



A Language Agnostic Multilingual Streaming On-Device ASR System

Bo Li, Tara N. Sainath, Ruoming Pang*, Shuo-yiin Chang, Qiumin Xu, Trevor Strohman, Vince Chen, Qiao Liang, Heguang Liu, Yanzhang He, Parisa Haghani, Sameer Bidichandani

Google LLC, USA

{boboli, tsainath}@google.com

Abstract

On-device end-to-end (E2E) models have shown improvements over a conventional model on English Voice Search tasks in both quality and latency. E2E models have also shown promising results for multilingual automatic speech recognition (ASR). In this paper, we extend our previous capacity solution to streaming applications and present a streaming multilingual E2E ASR system that runs fully on device with comparable quality and latency to individual monolingual models. To achieve that, we propose an Encoder Endpointer model and an End-of-Utterance (EOU) Joint Layer for a better quality and latency trade-off. Our system is built in a language agnostic manner allowing it to natively support intersentential code switching in real time. To address the feasibility concerns on large models, we conducted on-device profiling and replaced the time consuming LSTM decoder with the recently developed Embedding decoder. With these changes, we managed to run such a system on a mobile device in less than real time.

Index Terms: on-device, multilingual, code switching

1. Introduction

60% of the world’s population speaks more than one language [1]. Most commercial speech recognition systems, however, are designed and optimized in a monolingual setup. A common solution is to use multiple recognizers, one for each language, and output the hypothesis with the highest confidence score [2]. As each recognizer runs independently, the serving cost increases linearly with the number of languages needed, which is impractical for on-device applications. It is also challenging for intrasentential code-switching speech [3], where recognizers of different languages are needed to decode different parts of a single utterance. Alternatively, a language ID (LID) system can be used to detect and split the speech into monolingual segments before recognition [4], which incurs extra latency. Developing language agnostic multilingual ASR systems is thus critical to enable a natural human-computer speech interaction for the large multilingual user community.

The fast development of end-to-end (E2E) models [5–14] has largely improved automatic speech recognition (ASR) performance, especially for on-device applications. They run directly on users’ devices which saves network latency. Users’ audio data never leaves the device, which provides better privacy. There have been many successful efforts in building monolingual on-device streaming ASR systems with better quality and latency [15–21]. Multilingual systems, however, have been mainly focusing on non-streaming applications [22–26]. In [27], a streaming Recurrent Neural Network Transducer (RNN-T) [14] system was developed for Indian languages. To address the cross language variations, language dependent adapters was used, which requires knowing the target language information

*Work done in Google.

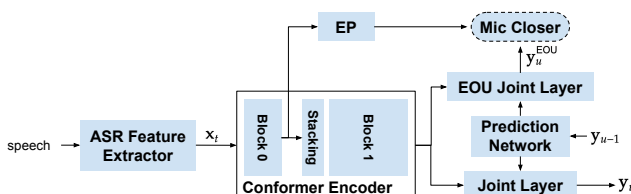


Figure 1: The proposed E2E multilingual streaming ASR system (EP: Endpointer; EOU: End-of-Utterance).

beforehand. [28] developed a configurable multilingual model that is trained once and can be configured as different language agnostic recognizers. Few works have touched on the latency aspect of such models. In this paper, we extend different work on improving latency [18] to multilingual ASR to fill in this gap. We further propose an Encoder Endpointer model and an End-of-Utterance (EOU) Joint Layer for a better quality and latency trade-off. Besides, we train our multilingual E2E ASR system by pooling data from all the languages without explicitly relying on language information. We find that given a sufficient model capacity, such multilingual systems can outperform monolingual models in both quality and latency. Moreover, these language agnostic multilingual models recognize code-switching speech with lower latency, different from existing work where LID is normally used [29–32].

2. System

In this section, we present the architecture of the proposed multilingual streaming ASR system (Figure 1).

2.1. E2E ASR Model

A Conformer [33] based RNN-T [14] model has recently been shown to outperform a strong hybrid ASR model on English [18, 21]. It consists of an encoder, a prediction network and a joint layer. The encoder consists of multiple Conformer layers [33] that are arranged into two blocks with a stacking layer in between (Figure 1). The stacking layer operates in the time dimension by concatenating two adjacent frames to achieve a 2x time reduction. The prediction network is a stack of LSTM layers that summarizes history predictions into a hidden representation. The feed-forward joint layer then combines the encoder and the prediction network outputs to predict a word-piece (WPM) token given the speech inputs. E2E models are often optimized without alignment information and tend to delay token predictions, which can be addressed by FastEmit [34].

