INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Improving End-to-End Modeling For Mandarin-English Code-Switching Using Lightweight Switch-Routing Mixture-of-Experts

Fengyun Tan, Chaofeng Feng, Tao Wei, Shuai Gong, Jinqiang Leng, Wei Chu, Jun Ma, Shaojun Wang, Jing Xiao

Ping An Technology, China

{tanfengyun813, fengchaofeng165, weitao820}@pingan.com.cn

Abstract

Code-switching is a common phenomenon in multilingual communities. In this paper, we study end-to-end model for Mandarin-English intra-sentential code-switching speech recognition. A lightweight Switch-Routing network is proposed, which includes two experts and a switch router. Two experts, representing Mandarin and English learners, implicitly provide language identification information and skillfully use monolingual data to assist code-switching task training, which solves the problem of data sparsity. In addition, our network is a lightweight structure, which makes use of the advantages of Switch Transformer and discards its weakness of increasing model capacity. Finally, we study the effect of using lightweight Switch Routing in different blocks of encoder and decoder. Compared with Bi-Encoder, proposed model has a better performance on the ASRU code-switching test set, and the most important thing is that it requires much less inference time with RTF decreasing by 31.39%.

Index Terms: code-switching, switch-transformer, mixture of experts, automatic speech recognition, end-to-end

1. Introduction

Code-switching (CS) refers to the phenomenon that more than one language may be used in a single conversation. It includes inter-sentential CS and intra-sentential CS, and widely exists in multilingual communities. However, CS task is challenging because it's difficult to get enough CS training data.

Many studies have been done for the CS speech recognition. With a hybrid speech recognition system, which consists of acoustic model, lexicon and language model, the common method is to design mixed phone set for the lexicon and acoustic model [1, 2]. However, the design of the phone set add the complexity of the speech recognition systems. Additionally, the unbalanced language distribution within CS utterances can lead to poor n-gram language model [3]. Recently, end-to-end systems which contain Connectionist Temporal Classification (CTC) [4, 5], recurrent neural network transducer (RNN-T) [6, 7], and attention-based encoder-decoder (AED) [8, 9, 10], are becoming more and more popular. Since they do not require explicit alignments and simplify the training of the model, many works for code-switching or multilingual are based on the end-to-end systems. As compared to hybrid systems, end-to-end (E2E) systems don't require the phone set design for lexicon, using the character or wordpiece as the output. It greatly simplifies the training of models for multilingual or CS, since we don't need to pay attention to the pronunciation of different languages. [11, 12, 13] has shown that building a single E2E model to recognize multilingual speech is possible by taking union over all the language-specific grapheme sets and training the model jointly on data from all the languages. The joint training model can also do CS task if the training corpus contain CS data. However, despite the simplistic design, E2E audio speech recognition (ASR) systems need large mounts of training data than hybrid based systems. Some data-augmentation techniques are used to improve the performance. [14, 15] propose to produce CS speech data with text to speech (TTS) systems to train acoustic models, and [16] leverages pointer-generator network to generate CS text corpus for language models. Another strategy is to leverage monolingual data to boost the performance. [17, 18, 19] proposes a multi-encoder network to leverage external monolingual data, and each encoder is initialized by the monolingual system.

More recently, [20, 21] has investigated to use mixture-of-experts (MoE) [22] in multilingual task. With the MoE architecture, each expert can capture the language specific feature and mitigate the conflicts between languages. The Bi-Encoder model proposed by [18] use two transformer encoders server as Mandarin and English experts, and the outputs of them are combined with interpolation coefficients produced by a gating network. Obviously, this approach increases the computational cost significantly.

In this paper, we propose a E2E CS ASR system. We use the WeNet [23] as the base framework, which is a very popular E2E speech recognition toolkit recently. Inspired by [24], we add lightweight Switch-Routing mixture-of-experts to the encoder and the decoder. The proposed model increases little extra computational cost compared to the monolingual system, and doesn't require any pre-training. The model can be jointly trained from scratch with the monolingual and CS training data.

