Published at IEEE/ACM Transactions on Audio, Speech, and Language Processing

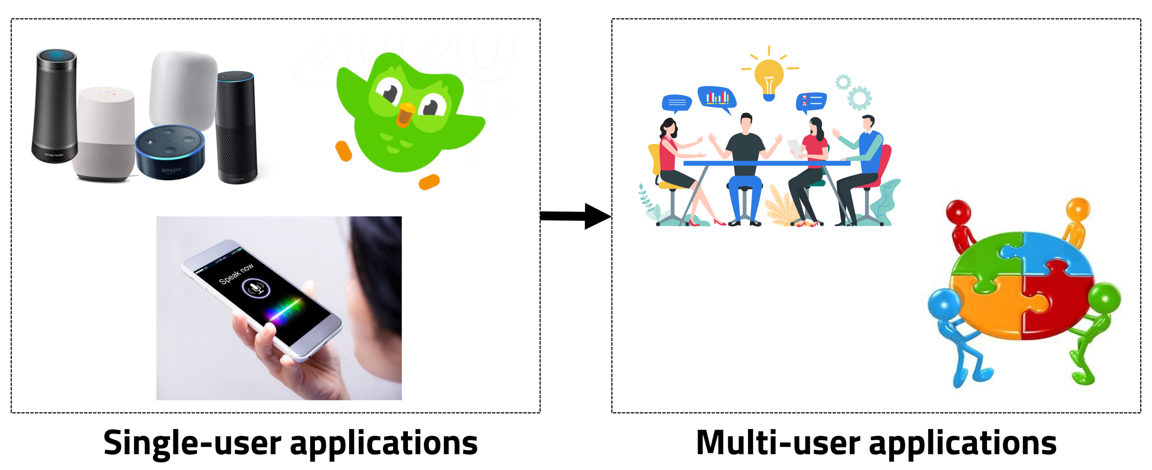
SURT 2.0: Advances in Transducer-based Multi-talker Speech Recognition

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

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

Motivation

- Existing ASR systems are mostly geared towards single-user applications.
- We want to build systems that answer "**who spoke what**" for free-flowing multi-party conversations, in real-time.
- How to train efficient **end-to-end** neural models for this task?



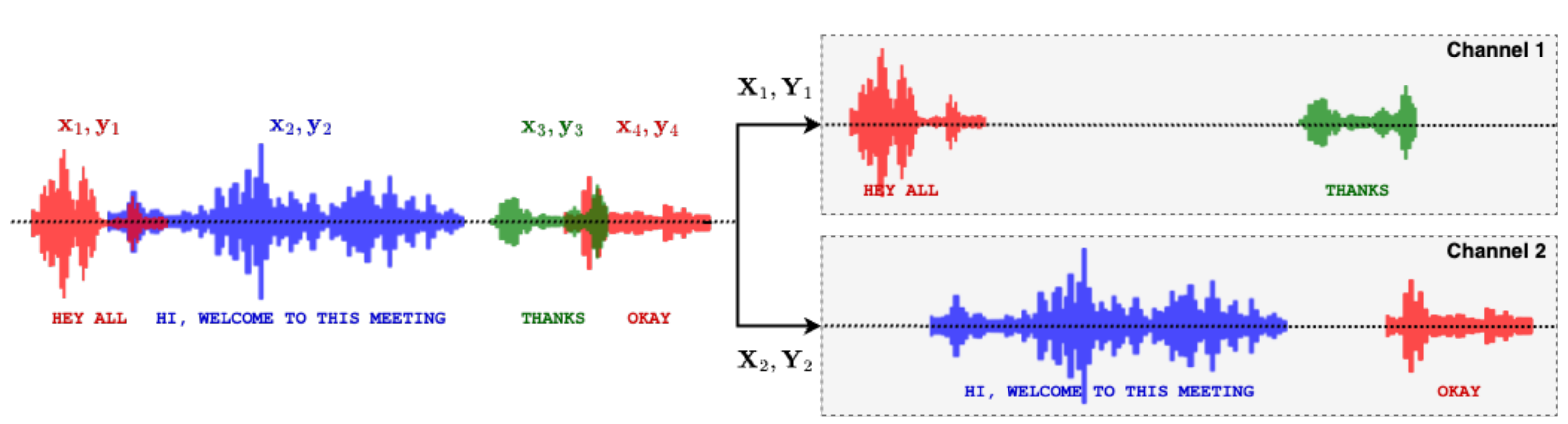
Challenges

Model — Multi-talker conversations contain **overlapping speech** and back-channels.

