Dual frequency- and block-permutation alignment for deep learning based block-online blind source separation

PADERBORN UNIVERSITY
The University for the Information Society
Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

Lukas Drude[1,2], Takuya Higuchi[2], Keisuke Kinoshita[2], Tomohiro Nakatani[2], Reinhold Haeb-Umbach[1]

[1]Department of Communications Engineering, Paderborn University, Paderborn, Germany
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

Motivation/ Setting

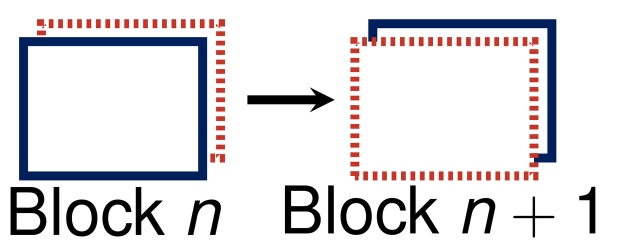
- Blind source separation
- Leverage Deep Attractor Network (DAN)
  → Problematic with long mixtures/ sessions
- Multiple microphones available

$$\mathbf{y}_{tf} = \sum_k \mathbf{h}_{fk}\, s_{tfk} + \mathbf{n}_{tf} = \sum_k \mathbf{x}_{tfk} + \mathbf{n}_{tf}$$

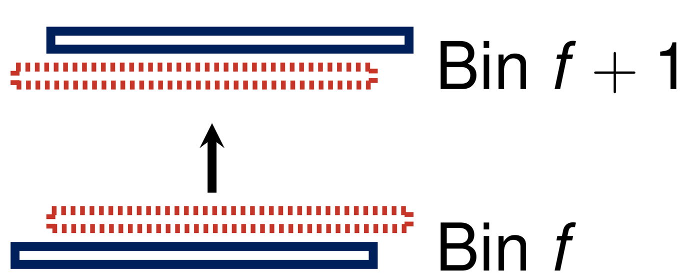
t: time frame index — k: class/ source index
f: frequency bin index — n: block index

Problem statement

- DAN generates embedding vectors $\mathbf{e}_{tf}$ indicative of which time frequency bin belongs to the same speaker.
- Embedding vectors $\mathbf{e}_{tf}$ can then be clustered.
- DAN not directly applicable to streaming data
  - Uses BLSTM
    → Split signal into blocks
  - Embedding space not fixed, centroid $\boldsymbol{\mu}_{nk}$ of each speaker changes from block to block (block permutation problem)



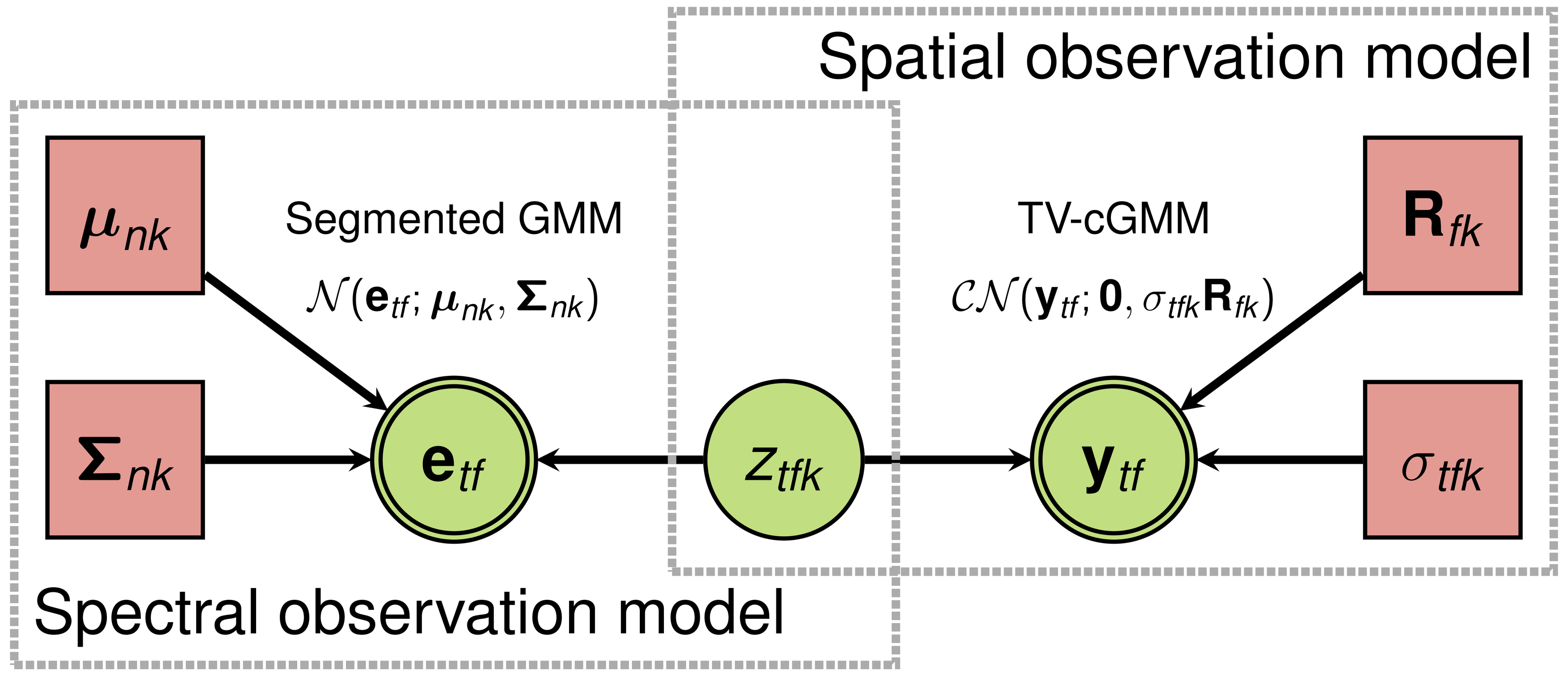
- Resort to spatial model (i.e. time variant complex GMM (TV-cGMM))
  - Independent solution on each frequency bin
  - Solutions not aligned (frequency permutation problem)



Idea

Here: Leverage...
- ...spatial info to solve block permutation,
- ...spectral info to solve frequency permutation problem.

