Speaker Diarization and Separation in Multi-talker Interaction Scenarios

Jun Du
University of Science and Technology of China (USTC)

Nov. 21, 2020

Outline

- Background
- Speaker Diarization (DIHARD I/II/III)
- Speech Separation (CHiME-5/CHiME-6)
- Speaker Diarization and Separation (CHiME-6/JSALT 2020)
- Summary

The Challenge of Cocktail Party Problem

- Raised by Colin Cherry (Cognitive Scientist) in 1953
- How to imitate the processing of multi-stream signals by human ears?

→ **Background Noises**

→ **Room Reverberation**

→ **Multiple Speakers**

→ **Lombard Effects**

One Ultimate Goal to Achieve Human-Level Auditory Perception

A Mathematical Perspective

Independent Research

- Background Noise

  $x \approx s + n$ → Speech Enhancement (SE)

  $s = f_1(x)$

- Reverberation

  $x \approx h * s$ → Speech Dereverberation (SD)

  $s = f_2(x)$

- Multiple Speakers

  $x \approx s_1 + s_2$ → Speech Separation (SS)

  $(s_1, s_2)= f_3(x)$

- Real Scenarios

  $x \approx h * (s_1 + s_2) + n$ → Speech Separation (SS)

  $(s_1, s_2)= f(x)$

Underdetermined Problem

Speaker Diarization vs. Speech Separation

The Gap Between Research and Reality

Timeline	2006	2007-2013	2015	2016	2018	2019	2020	
Research Task	SSC	SiSEC	SiSEC	SiSEC	CHiME-5	BabyTrain	CHiME-6	
	AMI		WSJ0-2mix		DIHARD-I	DIHARD-II	DIHARD-III	LibriCSS

Research Trend

- Single microphone → Microphone array → Microphone arrays
- Single environment factor → Multiple environment factors
- Small vocabulary → Medium/Large vocabulary
- Reading style → Conversional style
- Small-scale training data → Large-scale training data
- Simulation data → Realistic data

Outline

- Background
- Speaker Diarization (DIHARD I/II/III)
- Speech Separation (CHiME-5/CHiME-6)
- Speaker Diarization and Separation (CHiME-6/JSALT 2020)
- Summary

JSALT 2017: The Origin of DIHARD Challenge

Enhancement and Analysis of Conversational Speech

LDC
IBM
Apple
JHU
CMU
ENS
USTC
IISc

DIHARD-I Challenge (2018)

- Background

  Neville Ryant, Elika Bergelson, Kenneth Church, Alejandrina Cristia, Jun Du, et al. "ENHANCEMENT AND ANALYSIS OF CONVERSATIONAL SPEECH: JSALT 2017," ICASSP 2018.

- INTERSPEECH 2018 Special Session
  - The First DIHARD Speech Diarization Challenge
  - Challenge website: https://dihardchallenge.github.io/dihard1/

- Two tracks
  - Track 1: diarization beginning from gold speech segmentation
  - Track 2: diarization from scratch

JSALT 2019: Speaker Diarization Again

Speaker Detection in Adverse Scenarios with a Single Microphone

JHU
SRI
ENS
LIMSI
USTC
NEC
ETS

Leibny Paola Garcia, Jesus Villalba, Herve Bredin, Jun Du, Diego Castan, Alejandrina Cristia, et al. “Speaker Detection in the Wild: Lessons Learned from JSALT 2019,” Odyssey 2020.

DIHARD-II Challenge (2019)

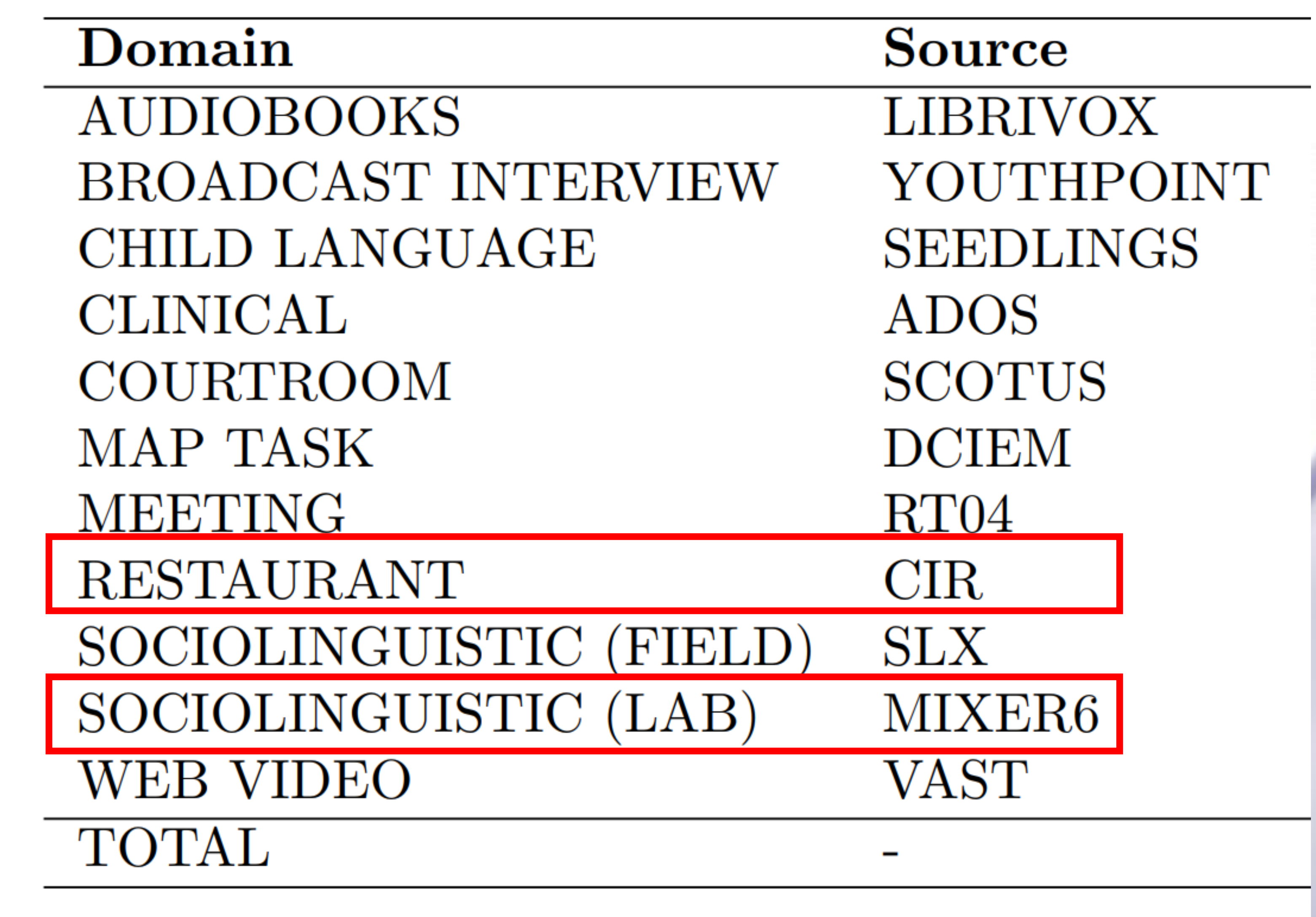
- Background

  Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, et al. “The Second DIHARD Diarization Challenge: Dataset, task, and baselines,” INTERSPEECH 2019.

