Towards a speaker diarization system for the CHiME 2020 dinner party transcription

Christoph Boeddeker[1], Tobias Cord-Landwehr[1], Jens Heitkaemper[1], Cătălin Zorilă[2], Daichi Hayakawa[3], Mohan Li[2], Min Liu[4], Rama Doddipatla[2], Reinhold Haeb-Umbach[1]

[1]Paderborn University, Department of Communications Engineering, Paderborn, Germany
[2]Toshiba Cambridge Research Laboratory, Cambridge, United Kingdom
[3] Toshiba Corporation Corporate R&D Center, Kawasaki, Japan
[4] Toshiba China R&D Center, Beijing, China

Overview

- Guided Source Separation (GSS)
  - Designed for CHiME-5
  - Baseline in CHiME-6 Track 1
  - Here: From Track 1 to Track 2 (i.e. drop human annotations)
- Acoustic models
- Results on Track 2
- Ongoing work
  - PIT Neural Speaker Diarization
  - Spatial features from Spatial Mixture Model (SMM)

CHiME-6 Dataset

For a complete description of the dataset see presentation:

S. Watanabe, M. Mandel, J. Barker, E. Vincent

"Overview of the 6th CHiME Challenge"

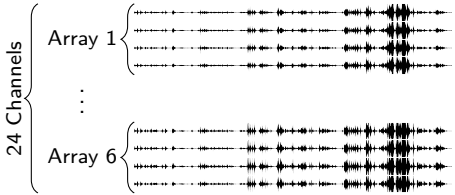
or publication:

S. Watanabe, M. Mandel, J. Barker, E. Vincent

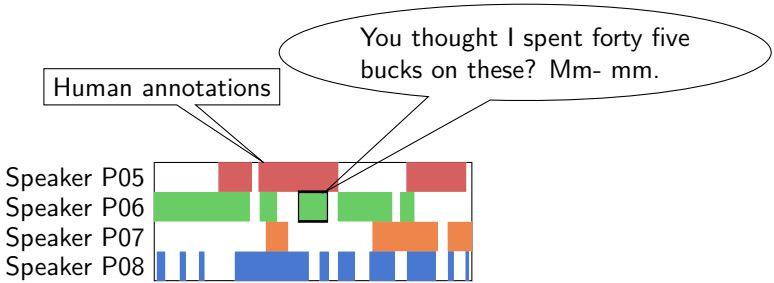
"CHiME-6 Challenge:Tackling Multispeaker Speech Recognition for Unsegmented Recordings"

arXiv preprint arXiv:2004.09249 (2020)

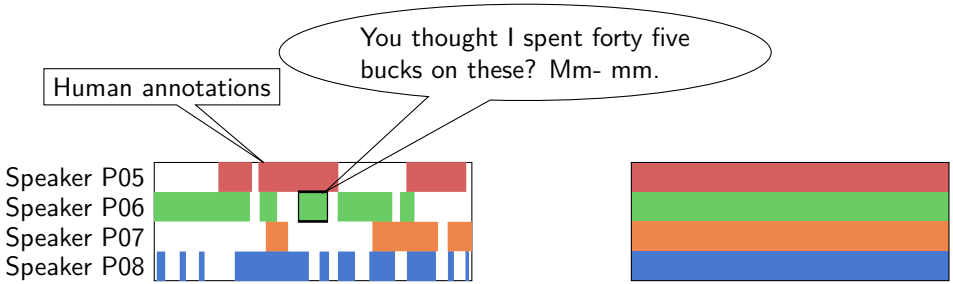
Track 1 vs Track 2



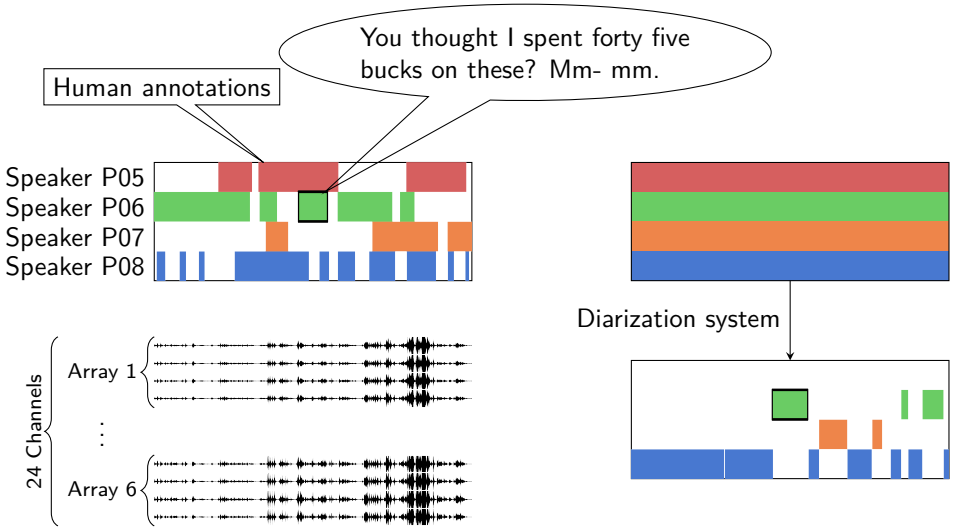
Track 1 vs Track 2



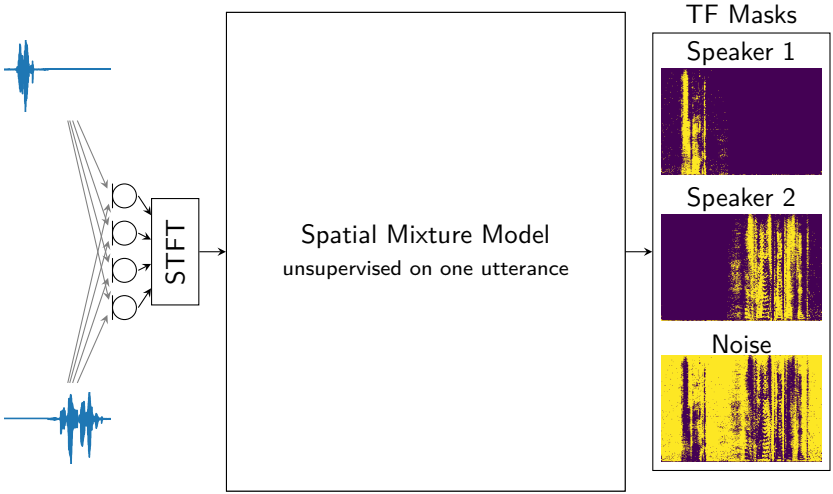
Track 1 vs Track 2



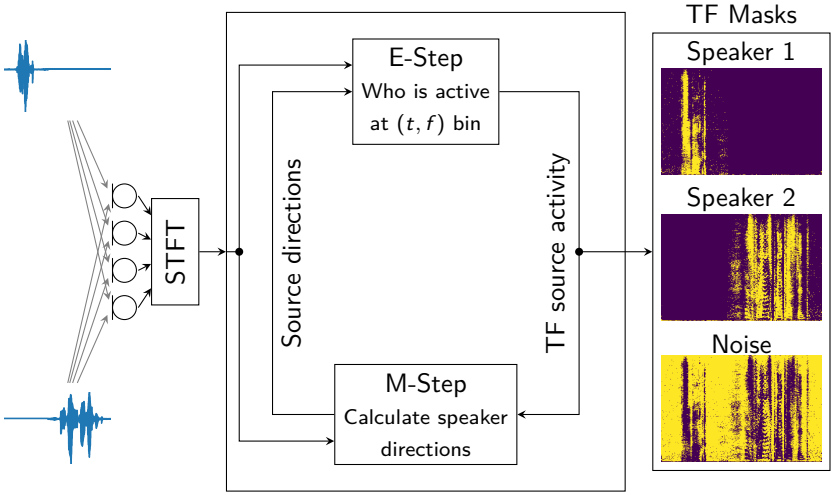
Track 1 vs Track 2



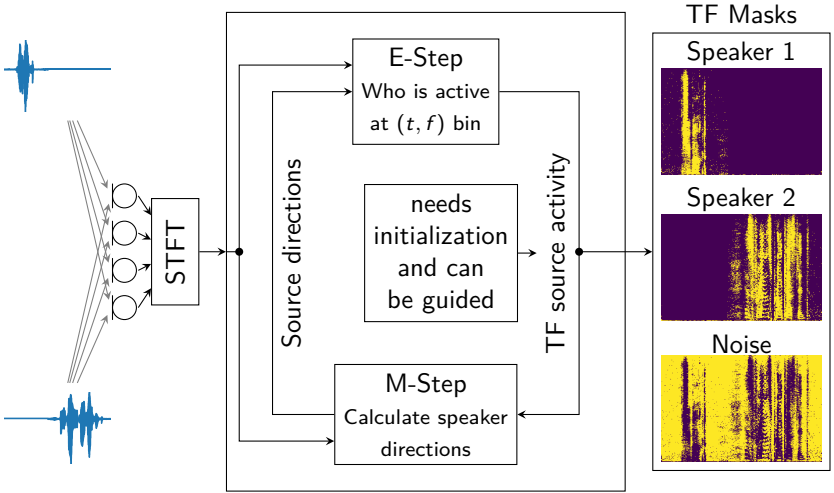
Spatial Mixture Model



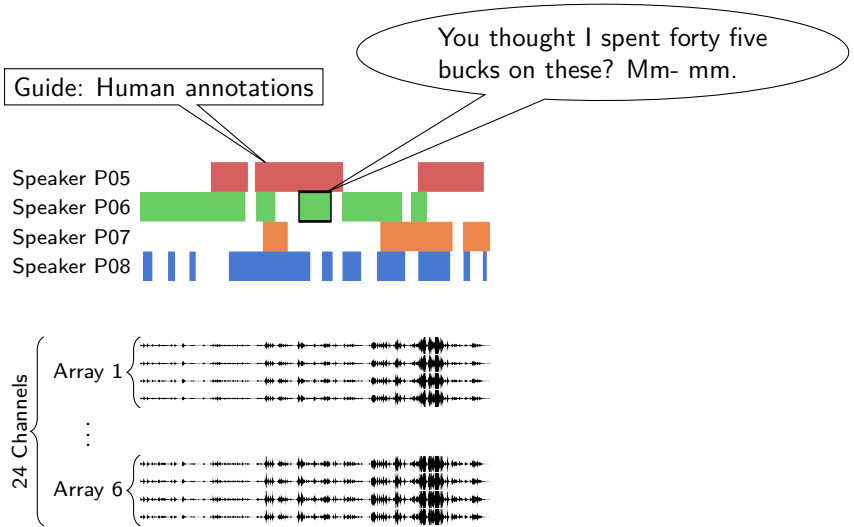
Spatial Mixture Model



