Latest Advances in End-to-End Speech Recognition

Tara N. Sainath

November 2, 2022

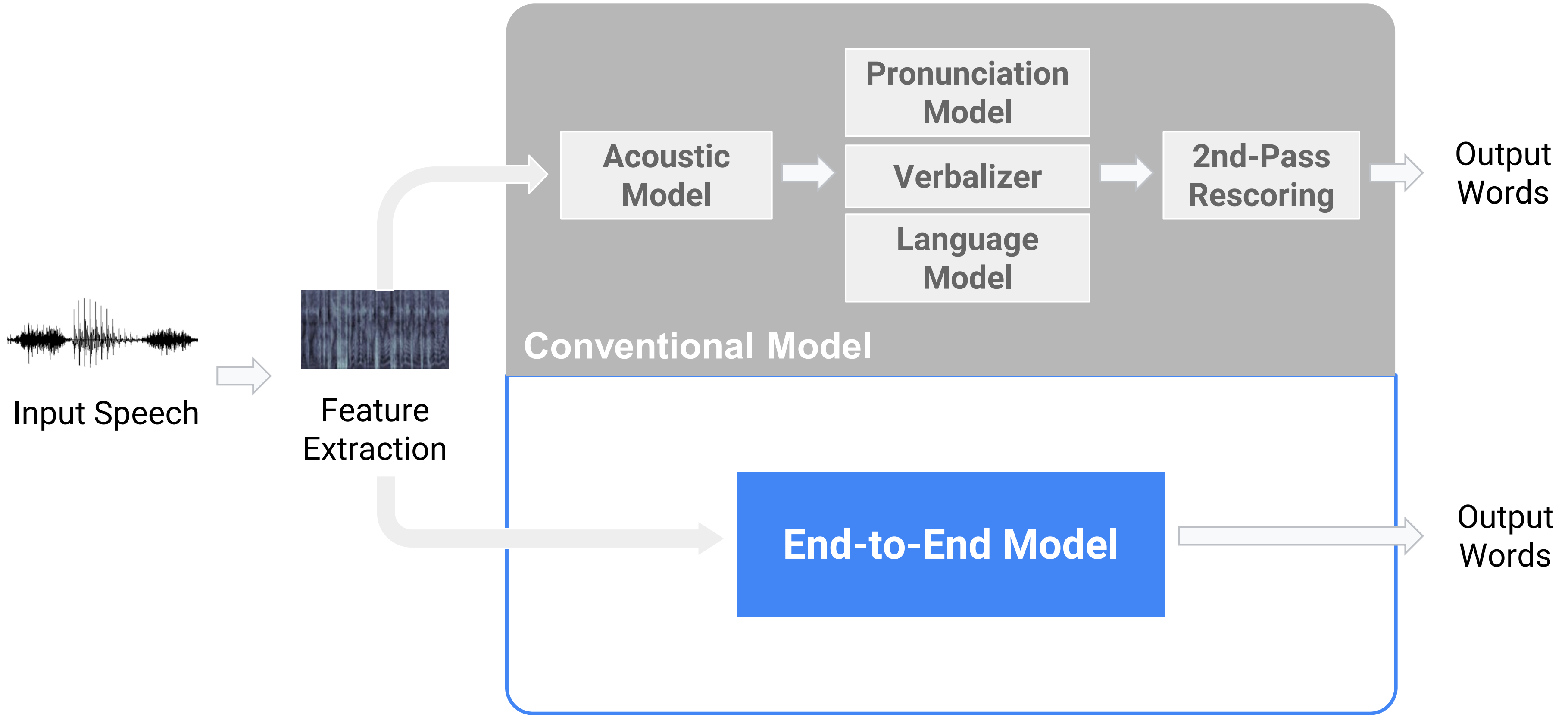
tsainath@google.com

A Joint Effort Between Google Brain, Hardware and Speech Teams

What is End-to-End ASR?

“A system which directly maps a sequence of input acoustic features into a sequence of graphemes or words.”

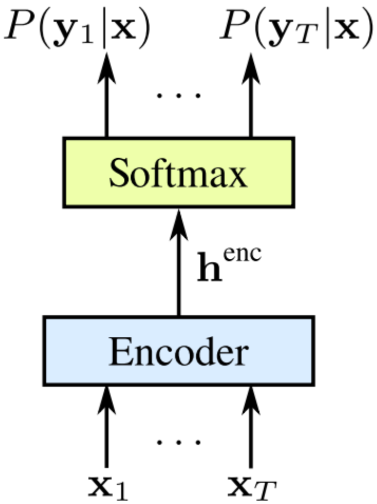
"A system which is trained to optimize criteria that are related to the final evaluation metric - word error rate."


Input Speech
Feature Extraction
Acoustic Model
Pronunciation Model
Verbalizer
Language Model
2nd-Pass Rescoring
Output Words
Conventional Model
End-to-End Model
Output Words

Historical Development of End-to-End ASR

Connectionist Temporal Classification (CTC)

[Graves et al., 2006]




```
B B c B B a a B B t
B c c B a B B B B t
        ...
B c B B a B B t t B
```

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\hat{\mathbf{y}} \in \mathcal{B}(\mathbf{y},\mathbf{x})} \prod_{t=1}^{T} P(\hat{y}_t|\mathbf{x})$$

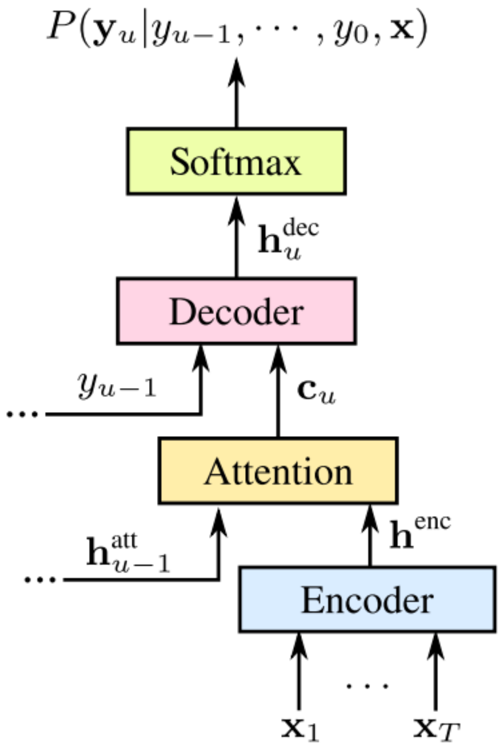
- CTC introduces a special symbol - blank (denoted by B) - and maximizes the total probability of the label sequence by marginalizing over all possible alignments.
- No frame-level alignment is needed.

CTC-Based End-to-End ASR

- [Graves & Jaitly, 2014] proposed a system with character-based CTC which directly output word sequences given input speech
- LM incorporated into first-pass decoding; easy integration with WFSTs
  - [Hannun et al., 2014] [Maas et al., 2015]: Direct first-pass decoding with an LM as opposed to rescoring as in [Graves & Jaitly, 2014]
- Large-scale GPU training; data augmentation; multiple languages
  - [Hannun et al., 2014; DeepSpeech] [Amodei et al., 2015; DeepSpeech2]: Large scale GPU training; Data Augmentation; Mandarin and English
- Using longer span units: words instead of characters
  - [Soltau et al., 2017]: Word-level CTC targets, trained on 125,000 hours of speech. Performance close to or better than a conventional system, even without using an LM!
  - [Audhkhasi et al., 2017]: Direct Acoustics-to-Word Models on Switchboard
- And many others ...

