INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Mixture-of-Expert Conformer for Streaming Multilingual ASR

Ke Hu, Bo Li, Tara N. Sainath, Yu Zhang, Francoise Beaufays

Google LLC, USA

huk@google.com

Abstract

End-to-end models with large capacity have significantly improved multilingual automatic speech recognition, but their computation cost poses challenges for on-device applications. We propose a streaming truly multilingual Conformer incorporating mixture-of-expert (MoE) layers that learn to only activate a subset of parameters in training and inference. The MoE layer consists of a softmax gate which chooses the best two experts among many in forward propagation. The proposed MoE layer offers efficient inference by activating a fixed number of parameters as the number of experts increases. We evaluate the proposed model on a set of 12 languages, and achieve an average 11.9% relative improvement in WER over the baseline. Compared to an adapter model using ground truth information, our MoE model achieves similar WER and activates similar number of parameters but without any language information. We further show around 3% relative WER improvement by multilingual shallow fusion.

1. Introduction

An end-to-end (E2E) multilingual automatic speech recognition (ASR) model is appealing because of its potential to recognize multiple languages using a single model. There has been significant effort in multilingual modeling using E2E models [1, 2, 3, 4, 5, 6, 7]. Previous studies on multilingual ASR have investigated different model structures such as connectionist temporal classification (CTC) models [2], long-short term memory [4], and attention based models [1, 3, 5, 6]. Among them, streaming models [6, 8] are promising candidates for on-device applications. By increasing the model capacity, [8] proposes an on-device streaming multilingual RNN-T model and achieves comparable recognition quality and latency compared to monolingual models. To further increase model capacity, it is crucial to keep computation low for on-device applications.

Other studies have also shown that increasing model capacity is a key factor in improving performance. By increasing a multilingual E2E model up to 1B size, [9] has improved quality of all variants of the multilingual model. In [10], the authors show that under a life-long learning strategy, the model performs consistently better as the capacity increases up to 1B parameters. The same trend has been observed in a two-pass multilingual deliberation model [11]. More recently, the Whisper model [12] and Google Universal Speech Model (USM) [13] have achieved human-approaching performance with the help of large-scale data and model sizes in billions of parameters.

Larger models come with more cost in training and inference. To improve the modeling efficiency, there have been several approaches in leveraging language-specific components for inference [6, 14, 15, 16]. However, how to predict language information reliably in a streaming fashion is a challenge itself. Others improve efficiency by neural network pruning [17] or mixture-of-expert type of models [18, 19, 20, 21, 22] (more discussion in Sect. 2).

In this work, we propose to use mixture-of-expert (MoE) layers [23, 24] to replace the feed-forward network (FFN) in the Conformer [25] for multilingual ASR. The MoE layer consists of multiple FFNs and a gating network [24]. The gating network is a softmax over the number of experts, and the outputs of the top two experts are combined in a weighted fashion as the final output. The proposed model is thus sparse when the number of total experts is greater than two. Such an MoE layer has been used in NLP [23] as well as shallow fusion [26] and achieved superior quality compared to their dense counterparts. By adding the MoE layers to the end FFN network of the Conformer layers, we improve the average WER of 12 languages by 11.9% relative compared to the baseline. In another comparison with a larger baseline (dense model) with a similar total size as the MoE model, we show that the proposed model achieves similar quality by activating only 53% of parameters during inference.

We also compare to an adapter model based on [4, 27]. By increasing the total number of experts while activating the top two experts during inference, we achieve similar quality compared to a ground truth language information based adapter model. The two models activate the same number of parameters for inference, however, we note that our model does not need any language information in inference and more straightforward in deployment. Finally, we further improve the MoE model performance by around 3% relative by shallow fusion using a multilingual neural LM.

2. Related Work

There have been several related mixture-of-expert approaches for ASR modeling, but our work differs from them in the following ways. Our MoE study is for streaming multilingual ASR compared to [18, 22], which are for monolingual ASR. We show that our Conformer based MoE structure is effective for multiple languages in Sect. 5.2. Compared to [18, 22], our MoE model also works without using any shared embedding network for expert routing. Other MoE models have been proposed for image processing or natural language processing (NLP) [20, 21]. In terms of model structure, DeepMoE [20] uses embedding network for expert routing and the Switch Transformer [21] activates only one expert layer layer. Further, the performance of DeepMoE and Switch Transformer in ASR is unclear. In multilingual ASR, a mixture of informed-expert model is proposed in [16]. However, it needs language information to select the expert, and the number of experts increases linearly with the number of languages. In a similar line, a per-language second pass model [14] and the adapter model [4, 27] can also be considered as informed-expert models since the language information is used to choose the corresponding module (cascaded encoder or adapter). Compared to [14], the proposed MoE model does not rely on any external language information for routing and is thus more generic. Although [15] predicts the language information for the second-pass, this increases both

