Expectation-Maximization (EM) Framework for Multiple Speaker Localization and Tracking

Sharon Gannot

Joint work with Ofer Schwartz, Yuval Dorfan & Gershon Hazan

Faculty of Engineering, Bar-Ilan University, Israel

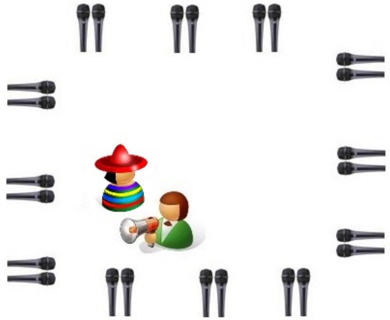
אוניברסיטת בר-אילן
Bar-Ilan University

The 3rd Annual Underwater Acoustics Symposium
Tel-Aviv University, June 19th, 2014

Preface

Multiple Speaker Localization using a Network of Microphone Pairs

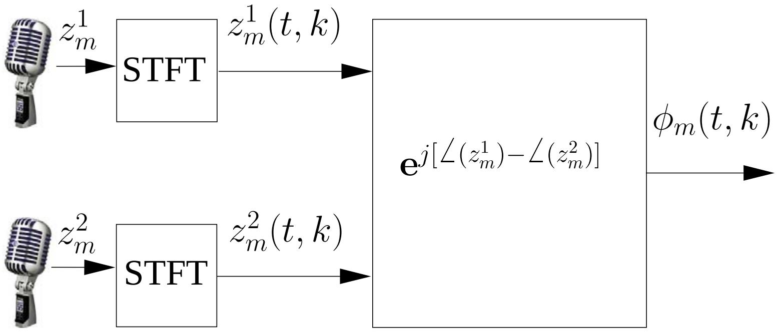
1. Tracking algorithm for moving sources (centralized processing).
2. Localization algorithm for static sources (distributed processing):
   - Constrained communication bandwidth.
   - Limited Computation capabilities at the nodes.



Outline

- Problem formulation & Maximum Likelihood (ML).
- Expectation-Maximization (EM).
- Recursive EM (REM).
- Distributed EM (DEM).
- Simulation results.

Received Data @microphone pair m



- z_m^1 & z_m^2 - Signals @microphone 1 & 2 of node m.
- $z_m^i(t,k) = \sum_{s=1}^{S} a_{sm}^i(t,k) \cdot b_s(t,k) + n_m^i(t,k)$.
- Pair-wise relative complex phase ratio (PRP): $\phi_m(t,k) \triangleq \frac{z_m^1(t,k)}{z_m^2(t,k)} \cdot \frac{|z_m^2(t,k)|}{|z_m^1(t,k)|}$.

Probabilistic Model @node m

Assumptions

- Define a grid of positions in the region of interest: $\mathbf{p} \in \mathcal{P}$.
- TDOA from any grid point to the microphone pair: $\tau_m(\mathbf{p}) \triangleq \frac{||\mathbf{p}-\mathbf{p}_m^2||-||\mathbf{p}-\mathbf{p}_m^1||}{c}$.
- Each T-F bin is solely dominated by one speaker (W-disjoint).

Phase @node m as Mixture of Gaussian (MoG)

$$f(\phi_m) = \prod_{t,k} \sum_{\tau_m} \psi_{\tau_m} \cdot \mathcal{N}^c(\phi_m(t,k); \tilde{\phi}_m^k(\tau_m), \sigma^2)$$

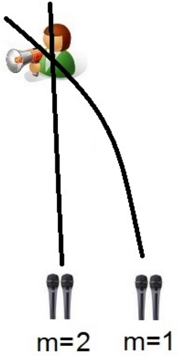
- $\tilde{\phi}_m^k(\mathbf{p})$ - Mean of phase differences pre-calculated for all grid positions $\mathbf{p}$.
- σ^2 - Known and constant variance of the Gaussians.
- ψ_{τ_m} - Probability that $\phi_m \triangleq \text{vec}_{t,k}(\{\phi_m(t,k)\})$ originates from TDOA τ_m.

