Modeling of motion dynamics and its influence on the performance of a particle filter for acoustic speaker tracking

Eric Lehmann, Anders Johansson and Sven Nordholm

Western Australian Telecommunications Research Institute
Perth, Western Australia
Eric.Lehmann@watri.org.au

THE UNIVERSITY OF WESTERN AUSTRALIA

WASPAA'07
New Paltz, NY
October 21–24, 2007

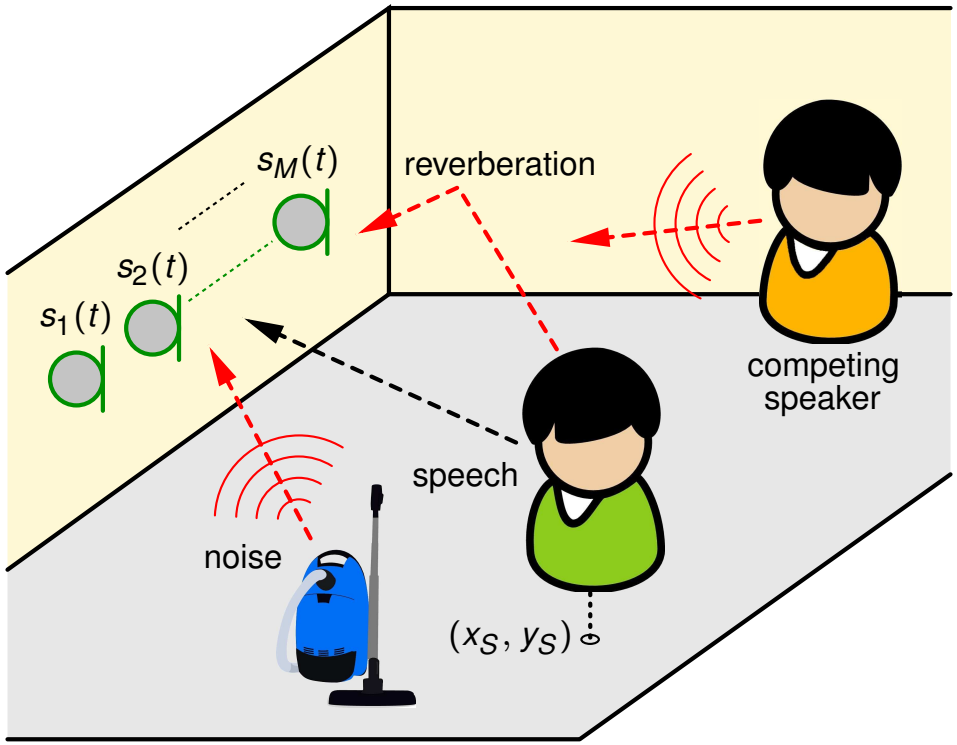
Outline

Acoustic Source Tracking

Purpose

Localizing and tracking an acoustic source using data collected at an array of M microphones.

- **Applications:**
  - teleconference
  - speech acquisition
  - noise reduction
  - surveillance
  - etc.
- **Disturbances:**
  - reverberation
  - background noise
  - competing speakers (multi-source tracking)
  - speech pauses



Bayesian Filtering/Tracking

Discrete-time state-space approach based on:

- *state variable* at time k: $\boldsymbol{X}_k = [x_k \;\; y_k \;\; \boldsymbol{s}_k]^{\mathrm{T}}$, where $\boldsymbol{s}_k$ contains additional state components (velocity, heading, etc.)
- *observation variable* (measurement data): $\boldsymbol{Y}_k$

Aim: compute the *posterior PDF* $p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k})$, from which an estimate of the state $\widehat{\boldsymbol{X}}_k$ can be obtained.