In this work, we adopt the same Conformer RNN-T model as [18] for the multilingual system. A truly multilingual system should not only recognize speech from different languages, but also deal well with language/code switching with lower latency. To achieve that, we drop the input language ID [22, 35] and use a global input mean and standard deviation normaliza-

tion and a 16K output WPM obtained from the training data pooled across all the languages. The large cross-language variation poses a challenge in multilingual training. With the same model capacity, multilingual models tend to perform worse than monolingual models due to the reduced per-language capacity. With larger models, full-context multilingual models can achieve similar qualities as monolingual ones [35]. In this study, we further verify this on streaming models.

2.2. Encoder Endpointer Model

A streaming ASR system usually uses an *Endpointer (EP)* model for better latency (Figure 1). In Voice Search, an *End-of-Utterance (EOU)* EP is normally used. It classifies each input speech frame to be either speech, initial silence, intermediate silence or final silence [36, 37]. *Mic Closer* then uses this decision to trigger the microphone closing event, which will then end the current speech recognition session and move to the following execution stage. This is referred to as acoustic endpointing [38, 39] and is critical for a low-latency speech experience.

EP and ASR models are usually developed and maintained independently. Scaling to more languages requires training and maintaining both models. A truly multilingual system requires both the recognizer and the EP model to be multilingual. To address this, we propose an *Encoder Endpointer* model that branches an EP model from the ASR encoder’s lower layers. Specifically, as depicted in Figure 1, the EP model sits on top of the ASR encoder’s block 0 before the stacking layer. In this way, EP predictions still have the same frame rate as the input frames while sharing the computation of block 0 with the ASR encoder. This synchronizes the model update for ASR and EP.

2.3. EOU Joint Layer

For better latency, RNN-T is normally extended to RNN-T EP [40] to explicitly emit an End-of-Utterance (EOU) token. This EOU token utilizes ASR’s prediction history, brings extra linguistic context information to acoustic EP and has been shown to yield better microphone closing decisions. This is referred to as decoder endpointing [38, 39].

Since the EOU token only appears at the sentence end during training, E2E model learns to stop emitting tokens after it. This can cause early cutoffs when inaccurate EOU predictions are made. Besides, we also find the latency benefit of the EOU prediction comes with slight quality degradations [40, 41]. In this paper, we propose a simple change to the RNN-T EP model by using an additional joint layer, namely the EOU Joint Layer. As illustrated in Figure 1, we keep the original joint layer unchanged and add an extra joint layer for the prediction of word-piece tokens together with EOU. To ensure the same recognition quality, we conduct the training in two stages. Firstly, we optimize all the encoder, prediction network and the original joint layer for recognition quality; then we freeze all those parameters and initialize a new EOU Joint Layer from the original joint layer and fine-tune only this EOU Joint Layer. During inference, we rely on the original joint layer for beam search decoding and only use the EOU Joint Layer to generate the EOU probability.

3. Experimental Setup

3.1. Data

We use a dataset consisting of speech from 9 Voice Search locales with a total number of 142.3M utterances. The total du-

Table 1: *Per language training data statistics. Utterance counts are in millions (M) and duration is in thousand (K) hours.*

Locale	Language	Counts(M)	Hours(K)
en-US	English (USA)	34.0	52.6
zh-TW	Chinese (Taiwan)	16.6	22.0
fr-FR	French	16.5	23.7
de-DE	German	15.3	23.4
ja-JP	Japanese	14.9	20.2
es-US	Spanish (USA)	14.2	23.8
es-ES	Spanish (Spain)	12.9	20.1
it-IT	Italian	11.8	19.8
en-GB	English (UK)	6.0	8.6
Total		142.3	214.2

ration is around 214.2K hours. The data is anonymized and human transcribed. The per language data distribution is listed in Table 1. The number of utterances for each language ranges from 6.0M to 34.0M, roughly corresponding to 8.6K to 52.6K hours of speech data. The test set for each language contains around 3.0K-15.4K utterances less than 5.5s in length (Table 2). They are also sampled from Voice Search traffic with no overlap from the training set. Similarly, they are anonymized and hand-transcribed for evaluation purpose.

3.2. Model Architecture

We use 80-dimensional(-dim) log Mel filterbank features that are computed on 32ms windows with a 10ms hop. Features from 3 contiguous frames are stacked and sub-sampled to form a 240-dim input representation with 30ms frame rate. SpecAugment [42] is used to improve models’ robustness against noise. Specifically, two frequency masks with a maximum length of 27 and two time masks with a maximum length of 50 are used.

Following [21], we use 512-dim Conformer layers in the ASR encoder. Causal convolution and left-context attention layers are used for the Conformer layer to strictly restrict the model to use no future inputs. 8-head attention is used in the self-attention layer and the Convolution kernel size used is 15. The encoder consists of 12 Conformer layers, separated into two blocks by a stacking layer. Block 0 consists of an input projection layer and 3 Conformer layers. The time stacking layer concatenates two adjacent frame outputs from block 0 with no overlap to form a 60ms frame. After that, block 1 first uses a 1024-dim Conformer layer, then a projection layer to reduce the model dim back to 512 and lastly 8 Conformer layers followed by a layer normalization. The RNN-T decoder consists of a prediction network with 2 LSTM layers with 2,048 units projected down to 640 output units, and a joint network with a single feed-forward layer of 640 units. All the E2E ASR models are trained to predict 16,384 word pieces generated from the transcripts pooled across all the languages. The final model has 110M parameters in the encoder and 33M parameters in the decoder. It is optimized by minimizing the RNN-T loss. FastEmit [34] regularization with a weight of 5e-3 is used to reduce prediction latency.

For the encoder endpointer model, we firstly project down the ASR encoder block 0’s output to 128-dim and then use a single layer of 128-dim Conformer layer. The output is further project down to 4-dim and layer normalized before the final softmax layer. The model has 449K additional parameters specific for endpointing. Per-frame cross entropy loss is used to train these endpointer specific parameters.

Table 2: WER (%) comparisons of different systems without Endpointer (EP).

System	Size (M)	Per Language WER (%)									Avg. WER (%)
		en-US	zh-TW	fr-FR	de-DE	ja-JP	es-US	es-ES	it-IT	en-GB	
Utterance Counts (K)	-	5.0	15.4	3.4	12.0	14.1	3.0	12.5	14.2	7.9	-
Monolingual Conformer RNN-T	140 * 9	6.3	4.7	12.5	12.9	10.5	7.8	7.3	7.2	7.2	8.49
S1 with input LID	140	6.1	5.2	12.6	13.3	11.9	7.0	6.9	7.7	6.1	8.53
S2 truly multilingual	140	6.3	5.4	12.8	13.6	12.0	7.2	7.1	8.1	6.3	8.76
S3 S2 with FastEmit	140	7.0	5.6	12.8	14.0	12.5	7.4	7.3	8.5	6.5	9.07
S4 S3 with EOU	140	7.5	6.1	13.2	14.4	13.0	7.9	7.6	9.0	7.0	9.52
S5 deeper and wider S3	500	6.1	4.9	12.7	13.2	11.1	6.9	6.8	7.2	5.9	8.31
S6 even wider S5	1000	5.6	4.8	12.6	12.6	10.4	7.0	6.5	6.6	5.6	7.97
S7 S6 with Embedding decoder	920	6.1	5.0	12.5	12.8	11.3	7.0	6.7	7.1	5.9	8.27

All the models are trained in Tensorflow using the Lingvo toolkit [43] on Google’s Tensor Processing Units (TPU) V3 [44] with a global batch size of 4,096 utterances. Models are trained with 512 TPU cores and optimized using synchronized stochastic gradient descent. We use the Adam optimizer [45] with parameters $\beta_1=0.9$ and $\beta_2=0.999$. A transformer learning rate schedule with peak learning rate $1.8e-3$ and 32K warm-up steps is used. Exponential moving average [46] has been used to stabilize the model weight updates.