Probabilistic Model from Array Perspective

Definitions & Relations

- $\phi = \text{vec}_m(\phi_m)$.
- Multiple source positions give rise to the same TDOA.
- $\psi_{\mathbf{p}}$ - Probability that ϕ originates from position $\mathbf{p}$.

$$\psi_{\tau_m} = \int_{\mathbf{p}' \to \tau_m} \psi_{\mathbf{p}'}\, \mathbf{p}' \approx \sum_{\mathbf{p}' \to \tau_m} \psi_{\mathbf{p}'}$$



Augmented Phase as Mixture of Gaussian (MoG)

$$f(\phi) = \prod_{t,k,m} \sum_{\mathbf{p}} \psi_{\mathbf{p}} \cdot \mathcal{N}^c(\phi_m(t,k); \tilde{\phi}_m^k(\tau_m(\mathbf{p})), \sigma^2)$$

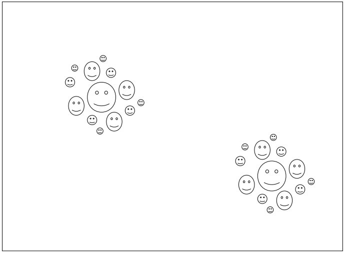
Maximum Likelihood

Straightforward ML

Let $\psi = \text{vec}_{\mathbf{p}}\left(\{\psi_{\mathbf{p}}\}\right)$:

$$f(\phi) = \prod_{t,k,m} \sum_{\mathbf{p}} \psi_{\mathbf{p}} \cdot \mathcal{N}^c(\phi_m(t,k); \tilde{\phi}_m^k(\mathbf{p}), \sigma^2)$$

$$\hat{\psi} = \underset{\psi}{\text{argmax}} \log f(\phi; \psi)$$



Goal

Estimate the most probable grid points that “explains” the received phases.

Iterative Solution using EM [Dempster et al., 1977]

Estimate-Maximize Procedure

- Solving the ML is a cumbersome task.
- Selecting a hidden data $\mathbf{x}$ that can simplify the solution.
- E-step: $Q(\psi|\hat{\psi}^{(\ell-1)}) \triangleq E\left\{\log\left(f(\phi, \mathbf{x}; \psi)\right)|\phi; \hat{\psi}^{(\ell-1)}\right\}$.
- M-step: $\hat{\psi}^{(\ell)} = \operatorname{argmax}_{\psi} Q(\psi|\hat{\psi}^{(\ell-1)})$.

Hidden Data [Mandel et al., 2007, Schwartz and Gannot, 2014]

- $x(t,k,\mathbf{p}) \sim I_{t,k}(\mathbf{p})$ (Speech sparsity assumption)
- $I_{t,k}(\mathbf{p})$ - Indicator that bin (t,k) belongs to a (single) speaker @position $\mathbf{p}$.

Batch EM

E-step

$$\mu^{(\ell-1)}(t,k,\mathbf{p}) \triangleq E\left\{x(t,k,\mathbf{p})|\boldsymbol{\phi}(t,k);\hat{\boldsymbol{\psi}}^{(\ell-1)}\right\}$$
$$=\frac{\hat{\psi}_{\mathbf{p}}^{(\ell-1)}\prod_m \mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2\right)}{\sum_{\mathbf{p}}\hat{\psi}_{\mathbf{p}}^{(\ell-1)}\prod_m \mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\mathbf{p}),\sigma^2\right)}$$

M-step

$$\hat{\psi}_{\mathbf{p}}^{(\ell)}=\frac{\sum_{t,k}\mu^{(\ell-1)}(t,k,\mathbf{p})}{T\cdot K}$$

T : # of frames and K : # of frequencies.

