Target-speaker Methods for Speech Recognition

Desh Raj

CLSP Seminar
March 27, 2023

Motivation

Single-user applications

Smart Assistants

Language Learning

Customer Service

Voice-based Search

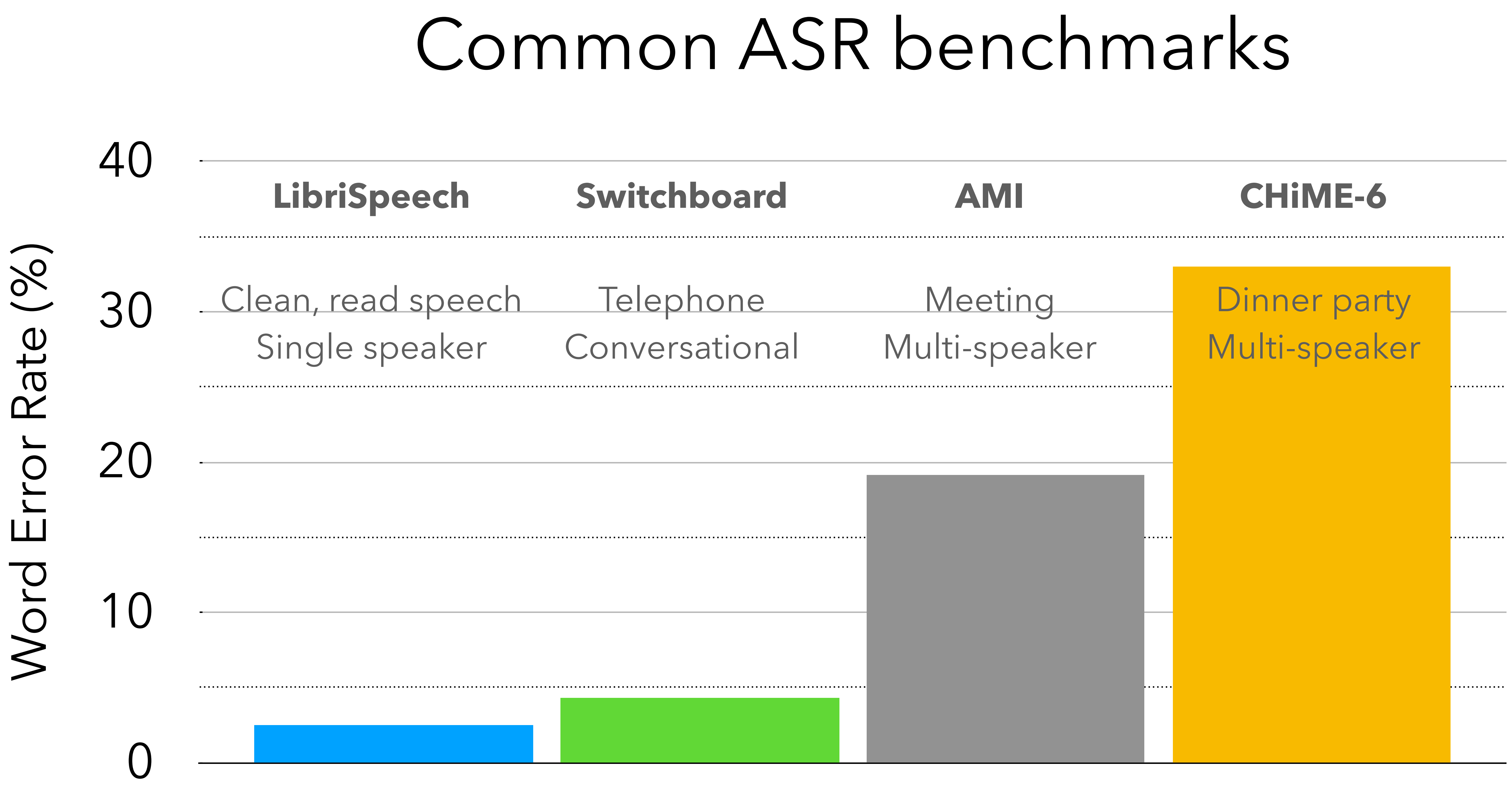
Multi-user applications

Meeting summaries

Collaborative Learning

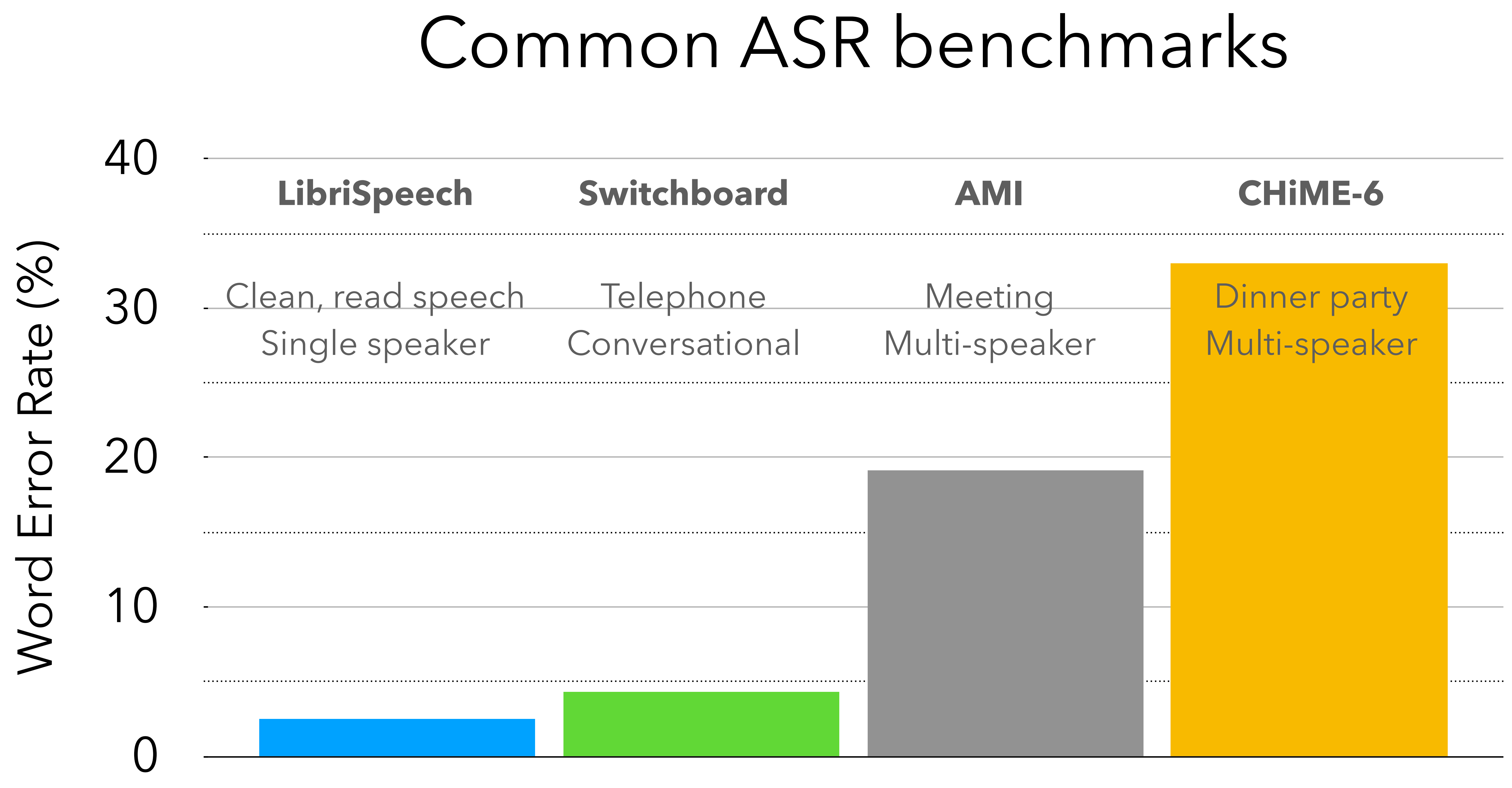
Cocktail-party Problem

Motivation



Common ASR benchmarks
Word Error Rate (%)
40
30
20
10
0
LibriSpeech
Switchboard
AMI
CHiME-6
Clean, read speech
Single speaker
Telephone
Conversational
Meeting
Multi-speaker
Dinner party
Multi-speaker

Motivation



What changed?

- Conversational speech
- Far-field audio: noise and reverberation
- Overlapping speakers

Biggest challenge for multi-talker ASR

Motivation

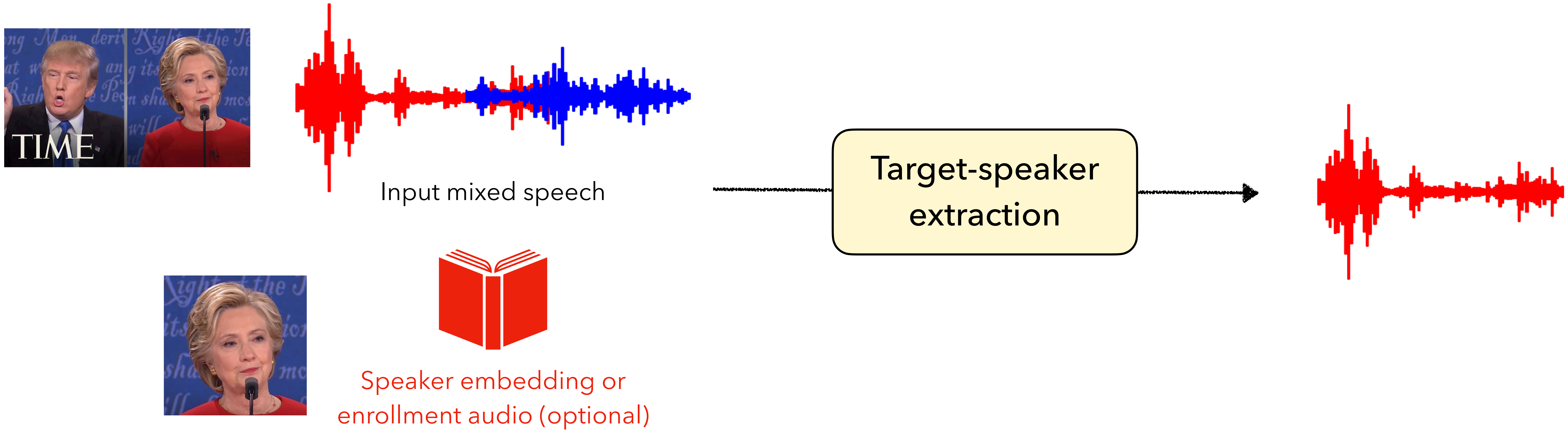
Overview

- What is target-speaker ASR?
- **Meeting transcription:** Offline, multi-channel TS-ASR with GSS
- **Voice-based assistant:** Real-time, wake-word based TS-ASR
- Bonus: TS-ASR + self-supervised models

What is target-speaker ASR?

