ICASSP2021 Tutorial: T-9

Distant conversational speech recognition and analysis: Recent advances, and trends towards end-to-end optimization

Keisuke Kinoshita (NTT Corporation)
Yusuke Fujita (LINE Corporation)
Naoyuki Kanda (Microsoft)
Shinji Watanabe (Carnegie Mellon University)

Tutorial presenters

Tutorial Presenters (1/4)

Keisuke Kinoshita

Distinguished Researcher, NTT Corporation

received the M. Eng. degree and Ph.D degree from Sophia University in Tokyo in 2003 and 2010, respectively. After joining NTT in 2003, he has been working on various topics related to distant conversational speech recognition such as 1ch/multi-channel speech enhancement, speaker diarization and robust speech recognition. He published more than 100 technical papers in refereed journals and conference proceedings, and also contributed to 5 book chapters. He was a Chief coordinator of the REVERB challenge 2014, and a member of IEEE AASP TC since 2018. He is honored to receive the 2006 IEICE Paper Award, the 2009 ASJ Outstanding Technical Development Prize, the 2011 ASJ Awaya Prize, the 2012 Japan Audio Society Award, the 2015 IEEE-ASRU Best Paper Award Honorable Mention.

Tutorial Presenters (2/4)

Yusuke Fujita

Senior Research Engineer, LINE Corporation

received his B.S. and M.S. degrees in computer science from Waseda University, Tokyo, Japan, in 2003 and 2005, respectively. He was a Senior Researcher at Hitachi Ltd. in Tokyo, Japan from 2005 to 2021, and was a Visiting Scholar at Johns Hopkins University, MD, USA (2018-2020). His research interests include speech recognition, speech separation, and speaker diarization. He has been working on end-to-end speaker diarization and distant speech recognition through the CHiME-5 challenge session chair, the CHiME-6 Scientific Committee, and the participation of CHiME-3,4 and CHiME-5 challenges.

Tutorial Presenters (3/4)

Naoyuki Kanda

Principal Researcher, Microsoft Corp.

received a B.S. in Engineering, M.S. in Informatics, and Ph.D. in Informatics from Kyoto University, Japan, in 2004, 2006, and 2014, respectively. He was a Senior Researcher at Hitachi Ltd. in Tokyo, Japan from 2006 to 2019, and held appointments as a Research Expert (2014-2016) and Cooperative Visiting Researcher (2016-2017) at the National Institute of Information and Communications Technology (NICT) in Kyoto, Japan. He has been actively working on the joint modeling of speech separation, speech recognition, and speaker identification / diarization. His algorithms and systems for the distant conversational speech recognition won the second prize at the CHiME-5 speech recognition competition in 2018.

Tutorial Presenters (4/4)

Shinji Watanabe

Associate Professor, Carnegie Mellon University

received his B.S., M.S., and Ph.D. degrees from Waseda University, Tokyo, Japan. He was a research scientist at NTT Communication Science Laboratories, Kyoto, Japan, from 2001 to 2011, and a Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA USA from 2012 to 2017. He has been actively working on the distant conversational speech recognition area through the CHiME speech separation and recognition challenge organization, a tutorial speaker about "Recent Advances in Distant Speech Recognition" in Interspeech'16, and a number of participations of the related project including the first place at the DIHARD challenge 2018.

Notations & Abbreviations

Abbreviations

AHC	Agglomerative Hierarchical Clustering
ASR	Automatic Speech Recognition
BLSTM	Bi-directional LSTM neural network
BCE	Binary Cross Entropy
BF	Beamformer
CNN	Convolutional Neural Network
CE	Cross Entropy
CTC	Connectionist Temporal Classification
DEV	Development set
DER	Diarization Error Rate
DNN	Deep Neural Network
E2E	End-to-End
EEND	End-to-End Neural Diarization
EVAL	Evaluation set
FIFO	First-In, First-Out
GMM	Gaussian Mixture Model
GRU	Gated Reccurent Unit
GSS	Guided Source Separation
HMM	Hidden Markov Model
LM	Language Model
LSTM	Long-Short Term Memory neural network
LLR	Log Likelihood Ratio
ML	Maximum Likelihood
MMI	Maximum Mututal Information
MMSE	Minimum Mean Squared Error
MSE	Mean Squared Error
MVDR	Minimum Variance Distortionless Response
NIN	Network In Network
NN	Neural Network
PIT	Permutation Invariant Training
PLDA	Probabilistic Linear Discriminant Analysis
PSD	Power Spectral Density
ReLU	Rectified Linear Unit
RIR	Room Impulse Response
RNN	Recurrent Neural Network
RNN-T	RNN-Tranceducer
RTF	Relative Transfer Function
RTTM	Rich Transcription Time Marked
SAD	Speaker Activity Detection
SC	Speaker Counting
SD	Speaker Diarization
SOT	Serialized Output Training
Spk	Speaker
STFT	Short-Time Fourier Transformation
SWER	Speaker-Attributed Word Error Rate
TF	Time-Frequency
TDNN	Time-Delay Neural Network
TS-VAD	Target-Speaker Voice Activity Detection
UIS-RNN	Unbounded Interleaved-State RNN
VAD	Voice Activity Detection
WER	Word Error Rate
WPE	Weighted Prediction Error
WSJ	Wall Street Journal

Notations (1/2)

Mathematical expressions and operations	
a	A scalar variable.
$\mathbf{a}$	A column vector.
$\mathbf{A}$	A matrix.
C	A constant.
$p(x)$	Probability density function
$\|\cdot\|_2$	Eucredean norm of a vector
$\mathbb{R}$ and $\mathbb{C}$	A set of real scalars, and a set of complex scalars.
$\mathbb{R}^M$ and $\mathbb{R}^{M\times M}$	A set of M dimentional real vectors, and a set of $M \times M$ dimentional real matrices. $\mathbb{C}^M$ and $\mathbb{C}^{M\times M}$ are defined similarly.

Notations (2/2)

Indices	
$\tilde{t}$, t, f, m	Indices of time samples, time frames, frequency bins, microphones.
k	Index for speakers.
i	Index for spoken words or sub-words.
$\tilde{T}$, T, F, M	Total numbers of time samples, time frames, frequency bins, microphones.
K	Total numbers of speakers.
I_k	Total numbers of spoken words or sub-words for the k-th speaker.
Symbols for Short Time Fourier Transformation (STFT) domain signals	
$s_{t,f,k} \in \mathbb{C}$	A clean speech signal for the k-th speaker.
$x_{m,t,f,k} \in \mathbb{C}$	A microphone image of the k-th speaker at the m-th microphone, i.e., noiseless reverberant signal for the k-th speaker at the m-th microphone.
$y_{m,t,f} \in \mathbb{C}$	An observed signal at the m-th microphone, i.e, noisy reverberant speech-mixture.
$n_{m,t,f} \in \mathbb{C}$	Diffuse noise.
Symbols for other domains	
$d_{t,k}$	Diarization label for the k-th speaker at the time frame t. A label for diarization systems.
$r_{n,k}$	The n-th spoken word or sub-word of the k-th speaker. A label for ASR systems.
$\mathbf{h}_t$	An embedding vector at the time frame t.
$\mathbf{h}_k^{(\mathrm{spk})}$	A speaker embedding vector for the k-th speaker.
$s_k[\tilde{t}]$	A time-domain clean speech signal for the k-th speaker. The other time-domain symbols are defined similarly.
o	Observed signal in a feature domain, e.g., Log-Mel feature domain.

Agenda

Part 1: Introduction
- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

25min

5min Q&A

Part 2: Current state-of-the-art systems
- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

25min

15min Q&A and Break

Part 3: A new research trend: Jointly optimal systems
- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

25min + 5min Q&A

25min + 5min Q&A

25min + 5min Q&A

Part 4: Summary and discussion
- 4.1. Strengths of each approach
- 4.2. Challenges and fundamental difficulties

5min + ?-min Q&A

Agenda

Part 1: Introduction

- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

Part 2: Current state-of-the-art systems

- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

Part 3: A new research trend: Jointly optimal systems

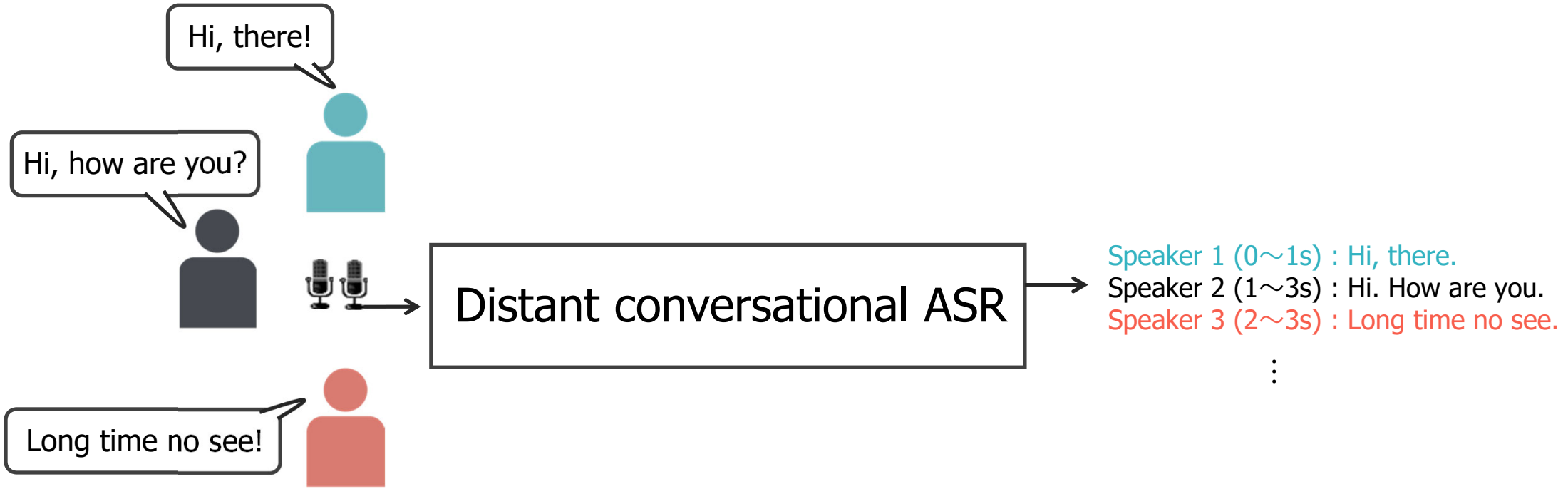
- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

Part 4: Summary and discussion

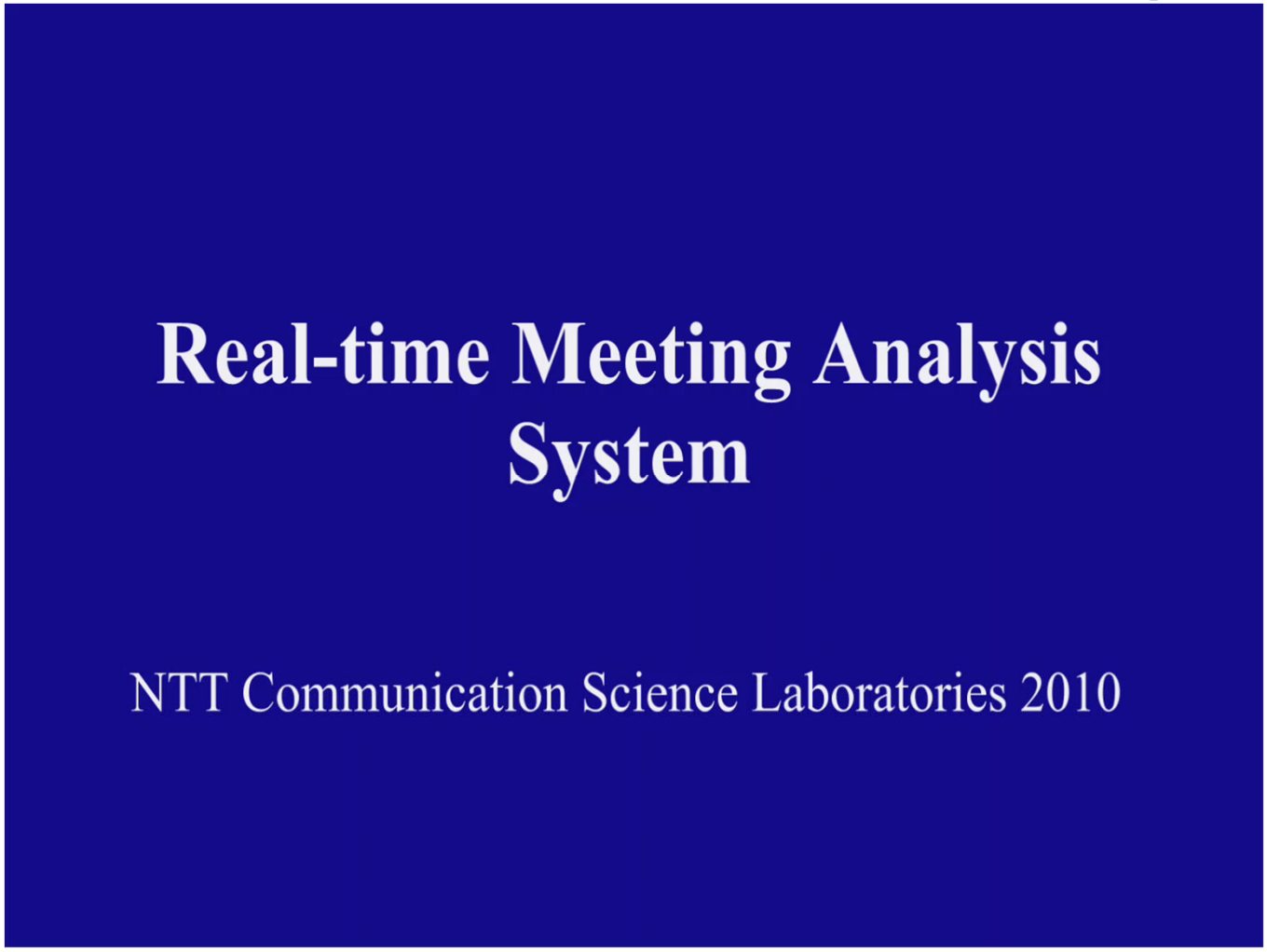
- 4.1. Strengths of each approach
- 4.2. Challenges and fundamental difficulties

What is distant conversational ASR and analysis?

- A task to obtain "who spoke what and when" information from natural conversation
  - For "who part", systems are not often required to output a name for the speakers, i.e., only a generic id which is unique within a meeting/conversation
- Sometimes, referred to as meeting recognition and meeting analysis



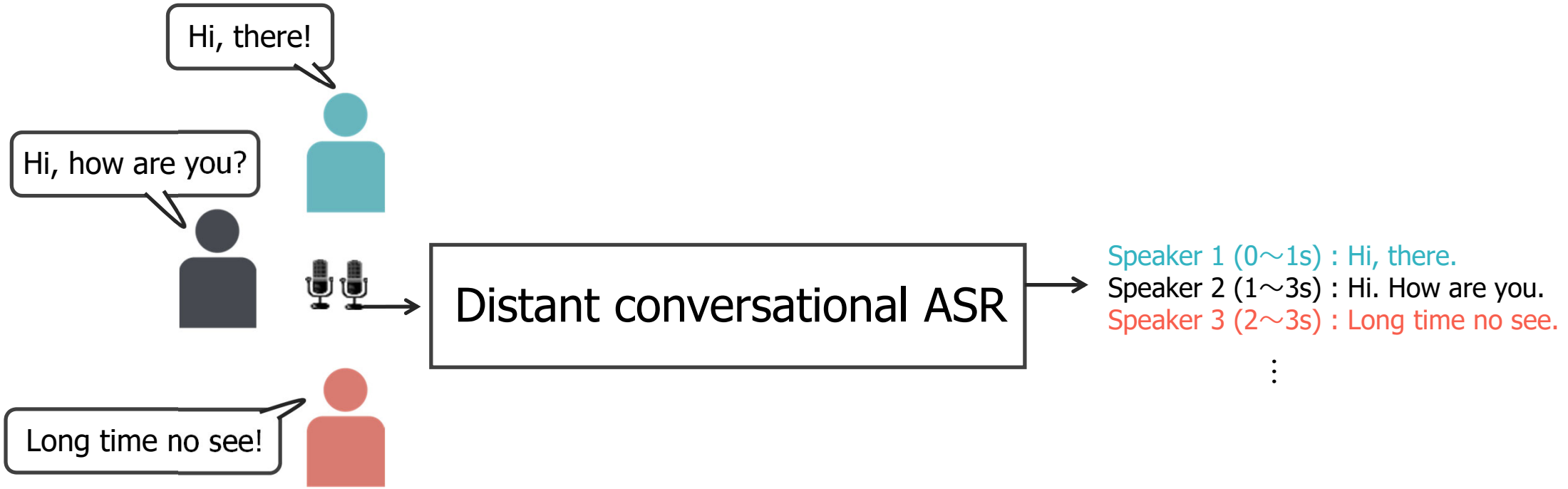
What is distant conversational ASR and anlysis?



T. Hori *et al.*, "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-Directional Camera," in IEEE TASLP, 2012

What is distant conversational ASR and analysis?

- A task to obtain "who spoke what and when" information from natural conversation
  - For "who part", systems are not often required to output a name for the speakers, i.e., only a generic id which is unique within a meeting/conversation
- Sometimes, referred to as meeting recognition and meeting analysis



Agenda

Part 1: Introduction

- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

Part 2: Current state-of-the-art systems

- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

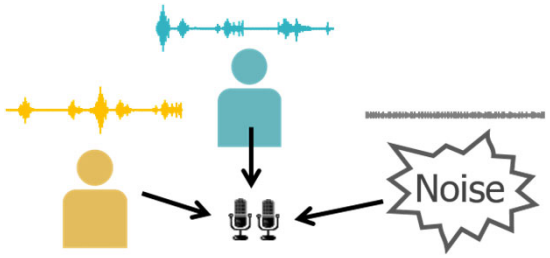
Part 3: A new research trend: Jointly optimal systems

- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

Part 4: Summary and discussion

- 4.1. Strength of each approach
- 4.2. Challenges and fundamental difficulties

Why is distant conversational ASR and analysis difficult?



1. Difficulties originated from the scenario
   - Acoustic point of view
     - Far-field effect (noise, reverberation)
     - Highly interactive, and thus overlapped speech
     - Varied recording environments (e.g., varied number of speakers)
     - Unsegmented (no explicit utterance timing information)
   - Lexical point of view
     - Spontaneous speech (e.g., variable articulation, inconsistent speaking rate, pause fillers, false-starts and self-edits)

Speech enhancement

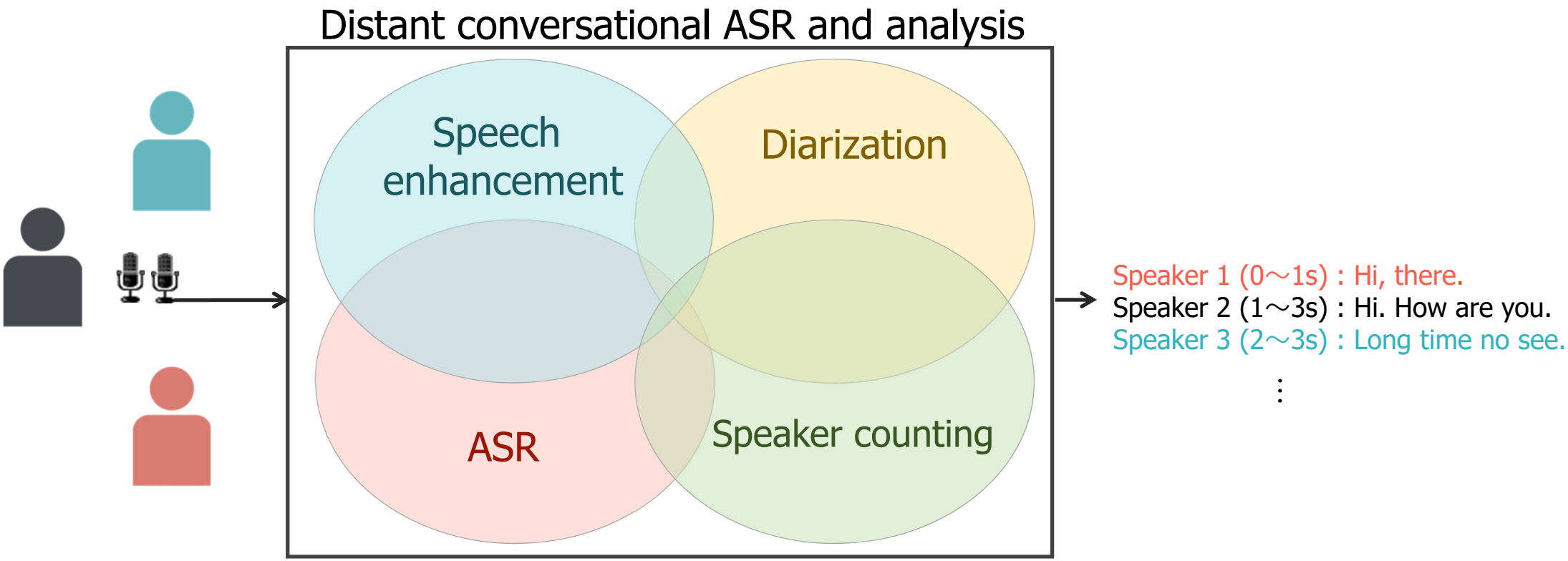
Speaker counting

Diarization

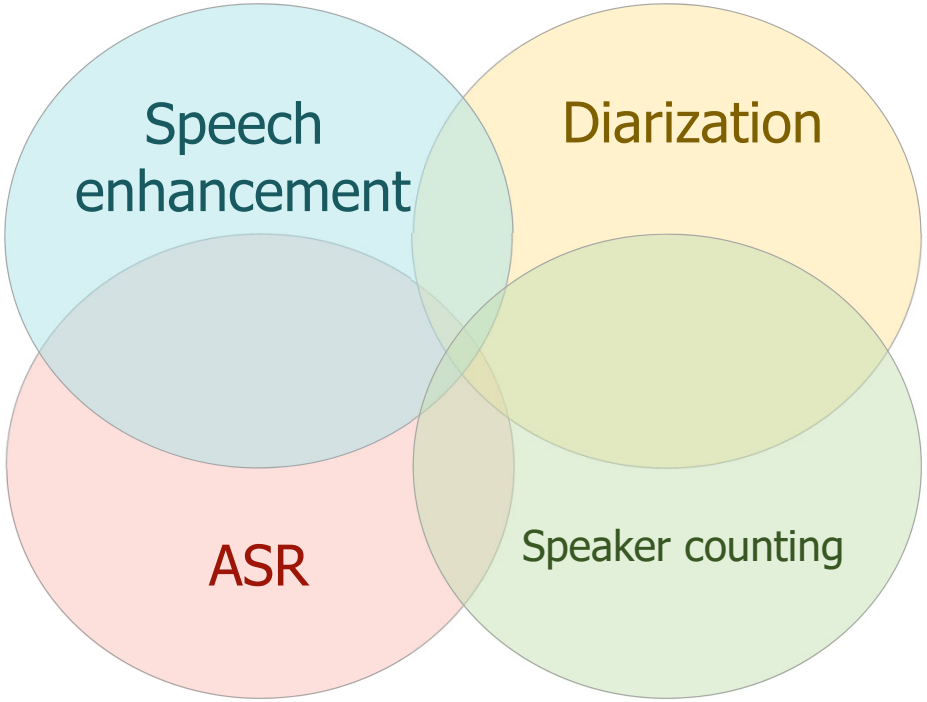
ASR

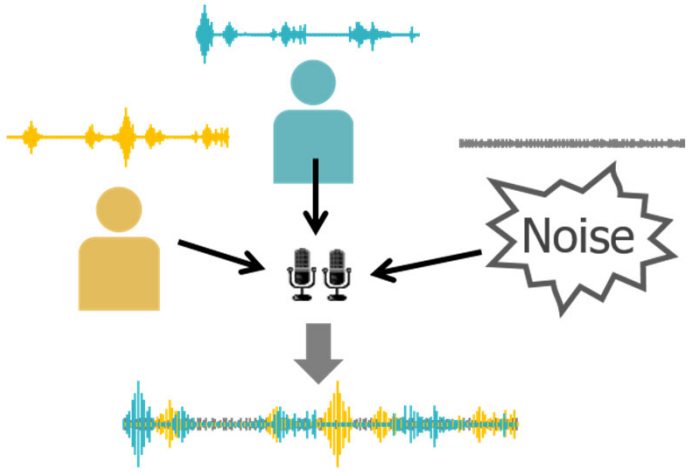
Introduction of sub-tasks

- Distant conversational ASR and analysis can comprises 4 sub-tasks, speech enhancement, ASR, diarization, and speaker counting.
- All these sub-tasks must be accomplished to achieve the goal.

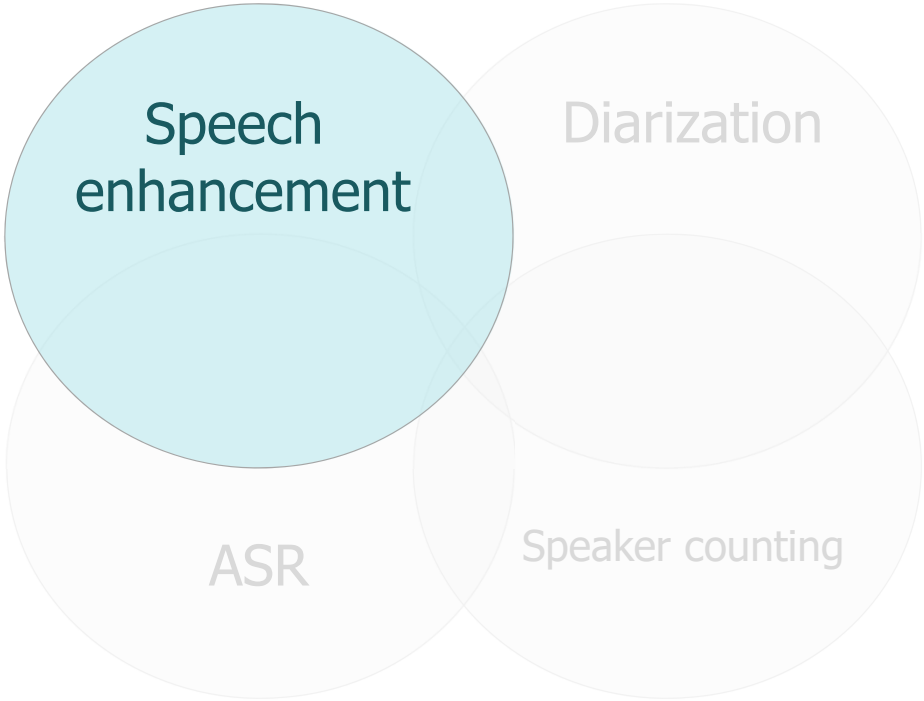


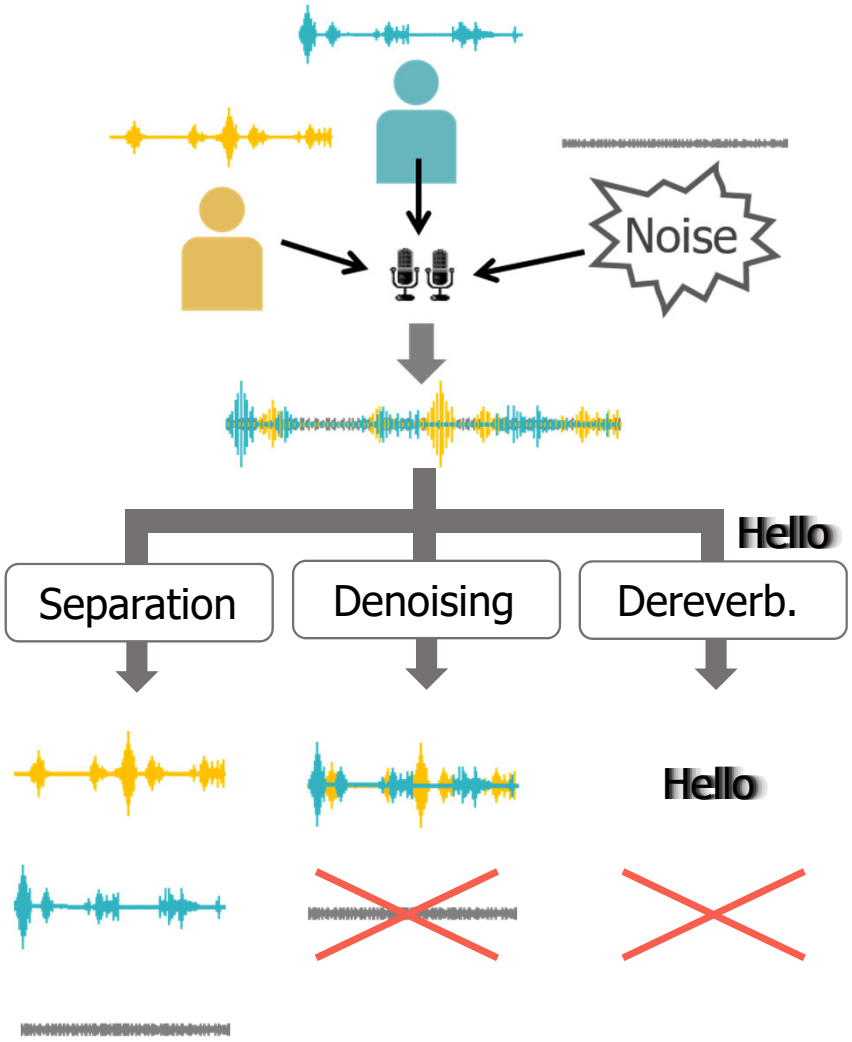
Introduction of sub-tasks



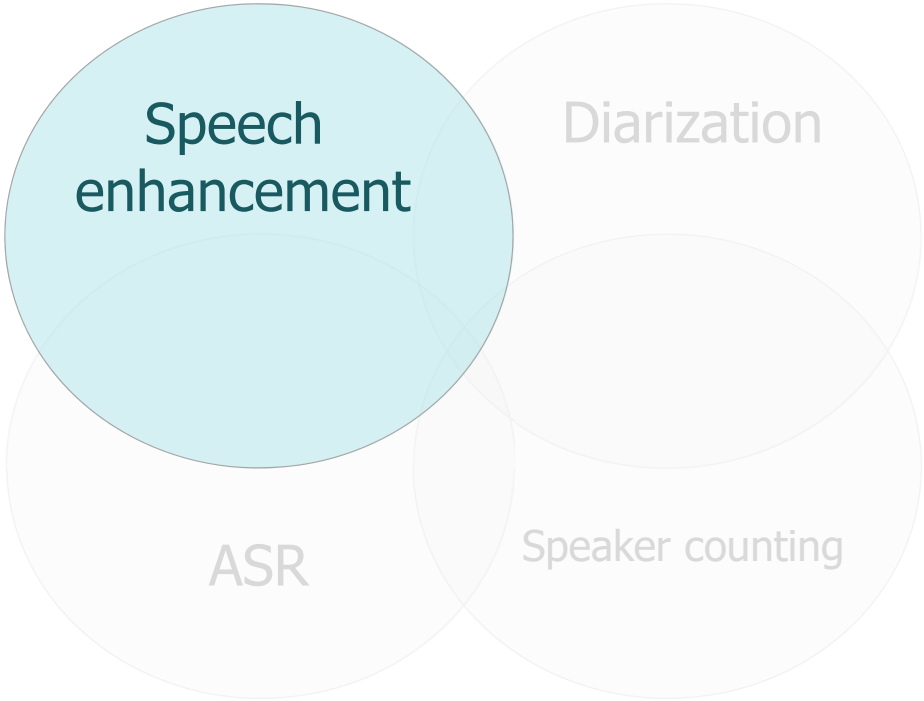


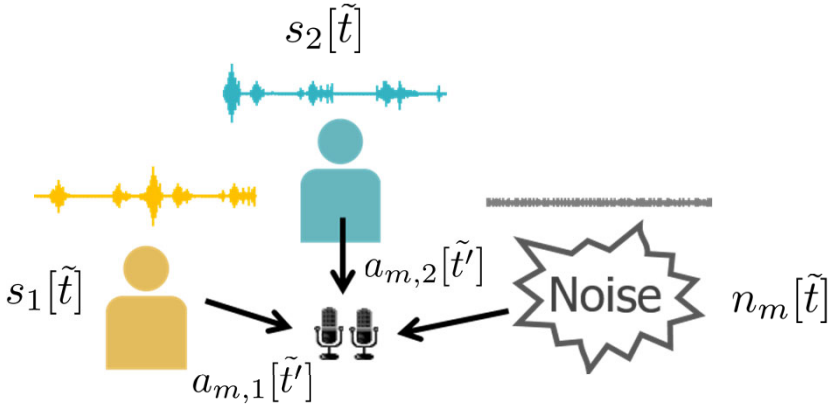
Introduction of sub-tasks





Introduction of sub-tasks





$$y_m[\tilde{t}] = \sum_{k=1}^{K} \sum_{\tilde{t}'} a_{m,k}[\tilde{t}'] s_k[\tilde{t} - \tilde{t}'] + n_m[\tilde{t}],$$

$$= \sum_{k=1}^{K} x_{m,k}[\tilde{t}] + n_m[\tilde{t}],$$

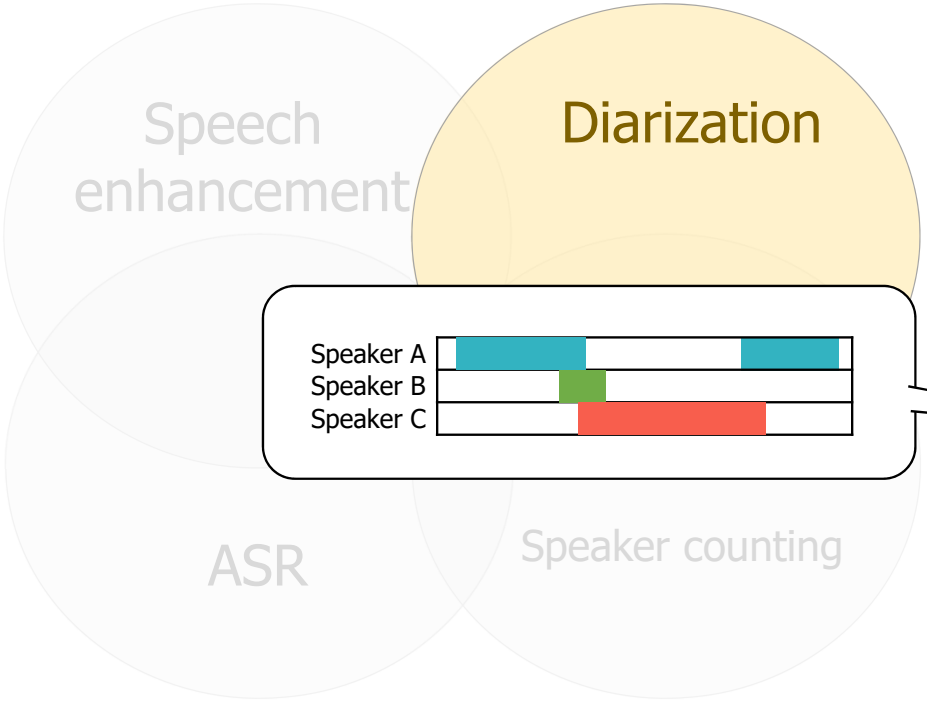
m : Microphone index
k : Speaker index

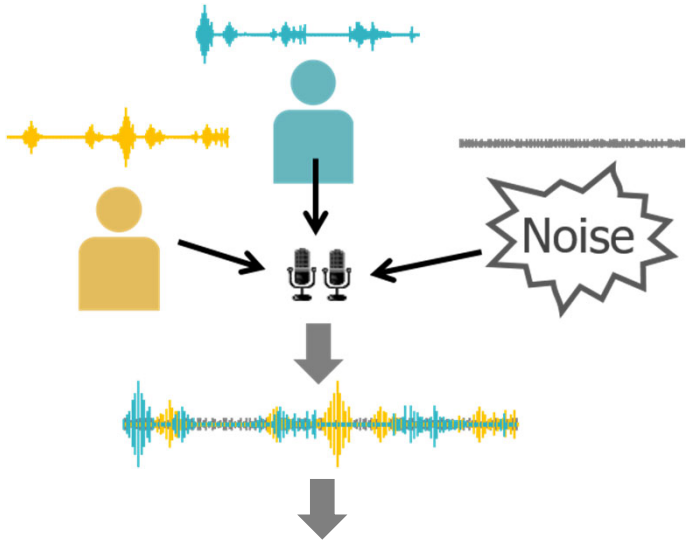
Speech separation: Extracting $s_k[\tilde{t}]$ or $x_{m,k}[\tilde{t}]$

Denoising: Reducing $n_m[\tilde{t}]$

Dereverberation: Reducing effect of $a_{m,k}[\tilde{t}']$

Introduction of sub-tasks





Estimating speech activity label $\mathbf{d}_t \in \{0,1\}^K$ for each speaker at each time t

Equivalently,

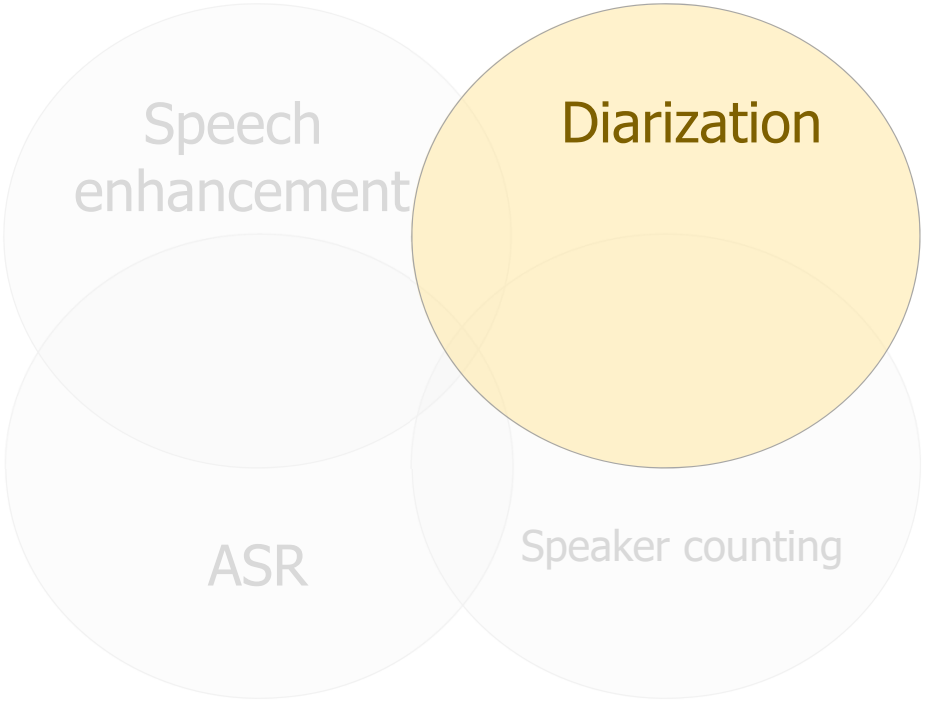
(RTTM file)

Spk ID	Start	Duration
A	0.5s	3.0s
B	2.5s	10.0s
C	14.5s	5.0s
A	20.0s	3.0s
⋮		

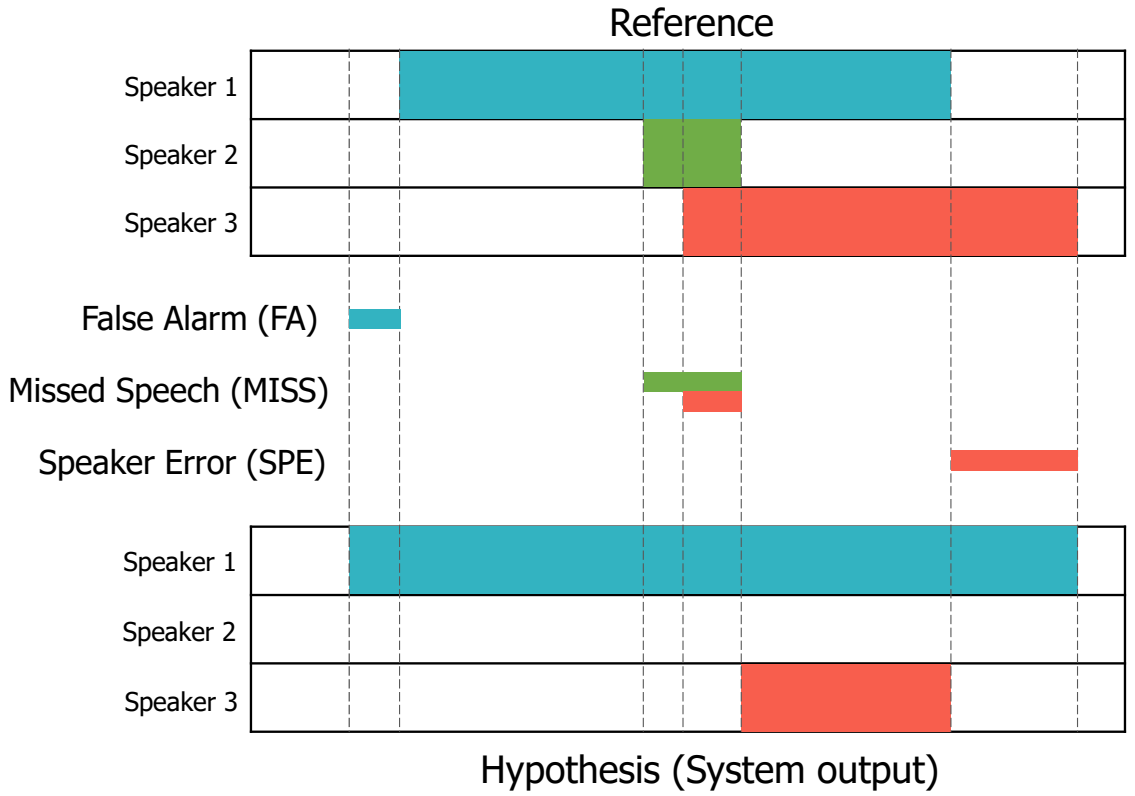
Estimating speech segments with speaker IDs, start timings, and their duration.

RTTM: Rich Transcription Time Marked

Introduction of sub-tasks

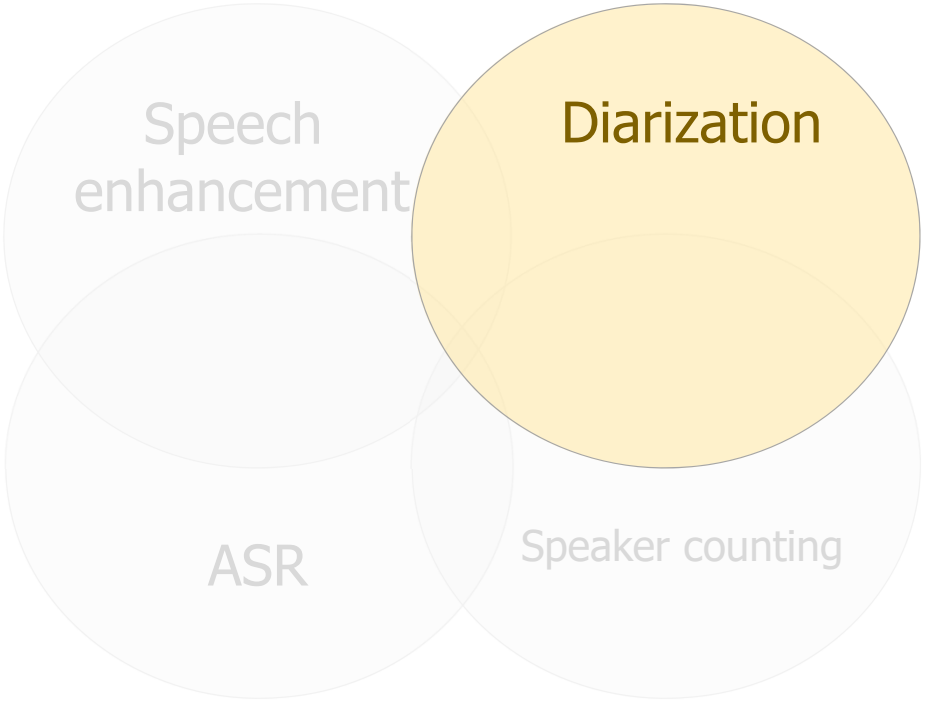


Evaluation metric: Diarization error rate (DER)

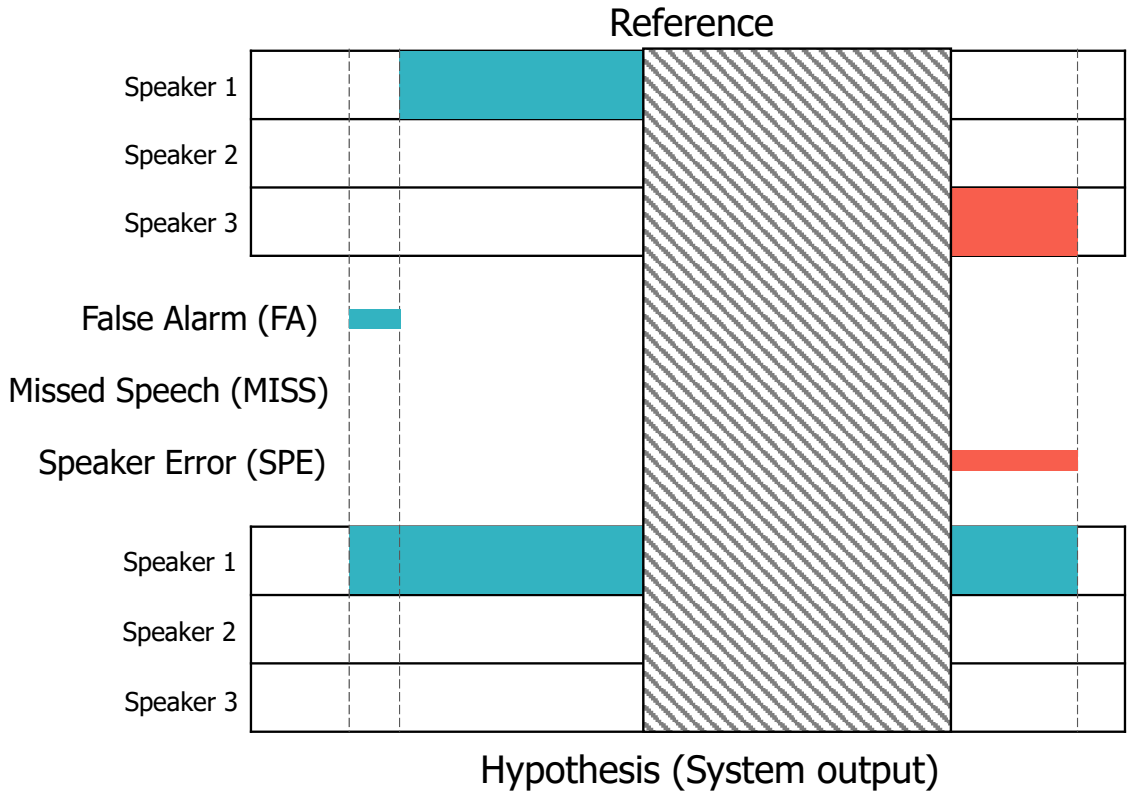


$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}} \quad \text{(\% of time)}$$

Introduction of sub-tasks

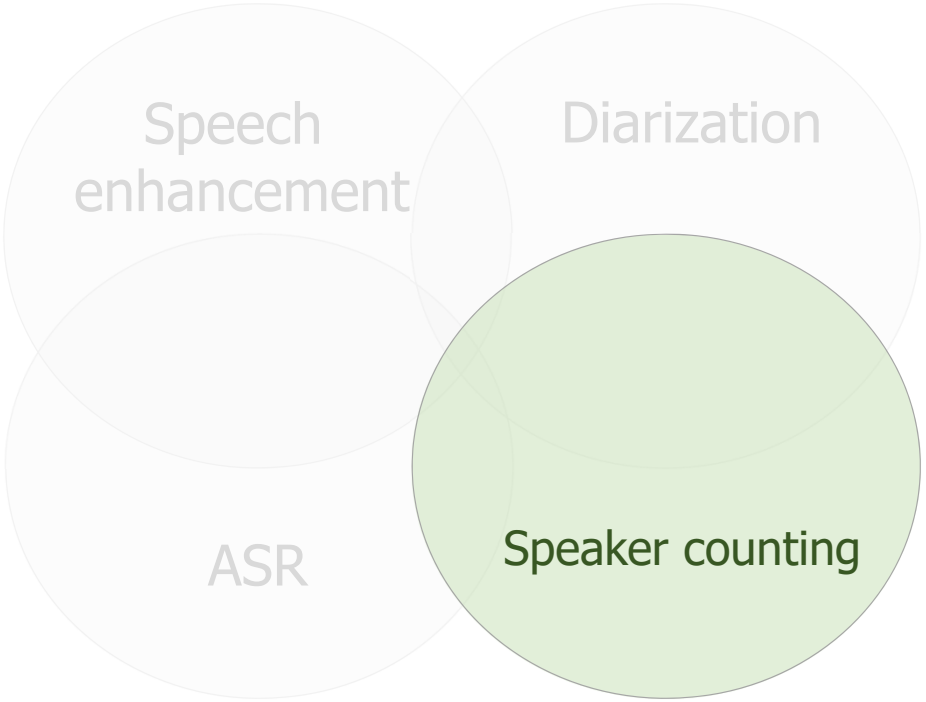


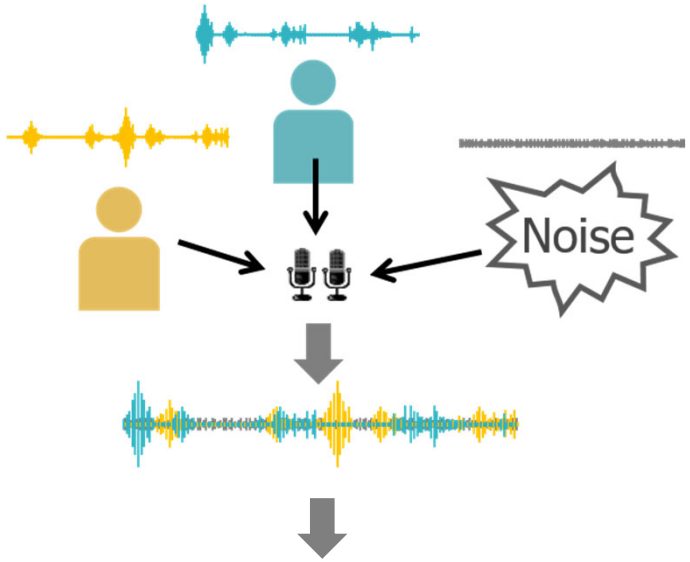
Evaluation metric: Diarization error rate (DER)



$$\text{(\% of time)}\quad \text{DER} = \frac{\text{FA} + \text{MISS} + \text{SPE}}{\text{Total Speech Duration}}$$

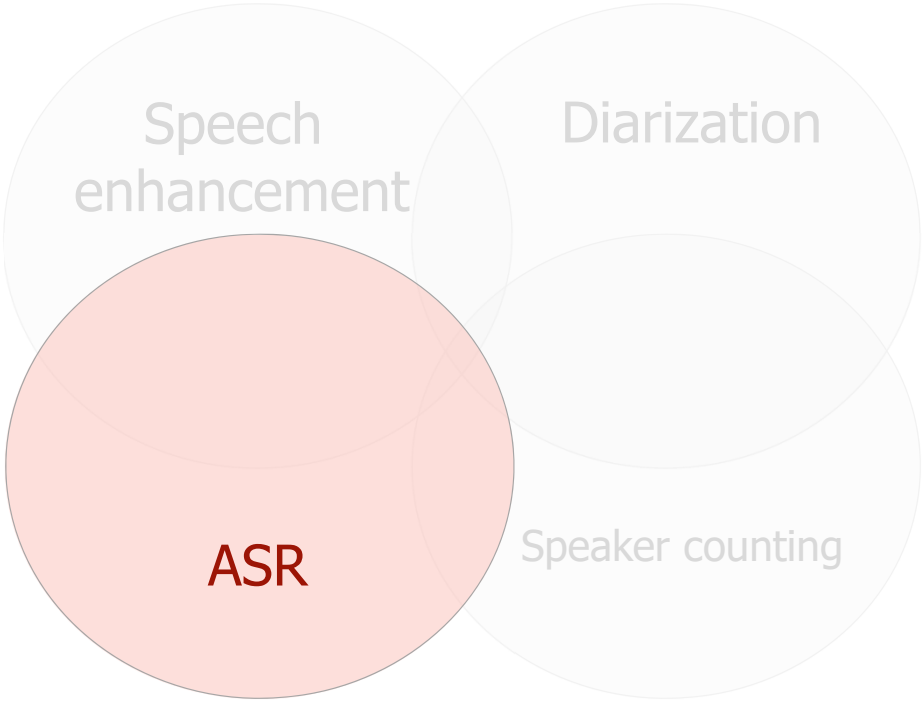
Introduction of sub-tasks

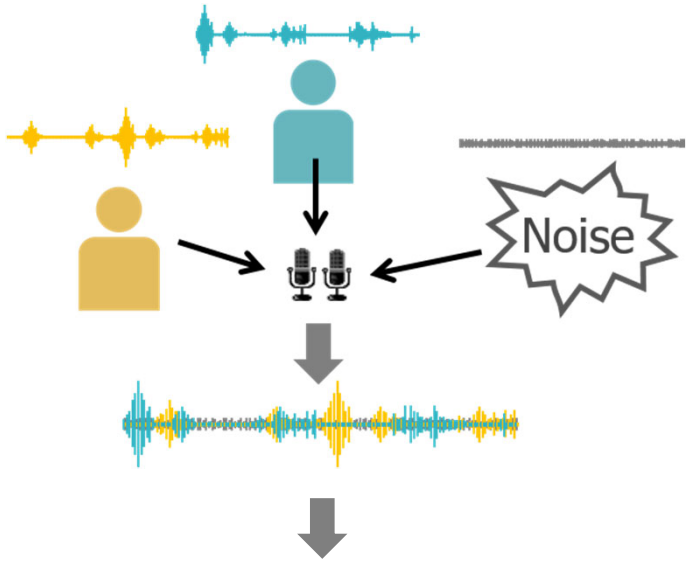




(number of speakers
in a segment/utterance/conversation)

Introduction of sub-tasks





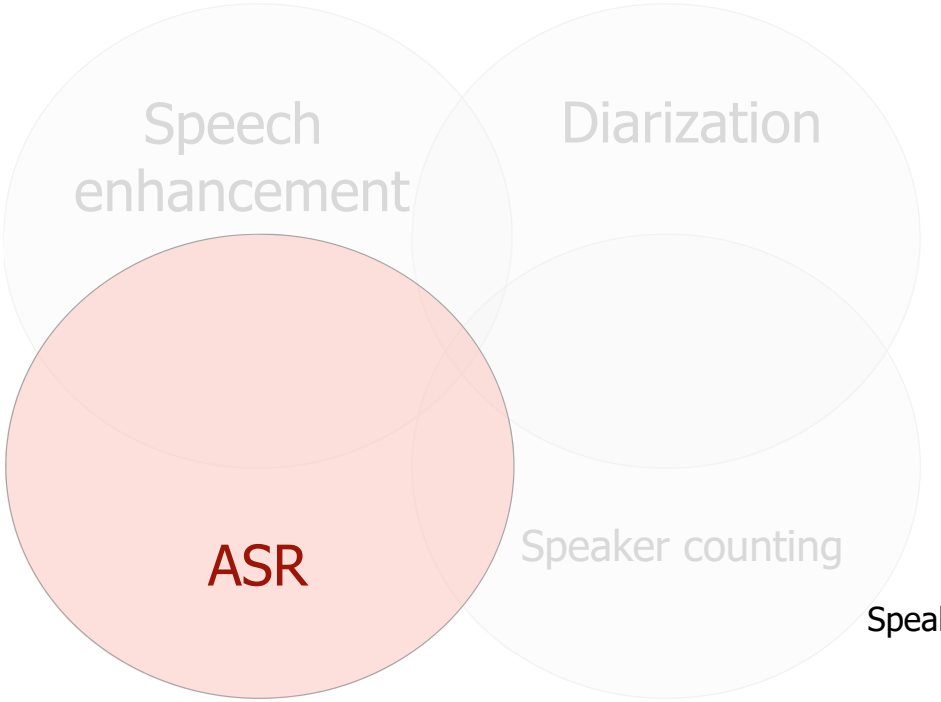
Finding the most probable word sequence,
typically, with the following Bayes decision theory:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{W}|\mathbf{O}),$$
$$= \underset{\mathbf{W}}{\operatorname{argmax}}\, p(\mathbf{O}|\mathbf{W})p(\mathbf{W}).$$

Word sequence

Observed feature sequence

Introduction of sub-tasks



Evaluation metric: Word Error Rate (WER)

$$WER = \frac{Deletion\ (D) + Insertion\ (I) + Substitution\ (S)}{Total}$$

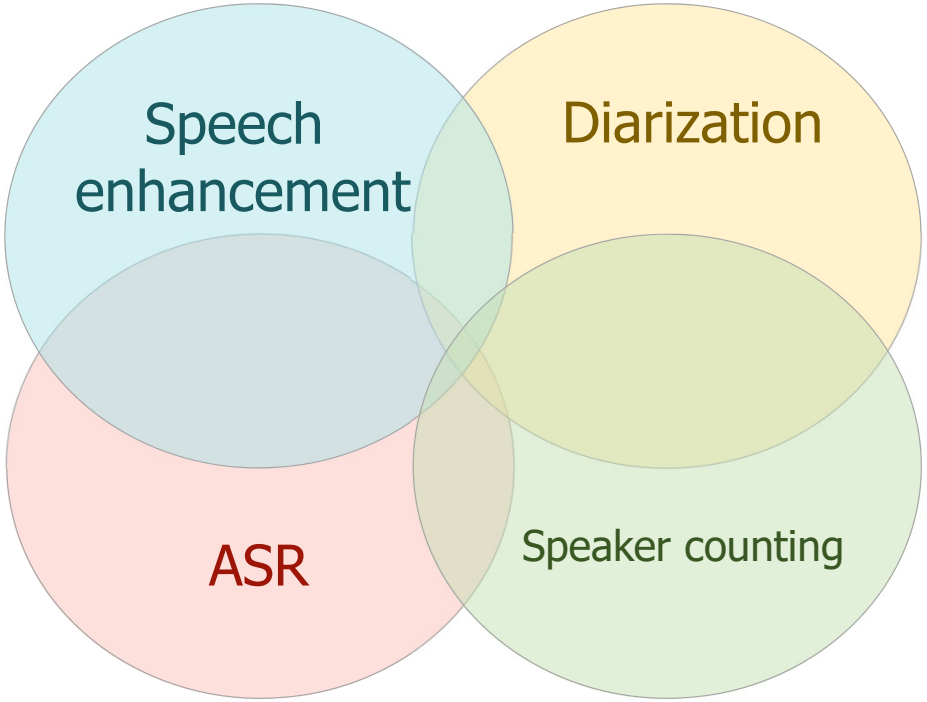
Reference:	A	B	C	D		F	G	H	I	J	K	L	M	N
Hypothesis:	A		C	D	E	F	G	H	I	J	K	M	M	N
Error type:		*D*			*I*							*S*		

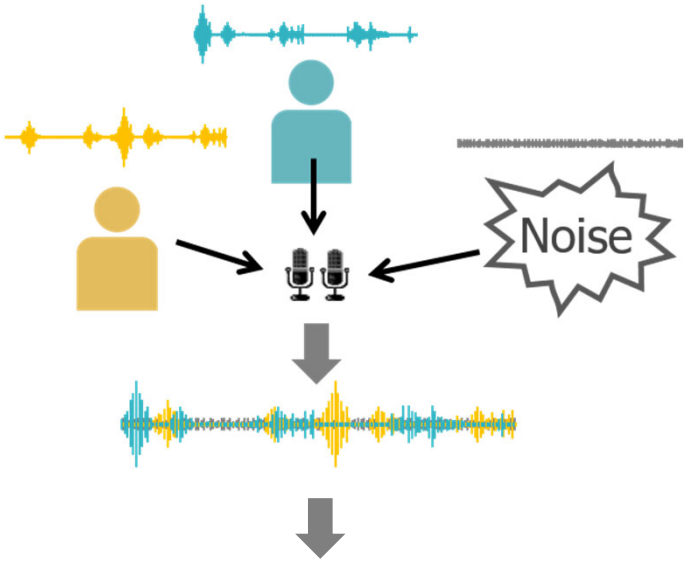
Evaluation metric: Speaker attributed Word Error Rate (**S**WER)

$$SWER = \frac{Deletion\ (D) + Insertion\ (I) + Substitution\ (S)}{Total}$$

Speaker 1	Reference:	A	B	C	D	E	F							
	Hypothesis:	A		C	D	E	F	G	H	I				
	Error type:		*D*					*I*	*I*	*I*				
Speaker 2	Reference:							G	H	I	J	K	L	M
	Hypothesis:										J	K	L	M	N
	Error type:							*D*	*D*	*D*				*I*

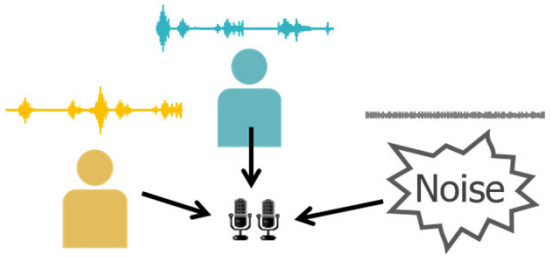
Introduction of sub-tasks





All these sub-tasks must be accomplished to achieve the distant conversational speech recognition and analysis.

Why is distant conversational ASR and analysis difficult?



1. Difficulties originated from the scenario
   - Acoustic point of view
     - Far-field effect (noise, reverberation)
     - Highly interactive, and thus overlapped speech
     - Varied recording environments (e.g., varied number of speakers)
     - Unsegmented (no explicit utterance timing information)
   - Lexical point of view
     - Spontaneous speech (e.g., variable articulation, inconsistent speaking rate, pause fillers, false-starts and self-edits)

Speech enhancement

Speaker counting

Diarization

ASR

2. Difficulty of system development
   - Simply collecting and assembling many required modules is already very challenging.
   - Optimization of the whole system is far more difficult.

Power of open source

End-to-end optimization

Agenda

Part 1: Introduction

- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

Part 2: Current state-of-the-art systems

- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

Part 3: A new research trend: Jointly optimal systems

- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

Part 4: Summary and discussion

- 4.1. Strength of each approach
- 4.2. Challenges and fundamental difficulties

Why is achieving this goal important?

- It can unfold a myriad of new speech applications, such as
  - Meeting minute and summary generation (Video by Microsoft)
  - Human-robot interaction (Video by Hitachi)
  - Communication agent that can understand, respond to and facilitate conversation
  - Conversation analysis and assessment
  - ⋮

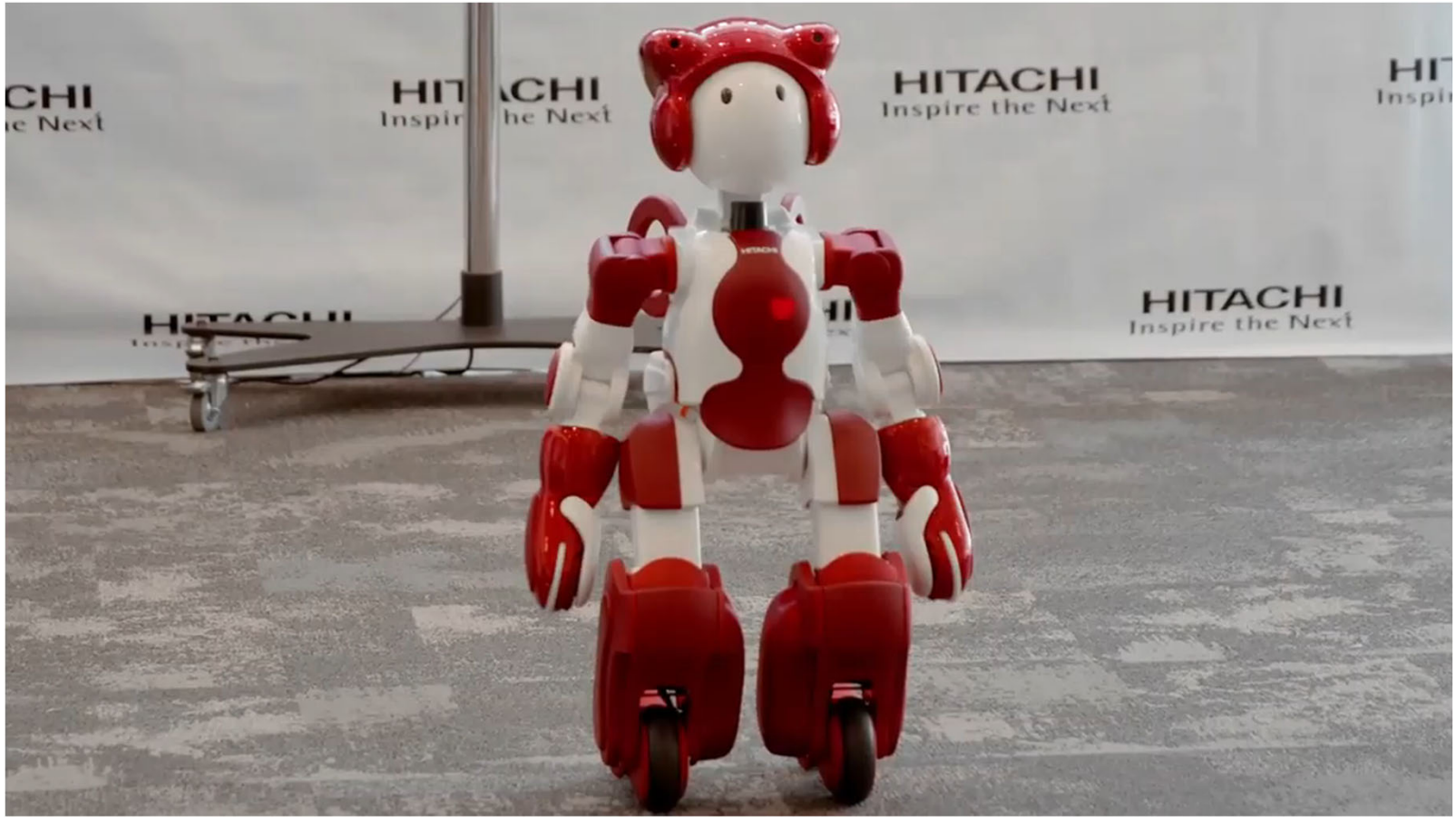
Why is achieving this goal important?

- Application: Meeting minute and summary generation (Video by Microsoft)

T. Yoshioka et al., Advances in Online Audio-Visual Meeting Transcription, ASRU, 2018 https://www.youtube.com/watch?v=ddb3ZgAp9TA

Why is achieving this goal important?

- Application: Human-robot interaction (Video by Hitachi)



https://www.youtube.com/watch?v=lPhA31DAO28

Why is achieving this goal important?

- It can unfold a myriad of new speech applications, such as
  - Meeting minute and summary generation (Microsoft)
  - Human-robot interaction (Hitachi)
  - Communication agent that can understand, respond to and facilitate conversation

  - Hearing assistance
  - Conversation analysis and assessment
  - ⋮

Agenda

Part 1: Introduction

- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

Part 2: Current state-of-the-art systems

- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

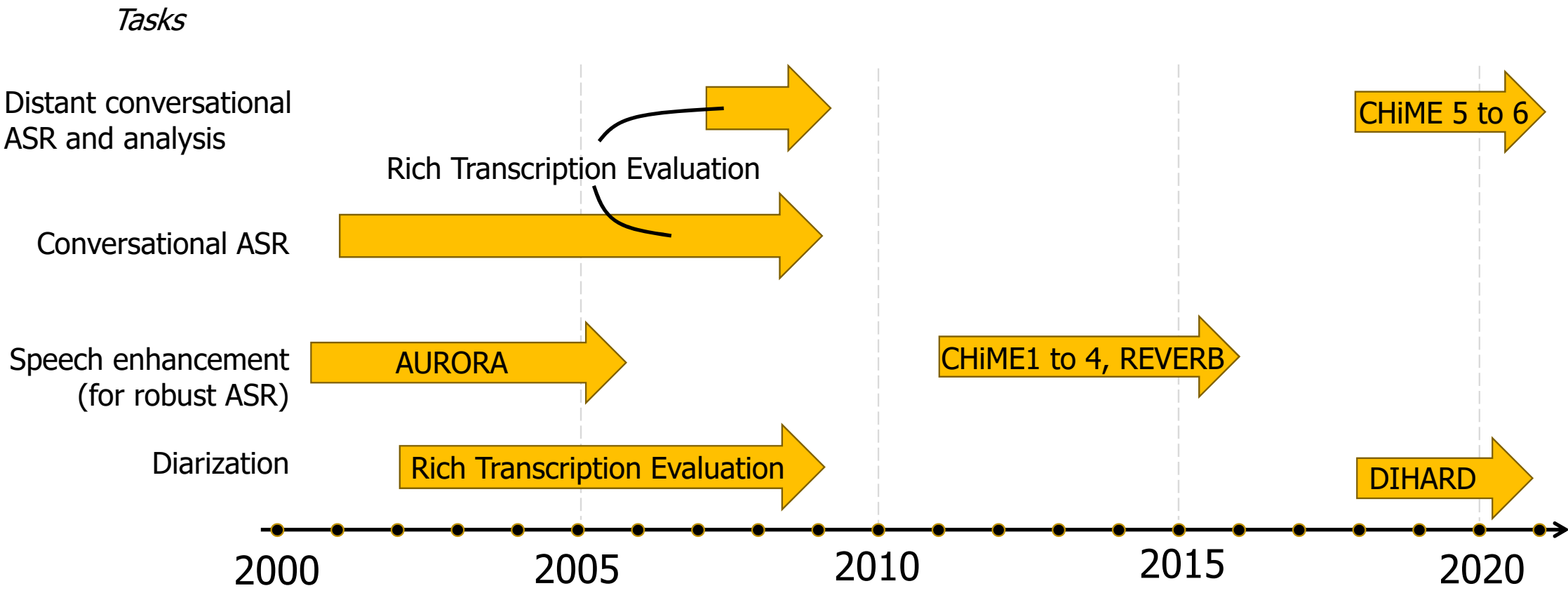
Part 3: A new research trend: Jointly optimal systems

- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

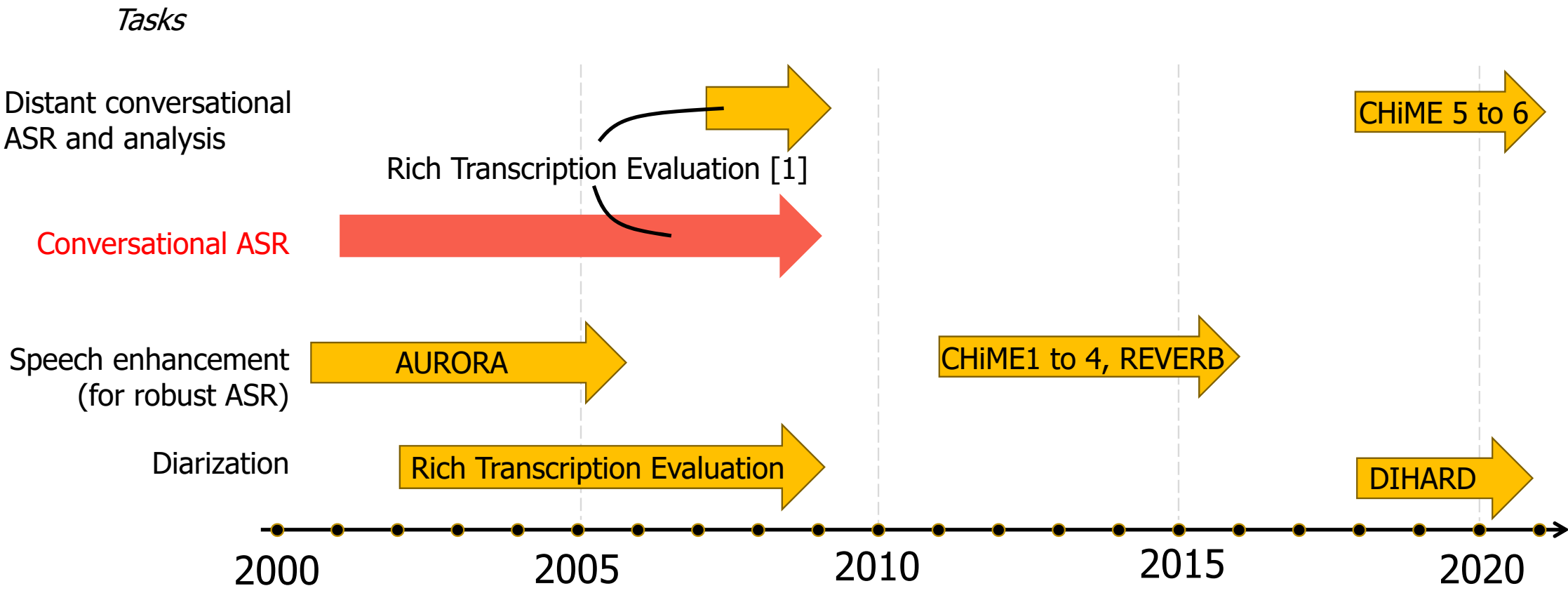
Part 4: Summary and discussion

- 4.1. Strength of each approach
- 4.2. Challenges and fundamental difficulties

History of distant conversational ASR and analysis research

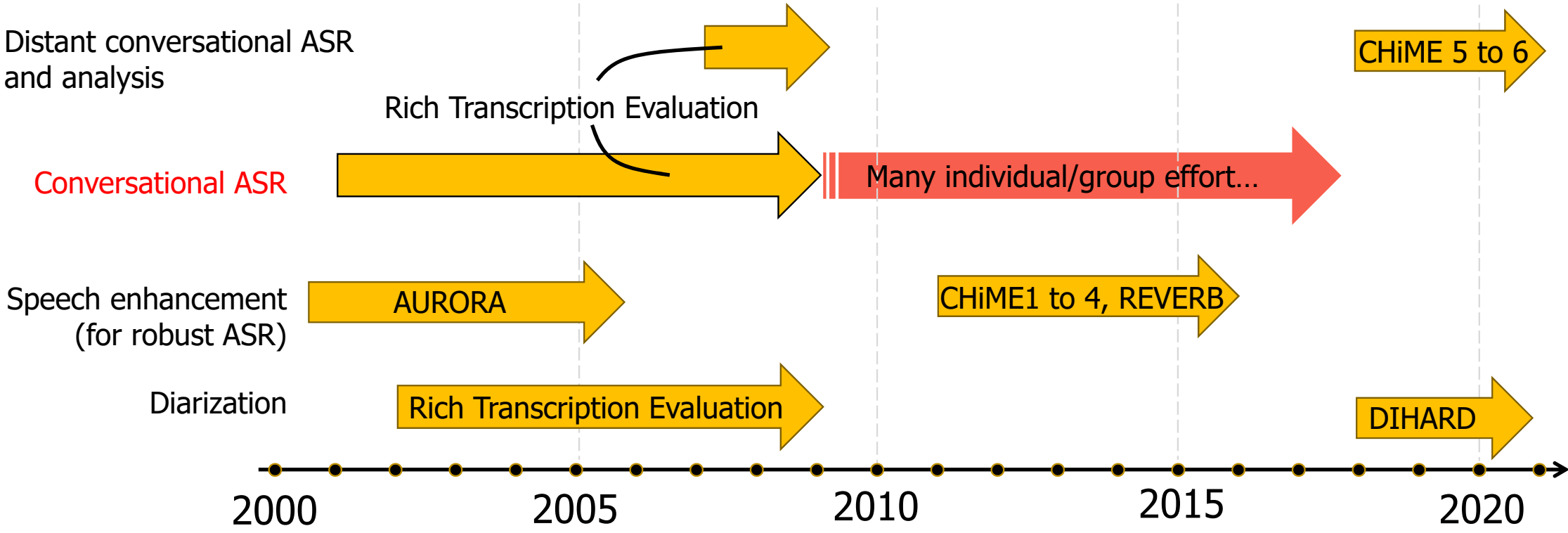


History of distant conversational ASR and analysis research



[1] J. G. Fiscus et al., The Rich Transcription 2007 Meeting Recognition, Multimodal Technologies for Perception of Humans, 2007

History of distant conversational ASR and analysis research

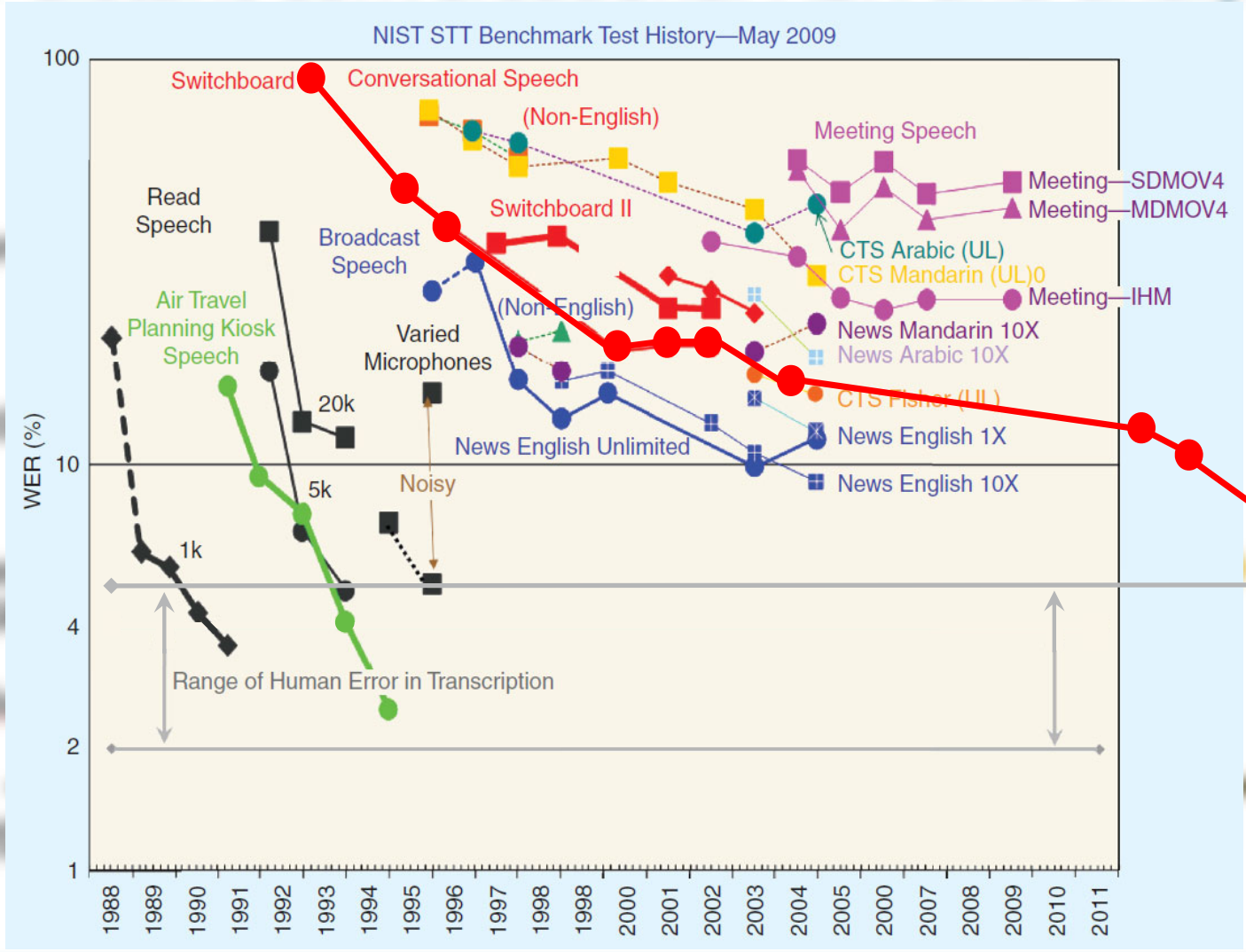


W. Xiong et al., Achieving Human Parity in Conversational Speech Recognition, arXiv, 2016
W. Xiong et al., The Microsoft 2017 Conversational Speech Recognition System, arXiv, 2017

A. Stolcke et al., Comparing human and machine errors in conversational speech transcription, Interspeech, 2017
G. Saon et al., English Conversational Telephone Speech Recognition by Humans and Machines, Interspeech,2017

