Continuous Streaming Multi-talker ASR with Dual-path Transducers

Desh Raj (Johns Hopkins University)

Collaborators: Liang Lu, Zhuo Chen, Yashesh Gaur, Jinyu Li

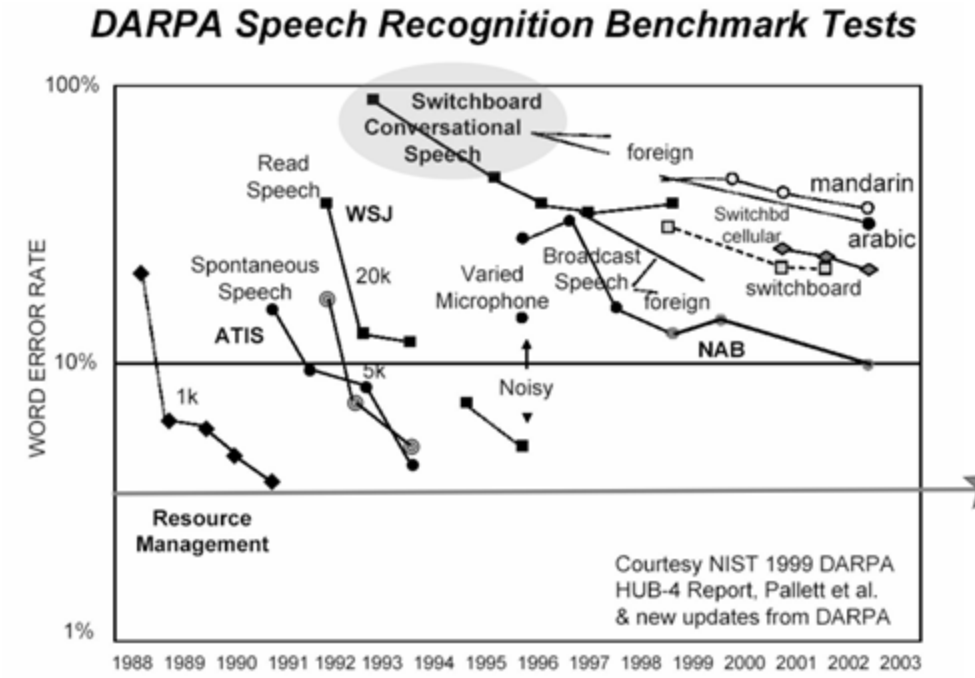
Overview of the talk

- Background
  - Multi-talker speech recognition
  - Streaming Unmixing and Recognition Transducer (SURT)
- From single to multi-turn sessions
  - HEAT vs. PIT
  - Multi-turn evaluation
- Dual-path modeling with LSTMs and Transformers
- Experimental Results

Background

Conversational speech recognition

Human parity on conversational speech using deep neural network models.



5.9% Human performance: Microsoft, October, 2016

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., & Zweig, G. (2016). Achieving Human Parity in Conversational Speech Recognition. *ArXiv, abs/1610.05256.*

Next big challenges:

- Cocktail party problem
- Far-field recordings

Cocktail party problem

Front-end

- Separation
- Denoising
- Dereverberation

Back-end

- Speech recognition
- Speaker counting
- Diarization

What is multi-talker ASR?

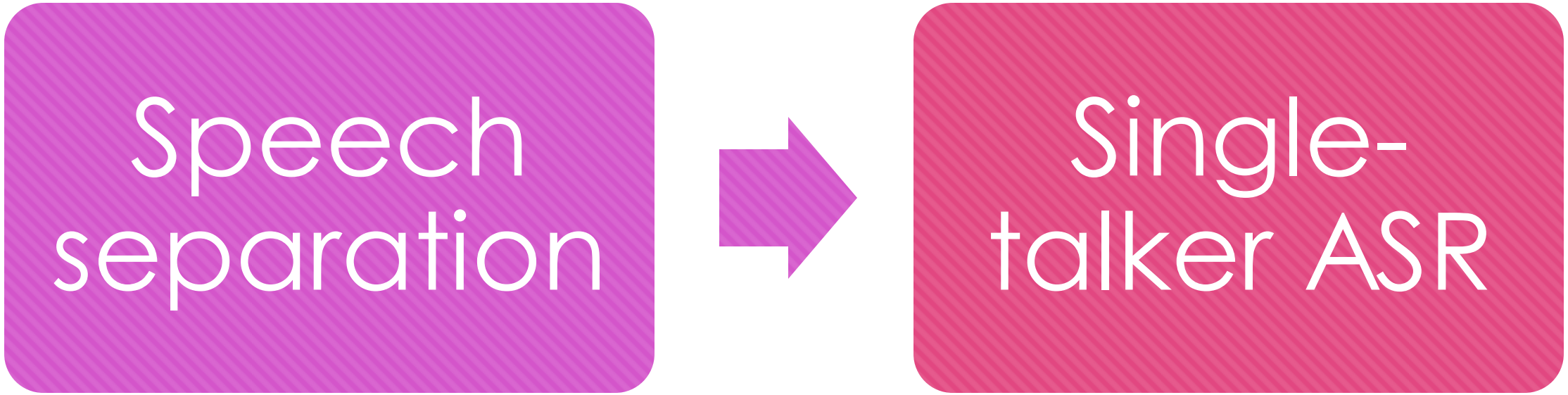
Recording contains >1 speaker

Overlapping speech

Common in meetings, dinner party etc.

The task is to transcribe all the speech in the recording.

The modular approach



Problems with the modular approach

- Modules are independently optimized
- Higher accumulated latency
- Requires engineering effort to maintain

The end-to-end promise

- End-to-end speech recognition has caught up with and outperformed hybrid ASR
  - Connectionist temporal classification (CTC)
  - Recurrent neural network transducer (RNN-T)
  - Attention-based encoder-decoder (AED)


Can we use end-to-end models to perform multi-talker speech recognition?

Previous Work

Hybrid HMM-DNN

Permutation invariant training

(Qian et al., 2018)

Attention-based encoder decoder

Permutation invariant training

(Settle et al., 2018)

Serialized output training

(Kanda et al., 2020)

RNN-Transducer

Masking + Embedding

(Tripathi et al., 2020)

