AN ONLINE MULTIPLE-SPEAKER DOA TRACKING USING THE CAPPÉ-MOULINES RECURSIVE EM ALGORITHM

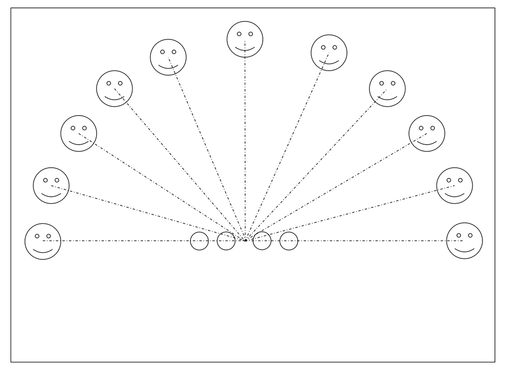
Koby Weisberg[1], *Ofer Schwartz*[2], *Sharon Gannot*[1]

[1] Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel
[2] CEVA-DSP Audio Department, Herzelia, Israel

Overview

- Multiple-speaker direction of arrival (DOA) tracking algorithm using microphone array based on the recursive EM (REM) paradigm proposed by Cappé and Moulines
- Using the Fisher-Neyman factorization, the scalar outputs of the minimum variance distortionless response (MVDR)-beamformer (BF) are shown to be the sufficient statistics for estimating the parameters

Problem formulation



- M : Number of DOA candidates
- s_m : mth speech candidate
- $\mathbf{v}$: Additive noise
- $\mathbf{g}_m$: steering vector of the mth candidate
- $\mathbf{z}$: Mixed signal
- d_m : Indicator for the active speaker

$$\mathbf{z}(t,k) = \sum_{m=1}^{M} d_m(t,k)\mathbf{g}_m(k)s_m(t,k) + \mathbf{v}(t,k)$$

Statistical Model

- Speech and noise distribution:

$$\mathbf{v}(t,k) \sim \mathcal{N}(\mathbf{v}(t,k), \mathbf{0}, \Phi_{\mathbf{v}}(k))$$
$$s_m(t,k) \sim \mathcal{N}(s_m(t,k), 0, \phi_{s,m}(t,k))$$

- The observation vectors are distributed as a mixture of M Gaussians:

$$P(\mathbf{z}(t,k)) = \sum_{m=1}^{M} \psi_m \mathcal{N}(\mathbf{z}(t,k), \mathbf{0}, \Phi_{\mathbf{z},m}(t,k))$$

where:

$$\Phi_{\mathbf{z},m}(t,k) = \mathbf{g}_m(k)\mathbf{g}_m^H(k)\phi_{s,m}(t,k) + \Phi_{\mathbf{v}}(k)$$

- $\phi_{s,m}(t,k)$ and ψ_m are the unknown parameters
- Distribution of the entire observation set:

$$P(\mathbf{z};\boldsymbol{\theta}) = \prod_{t,k}\sum_{m=1}^{M} \psi_m \mathcal{N}(\mathbf{z}(t,k), \mathbf{0}, \Phi_{\mathbf{z},m}(t,k))$$

- The maximum likelihood (ML) problem:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log P(\mathbf{z};\boldsymbol{\theta})$$

Factorization

- The MVDR-BF $\mathbf{w}_m = \frac{\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}{\mathbf{g}_m^H\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}$
- MVDR output power: $|\hat{s}_{m,\text{MVDR}}(t)|^2 \equiv |\mathbf{w}_m^H\mathbf{z}(t)|^2$
- The PSD of the noise: $\phi_{v,m} \equiv \frac{1}{\mathbf{g}_m^H\Phi_{\mathbf{v}}^{-1}\mathbf{g}_m}$
- Using F-N factorization, and several algebraic steps:

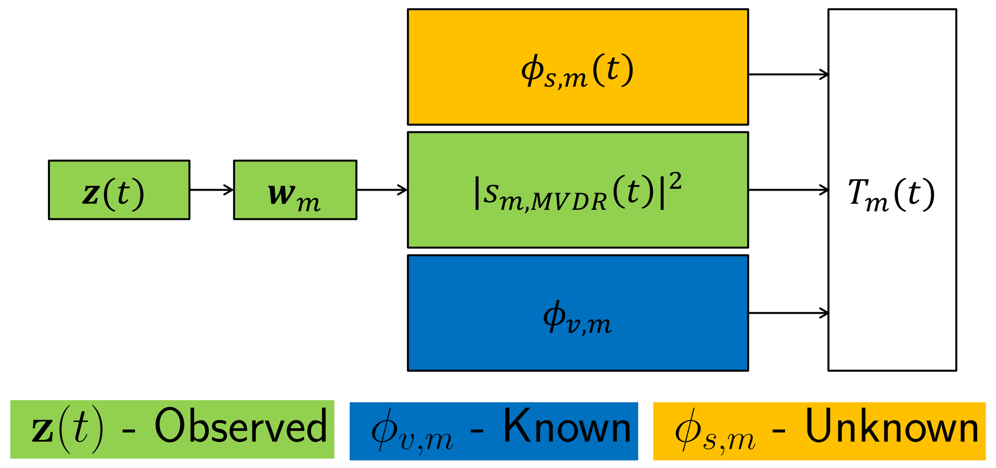
$$\mathcal{N}(\mathbf{z}(t), \mathbf{0}, \Phi_{\mathbf{z},m}(t)) = T_m(t;\phi_{s,m}(t))G(t)$$

- $G(t) = \mathcal{N}(\mathbf{z}(t), \mathbf{0}, \Phi_{\mathbf{v}})$ is independent of m

$$T_m(t;\phi_{s,m}(t)) = \frac{1}{1+\text{SNR}_m^{\text{pri}}(t)}\exp\left(\frac{\text{SNR}_m^{\text{pri}}(t)\text{SNR}_m^{\text{post}}(t)}{1+\text{SNR}_m^{\text{pri}}(t)}\right)$$

$$\text{SNR}_m^{\text{pri}}(t) = \frac{\phi_{s,m}(t)}{\phi_{v,m}} \qquad \text{SNR}_m^{\text{post}}(t) = \frac{|\hat{s}_{m,\text{MVDR}}(t)|^2}{\phi_{v,m}}$$

- $T_m(t)$ is a likelihood ratio test (LRT) to determine whether the m-th candidate direction is dominated by the signal or the noise



Batch EM algorithm

- Define $d_m(t,k)$ as the hidden data set
- The E-step:

$$\hat{d}_m^{(\ell-1)}(t) = \frac{\hat{\psi}_m^{(\ell-1)}T_m(t;\hat{\phi}_{s,m}^{(\ell-1)}(t))G(t)}{\sum_m \hat{\psi}_m^{(\ell-1)}T_m(t;\hat{\phi}_{s,m}^{(\ell-1)}(t))G(t)}$$

- $\hat{\psi}_m^{(\ell-1)}$ - Prior probability of the mth candidate
- $T_m(t;\hat{\phi}_{s,m}^{(\ell-1)}(t))$ - Is the m-th direction dominated by either speech signal or noise?
- The M-step:

$$\hat{\psi}_m^{(\ell)} = \frac{\sum_{t,k}\hat{d}_m^{(\ell-1)}(t,k)}{T\cdot K}$$
$$\hat{\phi}_{s,m}(t,k) = |\hat{s}_{m,\text{MVDR}}(t,k)|^2 - \phi_{v,m}(k).$$

$\Longrightarrow$ A priori and a posteriori SNRs are related:

$$\text{SNR}_m^{\text{pri}}(t) = \text{SNR}_m^{\text{post}}(t) - 1$$

- Equivalent LRT:

$$T_m(t;\hat{\phi}_{s,m}(t)) = \frac{1}{\text{SNR}_m^{\text{post}}(t)}\exp\left(\text{SNR}_m^{\text{post}}(t) - 1\right)$$