2. Related work and method

2.1. Mixture of Experts (MoE) and variants

The architecture of mixture of experts (MoE) proposed in [25] have been intensively investigated and found popularity in other tasks [22, 26]. MoE can effectively improve model accuracy by stacking multiple mixture of experts. The output of MoE module can be defined as follows:

$$\boldsymbol{y} = \sum_{i=1}^{K} G_i(\boldsymbol{x}) E_i(\boldsymbol{x}) \tag{1}$$

where $E_i(\boldsymbol{x})$ is the output of expert i, K is the set of selected top-k indices, $G_i(\boldsymbol{x})$ is the probability of the router layer and can be defined as:

$$G_i(\boldsymbol{x}) = \mathbf{Softmax}(KeepTopK(H(\boldsymbol{x}), k)) \tag{2}$$

The $H(\boldsymbol{x})$ adds tunable Gaussian noise to feature x.

10.21437/Interspeech.2023-1465

$KeepTopK(H(\boldsymbol{x}), k)$ keeps only the top k value, which is a strategy of sparsity to save the computation cost of module.

2.1.1. Switch transformer

Switch transformer is proposed in [24], which simplify MoE routing algorithm to yield training stability and computational benefits. The router routes each output of previous layer to the top-1 expert with largest router probability. The probability of router layer of switch transformer can be defined as Equation 3.

$$G(\boldsymbol{x}) = \mathbf{Max}(\mathbf{Softmax}(H(\boldsymbol{x}))) \tag{3}$$

Assuming that m is the index of largest router probability, then the output of MoE layer is defined as Equation 4.

$$\boldsymbol{y} = G_m(\boldsymbol{x})E_m(\boldsymbol{x}) \tag{4}$$

2.2. Bi-Encoder

[18] proposed Bi-Encoder transformer network based MoE architecture. In this work, two transformer encoders equivalent to a Mandarin expert and an English expert individually provide language-specific information. Meanwhile, a gated network in the MoE layer acts as a decision maker, weighting the expert output. For acoustic features $\boldsymbol{x}$, LID information can be given from the two different experts.

$$\boldsymbol{h}^{ch} = MandarinEncoder(\boldsymbol{x}) \tag{5}$$

$$\boldsymbol{h}^{en} = EnglishEncoder(\boldsymbol{x}) \tag{6}$$

At each frame t, interpolation coefficients α_t^{cn} and α_t^{en} of MoE is dynamically obtained from a gating network, which are utilized to weighted the two encoder output:

$$\boldsymbol{h}_t^{mix} = \alpha_t^{cn}\boldsymbol{h}_t^{cn} + \alpha_t^{en}\boldsymbol{h}_t^{en} \tag{7}$$

The interpolation coefficients $\boldsymbol{\alpha}_t = [\alpha_t^{cn}, \alpha_t^{en}]^T$ has two elements, α_t^{cn} and α_t^{en} range from [0, 1] and the sum of the coefficients equals to one for each frames.

$$\boldsymbol{\alpha}_t = \mathbf{Softmax}(\boldsymbol{W}_{coe}^{cn}\boldsymbol{h}_t^{cn} + \boldsymbol{W}_{coe}^{en}\boldsymbol{h}_t^{en} + \boldsymbol{b}_{coe}) \tag{8}$$

2.3. Lightweight Switch-Routing

2.3.1. Language experts

Scaling up to a larger model has been an effective way towards a flexible and powerful E2E ASR system. It has shown that a large Conformer [27] model can achieve state-of-the-art results across a wide variety of tasks. However, developing large models in real-world application is seriously hindered by the expensive computation cost of both training and inference time.

The guiding design principle for Switch Transformers [24] is to maximize the parameter count of a Transformer model in a simple and computationally efficient way. Inspired by Switch Transformers network, we propose a lightweight Switch-Routing (LSR) module. The LSR module consists of two FFN layers and a router which is equivalent to a switch-gated network. Specifically, one thing that differs from the Switch Transformer is that only two experts are required for the CS task, one for Mandarin and one for the English language specialist. As shown in Figure 1, the network architecture consists mainly of two parts, LSR module in encoder block and LSR module in decoder block. We extend the second Macaron-style FFN to the LSR module in the last conformer block of encoder and in the last transformer block of decoder respectively.