Attention-based Encoder-Decoder Models

[Chan et al., 2015][Chorowski et al., 2015]



- Encoder (analogous to AM):
  - Transforms input speech into higher-level representation
- Attention (alignment model):
  - Identifies encoded frames that are relevant to producing current output
- Decoder (analogous to PM, LM):
  - Operates autoregressively by predicting each output token as a function of the previous predictions

Comparing Various End-to-End Approaches

[Prabhavalkar et al., 2017]

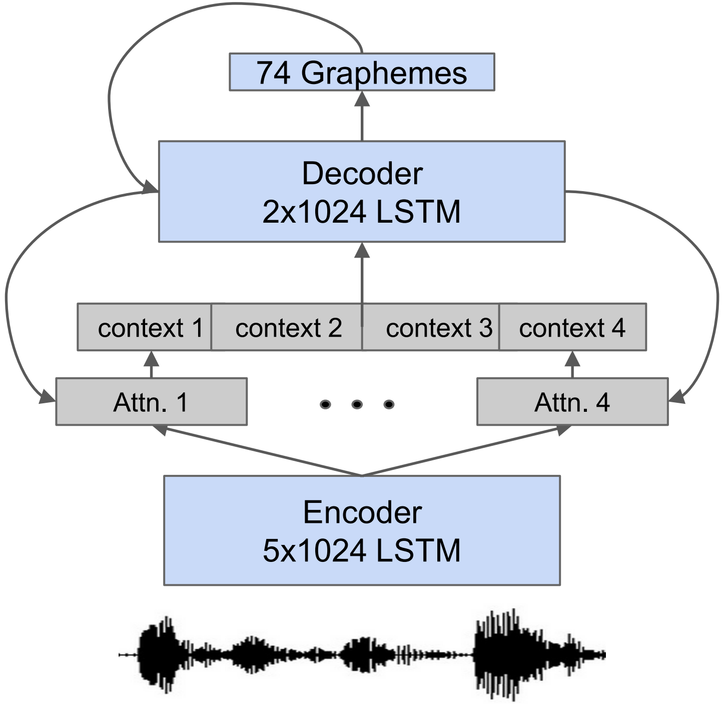
Model	Online/Offline	VoiceSearch Word Error Rate (%)
Conventional Model	online	9.9
	offline	8.6
CTC-Grapheme (no LM)	online	53.4
Attention-based Model	offline	11.7

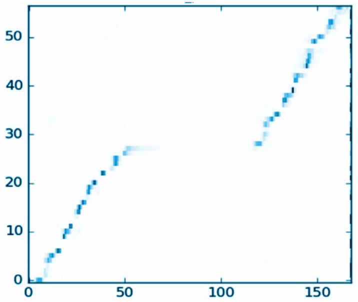
- Decoding CTC-grapheme models without an LM performs poorly.
- Attention-based model performs better, but still lags behind a conventional model.

2018: Further Improvements

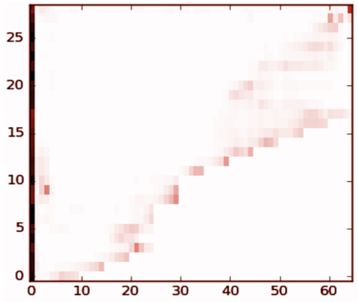
Multi-headed Attention

[A. Vaswani, 2017][Chiu et al., 2018]





Single Headed Attention



Multi-Headed Attention

- Multi-headed attention examines different parts of the utterance for each predicted label.
- Model looks predominantly towards previous frames.

Word Pieces

[Schuster, 2012][Chiu et al., 2018]

- We want to use subword units longer than graphemes:
  - Longer units have a lower LM perplexity
  - Longer units gives improved decoder efficiency
- Word pieces is a good longer-unit choice [Schuster, 2012]
  - Has shown good results for RNN-T [Rao, ASRU 2017]
- Word piece model (WPM) details
  - Trained to maximize LM likelihood on training data
  - Position dependent, determined determinstically
  - Units back off to characters → No OOVs

Good Afternoon → _go o d _aft er noon

Minimum Word Error Rate (MWER)

[Stolcke et al., 1997][Povey, 2003][Prabhavalkar et al., 2018]

- End-to-end models are typically trained by optimizing cross entropy loss (i.e., maximizing log-likelihood of the training data)

$$\mathcal{L}_{\text{CE}} = \sum_{(\mathbf{x},\mathbf{y}^*)} \sum_{u=1}^{L+1} -\log P(y_u^* | y_{u-1}^*, \cdots, y_0^* = \langle \text{sos} \rangle, \mathbf{x})$$

- Training criterion does not match metric of interest: Word Error Rate
- MWER optimizes the expected word error rates:

$$\mathcal{L}_{\text{werr}}(\mathbf{x}, \mathbf{y}^*) = \mathbb{E}[\boxed{\mathcal{W}(\mathbf{y}, \mathbf{y}^*)}] = \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}) \mathcal{W}(\mathbf{y}, \mathbf{y}^*)$$

Number of Word Errors

Optimization improvements
[Chiu et al., 2018]

- Scheduled Sampling [S. Bengio, 2015]
  - Feed back in the prediction from the model rather than the true previous prediction
  - Helps prevent overfitting
- Label smoothing [C. Szegedy, 2016]
  - Take the logit class with maximum probability and smooth it over the remaining labels
  - Helps prevent overfitting
- Sync training [P. Goyal, 2017]
  - Gradient updates between workers are synchronized
  - Leads to faster convergence and better model quality

Comparison to Conventional Model

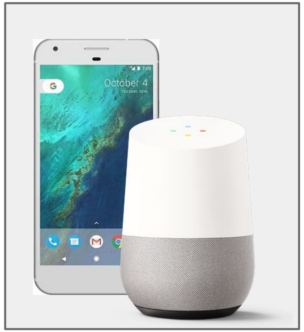
[Chiu et al., 2018]

Model	1st Pass Model Size	VoiceSearch Word Error Rate (%)
Conventional Server	0.1GB (AM) + 2.2 GB (PM) + 4.9 GM (LM) = 7.2 GB	6.7
Attention-based Model	**0.4 GB**	**5.6**

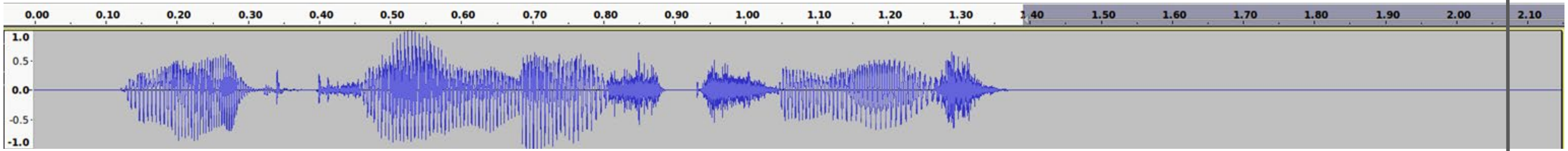
- **16%** relative performance improvement over conventional model
- **18X smaller** than conventional model in 1st pass
- Main drawback: model is **not streaming**

