INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-pass Training and Cross-information Fusion for Low-resource End-to-end Accented Speech Recognition

Xuefei Wang[1], *Yanhua Long*[1*], *Yijie Li*[2], *Haoran Wei*[3]

[1]Key Innovation Group of Digital Humanities Resource and Research,
Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co., Ltd., Beijing, China
[3]Department of ECE, University of Texas at Dallas, Richardson, TX 75080, USA

xuefei_wang@163.com, yanhua@shnu.edu.cn, liyijie@unisound.com, haoran.wei@utdallas.edu

Abstract

Low-resource accented speech recognition is one of the important challenges faced by current ASR technology in practical applications. In this study, we propose a Conformer-based architecture, called Aformer, to leverage both the acoustic information from large non-accented and limited accented training data. Specifically, a general encoder and an accent encoder are designed in the Aformer to extract complementary acoustic information. Moreover, we propose to train the Aformer in a multi-pass manner, and investigate three cross-information fusion methods to effectively combine the information from both general and accent encoders. All experiments are conducted on both the accented English and Mandarin ASR tasks. Results show that our proposed methods outperform the strong Conformer baseline by relative 10.2% to 24.5% word/character error rate reduction on six in-domain and out-of-domain accented test sets.

Index Terms: speech recognition, accented ASR, multi-pass training, cross-information fusion

1. Introduction

In recent years, the performance of automatic speech recognition (ASR) on high-resource languages has benefited enormously from neural models [1, 2, 3]. The excellent performance makes these ASR systems widely being used in variety of commercial speech recognition products [4, 5, 6]. However, it is well-known that ASR system performance degrades significantly when encountering accent speech, especially when these accents are not existing in the ASR training dataset [7]. Accent is a special way of pronunciation, which is influenced by the region, the speaking style and the level of speaker's education [8], etc. Such as in remote regions or villages of south China, those accents are very different and seriously affect the pronunciation of Mandarin. All these accent variations lead to extremely expensive and time-consuming for transcribing heavy accented speech. Therefore, in the literature, there is a serious data sparsity problem of non-mainstream accented speech, building high performance ASR system for low-resource accented speech is very important and fundamental.

In the past few years, many research works have been explored for improving accented speech recognition. Such as in [9, 10, 11], different model adaptation methods were proposed to handle the non-native speech recognition. In [12], they proposed a multi-accent deep acoustic model with an accent-specific top layer and shared bottom hidden layers. While in [13], authors explored to finetune a particular subset of neural network layers with limited accent data. Fine-tuning a pre-trained model using limited accent speech is a straightforward way to handle the accent issue [14, 15, 16]. In most recent years, many works focus on extracting representative accent embedding to improve the accented ASR, such as in [17], they extracted the accent embedding from a well-trained accent classifier to perform the layer-wise adaptation of end-to-end ASR model; In [18], they just used one-hot embedding to build a multi-dialect ASR system; And in [19], they designed a TTS auxiliary model to convert accent information into a global style embedding for improving the accent robustness of E2E ASR model. All these previous works have been greatly boosted the performance of accented speech recognition, however, most works still requires balanced or large amount of accented speech training data. The works about how to leverage the available large amount of non-accent training data to improve the low-resource accented ASR system are still very limited.

In this paper, we aim to investigate using large amount of non-accented training data to boost the performance of low-resource accented speech recognition. Three contributions are explored: 1) a unified architecture with both general and accent encoders are proposed. These encoders are designed to integrate the general acoustic information that learnt from large available non-accented training data and the accent-dependent acoustic information that extracted from the limited low-resource accented speech; 2) a multi-pass training is explored to build relatively stable accent and general encoders; 3) Three cross-information fusion methods are exploited to effectively combine the information from both general and accent encoders. All methods are improvements of the state-of-the-art Conformer [20] system. Experiments are performed on both accented English and Mandarin ASR tasks. Results show that, compared with the Conformer baseline, on the in-domain Indian and Guangdong accent test sets, our proposed methods can achieve 10.6% relative word error rate (WER) reduction and 15.6% relative character error rate (CER) reduction, respectively. On the out-of-domain accented test sets, we also obtain 10.2% to 24.5% performance improvements on both the English and Mandarin accented ASR tasks.

