INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Bi-encoder Transformer Network for Mandarin-English Code-switching Speech Recognition using Mixture of Experts

Yizhou Lu, Mingkun Huang, Hao Li, Jiaqi Guo, Yanmin Qian

MoE Key Lab of Artificial Intelligence
SpeechLab, Department of Computer Science and Engineering
AI Institute, Shanghai Jiao Tong University, Shanghai
{luyizhou4, mingkunhuang, lh575526, guojiaqi, yanminqian}@sjtu.edu.cn

Abstract

Code-switching speech recognition is a challenging task which has been studied in many previous work, and one main challenge for this task is the lack of code-switching data. In this paper, we study end-to-end models for Mandarin-English code-switching automatic speech recognition. External monolingual data are utilized to alleviate the data sparsity problem. More importantly, we propose a bi-encoder transformer network based Mixture of Experts (MoE) architecture to better leverage these data. We decouple Mandarin and English modeling with two separate encoders to better capture language-specific information, and a gating network is employed to explicitly handle the language identification task. For the gating network, different models and training modes are explored to learn the better MoE interpolation coefficients. Experimental results show that compared with the baseline transformer model, the proposed new MoE architecture can obtain up to 10.4% relative error reduction on the code-switching test set.

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

1. Introduction

Code-switching, including inter-sentential code-switching and intra-sentential code-switching, occurs when a speaker switches from one language to another. It's a common phenomenon in many multilingual communities [1, 2, 3]. Traditionally, automatic speech recognition (ASR) systems consist of acoustic, pronunciation and language models that are optimized independently [4]. In the scenario of code-switching, one challenge for traditional ASR based system is the requirement of hand-crafted components, such as a well-designed mixed phone set and the corresponding pronunciation lexicon [5].

End-to-end (E2E) models directly optimize the probability of output sequences given input speech observations with a single network, thus provide an elegant solution to build a ASR system. Recent work on E2E models can be categorized into three main approaches: Connectionist Temporal Classification (CTC) [6], RNN-Transducer [7, 8] and attention-based sequence-to-sequence models [9, 10]. Besides, joint CTC-attention model [11, 12] exploits the advantages from both CTC and sequence-to-sequence models within the multi-task learning framework, which leads to better performance and robustness. E2E ASR models have made promising progress in many areas including monolingual [13, 14], multilingual [15, 16] and multi-speaker [17] speech recognition task. It's also shown that attention-based sequence-to-sequence models are able to achieve state-of-the-art performance [13]. More recently, transformer network [18] that firstly proposed for Neural Machine Translation rapidly became the mainstream framework among other NLP tasks, and for ASR tasks it's shown to outperform RNN-based end-to-end models [19].

One major challenge for building a code-switching ASR system is the lack of code-switching training data, and this problem is even severe for E2E models. However, for Mandarin and English language where rich resources monolingual data are available, leveraging these external data can help alleviate this data sparsity issue. In prior work of multilingual ASR, it's observed that adding a one-hot language vector to condition the E2E model with specific language can boost the multilingual performance [15, 16, 20]. Similar strategy is also proposed in [21], where layer-wise gating mechanism is employed to adapt the model for specific language. But for intra-sentential code-switching task where languages change within a single utterance, obtaining the prior language identification (LID) information is not easy as it in multilingual ASR.

In this work, we study E2E approaches for Mandarin-English code-switching ASR task. To efficiently leverage the monolingual data, we propose a bi-encoder transformer network based mixture of experts (MoE) architecture. MoE models has been studied in many works including universal acoustic modeling [22] , multi-accent ASR [23] and language modeling [24]. Similar idea has also been applied in cluster adaptive training [25] for speaker adaptation. As for this Mandarin-English code-switching ASR task, two transformer encoders serve as Mandarin expert and English expert individually to provide different views, while a gating network is employed to weight the expert outputs. Unlike the normal transformer model, this MoE architecture enables the model to better capture language-specific information with separate encoders, and the language identification task is explicitly handled.

Moreover, we explore different gating network models and training modes to learn the MoE interpolation coefficients. We find that a single linear layer can well handle the LID task, and the MoE coefficients can be learned in an unsupervised mode. Experimental results show that the proposed bi-encoder based MoE architecture can obtain up to 10.4% relative error reduction over the baseline transformer model on a Mandarin-English code-switching test set. It's also observed that the code-switching performance can be further improved with an additional transfer learning stage.

The rest of this paper is organized as follows. In Section 2 we review the related work briefly. Then we describe the proposed MoE architecture for code-switching task. The proposed method is evaluated and results are analyzed in Section 4. Finally, we conclude the paper and discuss the future work.

Yanmin Qian is the corresponding author.

http://dx.doi.org/10.21437/Interspeech.2020-2485

2. Related work

Prior work on code-switching ASR is mainly in traditional hybrid systems [1, 2, 3]. Recently, motivated by the progress of E2E models, researchers have been interested in building E2E code-switching ASR system [5, 26, 27]. E2E CTC models for code-switching task was firstly explored in [26], where an additional LID classifier is introduced to adjust the posteriors of the initial CTC model. In [5, 27], LID based multitask learning was proposed to improve the performance of attention-based sequence-to-sequence models. Besides, there are also researches on augmenting output token set with LID tokens [28, 29].

In this work we focus on leveraging rich resources monolingual data to achieve a better code-switching ASR performance, and a new MoE structure is proposed to better handle Mandarin and English modeling.

3. Code-switching ASR with bi-encoder Mixture of Experts

In this section, we first give a brief review of the baseline transformer based E2E ASR, and then we describe the proposed bi-encoder transformer network based MoE architecture and related training strategies. The new approach mainly includes three parts: bi-encoder bilingual model pretraining, mixture of experts architecture construction and a gating network for MoE interpolation coefficients.

3.1. Revisit on transformer-based E2E ASR

Transformer network is a sequence-to-sequence structure that basically consists of encoder network and decoder network. The encoder network is composed of a stack of N identical layers, and each layer consists of multi-head self-attention and fully connected feed forward network [18]. It takes acoustic features $\mathbf{x}$ as input and maps $\mathbf{x}$ into high level representations $\mathbf{h}$. For ASR task, usually a front-end CNN network is adopted to do time-scale down-sampling [30].

$$\mathbf{h} = Encoder(\mathbf{x}) \tag{1}$$

The decoder network utilizes the encoded representation $\mathbf{h}$ with attention mechanism and outputs the predicted tokens auto-regressively. We denote the target sequence as $\mathbf{y}$, and at each decoding step, the decoder emits the posterior probabilities of the next token given previous outputs. We train the transformer model with joint CTC-attention [11, 12] framework to exploit the advantages from both CTC and S2S models. Denote $\mathcal{L}_{ctc}(\mathbf{y}|\mathbf{x})$ as the CTC objective loss, $\mathcal{L}_{s2s}(\mathbf{y}|\mathbf{x})$ as the S2S objective loss, the loss function of joint CTC-attention network is defined as:

$$\mathcal{L}_{jca}(\mathbf{y}|\mathbf{x}) = \lambda_{jca}\mathcal{L}_{ctc}(\mathbf{y}|\mathbf{x}) + (1-\lambda_{jca})\mathcal{L}_{s2s}(\mathbf{y}|\mathbf{x}) \tag{2}$$

with a tunable coefficient $\lambda_{jca} \in [0, 1]$ to control the contribution of each loss. Beam search decoding is adopted to predict the output sequence, where S2S scores together with CTC prefix scores are combined to make the decision.

For the modeling units, we combine Chinese characters and English BPE subwords [31] as final units. We also apply SpecAugment [14] for all data through out our experiments.

3.2. Pretrained bi-encoder bilingual model

We first pretrain a special bi-encoder bilingual model with only monolingual Mandarin and English data. Since language identity for monolingual data can be obtained in advance, we are able to decouple Mandarin and English language with two separate encoder[1]. As shown in the left part of Figure 1, when given acoustic features inputs, prior LID information is used to decide which encoder to use. Denote $\mathbf{X}_{cn}$ and $\mathbf{X}_{en}$ as the collection of all Mandarin inputs and English inputs separately, we formulate this procedure as:

$$\mathbf{h}^{enc} = \begin{cases} MandarinEncoder(\mathbf{x}) \text{ if } \mathbf{x} \in \boldsymbol{X}_{cn} \\ EnglishEncoder(\mathbf{x}) \text{ if } \mathbf{x} \in \boldsymbol{X}_{en} \end{cases} \tag{3}$$

The output embedding $\mathbf{h}^{enc}$ is further utilized in a CTC layer and a decoder network, which are shared across the two language. For Mandarin and English language where rich resources speech data are available, both the two encoder can be well trained with specific language data, without the disturbance from the other language domain. This separate modeling structure is much more flexible, which has the potential advantage of direct model structure adjustment to a specific language.

3.3. Mixture of Experts architecture for CS ASR

The pretrained bi-encoder bilingual model is able to handle both Mandarin and English modeling, however, it's unable to perform intre-sentential code-switching. Motivated by recent work on MoE [22, 23], bi-encoder transformer network based MoE architecture is explored to address code-switching ASR, and Mandarin and English encoders in the bilingual model are treated as two language experts. The pretrained bilingual model in the last subsection is used for initialization, and all monolingual Mandarin and English data together with code-switching data are utilized in this stage. Since LID information is not known in advance, we let the two experts in parallel provide two different expert views $\mathbf{h}^{cn}$ and $\mathbf{h}^{en}$:

$$\mathbf{h}^{cn} = MandarinEncoder(\mathbf{x}) \tag{4}$$

$$\mathbf{h}^{en} = EnglishEncoder(\mathbf{x}) \tag{5}$$

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

$$\mathbf{h}_t^{mix} = \alpha_t^{cn}\mathbf{h}_t^{cn} + \alpha_t^{en}\mathbf{h}_t^{en} \tag{6}$$

where the two scalar coefficients α_t^{cn} and α_t^{en} range from $[0, 1]$ and the sum of the coefficients equals to one for all frames. In the single language situation, e.g. pure Mandarin speech, we expect the model to attend more on the Mandarin encoder and so α_t^{cn} should be larger and even close to one while α_t^{en} is close to zero. In the scenario of code-switching, the MoE coefficients can control the language switch within the utterance.

3.4. Gating network for MoE interpolation

We develop a gating network to predict the MoE interpolation coefficients, and different methods are compared. One straight forward method is to train an external language identification (LID) classifier, and in our experiments we train a self attention network (SAN) based model for LID classification. We refer to this method as external LID method. In this

[1] It is noted that the two encoder can use different structures, but for simpleness we choose the identical transformer encoders in our experiments.

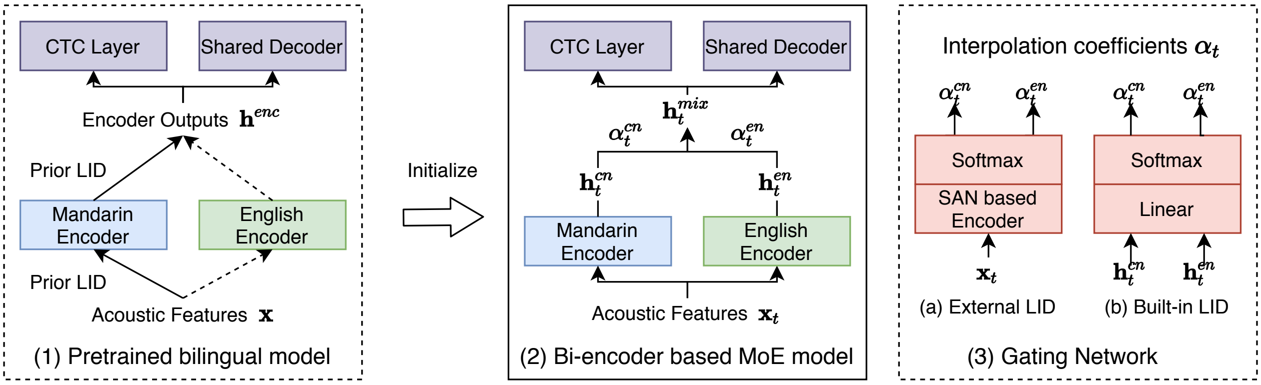


Figure 1: *The proposed bi-encoder transformer network based MoE architecture: (1) pretrained bi-encoder bilingual model; (2) mixture of experts architecture for code-switching ASR; (3) gating network for MoE interpolation coefficients;*

method, raw input features $\mathbf{x}$ are used to train the LID module in advance, and the output probability of each language is directly used to weight the experts outputs. The LID and ASR are trained independently. To improve the performance of LID classifier, we adopt transfer learning strategy and a pretrained CTC model is used for initialization.

In the second built-in LID method, we use the outputs of the two separate encoders as inputs to the gating network, thus the gating network is aware of expert outputs. We think that this high-level representation features are better for LID classification. For this build-in LID method, the ASR and LID modules in this MoE architecture can be trained jointly, and the objective loss is changed to:

$$\mathcal{L}_{mtl}(\mathbf{y}|\mathbf{x}) = \mathcal{L}_{jca}(\mathbf{y}|\mathbf{x}) + \lambda_{lid}\mathcal{L}_{lid}(\mathbf{y}_{lid}|\boldsymbol{\alpha}) \tag{7}$$

where $\mathbf{y}_{lid}$ is the LID target and $\boldsymbol{\alpha}$ the predicted MoE interpolation coefficients. This formulation includes two training modes of the gating networks: when $\lambda_{lid} > 0$ it means the supervised gating network training mode, and in contrast $\lambda_{lid} = 0$ means the unsupervised training mode.

Since the high-level representation $\mathbf{h}_t^{cn}$ and $\mathbf{h}_t^{en}$ already maintain rich linguistic information, the interpolation coefficients $\boldsymbol{\alpha}_t = [\alpha_t^{cn}, \alpha_t^{en}]^T$ can be modeled with a single linear layer:

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

4. Experiments

4.1. Experimental setup

Our experiments are conducted on ASRU 2019 Mandarin-English code-switching Challenge dataset, which consists of about 500 hours Mandarin data and 200 hours code-switching data. For English corpus, we choose a subset of 460 hours data from Librispeech corpus [32] to match the size of Mandarin data. Additional 20 hours code-switching data is reserved as development set. For system evaluation, we used three test sets: Mandarin test set (ZH), English test set (EN) and Mandarin-English code-switching test set (CS$_{eval}$).

For acoustic feature, 80 dimensional log-mel filterbanks are extracted with a step size of 10ms and window size of 25ms, and utterance-level CMVN is applied on the fbank features. As for modeling unit, we combine Chinese characters and English BPE subword units [31]. We choose Mandarin characters that occur more than 25 times in the training data, which results in 3003 characters, and the other characters are mapped into unk symbol. We generate 1000 BPE units for English, and there are totally 4006 tokens for modeling (with two extra tokens for $blank$ and sos/eos).

We report character error rate (CER) and word error rate (WER) for pure Mandarin and English test set respectively. As for the code-switching test set, we report Mandarin part CER, English part WER and the total mix error rate (MER), as those in ASRU2019 Challenge.

4.2. Performance evaluation of baseline systems

We use ESPnet toolkit [33] to train our baseline transformer model. We use 12-layer transformer in encoder and 6-layer in decoder, all with attention dimension 256. We apply SpecAugment [14] and fix λ_{jca} with 0.3 throughout our experiments. In the decoding stage, we use a beam search size of 8 and decoding CTC weight of 0.4.

Table 1: *Performance (CER/WER) (%) comparison of baseline systems trained with different data. "CHN", "ENG" and "CS" mean Chinese, English and code-switching training data respectively, "ALL" denotes using both code-switching and two monolingual training datasets. "n/a" means that results are not available for that system. Code-switching performance (CS$_{eval}$) is reported with Mandarin part CER, English part WER and the total MER.*

Model	TR-Data	ZH	EN	CS$_{eval}$		
				ZH	EN	MIX
Baseline	CHN	**2.93**	n/a	n/a	n/a	n/a
	ENG	n/a	**9.93**	n/a	n/a	n/a
	CS	n/a	n/a	9.60	30.18	11.84
	ALL	4.62	11.85	**8.79**	**28.14**	**10.89**

We present the performance of baseline systems in Table 1. It is observed that the monolingual system can obtain a low error rate on the monolingual test sets, but it cannot handle the cross-lingual or code-switching task. The system using only code-switching training data can perform code-switching, but the performance is not satisfactory due to the lack of code-switching training data, with a MER of 11.84%. The last line of Table 1 shows that pooling all the data together can make the system recognize all kinds of data, and the code-switching performance is significantly improved. However, on monolingual testing sets it performs much worse compared with the monolingual models. We hypothesize that the potential of these monolingual data is not fully exploited.