Data — There is limited amount of **real data** available for training end-to-end neural models.

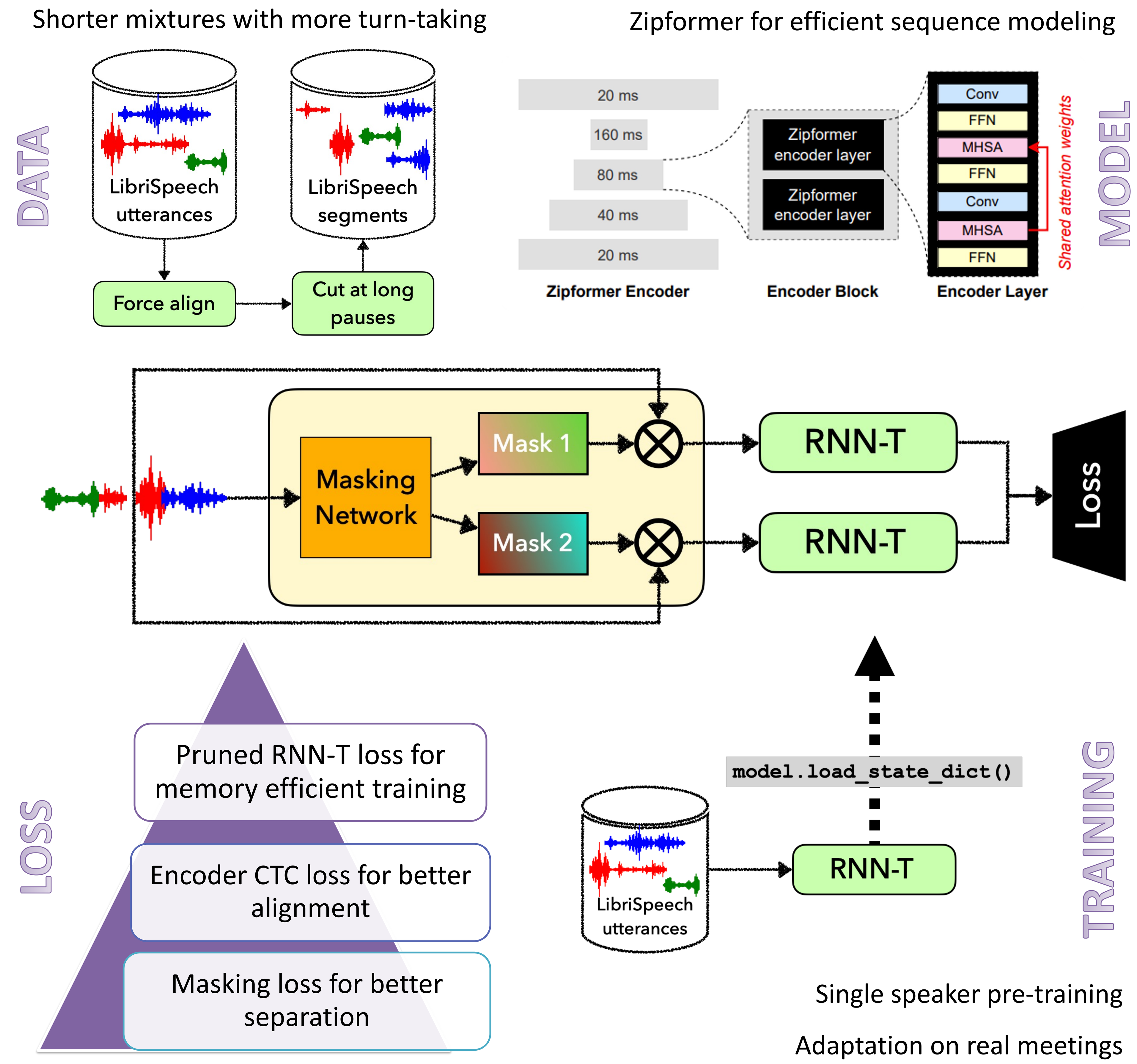
Compute — Training such systems requires **large computational resources.**

