Language-Routing Mixture of Experts for Multilingual and Code-Switching Speech Recognition

Wenxuan Wang, Guodong Ma, Yuke Li*, Binbin Du

NetEase Yidun AI Lab, Hangzhou, China

{*wangwenxuan, maguodong, liyuke, dubinbin*}@corp.netease.com

Abstract

Multilingual speech recognition for both monolingual and code-switching speech is a challenging task. Recently, based on the Mixture of Experts (MoE), many works have made good progress in multilingual and code-switching ASR, but present huge computational complexity with the increase of supported languages. In this work, we propose a computation-efficient network named **L**anguage-**R**outing **M**ixture **o**f **E**xperts (LR-MoE) for multilingual and code-switching ASR. LR-MoE extracts language-specific representations through the Mixture of Language Experts (MLE), which is guided to learn by a frame-wise language routing mechanism. The weight-shared frame-level language identification (LID) network is jointly trained as the shared pre-router of each MoE layer. Experiments show that the proposed method significantly improves multilingual and code-switching speech recognition performances over baseline with comparable computational efficiency.

Index Terms: mixture of experts, language identification, multilingual, code-switch, speech recognition

1. Introduction

Multilingualism is a widespread phenomenon in the world. Multilingual speakers often communicate in multiple languages simultaneously, such as interspersing English in Mandarin. Therefore, a practical multilingual speech recognition system needs to support the recognition of monolingual and code-switching utterances in multiple languages.

End-to-end (E2E) ASR systems [1–9] have become more and more popular recently due to the simple pipeline, excellent performance and less dependence on linguistic knowledge compared to traditional hybrid methods [10]. Prior works based on the E2E model have also made good progress in the field of multilingual ASR, including code-switching corpus synthesis [11–13], multi-task training with joint language identification [14, 15], self-supervised speech representation learning [16–18], cross-lingual transfer learning [15, 19], etc. The MoE architecture is an effective method to improve the performance of multilingual speech recognition in both monolingual and code-switching scenarios, which has been widely concerned and studied recently. The existing MoE-based methods [20–25] extract language-specific representations separately by independent encoders and fuse them to decode. However, the computational complexity of the models will increase significantly with the number of supported languages.

In this work, we propose a computation-efficient network named **L**anguage-**R**outing **M**ixture **o**f **E**xperts (LR-MoE) to improve the performance of the multilingual and code-switching ASR task. The LR-MoE architecture consists of a shared block

---

* Corresponding author

and a Mixture-of-Language-Experts (MLE) block. Unlike the sparsely-gated mixture of experts (sMoE) [26–29], the expert layers in the MLE block are language-dependent, which is called Language-Specific Experts (LSE). The shared block generates the global representation, while the LSE of the MLE block extracts language-specific representations. In the MLE block, we design a Frame-wise Language Routing (FLR) mechanism, which guides the expert layers to learn language specialization at the training stage. A weight-shared frame-level language identification (LID) network is jointly trained as the shared pre-router of each LSE layer, and the alignment of frame-wise LID will be used as the routing path of the LSE layers. We also compare utterance-wise and frame-wise language routing for LR-MoE in the multilingual and code-switching experiment. To distinguish them, we will name the two networks with different routing as ULR-MoE and FLR-MoE, respectively. Our contributions are summarized as follows:

- We propose a computation-efficient LR-MoE architecture, which is suitable to apply in more languages with little increase in computational complexity.

- We investigate multiple routing strategies of MoE and propose an FLR mechanism to guide the expert layers to learn language specialization, which is compatible with both multiple monolingual and code-switched ASR.

- In Mandarin-English code-switching and multilingual experiments, the proposed method significantly improves the performances of multilingual and code-switching speech recognition over the baseline with comparable computational efficiency and outperforms previous MoE-based methods with less computational complexity.

2. Related Works and Motivation

2.1. Previous MoE-based works

More recently, many works [20–24] focus on exploring MoE architectures to recognize monolingual and intra-sentence code-switching speech. The MoE-based methods mainly utilized language-specific expert encoders to generate parallel language-specific representations and fuse them, whose difference is primarily in the fusion mode of expert encoders and training strategy. For example, the Bi-encoder transformer network [21] uses a gated network to dynamically output the MoE interpolation coefficients to mix two encoding representations. The weights of expert encoders are initialized with the pretrained monolingual model, respectively. Conditional factorized neural transducer [22] defined the monolingual sub-tasks with label-to-frame synchronization to achieve unified modeling of code-switching and monolingual ASR. The language-aware encoder [23,24] learned language-specific representations through language-aware training with the language-specific auxiliary

loss instead of monolingual pretraining and used the frame-wise addition to fuse them.

2.2. Motivations

As mentioned above, the previous MoE-based works achieved considerable improvement in monolingual and code-switching ASR, but there are still the following problems:

- The approach needs to compute all language-specific blocks. However, only one works in the monolingual scene. It means a large amount of redundant computational overhead. And the more languages are supported, the more redundant computational overhead is.

- Language-specific blocks are isolated from each other and lack interaction. As a result, the cross-linguistic contextual information is easily lost in code-switching speech.

In order to alleviate the above two issues, we propose the LR-MoE architecture inspired by the sparsely-gated mixture of experts [26–29]. Please refer to Section 3 for more details.
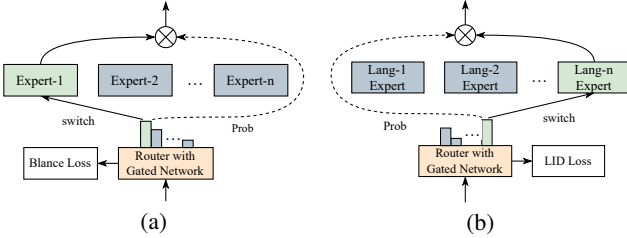

Figure 1: *Schematic diagram of MoE modules. (a) Sparsely-Gated Mixture of Experts (sMoE), (b) Mixture of Language Experts (MLE).*

3. Proposed Method

3.1. Sparsely-Gated Mixture of Experts

The sMoE module is shown in Fig. 1(a). As a representative, Switch Transformer [27] adopts a top-1 expert routing strategy in the MoE architecture to route the data samples to the expert model with the highest probability in the gated network. The computational complexity of the whole network increases slightly as the number of experts increases, and the extra computational overhead only comes from the gated network. The inputs of the expert layer and the gated network are the outputs of the previous non-expert layer o_{ne}. The router probability p can be expressed as follows:

$$p = Softmax(W_r \cdot o_{ne} + b_r) \tag{1}$$

where W_r and b_r is the weights and bias of router respectively. An auxiliary loss is added to guarantee load balance across the experts during training. The balance loss is expressed as:

$$\mathcal{L}_b = n \cdot \sum_{i=1}^{n} f_i \cdot p_i \tag{2}$$

where f_i is the fraction of samples dispatched to i-th expert, n is the number of experts.

3.2. Architecture of LR-MoE

In order to strengthen the interaction of cross-lingual context information, we further expand the shared parts, such as the attention layers in the Transformer [30] network. It is different from the separate language-specific encoder [20–23]. Fig. 2 shows the LR-MoE-based Transformer model. The shared block are
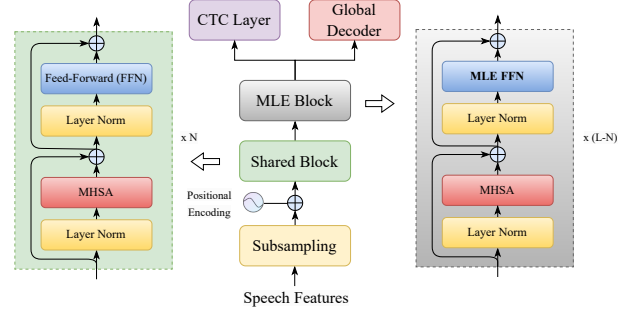

Figure 2: *The structure of the LR-MoE Transformer Model. N and $(L - N)$ are the number of layers of the shared block and the MLE block, respectively.*

stacked by standard transformer blocks. In contrast to the standard transformer block, we introduce the MLE FFN module as shown in Fig. 1(b) to strengthen language-specific representations in MLE block. All MLE modules share the same language router, which is a frame-level LID-gated network in front. According to the top-1 language predicted by the FLR, data samples are routed to the corresponding LSE layer. For each time frame, only one LSE will be routed, so the computational complexity of the model will not increase with more languages.

