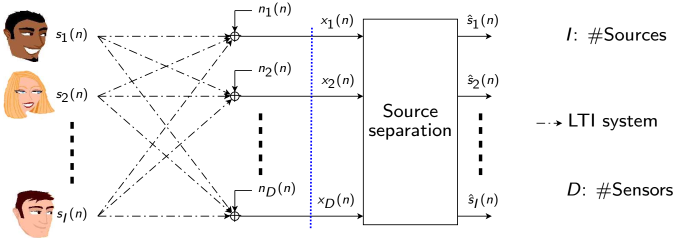
Blind Speech Separation Exploiting Temporal and Spectral Correlations Using 2D-HMMS

Dang Hai Tran Vu and Reinhold Haeb-Umbach

EUSIPCO 2013

Frequency domain blind source separation



Short-time Fourier-transform (STFT) domain mixing model

$$\mathbf{X}(m,k) \approx \sum_{i=1}^{I} \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k)$$

where m: frame index, k: frequency bin

Sparseness-based BSS

Generative model in STFT domain

- At most one source dominant in each time-frequency slot (m, k)
- Generative model:

$$\mathbf{X}(m,k) = \begin{cases} \mathbf{N}(m,k) & \text{if} \quad Z(m,k) = 0 \\ \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k) & \text{if} \quad Z(m,k) \in \{i; i = 1, \ldots, I\} \end{cases}$$

- $Z(m, k)$ hidden random variable indicating which source is active

Goal: Extract $S_i(m, k)$ solely from $\mathbf{X}(m, k)$

- By computation of source activity probability

$$P\left(Z(m,k) \middle| \mathbf{X}(m,k)\right)$$

Sparseness-based BSS

Generative model in STFT domain

- At most one source dominant in each time-frequency slot (m, k)
- Generative model:

$$\mathbf{X}(m,k) = \begin{cases} \mathbf{N}(m,k) & \text{if} \quad Z(m,k) = 0 \\ \mathbf{H}_i(k) S_i(m,k) + \mathbf{N}(m,k) & \text{if} \quad Z(m,k) \in \{i; i = 1, \dots, I\} \end{cases}$$

- $Z(m, k)$ hidden random variable indicating which source is active
- Temporal and spectral correlations in $\mathbf{X}(1..M, 1..K)$ present

Goal: Extract $S_i(m)$ solely from $\mathbf{X}(1..M, 1..K)$

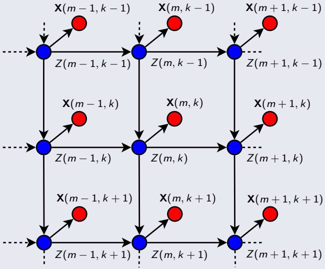
- By computation of source activity probability

$$P\left(Z(m,k) \middle| \mathbf{X}(1..M, 1..K)\right)$$

using temporal and spectral correlations!

Capturing temporal and spectral correlations

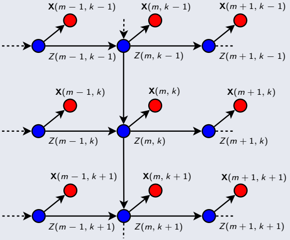
2D Hidden Markov Model



- Modeling correlations among adjacent time-frequency slots
- **Problem:** Exact inference is computational intractable!
  - **Solution:** Alternate inference in time and frequency direction

Decoding of a 2D-HMM

Modified forward-backward algorithm (FBA) in frequency direction



- Ignoring vertical dependencies in all other columns
- Modified forward-backward algorithm (FBA):