Continuous Streaming Multi-talker ASR

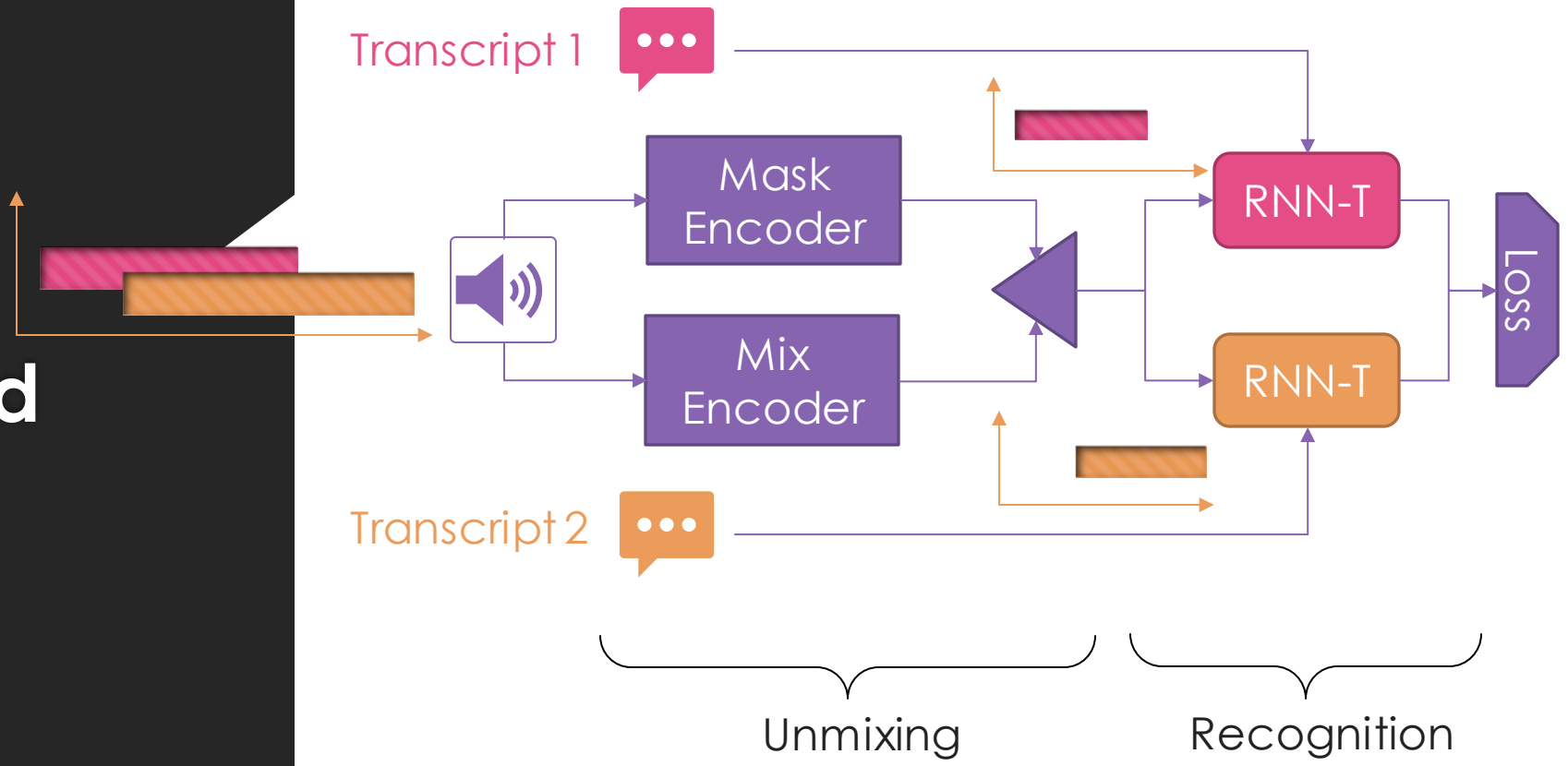
Continuous

- Does not rely on external segmentation for long-form audio

Streaming

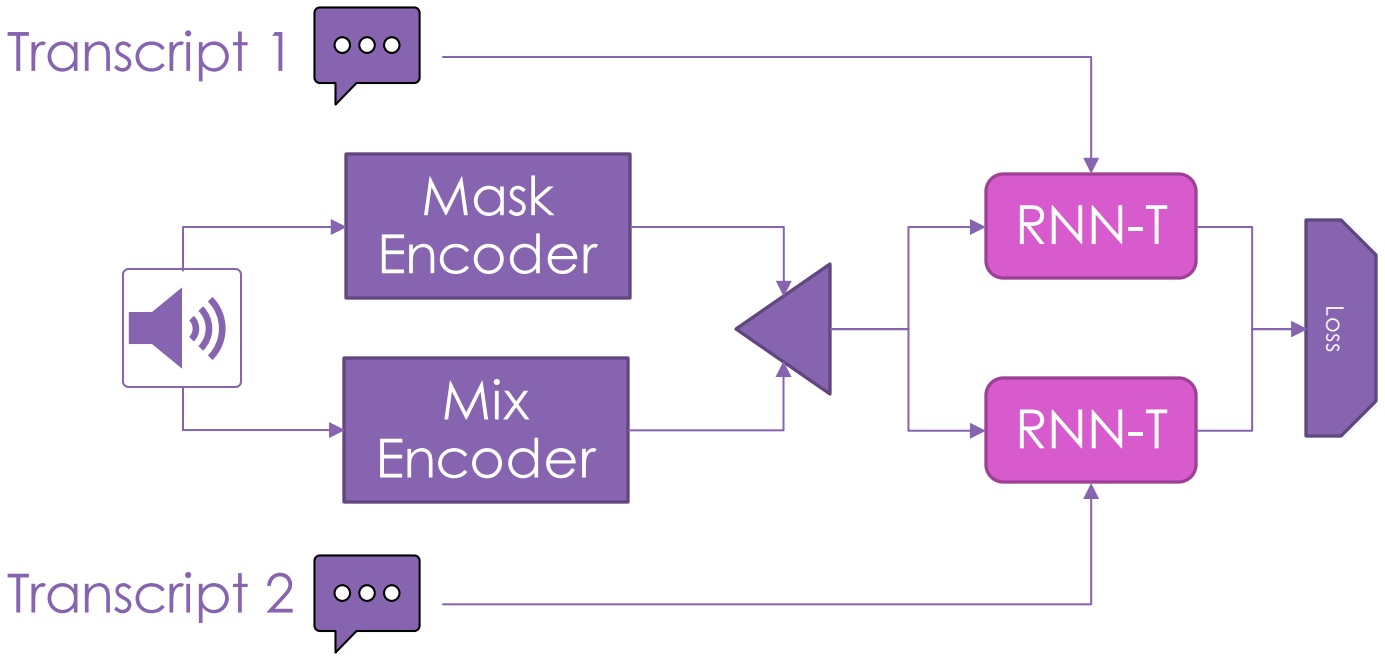
- Should not wait for first speaker to stop before transcribing overlapping speaker

Streaming Unmixing and Recognition Transducer



Lu, L., Kanda, N., Li, J., & Gong, Y. (2021). Streaming End-to-End Multi-Talker Speech Recognition. *IEEE Signal Processing Letters*.

Streaming Unmixing and Recognition Transducer

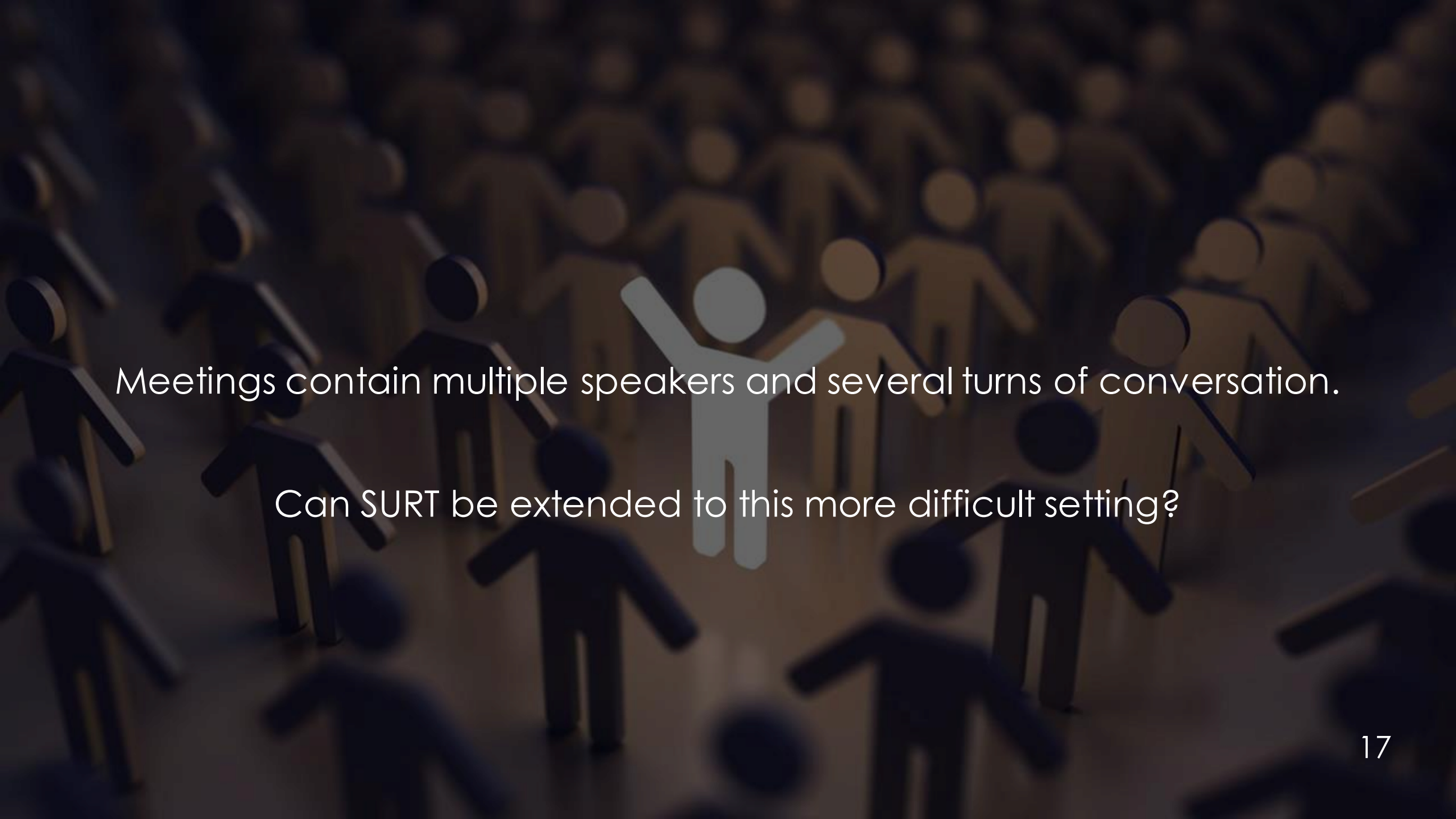


- **Assumption**: at most 2-speaker overlaps
- Each **output channel** transcribes one speaker
- Left context network (like **LSTM**) used in RNN-T
- Mask and Mix encoders are based on **conv nets**

Streaming Unmixing and Recognition Transducer

- Evaluated on 2-speaker **single-turn** sessions
- **Promising results**: 10.8% WER with 150 milliseconds latency

Model	Size	WER
PIT-S2S	161 M	11.1
SURT	81 M	10.8

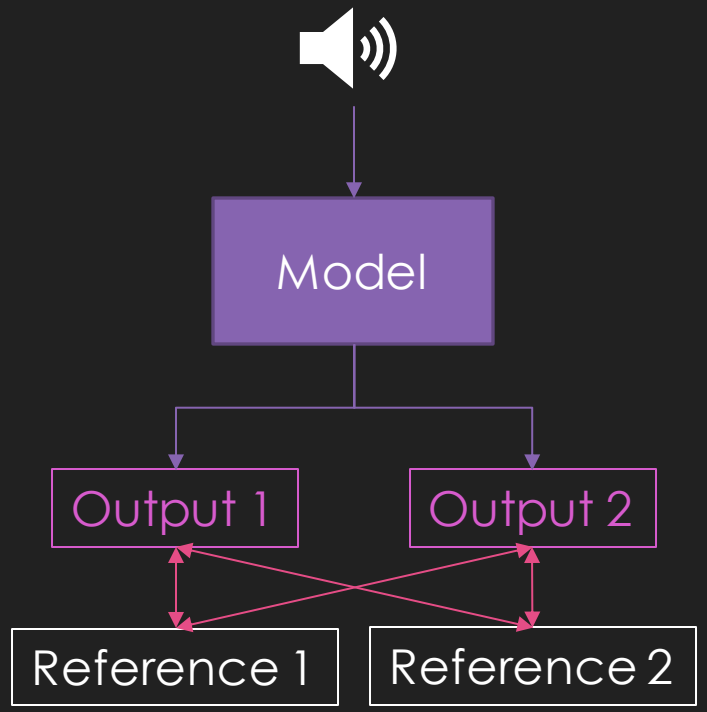


Meetings contain multiple speakers and several turns of conversation.

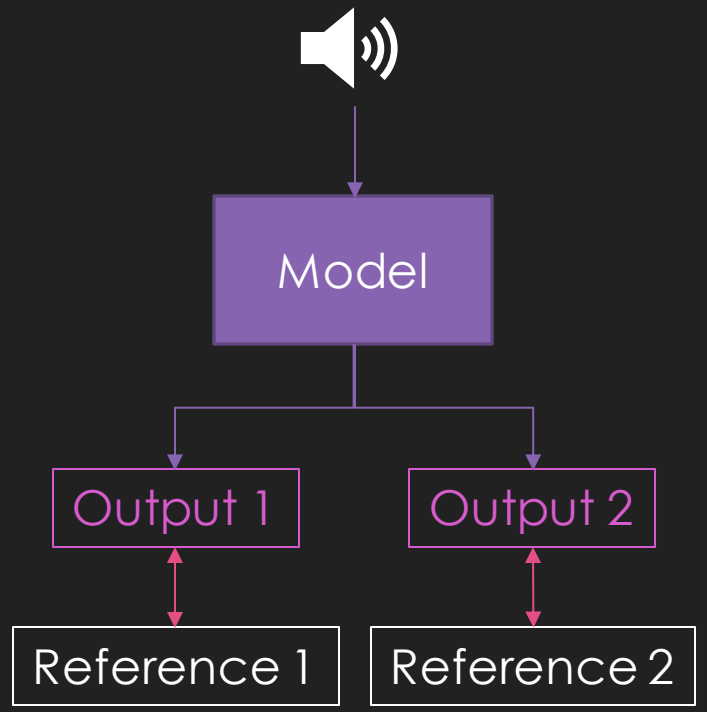
Can SURT be extended to this more difficult setting?

