Unsupervised Training of a Deep Clustering Model for Multichannel Blind Source Separation

Lukas Drude, Daniel Hasenklever, Reinhold Haeb-Umbach
Department of Communications Engineering, Paderborn University, Germany
{drude, haeb}@nt.upb.de

PADERBORN UNIVERSITY

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

Introduction

- Unsupervised training of source separation networks
  - Avoids parallel (clean + noisy) data
  - Does not **rely** on simulated data (e.g. Lombard effect often not simulated)
  - Leverages reverberant noisy recordings without ground truth
- Concept
  1. Apply unsupervised teacher to observation
  2. Use teacher result to train the student
- Signal model in STFT domain

Related work:

Tuesday	Seetharaman et al.
Wednesday	Tzinis et al.
This session	Aihara et al.

$$\mathbf{y}_{tf} = \sum_k \mathbf{h}_{kf} s_{ktf} + \mathbf{n}_{tf} = \sum_k \mathbf{x}_{ktf} + \mathbf{n}_{tf}$$

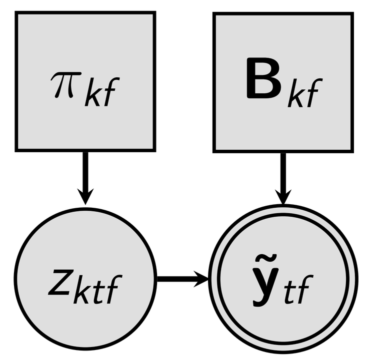
k source/ class index
t time frame index
f frequency bin index

Probabilistic model: cACGMM (teacher)

Complex angular central Gaussian mixture model (Ito et al., 2016)
Exploits **spatial** diversity to separate speakers:

- Multi-channel observations in STFT domain
- Complex random vectors $\tilde{\mathbf{y}}_{tf} = \mathbf{y}_{tf}/\|\mathbf{y}_{tf}\|$
- Relative acoustic transfer function captured by spatial correlation matrix

$$p(\tilde{\mathbf{y}}_{tf}; \boldsymbol{\theta}) = \sum_k \pi_{kf} \mathrm{cACG}(\tilde{\mathbf{y}}_{tf}, \mathbf{B}_{kf})$$

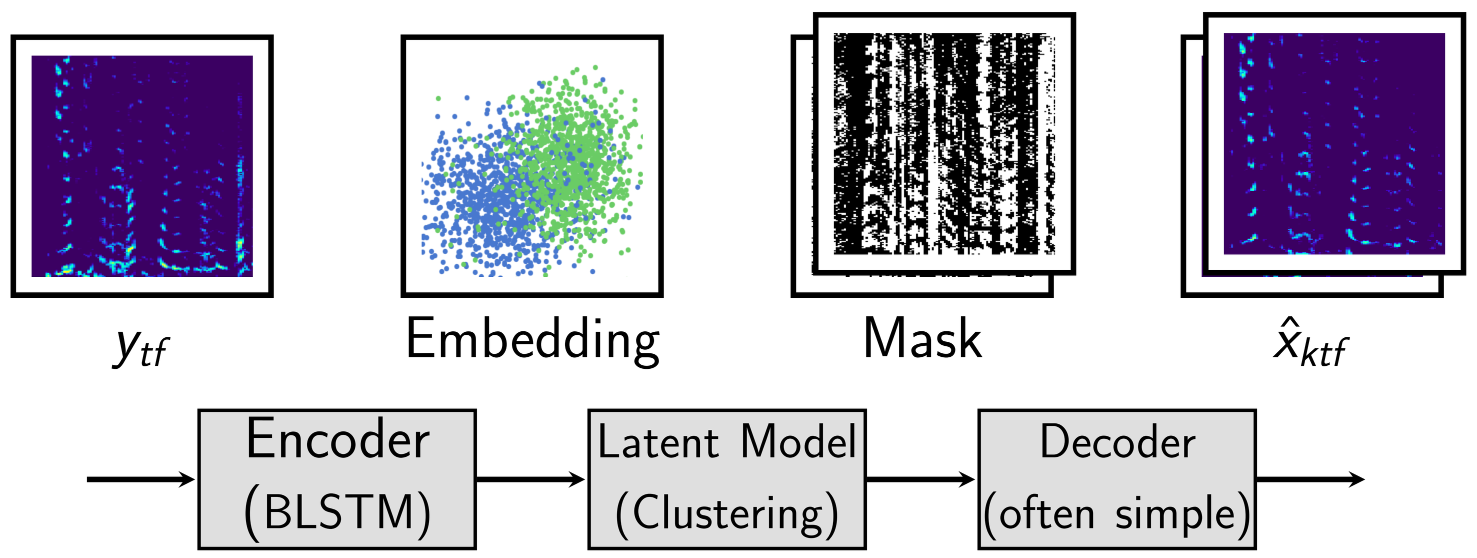


(Probabilistic dependencies)