2019-2020: Online End-to-End Models for Pixel 4

Streaming Speech Recognition



Finalize recognition &
Taking action / fetching the search results

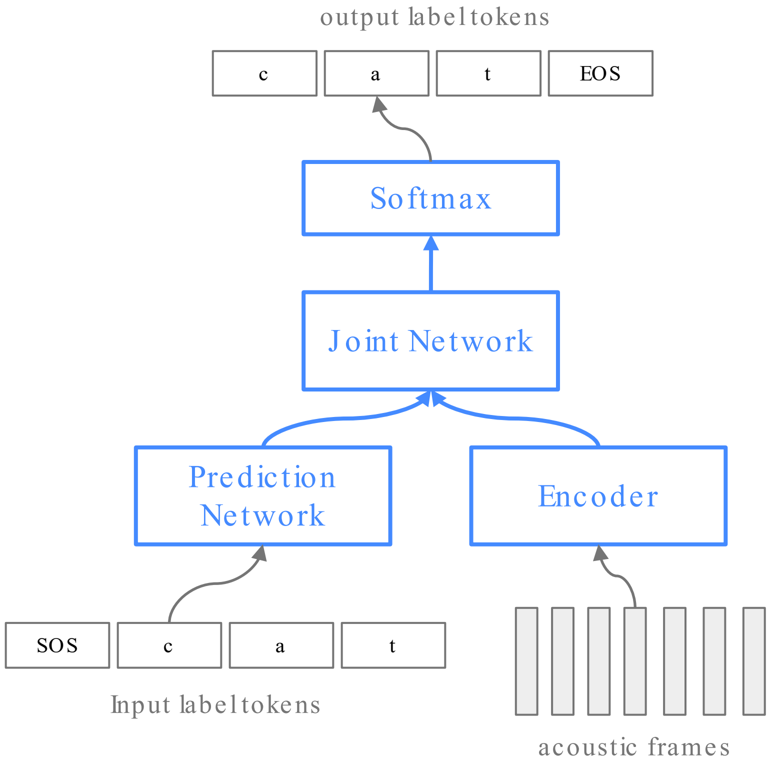


Recognize the audio

Recurrent Neural Network Transducer (RNN-T)

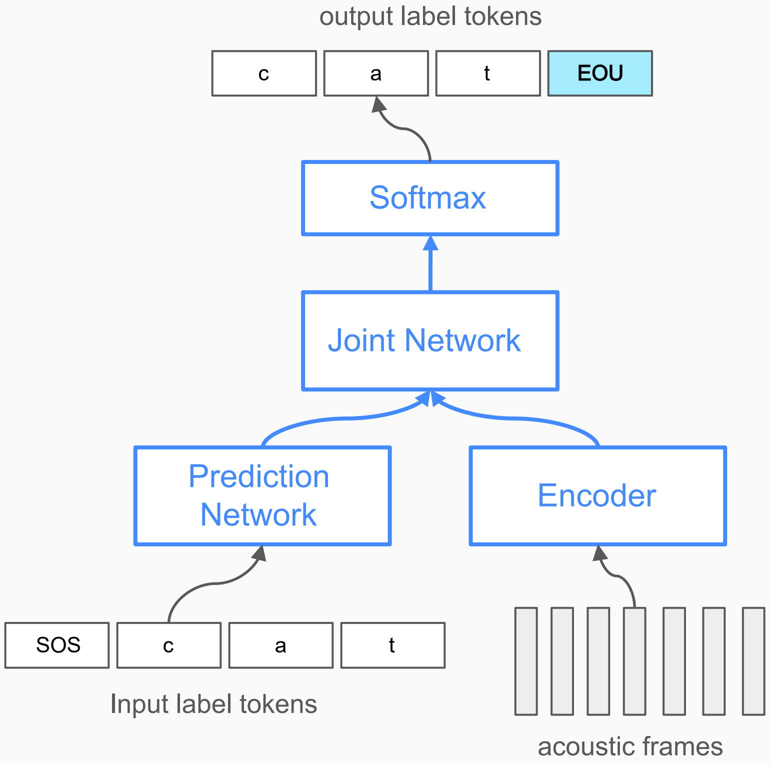
[Graves, 2012], [Rao et al., 2017], [He et al., 2018]

- Encoder Network
  - Set of recurrent layers (like am AM)
- Prediction Network
  - recurrent LM
- Joint Network
  - combines AM and LM predictions
- Jointly optimized end-to-end
- No alignment needed.
- Streaming model.



Low Latency RNN-T (RNN-T Endpointer)

[Li et al., 2020][Sainath et al., 2020]



OLD

EOU Modeling

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

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**.

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

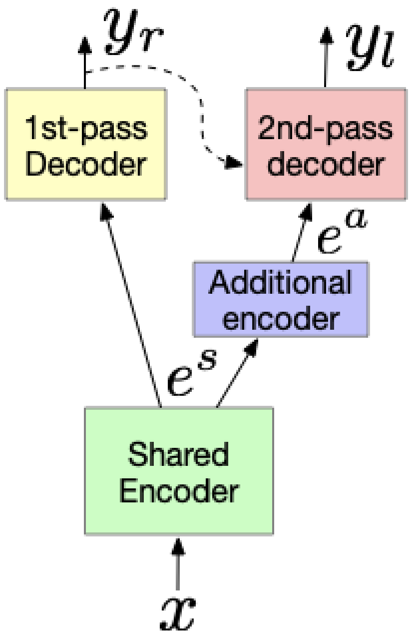
[2] Bo Li, et.al, "Towards Fast and Accurate Streaming End-to-End ASR", submitted to ICASSP 2020

Model	VoiceSearch Word Error Rate (%)	EOU Latency (90 percentile)
On-Device RNN-T + VAD	7.4%	860ms
On-Device RNN-T EP	**6.8%**	**790ms**

RNN-EP gives better WER and latency tradeoff compared to RNN-T + VAD

Second-pass LAS Rescoring

[Sainath et al. 2019][Sainath et al., 2020]



- 1st-pass RNN-T for streaming applications.
- 2nd-pass full-context attention-based LAS decoder for better quality.
- Shared encoder for a compact model.

Model	**VoiceSearch Word Error Rate (%)**
On-Device RNN-T EP	6.8%
+ LAS Rescoring	**6.1%**

Comparison to Conventional Model

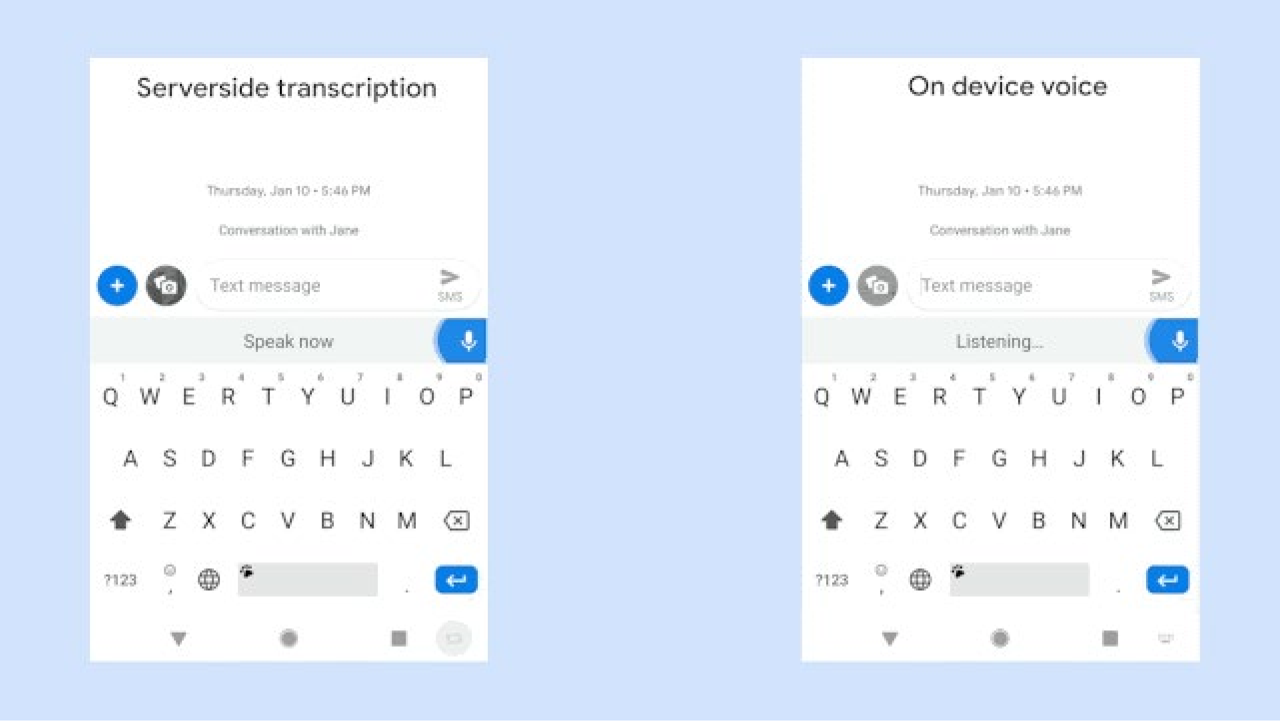
[Sainath et al., 2020]

Model	Size	VoiceSearch Word Error Rate (%)	EOU Latency (90 percentile)
Conventional Server	0.1GB (AM) + 2.2 GB (PM) + 4.9 GB (LM) + 80 GB (2nd-pass LM) = 87.2 GB	6.6%	870ms
On-device End-to-End	**0.18 GB**	**6.1%**	**780ms**

“ You can ask Assistant to do things that are **local to your phone**, and they'll happen **near instantaneously**...it is much faster and needs to rely on Google's server **much less**. ”

The Verge: [Google Pixel 4 and 4 XL Hands-on: this time, it's not about the camera.]

Gboard Demo


Serverside transcription
Thursday, Jan 10 • 5:46 PM
Conversation with Jane
Text message
SMS
Speak now
On device voice
Thursday, Jan 10 • 5:46 PM
Conversation with Jane
Text message
SMS
Listening...

- **Summary:**
  - Attention-based End-to-End models (LAS) **achieves state-of-the-art performance**, but is **not streaming**.
  - Recurrent Neural Network Transducer (RNN-T) provides an **accurate** and **fast on-device** speech recognition experience.
  - RNN-T EP + 2nd-pass LAS **surpasses** server-side conventional model in both **quality** and **latency**.

- **Challenges:**
  - Dealing with long tail words.
  - Further simplify the ASR system to build a single End-to-End model for multiple languages.

2021 On-Device Improvements

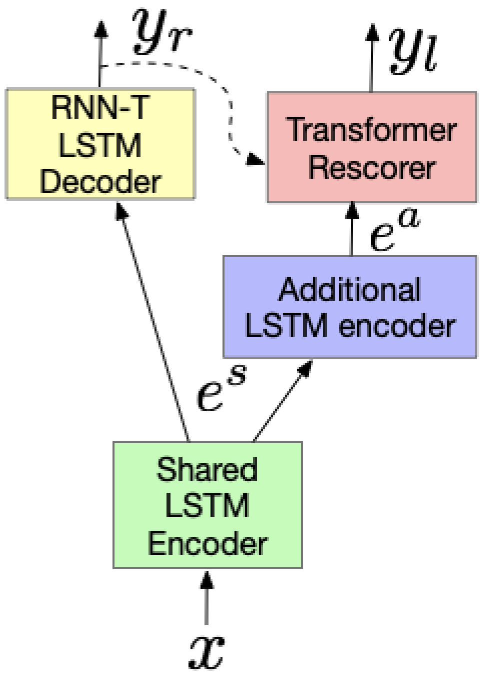
2021 End-to-End Goals with Pixel 6

Our goals in Pixel 6 are to develop an **on-device** end-to-end model that

- **Surpasses quality** (as measured by word error rate) of a conventional server-based model, on both general search traffic and long-tail phrases
- Is **faster** in terms of **latency** (endpointer, computational) compared to server-based model
- Consumes **lower power** than Pixel 4 / 5 for both short and long form audio.

Pixel 4/5 Model

- Model Details
  - 105M param LSTM Encoder
  - 20M param LSTM Decoder
  - 54M Transformer Rescorer + Additional LSTM Encoder
  - Model trained only on multidomain audio-text pairs



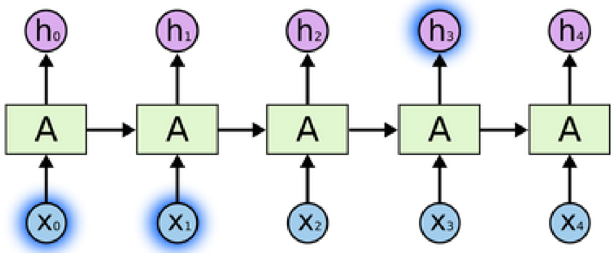
Pixel 6 Specifications

- Pixel 6 Google Tensor SoC Hardware
  - Google Tensor on-device edge TPU
  - 8 CPU cores
  - 8G (Pixel 6) or 12G (Pixel 6 Pro) DRAM
- We want an architecture where
  - **Encoder** can be **parallelized**
    - LSTM encoder → conformer encoder
  - **Decoder** that can **small** enough fit into SRAM
    - LSTM decoder → embedding decoder
  - **2nd-pass** is **streaming** for long-form
    - Transformer Rescorer → Multi-rate encoders
  - Does well on **long-tail named entities**
    - Additional Conformer Language Model

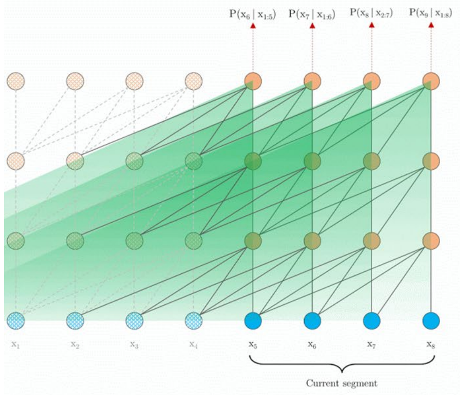
Latency Improvement: LSTM Encoder → Conformer Encoder

[Gulati et al, 2020][Zhang et al., 2020]

- LSTM
  - Sequential time dependency → not TPU friendly
  - Deletions in Long-form due to time dependency



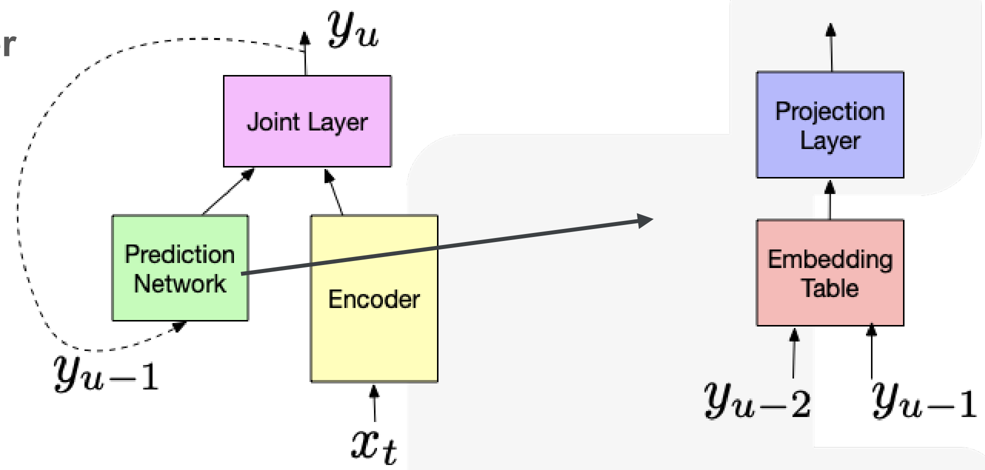
- Conformer/Transformer
  - Can compute multiple activations in parallel → more TPU friendly
  - Deletion issue less of a concern