Recuesive EM [Schwartz and Gannot, 2014]

Procedures

- Replace iteration index with time index.
- Execute one iteration per time index.
- Recursively estimate Q [Cappé and Moulines, 2009]:
  - $Q_R(\psi|\psi_R^{(t)}) = Q_R(\psi|\psi_R^{(t-1)}) + \gamma_t \left[Q(\psi|\psi_R^{(t)}) - Q_R(\psi|\psi_R^{(t-1)}) \right].$
  - $\psi_R^{(t+1)} = \text{argmax}_{\psi}\, Q_R(\psi|\psi_R^{(t)}).$
- Maximize using Newton's method [Titterington, 1984] (with constraints [Schwartz and Gannot, 2014]).

Solution (for both recursive procedures!))

$$\psi_R^{(t+1)} = \psi_R^{(t)} + \gamma_t(\psi^{(t+1)} - \psi_R^{(t)})$$

Distributed EM [Dorfan et al., 2014]

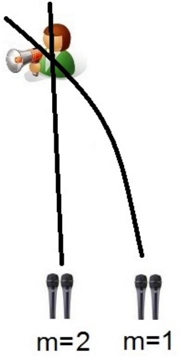
Centralized Computation

- Estimating the global hidden data depends on the availability of all PRPs in one point.
- Requires: powerful fusion center, communication bandwidth, ...

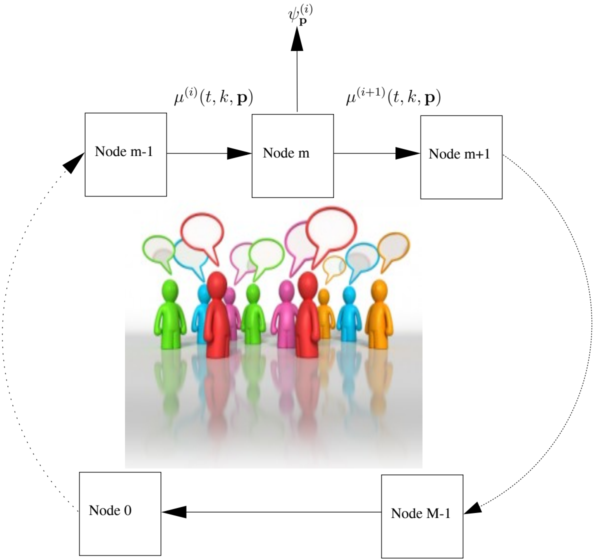
Local Hidden Data ⇔ Global Hidden Data

$$y(t,k,\tau_m(\mathbf{p})) \triangleq I_{t,k,m}(\tau_m(\mathbf{p}))$$
$$x(t,k,\mathbf{p}) \equiv \prod_m y(t,k,\tau_m(\mathbf{p}))$$

Multiple positions $\mathbf{p}$ can induce the same τ_m.



Incremental EM [Neal and Hinton, 1998] - Ring Topology



E-step: Global Hidden

$$\mu^{(i)}(t,k,\mathbf{p}) \triangleq E\left\{x(t,k,\mathbf{p})|\phi(t,k);\psi_{\mathbf{p}}^{(i-1)}\right\}.$$

$$i = (\ell-1)M+m;$$

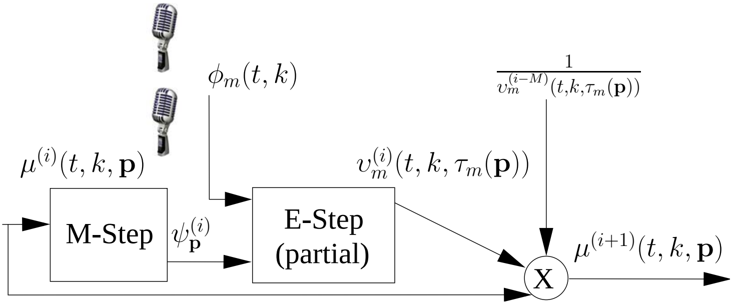
$$m = 0,\ldots,M-1.$$

Becomes sparse after few iterations.

