Tight integration of spatial and spectral features for BSS with Deep Clustering embeddings

Lukas Drude, Reinhold Haeb-Umbach

Department of Communications Engineering - Paderborn University
Prof. Dr.-Ing. Reinhold Haeb-Umbach

2017-08-20

Table of contents

Problem Statement

Goal:

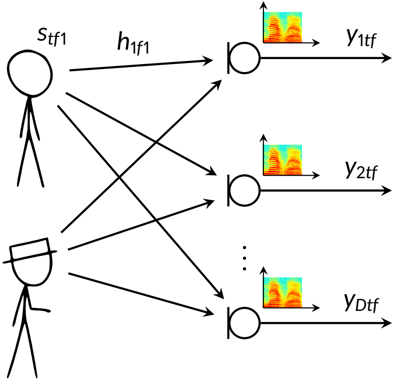
- Transcribe speech
  from an observed mixture

Application:

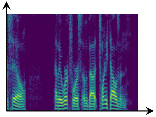
- Meeting transcription
- Home assistants/ automation
- (Surveillance)

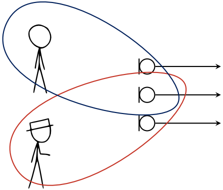
Focus:

- Robust source separation front-end
  exploiting spectral and spatial features



Motivation





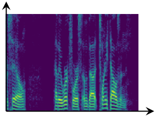
Spectral/ deep learning:

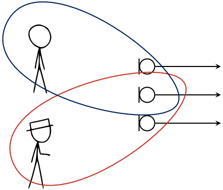
- + Leverage training data
- + Model speech characteristics
- – Overfit, possibly poor generalization

Spatial clustering (multi-channel):

- + Training data agnostic
- – No concept of human speech
- + Exploit spatial selectivity

Motivation





Spectral/ deep learning:

+ Leverage training data
+ Model speech characteristics
– Overfit, possibly poor generalization

Spatial clustering (multi-channel):

+ Training data agnostic
– No concept of human speech
+ Exploit spatial selectivity

Joint formulation desired!

Spectral: Deep Clustering [1]

Exploit speaker specific **spectral** characteristics for separation.

Concept:

- BLSTM yields an embedding vector $\mathbf{e}_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using k-means on $\mathbf{e}_{tf}$.

[1] Hershey et al., Deep clustering: Discriminative embeddings [...], ICASSP 2016

Spectral: Deep Clustering [1]

Exploit speaker specific **spectral** characteristics for separation.

Concept:

- BLSTM yields an embedding vector $\mathbf{e}_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using k-means on $\mathbf{e}_{tf}$.

Relevance:

- Speaker independent
- Number of speakers not fixed at training time

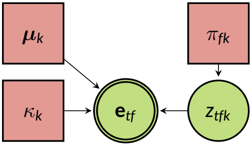
[1] Hershey et al., Deep clustering: Discriminative embeddings [...], ICASSP 2016

Spectral: Deep Clustering [1]

Exploit speaker specific **spectral** characteristics for separation.

Concept:

- BLSTM yields an embedding vector $\mathbf{e}_{tf}$ for each tf-bin.
- Encourage tendency to form clusters in embedding space.
- Cluster using ~~k-means~~ on $\mathbf{e}_{tf}$.
  von Mises Fisher MM

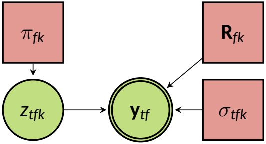


Relevance:

- Speaker independent
- Number of speakers not fixed at training time

[1] Hershey et al., Deep clustering: Discriminative embeddings [...], ICASSP 2016

Spatial: Time-Variant cGMM [2]



Exploit **spatial** diversity to separate speakers.

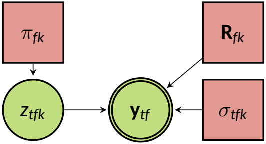
Concept:

- Complex random vectors $\mathbf{y}_{tf}$
- Acoustic transfer function captured by spatial correlation matrix.

$$p(\mathbf{y}_{tf}) = \sum_k \pi_{fk} \cdot \mathcal{N}_{\mathbb{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})$$

[2] Ito et al., Relaxed disjointness based clustering [...], IWAENC 2014

Spatial: Time-Variant cGMM [2]



Exploit **spatial** diversity to separate speakers.

Concept:

- Complex random vectors $\mathbf{y}_{tf}$
- Acoustic transfer function captured by spatial correlation matrix.