Latency Improvement: LSTM Decoder→ Embedding Decoder

[Variani et al, 2020][Botros et al., 2021]

- Motivation:
  - Bottleneck of TPU is the **latency transfer** of parameters between CPU and SRAM on the TPU
  - If decoder is small enough, it can fit inside local SRAM on TPU
- Replace the LSTM decoder (33M params) with a simple embedding decoder lookup table (~2M params)
- This gives **~30% computation speedup** with no accuracy degradation

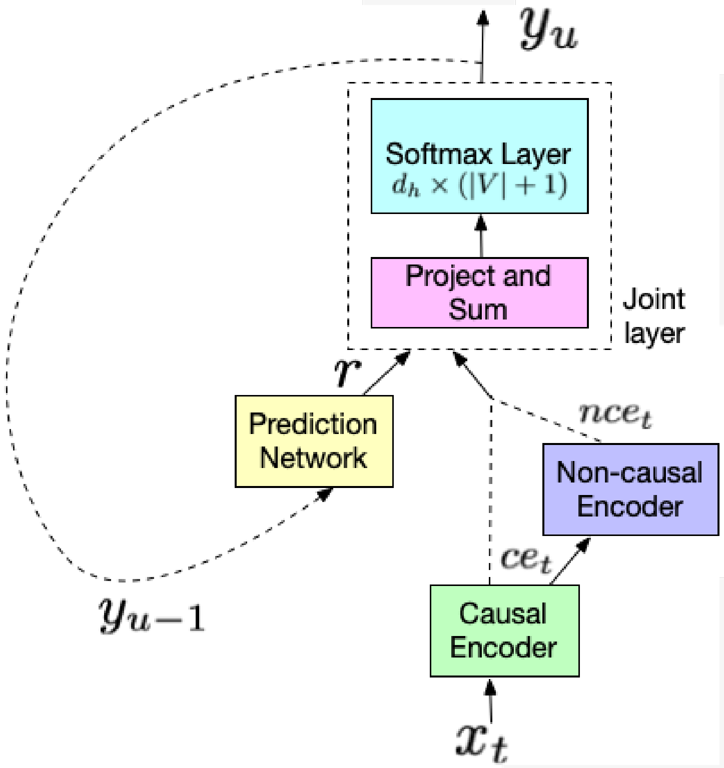


Quality Improvement: Multi-rate Encoders

[Tripathi et al, 2021][Yu et al., 2021][Narayanan et al., 2021]

- Segment and rescore (Pixel 5 architecture)
  - Not ideal for long-form
  - Beam search is better then rescoring

- Multi-rate encoders (Cascaded encoders)
  - 1st-pass causal encoder →decoder
  - 2nd-pass additional non-causal layers → decoder
  - Quality improvement with beam search while still being able to run in real time



Quality Improvement: Neural Language Model

[Variani et al, 2020]

- Incorporate Hybrid Autoregressive Transducer (HAT) factorization to better integrate language model

$$y^* = \arg\max_y \left[\log p(y|x) - \lambda_2 \log p_{ILM}(y) + \lambda_1 \log p_{LM}(y)\right]$$

- Language models:
  - Perform shallow fusion with a contextual biasing FST
  - Perform rescoring with a conformer LM trained on ~100 billion text utterances. This helps to address the long-tail proper noun issue.

ASR | WER for Short Form: en -US General Test Sets

[Sainath et al, 2021]

NGA: en-US with Conformer Cascaded Encoder + Neural LM is the best ASR we have built

Test Set / Vertical (en_us)	Server WER [Classic]	Pixel 4/5 WER [LSTM]	Pixel 6 WER [Conformer]
VS (Voice Search)	7.3	6.0	5.6
2018_VS	7.5	6.9	6.4
VS_NOISY	9.6	9.9	8.2
NUMERIC	5.7	5.4	5.2
VA_PLANNING	4.2	4.4	3.1
ASSISTANT_ON_ANDROID	6.8	8.0	5.9

ASR | WER for Short Form: en-US Biasing Test Sets

[Zhao et al, 2019]

- Biasing is an attempt to adapt the priors baked into the speech models to better model information gained between training and inference (aka context)
- Common uses cases include contacts, media and apps

Vertical / Use Cases	Server WER [Classic] w/ biasing (w/o biasing)	Pixel 4/5 WER [LSTM] w/ biasing (w/o biasing)	Pixel 6 WER [Conformer] w/ biasing (w/o biasing)
Contacts	9.7 (17.2)	6.1 (15.5)	3.3 (14.0)
Media	7.2 (7.8)	4.2 (9.1)	3.5 (8.7)
Open Apps	6 (6.1)	3.5	2.6 (5.0)

Biasing on conformer models further accelerated the quality improvements across verticals

ASR | Quality Wins on SxS Live Traffic

SxS: Live voice search queries are recognized by both on-device conformer and server. Then both results are sent to human raters for comparison. The on-device conformer model has cascaded encoder and neural LM rescorer.

On-device Conformer vs Server (SxS)				
Win	Loss	Neutral	p-Value	Impact
120	36	334	<0.1%	5.7e-2

Win/Loss ratio: 120 / 36 , which means new model is much better than current server model

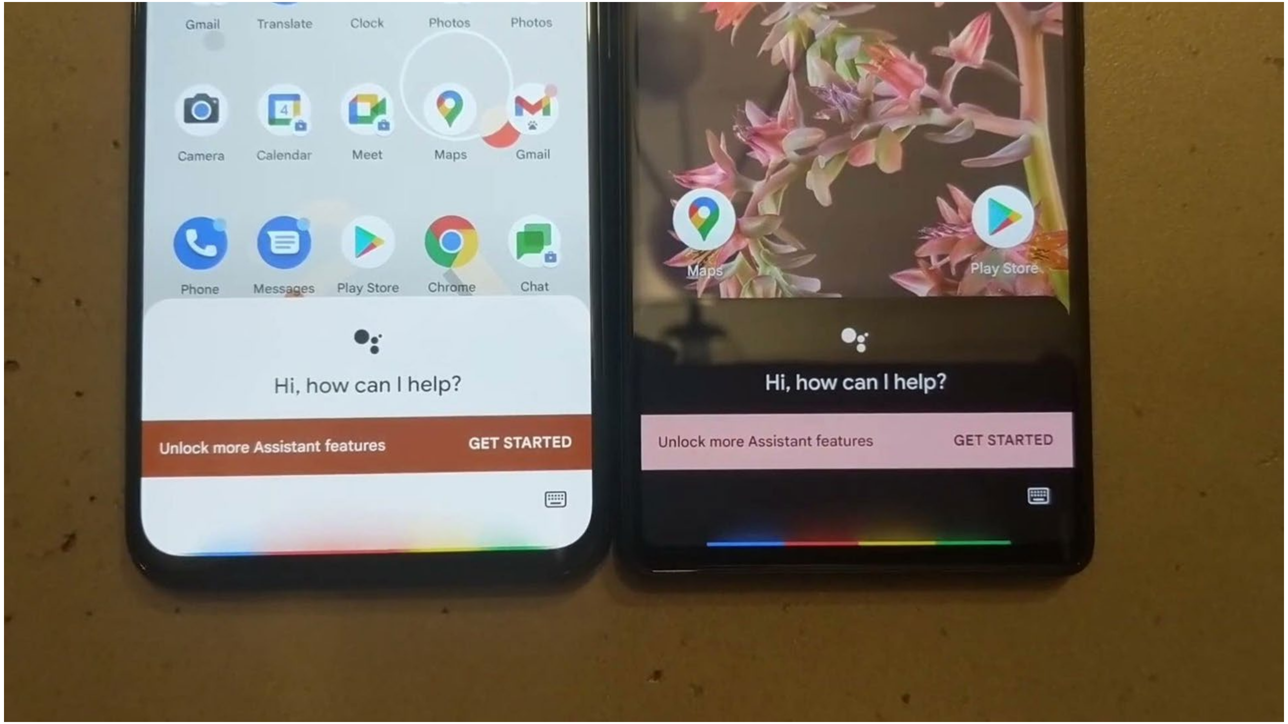
Video on Rare Words

Pixel4 (RNN-T + Transformer Rescorer)

