UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Computer Science, Electrical Engineering and Mathematics

NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

TOWARDS ONLINE SOURCE COUNTING IN SPEECH MIXTURES APPLYING A VARIATIONAL EM FOR COMPLEX WATSON MIXTURE MODELS

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 a model selection problem
- Directions learned by a complex Watson mixture model
- Observation selection based on power-quantile
- Comparison with DoA-based variational EM algorithm
- Proof of concept for an online algorithm

Modeling and feature extraction

Convolutive mixture model:

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

Phase, frequency and unit-norm normalization:

$$\tilde{X}_d(t,f) = |X_d(t,f)| \exp\left(\mathrm{j}\frac{\arg\left(X_d(t,f)X_1^*(t,f)\right)}{2 f_{\mathrm{real}} c^{-1} d_{\max}}\right)$$

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

$\Rightarrow$ Phase solely determined by source position

Statistical model

Complex Watson mixture model:

$$p(\mathbf{Y}|\mathbf{W}_{1:K+1}; \pi_{1:K+1}, \kappa_{1:K+1}) = \sum_{k=1}^{K+1} P(c(t,f)=k; \pi_{1:K+1}) \frac{1}{c_W(\kappa_k)} \mathrm{e}^{\kappa_k \left|\mathbf{W}_k^{\mathrm{H}} \mathbf{Y}(t,f)\right|^2}$$

$$P(c(t,f)=k; \pi_{1:K+1}) = \text{Categorical distribution}$$

$$p(\mathbf{W}_k; \mathbf{B}_k) = \text{Complex Bingham distribution as prior}$$

Arguments for a complex Watson mixture model:

- All spatial information preserved in observations
- A priori distribution available
  $\Rightarrow$ Variational EM (VEM) developed
- Distance measure $\mathbf{W}^{\mathrm{H}}\mathbf{Y}$ resembles a spatial correlation in the beamforming concept

Offline algorithm

- First iteration ($\nu = 1$), Quantile criterion:

$$A^{(1)}(t,f) = \begin{cases} \frac{1}{2} + \frac{1}{2}a, & \mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f) > P, \\ \frac{1}{2} - \frac{1}{2}a, & \mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f) < P, \end{cases} \quad P = \mathrm{quantile}(\mathbf{X}^{\mathrm{H}}(t,f)\mathbf{X}(t,f), q)$$

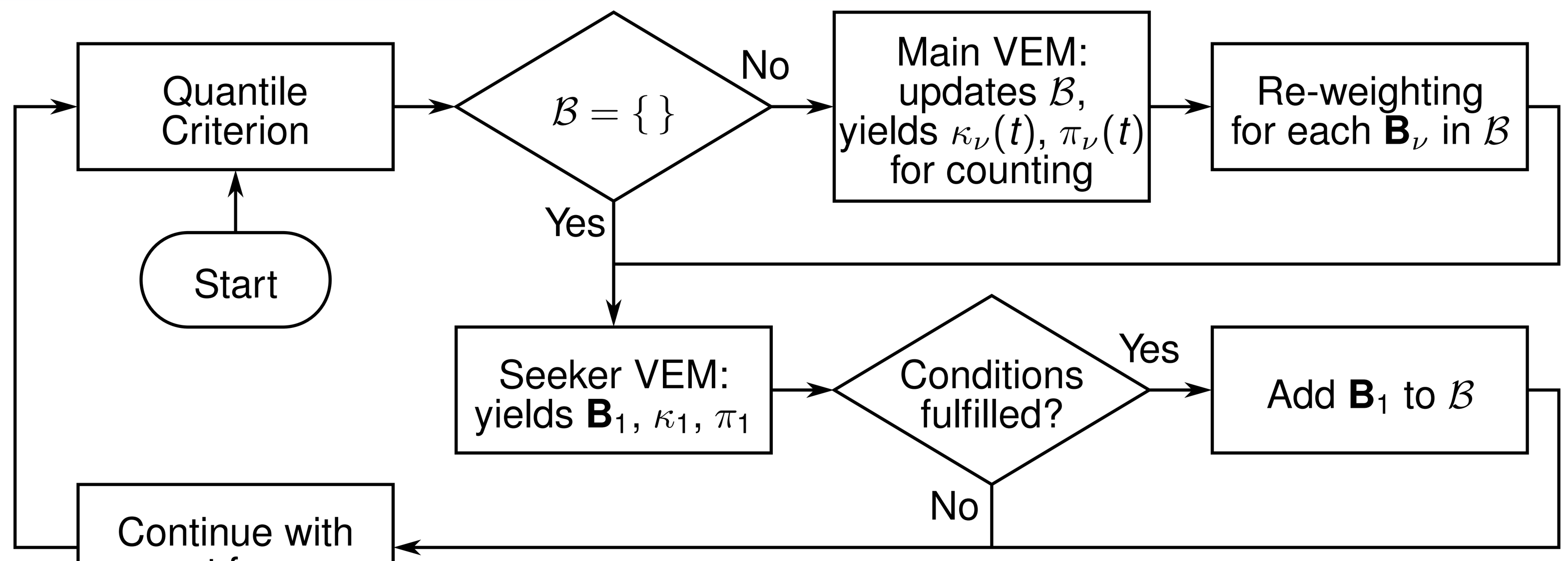
  $\Rightarrow$ Emphasize observations containing a dominant source
- Next iterations, updating observation weights:

