The USTC-NELSLIP Systems for DIHARD III Challenge

Maokui He, Yuxuan Wang, Shutong Niu, **Lei Sun**, Tian Gao, Xin Fang, Jia Pan, Jun Du, Chin-Hui Lee

National Engineering Lab for Speech and Language Information Processing (NELSLIP)

University of Science and Technology of China (USTC)

iFlytek Research

01/23/2021

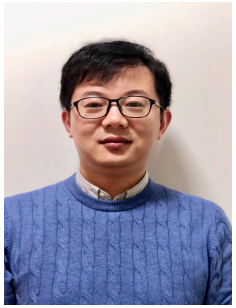
Team

Maokui He (USTC)

Yuxuan Wang (USTC)

Shutong Niu (USTC)

Lei Sun (iFlytek)



Tian Gao (iFlytek)

Xin Fang (iFlytek)

Jia Pan (iFlytek)

Jun Du (USTC)

Chin-Hui Lee (GIT)

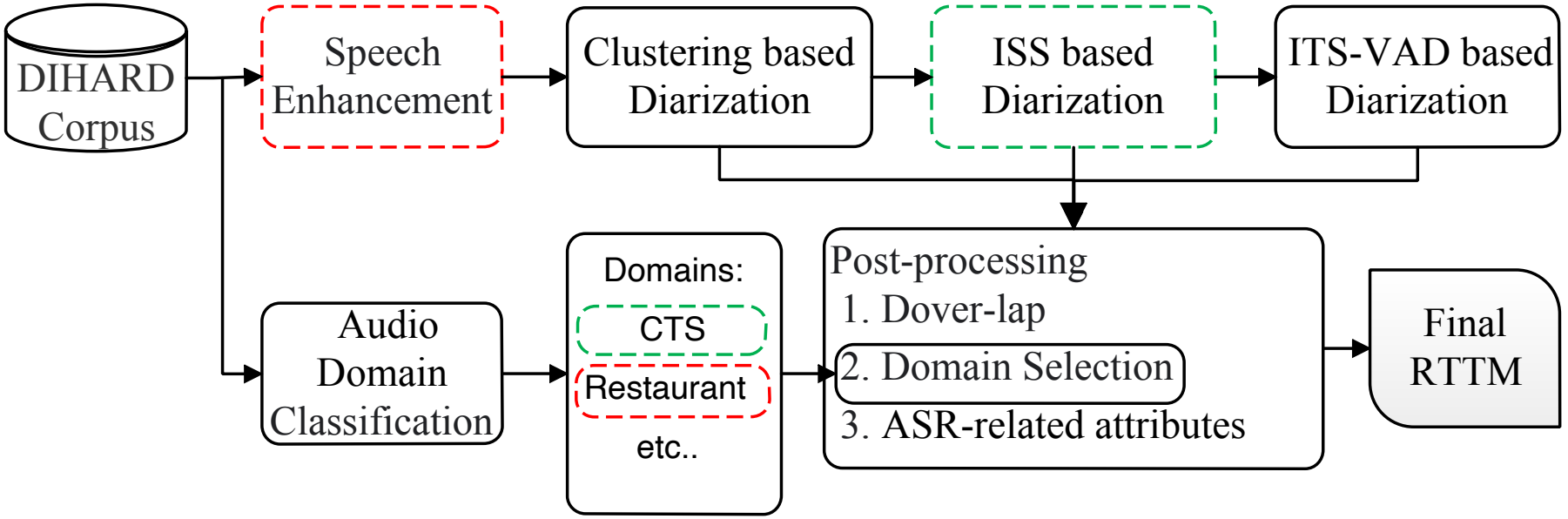
Motivation

- Pain points in DIHARD
  - Overlapped speech: detection, assignment, etc..
  - Diverse environments: telephone, cafe, street, etc..

Improving generalization ability

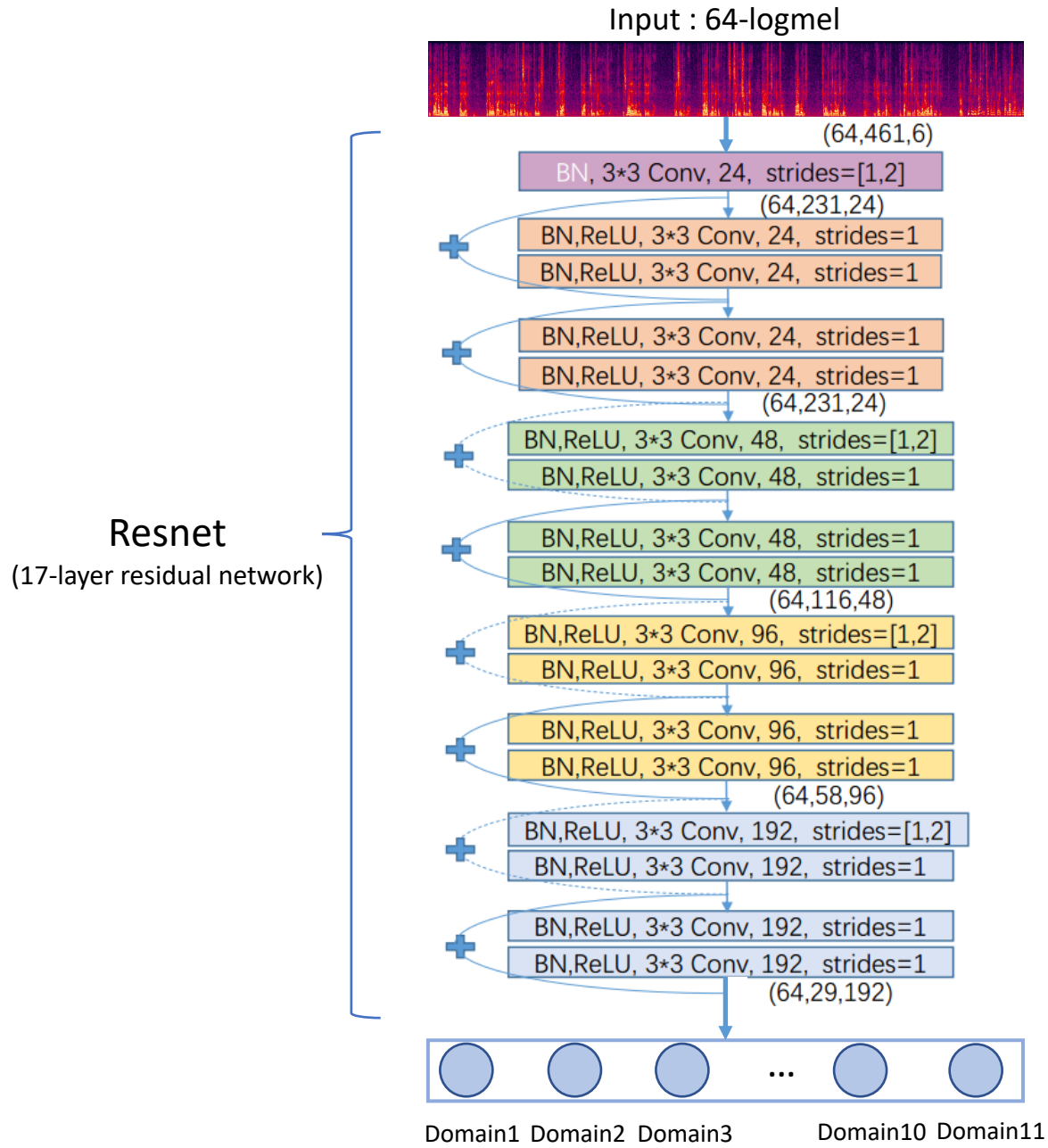
- Proposed main ideas
  - Iterative (multiple stages) strategy
  - Domain-dependent processing

