PADERBORN UNIVERSITY
The University for the Information Society

Communications Engineering

NT

Deep attractor networks for speaker re-identification and blind source separation

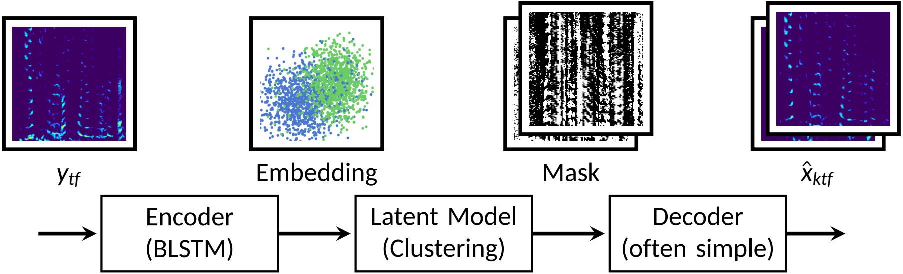
Lukas Drude, Thilo von Neumann, Reinhold Haeb-Umbach

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

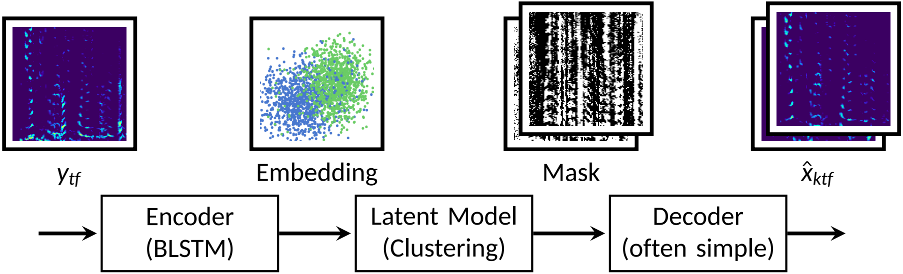
2018-04-17

Table of contents

Schematic overview: DC/ DAN

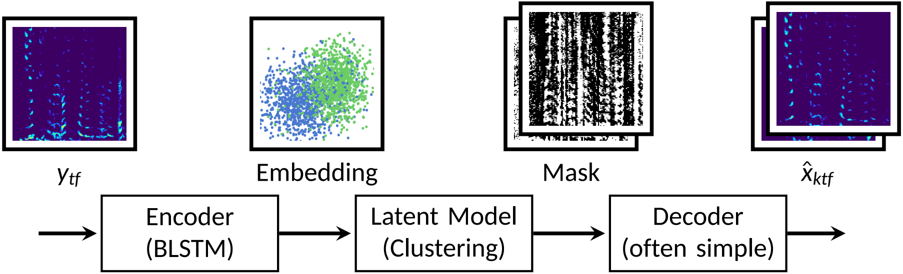


Schematic overview: DC/ DAN



- Deep Clustering (DC) [Hershey 2016]:
  - No assumption about the speaker at test time
  - Encoder network generates embedding vectors
  - Decoder just applies binary mask to observation
- Deep Attractor network (DAN) [Chen 2017]:
  - Different loss function allows end-to-end training
  - Decoder calculates soft mask first

Schematic overview: DC/ DAN



- Deep Clustering (DC) [Hershey 2016]:
  - No assumption about the speaker at test time
  - Encoder network generates embedding vectors
  - Decoder just applies binary mask to observation
- Deep Attractor network (DAN) [Chen 2017]:
  - Different loss function allows end-to-end training
  - Decoder calculates soft mask first

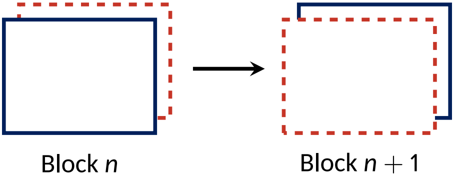
Developed for short mixtures.

Properties of the embeddings?

Identify speakers?