Full update model



M-step (Spatial observation model):

$$\sigma_{tfk} = \frac{1}{D}\mathbf{y}_{tf}^{\mathrm{H}}\mathbf{R}_{fk}^{-1}\mathbf{y}_{tf}, \qquad \Gamma_{fk} = \sum_t \gamma_{tfk},$$

$$\mathbf{R}_{fk} = \frac{1}{\Gamma_{fk}}\sum_t \gamma_{tfk}\frac{\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}}}{\sigma_{tfk}}$$

M-step (Spectral observation model):

$$\boldsymbol{\mu}_{nk} = \frac{1}{\Gamma_{nk}}\sum_{t\in\mathcal{T}_n, f}\gamma_{tfk}\mathbf{e}_{tf}, \qquad \Gamma_{nk} = \sum_{t\in\mathcal{T}_n, f}\gamma_{tfk},$$

$$\boldsymbol{\Sigma}_{nk} = \frac{1}{\Gamma_{nk}}\sum_{t\in\mathcal{T}_n, f}\gamma_{tfk}(\mathbf{e}_{tf}-\boldsymbol{\mu}_{nk})(\mathbf{e}_{tf}-\boldsymbol{\mu}_{nk})^{\mathrm{T}}$$

E-step with permutation alignment:

$$\Pi_f = \operatorname*{argmax}_{\Pi}\left\{\sum_{t,k}\overbrace{\frac{C_{tfk}(\Pi)}{\sum_{k'}C_{tfk'}(\Pi)}}^{\gamma_{tfk}}\cdot\ln C_{tfk}(\Pi)\right\},$$

$$\text{with } C_{tfk}(\Pi) = p(\mathbf{e}_{tf}; \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})\cdot p(\mathbf{y}_{tf}; \sigma_{tf,\Pi(k)}, \mathbf{R}_{f,\Pi(k)})$$

GEV beamforming

Source extraction by generalized eigenvalue decomposition of target and non-target covariance matrix:

$$\boldsymbol{\Phi}_{fk}^{\text{target}} = \frac{1}{\Gamma_{fk}}\sum_{t\in\mathcal{T}_n}\gamma_{tfk}\mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathrm{H}}, \qquad \boldsymbol{\Phi}_{fk}^{\text{non-target}} = \sum_{k\neq k'}\boldsymbol{\Phi}_{fk'}^{\text{target}}$$

Block-online algorithm

1: Split into N blocks and run model on the first block.
2: Apply GEV beamforming to the first block.
3: **for** n from 1 to N **do**
4: Forget all parameters but $\mathbf{R}_{n-1,fk}$ and $\boldsymbol{\Phi}_{n-1,fk}$.
5: Initialize σ_{tfk} with using $\mathbf{R}_{n-1,fk}$.
6: Initialize γ_{tfk} only with spatial observation model.
7: **while** not converged **do**
8: Obtain $\boldsymbol{\mu}_{nk}$ and $\boldsymbol{\Sigma}_{nk}$.
9: Incremental update for $\boldsymbol{R}_{nfk}$.
10: Calculate variance σ_{tfk}.
11: E-step with permutation alignment yields γ_{tfk} and Π_f.
12: Obtain spatial covariance matrices with incremental update.
13: Apply beamforming on current block.

Results

WSJ utterances + CHiME 3 noise + artificial RIRs
(baseline systems in gray)

	Active Speech Ratio:		73 %	60 %	47 %
	Average Duration:		16.9 s	27.0 s	36.4 s
	Model	Prior	SDR gain/dB		
Offline	TV-cGMM + Align	None	16.9	15.5	13.3
Offline	TV-cGMM + Align	fk	17.3	16.6	15.4
Offline	DAN + GMM + Oracle Align	None	15.8	15.8	15.5
Offline	DAN + GMM + Oracle Align	nk	15.7	15.6	15.4
Offline	DAN + Fixed Spatial Prior	tfk	17.3	16.7	15.1
Offline	DAN + Fixed Spectral Prior	tfk	17.1	17.0	16.1
Offline	DAN + Full Update Model	None	**17.9**	**17.4**	**16.3**
Offline	DAN + Full Update Model	fk	17.8	**17.4**	**16.3**
Block-online	TV-cGMM + Align	None	15.7	14.3	11.6
Block-online	TV-cGMM + Align	fk	16.1	15.0	12.5
Block-online	DAN + GMM + Oracle Align	None	13.7	14.0	13.7
Block-online	DAN + GMM + Oracle Align	nk	13.9	13.8	13.5
Block-online	DAN + Full Update Model	None	**17.4**	**16.5**	**14.0**
Block-online	DAN + Full Update Model	fk	17.1	16.2	13.4

{drude, haeb}@nt.upb.de, {higuchi.takuya, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp