Speech Processing in Realistic Environments ---From Speaker Diarization to Speech Recognition

Jun Du

University of Science and Technology of China

The 2nd International Symposium on Language Resources and Intelligence

2018.12.17 @Beijing

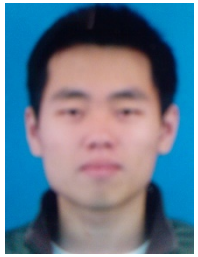
Key Contributors

- Collaborator

Prof. Chin-Hui Lee
Georgia Tech

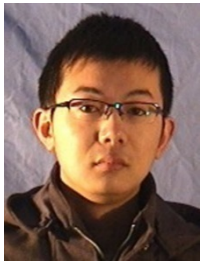
- My Team

Yong Xu



Yannan Wang

Qing Wang

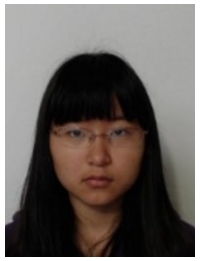


Tian Gao

Yanhui Tu

Lei Sun

Shixue Wen



Nana Fan

Li Chai

Xin Wang

Why Is Speech Processing Important?

Speech-to-Speech Translation

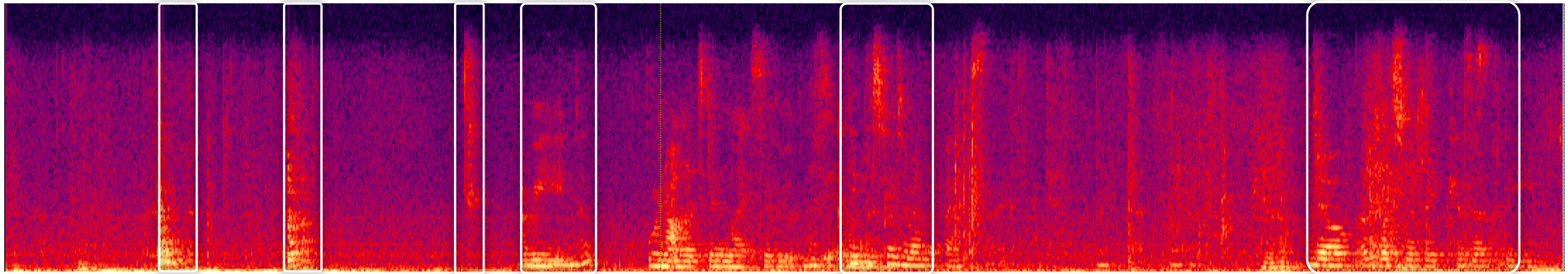
Human-Machine Dialogue System

Early Language Acquisition
(Child Speech)

Clinic Speech Analysis

Why Is Speech Processing Challenging?

AMI recording (speech-like noises, low-resolution, ambiguous, blurring speech segments)



Background sounds and noises
(Speech denoising)

Overlapping speech
(Speech separation)

Reverberation
(Speech dereverberation)

A Mathematical Perspective

- Background Noise → Speech Enhancement (SE)

  $x \approx s + n \longrightarrow s = f_1(x)$

- Reverberation → Speech Dereverberation (SD)