X Bear County Texas

Pixel6 (Conformer + Cascade Encoder + Neural LM)

✓ Bexar County Texas



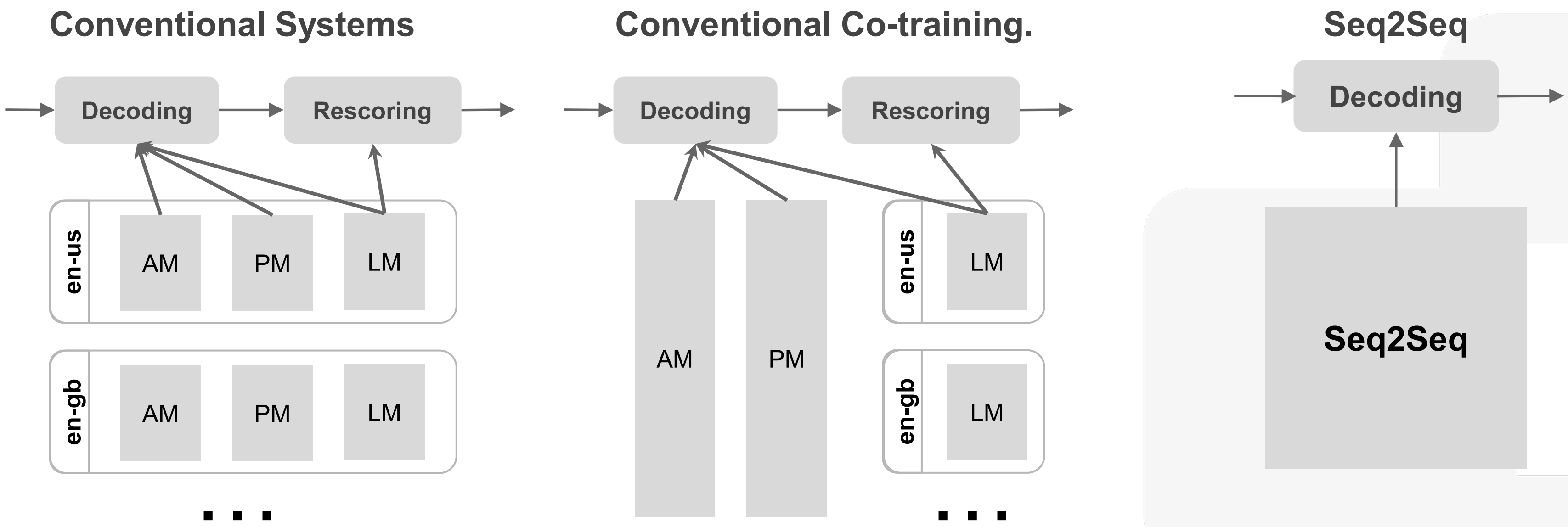
Pixel 6 Video

- **Summary:**
  - Attention-based End-to-End models (LAS) **achieves state-of-the-art performance**, but is **not streaming**.
  - Recurrent Neural Network Transducer (RNN-T) provides an **accurate** and **fast on-device** speech recognition experience.
  - [Pixel 4/5] RNN-T EP + 2nd-pass LAS **surpasses** server-side conventional model in both general search **quality** and **latency**.
  - [Pixel 6] Cascaded Encoder + neural LM **surpasses** server-side conventional model and Pixel 4/5 in both general search and long-tail **quality** and **latency**.

Future Challenges

E2E multi-dialect ASR

[Li et al., 2018, Toshniwal et al, 2018]



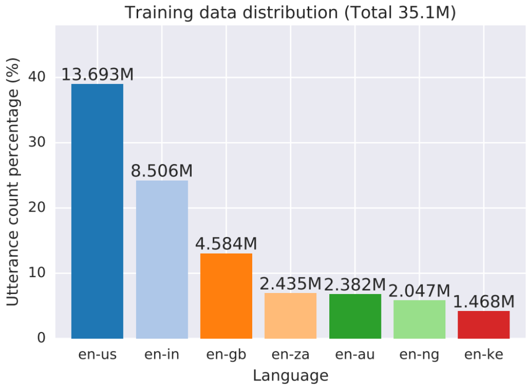
In conventional systems, languages/dialects, are handled with **individual AMs, PMs and LMs**. Upscaling is becoming challenging.

A single model for all.

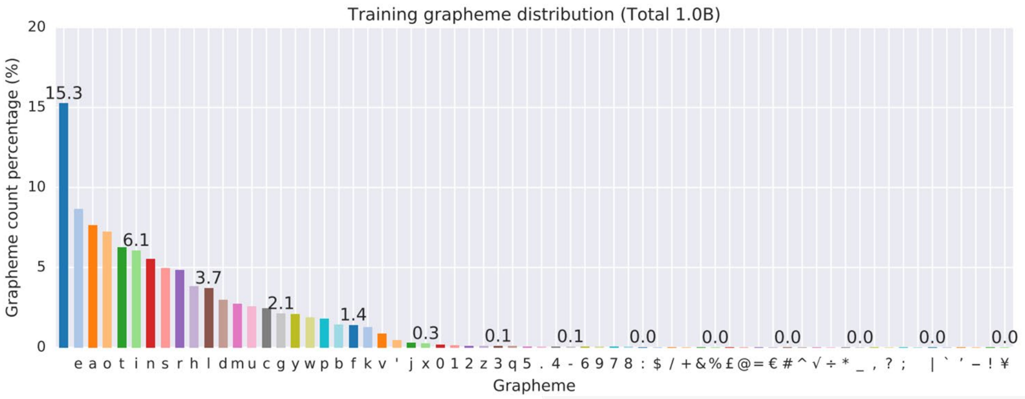
Task

[Li et al., 2018]

- **7 English dialects:** `US (America), IN (India), GB (Britain), ZA (South Africa), AU (Australia), NG (Nigeria & Ghana), KE (Kenya)`



