SOURCE COUNTING IN SPEECH MIXTURES USING A VARIATIONAL EM APPROACH FOR COMPLEX WATSON MIXTURE MODELS

UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Computer Science, Electrical Engineering and Mathematics
NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

Lukas Drude, Aleksej Chinaev, Dang Hai Tran Vu, Reinhold Haeb-Umbach, University of Paderborn, Germany
{drude, chinaev, tran, haeb}@nt.uni-paderborn.de, http://www-nt.uni-paderborn.de

Introduction

- Source counting treated as model selection problem
- Variational EM derived for directional statistics
- Source counting by thresholding VEM results and iteratively removing sources from the observations

Feature extraction

- Convolutive mixture in STFT domain:

$$\mathbf{X}(t,f) = \sum_{k=1}^{K} \mathbf{H}_k(f) S_k(t,f) + \mathbf{N}(t,f)$$

- Phase and frequency normalized observations:

$$\tilde{X}_d(t,f) = |X_d(t,f)| \exp\left(j \frac{\arg\left(X_d(t,f) X_1^*(t,f)\right)}{4f/F f_s c^{-1} d_{\max}} \right)$$

- Direction vector (amplitude normalized):

$$\mathbf{Y}(t,f) = \tilde{\mathbf{X}}(t,f) / \left\| \tilde{\mathbf{X}}(t,f) \right\|$$

Statistical modeling

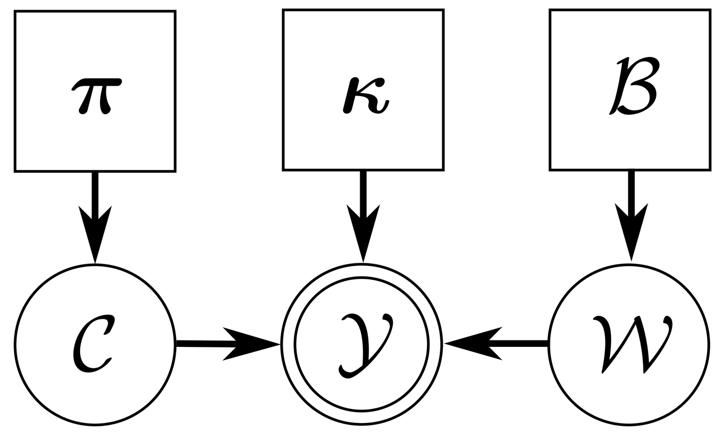
- Observations $\mathcal{Y} = \{\mathbf{Y}(t,f)\}$ form clusters on a complex unit-hypersphere $\rightarrow$ Complex Watson mixture model

$$p(\mathbf{Y}(t,f)|\mathcal{W},\boldsymbol{\kappa}) = \sum_{k=1}^{K} \pi_k \frac{1}{c_W(\kappa_k)} e^{\kappa_k \left|\mathbf{W}_k^H \mathbf{Y}(t,f)\right|^2}$$

$$p(\mathcal{Y},\mathcal{C},\mathcal{W}) = p(\mathcal{Y}|\mathcal{C},\mathcal{W};\boldsymbol{\kappa}) p(\mathcal{C};\boldsymbol{\pi}) p(\mathcal{W};\mathcal{B})$$

$$p(\mathbf{C}(t,f);\boldsymbol{\pi}) = \prod_{k=1}^{K+1} \pi_k^{c_k(t,f)}$$

$$p(\mathcal{W};\mathcal{B}) = \prod_{k=1}^{K+1} \frac{1}{c_B(\mathbf{B}_k)} e^{\mathbf{W}_k^H \mathbf{B}_k \mathbf{W}_k}$$