From Single to Multi-turn Evaluation

Heuristic Error Assignment Training (HEAT)

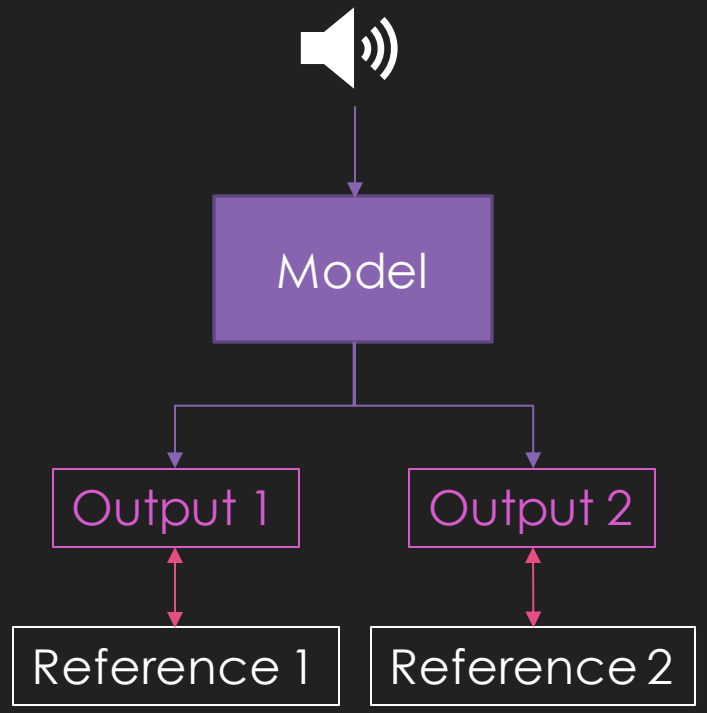


PIT



HEAT

Heuristic Error Assignment Training (HEAT)

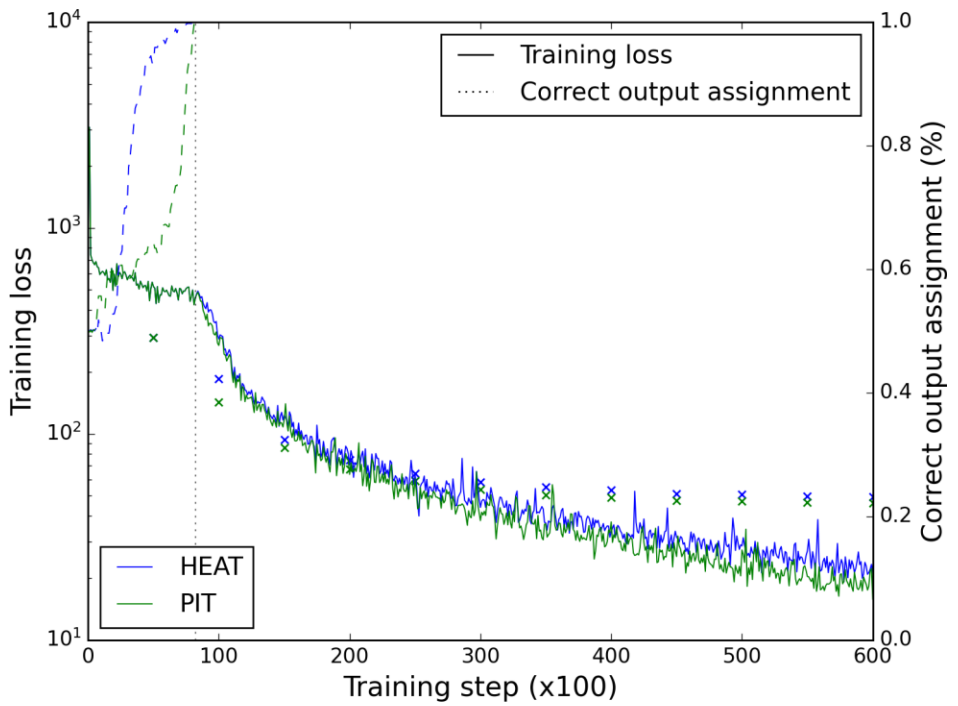


The order is specified by utterance start time

This works because meetings usually contain partial overlaps

HEAT versus PIT

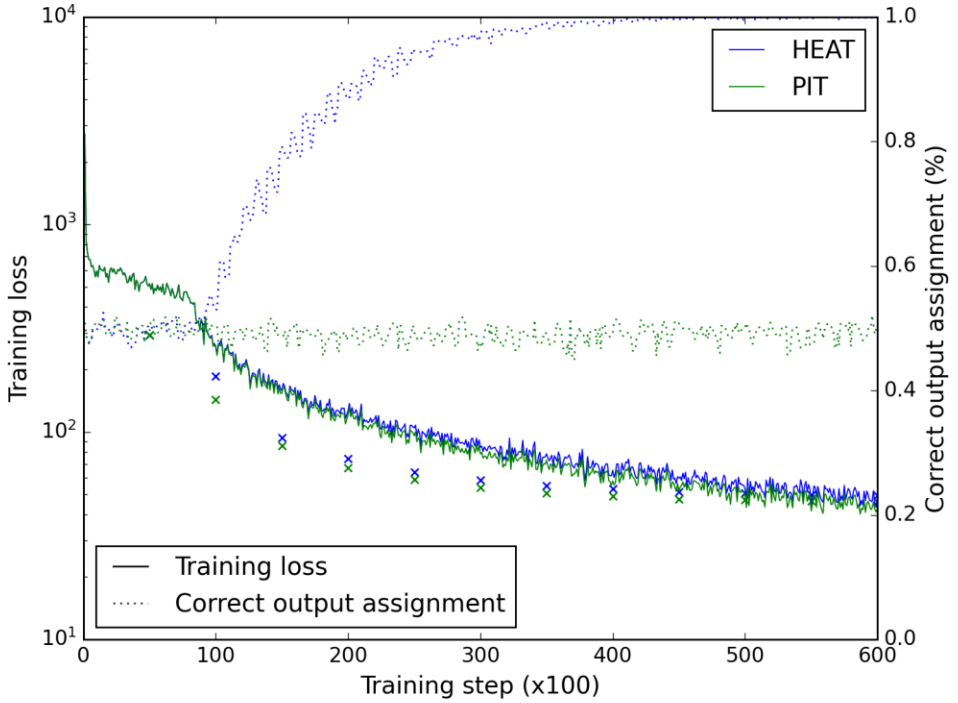
For sessions with non-zero utterance delay, PIT learns the same heuristic as HEAT.



Delay = 2.0 seconds

HEAT versus PIT

For sessions with zero utterance delay, HEAT converges to degenerate solution.



Delay = 0.0 seconds

Why should we use HEAT for multi-turn training?

For partial overlaps, PIT computes the same metric as HEAT

For session with *N* utterances → *N!* permutations: infeasible with RNN-T loss

Multi-turn evaluation data

1

Tier 1 (single-turn)

2 speakers
2 utterances

2

Tier 2 (short multi-turn)

2 speakers
2-4 utterances

3

Tier 3 (long multi-turn)

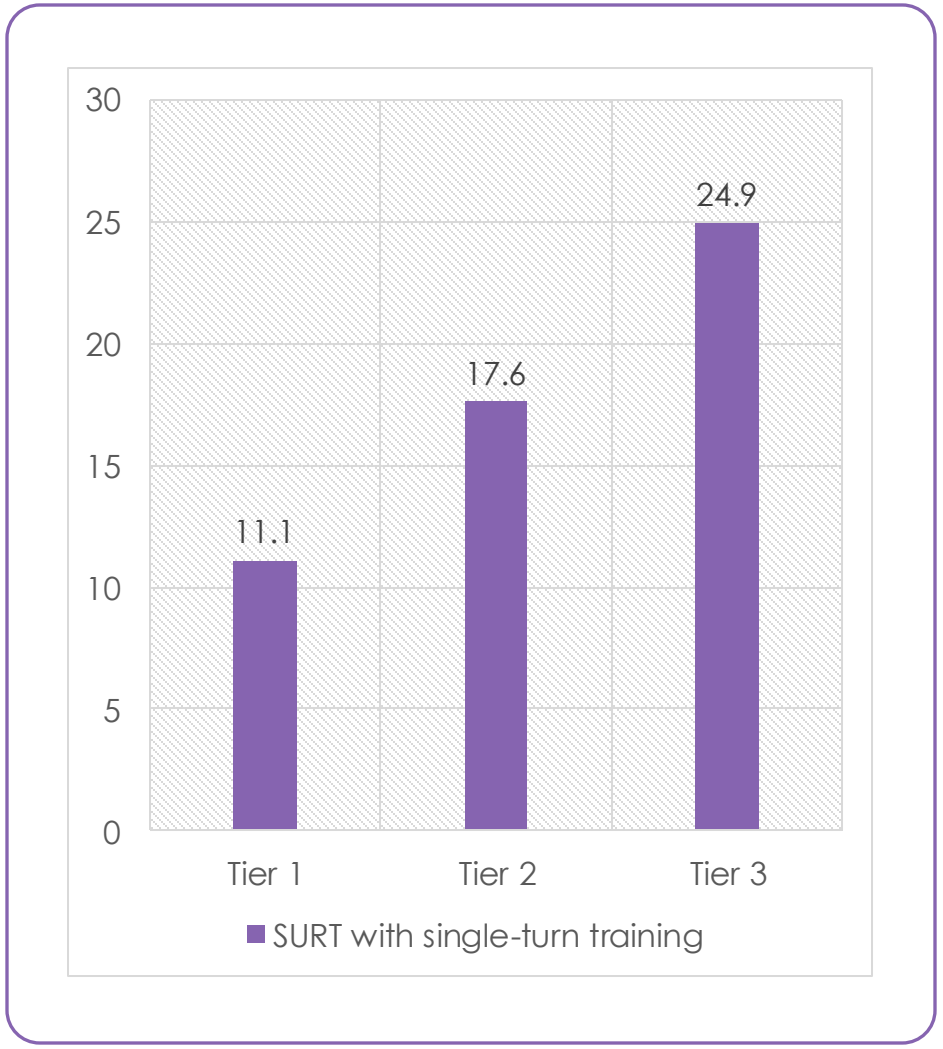
2-4 speakers
6-12 utterances

Sessions generated by mixing LibriSpeech utterances.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

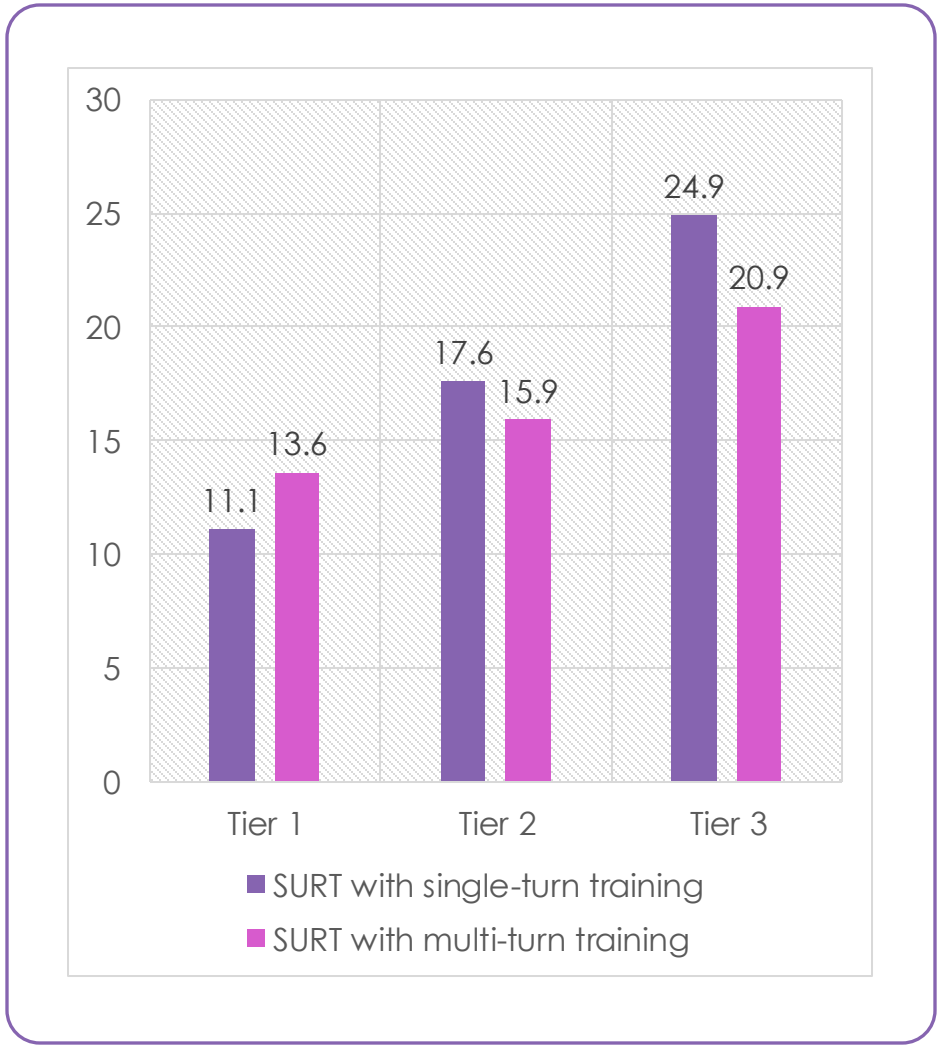
Does the “vanilla” SURT work?