Preliminary: Target speaker extraction

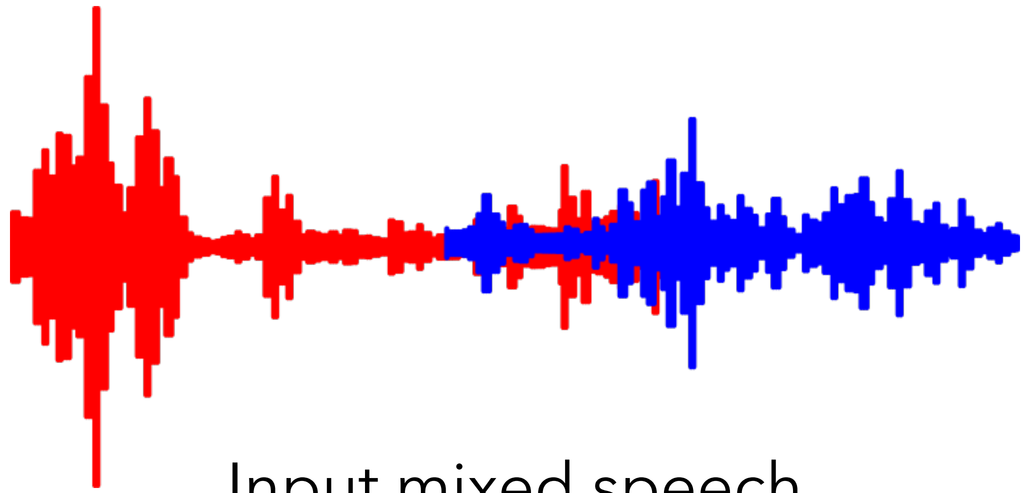
- Given an audio containing mixed speech, *extract* the speech of a **target speaker**
- Auxiliary information: enrollment audio or speaker embedding



What is target-speaker ASR?

Target speaker extraction + ASR

- Given an audio containing mixed speech, *transcribe* the speech of a **target speaker**
- Auxiliary information: enrollment audio or speaker embedding



Input mixed speech

Target-speaker ASR

We added a million new jobs…

Speaker embedding or enrollment audio (optional)

What is target-speaker ASR?

Methods

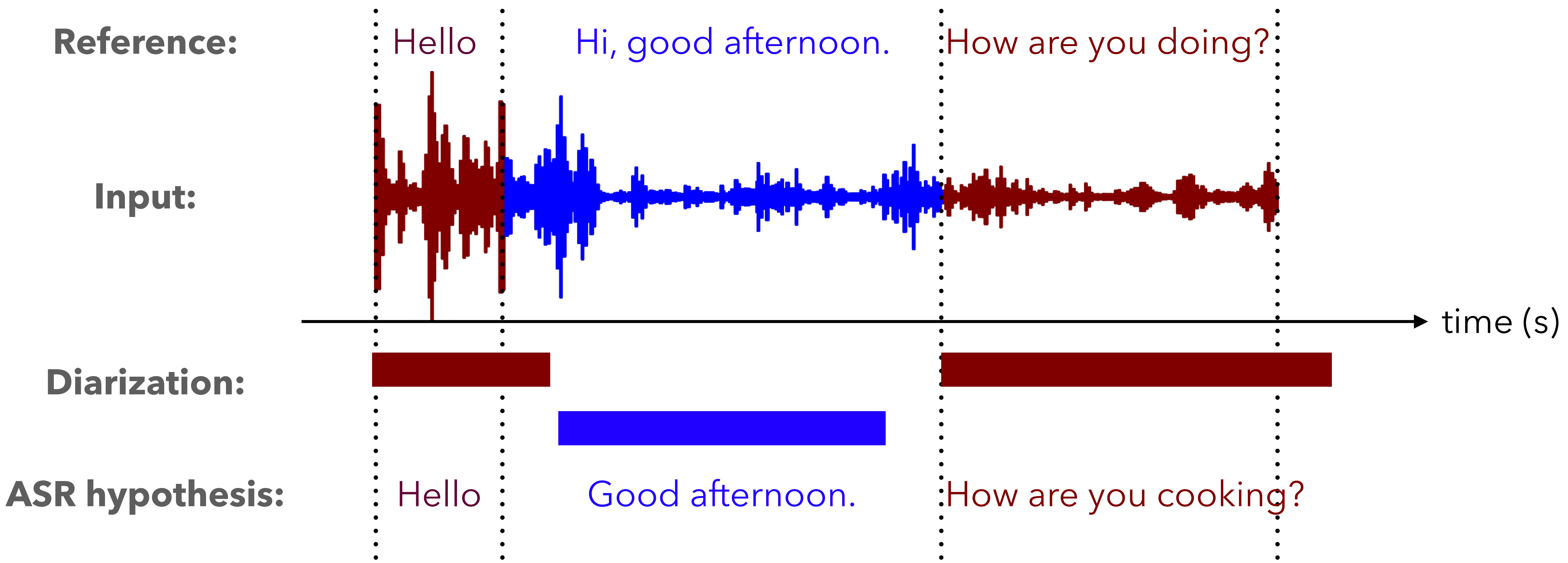
- Methods used for Target-speaker ASR depend on the application scenario.

Scenario	Meeting Transcription	Voice-based Assistant
Recording device	Multi-channel microphone array	Single microphone
Speakers	Multiple primary	1 primary + background
Wake-word	None	“Hey Siri”, “Alexa”, etc.
Real-time?	Optional	Required

Scenario 1: Meeting Transcription

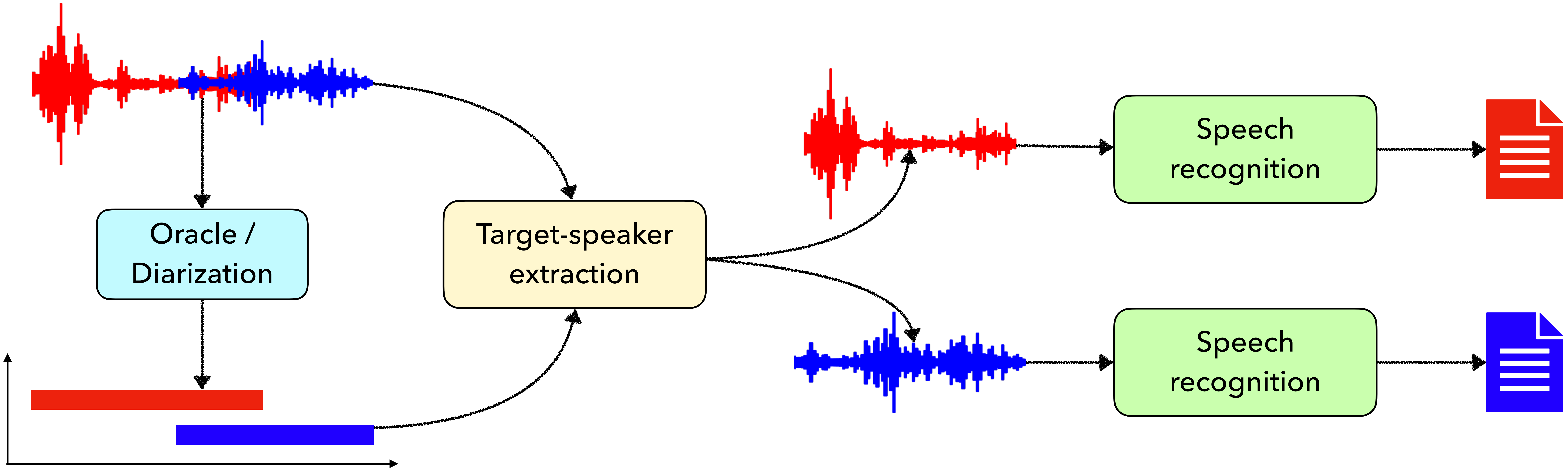
Meeting Transcription

Problem Statement



Meeting Transcription

Approach using target speaker methods

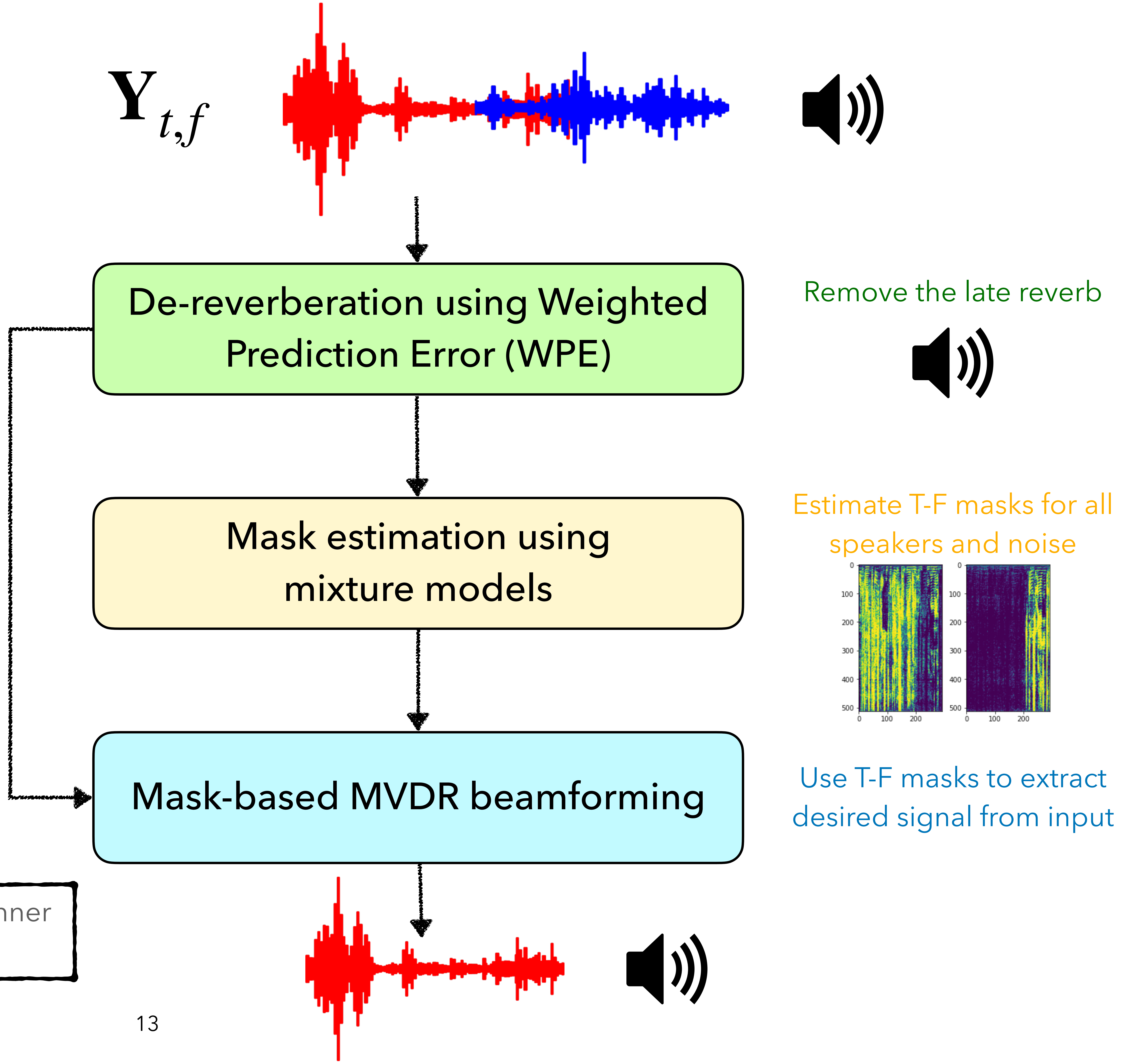


Guided source separation

Consists of 3 main steps

https://github.com/fgnt/pb_chime5

$$\mathbf{Y}_{t,f} = \underbrace{\sum_k \mathbf{X}^{\text{early}}_{t,f,k}}_{\text{Sum of speaker signals}} + \overbrace{\sum_k \mathbf{X}^{\text{tail}}_{t,f,k}}^{\text{Sum of reverb tails}} + \underbrace{\mathbf{N}_{t,f}}_{\text{Noise}}$$