10.21437/Interspeech.2023-480

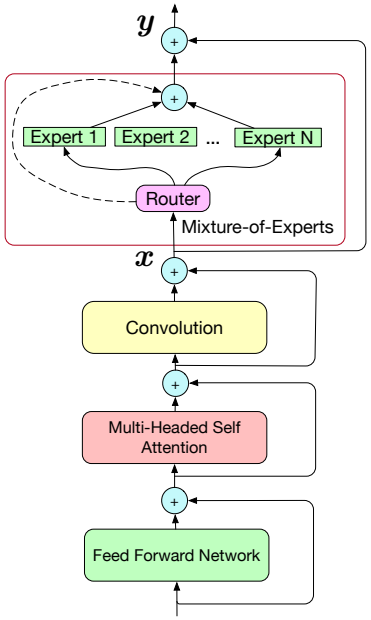

Figure 1: *Conformer layer with mixture of experts at the position of the end FFN layer.*

training and inference complexity and potential error propagation in practical applications.

In summary, the novelty of our work are mainly two folds: 1) We have proposed an MoE-based Conformer for multilingual ASR and have shown its effectiveness compared to dense models and adapter models, and 2) Our proposed MoE model is relatively simplistic, and does not require language information compared to [14, 16] and any shared embedding network for routing [18, 22].

3. Conformer with Mixture-of-Experts

Our model is based on [8] which uses a causal encoder and we add a non-causal cascaded encoder [28] for better quality. For both causal and non-causal encoders, we use Conformer as the main building block. We use separate decoders for causal and non-causal encoders and RNN-T loss for training (see Sect. 4.2.1 for more details).

A Conformer layer [25] consists of a multi-headed self-attention and a convolution-based layer sandwiched by two feed-forward networks (FFN). As shown in Fig. 1, to incorporate experts, we use an MoE layer [24, 23] to replace the end FFN in the Conformer layers. Similar to [24, 23], the MoE layer consists of a routing network and multiple experts, each of which is an FFN. To route a speech frame, we first use a softmax to estimate expert weights:

$$\boldsymbol{g}_l = Softmax(W_l \cdot \boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ is the output of the previous layer, and W_l is the weight matrix for the router at lth Conformer layer. Then the input is routed to the two experts with the highest weights (i.e., top 2 experts) and their outputs are weighted and summed to produce the final output:

$$\boldsymbol{y} = \sum_{i=1}^{2} g_{l,i} * \boldsymbol{e}_{l,i} \tag{2}$$

where $g_{l,i}$ is the weight for the top ith expert at the lth layer. $\boldsymbol{e}_{l,i}$ is the corresponding output of the expert. We use the top 2 experts for both training and inference. In Fig. 1, we replace the end FFN of the Conformer layer with the MoE layer. We have also tried replacing the start FFN layer or both of them (see more results in Sect. 5.1).

We use the RNN-T loss [29] for training the Conformer model with MoE layers. To ensure load balance across different experts, we use the same auxiliary loss as in [30]: $l_{aux} = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{S}\cdot m_i$, where m_i is the average number of times ith expert is selected over all frames, and c_i is the expert decision count for ith expert derived from the top-2 operation. We use the mean gates per expert $m_i \cdot (c_i/S)$ as a differentiable approximation for $(c_i/S)^2$ (more details in [30]).

Locale	Language	Counts (M)
en-US	English (USA)	18.1
zh-TW	Mandarin	0.5
fr-FR	French	10.8
de-DE	German	3.8
ja-JP	Japanese	10.9
es-US	Spanish (USA)	25.2
es-ES	Spanish (Spain)	20.3
ar-EG	Arabic	3.8
it-IT	Italian	13.0
hi-IN	Hindi	14.2
pt-BR	Portuguese	13.4
ru-RU	Russian	5.3
Total		139.4

Table 1: *Training data for 12 language locales. Utterance counts are in millions (M).*