Continuous Streaming Multi-talker ASR



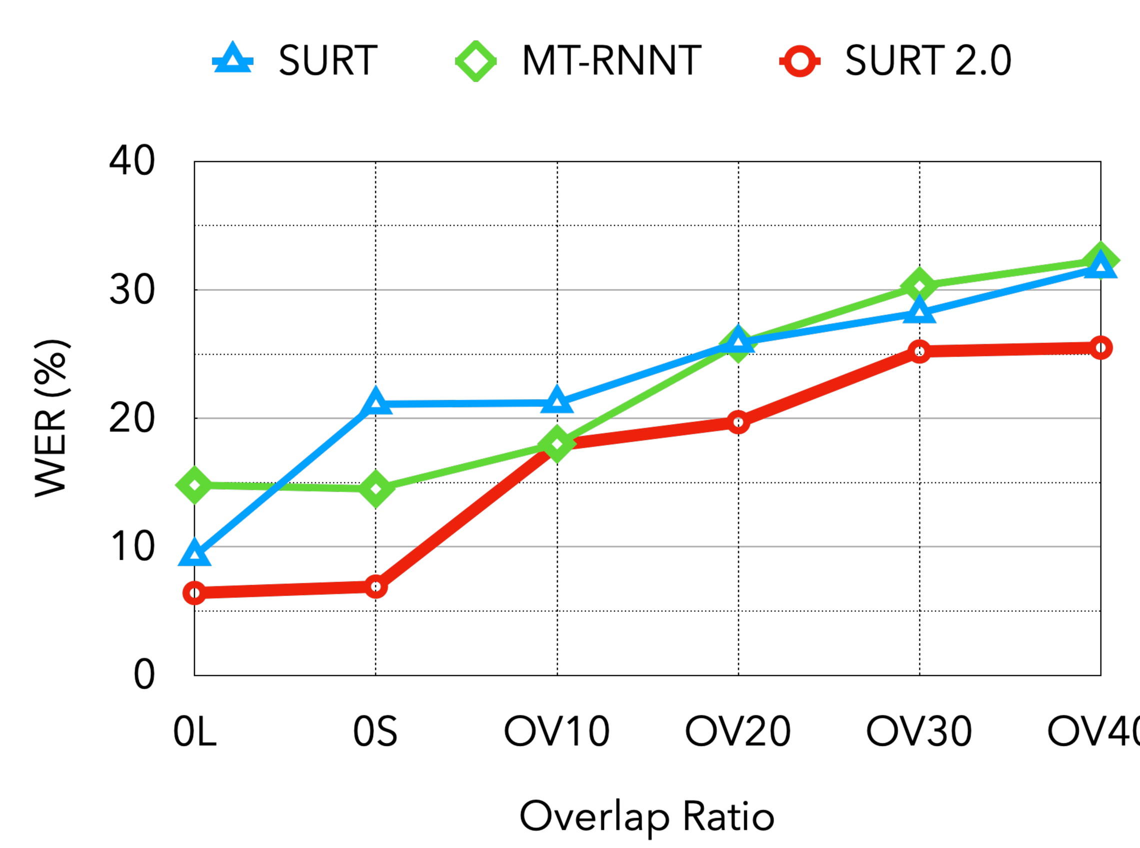
Continuous	Streaming
• No need of external segmentation	• Overlapping speakers transcribed simultaneously

Streaming Unmixing and Recognition Transducer



Results

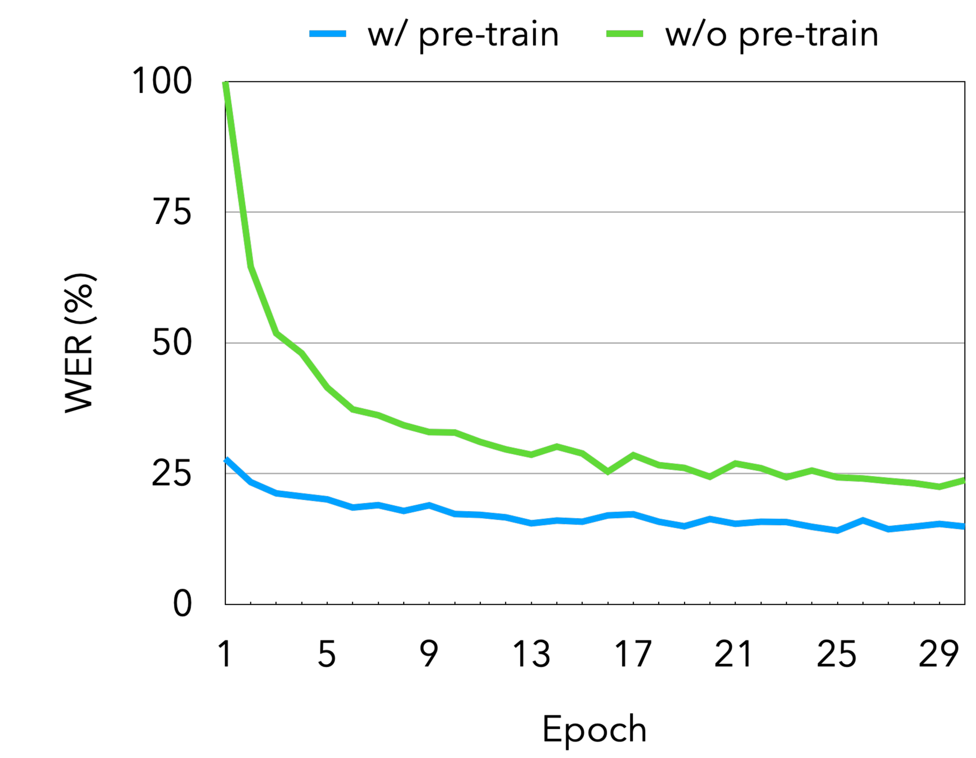
- Experiments on meeting corpora: LibriCSS, AMI, ICSI
- LibriCSS is "simulated"; AMI and ICSI are real meetings
- SURT 2.0 obtains 44.6% and 32.2% WER on real far-field meetings.