M-Step: Global Parameter Estimation

$$\psi_{\mathbf{p}}^{(i)} = \frac{\sum_{t,k}\mu^{(i)}(t,k,\mathbf{p})}{T\cdot K}$$

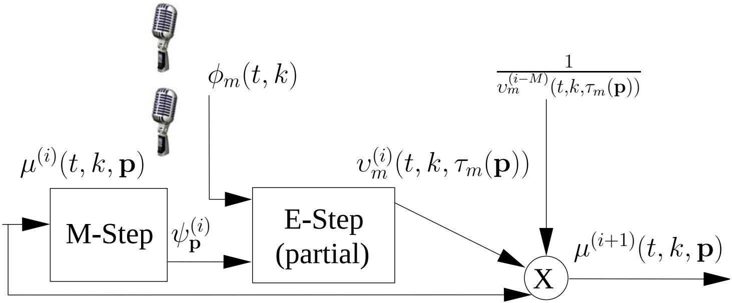
Increment @Node m



M-Step: Global Parameter Estimation (Reminder)

$$\psi_{\tau_m(\mathbf{p})}^{(i)} \triangleq \int_{\mathbf{p}'\to\tau_m(\mathbf{p})} \psi_{\mathbf{p}'}^{(i)}\, d\mathbf{p}'$$

Increment @Node m

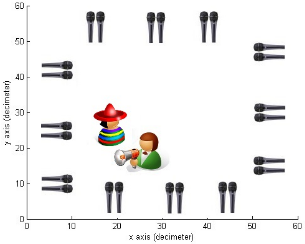


E-step: Local Hidden

$$\upsilon_m^{(i)}(t,k,\tau_m(\mathbf{p})) \triangleq E\left\{y(t,k,\tau_m(\mathbf{p}))|\phi_m(t,k);\psi_{\mathbf{p}}^{(i)}\right\}$$

$$= \frac{\psi_{\tau_m(\mathbf{p})}^{(i)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\tau_m(\mathbf{p})),\sigma^2\right)}{\sum_{\tau_m(\mathbf{p})}\psi_{\tau_m(\mathbf{p})}^{(i)}\mathcal{N}^c\left(\phi_m(t,k);\tilde{\phi}_m^k(\tau_m(\mathbf{p})),\sigma^2\right)}$$

Simulation Setup



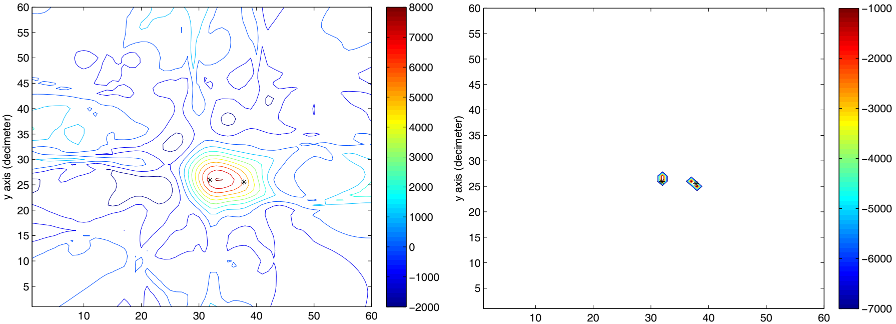
Tracking

- 2D setup: 10×10 cm grid.
- Trajectory: line, arc.
- 12 nodes.
- Inter-microphone pair: 20 cm.
- $T_{60} = 0.7$ Sec.
- Performance criterion: curve fit.