- INTERSPEECH 2019 Special Session
  - The Second DIHARD Speech Diarization Challenge
  - Challenge website: https://dihardchallenge.github.io/dihard2/
- Single-channel Tracks (Track1 and Track2)
  - Two more domains
- Multichannel Tracks (Track3 and Track4)
  - CHiME-5 corpus

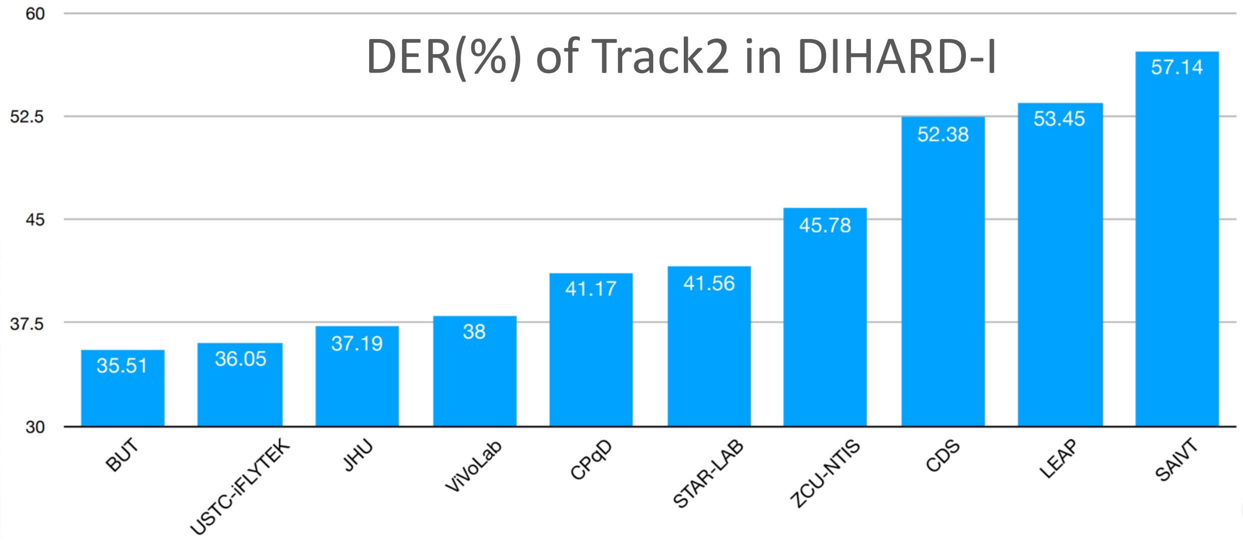
Application Scenairos

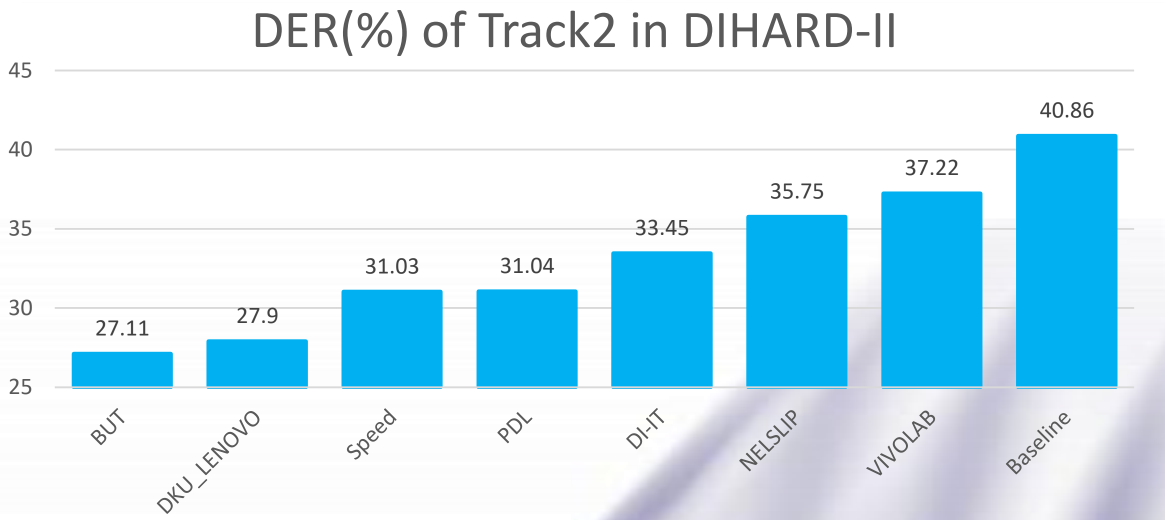
Single-channel Tracks



Domain	Source
AUDIOBOOKS	LIBRIVOX
BROADCAST INTERVIEW	YOUTHPOINT
CHILD LANGUAGE	SEEDLINGS
CLINICAL	ADOS
COURTROOM	SCOTUS
MAP TASK	DCIEM
MEETING	RT04
RESTAURANT	CIR
SOCIOLINGUISTIC (FIELD)	SLX
SOCIOLINGUISTIC (LAB)	MIXER6
WEB VIDEO	VAST
TOTAL	-

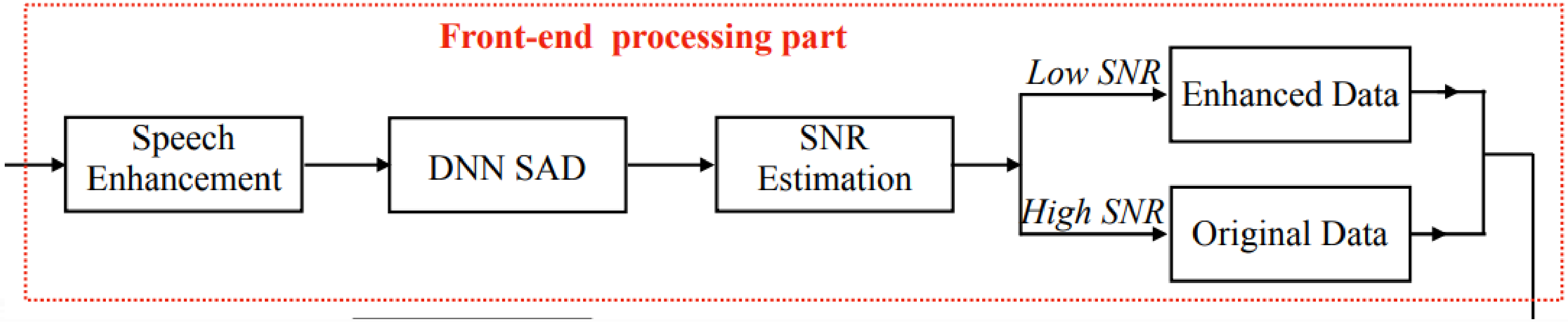
Leaderboard of DIHARD-I and DIHARD-II

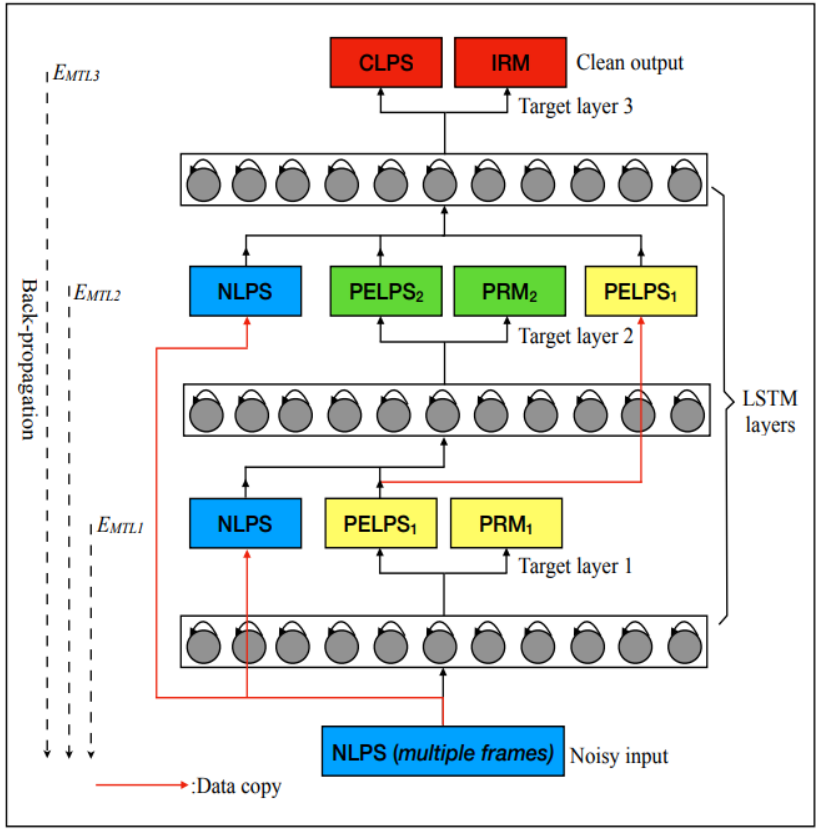




- Challenge of speaker diarization
  - Single-channel speech enhancement
  - Clustering: feature design, clustering algorithm, re-segmentation
  - Overlapped speech processing

