Towards Fast and Accurate Streaming End-To-End ASR

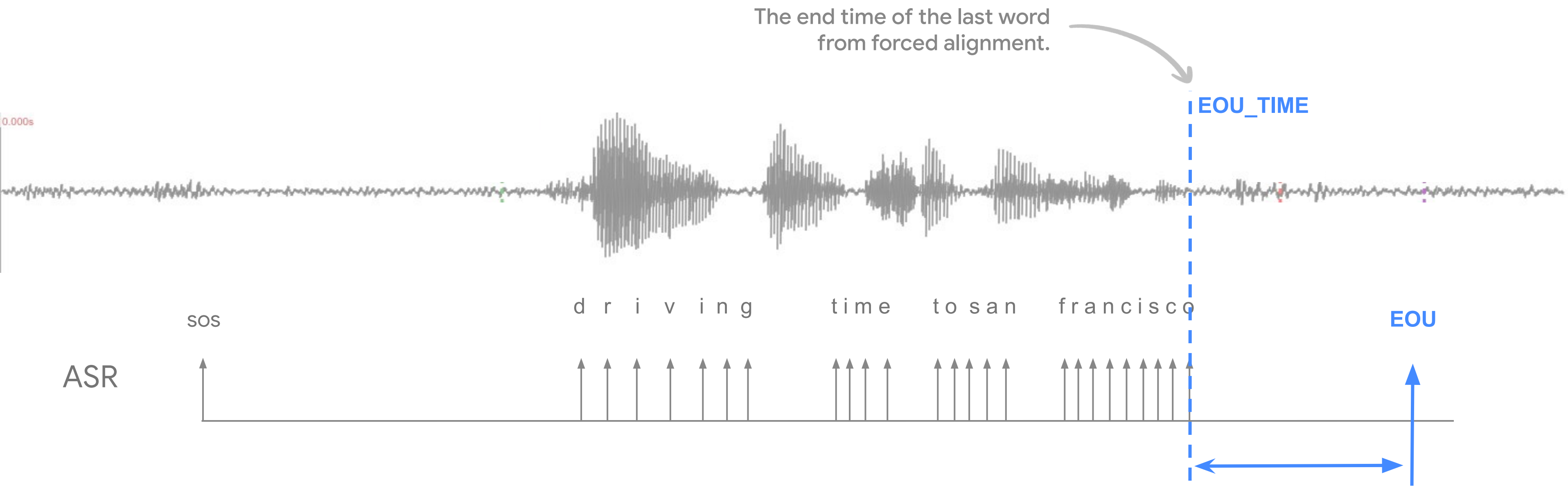
Presenter: Bo Li (boboli@google.com)

Authors: Bo Li, Shuo-yiin Chang, Tara N. Sainath, Ruoming Pang, Yanzhang He, Trevor Strohman, Yonghui Wu

ICASSP 2020

Google

Modeling End-Of-Utterance (EOU) jointly with ASR in RNN-T for better latency.

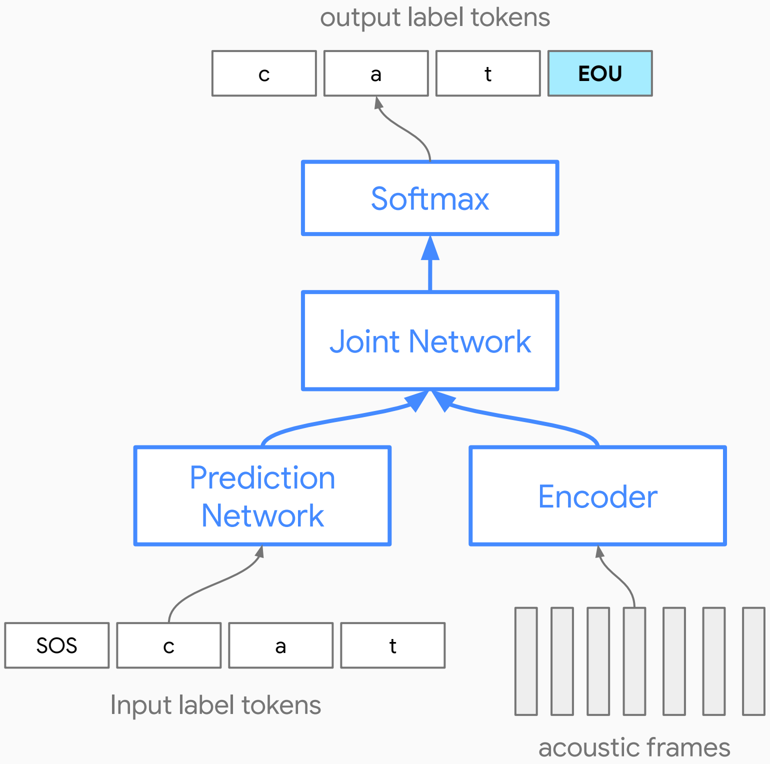


The time difference between the user finishes speaking (EOU_TIME) and the system generates the final hypothesis (EOU).