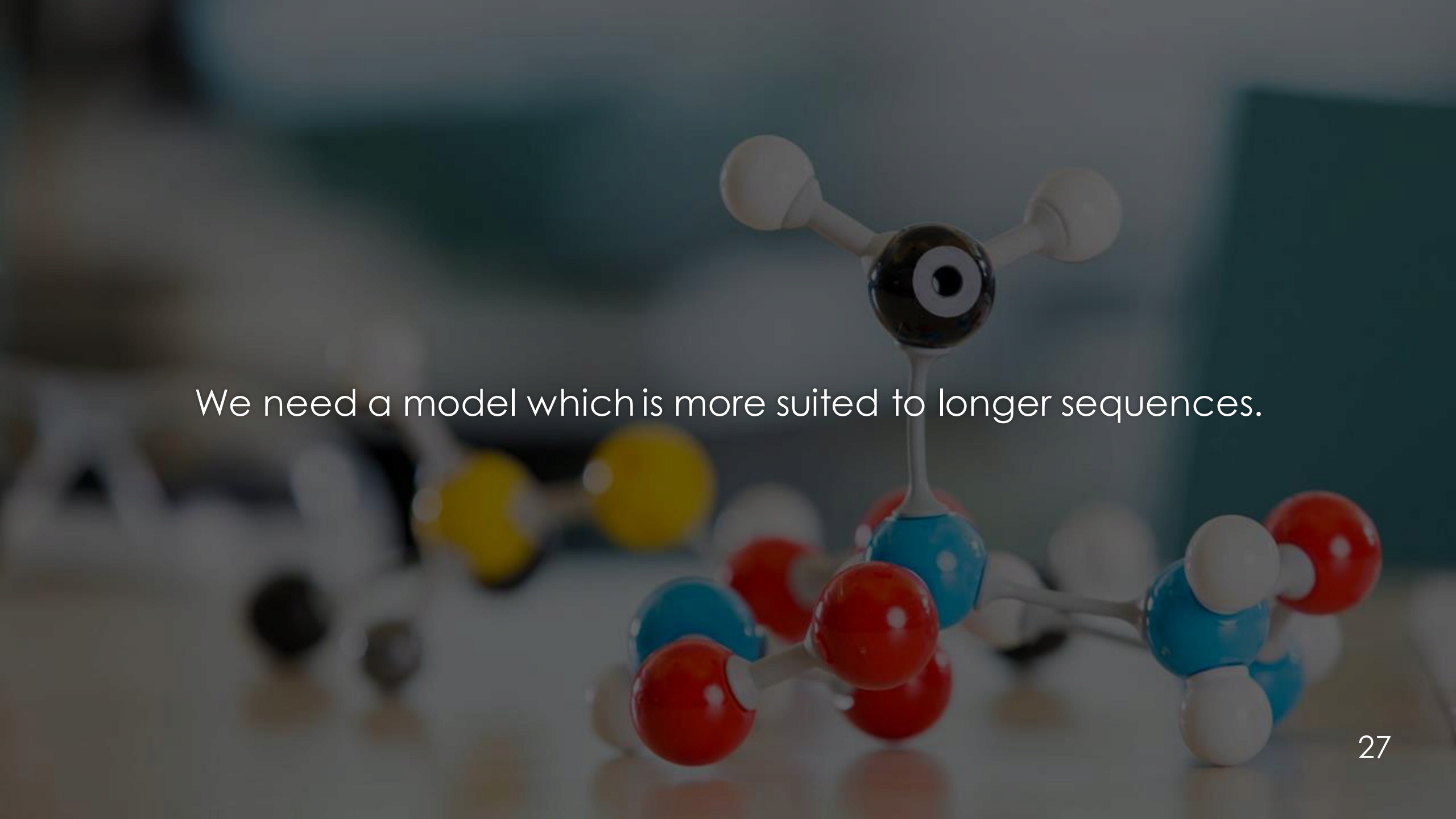
No, it does not work well for multi-turn sessions when trained only on single-turn data.



Does the “vanilla” SURT work when trained on multi-turn data?

- It improves on T2 and T3 (multi-turn) but degrades on T1 (single-turn).
- Training time is doubled since sessions are longer.

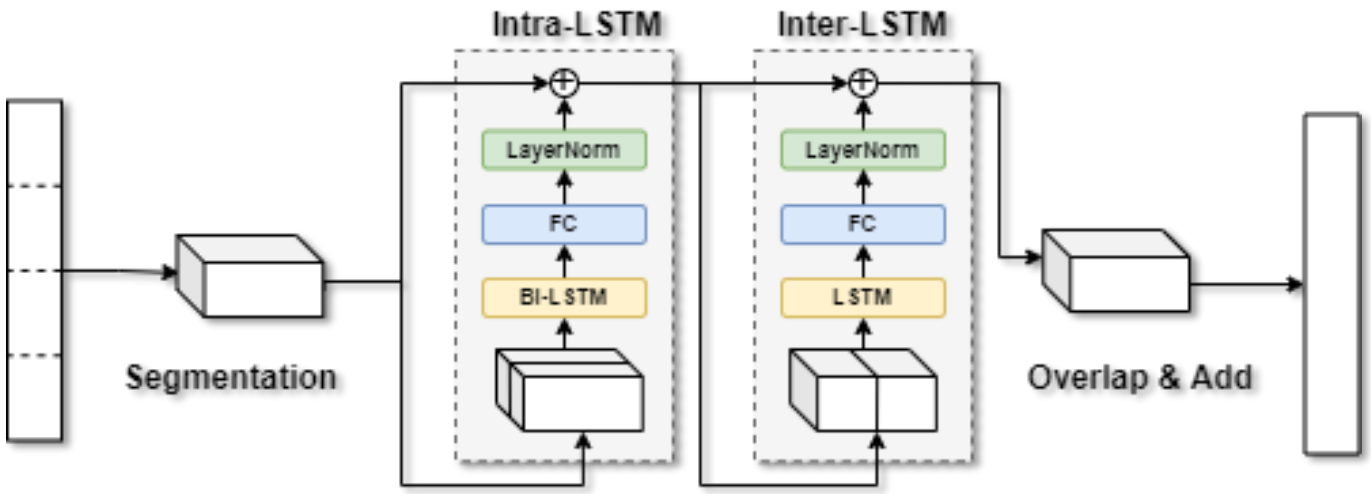




We need a model which is more suited to longer sequences.

Dual-path Modeling (with LSTMs and Transformers)

Dual-path RNN (DP-RNN)

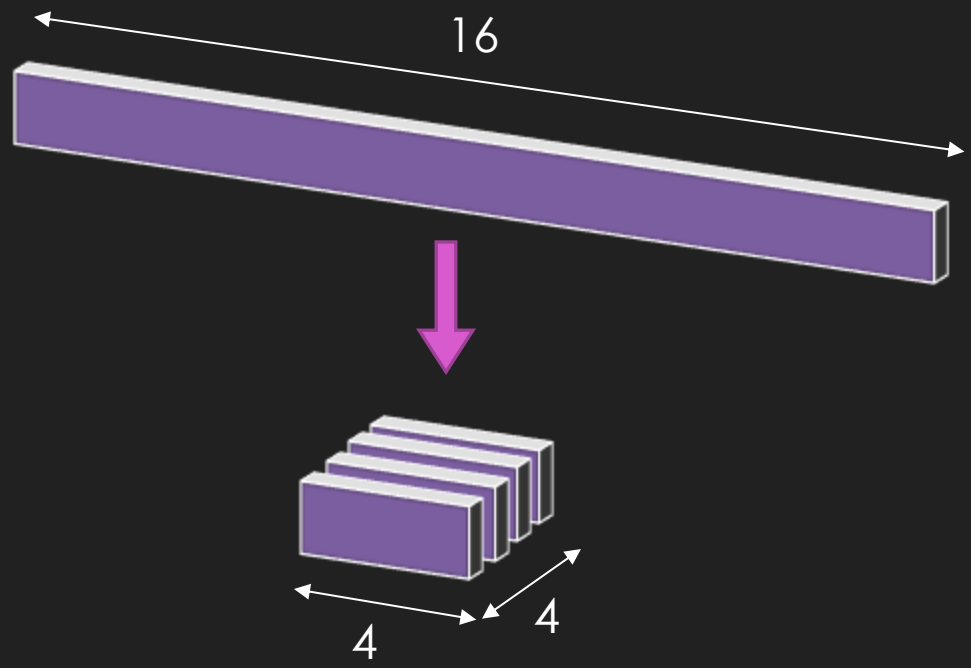


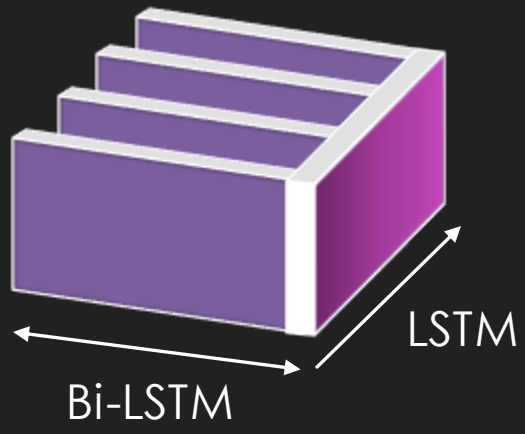
- Proposed for time-domain single-channel speech separation
- Each layer consists of an "**intra**" and an "**inter**" block
- Introduces latency equal to chunk stride

Luo, Y., Chen, Z., & Yoshioka, T. (2020). Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation. *ICASSP 2020*.