2. Conformer-based ASR

In this paper, all our contributions are based on the convolution-augmented Transformer (Conformer) E2E ASR model that has been recently proposed in [20]. In order to improve the ability to capture locality of a sequence, Conformer inserts a convolution layer into the transformer block [21]. Because of its consistent excellent performance over a wide range of ASR tasks [22], Conformer has been taken as the state-of-the-art E2E ASR tech-

Yanhua Long is the corresponding author, she is also with the Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University. The work is supported by the National Natural Science Foundation of China (Grant No.62071302).

10.21437/Interspeech.2023-142

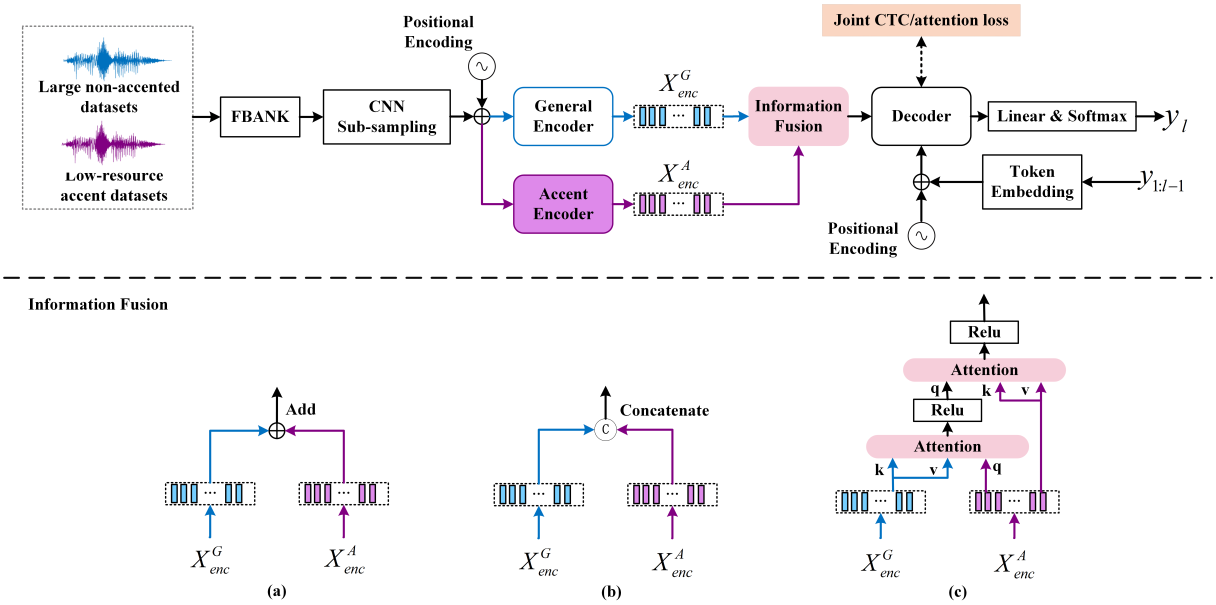

Figure 1: *System architecture of the proposed Aformer framework.*

nique, more and more Conformer variants have been explored in recent years [23]. The basic Conformer block consists mainly of four modules: the first feed-forward module (FFN1), the multi-head self-attention module (MHSA), the convolution module (Conv) and another feed-forward module (FFN2). Given an input sequence x, the output y of one Conformer block can be mathematically defined as:

$$\begin{aligned} x_{\mathrm{FFN}_1} &= x + \frac{1}{2}\mathrm{FFN}(x), \\ x_{\mathrm{MHSA}} &= x_{\mathrm{FFN}_1} + \mathrm{MHSA}(x_{\mathrm{FFN}_1}), \\ x_{\mathrm{Conv}} &= x_{\mathrm{MHSA}} + \mathrm{Conv}(x_{\mathrm{MHSA}}), \\ x_{\mathrm{FFN}_2} &= x_{\mathrm{Conv}} + \frac{1}{2}\mathrm{FFN}(x_{\mathrm{Conv}}), \\ y &= \mathrm{Layernorm}(x_{\mathrm{FFN}_2}) \end{aligned} \tag{1}$$