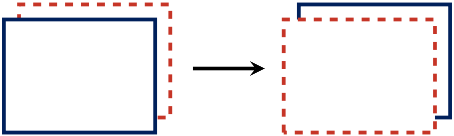
Tasks

Block permutation problem (tracing)



Tasks

Block permutation problem (tracing)



Block n Block $n+1$

Re-identification problem

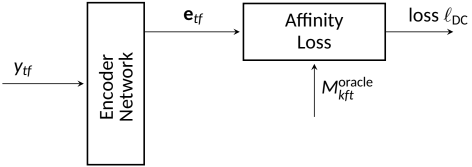
Blue speaker

Bag of K
separation results

Possible approaches

- Use i-vectors?
  $\rightarrow$ See results.
- Multichannel/ spatial cues?
  $\rightarrow$ AASP-P11.3:
  Drude et al., Dual Frequency- and Block-Permutation Alignment [...]
  Friday 13:30 – 15:30

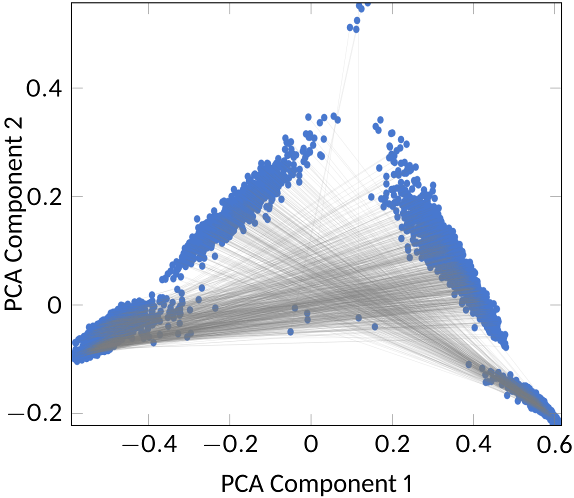
Deep Clustering



- Minimize difference between estimated and true affinity matrices:
  - Embedding vectors of same speaker co-linear
  - Embedding vectors of different speakers orthogonal

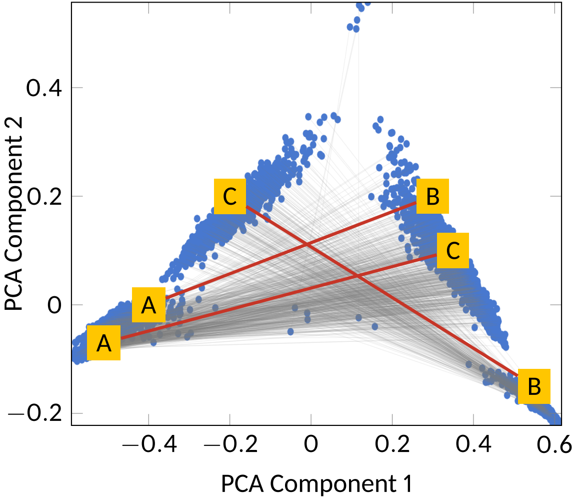
Deep Clustering – Centroids

- Each dot is an embedding **centroid** for each speaker
- Oracle mask used to visualize centroids

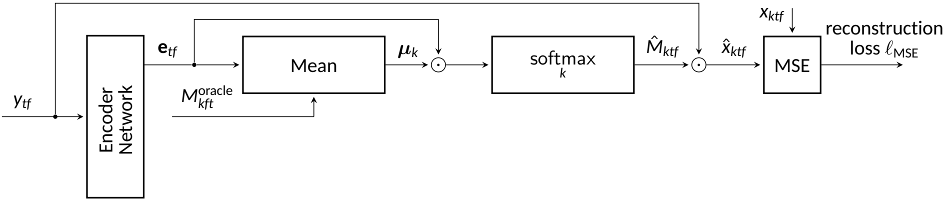


Deep Clustering – Centroids

- Each dot is an embedding **centroid** for each speaker
- Oracle mask used to visualize centroids