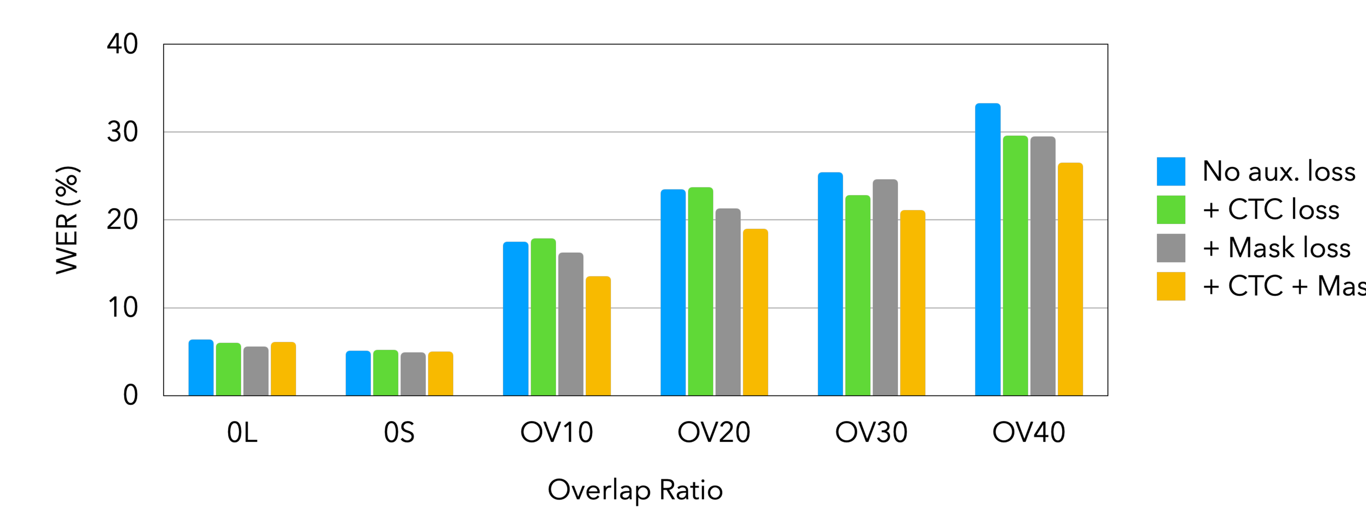
Model	# params (M)	WER (%)
MT-RNNT [1]	81.0	22.6
SURT [2]	42.9	22.9
SURT 2.0	**37.9**	**16.9**

Analysis

1. Single speaker pre-training is critical.



2. Auxiliary objectives improve performance on high-overlap conditions.



LHOTSE k2 ICEFALL

References

[1] I. Sklyar, A. Piunova, Y. Liu. "Streaming multi-speaker ASR with RNN-T." *IEEE ICASSP, 2021.*

[2] D. Raj, L. Lu, Z. Chen, Y. Gaur, J. Li. "Continuous streaming multi-talker ASR with dual-path transducers." *IEEE ICASSP 2022.*

r.desh26@gmail.com
@rdesh26
desh2608.github.io
rdesh26

SCAN ME