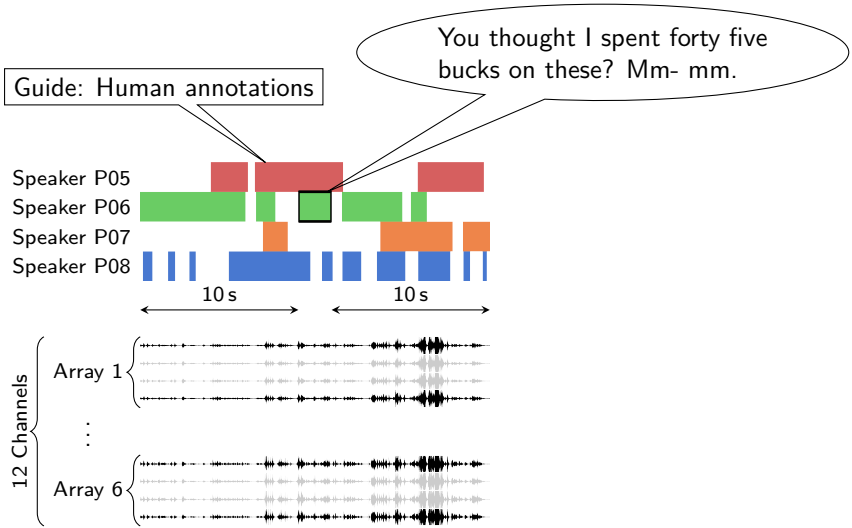
Spatial Mixture Model



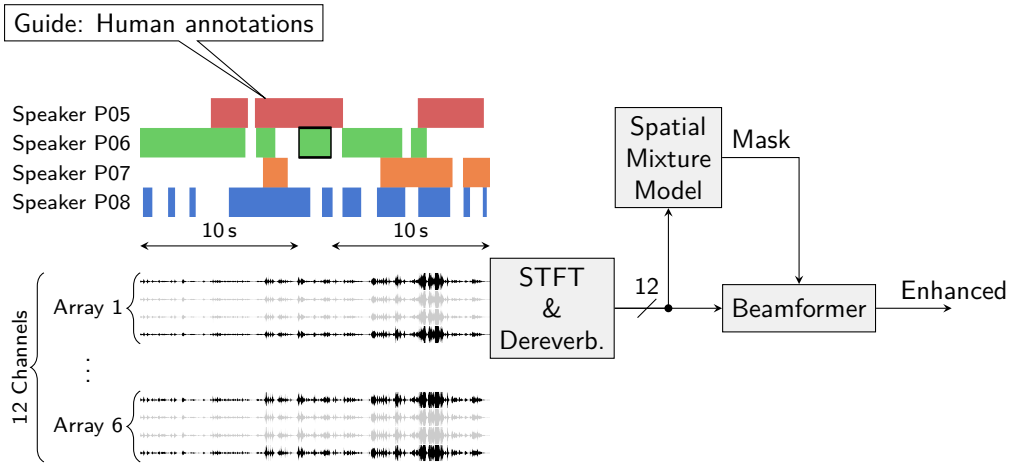
Guided Source Separation (GSS)



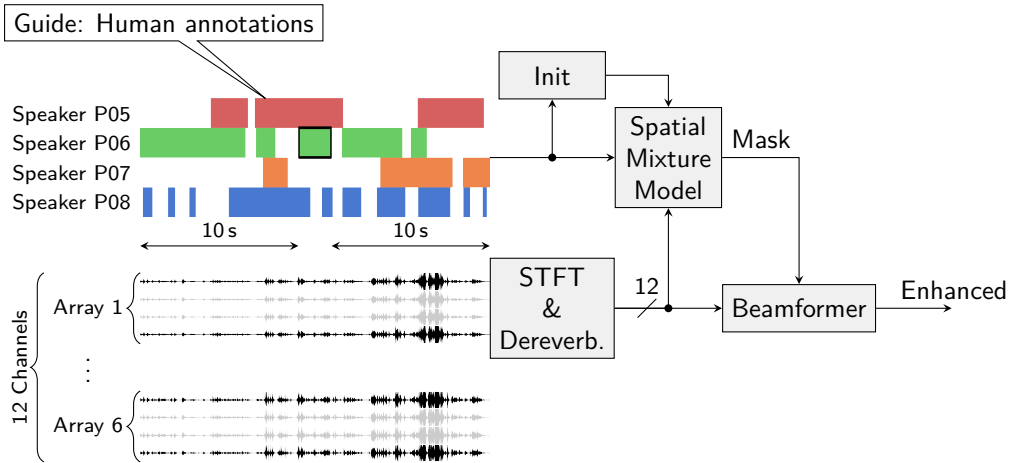
Guided Source Separation (GSS)