Single-channel Speech Enhancement



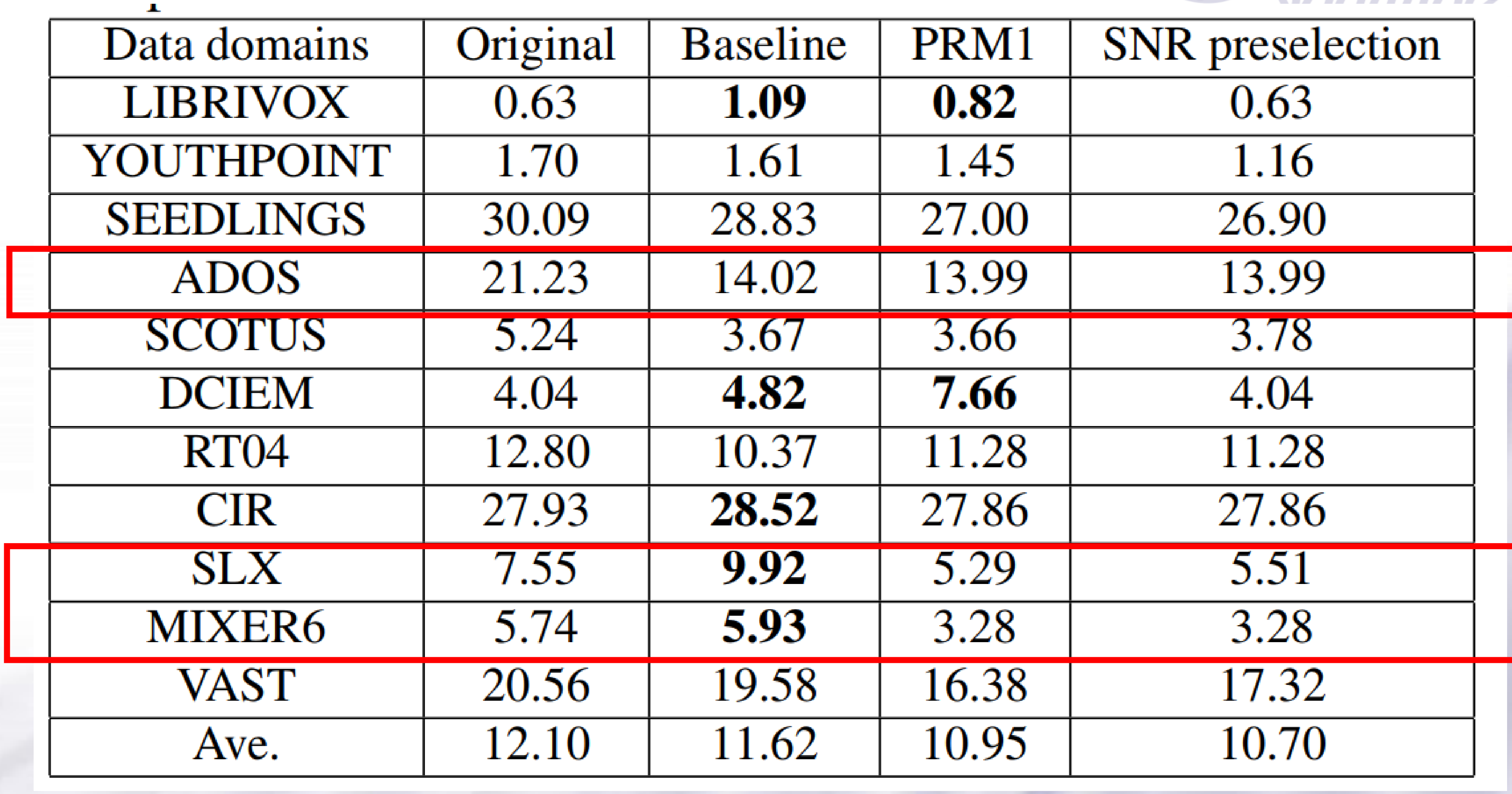


- SNR pre-selection
  - No processing for high-SNR cases
- Progressive ratio mask (PRM)
  - Additional intermediate targets

https://github.com/jsalt2019-diadet/speech_denoising_tools

[1] Lei Sun, Jun Du, etc., "A novel LSTM-based speech preprocessor for speaker diarization in realistic mismatch conditions," ICASSP 2018.
[2] Lei Sun, Jun Du, etc., "Speaker diarization with enhancing speech for the first DIHARD Challenge," INTERSPEECH 2018.
[3] Lei Sun, Jun Du, etc, "Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization," ICASSP 2020

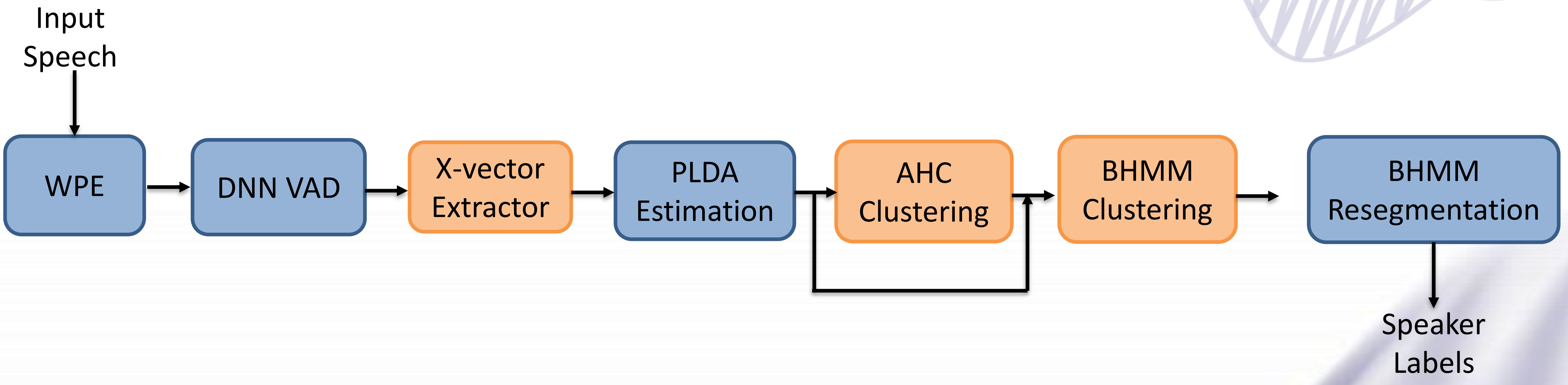
Single-channel Speech Enhancement



Data domains	Original	Baseline	PRM1	SNR preselection
LIBRIVOX	0.63	**1.09**	**0.82**	0.63
YOUTHPOINT	1.70	1.61	1.45	1.16
SEEDLINGS	30.09	28.83	27.00	26.90
ADOS	21.23	14.02	13.99	13.99
SCOTUS	5.24	3.67	3.66	3.78
DCIEM	4.04	**4.82**	**7.66**	4.04
RT04	12.80	10.37	11.28	11.28
CIR	27.93	**28.52**	27.86	27.86
SLX	7.55	**9.92**	5.29	5.51
MIXER6	5.74	**5.93**	3.28	3.28
VAST	20.56	19.58	16.38	17.32
Ave.	12.10	11.62	10.95	10.70

Achieving consistent DER reductions for all domains

BUT DIHARD-II System [4]