More details of the Conformer E2E ASR model can be referred to [20]. During Conformer training, the following multi-task criteria using interpolation of the CTC and attention cost is adopted [24],

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc} \tag{2}$$

Where the task weight λ is empirically set to 0.3 and fixed throughout the experiments of this paper.

3. Proposed Methods

In this section, we introduce the details of our proposed Aformer, which is specially designed for improving the performance of low-resource end-to-end accented speech recognition. The whole model architecture is presented in Section 3.1, followed by the description of multi-pass training in Section 3.2, and the cross-information fusion methods are described in Section 3.3.

3.1. Architecture

The whole architecture of our proposed Aformer is illustrated in Fig.1. Compared with the standard Conformer in Section 2, the only difference is the purple highlighted two blocks: the accent encoder and the information fusion. All other blocks are exactly the same as Conformer in Section 2, including the FBANK extraction, CNN sub-sampling, positional encoding, the general encoder, and the decoder, etc.

Assuming the outputs of two encoders are X^G_{enc} and X^A_{enc} for the general and accent encoder, respectively. The information fusion block is designed to effectively combine the different acoustic representations as

$$X^F_{enc} = \mathrm{Fusion}(X^G_{enc}, X^A_{enc}) \tag{3}$$

where the fusion methods are demonstrated in subfigure (a), (b) and (c) of Fig.1, and they will be presented in detail in Section 3.3. Finally, together with the token embedding, the combined acoustic embedding X^F_{enc} is then fed into the decoder module to get the decoding outputs.

The principle behind the design of Aformer is that, we aim to leverage the acoustic information in large amount of open-source non-accent training speech to boost the low-resource accented speech recognition. Therefore, in Aformer, we keep using the original Conformer encoder in [20] as the general encoder to extract the general accent-invariant acoustic context embedding, while adding a much simpler network as the additional accent encoder to learn the accent-dependent acoustic attributes from extremely low-resource accented data. After a proper embedding fusion, we expect these two different acoustic representatives will be well integrated to improve the final accented E2E ASR system.

3.2. Multi-pass Training

The proposal of multi-pass training of Aformer is motivated by the heavy data imbalance between non-accented and accented ASR training speech. Under the low-resource accented ASR condition, the labeled accented speech is normally a few minutes to tens of hours, these limited data is not enough to well train a separate accent encoder. Therefore, the multi-pass training is proposed, it aims to not only provide an good initialization of the general and accent encoders, but also to enable the

Aformer concentrate on well training the specific encoder at different training pass. The detail of Aformer multi-pass training can be divided into three passes as follows:

- `Pre-training`: Use the available large amount of non-accented training speech to train Aformer as the initialization model (A1).
- `Accent encoder adaptation`: Freeze the parameters of the general encoder, use the provided low-resource accent training data to only adapt the accent encoder of A1 to learn the specific accent acoustic characteristic. This adapted model is termed as A2.
- `Re-training`: Pooling both the non-accented and accent training data together to re-train the Aformer (A3) based on the adapted A2.

By performing the above three-pass model training, both the information in the limited accented and large amount of non-accented training data are effectively exploited, which helps the final Aformer A3 has the ability to well capture both the general acoustic context and accent-dependent acoustic information for improving the low-resource accented E2E ASR system performance.