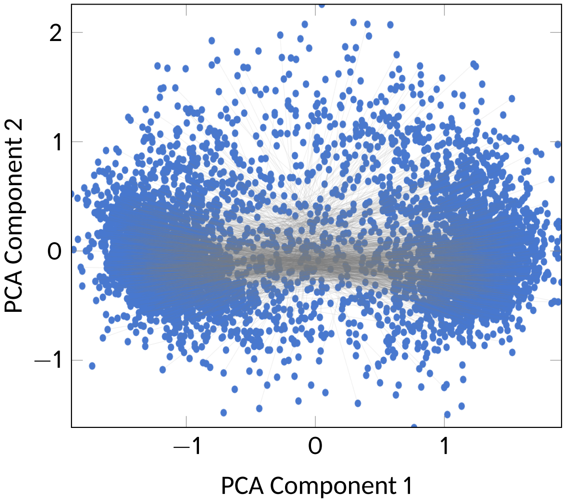
Deep Attractor Network



- Minimize reconstruction loss (MSE)
- Intuition:
  - Embedding vectors of same speaker in same direction
  - Embedding vectors of different speakers in opposite direction

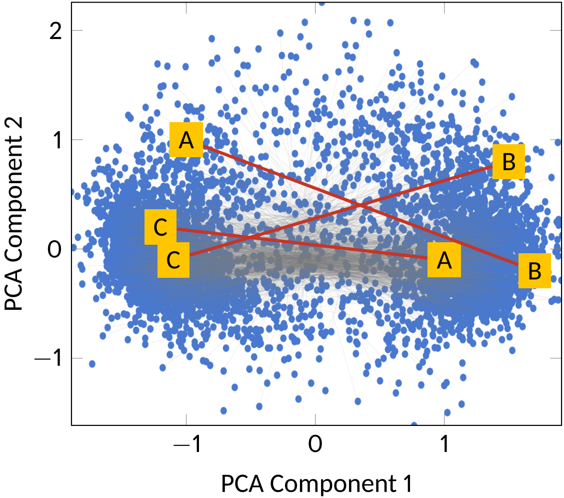
Deep Attractor Network – Centroids/Attractors

- Each dot is an embedding **centroid** for each speaker
- Oracle mask used to visualize centroids



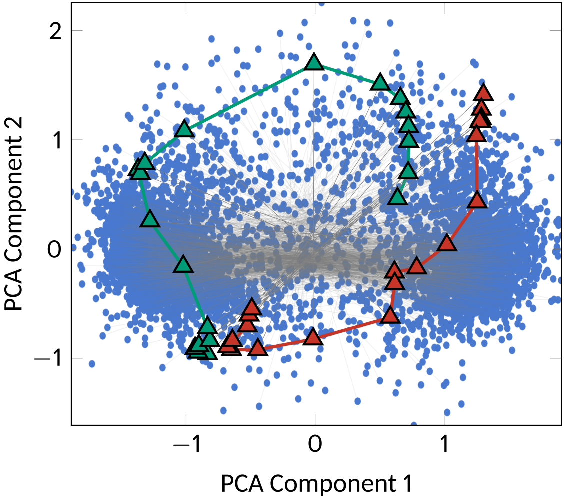
Deep Attractor Network – Centroids/Attractors

- Each dot is an embedding **centroid** for each speaker
- Oracle mask used to visualize centroids

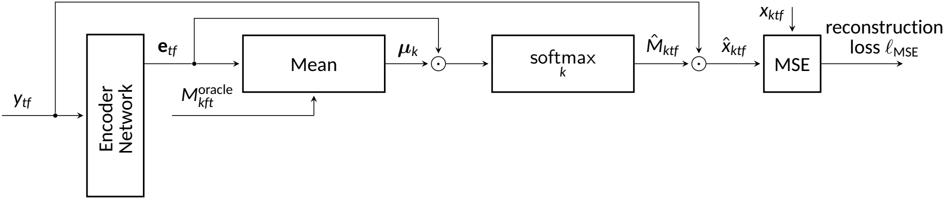


Deep Attractor Network – Centroids/Attractors

- Each dot is an embedding **centroid** for each speaker
- Oracle mask used to visualize centroids