4.3. Evaluation of the proposed bi-encoder MoE architecture

We evaluate of the proposed method in this section. A bi-encoder bilingual model with two identical 12-layer transformer encoder (dimension of 256) is pretrained, as introduced in Section 3.2. Later we use this pretrained model for initialization to train our MoE model, with different gating networks. For a fair comparison, we re-build a baseline with a larger encoder (dimension of 512), so that the mode size can be similar as the proposed MoE model. The encoder representation is projected down to the decoder dimension with a layer-normalized affine transformation, and the other condition are controlled the same as the previous baseline transformer models and MoE model.

Table 2: *Performance comparison (CER/WER) (%) of different systems trained with all monolingual and code-switching data. The middle part gives the performance of newly proposed MoE systems, with external LID (MoE-ext) and built-in LID (MoE-in) gating networks to learn the MoE coefficients respectively. For built-in gating networks, supervised and unsupervised modes correspond $\lambda_{lid} = 0.1$ and $\lambda_{lid} = 0.0$ respectively. Noted that the parameters of external LID classifier are counted for external LID method, thus the parameters are much larger.*

Model	#Params	ZH	EN	CS_{eval}		
				ZH	EN	MIX
Baseline	28.8M	4.62	11.85	8.79	28.14	10.89
Large Enc.	55.2M	4.30	11.34	8.50	27.69	10.58
MoE-ext	62.4M	**3.25**	**9.92**	7.87	26.91	9.94
MoE-in-sup	45.6M	3.28	9.99	7.75	26.76	9.82
MoE-in-unsup	45.6M	3.27	9.94	**7.70**	**26.64**	**9.76**
+ CS retrain	45.6M	5.44	27.96	7.34	25.13	9.27

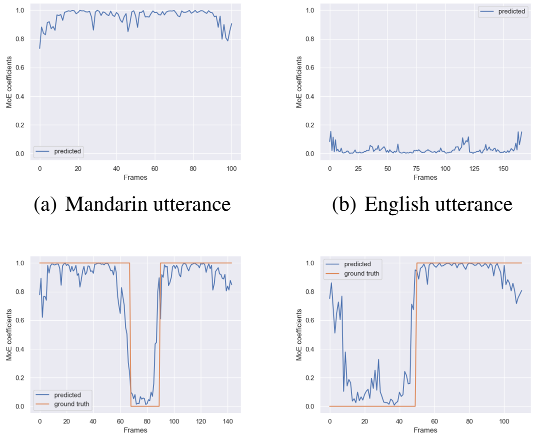
The performance of baselines and MoE systems are shown in the top and middle parts of Table 2. It is observed that the proposed bi-encoder based MoE model has better ability to leverage the monolingual data, and the performance on the monolingual sets even approaches that of the pure monolingual systems in Table 1, demonstrating the efficiency of separate encoder modeling. Besides, the proposed method also achieves a significant improvement on code-switching test sets, with up to 10.4% relative error reduction over baseline transformer model.

4.4. Evaluation of different gating networks

Moreover, we compare different LID gating networks for MoE coefficients, and the results are illustrated in the middle part of Table 2. For the external LID method, we train a 12-layer SAN based model for classifying the frame-level LID, and a pretrained CTC model is used as seed model for initialization to obtain higher LID accuracy. For the built-in LID method, only a single linear layer is employed for the LID task. We also tried to replace the linear layer with a more complicated LSTM structure, but no further improvements were obtained. It is observed that the built-in gating networks outperforms external one. For the two modes in built-in gating networks, the unsupervised mode outperforms the supervised one slightly.

To better improve the system performance on the code-switching data, we employ transfer learning strategy to retrain the MoE system with only code-switching training data, and the results are shown as the last line of Table 2. Re-finetune with the domain-dependent code-switching data can get an additional gain on the code-switching testing set, however there will be a large degradation on the domain-mismatched monolingual testing sets.