Normal Switch-Transformer [24] consists of a series of a large number of experts, which makes the model complex and difficult to train. However, in the code-switching task, only two FFN layers are required to serve as Mandarin expert and English expert respectively, to provide experience for their respective language learning.

$$\boldsymbol{E}_t^{ch} = MandarinFeedForward(\boldsymbol{h}_t) \tag{9}$$

$$\boldsymbol{E}_t^{en} = EnglishFeedForward(\boldsymbol{h}_t) \tag{10}$$

where the $\boldsymbol{h}_t$ is the output of the upper convolutional model, and $\boldsymbol{E}_t^{ch}$, $\boldsymbol{E}_t^{en}$ represents the output of Mandarin expert and English expert at frame t respectively. According to the routing decision, the current frame is routered to the FFN layer of the corresponding language for the next step.

2.3.2. Switching routing

As mentioned in [24], contrary to MoE routing, we instead use a simplified strategy, named Switching Routing, where we route to only a single expert. We show that this simplification reduces routing computation and performs better.

Switch layer takes a token representation x as an input and then routes this to the best expert, which selected from two language experts. Since the high-level representation $\boldsymbol{h}_t^{ch}$ and $\boldsymbol{h}_t^{en}$ already maintain rich linguistic information, the router coefficients $\boldsymbol{r}_t = [r_t^{cn}, r_t^{en}]^T$ can be modeled with a single linear layer as Equation 11. Where the two interpolation coefficients r_t^{cn} and r_t^{en} range from [0, 1] and the sum of the coefficients equals to one for each frames.

$$\boldsymbol{r}_t = \mathbf{Softmax}(\boldsymbol{W}_t\boldsymbol{h}_t + \boldsymbol{b}_t) \tag{11}$$

$$\boldsymbol{y}_t = \begin{cases} r_t^{ch}\boldsymbol{E}_t^{ch} & if(r_t^{ch} >= r_t^{en}) \\ r_t^{en}\boldsymbol{E}_t^{en} & if(r_t^{ch} < r_t^{en}) \end{cases} \tag{12}$$

where $\boldsymbol{y}_t$ represents the switch FFN layer output in Equation 12. With the introduction of the sparsely-gated switch FFN, LSR can dynamically route inputs to corresponding language expert, which enables us to satisfy training and inference efficiency by having sub-network activated on per-example basis.

3. Experiments

3.1. Dataset and experimental setup

Three data set are used in our experiments: 500 hours monolingual Mandarin, 460 hours monolingual English, and 200 hours Mandarin-English code-switching. The Mandarin-English and Mandarin corpus are obtained from ASRU 2019 Mandarin-English Code-Switching Challenge data set [28]. English is selected from LibriSpeech 460 hours clean data [29]. Among them, 10% of the mono data and additional 20 hours code-switching data are used as the development set. Meanwhile, we use three test sets: 5 hours Mandarin from ASRU (ASRU_CN), 5 hours English test_clean official test set from Librispeech (Libri) and 20 hours Mandarin-English code-switching official test set from ASRU (ASRU_CS). Performances are measured by character error rate (CER) for Mandarin and word error rate (WER) for English. As for the code-switching, we report Mandarin part CER, English part WER and the total mix error rate (MER), as those in ASRU2019 Challenge.

