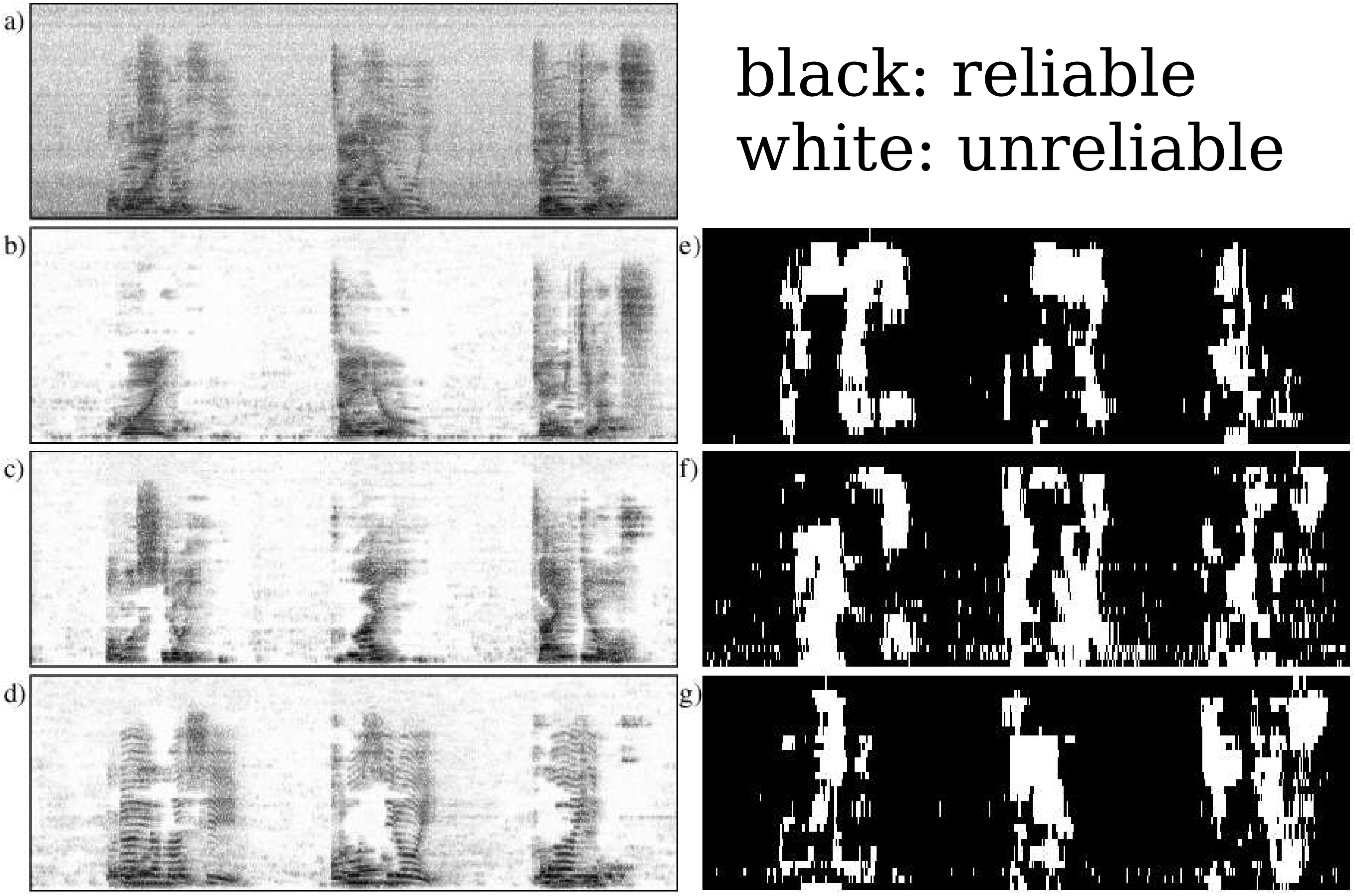
$$m_\ell(i) = \frac{S_\ell^{out}(i) + N_\ell(i)}{S_\ell^{in}(i)} \qquad M_\ell(i) = \begin{cases} 1, & m_\ell(i) > T_m \\ 0, & \text{otherwise} \end{cases}$$