$$A^{(\nu+1)}(t,f) = A^{(\nu)}(t,f)\left(1 - \mathrm{e}^{\kappa_{\mathrm{Re}}\left(\left|\hat{\mathbf{W}}_\nu^{\mathrm{H}}\mathbf{Y}(t,f)\right|^2 - 1\right)}\right)$$

  $\Rightarrow$ Deemphasize observations related to detected sources
- Learn one source and noise component for each ν:

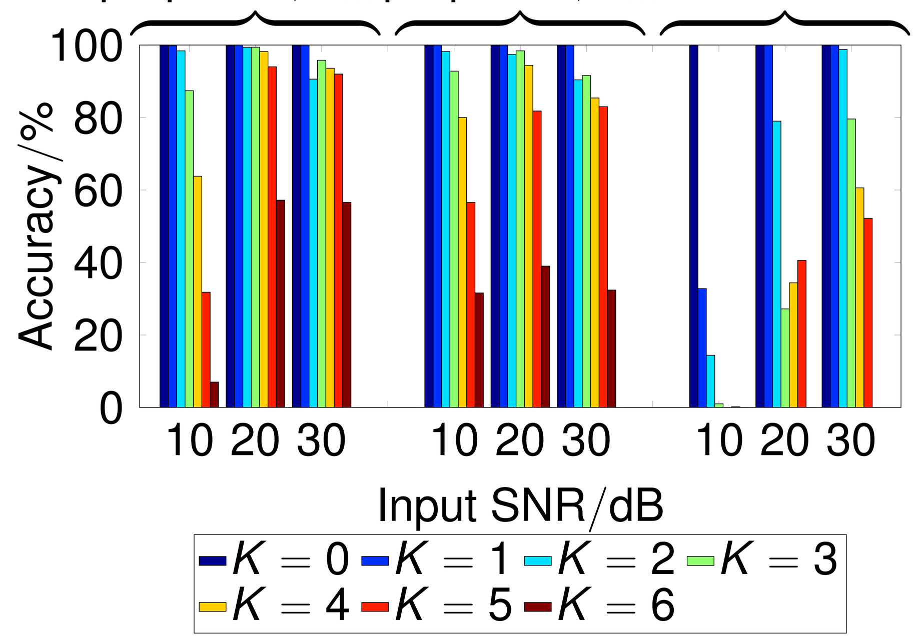
```
Calculate A^(1)(t,f)
for ν = 1, ..., ν_max do
    Use VEM with Y(t,f) and A^(ν)(t,f) to get B_ν and κ_ν
    Calculate principal component W_ν = P(B_ν)
    if ν < ν_max: then Reweight observations end if
end for
Calculate s_1 = 0, s_ν = max_{ν'=1...ν-1} |W_ν^H W_ν'| ∀ν ≠ 1
Count iterations where κ_ν > κ_Th ∧ s_ν < s_Th
```

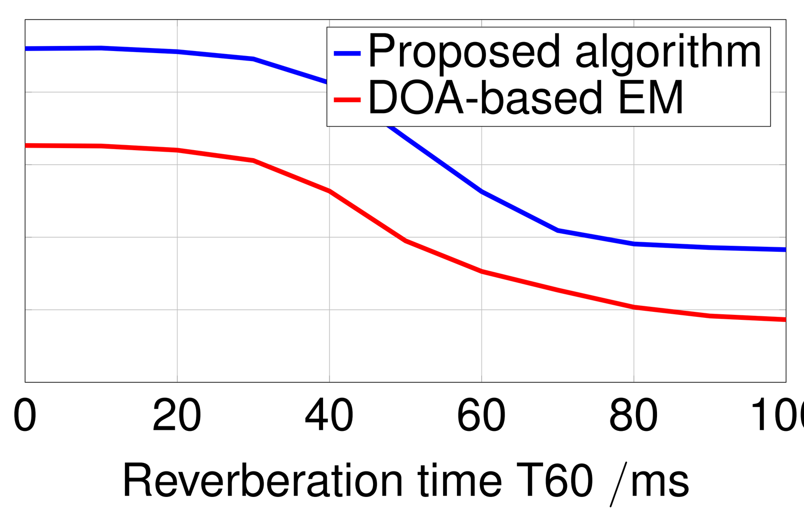
Online algorithm



- Frame-wise decision with low latency
- Seeker VEM needs one frame to find a candidate
- Main VEM needs one frame to validate a source
- Conditions to accept a new speaker similar to offline case