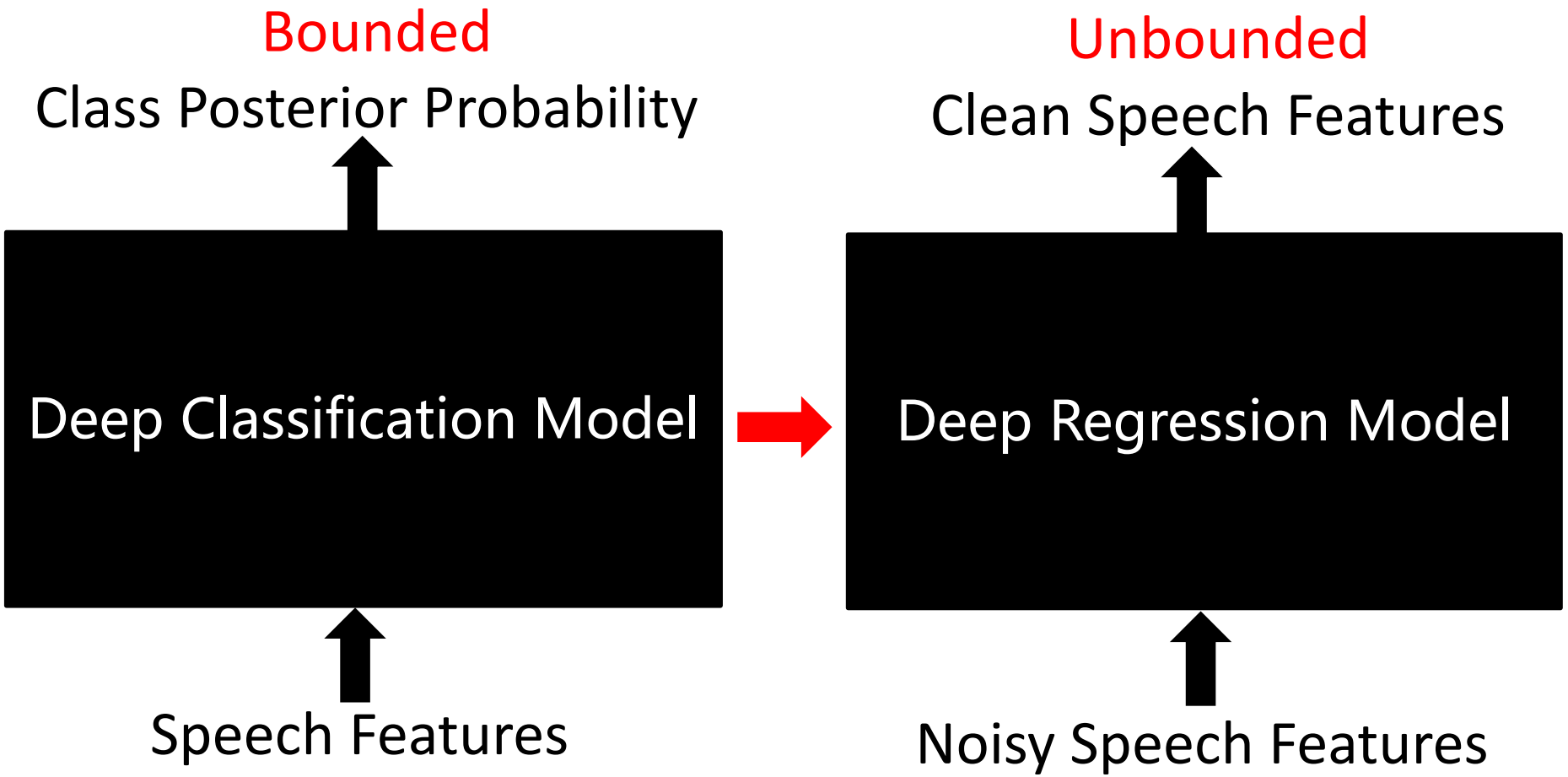
  $x \approx h * s \longrightarrow s = f_2(x)$

- Multiple Speakers → Speech Separation (SS)

  $x \approx s_1 + s_2 \longrightarrow (s_1, s_2) = f_3(x)$

- Real Scenarios: $x \approx h * (s_1 + s_2) + n$

Speech Processing As A Regression Problem



Yong Xu, Jun Du*, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 23, No. 1, pp.7-19, 2015. (2018 IEEE Signal Processing Society Best Paper Award)

Conventional Approaches vs. DL Approaches

- Conventional approaches
  - Strong model assumption

- Deep learning approaches
  - No model assumption but strong data assumption

Knowledge-Rich Deep Learning

- Neural Network Architecture Design
  - Open the black box: progressive learning with domain knowledge
- Objective Function Design
  - MMSE -> probabilistic framework
- Model Generalization
  - Transfer/Multitask/Ensemble Learning
- Training Data Augmentation and Clustering
  - Improve the diversity/homogeneity and reduce the redundancy
- Joint Optimization
  - SE/ASR, SE/SAD, SE/SS, SS/ASR, SD/ASR
- Multi-channel Problem
  - Incorporate spatial signal processing with deep learning

Speaker Diarization

From JSALT 2017 to DIHARD Challenge

JSALT Workshop 2017

- Why diarization? Is it challenging?
- Monologue to dialogue with a perfect ASR

Did you get any royalties since you'd already sold the song? Not really. {laughs} Not really. Would you- would you sing us a bit of the "Family Bible" and tell us what went into the writing of it? Well, this is sort of autobiographical, or practically 100 percent autobiographical.

TerryGross: Did you get any royalties since you'd already sold the song?

WillieNelson: Not really. {laughs} Not really.

TerryGross: Would you- would you sing us a bit of the "Family Bible" and tell us what went into the writing of it?

WillieNelson: Well, this is sort of autobiographical, or practically 100 percent autobiographical.