Dual-path RNN

- Choose chunk width as **square root** of session length
- Both intra and inter blocks process $\sqrt{N}$ length sequences

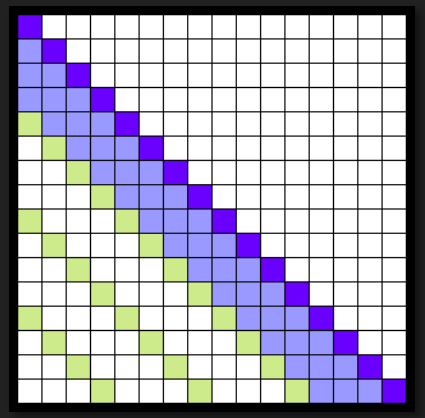




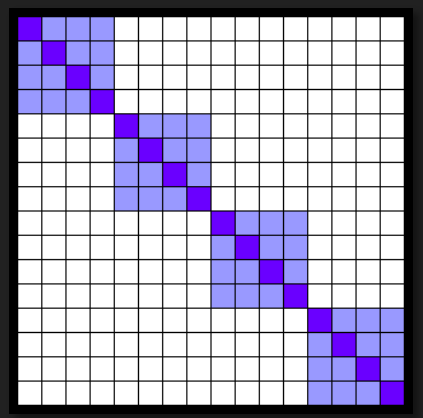
Dual-path Transformer

- LSTM blocks replaced with self-attention blocks
- Full attention for intra-block
- Causal attention for inter-block

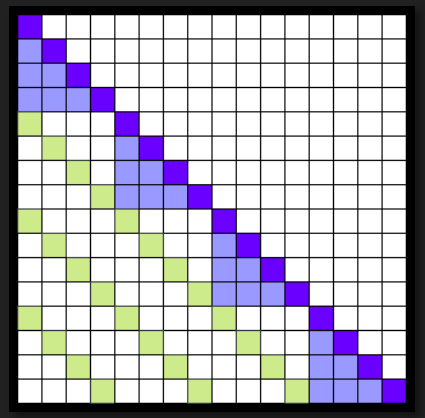
Chen, J., Mao, Q., & Liu, D. (2020). Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation. *ArXiv*.



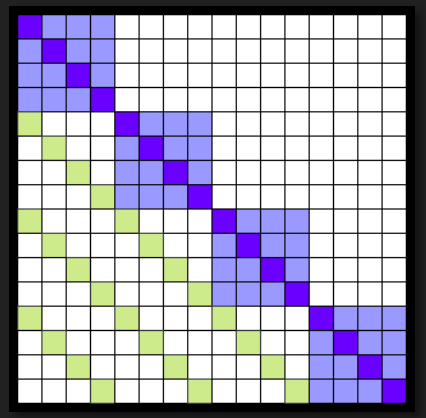
Strided attention
(Child et al., 2019)



Block attention
(Qiu et al., 2020)



Axial attention
(Ho et al., 2019)



DP-Transformer

Analysis of the DP-Transformer

DP-Transformer is a sparse transformer with complexity $O(N\sqrt{N})$

DP-Transformer is a universal function approximator

- Satisfies requirements for universal function approximator (Yun et al., 2020)
  1. Every token attends to itself.
  2. Directed attention graph has a Hamiltonian path through all tokens.
  3. Any token can directly/indirectly access all other tokens (in non-streaming mode).

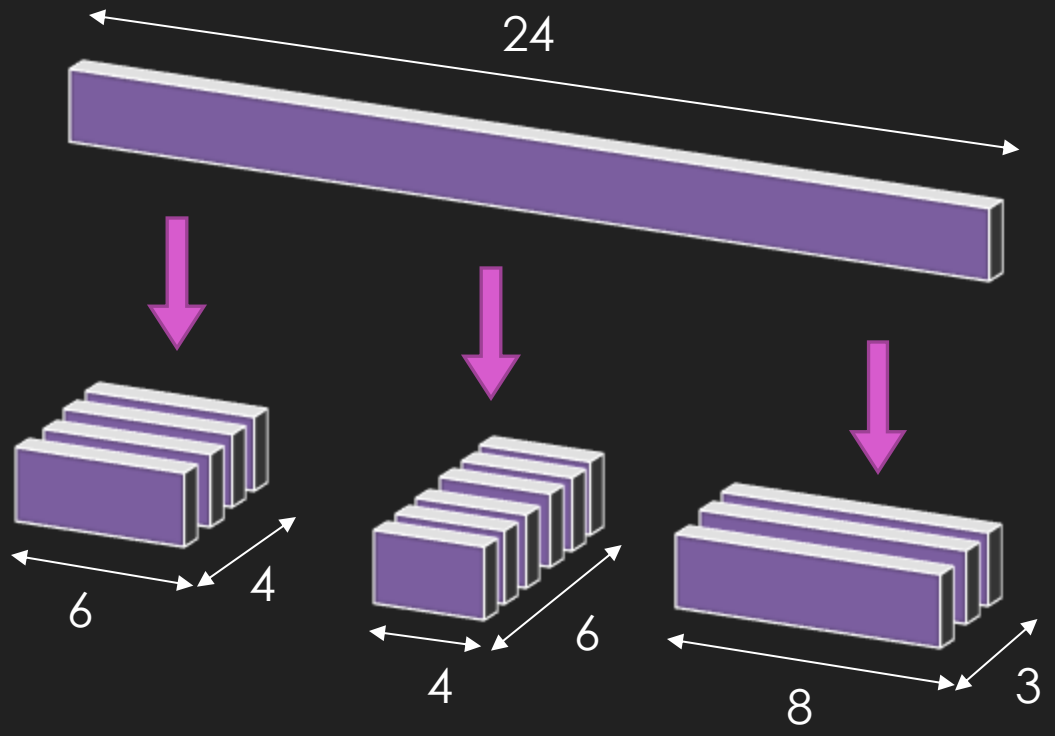
Yun, C., Chang, Y., Bhojanapalli, S., Rawat, A., Reddi, S.J., & Kumar, S. (2020). O(n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers. *ArXiv*.