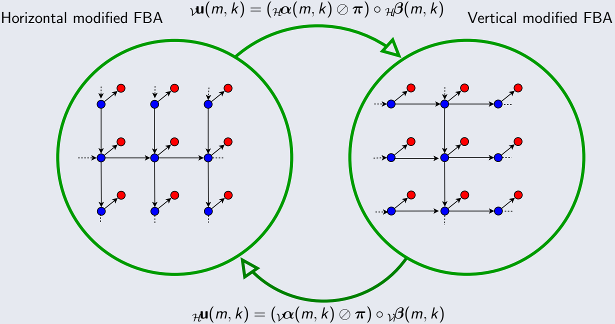
$$_{\mathcal{V}}\boldsymbol{\alpha}(m,k) = 1 \propto {_{\mathcal{V}}\mathbf{T}}^{\mathrm{T}} \left({_{\mathcal{V}}\boldsymbol{\alpha}}(m,k-1) \circ \mathbf{o}(m,k-1) \circ {_{\mathcal{V}}\mathbf{u}}(m,k-1) \right)$$
$$_{\mathcal{V}}\boldsymbol{\beta}(m,k) = \boldsymbol{\pi} \propto {_{\mathcal{V}}\mathbf{T}} \left({_{\mathcal{V}}\boldsymbol{\beta}}(m,k+1) \circ \mathbf{o}(m,k+1) \circ {_{\mathcal{V}}\mathbf{u}}(m,k+1) \right)$$
$$_{\mathcal{V}}\boldsymbol{\gamma}(m,k) = 1 \propto \mathbf{o}(m,k) \circ {_{\mathcal{H}}\mathbf{u}}(m,k) \circ {_{\mathcal{V}}\boldsymbol{\alpha}}(m,k) \circ {_{\mathcal{V}}\boldsymbol{\beta}}(m,k)$$

vertical forward prediction variable: $_{\mathcal{V}}\boldsymbol{\alpha}(m,k)$,
prior probabilities: $\boldsymbol{\pi}$,
observation likelihood: $\mathbf{o}(m,k)$,

vertical backward variable: $_{\mathcal{V}}\boldsymbol{\beta}(m,k)$,
vertical transition matrix: $_{\mathcal{V}}\mathbf{T}$,
vertical junction variable: $_{\mathcal{V}}\mathbf{u}(m,k)$

Decoding of a 2D-HMM

Iterative Turbo Decoding Scheme



- Based on modified forward-backward algorithm (FBA) along time and frequency direction
- Extrinsic information exchange between both directions

Polar model of $p\left(\mathbf{X}(m,k) \middle| Z(m,k)\right)$

Average *a-posteriori* SNR

$$\varphi(m,k) := \frac{1}{D}\mathbf{X}^{\mathrm{H}}(m,k)\mathbf{\Phi}_{\mathbf{NN}}^{-1}(k)\mathbf{X}(m,k)$$

where $\mathbf{\Phi}_{\mathbf{NN}} = \mathrm{E}\left[\mathbf{N}\mathbf{N}^{\mathrm{H}}\right]$

- Modeled by scaled chi-squared distribution

Frequency and unit-norm normalize observation vector

$$\tilde{Y}_j(m,k) := \left|X_j(m,k)\right| \exp\left\{\mathrm{i}\frac{\arg[X_j(m,k)X_1^*(m,k)]}{2(k-1)f_{\mathrm{S}}d_{\max}(Kc_v)^{-1}}\right\}$$

$$\mathbf{Y}(m,k) := \tilde{\mathbf{Y}}(m,k) \Big/ \left\|\tilde{\mathbf{Y}}(m,k)\right\|$$

- Modeled by complex Watson distribution

Assumption

$$p\left(\mathbf{X}(m,k) \middle| Z(m,k)\right) = p\left(\mathbf{Y}(m,k) \middle| Z(m,k)\right) \cdot p\left(\varphi(m,k) \middle| Z(m,k)\right)$$

Expectation Maximization algorithm for BSS

Parameters

- Set of unknown parameters $\mathbf{\Theta} = \{\mathbf{W}_1, \dots, \mathbf{W}_I\}$
- Parameters of 2D-HMM are pretrained with speech mixtures

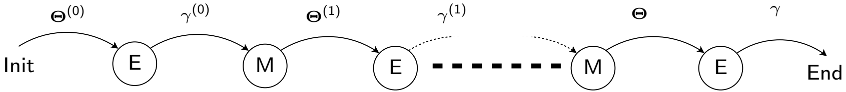
E-step: Estimate posterior speech activity probabilities

- Use proposed turbo decoding scheme to estimate posterior speech activity probabilities $\gamma_i^{(\nu)}(m,k) := P\left(Z(m,k) \middle| \mathbf{X}(1..M, 1..K)\,; \mathbf{\Theta}^{(\nu)}\right)$