4. Experiment Setup

4.1. Data

Our 12-language group consists of English (USA), Mandarin, French, German, Japanese, Spanish (USA), Spanish (Spain), Arabic, Italian, Hindi, Portuguese, and Russian (see Table 1 for more details). Our supervised training data of 12 language locales come from multiple domains such as Voice Search and YouTube. In total, they constitute around 139.4M utterances. The training data is anonymized and human transcribed. The per-language number of utterances ranges from 500K to 25.2M. For each language, we use a test set with utterances sampled from the Voice Search traffic ranging from 1.4K to 10K. The test sets do not overlap with the training set, and are also anonymized and human transcribed. We use word error rate (WER) for evaluation, and for languages such as zh-TW, the WER is computed based on characters. We are aware of the sensitive nature of the ASR research and other AI technologies used in this work. We thus ensure that this work abides by the Google AI Principles [31].

4.2. Modeling Details

4.2.1. Baseline Multilingual Model

We use a language agnostic multilingual model similar to [8] as the baseline. The baseline model consists of a 7-layer causal Conformer encoder and a 10-layer non-causal cascaded encoder. The causal encoder includes two blocks separated by a stacking layer. The first block consists of an input projection layer and 3 convolution layers. The stacking layer concatenates two neighboring encodings in time to form a 60-ms frame rate. The second block starts with a 1024-dim Conformer layer, and

Model	B1	E1	E2	E3
	No MoE	MoE-Start	MoE-End	MoE-Both
Total Size	180M	400M	400M	640M
Inf. Size	180M	211M	211M	246M
en-US	9.4	8.4	8.3	8.0
fr-FR	10.8	9.3	9.3	8.6
es-US	6.8	6.0	6.0	5.6
es-ES	6.6	5.2	4.9	4.8
ja-JP	15.5	13.1	12.9	12.0
zh-TW	9.4	9.2	9.1	9.2
de-DE	14.5	12.8	12.6	11.9
it-IT	10.4	8.8	8.8	8.1
ar-EG	12.3	10.9	10.8	10.3
pt-BR	7.6	7.0	6.9	6.5
ru-RU	13.0	11.0	10.9	10.4
hi-IN	19.6	19.5	19.2	19.1
Avg. WER	11.33	10.10	9.98	**9.54**

Table 2: *WERs (%) by placing the MoE layers at different places of the cascaded Conformer encoder.*

Model	E2	E4	E5
	8-Exp.	4-Exp.	2-Exp.
Total Size	400M	295M	211M
Inf. Size	211M	211M	211M
en-US	8.3	8.5	8.6
fr-FR	9.3	9.9	9.9
es-US	6.0	6.2	6.3
es-ES	4.9	5.7	5.8
ja-JP	12.9	14.2	14.5
zh-TW	9.1	8.7	9.4
de-DE	12.6	13.6	13.7
it-IT	8.8	9.2	9.2
ar-EG	10.8	11.2	11.3
pt-BR	6.9	6.8	7.2
ru-RU	10.9	11.3	11.5
hi-IN	19.2	19.5	19.5
Avg. WER	**9.98**	10.40	10.58

Table 3: *WERs (%) by reducing the number of MoE layers.*

then a projection layer to reduce the model dimension back to 512 for the rest of the causal layers. Note that the causal Conformer layers uses causal convolution and left-context attention and is thus strictly causal. Secondly, the non-causal layers are cascaded [28] to the causal encoder output. The 10 layers of non-causal Conformer layers have a dimension of 640, and a total right-context of 0.9 sec. We use separate decoders for causal and non-causal encoders to achieve the best quality.

Each transducer decoder consists of a prediction network and a joint network [29]. For the prediction network, we use two embedding layers to embed current and previous tokens separately and concatenate the embeddings as output. The joint network is a single feed-forward layer of 640 units. We use a hybrid autoregressive transducer (HAT) version of the decoder [32]. A softmax is used to predict 16,384 wordpieces. We generate the wordpieces using mixed transcripts pooled from all languages. The baseline multilingual transducer model has a total of 180M parameters.