Step-wise deletion algorithm

1: Calculate $A^{(1)}(t,f) = \|\tilde{\mathbf{X}}(t,f)\|$
2: **for** $\nu = 1 \dots \nu_{\max}$ **do**
3: **if** $\nu > 1$: **then** Apply re-weighting equation **end if**
4: Select observations $\mathcal{Y}^{(\nu)}$ with $A^{(\nu)} >$ quantile(q)
5: Use VEM algorithm with $\mathcal{Y}^{(\nu)}$ to calculate $\mathbf{B}_\nu$, κ_ν
6: Calculate principal component $\mathbf{W}_\nu = \mathcal{P}(\mathbf{B}_\nu)$
7: **end for**
8: Calculate angular distance $s_\nu = \max_{\nu'=1\dots\nu-1} |\mathbf{W}_\nu^H \mathbf{W}_{\nu'}| \forall \nu = 2 \dots \nu_{\max}$
9: Count iterations where $\kappa_\nu > \kappa_{\text{Th}} \wedge \pi_\nu > \pi_{\text{Th}} \wedge s_\nu < s_{\text{Th}}$

Variational EM for a cWMM

E-step: Class responsibilities:

$$\ln \gamma_k^{(i)}(t,f) = \kappa_k^{(i-1)} \mathbb{E}_{\mathbf{W}_k}\left\{ \mathbf{W}_k^H \mathbf{Y}(t,f) \mathbf{Y}^H(t,f) \mathbf{W}_k \right\} - \ln M(1,D,\kappa_k^{(i-1)}) + \ln(\pi_k^{(i-1)}) + \text{const.}$$

Mixture weights:

$$\pi_k^{(i)} = N_k^{(i)} / \sum_{k=1}^{K+1} N_k^{(i)}, \quad N_k^{(i)} = \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_k^{(i)}(t,f)$$

Bingham parameter matrix:

$$\mathbf{B}_k^{(i)} = \kappa_k^{(i-1)} N_k^{(i)} \boldsymbol{\Phi}_{YY,k}^{(i)} + \mathbf{B}_{0,k},$$

$$\boldsymbol{\Phi}_{YY,k}^{(i)} = \frac{1}{N_k^{(i)}} \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_k^{(i)}(t,f) \mathbf{Y}(t,f) \mathbf{Y}^H(t,f)$$

M-step: Concentration parameter:

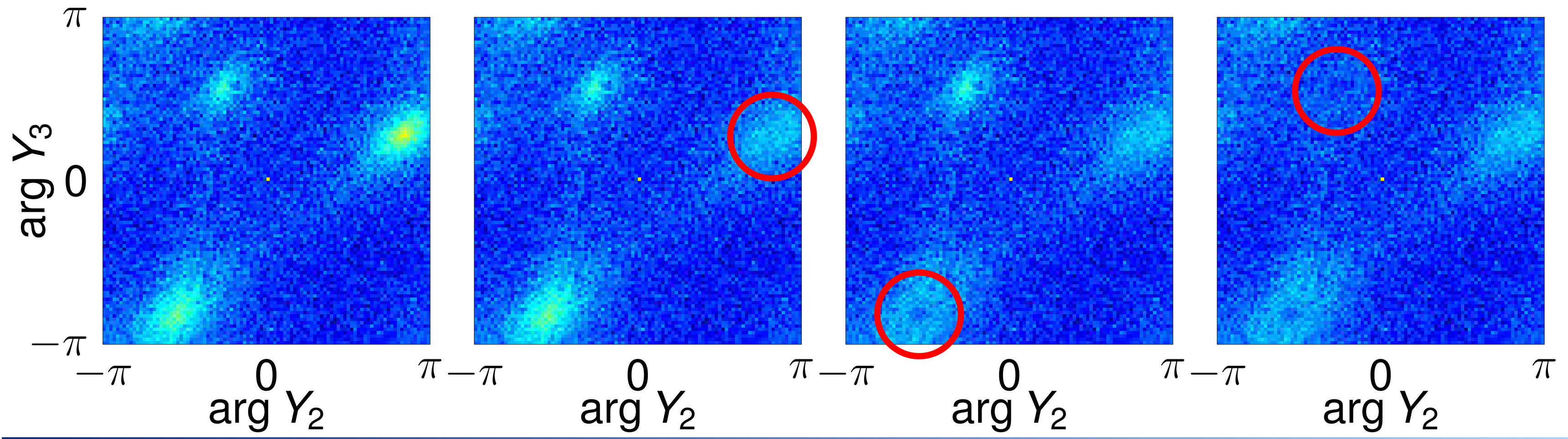
$$\frac{M(2, D+1, \kappa_k^{(i)})}{D \cdot M(1, D, \kappa_k^{(i)})} = \mathbb{E}_{\mathbf{W}_k}\left\{ \mathbf{W}_k^H \boldsymbol{\Phi}_{YY,k}^{(i)} \mathbf{W}_k \right\}$$