System Overview



- Three main diarization systems:
  - Clustering based Diarization
  - Iterative Speech Separation (ISS) based Diarization
  - Iterative Target-speaker VAD (ITS-VAD) based Diarization
- Several auxiliary techniques:
  - Audio Domain Classification
  - Speech Enhancement
  - Dover-lap for system fusion
  - ASR-related attributes

Audio Domain Classification



Resnet
(17-layer residual network)

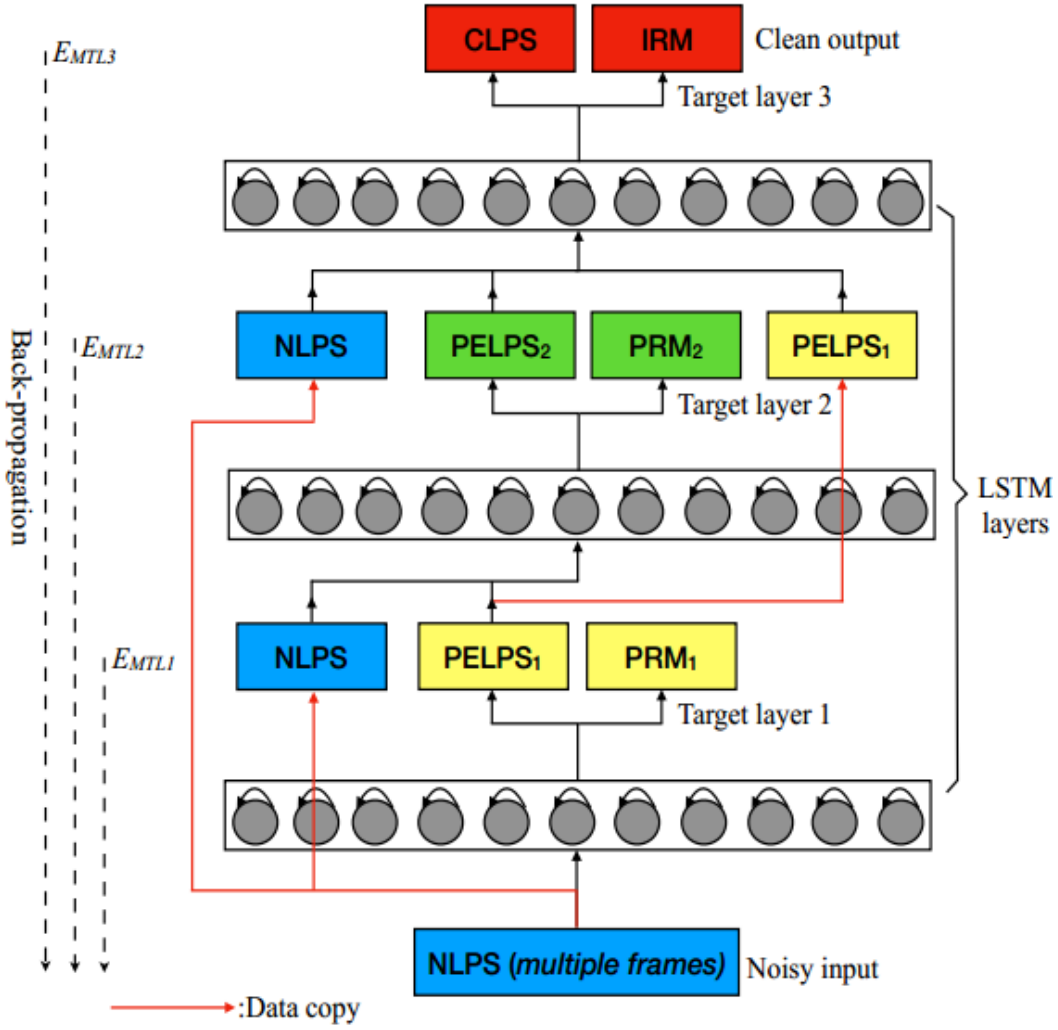
- Training set :

9/10 DIHARD III DEV set
(truncated into 10-second segments)

- Testing set :

another 1/10 DIHARD III DEV set
(sentence-level voting)

Speech Enhancement



PELPS$_1$[1] :