4. Results

4.1. A Truly Multilingual ASR System

To build a truly multilingual system (S2 in Table 2), we pool all the data from different languages together with its natural distribution. Different from previous work [22, 35], we do not feed any language information into the model. We use the architecture described in Section 3.2, specifically a 12-layer Conformer encoder and 2-layer LSTM decoder streaming RNN-T model. We compare with two baselines: 1) monolingual models (Monolingual in Table 2) and 2) a multilingual model with a 16-dim 1-hot LID vector concatenated to the filterbank features (S1 in Table 2). All the three systems use the same model architecture as described in Section 3.2, except that the monolingual models use language dependent WPM outputs. From the results in Table 2, multilingual models degrade on languages with the most data but outperforms monolingual models on es-US, es-ES and en-GB which have slightly less data. Oracle language information (S1) brings quality gains on monolingual test sets, but limits the capability in recognizing speech with code-switching phenomenon. For example, to recognize an utterance with different languages (Figure 2), S1 requires language ID information before recognition. However, with S2 we can natively switch language during the recognition. We hence take the multilingual model without language ID (S2) for the following studies.

Due to the lack of proper code-switching test sets, we conducted some preliminary user studies to understand S2’s capability in handling such speech. It performs surprisingly well despite the fact that we do not have code-mixing data during training. In one example as shown in Figure 2, S2 mistakenly outputs Chinese characters after the user switches from Chinese to Japanese (red colored texts in Figure 2), but it manages to correct itself after seeing more Japanese speech and outputs the correct transcript at the end. Future work will investigate intrasentential code switching in great details.

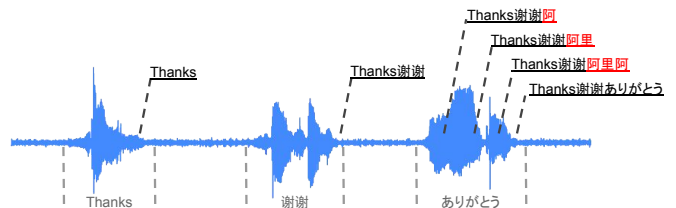


Figure 2: A code-switching example: a user speaking “thanks” in English, Chinese and Japanese. Texts below the waveform are the ground truth and those above are the top partial hypotheses at corresponding time (not all top partial hypotheses are plotted).

Table 3: Final silence accuracy (%) of different EOU Endpointer models.

Model	Architecture	Size(K)	Accuracy(%)
EP0	Standalone (3-layer LSTM)	462.5	91.6
EP1	Encoder Endpointer	2.0	88.3
EP2	EP1 + 3-layer LSTM	462.5	91.7
EP3	EP1 + 1-layer Conformer	448.7	93.8

4.2. Reducing Latency

Besides quality, latency is another critical metric for streaming applications. We compare different endpointer models for the streaming multilingual ASR system. We report only the frame accuracy for the final silence class which is used in the recognition pipeline for microphone closing decisions. We first train a conventional standalone 3-layer 128-dim LSTM endpointer (EP0 in Table 3) [37] as our baseline. For the proposed encoder endpointer, we explored a few different architectures including: a single projection layer (EP1), EP1 with additional 3-layer 128-dim LSTM (EP2), and EP1 with a comparable size but using a single Conformer layer (EP3). The per frame final silence prediction accuracy of these models together with the number of additional parameters brought by these EP models are listed in Table 3. EP1 has the lowest accuracy suggesting the need for endpointer specific model parameters. Simply projecting down the ASR encoder’s block 0 output for endpointing is not sufficient. Another potential solution is to jointly train the whole system with both ASR and endpointer targets. This way the encoder block 0 is aware of the endpointer task in training, which will be explored in future work. Next, with the same number of additional parameters (EP0 vs. EP2), encoder endpointer yields better accuracy, benefiting from the pretrained ASR encoder.