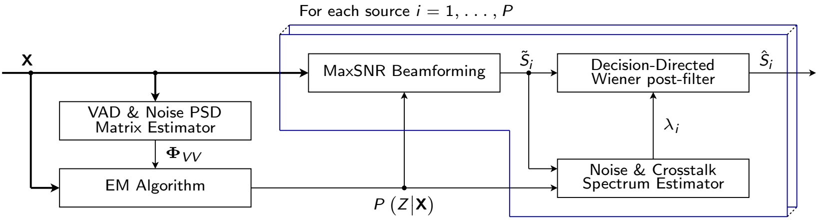
M-step: Update unknown parameter

- Principal eigenvector of source dependent matrix

$$\mathbf{\Phi}_{\mathbf{YY},i}^{(\nu)} := \frac{\sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_i^{(\nu)}(m,k)\mathbf{Y}(m,k)\mathbf{Y}^{\mathrm{H}}(m,k)}{\sum_{m=1}^{M} \sum_{k=1}^{K} \gamma_i^{(\nu)}(m,k)}$$



System overview



Properties

- Spatial filtering with generalized eigenvector (MaxSNR)-beamforming
- Spectral filtering with Wiener post-filter
- Adaptation control with source activity probability $P(Z|\mathbf{X})$

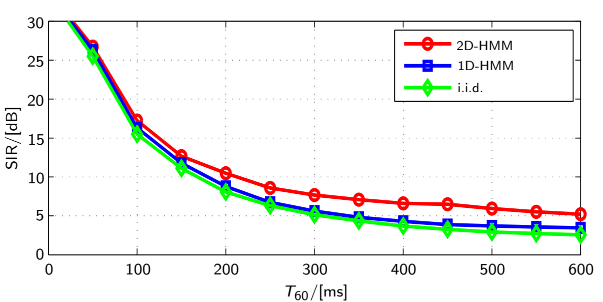
Evaluation

Setup

- Four sensor array arranged at the vertices of regular tetrahedron with lateral length of 2cm (No spatial aliasing!)
- 3 sources randomly positioned around the microphone
- Noise recordings of the fan noise of a video projector at $-10\,\text{dB}$
- White noise at $-20\,\text{dB}$
- Image method for reverberant room simulation

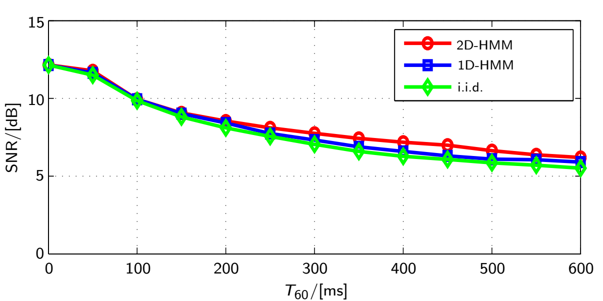
Evaluation

Gain in signal-to-interference-ratio (SIR)



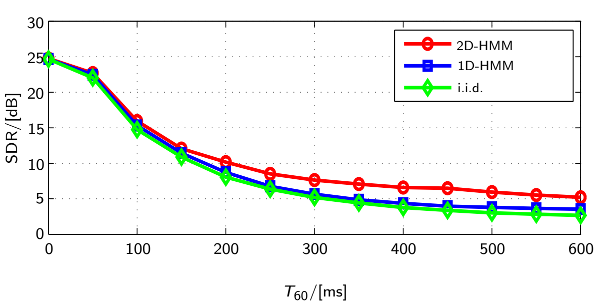
Evaluation

Gain in signal-to-noise-ratio (SNR)



Evaluation

Gain in signal-to-distortion-ratio (SDR)



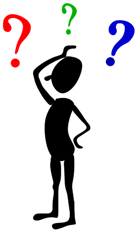
Summary and outlook

Summary

- Exploiting correlations of adjacent TF-slots for noisy BSS
  - 2D-HMM to capture temporal and spectral correlations
  - Iterative decoding scheme with modified FBA algorithm using extrinsic information exchange
- Improved performance in all cases and w.r.t. all measures
  - Advantages are evident especially in highly reverberant recording conditions

Outlook

- Exploitation of harmonic structures of speech
- Low latency block online implementation



Thank you for your attention!

Questions ?

Reinhold Haeb-Umbach

University of Paderborn
Department of Communications Engineering

haeb@nt.uni-paderborn.de
nt.uni-paderborn.de