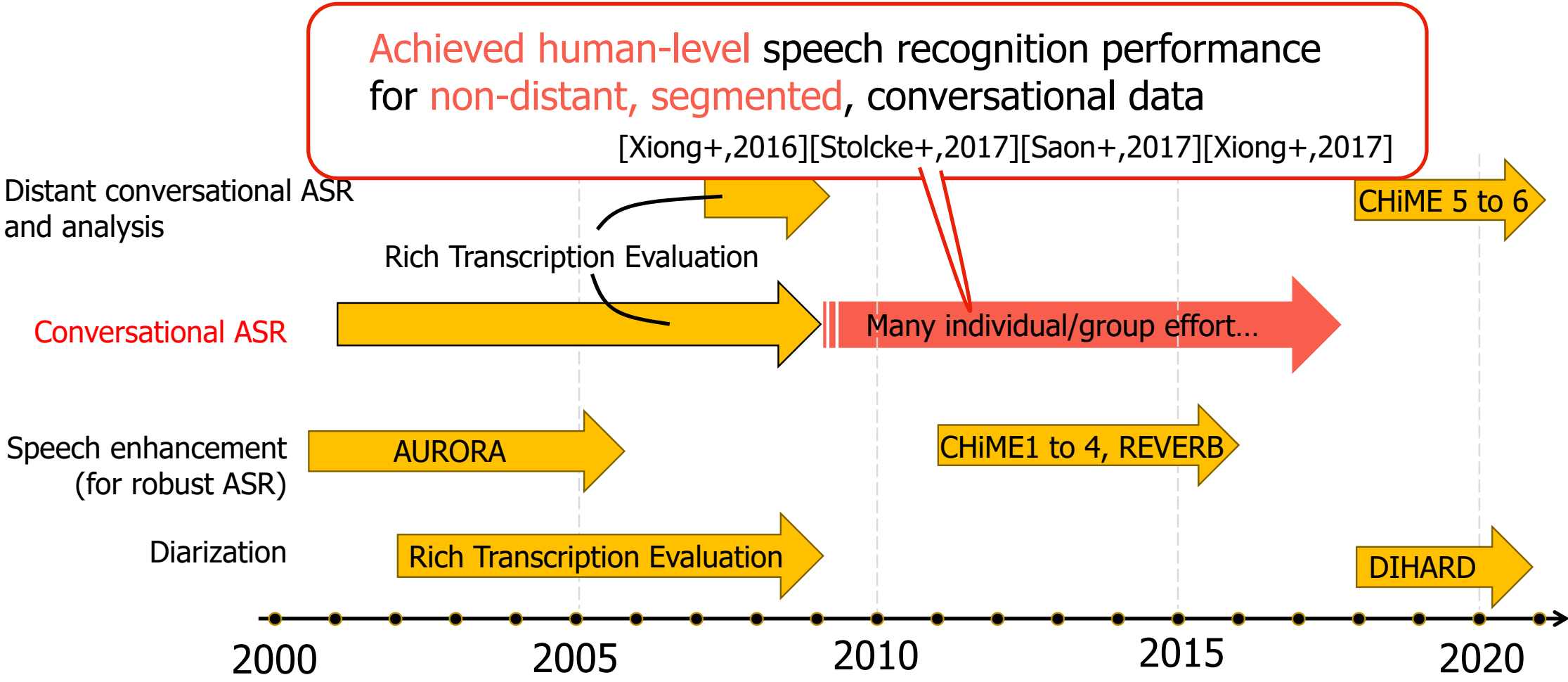
History of distant conversational ASR and analysis research

(Fig. from [1])



[1] Xiaodong He, Li Deng, Speech Recognition, Machine Translation, and Speech Translation —A Unified Discriminative Learning Paradigm, IEEE SP Magazine, 2011

History of distant conversational ASR and analysis research

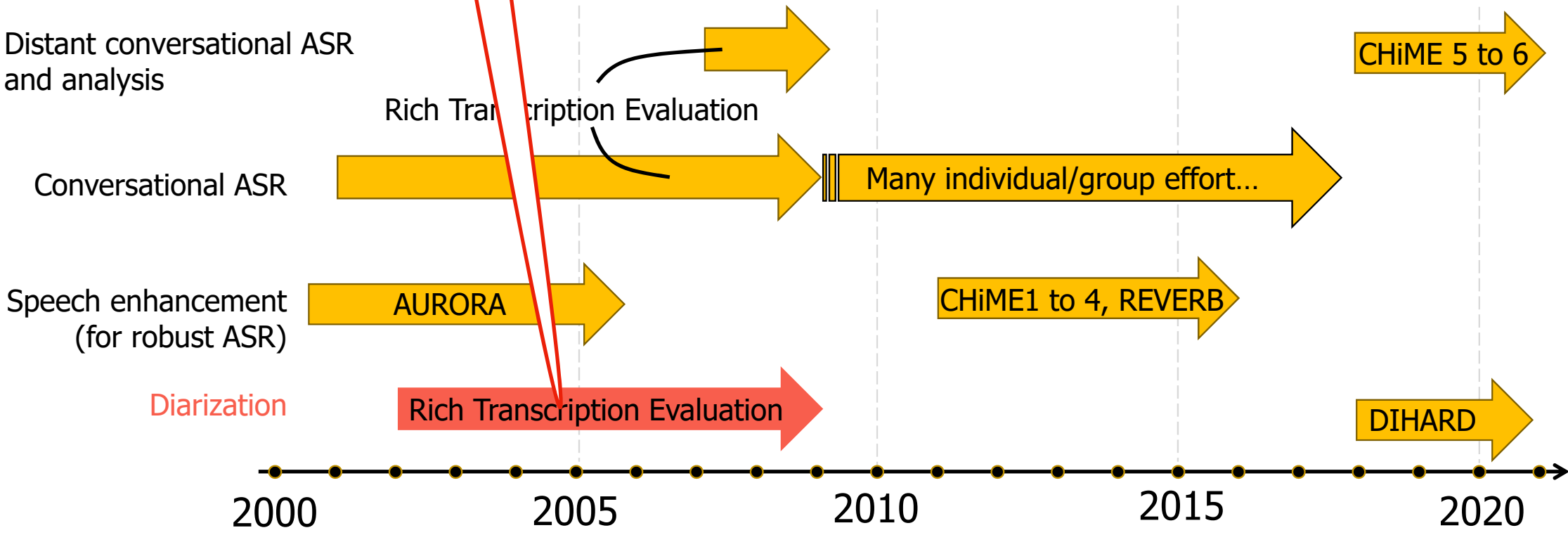


W. Xiong et al., Achieving Human Parity in Conversational Speech Recognition, arXiv, 2016
W. Xiong et al., The Microsoft 2017 Conversational Speech Recognition System, arXiv, 2017

A. Stolcke et al., Comparing human and machine errors in conversational speech transcription, Interspeech, 2017
G. Saon et al., English Conversational Telephone Speech Recognition by Humans and Machines, Interspeech,2017

History of distant conversational ASR and analysis research

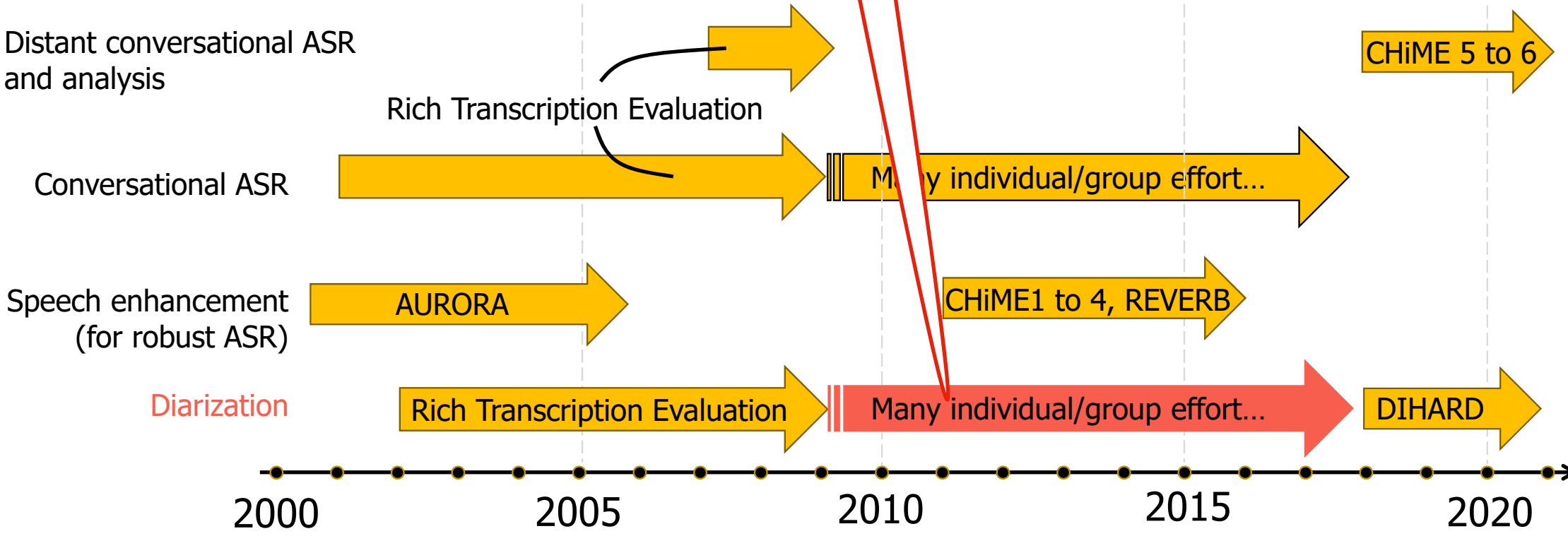
Developed basis of the current-state-of-the-art diarization system,
i.e., speaker embedding + clustering [Anguera+, 2012]



X. Anguera et al., Speaker Diarization: A Review of Recent Research, IEEE TASLP, 2012,

History of distant conversational ASR and analysis research

Solid improvement on embeddings, and clustering [Anguera+, 2012]

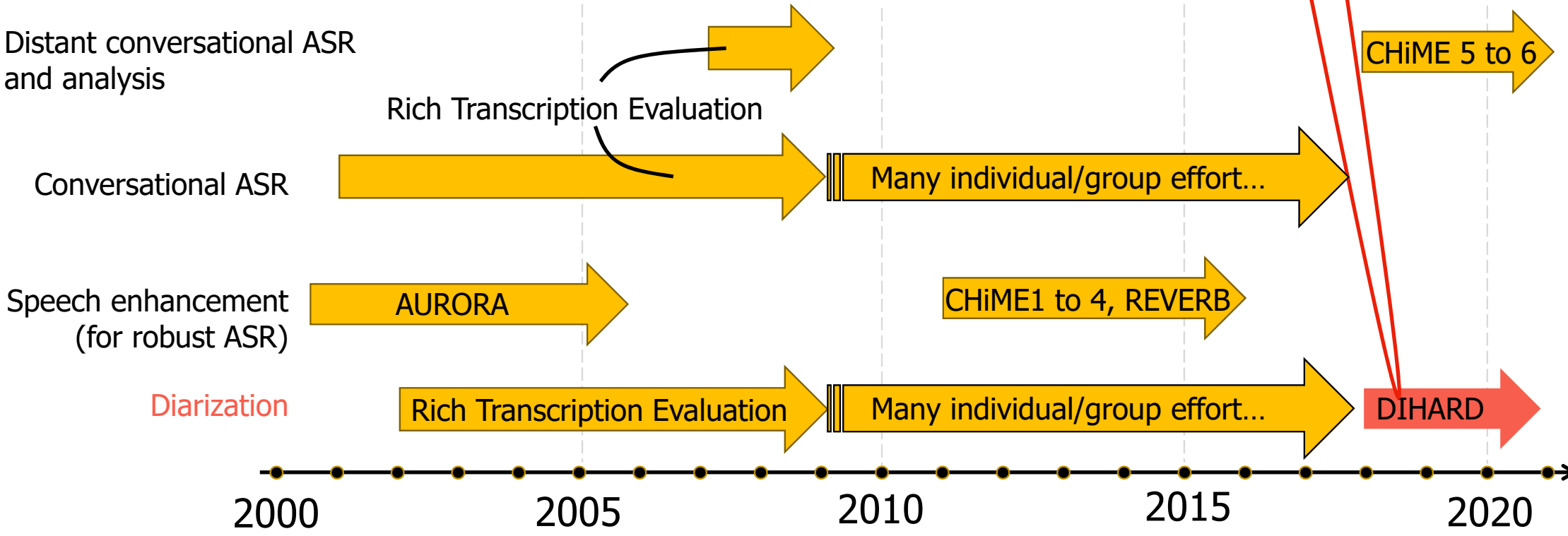


X. Anguera et al., Speaker Diarization: A Review of Recent Research, IEEE TASLP, 2012,

History of distant conversational ASR and analysis research

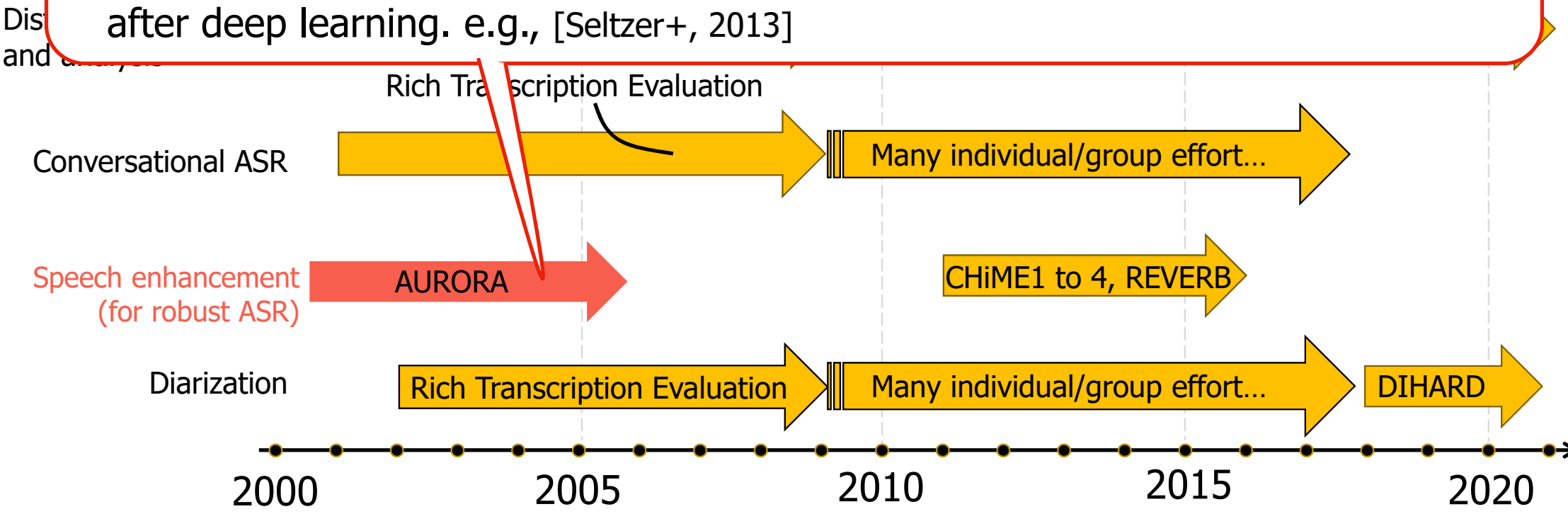
- Speaker embedding clustering
- End-to-end diarization

(Details will be covered in Part 3.1!)



History of distant conversational ASR and analysis research

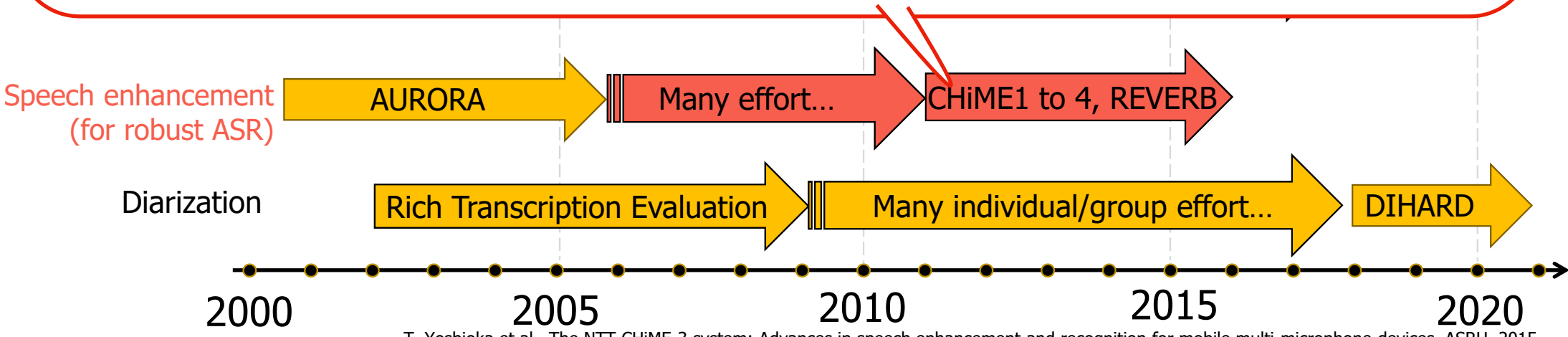
(Single-channel read-speech corpora for noise-robust ASR)
- Developed robust features, and feature enhancement algorithms [Li+, 2014]
- Most 1ch speech enhancement technologies became obsolete after deep learning. e.g., [Seltzer+, 2013]



J. Li et al., An Overview of Noise-Robust AutomaticSpeech Recognition, IEEE TASLP, 2014
M. Seltzer et al., An investigation of deep neural networks for noise robust speech recognition, ICASSP, 2013

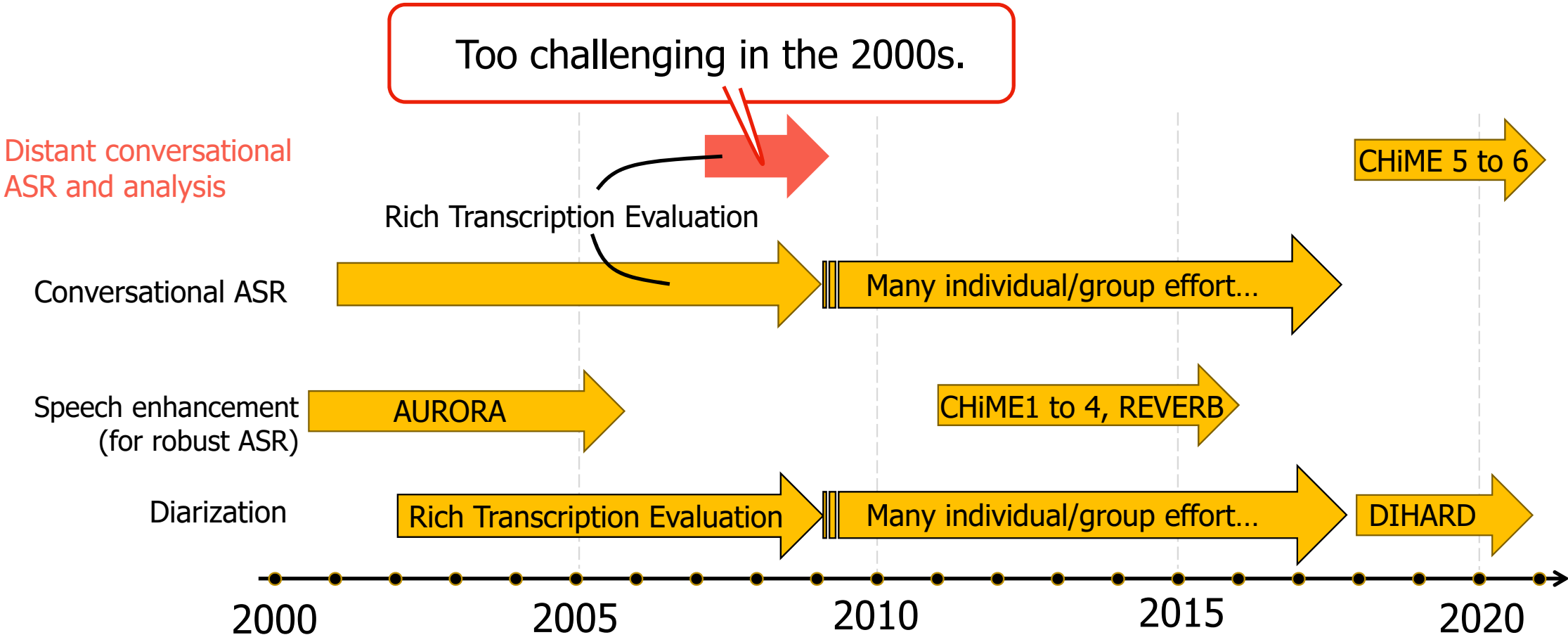
History of distant conversational ASR and analysis research

(Multichannel read-speech corpora for noise/reverb-robust ASR)
- Multichannel enhancement worked nicely with deep-learning ASR [Yoshioka+, 2015][Heymann+, 2016][Delcroix+, 2015]
- Achieved super-human recognition performance for noisy data [Amodei+, 2015].
- Techniques developed/used in these challenges are still employed as a part of the state-of-the-art systems in today's challenge systems. (will be introduced in Part 2!)



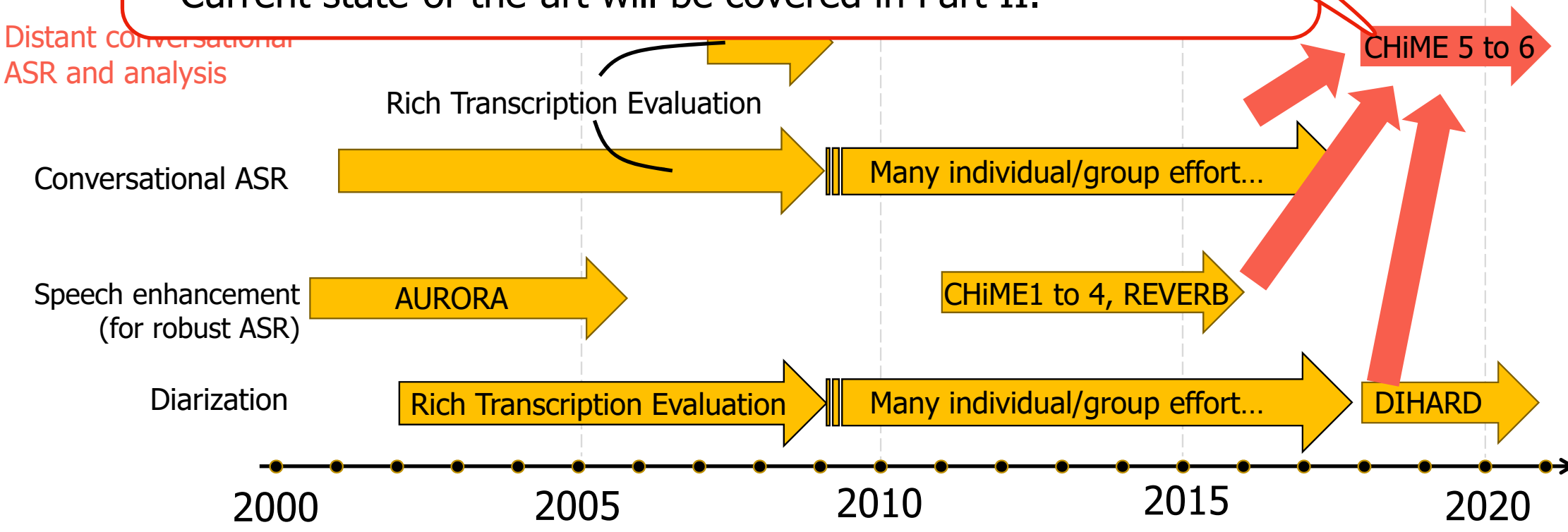
T. Yoshioka et al., The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices, ASRU, 2015
J. Heymann et al., Neural network based spectral mask estimation for acoustic beamforming, ICASSP, 2016
M. Delcroix et al., Strategies for distant speech recognitionin reverberant environments, EURASIP Journal on Advances in SP, 2015
D. Amodei, Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin, arXiv, 2015

History of distant conversational ASR and analysis research



History of distant conversational ASR and analysis research

- Now that we have human-level/super-human modules, we should be ready to tackle distant conversational ASR again!
- Current state-of-the-art will be covered in Part II.



Agenda

Part 1: Introduction

- 1.1. What is distant conversational ASR and analysis?
- 1.2. Why is it difficult?
- 1.3. Why is it important?
- 1.4. Its research history
- 1.5. Typical systems for distant conversational ASR and analysis

Part 2: Current state-of-the-art systems

- 2.1. Descriptions of the techniques
- 2.2. Reproducible baselines

Part 3: A new research trend: Jointly optimal systems

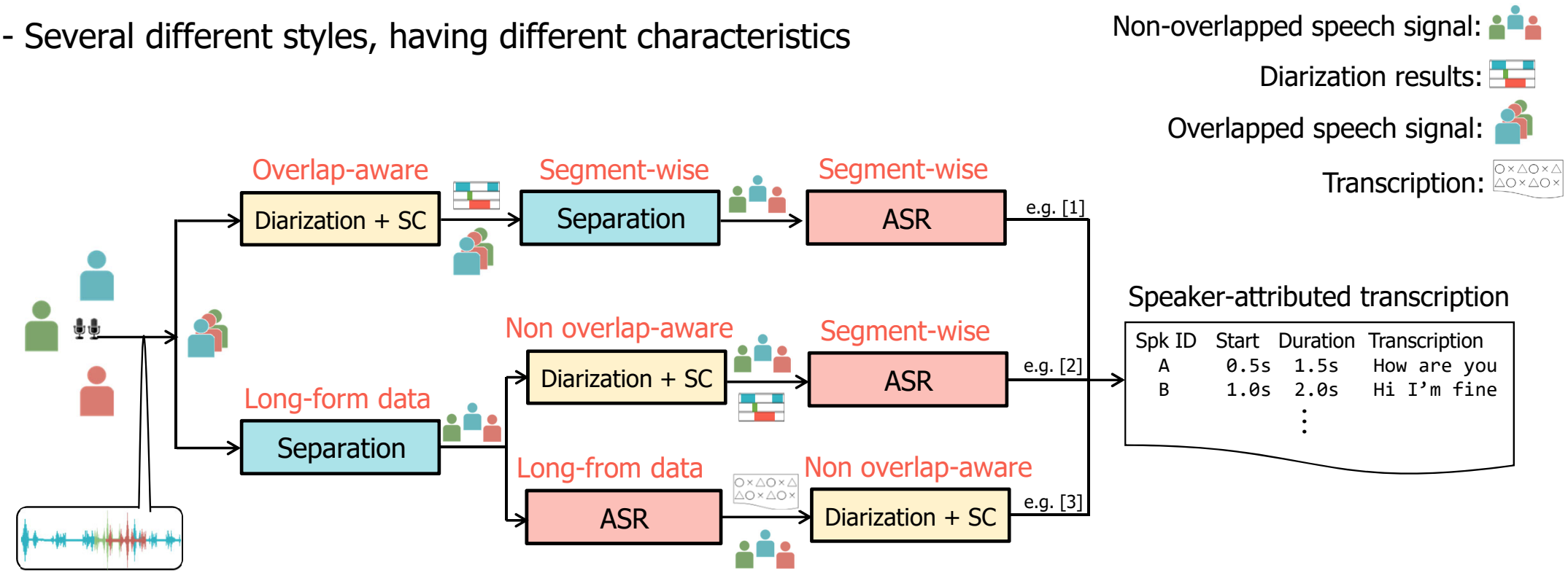
- 3.1. Diarization + x
- 3.2. Enhancement + x (x: other functionalities)
- 3.3. ASR + x

Part 4: Summary and discussion

- 4.1. Strength of each approach
- 4.2. Challenges and fundamental difficulties

Typical systems for distant conversational ASR and analysis

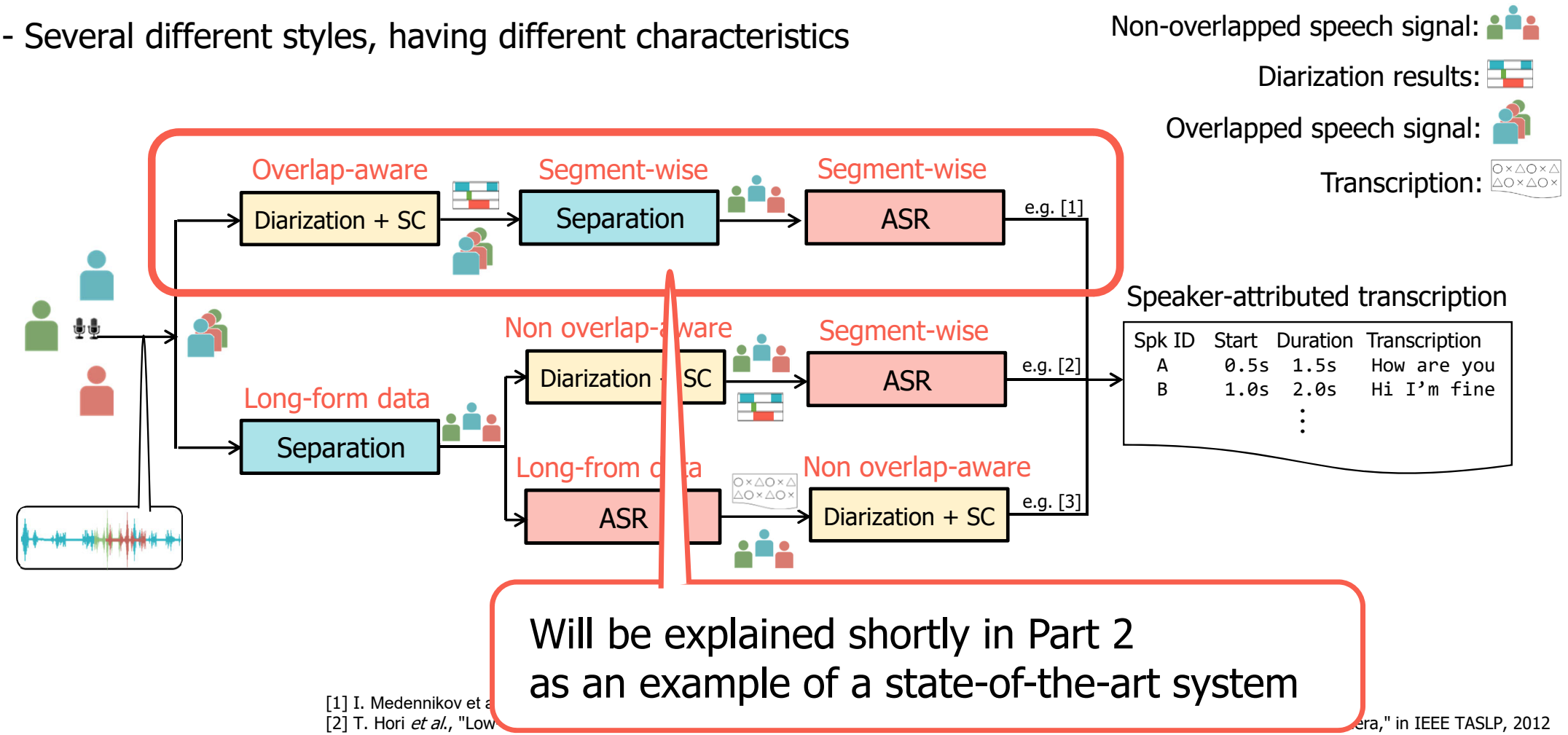
- Several different styles, having different characteristics



[1] I. Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020
[2] T. Hori *et al.*, "Low-Latency Real-Time Meeting Recognition and Understanding Using Distant Microphones and Omni-Directional Camera," in IEEE TASLP, 2012
[3] T. Yoshioka et al., "Advances in Online Audio-Visual Meeting Transcription", ASRU, 2019

Typical systems for distant conversational ASR and analysis

- Several different styles, having different characteristics

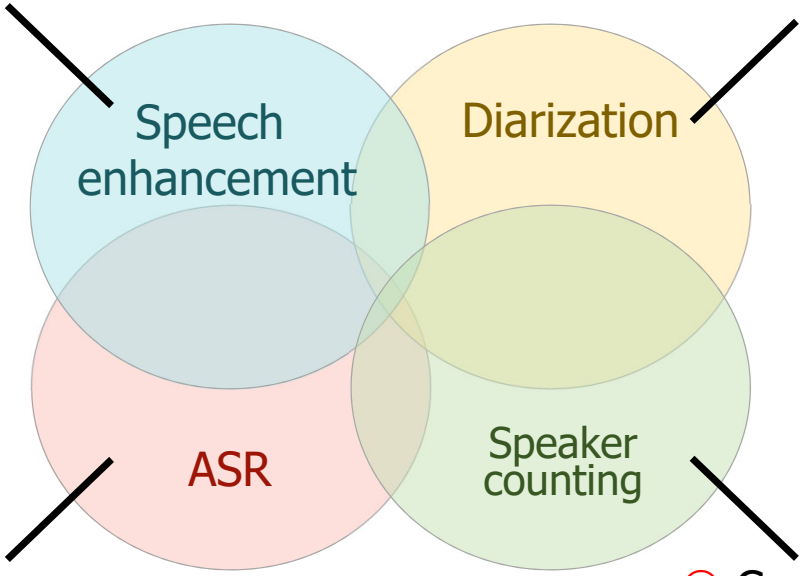


[1] I. Medennikov et a
[2] T. Hori *et al.*, "Low
era," in IEEE TASLP, 2012
[3] T. Yoshioka et al., "Advances in Online Audio-Visual Meeting Transcription", ASRU, 2019

Things we cover☺ and don't cover☹ in this tutorial

☺ State-of-the-art multichannel enhancement.
☺ Neural speech enhancement (mainly separation) and its combination with other modules such as ASR.

☺ State-of-the-art diarization systems
☺ Combination with other modules



☹ Multimodal approaches, e.g., audio-visual approaches [Afrouras+, 2018]

☺ State-of-the-art system
☺ End-to-end ASR combined with other modules
☹ Language model-based approaches, i.e., language model for conversation [Mikolov+, 2012].

☺ Combination with other modules
☹ Systems that perform only speaker counting, e.g. [Stöter+, 2019]

T. Mikolov et al., Context dependent recurrent neural network language model, SLT, 2012
F. Stöter et al., CountNet: Estimating the Number of Concurrent Speakers Using Supervised Learning Speaker Count Estimation, TASLP, 2019
T. Afouras et al., Deep audio-visual speech recognition, IEEE Trans. PAMI, 2018

2. Current state-of-the-art systems

2. Current state-of-the-art systems

2.1. Descriptions of the techniques

2.2. Reproducible baselines

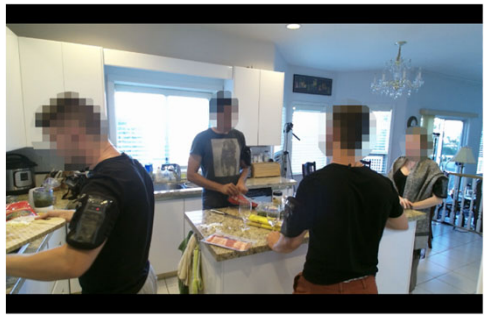
Distant conversational speech recognition

- We have to solve both separation, diarization, and ASR
- We mainly explain state-of-the-art techniques used in the **CHiME-6** systems because of most recent challenge activities
- Note that there are a lot of excellent benchmark activities including Rich transcriptions, AMI, LibriCSS, etc.
  - Many of the techniques introduced here were developed in these previous activities
  - Their effectiveness are verified in the CHiME-6 scenario

CHiME-6 Challenge

Dinner party scenario

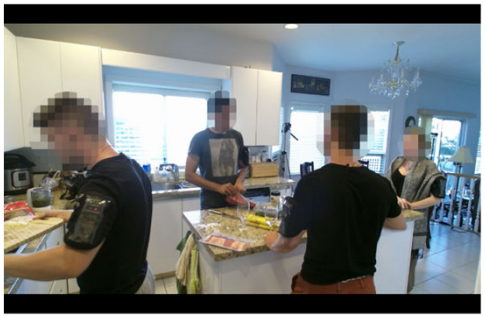
- Recorded in people's actual homes
- Parties of 4 - typically, two hosts and two guests
- Collection of 20 parties each lasting 2 to 3 hours

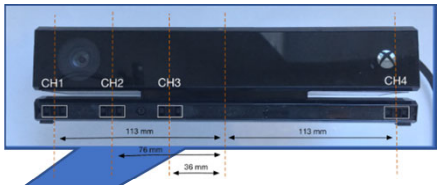



CHiME-6 Challenge

Dinner party scenario

- Recorded in people's actual homes
- Parties of 4 - typically, two hosts and two guests
- Collection of 20 parties each lasting 2 to 3 hours



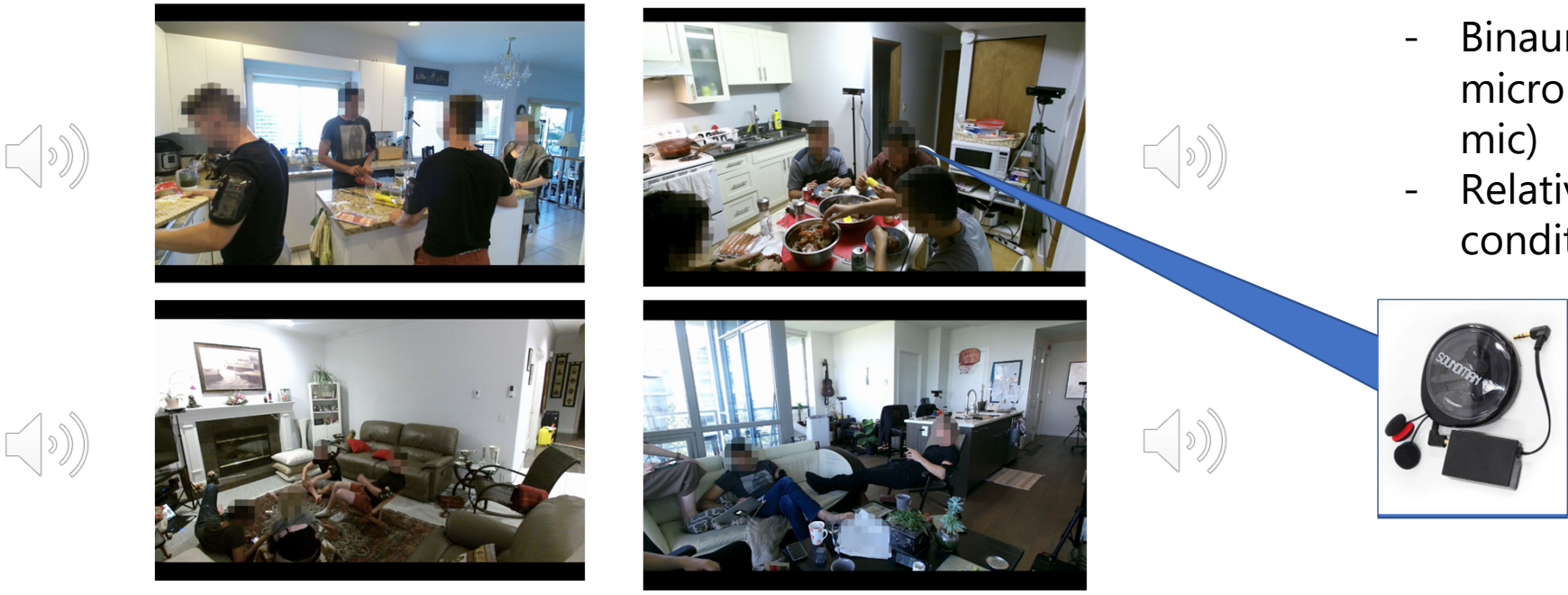



- Six separate Microsoft Kinect devices
- Two Kinects per living area (kitchen, dining, sitting)
- Channel: 6 x 4 audio

CHiME-6 Challenge

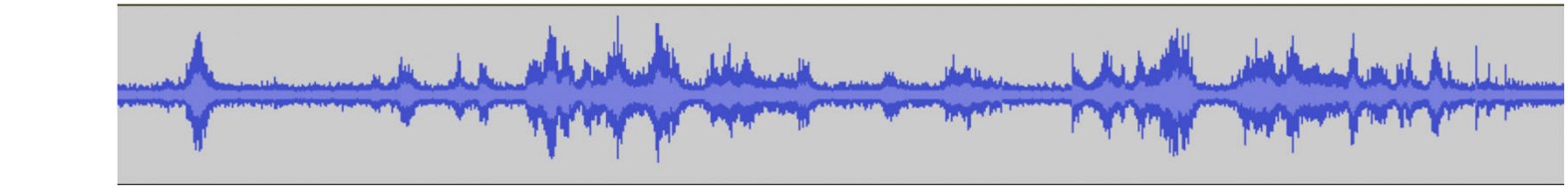
Dinner party scenario

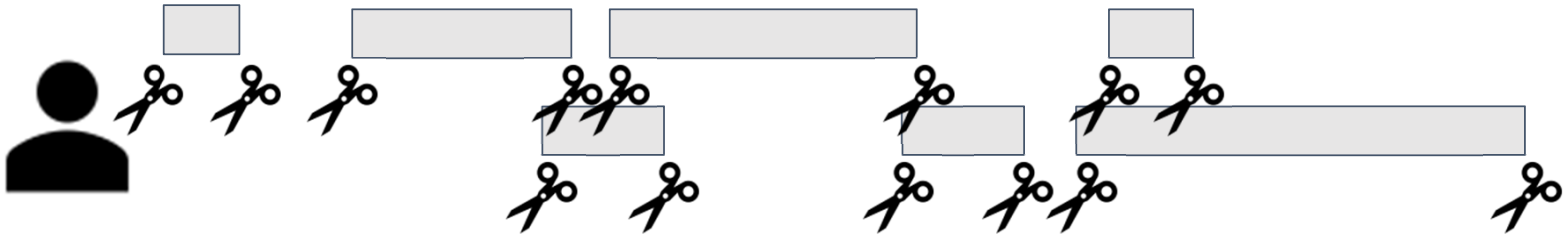
- Recorded in people's actual homes
- Parties of 4 - typically, two hosts and two guests
- Collection of 20 parties each lasting 2 to 3 hours



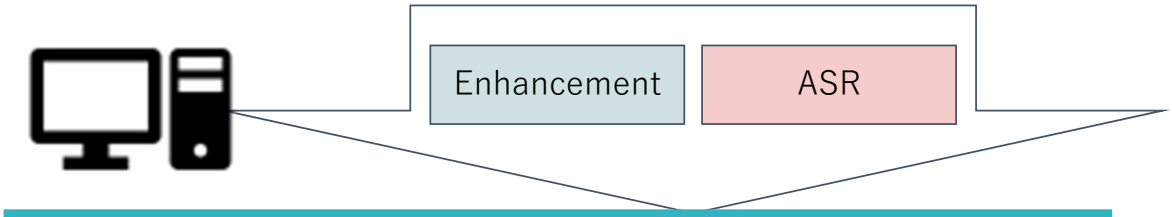
- Binaural in-ear microphones (worn mic)
- Relatively clean conditions

Classical ASR tasks (including previous CHiME challenges)





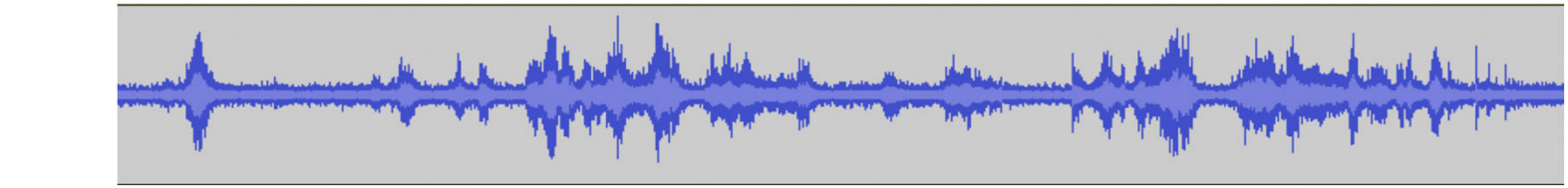
Human annotation

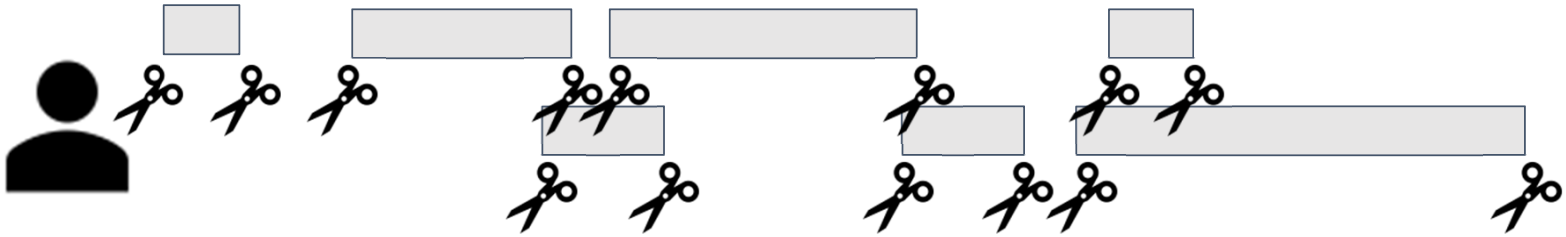


Given

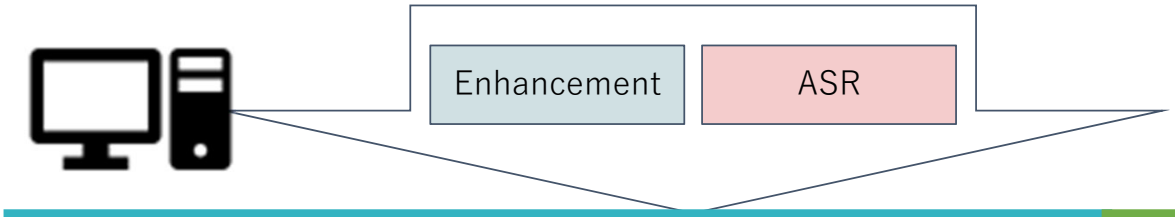
Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	
A-0008380-0009435	A	83.8	94.35	
B-0031242-0031700	B	312.42	317.0	

Classical ASR tasks (including previous CHiME challenges)





Human annotation

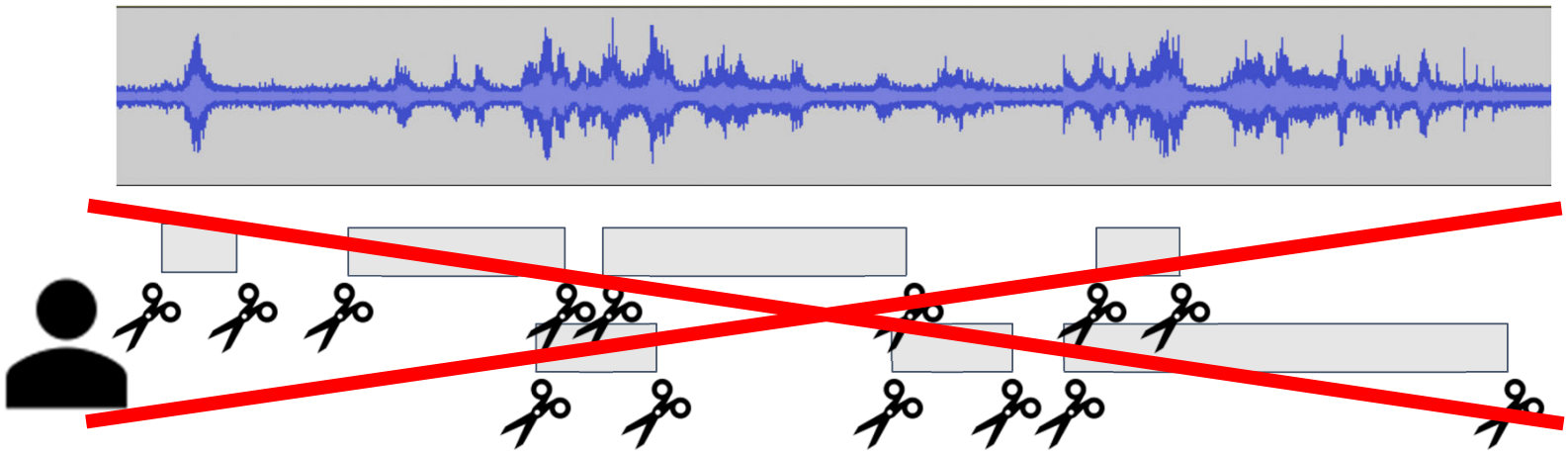


Given

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	yes
A-0008380-0009435	A	83.8	94.35	hey did you put…
B-0031242-0031700	B	312.42	317.0	hmm

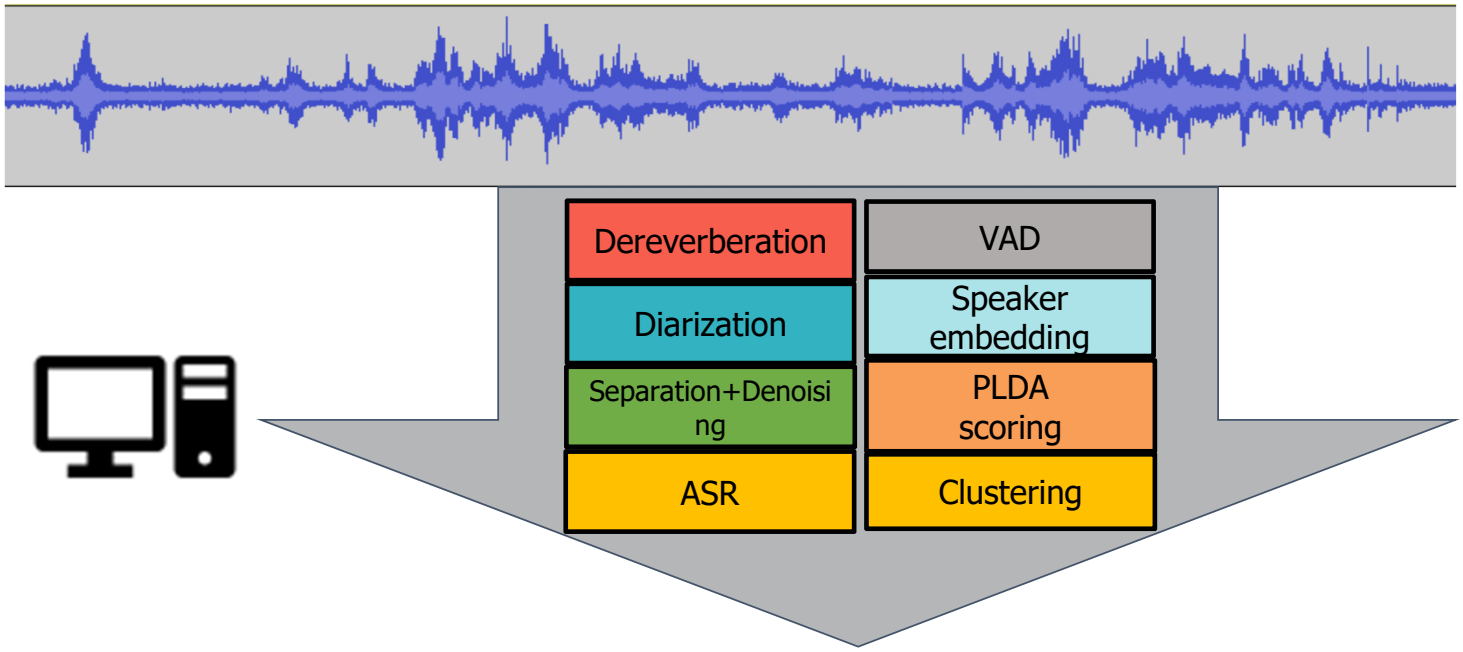
Predict

Distant conversational speech recognition and analysis



Human annotation

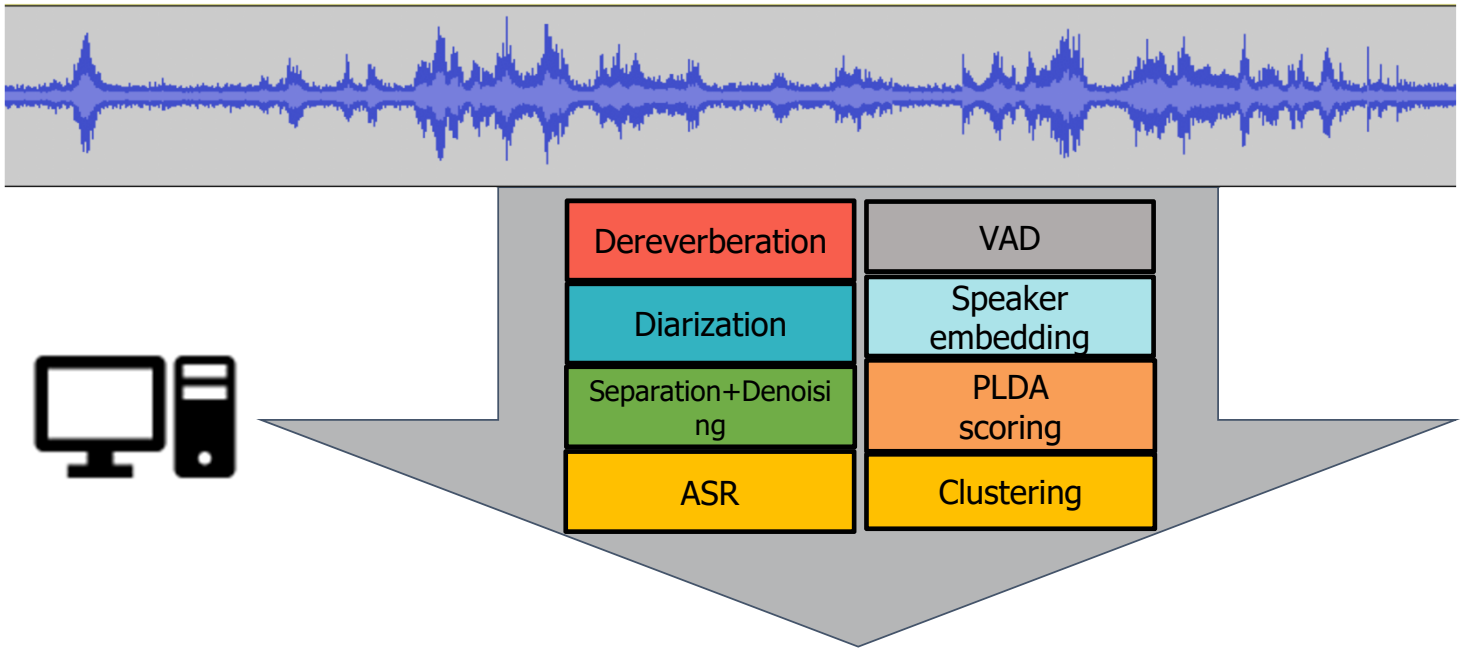
Distant conversational speech recognition and analysis
→ CHiME-6 track 2



Without any given information
(except for #speakers in CHiME-6)

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ

Distant conversational speech recognition and analysis
→ CHiME-6 track 2



Without any given information
(except for #speakers in CHiME-6)

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	yes
A-0008380-0009435	A	83.8	94.35	hey did you put…
B-0031242-0031700	B	312.42	317.0	hmm

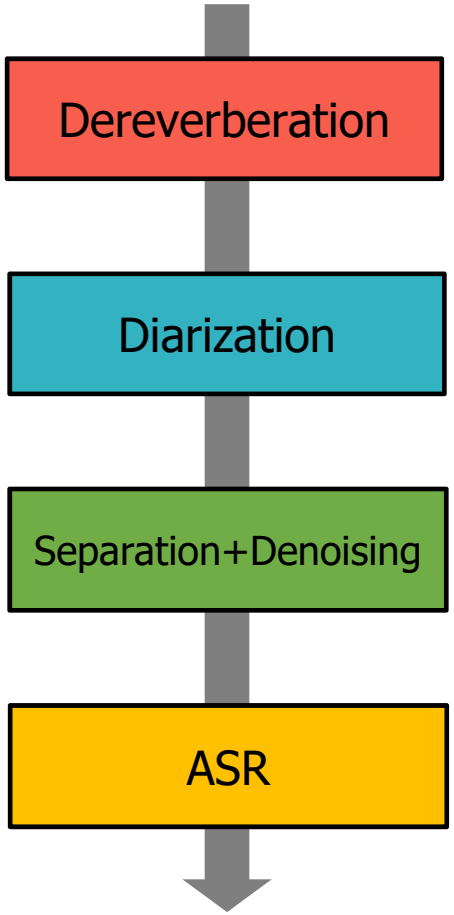
Predict everything

2. Current state-of-the-art systems

2.1. Descriptions of the techniques

2.2. Reproducible baselines

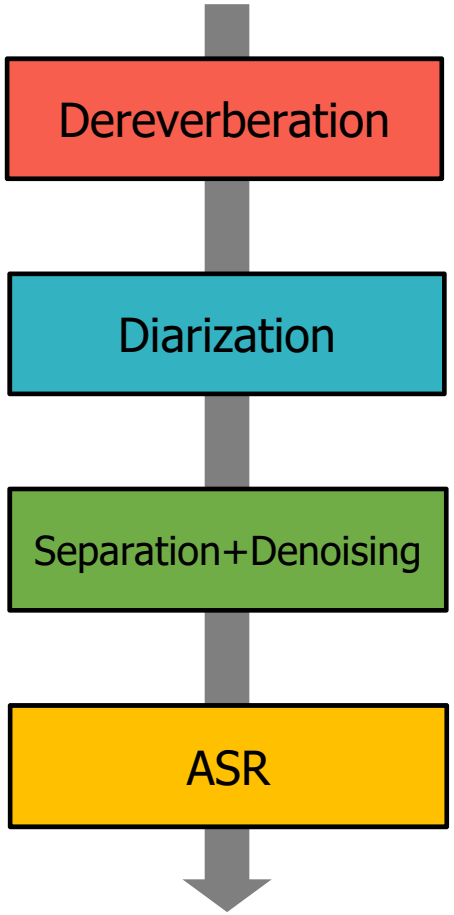
CHiME-6 pipeline



The process gradually fills out the missing RTTM information

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ

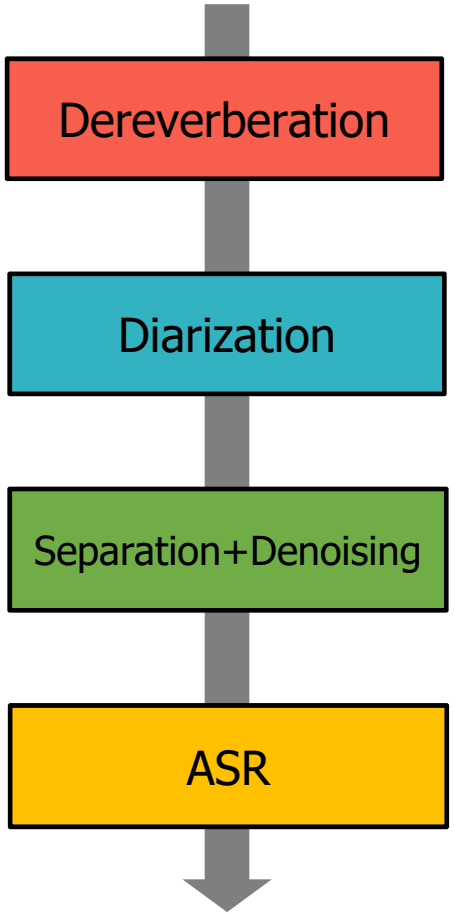
CHiME-6 pipeline



The process gradually fills out the missing RTTM information

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0007400-0007588		74.0	75.88	
0008380-0009435		83.8	94.35	
0031242-0031700		312.42	317.0	

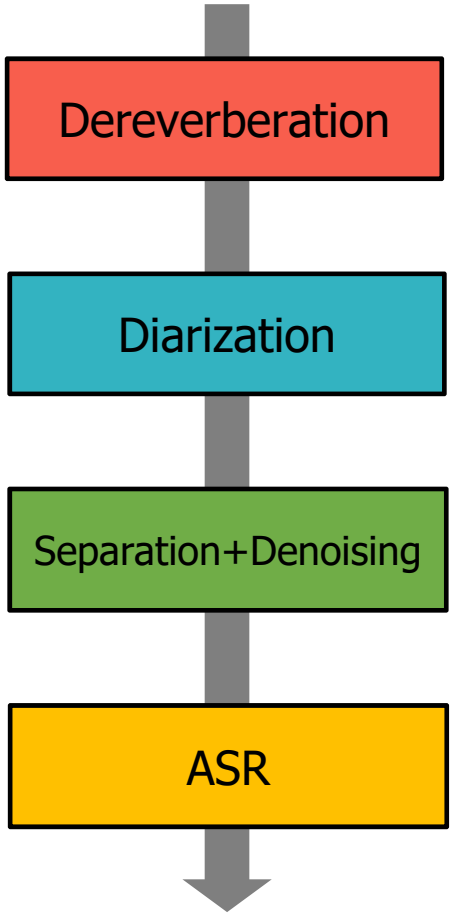
CHiME-6 pipeline



The process gradually fills out the missing RTTM information

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	
A-0008380-0009435	A	83.8	94.35	
B-0031242-0031700	B	312.42	317.0	

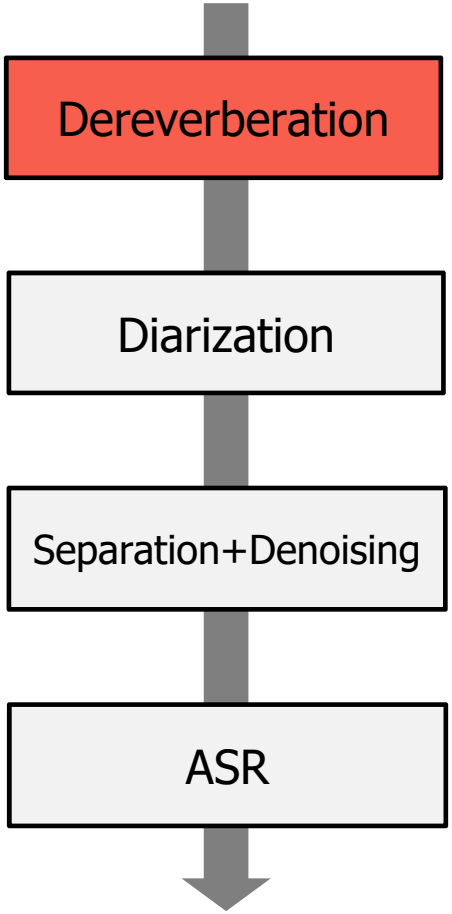
CHiME-6 pipeline



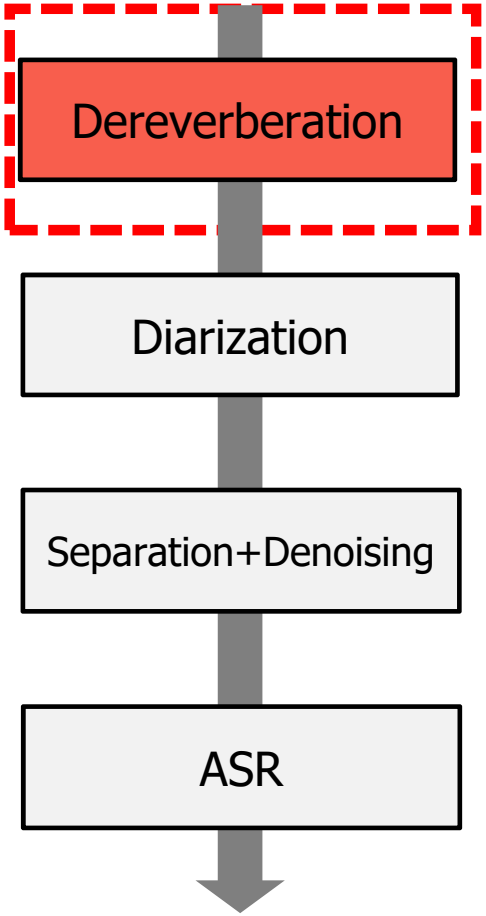
The process gradually fills out the missing RTTM information

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	yes
A-0008380-0009435	A	83.8	94.35	hey did you put…
B-0031242-0031700	B	312.42	317.0	hmm

Speech enhancement

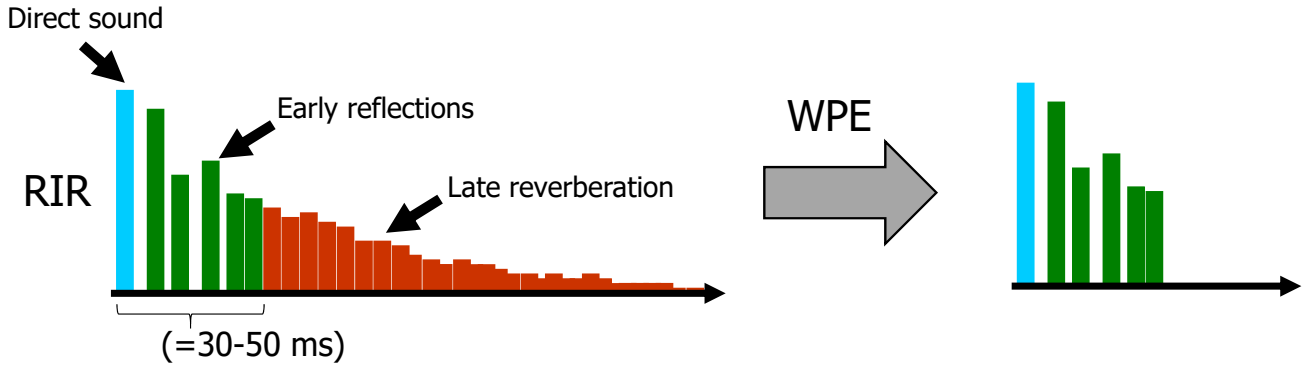


Speech enhancement (Dereverberation) [Nakatani+, 2010][Yoshioka, 2012]

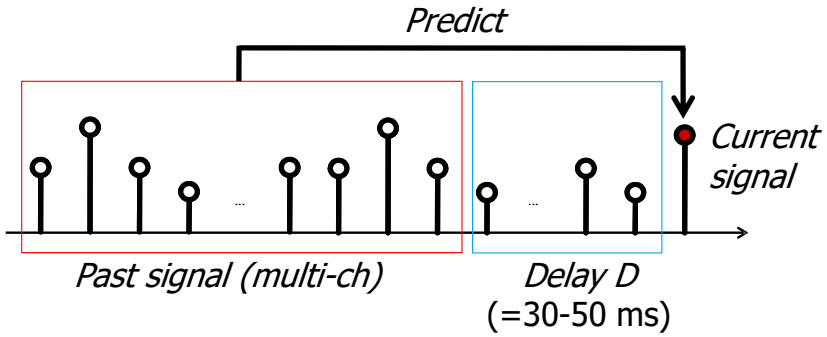


Weighted Prediction Error (WPE)-based dereverberation

-Effect: Reduce late reverberation harmful for ASR, separation, diarization



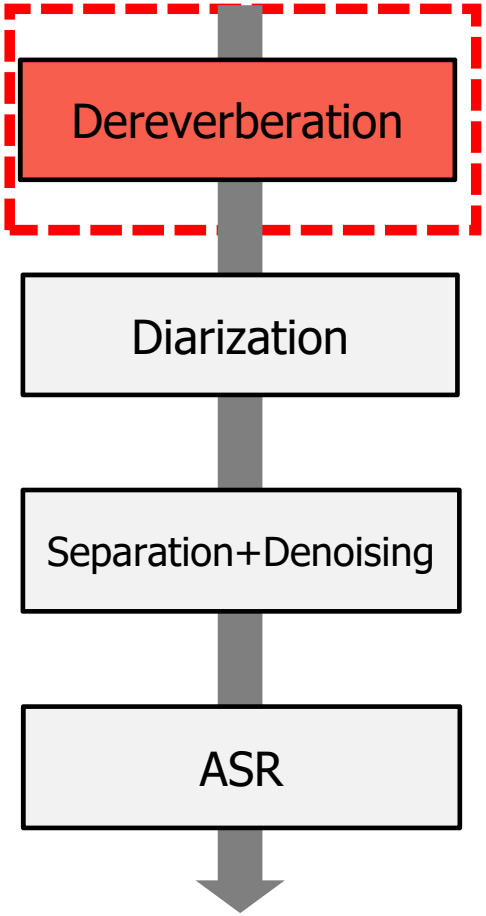
-How to estimate and reduce late reverberation



- Predictable component
  → late reverberation
- Unpredictable component
  → Direct sound

T. Nakatani et al., Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction, TASLP, 2010,
T. Yoshioka et al., Generalization of Multi-Channel Linear Prediction Methods for Blind MIMO Impulse Response Shortening, TASLP, 2012

Speech enhancement (Dereverberation) [Nakatani+, 2010]



Weighted Prediction Error (WPE)-based dereverberation

- How do we estimate WPE filter $\hat{\mathbf{W}}$ from M-channel observed signal $\mathbf{y}_{t,f} = \{y_{m,t,f}\}_{m=1}^{M} \in \mathbb{C}^M$?