Table 4: *Quality and Latency (EP50: 50-percentile and EP90: 90-percentil) comparisons between the proposed multilingual model (S3) and classical hybrid monolingual models.*

Metric	System	Per Language									Avg.
		en-US	zh-TW	fr-FR	de-DE	ja-JP	es-US	es-ES	it-IT	en-GB	
WER (%)	Monolingual classic	7.5	4.9	12.9	13.7	10.8	9.3	8.1	8.2	6.4	9.09
	S3	11.0	6.6	15.0	16.4	15.1	8.7	9.3	11.3	9.8	11.47
EP50 (ms)	Monolingual classic	430	370	380	300	510	440	440	540	510	436
	S3	330	450	470	200	420	320	450	530	350	391
EP90 (ms)	Monolingual classic	880	780	720	740	890	840	820	990	960	847
	S3	670	630	660	570	780	690	670	800	730	689

Lastly, after switching from LSTM to Conformer, we obtained the best 93.8% final silence prediction accuracy.

E2E models trained without explicit alignment constraints tend to delay output predictions, which increases the latency for streaming applications. To address that, we adopted FastEmit [34] in our multilingual training. Similar to existing findings [18], FastEmit hurts WER (comparing S3 to S2 in Table 2). Besides reducing RNN-T’s inherent token prediction latency, we also utilize decoder endpointer to reduce the microphone closing latency. We append an EOU token to the end of each training WPM sequence. On top of FastEmit, we found no additional late penalty [40] is needed for the EOU prediction. From the results (S4 in Table 2), we can see adding EOU token further degrades the recognition quality. To address this problem, we move the the EOU prediction into a separate EOU Joint Layer and keep the existing RNN-T model unchanged. In this way, we can maintain the same quality as S3 while still benefiting from the EOU prediction for a fast decoder endpointing.

4.3. Comparison with Monolingual Models

With all the above components, we conducted an end-to-end test with both endpointer and ASR models. For monolingual RNN-T models (Monolingual in Table 2), we need to conduct language specific parameters tuning to find the best quality and latency trade-offs for a fair comparison, which is a time consuming task. Instead, we compared with the current best production models for Voice Search, which are extensively tuned hybrid monolingual ASR systems (Monolingual in Table 4) consisting of acoustic models, lexicons and language models [21]. The final quality and latency metrics are depicted in Table 4. EP is crucial to streaming application latency but hurts WER (comparing the WER for S3 in Table 4 to Table 2). Our single 140M-param multilingual model (S3) lags behind monolingual models, except that on es-US, S3 achieves better quality. Moreover, the latency of the proposed multilingual system S3 is much better, especially in terms of the 90th percentile latency (EP90 in Table 4).

4.4. Further Quality Improvements

To address the quality gap between the proposed multilingual model and the monolingual models, we hypothesize the main problem is the limited model capacity [35]. The same number of model parameters as each monolingual model now needs to capture more variations across languages. Similar to [35], we increase S2’s encoder width from 512 to 640 and depth from 12 to 35, and decoder width from 2048 to 3072 and depth from 2 to 6, leading to a 500M-param model S5. By further increasing the encoder width to 1024, it brings us a 1B-param model

S6. From Table 2, with 1B-param model, we can achieve better or comparable performance to each monolingual models. In terms of endpointer latency, namely EP50 and EP90, the larger models perform similarly to S3 in Table 4.

4.5. On-Device Benchmark

To test the feasibility of deploying these large models to mobile devices, we conducted some preliminary benchmark studies on Google Pixel 6 phones with Google Tensor [47] on 100 utterances. We report the on-device real-time factor and memory usage to show the computation efficiency and resources needed to serve such models (Table 5). For the 1B-param model (S6), it is impractical for streaming applications. However, after detailed analysis, we found the LSTM decoder takes up most of the computation. Recent work [48,49] has found that stateless decoders are sufficient for E2E models. We hence swapped the LSTM decoder of S6 with a 33M-param Embedding decoder (S7 in Table 2 and Table 5). It increases the WER but is still better than most monolingual models (Table 2). More importantly, it largely reduces the real-time factor and enables the deployment of such large models to mobile devices.