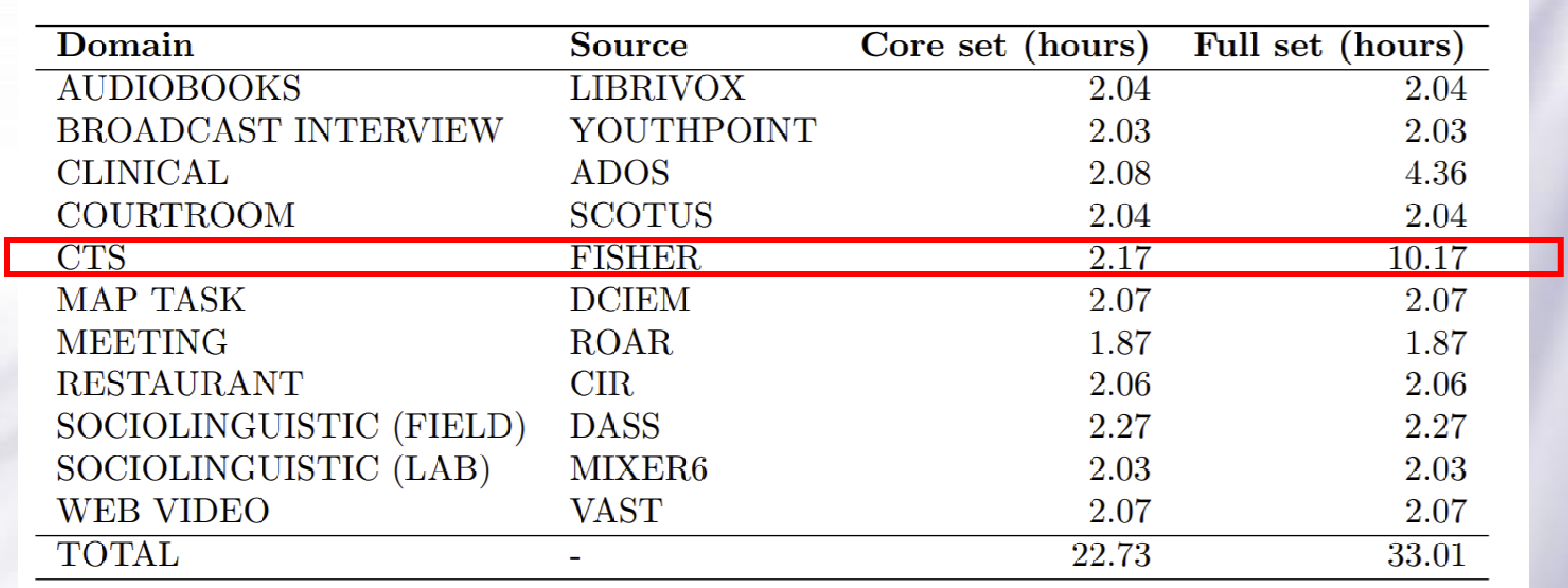


- X-vector extractor [1]: a higher frame-rate of 0.25s
- Two-stage clustering [2,3]: AHC over x-vectors, followed by the Bayesian HMM at frame level

[1] M. Diez, L. Burget, F. Landini, et al. “Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge,” ICASSP 2020
[2] M. Diez, L. Burget, F. Landini, et al. “Analysis of speaker diarization based on bayesian hmm with eigenvoice priors,” IEEE/ACM TASLP, 2019.
[3] M. Diez, L. Burget, F. Landini, et al. “Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge,” ICASSP 2020.
[4] F. Landini, S. Wang, M. Diez, et al. “BUT System for the Second DIHARD Speech Diarization Challenge,” ICASSP 2020.

DIHARD-III Challenge (2020)

- The Third DIHARD Speech Diarization Challenge is ongoing
  - Challenge website: https://dihardchallenge.github.io/dihard3/
- Hosted by NIST through the OpenSAT: https://sat.nist.gov/dihard3
- DIHARD workshop: Jan. 23, 2021 (after SLT)



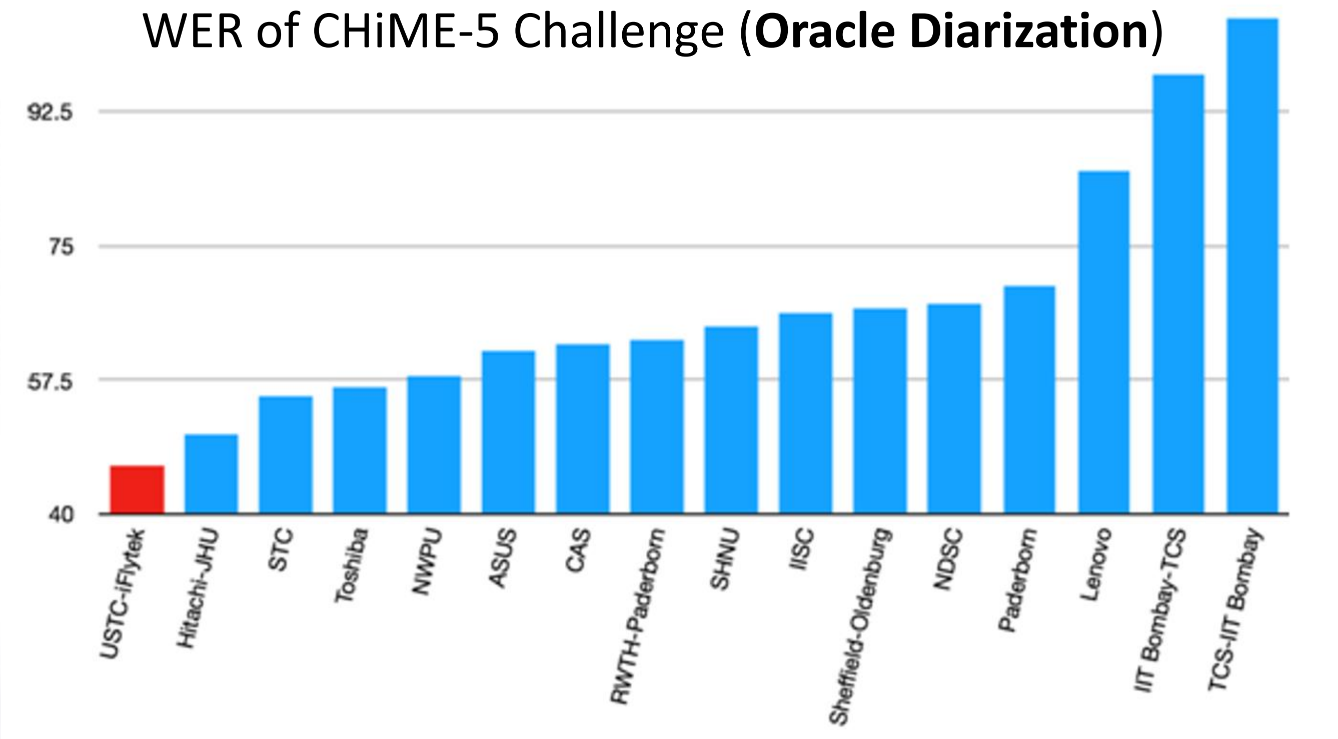
Domain	Source	Core set (hours)	Full set (hours)
AUDIOBOOKS	LIBRIVOX	2.04	2.04
BROADCAST INTERVIEW	YOUTHPOINT	2.03	2.03
CLINICAL	ADOS	2.08	4.36
COURTROOM	SCOTUS	2.04	2.04
CTS	FISHER	2.17	10.17
MAP TASK	DCIEM	2.07	2.07
MEETING	ROAR	1.87	1.87
RESTAURANT	CIR	2.06	2.06
SOCIOLINGUISTIC (FIELD)	DASS	2.27	2.27
SOCIOLINGUISTIC (LAB)	MIXER6	2.03	2.03
WEB VIDEO	VAST	2.07	2.07
TOTAL	-	22.73	33.01

Outline

- Background
- Speaker Diarization (DIHARD I/II/III)
- Speech Separation (CHiME-5/CHiME-6)
- Speaker Diarization and Separation (CHiME-6/JSALT 2020)
- Summary

