Towards Language-universal Mandarin-English Speech Recognition with Unsupervised Label Synchronous Adaptation

Song Li[1], *Haoneng Luo*[3], *Wenxuan Hu*[2], *Yuan Liu*[3], *Shiliang Zhang*[3], *Lin Li*[1], *Qingyang Hong*[2]

[1] School of Electronic Science and Technology, Xiamen University, China
[2]School of Informatics, Xiamen University, China
[3]Speech Lab, Alibaba Group

songli@stu.xmu.edu.cn

Abstract

End-to-end multilingual and code-switching speech recognition are two challenging tasks that are studied separately in many previous works. In this work, we jointly study multilingual and code-switching problems and present a novel unsupervised label synchronous adaptation algorithm for Mandarin-English speech recognition. Specifically, we use two parallel encoders to decompose the Mel-spectrum of speech into semantic information and other acoustic attributes, such as speaker identity, accents, pronunciation characteristics of different languages, etc. During the autoregressive decoding process of the speech recognition system, an adaptive decoder is used in parallel with the speech recognition decoder to generate an adaptive embedding for each character, so that the speech recognition model can be adaptive for Mandarin, English, and code-switching cases. Our experiments show that our proposed algorithm obtains 13.5% relative error reduction over a strong baseline in the code-switching case, and outperforms both the state-of-the-art Mandarin and English monolingual models.

Index Terms: speech recognition, multilingual, code-switching, unsupervised adaptation.

1. Introduction

As voice-driven interface to smart devices becomes mainstream, increasing the language coverage of speech recognition systems is particularly important [1]. There exists thousands of languages in human speech interaction, including various official languages and different dialects. Usually, language-specific automatic speech recognition (ASR) system is built for each language. However, as the number of supported languages continuously grows, it will dramatically increase the effort required to train, deploy, and maintain so many ASR systems in a production environment. Moreover, the code-switching phenomenon [2] that contains more than one language within an utterance is another great challenge to ASR service. Therefore, how to deal with these multilingual and code-switching problems have gained more and more attention in recent years.

With the rapid development of end-to-end speech recognition technology [3–8], more and more researchers are exploring end-to-end multilingual speech recognition. Previous research works can be classified into two main categories, the first one uses language information to guide the end-to-end speech recognition system to recognize speech in different languages. For example, Ref. [9] proposed to use a language identification (LID) model to obtain a language embedding as a bias, which can make the ASR model adaptive to different languages. Ref. [10] proposed to insert language IDs into the training label for joint training ASR and LID tasks. Another category of approach focuses on performing parameters decomposition. Ref. [11, 12] proposed to decompose the parameters of the Transformer into language-dependent and language-independent parameters, Ref. [13] set up a individual decoder for each language to recognize speech in different languages, and Ref. [14] used the mixture of experts (MOE) model to assign different encoders to each language.

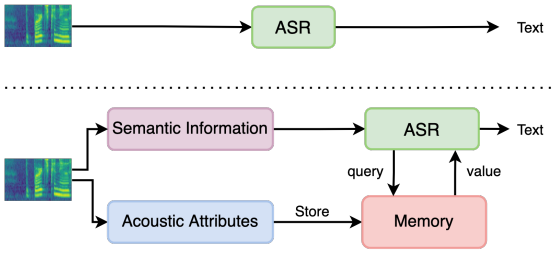

Figure 1: *The motivation of our proposed adaptive algorithms.*

Code-switching speech recognition has a greater challenge than multilingual speech recognition because the speech contains switching of multiple languages. Previous research works have explored two main aspects, the first one uses bi-encoder or bi-decoder approach [15, 16] to fuse Mandarin-English encoders or decoders to process code-switching speech. The other one [17–19] combines the language identification task with ASR, making the ASR model capable of recognizing language-switching points.