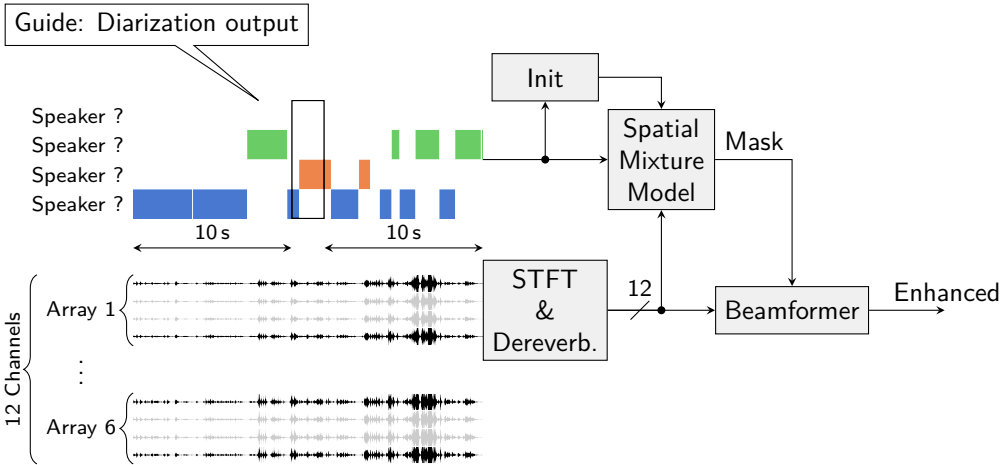
Guided Source Separation (GSS)



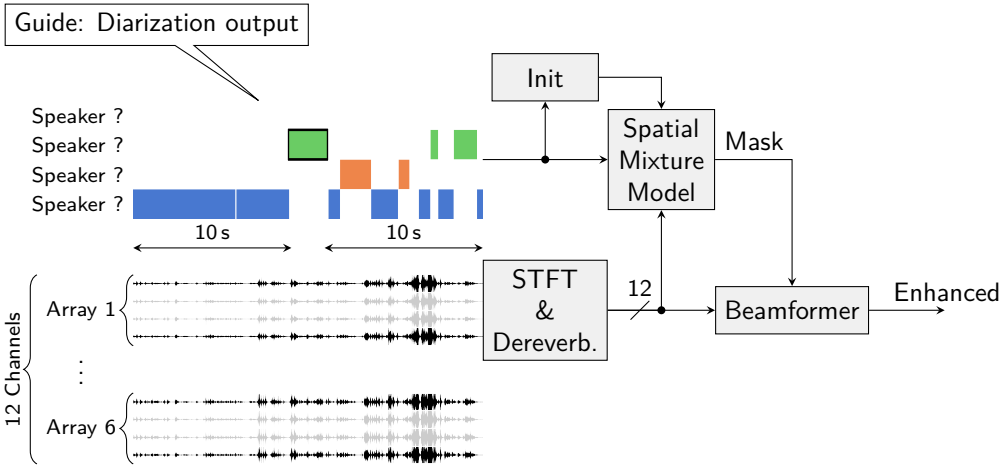
Guided Source Separation (GSS)



Guided Source Separation (GSS)



Guided Source Separation (GSS)



Implementation

- Guide replacement: Already implemented for refinement of human annotations
- Utterance boundary replacement: Missing
  - When forgotten, may increase or decrease WER
- After the Workshop: Publish code to support RTTM files for both

Acoustic model and results in Track 2 with baseline diarization system

	Acoustic model				Enh. in test	WER in %	
	Training Data	NN layers				DEV	EVAL
		CNN	TDNNF				
→ AM_0	unproc. & reverb.		15		WPE+BFIt	81.92	76.37
AM_0	unproc. & reverb.		15		GSS	78.12	73.06
AM_1	worn & WPE	10	9		GSS	76.44	72.04
AM_2	worn & GSS	10	9		GSS	74.74	71.27
AM_3	worn & GSS	10	9	DT	GSS	74.67	70.55
AM_4	worn & GSS	40 RESNET	9		GSS	74.05	70.47
AM_5	worn & GSS	40 RESNET	9	DT	GSS	74.73	70.14
AM_{0-5}					GSS	**73.50**	**68.96**