CHiME-5 Challenge

http://spandh.dcs.shef.ac.uk/chime_challenge/CHiME5

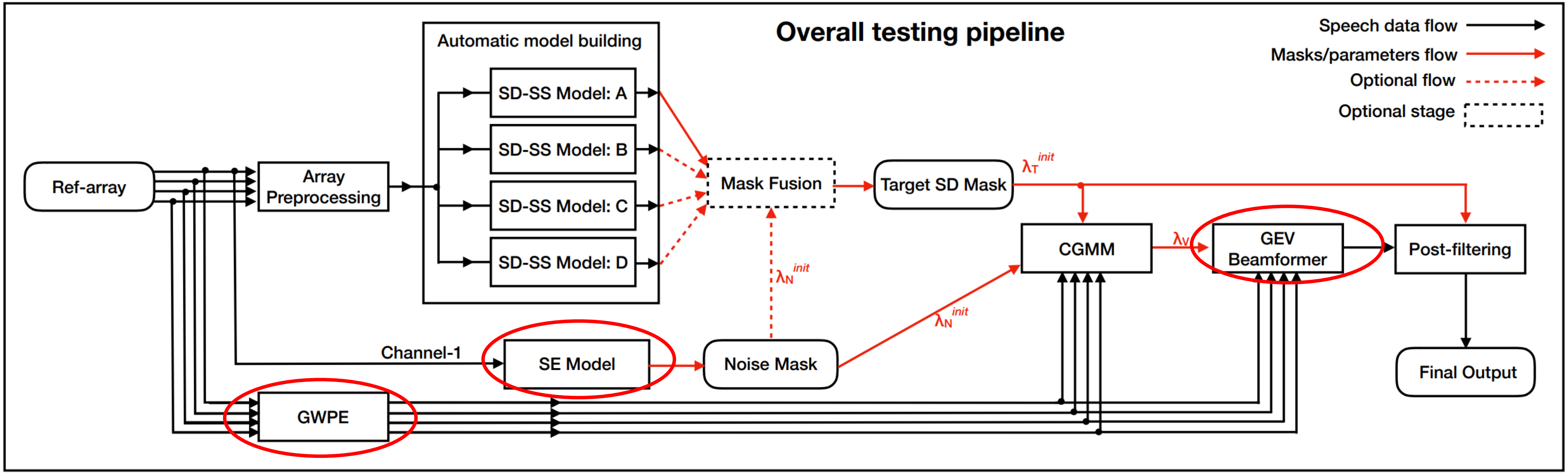


Dinner Party
(Far-field, Conversations, Multiple Speakers)

A small step towards solving the cocktail party problem

Our Front-End Solution for CHiME-5 Challenge

Multi-stage front-end by joint speech denoising, dereverberation and separation

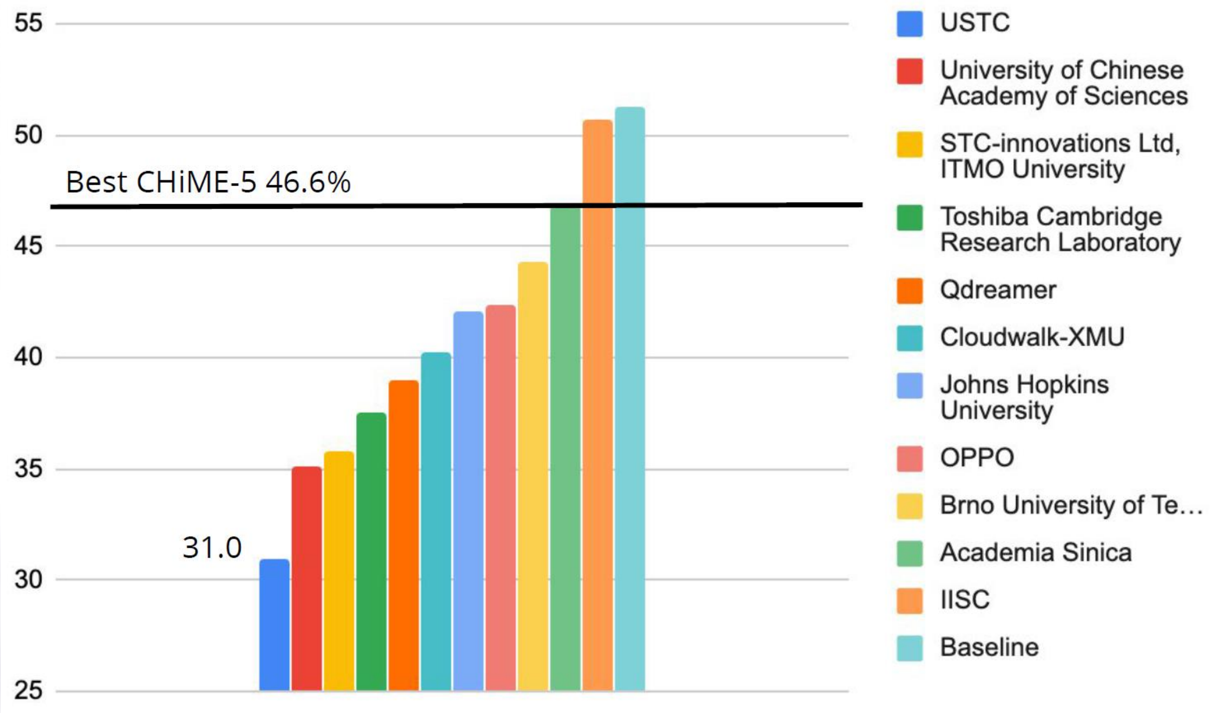


Yan-Hui Tu, Jun Du, Tian Gao, and Chin-Hui Lee, "A multi-target SNR-progressive learning approach to Regression Based Speech Enhancement," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 28, pp.1608-1619, 2020.

Lei Sun, Jun Du, etc., "A speaker-dependent single-channel/multichannel approach for front-end of CHiME-5 Challenge under far-field multi-talker scenario," *Journal of Selected Topics in Signal Processing*, Vol. 13, No. 4, pp. 827-840, 2019.

