# A Streaming On-Device End-to-End Model Surpassing Server-Side Conventional Model Quality and Latency

**Presenters**: Tara N. Sainath, Yanzhang (Ryan) He

**Authors:**
Tara N. Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, Shuo-yiin Chang, Wei Li, Raziel Alvarez, Zhifeng Chen, Chung-Cheng Chiu, David Garcia, Alex Gruenstein, Ke Hu, Minho Jin, Anjuli Kannan, Qiao Liang, Ian McGraw, Cal Peyser, Rohit Prabhavalkar, Golan Pundak, David Rybach, Yuan Shangguan, Yash Sheth, Trevor Strohman, Mirko Visontai, Yonghui Wu, Yu Zhang, Ding Zhao

ICASSP 2020

Google

# Outline

- **Motivation**
- Model Architecture
- Quality Improvements
  - Multi-Domain Data
  - Robustness to Accents
  - Learning Rates
- Latency Improvements
  - Joint RNN-T Endpointer
  - LAS Rescoring
- Experiments and Results

# Motivation

- E2E models are attractive for on-device [Y. He, T.N. Sainath *et al.* 2018][J. Li *et al.* 2019]
- Running ASR models on-device is **challenging**
  - Streaming recognition
  - Low latency
  - Quality comparable to server-side conventional model

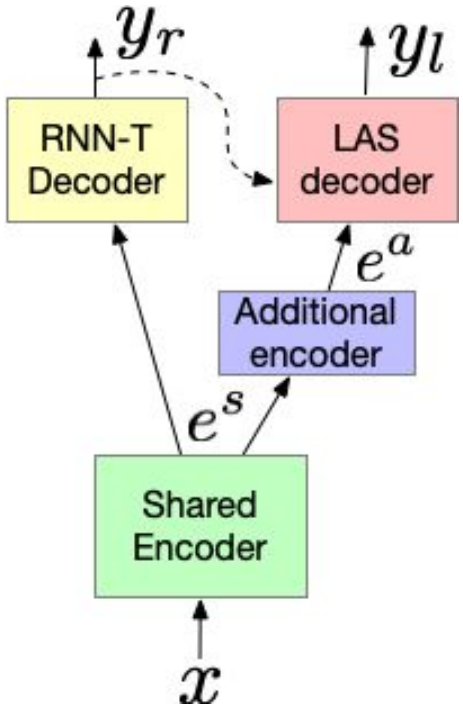
- Goal: Present a streaming E2E model that surpasses server-side conventional model in terms of both **quality** and **latency**

# Outline

- Motivation
- **Model Architecture**
- Quality Improvements
  - Multi-Domain Data
  - Robustness to Accents
  - Learning Rates
- Latency Improvements
  - Joint RNN-T Endpointer
  - LAS Rescoring
- Experiments and Results

# Two-Pass Model Overview

- Model described in: T.N. Sainath, R. Pang *et al.*, "Two-Pass End-to-End Speech Recognition", Proc. Interspeech, 2019.
- Training:
  - 1st-pass: shared-encoder + RNN-T decoder
  - 2nd-pass: additional encoder + LAS decoder
- Decoding:
  - 1st-pass RNN-T produces hypotheses
  - LAS decoder rescores hypotheses from RNN-T



# Outline

- Motivation
- Model Architecture
- **Quality Improvements**
  - **Multi-Domain Data**
  - **Robustness to Accents**
  - **Learning Rates**
- Latency Improvements
  - Joint RNN-T Endpointer
  - LAS Rescoring
- Experiments and Results

# Multi-Domain Data

- E2E models are trained on audio-text pairs, which is a *fraction* of data compared to a conventional language model
- Incorporate data from a variety of domains to increase vocabulary: Search, Farfield, Telephony, YouTube [A. Narayanan *et al.*, ASRU 2019]
- Mixing domains creates issues
  - Transcription convention ("$100" *versus* "one hundred dollars")
  - Different amounts of background/foreground speech
  - Different amounts of background noises
- Solution: Feed a 1-hot vector of the domain-id to the encoder (shown to be successful for multi-lingual/dialect ASR [B. Li *et al.*, ICASSP 2018])

# Robustness to Accents

- A common way conventional ASR systems handle accents is through a lexicon with multiple pronunciations
- Our E2E models emit wordpieces and decide *a-priori* how to break-up words based on training data
- To improve E2E accent robustness, we include additional English training data from Australia, New-Zealand, United Kingdom, Ireland, India, Kenya, Nigeria and South Africa (en-X)
- To handle spelling differences (color *versus* colour), all data is transliterated to US English spelling before training.
- **Goal:** Train a single E2E model to be robust to multiple en-X accents without knowing which accent the utterance comes from during inference. The transcripts are all recognized as US English spelling.

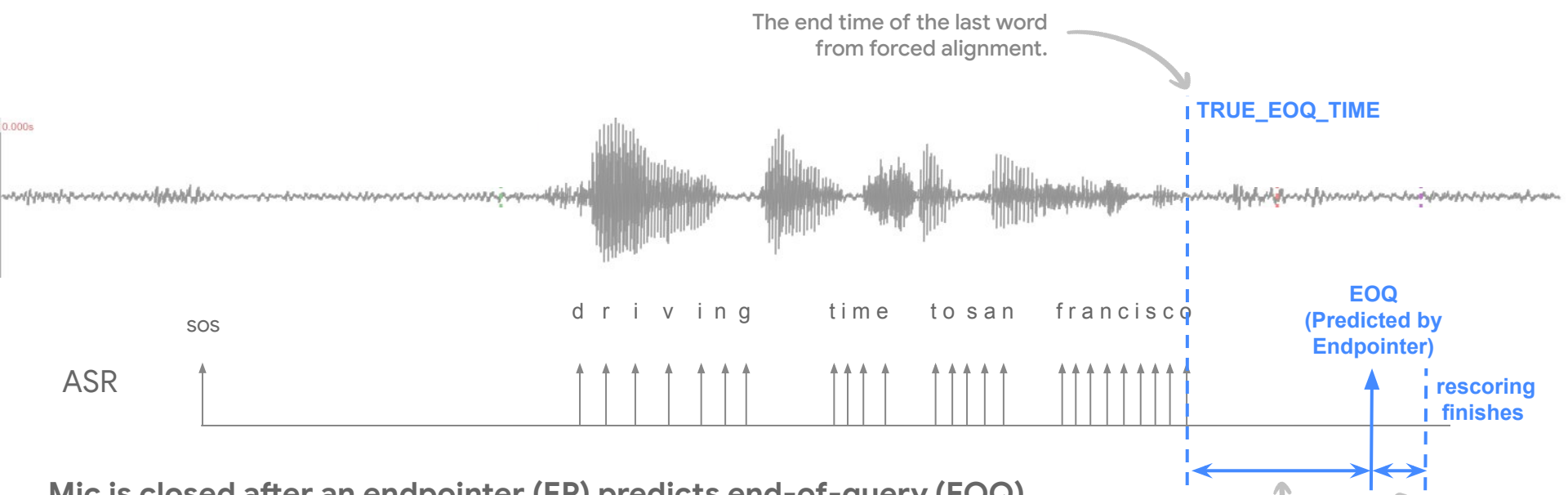
# Learning Rates

- With increased multi-domain and en-X data, we change the learning rate to ramp up and then be constant

- To help with training stability, we maintain  an exponential moving average (EMA) of the weights during training

# Outline

- Motivation
- Model Architecture
- Quality Improvements
  - Multi-domain Data
  - Robustness to Accents
  - Learning Rates
- **Latency Improvements**
  - **Joint RNN-T Endpointer**
  - **LAS Rescoring**
- Experiments and Results



