On Speaker Attribution with SURT

Desh Raj, Matthew Wiesner, Matthew Maciejewski, Paola Garcia, Daniel Povey, Sanjeev Khudanpur

Desh Raj
Meta

JOHNS HOPKINS UNIVERSITY

Motivation

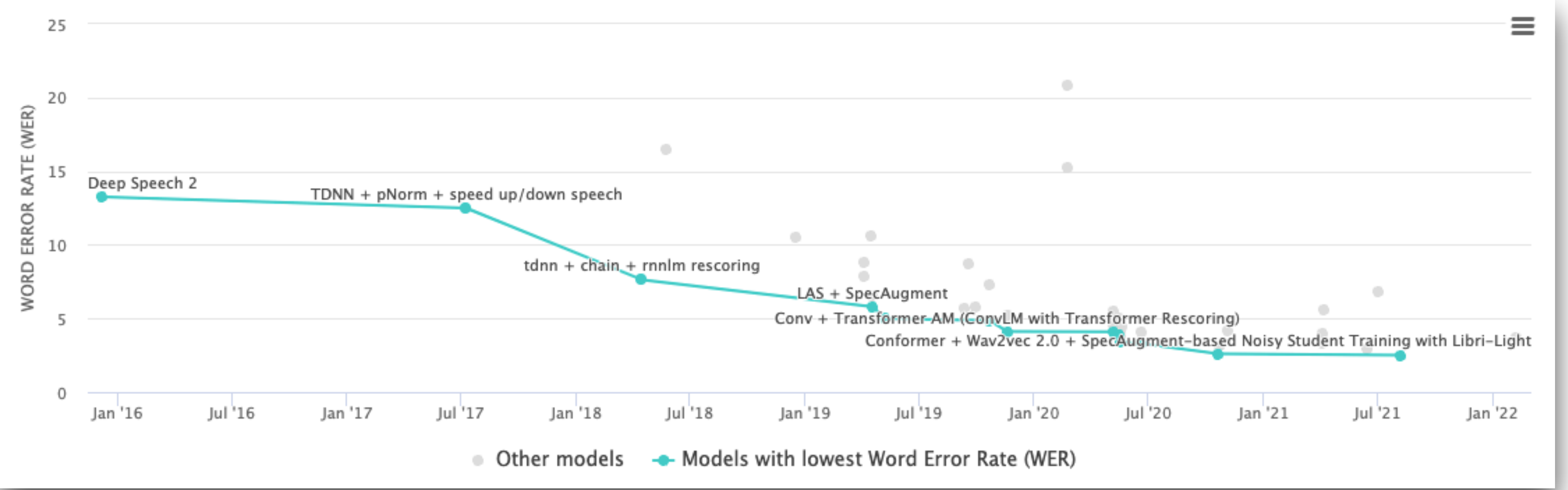
OCTOBER 20, 2020

AI outperforms humans in speech recognition

by Monika Landgraf, Karlsruhe Institute of Technology

Microsoft claims new speech recognition record, achieving a super-human 5.1% error rate

BY TODD BISHOP on August 20, 2017 at 7:44 pm



https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-other

Motivation

Single-user applications

Smart Assistants

Language Learning

Customer Service

Voice-based Search

Multi-user applications

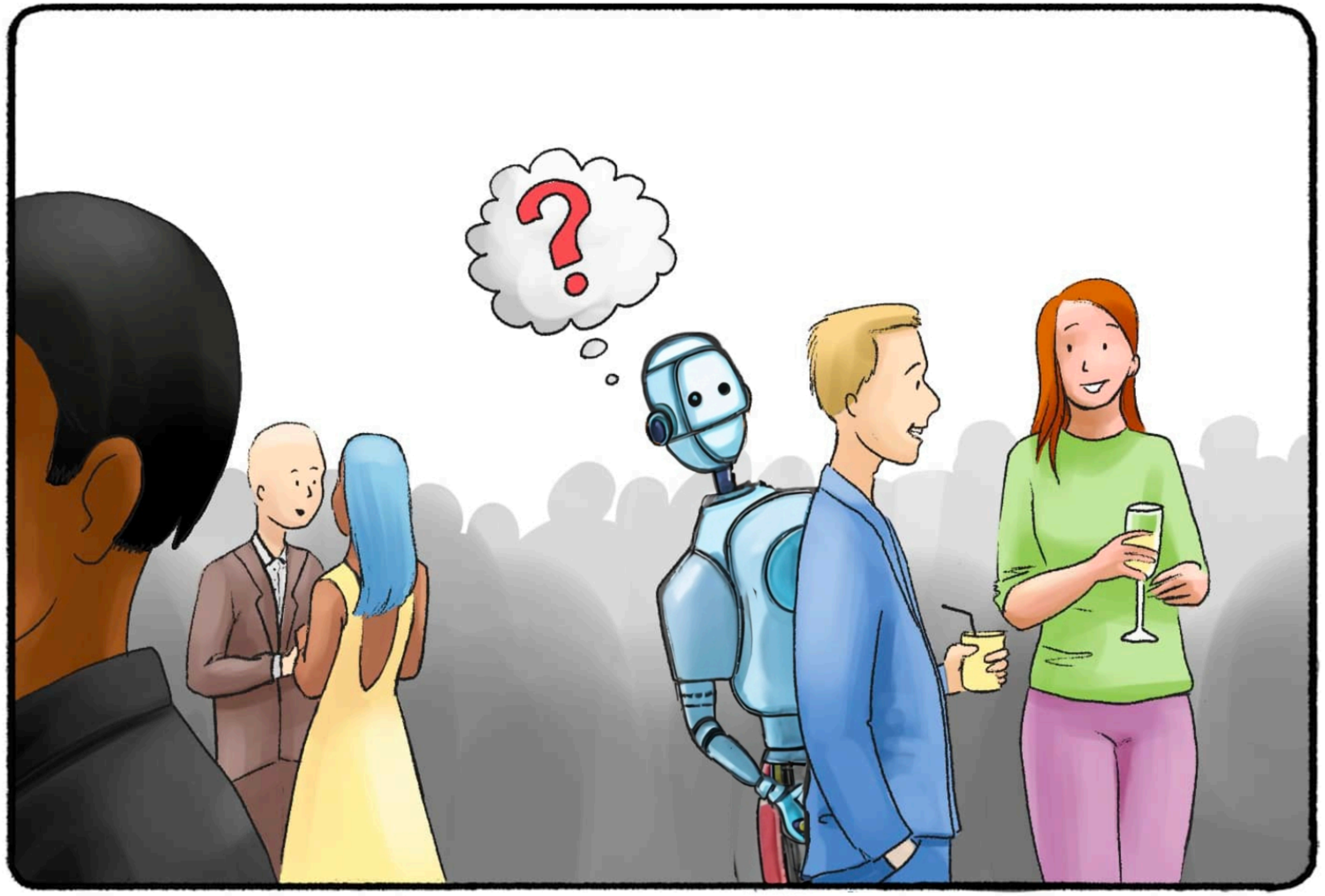
Meeting summaries

Collaborative Learning

Child language development

Motivation

The Cocktail Party Problem



Outline of the talk

1. Problem statement: “who spoke what?”
2. Modular system and its Limitations
3. Streaming Unmixing and Recognition Transducer (SURT)
4. Speaker-attributed transcription with SURT
5. Conclusion

Problem Statement

Multi-talker speaker-attributed ASR

- **Input:** long unsegmented (possibly multi-channel) recording containing multiple speakers.
- **Output:**
  - Transcription of the recording (speech recognition)
  - Speaker attribution (diarization)
  - Additional constraints: streaming, i.e., real-time transcription
- We specifically look at “meetings”: AMI, ICSI

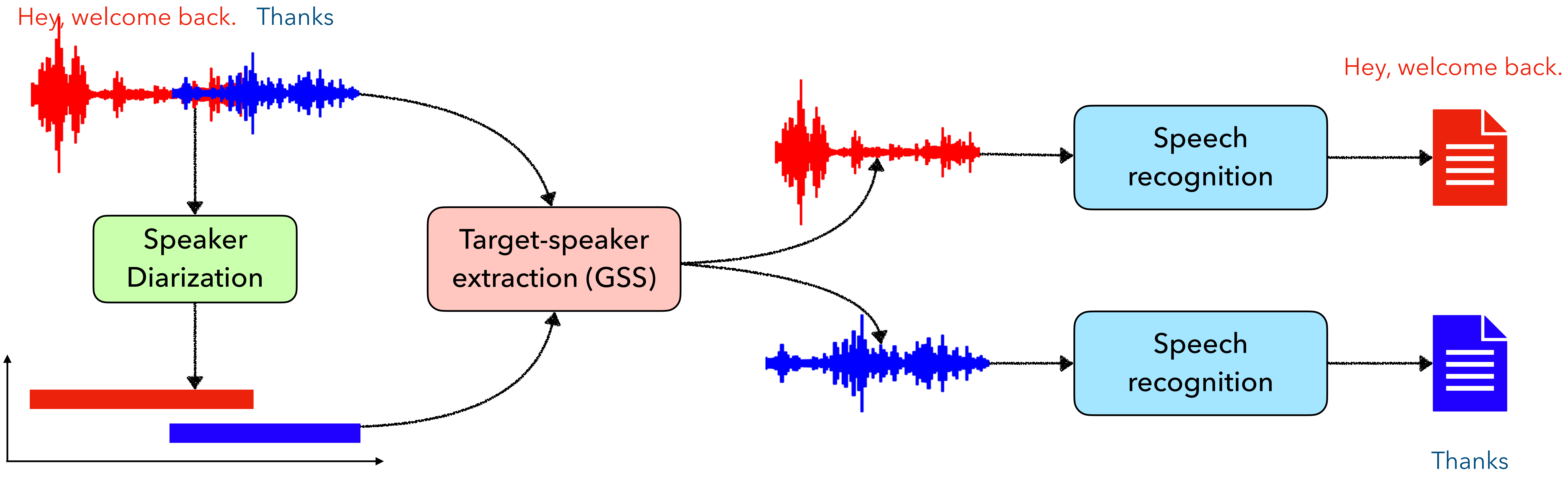
Problem Statement

Evaluation metrics

- *Speech Recognition*
  - Word error rate (**WER**) = insertion + deletion + substitution (Levenshtein distance)
- *Speaker Diarization*
  - Diarization error rate (**DER**) = missed speech + false alarm + speaker confusion
  - Word diarization error rate (**WDER**) = % of correctly recognized words attributed to the wrong speaker
- *Multi-talker ASR*
  - **ORC-WER**: WER for overlapping speech **without** speaker attribution
  - **cpWER**: WER for overlapping speech **with** speaker attribution

Modular system

Pipeline from the CHiME challenge



Shinji Watanabe, et al. CHiME-6 Challenge: Tackling Multi-speaker Speech Recognition for Unsegmented Recordings. *CHiME Workshop, 2020*.

Desh Raj, et al. GPU-accelerated Guided Source Separation for Meeting Transcription. *Interspeech, 2023*.

Modular system

Limitations

- Modules are independently optimized for different objectives
- Higher accumulated **latency**
- **Error propagation** through modules
- Requires more engineering efforts to maintain

- Cannot be used for streaming or single-channel inputs

Continuous, streaming, multi-talker ASR

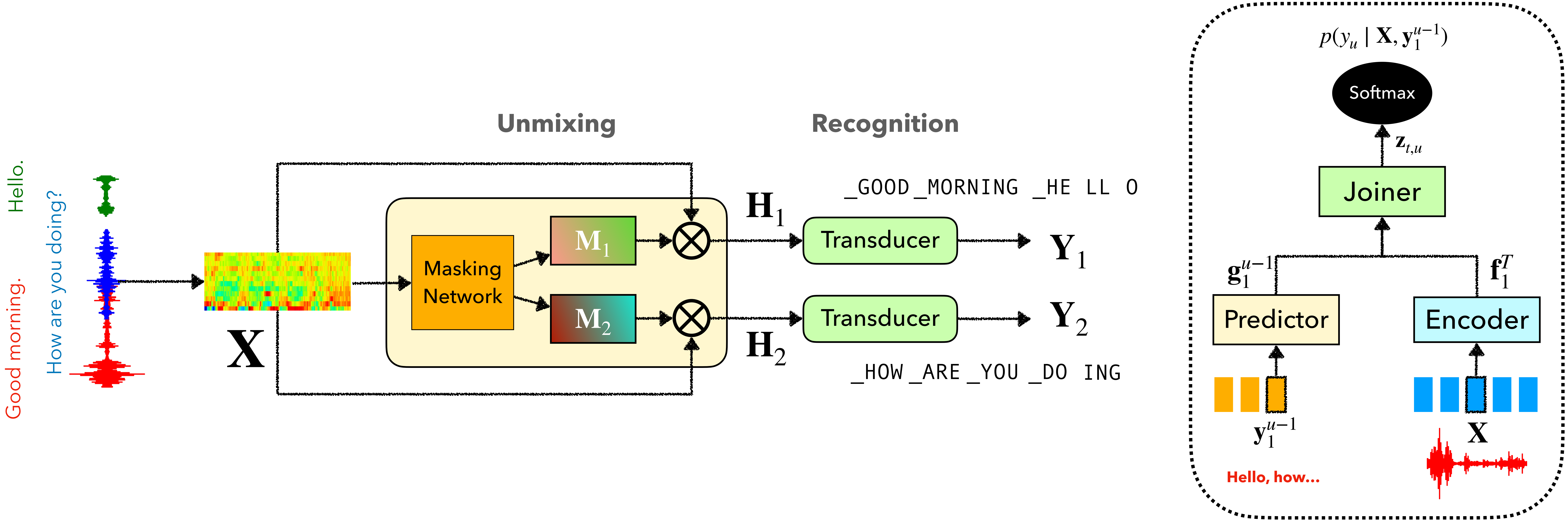
Definitions

- **Continuous:** does not rely on external segmentation
- **Streaming:** does not use right context; overlapping speech is transcribed simultaneously

Desh Raj, et al. Continuous Streaming Multi-Talker ASR with Dual-Path Transducers. *IEEE ICASSP, 2022*.

Desh Raj, et al. SURT 2.0: Advances in Transducer-Based Multi-Talker Speech Recognition. *IEEE/ACM TASLP, vol. 31, 2023*.

Streaming Unmixing and Recognition Transducer (SURT)



- To solve the **permutation problem**, assign utterances to first available channel in order of start time

$$\mathscr{L}_{\text{heat}}(\mathbf{y}_{1:N}, \mathbf{X}; \Theta) = -\log P_{\Theta}(\mathbf{Y}_1 \mid \mathbf{X}) - \log P_{\Theta}(\mathbf{Y}_2 \mid \mathbf{X})$$

Streaming Unmixing and Recognition Transducer (SURT)

Results on real meetings (AMI and ICSI)

AMI	
Close-talk WER (%)	35.1
Far-field WER (%)	44.6

Overlap ratio = 21.6%

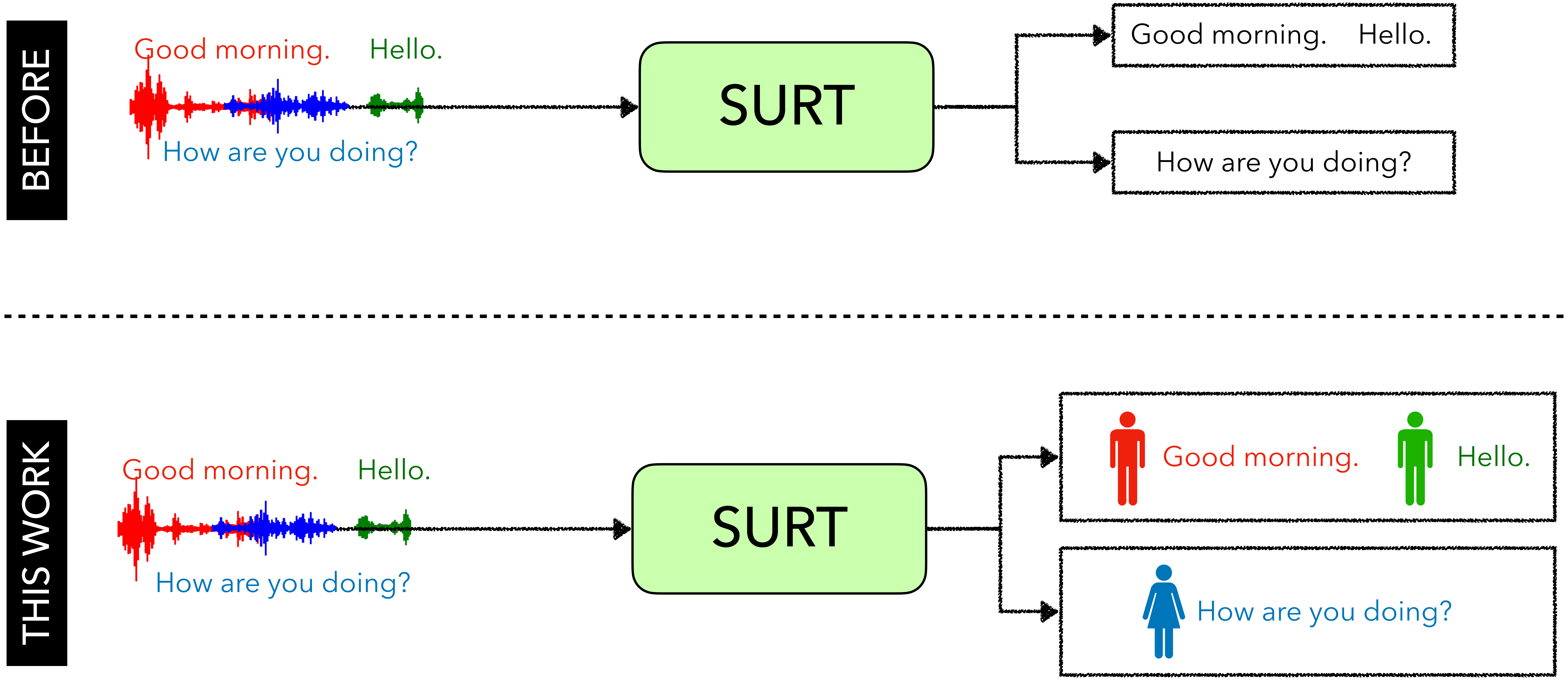
ICSI	
Close-talk WER (%)	24.4
Far-field WER (%)	32.2

Overlap ratio = 11.1%

- Results in terms of ORC-WER (speaker-agnostic).
- As a comparison, a single-speaker model for AMI gets ~18% (close-talk) and 32% (far-field).

Speaker attribution with SURT

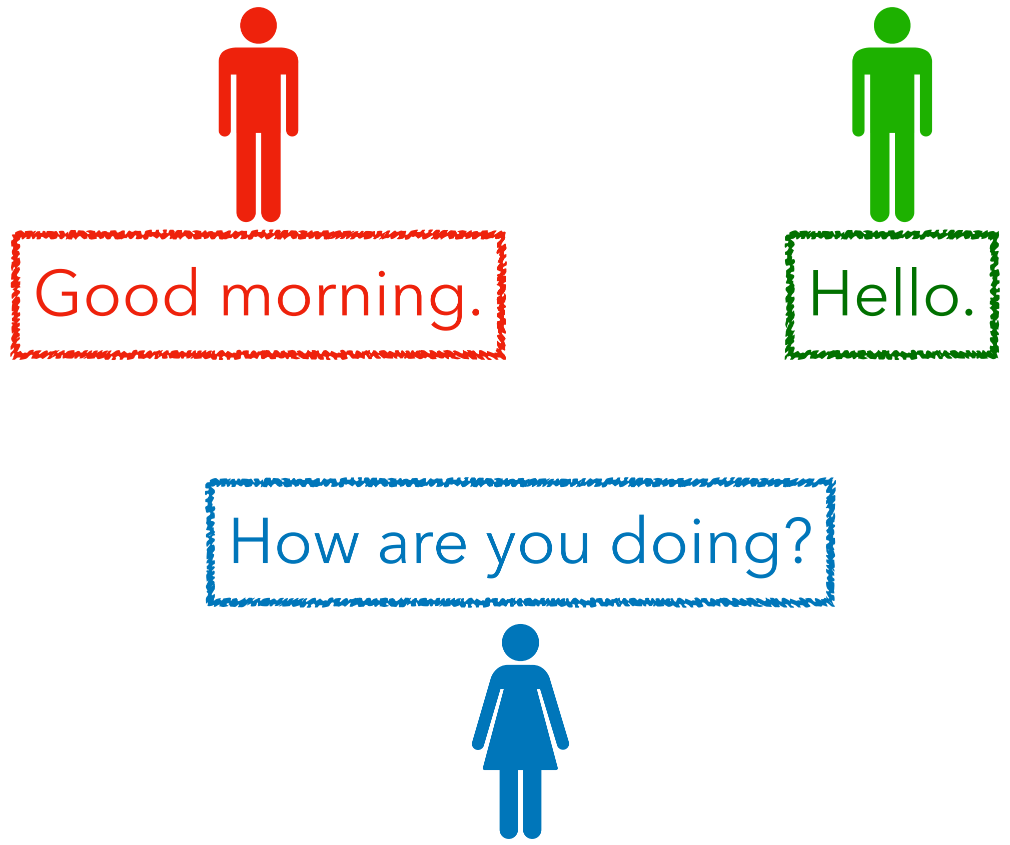
How to predict speaker labels with ASR tokens?



Speaker attribution with SURT

Heuristic error assignment training for speakers

- Use the same 2-branch strategy, but predict speaker labels instead of ASR tokens
- Speakers are ordered in their relative order of appearance
- *How to do both tasks jointly?*

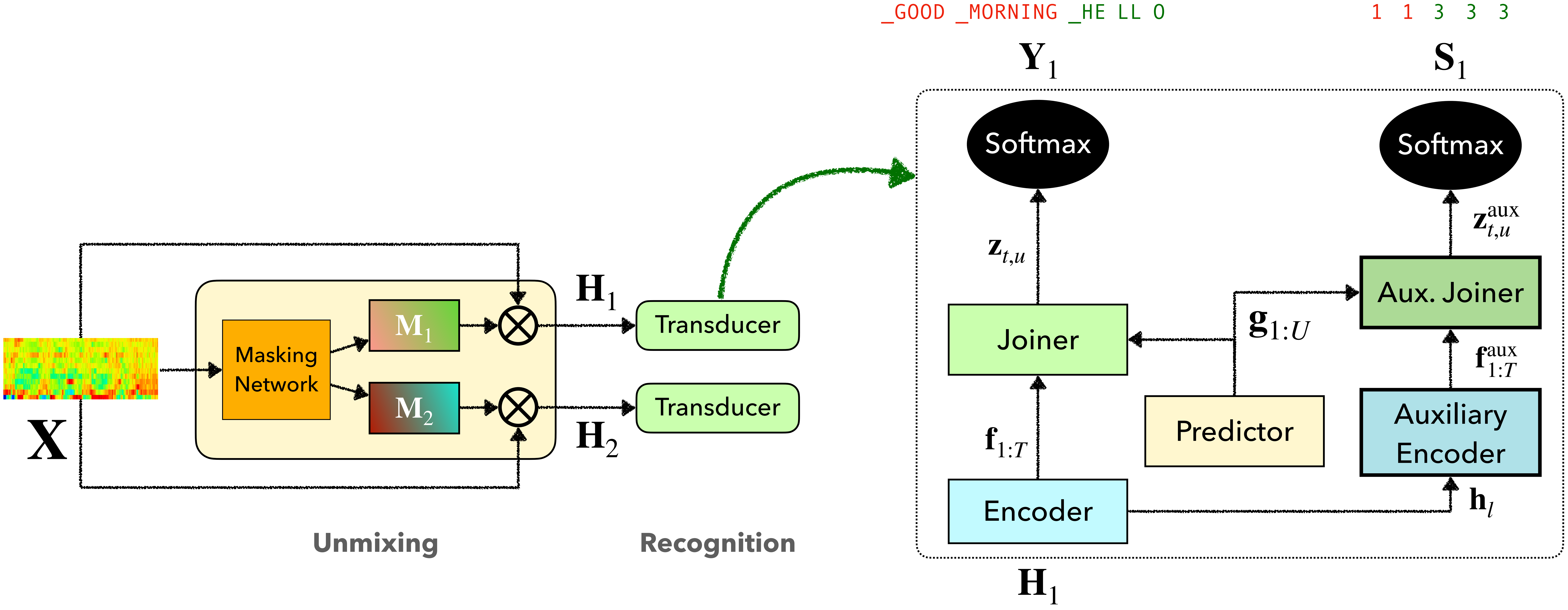


$\mathbf{Y}_1$	_GOOD	_MORNING	_HE	LL	O
$\mathbf{S}_1$	1	1	3	3	3

$\mathbf{Y}_2$	_HOW	_ARE	_YOU	_DO	ING
$\mathbf{S}_2$	2	2	2	2	2

Speaker attribution with SURT

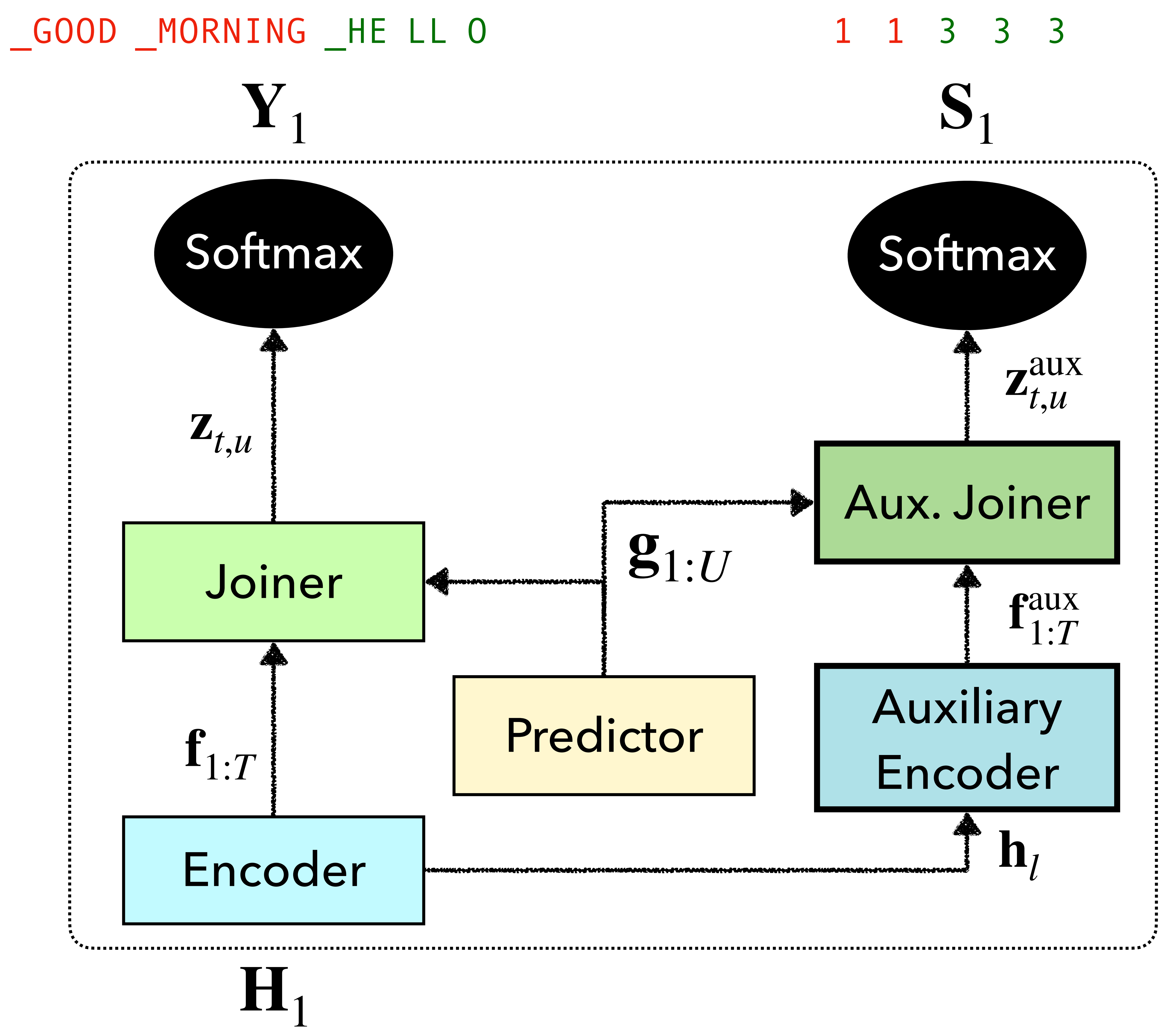
Auxiliary speaker encoder



Speaker attribution with SURT

Synchronizing speaker labels with ASR tokens

- At inference time, it is not necessary that both output streams emit same number of tokens.
- Even if they do, they may not be frame synchronous.



$\mathbf{Y}_1$	<blk>	_GOOD	_MORNING	<blk>	_HE	<blk>	LL	O
$\mathbf{S}_1$	<blk>	1	<blk>	1	<blk>	3	<blk>	3

Speaker attribution with SURT

Hybrid autoregressive transducer (HAT)

RNN-Transducer

$$P(\mathbf{a}_t \mid \mathbf{f}_1^t, \mathbf{g}_1^{u(t)-1}) = \mathrm{Softmax}(\mathbf{z}_{\mathrm{t,u}})$$

- Multinomial distribution over blank and non-blank tokens
- Cannot model blank probability separately

HAT

$$P(\mathbf{a}_t \mid \mathbf{f}_1^t, \mathbf{g}_1^{u(t)-1}) = \begin{cases} b_{t,u}, & \text{if } \mathbf{a}_t = \phi, \\ (1 - b_{t,u})\ \mathrm{Softmax}(\mathbf{z}_{\mathrm{t,u}}[1:]), & \text{otherwise} \end{cases} \qquad b_{t,u} = \sigma(\mathbf{z}_{t,u}[0])$$

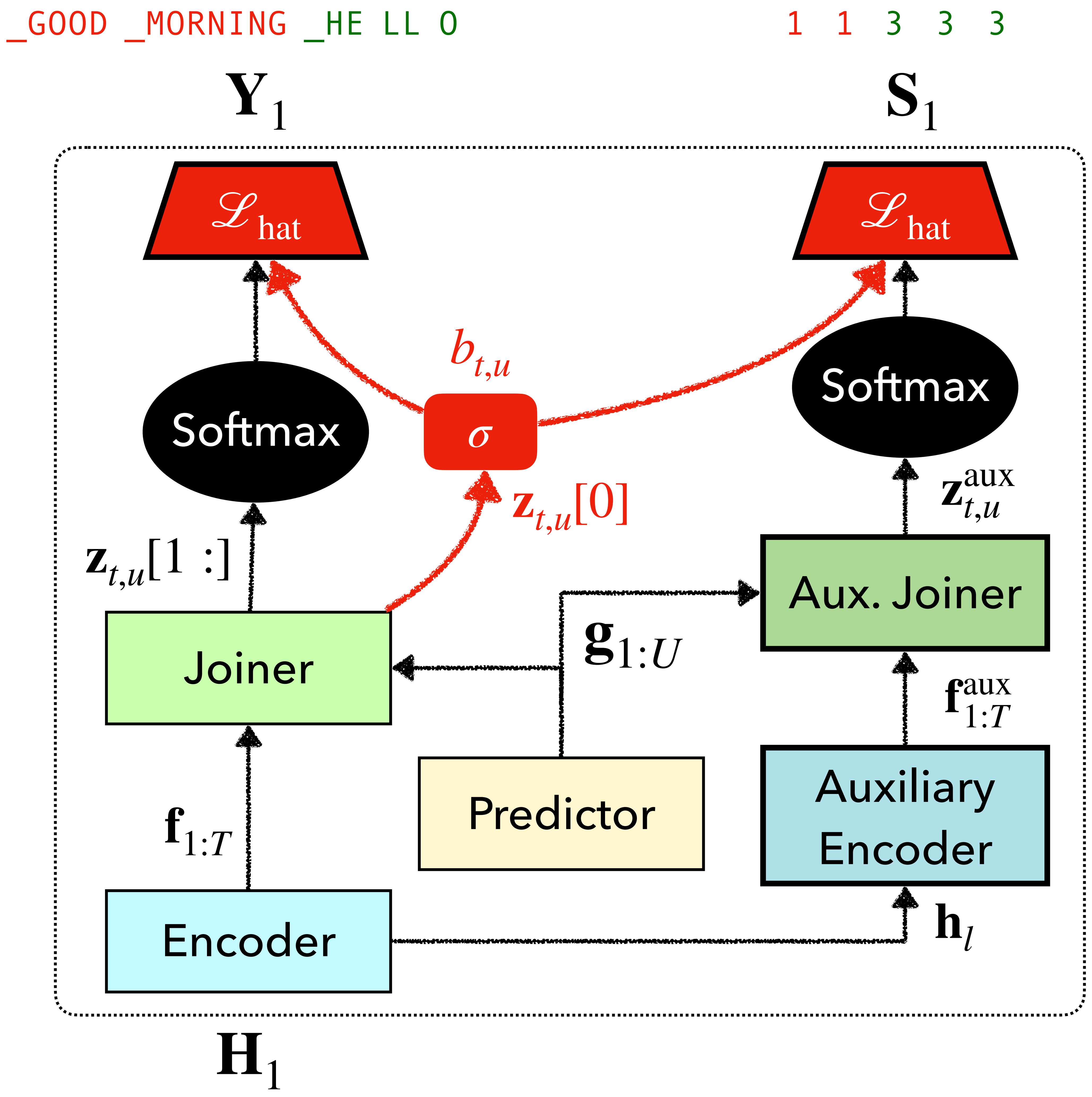
- Bernoulli distribution for blank; multinomial over non-blank tokens
- Probability of blank given directly by $b_{t,u}$

Ehsan Variani, et al. Hybrid Autoregressive Transducer (HAT). *IEEE ICASSP 2020*.

Speaker attribution with SURT

Synchronization by sharing `<blk>`

- If ASR branch emits `<blk>` do the same for speaker branch
- This is achieved by using HAT-style blank factorization, and sharing blank logit between ASR and speaker branch



Speaker attribution with SURT

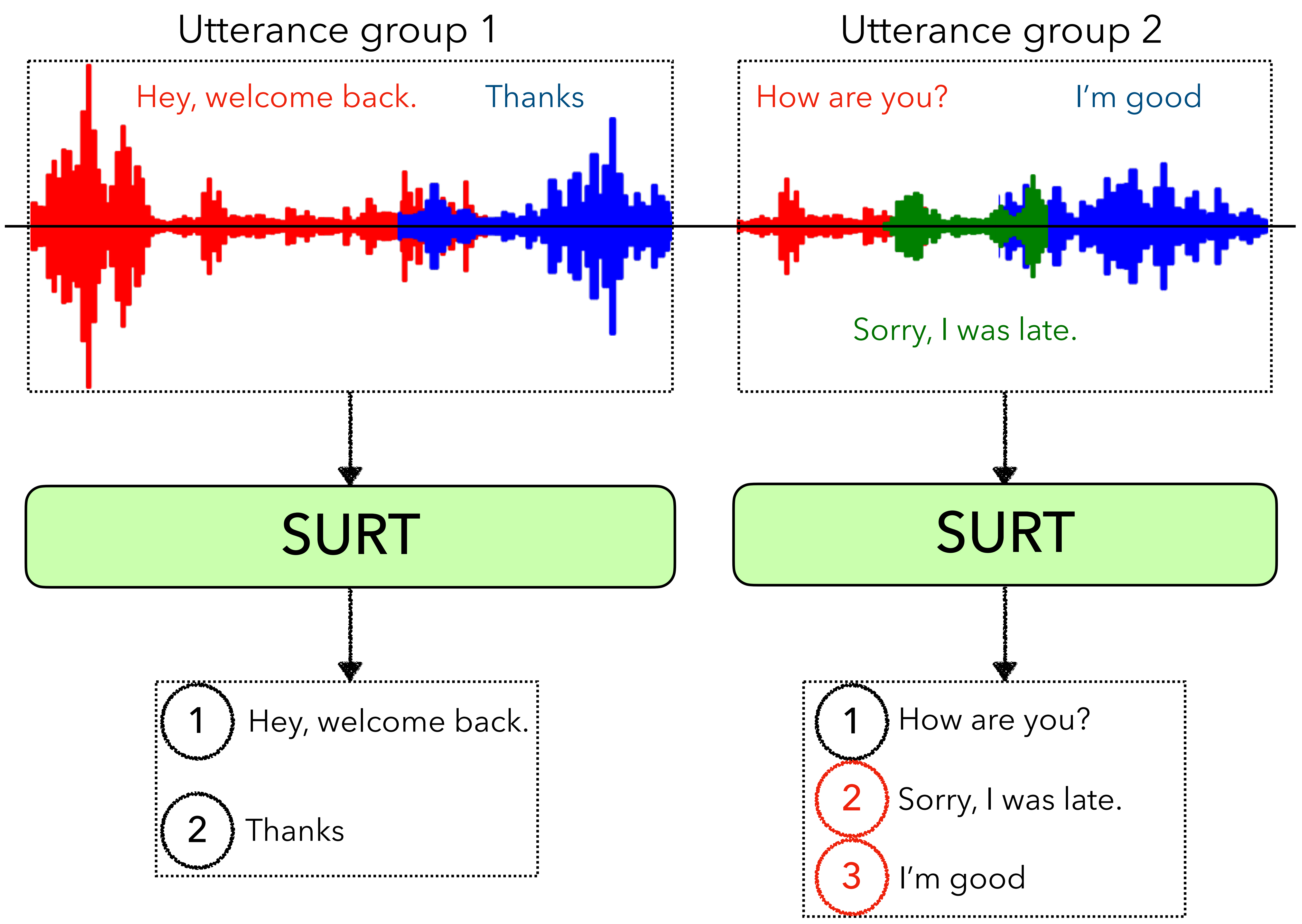
Results on AMI (evaluation on utterance groups)

Utterance group = set of utterances connected by overlaps or short pauses

	Streaming			Offline
Mic Setting	**ORC-WER**	**WDER**	**cpWER**	**Modular System cpWER**
Close-talk	34.9	9.3	42.3	–
Far-field	43.2	10.9	50.3	38.5

Speaker attribution with SURT

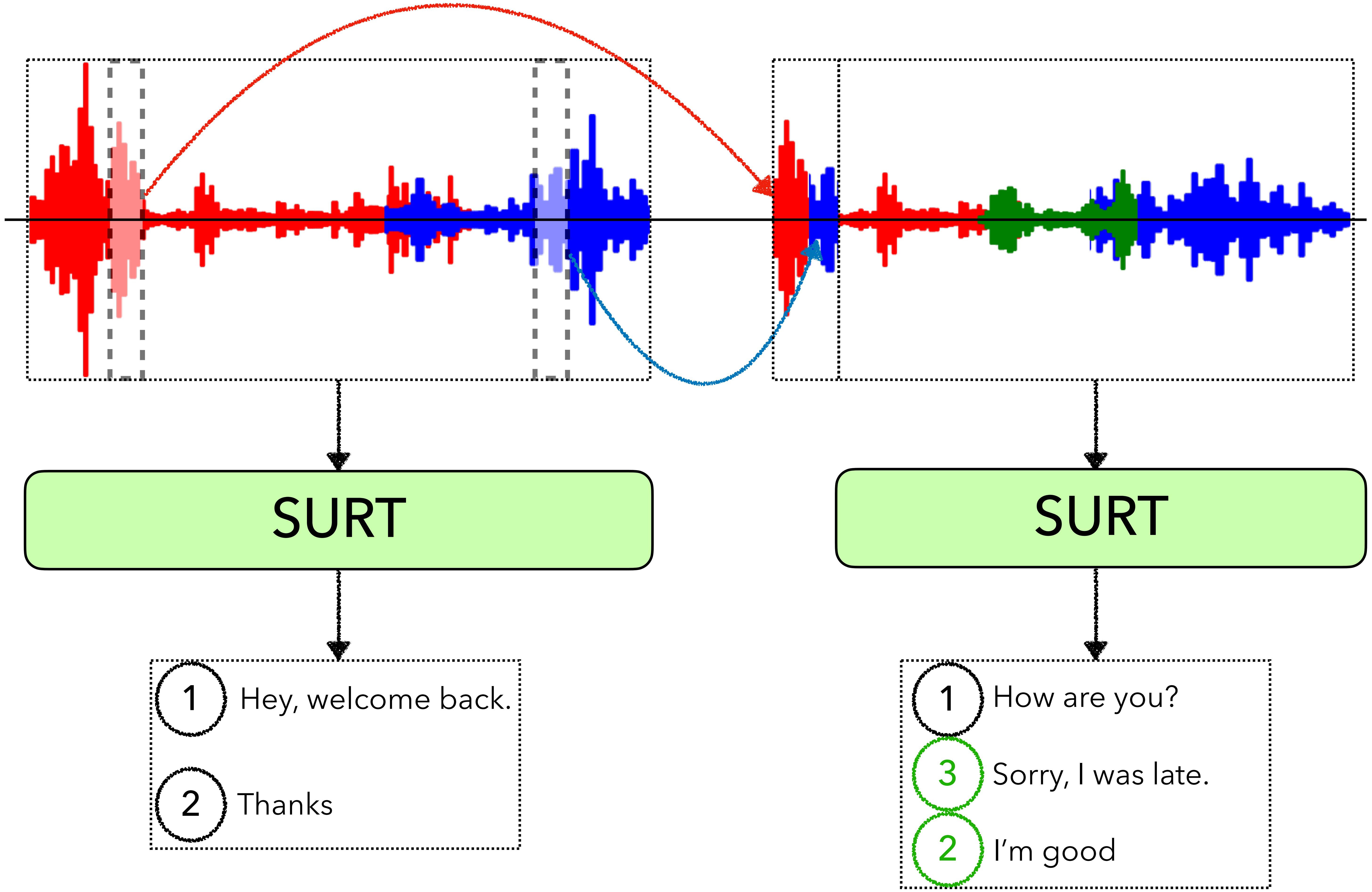
From utterance groups to full sessions



- How to maintain relative speaker labels when processing different utterance groups within the same session?

Speaker attribution with SURT

Speaker prefixing approach



- Extract high-confidence frames of predicted speakers and prefix them in front of current input.
- Remove prefixed part from encoder representation.

Summary

- **We showed that the same models that do transcription can also do speaker attribution with small changes!**
- For more results and analysis, please refer to our paper.
- Reviewer #5: *"I assume the authors are very eager to have these results published in Odyssey since a different (and longer) format would probably have suited this content better."*

@rdesh26 desh@meta.com

Extra Slides

Speaker attribution with SURT

Joint vs. sequential training

Experiments on simulated LibriSpeech mixtures

Method	ORC-WER	WDER	cpWER
✓ **Sequential**	8.5	**4.0**	**15.0**
Joint	**8.4**	4.5	**15.0**
Sequential + joint	9.2	4.3	15.3

Speaker attribution with SURT

Where to branch out of the main encoder?

Experiments on simulated LibriSpeech mixtures

Main Encoder Block	WDER	cpWER
Block 0 (after embedding layer)	5.4	16.7
Block 1	**4.0**	**15.0**
Block 2	6.7	19.6
Block 3	8.4	23.4

Speaker attribution with SURT

Evaluation on AMI IHM-Mix setting

"Enrollment" = using small chunk from speaker's enrollment speech for prefixing

Evaluation	Method	cpWER
Utterance group	SURT w/o speaker prefix	42.3
Full session	SURT w/o speaker prefix	100.1
	SURT w/ speaker prefix (128 frames = 1.28s per speaker)	82.8
	+ enrollment	53.8