Table 5: *On-device benchmark for different sized models (RT50: Real-time factor, 50th percentile, RT90: Real-time factor, 90th percentile and MEM: peak memory usage).*

	140M(S2)	500M(S5)	1B(S6)	920M(S7)
RT50	0.43	1.76	1.96	0.61
RT90	0.80	2.70	3.05	0.70
MEM(G)	0.38	1.08	1.63	1.03

5. Conclusions

In this paper, we presented a language agnostic multilingual streaming on-device end-to-end system with comparable quality and latency metrics to monolingual models. On the quality side, we found a larger model capacity is needed to obtain better qualities than monolingual models. Our on-device benchmark study showed that it is feasible to run large models in less than real time on a modern mobile device. On the latency side, we developed an Encoder Endpointer multitask model to unify the training of the multilingual endpointer and the speech recognizer. We further proposed to use a separate End-of-Utterance (EOU) Joint Layer which reduces the endpointing latency with EOU prediction while maintaining the recognition quality. With the fast advancement in both software and hardware, deploying high-quality and low-latency multilingual on-device speech experience would become practical in the near future.

6. References

- [1] “Multilingual People,” <http://ilanguages.org/bilingual.php>, accessed: 2021-09-28.
- [2] J. Gonzalez-Dominguez, D. Eustis, *et al.*, “A real-time end-to-end multilingual speech recognition architecture,” *IEEE Journal of selected topics in signal processing*, vol. 9, no. 4, 2014.
- [3] B. E. Bullock and A. E. Toribio, *The Cambridge handbook of linguistic code-switching*. Cambridge University Press, 2009.
- [4] J. Weiner, N. T. Vu *et al.*, “Integration of language identification into a recognition system for spoken conversations containing code-switches,” in *Proc. SLTU*, 2012.
- [5] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, 2013.
- [6] W. Chan, N. Jaitly *et al.*, “Listen, attend and spell,” *CoRR*, vol. abs/1508.01211, 2015.
- [7] J. Chorowski, D. Bahdanau *et al.*, “Attention-based models for speech recognition,” in *Proc. NIPS*, 2015.
- [8] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. ICASSP*, 2017.
- [9] E. Battenberg, J. Chen *et al.*, “Exploring neural transducers for end-to-end speech recognition,” in *Proc. ASRU*, 2017.
- [10] J. Li, R. Zhao *et al.*, “Improving RNN transducer modeling for end-to-end speech recognition,” in *Proc. ASRU*, 2019.
- [11] Y. He, T. N. Sainath *et al.*, “Streaming end-to-end speech recognition for mobile devices,” in *Proc. ICASSP*, 2019.
- [12] A. Zeyer, A. Merboldt *et al.*, “A new training pipeline for an improved neural transducer,” in *Proc. Interspeech*, 2020.
- [13] J. Li, Y. Wu, Y. Gaur *et al.*, “On the comparison of popular end-to-end models for large scale speech recognition,” in *Proc. Interspeech*, 2020.
- [14] A. Graves, “Sequence transduction with recurrent neural networks,” *arXiv preprint arXiv:1211.3711*, 2012.
- [15] Q. Zhang, H. Lu *et al.*, “Transformer Transducer: A streamable speech recognition model with Transformer encoders and RNN-T loss,” in *Proc. ICASSP*, 2020.
- [16] T. N. Sainath, Y. He *et al.*, “A streaming on-device end-to-end model surpassing server-side conventional model quality and latency,” in *Proc. ICASSP*, 2020.
- [17] D. Le, M. Jain, G. Keren *et al.*, “Contextualized streaming end-to-end speech recognition with Trie-based deep biasing and shallow fusion,” in *Proc. Interspeech*, 2021.
- [18] B. Li, A. Gulati, J. Yu *et al.*, “A better and faster end-to-end model for streaming ASR,” in *Proc. ICASSP*, 2021.
- [19] Y. Shi, Y. Wang, C. Wu *et al.*, “Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition,” in *Proc. ICASSP*, 2021.
- [20] X. Chen, Y. Wu, Z. Wang, S. Liu, and J. Li, “Developing real-time streaming Transformer transducer for speech recognition on large-scale dataset,” in *Proc. ICASSP*, 2012.