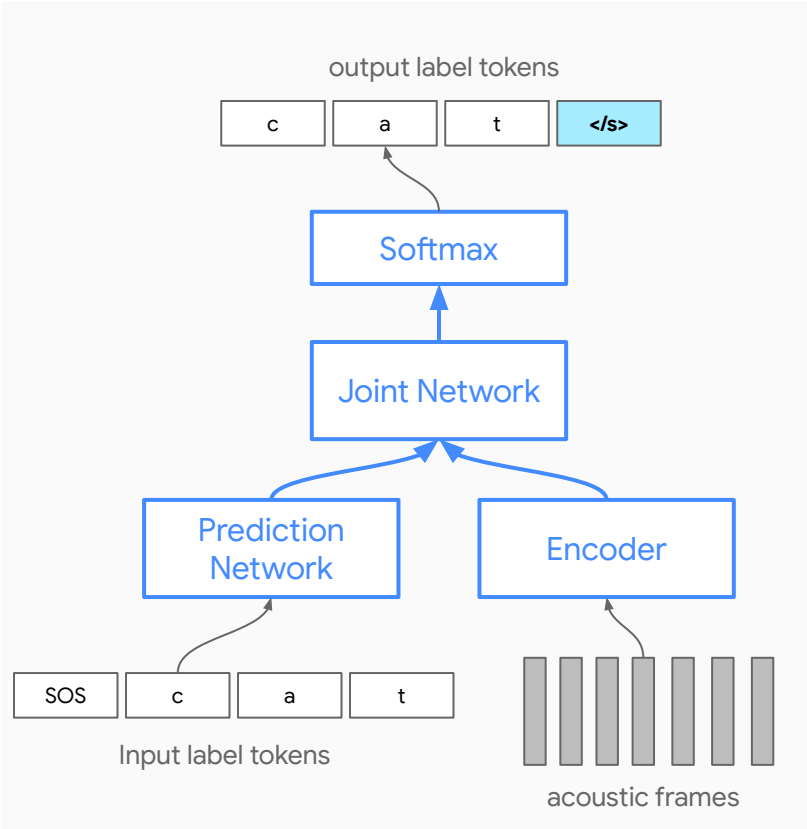

**Mic is closed after an endpointer (EP) predicts end-of-query (EOQ).**
More aggressive endpointer could cut off trailing words.

The ***audio time*** difference between when the user finishes speaking and when the endpointer generates EOQ. **Endpointer Latency**

The ***computation time*** a 2nd-pass rescorer needs to take after EOQ is generated. **Rescoring Latency**

# Joint RNN-T Endpointer

- Incorporate endpointing decision into RNN-T by having it predict </s>, i.e. the end-of-query (EOQ) token.
- The time for emitting </s> is constrained to be as close to the last word as possible



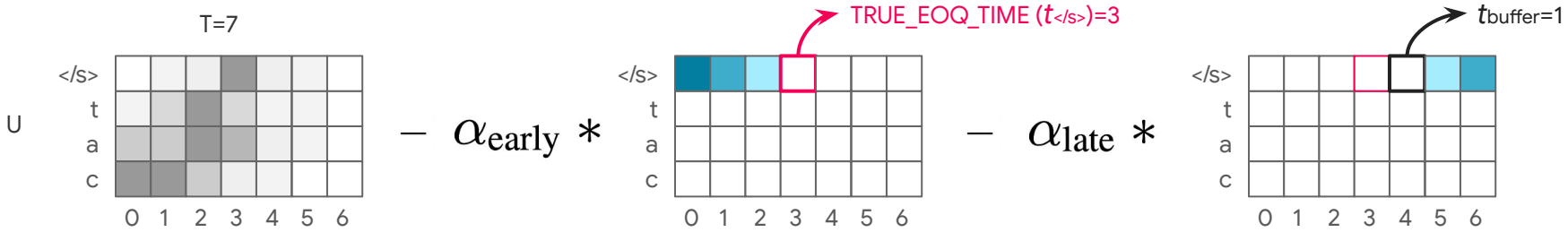
# </s> Penalties in Training

To help the model predict </s> (EOQ) as close to the end of the last word as possible.

Original U*T Matrix for RNN-T Log Posteriors

**Early Penalty**

**Late Penalty**




$$\log \mathrm{P}_{\text{RNN-T}}(y_U|\mathbf{x}_t) \quad - \quad \max(0,\ \alpha_{\text{early}} * (t_{</s>} - t)) \quad - \quad \max(0,\ \alpha_{\text{late}} * (t - t_{</s>} - t_{\text{buffer}}))$$

( $y_U$ is the last token, i.e. </s>)