Re-weighting equation

Already found speakers are removed from the signal by reweighting depending on the distance to the estimated mode vector:

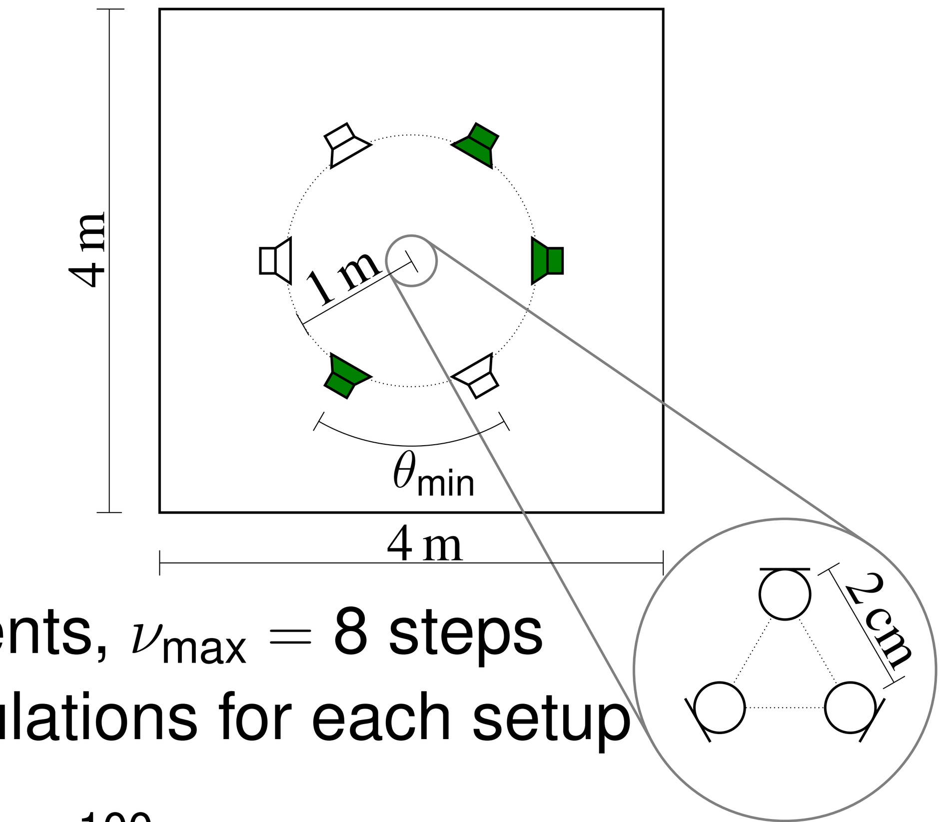
$$A^{(\nu)}(t,f) = A^{(\nu-1)}(t,f) \cdot \left(1 - e^{\left|\hat{\mathbf{W}}_{\nu-1}^H \mathbf{Y}(t,f)\right| - 1} \right)$$