→ Minimize the following prediction error with delay D, i.e., dereverberated signal $\hat{\mathbf{s}}_{t,f}$ is represented as

$$\hat{\mathbf{s}}_{t,f} = \mathbf{y}_{t,f} - \sum_{\tau=D}^{L} \hat{\mathbf{W}}_{\tau,f}^{\mathrm{H}} \mathbf{y}_{t-\tau,f}$$

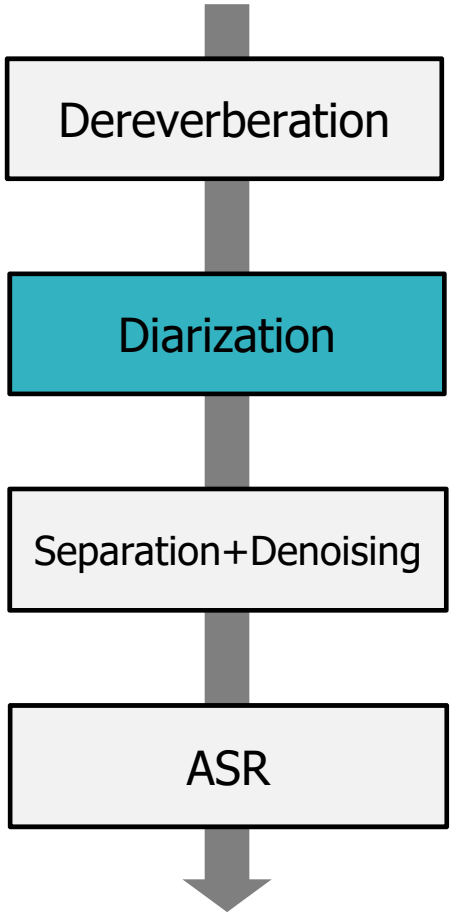
such that $\hat{\mathbf{s}}_{t,f}$ follows time-varying Gaussian distribution

$$p(\mathbf{s}_{t,f}; \theta) = \mathcal{N}_c(\mathbf{s}_{t,f}; 0, \sigma_{t,f}^2 \mathbf{I})$$

$\sigma_{t,f}^2$: power spectrum of ideal dereverberated speech

T. Nakatani et al., Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction, TASLP, 2010,

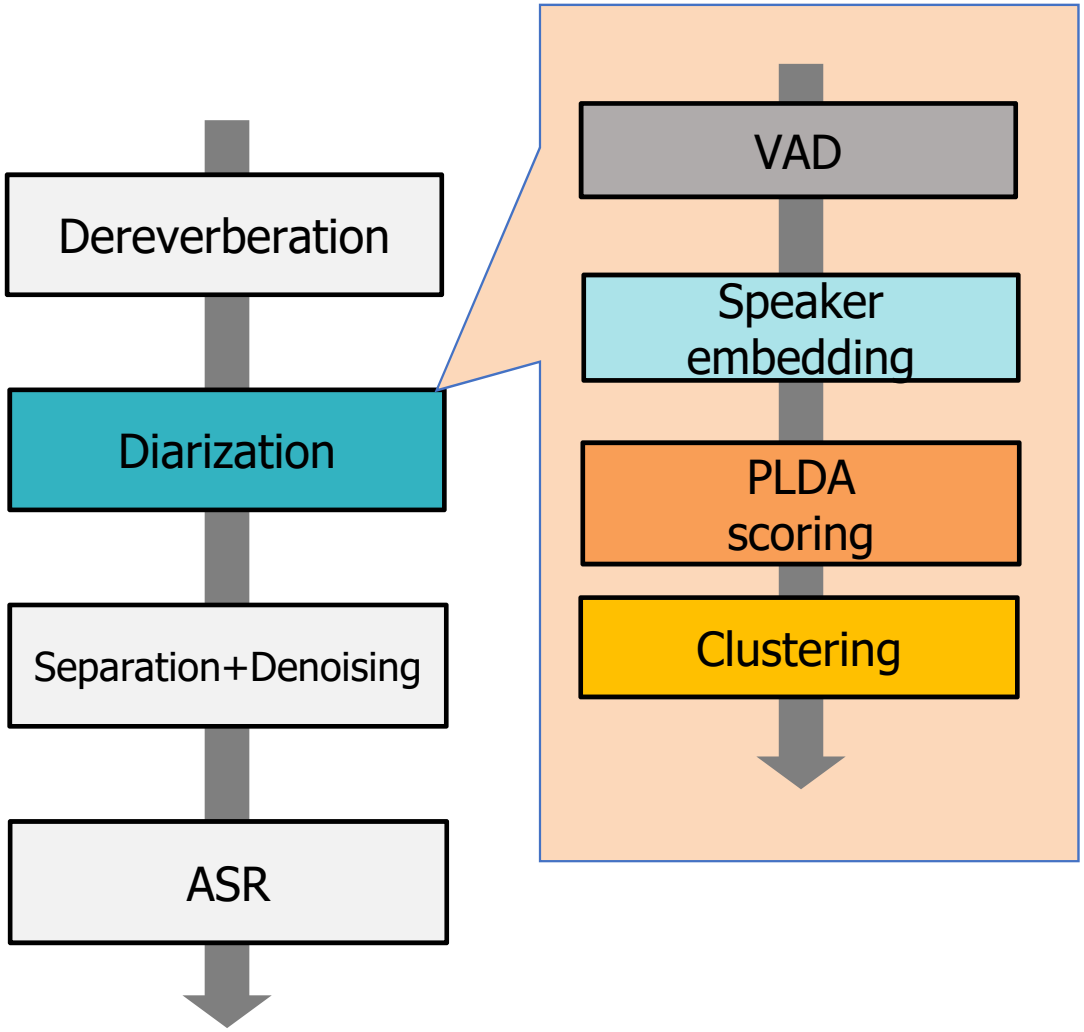
CHiME-6 pipeline



We will introduce two diarization systems

1. **Clustering based diarization**
   - Standard system in the speaker diarization studies
2. **Target-speaker voice activity detection based diarization**
   - Based on CHiME-challenge winning systems
   - Significantly reduce the diarization and ASR errors in the CHiME system

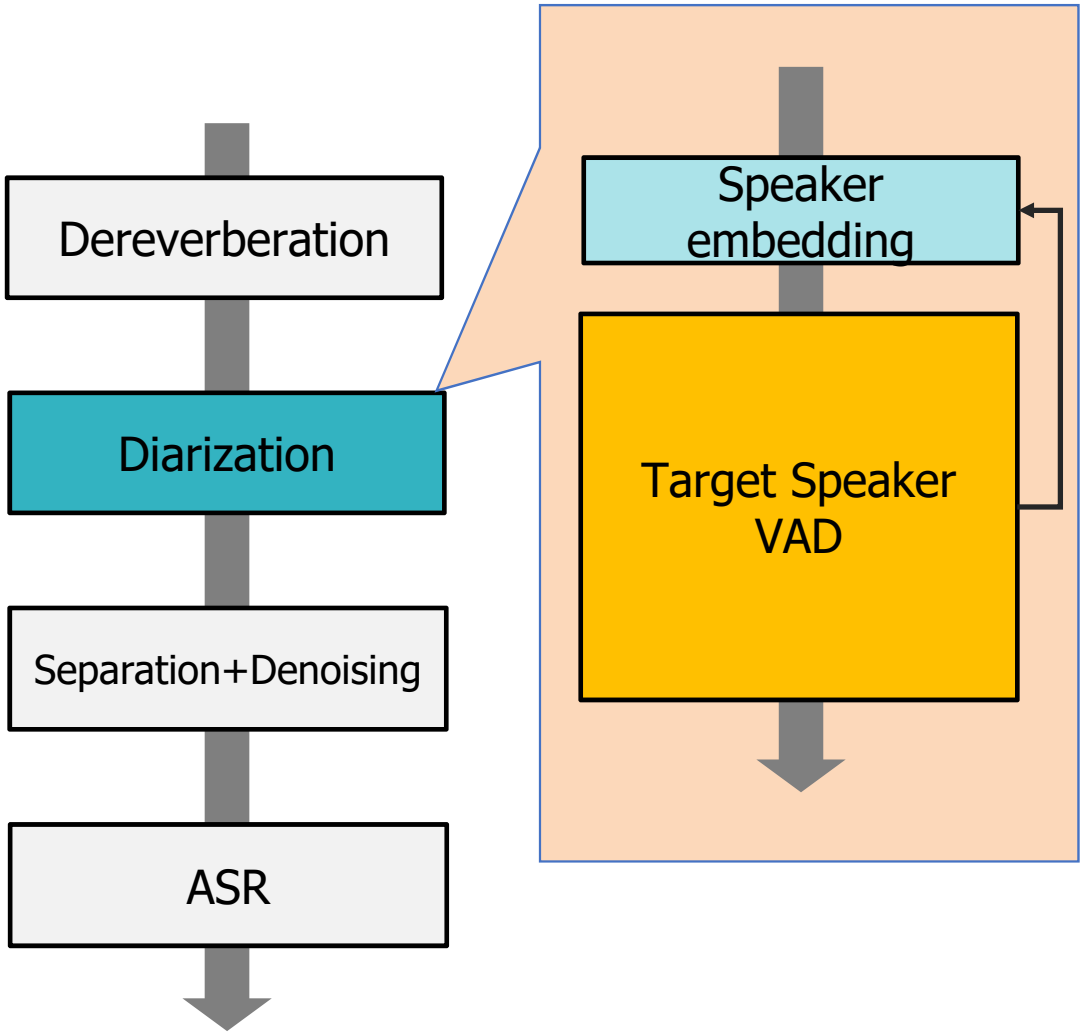
CHiME-6 pipeline



We will introduce two diarization systems

1. **Clustering based diarization**
   - Standard system in the speaker diarization studies
2. **Target-speaker voice activity detection based diarization**
   - Based on CHiME-challenge winning systems
   - Significantly reduce the diarization and ASR errors in the CHiME system

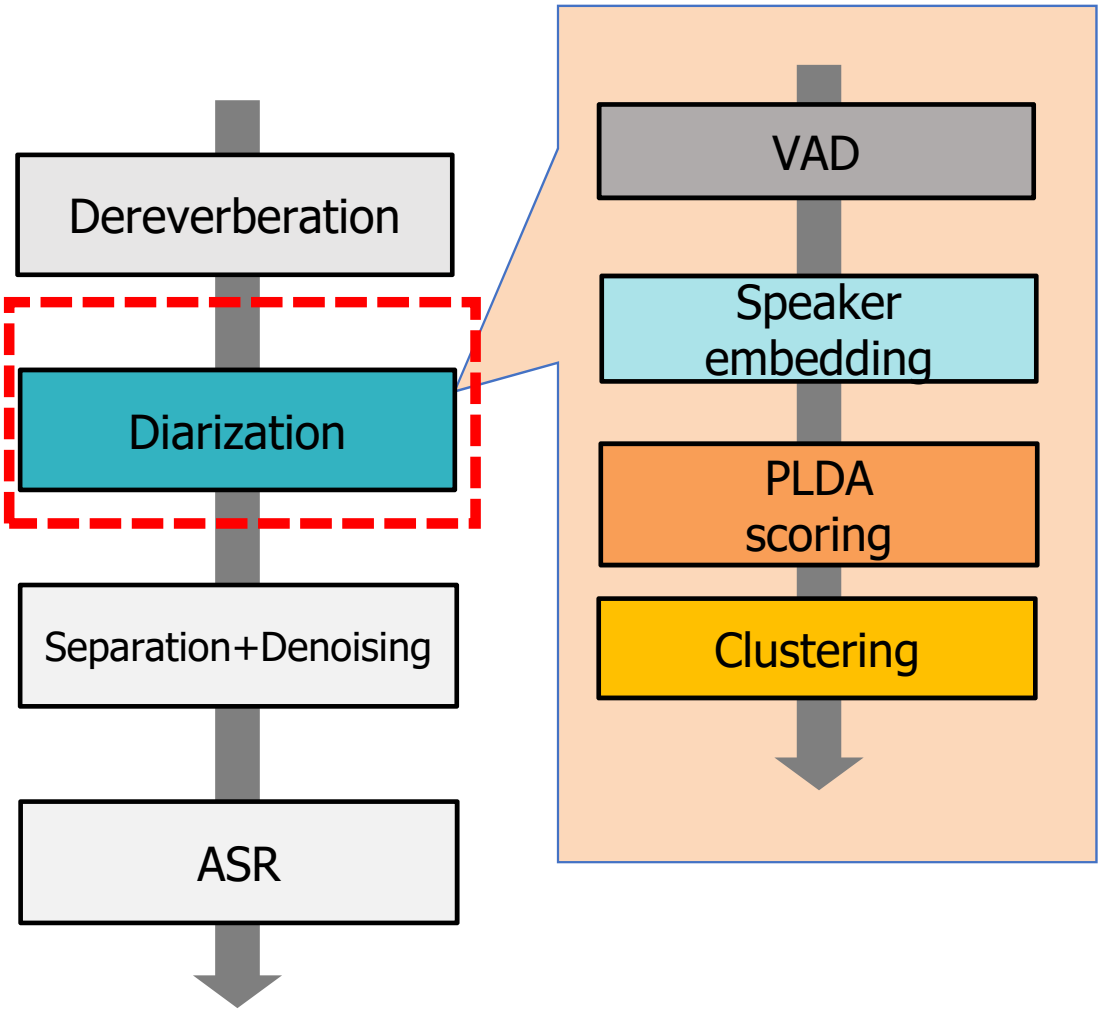
CHiME-6 pipeline



We will introduce two diarization systems

1. **Clustering based diarization**
   - Standard system in the speaker diarization studies
2. **Target-speaker voice activity detection based diarization**
   - Based on CHiME-challenge winning systems
   - Significantly reduce the diarization and ASR errors in the CHiME system

Diarization



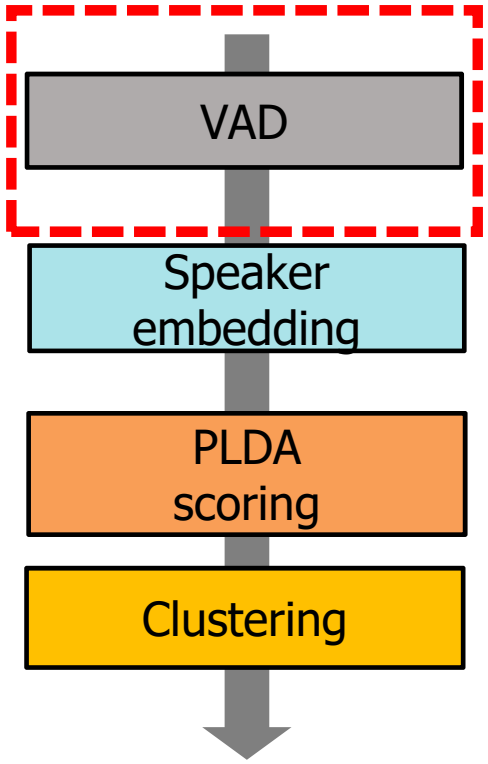
Input: Dereverberated speech
Output: Diarization output (e.g., RTTM)

Clustering based diarization is composed of

- **Voice Activity Detection (VAD)**, also called Speaker Activity Detection (SAD)
  - Generates speech segments
- **Speaker Embedding**
  - Compute a vector for each segment
- **Scoring** based on Probabilistic linear discriminant analysis (**PLDA**)
- **Clustering**

We will briefly discuss each module

VAD module in Diarization



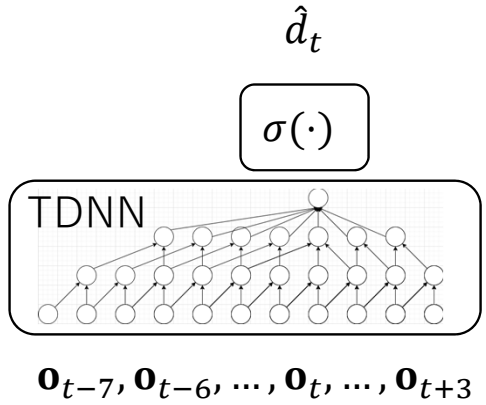
Input: Dereverberated speech
Output: Speech segment

Speech activity
$d_t \in \{0, 1\}$

- Predict speech activity $\hat{d}_t$ based on LSTM, SyncNet or Time-delay neural network (TDNN)

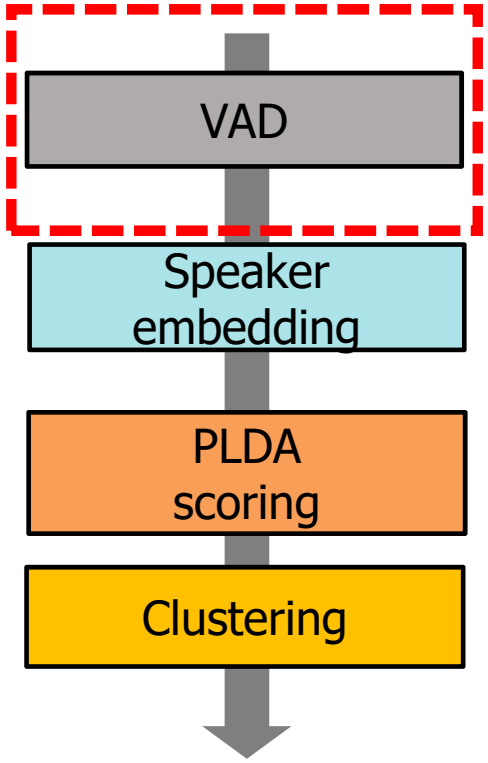
$$\hat{d}_t = \sigma\left(\mathrm{TDNN}(\mathbf{O})\right) \in (0, 1)$$

  - $\mathrm{TDNN}(\cdot)$: Time-delay neural network
  - $\mathbf{O}$: Enhanced speech feature
  - $\sigma(\cdot)$: Sigmoid function
- We use a threshold to obtain $d_t \in \{0, 1\}$

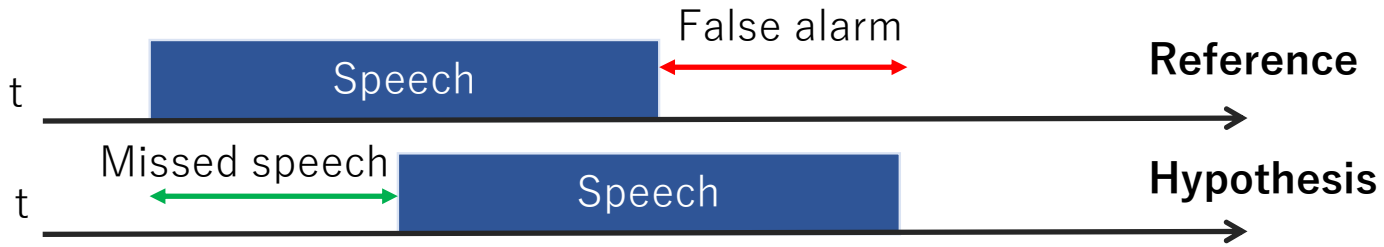


Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0007400-0007588		74.0	75.88	
0008380-0009435		83.8	94.35	

VAD module in Diarization



- VAD's performance metric



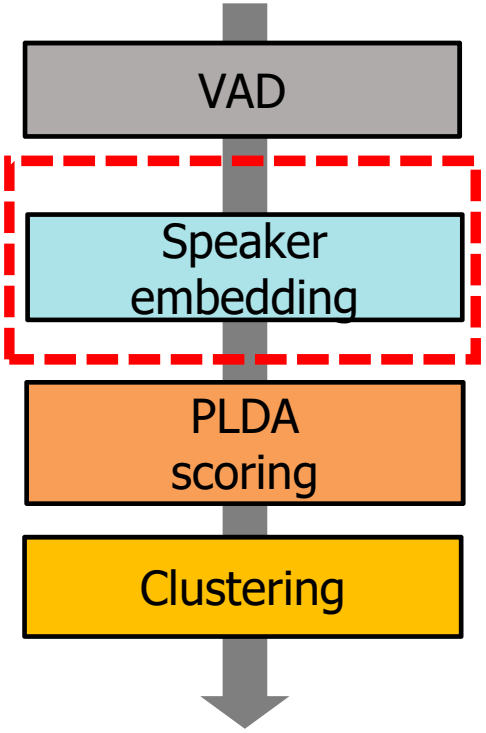
- VAD performance

Dev.			Eval.		
Missed speech	False alarm	Total error	Missed speech	False alarm	Total error
1.0%	1.0%	2.0%	2.0%	3.2%	5.2%

from https://github.com/desh2608/kaldi/blob/demo/egs/chime6/s5b_track2

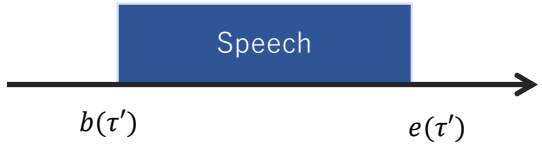
- VAD is fairly working well

Embedding and clustering in Diarization

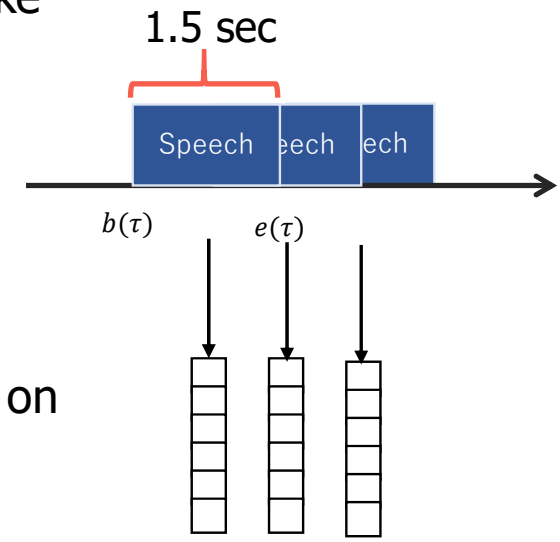


Input: dereverberated speech and VAD segment
Output: embedding vector for each segment

- Original VAD: $\mathbf{O}_{\tau'} = (\mathbf{o}_{t=b(\tau')}, \ldots, \mathbf{o}_{t=e(\tau')})$
  - τ'-th speech segment from VAD
  - $b(\tau')$: onset time, $e(\tau')$: offset time



- Further segment it with every **1.5sec** and make it the same length $e(\tau) - b(\tau) = 1.5$ sec
  - 0.25 sec shift



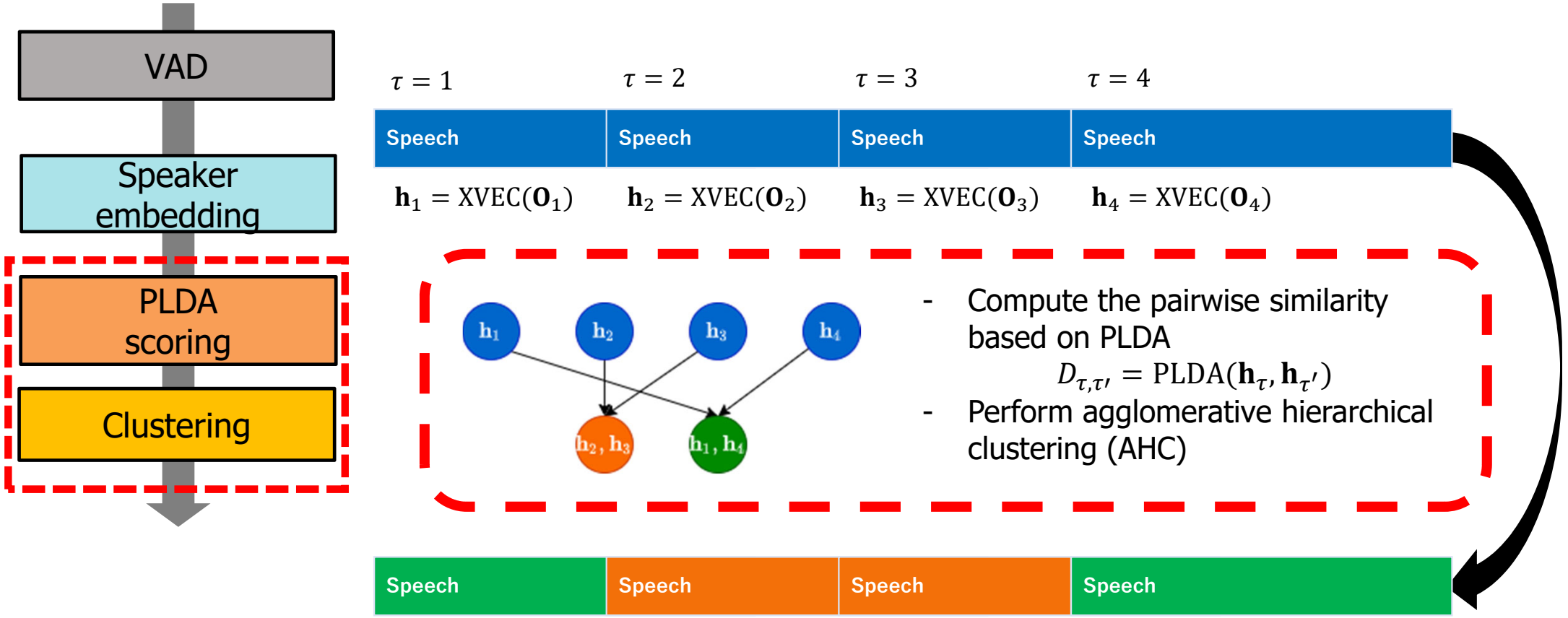
- x-vector $\mathbf{h}_{\tau}^{(\mathrm{spk})}$: Neural speaker embedding $\mathrm{XVEC}(\cdot)$ for segment τ

$$\mathbf{h}_{\tau}^{(\mathrm{spk})} = \mathrm{XVEC}(\mathbf{O}_{\tau}) = \mathrm{StatPool}(\mathrm{TDNN}(\mathbf{O}_{\tau}))$$

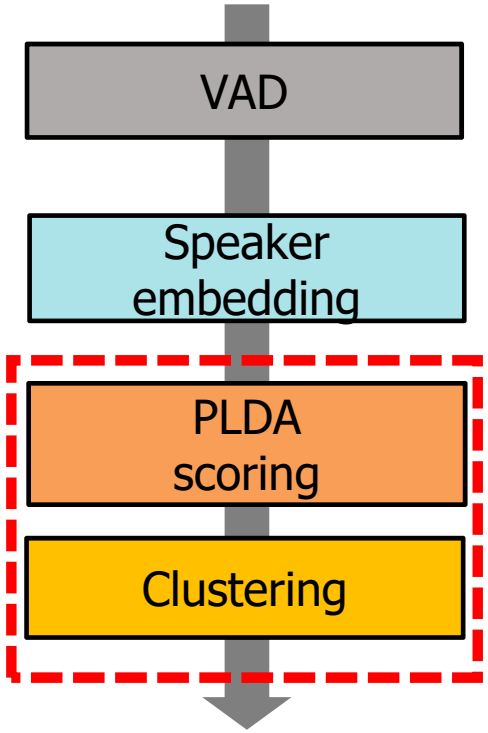
  - $\mathrm{StatPool}(\cdot)$: Statistical pooling layer based on mean and variance operations
  - $\mathrm{TDNN}(\cdot)$: Time delay neural network
  - ResNet is also used

Embedding and clustering in Diarization

Input: embedding vector for each segment based on VAD
Output: speaker diarization output (e.g., RTTM)



Embedding and clustering in Diarization

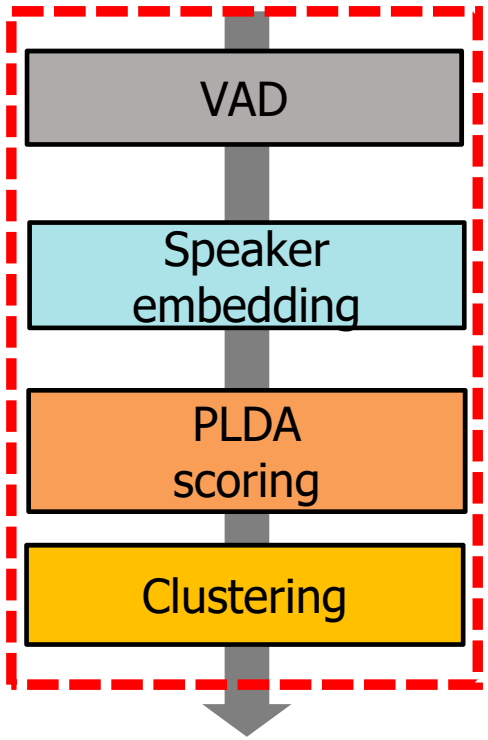


Input: embedding vector for each segment based on VAD
Output: speaker diarization output (e.g., RTTM)

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0007400-0007588		74.0	75.88	
0008380-0009435		83.8	94.35	
0031242-0031700		312.4	317.0	

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	
A-0008380-0009435	A	83.8	94.35	
B-0031242-0031700	B	312.4	317.0	

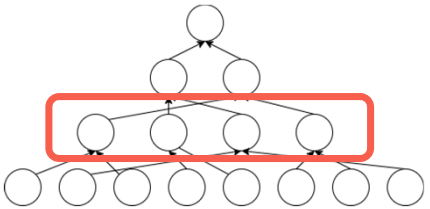
Embedding and clustering in Diarization



- x-vector neural speaker embedding model is trained with VoxCeleb
- PLDA model is trained with CHiME-6
- The number of speakers is **given** (=4)

Dev. DER	Eval. DER
63.4%	68.2%

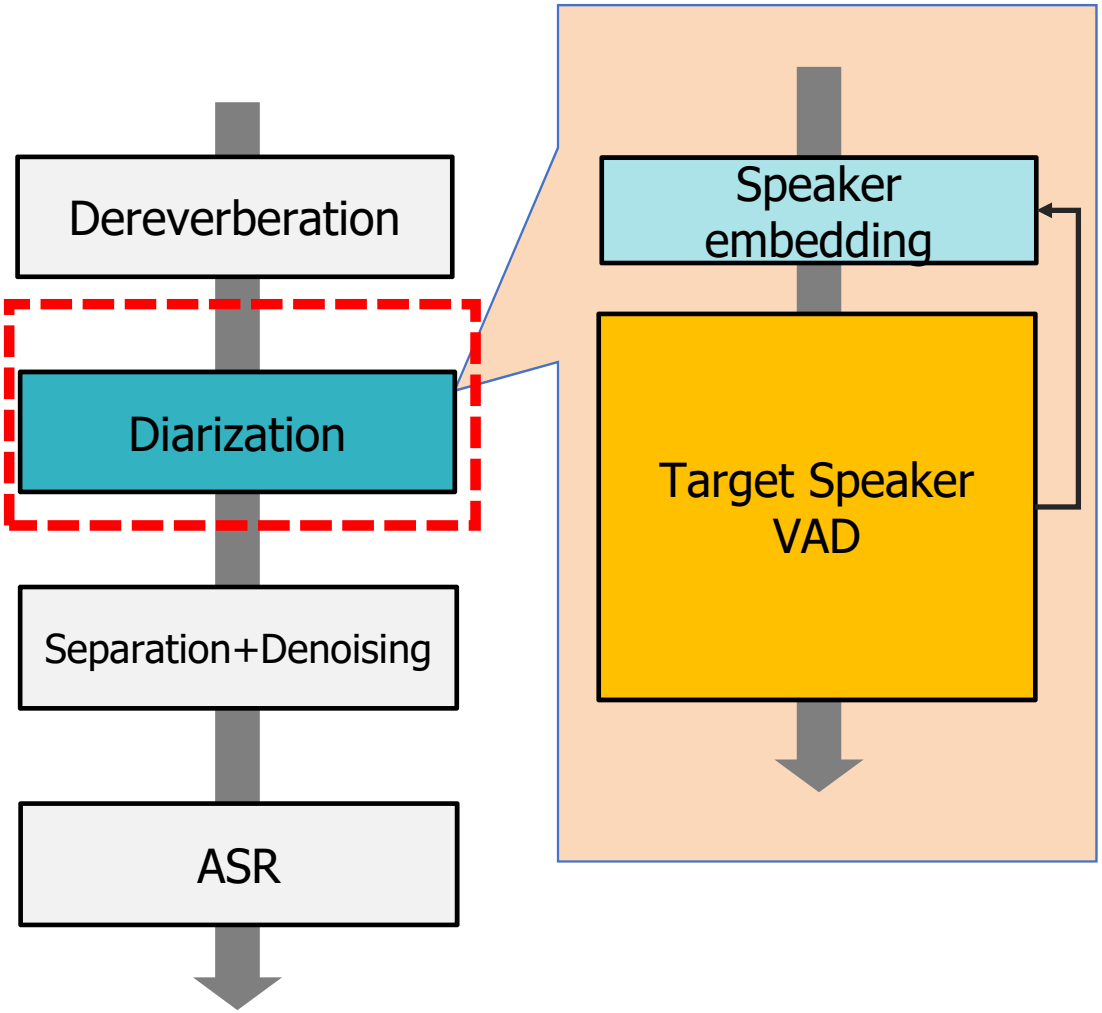
from https://chimechallenge.github.io/chime6/track2_instructions.html



Stop the clustering when #nodes become 4

- Diarization performance was poor in the CHiME scenario due to the large speaker overlap
- However, this shows very strong performance in the other scenarios including DIHARD

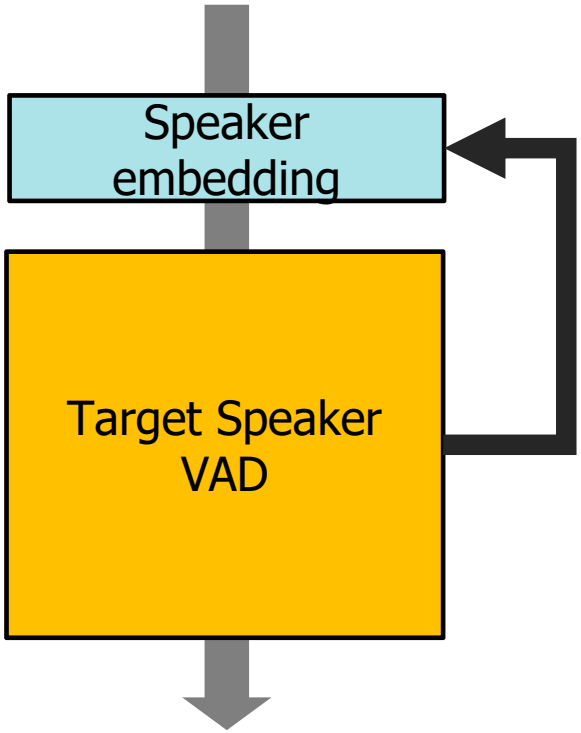
Diarization based on target speaker VAD



- Embedding vector extraction for each speaker
  - Compute an **i-vector** for each segment
- Target-speaker voice activity detection
  - **Multi-label classification** problems for each speaker
  - **Overlap can be handled** by joint label activation
  - Inspired by End-to-End Neural Diarization (c.f. Section 3.1)
  - i-vector input act as speaker-aware (target speaker) activity detection
- Performed **iteratively**

Medennikov et al., “The STC System for the CHiME-6 Challenge,” CHiME 2020

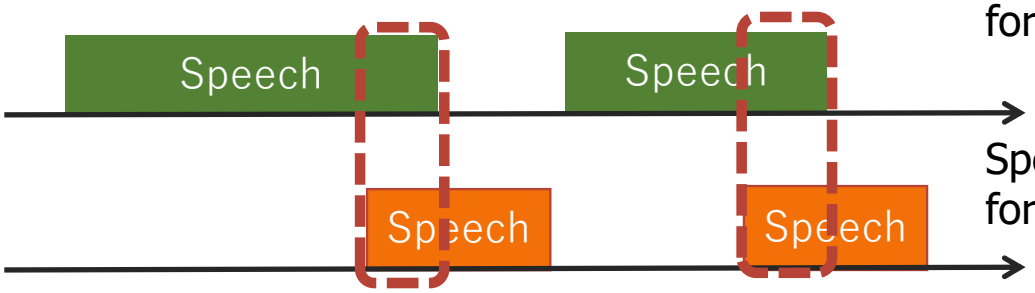
Diarization based on target speaker VAD



Target-speaker voice activity detection

- **Multi-label classification** problems for each speaker, i.e.,

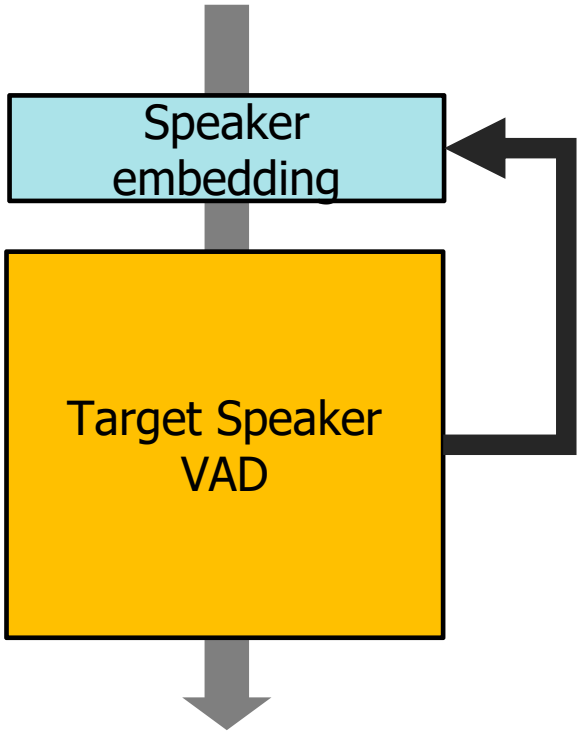
$$p\left(\{d_{k,t}\}_{k=1}^{K} | \mathbf{O}, \{\mathbf{h}_k\}_{k=1}^{K}\right)$$



- **Naturally handles overlap**, e.g., $d_{k,t} = d_{k't} = 1$ at frame t
- Note that this is very similar to VAD, i.e., $p(d_t|\mathbf{O})$ except that it is conditioned on the speaker-dependent variable

Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

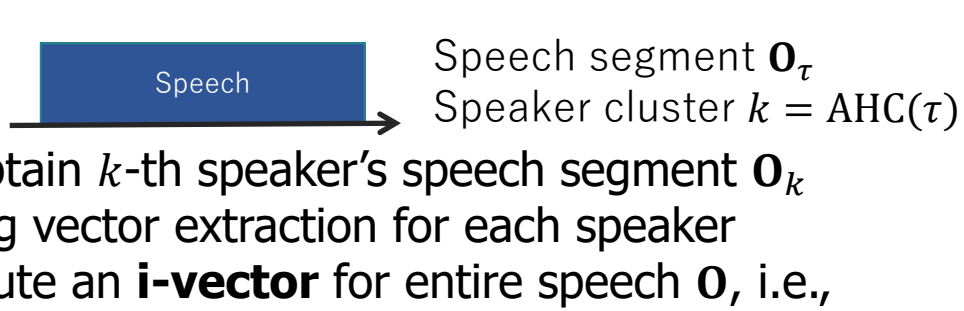
Diarization based on target speaker VAD



- How to make a speaker-dependent classifier $p\left(\{d_{k,t}\}_{k=1}^{K}|\mathbf{O},\{\mathbf{h}_k\}_{k=1}^{K}\right)$?

  We use speaker embedding vector $\mathbf{h}_k$

- From the **speaker diarization information**, we can obtain the speaker cluster index $k = \mathrm{AHC}(\tau)$ for each segment

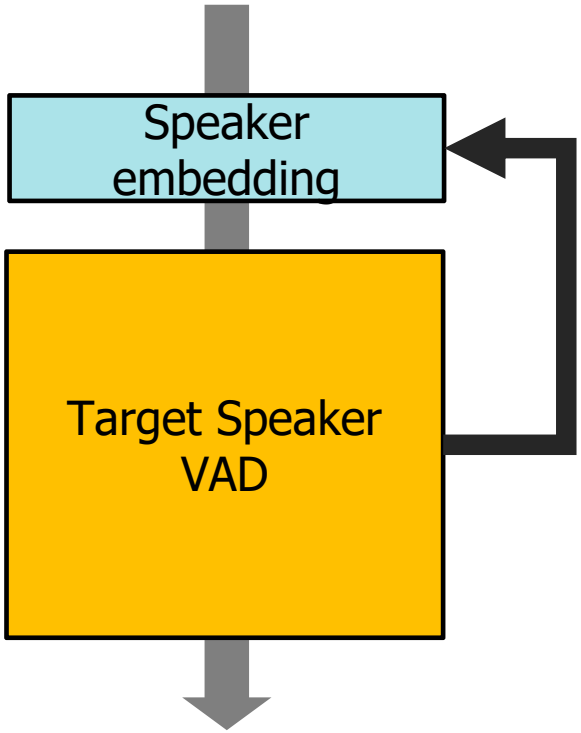


- We can obtain k-th speaker's speech segment $\mathbf{O}_k$
- Embedding vector extraction for each speaker
  - Compute an **i-vector** for entire speech $\mathbf{O}$, i.e.,

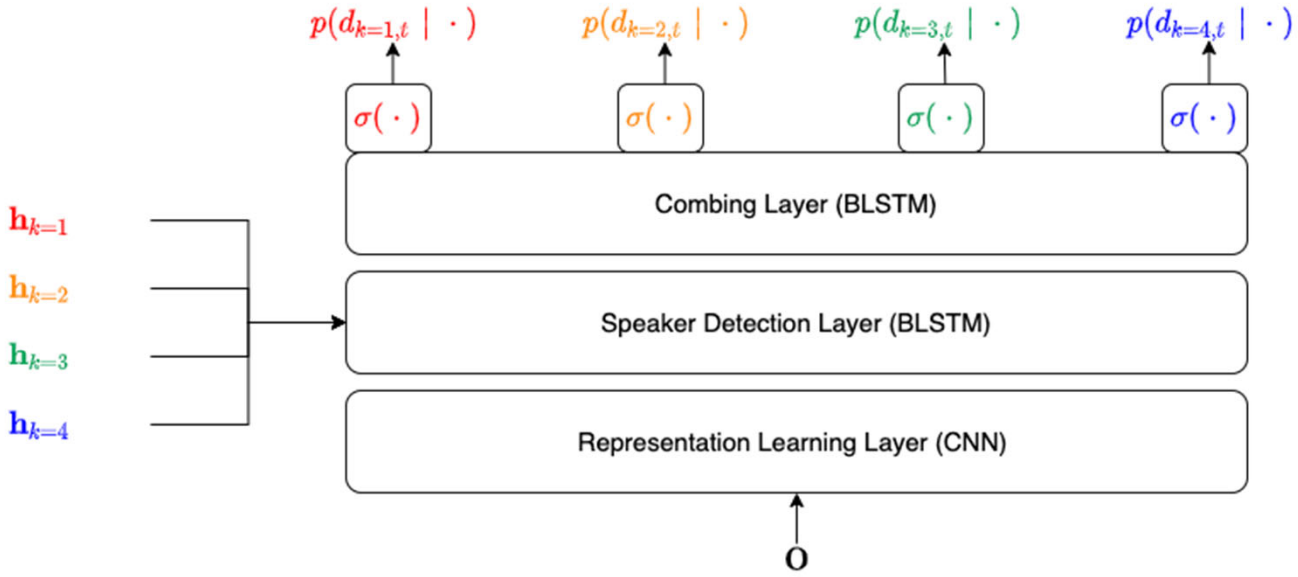
$$\mathbf{h}_k = \mathrm{IVEC}(\mathbf{O}_k)$$

Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

Diarization based on target speaker VAD

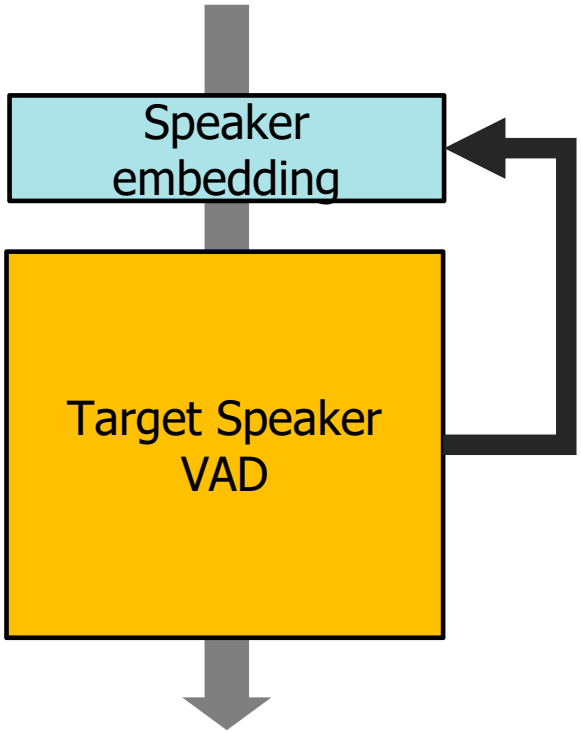


- **Multi-label classification** $p\left(\{d_{k,t}\}_{k=1}^{K}|\mathbf{O},\{\mathbf{h}_k\}_{k=1}^{K}\right)$ is obtained by feeding all speaker embedding information in addition to MFCC features $\mathbf{O}$



Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

Diarization based on target speaker VAD



Iterative processing

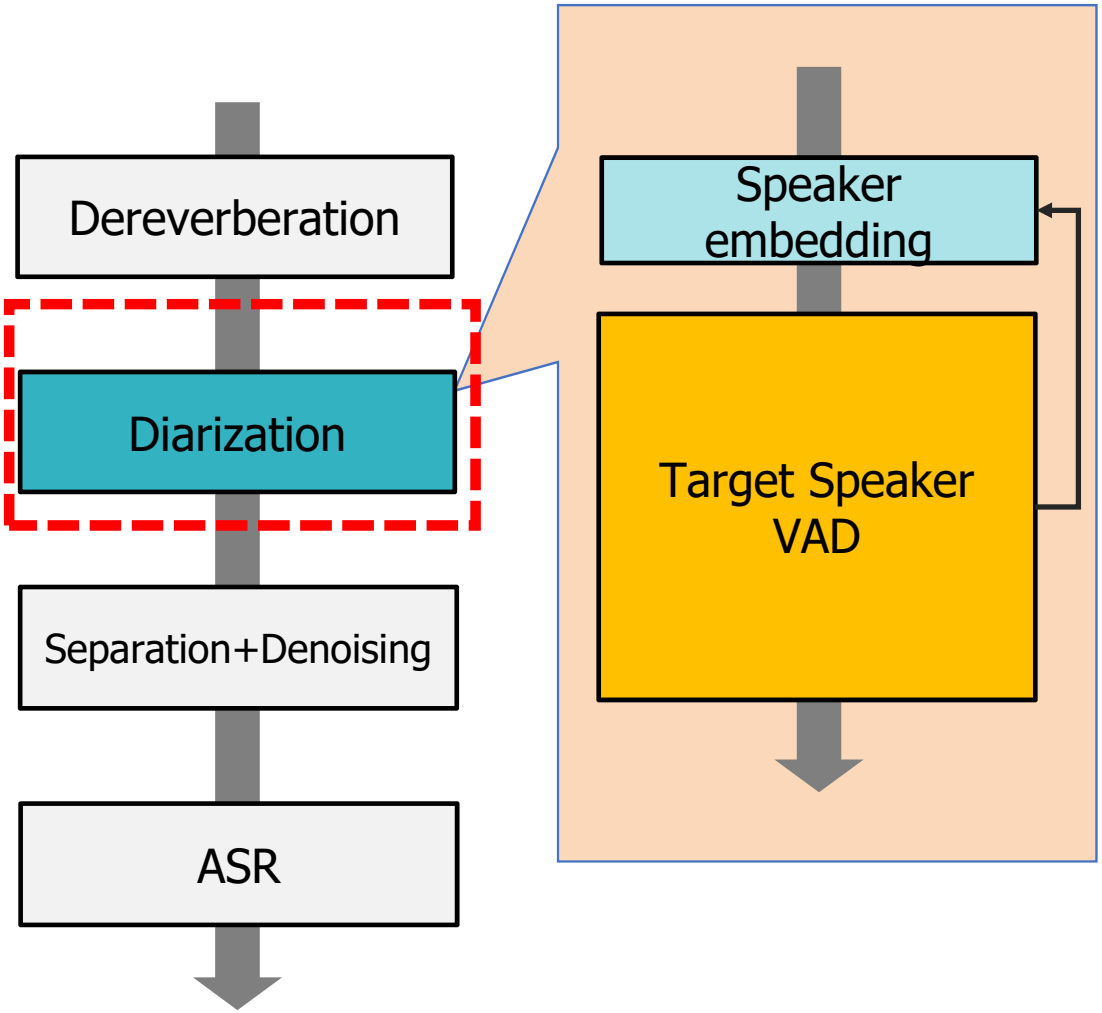
TS-VAD $p\left(\{d_{k,t}\}_{k=1}^{K}|\mathbf{O},\{\mathbf{h}_k\}_{k=1}^{K}\right)$

Speaker embedding $\mathbf{h}_k = \mathrm{IVEC}(\mathbf{O}_k)$

- Note that the initial diarization result is very critical (e.g., x-vector based on ResNet and spectral clustering)

Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

Diarization based on target speaker VAD



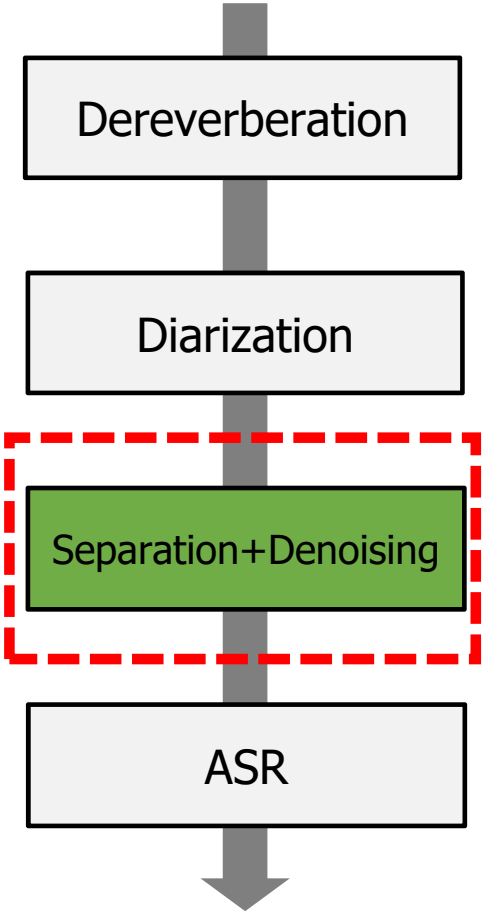
	Dev.	Eval.
	DER	DER
X-vector clustering	63.4%	68.2%
TS-VAD	32.8%	36.0%

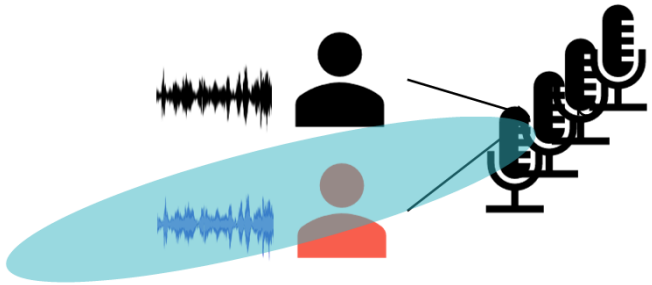
from Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

- Great improvements based on the explicit overlap handling and iterative procedure

Speech enhancement (Separation+Denoising)

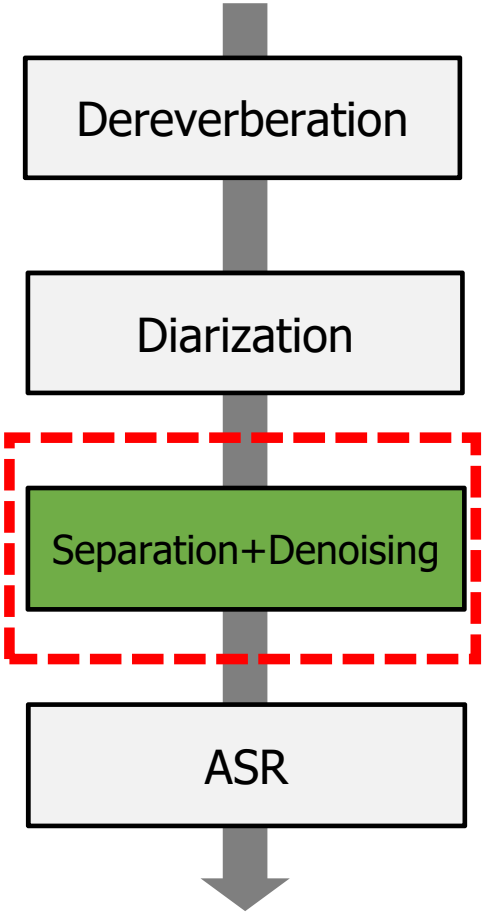
- Beamforming





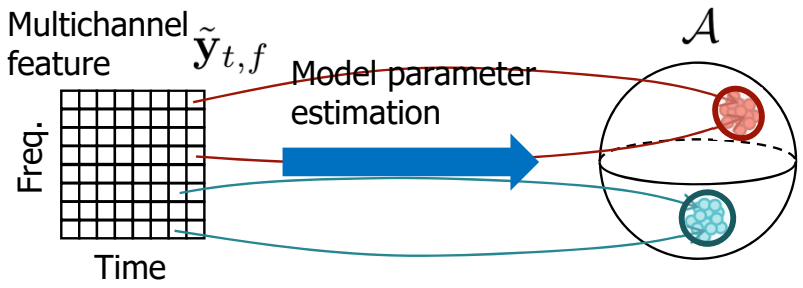
- Only extract a source based on the statistics of the source
- Time-frequency (TF) mask is used to estimate the statistics

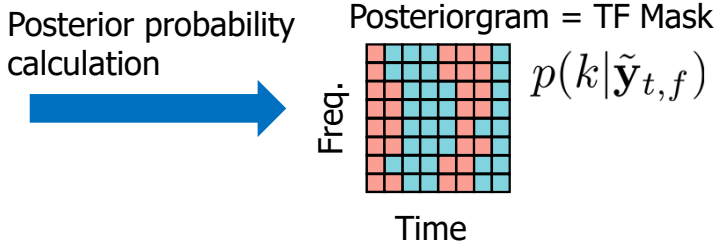
Speech enhancement (Separation+Denoising) [Ito+, 2016] [Boedekker+, 2018]



- Mask-based BF = Mask estimation + BF
  - TF Mask estimation: complex Angular Central (cAC)-GMM clustering
  - BF: Minimum variance distortionless response (MVDR) BF

- cAC-GMM clustering can provide TF-mask $p(k|\tilde{\mathbf{y}}_{t,f})$





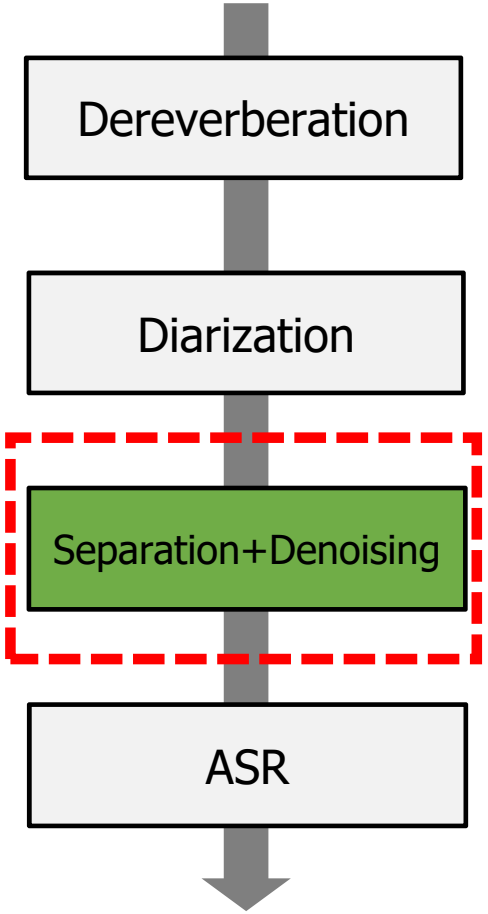
Step1:Parameter estimation of a mixture model
e.g., complex angular central GMM $\mathcal{A}$

$$p(\tilde{\mathbf{y}}_{t,f};\theta_f) = \sum_{k=1}^{K} \pi_{f,k}\mathcal{A}(\tilde{\mathbf{y}}_{t,f};\mathbf{B}_{f,k})$$

Step2: Calculate a mask for k-th speaker
as the posterior probability of k-th component:

$$p(k|\tilde{\mathbf{y}}_{t,f}) = \frac{\pi_{f,k}\mathcal{A}(\tilde{\mathbf{y}}_{t,f};\mathbf{B}_{f,k})}{\sum_{k=1}^{K} \pi_{f,k}\mathcal{A}(\tilde{\mathbf{y}}_{t,f};\mathbf{B}_{f,k})},$$

Speech enhancement (Separation+Denoising)

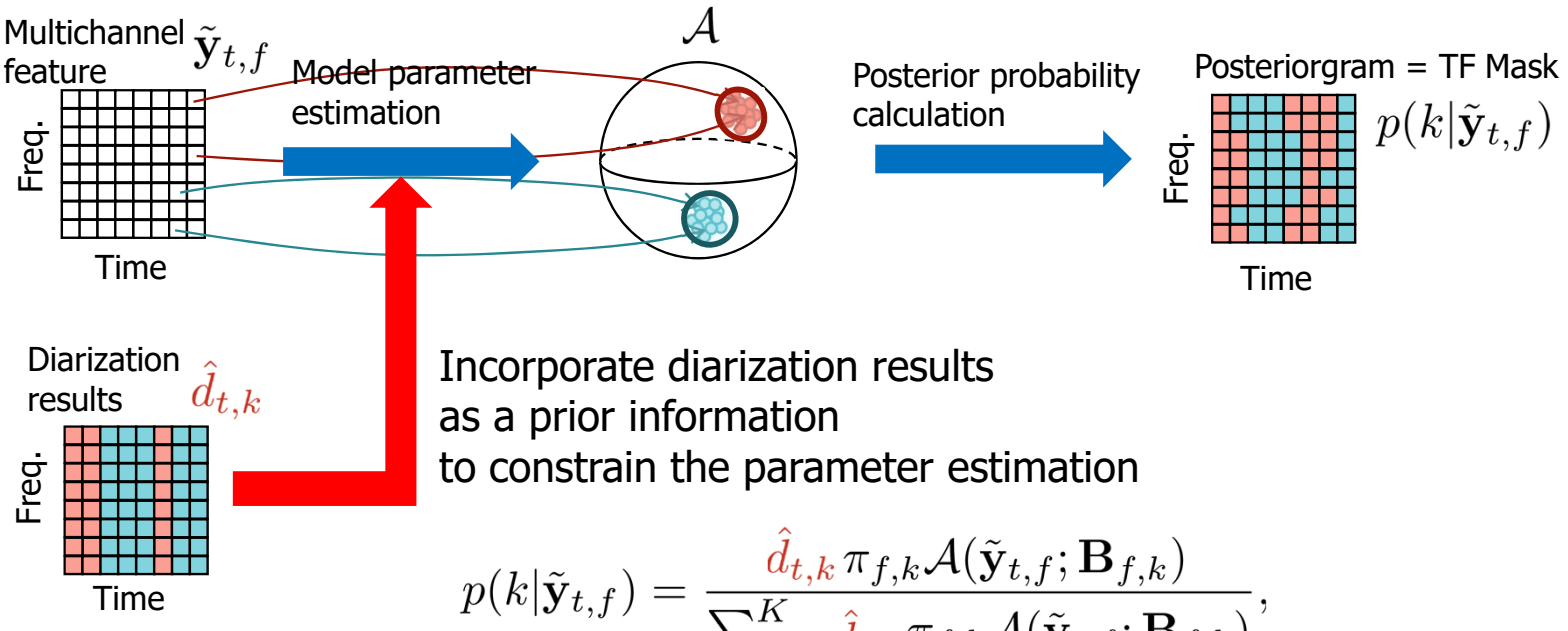


[Boedekker+, 2018]

- Mask-based BF = Mask estimation + BF
  - TF Mask estimation: complex Angular Central (cAC)-GMM clustering
  - BF: Minimum variance distortionless response (MVDR) BF

- How to incorporate diarization results into cAC-GMM clustering

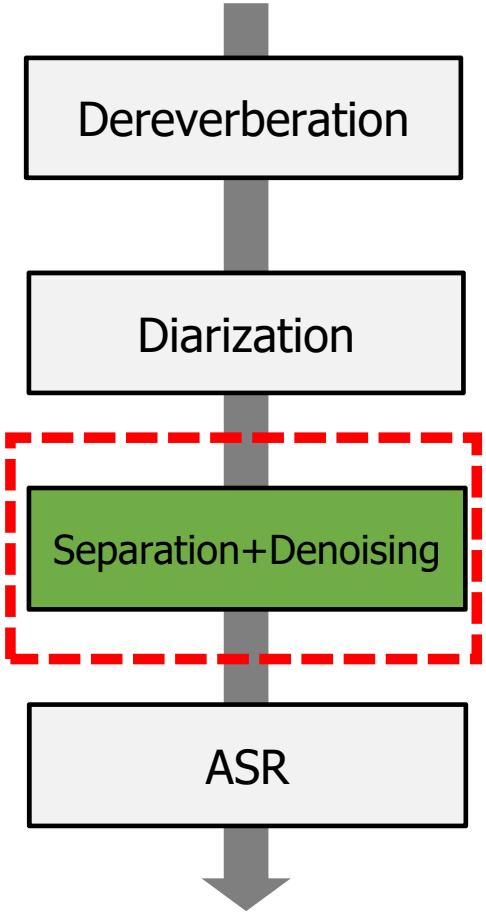
[Boedekker+, 2018]



$$p(k|\tilde{\mathbf{y}}_{t,f}) = \frac{\hat{d}_{t,k}\,\pi_{f,k}\mathcal{A}(\tilde{\mathbf{y}}_{t,f};\mathbf{B}_{f,k})}{\sum_{k=1}^{K}\hat{d}_{t,k}\,\pi_{f,k}\mathcal{A}(\tilde{\mathbf{y}}_{t,f};\mathbf{B}_{f,k})},$$

Speech enhancement (Separation+Denoising)

[Boedekker+, 2018]



- Mask-based BF = Mask estimation + BF
  - TF Mask estimation: complex Angular Central (cAC)-GMM clustering
  - BF: Minimum variance distortionless response (MVDR) BF

- How to calculate a BF filter based on the obtained mask $\hat{\mathbf{m}}_{t,f,k} = p(k|\tilde{\mathbf{y}}_{t,f})$?

Step1: Spatial covariance estimation based on the mask

$$\mathbf{\Phi}_{f,k} = \frac{1}{\sum_{t=1}^{T} \hat{m}_{t,f,k}} \sum_{t=1}^{T} \hat{m}_{t,f,k} \mathbf{y}_{t,f} \mathbf{y}_{t,f}^{H}$$

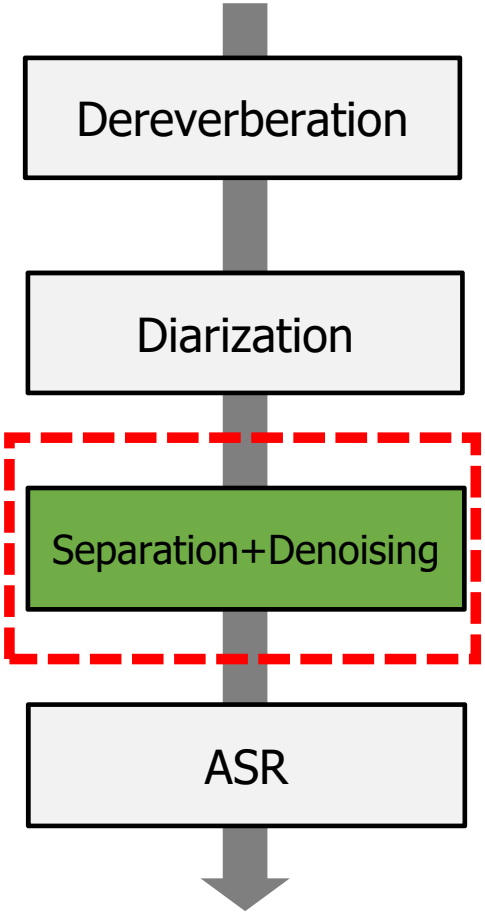
Step2: Filter calculation

$$\mathbf{g}_{f,k} = \frac{(\sum_{k' \neq k} \mathbf{\Phi}_{f,k'})^{-1} \mathbf{\Phi}_{f,k}}{\mathrm{Tr}((\sum_{k' \neq k} \mathbf{\Phi}_{f,k'})^{-1} \mathbf{\Phi}_{f,k})} \mathbf{u}$$

Step3: Filtering

$$\hat{x}_{t,f,k} = \mathbf{g}_{f,k}{}^{H} \mathbf{y}_{t,f}$$

Speech enhancement (Separation+Denoising)



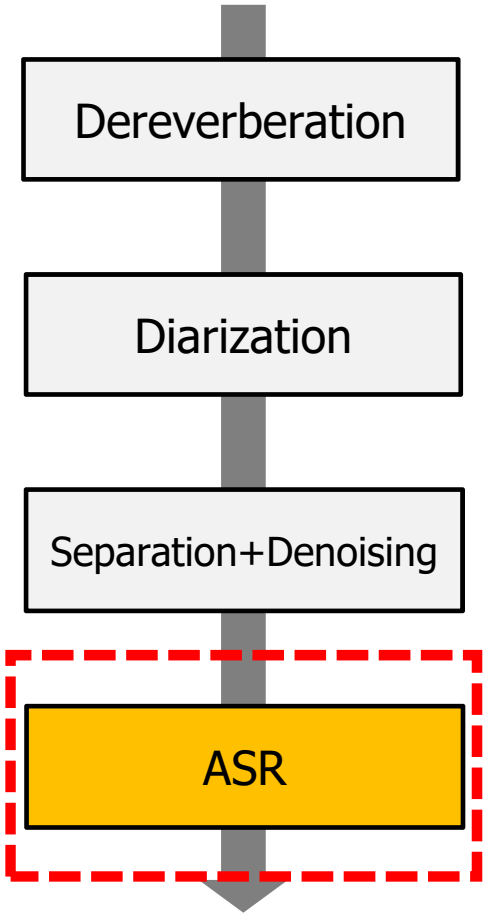
Guided source separation significantly improves the ASR performance

	Dev. SWER	Eval. SWER
w/o GSS	76.7%	72.6%
w/ GSS	**70.3%**	**69.0%**

from https://github.com/desh2608/kaldi/blob/demo/egs/chime6/s5b_track2

- This shows the effectiveness of the separation and denoising.

ASR



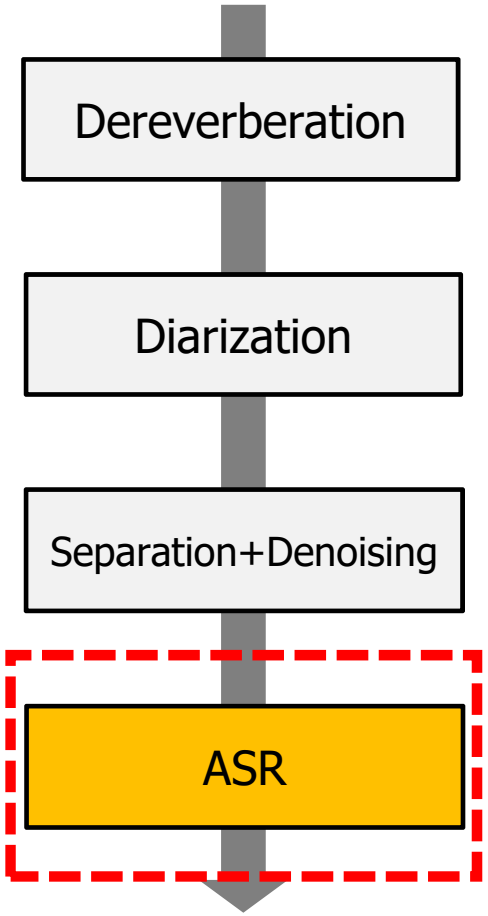
Input: Speaker diarization and enhanced speech for each speaker
Output: Transcriptions for each speaker at each segment

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	
A-0008380-0009435	A	83.8	94.35	
B-0031242-0031700	B	312.42	317.0	

Given the diarization process, it becomes a standard ASR task
In this part, we will explain a DNN-HMM hybrid system

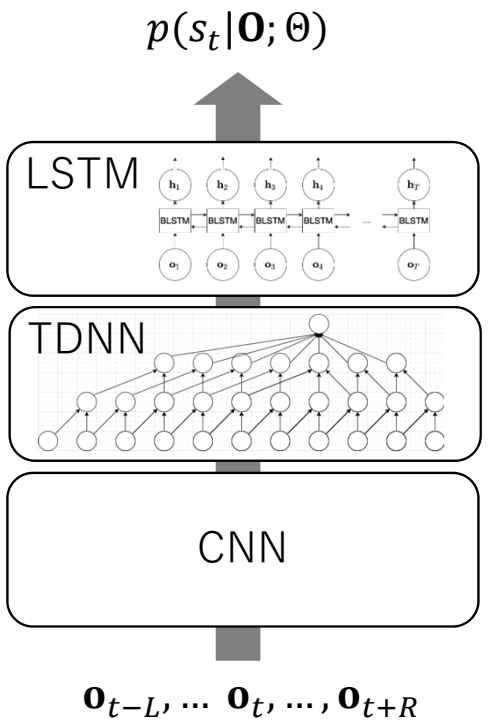
Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
A-0007400-0007588	A	74.0	75.88	yes
A-0008380-0009435	A	83.8	94.35	hey did you put…
B-0031242-0031700	B	312.42	317.0	hmm

ASR: Acoustic model

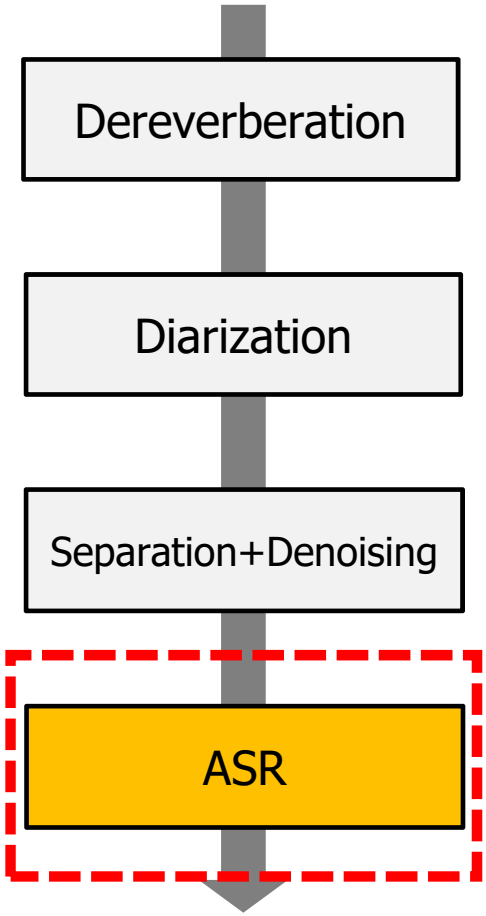


Acoustic model: CNN + TDNN + LSTM

- **Acoustic posterior** $p(s_t|\mathbf{O};\Theta)$ for HMM state s_t

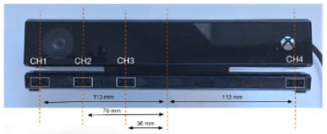


ASR: Acoustic model training



Acoustic model:

- **Four types of data**
  (totally 675 hours)
  1. Far-field array microphones $\mathbf{O}^{\text{kinnect}}$



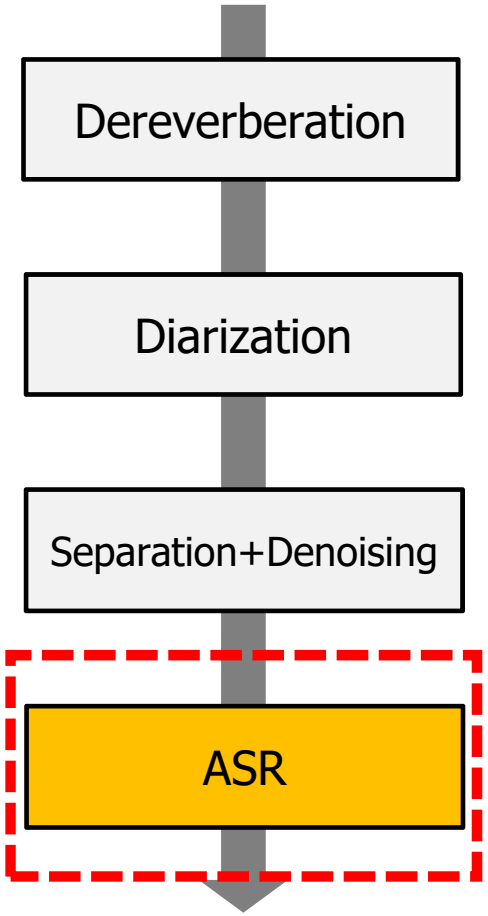
  2. Near-field (worn) microphones $\mathbf{O}^{\text{worn}}$

  3. Augmented data with noises and simulated RIRs
     $\mathbf{X}^{\text{nr}} = \text{RIR}(\mathbf{X}^{worn}) + \text{Noise}$
     $\text{RIR}(\cdot)$ is sampled from simulated RIRs
     Noise is sampled from CHiME-6 noise data
  4. Augmented data with enhanced signals obtained with GSS $\mathbf{O}^{\text{gss}}$

- **Lattice-free MMI-based loss**

$$\Theta^{\text{LF-MMI}} = \underset{\Theta}{\text{argmin}}\, \mathcal{L}(\mathbf{O}^{\text{kinnect}}, \mathbf{O}^{\text{worn}}, \mathbf{O}^{\text{nr}}, \mathbf{O}^{\text{gss}}; \Theta)$$

ASR results



	Diarization	Dev. SWER	Eval. SWER
Original baseline	Oracle	51.8%	51.3%
USTC system	Oracle	30.8%	30.5%
	Diarization	**Dev. WER**	**Eval. WER**
Original baseline	X-vector	84.3%	77.9%
STC system	TS-VAD	39.6%	42.7%

from Du et al, "The USTC-NELSLIP Systems for CHiME-6 Challenge," CHiME 2020
and Medennikov et al., "The STC System for the CHiME-6 Challenge," CHiME 2020

- Significantly improves the baseline by combining the all above techniques
- TS-VAD further improves the speaker attributed WER (SWER)
- Note that these CHiME challenge systems are based on iterative processing, intensive data augmentation, and system combinations, and quite complicated

2. Current state-of-the-art systems

2.1. Descriptions of the techniques

2.2. Reproducible baselines

Reproducible baselines

- **Reproducible (all open source) baseline**
  - Includes all modules, e.g., speech activity detection (SAD), speaker embedding, and speaker diarization in addition to ASR
  - **All-in-one recipe** including training and inference
- But maintain the **simplicity**
  - Excludes iterative processing, intensive data augmentation, and system combinations
  - Unfortunately, this will lead degradations from the state-of-the-art systems in the previous section (42.7% → 60.3%)

- This is one of the first baselines that integrate all multi speaker speech processing in this real scenario ☺

CHiME-6 command line demo

https://github.com/desh2608/kaldi/blob/demo/egs/chime6/s5b_track2/run_demo.sh
Special thanks to Desh Raj Ashish Arora, Aswin Subramanian, and Sanjeev Khudanpur

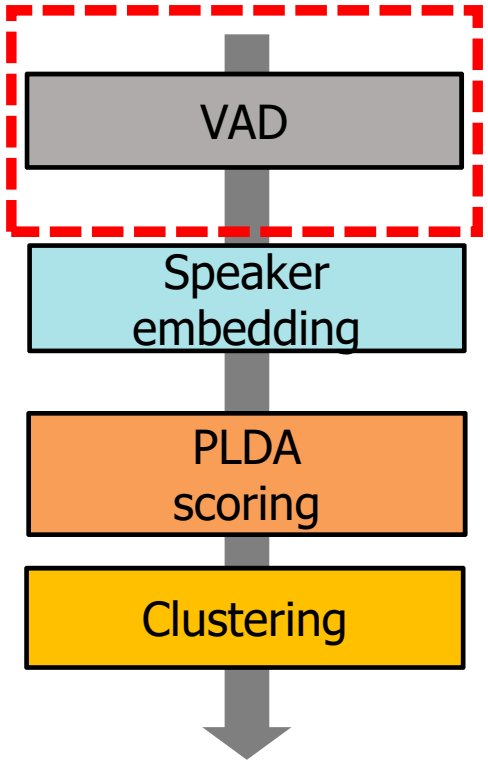
We have three systems in the Kaldi repository

- `s5_track2` Original Kaldi CHiME-6 baseline
- `s5b_track2` **Latest Kaldi system (x-vector clustering) ← mainly introduce this**
- `s5c_track2` **Latest Kaldi system (TS-VAD by STC)** thanks to help from the STC-innovation researchers

```
cd <your_kaldi>/kaldi/egs/chime6/s5b_track2
./run_demo.sh --stage-name prep
```

Preparation of pre-trained model download and installation of required tools

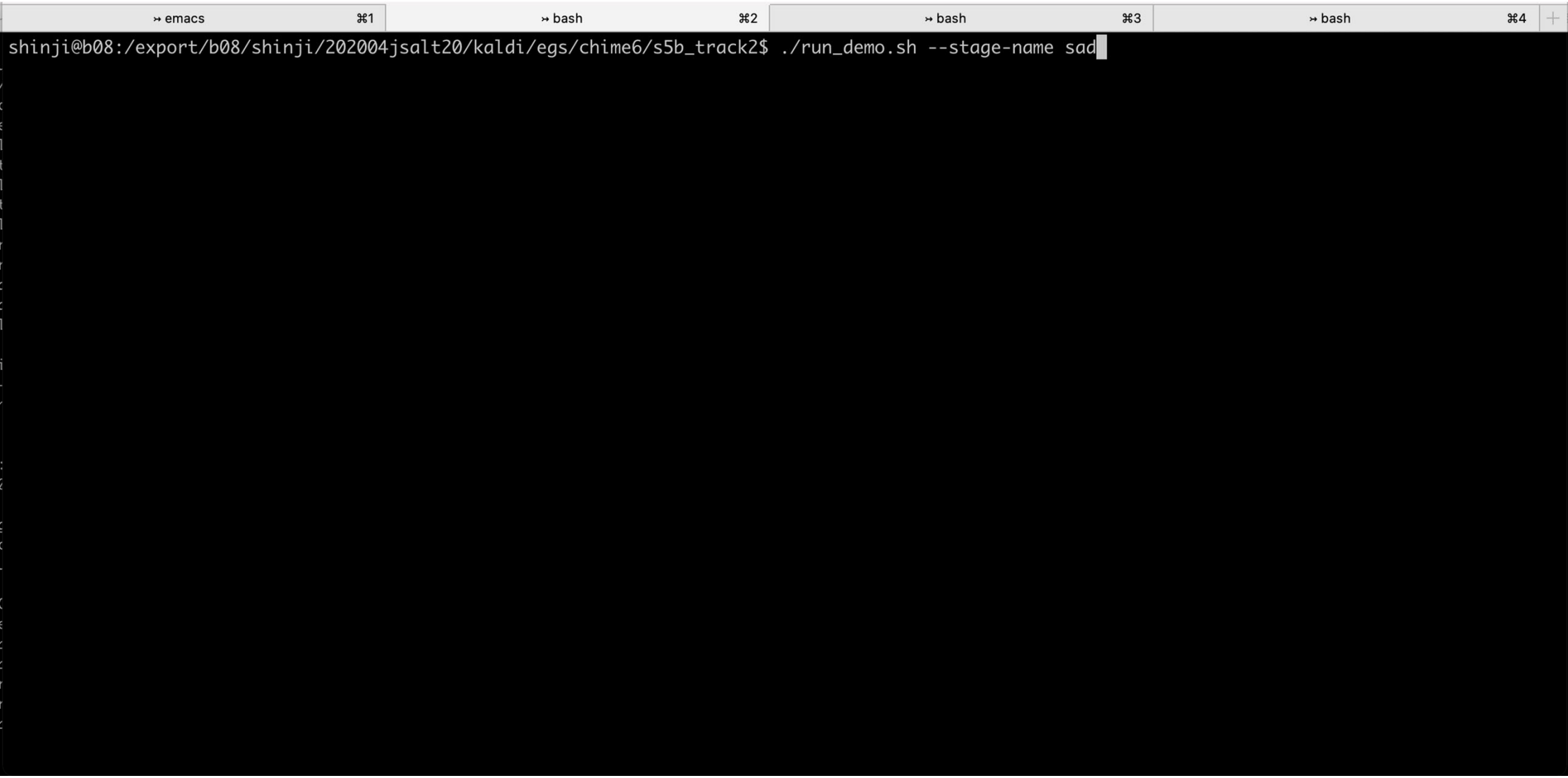
VAD



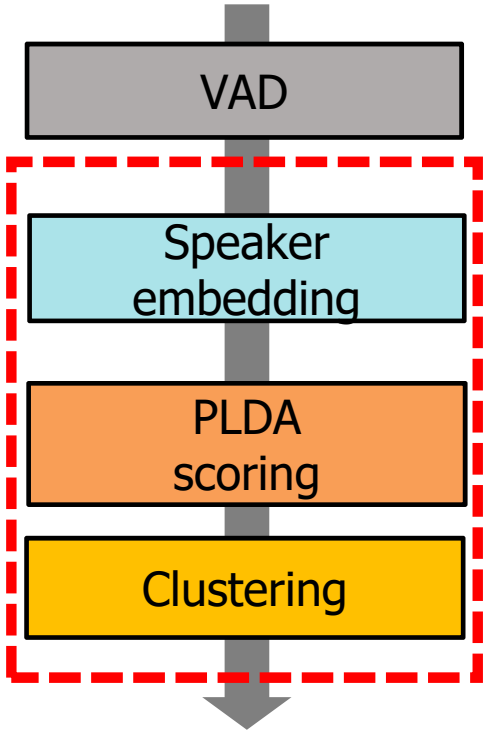

```
./run_demo.sh --stage-name sad
```

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0001		2.17	11.18	
0002		11.18	17.78	
0003		18.70	22.52	

VAD


↣ emacs ⌘1
↣ bash ⌘2
↣ bash ⌘3
↣ bash ⌘4
shinji@b08:/export/b08/shinji/202004jsalt20/kaldi/egs/chime6/s5b_track2$ ./run_demo.sh --stage-name sad

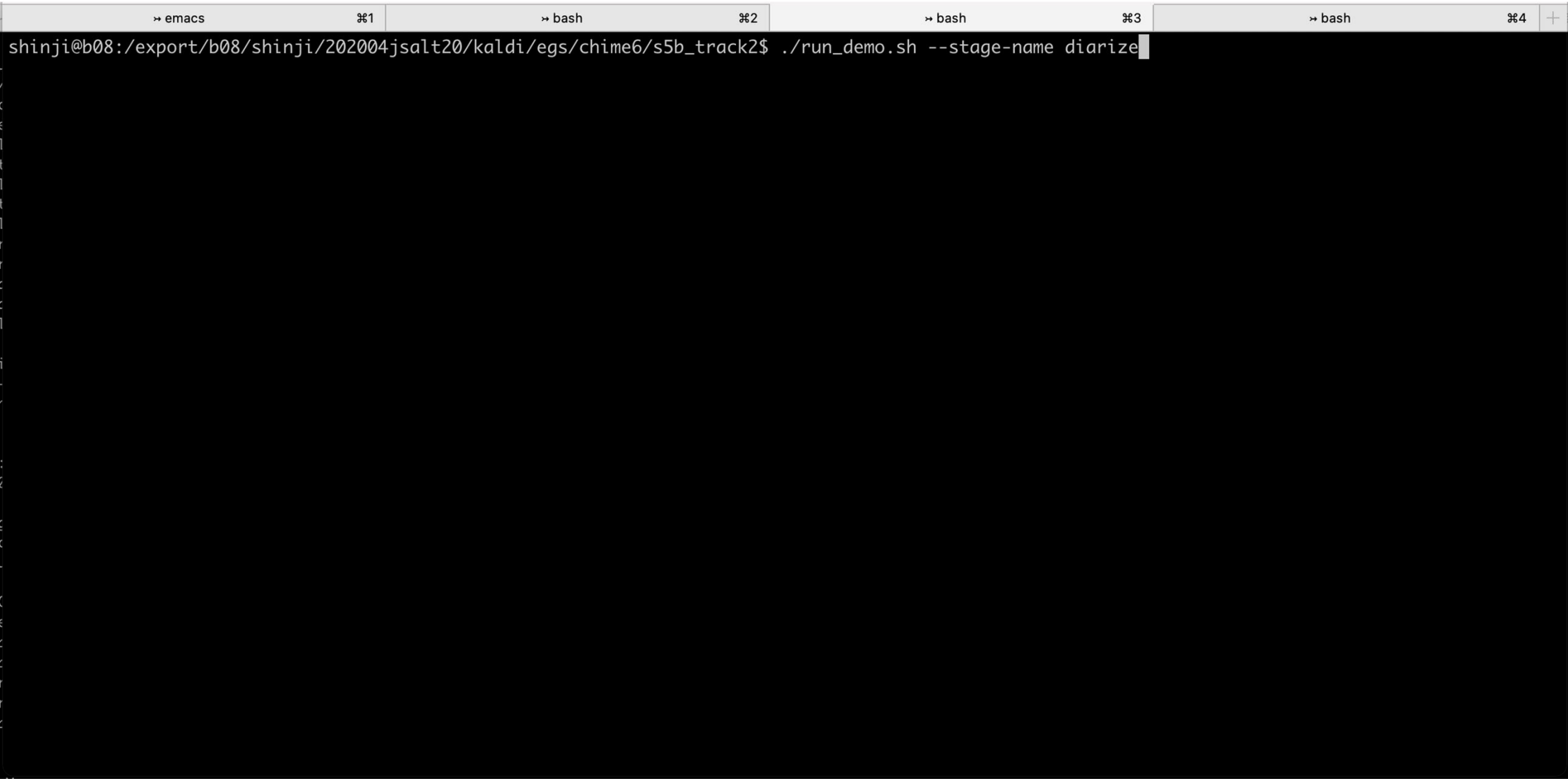
Diarization




```
./run_demo.sh --stage-name sad
./run_demo.sh --stage-name diarize
```

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0001	2	2.170	3.295	
0002	3	3.295	17.780	
0003	4	18.700	22.52	

Diarization

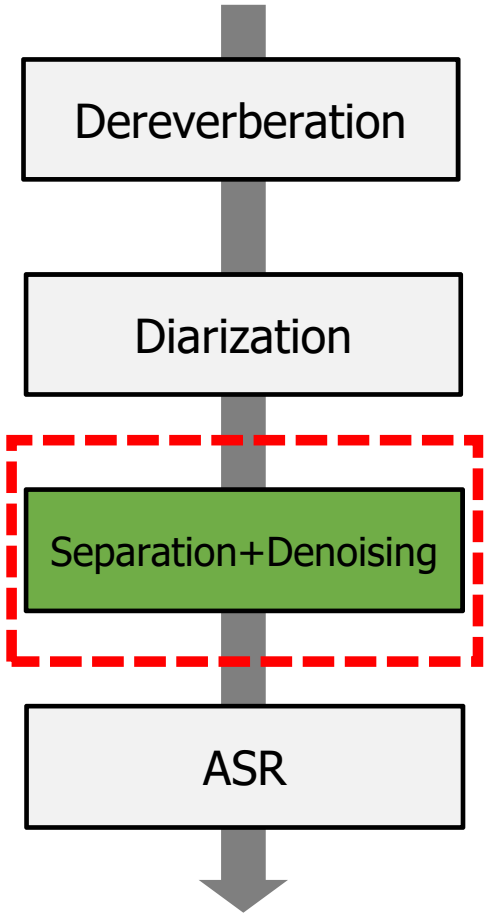


Diarization (after 30min.)