- [21] 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 *Interspeech*, 2021.
- [22] B. Li, T. N. Sainath, K. C. Sim *et al.*, “Multi-dialect speech recognition with a single sequence-to-sequence model,” in *Proc. ICASSP*, 2018.
- [23] V. Pratap, A. Sriram, P. Tomasello *et al.*, “Massively multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters,” in *Proc. Interspeech*, 2020.
- [24] W. Hou, Y. Dong, B. Zhuang *et al.*, “Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning,” in *Proc. Interspeech*, 2020.
- [25] O. Adams, M. Wiesner, S. Watanabe *et al.*, “Massively multilingual adversarial speech recognition,” *arXiv preprint arXiv:1904.02210*, 2019.
- [26] S. Zhou, S. Xu, and B. Xu, “Multilingual end-to-end speech recognition with a single transformer on low-resource languages,” *arXiv preprint arXiv:1806.05059*, 2018.
- [27] A. Kannan, A. Datta, T. N. Sainath *et al.*, “Large-scale multilingual speech recognition with a streaming end-to-end model,” in *Proc. Interspeech*, 2019.
- [28] L. Zhou, J. Li, E. Sun, and S. Liu, “A configurable multilingual model is all you need to recognize all languages,” *arXiv preprint arXiv:2107.05876*, 2021.
- [29] D.-C. Lyu and R.-Y. Lyu, “Language identification on code-switching utterances using multiple cues,” in *Proc. Interspeech*, 2008.
- [30] Z. Zeng, Y. Khassanov, V. T. Pham *et al.*, “On the end-to-end solution to mandarin-english code-switching speech recognition,” *arXiv preprint arXiv:1811.00241*, 2018.
- [31] K. Li, J. Li, G. Ye *et al.*, “Towards code-switching ASR for end-to-end CTC models,” in *Proc. ICASSP*, 2019.
- [32] S. Dalmia, Y. Liu *et al.*, “Transformer-transducers for code-switched speech recognition,” in *Proc. ICASSP*, 2021.
- [33] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [34] J. Yu *et al.*, “FastEmit: Low-latency streaming ASR with sequence-level emission regularization,” in *ICASSP*, 2021.
- [35] B. Li, R. Pang, T. N. Sainath *et al.*, “Scaling end-to-end models for large-scale multilingual ASR,” in *Proc. ASRU*, 2021.
- [36] S.-Y. Chang, B. Li, T. N. Sainath *et al.*, “Endpoint detection using grid long short-term memory networks for streaming speech recognition,” in *Proc. Interspeech*, 2017.
- [37] S.-Y. Chang, B. Li, and G. Simko, “A unified endpointer using multitask and multidomain training,” in *Proc. ASRU*, 2019.
- [38] M. Shannon *et al.*, “Improved end-of-query detection for streaming speech recognition,” in *Proc. Interspeech*, 2017.
- [39] S.-Y. Chang, B. Li *et al.*, “Low latency speech recognition using end-to-end prefetching,” in *Proc. Interspeech*, 2020.
- [40] B. Li, S.-Y. Chang, T. N. Sainath *et al.*, “Towards fast and accurate streaming end-to-end ASR,” in *Proc. ICASSP*, 2020.
- [41] S.-Y. Chang, R. Prabhavalkar, Y. He *et al.*, “Joint endpointing and decoding with end-to-end models,” in *Proc. ICASSP*, 2019.
- [42] D. S. Park, Y. Zhang, C.-C. Chiu *et al.*, “SpecAugment on large scale datasets,” in *Proc. ICASSP*, 2020.
- [43] J. Shen, P. Nguyen, Y. Wu *et al.*, “Lingvo: a modular and scalable framework for sequence-to-sequence modeling,” *arXiv:2005.08100*, 2019.
- [44] “Cloud Tensor Processing Units (TPUs),” <https://cloud.google.com/tpu/docs>, accessed: 2021-09-30.
- [45] N. Shazeer and M. Stern, “Adafactor: Adaptive learning rates with sublinear memory cost,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 4596–4604.
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [47] “Google Tensor is a milestone for machine learning,” <https://blog.google/products/pixel/introducing-google-tensor/>, accessed: 2022-03-17.
- [48] R. Prabhavalkar, Y. He, D. Rybach, S. Campbell, A. Narayanan, T. Strohman, and T. N. Sainath, “Less is more: Improved rnn-t decoding using limited label context and path merging,” in *Proc. ICASSP*. IEEE, 2021, pp. 5659–5663.
- [49] R. Botros, T. N. Sainath, R. David, E. Guzman, W. Li, and Y. He, “Tied & Reduced RNN-T Decoder,” in *Proc. Interspeech*, 2021.