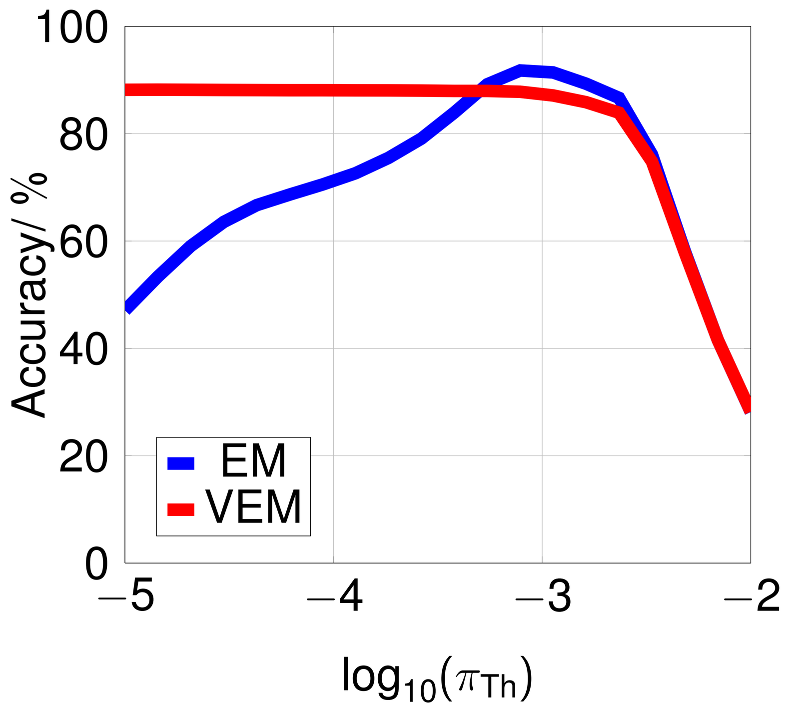
Visualization

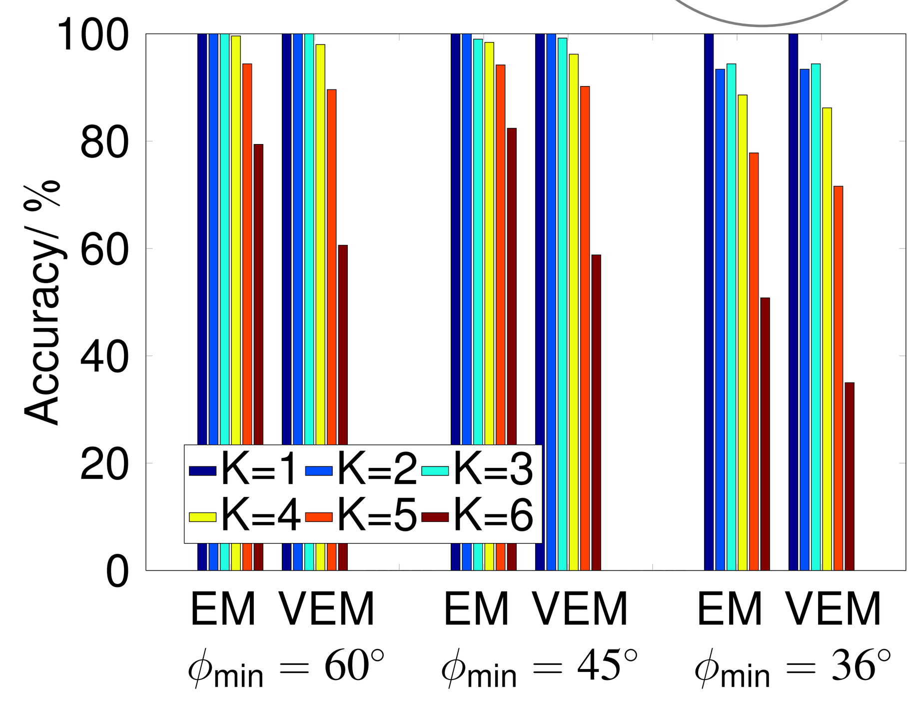


Experimental setup and results

- White Gaussian noise at 10 dB SNR
- Uninformative Bingham prior $\mathbf{B}_0^\nu = \mathbf{0}$
- $q = 90\,\%$, $s_{\text{Th}} = 0.7$, $\kappa_{\text{Th}} = 1$, $\pi_{\text{Th}} = 10^{-3}$
- $K = 2$ mixture components, $\nu_{\max} = 8$ steps
- Averages over 500 simulations for each setup







Conclusions

- Variational EM for complex Watson mixture models
- Stepwise deletion algorithm is able to count sources
- Uninformative Bingham prior increases robustness
- Key theoretical result: Analytical solution for $E_{\mathbf{W}_k}\left\{ \mathbf{W}_k^H \boldsymbol{\Phi}_{YY,k}^{(i)} \mathbf{W}_k \right\}$

Work supported by DFG contract no. Ha3455/8-1.