Continuous Streaming Multi-talker ASR with Dual-path Transducers

Desh Raj[1], Liang Lu[2], Zhuo Chen[2], Yashesh Gaur[2], Jinyu Li[2]

[1] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[2] Microsoft Corp., Redmond, USA

Objectives

- Use transducer-based models for **continuous** and **streaming** transcription, in the long-form multi-talker ASR task (e.g., LibriCSS [1]).
- Investigate Streaming Unmixing and Recognition Transducer (SURT) [2], which was previously evaluated on single-turn sessions.
- How to make the SURT model work for longer sessions containing multiple speakers?

Introduction

What is continuous streaming ASR?

- **Continuous:** Does not rely on external segmentation for long-form audio.
- **Streaming:** Overlapping speakers should be transcribed "simultaneously", instead of one-at-a-time.

Streaming Unmixing and Recognition Transducer (SURT)

- *Unmixer* extracts speaker-specific features from the mixed audio.
- *Recognizer* is a transducer model which transcribes the speaker stream.
- Model is trained end-to-end using RNN-T loss.

Evaluation Data

Table: Synthetic evaluation sets

Name	Description	# spk.	# utt.	dev	test
Tier-1	2-speaker single-turn	2	2	1355	1310
Tier-2	2-speaker multi-turn	2	2-4	892	885
Tier-3	Multi-speaker multi-turn	2-4	2-12	462	450

LibriCSS

- 10-minute sessions containing 8 speakers and 0-40% overlap
- Evaluated in single-channel setting

HEAT vs. PIT

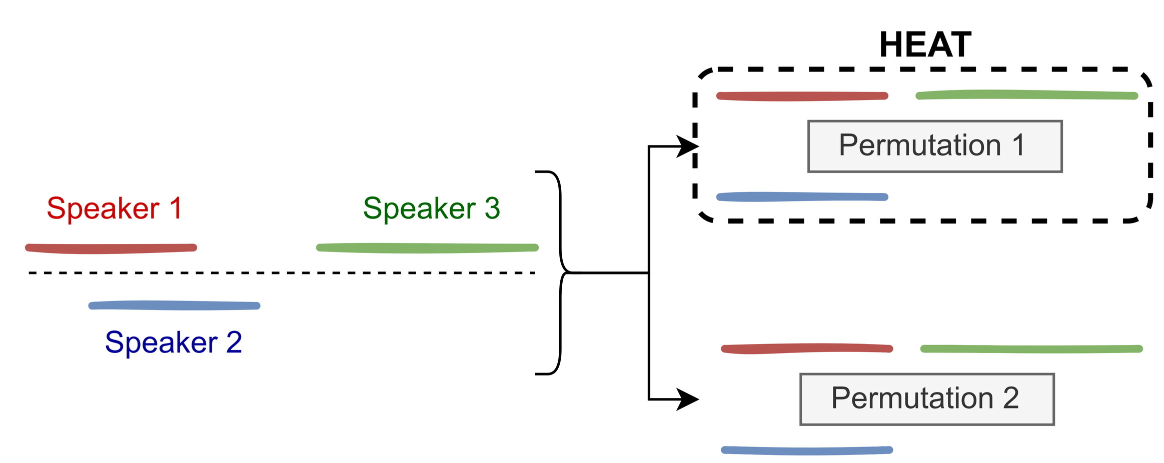
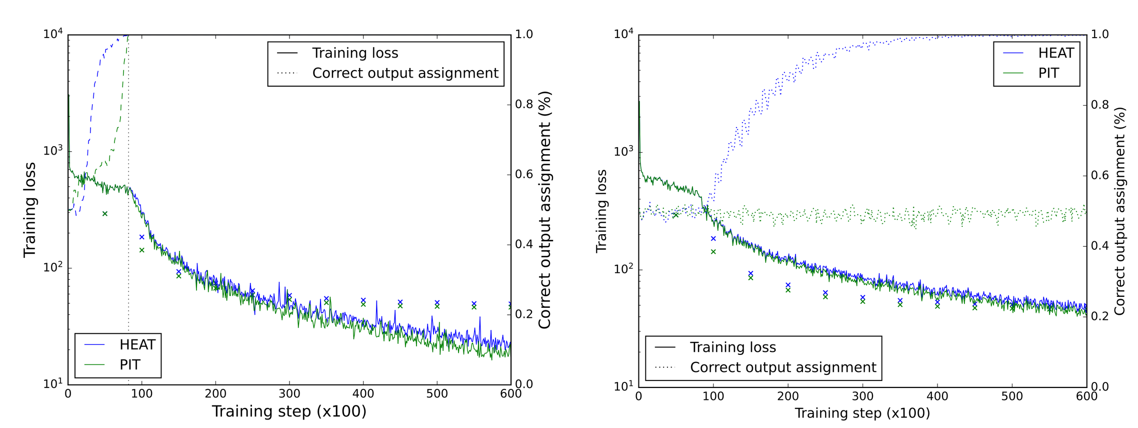