Offline results

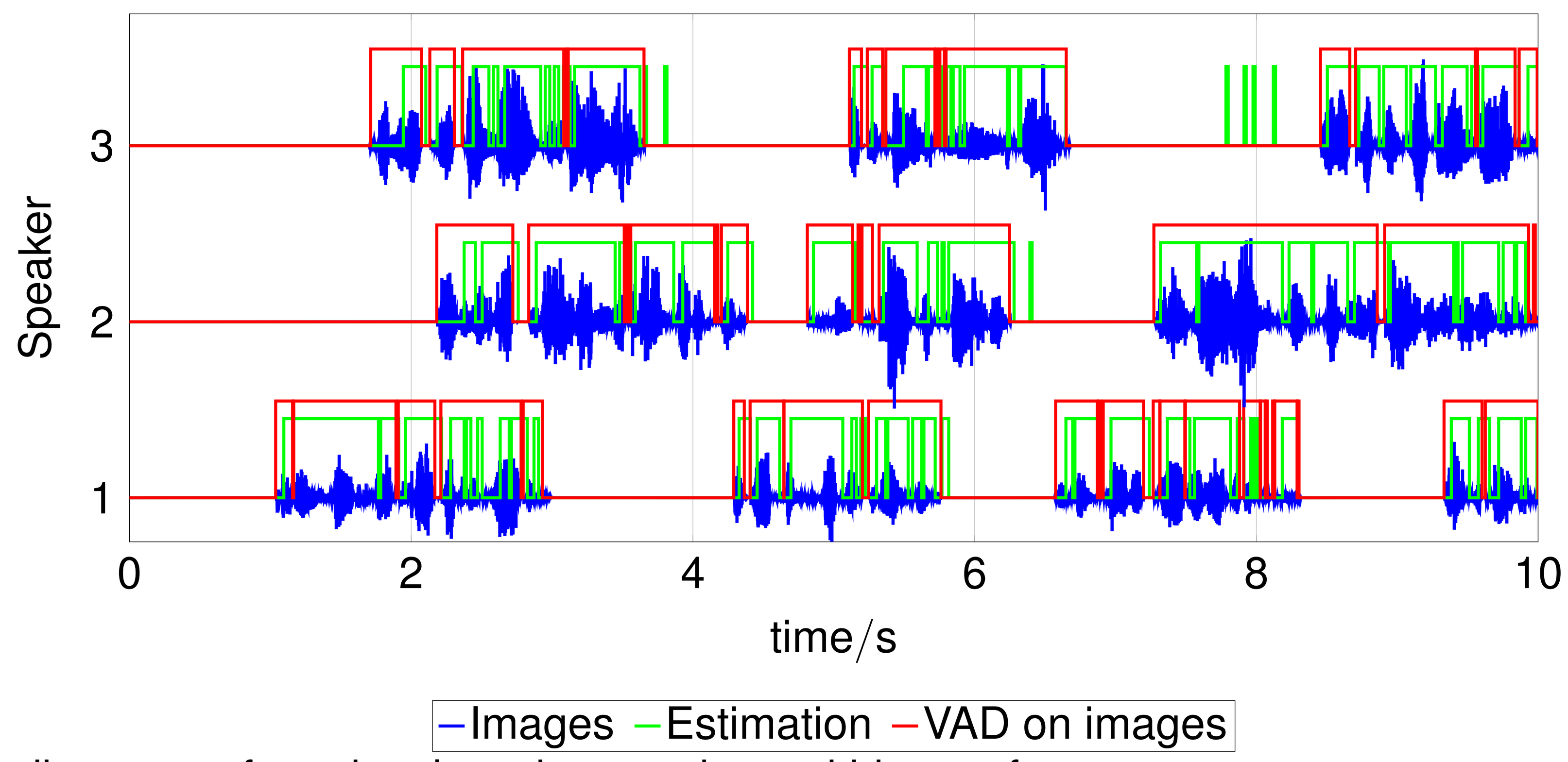




- Simulated room with image method
- White Gaussian sensor noise
- Uninformative complex Bingham prior

- Comparison: DoA-based VEM
- Proposed algorithm more noise robust
- Both algorithms suffer from reverberation

Online results



- Rediscovery of previously active speakers within one frame
- Diarization error rate 30 % of maximum of 187 %

Conclusions

- Robust observations emphasized
- Initialization problem relaxed by searching for single speaker at a time
- Low latency online algorithm
- Susceptible to reverberation because of frequency normalization $\Rightarrow$ Avoid frequency normalization at the cost of introducing the permutation problem.

IWAENC 2014, Juan les Pins

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