



# An Attention-Based Joint Acoustic and Text On-Device End-to-End Model

Tara N. Sainath, Ruoming Pang, Ron J. Weiss, Yanzhang (Ryan) He, Chung-cheng Chiu, Trevor Strohman

ICASSP 2020

May 6, 2020

# Outline

- **Motivation**
- Joint Acoustic and Text Decoder (JATD)
- Experiments
- Results

# Motivation

- E2E models are trained on audio-text pairs, which is a *fraction* of data compared to a conventional ASR model
- E2E models lag behind conventional model performance on *rare words* and *long-tail named entities*
- E2E long-tail performance can be improved with unpaired data

# Related Work Using Unpaired Data

- Language model fusion [[Chorowski2017](#), [Sriram2017](#), [Kannan2018](#)]
  - Requires additional model and not amenable to on-device
- Weak-distillation [[Li2019](#)]
  - Requires multiple training steps to incorporate unpaired data
- Synthesizing text-only data [[Sennrich2016](#)]
  - Some techniques have had limited success in ASR [[Li2019](#)]
  - Other techniques increase training steps [[Hori2019](#), [Re2019](#)]

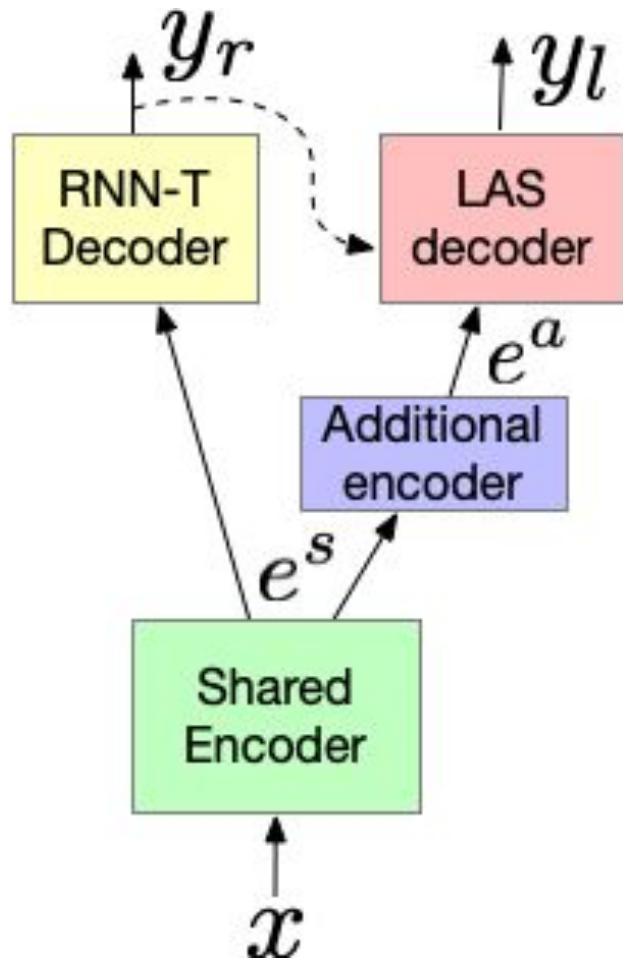
**Goal:** We seek to incorporate unpaired data without increasing *model size* or *training time* significantly so that the solution is *on-device* friendly.

# Outline

- Motivation
- **Joint Acoustic and Text Decoder (JATD)**
- Experiments
- Results

# Two-Pass Model Overview

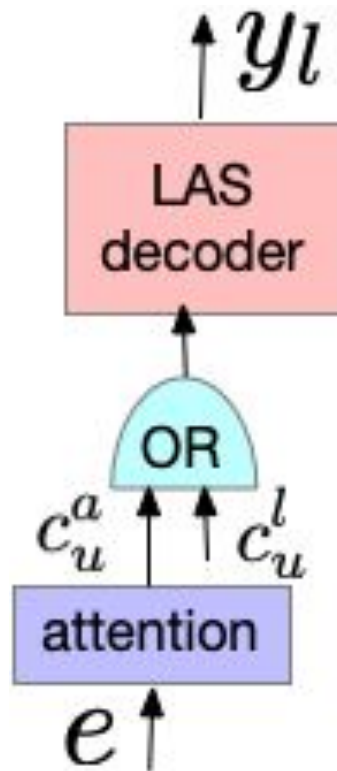
- 2-pass RNN-T + LAS E2E model has been shown to achieve similar quality to a conventional server-based model\*
- Training:
  - 1st-pass: shared-encoder + RNN-T decoder
  - 2nd-pass: additional encoder + LAS decoder
- Two Modes for Inference
  - 1st-pass RNN-T produces hypotheses
  - Rescoring: LAS rescors hypotheses from RNN-T
  - Beam-Search: LAS decoder runs a beam search ignoring RNN-T hypotheses



\* [T.N. Sainath, R. Pang et al, "[Two-Pass End-to-End Speech Recognition](#)", Proc. Interspeech, 2019.]