↣ emacs ⌘1 | ↣ bash ⌘2 | ↣ bash ⌘3 | ↣ bash ⌘4 | ↣ bash ⌘5

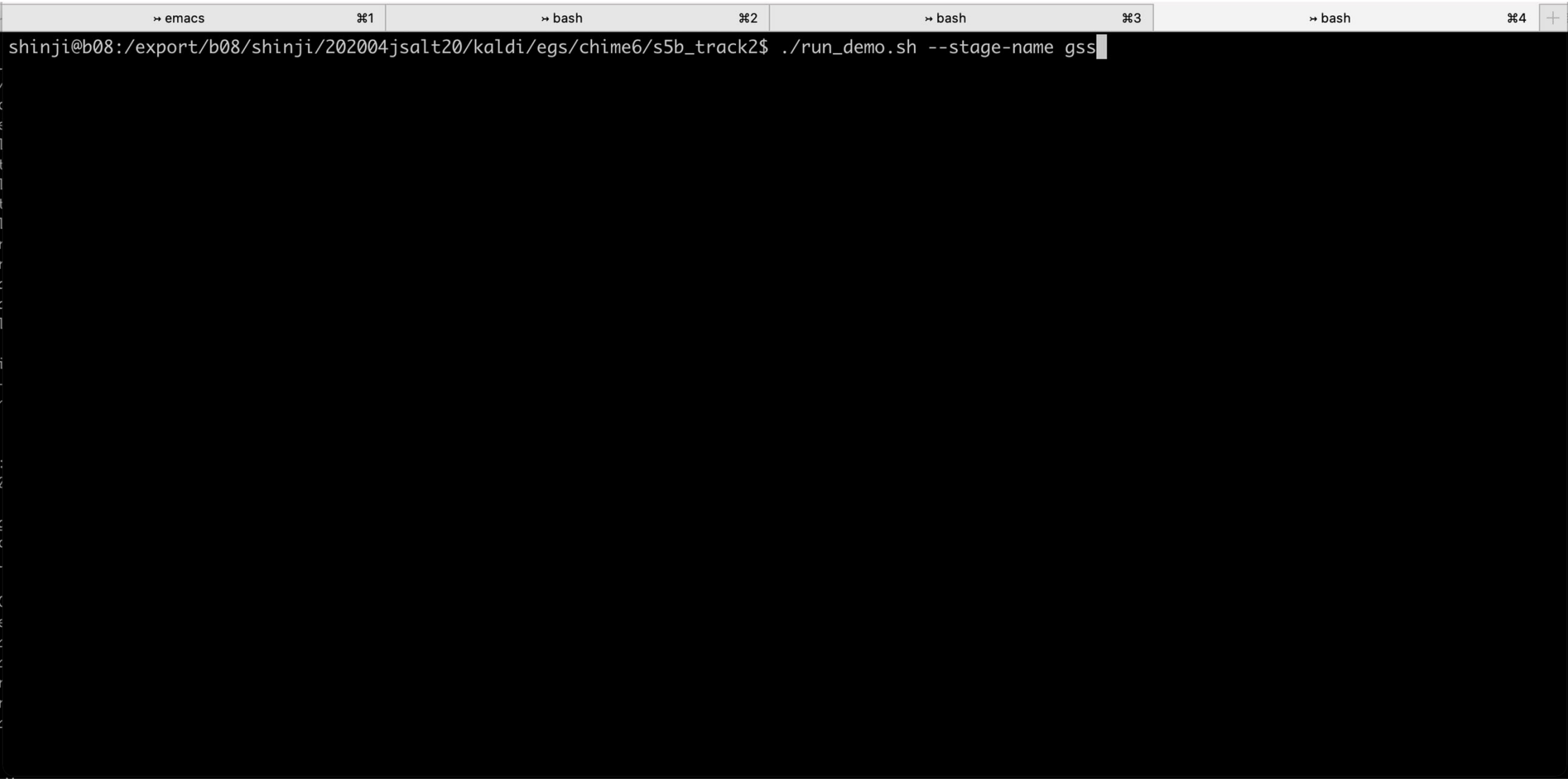
```
WARNING: Truncating turn overlapping non-scoring region. TURN: FILE: S02, SPEAKER: 3, ONSET: 54.850000, OFFSET: 68.350000, DUR: 13.500000
WARNING: Truncating turn overlapping non-scoring region. TURN: FILE: S02, SPEAKER: 3, ONSET: 70.090000, OFFSET: 71.000000, DUR: 0.910000
WARNING: Truncating turn overlapping non-scoring region. TURN: FILE: S02, SPEAKER: 3, ONSET: 73.390000, OFFSET: 74.050000, DUR: 0.660000
WARNING: Truncating turn overlapping non-scoring region. TURN: FILE: S02, SPEAKER: 4, ONSET: 74.050000, OFFSET: 76.300000, DUR: 2.250000
Checking for overlapping reference speaker turns...
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: P05
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: P06
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: P07
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: P08
Checking for overlapping system speaker turns...
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: 2
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: 3
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: 4
WARNING: Merging overlapping speaker turns. FILE: S02, SPEAKER: 5
Scoring...
WARNING: File "S01" missing in reference RTTMs.
WARNING: File "S01" missing in system RTTMs.
WARNING: File "S09" missing in reference RTTMs.
WARNING: File "S09" missing in system RTTMs.
WARNING: File "S21" missing in reference RTTMs.
WARNING: File "S21" missing in system RTTMs.
File              DER    JER    B3-Precision    B3-Recall    B3-F1    GKT(ref, sys)    GKT(sys, ref)    H(ref|sys)    H(sys|ref)    MI    NMI
--------------  -----  -----  --------------  -----------  -------  ---------------  ---------------  ------------  ------------  ----  -----
S01              0.00   0.00            1.00         1.00     1.00             1.00             1.00          0.00          0.00  0.00   1.00
S02             50.83  57.10            0.41         0.56     0.47             0.46             0.32          2.23          1.49  1.25   0.40
S09              0.00   0.00            1.00         1.00     1.00             1.00             1.00          0.00          0.00  0.00   1.00
S21              0.00   0.00            1.00         1.00     1.00             1.00             1.00          0.00          0.00  0.00   1.00
*** OVERALL ***  50.83  57.10            0.85         0.89     0.87             0.86             0.82          0.56          0.37  2.31   0.83
local/demo/run_diarize.sh: Saved audacity labels to demo/S02.diar
shinji@b08:/export/b08/shinji/202004jsalt20/kaldi/egs/chime6/s5b_track2$
```

Separation+Denoising

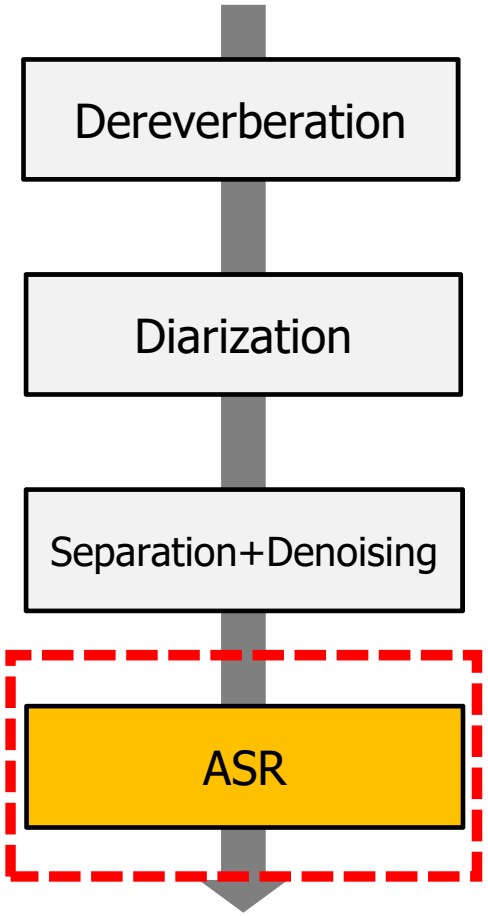



```
./run_demo.sh --stage-name sad
./run_demo.sh --stage-name diarize
./run_demo.sh --stage-name gss
```

Separation+Denoising



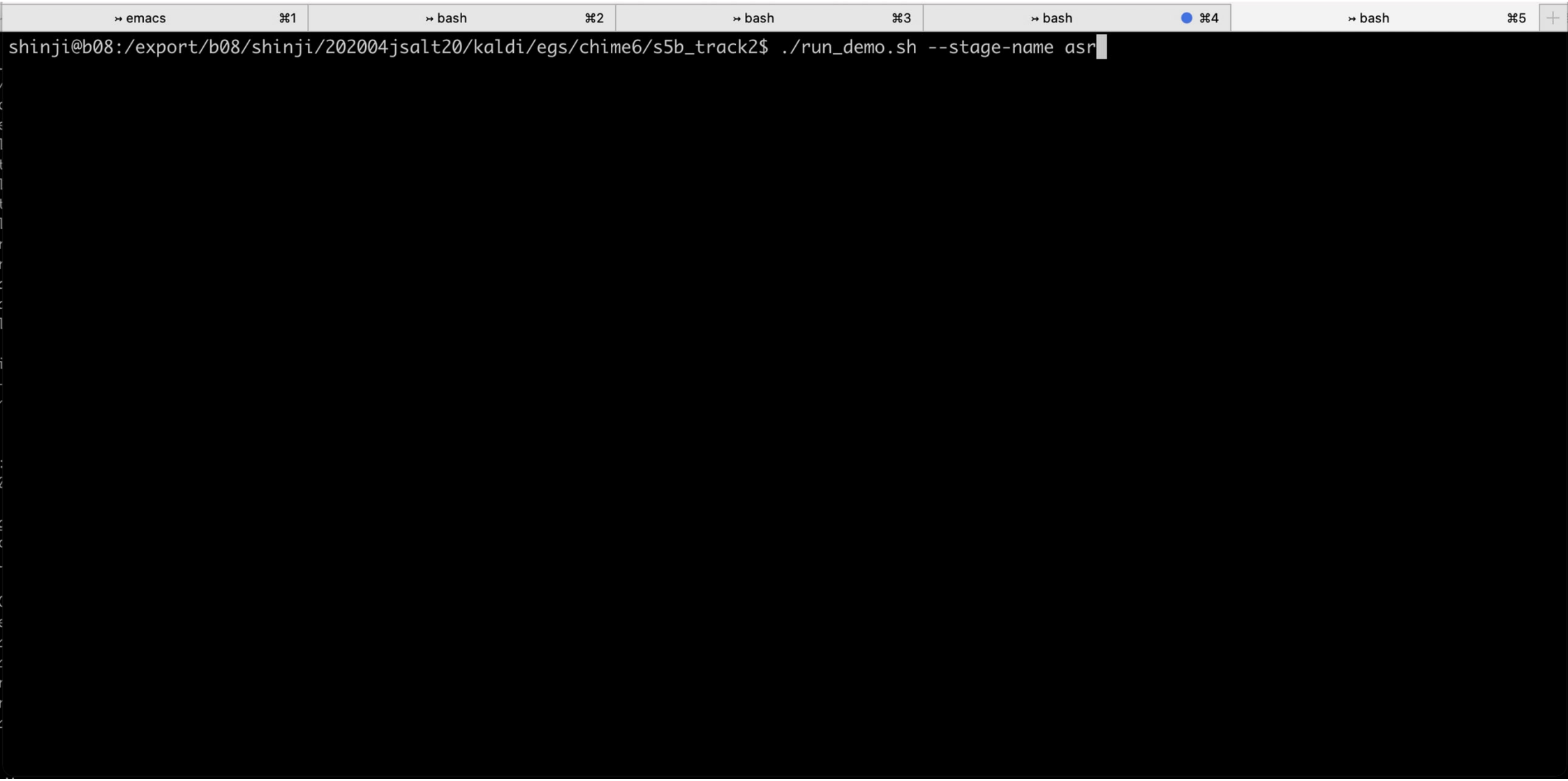
ASR



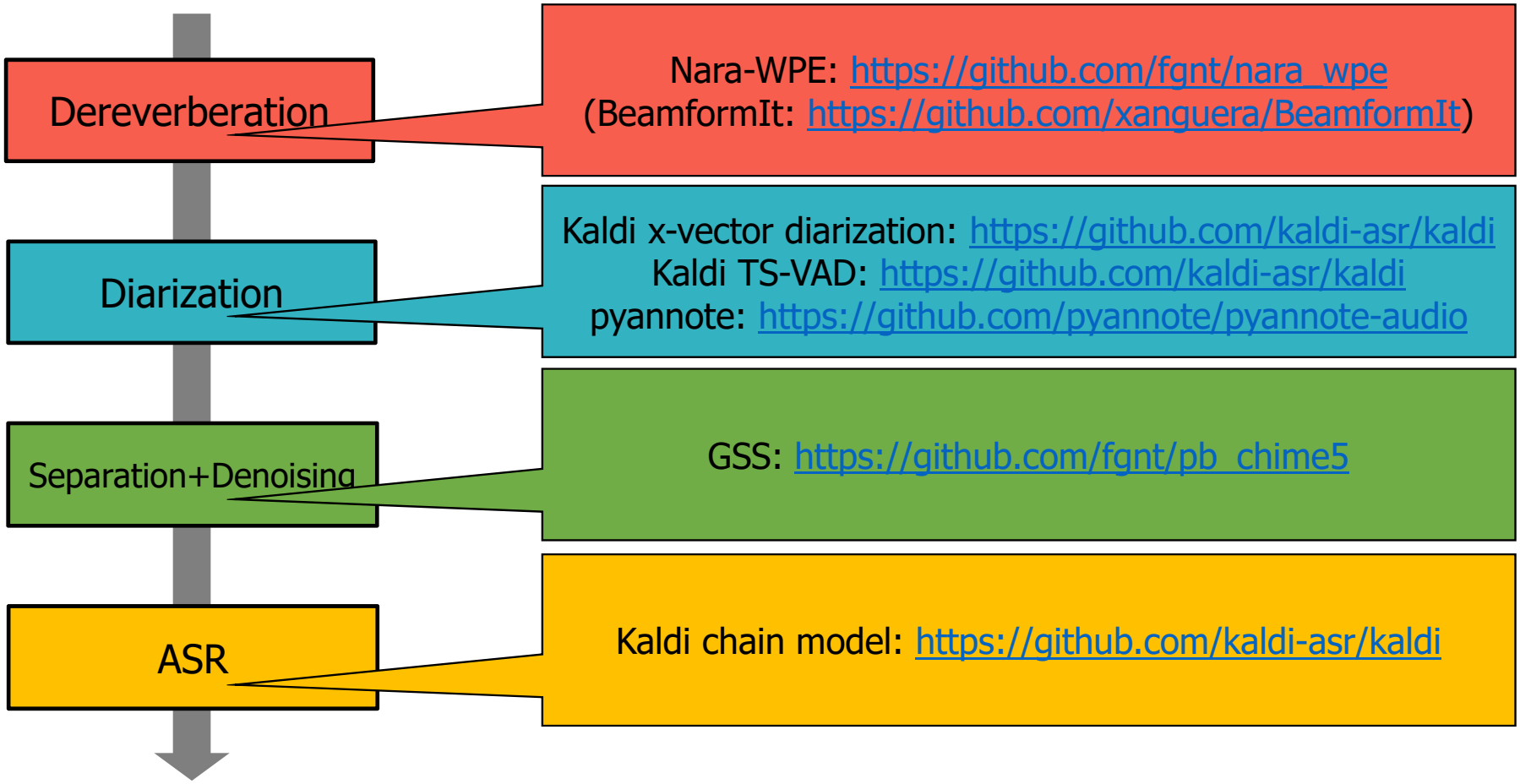

```
./run_demo.sh --stage-name sad
./run_demo.sh --stage-name diarize
./run_demo.sh --stage-name gss
./run_demo.sh --stage-name asr
```

Segment ID τ	Speaker Cluster k	Onset (sec) $b(\tau)$	Offset (sec) $e(\tau)$	Transcription W_τ
0001	2	2.170	3.295	can i can do yours the pie
0002	3	3.295	17.780	yeah you have to whisk it though right
0003	4	18.700	22.52	[laughs]

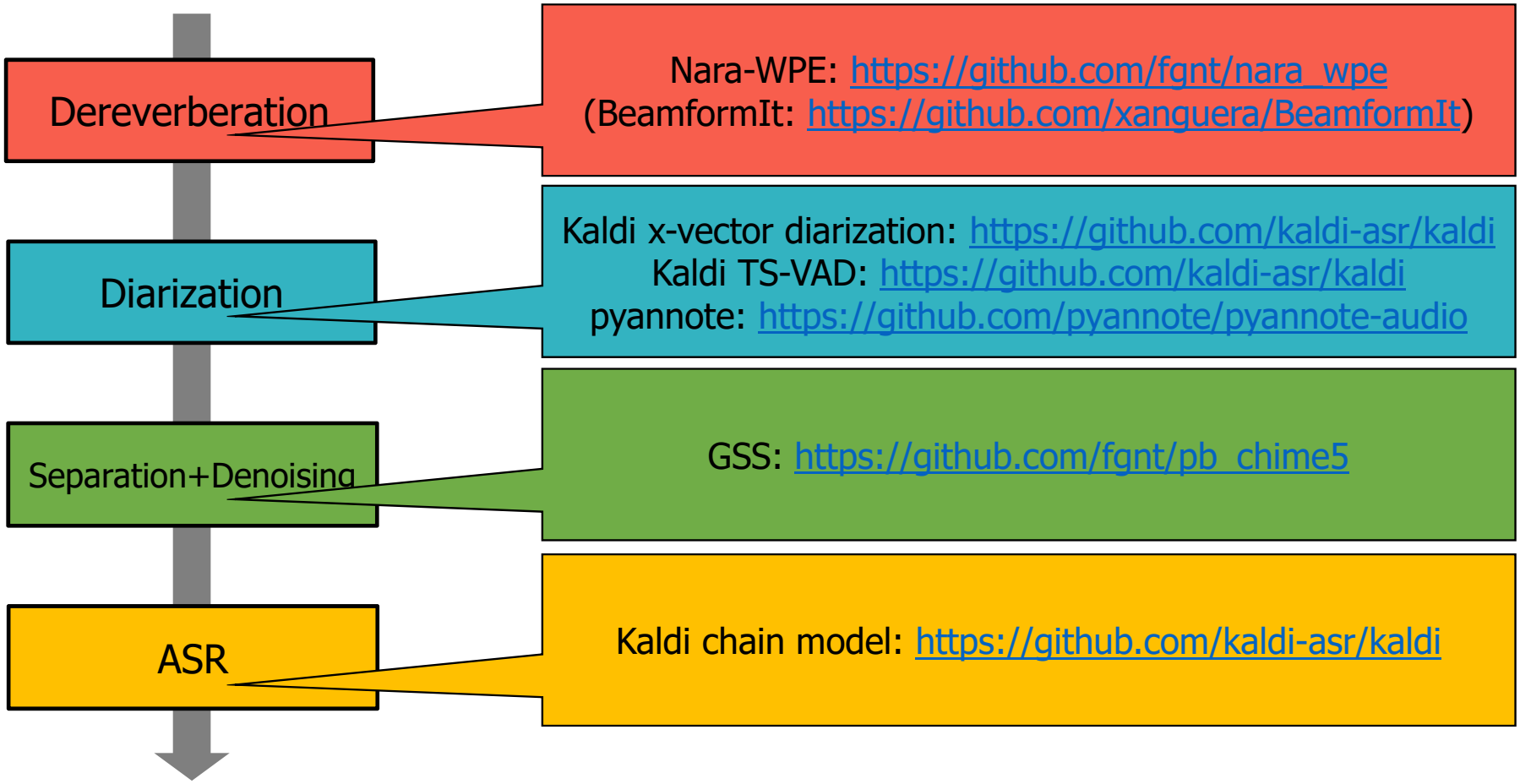
ASR


↣ emacs ⌘1
↣ bash ⌘2
↣ bash ⌘3
↣ bash ⌘4
↣ bash ⌘5
shinji@b08:/export/b08/shinji/202004jsalt20/kaldi/egs/chime6/s5b_track2$ ./run_demo.sh --stage-name asr

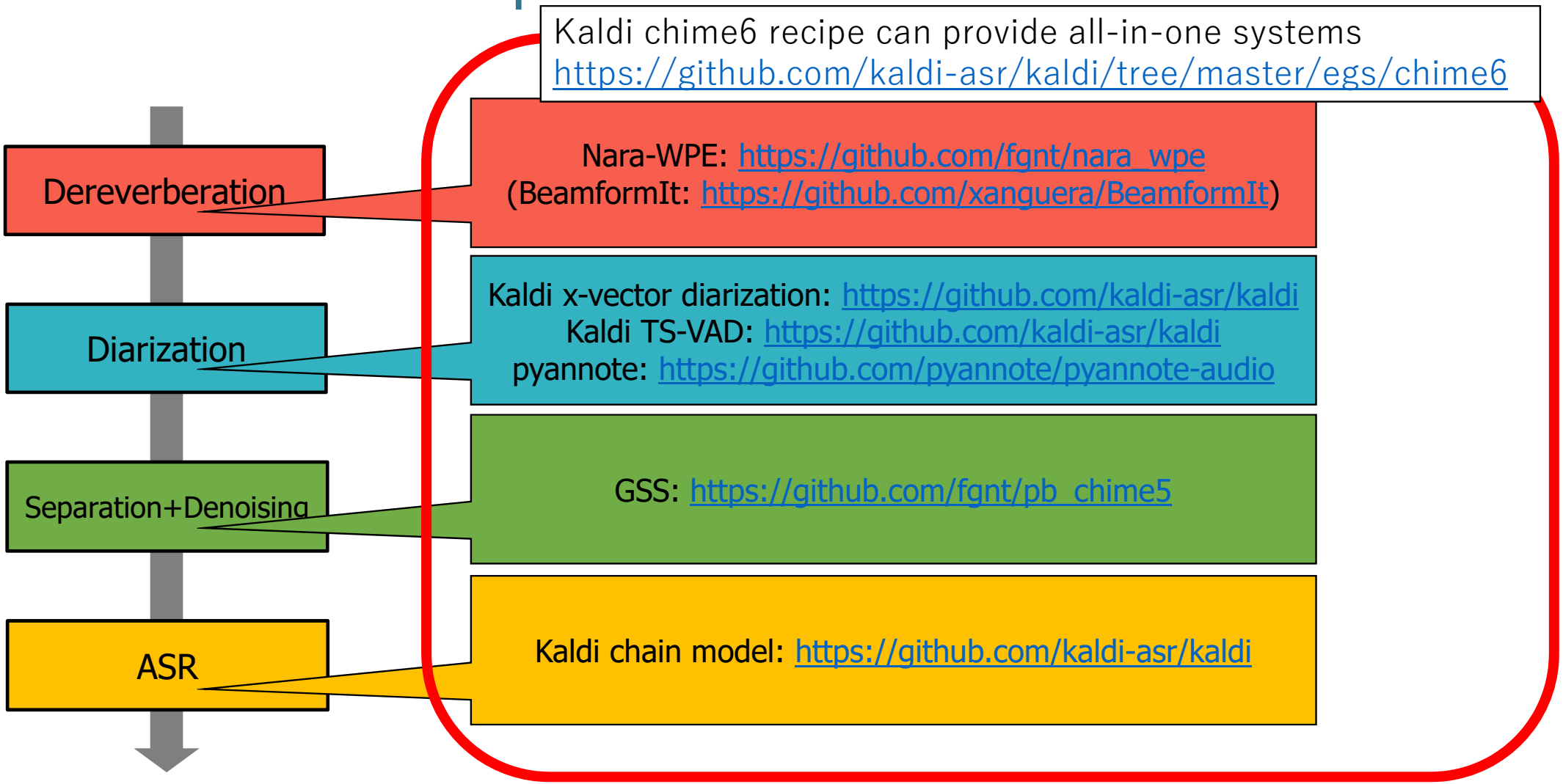
All are based on open source



All are based on open source



All are based on open source



Discussions

We have a great step toward distant conversational speech recognition and analysis

- The performance is significantly improved from the progress of each module
- Database and reproducible open-source systems (CHiME-6 as an example)

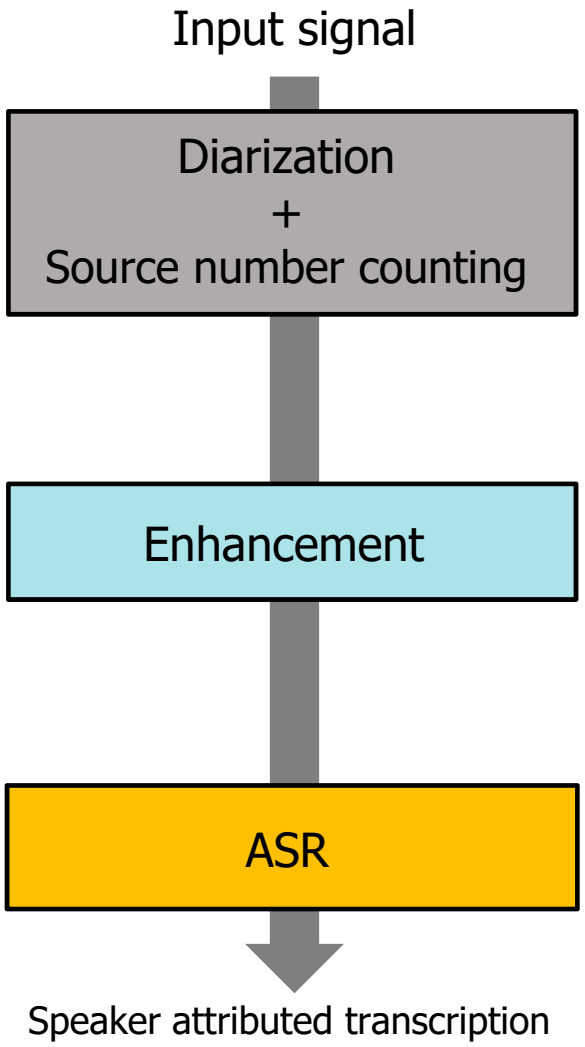
The system becomes highly complex

- Each module is optimized with different models, objectives, and tools

Jointly optimal systems

3. A new research trend: Jointly optimal systems

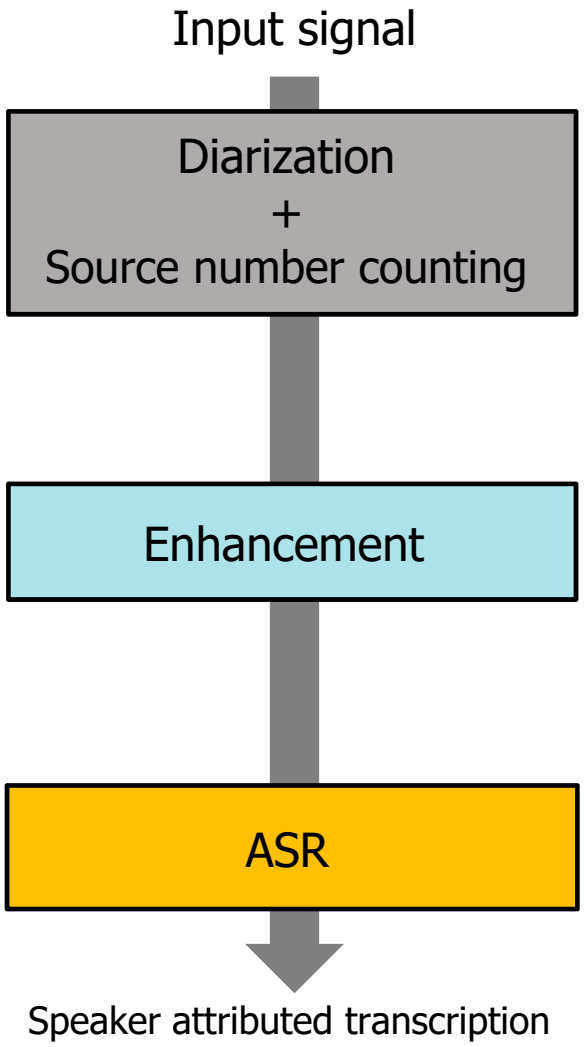
A typical pipeline assumed in challenges like CHiME-6



- Works for real data
- Large room for improvement

Can we do any better?

A typical pipeline assumed in challenges like CHiME-6



All modules are optimized independently,
and thus don't know distortions possibly included
in its input data, i.e., the system is not jointly optimal.

Trend toward jointly optimal systems

3. A new research trend: Jointly optimal systems

3.1. Diarization + x

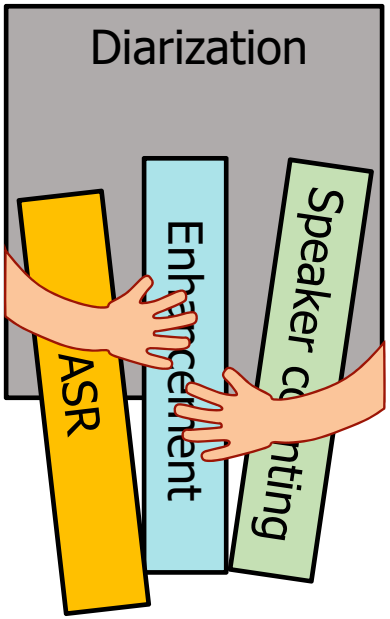
3.2. Enhancement + x

3.3. ASR + x

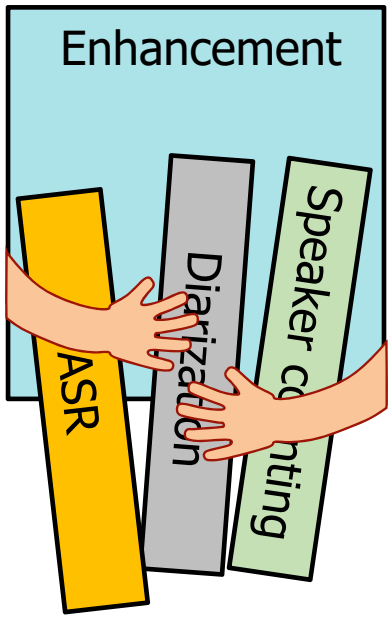
Approaches for optimal distant conversational ASR and analysis

- Current approaches can be roughly categorized into the following three.
- Currently proposed joint systems can be seen as extensions of a core module.
- Each approach has different strength. (This will be discussed in Part 4.)

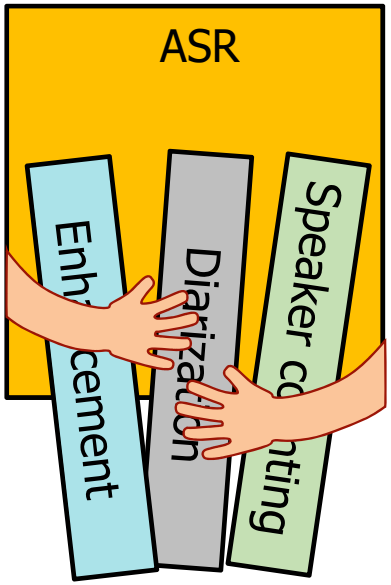
Diarization-originated approach



Enhancement-originated approach

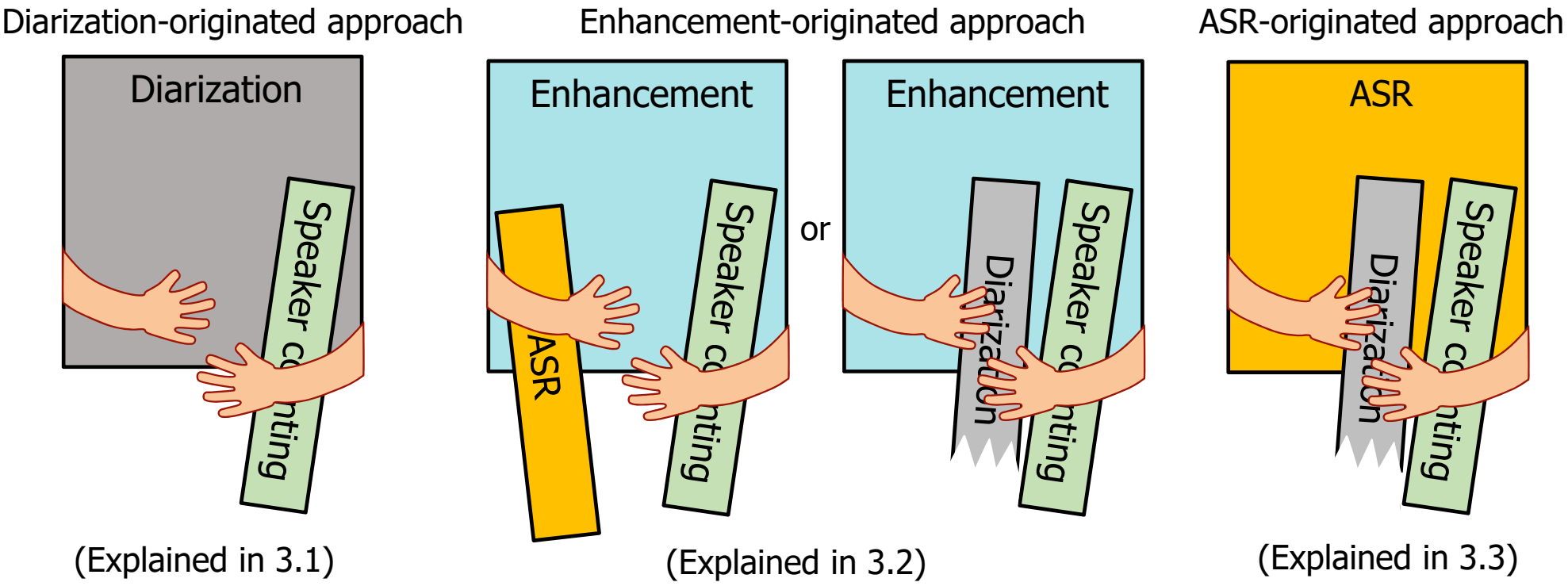


ASR-originated approach

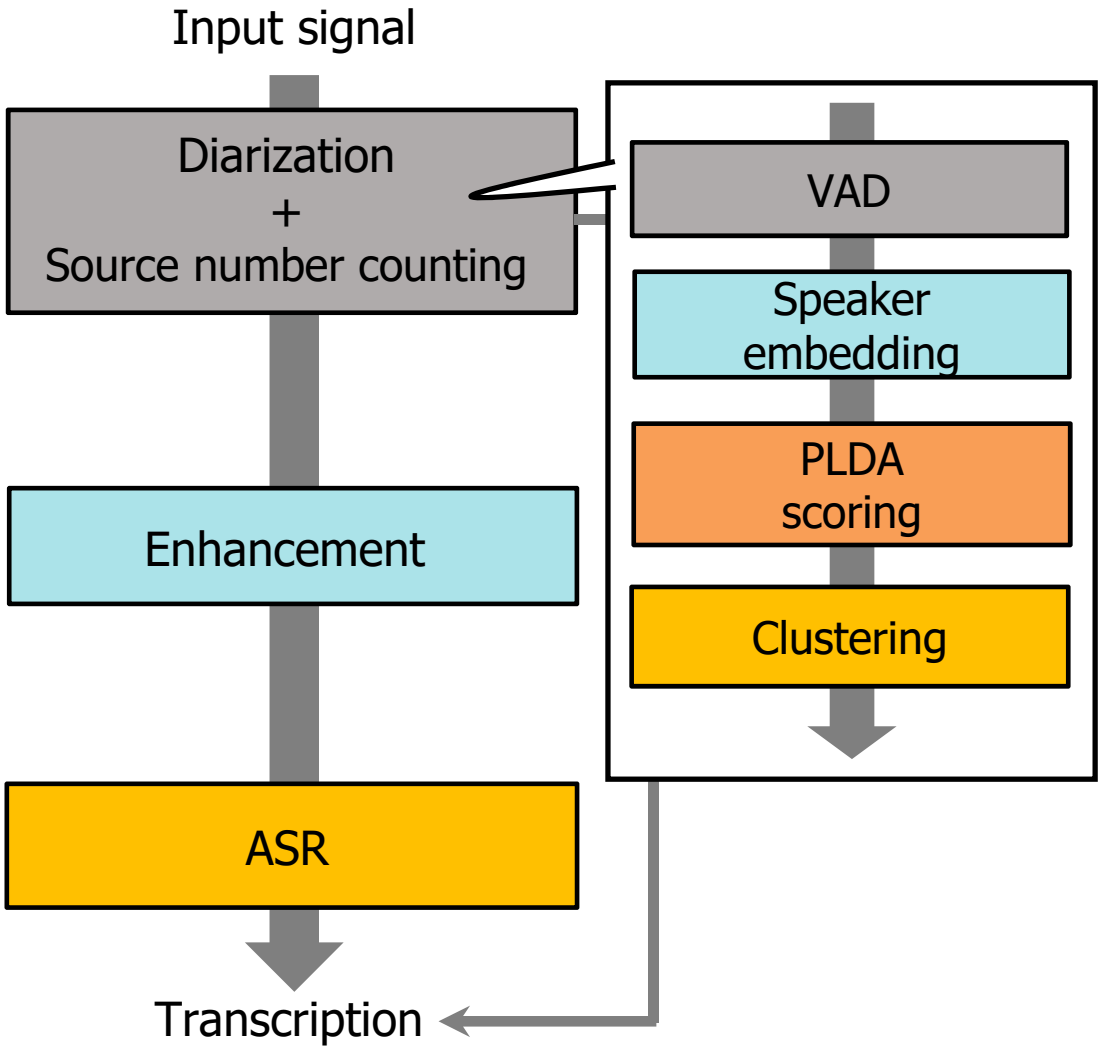


What has been achieved so far?

- None of them has yet achieved fully optimal distant conversational ASR and analysis.
- Their attainment levels are also different.
- Not yet clear which approach is most advantageous in the end.



Section 3.1: Diarization + X



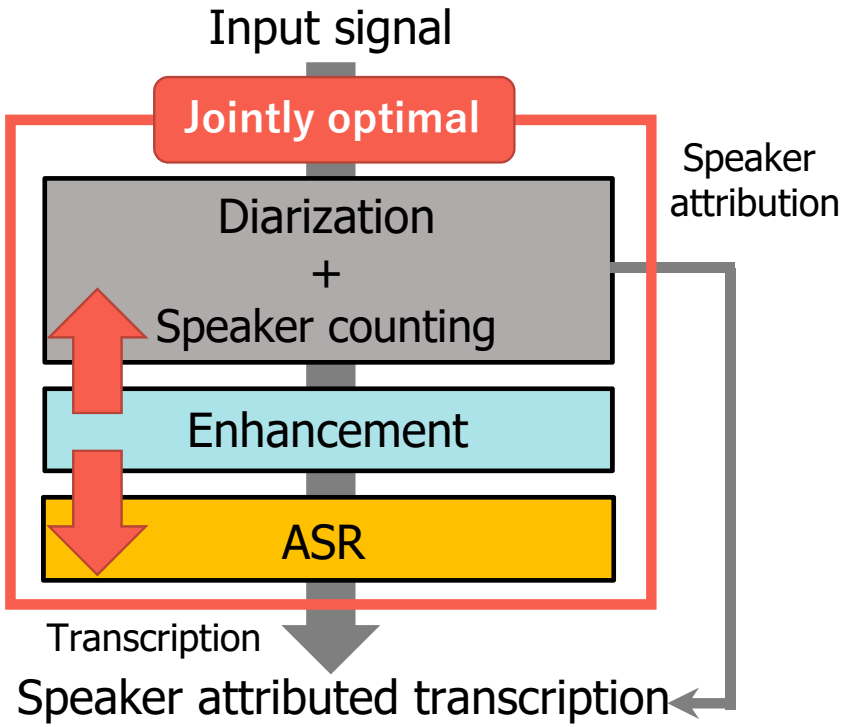
☹ Before talking about the joint optimization, the problem is the traditional diarization pipeline contains many independent modules.

→ Hard to optimize diarization systems
(a) for the diarization error minimization
(b) for combining it with other modules.

A new trend towards neural diarization!

Section 3.2: Enhancement + X

Trying to realize
jointly optimal modular systems,
by expanding enhancement module

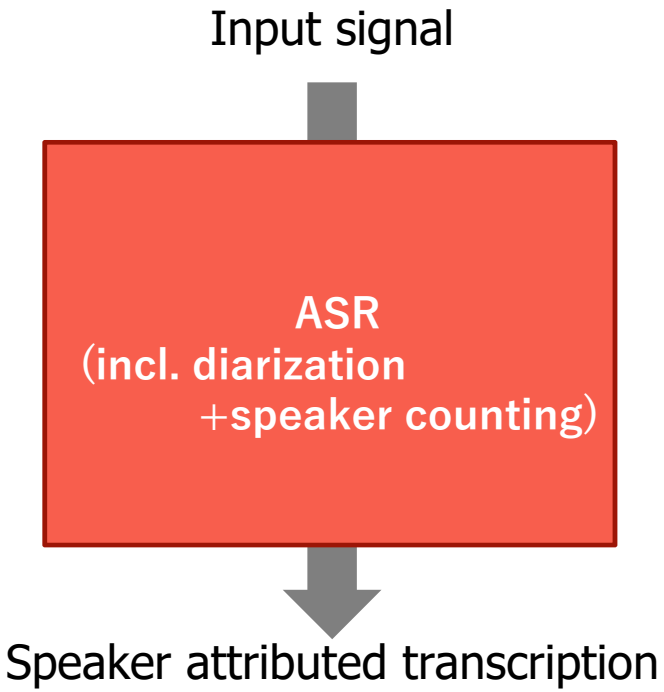


- Core system: Neural separation system

- Integration of the neural separation with
  (a) ASR
  (b) Multi-channel enhancement + ASR
  (c) Speaker counting + Diarization

Section 3.3: ASR + X

Trying to realize
a monolithic optimal system,
by extending the ASR module



- Core system: End-to-end ASR

- Integration of the end-to-end ASR with
  (a) Speaker counting
  (b) Diarzation

3. A new research trend: Jointly optimal systems

3.1. Diarization + x

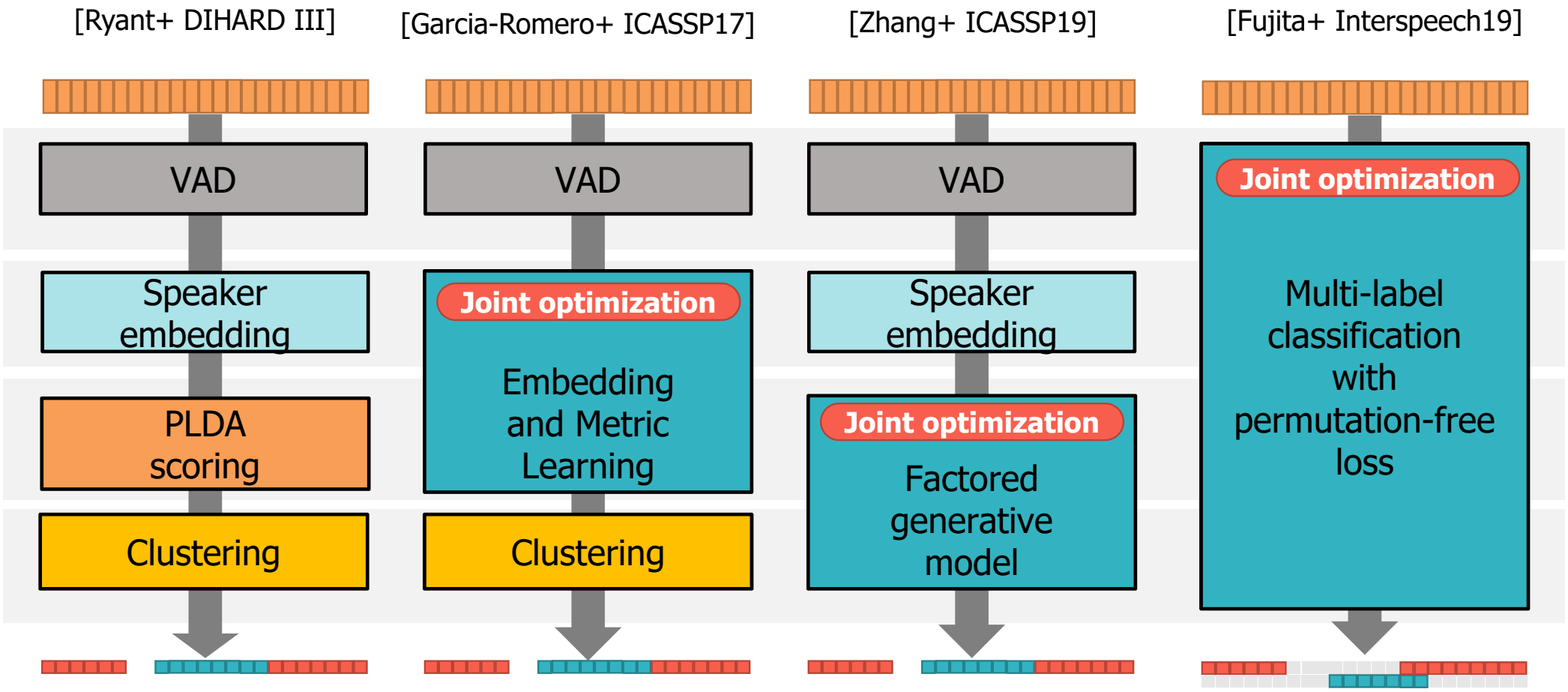
3.2. Enhancement + x

3.3. ASR + x

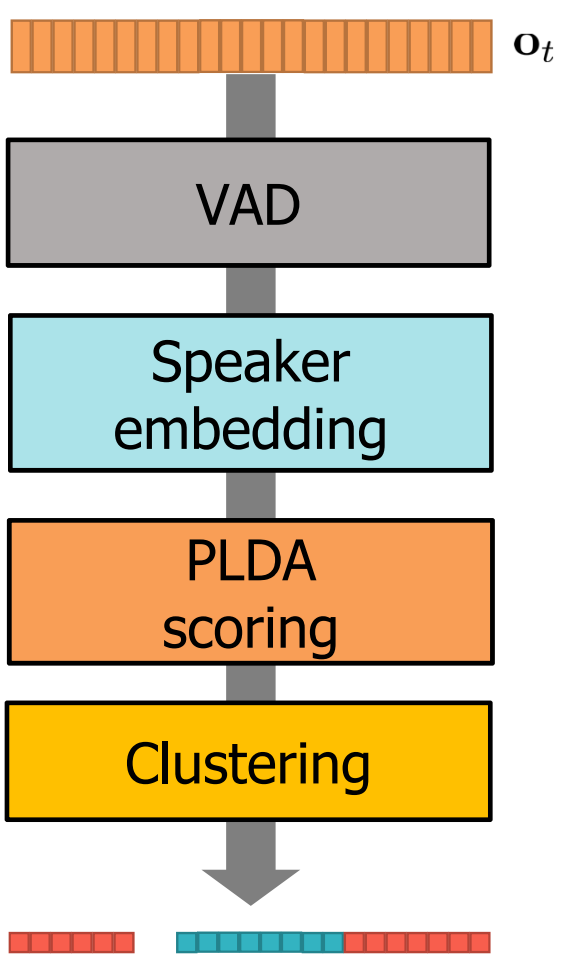
3.1 Diarization

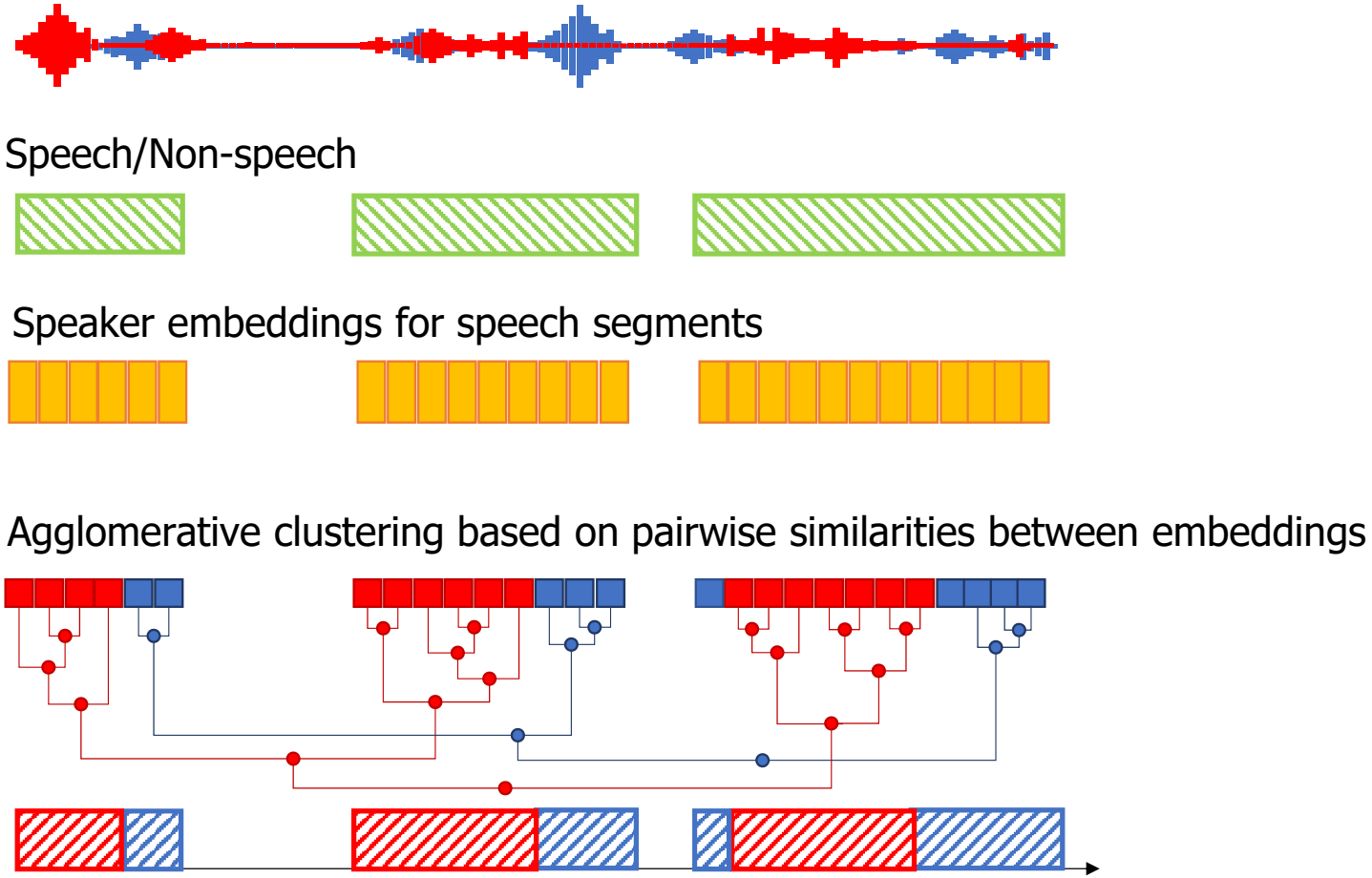
From clustering-based pipeline to end-to-end optimal systems

From Clustering-based to End-to-End



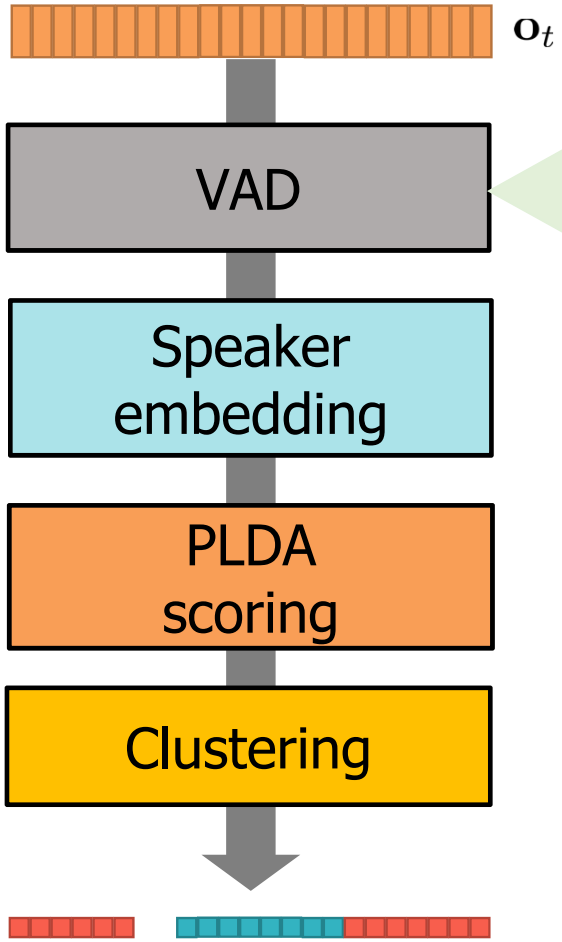
Typical Clustering-based Diarization



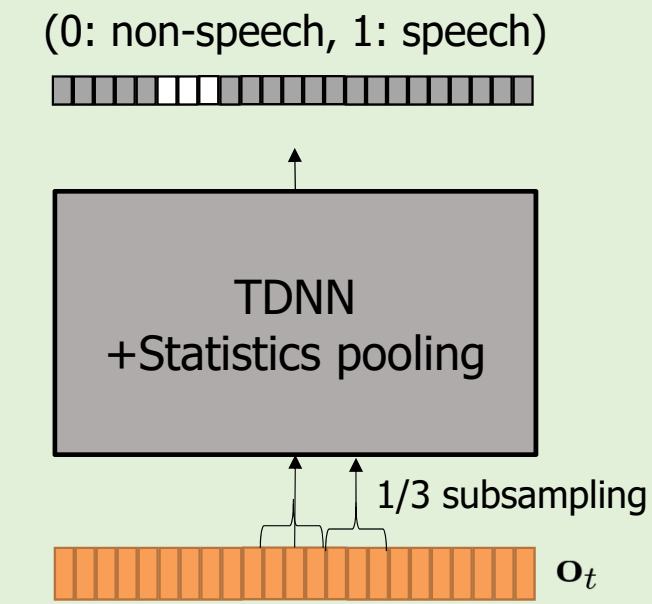


N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Typical Clustering-based Diarization

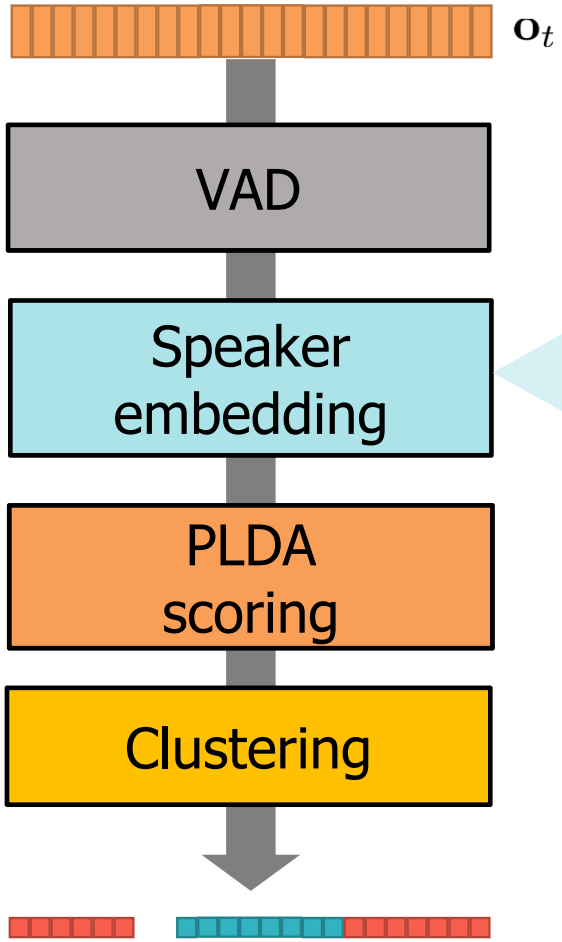


Voice activity detection using **a neural network**.

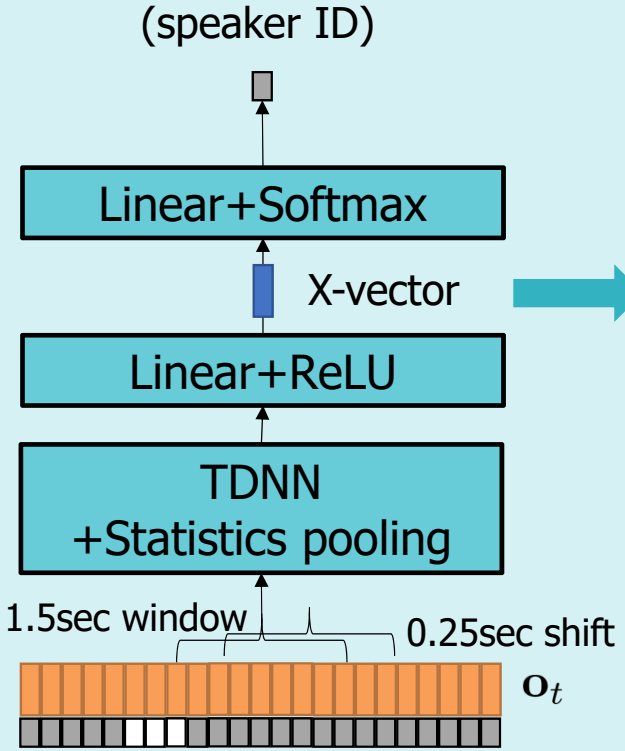


N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Typical Clustering-based Diarization

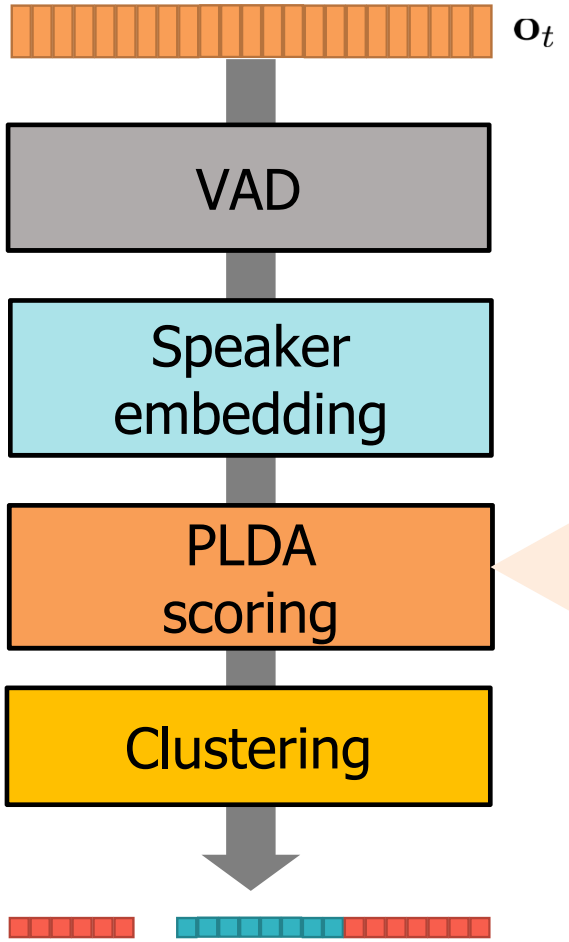


Speaker embedding requires **another neural network.**



N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Typical Clustering-based Diarization



Further, the system requires **another scoring model to be trained.**

$$\mathrm{LLR}(\mathbf{h}_{t_1}, \mathbf{h}_{t_2}) = \log \frac{p(\mathbf{h}_{t_1}, \mathbf{h}_{t_2} | H_{\mathrm{same}})}{p(\mathbf{h}_{t_1}, \mathbf{h}_{t_2} | H_{\mathrm{diff}})}$$

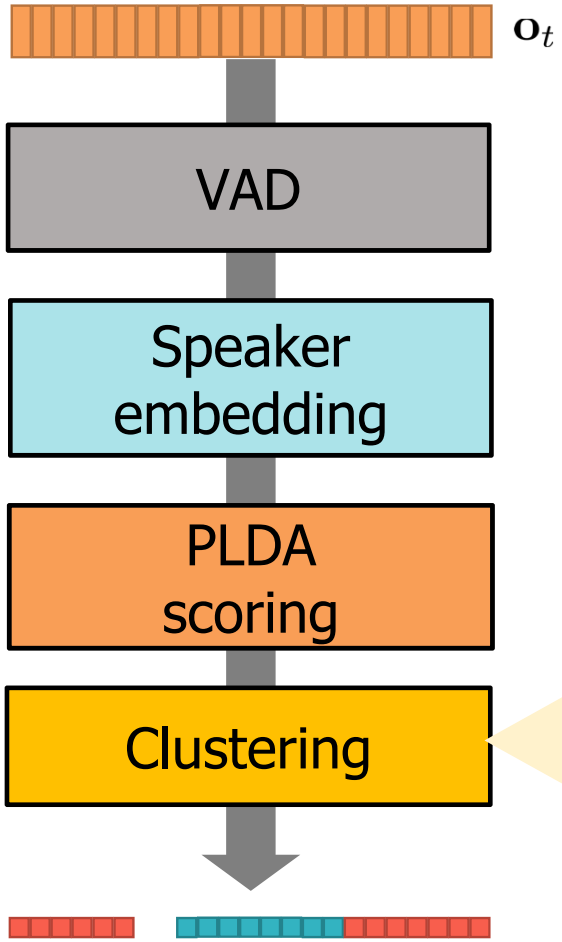
H_{same} : Hypothesis that two embeddings belong to the same speaker

H_{diff} : Hypothesis that two embeddings are from different speakers

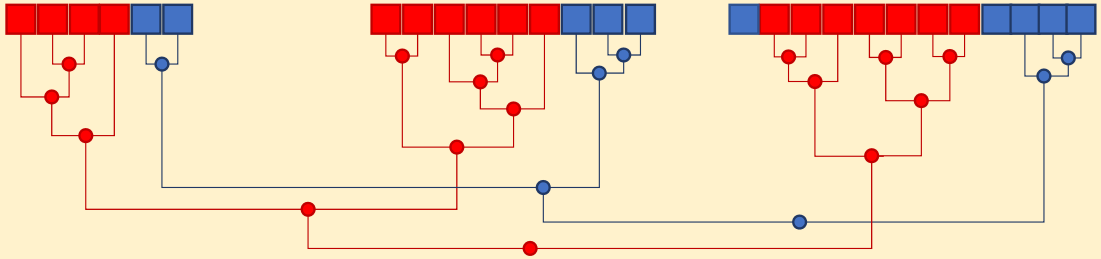
This log-likelihood ratio, known as **PLDA score**, is used for clustering.

N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Typical Clustering-based Diarization



Lastly, the system uses **unsupervised method** for grouping speakers.

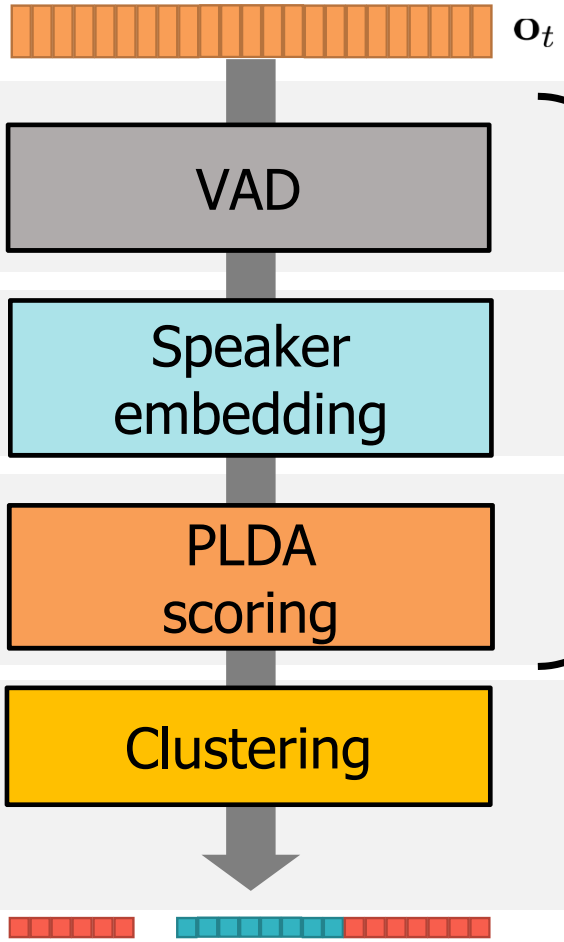


Agglomerative Hierarchical Clustering

Iteratively merge a pair of clusters that has the highest PLDA score, until the score met the threshold.

N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Typical Clustering-based Diarization



☹ Independent optimization

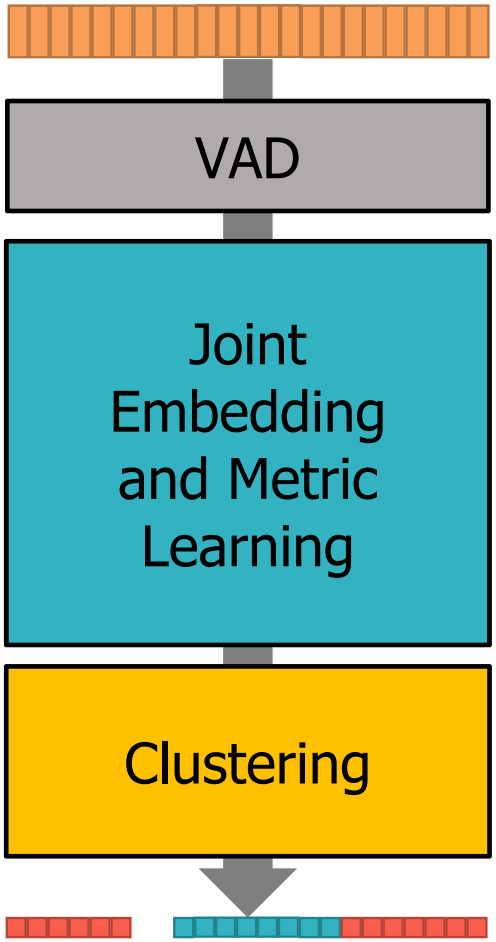
☹ Not directly optimized for diarization

☺ Speaker counting

☹ Can't handle overlapping speech

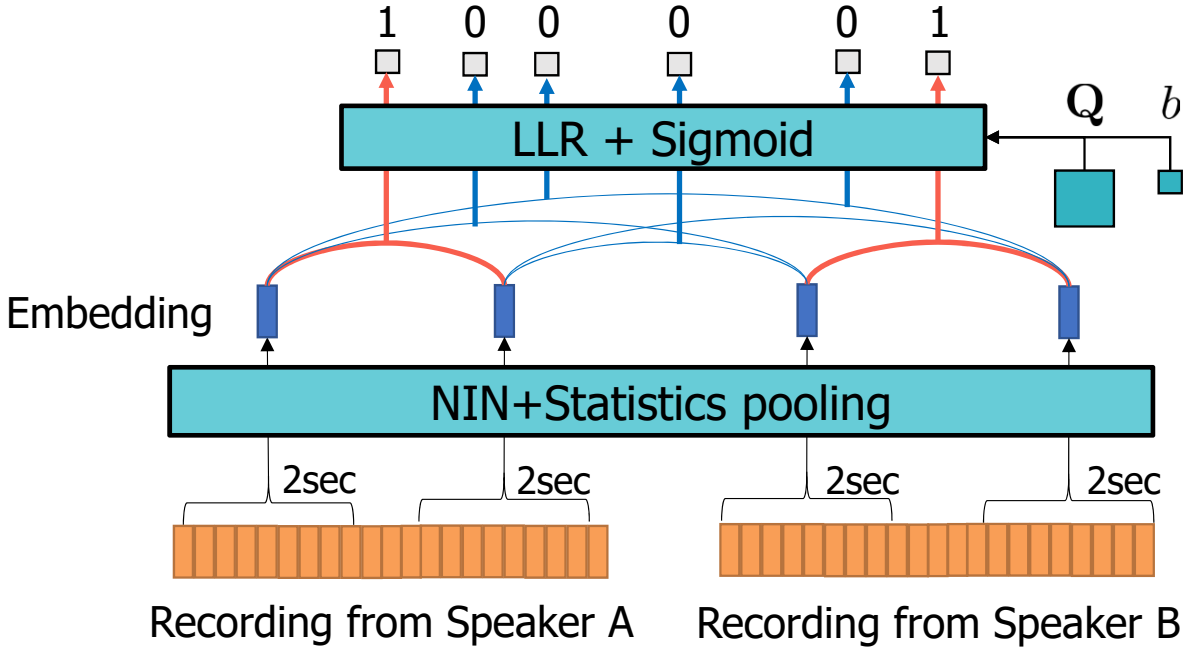
N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

Joint Embedding and Metric Learning



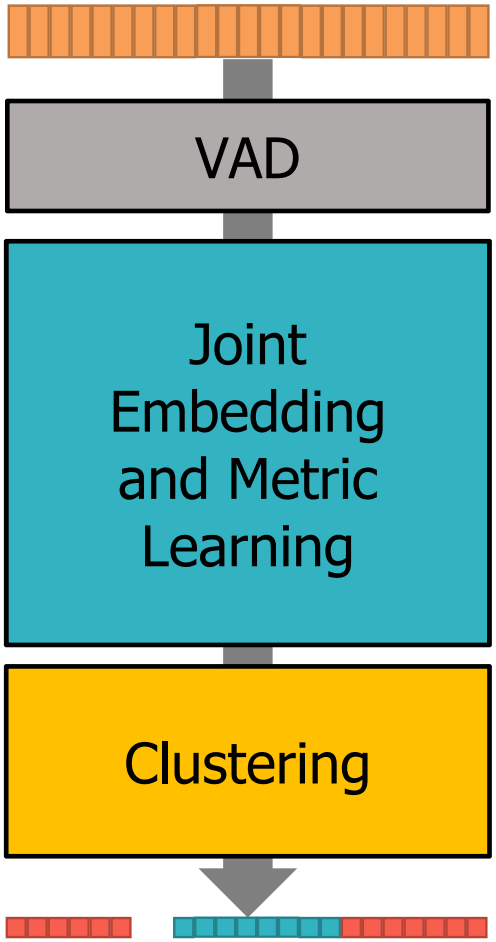
Speaker embedding and PLDA scoring matrix are jointly trained.

$$\mathrm{LLR}(\mathbf{h}_{t_1}, \mathbf{h}_{t_2}) = \mathbf{h}_{t_1}^{\top}\mathbf{Q}\mathbf{h}_{t_1} + \mathbf{h}_{t_2}^{\top}\mathbf{Q}\mathbf{h}_{t_2} + \mathbf{h}_{t_1}^{\top}\mathbf{h}_{t_2} + b$$



D. Garcia-Romero et al., "Speaker diarization using deep neural network embeddings," ICASSP 2017.

Joint Embedding and Metric Learning



DERs on CALLHOME 2-7 speakers **without VAD or overlap errors**.

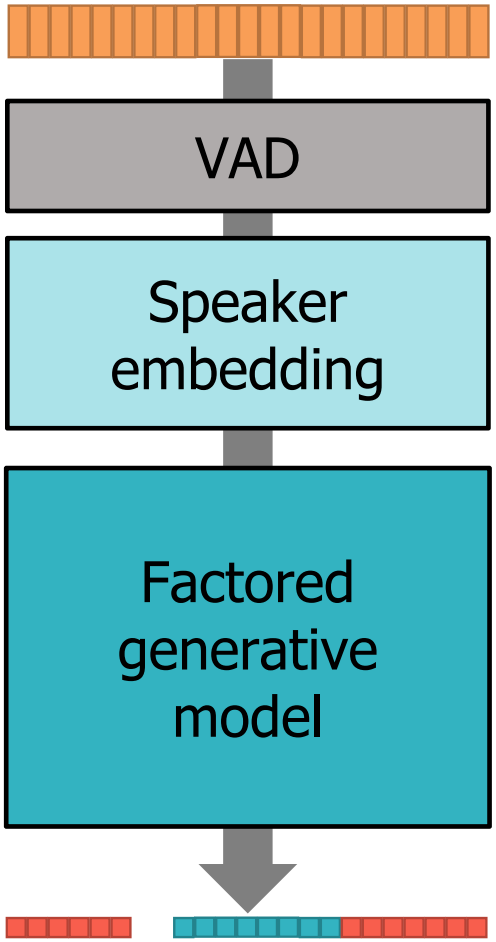
Speaker Embedding	Scoring	DER (%)
i-vector	PLDA	11.2
Senone DNN	PLDA	10.9
Joint Embedding and Metric Learning		**9.9**

☺ Joint optimization
☺ Speaker counting

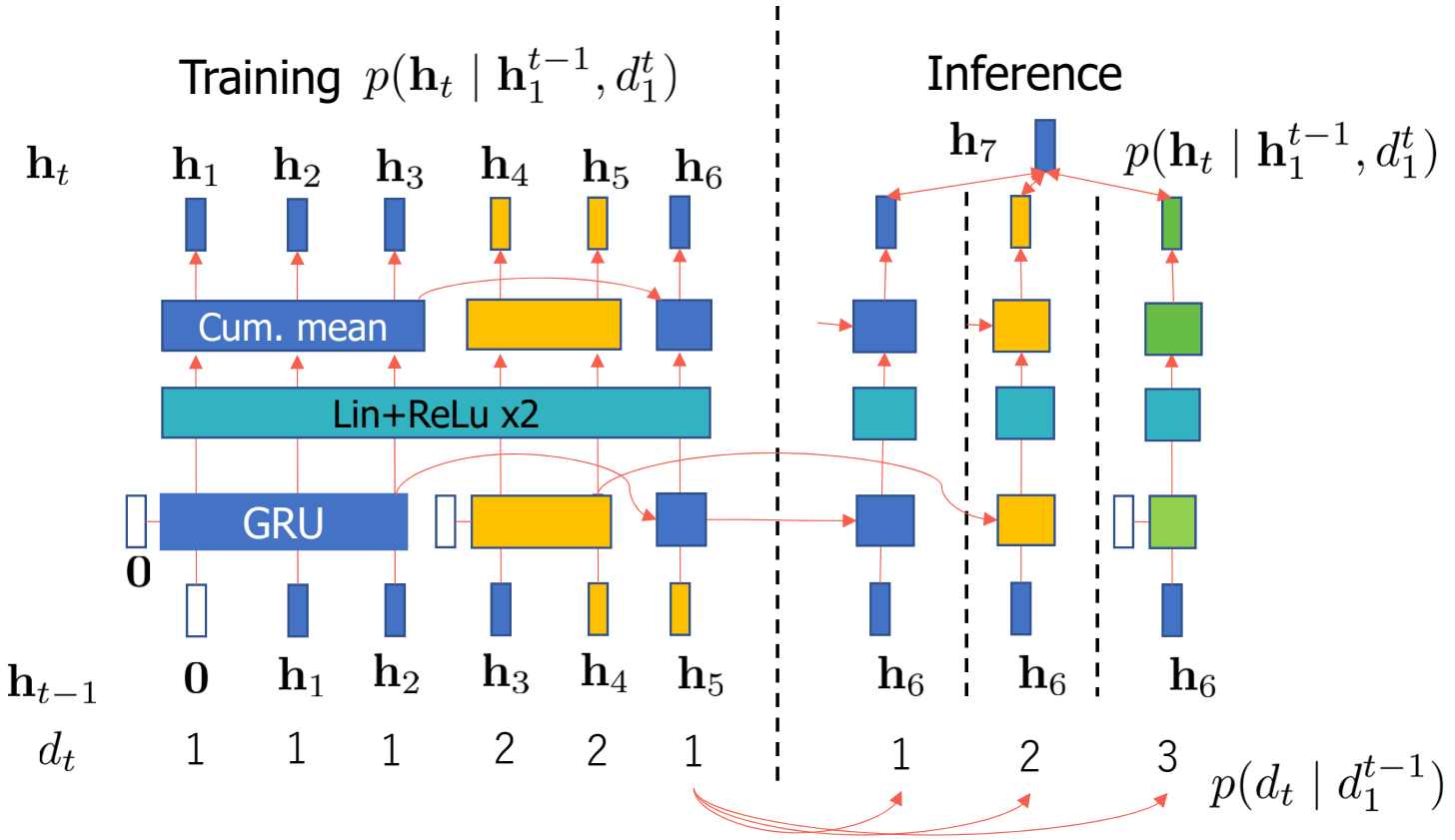
☹ Not directly optimized for diarization
☹ Can't handle overlapping speech

D. Garcia-Romero et al., "Speaker diarization using deep neural network embeddings," ICASSP 2017.

Fully-supervised Speaker Diarization

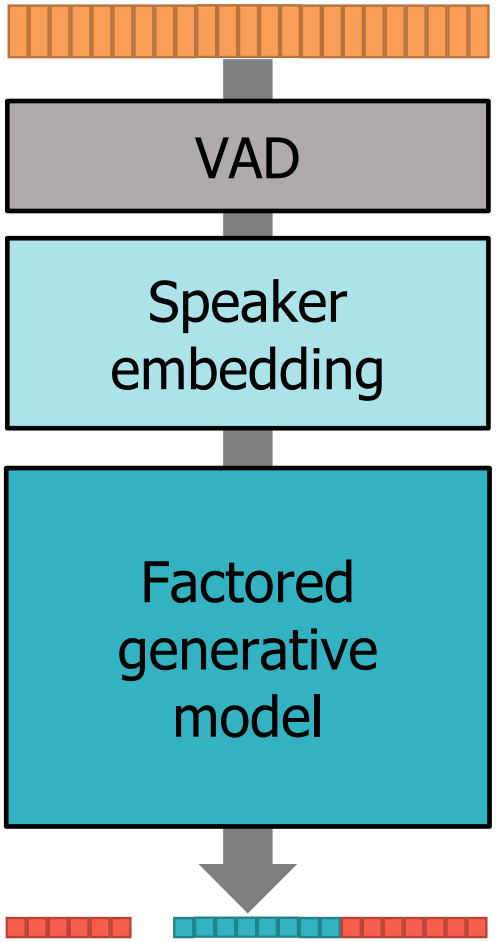


UIS-RNN is fully-supervised with speaker embeddings (d-vectors) and corresponding diarization.



A. Zhang et al., "Fully Supervised Speaker Diarization," ICASSP 2019.

Fully-supervised Speaker Diarization



DERs on CALLHOME 2-7 speakers **without VAD or overlap errors**.

Speaker Embedding	Method	DER (%)
d-vector V3	K-means clustering	12.3
d-vector V3	Spectral clustering	8.8
d-vector V3	**UIS-RNN**	**7.6**

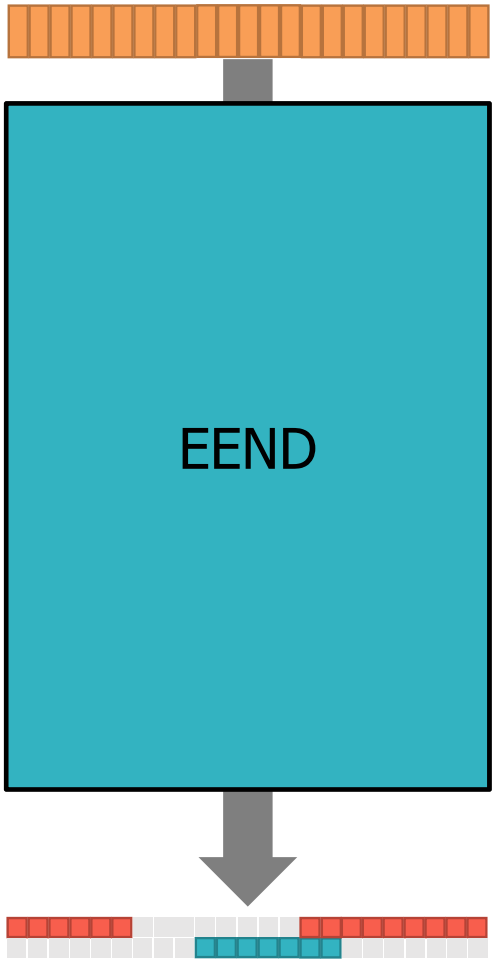
☺ Fully-supervised

☺ Speaker counting + Online

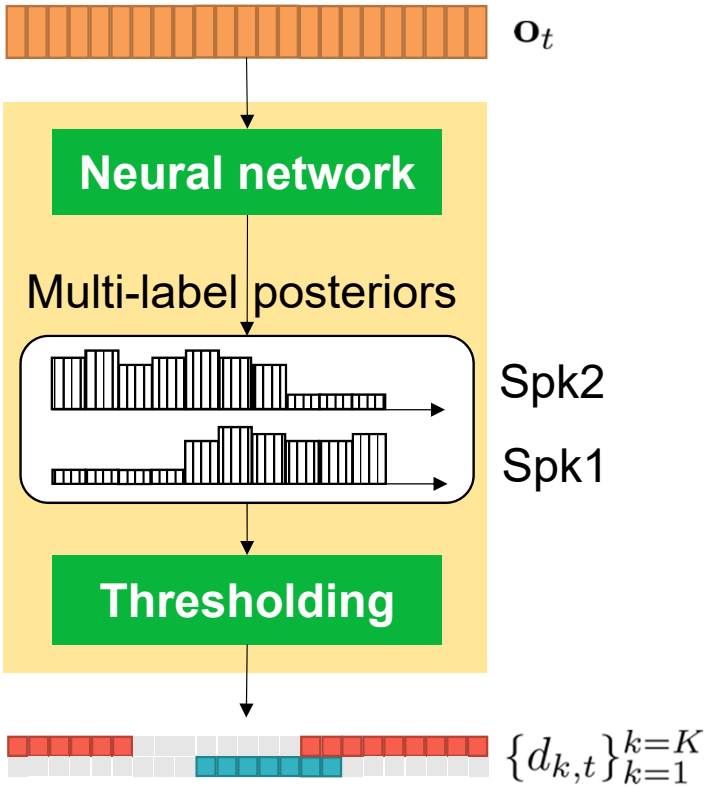
☹ Can't handle overlapping speech

A. Zhang et al., "Fully Supervised Speaker Diarization," ICASSP 2019.

End-to-End Neural Diarization

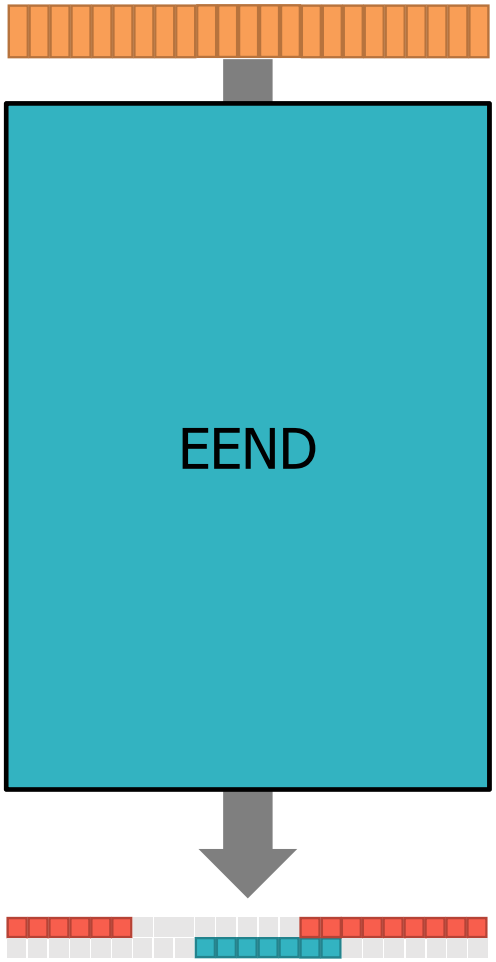


Direct estimation of diarization using **multi-label classification**

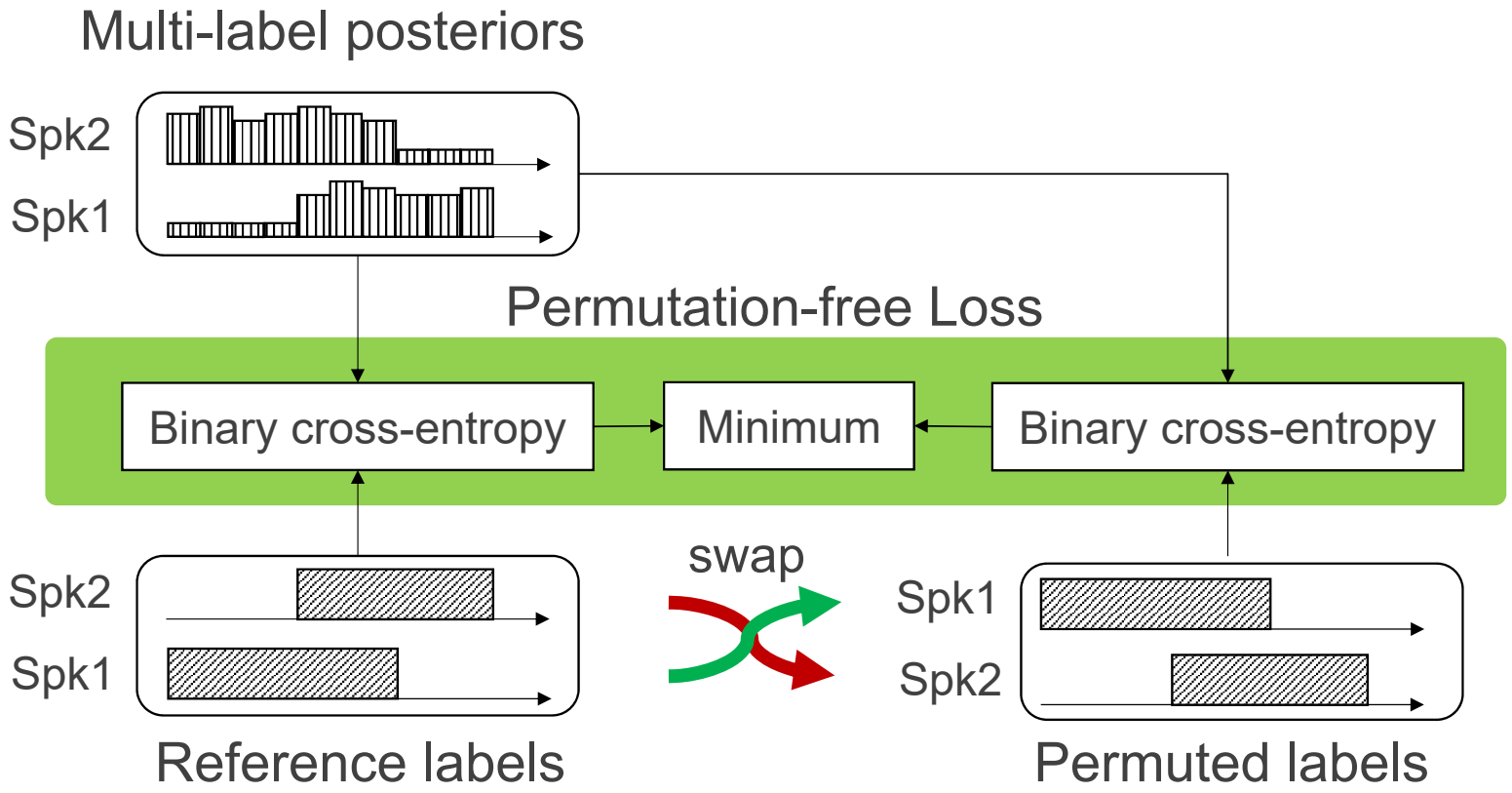


Fujita et al., "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," Interspeech 2019.