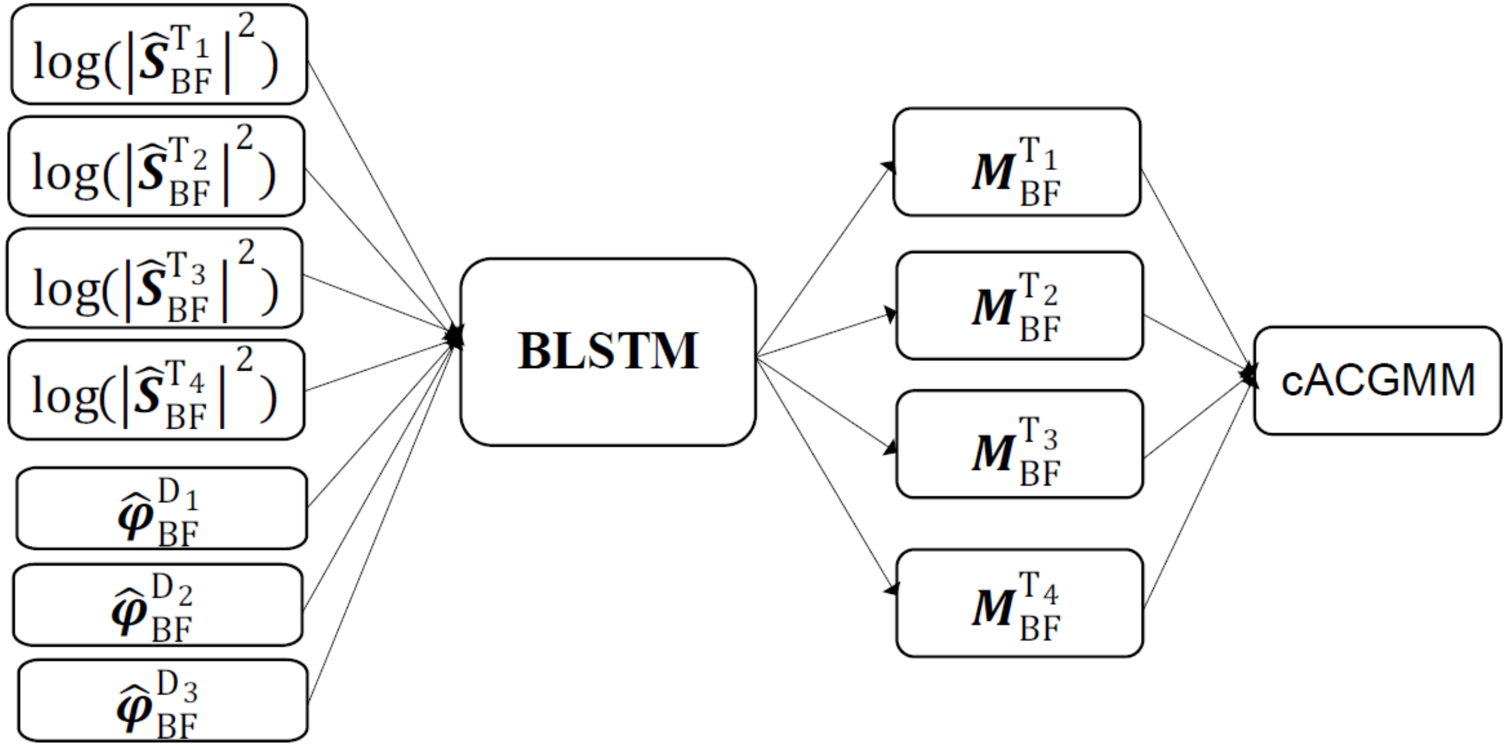
CHiME-6 Challenge

- What’s new (https://chimechallenge.github.io/chime6)
  - Better baseline results with new array synchronization
  - Guided source separation (GSS)
  - One more track (Track 2): diarization and recognition



Track 1: recognition with oracle speaker diarization

Track 1: Our Front-End Solution for CHiME-6


$\log(|\hat{S}_{BF}^{T_1}|^2)$
$\log(|\hat{S}_{BF}^{T_2}|^2)$
$\log(|\hat{S}_{BF}^{T_3}|^2)$
$\log(|\hat{S}_{BF}^{T_4}|^2)$
$\hat{\varphi}_{BF}^{D_1}$
$\hat{\varphi}_{BF}^{D_2}$
$\hat{\varphi}_{BF}^{D_3}$
BLSTM
$M_{BF}^{T_1}$
$M_{BF}^{T_2}$
$M_{BF}^{T_3}$
$M_{BF}^{T_4}$
cACGMM

Space-and-Speaker-Aware Iterative Mask Estimation

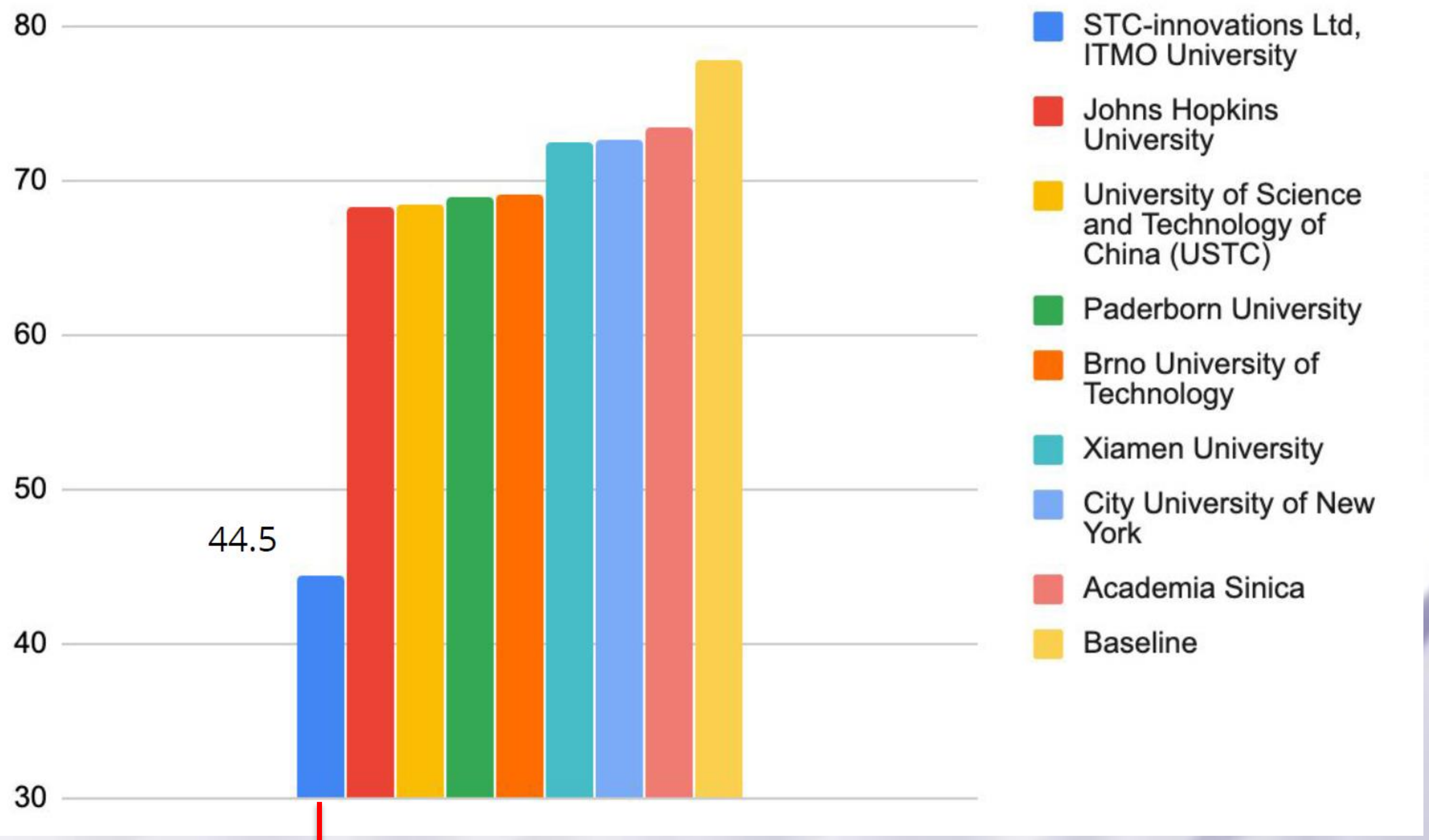
Yan-Hui Tu, Jun Du, Lei Sun, Feng Ma, Jia Pan, Chin-Hui Lee, "A space-and-speaker-aware iterative mask estimation approach to multi-channel speech recognition in the CHiME-6 Challenge," INTERSPEECH 2020.

Outline

- Background
- Speaker Diarization (DIHARD I/II/III)
- Speech Separation (CHiME-5/CHiME-6)
- Speaker Diarization and Separation (CHiME-6/JSALT 2020)
- Summary