# Joint Acoustic Text Decoder for LAS

- Multi-domain/dialect work [[Li2018](#)] has shown passing a context vector allows the model to be robust to different kinds of inputs
- We explore changing the context vector  $\mathbf{c}_u$  to the LAS decoder at decoding step  $u$ 
  - $\mathbf{c}_u^a$  is an acoustic context vector if input data is supervised audio-text pair
  - $\mathbf{c}_u^l$  is a fixed (but learnable) vector if input data is text-only unpaired data
- Putting two context-vectors into the model rather than interpolating two different models is amenable for *on-device*.




# Inference with JATD


- At each decoding step  $u$ , scores from acoustic and language context vector inputs are interpolated by  $\lambda$
- This is similar to an acoustic and language model

$$\lambda \log p(\mathbf{y}_u | \mathbf{x}, \mathbf{c}_u^a, \mathbf{y}_{u-1:1}) + (1 - \lambda) \log p(\mathbf{y}_u | \mathbf{c}_u^l, \mathbf{y}_{u-1:1})$$

Acoustic context  
vector score



Language context  
vector score





## (1) *Individual* Training Strategy

- Loss is computed either from the acoustic or language context vector, depending on the type of input data
- If paired data is used, update LAS decoder and acoustic context vector params
- If unpaired data is used, update LAS decoder and fixed context vector params

$$\mathcal{L} = \begin{cases} \log p(\mathbf{y}_u | \mathbf{x}, \mathbf{c}_u^a, \mathbf{y}_{u-1:1}), & \text{if paired example} \\ \log p(\mathbf{y}_u | \mathbf{c}_u^l, \mathbf{y}_{u-1:1}), & \text{if unpaired example} \end{cases}$$

## (2) Joint Training Strategy

- To better match inference, loss is computed by interpolating scores from acoustic and language context vectors during training
- Creating audio-text pairs:
  - $\mathbf{x}^a$ : supervised audio-text paired data
  - $\mathbf{x}^l$ : “created audio” from unpaired data (use unsupervised data in this work)
- If paired data is used, update LAS decoder and acoustic context vector params
- If unpaired data is used, update LAS decoder and fixed context vector params

$$\mathcal{L} = \begin{cases} \lambda \log p(\mathbf{y}_u | \mathbf{x}^a, \mathbf{c}_u^a, \mathbf{y}_{u-1:1}) + (1 - \lambda) \log p(\mathbf{y}_u | \mathbf{c}_u^l, \mathbf{y}_{u-1:1}), \\ \text{if paired example} \\ \lambda \log p(\mathbf{y}_u | \mathbf{x}^l, \mathbf{c}_u^a, \mathbf{y}_{u-1:1}) + (1 - \lambda) \log p(\mathbf{y}_u | \mathbf{c}_u^l, \mathbf{y}_{u-1:1}), \\ \text{if unpaired example} \end{cases}$$

# Outline

- Motivation
- Joint Acoustic and Text Decoder (JATD)
- **Experiments**
- Results

# Experimental Setup

- 1st-pass RNN-T model [[He et al. 2018](#)]: ~120M parameters, 4096 Word Piece Model
- 2nd-pass LAS model [[Sainath et al. 2019](#)]: ~33M parameters
- Train on Multi-domain utterances (Search, Farfield, Telephony, YouTube) [[Narayanan et al. 2019](#)]
- Data: anonymized utterances from Google traffic
  - Short Utterances (**SU**): <5.5s, ~14K Search utterances
  - Long Utterances (**LU**): >=5.5s, ~16K Search utterances
  - **SXS**: ~1K Search utterances where E2E has more losses compared to conventional model
  - Corrections (**Corr**): ~5K utterances where the user typed a query immediately after speaking
  - **App**: ~16K phrases of app interaction, synthesized
  - **Songs**: ~16K phrases of media requests, synthesized

# Outline

- Motivation
- Joint Acoustic and Text Decoder (JATD)
- Experiments
- **Results**

# Individual Log-Probability in Training

- Numbers are with JATD rescoring of RNN-T hypotheses for now
- WER as a function of  $\alpha$  (% of paired data seen in each training epoch)

Model	SU	LU	SXS	Corr	App	Songs
RNN-T	6.9	4.9	31.8	15.3	8.9	15.4
JATD, $\alpha=1.0$	5.9	3.7	29.6	14.3	8.7	13.2
JATD, $\alpha=0.9$	<b>6.0</b>	<b>3.6</b>	<b>29.4</b>	<b>14.1</b>	<b>8.7</b>	<b>12.8</b>
JATD, $\alpha=0.75$	6.1	3.7	29.5	14.2	8.6	12.8
JATD, $\alpha=0.5$	6.2	3.7	29.6	14.3	8.6	12.8