Boeddeker, Christoph et al. "Front-end processing for the CHiME-5 dinner party scenario." *CHiME Workshop, 2018* .

Guided source separation

Limitations with original implementation

- Several iterative parts, e.g., mask estimation using complex angular GMMs.
- All implementation on CPU (with NumPy).
- Example: Applying GSS on CHiME-6 *dev* set takes ~20h with 80 jobs!

Meeting Transcription

Approach using target speaker methods

GPU-accelerated Guided Source Separation for Meeting Transcription

Desh Raj[1], *Daniel Povey*[2], *Sanjeev Khudanpur*[1,3]

[1]CLSP & [3]HLTCOE, Johns Hopkins University, Baltimore, USA; [2]Xiaomi Corp., Beijing, China

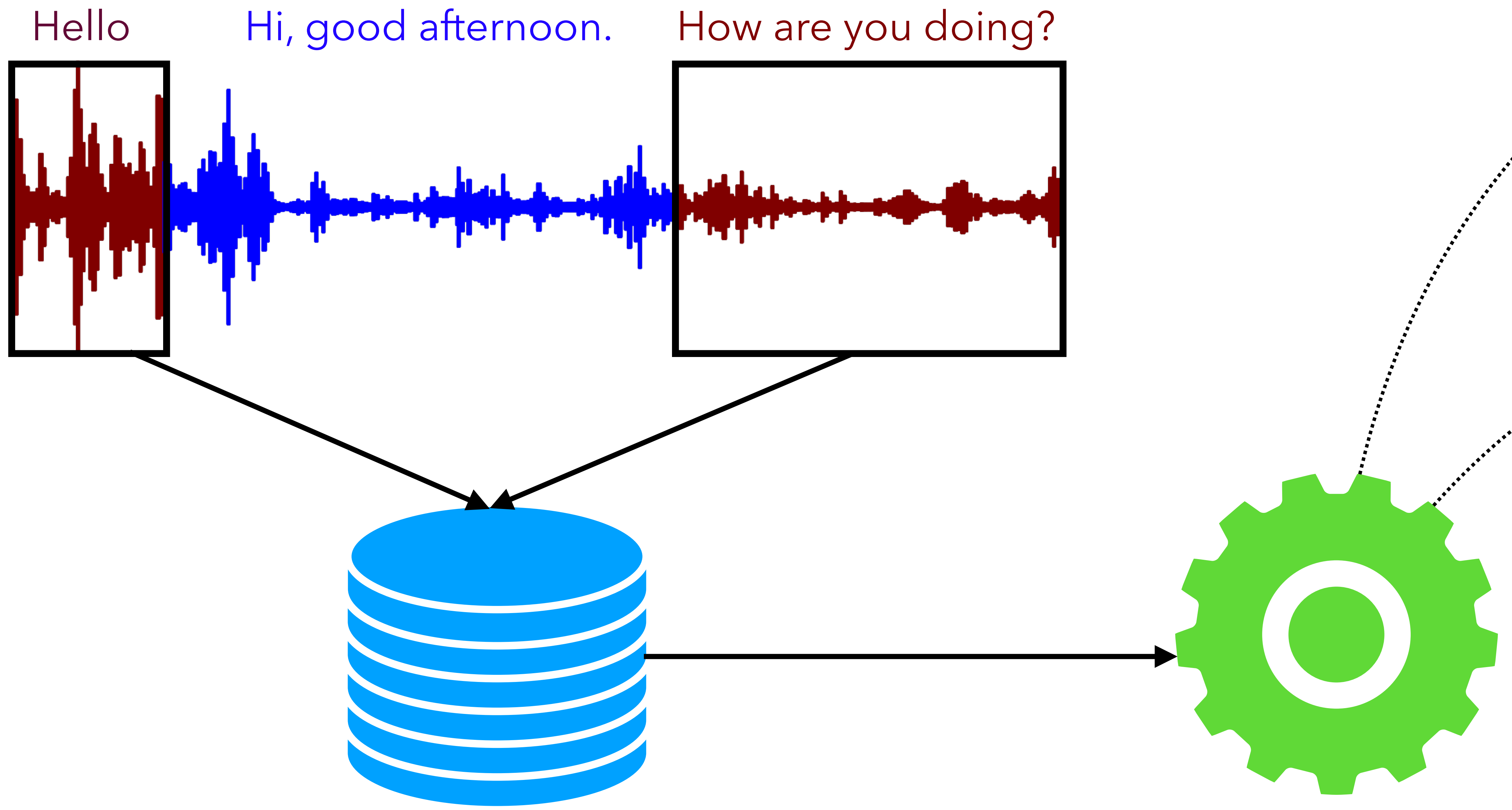
draj@cs.jhu.edu, dpovey@gmail.com, khudanpur@jhu.edu

Under review at InterSpeech 2023

Guided source separation

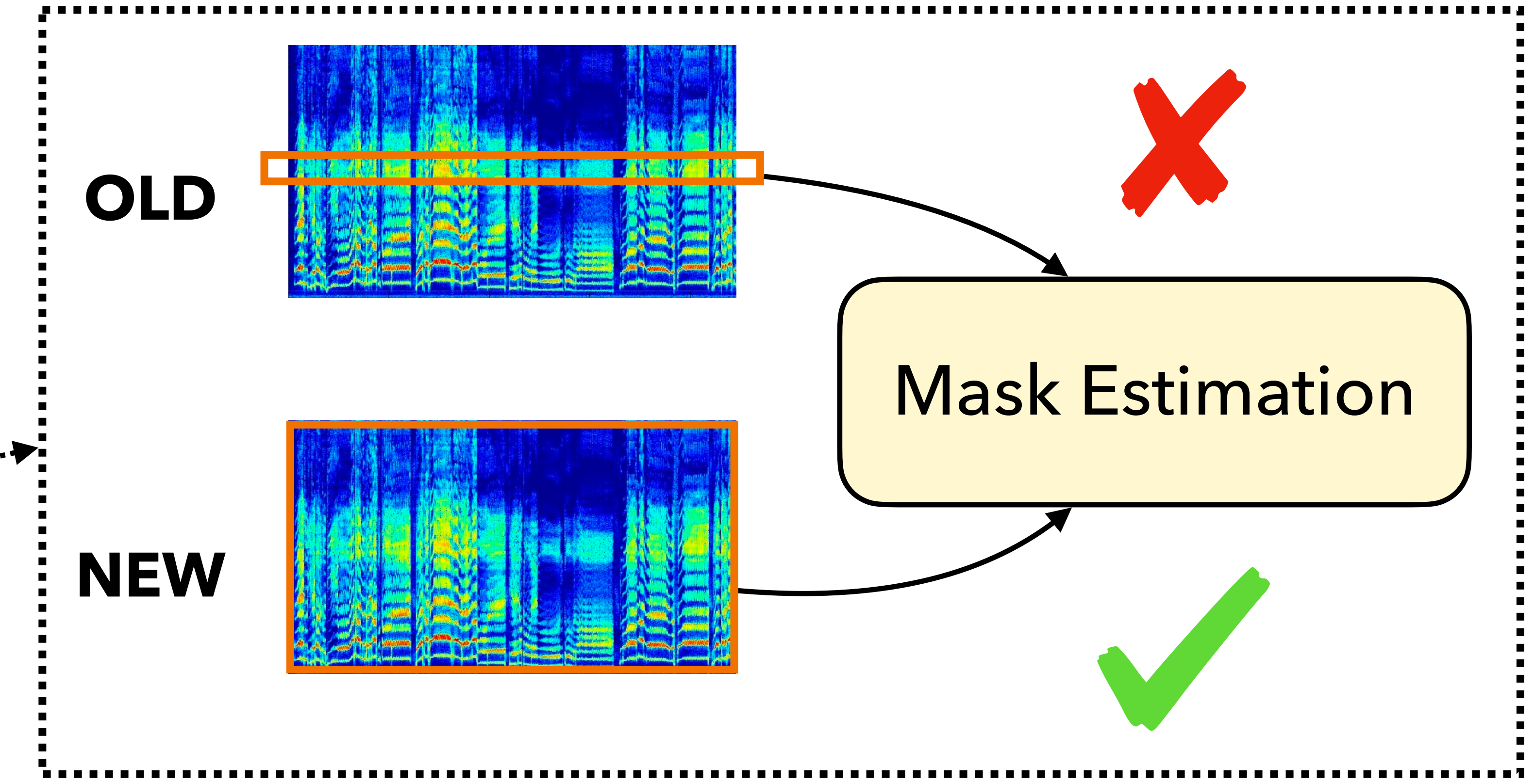
GPU-acceleration + engineering tricks

https://github.com/desh2608/gss



1. CPU-based data-loader performs smart batching of segments

2. STFT computation, WPE, mask estimation on GPU using CuPy



3. Batched processing of STFT frequency bins

```
covariance = D * cp.einsum(
    "...dn,...Dn,...n->...dD",
    y,
    y.conj(),
    (saliency / quadratic_form),
    optimize=einsum_path,
)
```

Cache optimized path on first iteration.

Use same path on subsequent iterations.