Most of the previous researchers focused on the challenges with limited training data, or low-resource data for their studies. However, in the case of a large amount of training data, some existing methods do not achieve performance improvement due to the interference of similar pronunciation in different languages [13] . Therefore, we focus on large training data to facilitate multilingual and code-switching speech recognition for industrial applications. As shown in Fig.1, pure end-to-end speech recognition model extracts semantic information from the Mel-spectrum in order to convert speech to text, and removes as many other acoustic attributes as possible. In fact, the Mel-spectrum contains many valuable acoustic attributes, such as speaker identity, accents, pronunciation characteristics of different languages, etc [20, 21]. In this paper, we propose an unsupervised label synchronous adaptation algorithm to extract the acoustic attributes contained in the Mel-spectrum and store them in the memory. When the ASR system generates a certain character, it can retrieve the acoustic attributes needed to generate this character from the memory. The acoustic at-

979-8-3503-9796-3/22/$31.00

tributes are used as auxiliary information for ASR system to be adaptive to different languages. As a plug-and-play module, our proposed algorithm does not introduce additional training steps and achieves good performance in Mandarin, English, and code-switching, facilitating the deployment of ASR in real-world environments.

The rest of the paper is organized as follows. In Section 2, the details of our proposed adaptive algorithm are described. In Section 3, we introduce specific experimental details. Experimental results are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Proposed technology

2.1. System overview

End-to-end speech recognition directly models the conditional probability $P(Y|X)$ of the input acoustic feature sequence X and the corresponding text sequence Y. To allow the end-to-end speech recognition model to be adaptive to different input features, we can provide it with a bias term as an additional condition to guide the model, i.e., model the probability $P(Y|X,B)$, where B is the bias term. In previous studies, B is usually obtained by extracting a embedding from a trained classifier, such as x-vector and language embedding, we refer to these embedding as adaptive embedding in this paper. In contrast, the unsupervised adaptive algorithms proposed in this paper enables the network to automatically learn the required adaptive embeddings, without the need to provide the corresponding classification labels to train a classifier network, which allowing the entire neural network to be trained in an end-to-end manner and greatly simplifying the training process. Our proposed unsupervised label synchronous adaptation algorithm generates individual adaptive embedding for each recognized character to perform fine-grained ASR adaptation. As a comparison, we also implement an unsupervised global adaptive algorithm which generates a global adaptive embedding for all characters. Finally, we introduce sequence-level multi-task learning techniques to further improve the speech recognition performance in the code-switching case.

2.2. Unsupervised global adaptation

The backbone ASR network in this paper is based on Transformer architecture, which mainly consists of two parts: encoder and decoder. The unsupervised global adaptation (UGA) extracts a global adaptive embedding from the input acoustic features and uses this adaptive embedding as a bias term to guide the speech recognition model at each decoding step. As shown in Fig.2, the green box diagram is the encoder and decoder of the Transformer-based ASR [22]. UGA employs an adaptive encoder parallel with the ASR encoder to extract a intermediate representation from the Mel-spectrum, and uses a statistical pooling layer [23] to generate a fixed-length vector as a query for the adaptive vector group to generate a global adaptive embedding. And then, this global adaptive embedding is used as a bias and added to the context vector of the src-attention [24] of the ASR decoder to achieve ASR adaption.

The adaptive vector group is a key component of our adaptive algorithms. As shown in Fig.3, an adaptive vector group block consists of N learnable vectors. By computing the attention weights of the query provided by the adaptive encoder with the N learnable vectors, we obtain the corresponding adaptive embedding. The textual semantic information contained in the Mel-spectrum is essentially a continuous information, which