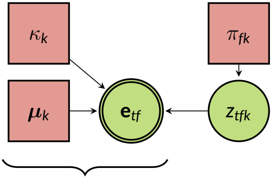
$$p(\mathbf{y}_{tf}) = \sum_k \pi_{fk} \cdot \mathcal{N}_{\mathbb{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})$$

Relevance:

- Competitive: Winning system of CHiME 3 and CHiME 4 challenges

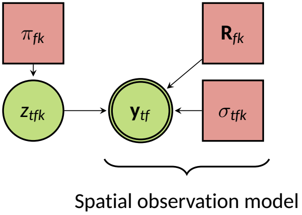
[2] Ito et al., Relaxed disjointness based clustering [...], IWAENC 2014

Integrated Model

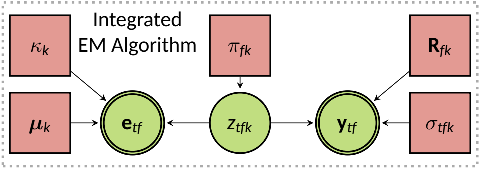


Spectral observation model

Integrated Model



Integrated Model

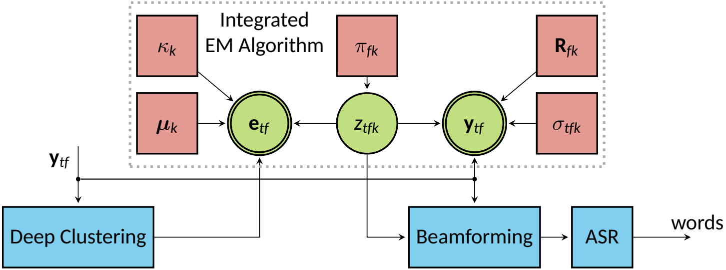


- Statistical model:

$$p(\mathbf{e}_{tf}, \mathbf{y}_{tf}) = \sum_k \pi_{fk} \cdot \mathrm{vMF}(\mathbf{e}_{tf}; \boldsymbol{\mu}_k, \kappa_k) \cdot \mathcal{N}_{\mathbb{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})$$

- Observations independent, given class label
- Estimate of all parameters better, when estimated jointly
- Risk/ opportunity: Weighting between both models

Integrated Model

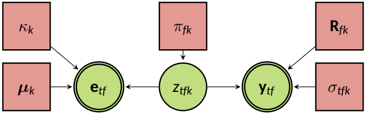


- Statistical model:

$$p(\mathbf{e}_{tf}, \mathbf{y}_{tf}) = \sum_k \pi_{fk} \cdot \mathrm{vMF}(\mathbf{e}_{tf}; \boldsymbol{\mu}_k, \kappa_k) \cdot \mathcal{N}_{\mathbb{C}}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})$$

- Observations independent, given class label
- Estimate of all parameters better, when estimated jointly
- Risk/ opportunity: Weighting between both models

Update equations: E-step



- Update class affiliation posterior:

$$\gamma'_{tfk} = \underbrace{\pi_{fk}}_{\substack{\text{Prior}\\ p(z_{tf})}} \cdot \underbrace{\mathrm{vMF}^{\alpha}(\mathbf{e}_{tf}; \boldsymbol{\mu}_k, \kappa_k)}_{\substack{\text{Spectral Model}\\ p(\mathbf{e}_{tf}|z_{tf})}} \cdot \underbrace{\mathcal{N}_{\mathbb{C}}^{(1-\alpha)}(\mathbf{y}_{tf}; \mathbf{0}, \sigma_{tfk} \cdot \mathbf{R}_{fk})}_{\substack{\text{Spatial Model}\\ p(\mathbf{y}_{tf}|z_{tf})}}$$

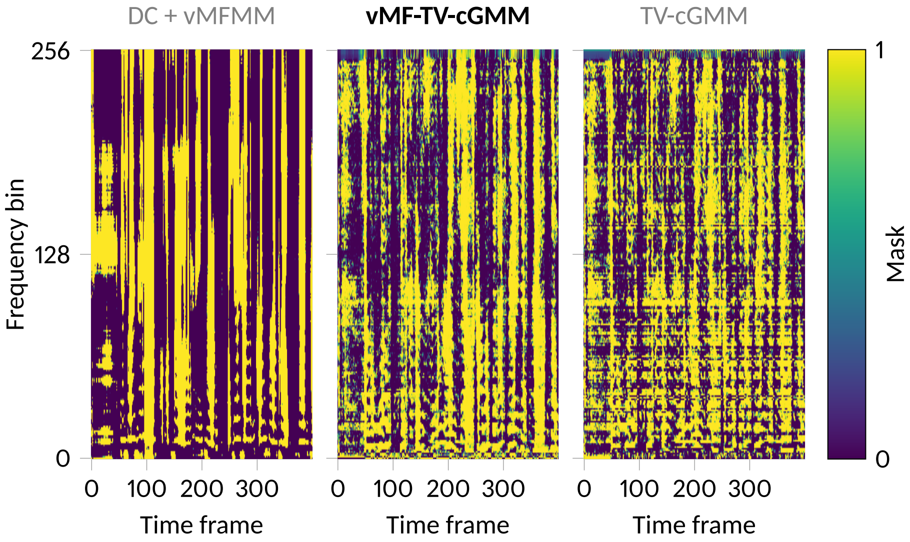
$$\gamma_{tfk} = \gamma'_{tfk} / \sum_k \gamma'_{tfk}$$