Chunk width randomization

Training with fixed chunk width may be prohibitive

Randomize chunk width during training

Creates diverse sequence lengths for inter-block

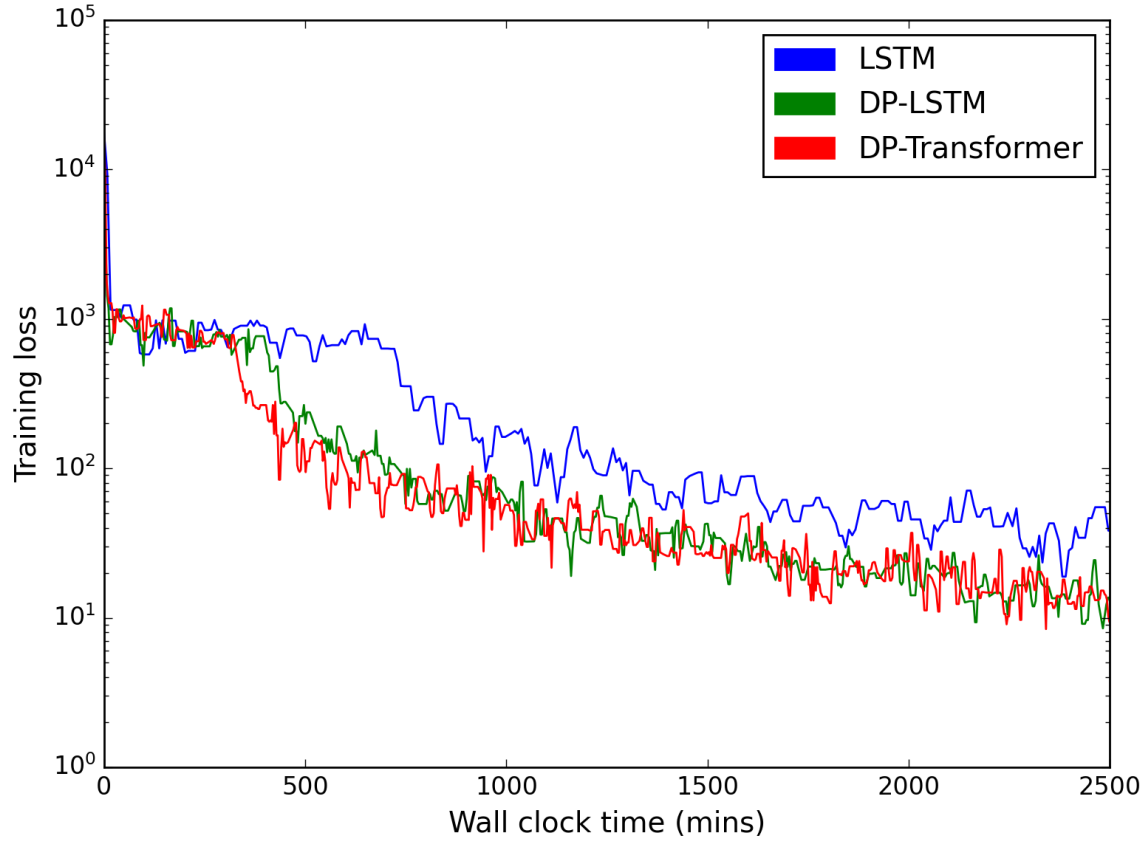


Experimental Results

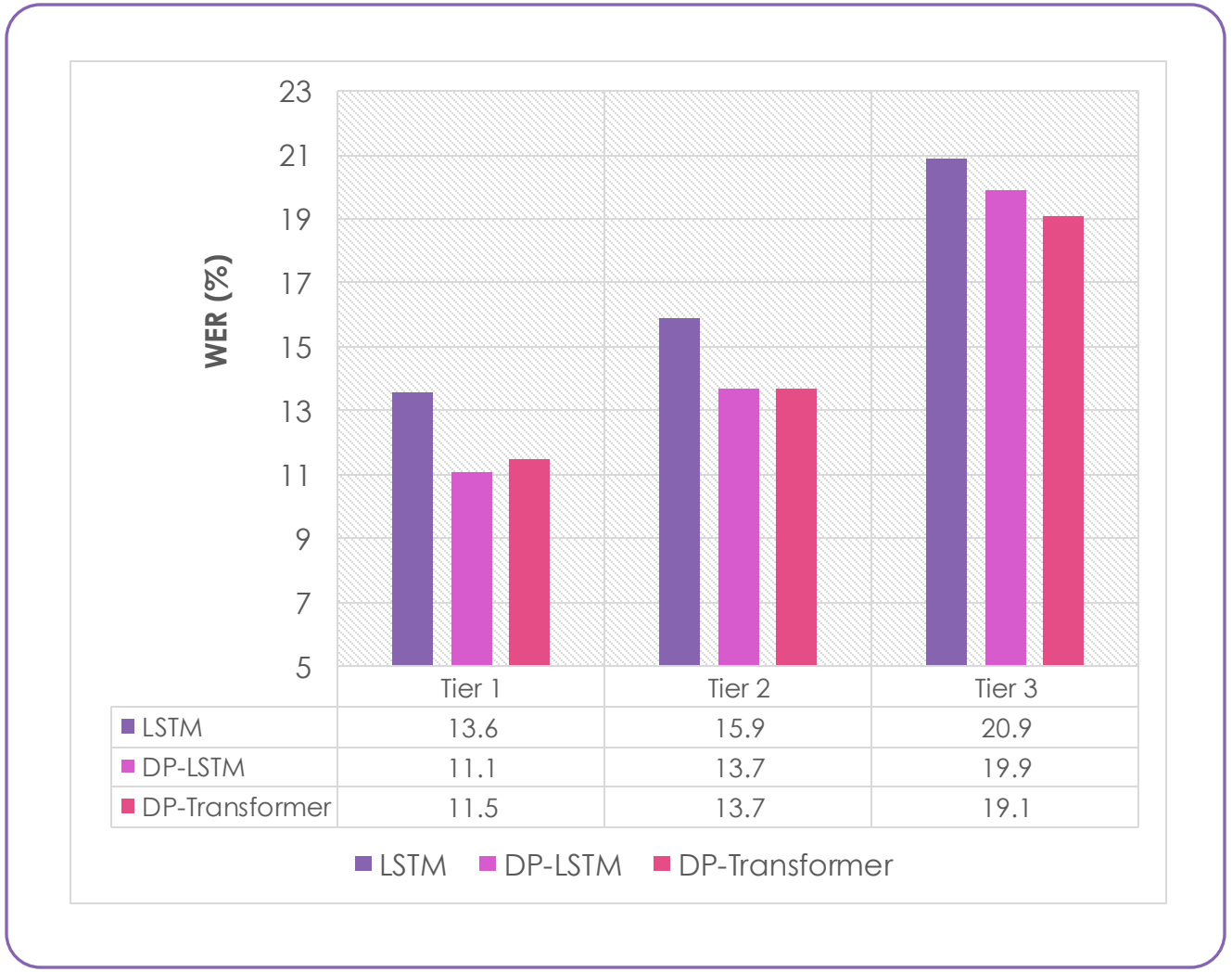
Models

	LSTM	DP-LSTM	DP-Transformer
Encoder	6-layer 1024-dim LSTM	6-layer 512-dim DP-LSTM	12-block DP-Transformer
Decoder	2-layer 1024-dim LSTM	2-layer 1024-dim LSTM	2-layer 1024-dim LSTM
Model size	75.6 M	65.4 M	42.9 M

Dual-path models converge **faster** and to a **better** minima

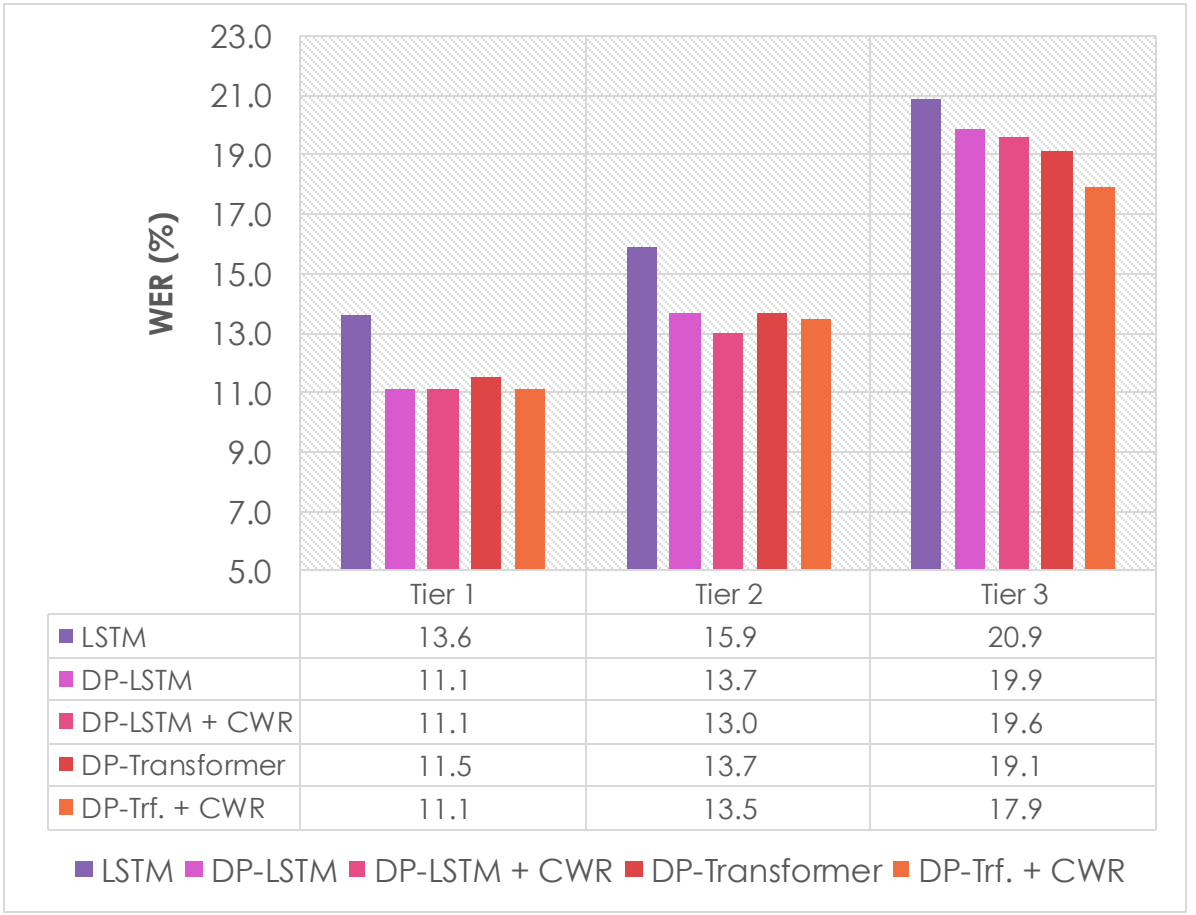


Multi-turn evaluation



Multi-turn evaluation

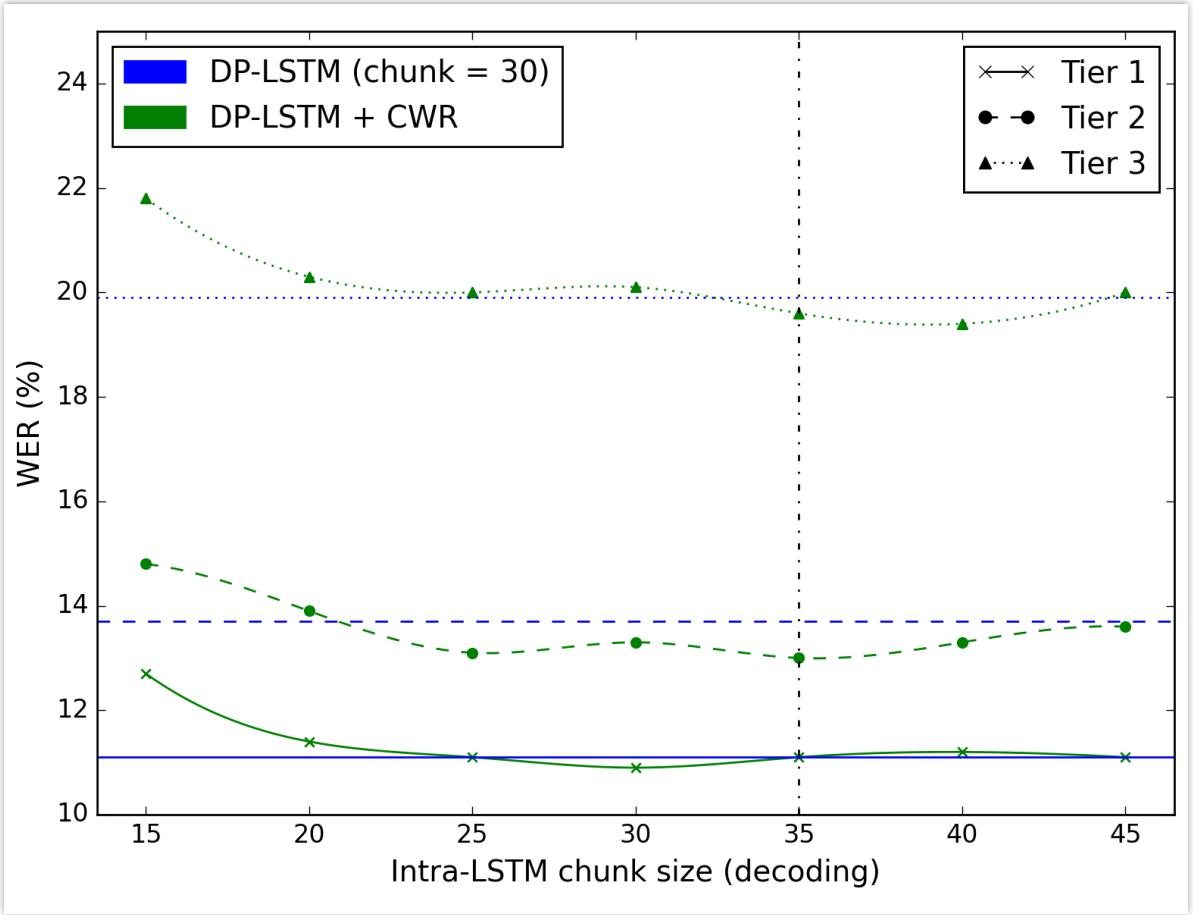
Chunk width randomization (CWR) improves performance across tiers



	Tier 1	Tier 2	Tier 3
LSTM	13.6	15.9	20.9
DP-LSTM	11.1	13.7	19.9
DP-LSTM + CWR	11.1	13.0	19.6
DP-Transformer	11.5	13.7	19.1
DP-Trf. + CWR	11.1	13.5	17.9