Latency

The closer EOU is predicted to EOU_TIME, the better the latency is.

RNN-T EP



OLD

NEW

Accurate EOU Timing

Based on **time alignment** of the end of last word.
Adding **early and late penalties** for EOU predictions.

Reducing Premature EOU

EOU terminates beam search paths during inference.
Sequence training with **MWER**.

Trading WER for Latency

Sacrifice WER for latency in the 1st pass RNN-T decoding.
Recover WER gains via **2nd pass LAS rescoring**.

[1] Shuo-Yiin Chang, et.al, "Joint Endpointing and Decoding with End-to-end Models", ICASSP 2019

[2] Bo Li, et.al, "Towards Fast and Accurate Streaming End-to-End ASR", submitted to ICASSP 2020

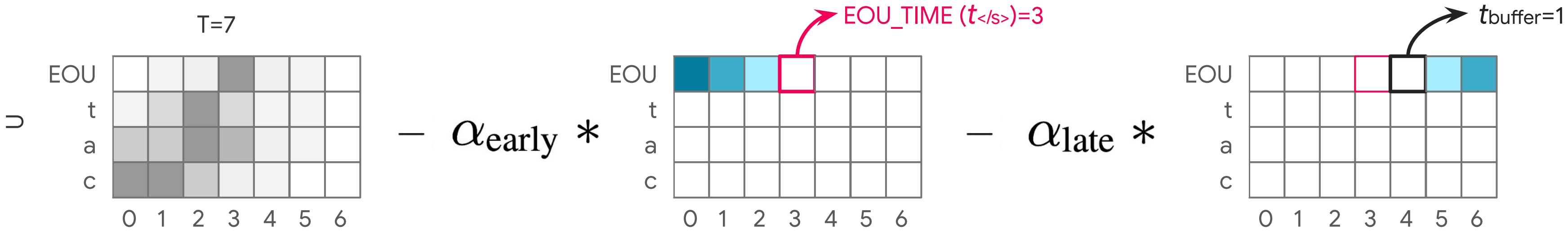
Accurate EOU Timing

To help the model predict EOU as close to the end of the last word as possible.