We select 3,003 Mandarin characters, 1,000 English BPE subwords, along with three other token(unk, $blank$ and sos/eos) to form the 4,006 Character-BPE modeling units. The

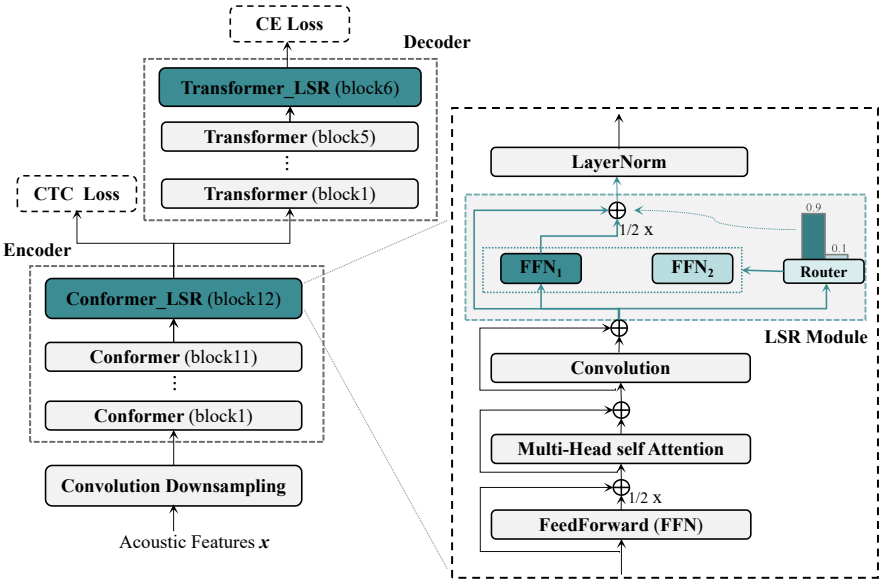


Figure 1: *Proposed architecture. The blue-green dotted box represents the Lightweight Switch-Routing(LSR) module.*

80 dimensional log-mel filterbanks acoustic features are extracted with 25ms windowing and 10ms frame shift, and global Cepstral Mean and Variance Normalization (CMVN) computed using the training set is applied on the fbank features. SpecAugment [30] is applied for data augmentation during all model training. Monolingual and Multilingual baseline use 12-layer conformer in encoder and 6-layer transformer in decoder. We initialize the encoders of Bi-Encoder model by encoder of pre-trained monolingual Mandarin and English models separately. It is worth noting that proposed model only needs random initialization and does not need complex pre-training. The attention dimension is 256 and the dimensionality of inner-layer in FFN is 2048. And 4 heads are used for multi head attention in all attention sub-layers.

3.2. Result

3.2.1. Proposed model and baselines

We present the performance of baseline and proposed model in Table 1. It is observed that the monolingual model can obtain a low error rate on the monolingual test sets, but it can handle neither inter-sentential CS task nor intra-sentential CS task. The model using only CS training data can handle CS tasks, but the MER is 13.52% due to insufficient CS training data with annotations. We train a multilingual model with monolingual data and CS data together, which can be compatible with multiple scene data recognition and improves CS performance. Second to last row of Table 1 shows the Bi-Encoder model reproduced according to [18]. Compared with multilingual model, the WER of Bi-Encoder model is reduced on both monolingual and CS test sets, which is consistent with [18]. The last line of Table 1 shows that proposed model can make the system recognize all kinds of data, and the mono Mandarin the CS performance is significantly improved.

In Table 1, we calculate the RTF by attention decoding with the Mandarin, English, and code-switching test sets together. It is worth nothing that the RTF of the proposed model is reduced by 31.39% compared with Bi-Encoder, which greatly saves the computing resources and improves the decoding speed.

Table 1: *A comparison of (CERs/WERs/MERs) (%) and Real Time Factors (RTFs) on testsets.* ***ASRU_CN/Libri/ ASRU_CS****: 5/5/20 hrs Mandarin-only test set from ASRU/official test_clean test set from LibriSpeech/official Mandarin-English CS test set from ASRU, respectively.* ***Man/Eng/CS/Multi****: the same network structure, trained with mono Mandarin, mono English, code-switching and all data, respectively.* ***Bi-Enc****: Bi-Encoder baseline;* ***LSR****: proposed Lightweight Switch-Routing.*