Recursive EM

- To allow for a smooth estimate of the speech power spectral density (PSD), we introduce time-dependency between frames, i.e. $\hat{\phi}_s(t)$ depends on a set of frames
- The (smooth) time-variations of the speech PSD will be naturally obtained by the recursive nature of the algorithm
- Applying batch EM with this assumption would yield:

$$\hat{\phi}_{s,m}^{(\ell)}(t) = \frac{\sum_{t'}\hat{d}_m^{(\ell-1)}(t')\,|\hat{s}_{m,\text{MVDR}}(t')|^2}{\sum_{t'}\hat{d}_m^{(\ell-1)}(t')} - \phi_{v,m}$$

- Use the Cappé-Moulines variant of the recursive EM for online parameter estimation:

$$Q_R(t;\boldsymbol{\theta}) = (1-\gamma)Q_R(t;\boldsymbol{\theta}) + \gamma Q(\boldsymbol{\theta}|\boldsymbol{\theta}(t-1))$$

- The E-step:

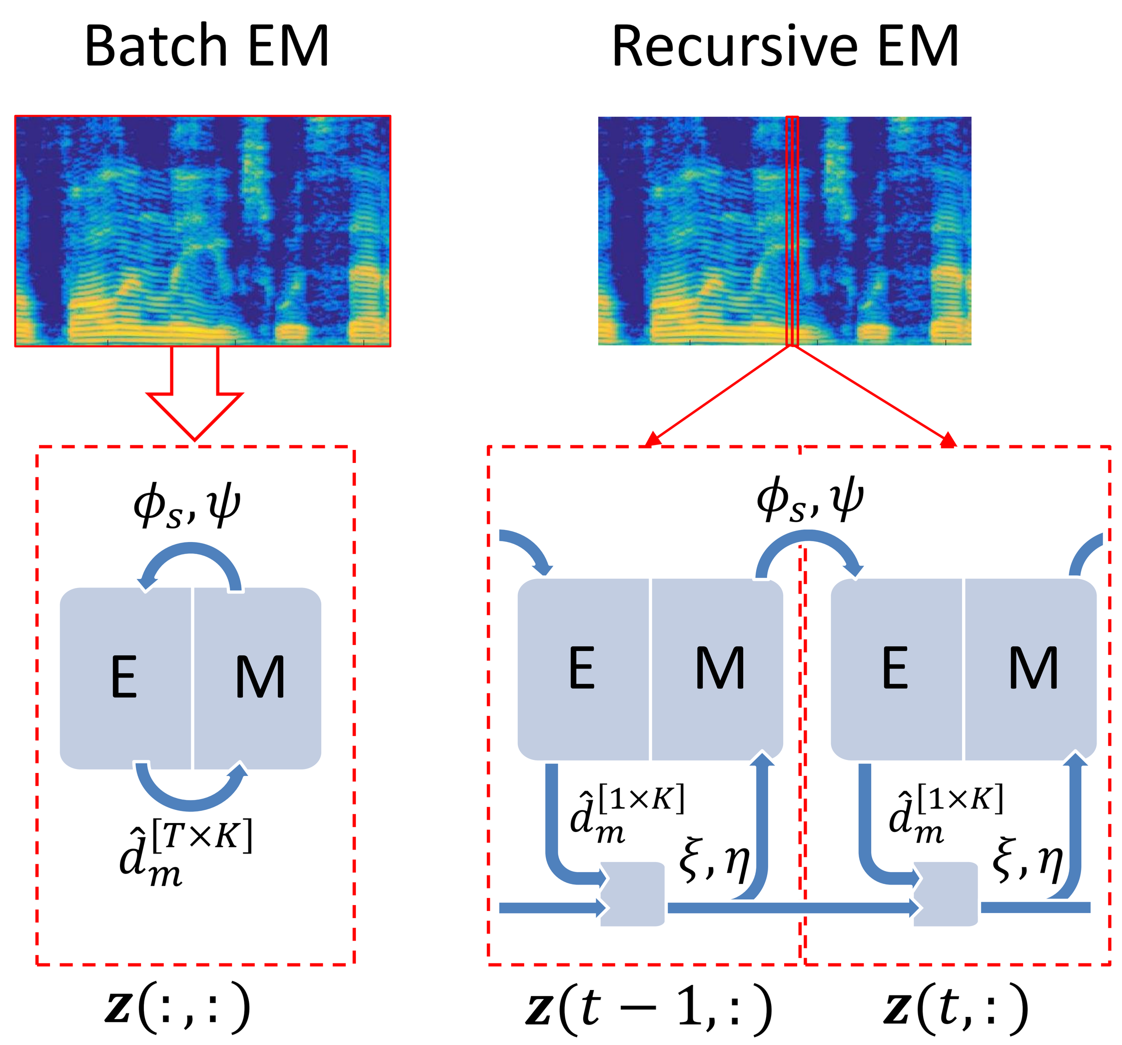
$$\hat{d}_m(t) = \frac{\hat{\psi}_m(t-1)T_m(t;\hat{\phi}_{s,m}(t-1))}{\sum_m \hat{\psi}_m(t-1)T_m(t;\hat{\phi}_{s,m}(t-1))}$$