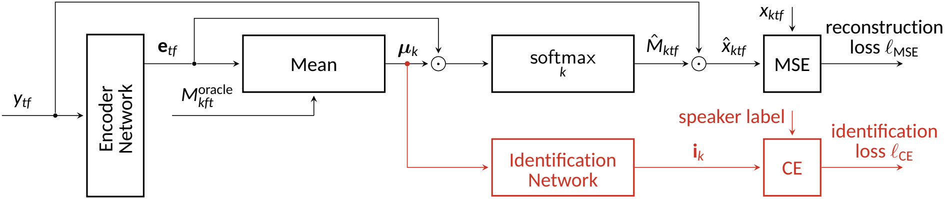


Solution: Identification loss



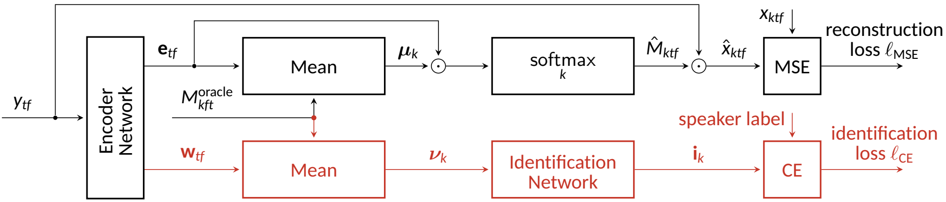
- Upper branch: Vanilla DAN

Solution: Identification loss



- Upper branch: Vanilla DAN
- Lower branch: Identification network + loss
  - Loss during training
  - Just use corresponding centroid at test time
- Multi-task learning: $\ell_{\text{total}} = \ell_{\text{MSE}} + \alpha \ell_{\text{CE}}$

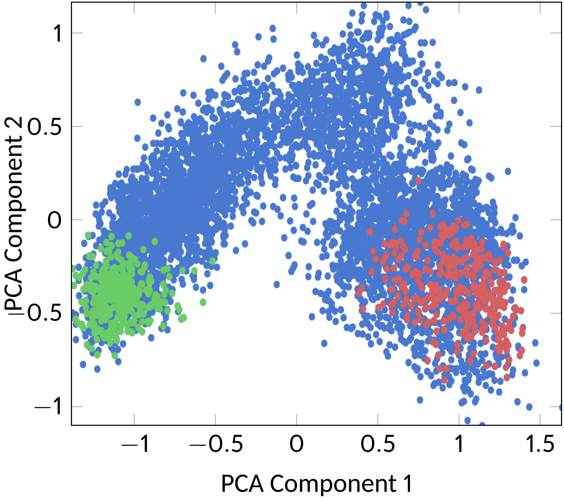
Solution: Identification loss



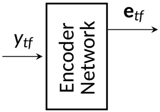
- Upper branch: Vanilla DAN
- Lower branch: Identification network + loss
  - Loss during training
  - Just use corresponding centroid at test time
- Multi-task learning: $\ell_{\text{total}} = \ell_{\text{MSE}} + \alpha \ell_{\text{CE}}$

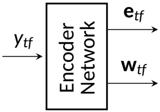
Solution: Identification loss

- Location of **identification attractors** tend to form clusters



Source separation performance

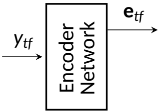


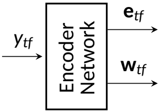


	α	SDR/dB
DAN		9.4
DAN + ID loss	0.001	**10.1**
	0.01	9.9
	0.1	9.9
	1	9.7
	10	8.9
DAN + ID emb.	0.001	9.9
	0.01	9.8
	0.1	10.0
	1	**10.1**
	10	9.2

Permutation/re-identification performance

Error Rate / %:	α	Permutation	Identification
Chance level		50.0	50.0
i-vector with VAD		8.0	9.7
DC		7.3	33.4
DAN		5.8	31.5
DAN + ID loss	0.001	6.7	32.7
	0.01	6.0	31.1
	1	5.0	20.1
	10	4.0	9.3
DAN + ID emb.	0.001	4.7	9.9
	0.01	3.7	7.7
	1	4.2	8.5
	10	**3.1**	**6.4**





Summary

- Embedding topology only valid for one mixture
  - Limitations in changing mixing conditions
  - Limitations for re-identification
- Extract speaker information with same encoder network
  - Multi-task learning helps both objectives
- Ways for speaker tracing/ identification...
  - i-vectors
  - Multichannel/ spatial cues (Drude et al., Friday, AASP-P11.3)
  - Embedding network provides ID embeddings