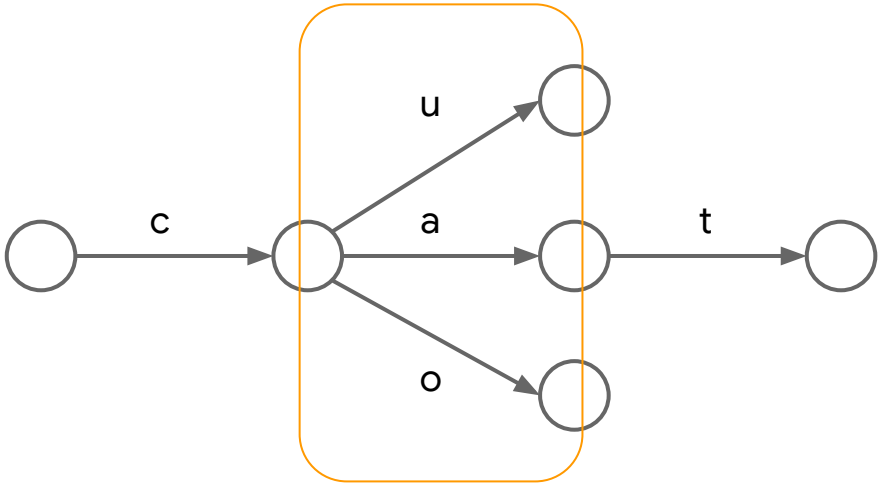


Bo Li *et al.*, "Towards Fast and Accurate Streaming End-to-End ASR", ICASSP 2020

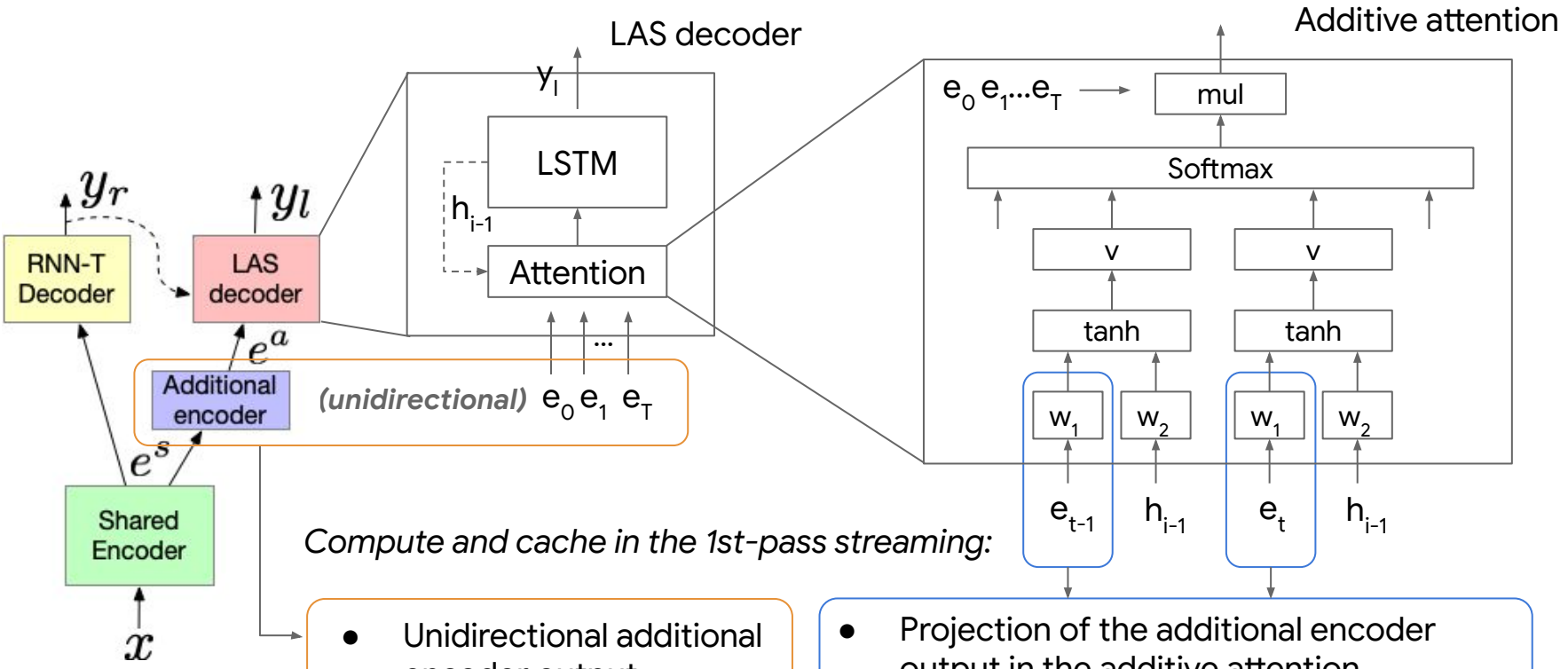
# LAS Rescoring

- Apply LAS rescoring to ***lattice*** instead of N-best list to avoid duplicate computation on the common prefixes between hypotheses.