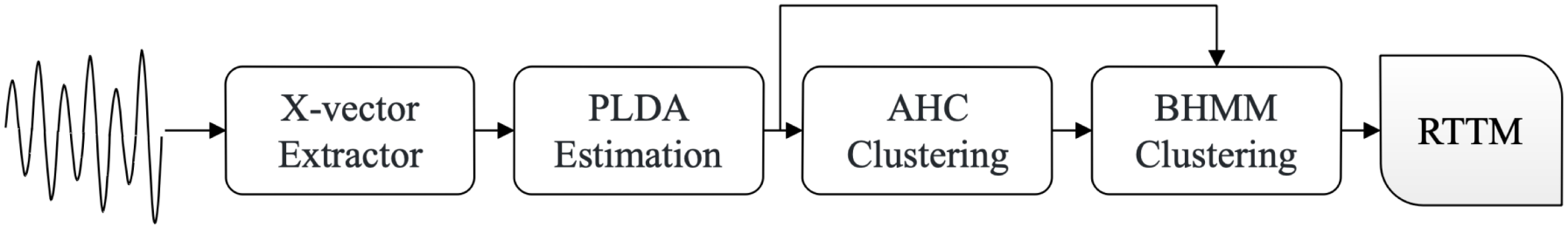
- Progressively Enhanced LPS at target layer 1
- 10dB increasing between 2 adjacent targets

PELPS$_1$ enhanced speech applied on :

- RESTAURANT domain
- TRACK2 SAD

[1] L. Sun, J. Du, X. Zhang, T. Gao, X. Fang, and C.-H. Lee, “Progressive multi-target network based speech enhancement with snr-preselection for robust speaker diarization,” ICASSP, 2020.

Clustering Based Diarization System



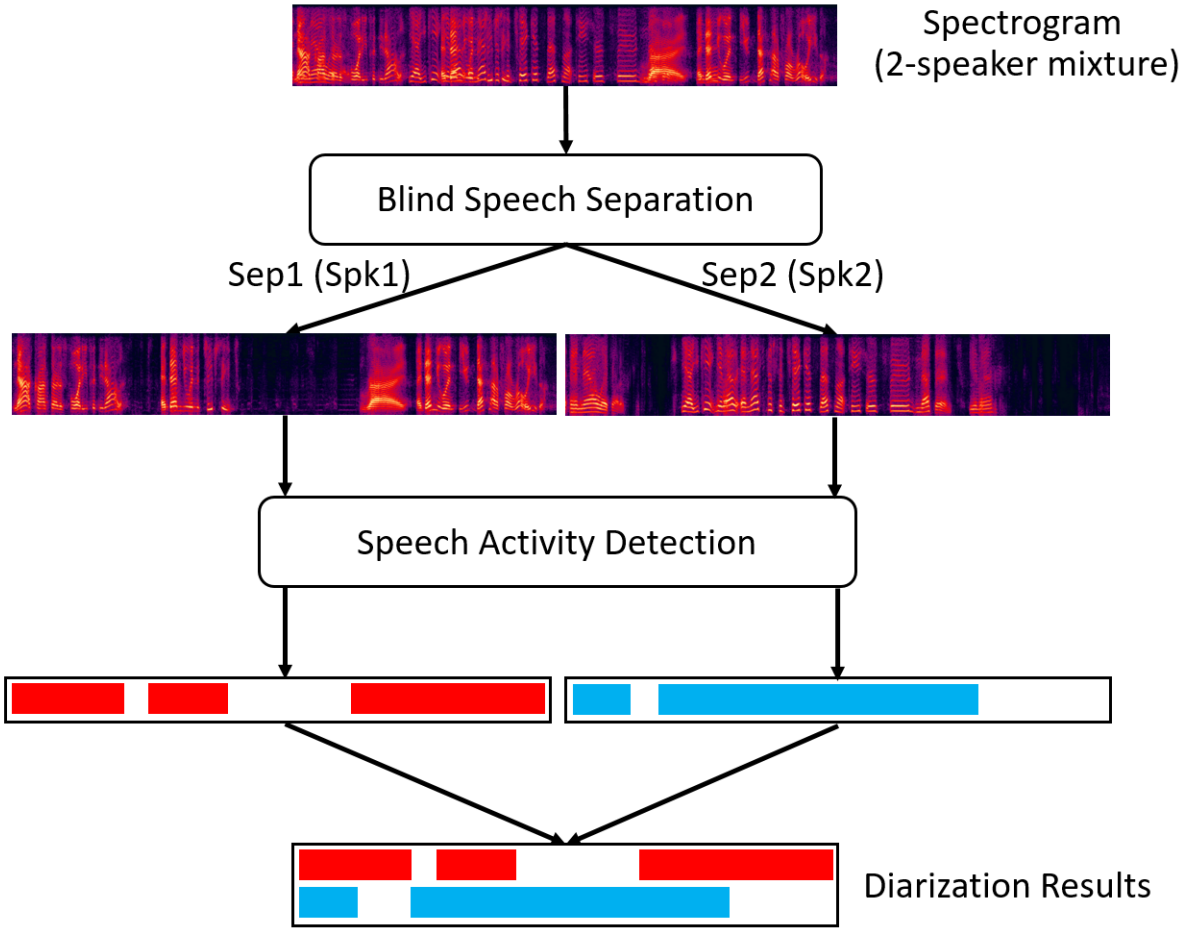
DER (%) on Track1 Development Set							
Full				Core			
Miss	FA	SpkErr	DER	Miss	FA	SpkErr	DER
10.92	0	4.98	15.9	10.94	0	5.18	16.12