4.5. Analysis of MoE coefficients



(a) Mandarin utterance

(b) English utterance

(c) Code-switching utterance-1

(d) Code-switching utterance-2

Figure 2: *Visualization of the unsupervised learned MoE coefficients α_{cn} of built-in LID method.*

We visualize the unsupervised learned MoE coefficients for differnt utterances, including monolingual Mandarin and English utterances and code-switching utterances. As shown in Figure 2(a) and 2(b), when the input utterance is pure Mandarin or English, in most of the frames the MoE coefficients α_{cn} are close to 1 or 0 for Mandarin and English respectively. From Figure 2(c) and 2(d), it is observed that the MoE coefficients from the gating network can perform code-switching well following the real switch point, which further demonstrates the effectiveness of the proposed new architecture for code-switching E2E ASR task.

5. Conclusion and future work

In this paper we proposed a bi-encoder transformer network based MoE architecture to improve E2E based Mandarin-English code-switching speech recognition. The proposed new model has better capacity to leverage monolingual data, which contributes to its code-switching performance. Besides, we developed different approaches to learn the MoE interpolation coefficients. We also employ transfer learning strategy to better improve the code-switching performance.

In the future, we plan to study hierarchical attention network to further improve the proposed bi-encoder based MoE system on the code-switching ASR. we also plan to further improves code-switching ASR with some knowledge distillation approaches from the monolingual systems.

6. Acknowledgements

This work has been supported by the National Key Research and Development Program of China (Grant No.2017YFB1302402). Experiments have been carried out on the PI supercomputers at Shanghai Jiao Tong University.

7. References

[1] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4889–4892.

[2] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in french/algerian arabic speech," in *Interspeech 2017*, 2017, pp. 62–66.

[3] E. Yılmaz, A. Biswas, E. van der Westhuizen, F. de Wet, and T. Niesler, "Building a unified code-switching asr system for south african languages," *arXiv preprint arXiv:1807.10949*, 2018.

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "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.

[5] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Investigating end-to-end speech recognition for mandarin-english code-switching," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6056–6060.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

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

[8] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[10] 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, pp. 4960–4964.

[11] 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.

[12] 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," *arXiv preprint arXiv:1706.02737*, 2017.

[13] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[14] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[15] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[16] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.

[17] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," *arXiv preprint arXiv:1910.06522*, 2019.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[19] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on transformer vs rnn in speech applications," *arXiv preprint arXiv:1909.06317*, 2019.

[20] 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," *arXiv preprint arXiv:1909.05330*, 2019.

[21] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4914–4918.

[22] A. Das, J. Li, C. Liu, and Y. Gong, "Universal acoustic modeling using neural mixture models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5681–5685.

[23] A. Jain, V. P. Singh, and S. P. Rath, "A multi-accent acoustic model using mixture of experts for speech recognition," *Proc. Interspeech 2019*, pp. 779–783, 2019.

[24] K. Irie, S. Kumar, M. Nirschl, and H. Liao, "Radmm: recurrent adaptive mixture model with applications to domain robust language modeling," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6079–6083.

[25] T. Tan, Y. Qian, M. Yin, Y. Zhuang, and K. Yu, "Cluster adaptive training for deep neural network," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4325–4329.

[26] K. Li, J. Li, G. Ye, R. Zhao, and Y. Gong, "Towards code-switching asr for end-to-end ctc models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6076–6080.

[27] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," *arXiv preprint arXiv:1811.00241*, 2018.

[28] M. S. Mary N J, V. M. Shetty, and S. Umesh, "Investigation of methods to improve the recognition performance of tamil-english code-switched data in transformer framework," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7889–7893.

[29] S. Zhang, J. Yi, Z. Tian, J. Tao, and Y. Bai, "Rnn-transducer with language bias for end-to-end mandarin-english code-switching speech recognition," *arXiv preprint arXiv:2002.08126*, 2020.

[30] S. Karita, N. E. Y. Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. Interspeech 2019*, 2019, pp. 1408–1412. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1938

[31] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[32] 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.

[33] 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.