Bayesian Recursion

$$p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k-1}) = \int p(\boldsymbol{X}_k|\boldsymbol{X}_{k-1})\, p(\boldsymbol{X}_{k-1}|\boldsymbol{Y}_{1:k-1})\, \mathrm{d}\boldsymbol{X}_{k-1}$$

$$p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k}) \propto p(\boldsymbol{Y}_k|\boldsymbol{X}_k)\, p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k-1})$$

$p(\boldsymbol{X}_k|\boldsymbol{X}_{k-1})$: transition PDF

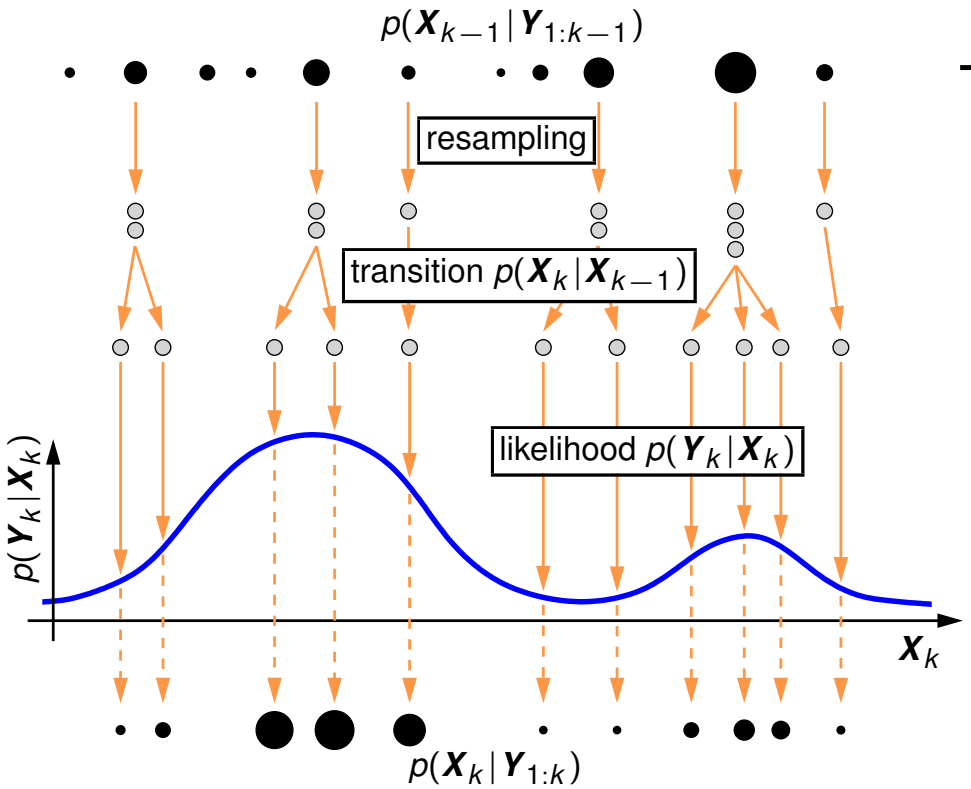
$p(\boldsymbol{Y}_k|\boldsymbol{X}_k)$: likelihood function

Particle Filtering (PF)

Sequential Monte Carlo Technique

Approximate Bayesian solution based on a discrete representation of the posterior PDF with *N particles* and *weights*:

$$\{(\boldsymbol{X}_k^{(n)}, w_k^{(n)})\}_{n=1}^{N} \sim p(\boldsymbol{X}_k|\boldsymbol{Y}_{1:k})$$



Two main concepts:

- **likelihood function** $p(\boldsymbol{Y}_k|\boldsymbol{X}_k)$ based on the measurement data $\boldsymbol{Y}_k$
- **transition equation** $\boldsymbol{X}_k = g(\boldsymbol{X}_{k-1}, \boldsymbol{u}_k)$ defined by a model of the target dynamics

PF for Acoustic Source Tracking

Currently available literature focuses on:

- development of improved PF algorithms: new proposal distributions, multiple-source tracking, etc.
- enhancement of the measurement PDF (likelihood)

Little attention is given to the *dynamics model!* A typical model usually implemented relies on the so-called *Langevin* dynamics.