- Clustering based diarization system[1] can't well handle overlapping speech

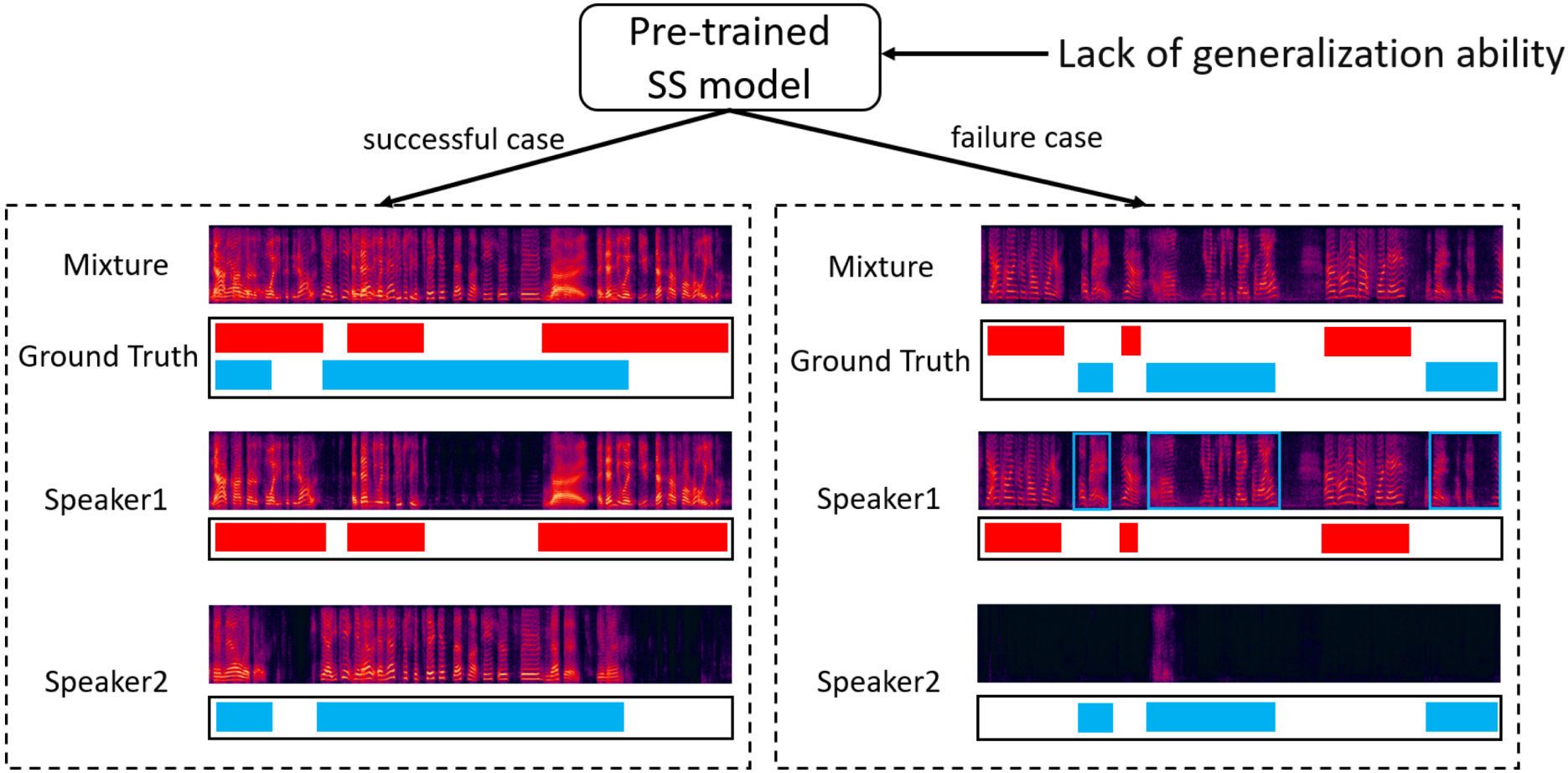
[1] M. Diez, L. Burget, F. Landini, et al. "Optimizing Bayesian HMM based x-vector clustering for the second DIHARD speech diarization challenge," ICASSP, 2020.

Speech Separation Based Diarization

- Solving diarization via speech separation
  - Two parts: separation and detection
  - Well handling overlapped regions in detection part