3.3. Cross-information Fusion

In the proposed Aformer, how to combine the outputs of its two different encoders is very important. In this study, we investigate three ways to perform the information fusion for validating the complementarity between two encoders' outputs. Details are shown in Fig.1 (a) to (c). Specifically, (a) and (b) are two traditional information fusion methods that defined in Eq.(4) and (5): the linear addition and the concatenation.

$$X_{enc}^{F} = \mathrm{Add}(X_{enc}^{G}, X_{enc}^{A}) \tag{4}$$

$$X_{enc}^{F} = \mathrm{Concat}(X_{enc}^{G}, X_{enc}^{A}) \tag{5}$$

Different from (a) and (b), in Fig.1(c), we propose a two-layer cross-attention structure to combine the embeddings of the general and accent encoders as,

$$\begin{aligned} X_{enc}^{M} &= \mathrm{Relu}(\mathrm{Softmax}(\frac{Q_{enc}^{A}(K_{enc}^{G})^{T}}{\sqrt{d_{att}}})V_{enc}^{G}), \\ X_{enc}^{F} &= \mathrm{Relu}(\mathrm{Softmax}(\frac{Q_{enc}^{M}(K_{enc}^{A})^{T}}{\sqrt{d_{att}}})V_{enc}^{A}), \\ Q_{enc}^{i} &= W_{i}^{Q}X_{enc}^{i}, K_{enc}^{i} = W_{i}^{K}X_{enc}^{i}, V_{enc}^{i} = W_{i}^{V}X_{enc}^{i} \end{aligned} \tag{6}$$

Where $X_{enc}^{i} \in \{X_{enc}^{G}, X_{enc}^{A}, X_{enc}^{M}\}$, d_{att} is the attention dimension. Q_{enc}^{i}, K_{enc}^{i} and V_{enc}^{i} are the query, key and value linear projections on the encoder output X_{enc}^{G}, X_{enc}^{A} and the output of the first layer cross-attention X_{enc}^{M}, respectively. The projections are parameter matrices W_{i}^{Q}, W_{i}^{K} and W_{i}^{V}. This cross-attention structure is used to highlight the complementary information between the general and accent-dependent acoustic embeddings, making the combined encoder output X_{enc}^{F} more robust and representative.

4. Experiments

4.1. Datasets

In this study, two language datasets are used to examine the effectiveness of our proposed methods, one is English and the other is Mandarin. To simulate low-resource accented ASR tasks, we use the "train-clean-360" [25] as the non-accented English training data, and randomly select 20 hours (hrs) English data with Indian (IN) accent from the publicly available Common Voice [26] as the limited accented training data. For Mandarin task, the open-source Aishell [27] is taken as the non-accented training data, while the accented Mandarin datasets are all collected from a live speech service system of Unisound corporation in China (https://www.unisound.com/). 20 hrs Mandarin with Guangdong (GD) accent is selected as the training data. Six accent test sets are used for system evaluation, including the IN and GD in-domain test sets, and four out-of-domain test sets with England (EN), Canada (CA), Sichuan (SC) and Hunan (HN) accents. More details are shown in Table 1.

Table 1: *Details of both non-accented and accented English and Mandarin datasets.*

English (#hrs)			Mandarin (#hrs)		
	Train	Test		Train	Test
LibriSpeech	360	5.4	Aishell	164	10
Indian	20	3.8	Guangdong	20	2.0
Canada	-	2.2	Hunan	-	2.0
England	-	1.9	Sichuan	-	2.0

4.2. Experimental Setups

We use 80-dimensional log Mel-filter bank (FBANK) plus one-dimensional pitch as input acoustic feature. They are computed using 25ms windows with a 10ms hop. The utterance-level cepstral mean and variance normalization (CMVN) computed using the training set is applied on the FBANK for feature normalization. All our experiments are implemented with the ESPnet [28] end-to-end speech processing toolkit. No data augmentation and no extra language model are applied.

For the acoustic encoder of Aformer, the input features are first sub-sampled by the convolution subsampling module which contains two 2-D convolutional layers with stride 2. The general encoder of Aformer contains 12 conformer encoder layers with 2048-dimension feed-forward and 256-dimension attention with 4 self-attention heads. Two structures of accent encoder are investigated, one is with 4 transformer encoder layers, the other is with 2 layers 256-dimensional LSTM. All models are trained with the Adam optimizer [29]. The warmup learning schedule [30] is used for our first 25K training iterations, and both label smoothing [31] weight and dropout is set to 0.1 for model regularization.