3.3. Language Routing

3.3.1. LID-Gated Network

Considering that LID can be regarded as a low-dimensional subtask of ASR and the output of the non-expert layer o_{ne} already contains rich high-dimensional linguistic information, we model the frame-level LID task with a linear layer as follows:

$$r = W_r \cdot o_{ne} + b_r \tag{3}$$

A frame-level LID auxiliary loss is added to jointly train the LID and ASR tasks at the training stage. We get the auxiliary labels $\mathcal{Y}_{lid}$ by replacing the tokens in the text labels with the corresponding language IDs. Then, based on r, the LID-CTC loss is adopted to get the token-to-frame alignment, as shown in Eq. (4).

$$\mathcal{L}_{lid-ctc} = -\log P_{CTC}(\mathcal{Y}_{lid}|r) \tag{4}$$

Due to the sparse spike property of the Connectionist Temporal Classification (CTC), the greedy decoding result of the LID output z_t will contain a large amount of blank. Therefore, we adopt a simplified alignment strategy to get dense frame-wise language routing information as follows:

$$z_t = \begin{cases} z_f, & \text{if } t = 0 \\ z_t, & \text{if } z_t \neq \phi \\ z_{t-1}, & \text{if } z_t = \phi \end{cases} \tag{5}$$

where $z_t \in \{language\ ids\} \cup \phi, t = 0, 1, 2, \ldots, T$, z_f is the first non-blank element.

Besides, we also use an utterance-wise LID-gated network with the cross entropy (CE) loss for comparison. The utterance-wise LID loss is as follows:

$$\mathcal{L}_{lid-utt} = CE(r_u, \mathcal{U}_{lid}) \tag{6}$$

where r_u is the time-dimension global average pooling of r, $\mathcal{U}_{lid}$ is the language ID of utterance.

The final object loss function is shown in Eq. (7):

$$\mathcal{L}_{mtl} = \mathcal{L}_{asr} + \lambda_{lid}\mathcal{L}_{lid} \tag{7}$$

where λ_{lid} is selected by hand, $\mathcal{L}_{lid}$ of ULR and FLR correspond to $\mathcal{L}_{lid-utt}$ and $\mathcal{L}_{lid-ctc}$, respectively.

3.3.2. Shared Router

Unlike sMoE [26–29], we use a shared router instead of the independent router for each MLE layer, mainly due to the following considerations: The independent router of each MoE layer in sMoE is helpful in obtaining more diverse routing paths and larger model capability. However, the expert layers are language-specific and the desired routing paths are determined with a priori in LR-MoE. Therefore, the shared LID router might be helpful to reduce additional computation and the multi-level error accumulation caused by the alignment drift of the language routing.

3.4. Pretrained Shared Block

The bottleneck features of ASR are effective in transfer learning for LID [31]. Therefore, we utilize a pretrained shared block to speed up the convergence of the LID-gated network and reduce the bad gradient back-propagation caused by erroneous routing paths, especially at the early training stage.

4. Experiments

4.1. Datasets

Our experiments are conducted on ASRU 2019 Mandarin-English code-switching Challenge dataset [32] and a four-language dataset including Aishell-1 (CN) [33], train-clean-100 subset of Librispeech [34] (EN), Japanese (JA), Zeroth-Korean (KR) [1] and Mandarin-English code-switching (CN-EN) data. JA and CN-EN are collected from Datatang [2]. Table 1 and 2 show the details of all experimental datasets.

Table 1: *Details of Mandarin-English Code-Switching Dataset*

Lang	Corpora	Dur. (Hrs)		Utterance(k)	
		Train	Eval	Train	Eval
CN	ASRU-Man [32]	482.6	14.3	545.2	16.6
EN	Librispeech [34]	464.2	10.5	132.5	5.6
CN-EN	ASRU-CS [32]	199.0	20.3	186.4	16.2

Table 2: *Details of Multilingual and Code-Switching Dataset*

Split	Information	CN	EN	JA	KR	CN-EN
Train	Dur. (Hrs.)	151.2	100.6	93.7	51.6	93.1
	Utterances(k)	120.1	28.5	75.5	22.3	84.0
Eval	Dur. (Hrs.)	9.7	5.4	6.3	1.2	6.9
	Utterances(k)	7.2	2.6	5.2	0.5	6.6

For all the experiments, the acoustic features are 80-dimensional log filter-bank energy extracted with a stride size of 10ms and a window size of 25ms. SpecAugment [35] is applied during all training stages. The Mandarin-English vocabulary and the vocabulary of 4 languages consist of 12064 and 15492 unique characters and BPE [36] tokens.

4.2. Experimental setup

The experiments are conducted on the ESPnet toolkit [37]. We use the Transformer CTC/Attention model with a 12-layer encoder and a 6-layer decoder as our baseline, called the Vallina model. The LR-MoE encoder consists of a 6-layer shared block

---

and a 6-layer MLE block in experiments. We also compare various MoE-based encoders with our proposed method, including Bi-Encoder [21], LAE [24]. The Bi-Encoder contains the 12-layer encoder for each language. The LAE contains a 9-layer shared block and a 3-layer language-specific block for each language. Besides, we implement 12-layer sMoE [29] with 4 experts in each MoE block for comparison in the multilingual and code-switching experiment. All encoders and decoders are stacked transformer-based blocks with attention dimension of 256, 4 attention heads and feed-forward dimension of 2048. We implement multi-task learning with $\lambda_{ctc} = 0.3$ and $\lambda_{lid} = 0.3$ for ASR and LID at the training stage of the LR-MoE model.

We use the Adam optimizer with a transformer-lr scale of 1 and warmup steps of 25k to train 100 epochs on 8 Tesla V100 GPUs. The training process adopts a dynamic batch size strategy with a maximum batch size of 128. We train a 4-gram language model with all training transcriptions and adopt the CTC prefix beam search for decoder with a fixed beam size of 10.

4.3. Experimental Results

4.3.1. Results on Mandarin-English ASR

Table 3: *Performance of models in the CTC-based and AED-based Mandarin-English ASR system*[*]. *"CN", "EN" and "ALL" mean the character error rate (CER) of monolingual Mandarin, the word error rate (WER) of monolingual English and the total mix error rate (MER) of code-switching test set respectively.*

Model	Params	Mono		Code-Switch		
		CN	EN	ALL	CN	EN
Vallina CTC	19.8M	7.1	12.4	12.2	9.0	38.9
Bi-Encoder CTC	36.6M	5.3	10.3	10.7	7.9	33.3
LAE CTC	26.5M	5.3	10.5	10.8	8.0	33.7
FLR-MoE CTC	25.8M	**5.1**	**10.1**	**10.5**	**7.7**	**33.1**
Vallina AED	34.7M	6.3	11.7	11.2	8.6	32.5
Bi-Encoder AED	51.5M	4.9	9.8	9.9	7.6	28.9
LAE AED	41.4M	5.0	9.9	10.0	7.7	29.2
FLR-MoE AED	40.7M	**4.7**	**9.6**	**9.7**	**7.4**	**28.4**

[*] Note: The results of the Bi-Encoder and LAE are achieved with our experimental configuration.

As shown in Table 3, our proposed method outperforms the previous MoE-based methods with comparable or fewer parameters in the Mandarin-English ASR system, including CTC and Attention-based Encoder-Decoder (AED) models. The proposed method achieves significant performance improvement over the baseline. The relative improvements on mono-Mandarin, mono-English and code-switch evaluation sets are 28.2%, 18.5% and 13.9% in the CTC-based model and 25.4%, 17.9% and 13.4% in the AED-based model, respectively.

4.3.2. Results on multilingual ASR

Table 4 shows the results of models in the CTC-based multilingual ASR system. Compared with previous MoE-based methods, our proposed method achieves significant performance improvement in both monolingual and code-switching scenarios. Regarding FLOPs, the proposed architecture's computational complexity increases little with the increase of languages, and it is easier to scale in multilingual ASR.

The proposed method significantly improves performance over the baseline with comparable computational complexity. The relative average improvements on monolingual and code-switch evaluation sets are 28.4% and 26.8%, respectively.

Table 4: ***Performance of models in the CTC-based multilingual ASR system.*** *Multi-Encoder is a multilingual extension of the method of Bi-Encoder [21] with the unsupervised gating network. "CN", "EN", "JA", "KR" and "ALL" mean CER of Mandarin, WER of English, CER of Japanese, CER of Korean and MER of Mandarin-English code-switch, respectively. We also use the number of parameters and floating-point operations (FLOPs) of a 30s input to evaluate the computational complexity of the model.*

Model	Shared Router	Params	GFLOPs	Monolingual					Code-Switch		
				CN(CER)	EN(WER)	JA(CER)	KR(CER)	Avg	ALL(MER)	CN	EN
Vallina CTC	-	19.8M	55.4	6.3	13.1	10.1	5.6	8.8	10.4	6.9	29.7
Multi-Encoder	-	71.0M	146.9	5.9	10.9	9.1	2.6	7.1	9.4	6.6	24.2
LAE	-	35.7M	78.3	5.6	11.3	8.8	2.4	7.0	8.6	6.1	22.5
sMoE	w/o	57.4M	55.4	5.5	10.7	8.6	3.2	7.0	9.3	6.5	25.1
ULR-MoE	w/o	38.6M	55.4	5.3	10.4	8.4	2.0	6.5	8.6	5.9	23.5
FLR-MoE	w/o	38.6M	55.4	5.2	10.5	8.5	2.1	6.6	8.4	5.9	22.4
FLR-MoE (ours)	w	38.6M	55.4	**5.2**	**10.0**	**8.2**	**1.9**	**6.3**	**8.2**	**5.8**	**21.8**

We also compare different routing strategies of LR-MoE. Experiments show that FLR-MoE achieves slight performance improvement compared to ULR-MoE, especially in code-switching scenarios. The shared language routing strategy for FLR-MoE outperforms the layer-wise language routing strategy on both monolingual and code-switching evaluation sets.

4.4. Ablation study and analysis

4.4.1. Position of MLE

Table 5: ***Ablation study on the position of the MLE FFN layers.*** *We compare the average token error rate (TER) of four monolingual evaluation sets, the total mix error rate (MER) of the code-switching evaluation set and the average utterance-level LID accuracy for the different positions of the MLE module.*

Position of MLE	Shared layers	Monolingual (Avg TER)	Code-Switch (MER)	LID Acc
-	12	8.8	10.4	98.8
{10-12}	9	7.0	8.4	**99.5**
{7-12}	6	**6.3**	**8.2**	99.4
{4-12}	3	6.8	9.3	99.1
{1-12}	0	11.6	16.2	97.7

As shown in Table 5, we explore the effect of the location the position of the MLE layers on performance. According to our analysis, the deeper the shared block, the more accurate the LID and the weaker the language-specific representation in the fix-depth model of FLR-MoE, which is a trade-off of the model design. Experiments show the proposed method has a strong ability to distinguish languages, and we achieve the best results with the language router in the middle position.

4.4.2. LID and Routing Analysis

We summarize the results of utterance-level language classification based on the frame-level LID routing information and the results of ASR. The proposed method's language classification accuracies and confusion matrix are shown in Fig. 3. Compared to the 98.8% average LID accuracy of the baseline, the proposed method's average LID accuracy is 99.4%. This shows that the proposed method can effectively reduce the confusion between languages.

As shown in Fig. 4, for the code-switching input, the proposed method obtains the routing of the language expert by token-to-frame LID alignment and routes the layer inputs of the language segments to the corresponding language experts, which demonstrates the effectiveness of the proposed method for code-switching ASR.
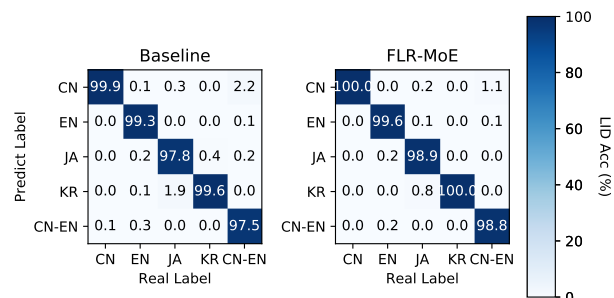

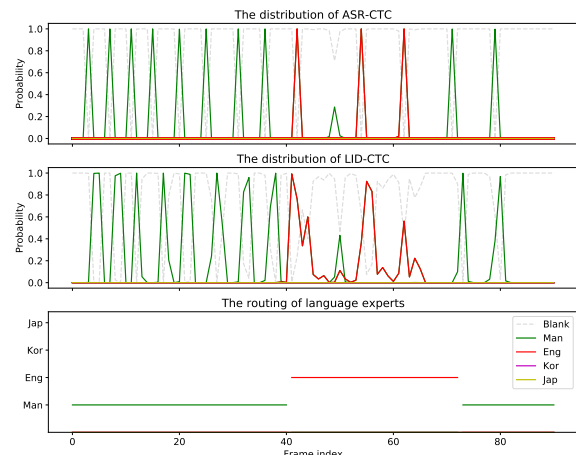Figure 3: *Confusion matrix of language over the multilingual evaluation sets. left: Baseline, right: FLR-MoE.*