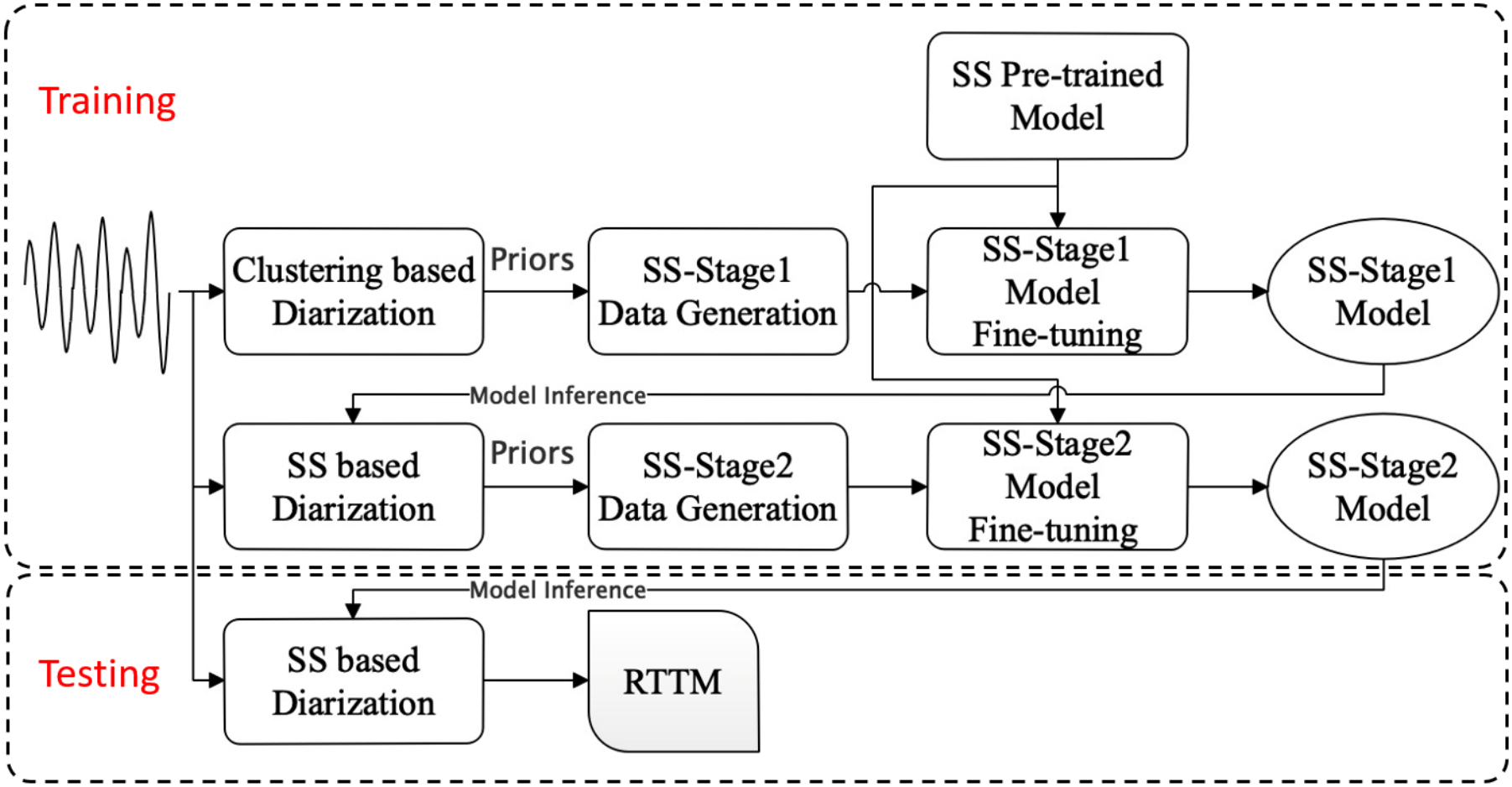


Problem of Blind Speech Separation



Iterative Speech Separation Based Diarization

- Improving generalization ability by multi-stage process
- Improving performance via more accurate priors in iterative process



Iterative Speech Separation Based Diarization

Experimental setup

Pre-trained model:

- Use the Librispeech dataset to simulate 250 hours training data;
- Train a fully convolutional time-domain audio separation network (Conv-TasNet)[1,2] model;

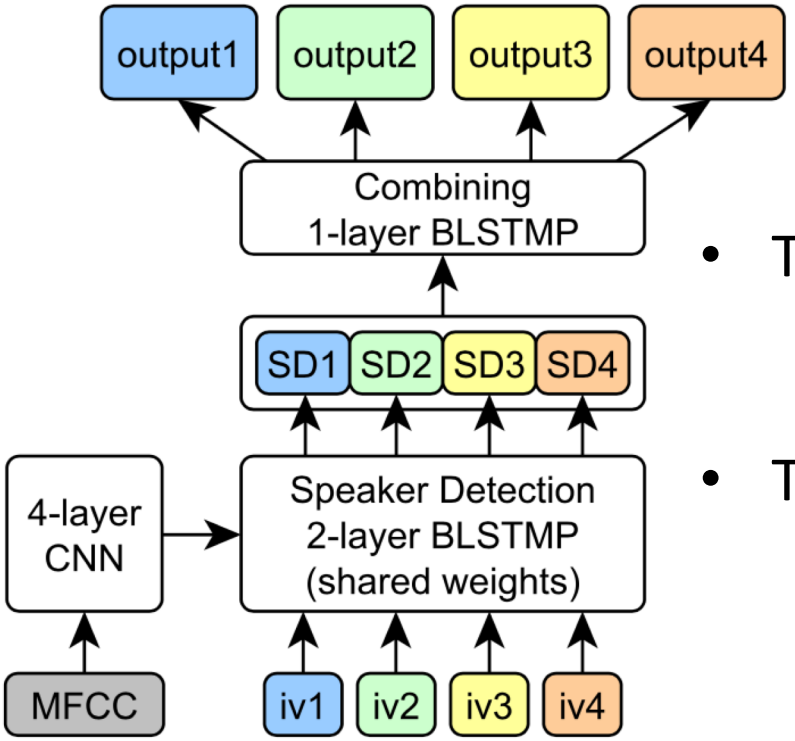
Fine-tuned model:

- Simulate 5000 mixed audios (about 2-3 hours) for each session in CTS;

System \ DER (%)	CTS	FULL	CORE
Clustering based diarization	16.22	15.78	15.94
ISS based diarization	8.31	13.11	15.11

[1]Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation.” IEEE/ACM transactions on audio, speech, and language processing, 2019.
[2]https://github.com/asteroid-team/asteroid

Target-Speaker Voice Activity Detection

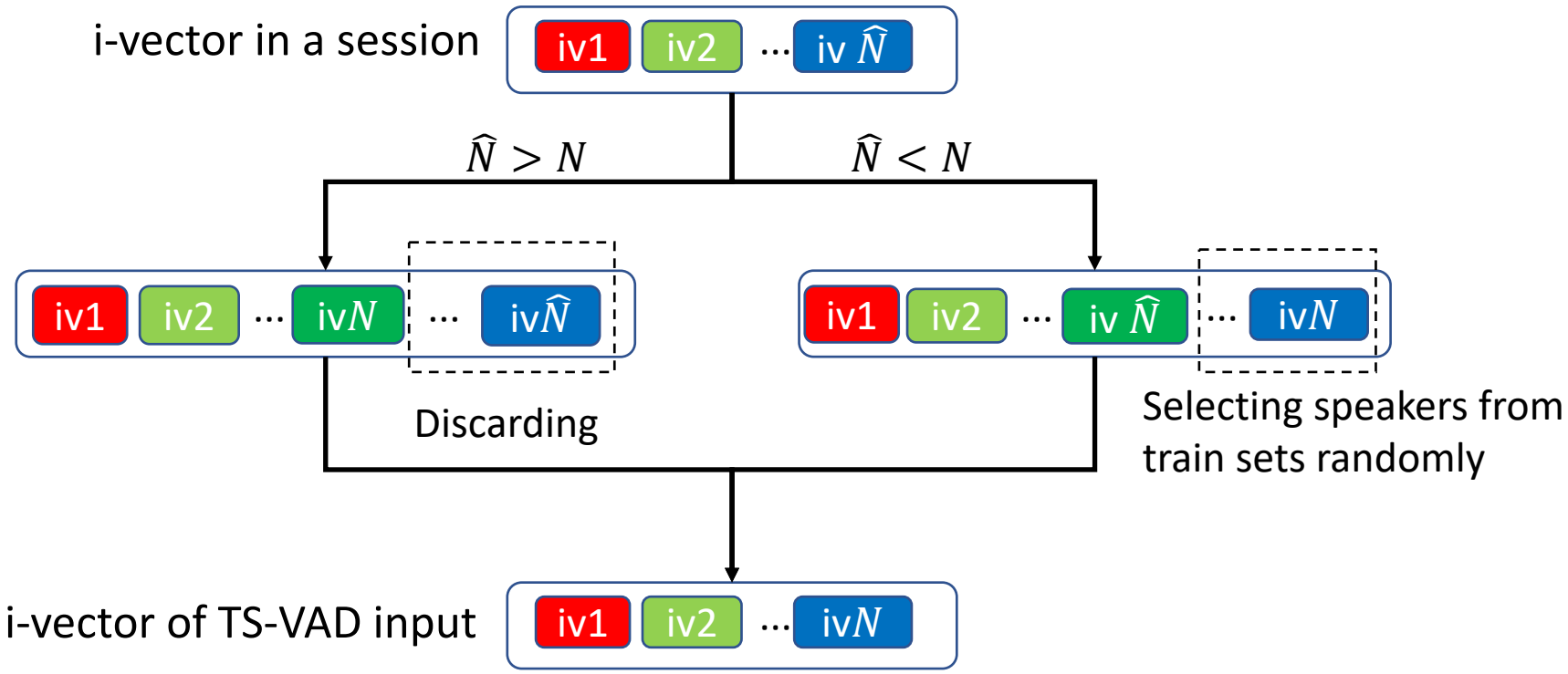


- TS-VAD[1]
  - Handling overlapping speech
  - Obtaining great performance on CHiME-6
- TS-VAD problems
  - Only handling session of fixed speaker number
  - Poor generalization ability to diverse environments

[1]Ivan Medennikov, et al. “Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario”, Interspeech, 2020.

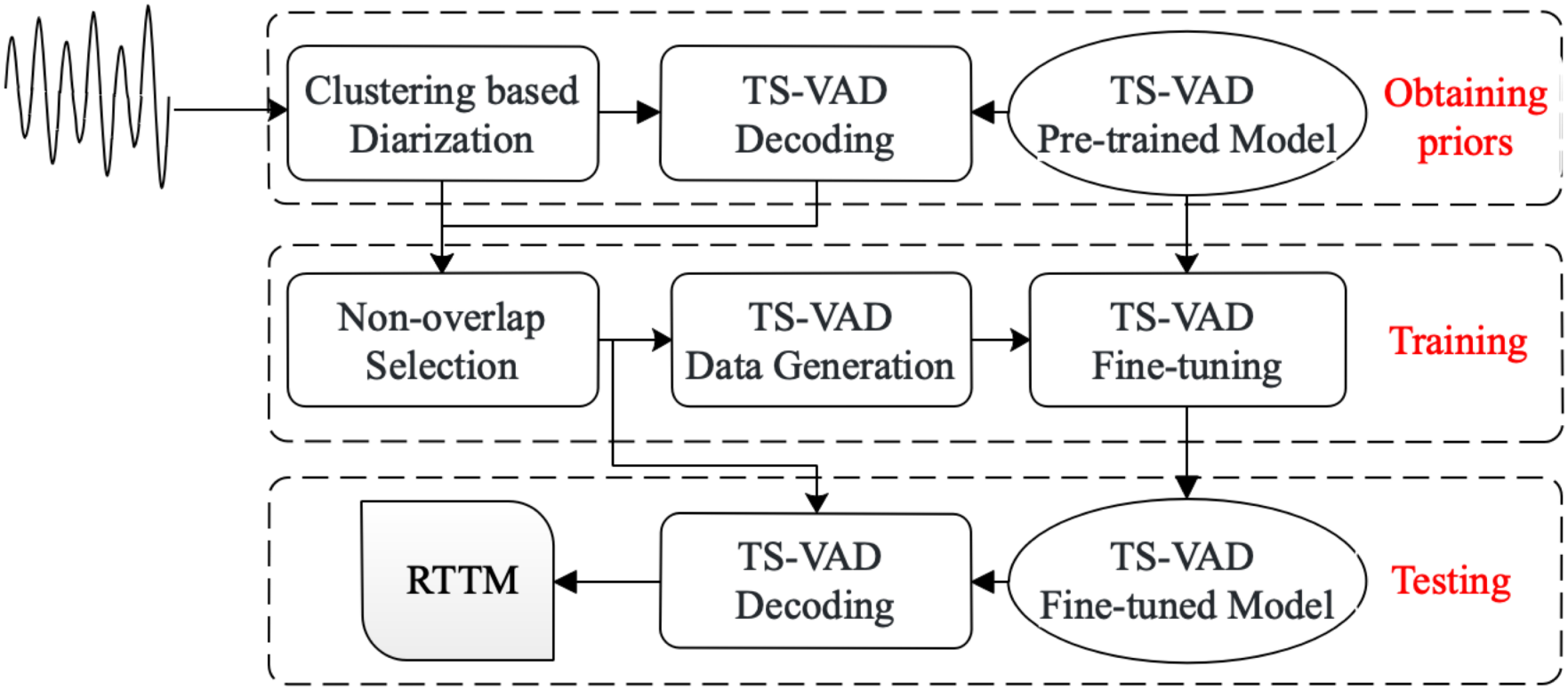
TS-VAD for Variable Number of Speakers

- Keeping the original TS-VAD structure and taking output speaker $N = 8$
- When session speaker number $\widehat{N}! = N$ in training and testing



Iterative TS-VAD for Variable Number of Speakers

- Iterative TS-VAD is proposed to solve mismatch between training and testing set
- Fine-tuning TS-VAD pre-trained model for each session



Experiments on Track1

Training data

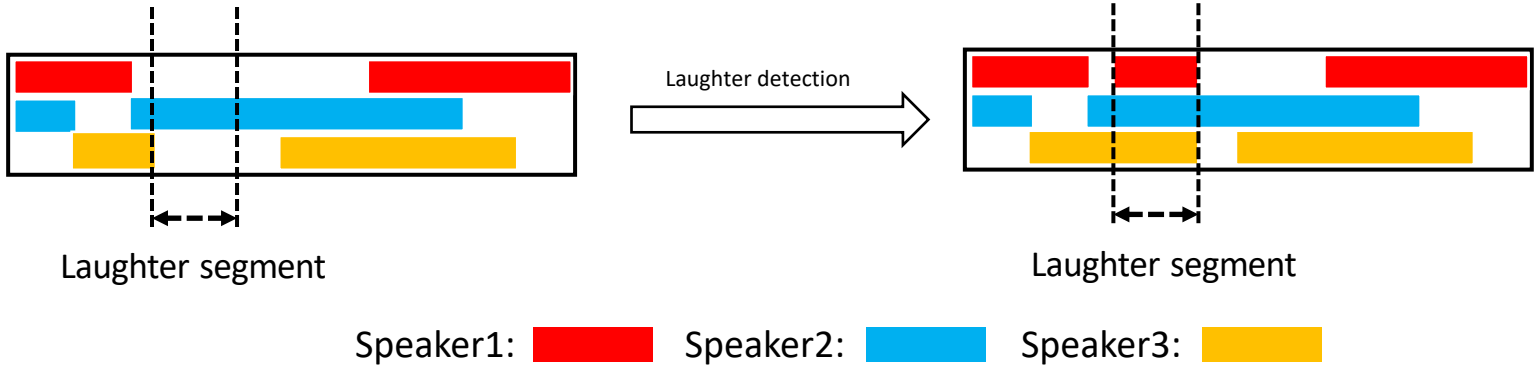
- i-vector extractor
  - Voxceleb 1 and 2
- TS-VAD pre-trained model (Total: 2500 hours)
  - Switchboard-2, AMI Meeting Corpus, Voxconverse DEV
  - Simulated multiple speaker dialogues with LibriSpeech
- Iterative TS-VAD finetuned model (4 hours for each session)
  - Simulated multiple speaker dialogues with non-overlap speaker segments

DOMAIN	MAPTASK	BROADC.	COURT.	SOC. LAB	CTS	CLINICAL	SOC. FIELD	MEETING	WEBVIDEO	RESTAURANT
Clustering based diarization	5.02	2.60	2.95	7.97	16.22	10.97	11.87	26.41	35.02	38.14
TS-VAD	6.71	2.94	3.15	8.81	10.21	16.48	13.79	24.72	36.73	47.71
Iterative TS-VAD	2.27	2.37	2.46	5.17	7.76	9.83	10.74	23.05	35.55	39.77

- TS-VAD
  - Performing better on well matched domains
- Iterative TS-VAD (ITS-VAD)
  - Greatly improving generalization abilities on most domains
  - Still cannot handle complex environments

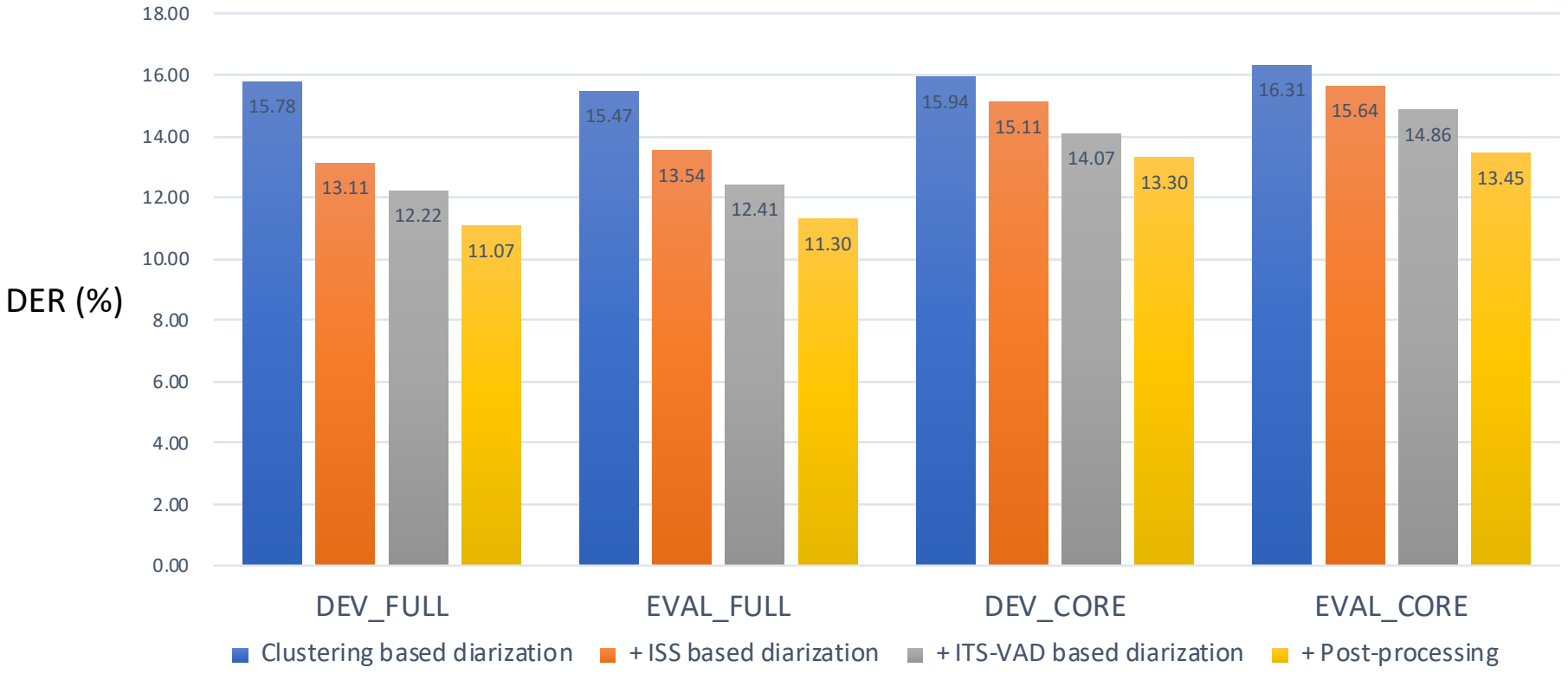
Post-processing

- Diarization Systems
  - Clustering based diarization
  - ISS based diarization
  - Iterative TS-VAD based diarization with different priors
- System Fusion
  - Dover-lap [1] of above systems
- Domain Selection
  - Selecting the best system for each domain according to DEV sets.
- ASR-related attributes
  - laughter detection



[1] D. Raj, L. P. Garcia-Perera, Z. Huang, et al. “DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs.” arXiv preprint arXiv:2011.01997, 2020.

Track1 Results



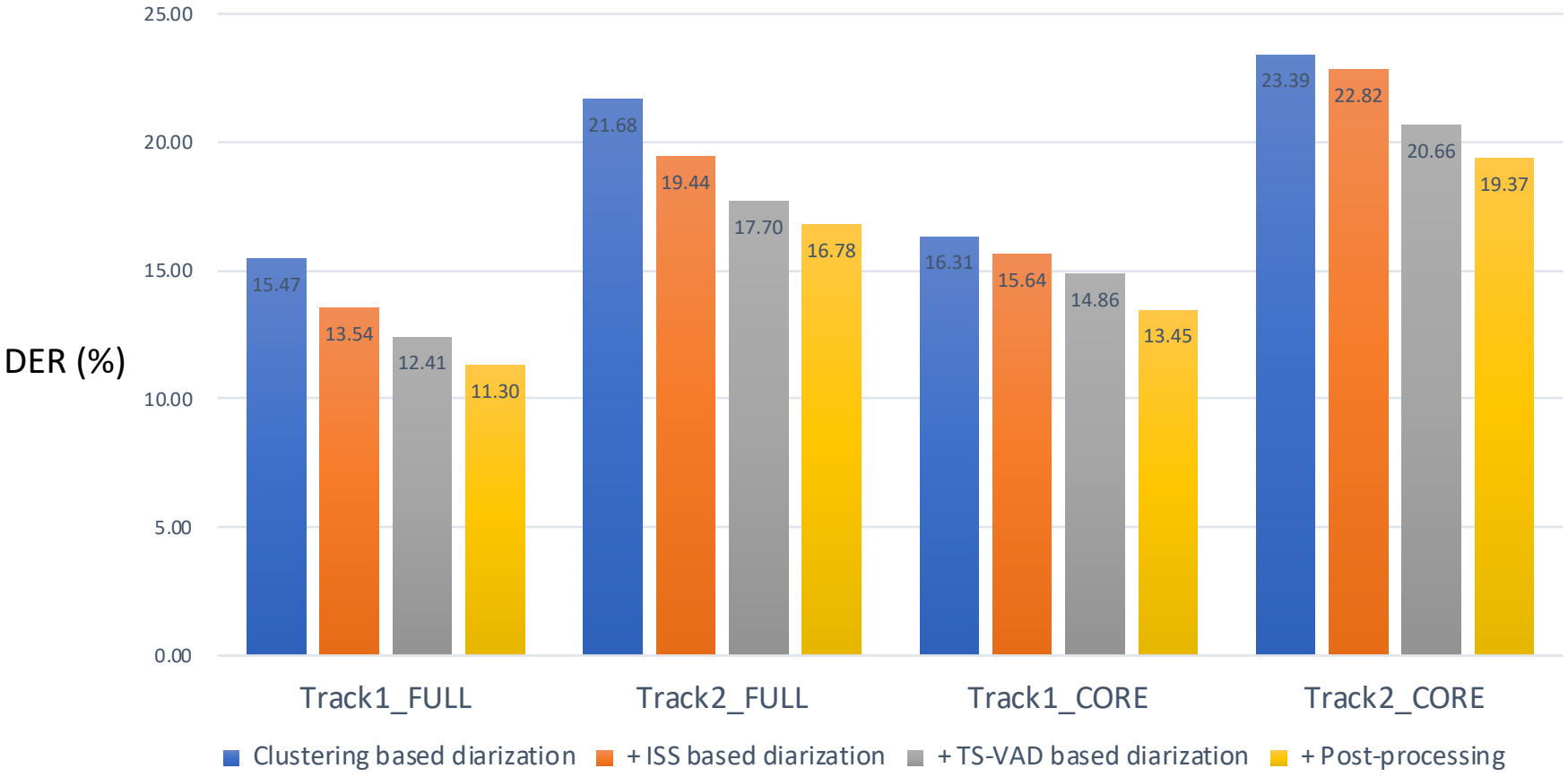
- We ranked 1st on both FULL and CORE sets of Track1.

Track2 SAD

- Network structures
  - DNN (195-256-128-2)
  - CNN-LSTM-DNN (2 CNN layers, 2 LSTM layers, 2 DNN layers)
  - TDNN[1,2]
- Enhanced speech for fine-tuning and testing
- Fusion: voting from the three systems

[1] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” ISCA, 2015.
[2] P. Ghahremani, V. Manohar, D. Povey, S. Khudanpur, “Acoustic Modelling from the Signal Domain Using CNNs,” Interspeech, 2016.

Track2 Results



- We ranked 1st on both FULL and CORE sets of Track2.

Acknowledgement

- JSALT 2017
  Team: Enhancement and Analysis of Conversational Speech
- JSALT 2019
  Team: Speaker Detection in Adverse Scenarios with a Single Microphone
- JSALT 2020
  Team: Speech Recognition and Diarization for unsegmented Multi-talker recordings with Speaker Overlaps
- DIHARD I, II, III
  All organizers and contributors
- ALL Colleges In Speech family！

JSALT 2017

JSALT 2019

The First DIHARD Speech Diarization Challenge

The Second DIHARD Speech Diarization Challenge

The Third DIHARD Speech Diarization Challenge

.....

Take-home Messages

- Iterative multi-stage processing is important
  - Speaker information can be updated stage-by-stage
- Speech separation is a promising direction:
  - Currently useful for simple telephone data
  - The generalization ability needs to be improved
- Domain dependent methods can achieve better results
  - Auxiliary techniques should be used flexibly (e.g. Speech enhancement)

Thanks

Q&A