End-to-End Neural Diarization

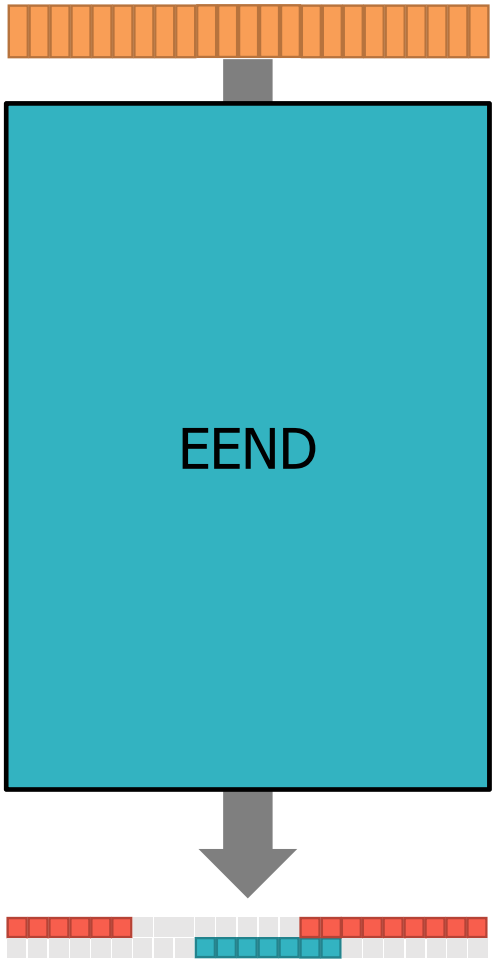


Permutation-free loss

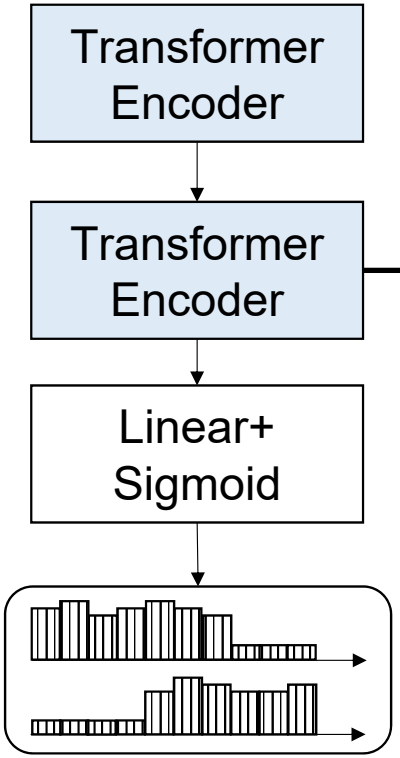


Y. Fujita et al., "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," Interspeech 2019.

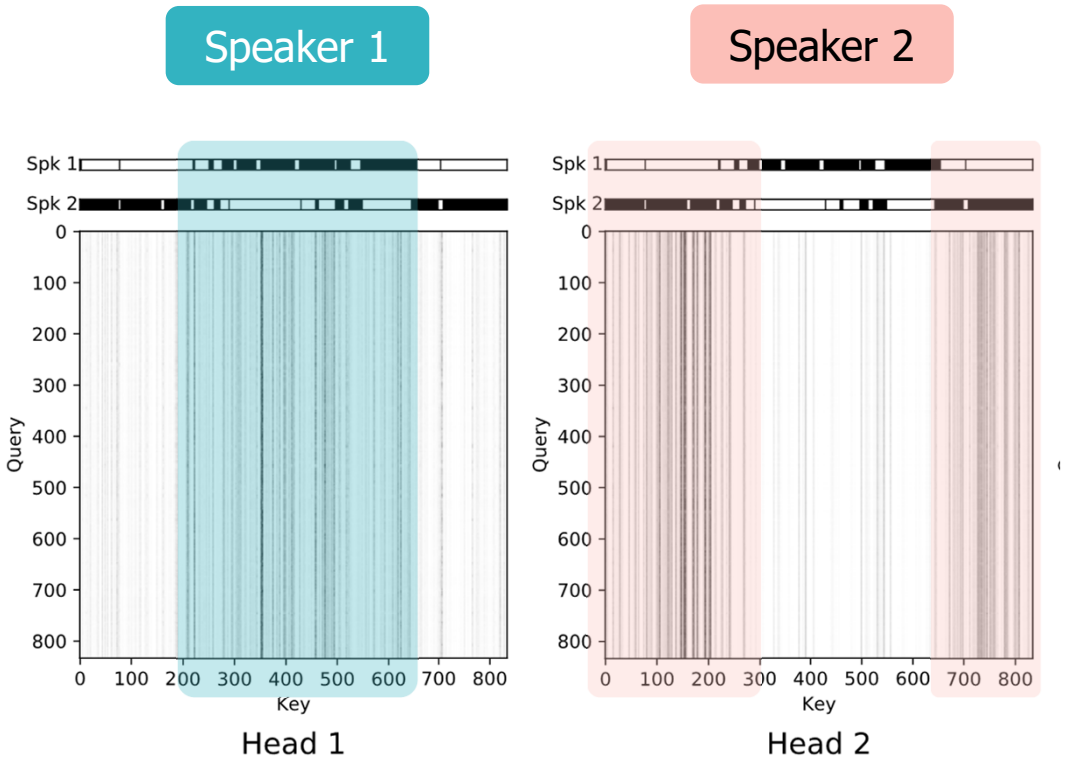
End-to-End Neural Diarization



Self-attention based network

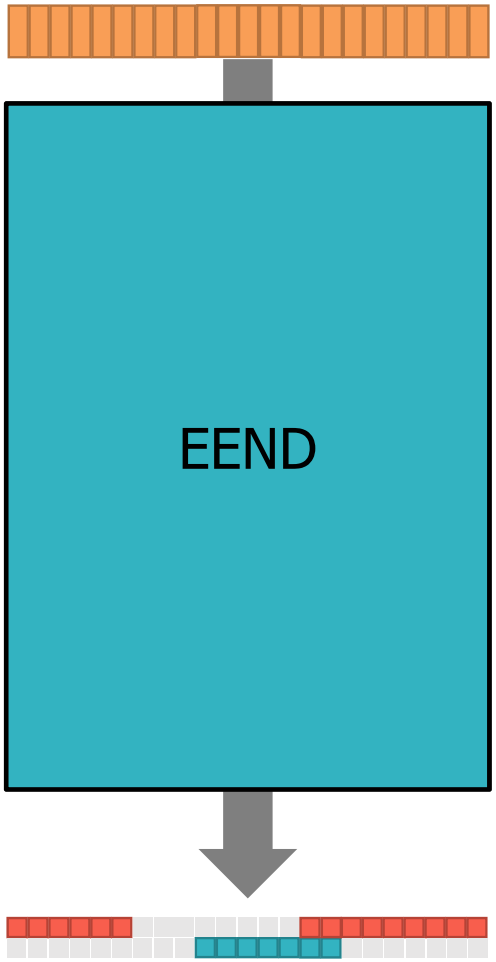


Attention weight matrix shows that one attention head focused on one speaker's activity.



Y. Fujita et al., "End-to-End Neural Speaker Diarization with Self-attention," ASRU 2019.

End-to-End Neural Diarization

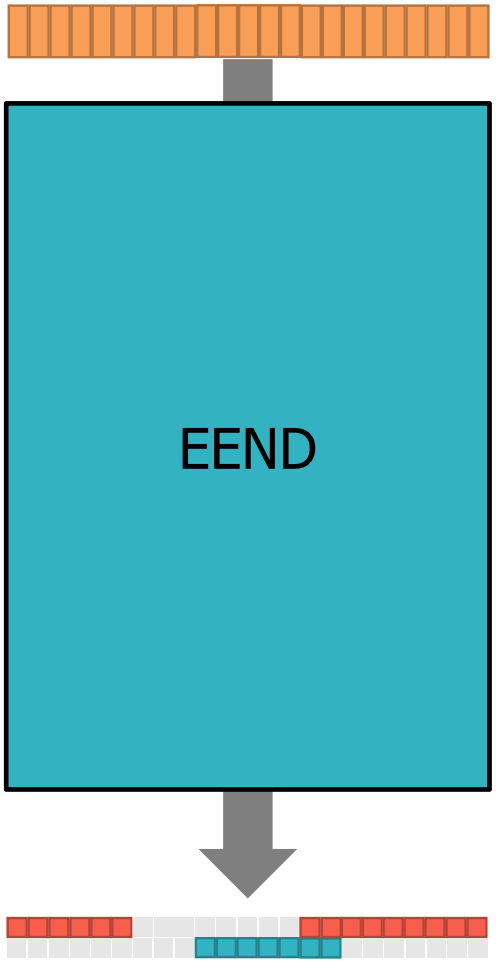


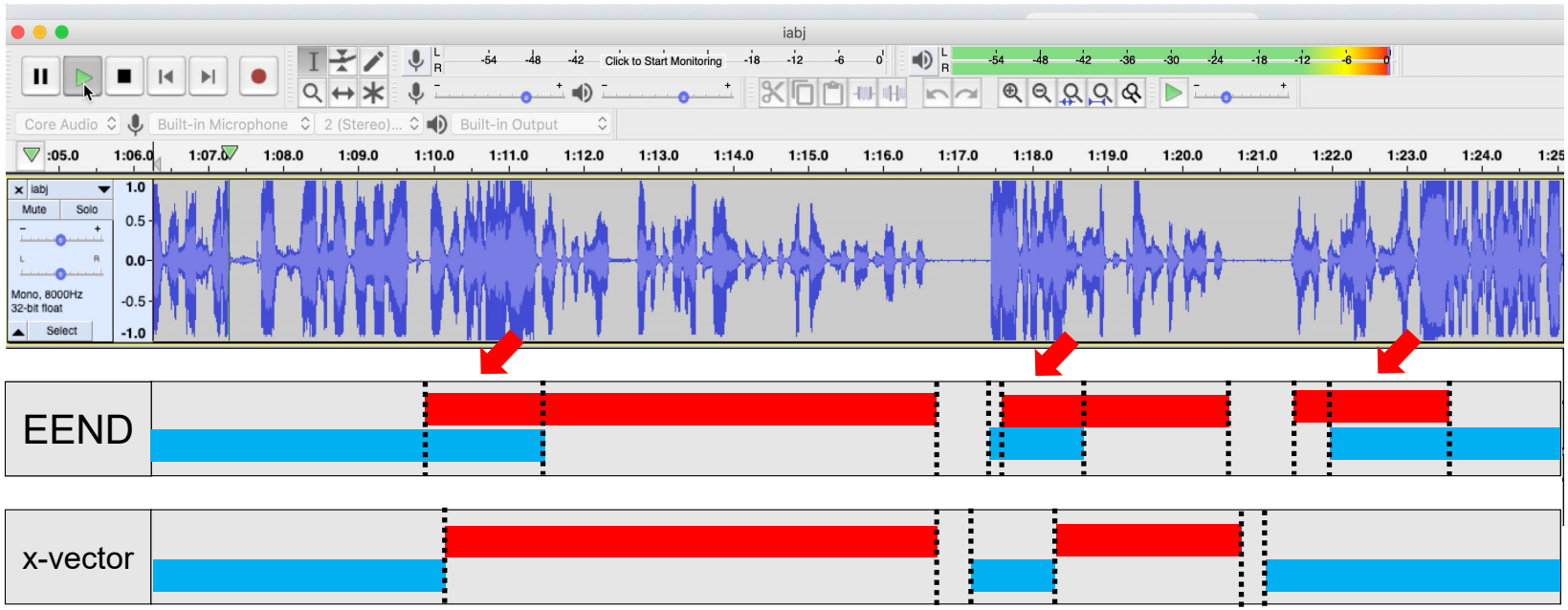
DERs on CALLHOME 2-speakers **with VAD and overlap errors.**

Method		Training Data	DER(%)
Clustering-based	i-vector	SWB2&SRE	12.10
	x-vector	SWB2&SRE	11.53
EEND	4-layer Self-attention	SimTrain (34.4%) → adaptation	**9.54**

Y. Fujita et al., "End-to-End Neural Speaker Diarization with Self-attention," ASRU 2019.

End-to-End Neural Diarization





☺ End-to-end training

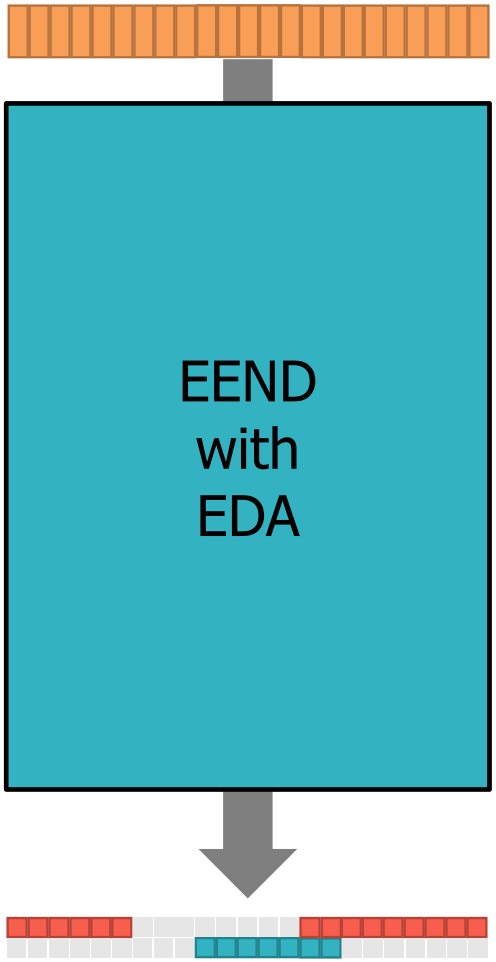
☺ Can handle overlapping speech

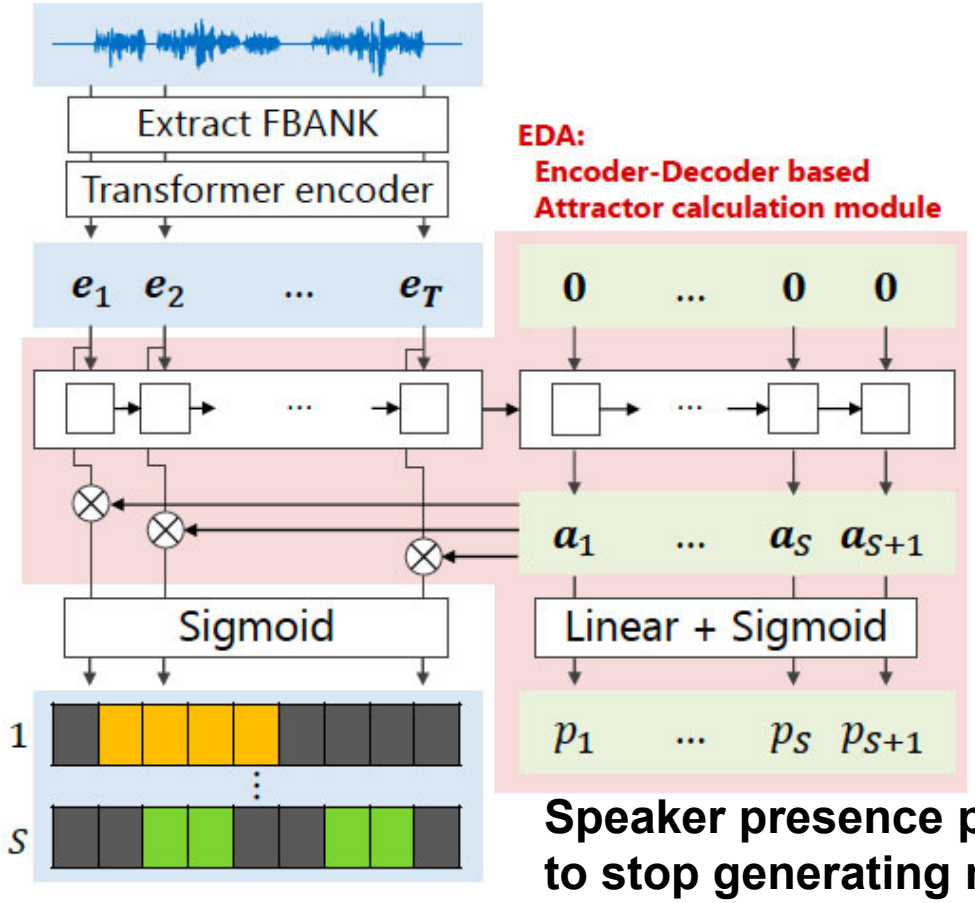
☹ Speaker counting

Top: waveform
Middle: EEND
Bottom: x-vector clustering

Y. Fujita et al., "End-to-End Neural Speaker Diarization with Self-attention," ASRU 2019.

EEND with Encoder-Decoder Attractor (EDA)



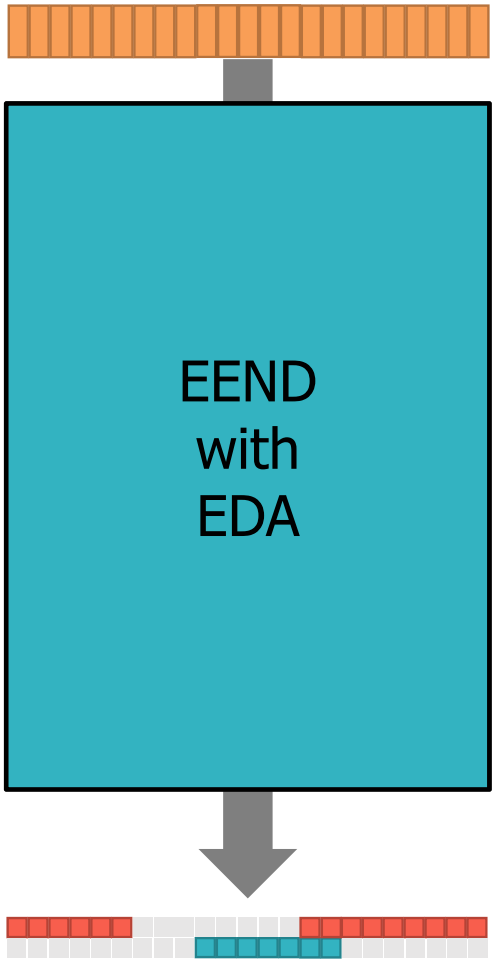


Speaker-wise generation of attractor (centroid) vectors

Speaker presence prob. to stop generating new speakers

S. Horiguchi et al., "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," Interspeech 2020.

EEND with Encoder-Decoder Attractor (EDA)



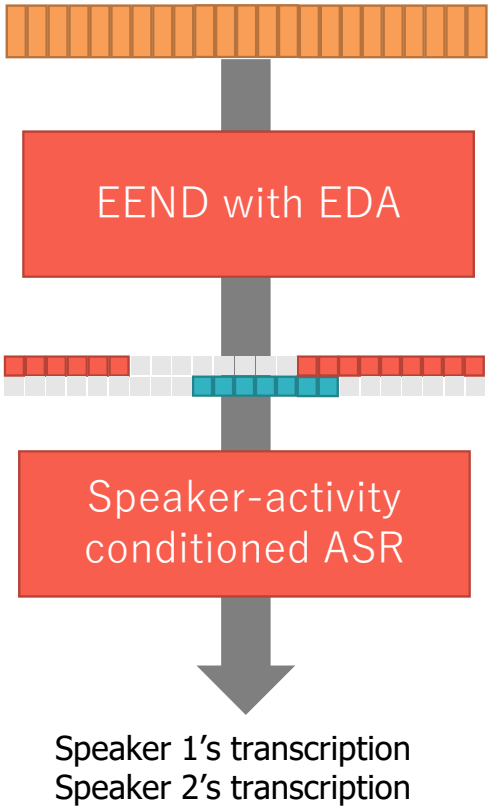
DERs on CALLHOME **2-6 speakers with VAD and overlap errors**

	Number of speakers					
Model	2	3	4	5	6	All
X-vector	15.45	18.01	22.68	**31.40**	**34.27**	19.43
EEND-EDA	**8.50**	**13.24**	**21.46**	33.16	40.29	**15.29**

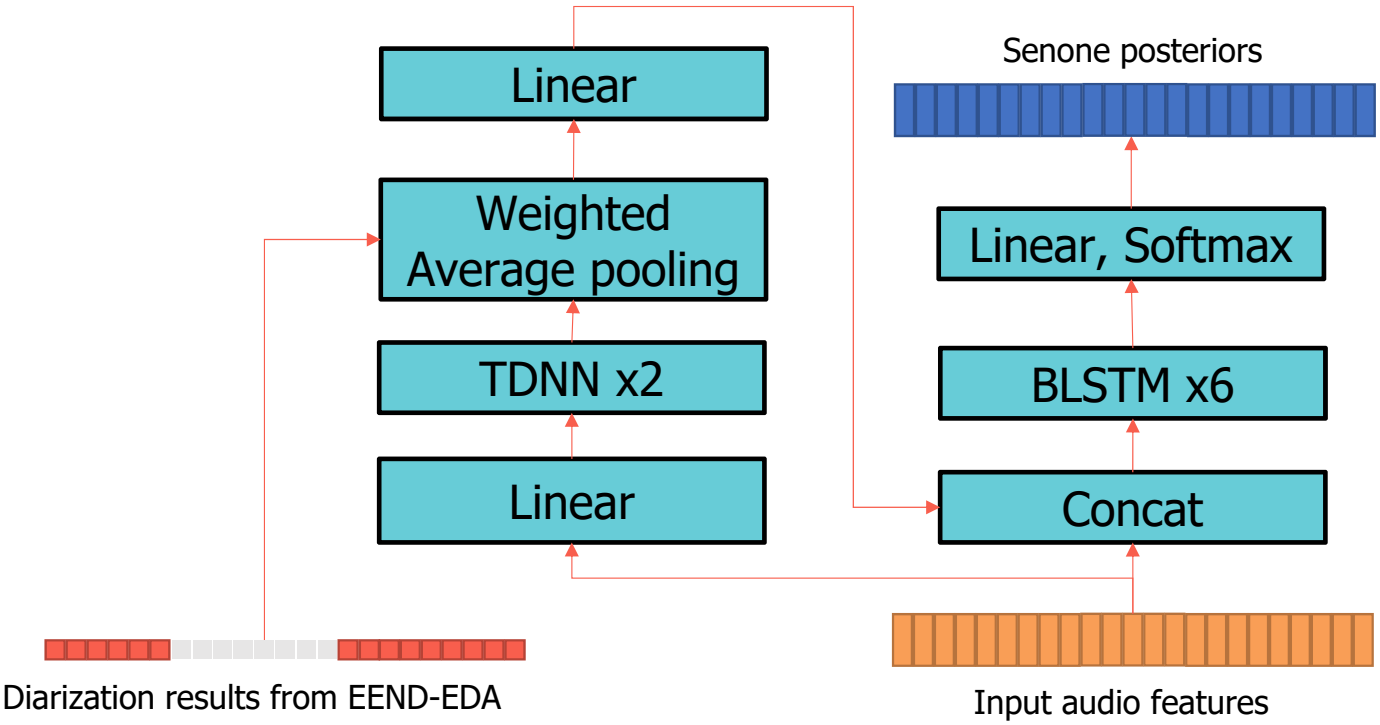
- ☺ End-to-end training
- ☺ Can handle overlapping speech
- ☺ Speaker counting

S. Horiguchi et al., "End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors," Interspeech 2020

Diarization-assisted ASR

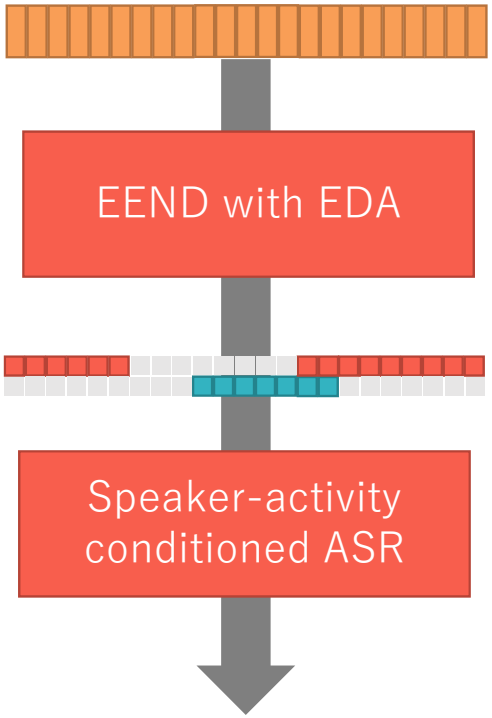


Diarization results are fed into the auxiliary network for conditioning ASR.



S. R. Chetupalli et al., "Speaker diarization assisted ASR for multi-speaker conversations," arXiv 2021.

Diarization-assisted ASR



SWERs on Hub 5 English evaluation speech

Model	w/ Ground-truth activity	w/ EEND-EDA
Single-speaker ASR	38.3	37.7
Diarization-assisted ASR	**25.8**	**26.5**

S. R. Chetupalli et al., "Speaker diarization assisted ASR for multi-speaker conversations," arXiv 2021.

DIHARD III challenge

The third speaker diarization challenge focusing on **"hard" diarization.**
Evaluation sets are drawn from a diverse conditions including
broadcast interviews, meeting speech, speech in restaurants, and YouTube videos.

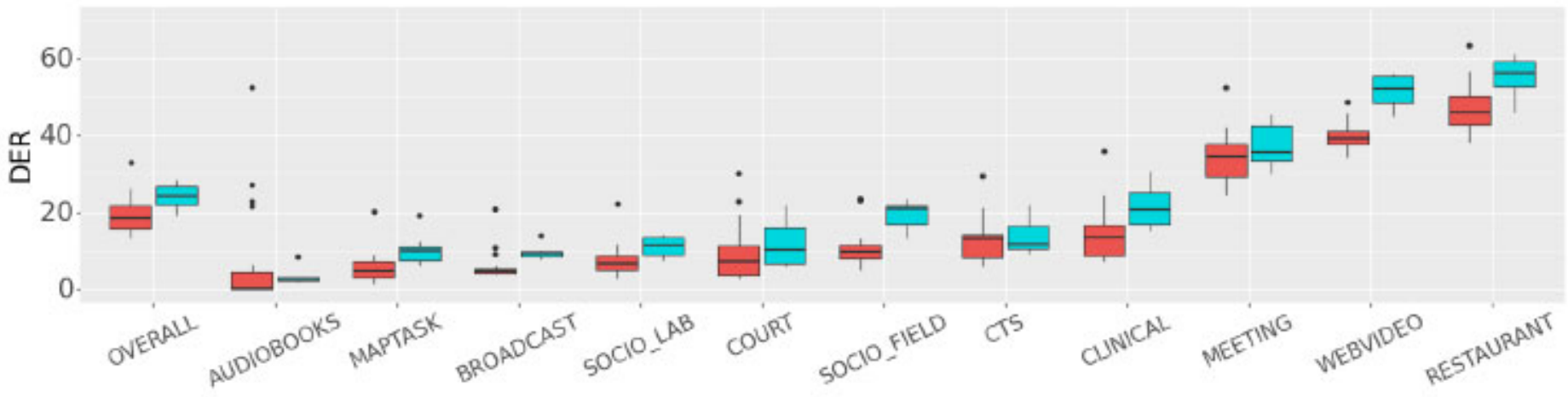
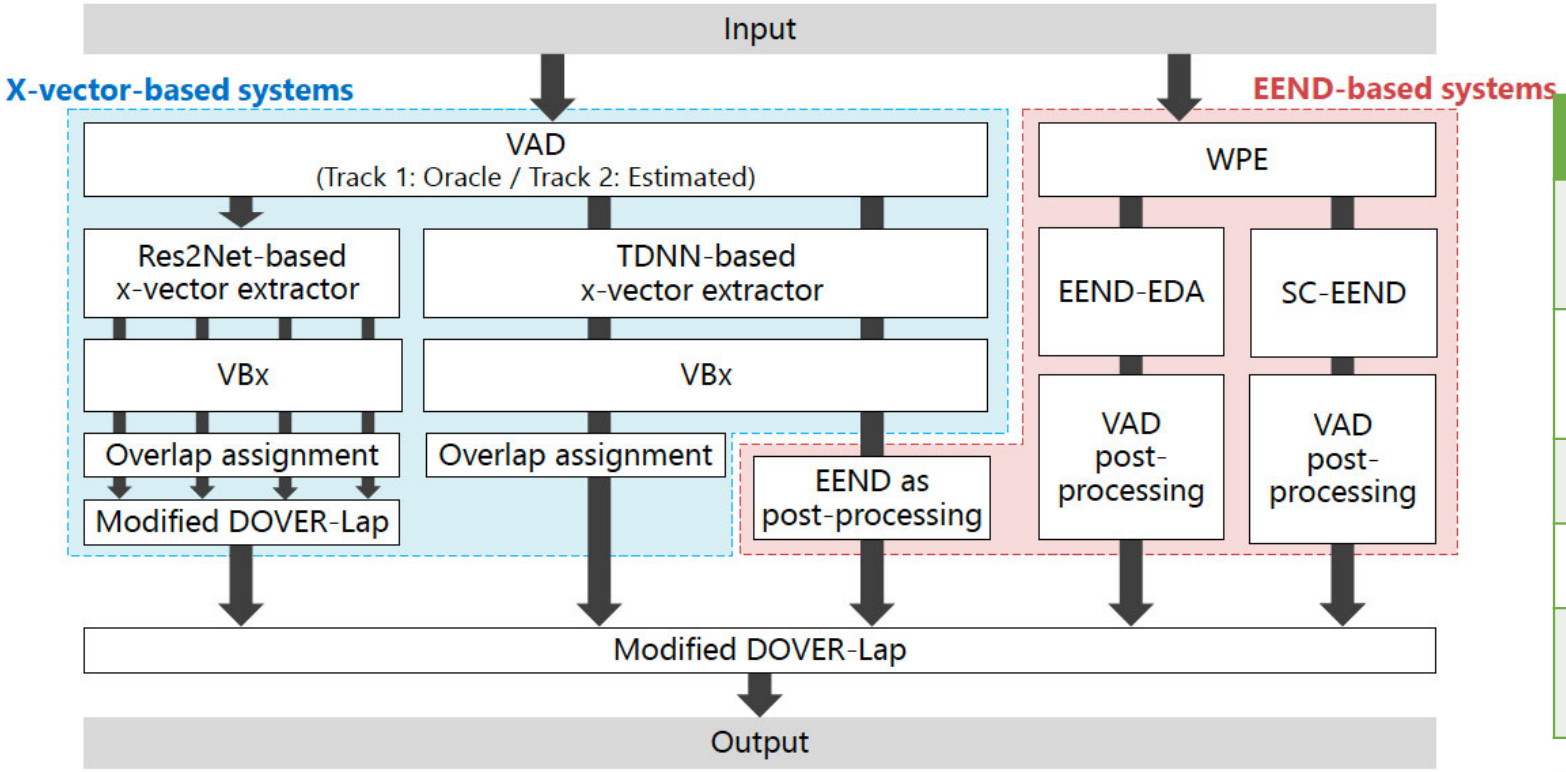


Figure 1: *Track 1 (red) and track 2 (blue) DER of primary submissions by domain on the core EVAL set.*

N. Ryant et al., "The Third DIHARD Diarization Challenge," arXiv 2021.

DIHARD III challenge

Hitachi/JHU system combines **clustering-based** and **end-to-end neural** diarization subsystems.



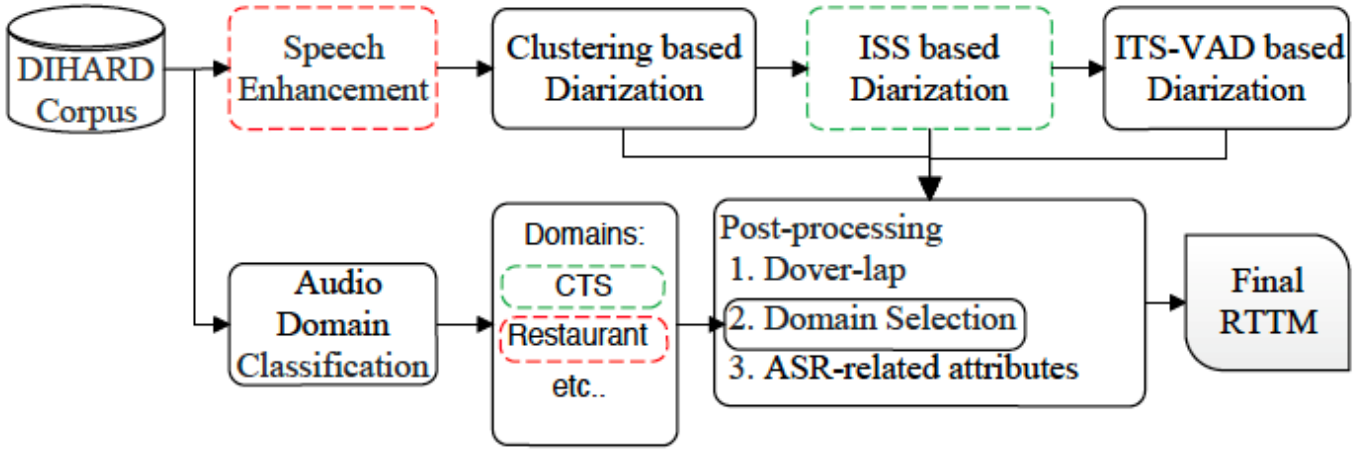
DERs on core set (Task 2)
with VAD and overlap errors

Model	Dev	Eval
DIHARD III baseline	22.28	27.34
Res2Net x-vector clustering	18.39	24.64
EEND-EDA	18.50	22.84
...		
DOVER-Lap of 5 systems	**15.81**	**20.01**

S. Horiguchi et al., "The Hitachi-JHU DIHARD III System: Competitive End-to-End Neural Diarization and X-Vector Clustering Systems Combined by DOVER-Lap, " DIHARD III workshop, 2021

DIHARD III challenge

USTC-NELSIP system trains **TS-VAD** and **source separation** models initialized by **clustering-based** diarization.



DERs on core set (Task 2)
with VAD and overlap errors

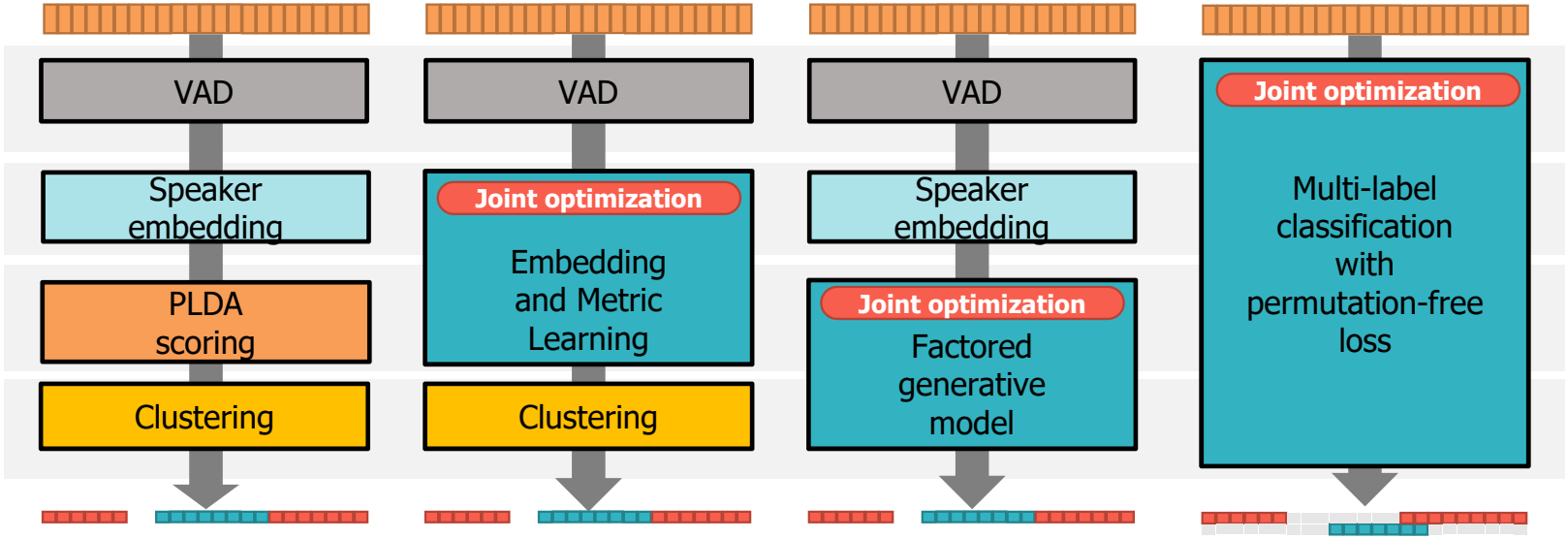
Model	Dev	Eval
DIHARD III baseline	22.28	27.34
USTC-NELSLIP	**14.49**	**19.37**

Domain Selection: Performance comparison among three methods.

Method	AUDIOBOOK	CTS	Restaurant	WebVideo	Others
Clustering	-	☹	☺	☺	☹
Source separation	-	☺	☹	☹	☹
TS-VAD	-	☺	☹	☹	☺

Y. Wang et al., "USTC-NELSLIP System Description for DIHARD III Challenge," DIHARD III workshop, 2021

Short summary of "Diarization + X"



- Reviewed recent techniques focusing on joint optimization in diarization systems.
- End-to-end training was made possible by replacing clustering module with a neural network.
- End-to-end systems exhibit strong performance in various domains of DIHARD III challenge.

- Problems on large number of speakers with frequent overlaps remain far from solved.
- Diarization→ASR: further joint optimal systems are possible.

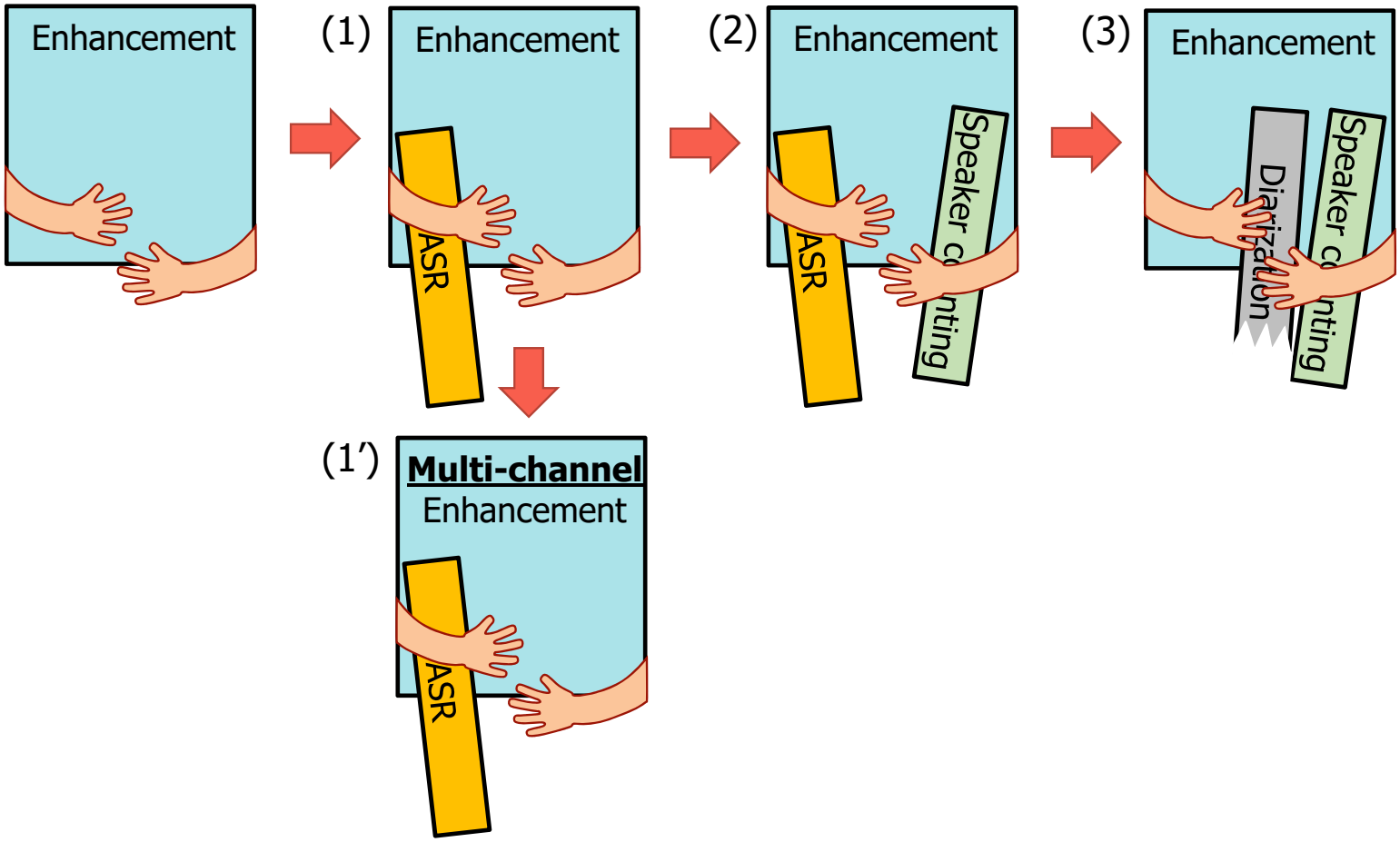
3. A new research trend: Jointly optimal systems

3.1. Diarization + x

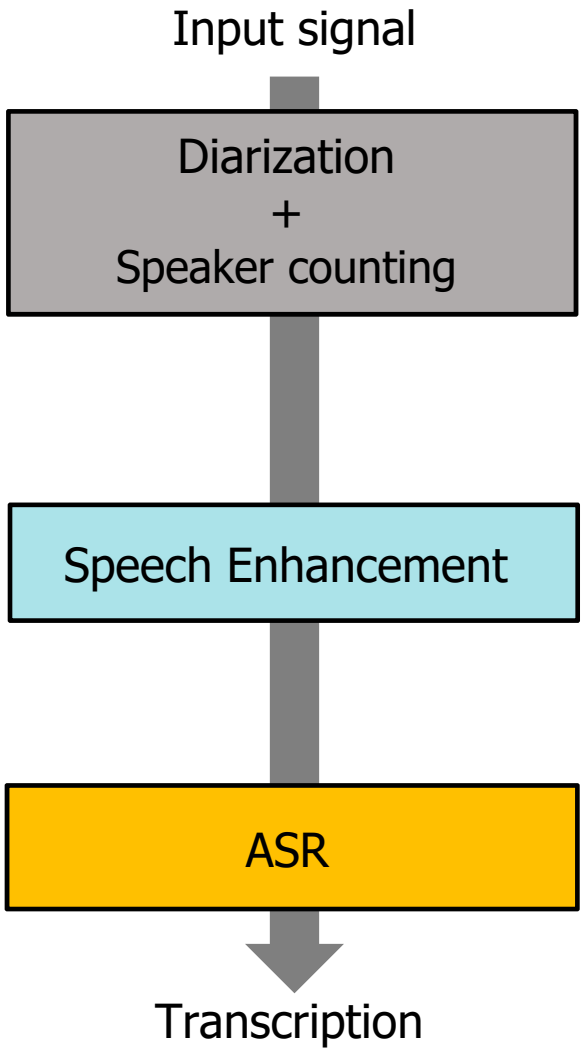
3.2. Enhancement + x

3.3. ASR + x

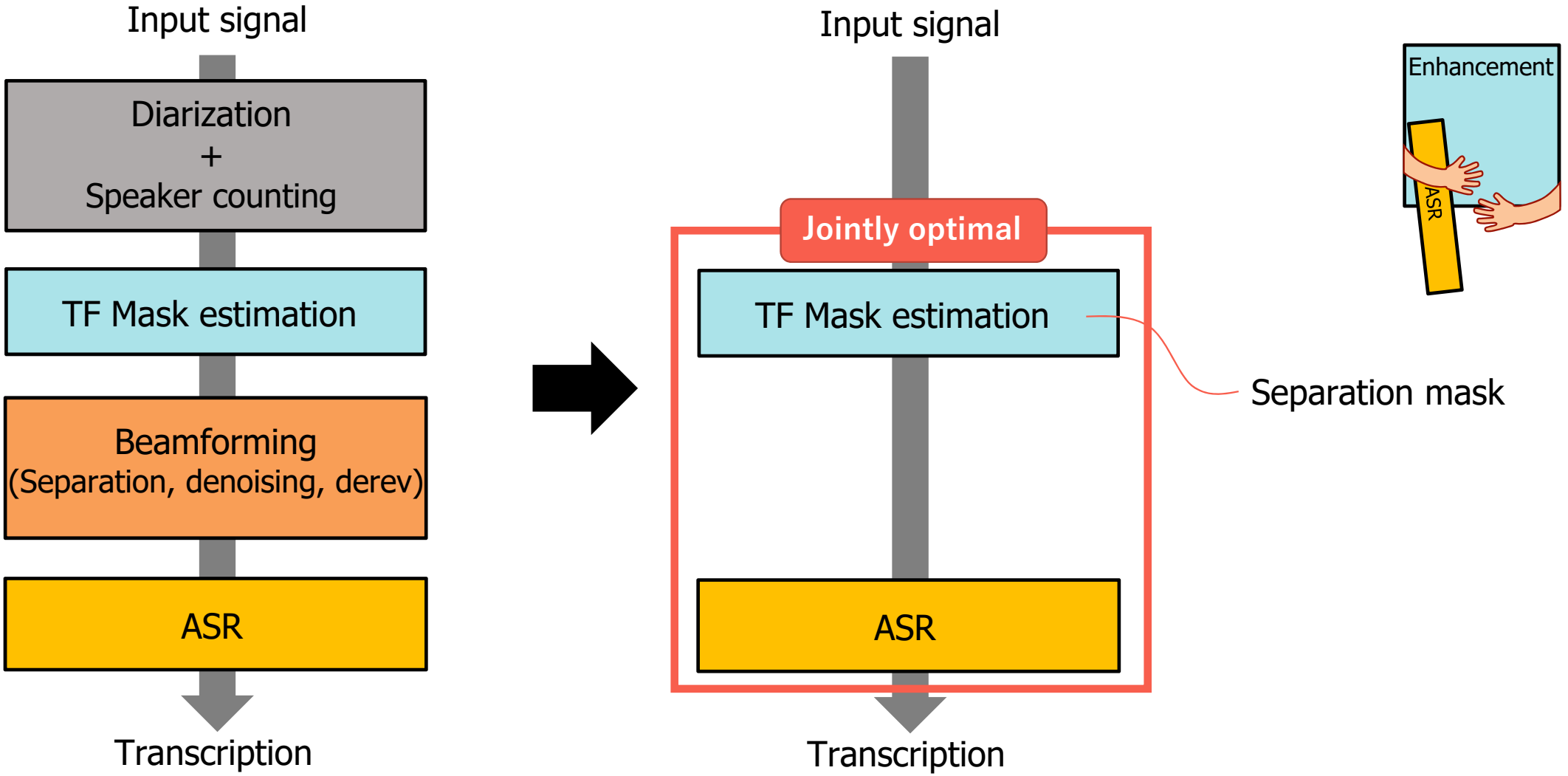
What has been achieved in "Enhancement + X"



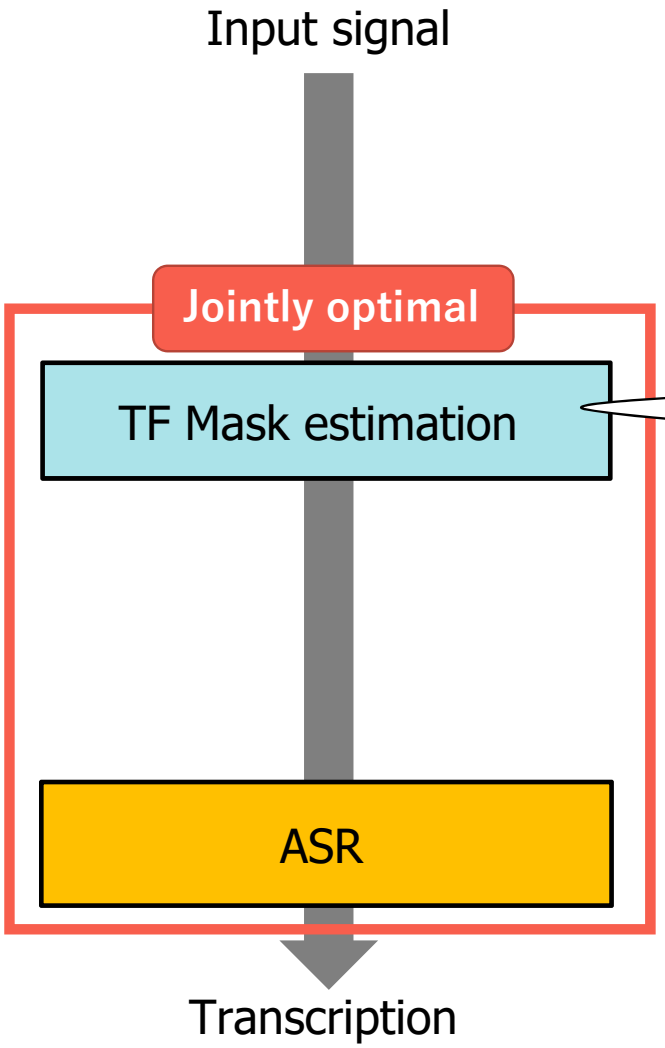
(1) TF Mask estimation + ASR



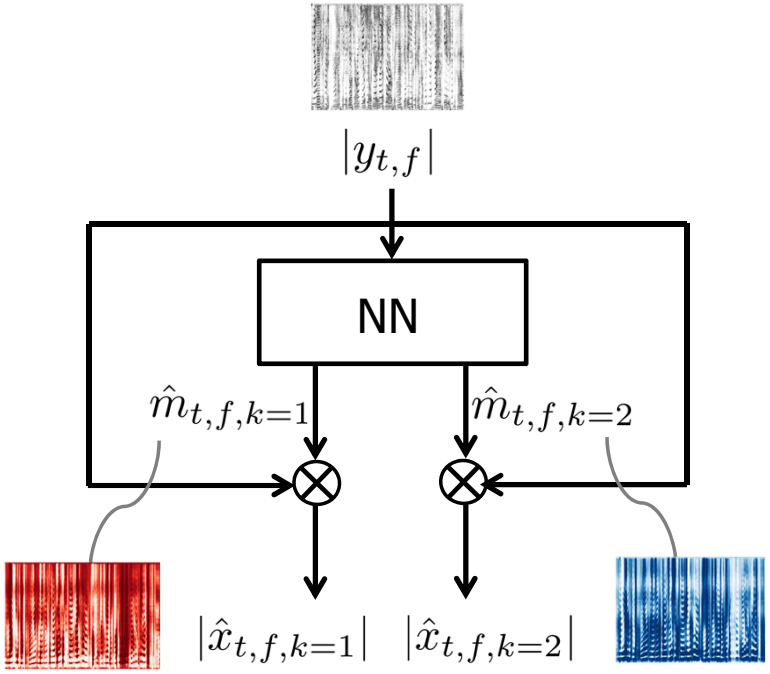
(1) TF Mask estimation + ASR



(1) TF Mask estimation + ASR: NN-based mask estimator



- Estimating TF masks (i.e., STFT-domain approach) [Yu+, 2017],
  or separated waveforms (i.e., time-domain approach) [Luo+, 2018]
  → NN-based source separation

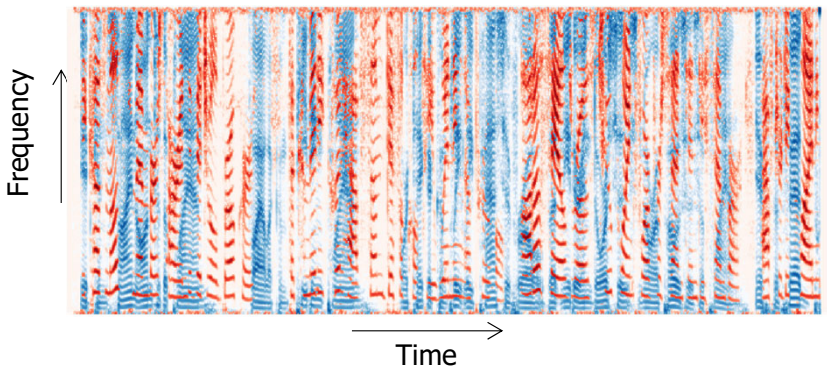


D. Yu et al., Permutation invariant training of deep models for speaker-independent multi-talker speech separation, ICASSP, 2017
Y. Luo et al., TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," ICASSP, 2018

(1) TF Mask estimation + ASR: NN-based mask estimator

Input signal

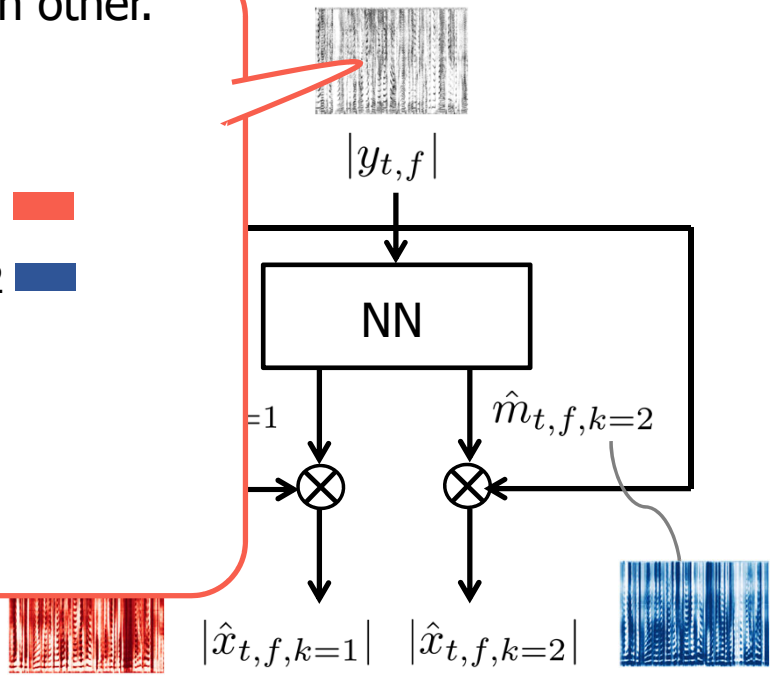
Speech signals are sparse, and rarely overlap each other.



ASR

Transcription

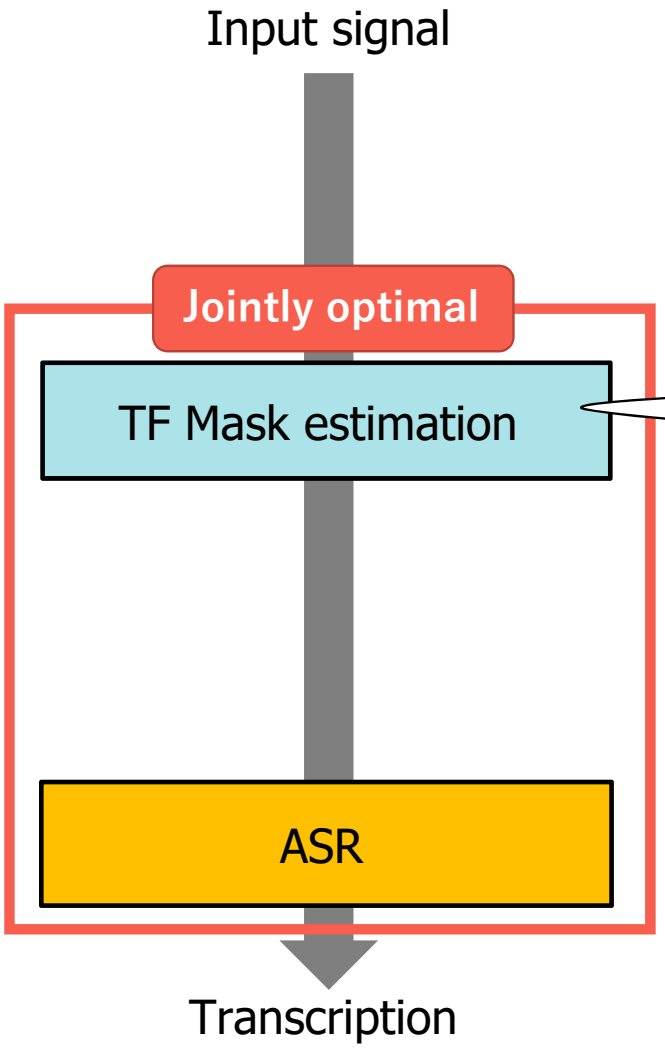
- Estimating TF masks (i.e., STFT-domain approach) [Yu+, 2017],
  or separated waveforms (i.e., time-domain approach) [Luo+, 2018]
  → NN-based source separation



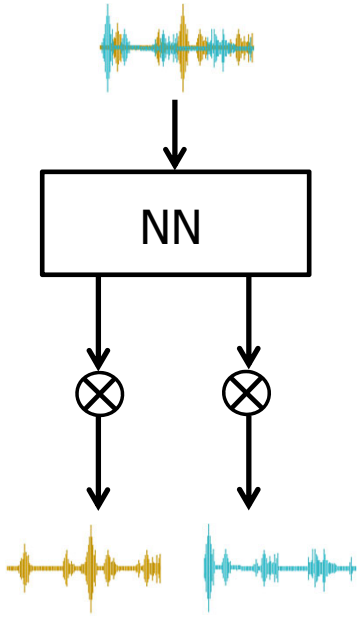
D. Yu et al., Permutation invariant training of deep models for speaker-independent multi-talker speech separation, ICASSP, 2017

Y. Luo et al., TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," ICASSP, 2018

(1) TF Mask estimation + ASR: NN-based mask estimator

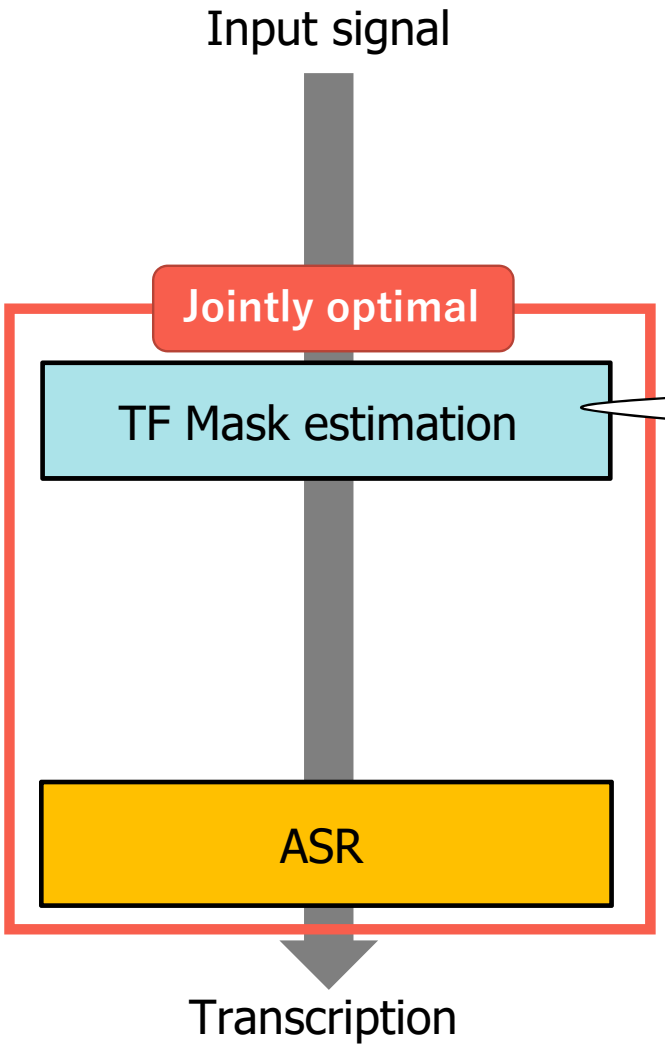


- Estimating TF masks (i.e., STFT-domain approach) [Yu+, 2017],
  or separated waveforms (i.e., time-domain approach) [Luo+, 2018]
  → NN-based source separation



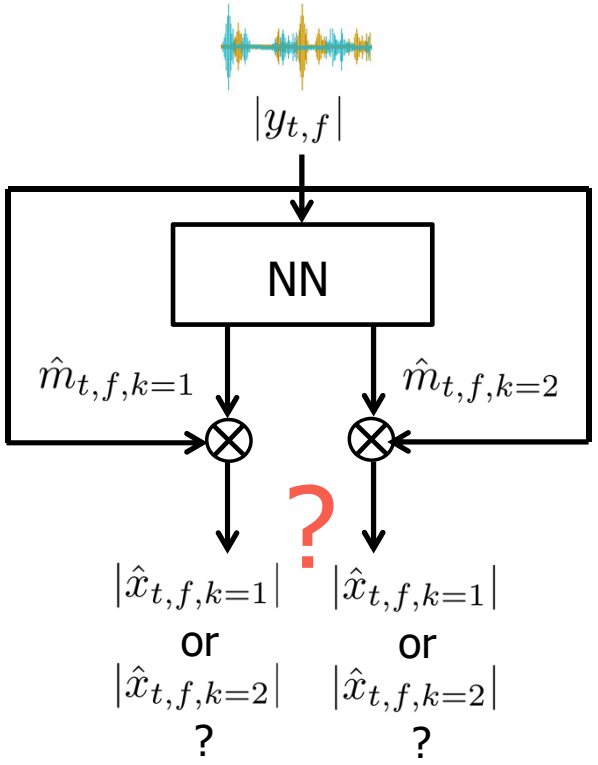
D. Yu et al., Permutation invariant training of deep models for speaker-independent multi-talker speech separation, ICASSP, 2017
Y. Luo et al., TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation," ICASSP, 2018

(1) TF Mask estimation + ASR: Label ambiguity problem



Output label ambiguity problem [Yu+, 2017] during the NN training

→ Permutation-invariant/Permutation-free training



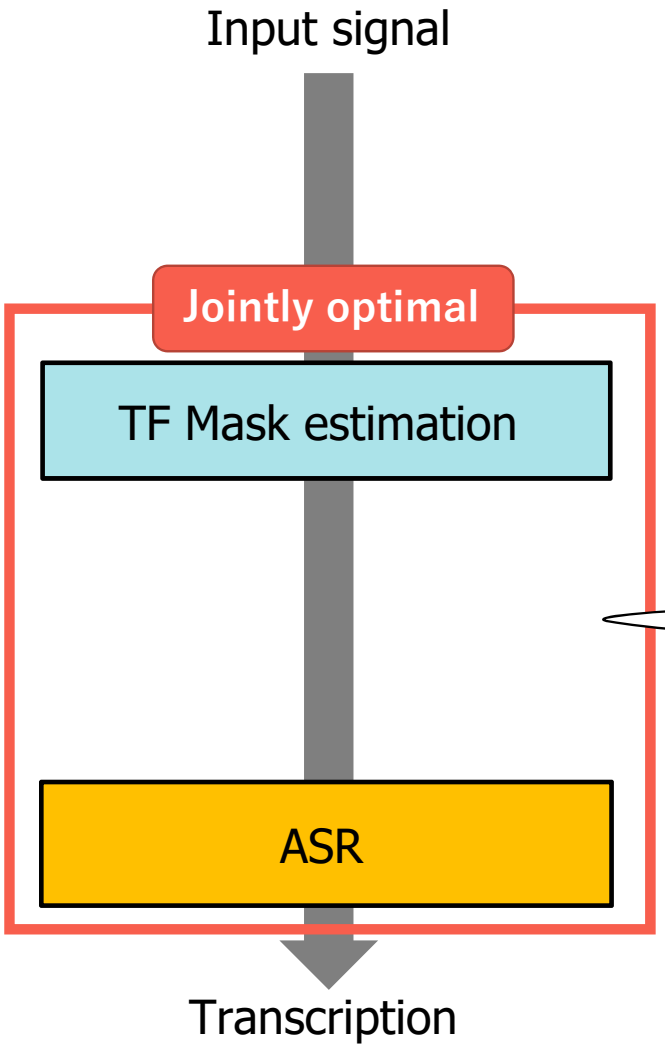
- Compute all permutations between the targets and the estimated sources, and find permutation $\phi^\star$ that minimizes the loss, e.g.,

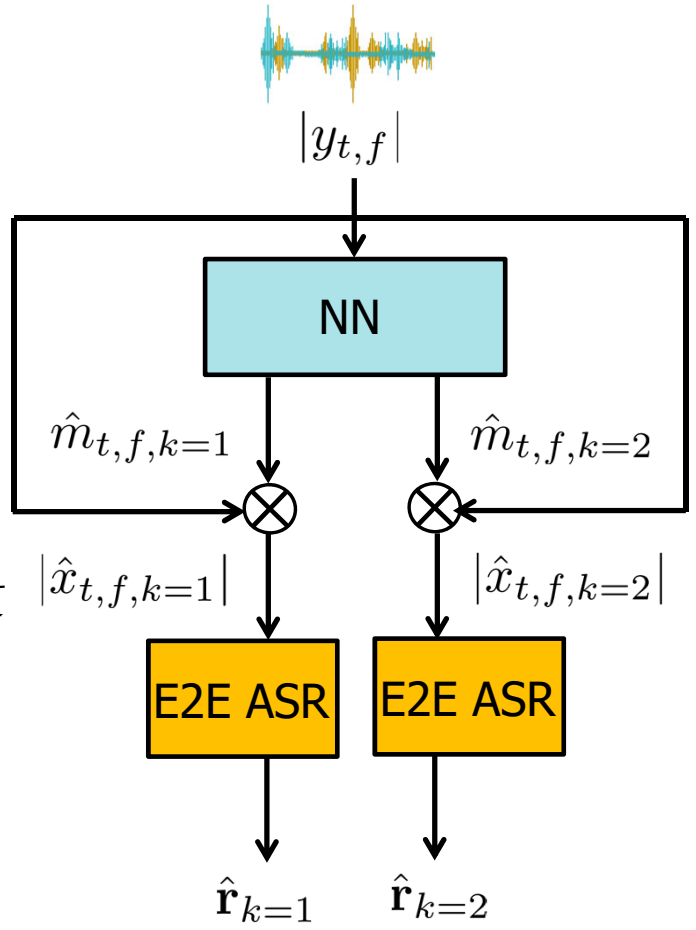
$$J^{\mathrm{ENH}}_{\phi^\star} = \min_{\phi \in \mathcal{P}} \sum_{t,f,k} \left| \hat{m}_{t,f,k} |y_{t,f}| - |x_{t,f,\phi(k)}| \right|^2$$

- The label ambiguity problem is common for various systems that handles multiple speakers simultaneously, such as ASR and diarization systems
- The above idea can be applied to to other loss functions, e.g., CER.

D. Yu et al., Permutation invariant training of deep models for speaker-independent multi-talker speech separation, ICASSP, 2017

(1) TF Mask estimation + ASR: Label ambiguity problem





[Training procedure]

Step 1:
Fix permutation based on signal-level loss, i.e.,

$$J^{\mathrm{ENH}}_{\phi^\star} = \min_{\phi\in\mathcal{P}} \sum_{t,f,k} \left| \hat{m}_{t,f,k}|y_{t,f}| - |x_{t,f,\phi(k)}| \right|^2$$

or,
on ASR transcription-level loss,

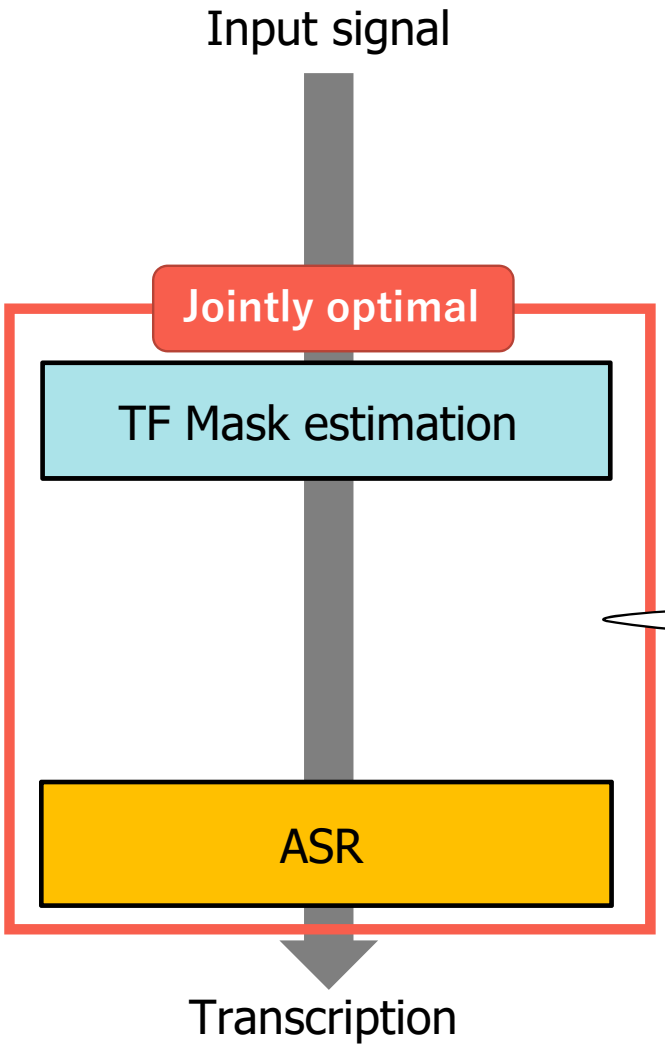
$$J^{\mathrm{ASR}}_{\phi^\star} = \min_{\phi\in\mathcal{P}} \sum_{k} J^{\mathrm{ASR}}_{\phi(k)}$$

Step 2:
Calculate the total loss with the obtained permutation $\phi^\star$:

$$J = \alpha J^{\mathrm{ASR}}_{\phi^\star} + \beta J^{\mathrm{ENH}}_{\phi^\star},$$

ASR system is aware of typical distortions included in the input feature.

(1) TF Mask estimation + ASR: Effectiveness



- Training and test data: WSJ-2mix (2 speaker full-overlap, noiseless, anechoic)

	Optimized part in joint system		SWER(%)
	front-end	back-end	
No proc.	-	-	79.1
STFT-domain separation	-	-	23.1
STFT-domain separation	-	✔	17.9
STFT-domain separation	✔	✔	13.2
Time-domain separation	-	-	22.9
Time-domain separation	-	✔	11.7
Time-domain separation	✔	✔	11.0

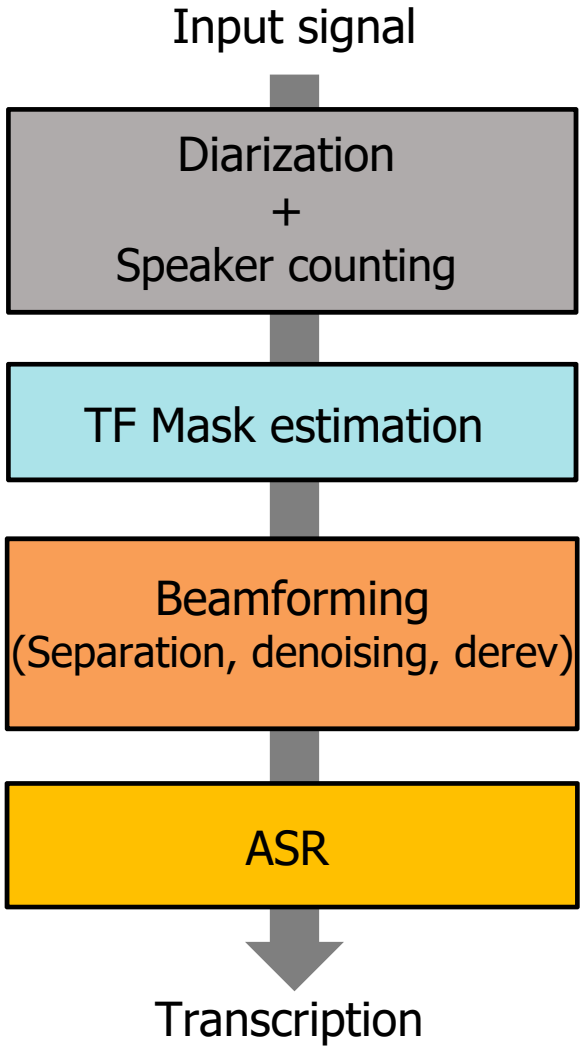
STFT-domain separation: Deep clustering+Mask inference [Settle+, 2018]
Time-domain separation: Conv-TasNet [von Neumann+, 2020]

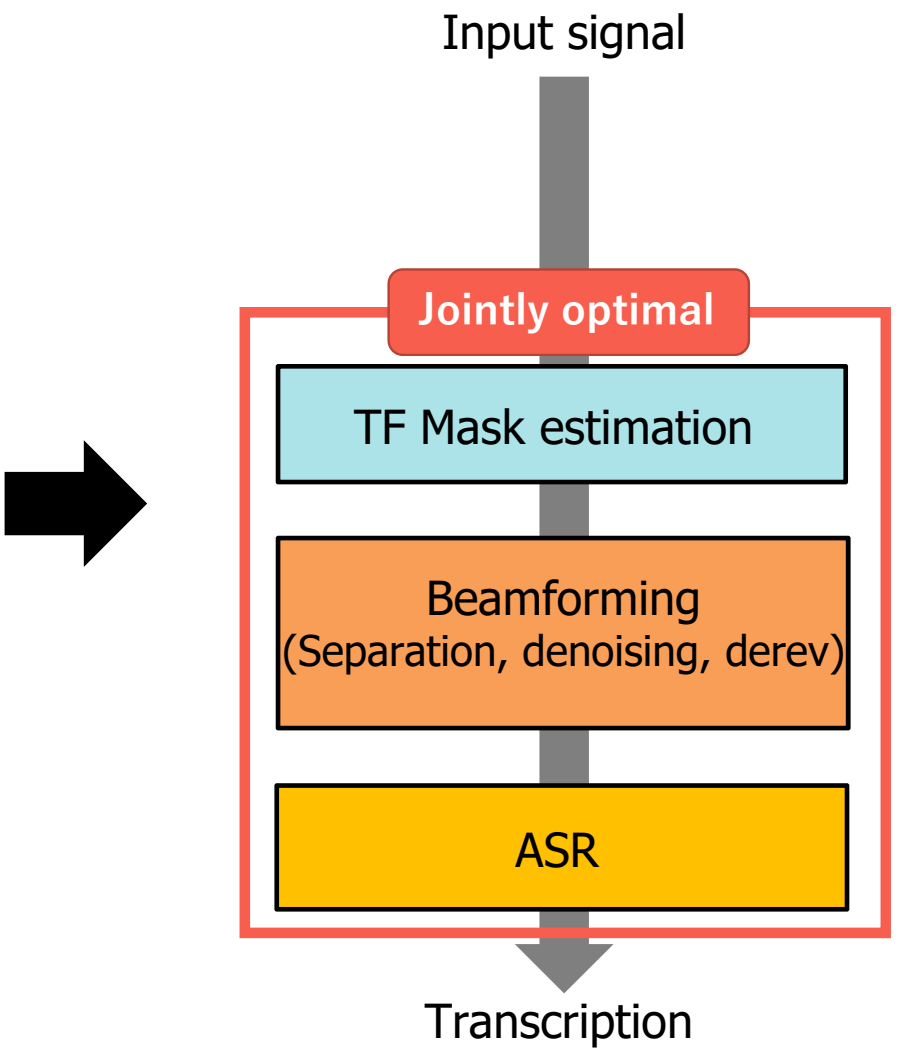
- Best results are obtained by jointly optimizing both front-end and back-end

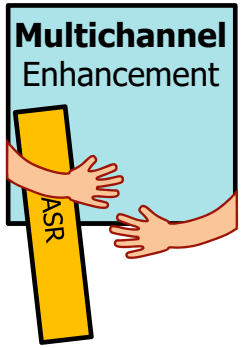
✔ Jointly optimal: 1ch separation + ASR

S. Settle et al., End-to-End Multi-Speaker Speech Recognition, ICASSP, 2018
T. von Neumann et al., End-to-end training of time domain audio separation and recognition, ICASSP, 2020

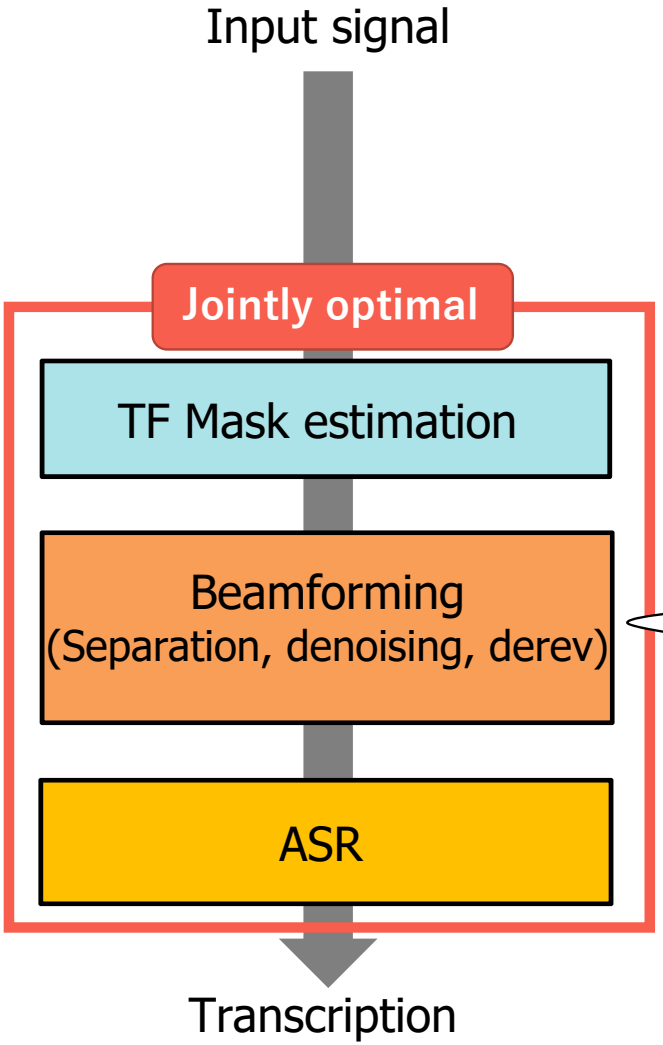
(1') TF Mask estimation + **Beamforming** + ASR

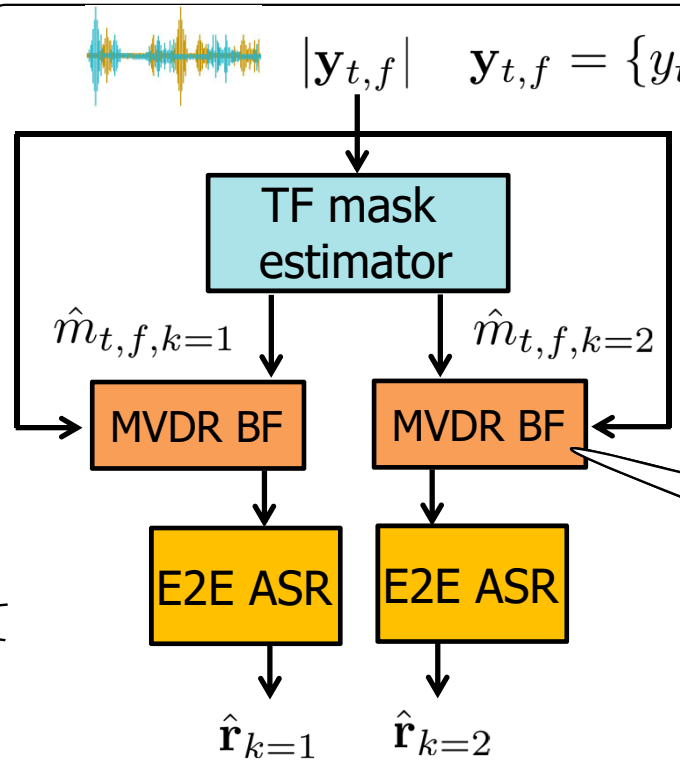






(1') TF Mask estimation + Beamforming + ASR





Step1: Spatial covariance estimation

$$\mathbf{\Phi}_{f,k} = \frac{1}{\sum_{t=1}^{T} \hat{m}_{t,f,k}} \sum_{t=1}^{T} \hat{m}_{t,f,k} \mathbf{y}_{t,f} \mathbf{y}_{t,f}^H$$

Step2: Filter calculation

$$\mathbf{g}_{f,k} = \frac{(\sum_{k' \neq k} \mathbf{\Phi}_{f,k'})^{-1} \mathbf{\Phi}_{f,k}}{\mathrm{Tr}((\sum_{k' \neq i} \mathbf{\Phi}_{f,k'})^{-1} \mathbf{\Phi}_{f,k})} \mathbf{u}$$

Step3: Filtering

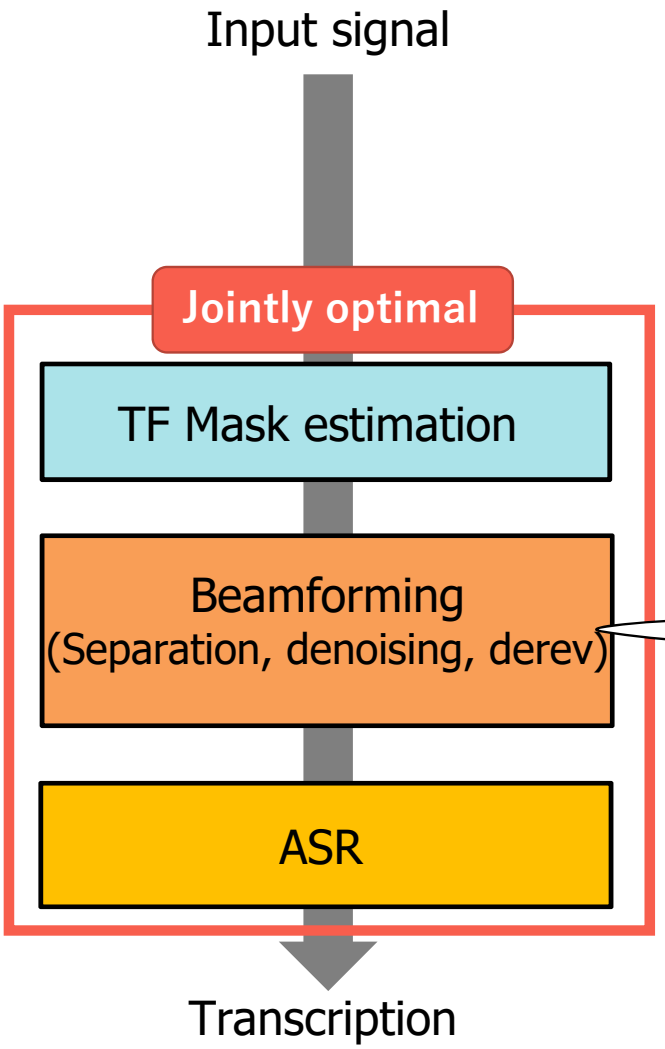
$$\hat{x}_{t,f,k} = \mathbf{g}_{f,k}{}^H \mathbf{y}_{t,f}$$

MVDR BF contains only linear operations and thus is differentiable with respect to the estimated mask $\hat{m}_{t,f,k}$

➡ Both ASR and enhancement modules (mask estimator + BF) can be jointly optimized with an ASR criterion.