In English tasks, 3000 byte-pair-encoder (BPE) [32] units generated by SentencePiece [33] are taken as the decoder outputs. In Mandarin tasks, we use 4231 Chinese characters as the Aformer modeling units. The CTC weight is set to 0.3 during both model training and inference. The word error rates (WER) and character error rate (CER) are used for evaluating the ASR performance for English and Mandarin tasks, respectively.

4.3. Results and Discussions

4.3.1. Results on English Accented-ASR

Table 2 presents the performance of low-resource English accented ASR systems. E1 is the Conformer system trained only on 360 hrs LibriSpeech data, and E2 is finetuned from E1 using 20 hrs Indian accent speech. E3 to E6 are all Aformer systems that trained using the combined LibriSpeech and Indian training data with the proposed multi-pass training. However, in E3,

Table 2: *WER(%) for low-resource English accented ASR task.*

ID	Train	System	Fusion	Test Set		
				IN	EN	CA
E1	LibriSpeech	Conformer	-	78.1	36.4	20.7
E2	Indian	+Finetune	-	**36.7**	**40.6**	**23.2**
E3	LibriSpeech +Indian	Aformer	Add-LSTM	34.0	40.2	22.7
E4			Add	33.1	32.5	17.6
E5			Concat	33.2	32.5	17.9
E6			Cross-attention	**32.8**	**32.2**	**17.5**

the accent encoder is 2-layers LSTM, while in E4 to E6, their accent encoder are the 4 transformer layers but with different encoder output fusion methods.

Comparing the results of E1 system, it's clear that the Conformer performance is significantly degraded when it meets accented test speech, especially when the heavy accent deviates far from the training data. E.g. on Indian test set, the WER reaches to 78.1%, while on test sets with England and Canada accents, the WER numbers are greatly reduced to 36.4% and 20.7%. This is because compared with IN accent, the EN and CA are more close to the speaking style or acoustics of the data in LibriSpeech. When comparing E1 and E2, we see 53% WER reduction on IN test set, even there are slight performance degradation on the out-of-domain EN and CA test sets. It indicates that the traditional finetuning is very effective to obtain good results on low-resource accented ASR task. Thus, we take E2 as our baseline.

E3 and E4 are used to compare different accent encoder structures. We see that, E4 is much better than E3, this may due to the acoustic modeling ability of transformer is much stronger than LSTM, and 4-layers transformer has more parameters than the 2-layers LSTM. Difference between E4 to E6 are only the three information fusion methods to combine the embeddings of general and accent encoders. It's clear that, there is no big performance gap between using linear addition (Add) and concatenation (Concat), and the proposed cross-attention fusion achieves the best results. In conclusion, the proposed Aformer with all three fusion methods can obtain much better results than baseline system of E2, both on the in-domain and out-of-domain accented test sets. And compared with E2, the best system E6 achieves 10.6%, 20.6% and 24.5% relative WER reduction on IN, EN and CA accented test set, respectively. It means that, our proposed Aformer is more effective and has stronger generalization ability to out-of-domain accented speech than the conventional finetuning.

4.3.2. Results on Mandarin Accented-ASR

Table 3: *CER(%) on Mandarin accented ASR task.*

ID	Train	System	Fusion	Test Set		
				GD	HN	SC
M1	Aishell	Conformer	-	53.0	54.8	51.6
M2	Guangdong	+Finetune	-	**29.4**	**38.7**	**40.1**
M3	Aishell +Guangdong	Aformer	Add-LSTM	26.6	35.1	37.5
M4			Add	25.7	34.7	35.6
M5			Concat	25.3	34.7	36.2
M6			Cross-attention	**24.8**	**34.0**	**36.0**

Table 3 shows the performance on the low-resource Mandarin accented ASR task. Similar as the system E1 to E6 in Table 2, M1 and M2 are taken as the baseline systems, M3 to M6 are the proposed Aformer with different accent encoder and information fusion methods. Different from the observation in E1 results, the CERs of M1 on the GD, HN and SC are almost at the same level, it tells us that, all these three accents deviates far from the non-accented Aishell training data. When M1 is finetuned by 20 hrs Guangdong accent data, the performance on all three accented test sets are greatly improved, even the improvement gain on in-domain GD is much larger than the ones on other two out-of-domain test sets.