Langevin Model (Vermaak & Blake, 2001)

State variable: $\boldsymbol{X}_k = [x_k \;\; y_k \;\; \dot{x}_k \;\; \dot{y}_k]^{\mathsf{T}}$

Transition equation (similar for y coordinate):

$$\dot{x}_k = a \cdot \dot{x}_{k-1} + b \cdot u_k, \quad u_k \sim \mathcal{N}(0,1)$$
$$x_k = x_{k-1} + \Delta T \cdot \dot{x}_k$$

with $a = \exp(-\beta \cdot \Delta T)$, $b = \bar{v} \cdot \sqrt{1 - a^2}$, $\beta = 10\text{Hz}$, $\bar{v} = 1\text{m/s}$.

PF for Acoustic Source Tracking

PF-VAD: PF algorithm for acoustic source tracking...

- Observation $\boldsymbol{Y}_k$ computed from sensor array data as the steered beamformer output for a given location $\ell = [x \;\; y]^{\mathsf{T}}$:

$$\mathcal{P}(\ell) = \int \left| \sum_{m=1}^{M} W_m(\omega)\, S_m(\omega)\, \mathrm{e}^{j\omega \|\ell - \ell_m\|/c} \right|^2 \mathrm{d}\omega$$

  with PHAT weighting $W_m(\omega) = \left|S_m(\omega)\right|^{-1}$.

- Likelihood function defined as mixture PDF integrating voice activity detection (VAD) data:

$$p(\boldsymbol{Y}_k|\boldsymbol{X}_k) \propto q_k \cdot \mathcal{U}(\boldsymbol{X}_k) + (1 - q_k) \cdot \mathcal{P}(\boldsymbol{X}_k)$$

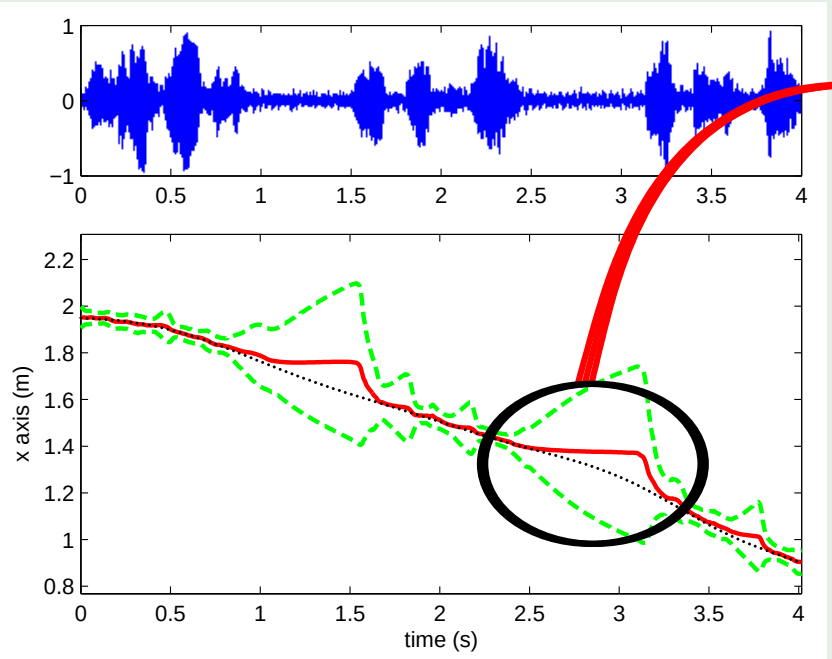
  where $\mathrm{q}_k = 1 - \mathrm{VAD}_k$.

- Standard Langevin model as transition PDF $p(\boldsymbol{X}_k|\boldsymbol{X}_{k-1})$.