J. Heymann et al., BeamNet: End-to-end training of a beamformer-supported multi-channel ASR system, ICASSP, 2017
X. Chang et al., MIMO-Speech: End-toend multi-channel multi-speaker speech recognition, ASRU, 2019

(1') TF Mask estimation + Beamforming + ASR



- Training and test data: noiseless spatialized WSJ-2mix (full-overlap, anechoic)

Enhancement	ASR model	# of Mics	Joint optimization	SWER (%)
Adaptive BF (BeamformIt)	1-speaker ASR model	2	-	98.0
MVDR BF	1-speaker ASR model	2	✔	7.6

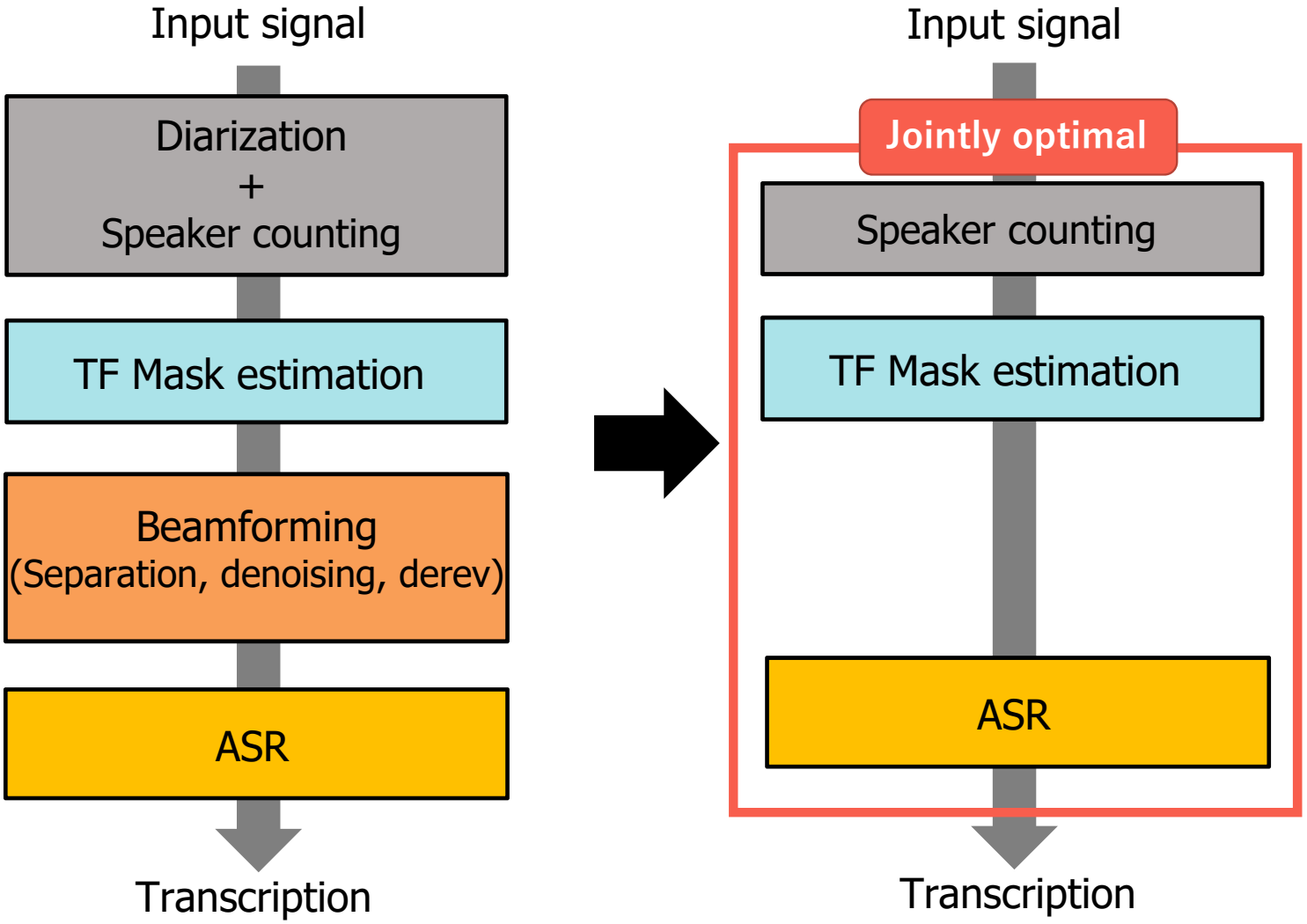
- Multichannel jointly optimal system is significantly more advantageous in comparison with multichannel non-optimal cascade systems. [Chang+, 2019]

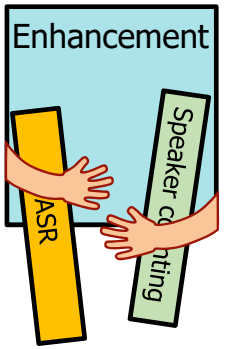
- Similar results are obtained with reverberant WSJ-2mix, with a jointly optimal multichannel derev. + MVDR separation + ASR system [Zhang+, 2020]

✔ Jointly optimal: multi-ch enhancement + ASR

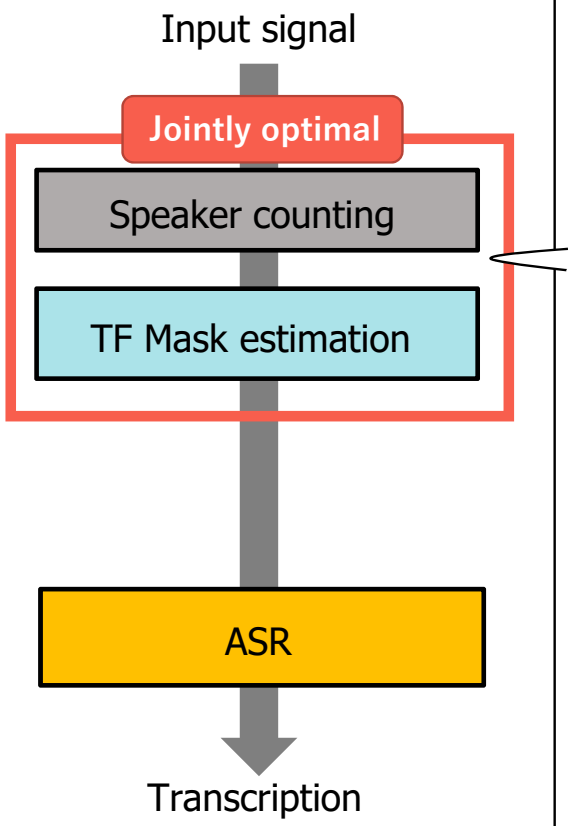
X. Chang et al., MIMO-Speech: End-toend multi-channel multi-speaker speech recognition, ASRU, 2019
W. Zhang et al. End-to-end dereverberation, beamforming, and speech recognition with improved numerical stability and advanced frontend, ICASSP, 2020

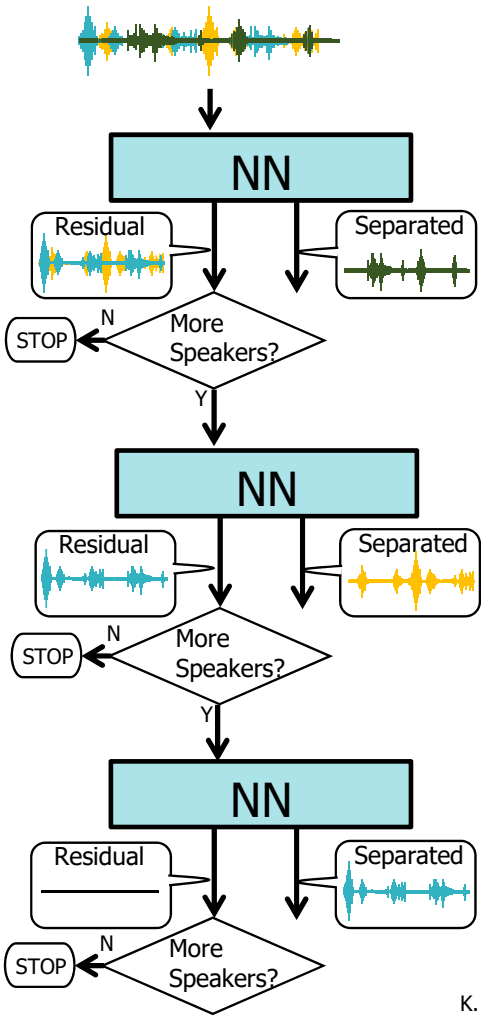
(2) Speaker counting + TF Mask estimation + ASR





(2) Speaker counting + TF Mask estimation + ASR





- Conventional NN source separation
  → Separation in one-go
  → Number of speakers it can deal with is bounded by the number of output nodes

- Extracting speakers recursively to perform jointly optimal separation and speaker counting. [Kinoshita+ 2018]

- Training of the NNs e.g., in case of time-domain separation [Takahashi+ 2019]

$$J_{\phi^\star}^{\text{Enh+Count}} = J_{\phi^\star}^{\text{Enh}} + J^{\text{Count}}$$

$$J_{\phi^\star}^{\text{ENH}} = \min_{\phi \in \mathcal{P}} \mathcal{L}^{\text{SDR}}(x_{\phi(k)}[\tilde{t}], \hat{x}_1[\tilde{t}]) + \mathcal{L}^{\text{SDR}}\Big(\sum_{k' \neq \phi(k)} x_{k'}[\tilde{t}], \hat{x}_2[\tilde{t}]\Big)$$

Reconstruction of separated signal

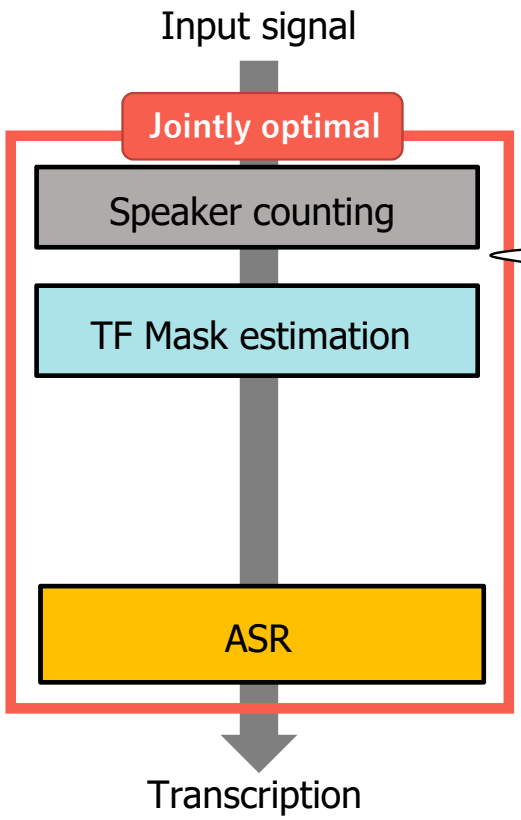
Reconstruction of residual signal

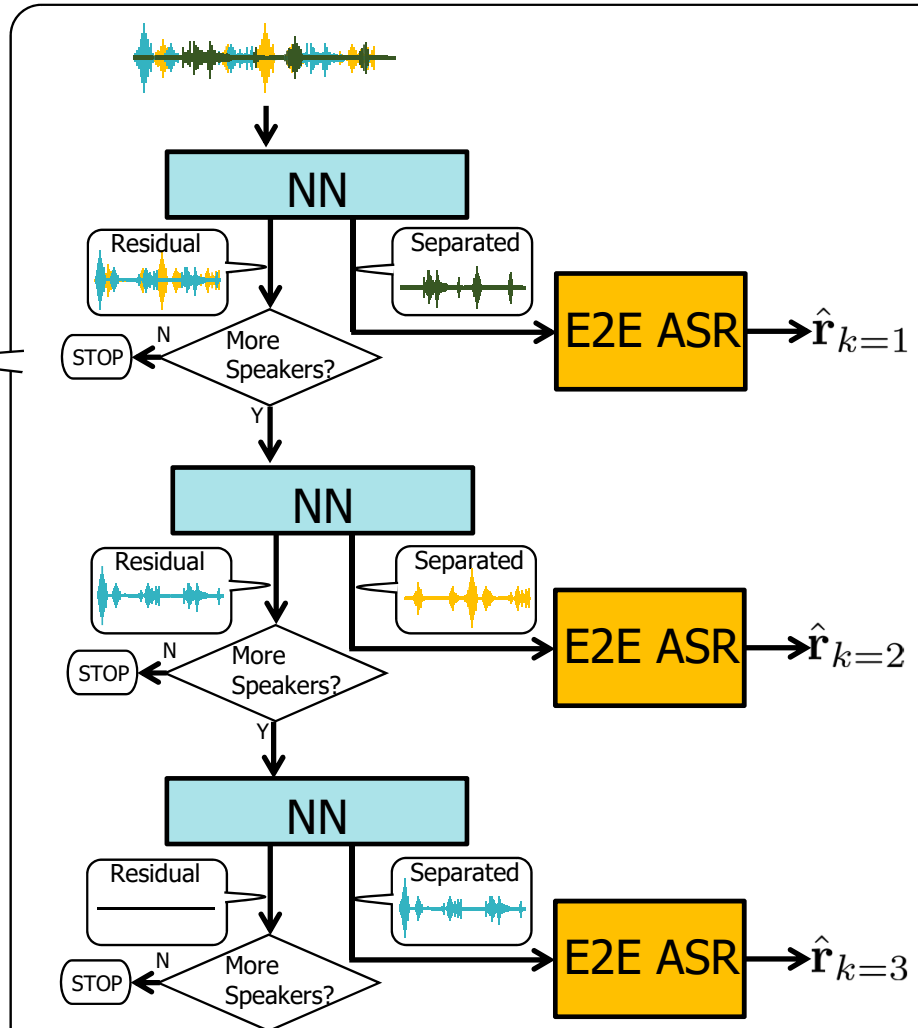
$$J^{\text{Count}} = -f \log(\hat{f}) - (1-f)\log(1-\hat{f}) \qquad f = \begin{cases} 1 & (no\ speech) \\ 0 & (speech) \end{cases}$$

A stop flag loss to tell whether the residual contains speech

K. Kinoshita et al., Listening to each speaker one by one with recurrent selective attention network, ICASSP, 2018
N. Takahashi et al., Recursive speech separation for unknown number of speakers, Interspeech, 2019

(2) Speaker counting + TF Mask estimation + ASR





[Training procedure]

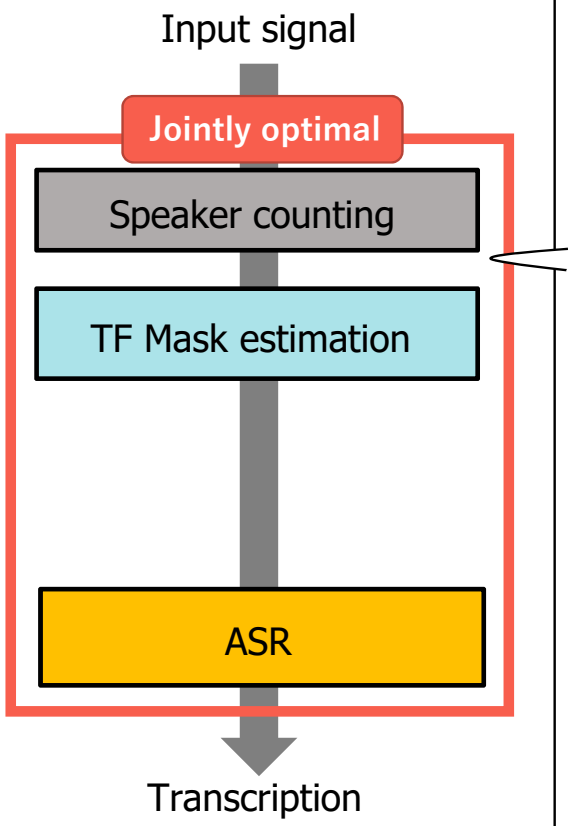
Step 1:
Fix permutation based on signal-level loss,
or,
on ASR transcription-level loss.

Step 2:
Calculate loss with the obtained permutation as:

$$J = \alpha J^{\mathrm{ASR}}_{\phi^\star} + \beta J^{\mathrm{Enh+Counting}}_{\phi^\star},$$

T. von Neumann et al., Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR, Interspeech, 2020

(2) Speaker counting + TF Mask estimation + ASR



- Data: simulated mixture. 2- to 4-speaker fully overlapped speech.
  No noise and reverb.

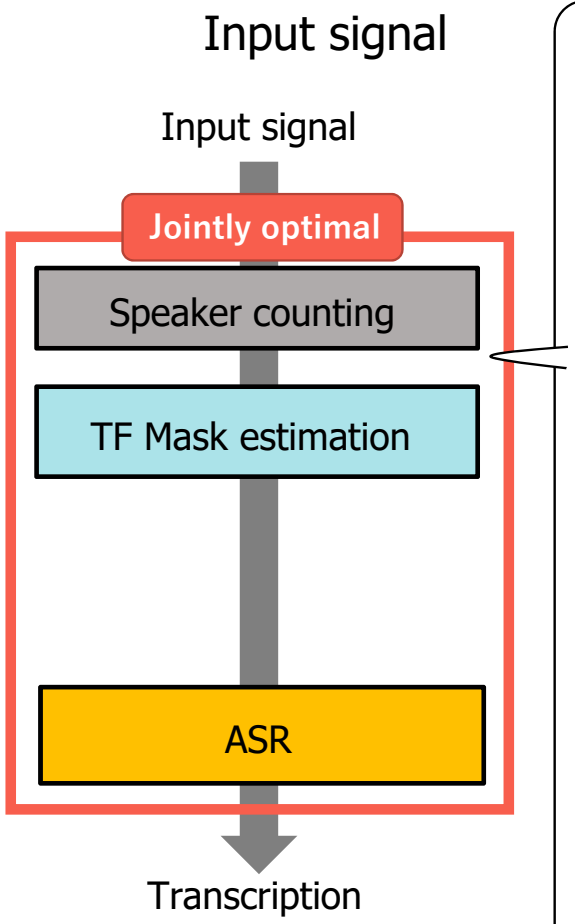
[Source counting accuracy (%)]

Model	Training data	Number of speakers			
		1	2	3	4
One-pass model (3 speaker output)	2 to 3 speakers	0.0	99.9	99.4	-
Recursive model	2 to 3 speakers	100.0	100.0	96.7	91.9

- Recursive separation
  - Accurate counting accuracy for seen cases i.e., 1- and 3-speaker cases
  - Could also handle unseen cases (4-speaker cases)

T. von Neumann et al., Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR, Interspeech, 2020

(2) Speaker counting + TF Mask estimation + ASR



- Data: simulated mixture. 2- to 4-speaker fully overlapped speech.
  No noise and reverb.

[ASR performance (SWER (%))]

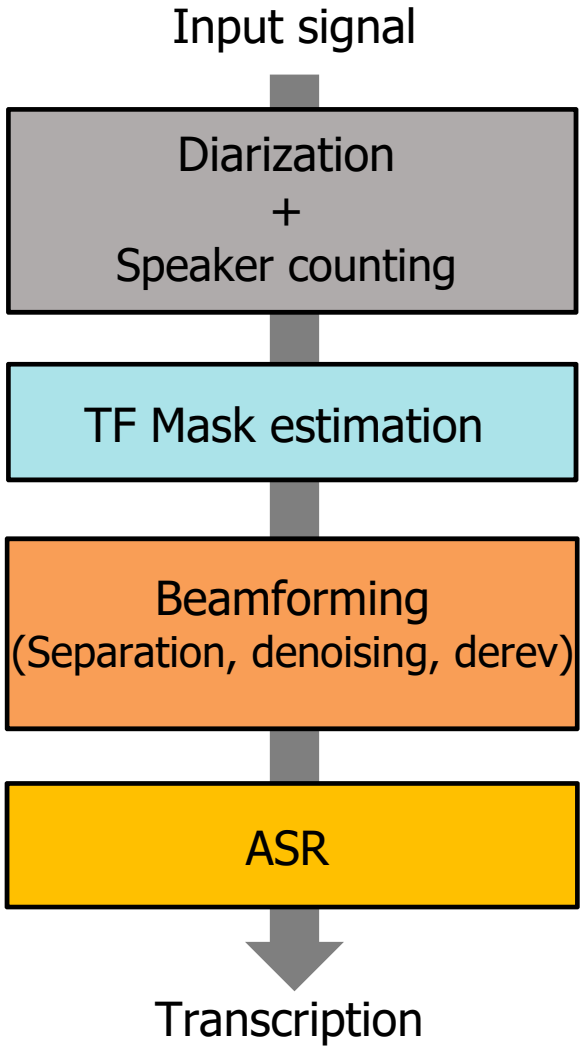
Model	Training data	Number of speakers		
		2	3	4
One-pass model (3 speaker output)	2 to 3 speakers	10.5	34.9	-
Recursive model	2 to 3 speakers	8.7	20.5	45.2

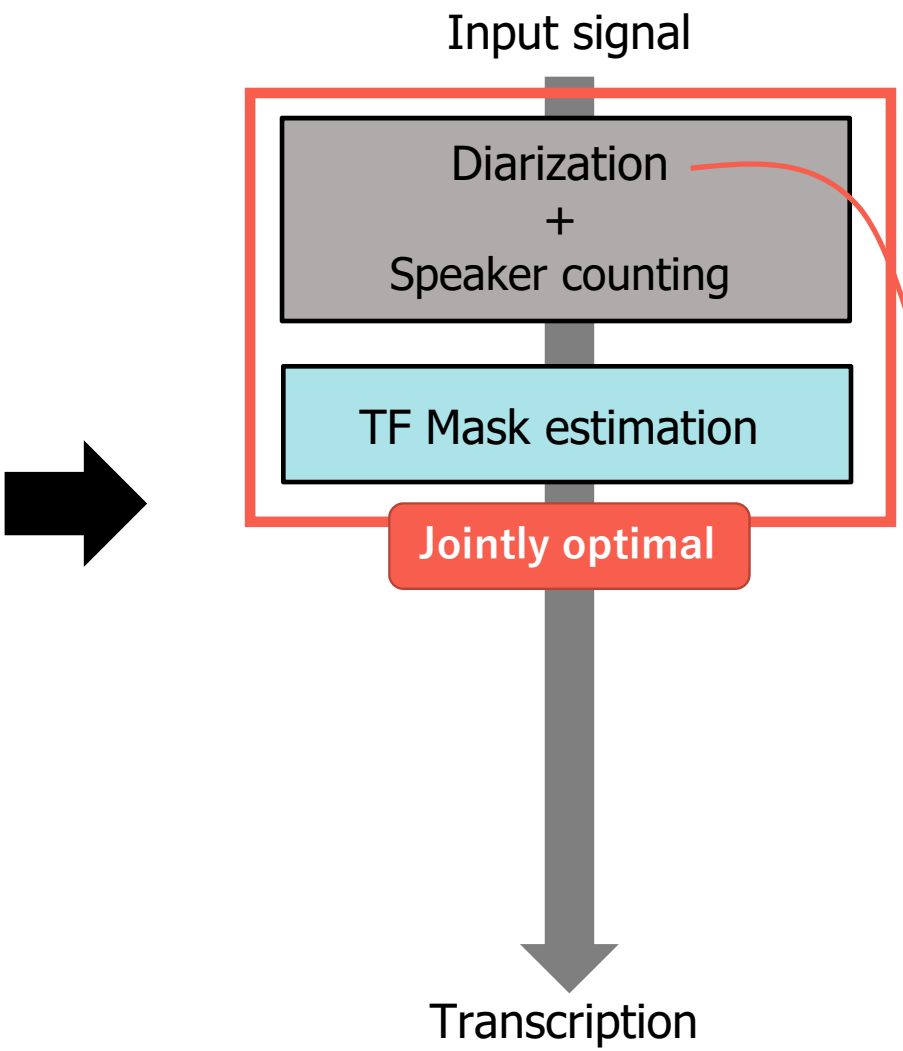
- ASR, separation, and speaker counting functionalities in one model.

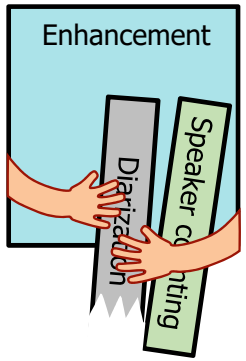
☺ Jointly optimal: 1-ch enhancement + speaker counting + ASR

T. von Neumann et al., Multi-talker ASR for an unknown number of sources: Joint training of source counting, separation and ASR, Interspeech, 2020

(3) Diarization + Speaker counting + TF Mask estimation

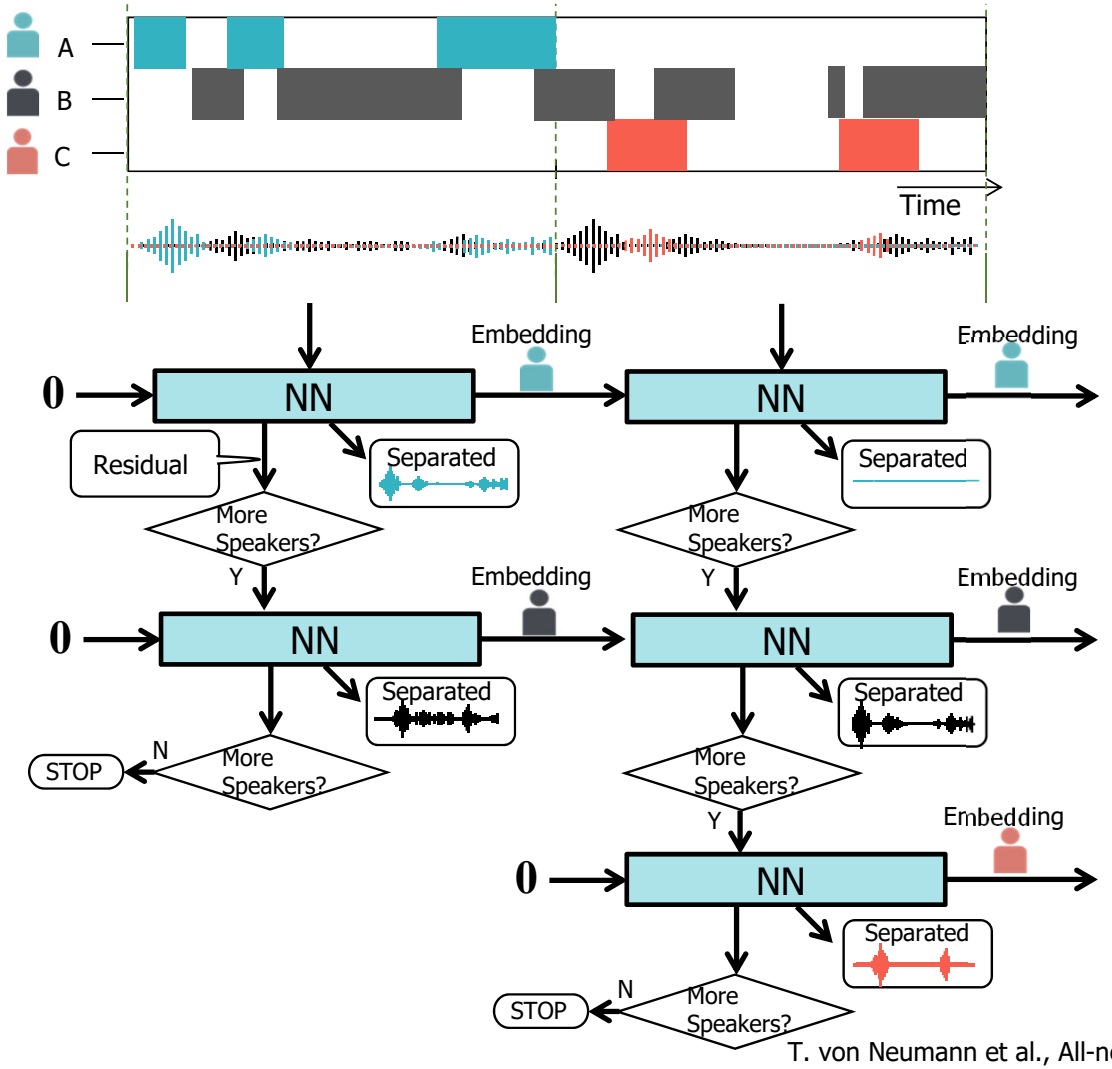






(Strictly speaking)
Speaker tracking

(3) Diarization + Speaker counting + TF Mask estimation

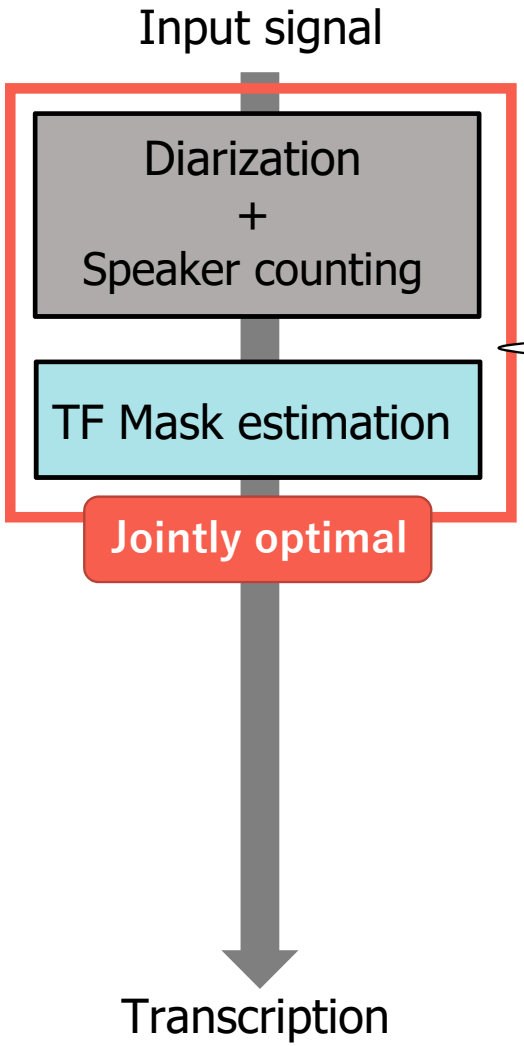


- Extension of recursive source separation [von Neumann+ 2019]
- At each block, perform recursive separation + embedding estimation
- Diarization can be performed with simple VAD
- An online model of an entire meeting
- Can be trained in a similar way as recursive source separation, i.e.,

$$J_{\phi^\star}^{\text{Enh+Count}} = J_{\phi^\star}^{\text{Enh}} + J^{\text{Count}}$$

T. von Neumann et al., All-neural online source separation, counting, and diarization for meeting analysis, ICASSP, 2019

(3) Diarization + Speaker counting + TF Mask estimation

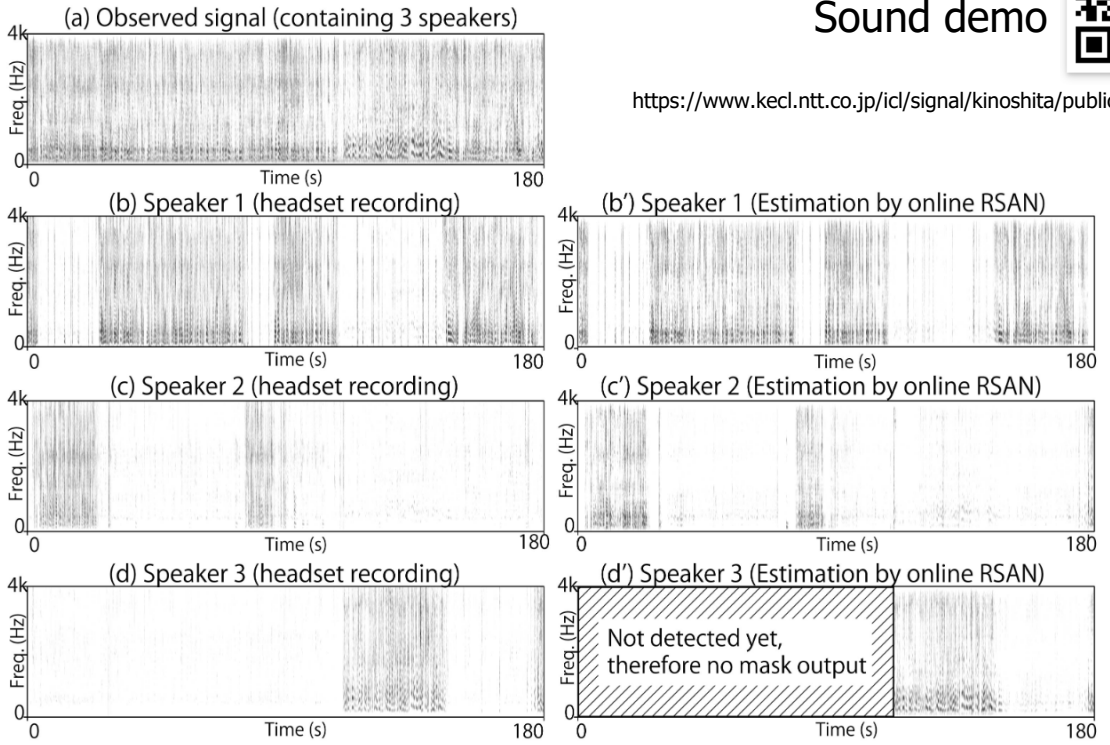


Data: Real meetings (4 to 6 speakers, about 15 min.),
Significant noise and reverberation, overlap-ratio

[Source separation + speaker tracking results]

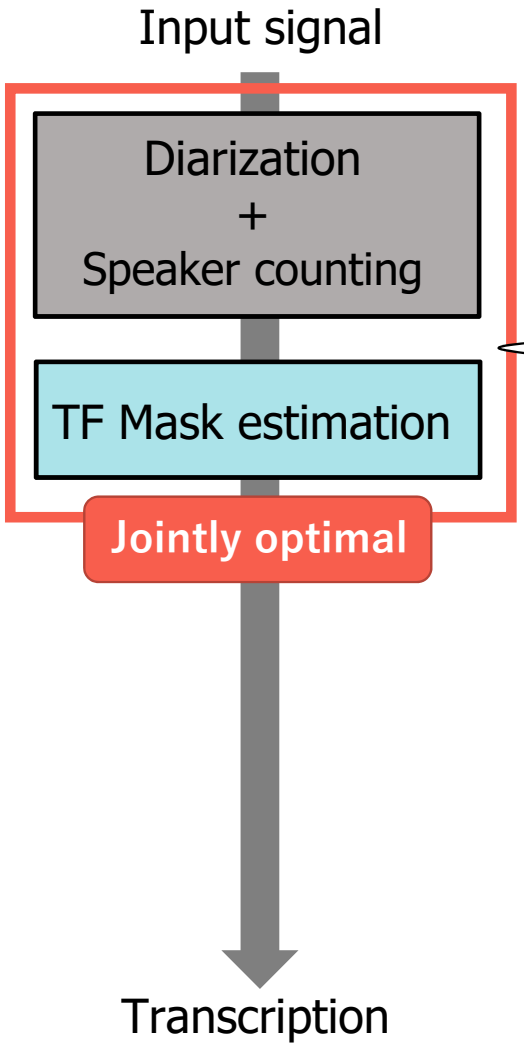
Sound demo

https://www.kecl.ntt.co.jp/icl/signal/kinoshita/publications/ICASSP20/onlineRSAN/index.html



K. Kinoshita et al., Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system, ICASSP, 2020

(2') Diarization + Speaker counting + TF Mask estimation



Data: Real meetings (4 to 6 speakers, each meeting is about 15 min.),
Significant noise and reverberation

[Diarization results]

Method	DER (%)
X-vector clustering	65.0
Separation+Speaker tracking	51.4

☺ Outperformed x-vector clustering system

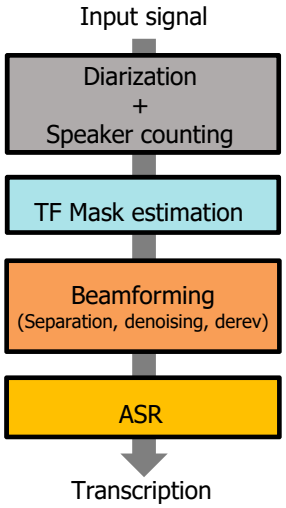
☹ Not yet combined with ASR, partially because the model is already too big.

☺ Jointly optimal: 1-ch enhancement + speaker counting +speaker tracking
(similar to diarization)

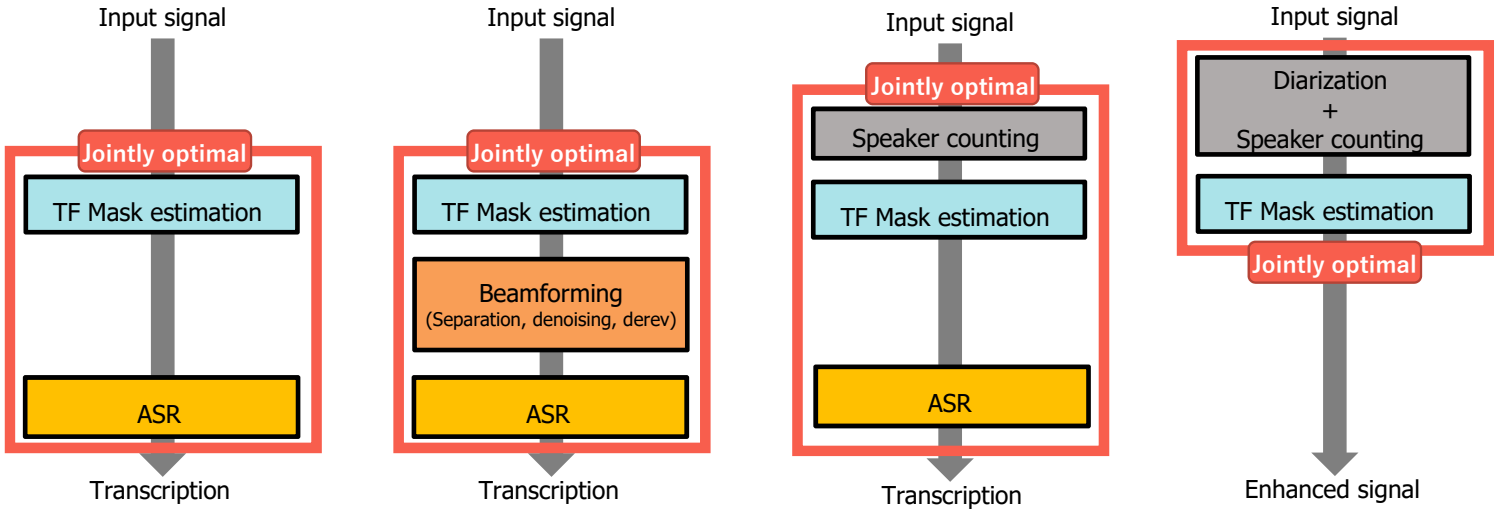
K. Kinoshita et al., Tackling real noisy reverberant meetings with all-neural source separation, counting, and diarization system, ICASSP, 2020

Short summary of "Enhancement + X"

Baseline system



Jointly optimal modular systems



☺ Jointly optimal systems successfully outperformed cascaded systems.

☹ Each module is relatively large and sometimes computationally inefficient, which makes it difficult to train the system with large amount of data.
→ Need more efficient training schemes, or more powerful computation resources.

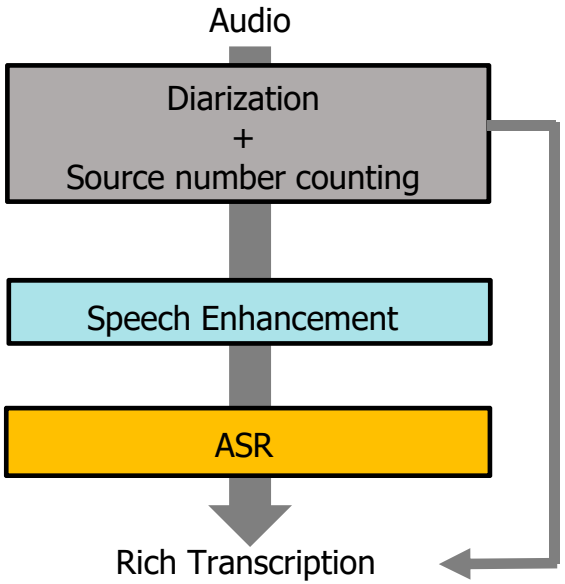
3. A new research trend: Jointly optimal systems

3.1. Diarization + x

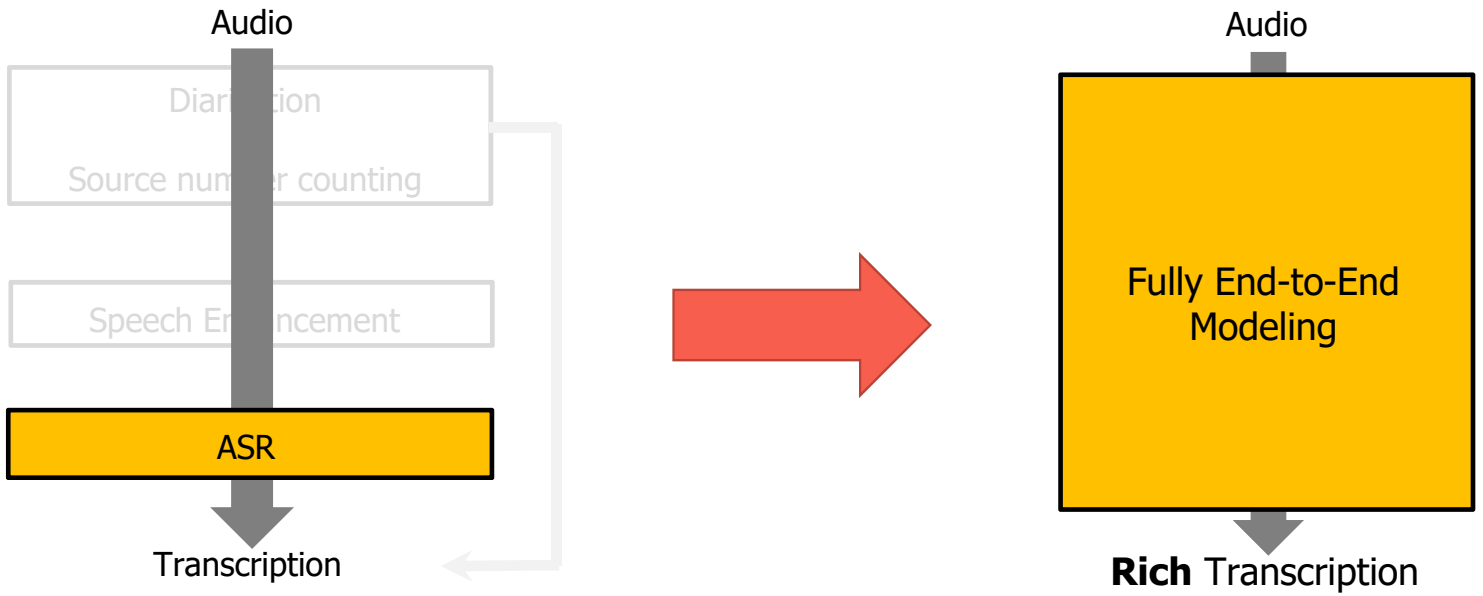
3.2. Enhancement + x

3.3. ASR + x

ASR + x: Enriching ASR System towards Rich Transcription

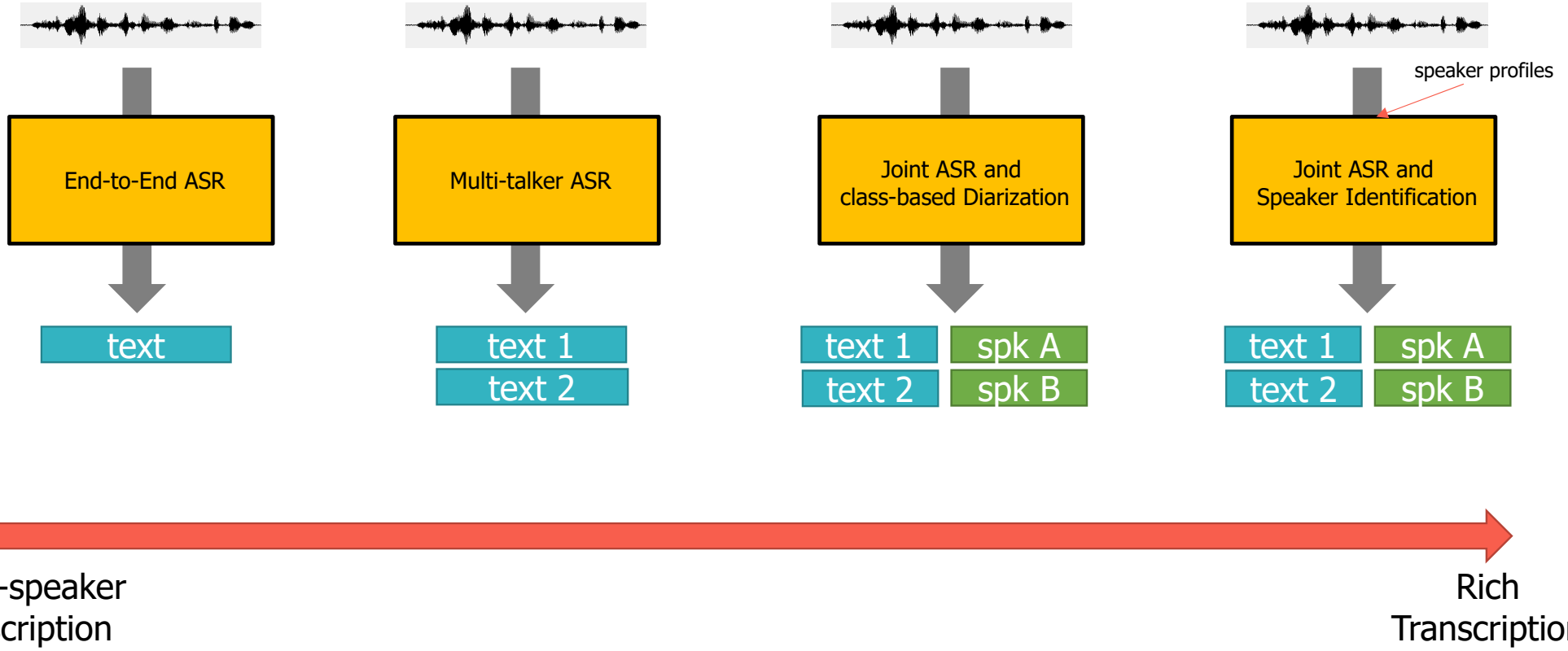


ASR + x: Enriching ASR System towards Rich Transcription

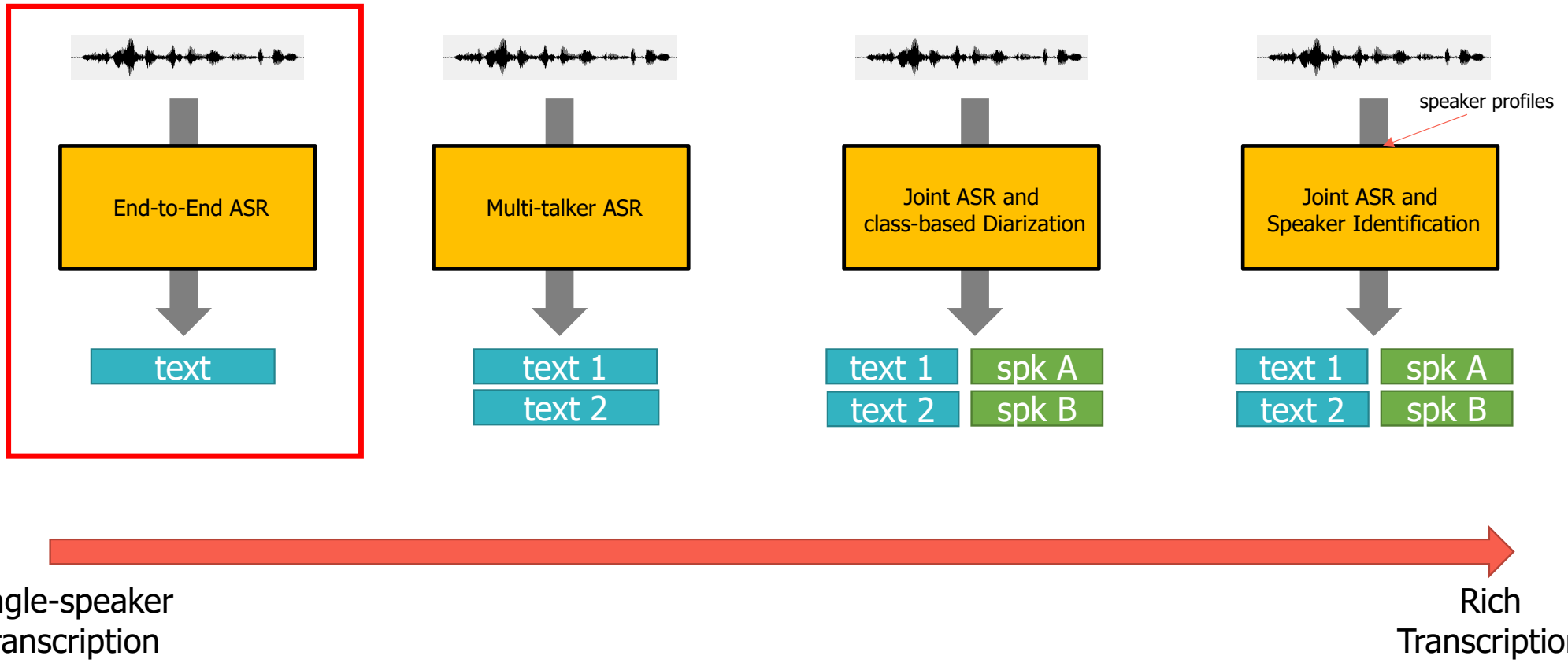


☺ Can keep the whole system simpler and smaller (i.e., computationally efficient, less parameters).

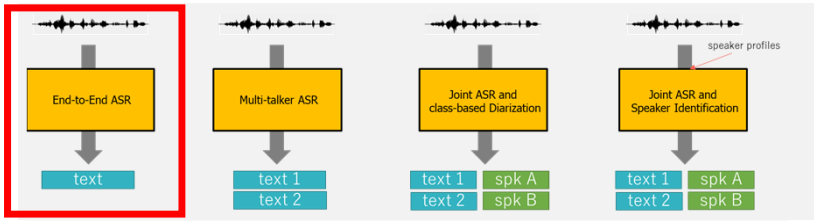
What has been achieved for "ASR + x"



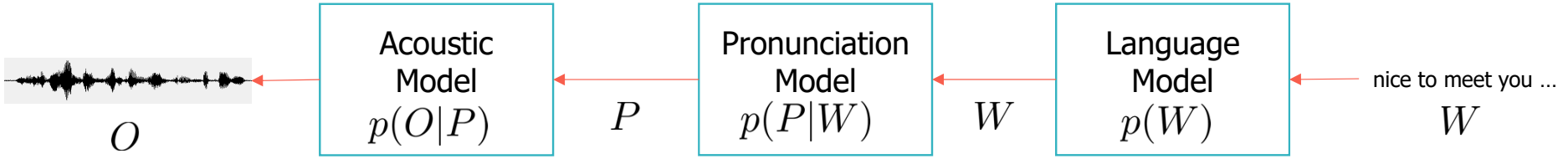
What has been achieved for "ASR + x"



Hybrid ASR and End-to-End ASR



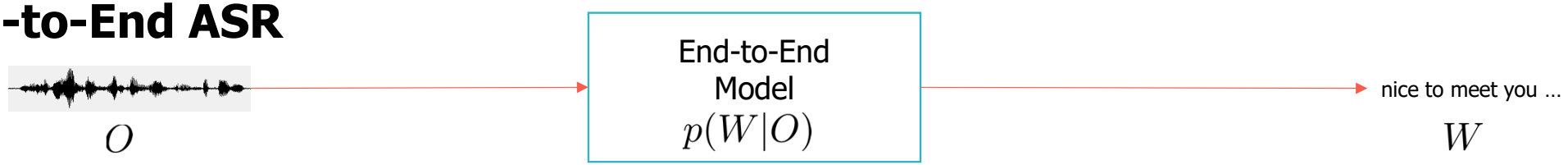
Hybrid ASR



$$\tilde{W} = \text{argmax}_W p(W|O) = \text{argmax}_W p(O|P)p(P|W)p(W)$$

- Flexible for customization (change LM, adding pronunciation)

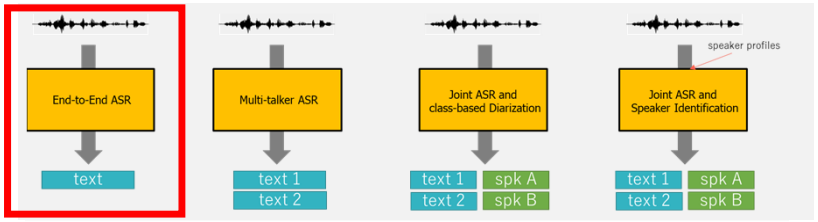
End-to-End ASR

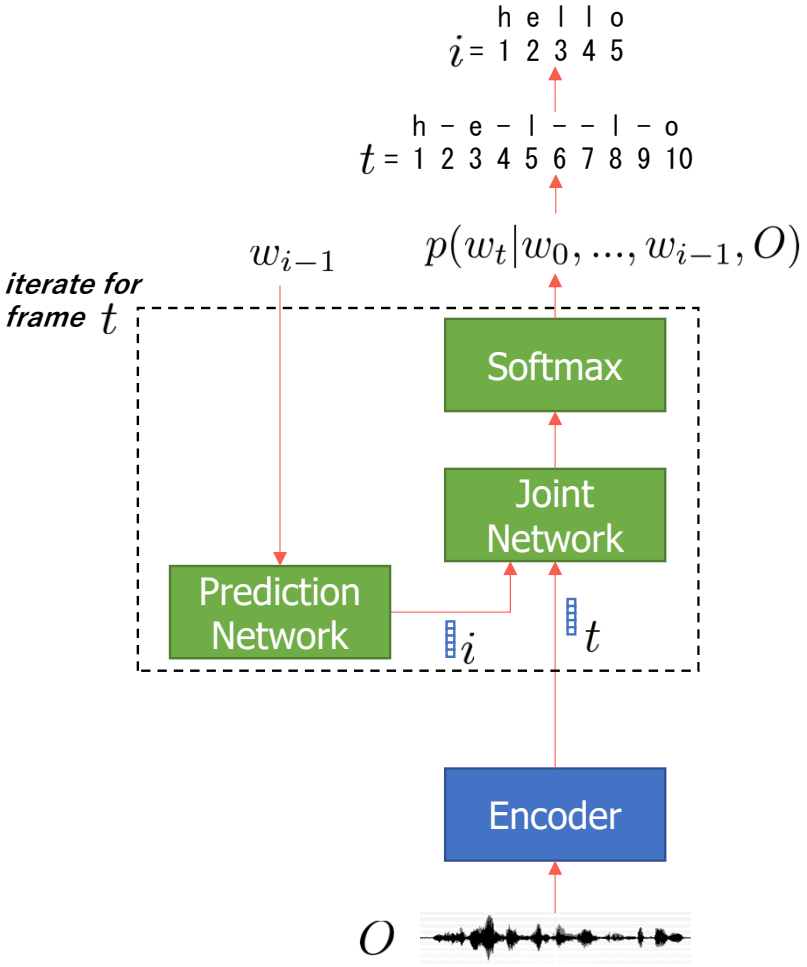


$$\tilde{W} = \text{argmax}_W p(W|O)$$

- End-to-end optimization
- Easy to incorporate additional input / output in the model

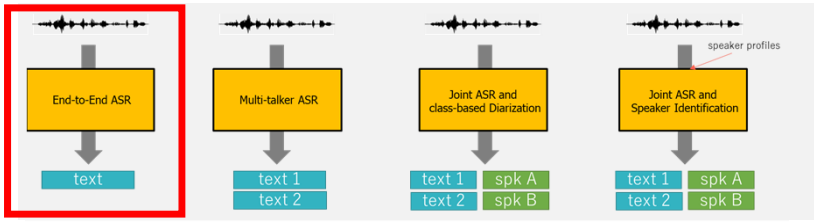
RNN Transducer and Attention Encoder Decoder

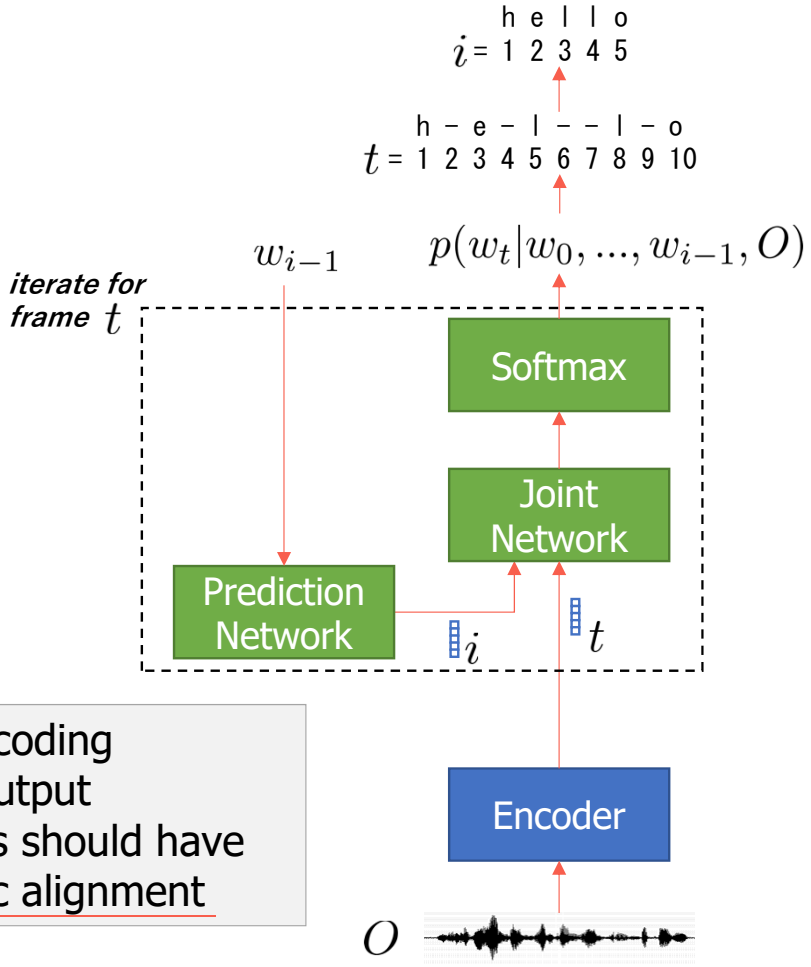




RNN Transducer (RNN-T)

RNN Transducer and Attention Encoder Decoder



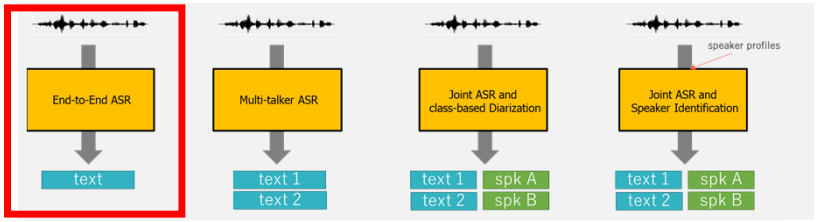


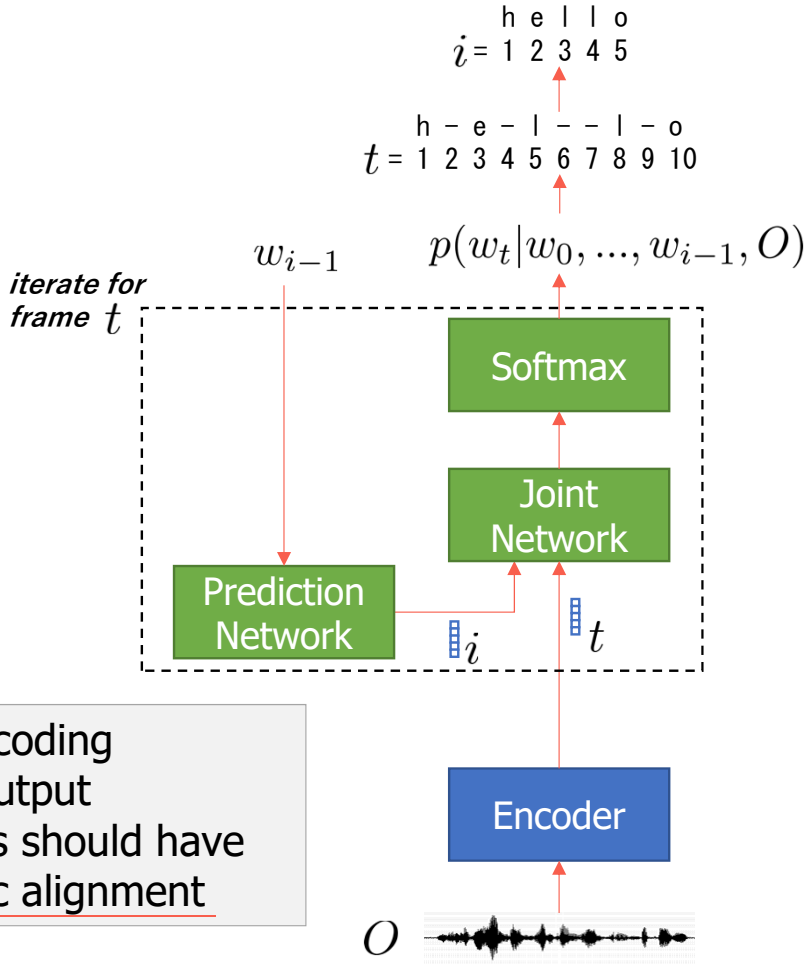
- Online decoding
- Input & output sequences should have monotonic alignment

Eg)
- 😀 one hundred dollar
- 🤨 $100

RNN Transducer (RNN-T)

RNN Transducer and Attention Encoder Decoder

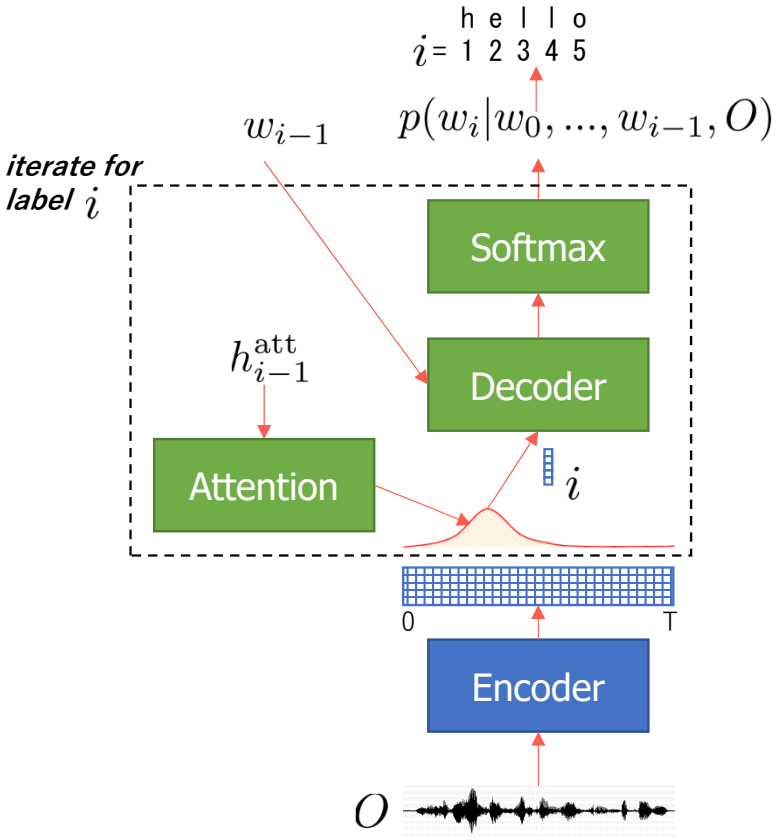




- Online decoding
- Input & output sequences should have monotonic alignment

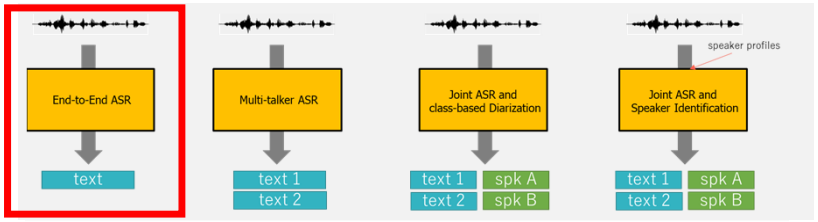
Eg)
- one hundred dollar
- $100

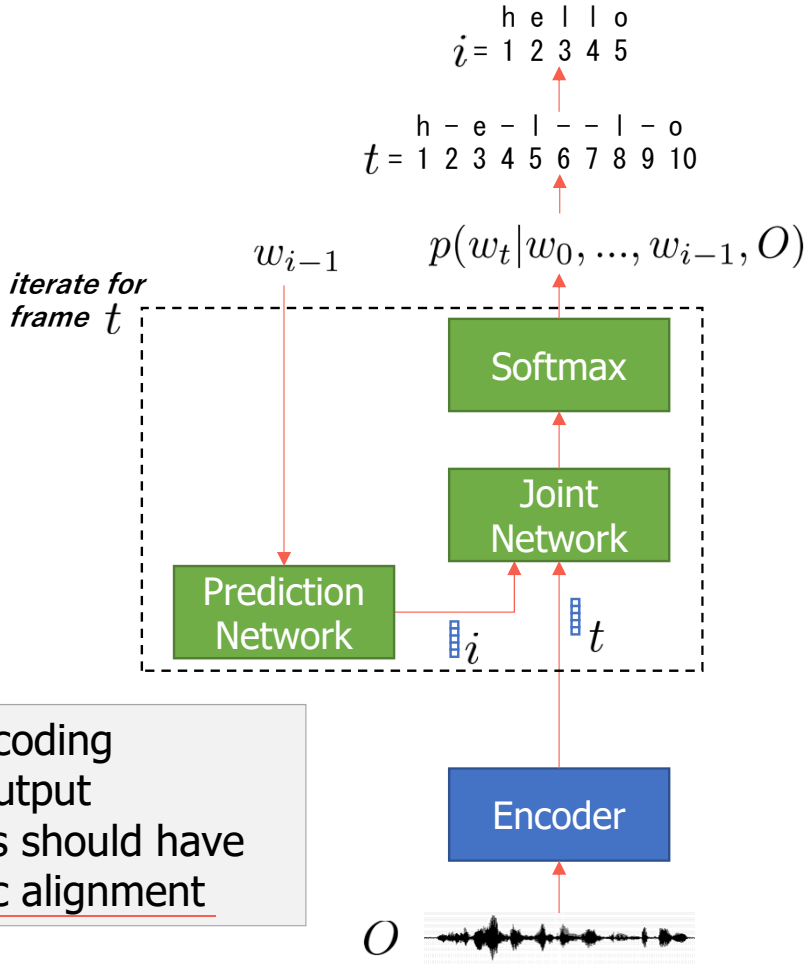
RNN Transducer (RNN-T)



Attention Encoder Decoder (AED)

RNN Transducer and Attention Encoder Decoder

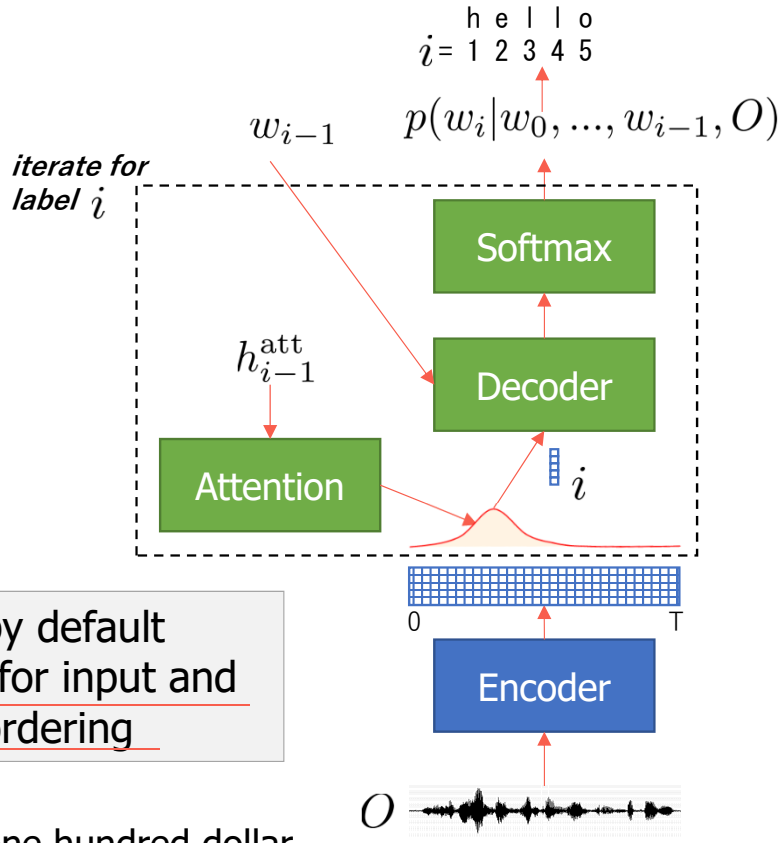




- Online decoding
- Input & output sequences should have monotonic alignment

Eg)
- one hundred dollar
- $100

RNN Transducer (RNN-T)

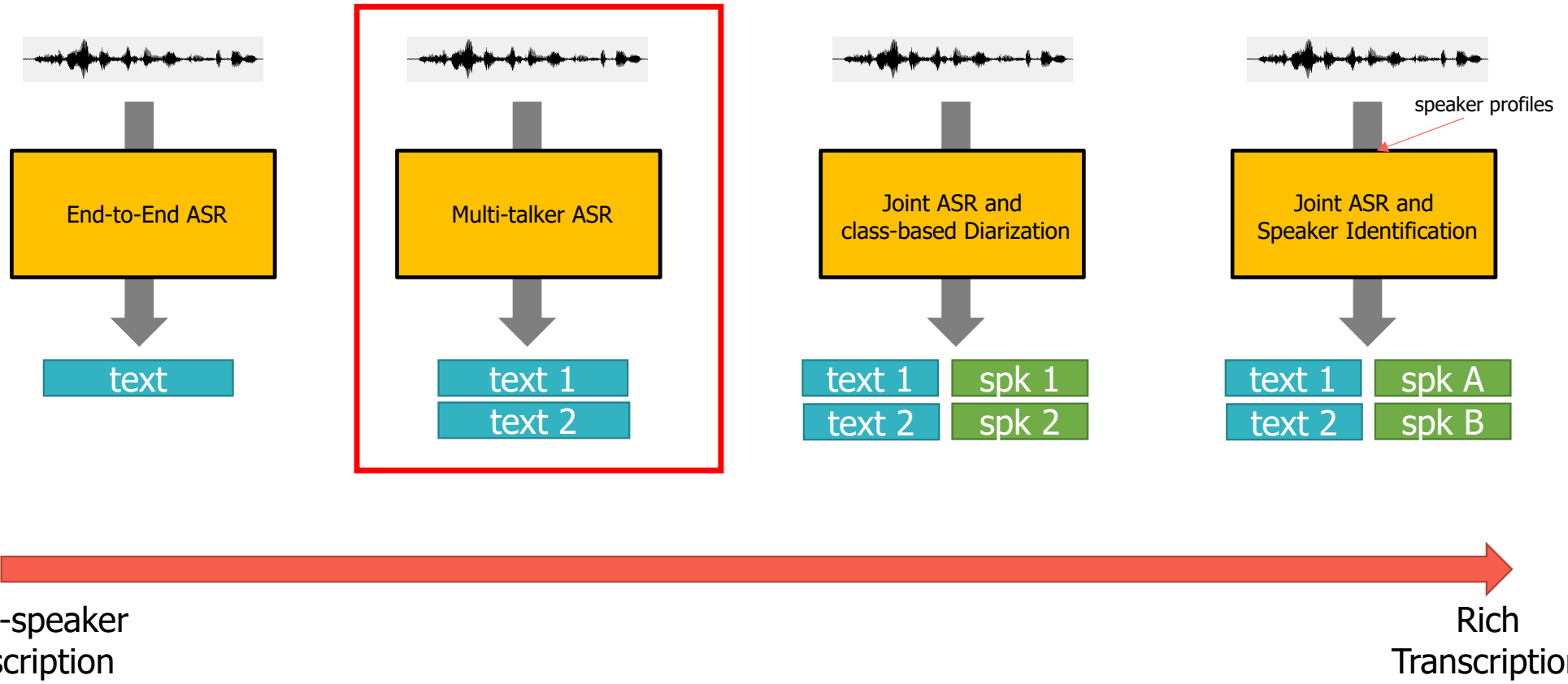


- Offline by default
- Flexible for input and output ordering

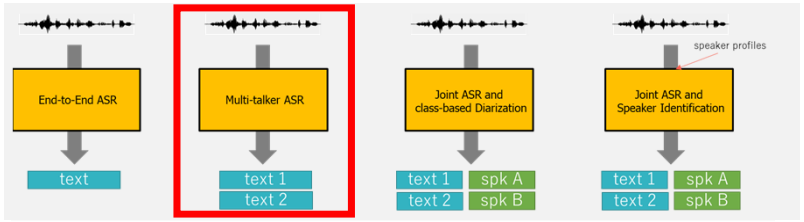
Eg)
- one hundred dollar
- $100

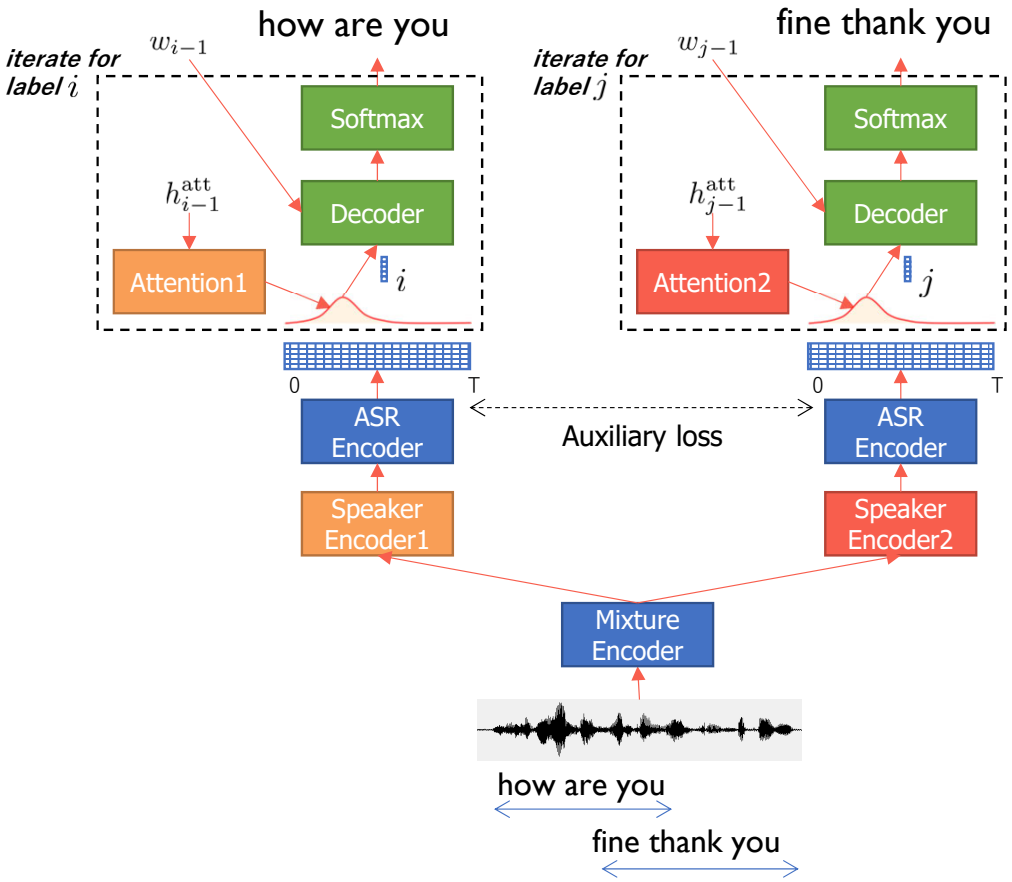
Attention Encoder Decoder (AED)

What has been achieved for "ASR + x"



A Purely End-to-end System for Multi-speaker Speech Recognition [Seki+ 2018]



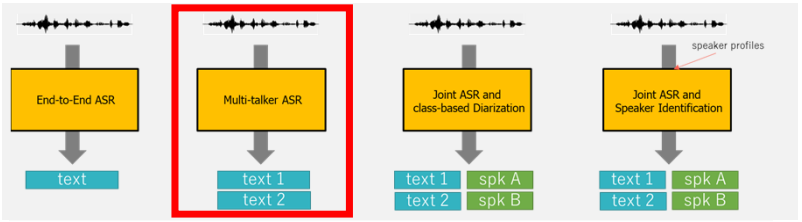


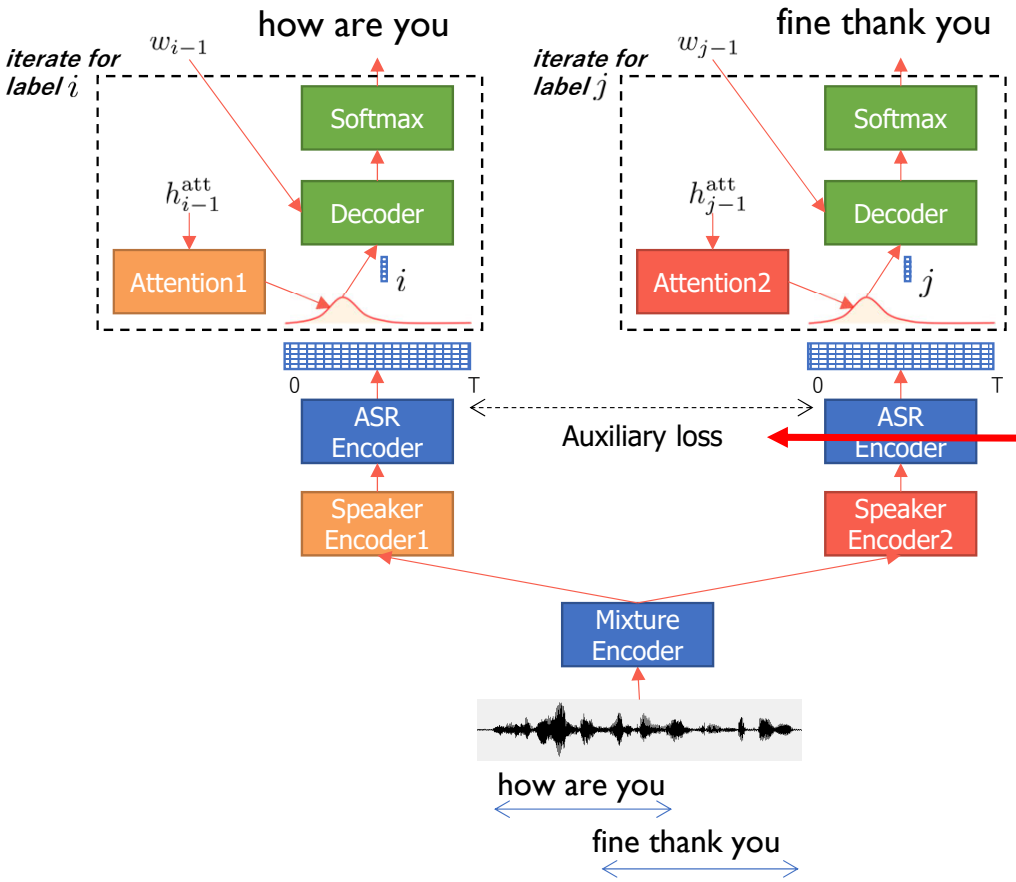
Two output model with *speaker-dependent encoder*

- Simpler training.
- No need for noisy-clean audio pair for training.