Figure 4: *Visualization of the distribution of ASR-CTC, LID-CTC and language routing for the Mandarin-English code-switching speech.*

5. Conclusions

This paper proposes the LR-MoE architecture to improve multilingual ASR system for monolingual and code-switching situations. Based on the frame-wise language-routing (FLR) mechanism, the proposed LR-MoE can switch the corresponding language expert block to extract language-specific representations adaptively and efficiently. Experiments show the LR-MoE significantly improves multilingual and code-switching ASR system over the standard Transformer model with comparable computational complexity and outperforms the previous MoE-based methods with less computational complexity. In the future, we will explore more efficient MoE routing mechanisms for multilingual and code-switching speech recognition.

6. References

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

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

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*.

[4] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Neural Information Processing Systems*, 2015.

[5] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[7] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint ctc-attention based end-to-end speech recognition with a deep cnn encoder and rnn-lm," *Proc. Interspeech 2017*, pp. 949–953, 2017.

[8] G. Ma, P. Hu, J. Kang, S. Huang, and H. Huang, "Leveraging Phone Mask Training for Phonetic-Reduction-Robust E2E Uyghur Speech Recognition," in *Proc. Interspeech 2021*, pp. 306–310.

[9] G. Ma, P. Hu, N. Yolwas, S. Huang, and H. Huang, "PM-MMUT: Boosted Phone-mask Data Augmentation using Multi-Modeling Unit Training for Phonetic-Reduction-Robust E2E Speech Recognition," in *Proc. Interspeech 2022*, pp. 1021–1025.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[11] A. Pratapa, "Language modeling for code-mixing: The role of linguistic theory based synthetic data," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.

[12] G. Lee, X. Yue, and H. Li, "Linguistically motivated parallel data augmentation for code-switch language modeling," in *Interspeech 2019*.

[13] M. Ma, B. Ramabhadran, J. Emond, A. Rosenberg, and F. Biadsy, "Comparison of data augmentation and adaptation strategies for code-switched automatic speech recognition," in *ICASSP*, 2019.

[14] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.

[15] 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 2019*, pp. 2130–2134.

[16] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[17] C. Jacobs, Y. Matusevych, and H. Kamper, "Acoustic word embeddings for zero-resource languages using self-supervised contrastive learning and multilingual adaptation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*.

[18] R. Lahiri, K. Kumatani, E. Sun, and Y. Qian, "Multilingual speech recognition using knowledge transfer across learning processes," *arXiv preprint arXiv:2110.07909*, 2021.

[19] W. Hou, Y. Wang, S. Gao, and T. Shinozaki, "Meta-adapter: Efficient cross-lingual adaptation with meta-learning," in *ICASSP*, 2021.

[20] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, "Transformer-transducers for code-switched speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5859–5863.

[21] 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.

[22] B. Yan, C. Zhang, M. Yu, S.-X. Zhang, S. Dalmia, D. Berrebbi, C. Weng, S. Watanabe, and D. Yu, "Joint modeling of code-switched and monolingual asr via conditional factorization," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6412–6416.

[23] T. Song, Q. Xu, M. Ge, L. Wang, H. Shi, Y. Lv, Y. Lin, and J. Dang, "Language-specific characteristic assistance for code-switching speech recognition," *ArXiv*, vol. abs/2206.14580, 2022.

[24] J. Tian, J. Yu, C. Zhang, Y. Zou, and D. Yu, "LAE: Language-Aware Encoder for Monolingual and Multilingual ASR," in *Proc. Interspeech 2022*, 2022, pp. 3178–3182.

[25] 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)*, pp. 6234–6238.

[26] 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.

[27] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[28] 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.

[29] K. Kumatani, R. Gmyr, F. C. Salinas, L. Liu, W. Zuo, D. Patel, E. Sun, and Y. Shi, "Building a great multi-lingual teacher with sparsely-gated mixture of experts for speech recognition," *ArXiv*, vol. abs/2112.05820, 2021.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] D. Wang, S. Ye, X. Hu, S. Li, and X. Xu, "An end-to-end dialect identification system with transfer learning from a multilingual automatic speech recognition model." in *Interspeech*, 2021, pp. 3266–3270.

[32] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.

[33] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*. IEEE, pp. 1–5.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[35] 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 2019*, pp. 2613–2617.

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *arXiv preprint arXiv:1508.07909*, 2015.

[37] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.