★ **unbalanced** dialect data



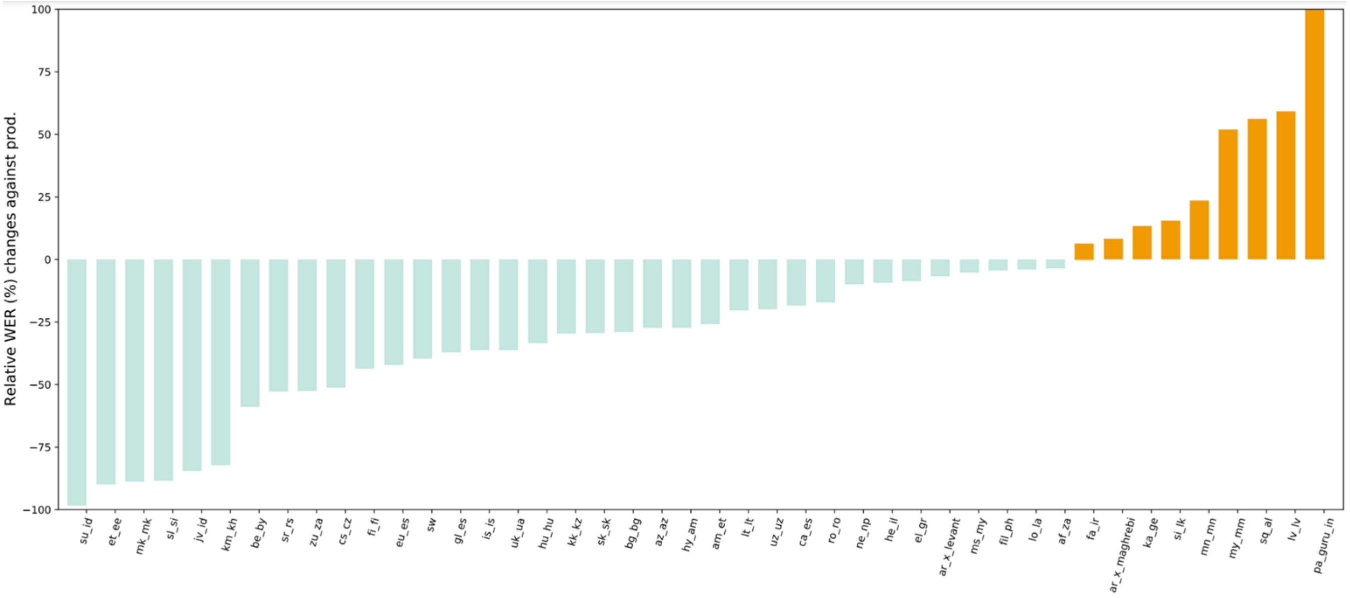
★ **unbalanced** target classes

E2E With Dialect as Input Features

Dialect	US	IN	GB	ZA	AU	NG	KE
Baseline (dialect-dep.)	9.7	16.2	12.7	11.0	12.1	33.4	19.0
encoder	9.6	16.4	11.8	10.6	10.7	31.6	18.1
decoder	9.4	16.2	**11.3**	10.8	10.9	32.8	18.0
both	**9.1**	**15.7**	11.5	**10.0**	**10.1**	**31.3**	**17.4**

★ feeding dialect to **both encoder and decoder** gives the largest gains

Question : Can we train **one model** on 44 low-resource languages?



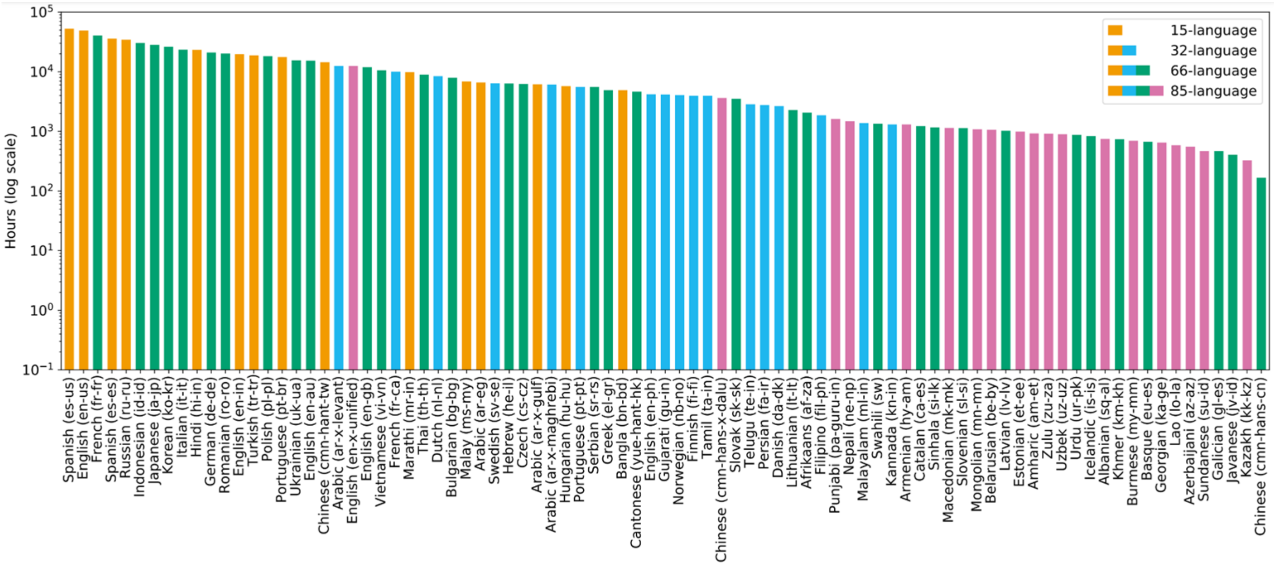
Key Result : Comparing to E2E trained on each individual language, 1 MMASR E2E model **wins** over **35 languages** .

Cross-lingual sharing boosts quality.

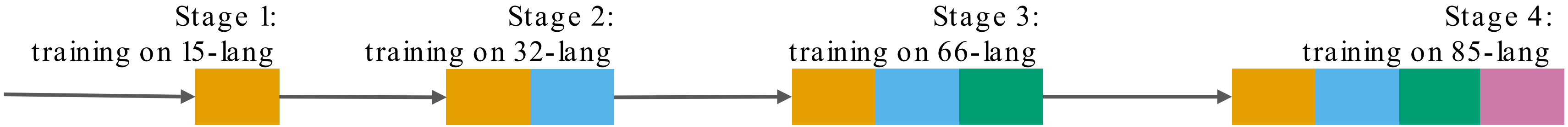
Infrastructure **simplified** with one model.

Task

Building **high quality** teacher models across languages.



With a fixed 1B-param model size, **lifelong learning** resolves the quality regressions.



Exp.	en-us	Avg. WER(%)			
		15-lang	32-lang	66-lang	85-lang
Monolingual	4.6	9.3	11.9	-	-
Training from scratch	5.4	10.4	13.3	11.5	12.3
Lifelong learning	**4.2**	**8.8**	**11.5**	**9.9**	**10.9**

[1] Massively Multilingal ASR: A Lifelong Learning Solution

Research Challenges

- How can we scale multi-lingual E2E for more languages?
- How can we maintain quality when the model is not fed a language-id?
- How do we handle code-switching within the utterance?
- How can we do this at an appropriate model size for on-device?

Thank You

Google

References

[Audhkhasi et al., 2017] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, D. Nahamoo "Direct Acoustics-to-Word Models for English Conversational Speech Recognition," Proc. of Interspeech, 2017.
[Bahdanau et al., 2017] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, Y. Bengio, "An Actor-Critic Algorithm for Sequence Prediction," Proc. of ICLR, 2017.
[Battenberg et al., 2017] E. Battenberg, J. Chen, R. Child, A. Coates, Y. Gaur, Y. Li, H. Liu, S. Satheesh, D. Seetapun, A. Sriram, Z. Zhu, "Exploring Neural Transducers For End-to-End Speech Recognition," Proc. of ASRU, 2017.
[Botros et al, 2021] R. Botros et al, "Tied & Reduced RNN-T Decoder," in Proc. Interspeech, 2021.
[Chan et al., 2015] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," CoRR, vol. abs/1508.01211, 2015.
[Chang et al., 2017] S-Y. Chang, B. Li, T. N. Sainath, G. Simko, C. Parada, "Endpoint Detection using Grid Long Short-Term Memory Networks for Streaming Speech Recognition," Proc. of Interspeech, 2017.
[Chang et al., 2018] S-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, O. Vinyals, "Temporal Modeling Using Dilated Convolution and Gating for Voice-Activity-Detection," Proc. of ICASSP, 2018.
[Chiu and Raffel, 2017] C.-C. Chiu, C. Raffel, "Monotonic Chunkwise Alignments," Proc. of ICLR, 2017.
[Chiu et al., 2018] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, M. Bacchiani, "State-of-the-art Speech Recognition With Sequence-to-Sequence Models," Proc. of ICASSP, 2018.
[Chorowski et al., 2015] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in Proc. of NIPS, 2015.
[Graves et al., 2006] A. Graves, S. Fernandez, F. Gomez, J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," Proc. of ICML, 2006.
[Graves, 2012] A. Graves, "Sequence Transduction with Recurrent Neural Networks," Proc. of ICML Representation Learning Workshop, 2012.
[Graves et al., 2013] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Neural Networks," in Proc. ICASSP, 2013.