- H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. *ACL*, 2018.
- X. Chang et al., End-to-End Monaural Multi-speaker ASR System without pretraining, *ICASSP* 2019.

A Purely End-to-end System for Multi-speaker Speech Recognition [Seki+ 2018]





Two output model with *speaker-dependent encoder*

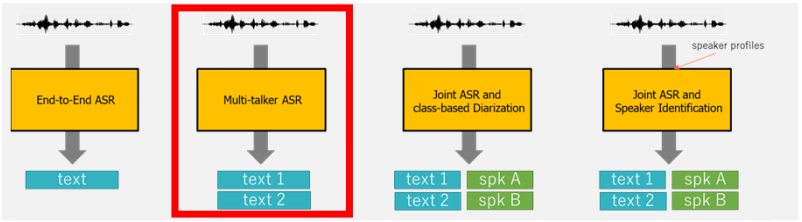
☺ Simpler training.
No need for noisy-clean audio pair for training.

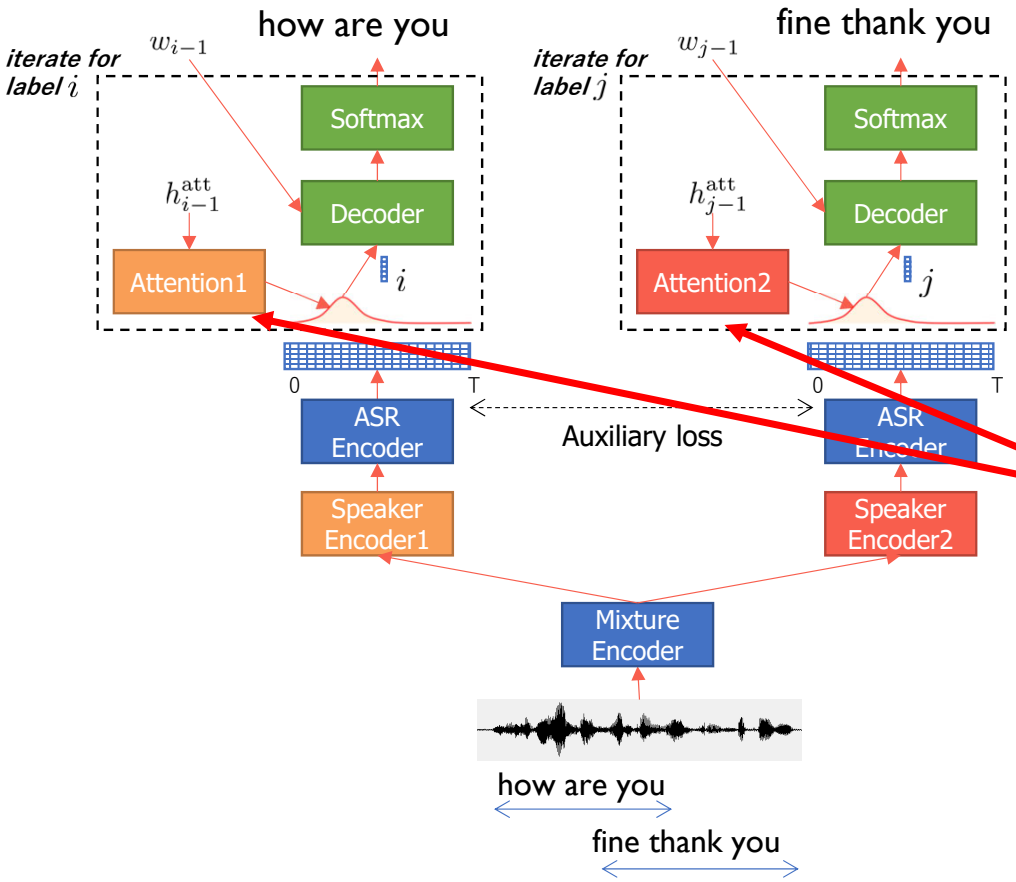
Techniques

- KL-divergence-based auxiliary loss to make the distance between outputs of ASR encoders different [Seki+ 2018]

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. *ACL*, 2018.
❑ X. Chang et al., End-to-End Monaural Multi-speaker ASR System without pretraining, *ICASSP* 2019.

A Purely End-to-end System for Multi-speaker Speech Recognition [Seki+ 2018]





Two output model with *speaker-dependent encoder*

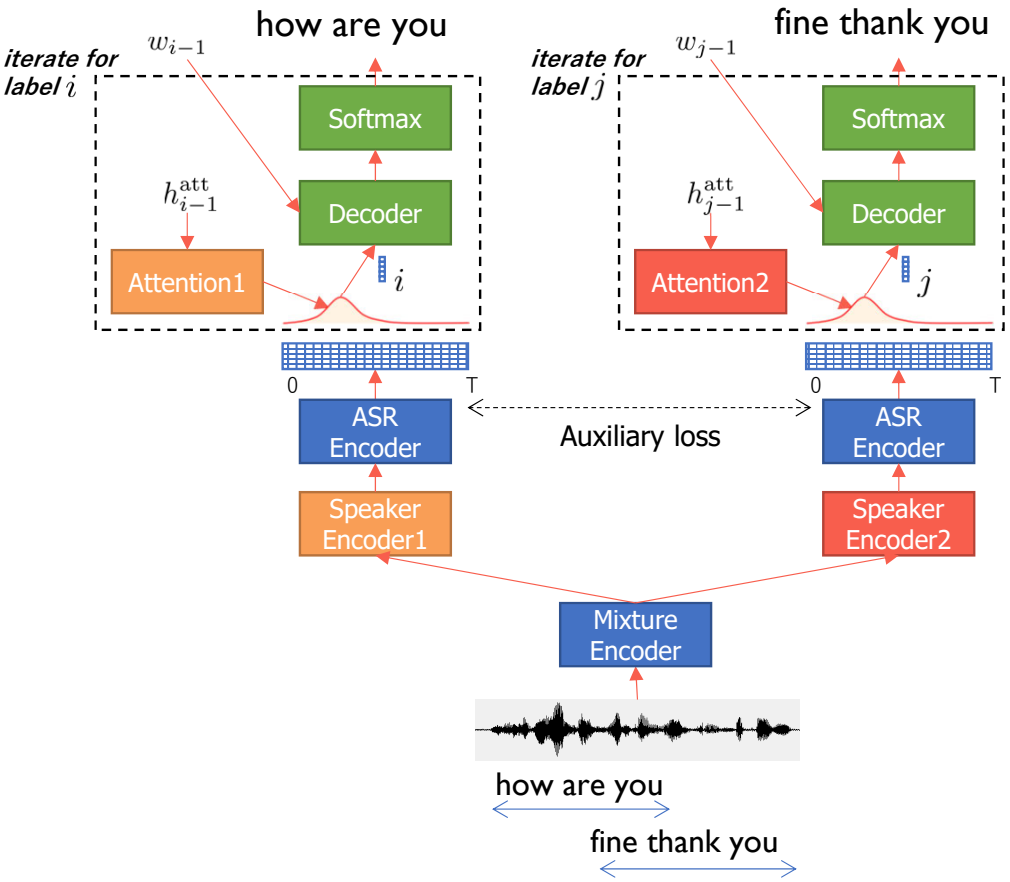
☺ Simpler training.
No need for noisy-clean audio pair for training.

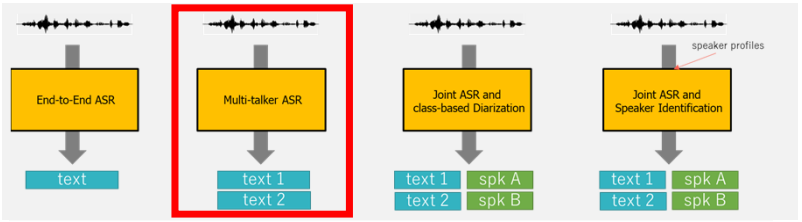
Techniques

- KL-divergence-based auxiliary loss to make the distance between outputs of ASR encoders different [Seki+ 2018]
- Speaker-specific attention layers are better [Chang+ 2019]

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. *ACL*, 2018.
❑ X. Chang et al., End-to-End Monaural Multi-speaker ASR System without pretraining, *ICASSP* 2019.

A Purely End-to-end System for Multi-speaker Speech Recognition [Seki+ 2018]





Two output model with *speaker-dependent encoder*

☺ Simpler training.
No need for noisy-clean audio pair for training.

Techniques

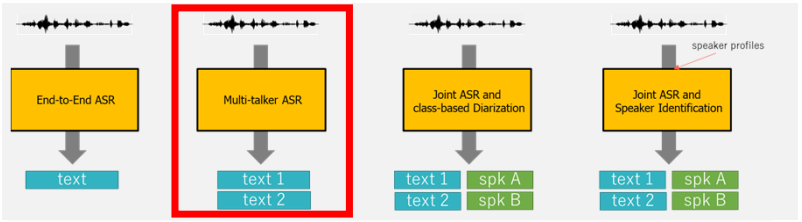
- KL-divergence-based auxiliary loss to make the distance between outputs of ASR encoders different [Seki+ 2018]
- Speaker-specific attention layers are better [Chang+ 2019]

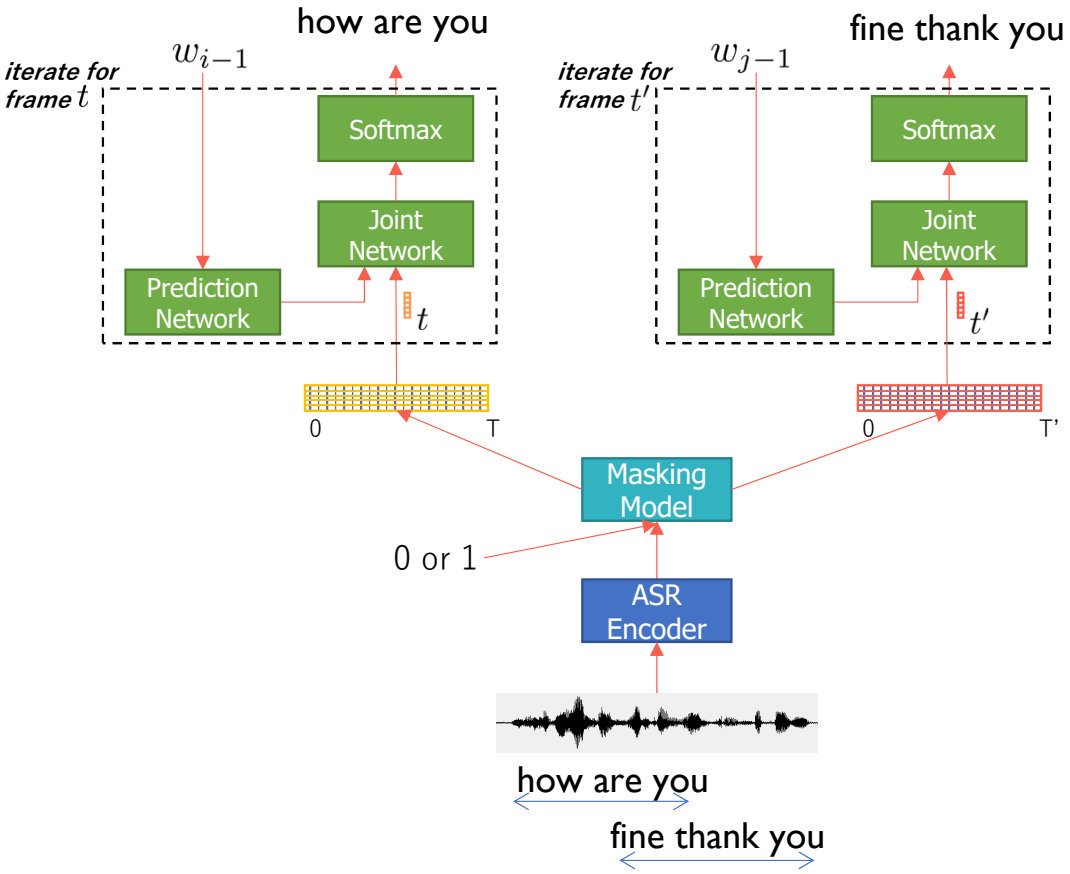
SWER (%) on 2-speaker mixed WSJ0 corpus (from [Chang+ 2019])

Model	Avg. WER
DPCL+ASR [17]	30.8
PIT-ASR [24]	28.2
End-to-end ASR (Char/Word-LM) [26]	28.2
Proposed End-to-end ASR with SPA (Word LM)	**25.4**

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. *ACL*, 2018.
❑ X. Chang et al., End-to-End Monaural Multi-speaker ASR System without pretraining, *ICASSP* 2019.

End-to-end Multi-talker Overlapping Speech Recognition [Tripathi+ 2020]

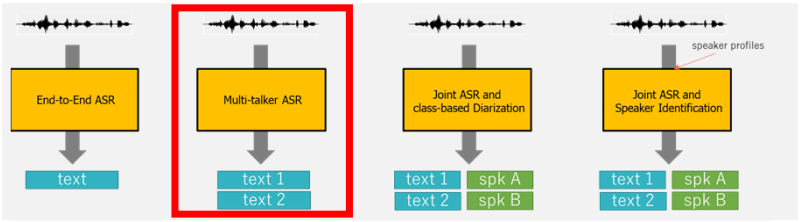


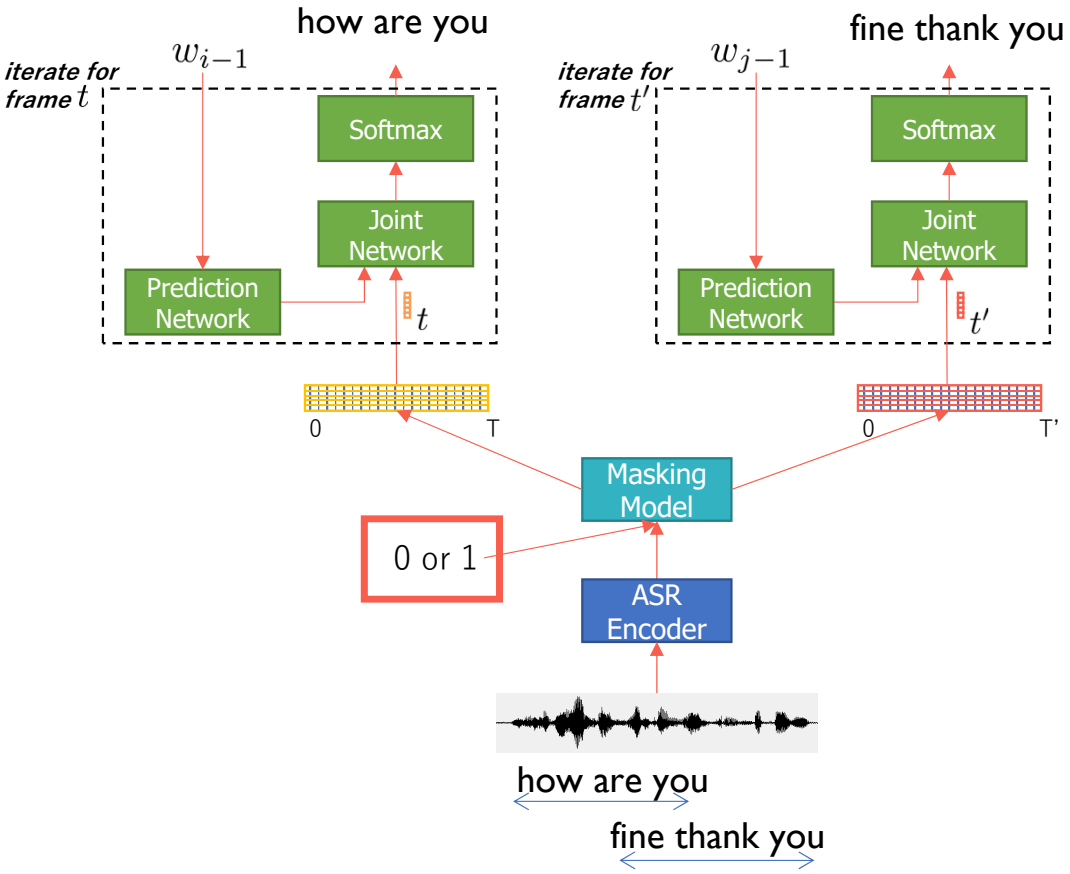


- Extension for RNN-T (yet offline model)
- Masking model to recognize two speakers

❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.

End-to-end Multi-talker Overlapping Speech Recognition [Tripathi+ 2020]

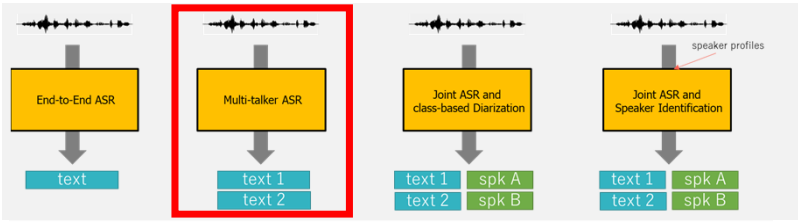


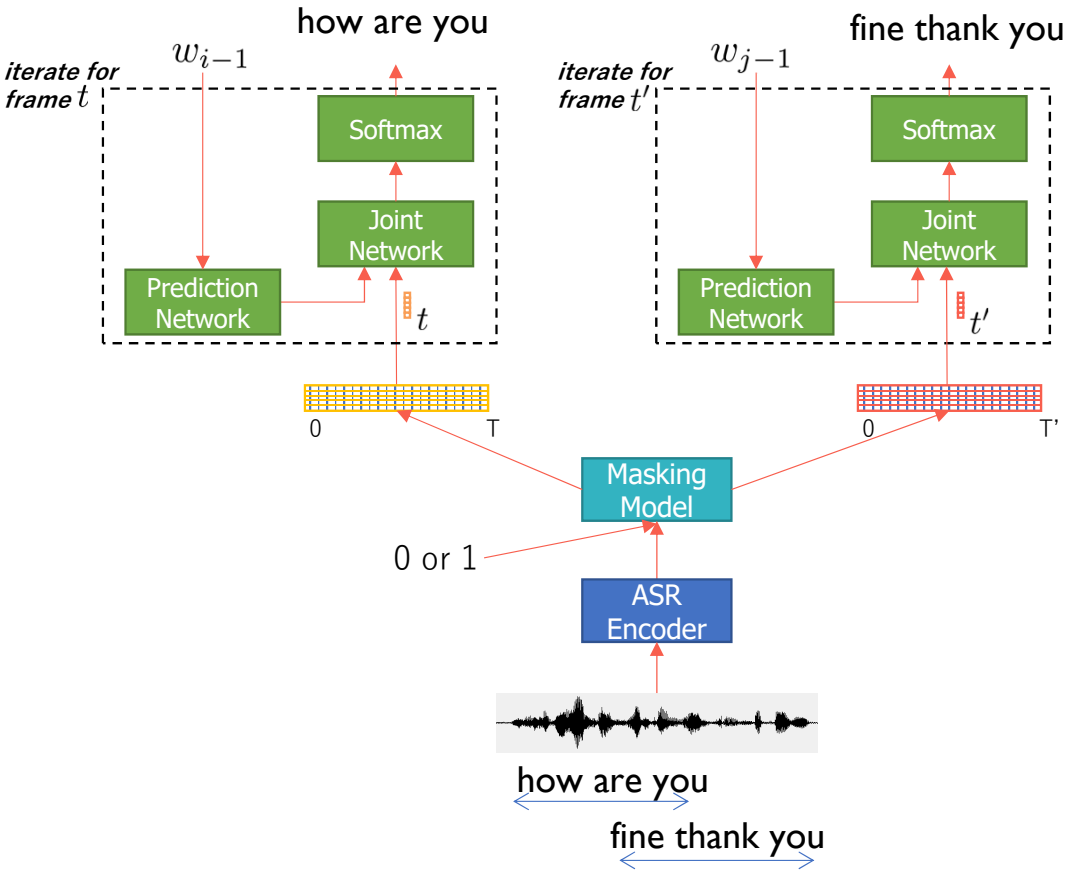


- Extension for RNN-T (yet offline model)
- Masking model to recognize two speakers

❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.

End-to-end Multi-talker Overlapping Speech Recognition [Tripathi+ 2020]





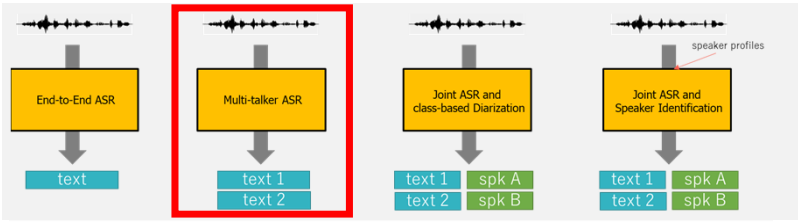
- Extension for RNN-T (yet offline model)
- Masking model to recognize two speakers

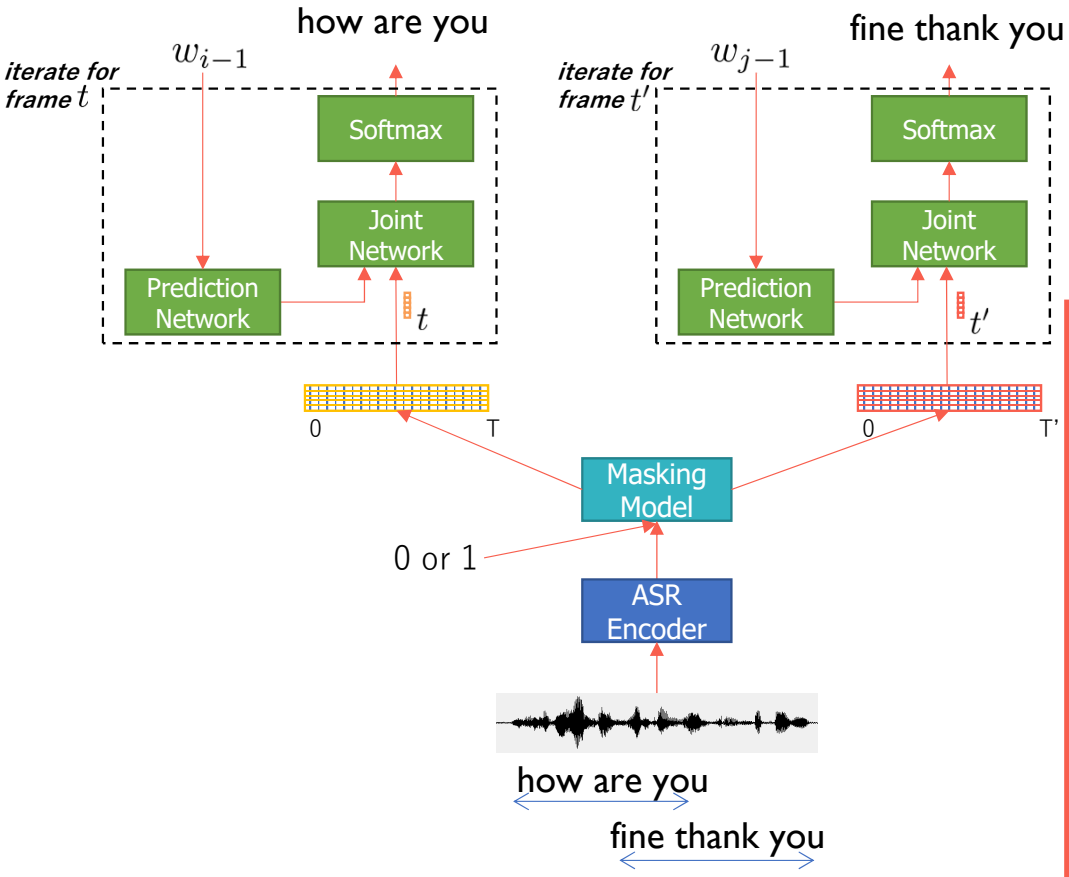
😀 Leverage well-trained RNN-T parameters as an initial params.

😀 Very good result compared to human transcribers.

❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.

End-to-end Multi-talker Overlapping Speech Recognition [Tripathi+ 2020]

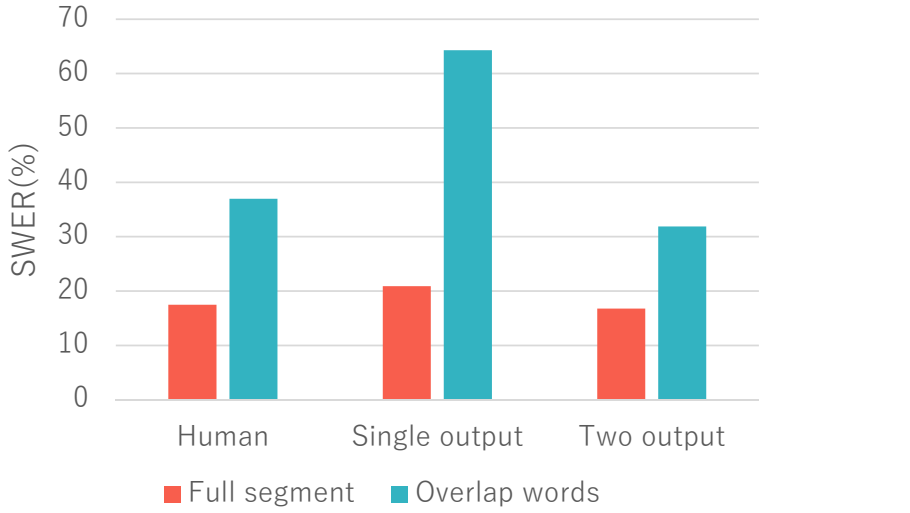




- Extension for RNN-T (yet offline model)
- Masking model to recognize two speakers

😀 Leverage well-trained RNN-T parameters as an initial params.

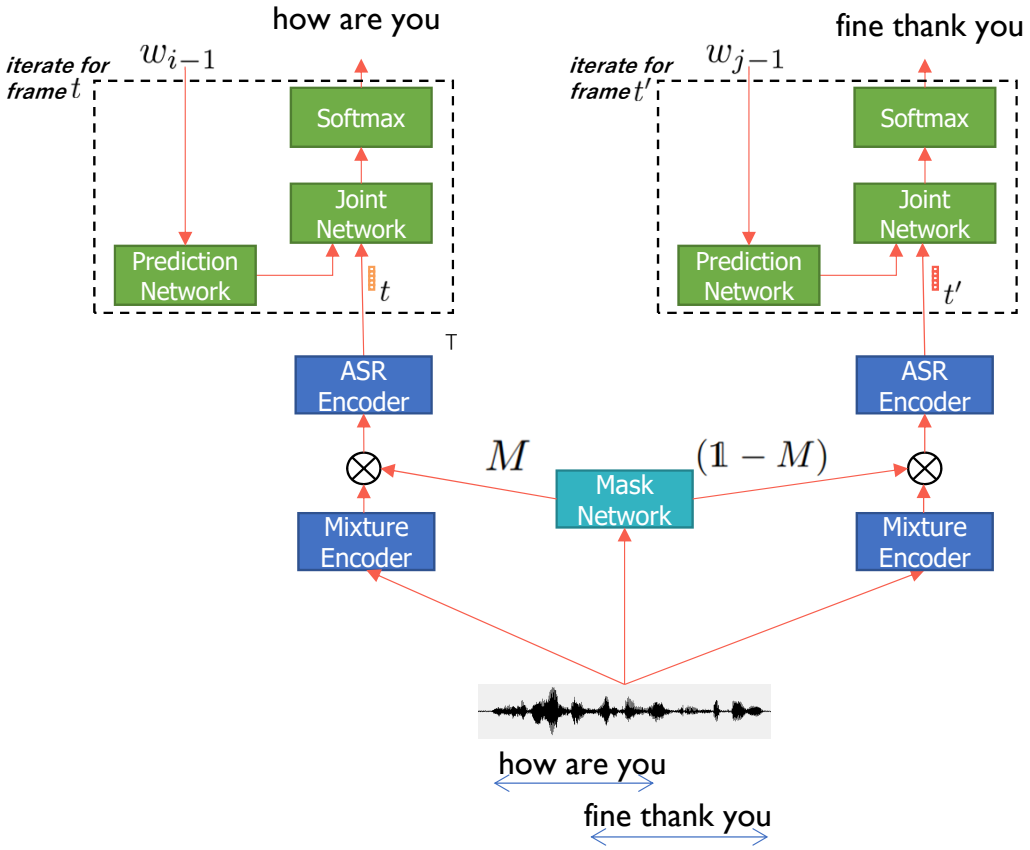
😀 Very good result compared to human transcribers.



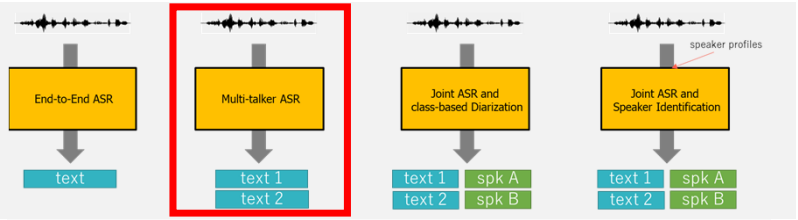
SWER (%) on 2-speaker mixed simulation data (from [Tripathi+ 2019])

❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.

Streaming end-to-end multi-talker speech recognition [Lu+ 2021][Sklyar+ 2021]



Streaming Unmixing and Recognition Transducer (SURT) with mask-based unmixing model [Lu+ 2021]

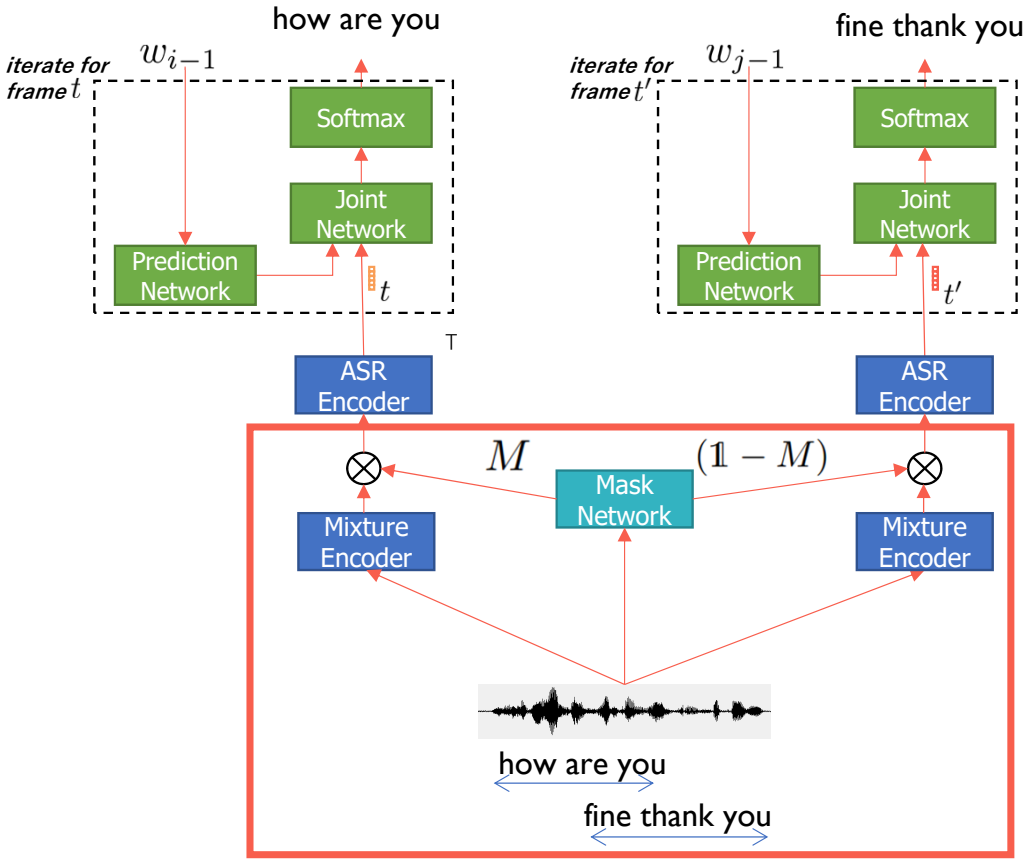


Streaming model based on RNN-T

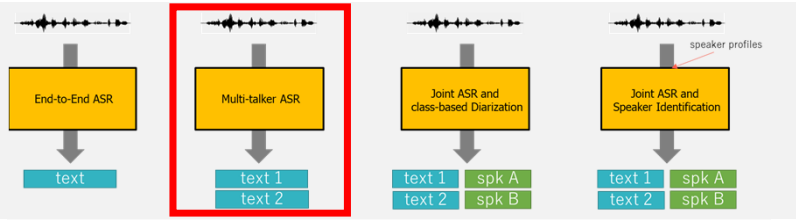
- ❑ [Lu+ 2021] Mask-based model or SD-encoder
- ❑ [Sklyar+ 2021] SD-encoder

- ❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
- ❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Streaming end-to-end multi-talker speech recognition [Lu+ 2021][Sklyar+ 2021]



Streaming Unmixing and Recognition Transducer (SURT) with mask-based unmixing model [Lu+ 2021]

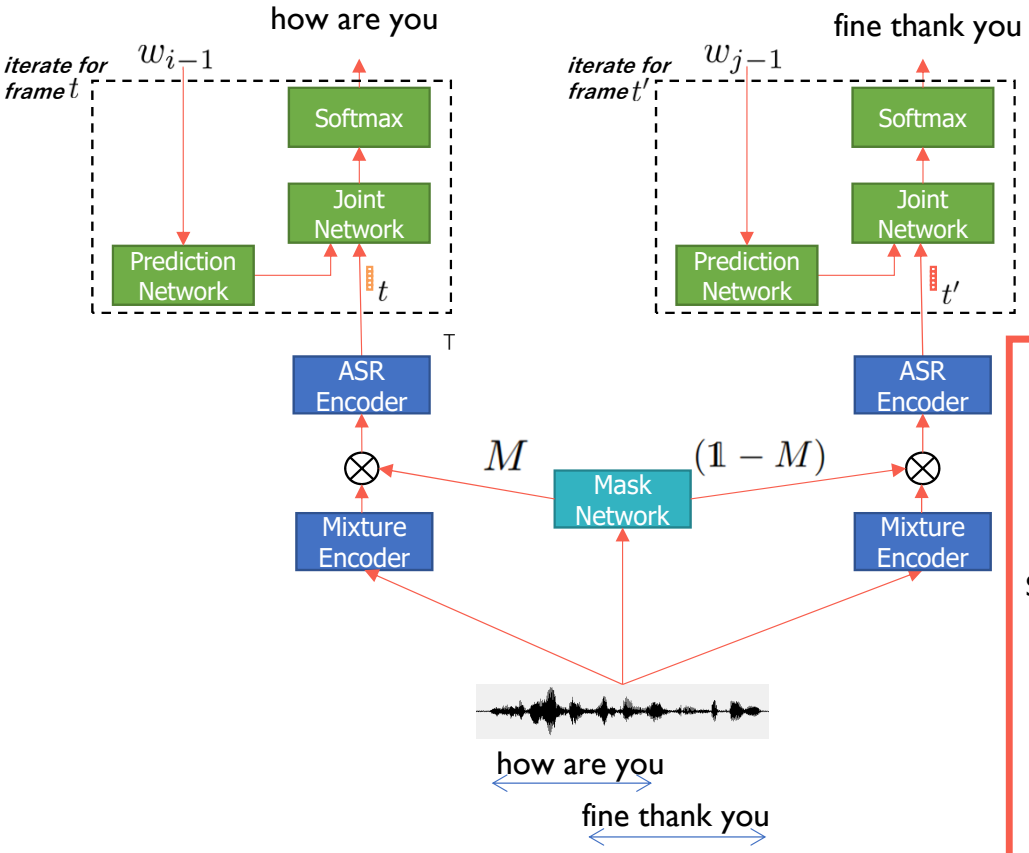


Streaming model based on RNN-T

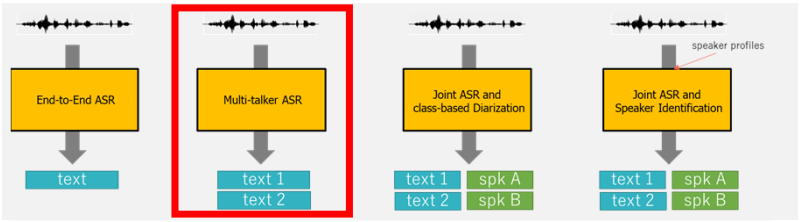
- ❑ [Lu+ 2021] Mask-based model or SD-encoder
- ❑ [Sklyar+ 2021] SD-encoder

- ❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
- ❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Streaming end-to-end multi-talker speech recognition [Lu+ 2021][Sklyar+ 2021]



Streaming Unmixing and Recognition Transducer (SURT) with mask-based unmixing model [Lu+ 2021]



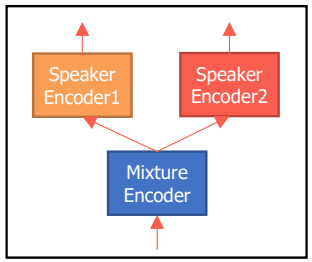
Streaming model based on RNN-T

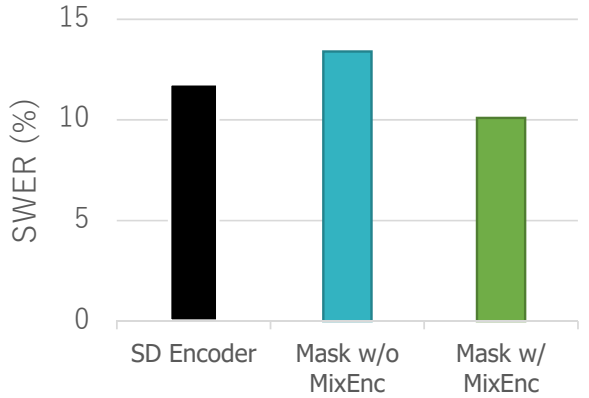
- ❑ [Lu+ 2021] Mask-based model or SD-encoder
- ❑ [Sklyar+ 2021] SD-encoder

Techniques

- Mask-based model (left figure) was better than SD-encoder-based model [Lu+ 2021]

Speaker-dependent (SD)-encoder

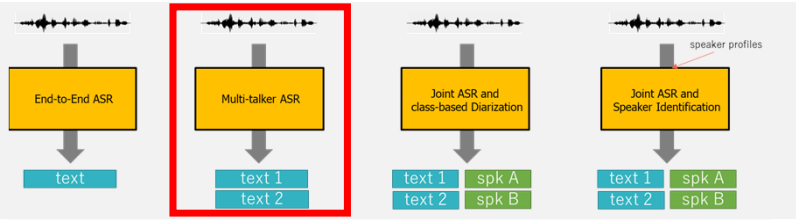




SWER (%) on 2-speaker mixed LibriSpeechMix (from [Lu+ 2021])

❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters*, 2021.
❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

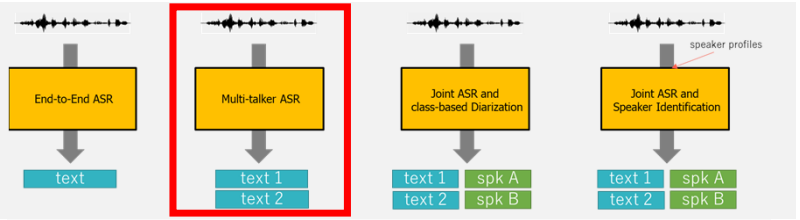
Reducing the computational cost of Permutation Invariant Training (PIT)



- A multi-output branch model usually uses PIT, which requires $O(S^3)$ computation w.r.t. the number of outputs S

- H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. ACL, 2018.
- A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.
- L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
- I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Reducing the computational cost of Permutation Invariant Training (PIT)

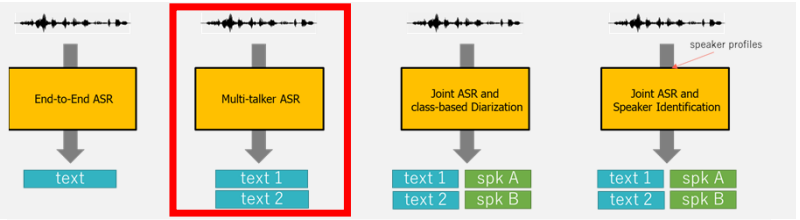


- A multi-output branch model usually uses PIT, which requires $O(S^3)$ computation w.r.t. the number of outputs S

- Use auxiliary CTC loss to determine the speaker order [Seki+ 2018]
  - Still $O(S^3)$, but CTC-loss is much less expensive than AED-loss.

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. ACL, 2018.
❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.
❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Reducing the computational cost of Permutation Invariant Training (PIT)

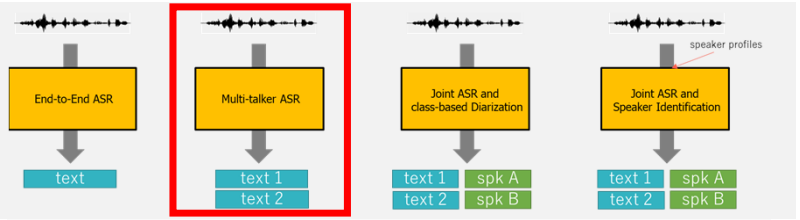


- A multi-output branch model usually uses PIT, which requires $O(S^3)$ computation w.r.t. the number of outputs S

- Use auxiliary CTC loss to determine the speaker order [Seki+ 2018]
  - Still $O(S^3)$, but CTC-loss is much less expensive than AED-loss.
- Use start time of each utterance to determine the order → $O(S)$

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. ACL, 2018.
❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.
❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Reducing the computational cost of Permutation Invariant Training (PIT)

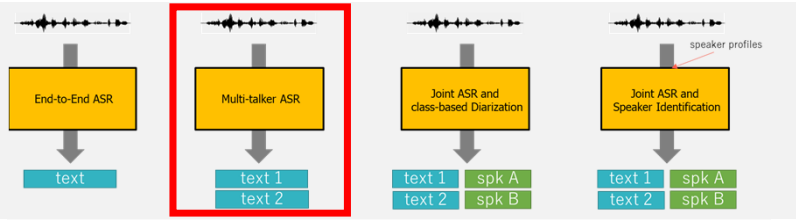


- A multi-output branch model usually uses PIT, which requires $O(S^3)$ computation w.r.t. the number of outputs S

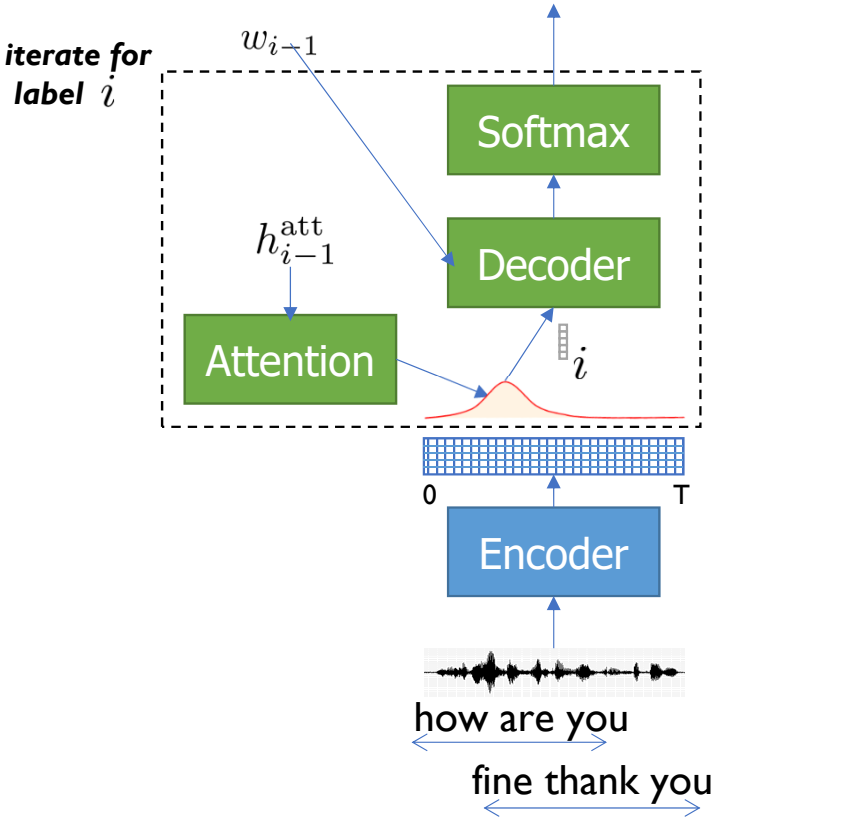
- Use auxiliary CTC loss to determine the speaker order [Seki+ 2018]
  - Still $O(S^3)$, but CTC-loss is much less expensive than AED-loss.
- Use start time of each utterance to determine the order → $O(S)$
  - Used in [Tripathi+ 2020] without comparison.
  - Later, [Lu+ 2021] and [Sklyar+ 2021] concurrently showed this heuristic error assignment is **as good as PIT**
    - Heuristic Error Assignment Training (HEAT) [Lu+ 2021]; Reported slightly better result than PIT
    - Deterministic Assignment Training (DAT) [Sklyar+ 2021]; Reported slightly worse result than PIT

❑ H. Seki et al., A Purely End-to-end System for Multi-speaker Speech Recognition. ACL, 2018.
❑ A. Tripathi et al., End-to-end multi-talker overlapping speech recognition. In: *ICASSP*. 2020. p. 6129-6133.
❑ L. Lu et al., Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters, 2021.*
❑ I Sklyar et al., Streaming Multi-speaker ASR with RNN-T. *ICASSP*, 2021.

Serialized Output Training (SOT) for End-to-End Overlapped Speech Recognition [Kanda+ 2020]

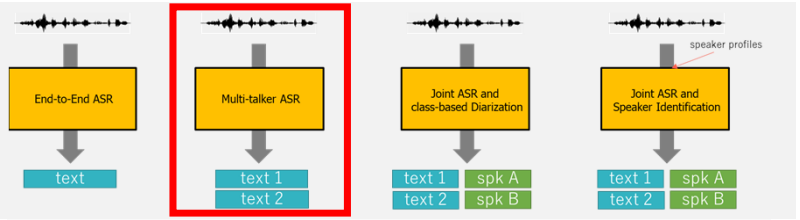


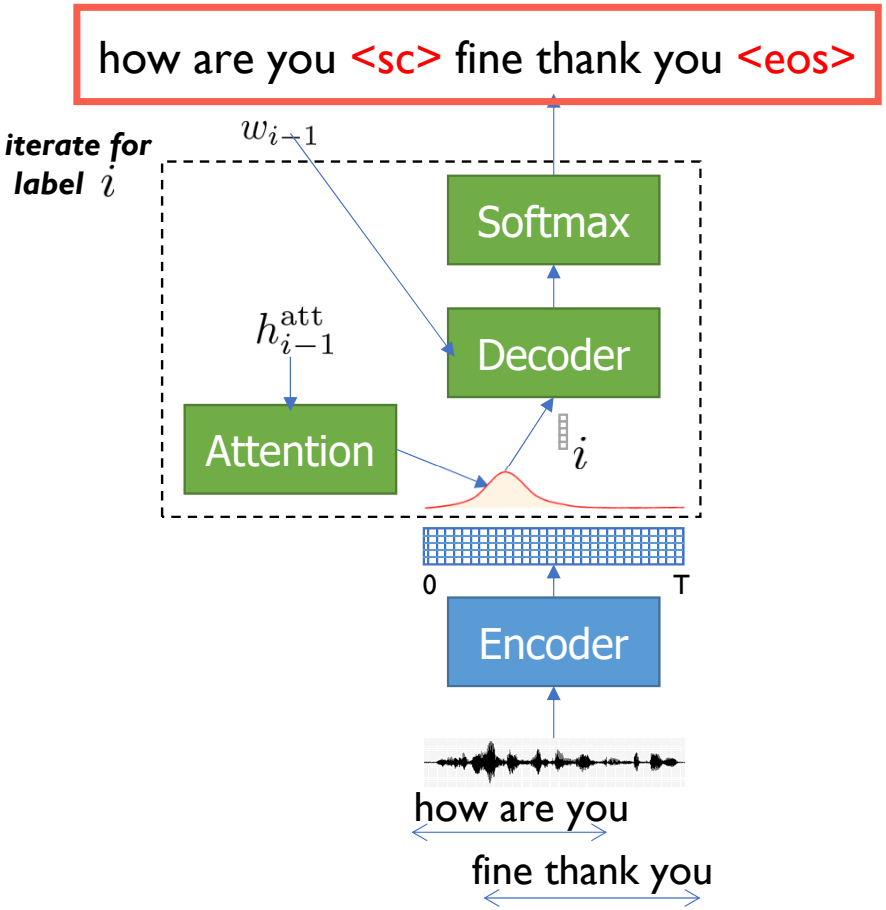
how are you <sc> fine thank you <eos>



N. Kanda et al., Serialized Output Training for End-to-End Overlapped Speech Recognition. Interspeech, 2020.

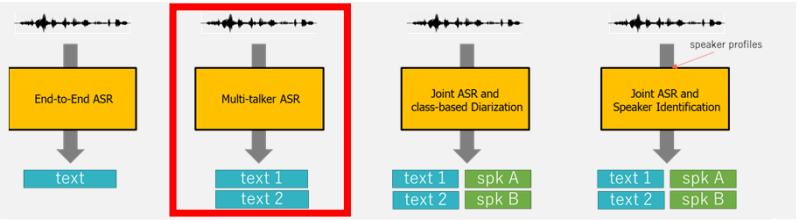
Serialized Output Training (SOT) for End-to-End Overlapped Speech Recognition [Kanda+ 2020]



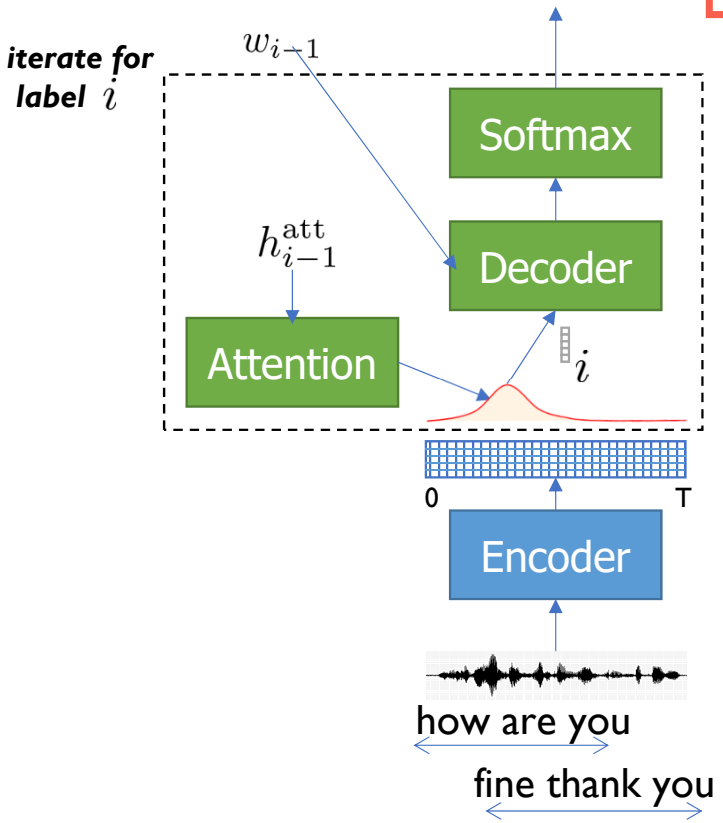


N. Kanda et al., Serialized Output Training for End-to-End Overlapped Speech Recognition. Interspeech, 2020.

Serialized Output Training (SOT) for End-to-End Overlapped Speech Recognition [Kanda+ 2020]



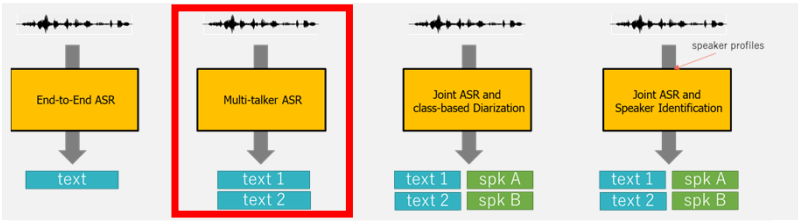
how are you <sc> fine thank you <eos>



- Can recognize **any number of speakers**
- Can **count the number of speakers**

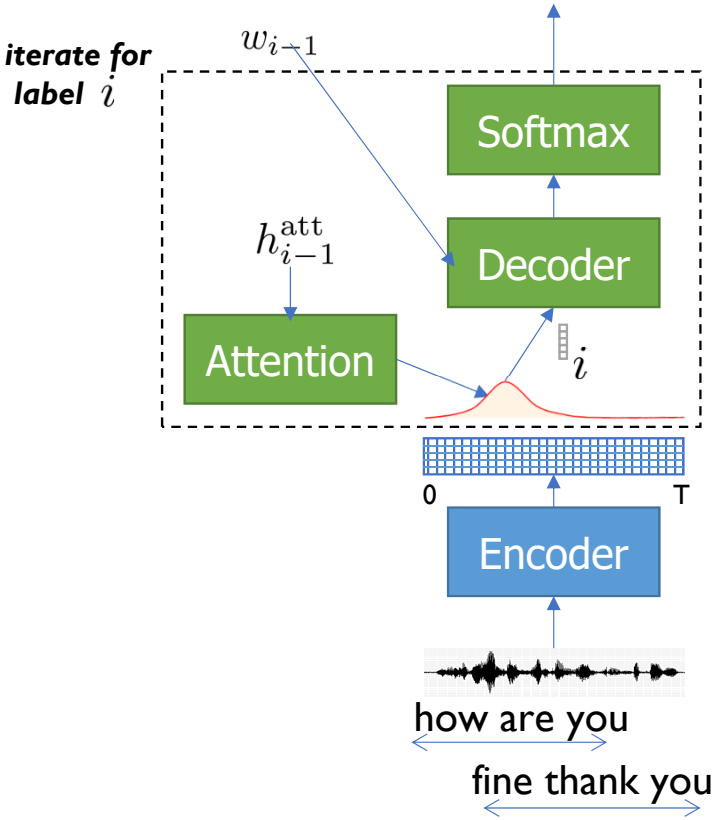
N. Kanda et al., Serialized Output Training for End-to-End Overlapped Speech Recognition. Interspeech, 2020.

Serialized Output Training (SOT) for End-to-End Overlapped Speech Recognition [Kanda+ 2020]



First-In, First-Out (FIFO)

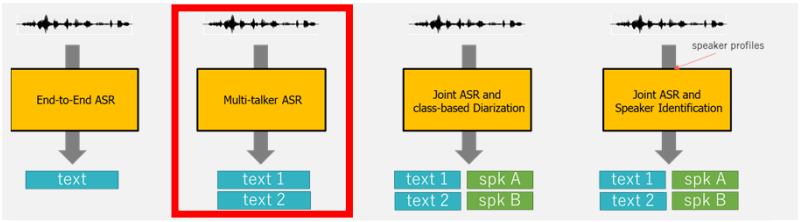
how are you <sc> fine thank you <eos>



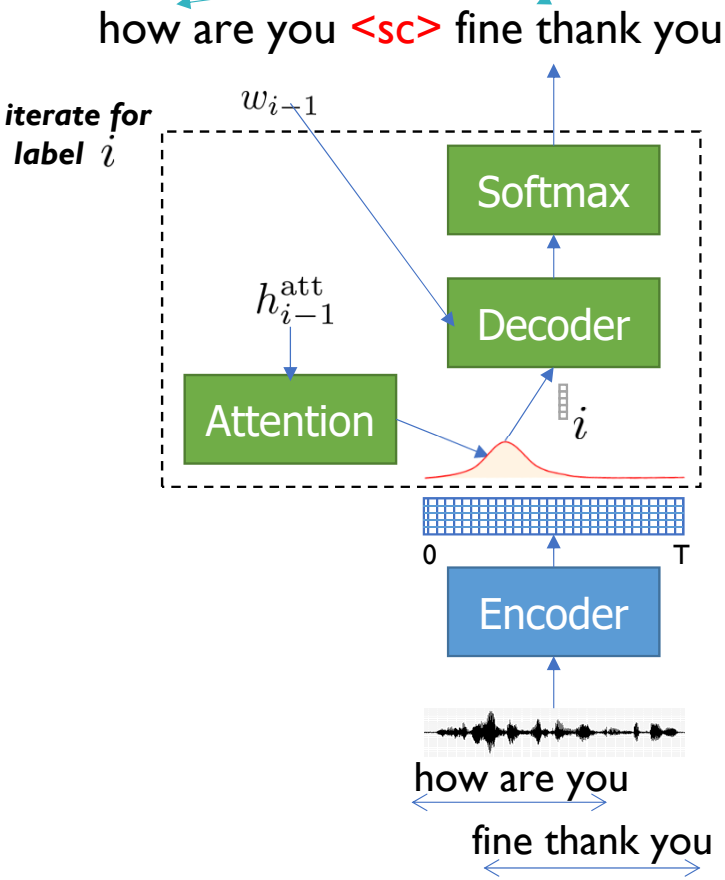
- 😀 Can recognize **any number of speakers**
- 😐 Can **count the number of speakers**
- 😀 FIFO training for O(S) training (Proposed concurrently with [Tripathi+ 2020])

N. Kanda et al., Serialized Output Training for End-to-End Overlapped Speech Recognition. Interspeech, 2020.

Serialized Output Training (SOT) for End-to-End Overlapped Speech Recognition [Kanda+ 2020]



First-In, First-Out (FIFO)



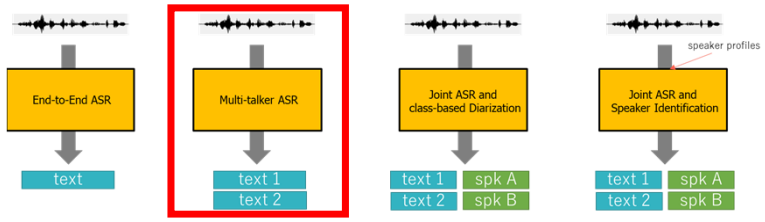
- Can recognize **any number of speakers**
- Can **count the number of speakers**
- FIFO training for O(S) training (Proposed concurrently with [Tripathi+ 2020])

SWER (%) on LibriSpeechMix (from [Kanda+ 2020])

Model (1024-dim)	# of Speakers in Training Data	# of Params.	# of Speakers in Test Data 1	2	3
2-output PIT	2	160.7M	(80.6)	11.1	(52.1)
2-output PIT	1,2	160.7M	6.7	11.9	(52.3)
SOT	1,2,3	135.6M	4.6	11.2	24.0

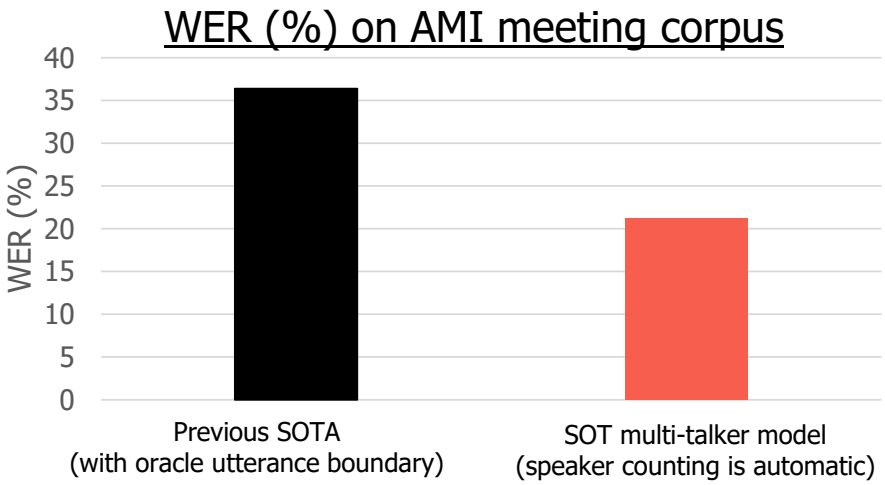
N. Kanda et al., Serialized Output Training for End-to-End Overlapped Speech Recognition. Interspeech, 2020.

Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone [Kanda+ 2021]



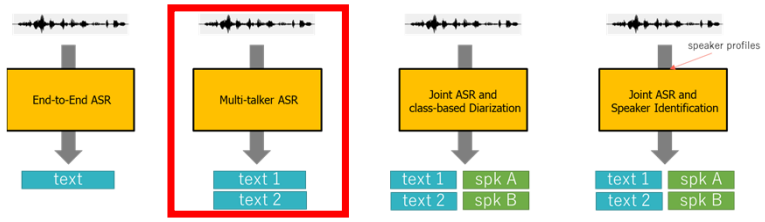
Does multi-talker model really work for real conversation?

→ Yes, SOT model achieved SOTA on AMI single distant microphone test set



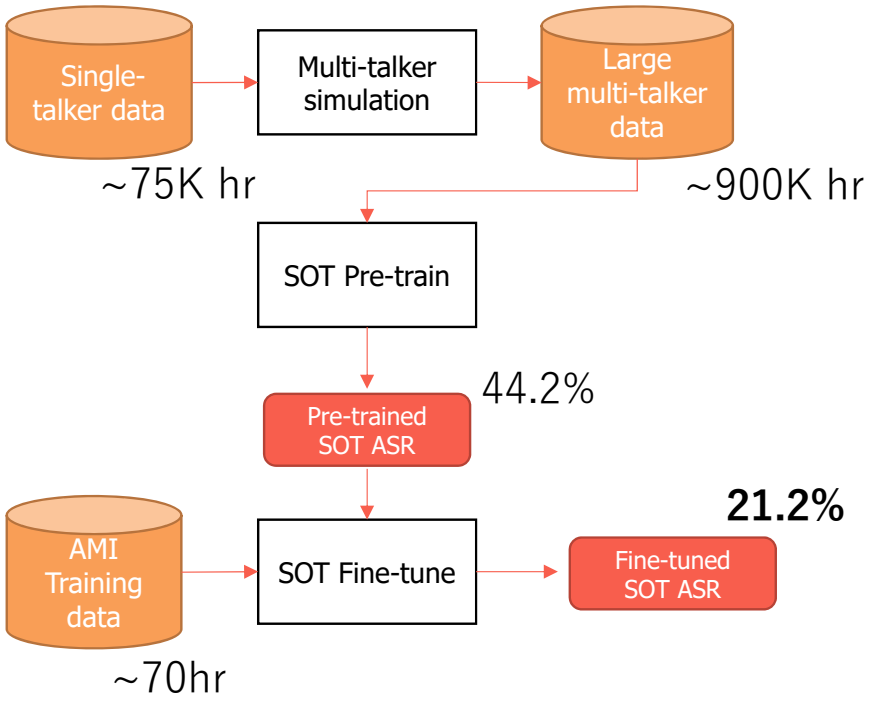
N. Kanda et al., Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. arXiv preprint arXiv:2103.16776, 2021.

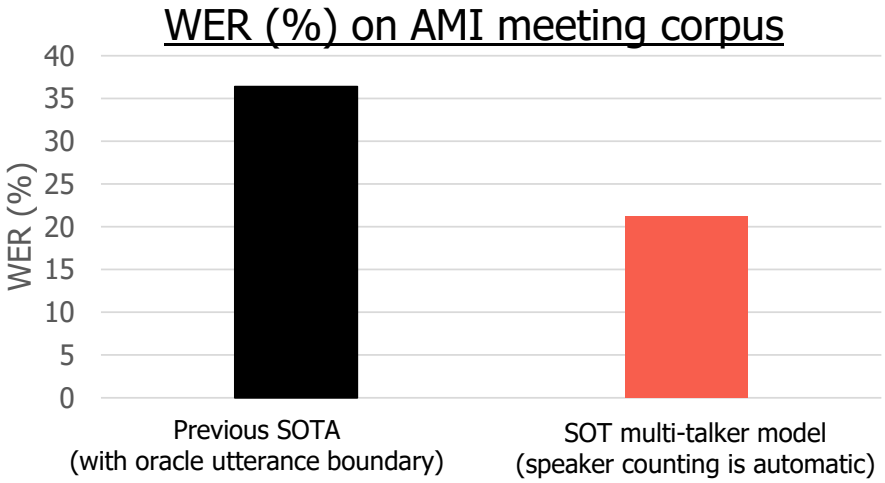
Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone [Kanda+ 2021]



Does multi-talker model really work for real conversation?

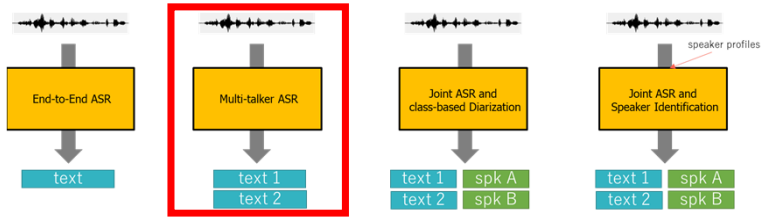
→ Yes, SOT model achieved SOTA on AMI single distant microphone test set





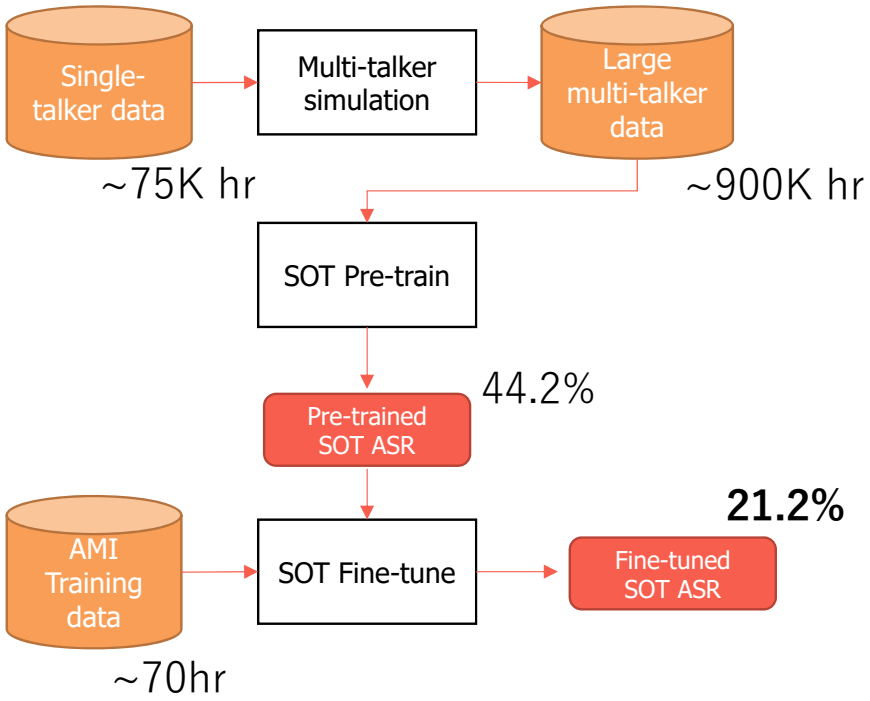
N. Kanda et al., Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. arXiv preprint arXiv:2103.16776, 2021.

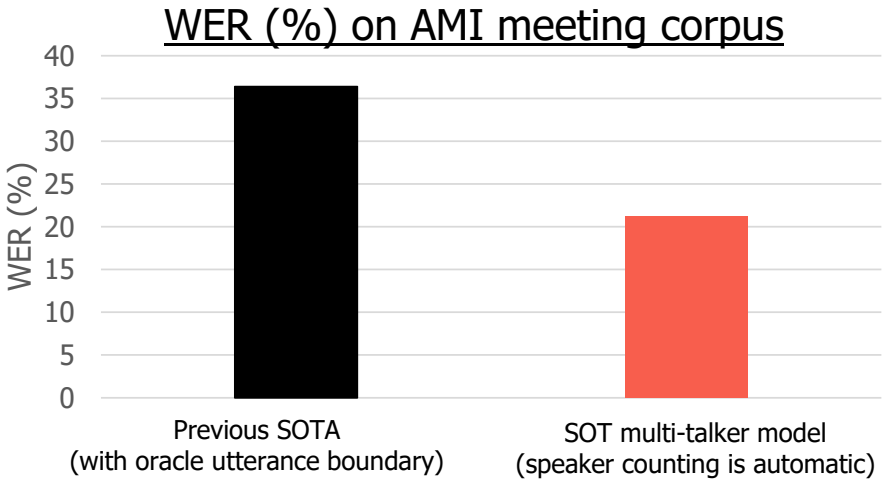
Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone [Kanda+ 2021]



Does multi-talker model really work for real conversation?

→ Yes, SOT model achieved SOTA on AMI single distant microphone test set



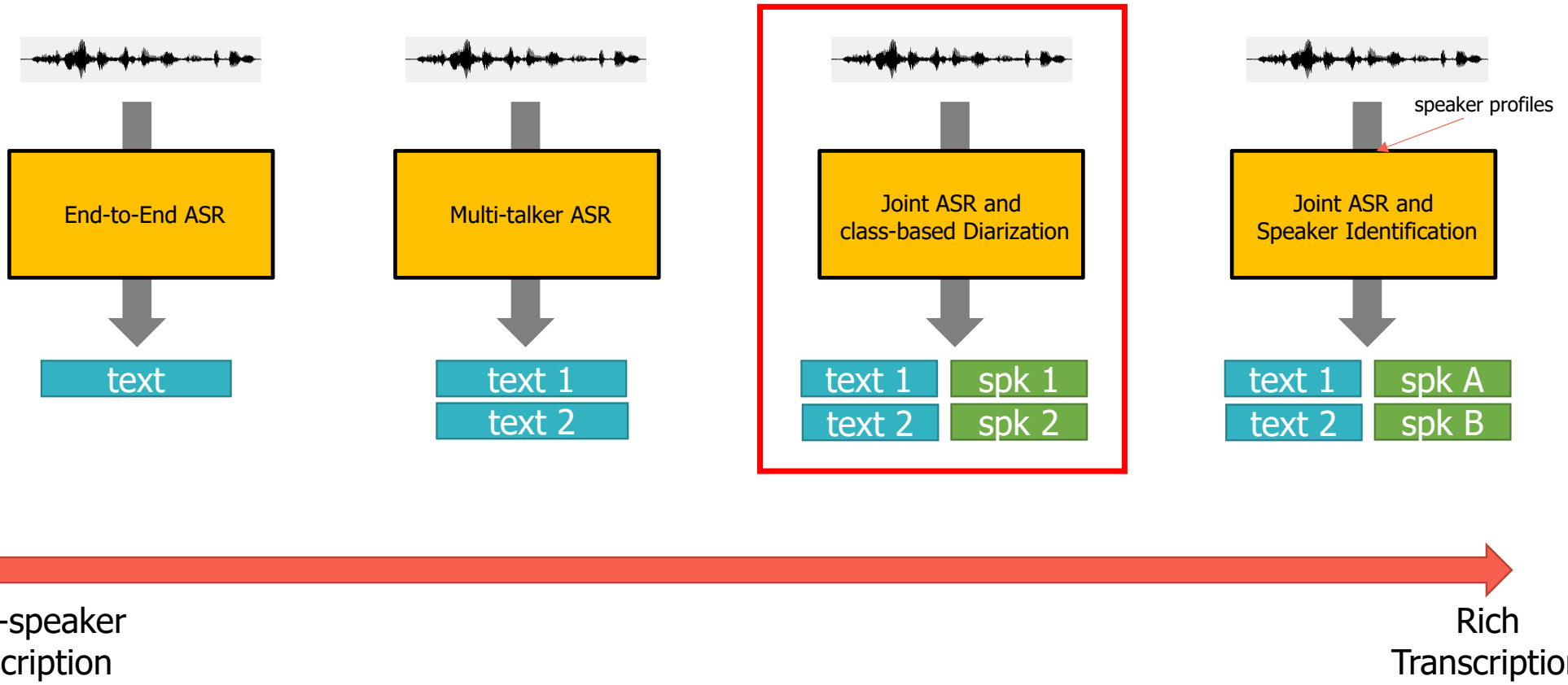


Speaker Counting Accuracy

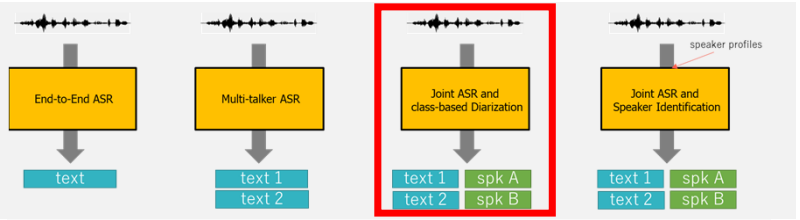
		Estimated # of speakers					
		0	1	2	3	4	5
Actual # of speakers	1	0.2	**97.2**	2.5	0.1	0.0	0.0
	2	0.0	13.7	**80.5**	5.9	0.0	0.0
	3	0.0	2.4	32.6	**60.2**	4.8	0.0
	4	0.0	0.0	9.9	51.2	**38.9**	0.0

N. Kanda et al., Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. arXiv preprint arXiv:2103.16776, 2021.

What has been achieved for "ASR + x"



Joint ASR and ***Class-based*** Diarization

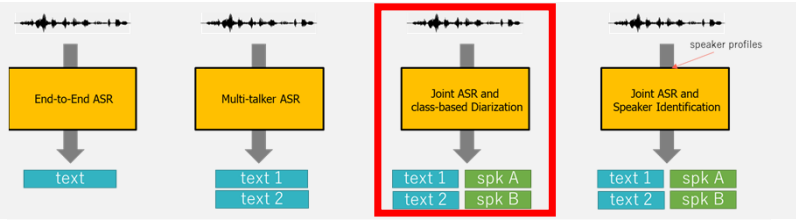


how are you going <**class1**> i have a headache <**class2**>

Insert a ***special symbol that represents speaker*** at the end of each utterance.

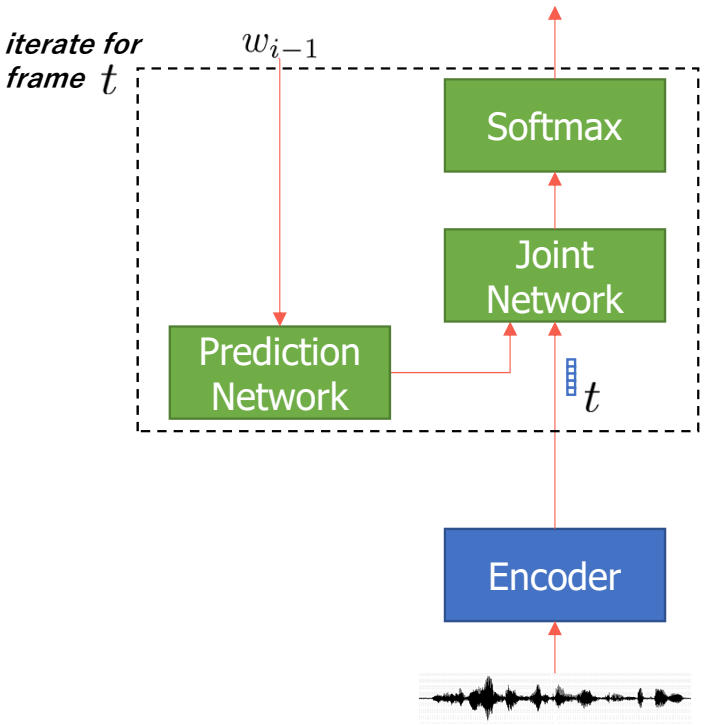
- 😀 Simple. No need to change ASR code.
- 😟 Cannot cope with a class that does not appear in training data.

Joint Speech Recognition and Speaker Diarization via Sequence Transduction [Shafey+ 2019]



speaker role tag

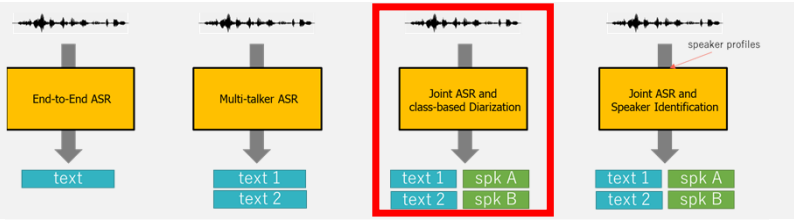
how are you going <doctor> i have a headache <patient>



Insert a ***speaker-role*** as one of output tokens.

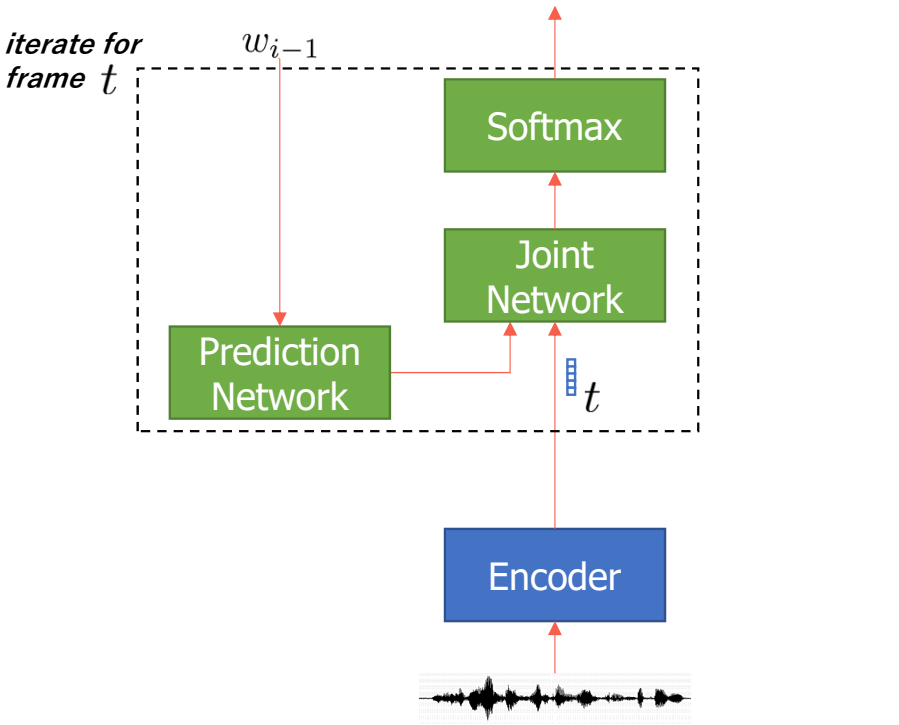
L. El Shafey et al., Joint Speech Recognition and Speaker Diarization via Sequence Transduction. *Proc. Interspeech 2019*, 2019, 396-400.