Evaluation: Setup

- Train/CV deep clustering on single channel WSJ utterances [1, 3]
- GMM-HMM recognizer trained on clean, 3-Gram LM [3]
- 3000 test mixtures, 2 speakers per mixture:
  - Random source and array positions (6 sensors)
  - Image method to generate room impulse responses

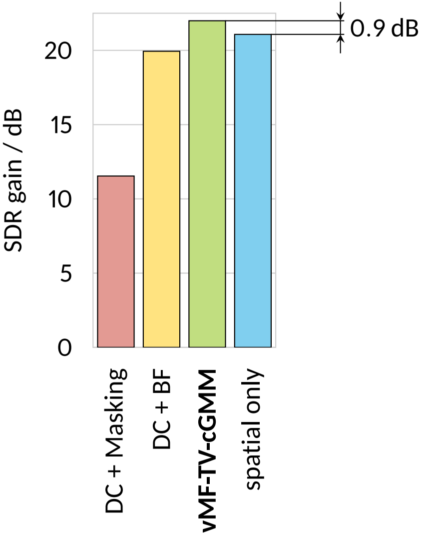
[3] Isik et al., Single-Channel Multi-Speaker Separation using Deep Clustering, Interspeech 2016

Example: Posterior Masks (200 ms $< T_{60} <$ 300 ms)

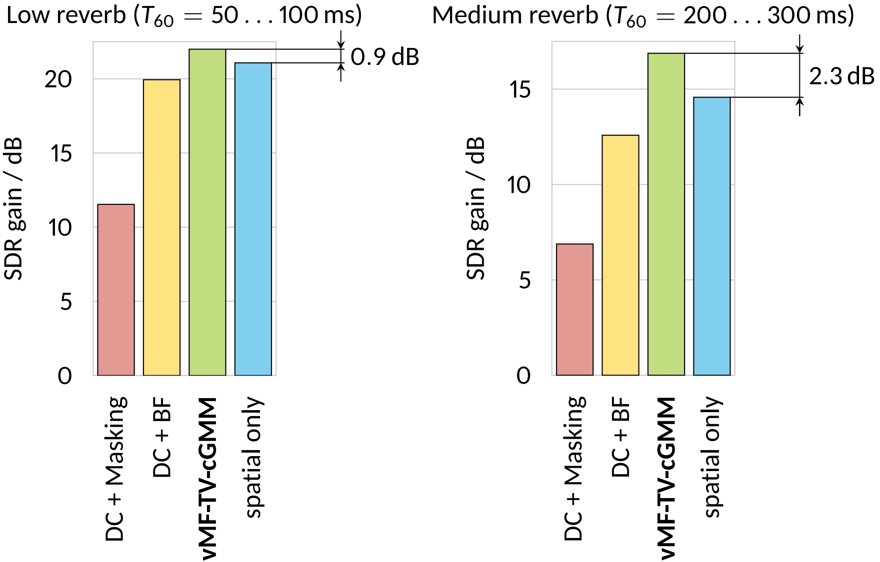


Evaluation: Signal to Distortion Ratio Gain

Low reverb ($T_{60} = 50 \ldots 100$ ms)



Evaluation: Signal to Distortion Ratio Gain



Word Error Rates ($T_{60} = 50 \dots 100$ ms)

Features	Model	Extraction	WER / %
spectral	DC [3]	Masking	65.8
spectral	DC	Beamforming	42.4
integrated	**vMF-TV-cGMM**	**Beamforming**	**29.7**
spatial	TV-cGMM [2]	Beamforming	33.6

[2] Ito et al., Relaxed disjointness based clustering [...], IWAENC 2014

[3] Isik et al., Single-Channel Multi-Speaker Separation using Deep Clustering, Interspeech 2016

Audio Example

(Switch to external audio player.)

- Observed mixture: y.wav
- Estimates: z_1.wav, z_2.wav
- Clean speech: x_1.wav, x_2.wav

Conclusion and Future Work

- EM algorithm for joint model

Spectral/ deep learning:

+ Leverages training data
+ Models speech characteristics

Spatial clustering:

+ Generalizes to unseen conditions
+ Exploits spatial selectivity

- Outperforms both very different baselines
- Bridges model-based research and deep learning research

Future Work:

- Joint training
- Real recordings