PF for Acoustic Source Tracking

Example Result

- real audio, 8 mic. array
- $T_{60} \approx 300$ms, $\sim$15dB SNR
- red: x-dim. estimate
  green: standard deviation of particles in x-dim.



Outcomes:

- accurate tracking during speech segments
- during silence (i.e. no valid observation data):
  - PF estimate is “frozen”
  - particles spread out

... can we do better?

Current assumption: speaker can change direction and/or velocity *abruptly*!

More relevant (for speaker tracking): propagate the last-observed target motion during silence $\Rightarrow$ *momentum!*

Dynamics Modeling for Speaker Tracking

Motivation of Current Research

Aim: *preliminary experiments* to provide insight into the influence of dynamics modeling on the PF tracking performance

- How important/relevant is dynamics modeling?
- Can a PF "track" the target's velocity (momentum)?
- Is an acceleration component necessary?

Design objectives:

- *bulk of particles* should follow the speaker during silence
- allow the particles to *spread* during silence

This work is *not* about determining the best type of dynamics model for speaker tracking applications (requires a large data set) $\Rightarrow$ *object of ongoing research*...

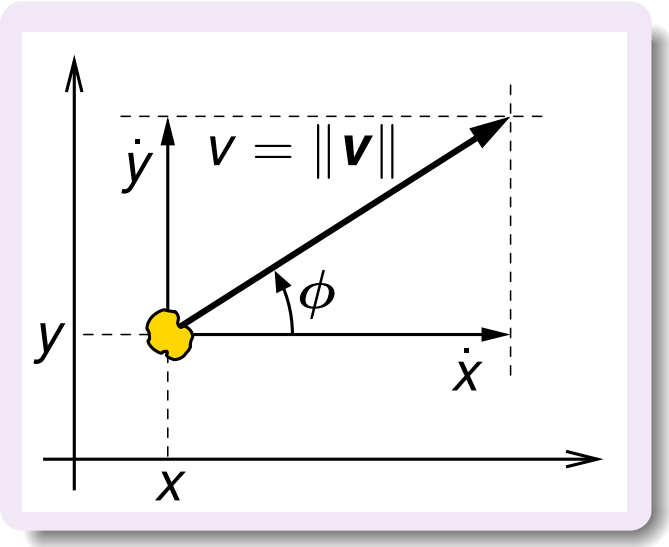
Dynamics Model Types

Two main model classes...

- **Coordinate-uncoupled (CU):**
  - state vector given in a Cartesian coordinate system, e.g.:

$$\boldsymbol{X}_k = [x_k \;\; y_k \;\; \dot{x}_k \;\; \dot{y}_k \;\; \ddot{x}_k \;\; \ddot{y}_k]^{\mathsf{T}}$$

  - no coupling between the x and y coordinates



- **Curvilinear (CL):**
  - state vector defined in a polar coordinate system:

$$\boldsymbol{X}_k = [x_k \;\; y_k \;\; v_k \;\; \phi_k]^{\mathsf{T}}$$

  - coupled x and y motions
  - target position:

$$x_k = x_{k-1} + \Delta T \cdot v_k \cdot \cos(\phi_k)$$
$$y_k = y_{k-1} + \Delta T \cdot v_k \cdot \sin(\phi_k)$$

Dynamics Model Types

Two types of process noise, for any state variable (here using v_k as example)...

- **Random-walk (RW):**

$$v_k = v_{k-1} + \sigma_v \cdot u_k, \quad u_k \sim \mathcal{N}(0,1)$$

- **Time-correlated (TC):**

$$v_k = \mathrm{e}^{-\beta_v \cdot \Delta T} \cdot v_{k-1} + \sigma_v \sqrt{1 - \mathrm{e}^{-2\beta_v \cdot \Delta T}} \cdot u_k, \quad u_k \sim \mathcal{N}(0,1)$$

Finally, various models also result from different model orders (velocity, acceleration, etc.).

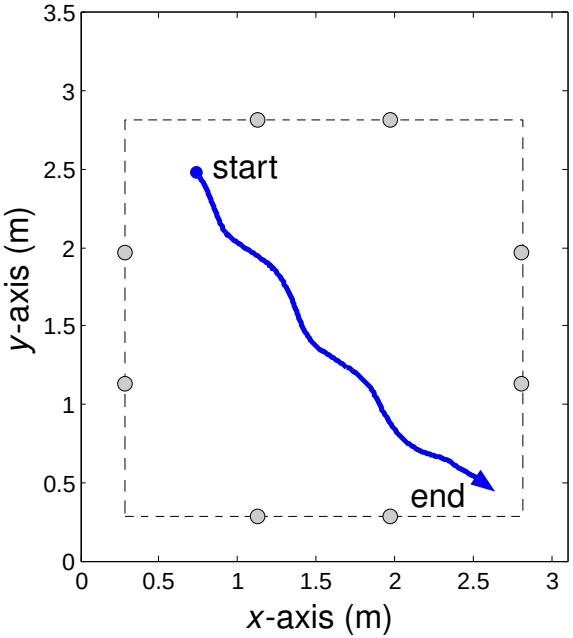
$\Rightarrow$ *... many different combinations can be considered!*

In this work, only a small subset are implemented, involving CU, CL, RW, TC, 1st order and 2nd order models.

Experimental Simulation

Setup details:

- real room recordings
- room size: 3.1m $\times$ 3.5m $\times$ 2.2m
- reverberation: $T_{60} \approx 0.3$s
- microphone array: $M = 8$ sensors, omnidirectional
- sampling frequency: $F_s = 16$kHz



Parameter optimization:

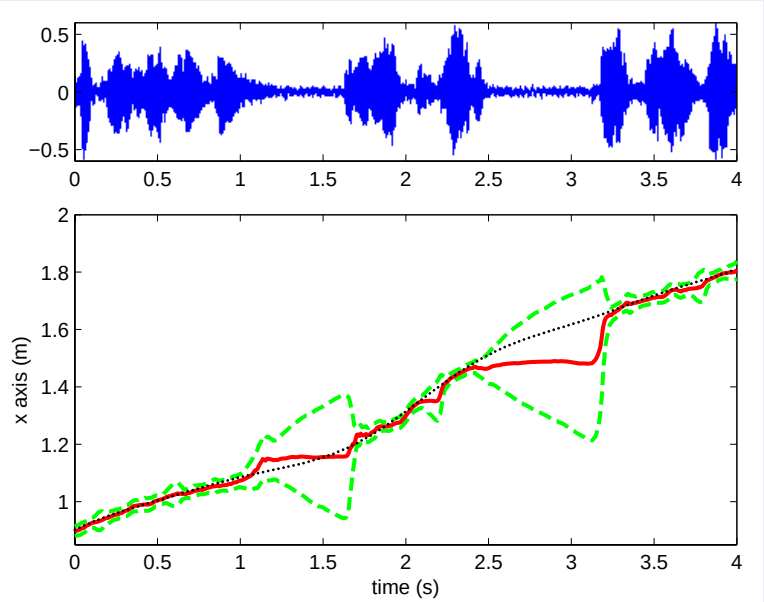
- dynamics models implemented within the PF-VAD algorithm
- model parameters coarsely optimized using a variety of speakers (male and female) and trajectories

Experimental results:

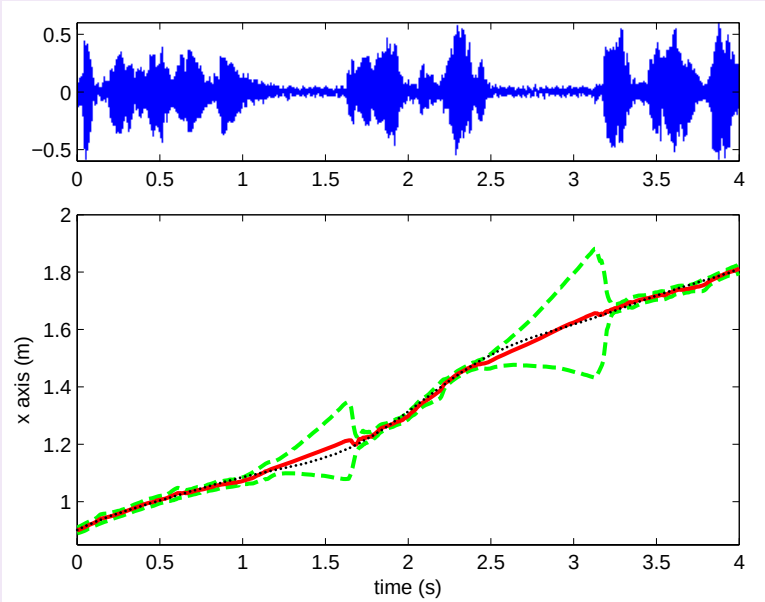
Compare the resulting PF tracking performance obtained for each optimized model $\rightarrow$ see WASPAA paper for full detail.

Comparative Results

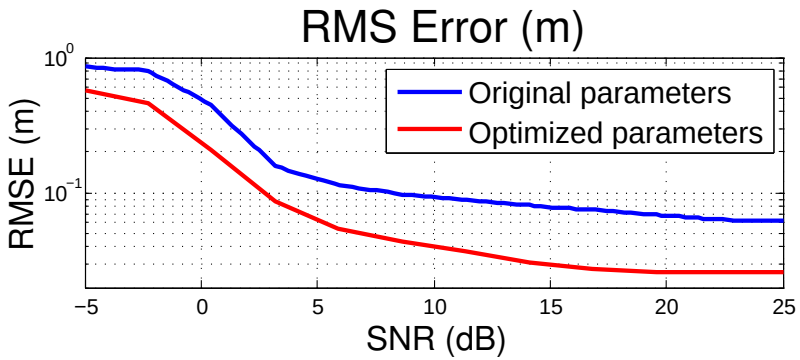
Original Langevin Model

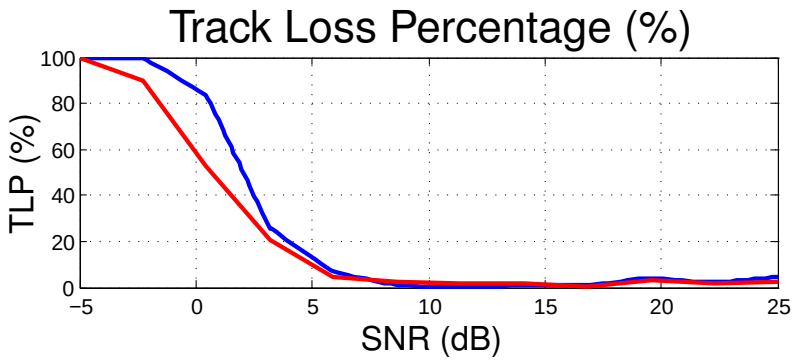


Optimized Langevin Model



Average performance vs. SNR:





Conclusion

Main Outcomes

- Dynamics modeling has significant impact on PF tracking
- Proper dynamics modeling leads to increased robustness against disturbances and/or with multiple speakers
- Model type as well as parameter optimization are *both* crucial issues
- All tested model types (CL vs. CU, RW vs. TC, etc.) are of potential interest:
  - a few combinations proved unsuccessful
  - 2nd order models (using an acceleration variable) are not necessary to achieve improved tracking performance

Future/current research:

- larger database of typical speaker trajectories
- “automatic” learning of target dynamics (genetic algorithm)

The End!

Thanks for your attention!...

Typeset by pdfLaTeX using the BEAMER document class