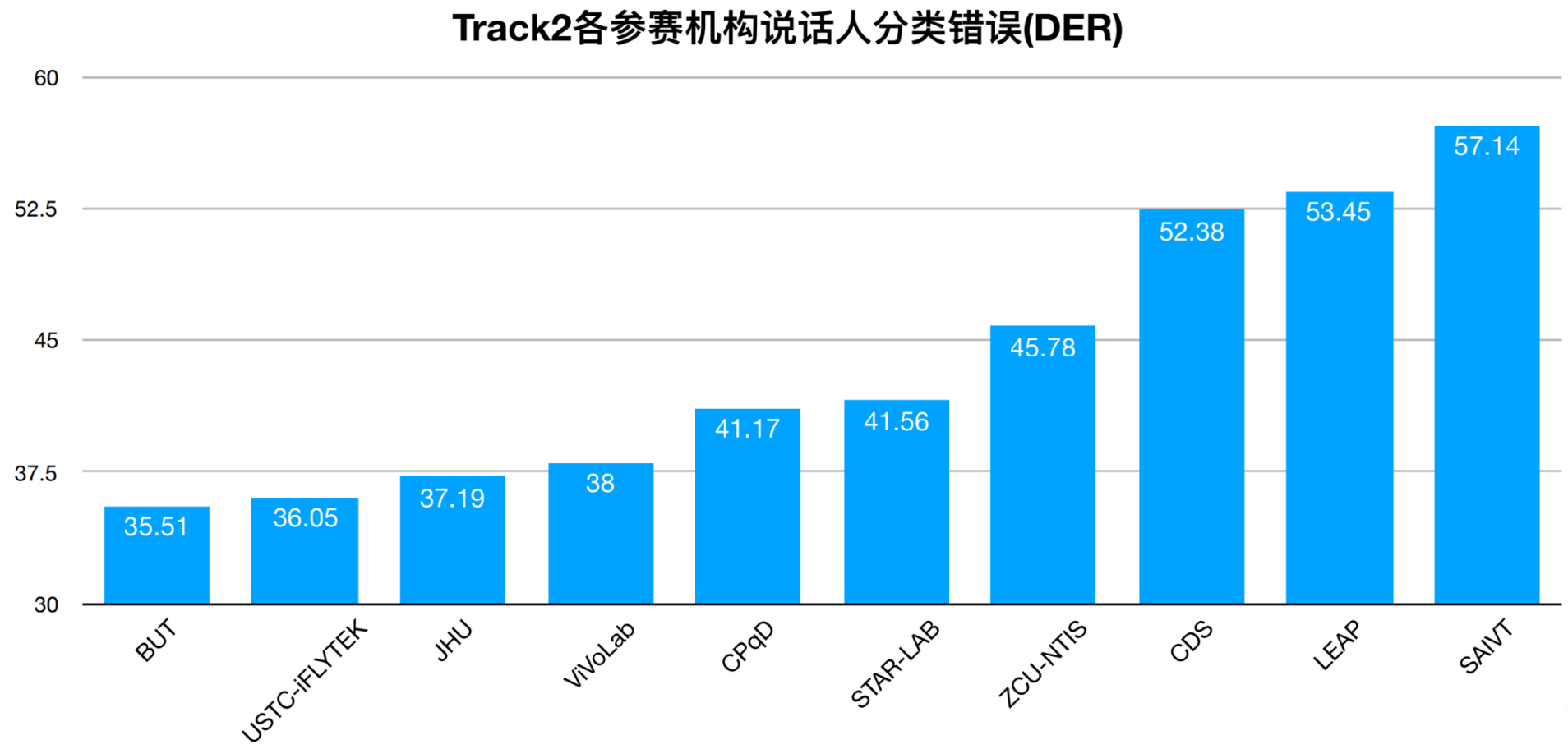
DIHARD Challenge

- Background

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

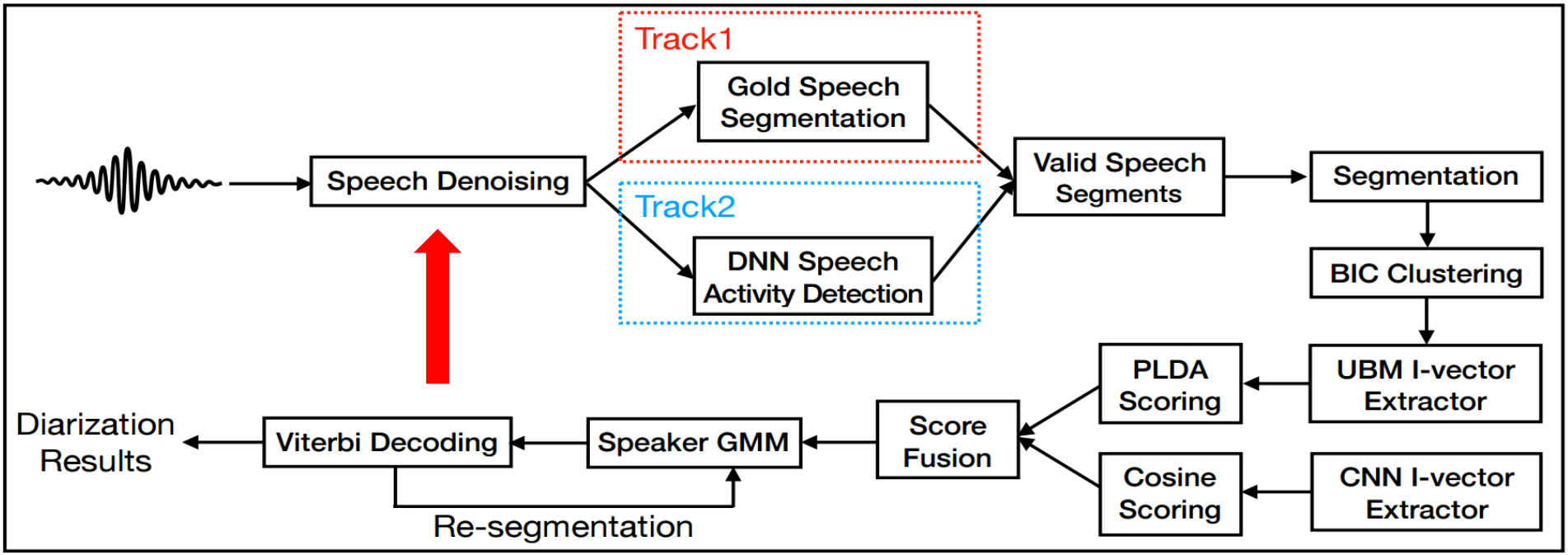
- INTERSPEECH 2018 Special Session
  - The First DIHARD Speech Diarization Challenge
  - Challenge website: https://coml.lscp.ens.fr/dihard/index.html
- Scenarios
  - Clinical interviews
  - Extended child language acquisition recordings
  - YouTube videos
  - Speech in the wild (e.g., recordings in restaurants)
- Two tracks
  - Track 1: diarization beginning from gold speech segmentation
  - Track 2: diarization from scratch

DIHARD Challenge Leaderboard



A long way to go…

Our Diarization System

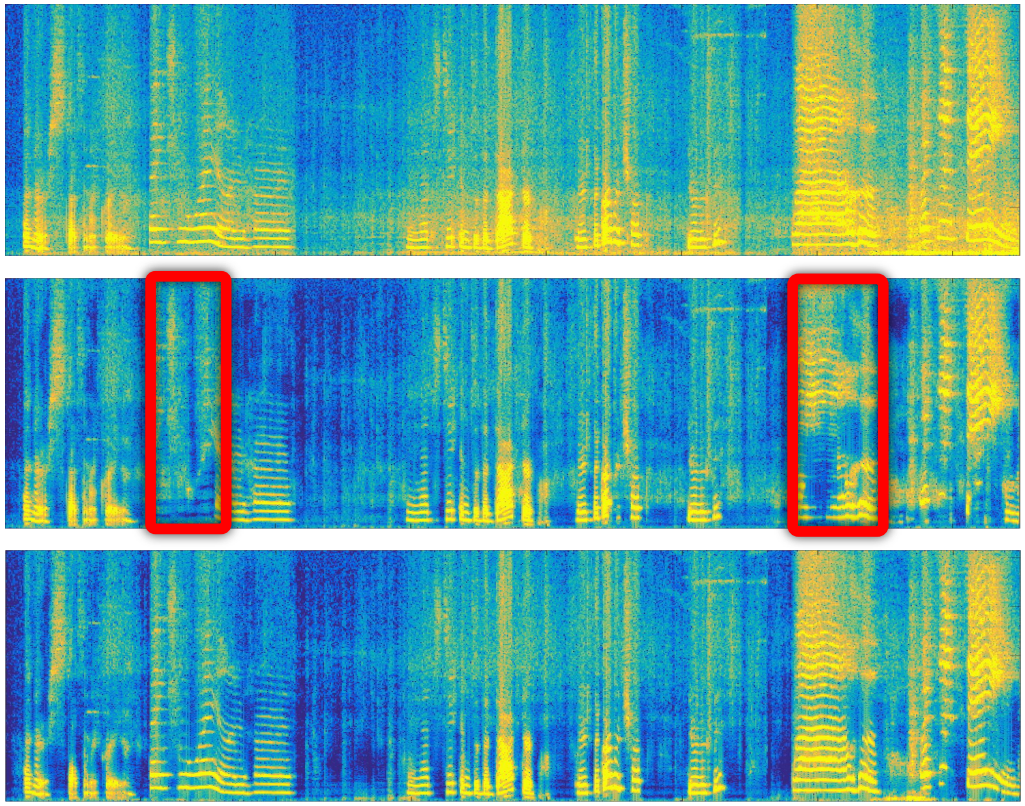


L. Sun, J. Du, etc., "A Novel LSTM-based Speech Preprocessor For Speaker Diarization In Realistic Mismatch Conditions," ICASSP 2018.

L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin and C.-H. Lee, "Speaker Diarization with Enhancing Speech for the First DIHARD Challenge," INTERSPEECH 2018.

Improved Speech Enhancement

- Adding a 50-hour Chinese speech corpus
- Increasing the data amount from 36h to 400h
- Well preserving the child speech



Noisy speech

The ICASSP 2018 approach

The INTERSPEECH 2018 approach

Speech Enhancement for Diarization

Track 1: Using oracle SAD

DER(%)	Seedlings	SCOTUS	DCIEM	ADOS	YP	SLX	RT04S	LIBRIBOX	VAST	TOTAL
Original	40.58	7.65	7.27	22.9	3.32	18.88	36.55	0.6	36.46	20.26
Denoised	37.86	7.10	7.58	21.73	3.12	16.94	35.93	1.9	36.08	19.68

Track 2: Diarization from scratch

DER(%)	Seedlings	SCOTUS	DCIEM	ADOS	YP	SLX	RT04S	LIBRIBOX	VAST	TOTAL
Original	65.72	15.75	17.49	39.46	12.97	31.45	49.42	8.77	53.16	33.2
Denoised	61.89	15.25	20.02	34.55	12.06	27.81	45.00	9.07	44.64	30.4

Speech Recognition

Environmentally Robust ASR in The Past Decade

2000	2001	2003	2006	2011	2013	2015	2016	2018
Aurora 2/3/4			AMI	CHiME-1	CHiME-2	CHiME-3	CHiME-4	CHiME-5
			SSC		REVERB			DIHARD-1

Single Microphone → Microphone Array → Microphone Arrays

Single environment factor → Multiple environment factors

Single speaker → Multiple speakers

Small vocabulary → Medium/Large vocabulary

Reading style → Conversional style

Small-scale training data → Large-scale training data

CHiME-3 & CHiME-4 (1)

Sitting in a cafe (**CAF**)

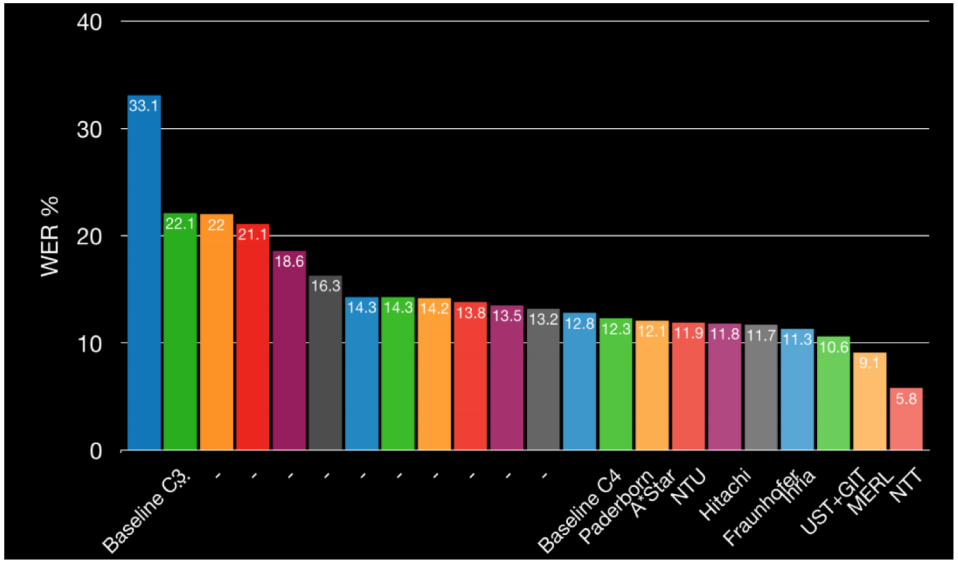
Standing at a street junction (**STR**)

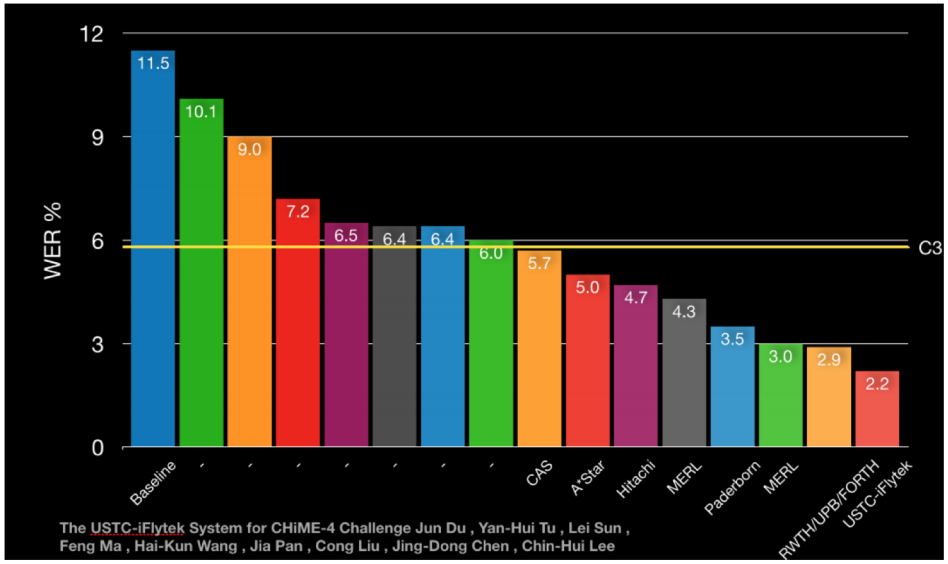
Travelling on a bus (**BUS**)

In a pedestrian area (**PED**)

Near Field, Single Speaker, Reading Style

CHiME-3 & CHiME-4 (2)



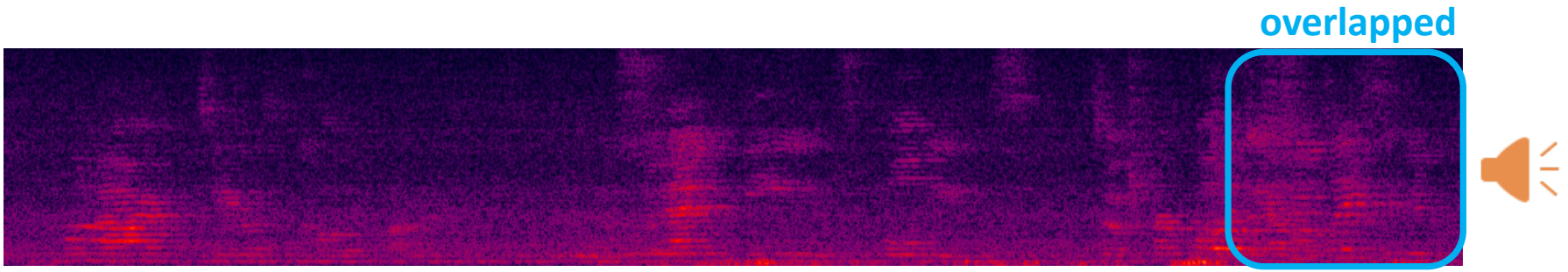


Deep Learning + Conventional Beamforming

Y.-H. Tu, J. Du, etc., "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Communication*, pp. 31-43, 2019.

How Challenging is CHiME-5? (1)

- Dinner party scenario
- Conversational speech, multi-talkers, far-field
- The baseline WER of DEV set is more than 80%
  - Using oracle diarization information



"Well, now I have potatoes and cauliflower and beans, pole beans, so purple **and yellow**."

How Challenging is CHiME-5? (2)

- MVDR + CGMM/Music/estnoiseg mask ✗
- DeepBeam [Qian, 2018] ✗
- GEV + BLSTM mask [Heymann, 2016] ✗
- Denoising with CGMM mask ✗
- Denoising Wavenet [Rethage, 2017] ✗
- Deep Clustering ✗
- Permutation invariant training (PIT) ✗
- WPE ✓

http://spandh.dcs.shef.ac.uk/chime_workshop/presentations/CHiME_2018_Medennikov_oral.pdf

USTC-iFlytek Team for CHiME-5

Jun Du (USTC)

Tian Gao (USTC)

Lei Sun (USTC)

Chin-Hui Lee (GIT)

Feng Ma (iFlytek)

Yi Fang (iFlytek)

Di-Yuan Liu (iFlytek)

Qiang Zhang (iFlytek)

Xiang Zhang (iFlytek)

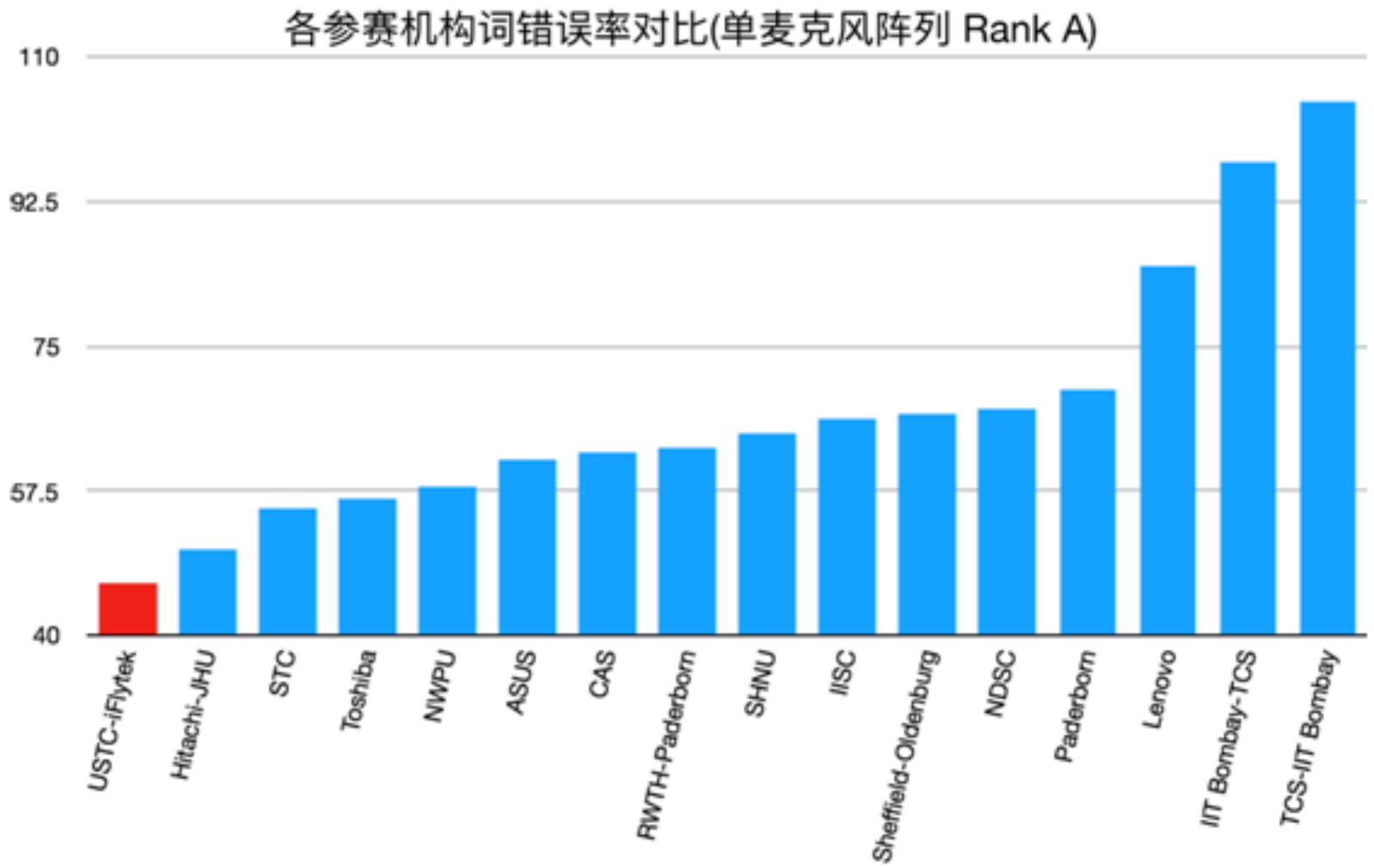
Hai-Kun Wang (iFlytek)

Jia Pan (iFlytek)

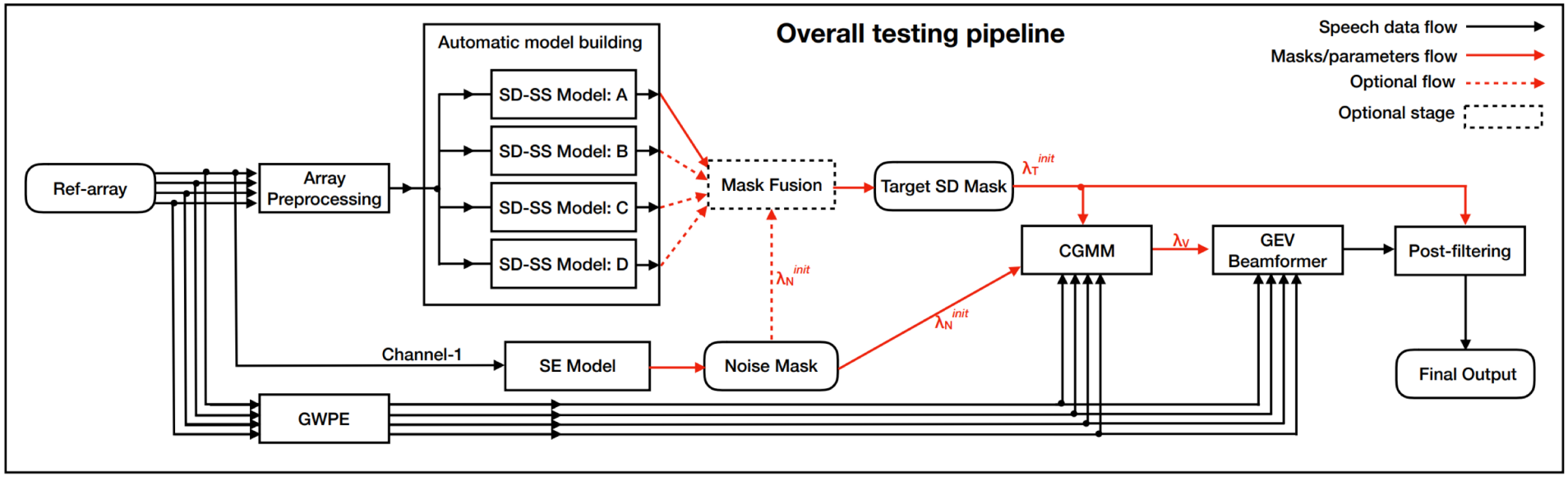
Jian-Qing Gao (iFlytek)