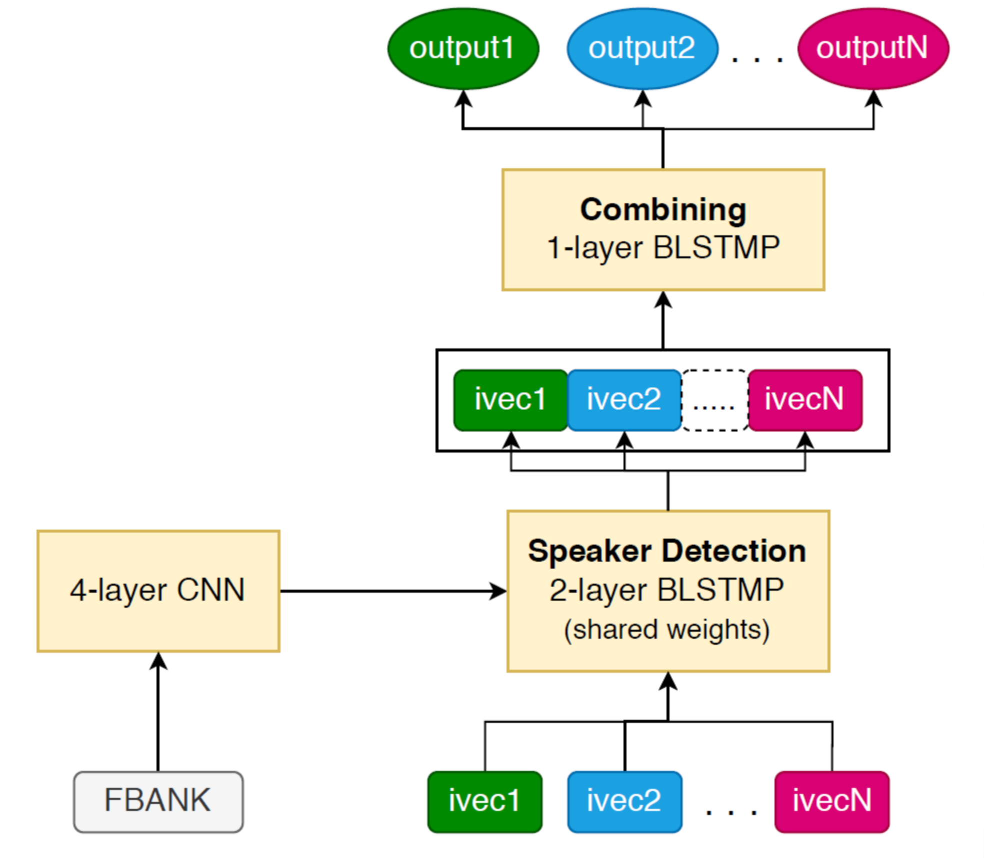
CHiME-6 Challenge Track 2

Track 2: diarization and recognition



The performance for speaker diarization is the key for subsequent speech separation

Track 2: TS-VAD from STC Team

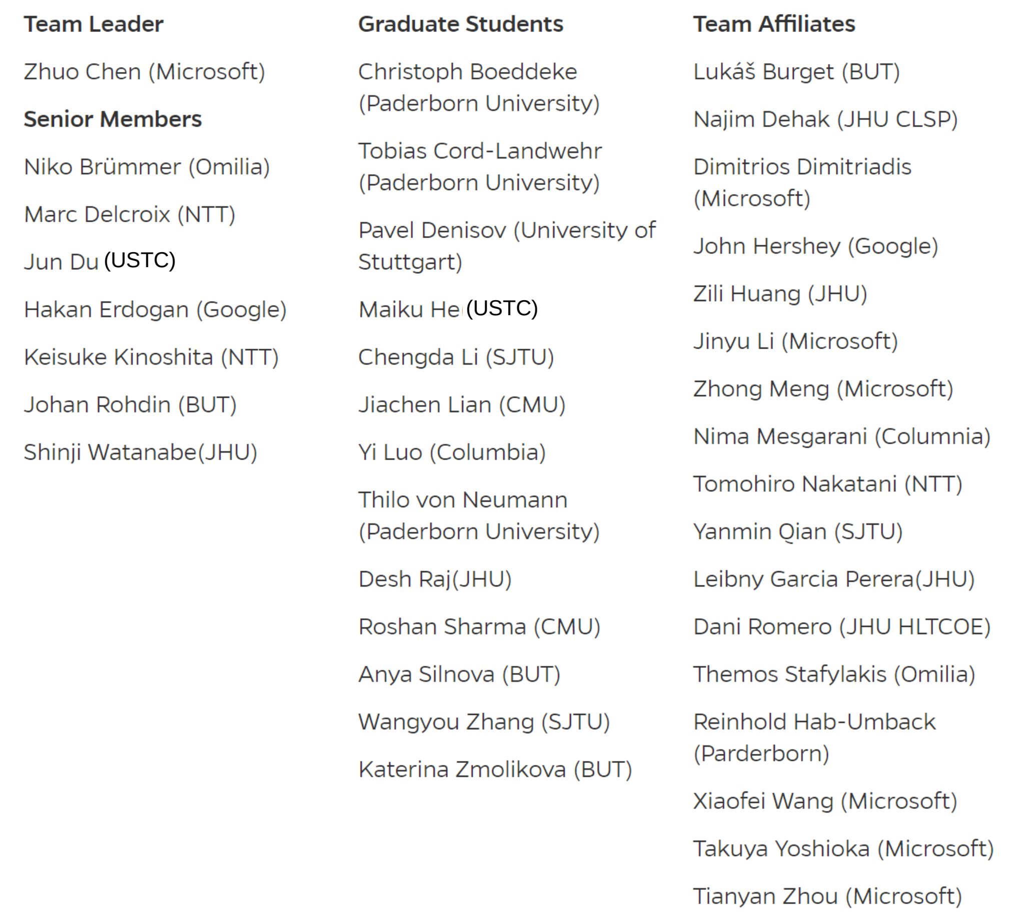


- Supervised approach for speaker diarization
- Iterative diarization with significant gains
- Problem 1: fixed number of speakers
- Problem 2: generalization capability

Ivan Medennikov, Maxim Korenevsky, Tatiana Prisyach, Yuri Y. Khokhlov, Mariya Korenevskaya, Ivan Sorokin, Tatiana V. Timofeeva, Anton Mitrofanov, Andrei Andrusenko, Ivan Podluzhny, Aleksandr Laptev, and Aleksei Romanenko, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," ArXiv, vol. abs/2005.07272, 2020.

JSALT 2020 (Virtual Workshop)

Speech Recognition and Diarization for Unsegmented Multi-talker Recordings with Speaker Overlaps



Team Leader

Zhuo Chen (Microsoft)

Senior Members

Niko Brümmer (Omilia)

Marc Delcroix (NTT)

Jun Du (USTC)

Hakan Erdogan (Google)

Keisuke Kinoshita (NTT)

Johan Rohdin (BUT)

Shinji Watanabe(JHU)

Graduate Students

Christoph Boeddeke (Paderborn University)

Tobias Cord-Landwehr (Paderborn University)

Pavel Denisov (University of Stuttgart)

Maiku He (USTC)

Chengda Li (SJTU)

Jiachen Lian (CMU)

Yi Luo (Columbia)

Thilo von Neumann (Paderborn University)

Desh Raj(JHU)

Roshan Sharma (CMU)

Anya Silnova (BUT)

Wangyou Zhang (SJTU)

Katerina Zmolikova (BUT)

Team Affiliates

Lukáš Burget (BUT)

Najim Dehak (JHU CLSP)

Dimitrios Dimitriadis (Microsoft)

John Hershey (Google)

Zili Huang (JHU)

Jinyu Li (Microsoft)

Zhong Meng (Microsoft)

Nima Mesgarani (Columnia)

Tomohiro Nakatani (NTT)

Yanmin Qian (SJTU)

Leibny Garcia Perera(JHU)

Dani Romero (JHU HLTCOE)