Model	E2	E6	E7
	MoE-End	MoE-End-Odd	MoE-Conf1
Total Size	400M	295M	203M
Inf. Size	211M	196M	183M
en-US	8.3	8.5	8.8
fr-FR	9.3	10.1	10.6
es-US	6.0	6.4	6.5
es-ES	4.9	5.9	6.2
ja-JP	12.9	14.6	14.8
zh-TW	9.1	8.6	8.9
de-DE	12.6	13.7	14.1
it-IT	8.8	9.3	9.9
ar-EG	10.8	11.4	12.0
pt-BR	6.9	6.8	7.2
ru-RU	10.9	11.3	12.0
hi-IN	19.2	19.4	19.5
Avg. WER	**9.98**	10.50	10.88

Table 4: *WERs (%) by reducing the number of MoE layers.*

4.2.2. MoE Conformer

We replace the start, the end, or both FFNs of the Conformer layers by MoE layers (see experiments in Sect. 5.1) in the cascaded encoder of the baseline model. We use up to 24 experts in our experiments and dynamically choose the top 2 for training and inference. The expert FFNs have the same structure as the Conformer FFNs. We use the auxiliary loss in Sect. 3 to encourage load balance between experts. In training, we compute an over-capacity ratio which tracks whether certain experts are overloaded. It is calculated as, for each batch, the percentage of tokens going through a specific expert above a threshold. Our training shows that most experts have over capacity ratios ranging from 0.01 to 0.2 and only 10% of them range from 0.2 to a maximum value of 0.35, which is quite balanced. The model is trained to predict the same 16,384 wordpieces as the baseline.

We divide the input speech using 32-ms windows with a frame rate of 10 ms. 128D log-Mel filterbank features are extracted from each frame and then stacked together from 3 previous continuous frames to form a 512D input vector. These input vectors are further downsampled to have a 30-ms frame rate. We use SpecAug [33] to improve model robustness against noise. Two frequency masks with a maximum length of 27 and two time masks with a maximum length of 50 are used.

5. Results and Comparisons

5.1. Ablation Studies

5.1.1. Place of MoE Layers

In Table 2, we add the MoE layers to different places of the Conformer layer and each MoE layer has 8 experts. We show in Table 2 that using MoE to replace the start FFN (i.e. MoE-Start) in a Conformer layer improves the baseline (B1) significantly by around 10.9% relative on average. We note the improvement is uniform and significant for all languages. To ablate on the location of where MoE layers are added, we also add MoE layers to the end FFN of the Conformer layer (`E2`, MoE-End), or both start and end FFNs (`E3`, MoE-Both). We see in Table 2 that using MoE at the end FFN is slightly better than at the start. Using MoE for both start and end FFN works best but it also increases the inference model size because we have more MoE layers at inference. We are aware that the improvement of MoE models in Table 2 may be due to the increased inference model

ID	Model	Total Size	Inf. Size	WER (%)												Avg. WER (%)
				en-US	fr-FR	es-US	es-ES	ja-JP	zh-TW	de-DE	it-IT	ar-EG	pt-BR	ru-RU	hi-IN	
B1	Multi. Cas.	180M	180M	9.4	10.8	6.8	6.6	15.5	9.4	14.5	10.4	12.3	7.6	13.0	19.6	11.33
E2	MoE-End	400M	211M	8.3	9.3	6.0	4.9	12.9	9.1	12.6	8.8	10.8	6.9	10.9	19.2	9.98
B2	Larger B1	400M	400M	8.1	9.1	6.2	5.4	14.5	9.3	12.4	8.1	10.7	6.4	10.6	**19**	9.98
B3	B1+Adapter	280M	187M	**7.3**	9.4	6.5	5.5	14.4	**8.5**	13.2	8.7	**10.0**	6.8	11.2	19.1	10.05
E8	End-3-8	336M	187M	8.5	9.3	6.2	5.5	14.1	8.9	13.4	9.1	11.1	6.8	11.2	19.1	10.27
E9	End-3-16	532M	187M	8.1	9.6	6.0	5.3	13.5	8.9	12.8	8.9	11.2	6.7	11.0	19.4	10.12
E10	End-3-24	729M	187M	7.9	9.8	6.1	5.3	13.4	9.2	12.7	8.9	11.1	6.6	10.7	19.4	10.09
E11	E2+SF	+128M LM	+128M LM	8.1	**8.0**	**5.7**	**4.8**	**12.4**	10.4	**11.6**	**7.7**	10.6	**6.1**	**10.3**	20.4	**9.68**

Table 5: *Comparison of multilingual baseline models, adapter models, and MoE models. The numbers in bold represent best WER for any language.*

size. In Sect. 5.2, we will compare to a larger baseline which has a similar size as the MoE model at the inference time.

We have also tried adding MoE layers to the causal encoder, and the average WER is significantly worse. In following sections, we use MoE-End given the slightly better performance and efficiency.