- The recursive M-step:

$$\eta_m(t) = (1-\gamma)\eta_m(t-1) + \gamma\hat{d}_m(t)$$
$$\xi_m(t) = (1-\gamma)\xi_m(t-1) + \gamma\hat{d}_m(t)\,|\hat{s}_{m,\text{MVDR}}(t)|^2$$

$$\hat{\psi}_m(t) = \frac{\sum_k \eta_m(t,k)}{K}$$
$$\hat{\phi}_{s,m}(t,k) = \frac{\xi_m(t,k)}{\eta_m(t,k)} - \phi_{v,m}(k)$$

- Smoothness control by γ



Experimental Study: Simulation

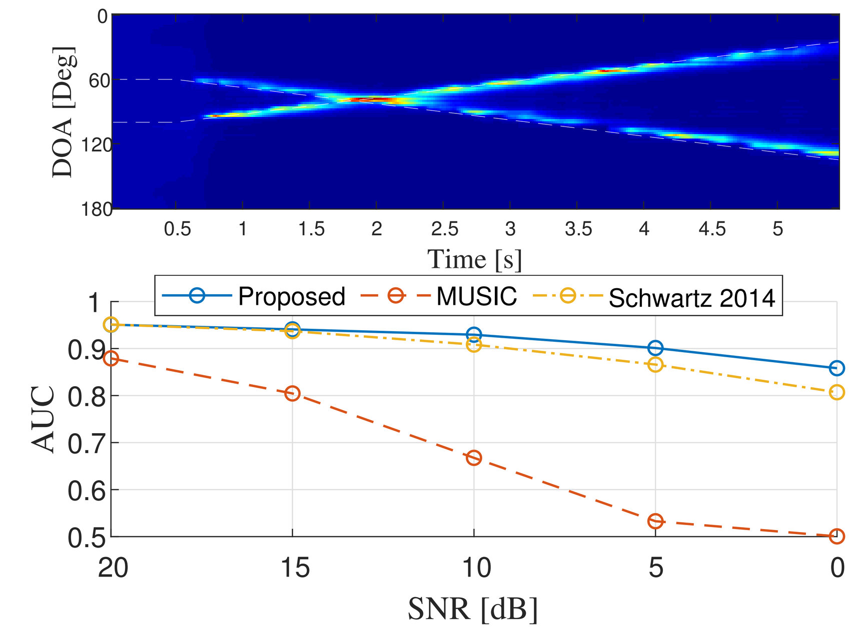
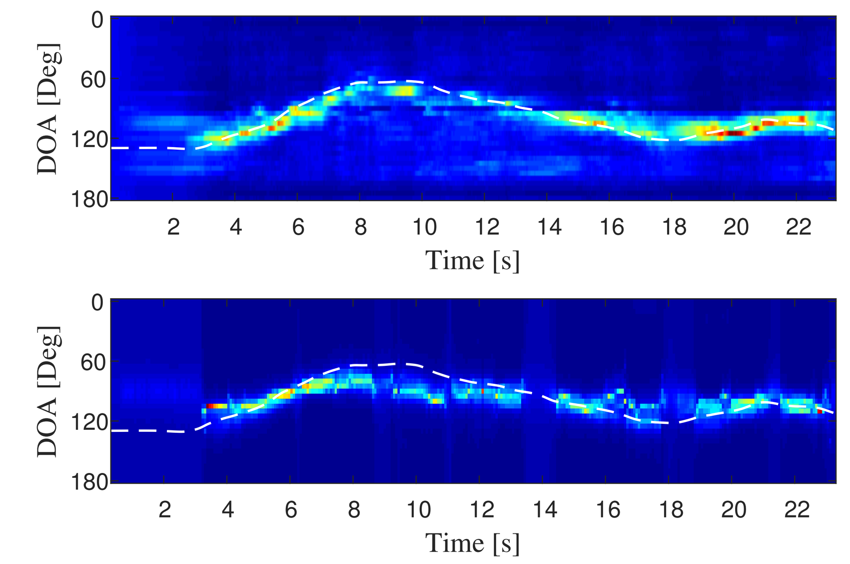
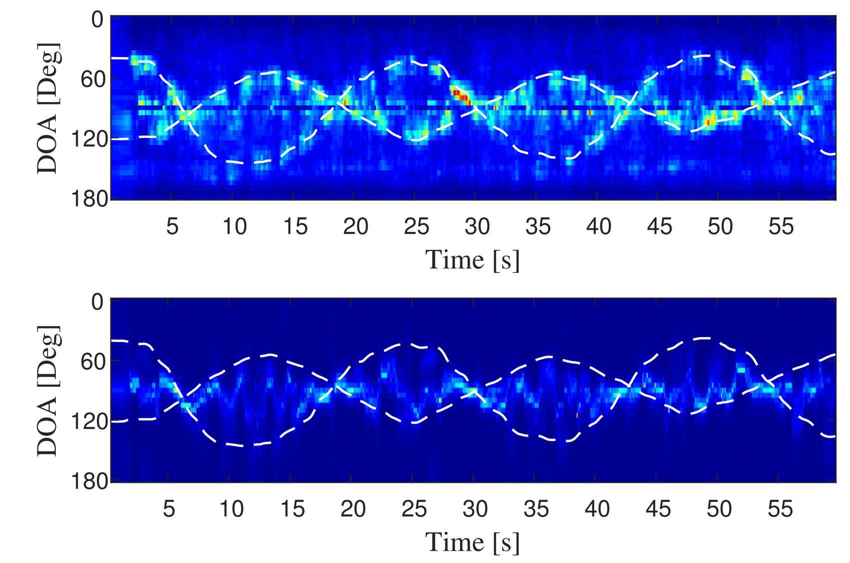


Figure: The area under the ROC curve (AUC). Simulated two moving sources. Detection in the range around the true DOA is considered true positive. Results averaged over 30 Monte-Carlo trials. One trial depicted on the top panel.

Experimental Study: LOCATA Challenge



(a) Single speaker. Proposed (top) and MUSIC (bottom)



(b) Two speakers. Proposed (top) and MUSIC (bottom)

Summary

- ✓ A computationally efficient tracking algorithm based on Cappé- Moulines REM
- ✓ Set of MVDR outputs as features
- ✓ High tracking capabilities, compared with baseline methods