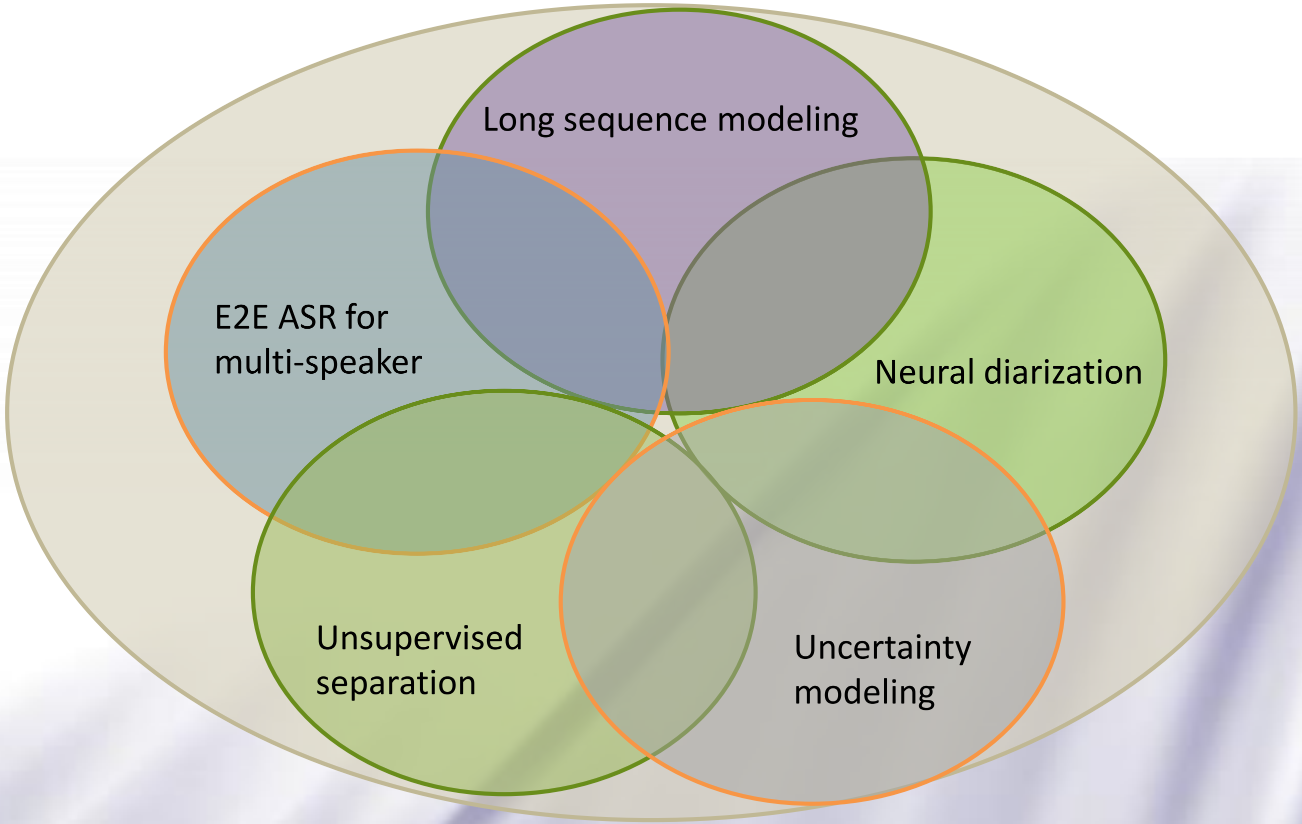
Themos Stafylakis (Omilia)

Reinhold Hab-Umback (Parderborn)

Xiaofei Wang (Microsoft)

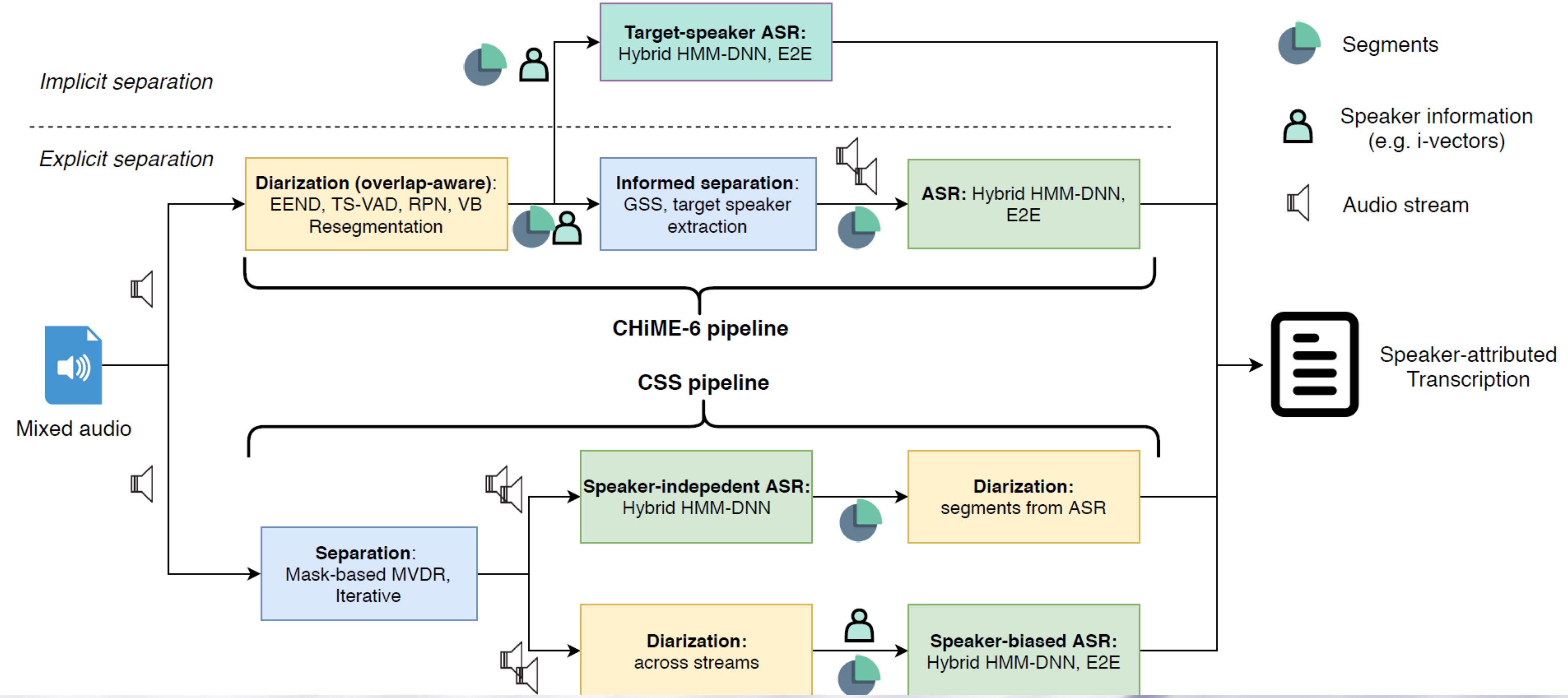
Takuya Yoshioka (Microsoft)

Tianyan Zhou (Microsoft)



Deep collaborations between industry (17 members) and academia (33 members)

Overall Pipeline


Implicit separation
Explicit separation
Target-speaker ASR: Hybrid HMM-DNN, E2E
Diarization (overlap-aware): EEND, TS-VAD, RPN, VB Resegmentation
Informed separation: GSS, target speaker extraction
ASR: Hybrid HMM-DNN, E2E
CHiME-6 pipeline
CSS pipeline
Mixed audio
Separation: Mask-based MVDR, Iterative
Speaker-indepedent ASR: Hybrid HMM-DNN
Diarization: segments from ASR
Diarization: across streams
Speaker-biased ASR: Hybrid HMM-DNN, E2E
Segments
Speaker information (e.g. i-vectors)
Audio stream
Speaker-attributed Transcription

Desh Raj, Pavel Denisov, Zhuo Chen, Hakan Erdogan, Zili Huang, Maokui He, Shinji Watanabe, Jun Du, Takuya Yoshioka, Yi Luo, Naoyuki Kanda, Jinyu Li, Scott Wisdom, and John R. Hershey, “Integration of speech separation, diarization, and recognition for multi-speaker meetings: system description, comparison, and analysis,” SLT 2021.

Outline

- Background
- Speaker Diarization (DIHARD I/II/III)
- Speech Separation (CHiME-5/CHiME-6)
- Speaker Diarization and Separation (CHiME-6/JSALT 2020)
- Summary

Summary

- Speaker diarization in adverse environments
  - Preprocessing, speaker embedding, BHMM, TSVAD, …
  - Combining different unsupervised and supervised approaches
- Speech separation in adverse environments
  - Joint modeling of multiple factors (noises, reverberation, interfering speakers)
  - One-stage approach (or end-to-end) vs. multi-stage approach (or iterative)
- Speaker diarization and separation
  - Overlap detection and separation
  - Multi-stage approach to combine diarization and separation
  - Joint optimization with the downstream tasks

Computer Speech and Language Special Issue on Separation, Recognition, and Diarization of Conversational Speech

Thank You!

Q&A