Original U*T Matrix

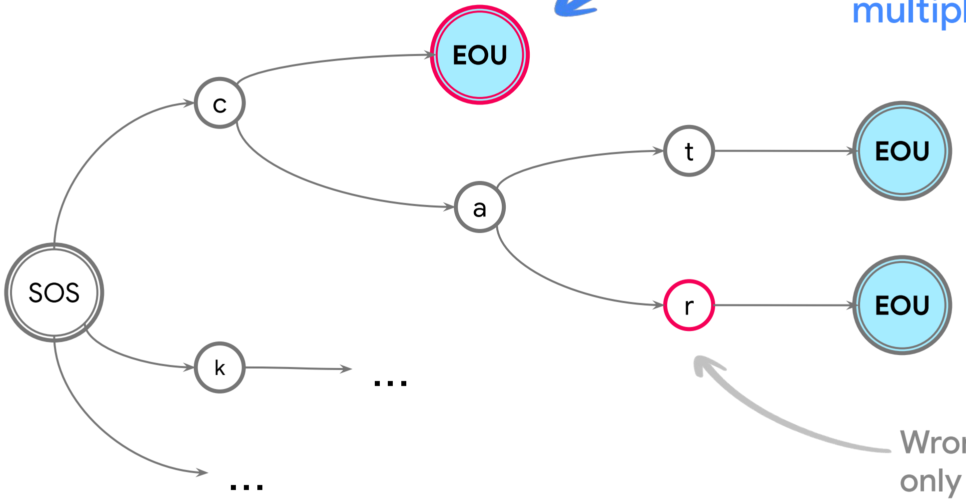
Early Penalty

Late Penalty



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

Reducing Premature EOU



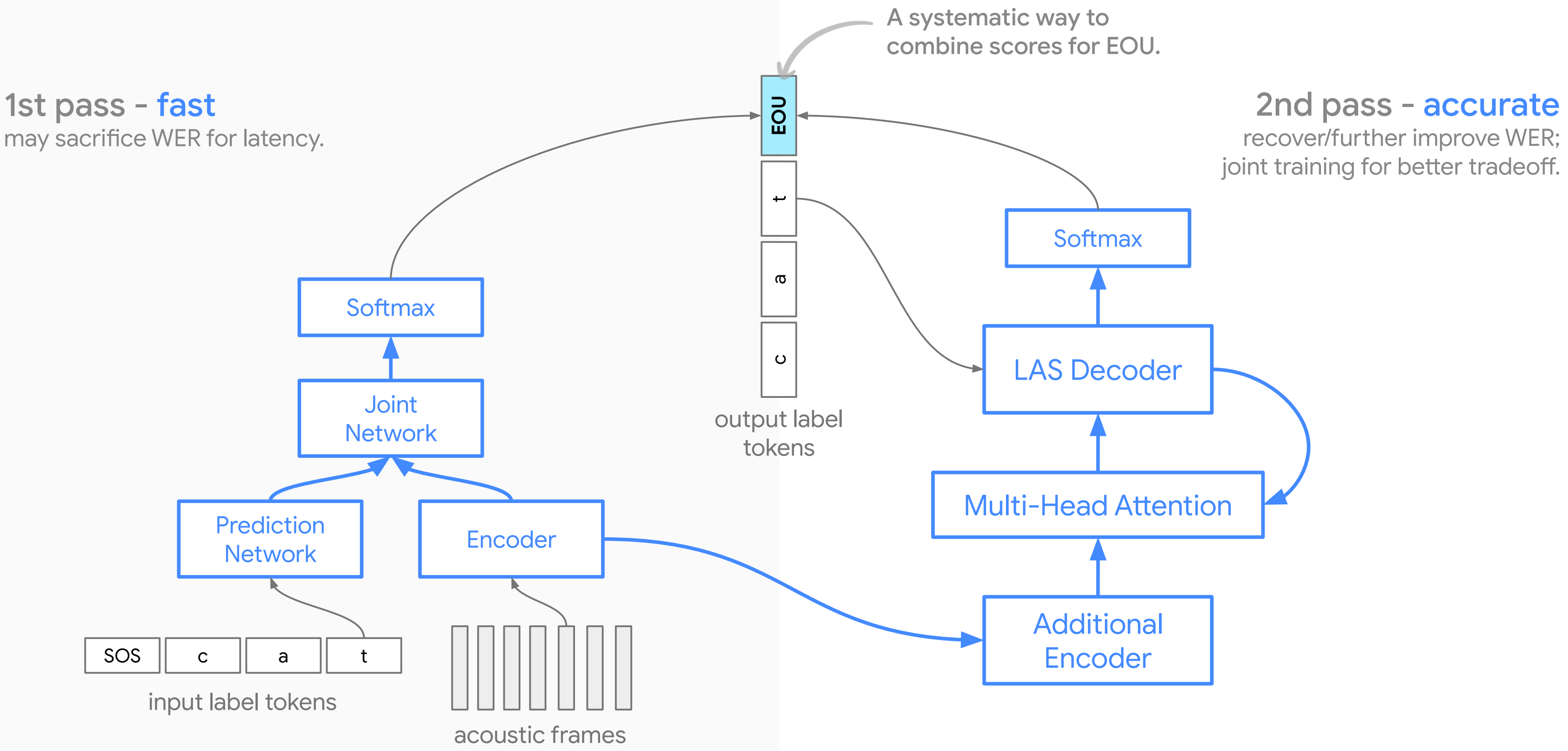
Premature EOU prediction can lead to multiple deletion errors.

Wrong prediction of non EOU token only cause a single error.

Sequence training using Minimum Word Error Rate (MWER) is adopted to address this issue.

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

Trading WER for Latency



[1] Tara N. Sainath, et.al, "Two-Pass End-to-End Speech Recognition", Interspeech 2019

Experiment Setup

- Dataset:
  - Human transcribed audio-text pairs from a variety of domains: Search, Farfield, Telephony, YouTube [A. Narayanan et al., ASRU 2019]
- Features:
  - Log-mel Filterbanks together with a 1-hot vector of the domain-id to help with modeling domain variations [B. Li *et al.*, ICASSP 2018].
- Models:
  - 1st-pass RNN-T model [Y. He *et al.* 2018]: 120M parameters, 4096 Word Piece Model
  - 2nd-pass LAS model [T.N. Sainath *et al.* 2019]: 33M parameters
- Metrics:
  - Word error rate (WER)
  - Median latency (EP50) and 90-percentile latency (EP90)

Baselines

	RNN-T	+EP
WER	**7.2**	7.5
EP50	540	**410**
EP90	910	**710**

increase in deletion errors

Joint modeling of EOU in RNN-T with ASR helps reducing latency but hurts quality.

Early & Late Penalties

	RNN-T	+EP	+Early&Late
WER	7.2	7.5	**7.2**
EP50	540	410	**380**
EP90	910	710	**850**

Constraining EOU prediction time during training via early and late penalties helps both quality and latency, although EP90 gain is relatively small.