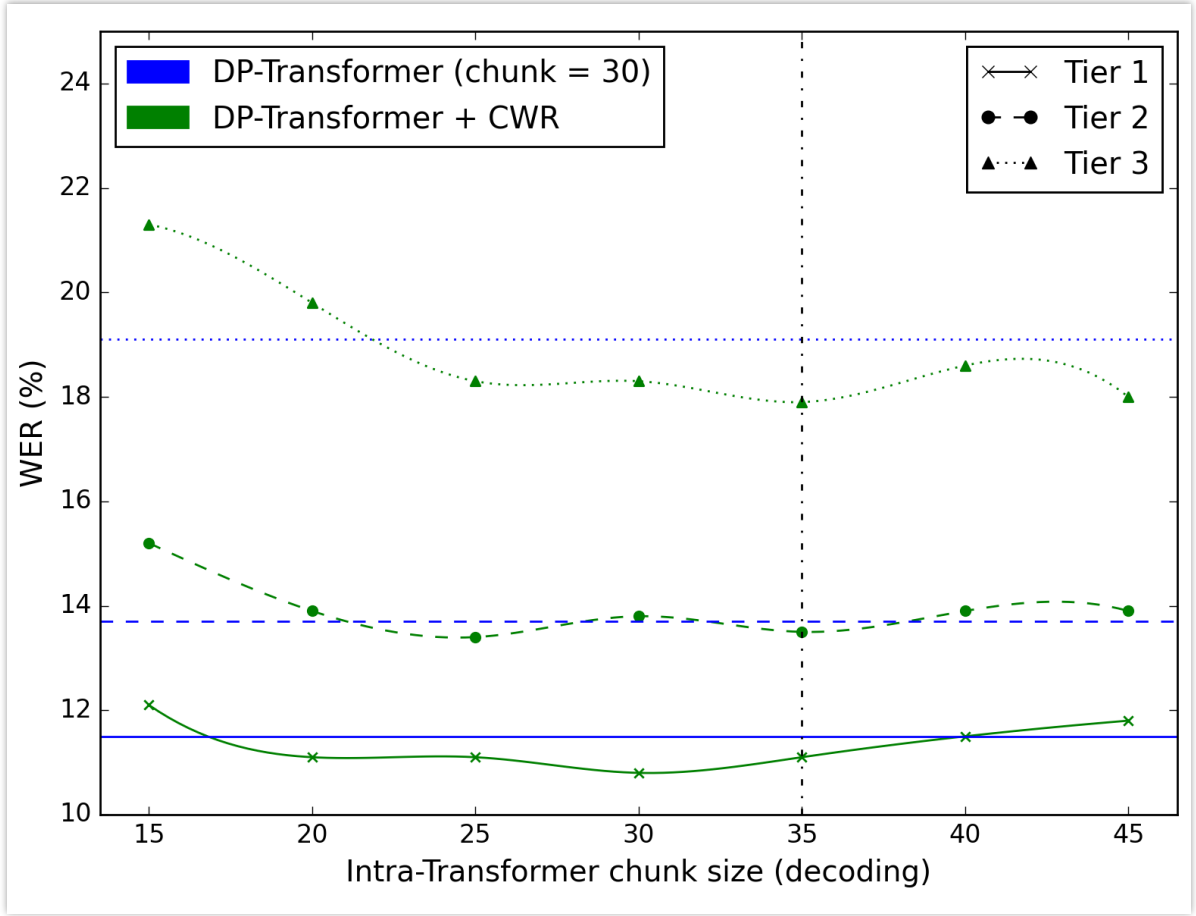
Accuracy vs. latency

- Smaller chunk width → lower latency at the cost of WER degradation
- Large chunk width → not enough context for inter-block

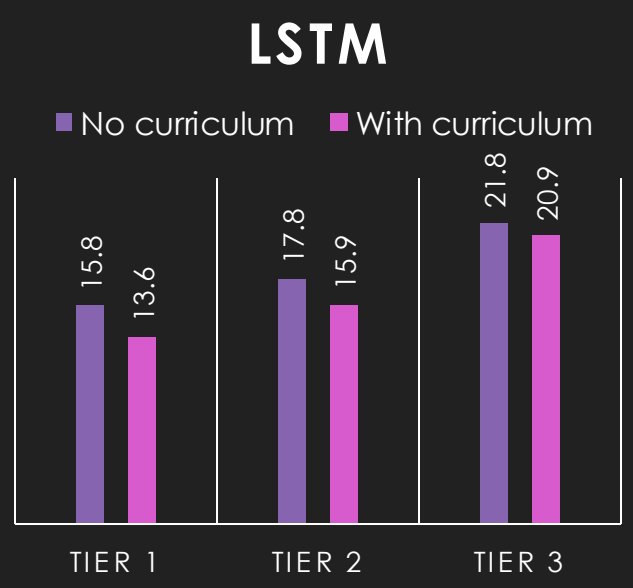


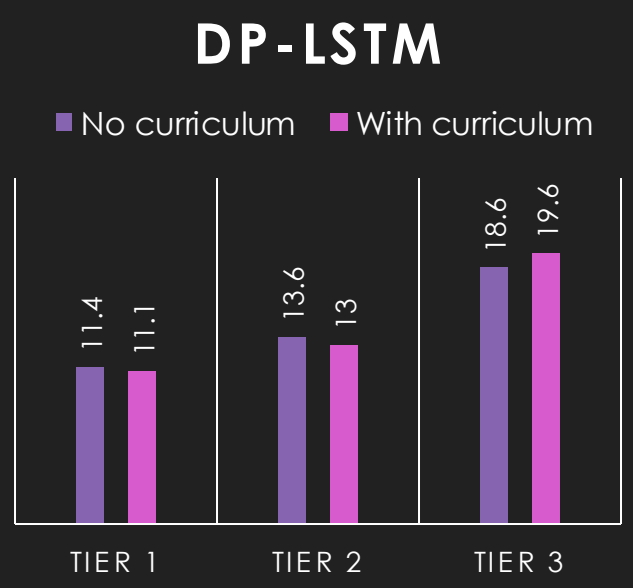
Accuracy vs. latency

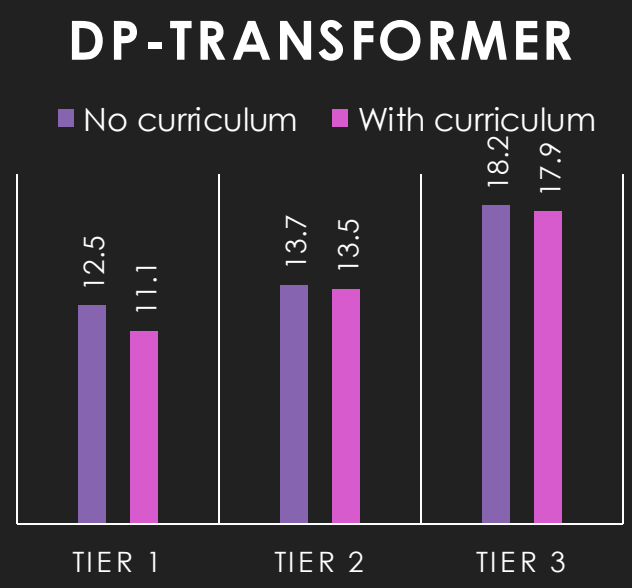
- Similar observations for DP-Transformer with chunk width randomization
- Both models better than the best fixed chunk counterparts

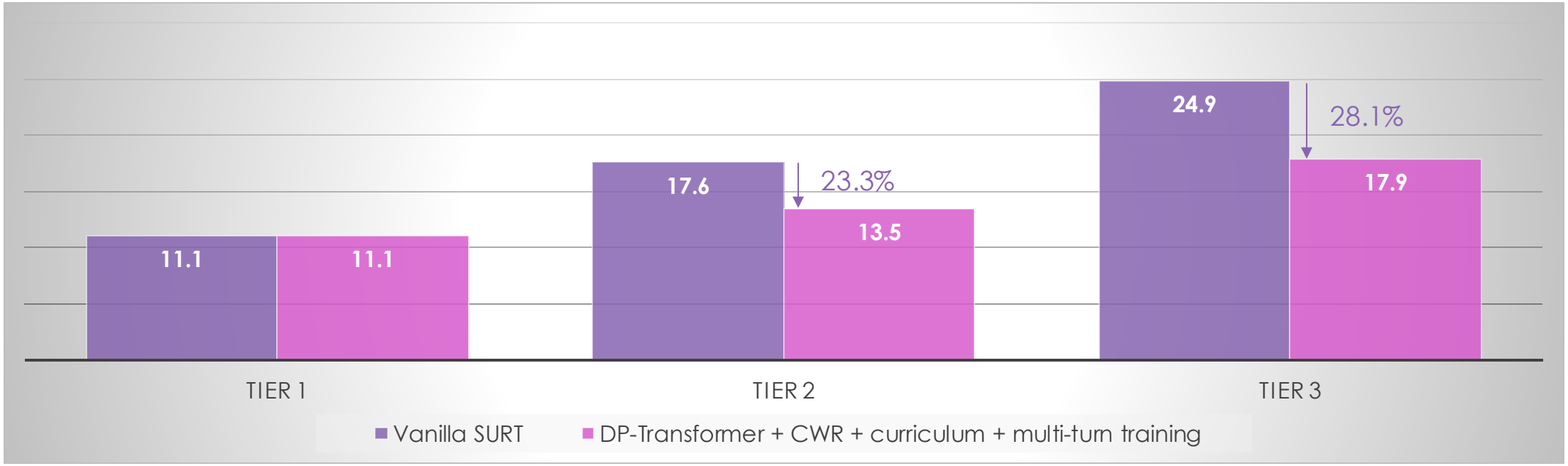


Importance of curriculum learning

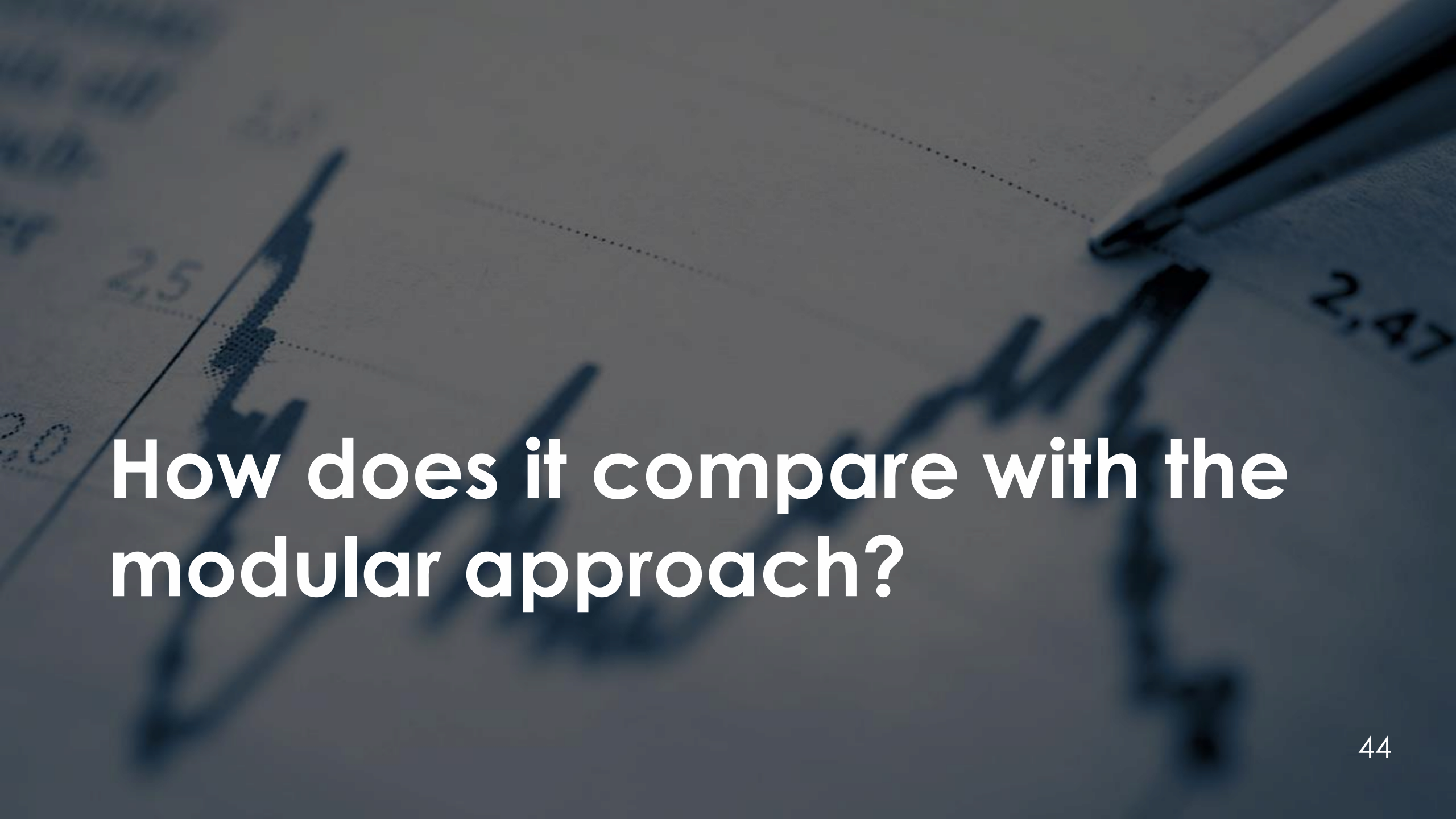








Putting it all together



How does it compare with the modular approach?