Neural network: Deep Clustering (student)

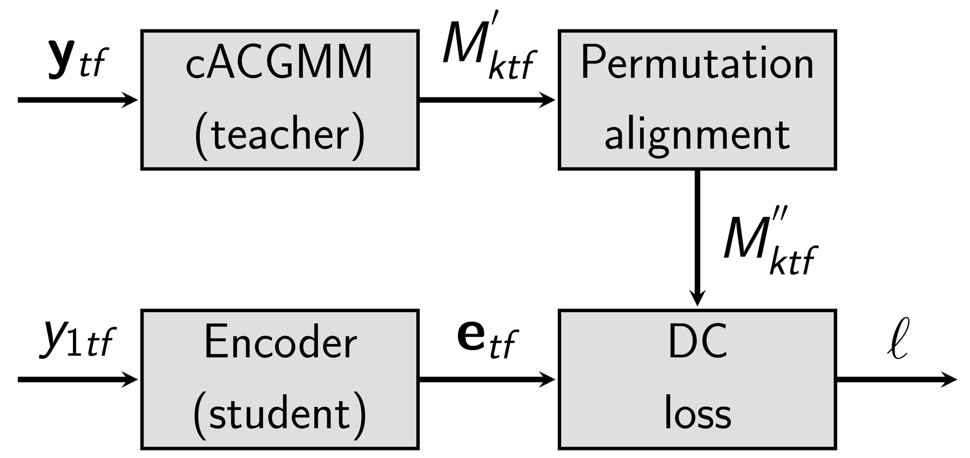
Exploit speaker specific **spectral** characteristics for separation:

- Single-channel observations in magnitude spectrum domain
- Encoder (BLSTM) yields embedding vectors $\mathbf{e}_{tf}$
- Training encourages tendency to form clusters in embedding space
- Cluster using k-means on $\mathbf{e}_{tf}$