Sequence Training

	RNN-T	+EP	+Early&Late	+MWER	-Early
WER	7.2	7.5	7.2	7.2	**6.9**
EP50	540	410	380	430	**380**
EP90	910	710	850	630	**580**

MWER already penalizes premature EOU prediction, rendering early penalty unnecessary.

MWER without early penalty improves both qualty and latency:

WER: rel. 4.2%
EP50: 160 ms
EP90: 330 ms

2nd Pass Rescoring

	RNN-T	+EP	+Early&Late	+MWER	-Early
WER	7.2	7.5	7.2	7.2	6.9
EP50	540	410	380	430	380
EP90	910	710	850	630	580

LAS largely improves quality,
11.1% rel. WER reduction.

	+LAS	+ignore RNN-T EOU score
WER	6.4	6.6
EP50	380/**370**	370
EP90	850/**740**	740

This simulates RNN-T + LAS. RNN-T EP + LAS is a more systematic way of combining EOU scores.

Final System

	RNN-T	+EP	+Early&Late
WER	7.2	7.5	7.2
EP50	540	410	380
EP90	910	710	850

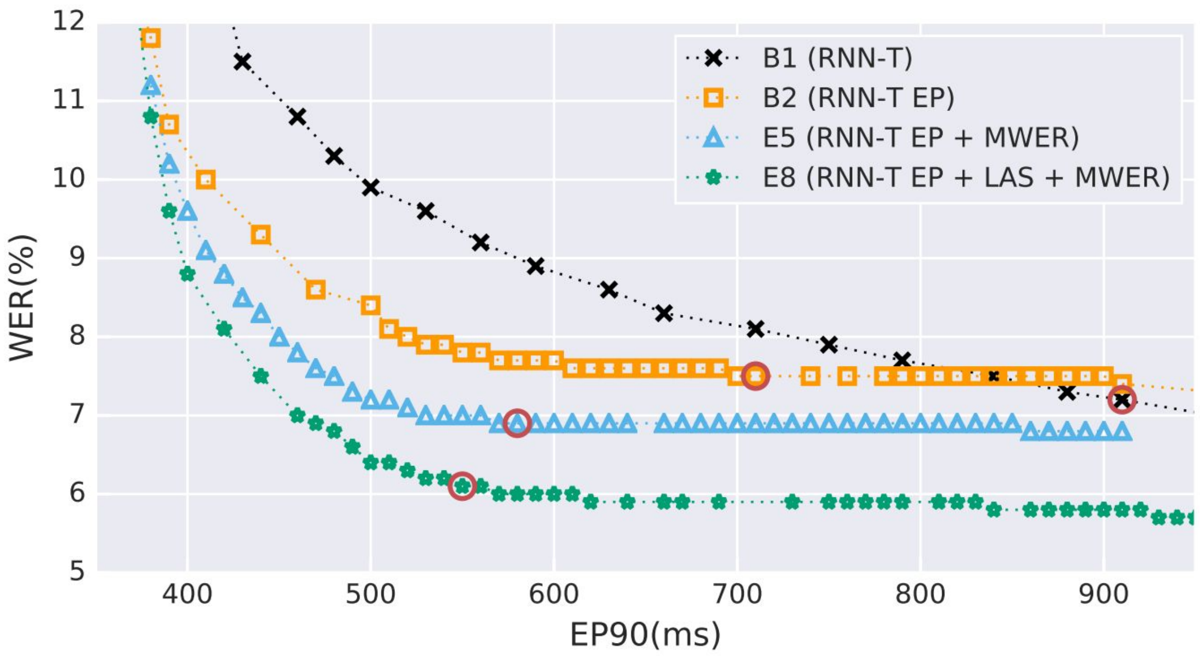
MWER training of both the RNN-T and LAS gives the best quality and latency:

WER: rel. 15.3%
EP50: 170 ms
EP90: 360 ms

	+LAS	+MWER LAS	+MWER ALL
WER	6.4	6.2	**6.1**
EP50	380/370	350	**370**
EP90	850/740	620	**550**

Analysis

The proposed systems are consistently better:



Summary

Accurate EOU Timing through early and late penalties.

Based on **time alignment** of the end of last word.
Adding **early and late penalties** for EOU predictions.

Reducing Premature EOU via MWER sequence training.

EOU terminates beam search paths during inference.
Sequence training with **MWER**.

Trading WER for Latency via 2nd pass LAS rescoring.

Sacrifice WER for latency in the 1st pass RNN-T decoding.
Recover WER gains via **2nd pass LAS rescoring**.

Thank you!

Contacts: {boboli, shuoyiin}@google.com