Track 2 baseline

Simular to C. Zorila, M. Li, D. Hayakawa, M. Liu, N. Ding and R. Doddipatla "Toshiba's speech recognition system for the CHiME 2020 Challenge" but used on Track 2 here.
DT: Discriminative Training, TDNNF: Factorized Time Delay Neural Network

Acoustic model and results in Track 2 with baseline diarization system

	Acoustic model				Enh. in test	WER in %	
	Training Data	NN layers				DEV	EVAL
		CNN	TDNNF				
Track 2 baseline →	AM_0	unproc. & reverb.	15		WPE+BFIt	81.92	76.37
	AM_0	unproc. & reverb.	15		GSS	78.12	73.06
	AM_1	worn & WPE	10	9	GSS	76.44	72.04
	AM_2	worn & GSS	10	9	GSS	74.74	71.27
	AM_3	worn & GSS	10	9 DT	GSS	74.67	70.55
	AM_4	worn & GSS	40 RESNET	9	GSS	74.05	70.47
	AM_5	worn & GSS	40 RESNET	9 DT	GSS	74.73	70.14
	AM_{0-5}				GSS	**73.50**	**68.96**

Simular to C. Zorila, M. Li, D. Hayakawa, M. Liu, N. Ding and R. Doddipatla "Toshiba's speech recognition system for the CHiME 2020 Challenge" but used on Track 2 here.
DT: Discriminative Training, TDNNF: Factorized Time Delay Neural Network

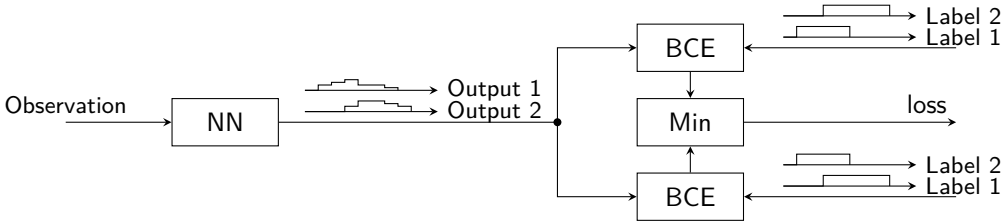
Ongoing work

Track 2: Ongoing work

- Not conforming to challenge regulations

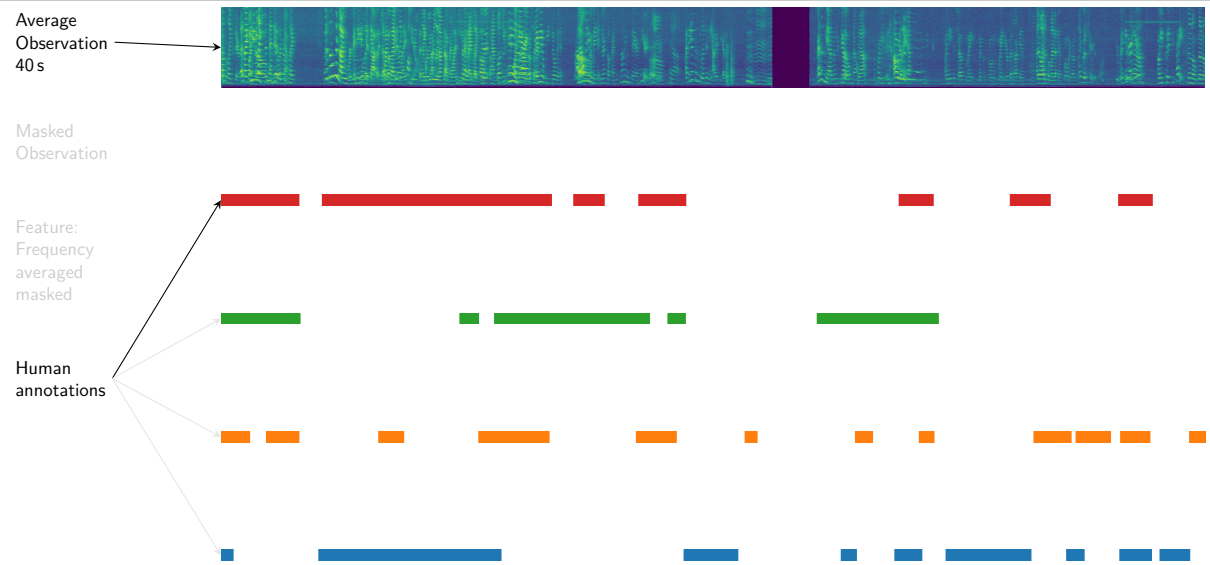
PIT Neural Speaker Diarization (NSD)