Jing-Dong Chen (NWPU)

CHiME-5 Challenge Leaderboard



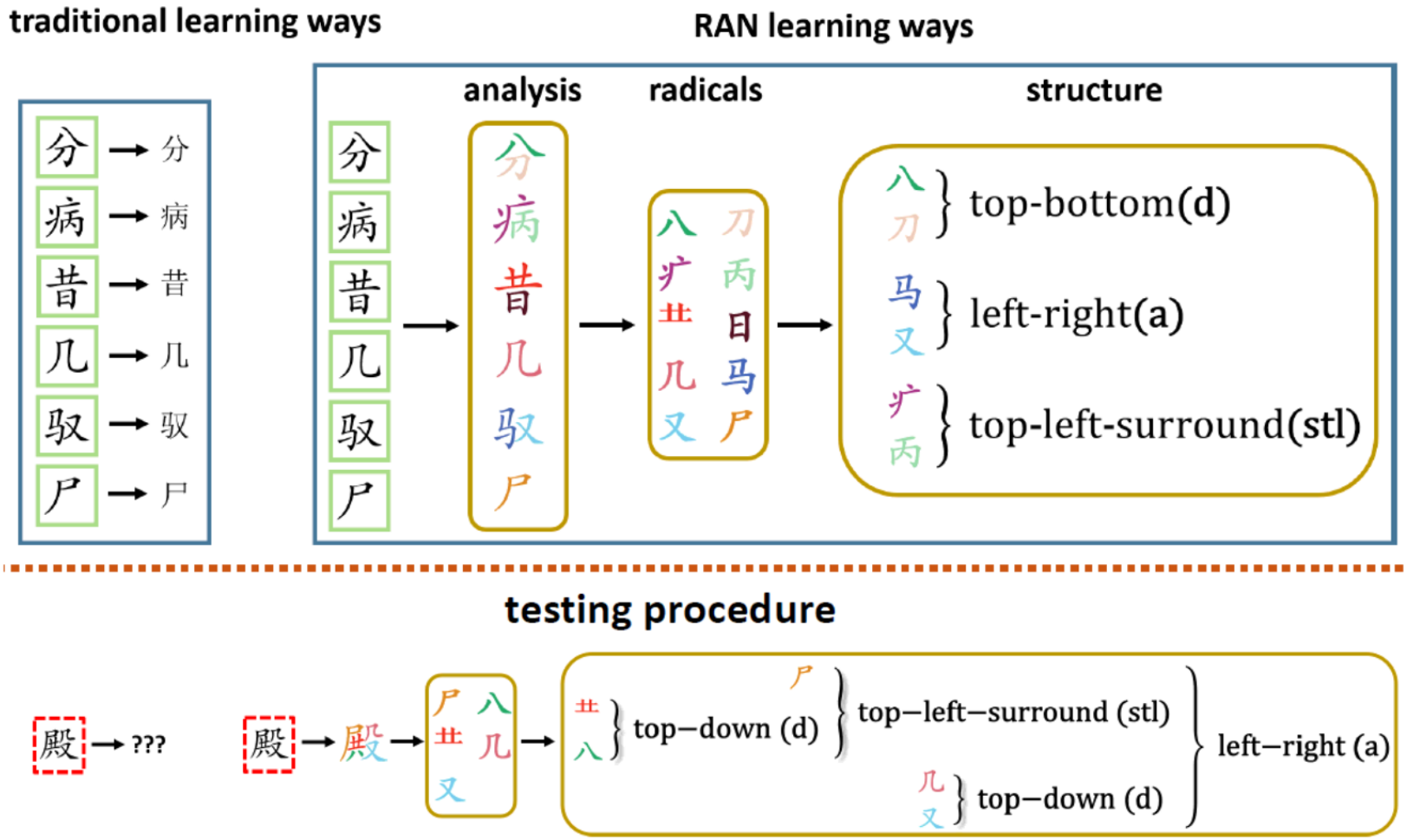
Our Front-End Solution for CHiME-5



Advanced acoustic/language models did not work without a well-designed front-end!

L. Sun, J. Du, etc., "A speaker-dependent single-channel/multichannel approach for front-end of CHiME-5 Challenge under far-field multi-talker scenario," *Submitted to Journal of Selected Topics in Signal Processing.*

Radical Analysis Network for Chinese OCR



Thanks!