Model	#Params	RTF	ASRU _CN	Libri	ASRU_CS		
					CH	EN	MIX
Man	49.2M	-	2.34	-	-	-	-
Eng	49.2M	-	-	6.53	-	-	-
CS	49.2M	-	-	-	11.22	32.42	13.52
Multi	49.2M	0.145	2.73	6.83	8.52	27.89	10.62
Bi-Enc	84.2M	0.223	2.64	**6.48**	8.50	**27.52**	10.57
LSR	51.3M	**0.153**	**2.47**	**6.48**	**8.20**	27.54	**10.30**

3.2.2. Ablation study

We conduct a number of experiments in an ablation study. We tried to modify the blocks of encoder and decoder, and compared the performance of LSR module in different blocks. The training data used by these models in Table 2 are consistent, including monolingual and code-switching data. Comparing LSR_Enc-1layer and LSR_Enc-2layers from Table 2, it is found that the CER of monolingual Mandarin test set decreases by only 0.2%, and the performance of other test sets does not improve significantly. Comparing LSR_Enc-2layers and LSR_Enc-all, except for the 0.13% difference in CER on the monolingual Mandarin test set, the comparison results of the other two test sets show that there is no significant gain in applying LSR module in more blocks of encoder. Similarly, it can be concluded that applying the LSR module in more blocks of decoder will not bring much gain to the model.

However, it can be seen that LSR_Enc_Dec has the best per-

Table 2: *Performance comparison (CER/WER/MER) (%) of LSR module in different block of encoder and decoder.* ***LSR_Enc-1layer****: replaces macaron FFN with LSR module in the last block of encoder.* ***LSR_Enc-2layers****: replaces macaron FFN with LSR module in the last two blocks of encoder.* ***LSR_Enc-all****: replaces macaron FFN with LSR module in all blocks of encoder.* ***LSR_Enc_Dec****: replaces macaron FFN with LSR module in the last block of encoder and decoder respectively.*

Model	Params	ASRU	Libri	ASRU_CS		
		_CN		CH	EN	MIX
LSR_Enc-1layer	50.2M	2.79	6.61	8.41	27.60	10.49
LSR_Enc-2layers	51.3M	2.59	6.54	8.28	27.94	10.42
LSR_Enc-all	61.8M	**2.46**	6.55	8.47	27.86	10.58
LSR_Enc_Dec	51.3M	2.47	**6.48**	**8.20**	**27.54**	**10.30**

formance compared to the first three models. In particular, compared with LSR_Enc-1layer, it decreases CER/WER/MER by 11.47%, 1.97% and 1.81% respectively on monolingual Mandarin, monolingual English and CS test sets. This proves that it is effective to add LSR module to the last block of encoder and decoder respectively.

In this section, we investigate the effect of increasing the number of experts to prove that the improvement of model performance is due to the excellent ability of the two experts to learn specific information, not the increase in model capacity. Table 3 shows the performance comparison on different number of experts with LSR. Line 2 presents the results of our proposed optimal LSR model which are maked as MoE-2e. The two experts learn specific linguistic information separately. Line 3 presents the results of LSR model which are maked as MoE-3e. The results clearly show that performance do not get better as the number of experts increases. Specially, the performance of LSR_Enc_Dec-2e achieves up to 6.4% relative CER improvement over the LSR_Enc_Dec-3e on the ASRU_CN. Similarly, on the Libri English clean test set, compared with LSR_Enc_Dec-3e, WER of LSR_Enc_Dec-2e was relatively reduced by 0.8%. The most important thing is that the proposed LSR improves the overall MER by 1.2% on the ASRU_CS test set. Although the WER increase of 0.34% in the English part of the ASRU code-switching test set may be due to the small proportion of English parts in the code-switching test set. It is obvious that LSR_Enc_Dec-3e increases the capacity of the model as the number of experts increases, but the model performance does not improve. It turns out that two experts of LSR_Enc_Dec-2e do learn specific language information, which leads to further improvement of the model effect. This conclusion will also be presented more intuitively in graphical form below.