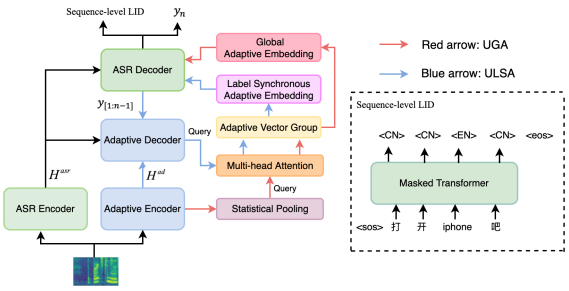

Figure 2: *The description of our proposed adaptive algorithms.*

needs to be represented using a continuous feature sequence. In contrast, the non-semantic information contained in the Mel spectrum, such as speaker identity, accents, pronunciation characteristics of different languages, etc., is inherently discrete global information. Thus, the N learnable vectors contained in the adaptive vector group are responsible for discretizing the extracted depth features, so that our adaptive algorithms can extract non-semantic information instead of extracting textual semantic information that is duplicated by the ASR encoder. Let's go back to Fig.1, these learnable vectors in the adaptive vector group serve as the memory and are responsible for storing the non-semantic information (acoustic attributes) contained in the Mel-spectrum. The semantic information, on the other hand, is provided by the ASR encoder, so our adaptive algorithms essentially performs the disentanglement of semantic and non-semantic information contained in the Mel-spectrum. In addition, these learnable vectors are trained with the loss function of speech recognition and optimized by the back propagation algorithm [25], without introducing additional acoustic attribute classification labels, so they belong to unsupervised learning.

Moreover, we try to design a hierarchical modeling approach to store this non-semantic information in a hierarchical manner. Specifically, we use the output of the previous adaptive vector group as the input of the next adaptive vector group and employ residual connections to make the stored information in different groups recursive rather than repetitive.

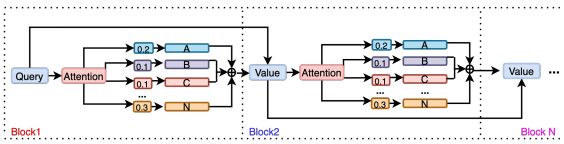

Figure 3: *The structure of the hierarchical adaptive vector group.*

2.3. Unsupervised label synchronous adaptation

The UGA algorithm performs adaptation by generating a global adaptive embedding as the bias of ASR. However, a global embedding is not fine-grained enough to accurately perform fine adaption for each character generated by ASR. Therefore, we propose the unsupervised label synchronous adaptation (ULSA) algorithm. As shown in Fig.2, compared to UGA, the ULSA additionally introduces an adaptive decoder to synchronize with the ASR decoder. In the autoregressive decoding process of the Transformer-based ASR model, the adaptive decoder generates

a query for the character generated in the previous step, and then uses the query to get the adaptive embedding corresponding to the character from the adaptive vector group to perform character-level (label synchronous) ASR adaptation.

We carefully design the adaptive decoder so that the generation of the adaptive embedding can be synchronized with ASR decoding. Compared with the standard Transformer decoder [24], our adaptive decoder extends an additional src-attention to synchronize the ASR decoder and the adaptive decoder. The calculation process of the adaptive decoder is as follows:

$$Query^1 = MHA(H^{asr}, F(y_{n-1}), H^{ad}) \quad (1)$$
$$Query^2 = MHA(Query^1, Query^1, Query^1) \quad (2)$$
$$Query^3 = MHA(Query^2, H^{ad}, H^{ad}) \quad (3)$$

where H^{asr} and H^{ad} are the outputs of the ASR encoder and the adaptive encoder respectively, y_{n-1} is the output of the previous decoding step of ASR, and F is the masked self-attention in the ASR decoder. For simplicity, the feedforward network and residual connection of Transformer are omitted. The first multi-head attention (MHA) of the adaptive decoder multiplexes the query and key of the src-attention of the ASR decoder, and uses H^{ad} as the value, which enables the attention windows of the two decoders to be synchronized. The second MHA is responsible for further extraction of high-level features on $Query^1$, and the last MHA is responsible for further extracting the adaptive context vector corresponding to y_{n-1} from H^{ad} using $Query^2$ as the query.