- Apply ***batch inference*** on the arcs expanded from the same node with dynamic batch size, to utilize matrix-matrix multiplication more efficiently.



# Offload Part of LAS Computation to First-Pass




*Compute and cache in the 1st-pass streaming:*

- Unidirectional additional encoder output.
- Projection of the additional encoder output in the additive attention.

# Outline

- Motivation
- Model Architecture
- Quality Improvements
  - Multi-domain Data
  - Robustness to Accents
  - Learning Rates
- Latency Improvements
  - Joint RNN-T Endpointer
  - LAS Rescoring
- **Experiments and Results**

# Experimental Setup

- 1st-pass RNN-T model [Y. He *et al.* 2018]: 120M parameters, 4096 Word Pieces
- 2nd-pass LAS model [T.N. Sainath *et al.* 2019]: 33M parameters
- Baseline conventional server model:
  - Low-frame-rate AM, 1st-pass 5-gram LM, MaxEnt LM 2nd-pass rescoring, 80GB+.
- Data: anonymized utterances from Google traffic
  - Transcripts are normalized: lower-cased, punctuations removed
  - Test data:
    - Search: ~14,000 Search utterances
    - Numeric:~4,000 numeric utterances
    - Multi-talker: ~6,000 multi-talker interfering speech
- Evaluation metrics:
  - Word error rate (WER).
  - Median and 90-percentile endpointer latency (EP50 / EP90).
  - Median and 90-percentile LAS computational latency.

# Conventional Server Model

Without endpointer

# RNN-T Only

Without endpointer

Without LAS rescoring

# Domain-ID Models

| Model | Search | Numeric | Multi-talker |
|---|---|---|---|
| Conventional Server | 6.3 | 13.3 | 8.4 |
| RNN-T (Search) | 6.8 | **10.1** | 10.4 |
| RNN-T (Multi-Domain) | 6.7 | 11.7 | 8.0 |
| RNN-T (Multi-Domain + Domain-id) | **6.6** | 10.4 | **7.7** |

Feeding domain-id improves *robustness* to numerics and background speech babble

# Robustness to Accents

| Search | Conventional Server | RNN-T (Multi-domain+ Domain-id) | +en-X |
|---|---|---|---|
| en-us | 6.3 | 6.6 | 6.7 |
| en-au | 12.1 | 12.6 | **10.3** |
| en-gb | 11.2 | 10.9 | **9.1** |
| en-in | 23.9 | 24.7 | **17.8** |
| en-ke | **27.2** | 28.3 | **27.2** |
| en-ng | 25.6 | 23.6 | **22.8** |
| en-za | 14.3 | 15.7 | 14.8 |

Note: it is assumed that the accent information of each utterance during inference is **NOT** available to all systems.

All the en-X test sets reference transcripts are converted to US English spelling.

For the technique and performance when the accent information **IS** available, see [B. Li *et al.*, ICASSP 2018].

Training with additional en-X data improves accented performance

# Learning Rates

| Model | Search | Numeric | Multi-talker |
|---|---|---|---|
| Conventional Server | 6.3 | 13.3 | 8.4 |
| RNN-T (Decay Learning Rate) | 6.7 | **10.4** | 7.7 |
| RNN-T (Constant Learning Rate + EMA) | **6.2** | 10.5 | **7.1** |

Changing learning rate schedule given increased amount of data improves results by 7% relative

## Conventional Server Model

Without endpointer

# Joint RNN-T EP

Without LAS rescoring

# Endpointer Latency

| Model + Endpointer | WER (Search) | EP50 (ms) | EP90 (ms) |
|---|---|---|---|
| RNN-T without EP | 6.2 | N/A | N/A |
| RNN-T + EOQ EP [1] | 7.4 | 450 | 860 |
| Joint RNN-T EP | **6.8** | **430** | **790** |

Joint RNN-T EP achieves much better WER vs. latency trade-off than a separate EOQ EP, with 8% relative improvement on both WER and EP90.