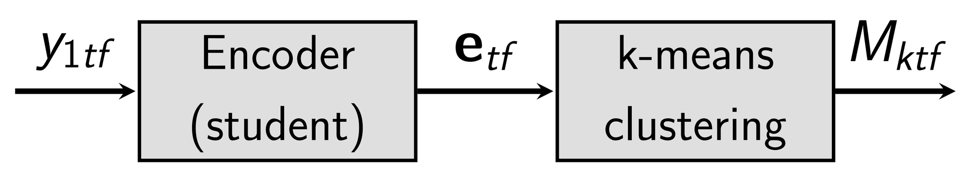


Proposed training scheme

Unsupervised Deep Clustering: Knowledge transfer across domains

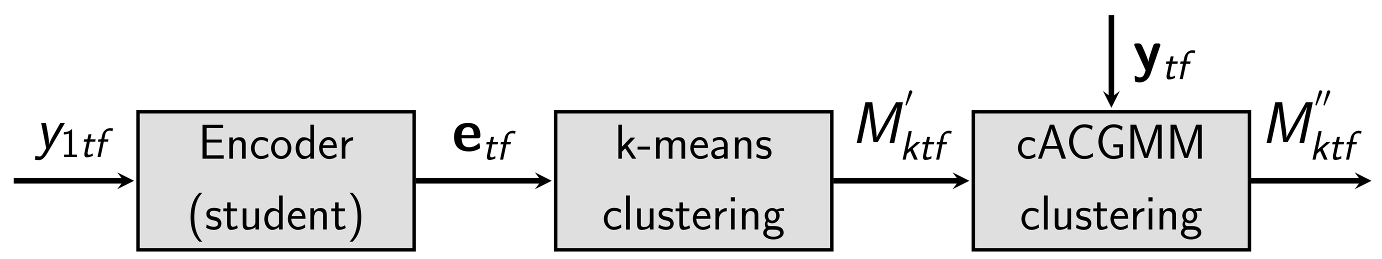


Prediction variant: k-means only



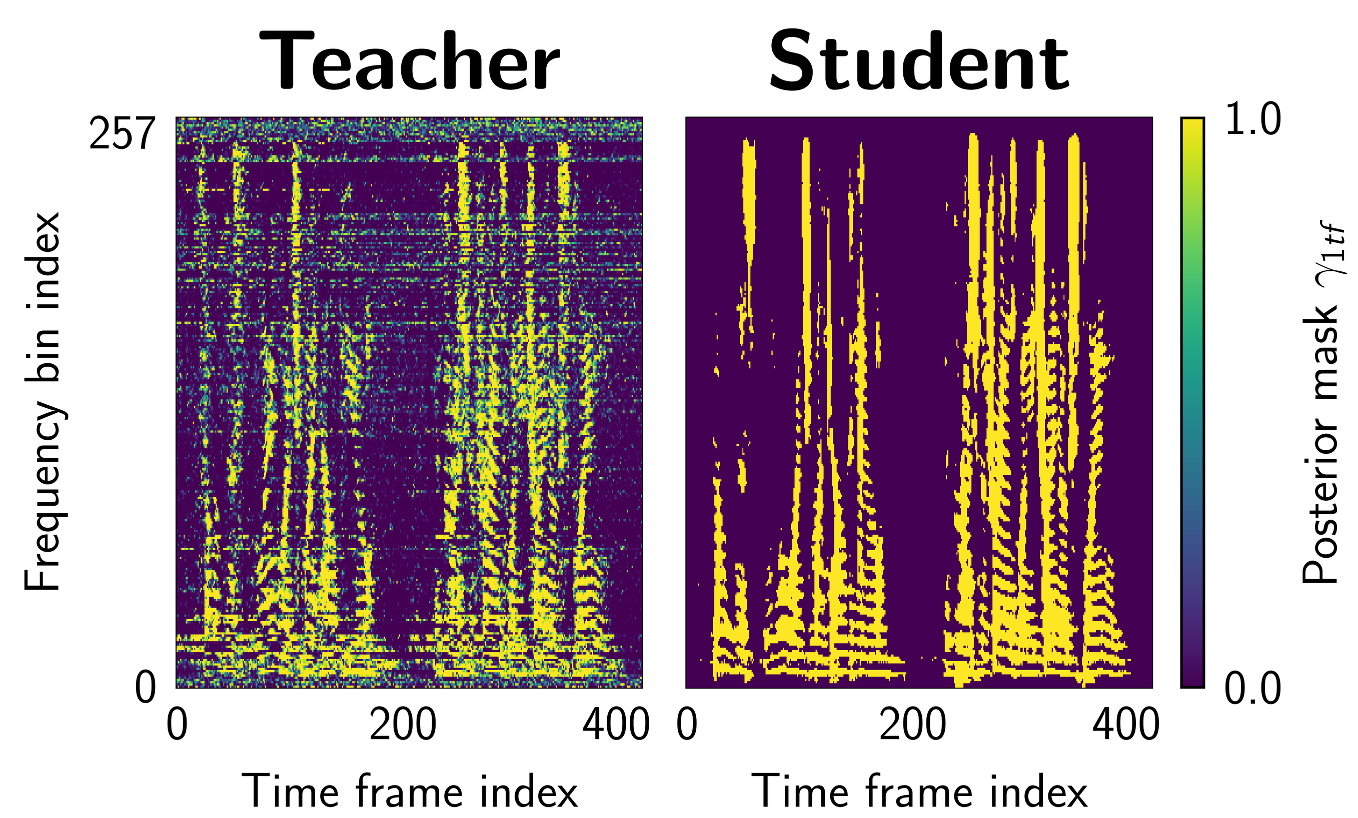
- Applicable in a single channel scenario
- Yields hard masks

Prediction variant: k-means → cACGMM



- Needs multi-channel observations to leverage spatial cues
- On average better performance
- Yields soft masks

Example masks



Datasets

- 30000, 500, and 1500 six channel mixtures from 3 WSJ sets
- White background noise: 20 dB – 30 dB
- Reverberation time T_{60}: 200 ms – 500 ms (image method)

Source extraction by masking

Model	SDR gain/dB		PESQ	STOI	WER
	BSS-Eval	Invasive	gain	gain	/%
cACGMM only	7.2	10.4	0.17	0.11	38.4
Student → k-means	5.5	9.4	−0.42	0.04	75.1
Student → k-means → cACGMM	9.5	13.2	0.40	0.18	29.3
Superv. → k-means	5.9	9.5	−0.25	0.06	75.8
Superv. → k-means → cACGMM	9.1	12.6	0.37	0.16	31.0
Oracle IBM → cACGMM	9.7	13.3	0.48	0.14	28.9

Source extraction by beamforming

Model	SDR gain/dB		PESQ	STOI	WER
	BSS-Eval	Invasive	gain	gain	/%
cACGMM only	5.1	12.7	0.37	0.09	28.0
Student → k-means	5.7	13.6	0.43	0.11	29.0
Student → k-means → cACGMM	6.4	15.3	0.52	0.13	20.7
Superv. → k-means	5.9	14.2	0.47	0.12	26.5
Superv. → k-means → cACGMM	6.1	14.9	0.50	0.12	21.1
Oracle IBM → cACGMM	6.4	15.5	0.78	0.12	19.9

Conclusions

- Deep Clustering can be trained from scratch without supervision
- No need for parallel data or simulated data
- Student outperforms the probabilistic model-based teacher
- Unsupervised system able to outperform supervised system

Concept generalizes to other applications:

Teacher	Student
IMCRA	Neural noise mask estimator
cACGMM	Neural network-supported beamforming
SRP-Phat	Neural DoA estimator

Interspeech: Unsupervised training of neural mask-based beamforming

Computational resources were provided by the Paderborn Center for Parallel Computing.