Distributed Localization

- 2D setup: 10×10 cm grid.
- Randomly located sources.
- 12 nodes.
- Inter-microphone pair: 50 cm.
- $T_{60} = 0.3$ Sec.
- Performance criteria:
  - Detection rate.
  - False Alarm (FA) rate.
  - Mean Square Error (MSE).

Simulation Results: Distributed EM

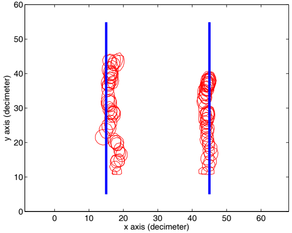


(a) Delay & Sum BF (b) Distributed EM

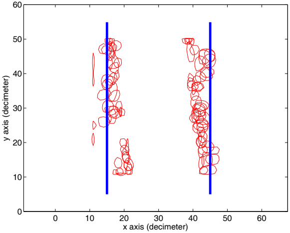
# Sources	Detection[%]	FA[%]	MSE[cm]
1	100	22	3.9
2	98	6.5	7.1

Table : Results for 100 Monte-Carlo simulations

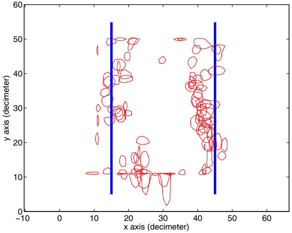
Simulation Results: Recursive EM



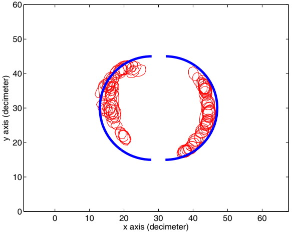
(c) $\gamma = 0.1$



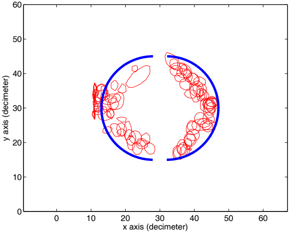
(d) $\gamma = 0.5$



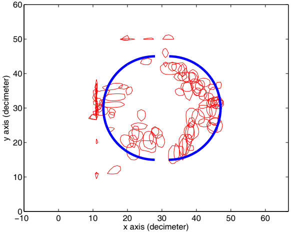
(e) $\gamma = 1$



(f) $\gamma = 0.1$



(g) $\gamma = 0.5$



(h) $\gamma = 1$

Summary

Recursive EM Algorithm for Tracking

1. Speech sparsity utilized to derive EM-based Localization.
2. Two versions of tracking algorithms were proposed based on [Cappé and Moulines, 2009],[Titterington, 1984].
3. A Constrained version of [Titterington, 1984] was derived.

Distributed EM Algorithm for Localization

1. No central processing unit required.
2. Decomposing the global hidden data to local hidden data is the key step in distributed algorithm derivation.
3. Detection and localization of multiple concurrent sources with minimal a priori information.
4. Only two global iterations required in our simulations.
5. No significant dependency on initial conditions observed.

References I

Cappé, O. and Moulines, E. (2009).
On-line expectationmaximization algorithm for latent data models.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(3):593–613.

Dempster, A., Laird, N., and Rubin, D. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

Dorfan, Y., Hazan, G., and Gannot, S. (2014).
Multiple acoustic sources localization using distributed Expectation-Maximization algorithm.
In *The 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France.
best student paper award.

Mandel, M., Ellis, D., and Jebara, T. (2007).
An EM algorithm for localizing multiple sound sources in reverberant environments.
Advances in Neural Information Processing Systems, 19:953.

Neal, R. and Hinton, G. (1998).
A view of the EM algorithm that justifies incremental, sparse, and other variants.
Learning in graphical models, 89:355–368.

Schwartz, O. and Gannot, S. (2014).
Speaker tracking using recursive EM algorithms.
IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(2):392–402.

Titterington, D. (1984).
Recursive parameter estimation using incomplete data.
J. Roy. Statist. Soc. Ser. B, 46:257–267.