Modular systems (CSS + ASR)

System	CSS Front-end	ASR Back-end	Language Model	Streaming
BLSTM CSS + Hybrid ASR	3-layer 1024-dim BLSTM	3-layer 512-dim BLSTM	4-gram	No
Conformer CSS + E2E ASR	18-block Conformer	24-layer encoder + 12-layer decoder	RNNLM fusion + rescoring	No

Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., Luo, Y., Wu, J., & Li, J. (2020). Continuous Speech Separation: Dataset and Analysis. *ICASSP 2020*.

Chen, S., Wu, Y., Chen, Z., Li, J., Wang, C., Liu, S., & Zhou, M. (2021). Continuous Speech Separation with Conformer.*ICASSP 2021*..

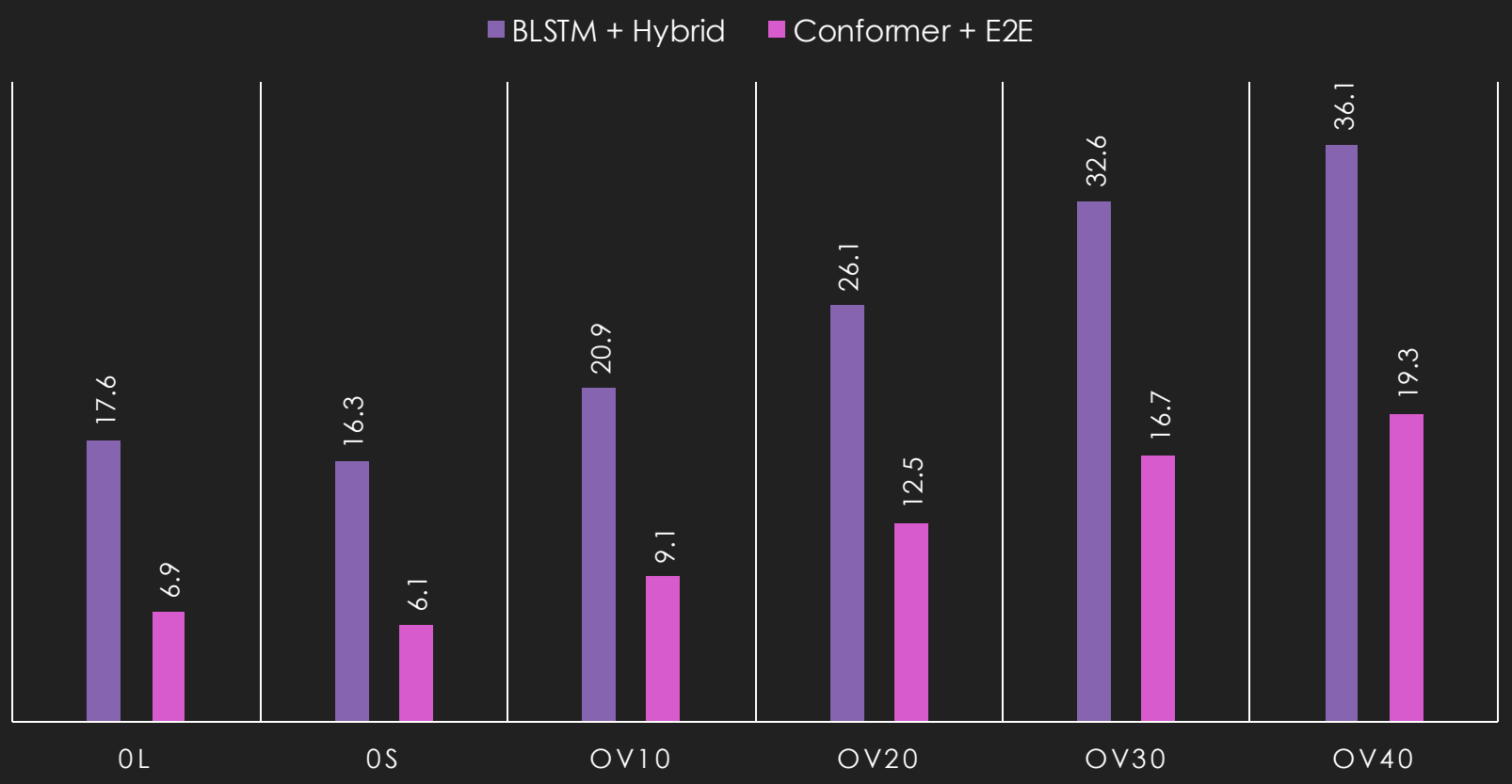
LibriCSS

Obtained from mixing LibriSpeech utterances and replaying in meeting room

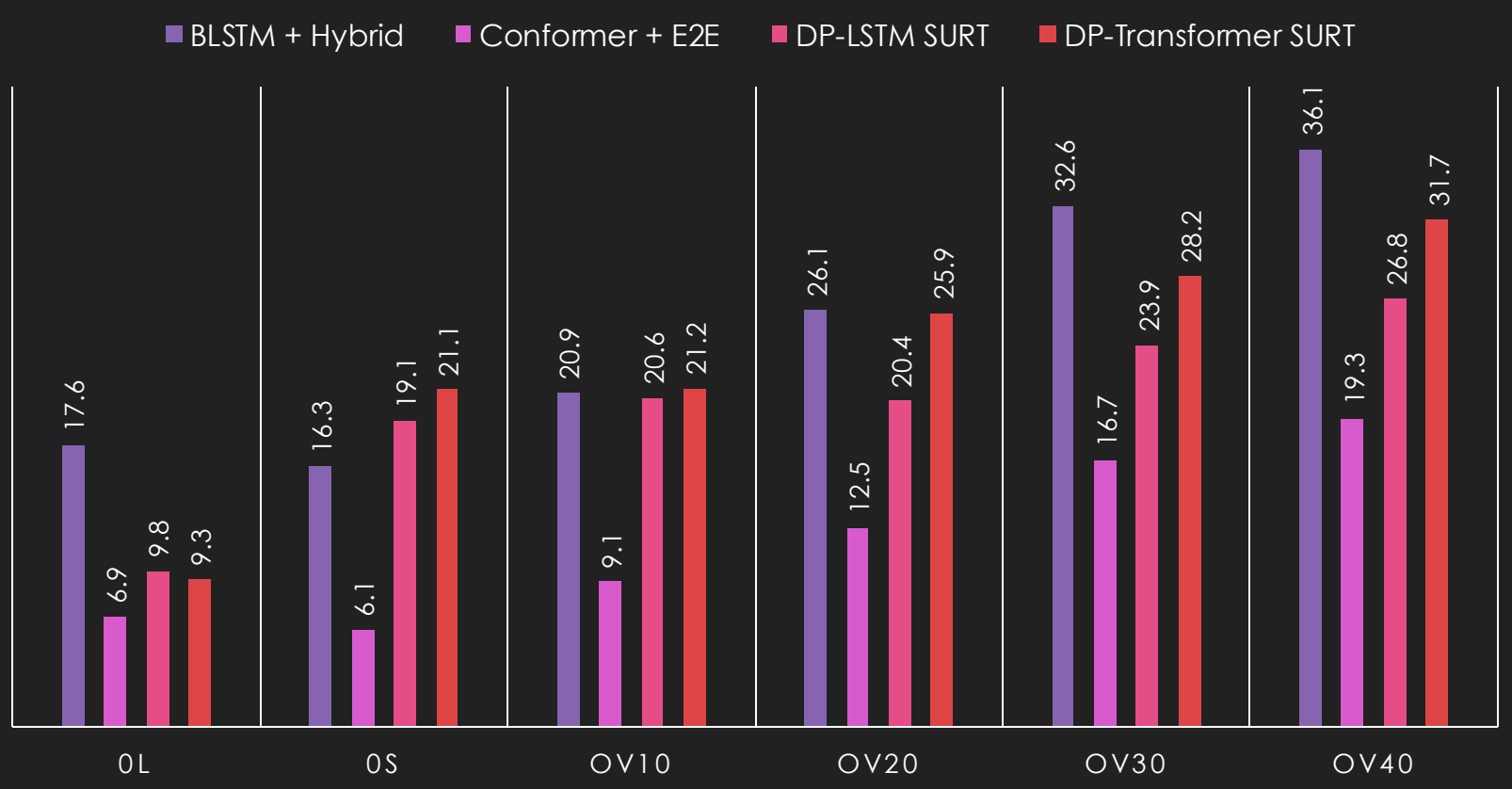
10-min long mini-sessions with specified overlaps between 0% and 40%

Evaluation in single-channel setting

SURT is competitive* with modular systems



SURT is competitive* with modular systems



Sources of Error

Leakage

- Both channels predict output in single-speaker region → insertion errors

Omission (0S)

- No channel transcribes entire utterance → deletion errors

Summary and Next Steps

Key Take-aways

SURT is a viable model for **continuous streaming** multi-talker ASR

Encoder architectures suitable for **long sequence modelling** provide WER gains with fast training and inference

What's next for SURT?

Mitigating problems of leakage and omission

Speaker-attributed transcription

Utterance segmentation

Joint speaker identification

Lu, L., Kanda, N., Li, J., & Gong, Y. (2021). Streaming Multi-talker Speech Recognition with Joint Speaker Identification. *INTERSPEECH 2021*.

Thank You for Listening!

Find me on:

Twitter: @rdesh26

LinkedIn: rdesh26

Webpage: desh2608.github.io

Acknowledgements: Naoyuki Kanda, Lyle Corbin