Joint Speech Recognition and Speaker Diarization via Sequence Transduction [Shafey+ 2019]



speaker role tag

how are you going <doctor> i have a headache <patient>



Insert a ***speaker-role*** as one of output tokens.

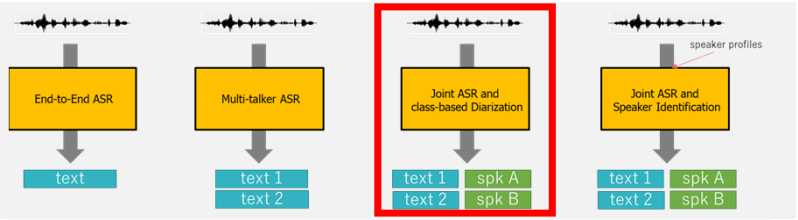
- 😀 Work well for diarization with marginal degradation of ASR accuracy

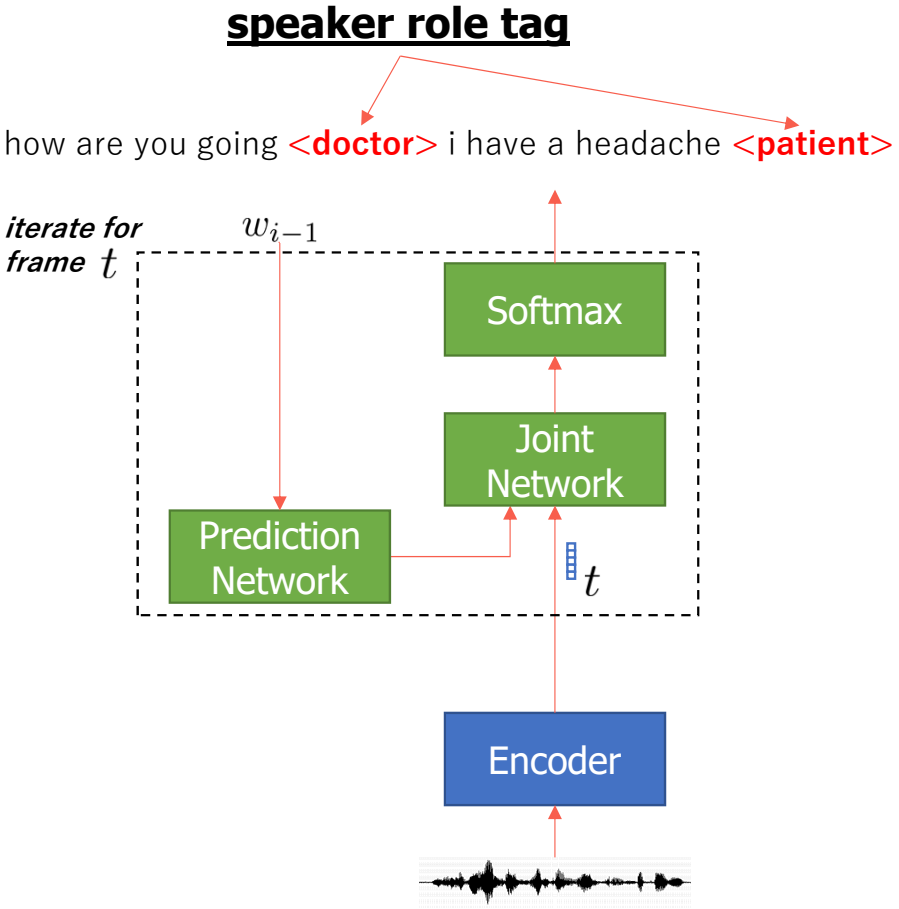
Table 1: *Word Diarization Error Rate* (WDER), *Word Error Rate* (WER) *and its decomposition in Deletion / Insertion / Substitution errors* (D/I/S) *on the evaluation set.*

	Baseline	Joint ASR+SD
WDER	15.8%	**2.2%**
WER	**18.7%**	19.3%
D/I/S	7.2%/2.1%/9.4%	6.8%/2.8%/9.7%

L. El Shafey et al., Joint Speech Recognition and Speaker Diarization via Sequence Transduction. *Proc. Interspeech 2019*, 2019, 396-400.

Joint Speech Recognition and Speaker Diarization via Sequence Transduction [Shafey+ 2019]





Insert a ***speaker-role*** as one of output tokens.

- 😀 Work well for diarization with marginal degradation of ASR accuracy

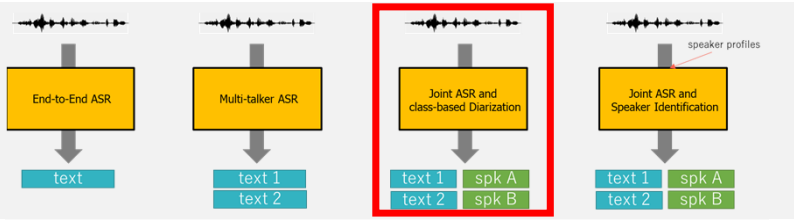
Table 1: *Word Diarization Error Rate* (WDER), *Word Error Rate* (WER) *and its decomposition in Deletion / Insertion / Substitution errors* (D/I/S) *on the evaluation set.*

	Baseline	Joint ASR+SD
WDER	15.8%	**2.2%**
WER	**18.7%**	19.3%
D/I/S	7.2%/2.1%/9.4%	6.8%/2.8%/9.7%

- 😕 Work only for trained speaker-roles (such as doctor and patient)

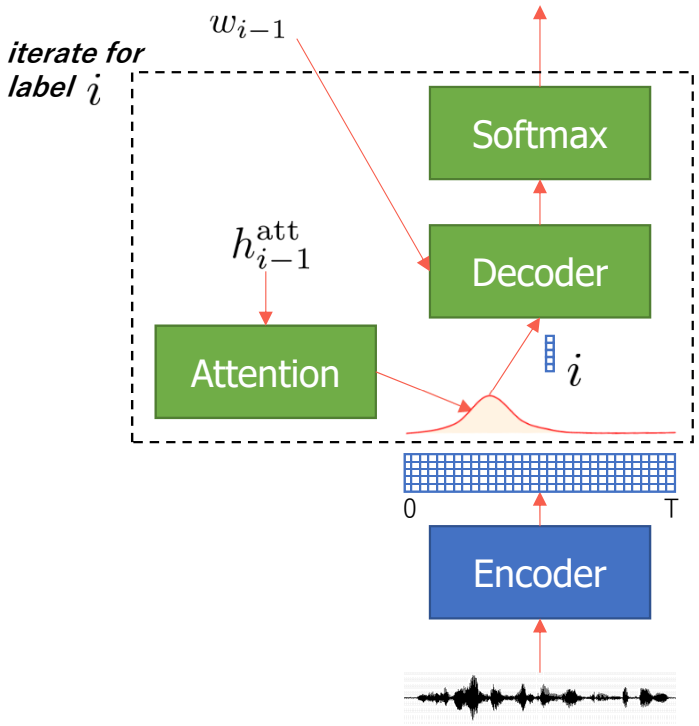
L. El Shafey et al., Joint Speech Recognition and Speaker Diarization via Sequence Transduction. *Proc. Interspeech 2019*, 2019, 396-400.

Speech Recognition and Multi-Speaker Diarization of Long Conversations [Mao+ 2020]



speaker ID

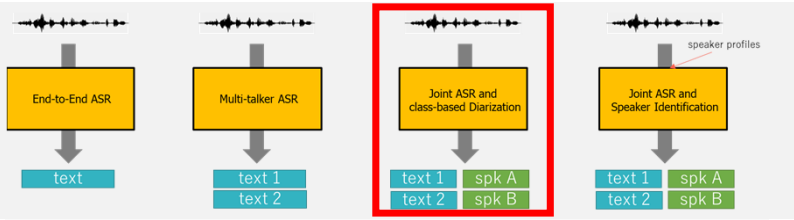
how are you going <speaker 91> i have a headache <speaker 15>



Insert a ***speaker-id*** as one of output tokens.

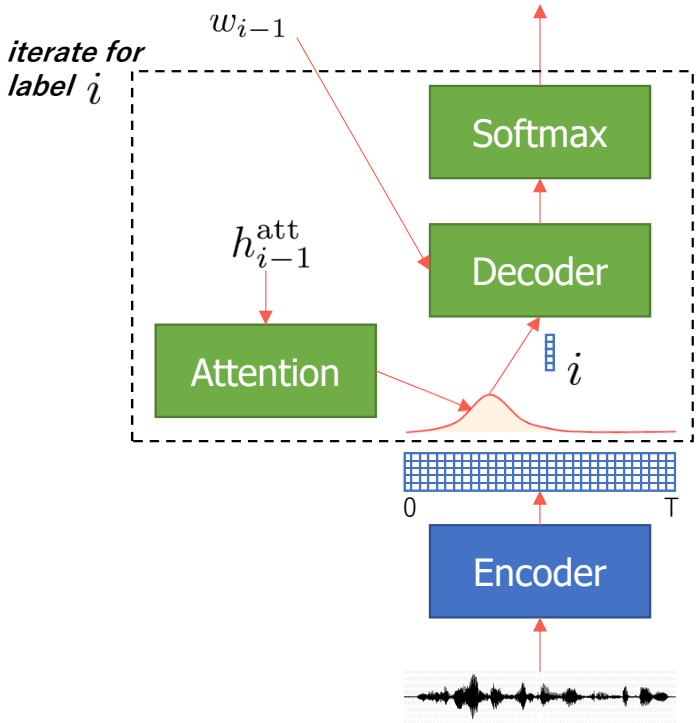
H. H. Mao et al., Speech Recognition and Multi-Speaker Diarization of Long Conversations. *Proc. Interspeech 2020*, 2020, 691-695.

Speech Recognition and Multi-Speaker Diarization of Long Conversations [Mao+ 2020]



speaker ID

how are you going <speaker 91> i have a headache <speaker 15>



Insert a ***speaker-id*** as one of output tokens.

- 😀 Work well for a difficult case where utterance boundary is not given.
- 😟 Degradation when utterance boundary is given.

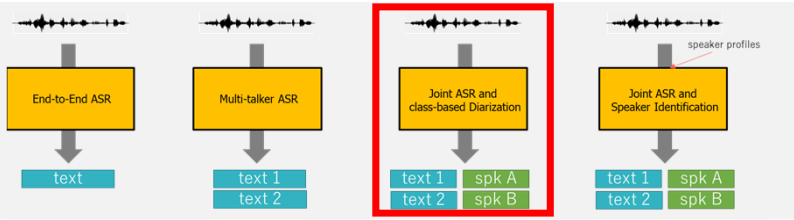
Result for "This American Life" radio program test set (from [Mao+ 2020])

	Utt. boundary is given		Utt. boundary is not given	
	WER	DER	WER	DER
Separate	24.3	15.4	58.3	91.3
Joint(*)	25.4	31.9	58.2	62.2

(*) Mao et al. also proposed pre-training and data augmentation which showed a better result, but we extract the results without these techniques to highlight the pure difference between the separate model and joint model.

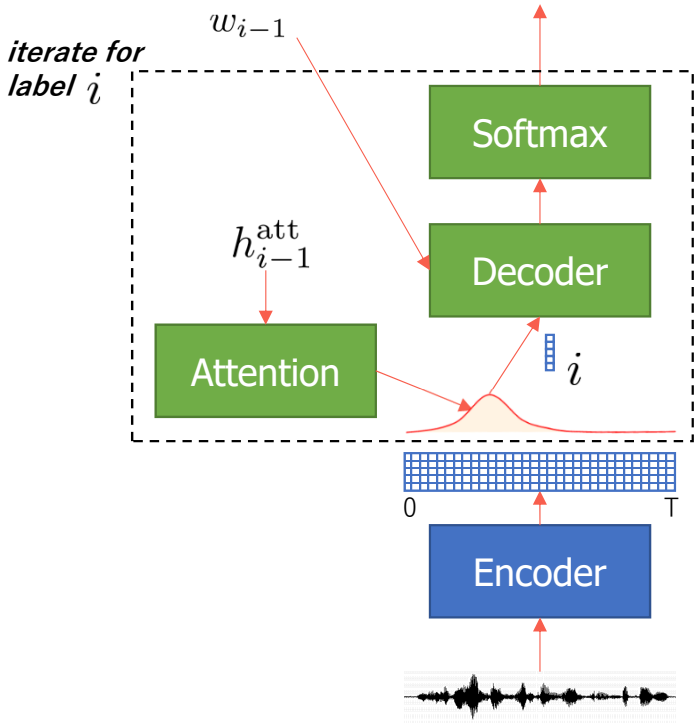
H. H. Mao et al., Speech Recognition and Multi-Speaker Diarization of Long Conversations. *Proc. Interspeech 2020*, 2020, 691-695.

Speech Recognition and Multi-Speaker Diarization of Long Conversations [Mao+ 2020]



speaker ID

how are you going <speaker 91> i have a headache <speaker 15>



Insert a ***speaker-id*** as one of output tokens.

- 😀 Work well for a difficult case where utterance boundary is not given.
- 😕 Degradation when utterance boundary is given.

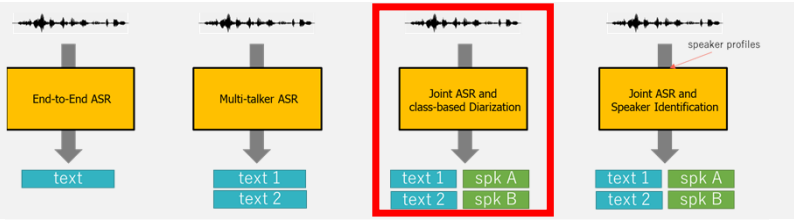
Result for "This American Life" radio program test set (from [Mao+ 2020])

	Utt. boundary is given		Utt. boundary is not given	
	WER	DER	WER	DER
Separate	24.3	15.4	58.3	91.3
Joint(*)	25.4	31.9	58.2	62.2

(*) Mao et al. also proposed pre-training and data augmentation which showed a better result, but we extract the results without these techniques to highlight the pure difference between the separate model and joint model.

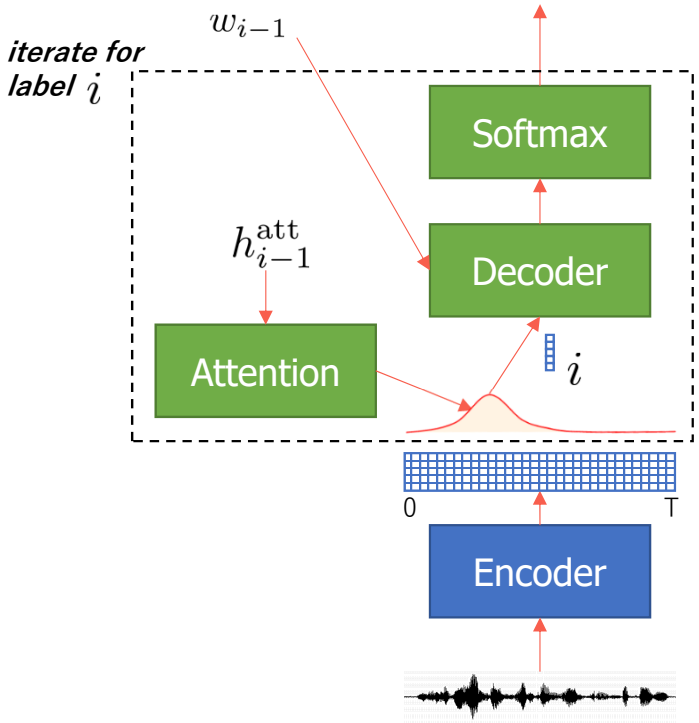
H. H. Mao et al., Speech Recognition and Multi-Speaker Diarization of Long Conversations. *Proc. Interspeech 2020*, 2020, 691-695.

Speech Recognition and Multi-Speaker Diarization of Long Conversations [Mao+ 2020]



speaker ID

how are you going <speaker 91> i have a headache <speaker 15>



Insert a ***speaker-id*** as one of output tokens.

- 😀 Work well for a difficult case where utterance boundary is not given.
- 😕 Degradation when utterance boundary is given.

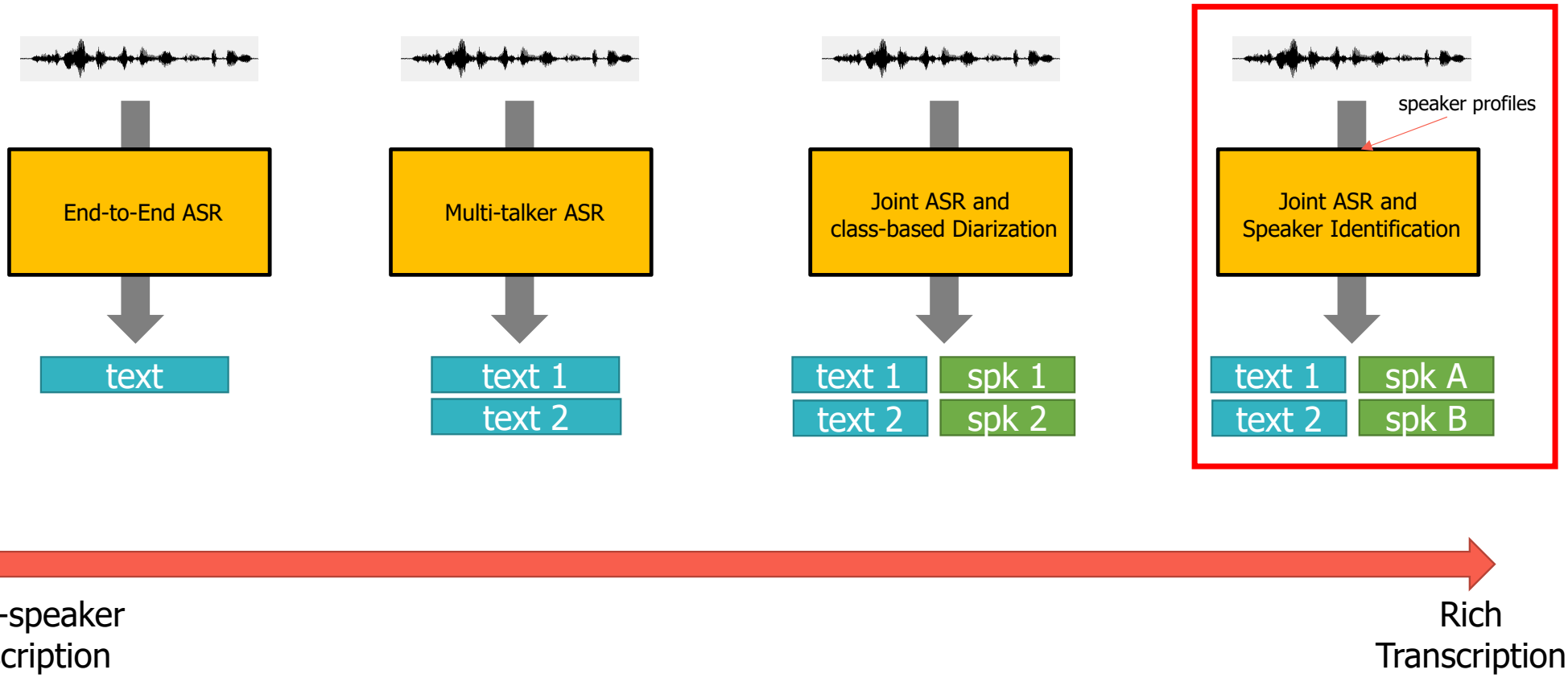
Result for "This American Life" radio program test set (from [Mao+ 2020])

	Utt. boundary is given		Utt. boundary is not given	
	WER	DER	WER	DER
Separate	24.3	15.4	58.3	91.3
Joint(*)	25.4	31.9	58.2	62.2

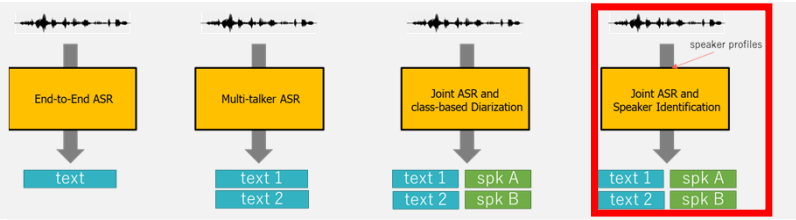
(*) Mao et al. also proposed pre-training and data augmentation which showed a better result, but we extract the results without these techniques to highlight the pure difference between the separate model and joint model.

H. H. Mao et al., Speech Recognition and Multi-Speaker Diarization of Long Conversations. *Proc. Interspeech 2020*, 2020, 691-695.

What has been achieved for "ASR + x"



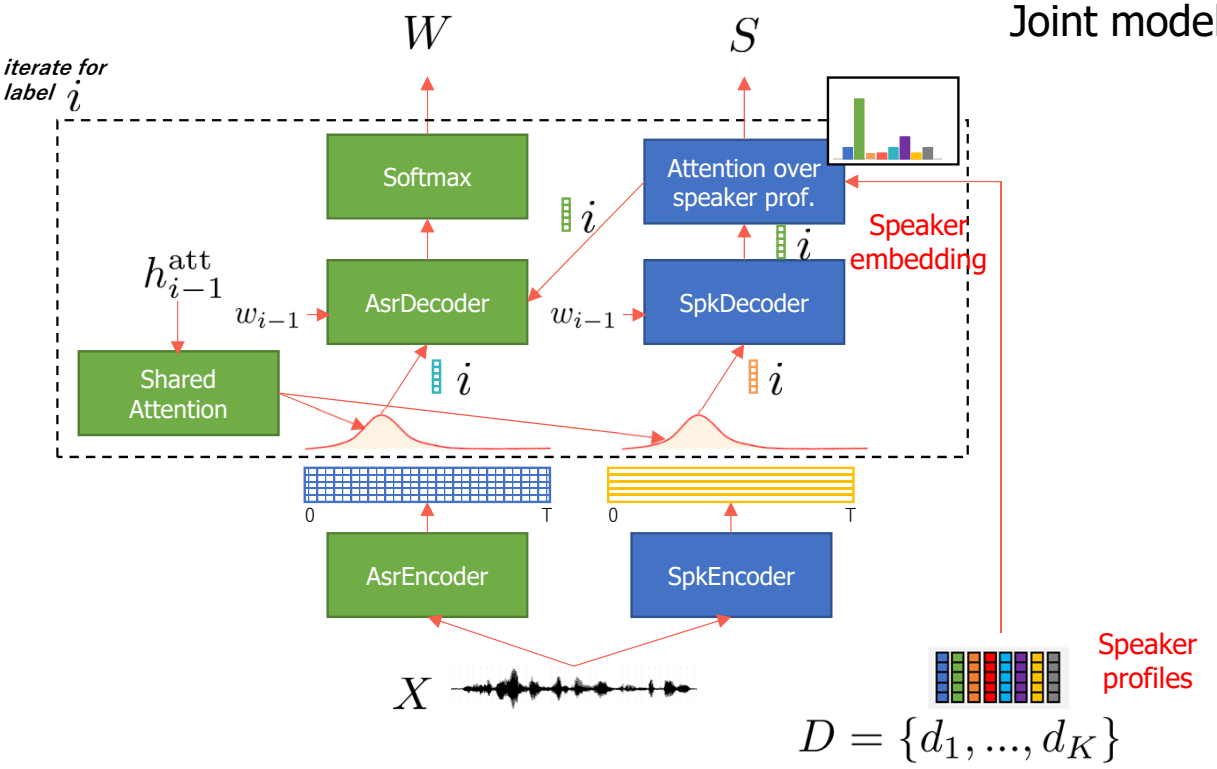
Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



W how are you <sc> i am fine thank you <eos>

S 2 2 2 2 5 5 5 5 5 5

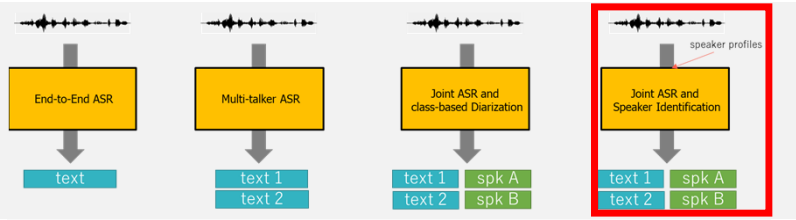
Joint model of ASR and ***speaker identification***



End-to-End Speaker-Attributed ASR

N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

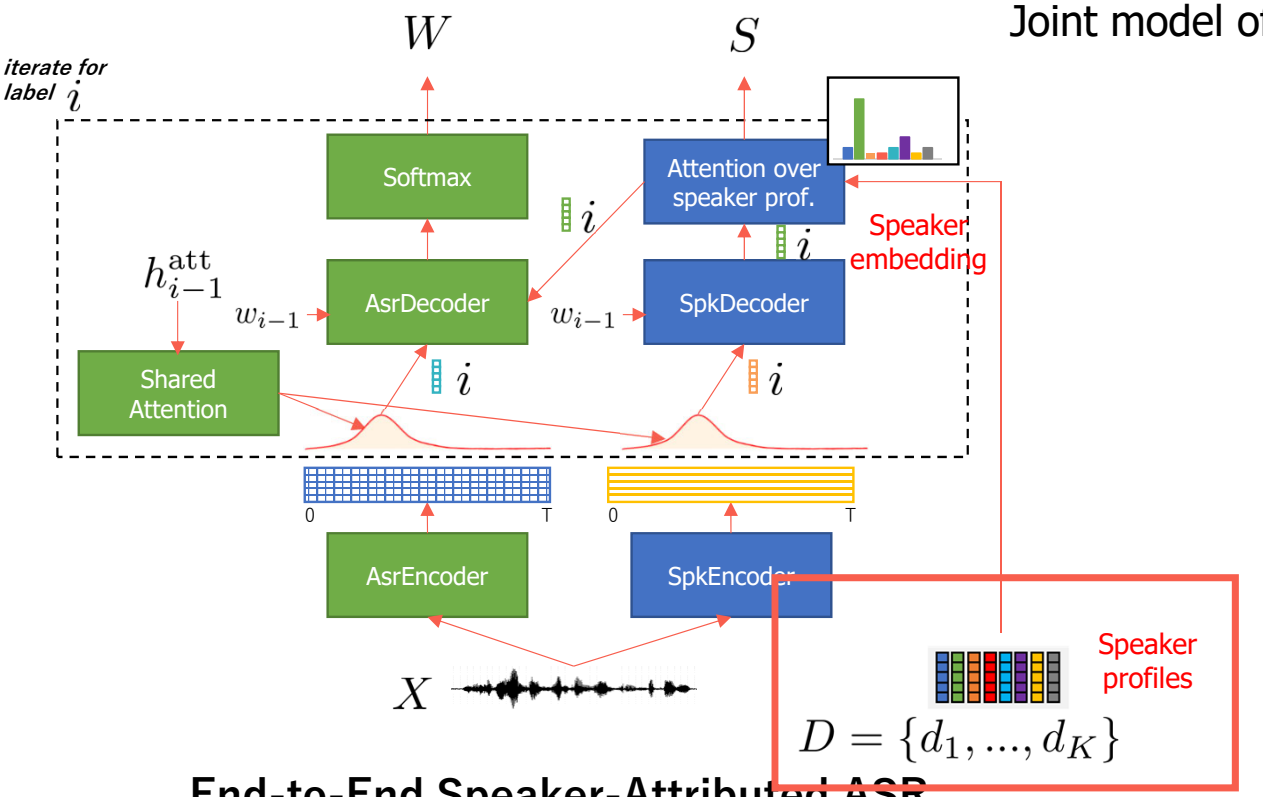
Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



W how are you <sc> i am fine thank you <eos>

S 2 2 2 2 5 5 5 5 5 5

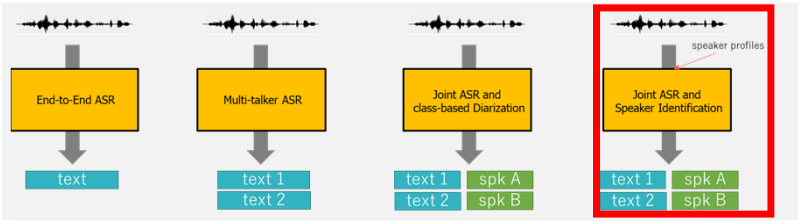
Joint model of ASR and ***speaker identification***

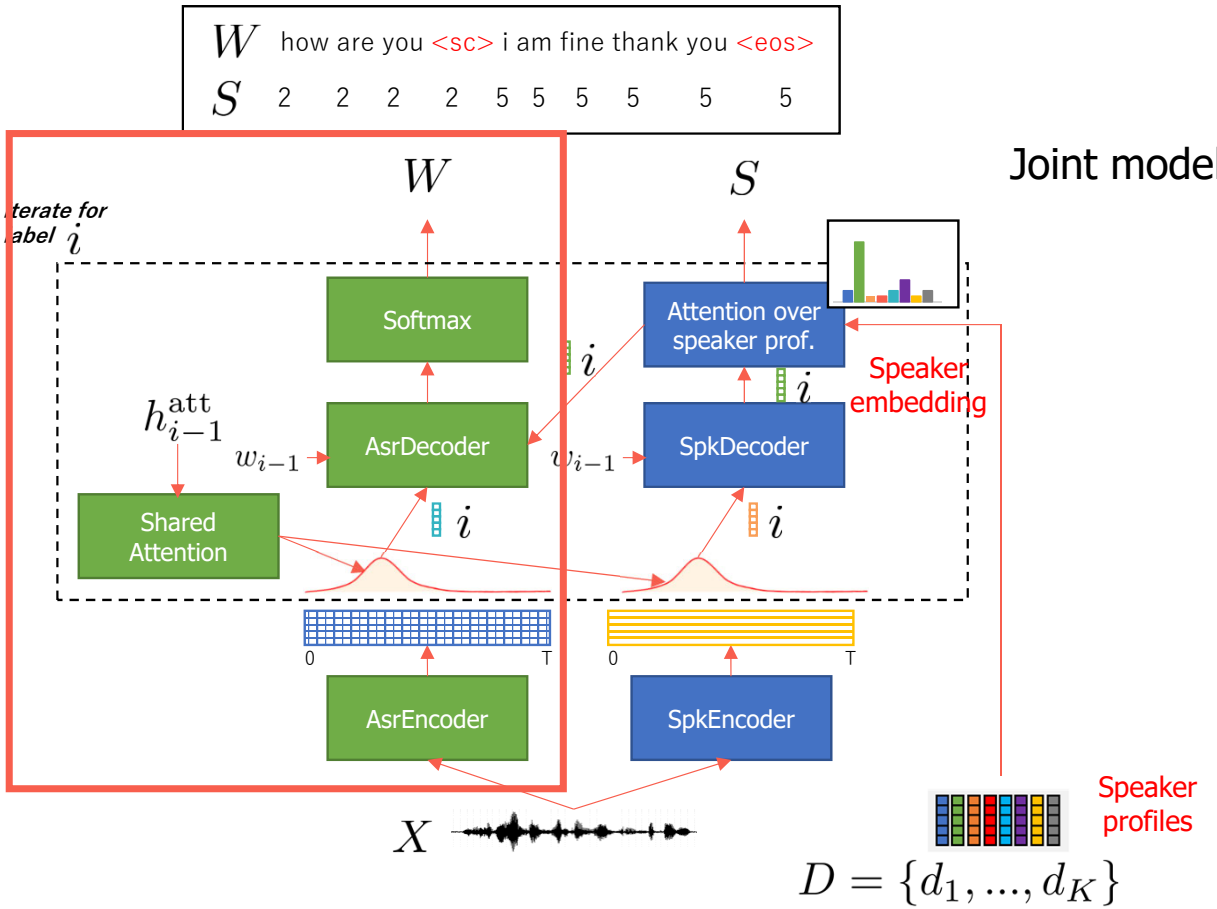


End-to-End Speaker-Attributed ASR

N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



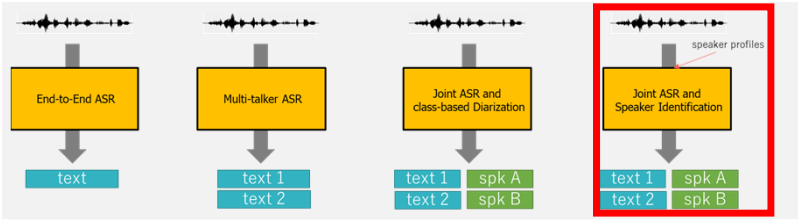


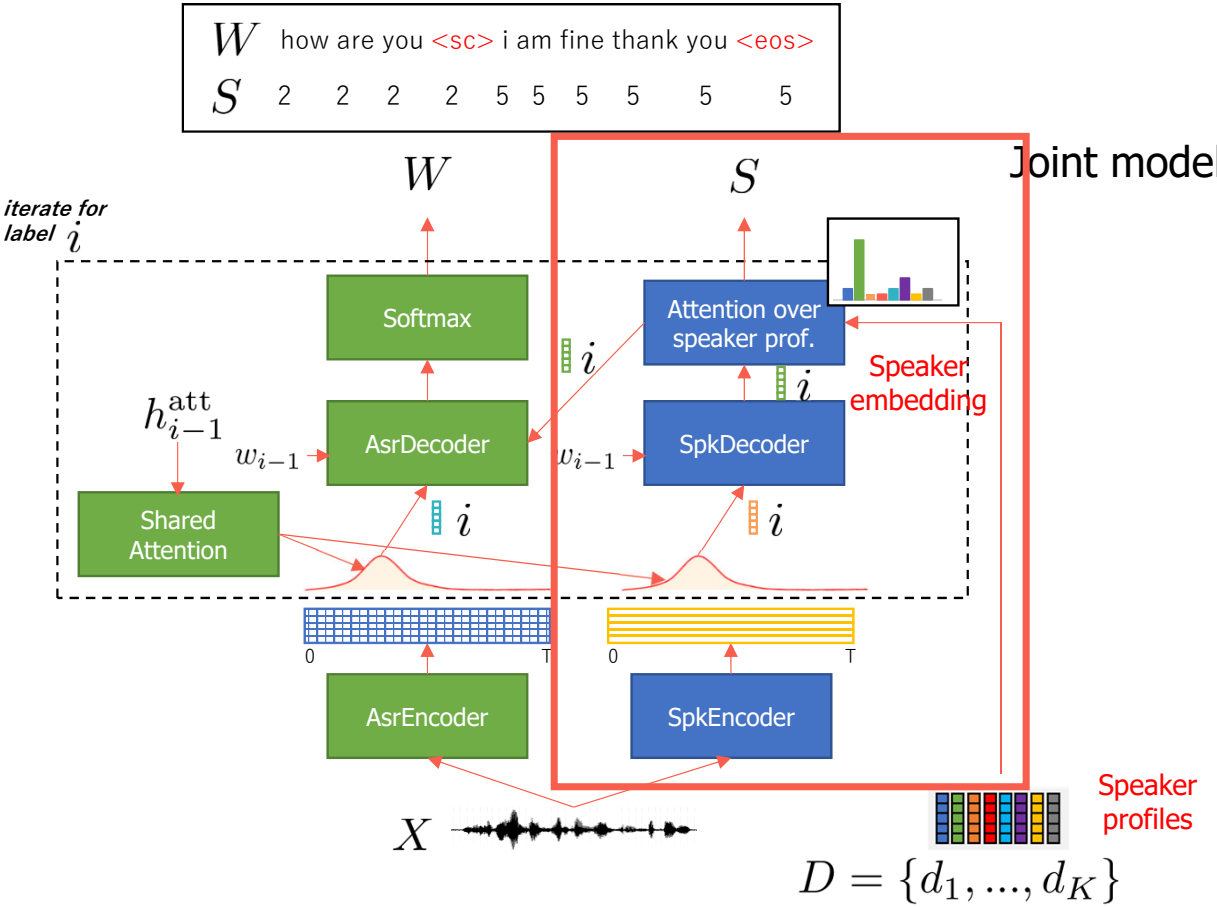
Joint model of ASR and ***speaker identification***

End-to-End Speaker-Attributed ASR

N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



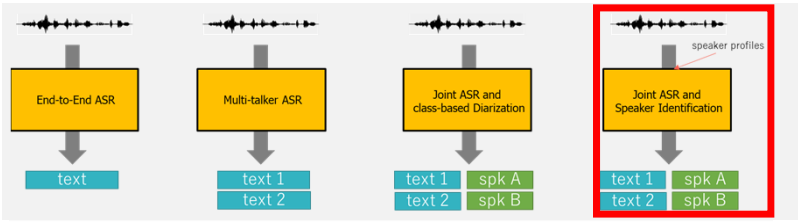


Joint model of ASR and ***speaker identification***

End-to-End Speaker-Attributed ASR

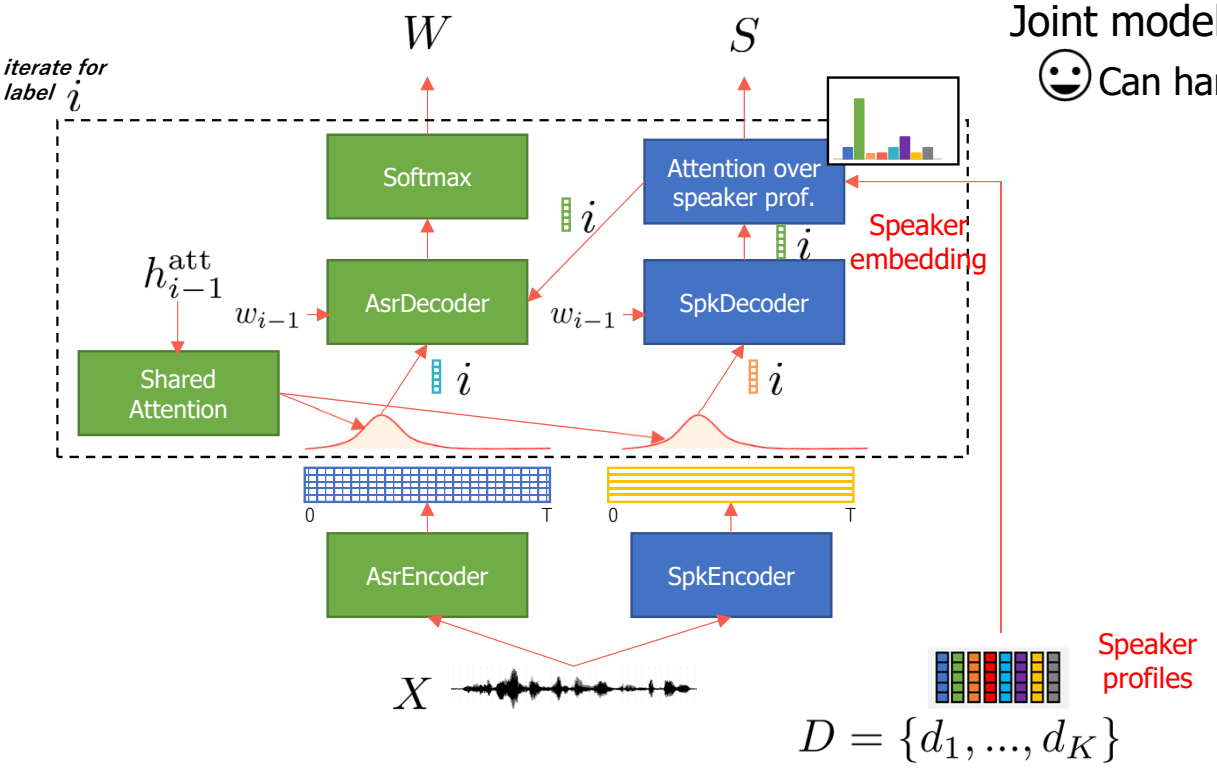
N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



W how are you <sc> i am fine thank you <eos>

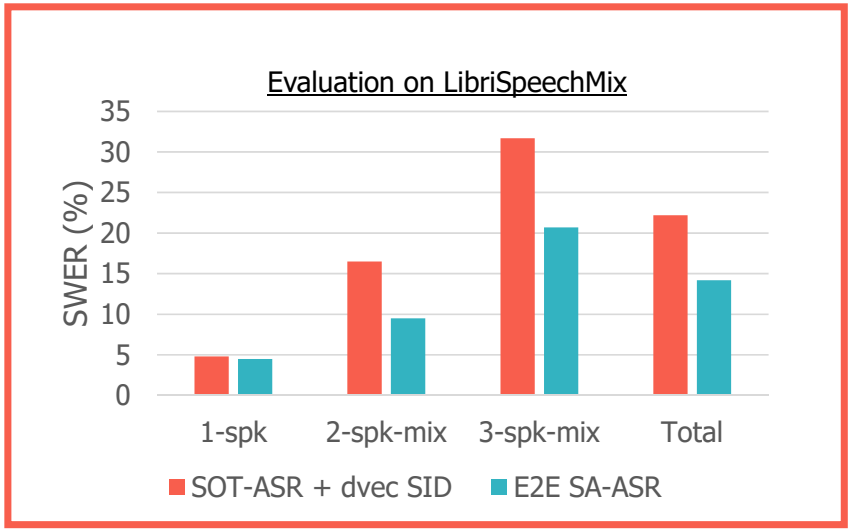
S 2 2 2 2 5 5 5 5 5 5



End-to-End Speaker-Attributed ASR

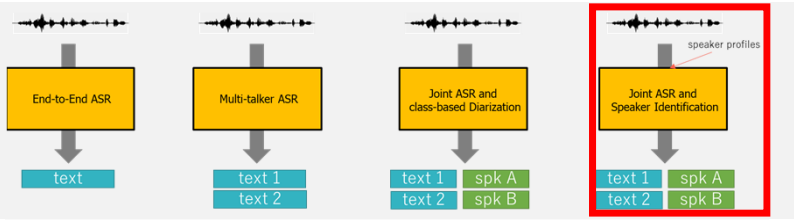
Joint model of ASR and ***speaker identification***

☺ Can handle any number of speakers from overlapped speech



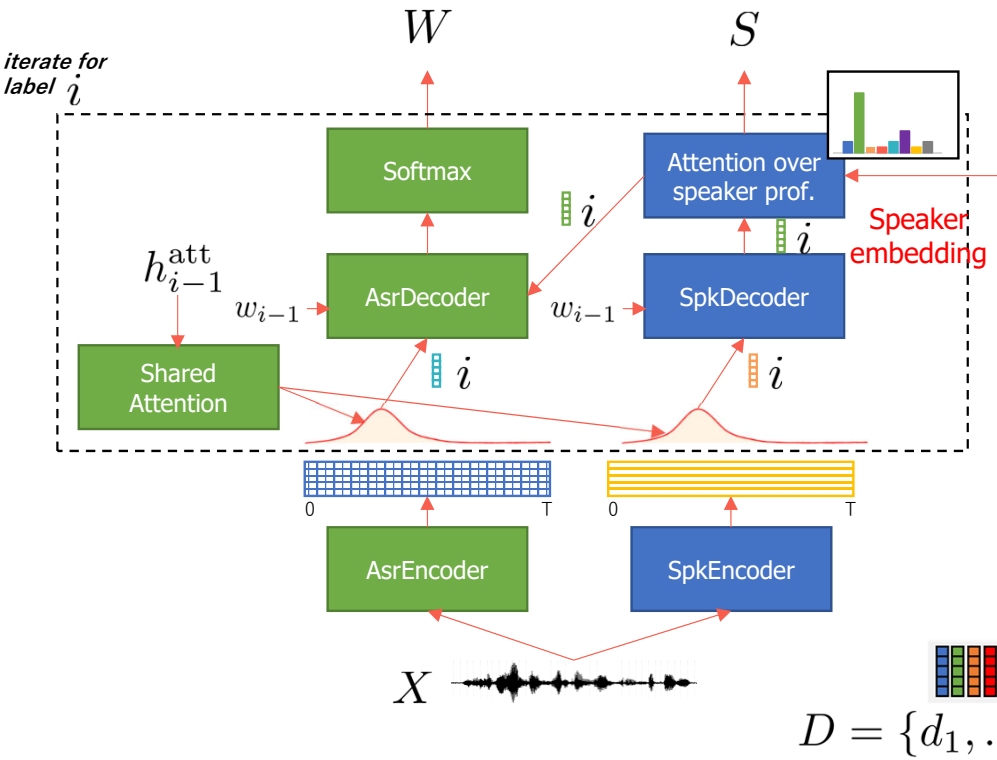
N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



W how are you <sc> i am fine thank you <eos>

S 2 2 2 2 5 5 5 5 5 5

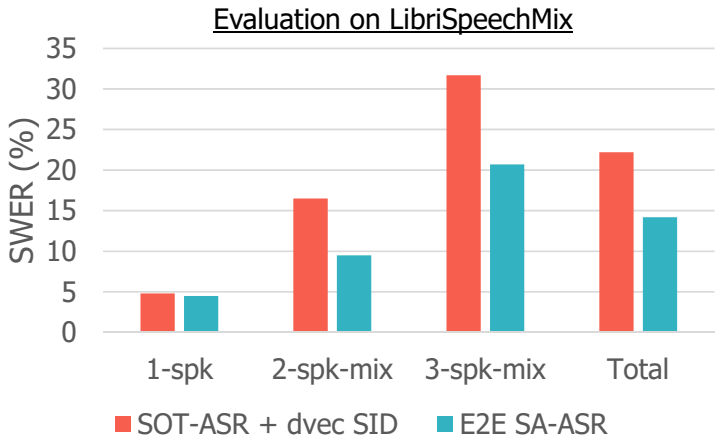


Speaker profiles

$D = \{d_1, ..., d_K\}$

Joint model of ASR and ***speaker identification***

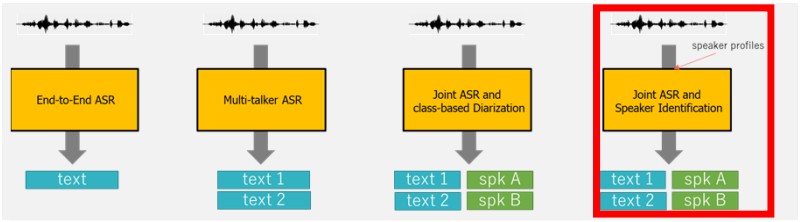
- Can handle any number of speakers from overlapped speech
- Can be optimized by minimizing expected SWER [Kanda+ 2021]
- *Streaming* model is also recently proposed [Lu+ 2021]



End-to-End Speaker-Attributed ASR

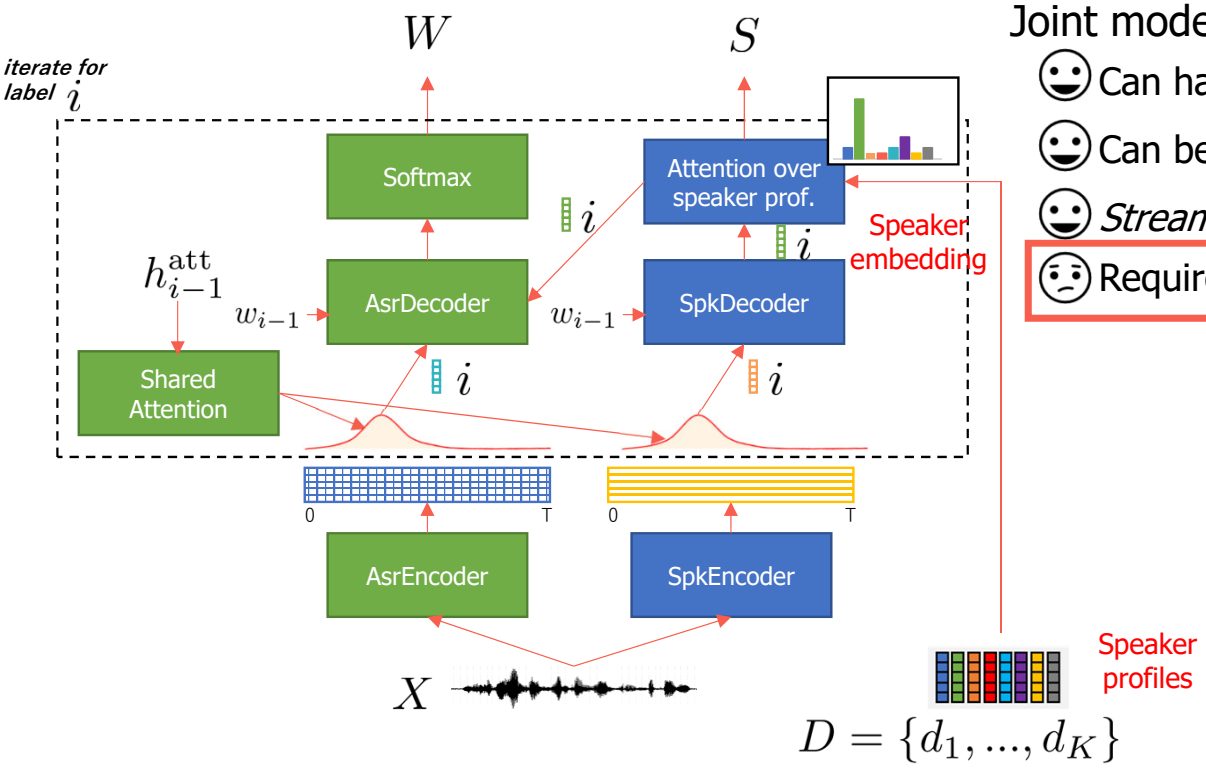
N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers [Kanda+ 2020]



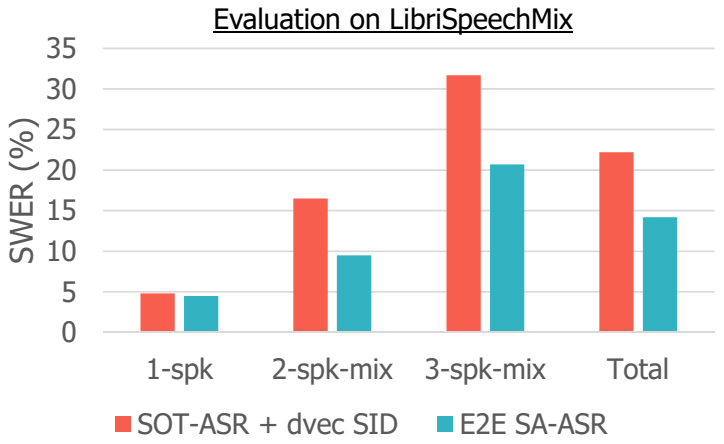
W how are you <sc> i am fine thank you <eos>

S 2 2 2 2 5 5 5 5 5 5



Joint model of ASR and ***speaker identification***

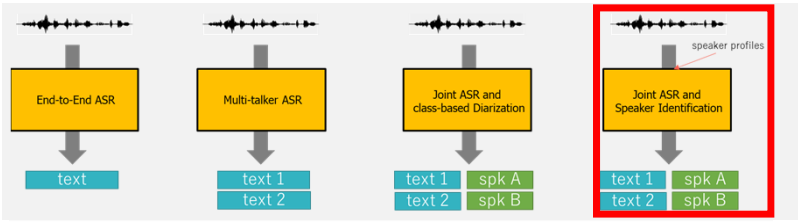
- Can handle any number of speakers from overlapped speech
- Can be optimized by minimizing expected SWER [Kanda+ 2021]
- *Streaming* model is also recently proposed [Lu+ 2021]
- Requires speaker profiles as an auxiliary input (not pure diarization)



End-to-End Speaker-Attributed ASR

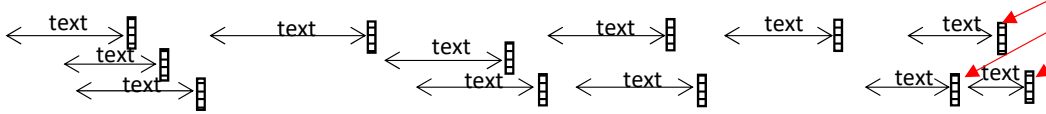
N. Kanda et al., Joint Speaker Counting, Speech Recognition, and Speaker Identification for Overlapped Speech of Any Number of Speakers, Interspeech 2020.
N. Kanda et al., Minimum Bayes Risk Training for End-to-End Speaker-Attributed ASR, ICASSP 2021.
L. Lu et al., Streaming Multi-talker Speech Recognition with Joint Speaker Identification, arXiv, 2021.

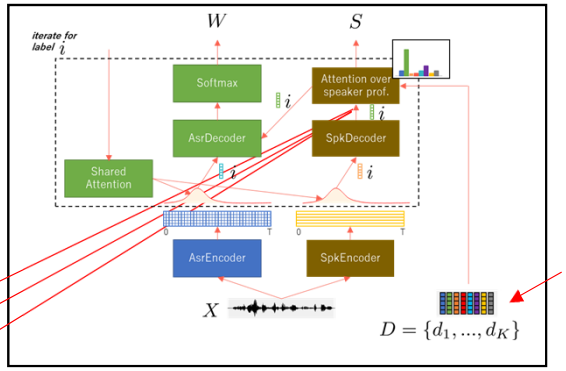
E2E SA-ASR with speaker clustering [Kanda+ 2021]



(1) Segmentation with silence detection

(2) Apply E2E SA-ASR with *dummy* speaker profiles

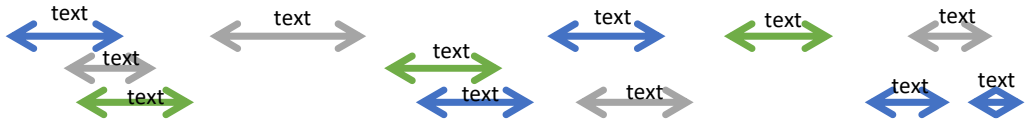




Feed ***dummy*** speaker profiles extracted from training data

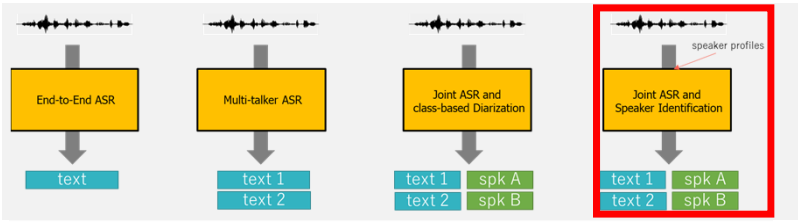
Remember averaged speaker embeddings for each utterance

(3) Speaker counting and clustering based on internal embeddings of E2E SA-ASR



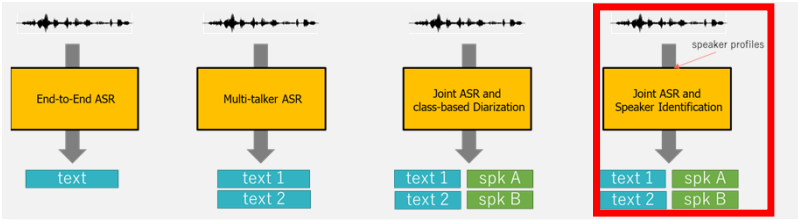
N. Kanda et al., Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, SLT, 2021.
N. Kanda et al., End-to-End Speaker-Attributed ASR with Transformer, ArXiv, 2021.

E2E SA-ASR with speaker clustering [Kanda+ 2021]



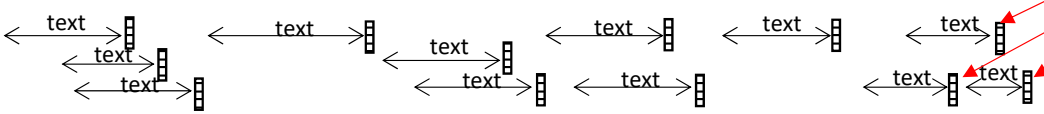
N. Kanda et al., Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, SLT, 2021.
N. Kanda et al., End-to-End Speaker-Attributed ASR with Transformer, ArXiv, 2021.

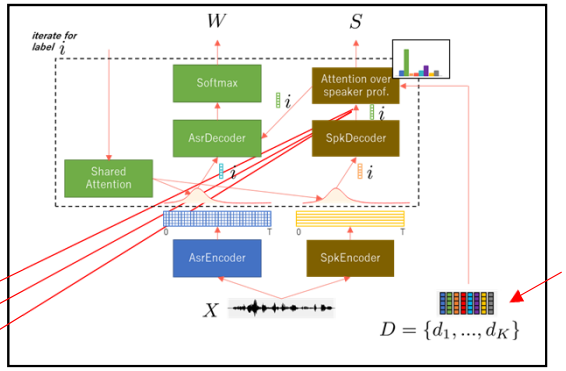
E2E SA-ASR with speaker clustering [Kanda+ 2021]



(1) Segmentation with silence detection

(2) Apply E2E SA-ASR with *dummy* speaker profiles



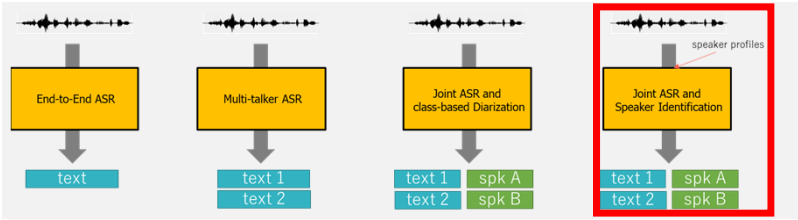


Feed ***dummy*** speaker profiles extracted from training data

Remember averaged speaker embeddings for each utterance

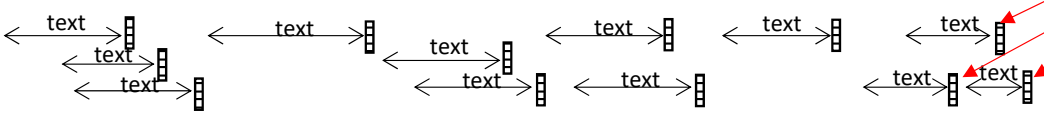
N. Kanda et al., Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, SLT, 2021.
N. Kanda et al., End-to-End Speaker-Attributed ASR with Transformer, ArXiv, 2021.

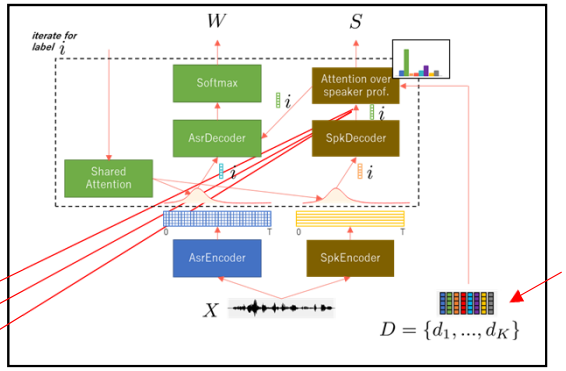
E2E SA-ASR with speaker clustering [Kanda+ 2021]



(1) Segmentation with silence detection

(2) Apply E2E SA-ASR with *dummy* speaker profiles

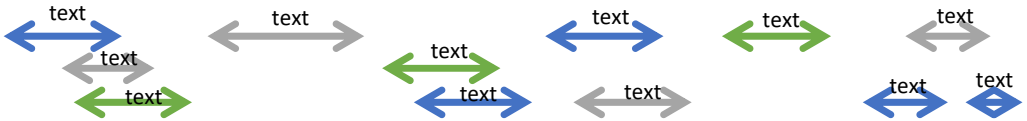




Feed ***dummy*** speaker profiles extracted from training data

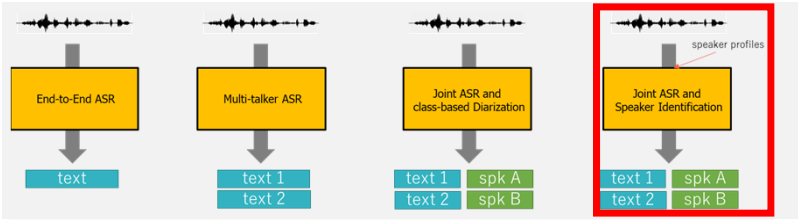
Remember averaged speaker embeddings for each utterance

(3) Speaker counting and clustering based on internal embeddings of E2E SA-ASR



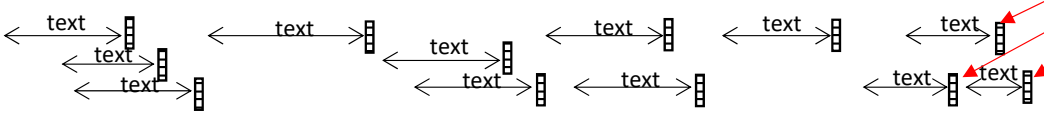
N. Kanda et al., Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, SLT, 2021.
N. Kanda et al., End-to-End Speaker-Attributed ASR with Transformer, ArXiv, 2021.

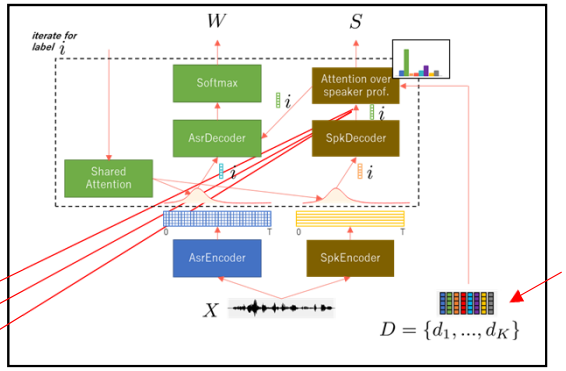
E2E SA-ASR with speaker clustering [Kanda+ 2021]



(1) Segmentation with silence detection

(2) Apply E2E SA-ASR with *dummy* speaker profiles

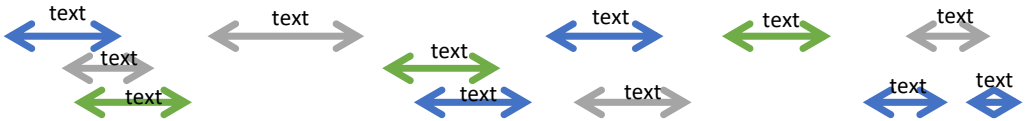




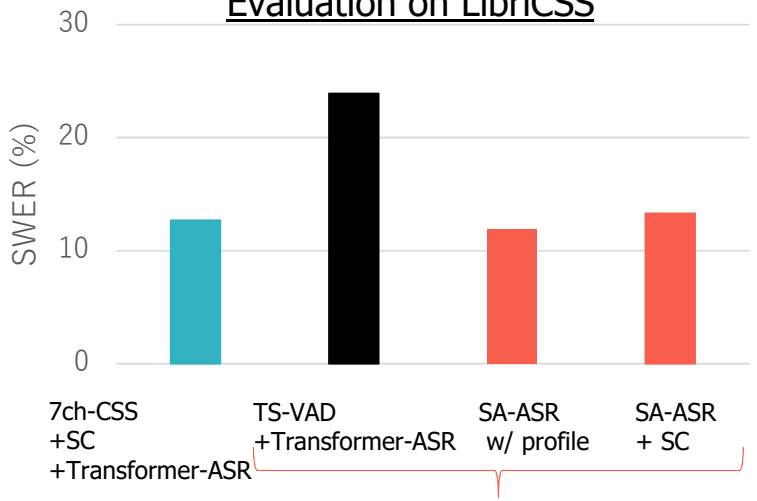
Feed ***dummy*** speaker profiles extracted from training data

Remember averaged speaker embeddings for each utterance

(3) Speaker counting and clustering based on internal embeddings of E2E SA-ASR



Evaluation on LibriCSS

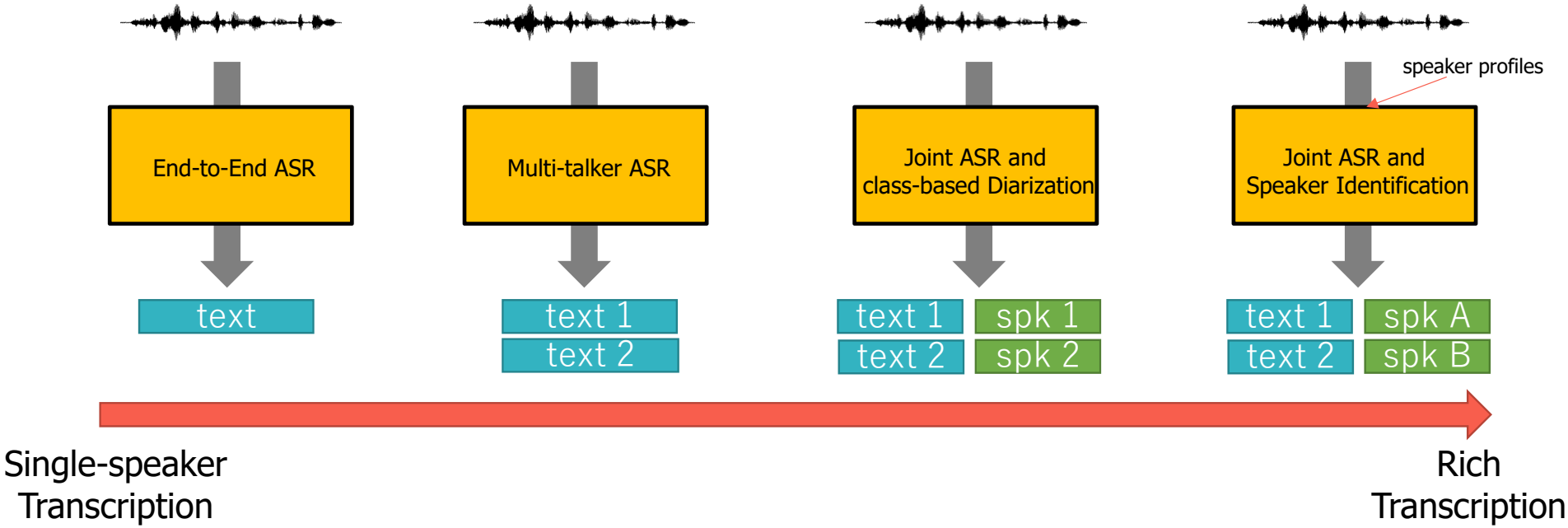


N. Kanda et al., Investigation of end-to-end speaker-attributed ASR for continuous multi-talker recordings, SLT, 2021.
N. Kanda et al., End-to-End Speaker-Attributed ASR with Transformer, ArXiv, 2021.

Short summary of ASR + x

Review the recent progress from the perspective of the extension of the E2E ASR.

- Multi-talker ASR
  - Two output-branch model
  - Serialized Output Training (SOT) for any number of speakers
- Joint ASR and class-based diarization
- Joint ASR and speaker identification (+ speaker clustering)

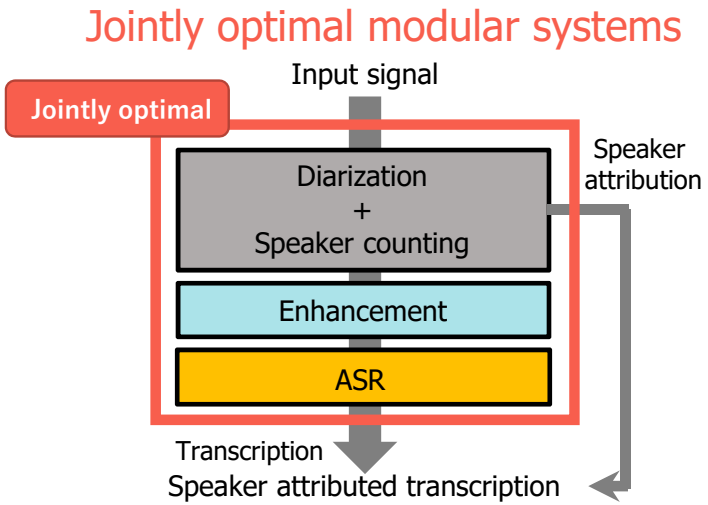


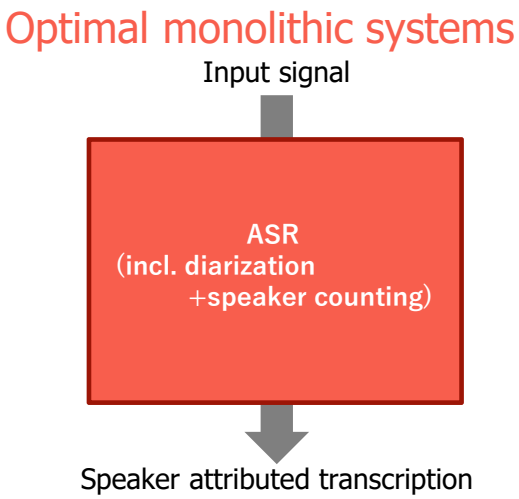
4. Discussion and summary

4.1. Strengths of each approach
4.2. Current challenges and possible future directions
4.3. Take home message + Starting points.

Strength of each approach

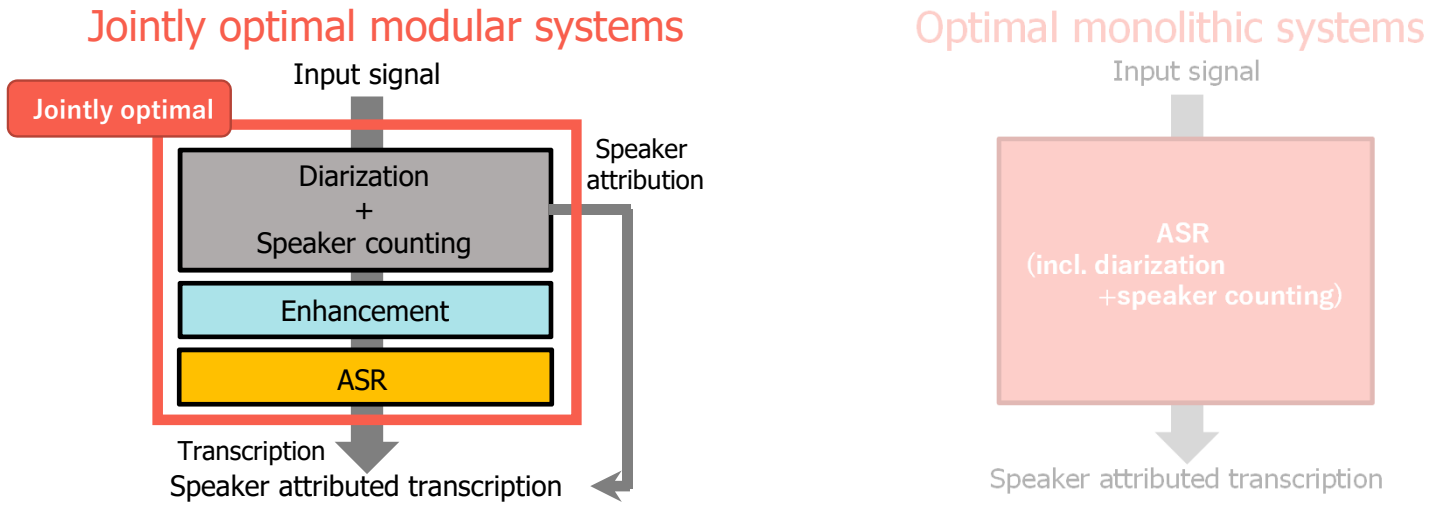
The enhancement- and diarization-originated approaches tend to result in jointly optimal modular systems, while the ASR-originated approach tend to result in optimal monolithic systems.





Strength of each approach

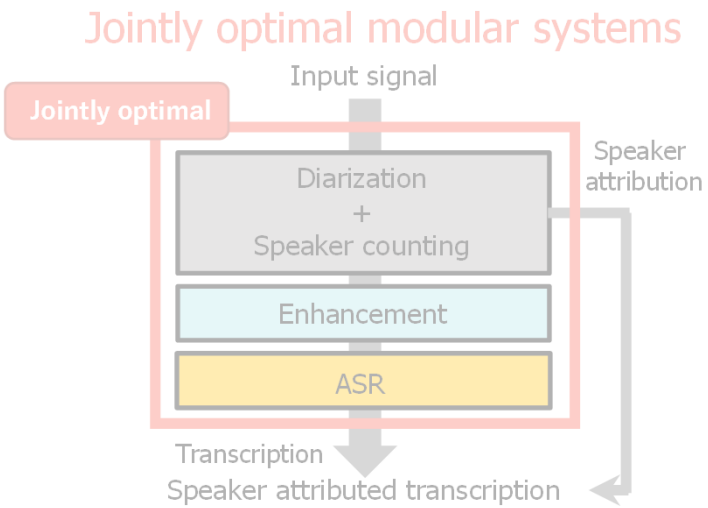
The enhancement- and diarization-originated approaches tend to result in jointly optimal modular systems, while the ASR-originated approach tend to result in optimal monolithic systems.

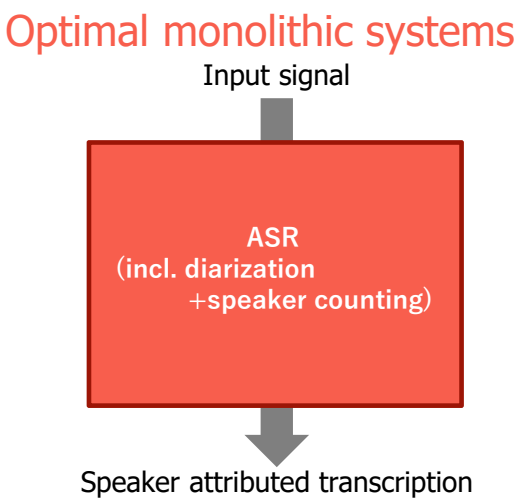


- ☺ Interpretable. → Easier to debug and to find a source of performance degradation.
- ☺ Can incorporate well-studied/already-developed modules, e.g., single-speaker streaming ASR.
- ☺ Flexible. Can significantly alter the functionality of the total system by modifying a module. E.g., multi-talker ASR to single-talker ASR. Command recognition to utterance recognition.

Strength of each approach

The enhancement- and diarization-originated approaches tend to result in jointly optimal modular systems, while the ASR-originated approach tend to result in optimal monolithic systems.





- ☺ Can keep the whole system simpler and smaller (i.e., computationally efficient, less parameters).
- ☺ Dedicated system. Can be completely tuned to certain purpose, e.g., distant conversational ASR.
- ☺ Possible to prepare reference labels i.e., transcription, even from real recordings.
  → Possibility to adapt the system to unseen/real test data.

4. Discussion and summary

4.1. Strengths of each approach

4.2. Current challenges and possible future directions

4.3. Take home message + Starting points

Discussion on challenges and fundamental difficulties

- Currently available solution/systems have not achieved close-talking-microphone ASR performance, i.e., ASR performance obtained without effects of overlaps, noise, reverb, Lombard effect, etc.

- Why? Because we still have many issues such as:

  - Unsolved subtasks: Some issues in each subtask are not yet fully solved, e.g.,
    - Diarization (and separation): An arbitrary (potentially large) number of speakers [Horiguchi+,2020]
    - ASR: Long-form audio [Chang+, ICASSP2021].
    - Enhancement: Noise and reverb still degrade conversational ASR performance a lot [Kanda+, 2021].

  - Optimality of the total system: Not yet achieved fully optimal distant conversational ASR system.

  - Data issue: Not yet clear what kind of and how much data is necessary
    E.g., Simply mixing many different corpora can boost conversational ASR performance [Chan+, 2021]

  - Labels and eval metrics: Preparation of "correct" transcription is difficult when the input data is spontaneous speech, including fillers, self-edits etc [Saon+, 2017].
    Whether WER is really appropriate or not, can be also debatable.

S. Horiguchi et al. End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors, arXiv, 2020
X. Chang et al., Hypothesis Stitcher for End-to-End Speaker-attributed ASR on Long-form Multi-talker Recordings, ICASSP, 2021
N. Kanda et al. Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone, arXiv, 2021
W. Chan et al. SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network, arXiv, 2021,
G. Saon et al., English Conversational Telephone Speech Recognition by Humans and Machines, Interspeech,2017

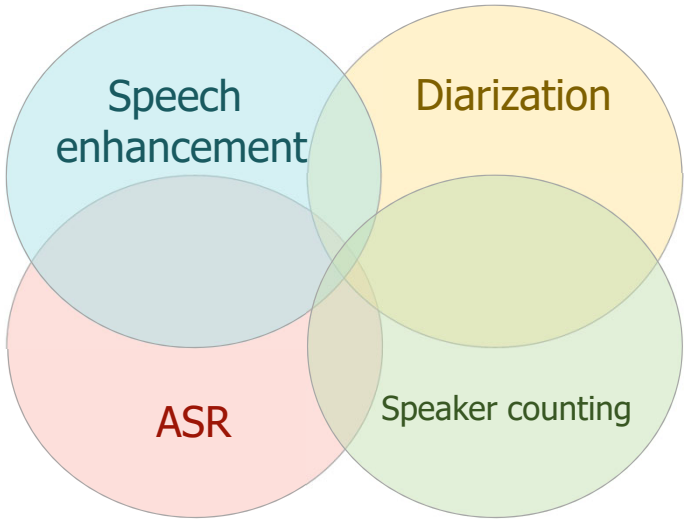
4. Discussion and summary

4.1. Strengths of each approach
4.2. Current challenges and possible future directions
4.3. Take home message + Starting points

Take-home message

- The distant conversational ASR and analysis is a core of next-generation speech interfaces.
- Many academic institutes and industry are actively working on this task.
- Lots of interesting research problems, increasingly active area, many OSS and data.

Let's work together on this topic and solve it!



Starting points (useful software, database, etc) (1/2)

- Open source software
  - Enhancement
    - Asteroid (Neural source separation) [1]
  - Diarization
    - pyannote.audio (Neural building blocks for speaker diarization) [2]
  - ASR
    - Kaldi (DNN-hybrid ASR) [3]
      - Upgraded CHiME6 baseline system containing diarization, multichannel enh.
      - LibriCSS baseline system containing separation, diarization, multichannel enh.
    - ESPnet (E2E ASR, neural speech enhancement, neural diarization) [4]
    - SpeechBrain (E2E ASR, Speaker diarization/recognition/identification, neural speech enhancement, beamforming) [5]
    - RETURNN (E2E ASR) [6]

[1] https://github.com/asteroid-team/asteroid [2] https://github.com/pyannote/pyannote-audio
[3] https://github.com/kaldi-asr/kaldi [4] https://github.com/espnet/espnet
[5] https://github.com/speechbrain/speechbrain [6] https://github.com/rwth-i6/returnn

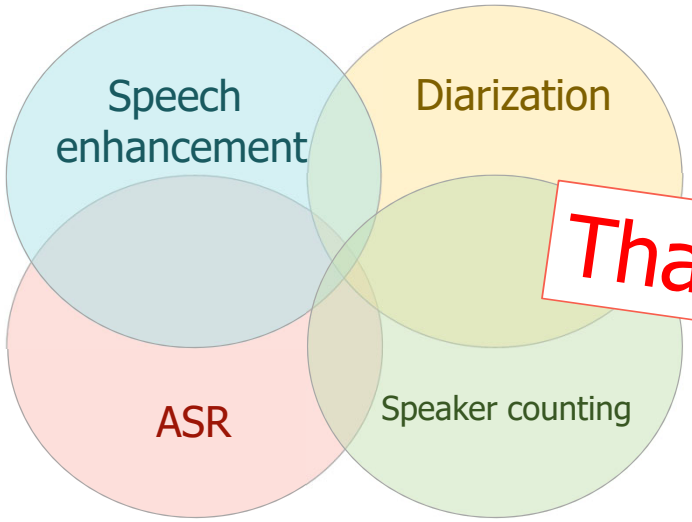
Starting points (useful software, database, etc) (2/2)

- Corpora containing real conversational speech data
  - CALLHOME (telephone speech, single-microphone setup)
  - AMI (3 to 5 (mostly, 4) speaker meetings, noise & reverb, multi-microphone setup)
  - CHiME-5 data (4-speaker very spontaneous meetings, noise & reverb, multiple microphone-array setup)
  - AISHELL-4 (4 to 8 speaker meetings, noise & reverb, multi-microphone setup)
  - CHiME 7 (Coming soon)

Take-home message

- The distant conversational ASR and analysis is a core of next-generation speech interfaces.
- Many academic institutes and industry are actively working on this task.
- Lots of interesting research problems, increasingly active area, many OSS and data.

Let's work together on this topic and solve it!



Thank you very much!

Acknowledgement

Desh Raj, Ashish Arora, Aswin Subramanian, Sanjeev Khudanpur, Wangyou Zhang, Xuankai Chang, Marc Delcroix, Tomohiro Nakatani, Reinhold Haeb-Umbach