We obtain the adaptive embedding (B_{n-1}) with respect to y_{n-1} by inputting $Query^3$ into the adaptive vector group for attention calculation, and add to the context vector of the src-attention of the ASR decoder to achieve ASR adaption. To further improve the performance, as shown in Fig.4, we use a linear layer to assign the attention weights of the adaptive embedding and the src-attention context vector of the ASR decoder and perform a weighted summation, which we call the adaptive weight assignment. Finally, the label synchronous adaptation algorithm can synchronize with the ASR decoding process to generate the adaptive embedding for each character to achieve a more refined adaption. Specifically, when Transformer-based ASR performs autoregressive decoding, the next character to be generated in each decoding time step is no longer determined by the context vector only, but is determined by both the context vector and the adaptive embedding of the previous character, i.e., modeling $P(y_n|y_1, y_2, ..., y_{n-1}, X, B_{n-1})$.

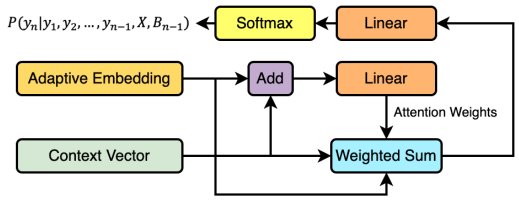

Figure 4: *Description of the adaptive weight assignment.*

2.4. Sequence-level multi-task learning

For multilingual and code-switching speech recognition systems, language identification capability is important for performance improvement. Unlike previous papers [9, 10, 19] , in which a separate language classifier is trained or language IDs are inserted in the training labels. We propose to use sequence-level multi-task learning (SMTL) for parallel language identification with ASR. As shown in Fig.2, we generate a language ID for each word for the training labels, where $<CN>$ represents Mandarin and $<EN>$ represents English. During the training process, we train the sequence-level language identification task and the ASR task simultaneously, so that the ASR can determine the language of each character while generating it during ASR decoding. The total loss function is shown as follows:

$$L_{SMTL} = L_{ASR} + \alpha L_{LID} \quad (4)$$

where L_{ASR} is the loss function of ASR and L_{LID} is the cross-entropy loss function of the language identification task. α is a hyperparameter to balance the two losses. In the inference stage, we no longer need the language identification branches, so no additional number of parameters is added.

3. Experimental setup

3.1. Data sets

We conduct our experiments on two large Mandarin (contains a small number of code-switching utterances) and English corpora that consist of about 20,000 hours and 19,000 hours data respectively. We divide the data into training set and development set, which contain of 97% and 3% data, respectively. These monolingual Mandarin corpus and English corpus are further mixed to form the bilingual Mandarin-English corpus, which is used to train bilingual and code-switching ASR models. For model inference, we have built three test sets: Mandarin test set (about 30 hours), English test set (about 10 hours) and Mandarin-English code-switching test set (about 7 hours). Acoustic features are 80-dimensional energy-based Log Mel-filterbanks (FBank) computed on a window of 25ms with a 10ms shift. A low frame rate (LFR) [26] is made by stacking consecutive frames into a size of seven context window (3+1+3), and then down-sampling the input frame rate to 60ms. For Mandarin, characters are used as the modeling units. For English, we use Byte Pair Encoding [27] to generate 1000 subwords as the modeling units.

3.2. Implementation details

To evaluate the performance of our proposed adaptive algorithm comprehensively, we set up three sets of strong baselines: monolingual models in Mandarin and English respectively, and multilingual models trained with mixed Mandarin and English data. For ASR model, we used the DFSMN-augmented Transformer: SAN-M [28], and we increased its encoder to 40 layers and its decoder to 12 layers, which makes our baseline models strong enough to fully evaluate the performance of different algorithms. In addition, we implemented some mainstream multilingual and code-switching speech recognition algorithms based on SAN-M for comparison.

For both adaptive algorithms, we use 6-layer SAN-M as the adaptive encoder and explore the impact of different adaptive vector groups in terms of the number (N) of vectors and vector dimensions (D) on the ASR performance. For the label synchronous adaptation algorithm, the adaptive decoder needs to be synchronized with the ASR decoder, so we explored different placement (P) of the adaptive decoder. Finally, we design hierarchical unsupervised label synchronous adaptation (H-ULSA) algorithm and explore the impact of different adaptive vector

group layers (L) on ASR performance.