[1] Shuo-Yiin Chang *et al.*, "A Unified Endpointer Using Multitask and Multidomain Training", ASRU 2019

# Conventional Server Model

With endpointer

# Joint RNN-T EP + LAS

# LAS Rescoring Computational Latency

| Lattice Rescoring with LAS | 50% latency (ms)* | 90% latency (ms)* |
|---|---|---|
| Without batch inference over arcs | 86 | 145 |
| With batch inference over arcs | **58** | **97** |

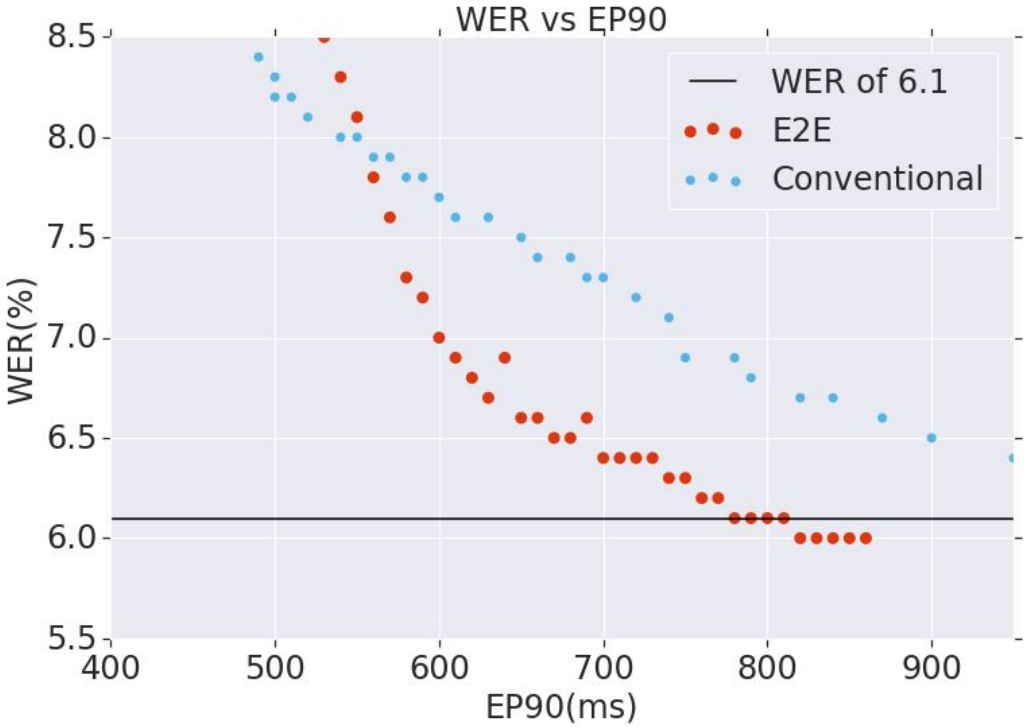
*Benchmarked Search utterances on a Google Pixel4 phone with CPU.

Dynamic batching inference in lattice rescoring with LAS reduces the computational latency by >30%.

# E2E Versus Server

| Model | WER | EP50 | EP90 |
|---|---|---|---|
| Joint RNN-T EP | 6.8 | 430 | 790 |
| + LAS, MWER[1] | **6.1** | **430** | **780** |
| Conventional Server | 6.6 | 460 | 870 |

Joint RNN-T EP + LAS has better WER and latency than server.




[1] Rohit Prabhavalkar *et al.*, "Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models", ICASSP 2018

# Conclusions

- Presented a streaming on-device 1st-pass RNN-T + 2nd-pass LAS model
  - 400 times smaller, 8% lower WER, 10% lower EP latency than conventional server model.
- Model offers improved quality over conventional server model
  - Using domain-id with multi-domain data
  - Using en-X data for accent robustness
  - Constant learning rate with EMA to handle increased data
- Model offers improved latency over conventional server model
  - E2E endpointer to predict end of sentence
  - LAS rescoring: batch inference on lattice arcs, offload some computation into the 1st-pass
- Current on-device model available on Google Pixel phones and the new Google Nest Mini speakers for multiple domains/applications!
  - Next Generation Assistant (*voice search/actions*), Gboard (*short message dictation*)
  - Call Screen (*telephony speech*), Recorder (*long-form speech*)
  - Live Caption (*video/audio captioning*), Assistant on Nest Mini (*farfield/noisy speech*)

# Thank you!

Contacts: {tsainath, yanzhanghe}@google.com