References

[Gulati et al., 2020] Conformer: Convolution-augmented Transformer for Speech Recognition, in Proc. ICASSP, 2020.
[Gulcehre et al., 2015] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, Y. Bengio, "On Using Monolingual Corpora in Neural Machine Translation", CoRR, vol. abs/1503.03535, 2015.
[Hannun et al., 2014] A. Hannun, A. Maas, D. Jurafsky, A. Ng, "First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs," CoRR, vol. abs/1408.2873, 2014.
[He et al., 2017] Y. He, R. Prabhavalkar, K. Rao, W. Li, A. Bakhtin and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," Proc. of ASRU, 2017.
[He et al., 2018] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S-Y. Chang, K. Rao, A. Gruenstein, "Streaming End-to-end Speech Recognition For Mobile Devices," CoRR, vol. abs/1811.06621, 2018.
[Jaitly et al., 2016] N. Jaitly, D. Sussillo, Q. V. Le, O. Vinyals, I. Sutskever, S. Bengio, "An Online Sequence-to-Sequence Model Using Partial Conditioning," Proc. of NIPS, 2016.
[Kannan et al., 2018] A.Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," Proc. of ICASSP, 2018.
[Kim and Rush, 2016] Y. Kim and A. M. Rush, "Sequence-level Knowledge Distillation," Proc. of EMNLP, 2016.
[Kim et al., 2017] S. Kim, T. Hori and S. Watanabe, "Joint CTC-attention based End-to-End Speech Recognition using Multi-Task Learning," Proc. of ICASSP, 2017.
[Kingsbury, 2009] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," Proc. of ICASSP, 2009.
[Li et al., 2018] B. Li, T. N. Sainath, K. C. Sim, M. Bacchiani, E. Weinstein, P. Nguyen, Z. Chen, Y. Wu, K. Rao, "Multi-Dialect Speech Recognition With A Single Sequence-To-Sequence Model," Proc. of ICASSP, 2018.
[Li et al., 2020] B. Li, S. Y. Chang, T. N. Sainath, R. Pang, Y. He, T. Strohman, Y. Wu, "Towards Fast and Accurate Streaming End-To-End ASR" Proc. of ICASSP, 2020.

References

[Maas et al., 2015] A. Maas, Z. Xie, D. Jurafsky, A. Ng, "Lexicon-Free Conversational Speech Recognition with Neural Networks," Proc. of NAACL-HLT, 2015.

[McGraw et al., 2016] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, K. Rao, D. Rybach, O. Alsharif, H. Sak, A. Gruenstein, F. Beaufays, C. Parada, "Personalized speech recognition on mobile devices", Proc. of ICASSP, 2016

[Narayanan et al., 2021] A. Narayanan et al., "Cascaded encoders for unifying streaming and non-streaming ASR," in Proc. of ICASSP, 2021.

[Rabiner, 1989] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proc. of IEEE, 1989.

[Prabhavalkar et al., 2017] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, N. Jaitly, "A Comparison of Sequence-to-Sequence Models for Speech Recognition," Proc. of Interspeech, 2017.

[Prabhavalkar et al., 2018] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, A. Kannan, "Minimum Word Error Rate Training for Attention-based Sequence-to-Sequence Models," Proc. of ICASSP, 2018.

[Povey, 2003] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition", Ph.D. thesis, Cambridge University Engineering Department, 2003.

[Pundak et al., 2018] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, D. Zhao, "Deep context: end-to-end contextual speech recognition," Proc. of SLT, 2018.

[Rao et al., 2017] K. Rao, H. Sak, R. Prabhavalkar, "Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer", Proc. of ASRU, 2017.

[Ranzato et al., 2016] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," Proc. of ICLR, 2016.

[Sainath et al., 2018] T. N. Sainath, C.-C. Chiu, R. Prabhavalkar, A. Kannan, Y. Wu, P. Nguyen, Z. Chen, "Improving the Performance of Online Neural Transducer Models," Proc. of ICASSP, 2018.

[Sainath et al., 2019] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.C Chiu,, "Two-Pass End-to-End Speech Recognition," Proc. of Interspeech, 2019.

References

[Sainath et al, 2021] T.N. Sainath, Y. He, A. Narayanan, et al "An Efficient Streaming Non-Recurrent On-Device End-to-End Model with Improvements to Rare-Word Modeling," in Proc. Interspeech, 2021.

[Sainath et al., 2020] T. N. Sainath, Y. He, et. al., "A streaming on-device End-To-End model surpassing server-side conventional model quality and latency," Proc. of ICASSP, 2020.

[Sak et al., 2015] Hasim Sak, Andrew Senior, Kanishka Rao, Francoise Beaufays, "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition," Proc. of Interspeech, 2015.

[Sak et al., 2017] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in Proc. of Interspeech, 2017.

[Schuster & Nakajima, 2012] M. Schuster and K. Nakajima, "Japanese and Korean Voice Search," Proc. of ICASSP, 2012.

[Shannon, 2017] M. Shannon, "Optimizing expected word error rate via sampling for speech recognition," in Proc. of Interspeech, 2017.

[Sim et al., 2017] K. Sim, A. Narayanan, T. Bagby, T. N. Sainath, and M. Bacchiani, "Improving the Efficiency of Forward-Backward Algorithm using Batched Computation in TensorFlow," Proc. of ASRU, 2017.

[Sriram et al., 2018] A. Sriram, H. Jun, S. Satheesh, A. Coates, "Cold Fusion: Training Seq2Seq Models Together with Language Models," Proc. of ICLR, 2018.

[Stolcke et al., 1997] A. Stolcke, Y. Konig, M. Weintraub, "Explicit word error minimization in N-best list rescoring," Proc. of Eurospeech, 1997.

[Su et al., 2013] H. Su, G. Li, D. Yu, and F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription," Proc. of ICASSP, 2013.

[Szegedy et al., 2016] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Proc. of CVPR, 2016.

References

[Toshniwal et al., 2018] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, K. Rao, “Multilingual Speech Recognition With A Single End-To-End Model,” Proc. of ICASSP, 2018.

[Tripathi et al., 2021] A. Tripathi et al., “Transformer transducer: One model unifying streaming and non-streaming speech recognition,” Proc. of ICASSP, 2021.

[Variani et al., 2020] E. Variani et al, “Hybrid Autoregressive Transducer,” in Proc. ICASSP, 2020.

[Vaswani et al., 2017] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” Proc. of NIPS, 2017.

[Wiseman and Rush, 2016] S. Wiseman and A. M. Rush, “Sequence-to-Sequence Learning as Beam Search Optimization,” Proc. of EMNLP, 2016.

[Yu et al., 2021] J. Yu et al., “Dual-mode ASR: Unify and improve streaming ASR with full-context modeling”, Proc. of ICLR, 2021.

[Zhao et al., 2019] D. Zhao, T.N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li and R. Pang, "Shallow-Fusion End-to-End Contextual Biasing," in Proc. Interspeech, 2019.

[Zhang et al., 2020] Q. Zhang et al, “Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss,” in Proc. ICASSP, 2020.