All experiments are conducted based on the Tensorflow toolkit [29]. During training, we adopt the LazyAdam optimizer [30] with β_1= 0.9, β_1 = 0.998, and a Noam learning rate decay strategy [31] with d = 320, $warmup\_n$= 8000, and k = 1. In addition, label smoothing and dropout regularization rate of 0.1 are employed to prevent over-fitting.

4. Results

4.1. The effects of SMTL

As seen in Table 1, the performance of the Mandarin-English mixed baseline becomes worse on the monolingual test set compared to the monolingual baseline, because of the interference from similar pronunciations in both languages. Therefore, we used SMTL to enhance the linguistic discrimination ability of ASR. As described in Section 2.3, we tried several values of α and found that optimal performance was achieved when α was set to 0.2, but it still did not meet our expectation, so we continued to explore UGA, ULSA and H-ULSA.

Table 1: *Comparison of WERs(%) for the three strong baselines and SMTL-augmented baselines.*

Model	Mandarin	English	Code-switching
Mandarin Baseline	**8.76**	-	22.01
English Baseline	-	**10.97**	-
Mandarin-English Mixed Baseline	9.96	11.44	16.05
+SMTL (α =0.1)	9.91	11.38	15.43
+SMTL (α =0.2)	9.41	11.05	**15.14**
+SMTL (α =0.3)	9.41	11.15	15.34
+SMTL (α =0.4)	9.57	11.21	15.44
+SMTL (α =0.5)	9.62	11.24	15.27

4.2. The effects of UGA

As discussed in section 3.2, we explored the effects of different N and D of UGA on ASR performance, and we found that optimal performance was achieved for N=5, D=256. As shown in Table 2, UAG brings ASR performance improvements in all three cases, especially in the English and code-switching cases. In addition, the performance of UGA has been further improved by introducing SMTL with α set to 0.2.

Table 2: *Comparison of WERs(%) for different UGA configurations.*

Model	Mandarin	English	Code-switching
UGA (N=5, D=128)	9.52	10.78	15.23
UGA (N=5, D=256)	9.39	10.75	15.13
+SMTL (α =0.2)	**9.25**	**10.53**	**14.77**
UGA (N=5, D=512)	9.72	10.65	15.58
UGA (N=10, D=256)	9.82	10.97	15.37
UGA (N=15, D=256)	10.64	11.11	15.88

4.3. The effects of ULSA

UGA did not achieve significant improvement on the Mandarin test set due to the reason that a global embedding cannot be finely adaptive for all languages. As shown in Table 3, for ULSA, we additionally explored the placement of the adaptive decoder, and we found that the adaptive decoder only needs to be synchronized with the first ASR decoder layer to obtain optimal performance. In addition, we explored hierarchical ULSA to further improve the performance and found that optimal performance was achieved when L = 3, which indicates that hierarchical modeling can extract more representational non-semantic features than single-level modeling. Finally, by introducing adaptive weight assignment and SMTL, our H-ULSA achieves better performance than the monolingual model on the monolingual test set, and a relative improvement of 13.2 % over the mixed baseline on the code-switching test set.

Table 3: *Comparison of WERs(%) for different ULSA and H-ULSA configurations.*

Model	Mandarin	English	Code-switching
ULSA (N=5, D=256,P=1)	9.48	10.72	14.78
ULSA (N=10, D=256,P=1)	9.08	10.64	14.59
ULSA (N=15, D=256,P=1)	10.72	12.92	16.19
ULSA (N=5, D=256,P=1,2)	10.21	11.55	16.78
ULSA (N=5, D=256,P=1,2,4)	10.54	11.82	17.21
ULSA (N=5, D=256,P=1,2,4,6)	11.32	12.12	17.75
H-ULSA (N=10, D=256,P=1,L=2)	8.98	10.42	14.27
H-ULSA (N=10, D=256,P=1,L=3)	8.80	10.32	14.11
+ Adaptive Weight Assignment	8.75	10.21	14.01
+SMTL (α =0.2)	**8.70**	**10.18**	**13.93**
H-ULSA(N=10, D=256,P=1,L=4)	9.01	10.79	14.36

4.4. Comparison with other algorithms

The label synchronous embedding of ULSA belongs to local adaptive embedding, while the embedding of UGA belongs to global adaptive embedding. Therefore, as shown in Table 4, we combine H-ULSA and UGA to achieve the optimal performance in this paper. In addition, we reproduce some other mainstream multilingual and code-switching speech recognition algorithms, and we found that our proposed adaptive algorithms can achieve better performance and has the advantage of being plug-and-play for all attention-based ASR models.