5.1.2. *Number of Experts*

To reduce the total model size, we tried varying the number of experts in the MoE layer in Table 3. When we reduce the number of experts, the model total size decreases while the inference size stays the same because we always use the top 2 experts during inference. We see that the performance drops significantly when we decrease the expert number from 8 to 4, and relatively slowly from 4 to 2. This shows the model has been able to utilize the capacity from all experts. We also note that `E5` degenerates to a dense model and the performance difference between `B1` and `E5` is due to model size. We have also tried further increasing the number of experts and obtained better performance in Sect. 5.2.

5.1.3. *Reduce Number of MoE Layers*

Since reducing the number of experts reduces total model size but not inference model size, we have also tried reducing the number of MoE layers to make inference more efficient. In Table 4, we reduce the number of MoE layers by only adding it to every other Conformer layer (MoE-End-Odd), or only the first Conformer layer (MoE-Conf1). We see that by reducing the number of MoE layers, the model performance drops significantly. However, we note that even adding one MoE layer at the first Conformer layer is helpful. Although it increases the baseline size by only 3M parameters, it reduces the average WER from 11.33% to 10.88%.

5.2. Comparisons

In Table 5, we first compare the multilingual cascaded encoder baseline (`B1`) to the multilingual MoE-End model (`E2`), and `E2` reduces the average WER by 11.9% relative compared by `B1`. We then increase the cascaded encoder of `B1` to a 896-D 17-layer Conformer to have a total size of 400M (i.e. `B2`), which is the same size as the multilingual MoE model (`E2`). We show that `B2` and `E2` have the same average WERs: 9.98%, but the MoE model is around 47% more efficient (211M vs 400M) in terms of model parameters activated for inference. We note that in practice, one needs to implement inference in a way that only activated experts are selected for forward propagation in order to achieve this efficiency.

We also compare to an adapter model (`B3`) which adds residual adapters [34] to the baseline. The adapters we use follow a similar structure in [4, 27], i.e., we insert a 512-D residual adapter after every Conformer layer in the cascaded encoder. In both training and decoding, ground truth language information is used to select the corresponding adapter. To compare to the adapter model fairly in inference (same activated parameters), we first reduce the FFN multiplier in the MoE layer from 4 to 3 and then increase the number of experts from 8, 16, to 24. The results in Table 5 show that when the number of experts increases from 8 to 16, the MoE model improves because of increased model capacity. When we increase the experts to 16 or 24, the MoE models (`E9`, `E10`) perform similarly to the adapter model on average (10.12% or 10.09% vs 10.05%). However, we note that we do not explicitly use any language information in inference for MoE models while the adapter model uses ground truth language information. We also note that the improvement from 16-expert to 24-expert model (`E9` vs `E10`) is slight. This is probably because we only have 12 languages and the total number of experts may have exceed the needed capacity.

5.3. Further Improvement by Shallow Fusion

We further train a 128M multilingual LM by pooling the text portion of the supervised training data and text-only data from the 12 languages. Our text-only data covers 12 languages, and the sentences for each language ranges from 3.9B to 451B. The sentences are sampled from anonymized search traffic across multiple domains such as Web, Maps, News, Play, and YouTube. We use a 12-layer Conformer LM for shallow fusion (SF). The Conformer LM has a model dimension of 768 and a feedforward layer dimension of 2048. A left context of 31 tokens is used to attend to the previous tokens. We use 6 heads for self-attention. The total model size is around 140M. We use the same wordpiece model as the MoE model.

As shown in Table 5, shallow fusion further improves WER for almost all languages. The average WER reduction is around 3% relative, with the largest improvement of around 14.0% for fr-FR. We got regression for a couple of languages: zh-TW and hi-IN. The regression on zh-TW is probably because some of our text-only data contain Cantonese transcripts which is different from zh-TW. As for hi-IN, it has the largest amount of text-only data (around 451B sentences) and we may need to research filtering technique to better match the Search domain.

6. Conclusion

We propose a streaming multilingual Conformer model with mixture-of-expert (MoE) layers. By adding the MoE layers to the end FFN network of the Conformer, we improve the average WER of 12 languages by 11.9% relative compared to the baseline. The proposed model is also efficient: Activating only 53% of parameters during inference compared to a large baseline with similar quality. We also achieve similar quality compared to a ground truth language information based adapter model with increased experts but activating the same number of parameters in inference and not using language information. Further improvements have been obtained by multilingual shallow fusion.