Figure: The HEAT and PIT objectives



(a) Delay = 2.0 s (b) Delay = 0.0 s

Figure: Training dynamics for HEAT versus PIT based loss for different utterance delays: (a) 2.0 s, and (b) 0.0 s.

Streaming Dual-path Transducer

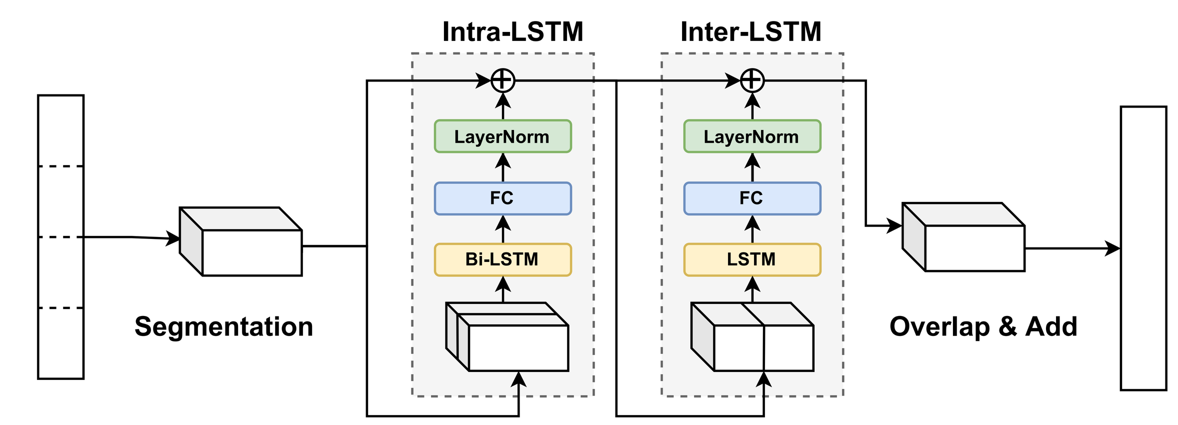


Figure: Dual path RNN

Table: WER results with regular and dual-path encoders.

Encoder	Size (M)	Tier-1 dev	Tier-1 test	Tier-2 dev	Tier-2 test	Tier-3 dev	Tier-3 test
LSTM	75.6 M	13.6	13.8	15.9	17.1	20.9	21.0
DP-LSTM	65.4	**11.1**	**11.4**	**13.0**	**14.1**	19.6	19.6
DP-Transformer	42.9	**11.1**	12.2	13.5	14.5	**17.9**	**18.6**

Important training tricks:

- **Chunk width randomization**
- **Curriculum learning**

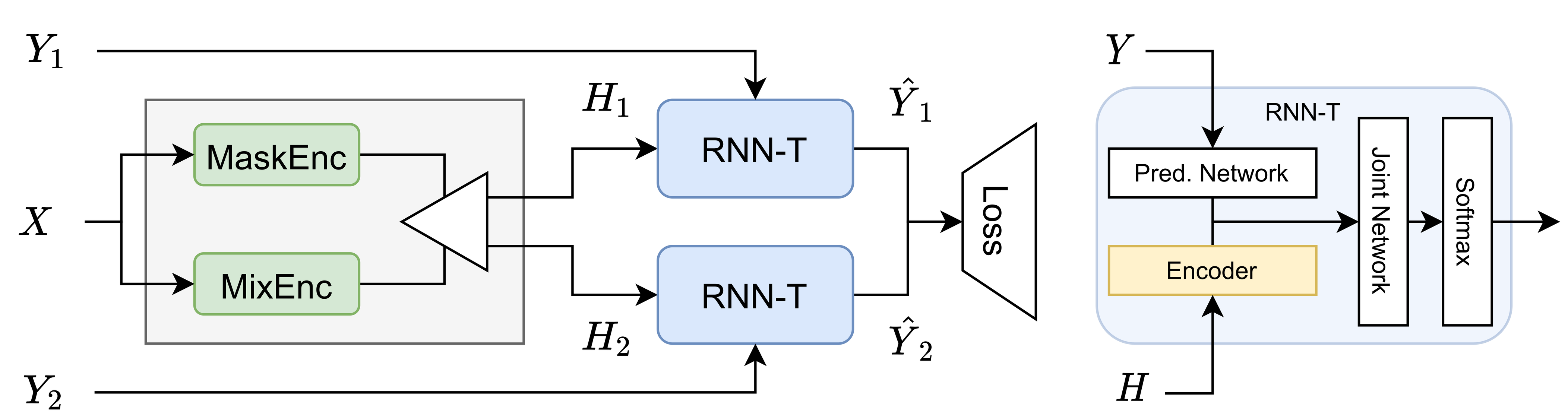


Figure: Streaming Unmixing and Recognition Transducer (SURT).

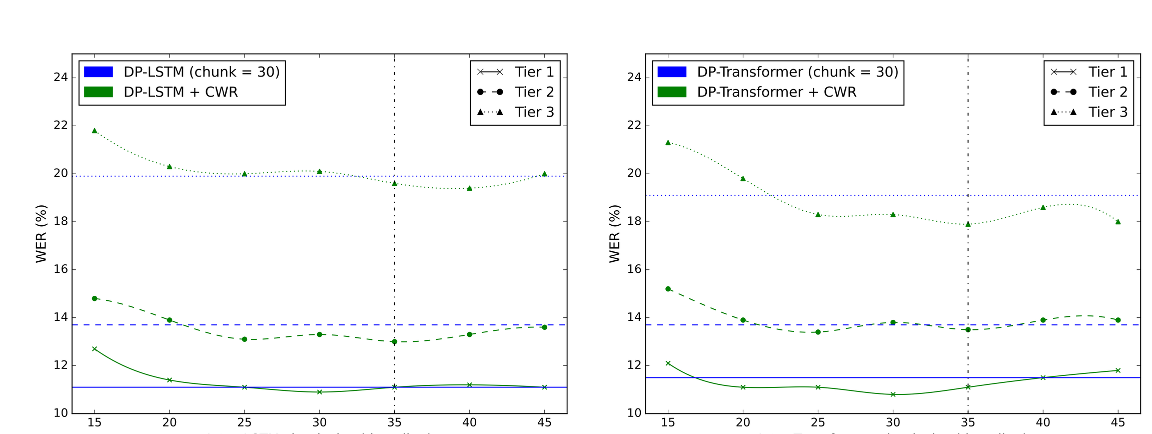
Limitation with the vanilla SURT

LSTM-based SURT models trained on single-turn sessions cannot generalize to multi-turn sessions.

Train \Eval	Tier-1	Tier-2	Tier-3
Single-turn	11.1	17.6	24.9
Multi-turn	13.6	15.9	20.9

Idea: How can we train with multi-turn sessions?

Accuracy vs. Latency



(a) DP-LSTM (b) DP-Transformer

Figure: Accuracy vs. latency trade-off for dual-path models.

Results on LibriCSS

Model	0L	0S	10	20	30	40
BLSTM CSS + Hybrid ASR [1]	16.3	17.6	20.9	26.1	32.6	36.1
Conformer CSS + E2E ASR	6.1	6.9	9.1	12.5	16.7	19.3
SURT w/ DP-LSTM	9.8	19.1	20.6	20.4	23.9	26.8
SURT w/ DP-Transformer	9.3	21.1	21.2	25.9	28.2	31.7

(Overlap ratio in %)

The main sources of errors were

- *leakage* in single-speaker regions, and
- *omissions*, where some utterances were missed by both channels.

References

[1] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, and J. Li. Continuous speech separation: Dataset and analysis. In *IEEE ICASSP*, 2020.

[2] L. Lu, N. Kanda, J. Li, and Y. Gong. Streaming end-to-end multi-talker speech recognition. *IEEE Signal Processing Letters*, 2021.

[3] Y. Luo, Z. Chen, and T. Yoshioka. Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *IEEE ICASSP*, 2020.

Acknowledgements

We thank Naoyuki Kanda and Takuya Yoshioka for providing some of the data simulation scripts and insightful discussions.

Contact Information

- Web: `https://desh2608.github.io`
- Email: `r.desh26@gmail.com`
- Twitter: `@rdesh26`

JOHNS HOPKINS UNIVERSITY Microsoft