Missing Feature Mask

Interference: unreliable

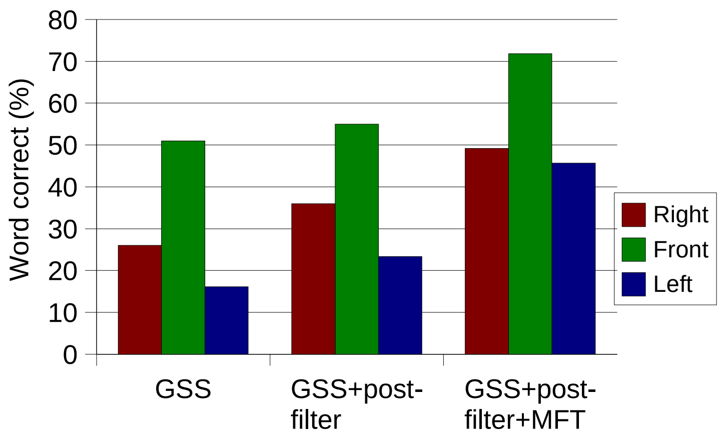
Stationary noise: reliable

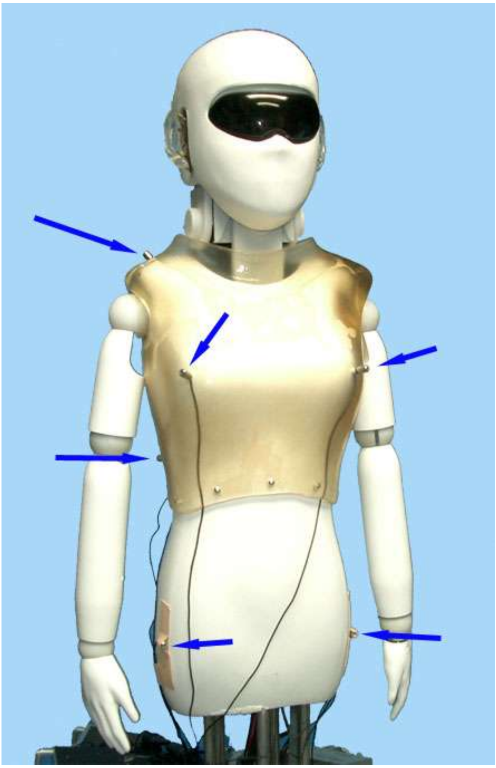
black: reliable
white: unreliable



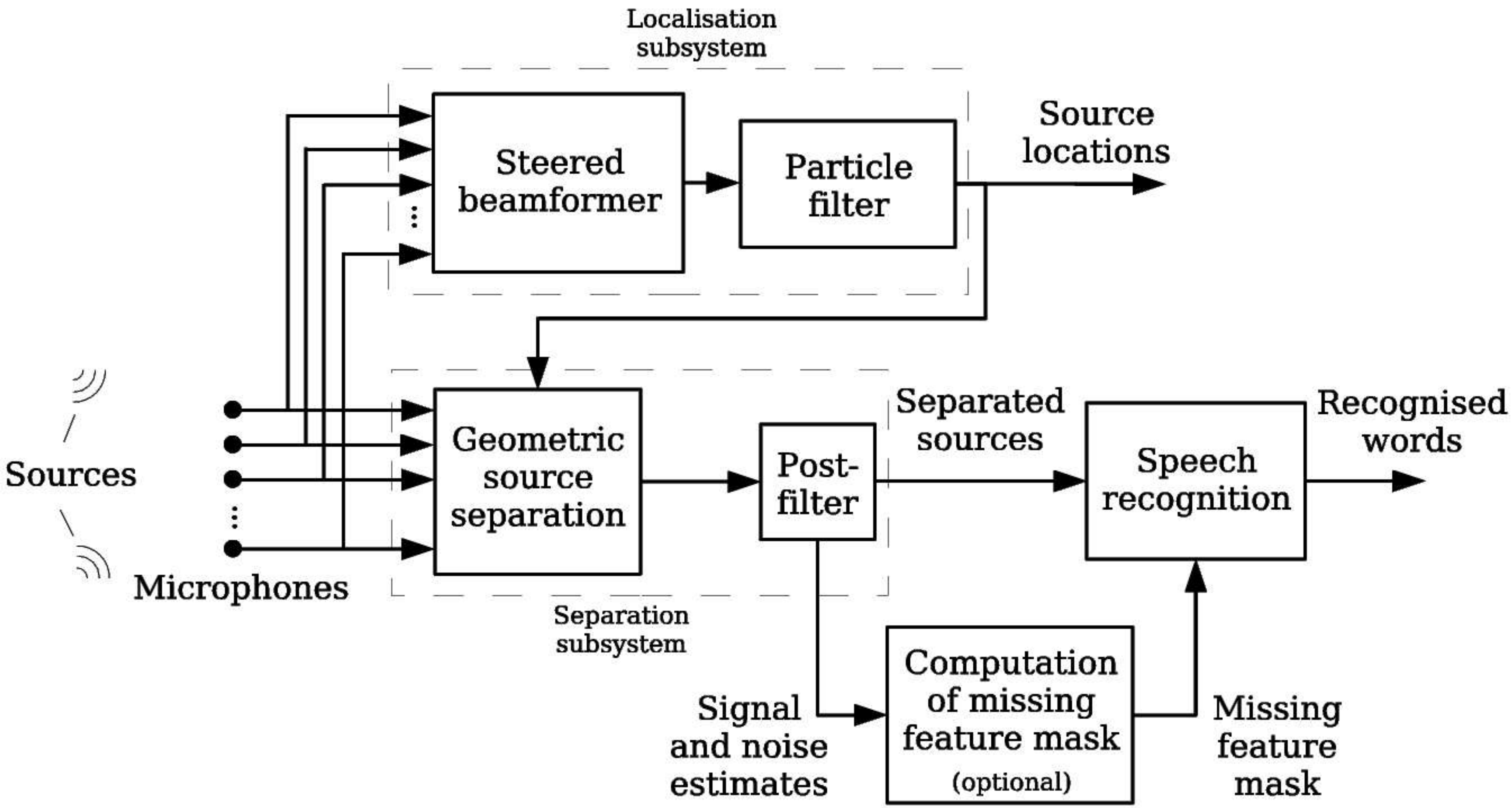
Results (MFT)

- Japanese isolated word recognition (SIG2 robot, CTK)
  - 3 simultaneous sources
  - 200-word vocabulary
  - 30, 60, 90 degrees separation





Summary of the System



Conclusion

- What have we achieved?
  - Localisation and tracking of sound sources
  - Separation of multiple sources
  - Robust basis for human-robot interaction
- What are the main innovations?
  - Frequency-domain steered beamformer
  - Particle filtering source-observation assignment
  - Separation post-filtering for multiple sources and reverberation
  - Integration with missing feature theory

Where From Here?

- Future work
  - Complete dialogue system
  - Echo cancellation for the robot's own voice
  - Use human-inspired techniques
  - Environmental sound recognition
  - Embedded implementation
- Other applications
  - Video-conference: automatically follow speaker with a camera
  - Automatic transcription

Questions? Comments?