7. References

[1] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 265–271.

[2] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *ICASSP*. IEEE, 2018, pp. 4914–4918.

[3] 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.

[4] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *Proc. Interspeech*, pp. 2130–2134, 2019.

[5] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Interspeech*, 2020, pp. 1037–1041.

[6] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur *et al.*, "Multilingual speech recognition with self-attention structured parameterization." in *Proc. Interspeech*, 2020, pp. 4741–4745.

[7] L. Zhou, J. Li, E. Sun, and S. Liu, "A configurable multilingual model is all you need to recognize all languages," in *ICASSP*. IEEE, 2022, pp. 6422–6426.

[8] B. Li, T. N. Sainath, R. Pang, S.-y. Chang, Q. Xu, T. Strohman, V. Chen, Q. Liang, H. Liu, Y. He, P. Haghani, and S. Bidichandani, "A language agnostic multilingual streaming on-device ASR system," in *Interspeech*, 2022, pp. 3188–3192.

[9] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *Proc. Interspeech*, pp. 4751–4755, 2020.

[10] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohman, P. Haghani, Y. Zhu, B. Farris, N. Gaur, and M. Prasad, "Massively multilingual ASR: A lifelong learning solution," in *ICASSP*. IEEE, 2022, pp. 6397–6401.

[11] K. Hu, B. Li, and T. N. Sainath, "Scaling up deliberation for multilingual ASR," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 771–776.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[13] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[14] S. Mavandadi, B. Li, C. Zhang, B. Farris, T. N. Sainath, and T. Strohman, "A truly multilingual first pass and monolingual second pass streaming on-device ASR system," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 838–845.

[15] C. Zhang, B. Li, T. Sainath, T. Strohman, S. Mavandadi, S.-y. Chang, and P. Haghani, "Streaming end-to-end multilingual speech recognition with joint language identification," *arXiv preprint arXiv:2209.06058*, 2022.

[16] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, "Mixture of informed experts for multilingual speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6234–6238.

[17] M. Yang, A. Tjandra, C. Liu, D. Zhang, D. Le, J. H. Hansen, and O. Kalinli, "Learning ASR pathways: A sparse multilingual ASR model," *arXiv preprint arXiv:2209.05735*, 2022.

[18] Z. You, S. Feng, D. Su, and D. Yu, "SpeechMoE: Scaling to large acoustic models with dynamic routing mixture of experts," *arXiv preprint arXiv:2105.03036*, 2021.

[19] Y. Lu, M. Huang, H. Li, J. Guo, and Y. Qian, "Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts." in *Interspeech*, 2020, pp. 4766–4770.

[20] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez, "Deep mixture of experts via shallow embedding," in *Uncertainty in artificial intelligence*. PMLR, 2020, pp. 552–562.

[21] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res*, vol. 23, pp. 1–40, 2021.

[22] Z. You, S. Feng, D. Su, and D. Yu, "SpeechMoE2: Mixture-of-experts model with improved routing," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7217–7221.

[23] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat *et al.*, "GLaM: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning*. PMLR, 2022, pp. 5547–5569.

[24] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017.

[25] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, pp. 5036–5040, 2020.

[26] K. Hu, T. N. Sainath, B. Li, N. Du, Y. Huang, A. M. Dai, Y. Zhang, R. Cabrera, Z. Chen, and T. Strohman, "Massively multilingual shallow fusion with large language models," *arXiv preprint arXiv:2302.08917*, 2023.

[27] B. Li, D. Hwang, Z. Huo, J. Bai, G. Prakash, T. N. Sainath, K. C. Sim, Y. Zhang, W. Han, T. Strohman *et al.*, "Efficient domain adaptation for speech foundation models," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[28] A. Narayanan, T. N. Sainath, R. Pang, J. Yu, C.-C. Chiu, R. Prabhavalkar, E. Variani, and T. Strohman, "Cascaded encoders for unifying streaming and non-streaming ASR," in *IEEE ICASSP*, 2021, pp. 5629–5633.

[29] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[30] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen, "GShard: Scaling giant models with conditional computation and automatic sharding," in *International Conference on Learning Representations*, 2021.

[31] "Artificial intelligence at google: Our principles." https://ai.google/principles/, accessed: 2022-07-20.

[32] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (HAT)," in *IEEE ICASSP*, 2020, pp. 6139–6143.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech*, pp. 2613–2617, 2019.

[34] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.