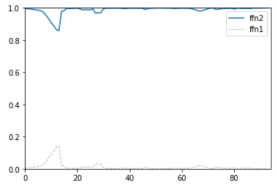
Table 3: *Performance comparison (CER/WER/MER) (%) of increasing the number of experts on LSR module.* ***LSR_Enc_Dec-2e****: our proposed optimal LSR model which has two experts in the last block of encoder and decoder respectively.* ***LSR_Enc_Dec-3e****: LSR module with three experts in the last block of encoder and decoder respectively.*

Model	Params	ASRU	Libri	ASRU_CS		
		_CN		CH	EN	MIX
LSR_Enc_Dec-2e	51.3M	**2.47**	**6.48**	**8.20**	27.54	**10.30**
LSR_Enc_Dec-3e	53.4M	2.64	6.53	8.37	**27.20**	10.42

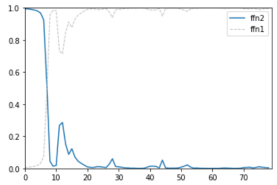
In Table 4, only the training data differs between the two models. LSR_Enc_Dec trained with all data achieves a significant improvement on CS, with up to 24.26% relative mix error reduction over LSR_Enc_Dec trained witch only code-switching data. It is observed that the LSR module has better ability to leverage the monolingual data, demonstrating the efficiency of LSR module.

Table 4: *Performance comparison (CER/WER/MER)(%) of* ***LSR_Enc_Dec*** *trained with different data.* ***CS/Mono+CS****: replaces only code-switching train set/Mandarin, English and code-switching train set together.*

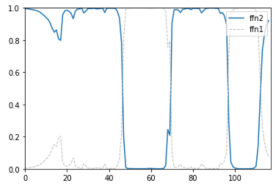
Model	TR-Data	ASRU	Libri	ASRU_CS		
		_CN		CH	EN	MIX
LSR_Enc_Dec	CS	-	-	11.3	32.46	13.60
LSR_Enc_Dec	Mono+CS	2.47	6.48	**8.20**	**27.54**	**10.30**



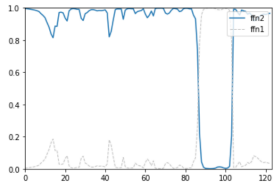
(a) 没文化真可怕有文化真可怕.



(b) SO WE NEED TO REMEMBER THAT WHEN.



(c) 形容自己叫SNOBBISH形容别人叫CHEEKY.



(d) 期望在恋人身上看到CONSCIOUS品质.

Figure 2: *Visualization of the router coefficients r^{ch} and r^{en} of proposed model. The blue line and gray line represents routing probability of Mandarin and English experts respectively.*

We visualize the router coefficients of lightweight Switch-Routing for different utterances. As shown in Figure 2, switch routing can accurately determine the language category for pure Mandarin or English utterance, while improving the ability of switching point conversion in intra-sentential code-switching task. It further proves the effectiveness of the proposed LSR code-switching E2E architecture.

4. Conclusions

In this paper, we apply lightweight Switch-Routing on the CTC/AED E2E ASR framework. The RTF value of proposed model reduces by 31.39% compared to the Bi-Encoder, which accelerates the model decoding speed. LSR module includes two parallel FFN, acting as Mandarin and English experts respectively, which is equivalent to providing LID information invisibly. In addition, we increased the number of experts, and the results showed that the model did not gain from it. Finally, we demonstrate the effectiveness of our method on both monolingual and CS data set. In the future, we will try to increase a larger number of experts to train a LSR model.

5. References

[1] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, and E.-S. Chng, "A first speech recognition system for mandarin-english codeswitch conversational speech," in *Proc. ICASSP 2012 – 37th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2012.

[2] C.-F. Y. L.-S. Lee, "Transcribing code-switched bilingual lectures using deep neural networks with unit merging in acoustic modeling," in *Proc. ICASSP 2014 – 39th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2014.

[3] H. Adel, H. Adel, and H. Adel, "Combination of recurrent neural networks and factored language models for code-switching language modeling," in *Proc. ACL 2013 – 50th Annual Conference of the Association for Computational*, 2013.

[4] G. Alex, F. Santiago, G. Faustino, and S. Jürgen, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning*, 2006.