Proposed in: Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe "End-to-end neural speaker diarization with permutation-free objectives.", Interspeech (2019)

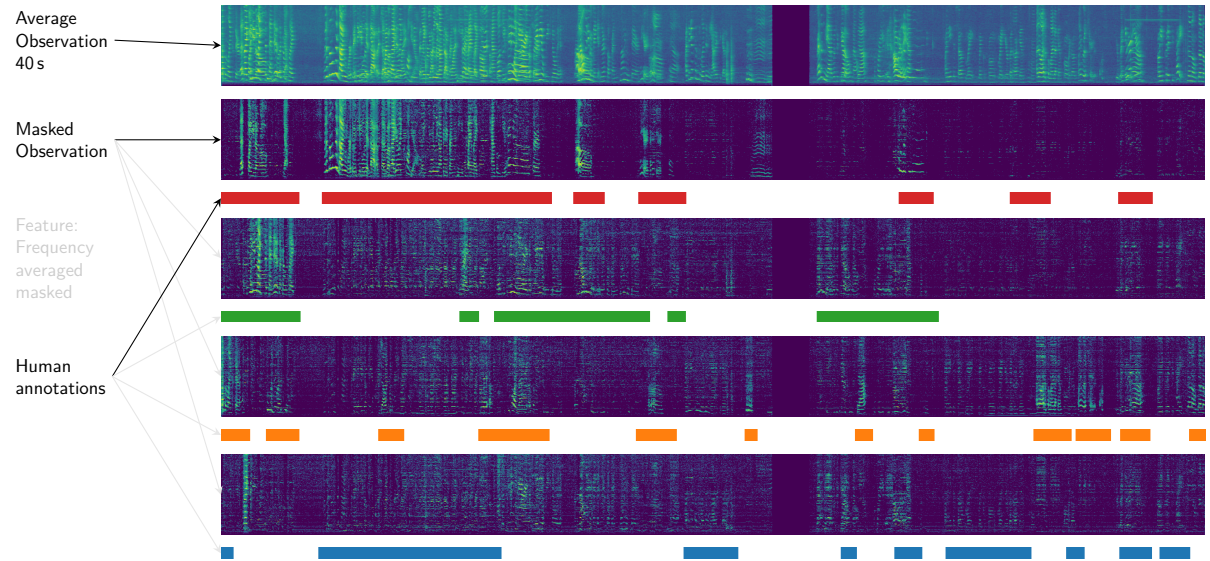


Onset and offset from simple Viterbi decoding on an HMM as used in the SAD baseline system