Very small improvements when incorporating text-only data ( $\alpha < 1.0$ )

# Joint Log-Probability in Training

Model	SU	LU	SXS	Corr	App	Songs
JATD, $\alpha=1.0$	5.9	3.7	29.6	14.3	8.7	13.2
JATD, $\alpha=0.9$	6.0	<b>3.6</b>	28.6	14.1	8.3	12.4
JATD, $\alpha=0.75$	<b>6.0</b>	3.7	<b>28.2</b>	<b>13.9</b>	<b>8.2</b>	<b>11.9</b>
JATD, $\alpha=0.5$	6.1	3.7	28.5	14.0	8.2	12.2
JATD, $\alpha=0.25$	6.3	3.8	28.6	14.2	8.1	12.1

With joint log-probability 3-10% relative improvement ( $\alpha=0.75$ ) compared to no text data ( $\alpha=1.0$ )

# Increasing Model Capacity

- Since the JATD model has two functions (i.e., an AM and LM), we explore increasing model capacity of decoder from 33M parameters (S) to 66M (L)

Model	SU	LU	SXS	Corr	App	Songs
JATD, $\alpha=1.0$ , S	5.9	3.7	29.6	14.3	8.7	13.2
JATD, $\alpha=1.0$ , L	5.8	3.6	29.2	14.3	8.7	13.2
JATD, $\alpha=0.75$ , S	6.0	3.7	28.2	13.9	8.2	11.9
JATD, $\alpha=0.75$ , L	5.8	3.7	<b>27.3</b>	13.9	8.1	11.9
JATD, $\alpha=0.5$ , L	<b>5.8</b>	<b>3.6</b>	27.4	<b>13.8</b>	<b>8.0</b>	<b>11.9</b>
JATD, $\alpha=0.25$ , L	6.0	3.6	27.4	14.1	8.1	11.9

Increasing model size allows us to use more text-only data (smaller  $\alpha$ ) with further improvements



# Using JATD for Beam Search

Model	SU	LU	SXS	Corr	App	Songs
RNN-T	6.9	4.9	31.8	15.3	8.9	15.4
JATD, $\alpha=1.0$ , L	5.6	3.5	26.5	14.4	8.6	12.5
JATD, $\alpha=0.75$ , L	5.7	3.5	24.3	14.0	8.1	11.3
JATD, $\alpha=0.5$ , L	<b>5.6</b>	<b>3.5</b>	<b>23.2</b>	<b>14.0</b>	<b>7.9</b>	<b>11.1</b>
JATD, $\alpha=0.25$ , L	5.8	3.7	22.8	14.0	7.9	11.1

Beam search JATD gives 3-12% relative improvement ( $\alpha=0.5$ ) compared to no text data ( $\alpha=1.0$ )

# Comparison to Other Techniques

- Compare JATD to techniques which do not increase model size/training time
- All models run in beam-search mode

Model	Training Data for LAS Decoder	SU	LU	SXS	Corr	App	Songs
LAS - P	100% Paired	5.6	3.5	26.5	14.4	8.6	12.5
LAS - P+U	50% Paired, 50% Unpaired	6.5	6.3	26.4	15.0	8.9	14.7
JATD	50% Paired, 50% Unpaired	<b>5.6</b>	<b>3.5</b>	<b>23.2</b>	<b>14.0</b>	<b>7.9</b>	<b>11.1</b>

JATD shows the best performance compared to other techniques that do not increase model size or training time

# Wins of JATD

- Table shows JATD wins (green) and LAS - P errors (in red)

LAS - P	JATD
How do you <b>bake</b> a hook for bass fishing	How do you <b>bait</b> a hook for bass fishing
Peanut butter cookies <b>and</b> scratch	Peanut butter cookies <b>from</b> scratch
Houston <b>Astro</b> cap	Houston <b>Astros</b> cap
What is <b>ligonberry</b>	What is <b>lingonberry</b>

JATD fixes both language modeling and proper noun errors

# Rare Word Analysis

- Rare word: word with count of  $< 10$  in training

Model	SU	LU	SXS	Corr	App	Songs
LAS - P	6.4	4.9	7.6	17.6	10.0	6.3
JATD	4.5	3.9	5.0	11.7	4.7	5.1

% errors due to rare words decreases by more than 20% relative with JATD model

# Conclusions

- Presented a joint acoustic and text decoder (JATD) within the LAS 2nd-pass framework to use both paired and unpaired data.
- Model is efficient for on-device
  - Does not increase model size
  - Does not increase training time
- JATD model gives a 3-12% relative improvement across a variety of proper noun test sets compared to an LAS model trained on paired data only.

Thank you!

Contact: [tsainath@google.com](mailto:tsainath@google.com)