Table 4: *Comparison of WERs(%) for our adaptive algorithms and other algorithms.*

Model	Mandarin	English	Code-switching
Mandarin-English Mixed Baseline	9.96	11.44	16.05
Language ID MTL [19]	11.23	13.44	18.12
Bi-encoder [15]	9.57	11.31	15.29
Bi-encoder-attention [16]	9.23	11.01	15.12
UGA (best)	9.25	10.75	14.77
H-ULSA (best)	8.70	10.18	13.93
UGA (best) + H-ULSAE (best)	8.69	10.01	13.87

5. Conclusions

We propose an unsupervised label synchronous adaptation algorithm that decomposes non-semantic information, such as speaker identity, accents and pronunciation characteristics of different languages from the Mel-spectrum synchronously during ASR decoding, and encodes them as adaptive embeddings to achieve multilingual and code-switching speech recognition adaptation. In addition, we propose a sequence-level multi-task learning algorithm, which enables the end-to-end speech recognition system to distinguish different languages and further improve the recognition performance.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61876160 and No.62001405).

7. References

[1] A. I. Rudnicky, "The design of voice-driven interfaces," CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, Tech. Rep., 1989.

[2] P. Auer, *Code-switching in conversation: Language, interaction and identity*. Routledge, 2013.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[6] 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.

[7] H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan, "Transformer-based online CTC/Attention end-to-end speech recognition architecture," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6084–6088.

[8] 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," *arXiv preprint arXiv:2005.08100*, 2020.

[9] S. Punjabi, H. Arsikere, Z. Raeesy, C. Chandak, N. Bhave, A. Bansal, M. Müller, S. Murillo, A. Rastrow, S. Garimella *et al.*, "Streaming end-to-end bilingual asr systems with joint language identification," *arXiv preprint arXiv:2007.03900*, 2020.

[10] 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.

[11] 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 *INTERSPEECH*, 2020, pp. 4741–4745.

[12] N.-Q. Pham, T.-N. Nguyen, S. Stueker, and A. Waibel, "Efficient weight factorization for multilingual speech recognition," *arXiv preprint arXiv:2105.03010*, 2021.

[13] 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," *arXiv preprint arXiv:2007.03001*, 2020.

[14] 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.

[15] 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.

[16] X. Zhou, E. Yılmaz, Y. Long, Y. Li, and H. Li, "Multi-encoder-decoder transformer for code-switching speech recognition," *arXiv preprint arXiv:2006.10414*, 2020.

[17] 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.

[18] 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.

[19] 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," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[20] C. Sobin and M. Alpert, "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy," *Journal of psycholinguistic research*, vol. 28, no. 4, pp. 347–365, 1999.

[21] A.-Y. Hung and Y. Cheng, "Sex differences in preattentive perception of emotional voices and acoustic attributes," *Neuroreport*, vol. 25, no. 7, pp. 464–469, 2014.

[22] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[24] 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.

[25] M. Cilimkovic, "Neural networks and back propagation algorithm," *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, vol. 15, no. 1, 2015.

[26] S. Zhang and M. Lei, "Acoustic modeling with dfsmn-ctc and joint ctc-ce learning." in *INTERSPEECH*, 2018, pp. 771–775.

[27] Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa, "Byte pair encoding: A text compression scheme that accelerates pattern matching," 1999.

[28] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "San-m: Memory equipped self-attention for end-to-end speech recognition," *arXiv preprint arXiv:2006.01713*, 2020.

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher, "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," *arXiv preprint arXiv:1810.13243*, 2018.