4. einsum path caching

Guided source separation

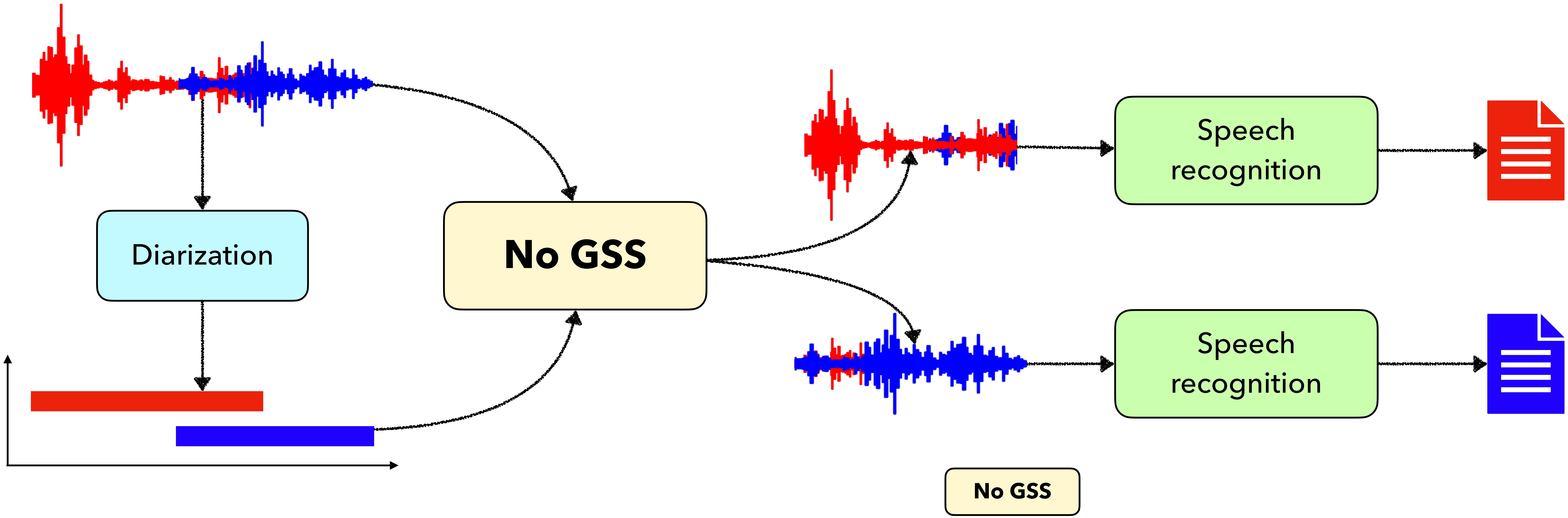
Speed-up

- Comparison on CHiME-6 dev set
- Old GSS: Takes **19.3** hours using 80 jobs
- New GSS: Takes **1.3** hours using 4 GPUs

- Part of the official baseline in CHiME-7 DASR challenge: https://www.chimechallenge.org/current/task1/index

Meeting Transcription

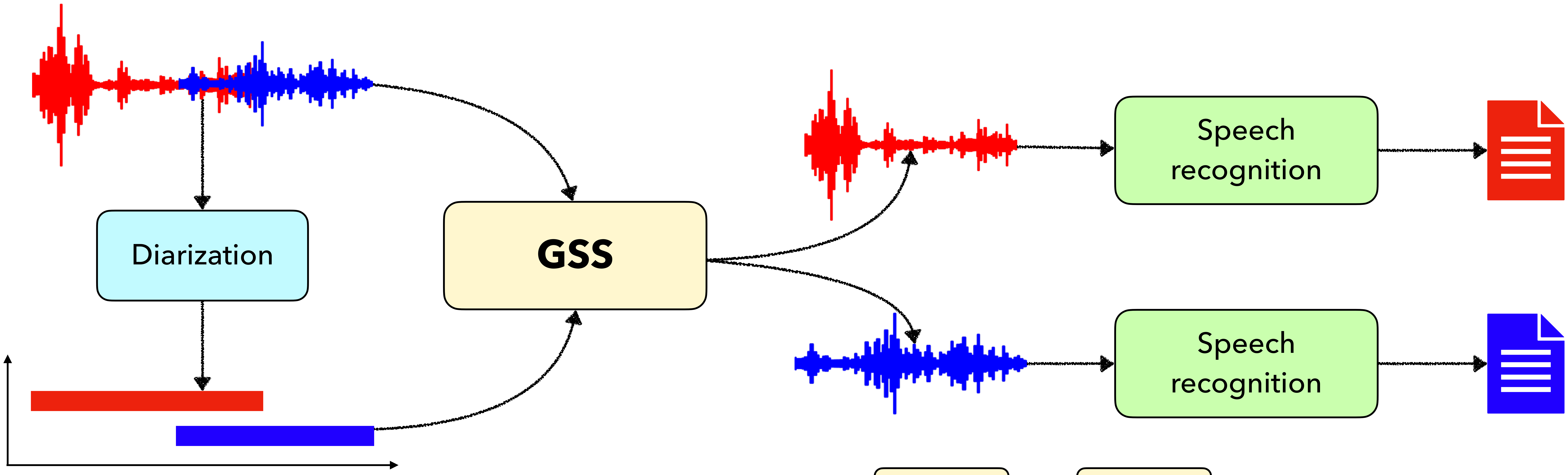
Results on AMI without GSS



Diarizer	DER (%)	WER (%)
Oracle	0.0	32.1
Clustering	23.7	38.5

Meeting Transcription

Results on AMI with GSS

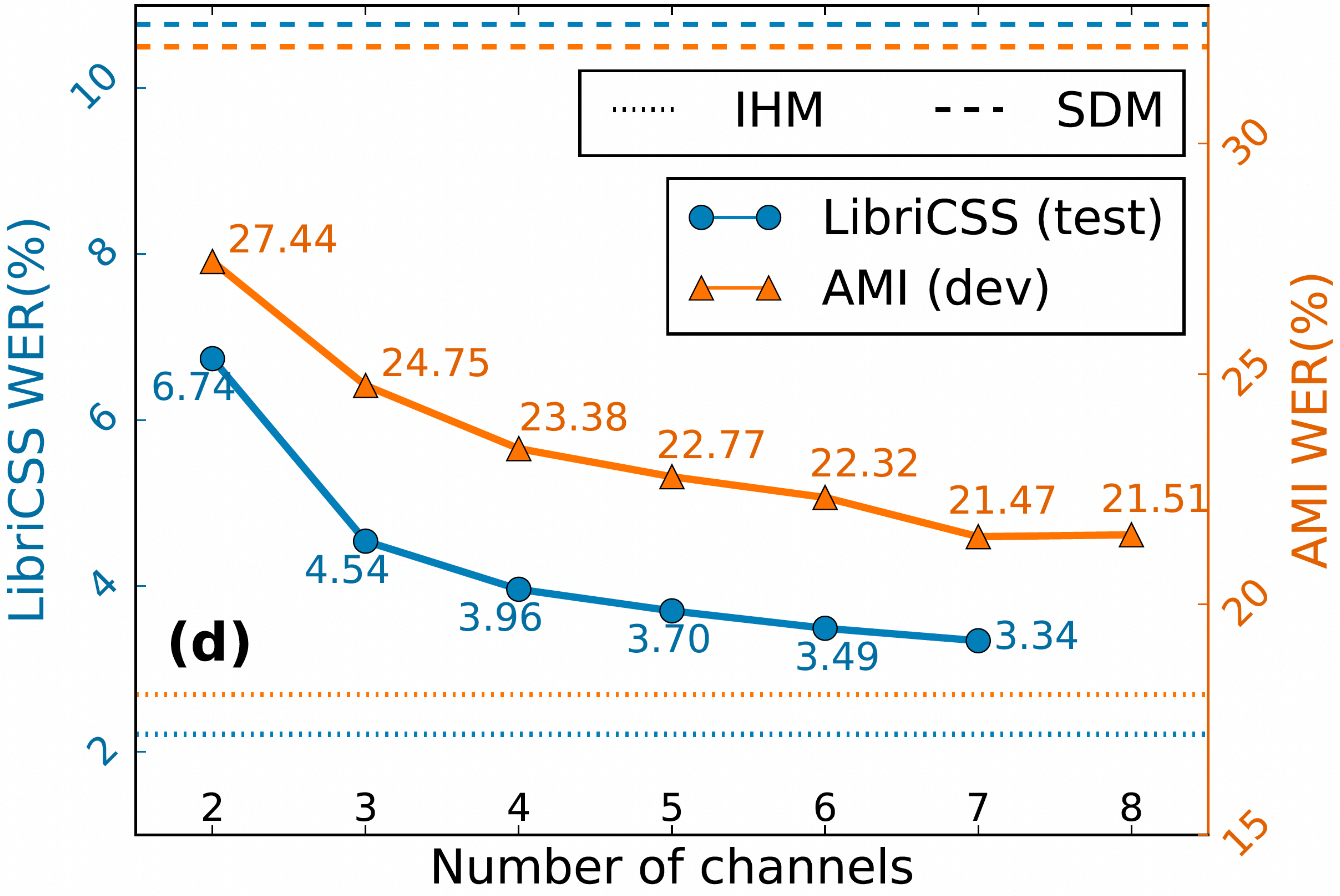


		No GSS	GSS
Diarizer	**DER (%)**	**WER (%)**	**WER (%)**
Oracle	0.0	32.1	22.8
Clustering	23.7	38.5	31.0

29.0%
19.5%

Guided source separation

Effect of number of channels



LibriCSS example

No GSS

REFERENCE:

Paul declares that the false apostles were called or sent neither by men nor by man

2 channels

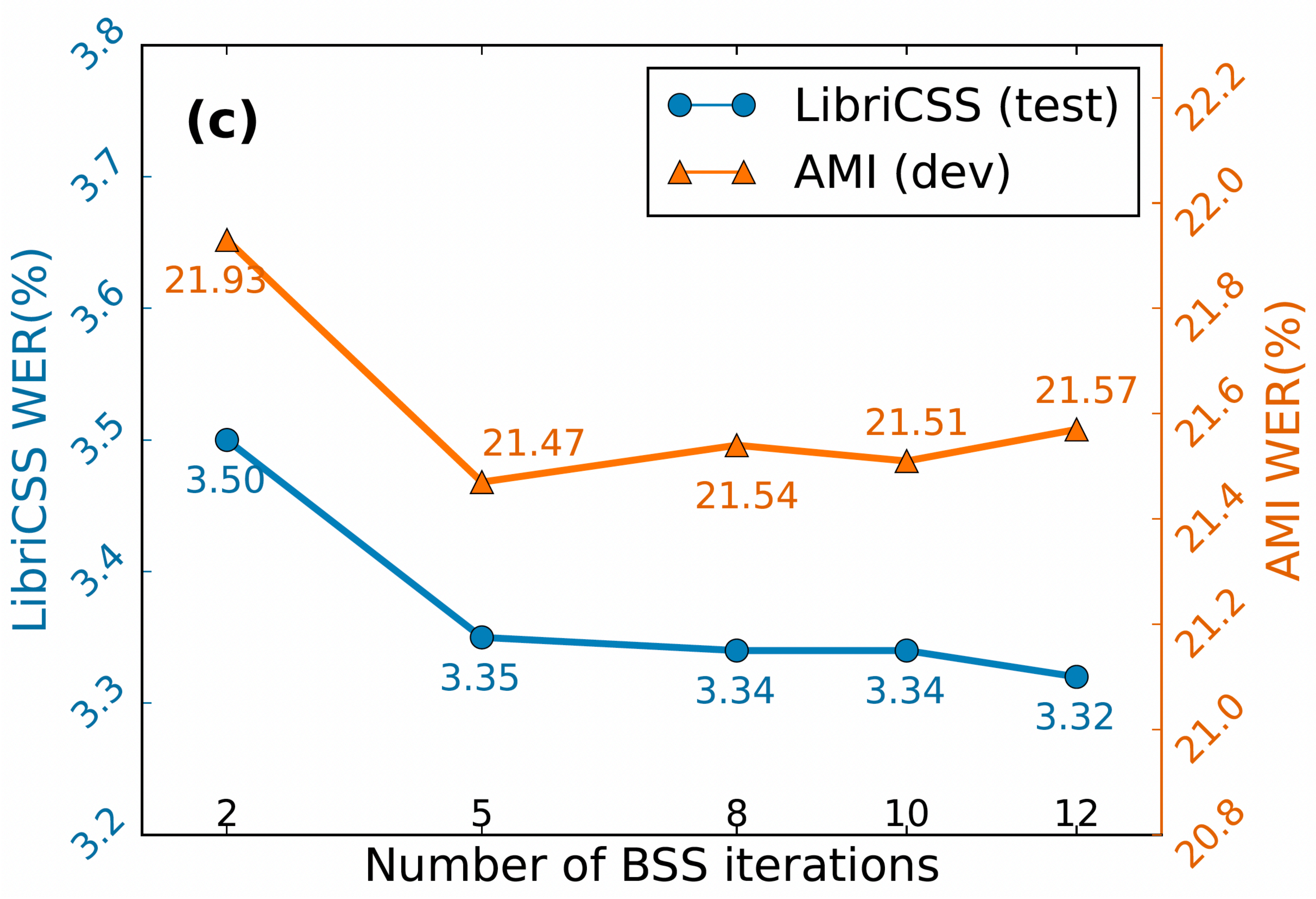
All declares of the false apostles [were] recalled or sent neither by men [nor by man]

7 channels

All declares that the false apostles were called or sent neither by men nor by man

Guided source separation

Effect of number of iterations for mask estimation



Scenario 2: Voice-based Assistant

Recall from earlier...

Very different from meeting transcription

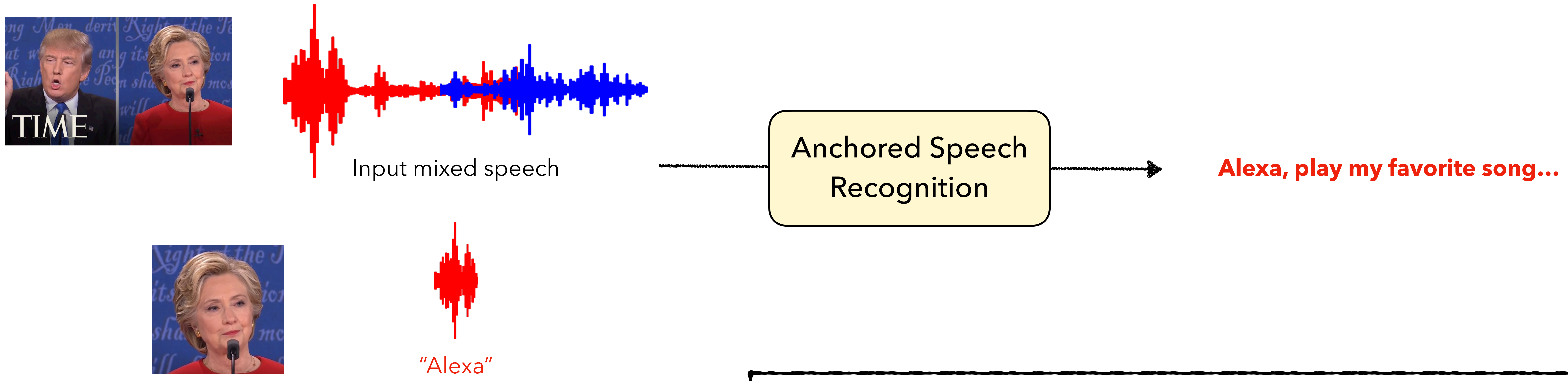
- Methods used for Target-speaker ASR depend on the application scenario.

Scenario	Meeting Transcription	Voice-based Assistant
Recording device	Multi-channel microphone array	Single microphone
Speakers	Multiple primary	1 primary + background
Wake-word	None	"Hey Siri", "Alexa", etc.
Real-time?	Optional	Required

From GSS to anchored speech recognition

“Anchor” = wake-word

- **“Alexa,** play my favorite song.”
- Auxiliary information: “Alexa”



Wang, Yiming et al. “End-to-end Anchored Speech Recognition.” IEEE ICASSP, 2019.

Voice-based Assistant

Approach using target speaker methods

ANCHORED SPEECH RECOGNITION WITH NEURAL TRANSDUCERS

Desh Raj[*1], *Junteng Jia*[2], *Jay Mahadeokar*[2], *Chunyang Wu*[2], *Niko Moritz*[2], *Xiaohui Zhang*[2], *Ozlem Kalinli*[2]

[1]Center for Language and Speech Processing, Johns Hopkins University, USA, [2]Meta AI, USA

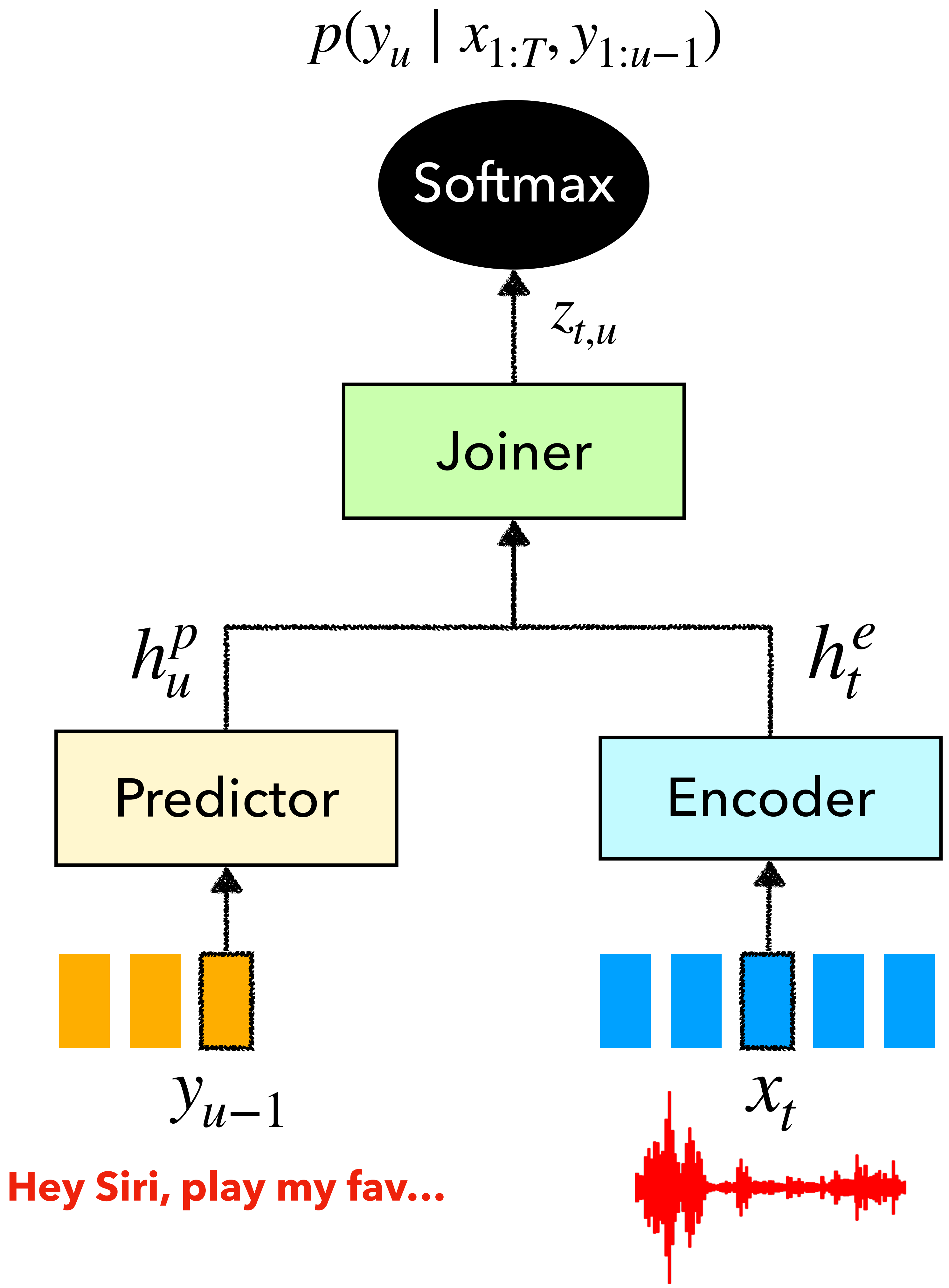
To appear at **IEEE ICASSP 2023**

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Meta

Voice-based Assistant

The basic ASR system: Neural transducer

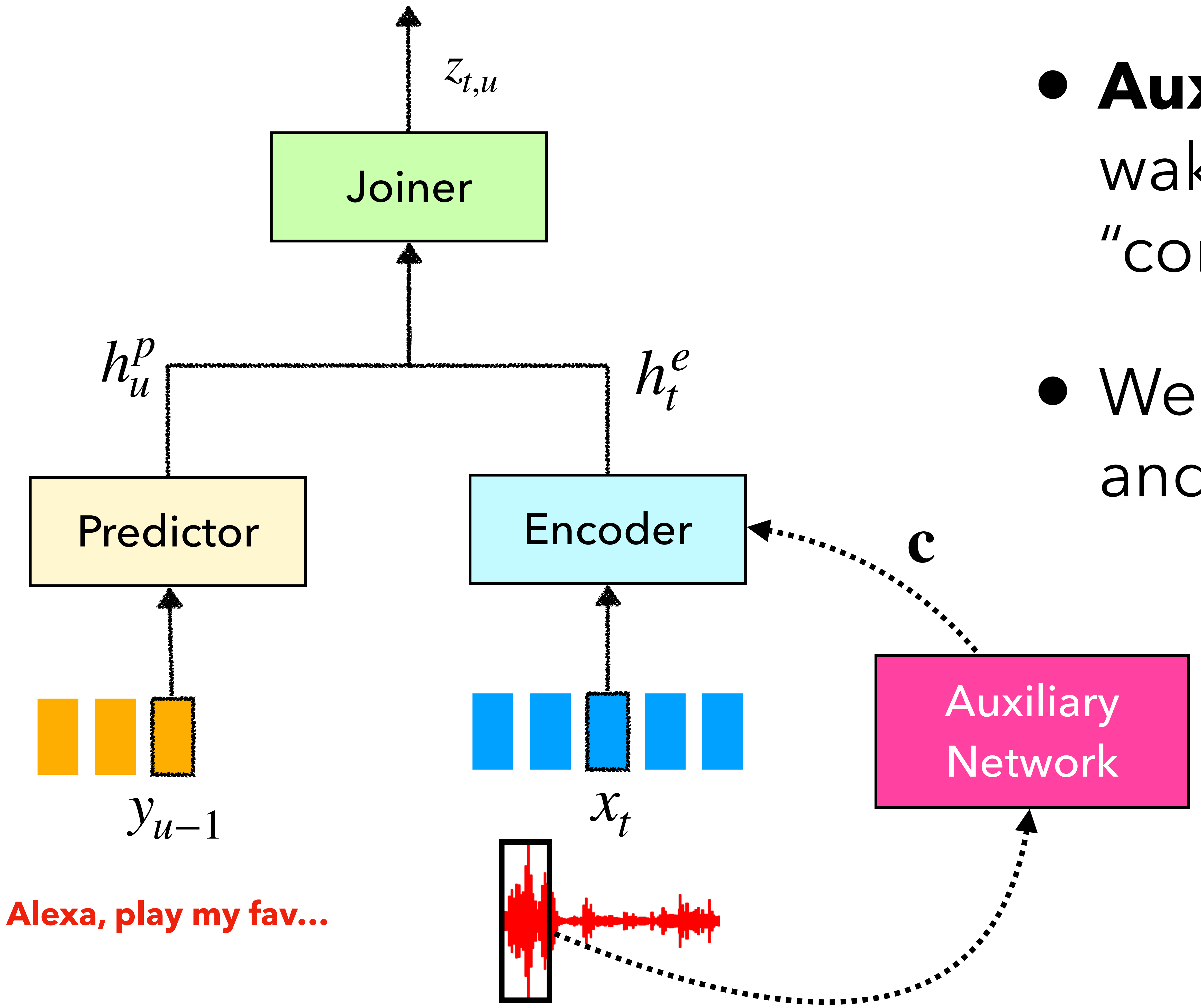


- **Encoder** converts input *audio* to high-dimensional representation
- **Predictor** is an autoregressive model that encodes input *text*
- **Joiner** combines audio and text representations to predict next token

Voice-based Assistant

1. Biasing the encoder with context

Encoder can use context embedding to suppress background speech.



*Only 1% of the model size

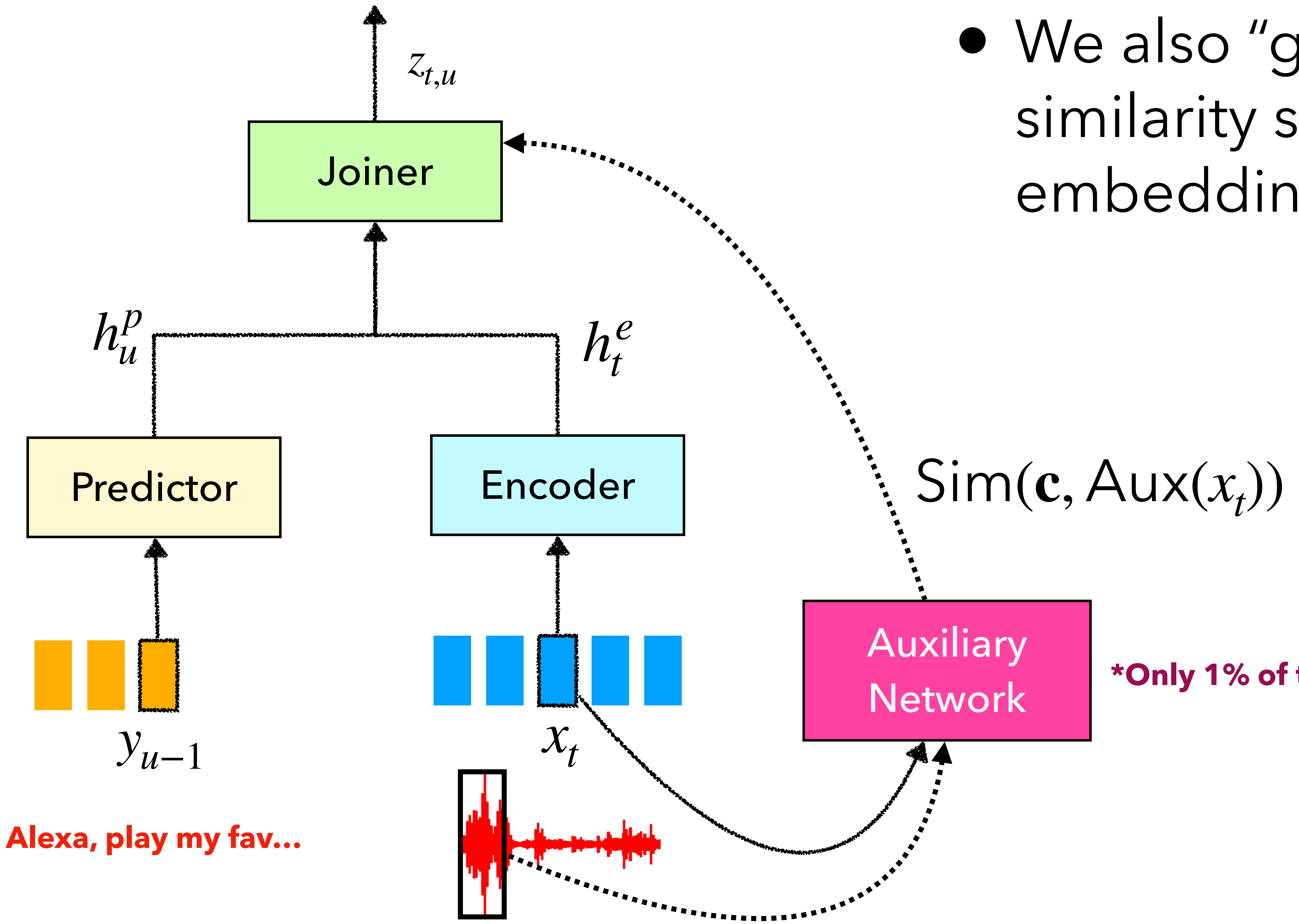
- **Auxiliary network** encodes the wake-word segment into a "context" embedding
- We concat this to input features and project to original dimension

Voice-based Assistant

2. Gating the joiner

Boost the logits for blank tokens when speaker is different from wake-word segment.

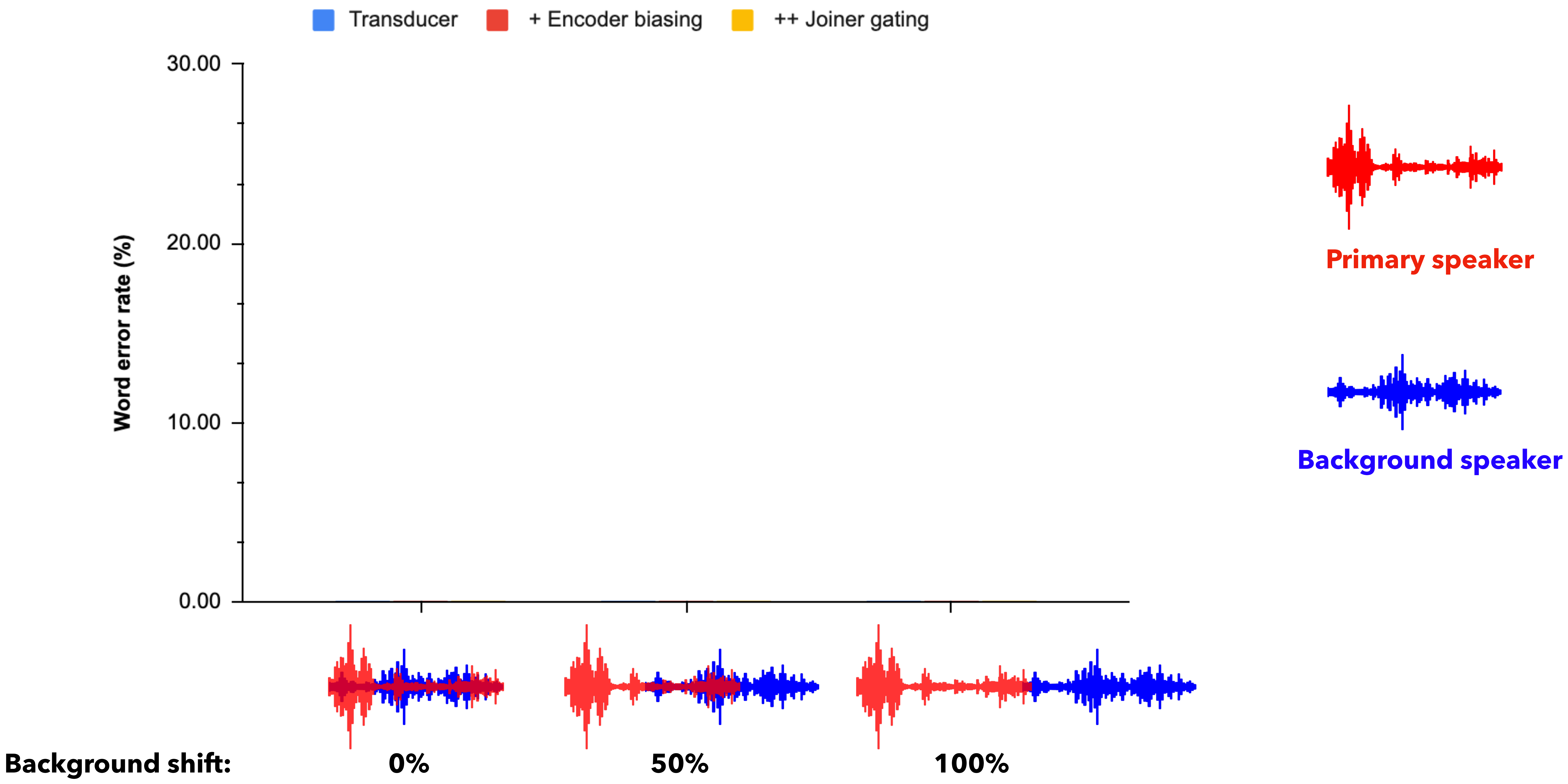
- We also "gate" the logits with the similarity score between context embedding and segments.



*Only 1% of the model size

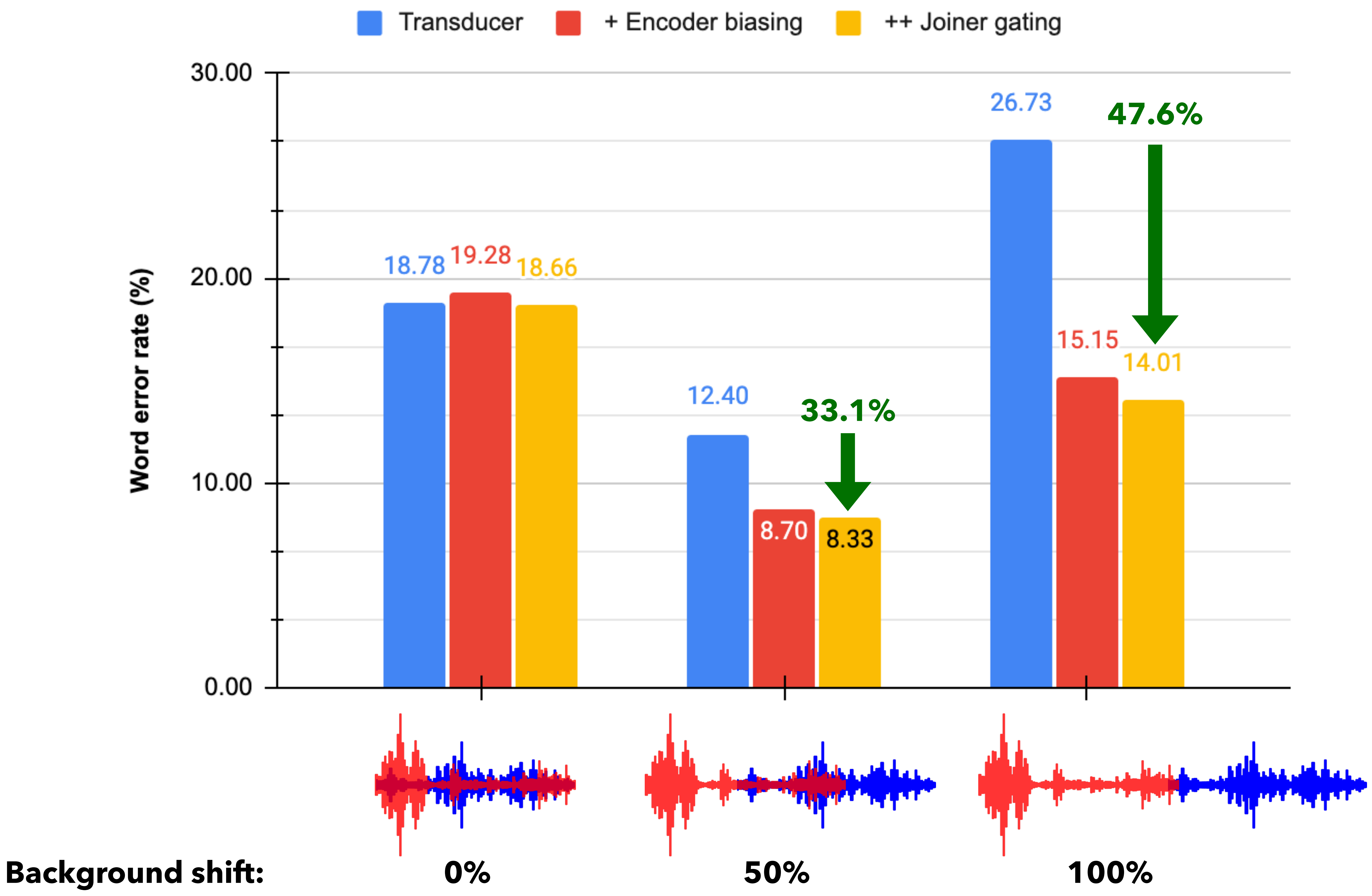
Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)



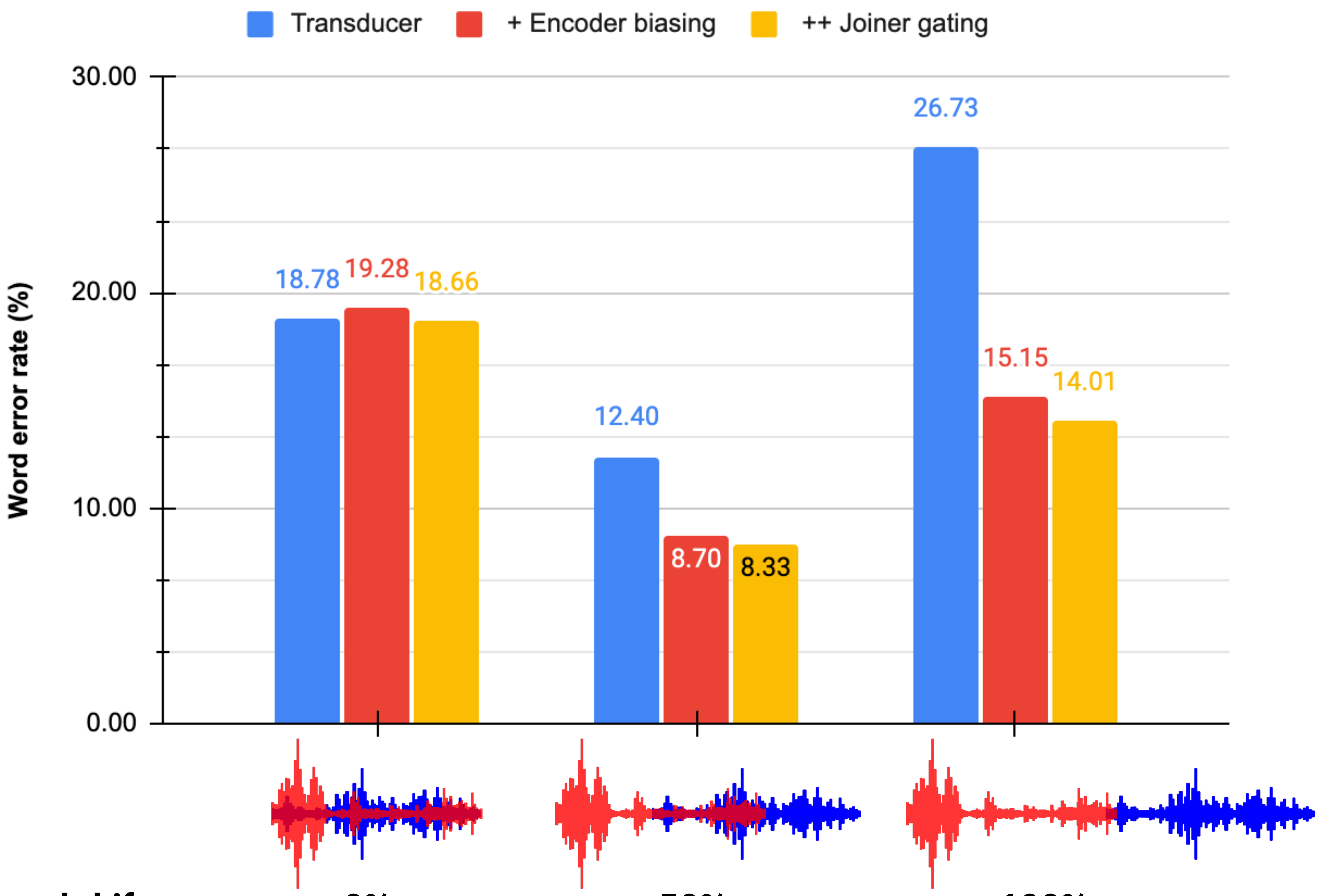
Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)



Effect of TS-ASR

WER on LibriSpeech mixtures (average over SNRs 1~20 dB)



Background shift: 0% 50% 100%

REFERENCE:

Then they seemed to spring from every part of the country

TRANSDUCER:

Then they seemed to spring from every part of the country [hastened may be very much modified to dogmas]

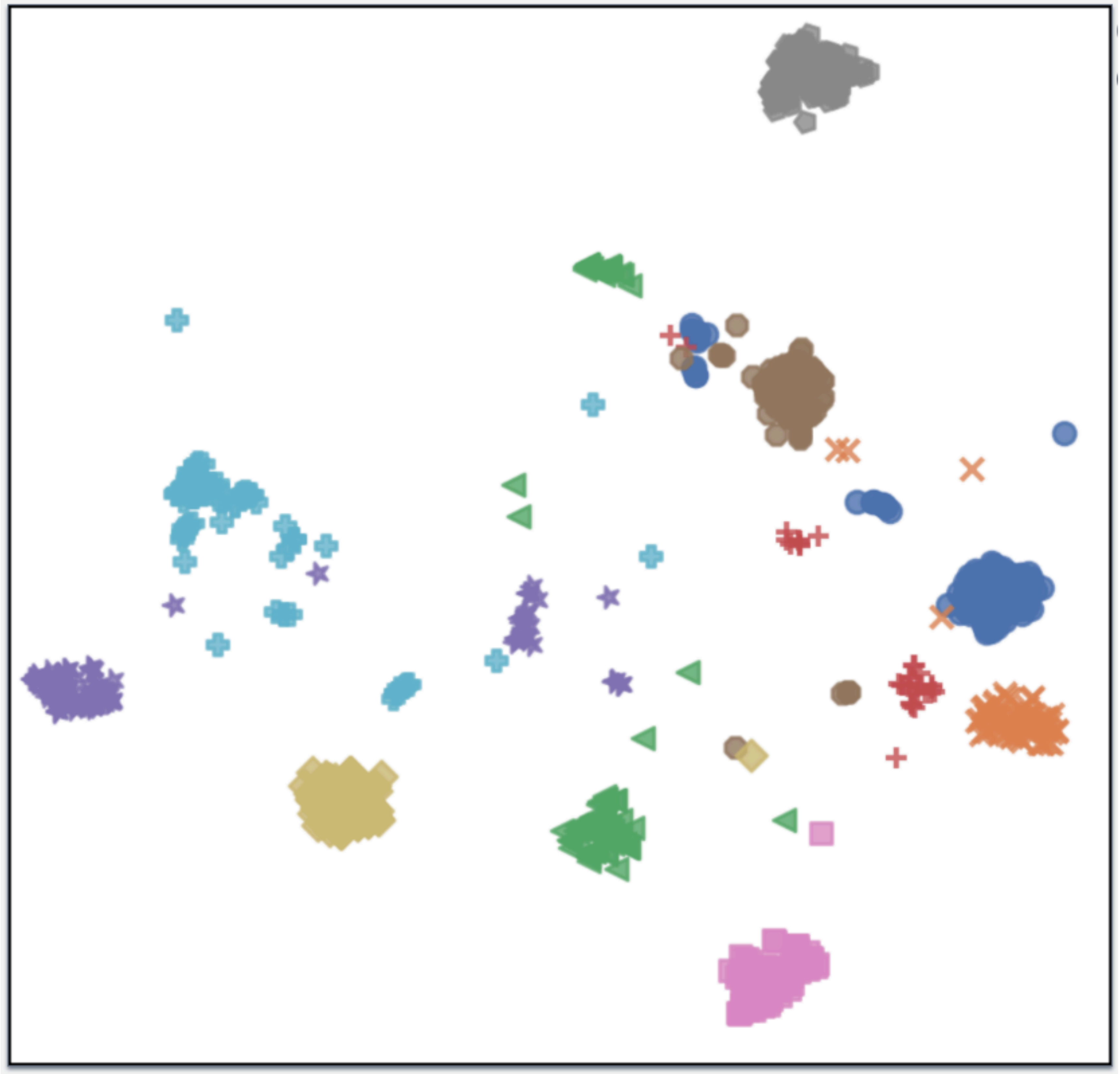
TS–ASR:

Then they seemed to spring from every part of the country

Voice-based Assistant

Disentangling style from content in the context embedding

- How do we ensure that **c** only contains the speaker characteristics, and not lexical content?
- Auxiliary training objectives:
  - Feature reconstruction
  - VIC regularization (self-supervised)
- See the paper for details!



What about self-supervised models?

SSL + TS-ASR

using speaker embeddings

ADAPTING SELF-SUPERVISED MODELS TO MULTI-TALKER SPEECH RECOGNITION USING SPEAKER EMBEDDINGS

Zili Huang, Desh Raj, Paola García, Sanjeev Khudanpur

Center for Language and Speech Processing and Human Language Technology Center of Excellence,
Johns Hopkins University, Baltimore, USA

To appear at **IEEE ICASSP 2023**

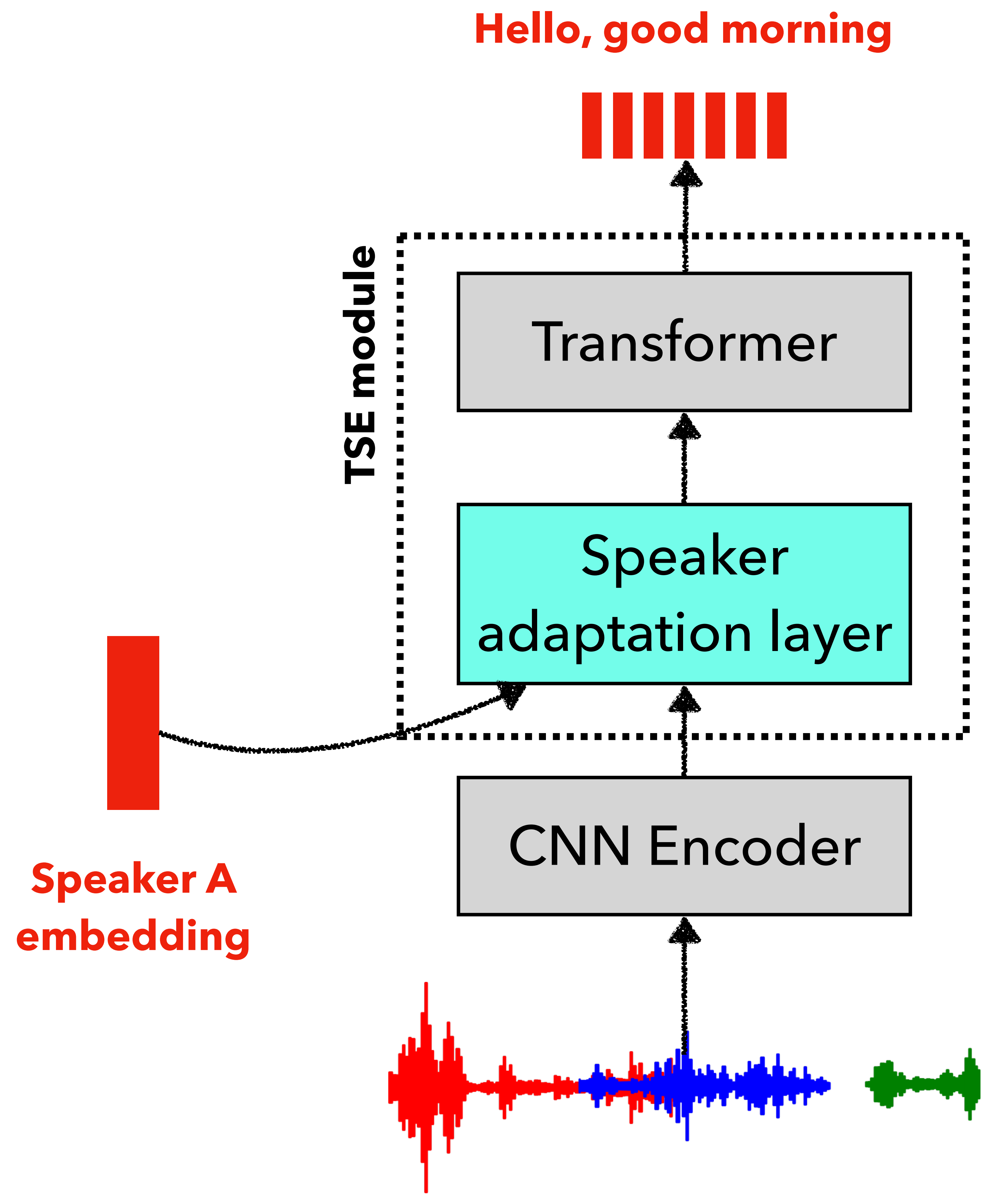
Zili **Huang**

CENTER FOR LANGUAGE AND SPEECH PROCESSING

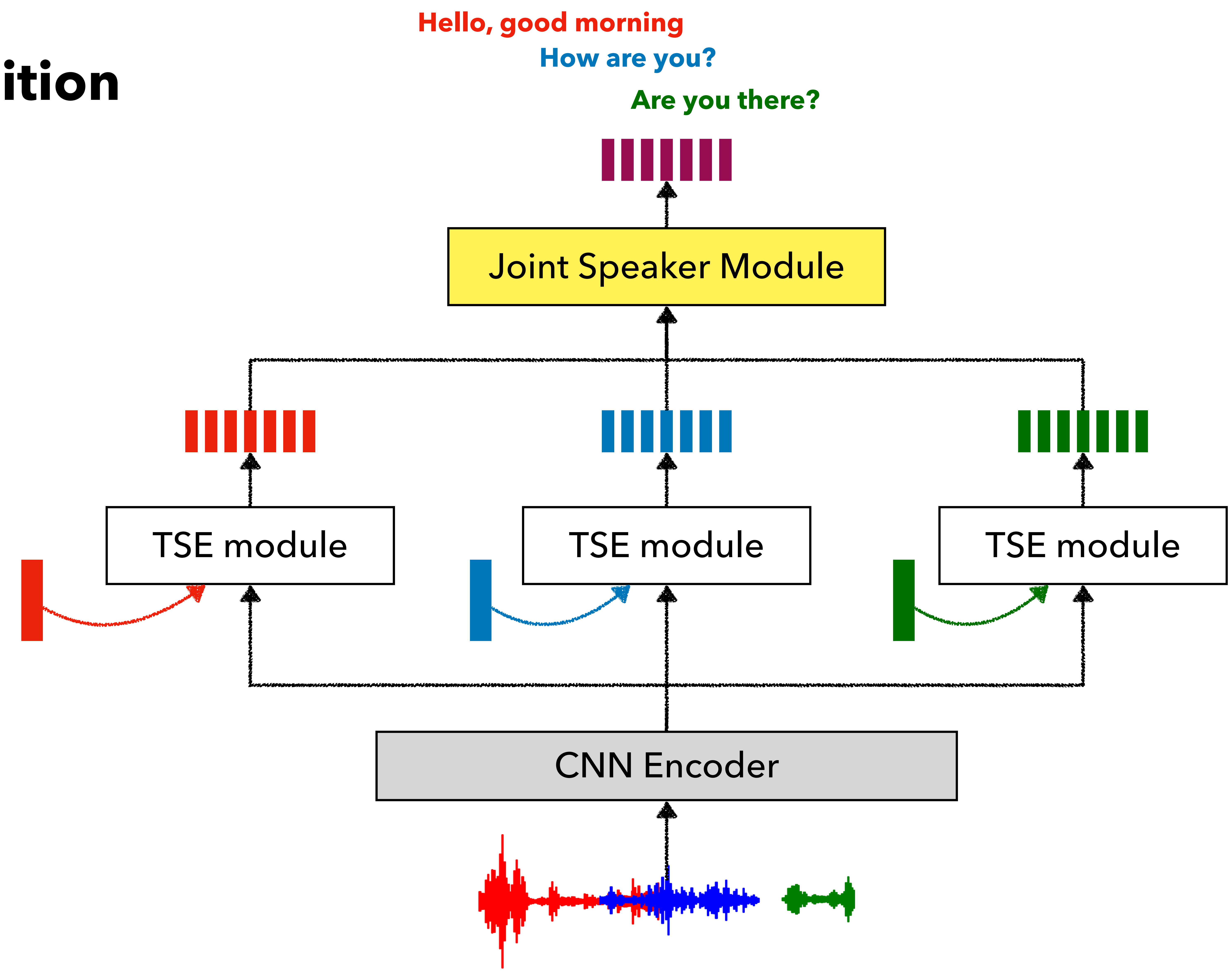
SSL pretrained components

SSL + TS-ASR

Multi-talker speech recognition



$P(W_A, W_B | \mathbf{X}, \mathbf{s}_A, \mathbf{s}_B) \sim P(W_A | \mathbf{X}, \mathbf{s}_A)\ P(W_B | \mathbf{X}, \mathbf{s}_B)$



SSL + TS-ASR

Results on AMI (unsegmented)

Model	WER (%)
WavLM (no fine-tuning)	100.3
WavLM + TSE (iterative)	49.5
WavLM + TSE + JSM	**28.4**

Summary

- Target-speaker ASR comes in **different flavors**, depending on the use-case.

Method	Application	Auxiliary information	# channels	Streaming?
GSS	Meeting transcription	Context (implicit)	Multi-channel	No
Anchored ASR	Voice-based assistants	Wake-word	Single-channel	Yes
WavLM + TSE + JSM	Meeting transcription	Speaker embedding	Single-channel	No