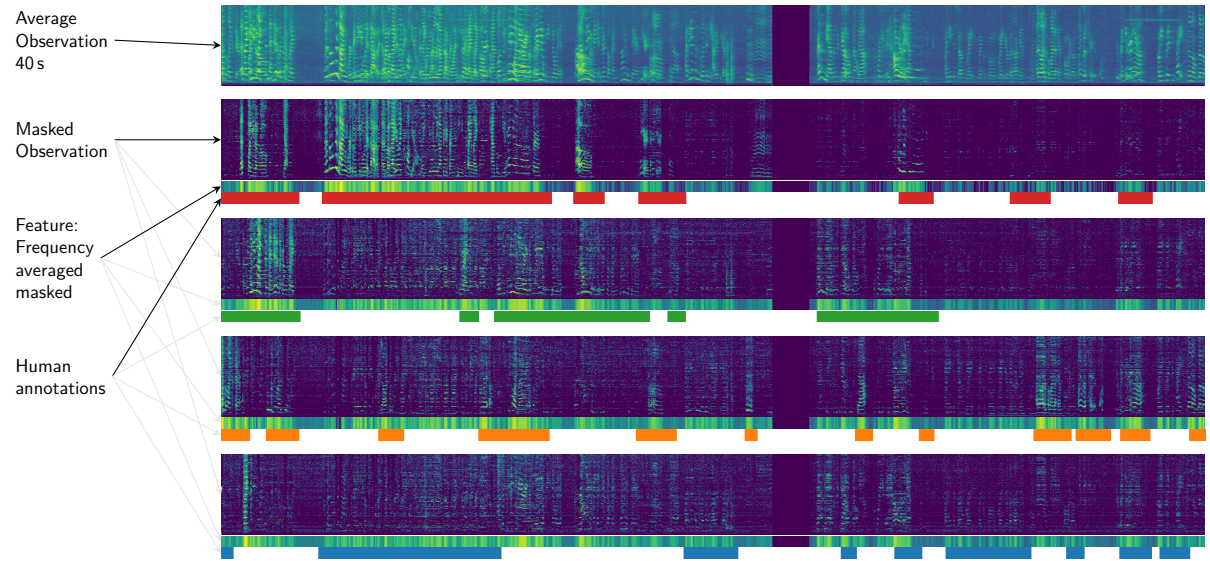
Unsupervised Spatial Mixture Model → spatial features



Unsupervised Spatial Mixture Model → spatial features



Unsupervised Spatial Mixture Model → spatial features



PIT Neural Speaker Diarization (NSD) results

Diarization	Mel bins	SMM spat. features	Dropout	DER		WER	
				TRAIN	DEV+EVAL	DEV	EVAL
Baseline[1]	-	-	-	-	60.87	77.49	71.92
2 BLSTM	80	No	No	34.28	60.09	72.68	71.25
2 BLSTM	80	Yes	No	2.52	53.75	70.05	69.50
1 BLSTM	**24**	**Yes**	**0.25**	**7.03**	**48.89**	**63.82**	**63.47**
Oracle	-	-	-	-	0	47.67	49.54 ← Track 1 baseline

Enhancement in test: GSS

Problems

However: Not conforming to challenge regulations

- Applied to 40 s segments
- Intra segment speaker permutation solved by oracle

Why didn't we address this?

- Focused on reduce overfitting first
- Overfitting source:
  - Training dataset has only 32 speakers
  - X-Vector baseline system: more than 7000 speakers from VoxCeleb

Summary

- Guided Source Separation: Replaced human annotations with diarization output
- Submitted to Track 2: Baseline Diarization System + GSS + 6 AMs
  - Category A: DEV: 73.50 %, EVAL 68.96 %
  - Category B: DEV: 73.05 %, EVAL: 68.45 %
- Neural Speaker Diarization (NSD) trained on CHiME-6 data
  - Proposed spatial feature from spatial mixture model
  - Overfitting problem
  - ToDo: Segment permutation problem solver

Thank you for your attention! Questions?

Contact me via email: boeddeker@nt.upb.de

Summary

- Guided Source Separation: Replaced human annotations with diarization output
- Submitted to Track 2: Baseline Diarization System + GSS + 6 AMs
  - Category A: DEV: 73.50 %, EVAL 68.96 %
  - Category B: DEV: 73.05 %, EVAL: 68.45 %
- Neural Speaker Diarization (NSD) trained on CHiME-6 data
  - Proposed spatial feature from spatial mixture model
  - Overfitting problem
  - ToDo: Segment permutation problem solver

Thank you for your attention! Questions?

Contact me via email: boeddeker@nt.upb.de