[5] A. Dario, A. Sundaram, A. Rishita, B. Jingliang, B. Eric, C. Carl, C. Jared, C. Bryan, C. Qiang, and C. Guoliang, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *Proc. ICML 2016 – International conference on machine learning*, 2016.

[6] G. Alex, "Sequence transduction with recurrent neural networks," in *Proc. ICML 2012 – Internation Conference of Asian Conference on Machine Learning Representation learning workshop*, 2012.

[7] G. Alex, M. Abdel-rahman, and H. Geoffrey, "Speech recognition with deep recurrent neural networks," in *Proc. ICASSP 2013 – 38th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2013.

[8] C. Jan, B. Dzmitry, C. Kyunghyun, and B. Yoshua, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *NIPS 2014 Workshop on Deep Learning*, 2014.

[9] C. J. K, B. Dzmitry, S. Dmitriy, C. Kyunghyun, and B. Yoshua, "Attention-based models for speech recognition," *Advances in neural information processing systems*, 2015.

[10] C. William, J. Navdeep, L. Quoc, and V. Oriol, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP 2016 – 41st Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2016.

[11] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP 2018 – 43rd Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2018.

[12] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, and V. Liptchinsky, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Oct. 2020.

[13] B. Li, R. M. Pang, T. N. Sainath, A. Gulati, Y. Zhang, J. Qin, P. Haghani, W. R. Huang, M. Ma, and J. Bai, "Scaling end-to-end models for large-scale multilingual asr," in *Proc. ASRU 2021 – Automatic Speech Recognition and Understanding Workshop*, 2021.

[14] Y. Sharma, B. Abraham, K. Taneja, and P. Jyothi, "Improving low resource code-switched asr using augmented code-switched tts," in *Proc. Interspeech*, 2020.

[15] C. Du, H. Li, Y. Lu, L. Wang, and Y. Qian, "Data augmentation for end-to-end code-switching speech recognition," in *Proc. SLT 2021 – IEEE Spoken Language Technology Workshop*, 2021.

[16] X. Hu, Q. Zhang, L. Yang, B. Gu, and X. Xu, "Data augmentation for code-switch language modeling by fusing multiple text generation methods," in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Oct. 2020.

[17] S. Zhang, Y. Liu, M. Lei, B. Ma, and L. Xie, "Towards language-universal mandarin-english speech recognition," in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, 2019.

[18] 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 *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Oct. 2020.

[19] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, "Transformer-transducers for code-switched speech recognition," in *Proc. ICASSP 2021 – 40th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2021.

[20] A. Das, K. Kumar, and J. Wu, "Multi-dialect speech recognition in english using attention on ensemble of experts," in *Proc. ICASSP 2021 – 40th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2021.

[21] 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 *Proc. ICASSP 2021 – 40th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2021.

[22] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, , and J. Dean, "The sparsely gated mixture-of-experts layer," in *Proc. ICASSP 2021 – 40th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2021.

[23] Y. Zhuoyuan, W. Di, W. Xiong, Z. Binbin, Y. Fan, Y. Chao, P. Zhendong, C. Xiaoyu, X. Lei, and L. Xin, "WeNet: Production oriented Streaming and Non-streaming End-to-End Speech Recognition Toolkit," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, 2021.

[24] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," in *Proc. 2022 Journal of Machine Learning Research*, 2022.

[25] R. Jacobs, M. Jordan, S. Nowlan, and G. Hinton, "Adaptive mixtures of local experts," *Neural Computure*, vol. 3, no. 1, pp. 79–87, 1991.

[26] S. Papi, E. Trentin, R. Gretter, M. Matassoni, and D. Falavigna, "Mixtures of deep neural experts for automated speech scoring," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, 2021.

[27] A. Gulati, J. Qin, C. C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, and Y. Wu, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. ACL 2013 – 50th Annual Conference of the Association for Computational*, 2020.

[28] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," 2020.

[29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP 2015 – 40th Annual Conference of IEEE international conference on acoustics, speech and signal processing*, 2015.

[30] 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," in *arXiv preprint arXiv:1904.08779*, 2019.