The findings in M3 to M6 are consistent with the ones that observed in E3 to E6 from Table 2, such as, the best results are also achieved from the Aformer (M6) using transformer accent encoder and cross-attention method for information fusion. Compared with M2, system M6 obtains a relative CER reduction of 15.6%, 12.1% and 10.2% on the GD, HN and SC accented test sets, respectively.

4.3.3. Ablation of Multi-pass Training

Fig. 2 shows our ablation experimental results to verify the effectiveness of the proposed multi-pass training. All these experiments are performed on E4 and M4. In Fig. 2, four bars means using the three different training stages in multi-pass training that described in Section 3.2. It is worth mentioning that (a1) is the "Pre-training" described in Section 3.2 with only non-accented data, and (a2) is the pre-training with mixing all the accented and non-accented data together to train the Aformer structure. (b) and (c) are exactly the same as described in Section 3.2. The white four bars are WER% on the Indian test set of E4, while the pink four ones are CER% on the Guangdong test set of M4. It's clear that, the re-trained Aformer achieves the best results on both English and Mandarin accented ASR tasks. It means that, the multi-pass training is effective than only using pooling data to train the Aformer, and the Aformer that finetuned using the accented training data.

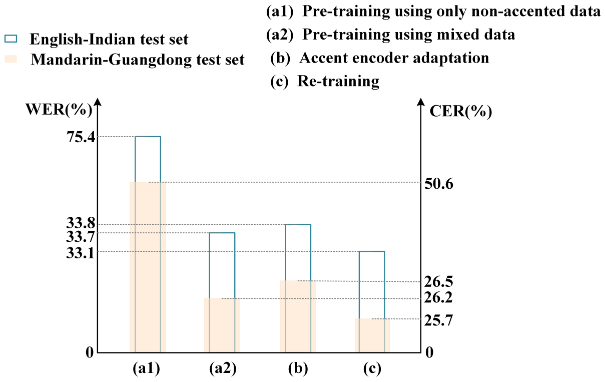

Figure 2: *Ablation study of multi-pass training method.*

5. Conclusion

In this study, we explore the approach of leveraging large amount of non-accented training data to enhance the low-resource accented end-to-end ASR system. Based on the standard Conformer ASR architecture, we propose an Aformer to capture both the general acoustic context and accent-dependent acoustic information. Moreover, a multi-pass training and different cross-information fusion methods are also investigated to further improve the Aformer. Results on both the low-resource accented English and Mandarin ASR tasks show that, the proposed methods outperform the finetuned Conformer significantly, either on the in-domain or the out-of-domain accented speech test sets.

6. References

[1] W. Chan, N. Jaitly, and Q. Le, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[2] F. Weninger, M. Gaudesi, and M. A. Haidar, "Conformer with dual-mode chunked attention for joint online and offline asr," in *Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 18–22.

[3] C. Chiu, T. N. Sainath, and Y. Wu, "State-of-the-art speech recognition with sequence-to-sequence models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4774–4778.

[4] A. Baevski, H. Zhou, and A. Mohamed, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, pp. 12 449–12 460.

[5] Y. Zhang, J. Qin, and D. S. Park, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv, preprint arXiv:2010.10504*, 2020.

[6] W. N. Hsu, B. Bolte, and H. Tsai, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] C. Huang, T. Chen, and E. Chang, "Accent issues in large vocabulary continuous speech recognition: Special double issue on chinese spoken language technology," *International Journal of Speech Technology*, vol. 7, pp. 141–153, 2004.

[8] D. Bahdanau, J. Chorowski, and D. Serdyuk, "End-to-end attention-based large vocabulary speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4945–4949.

[9] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2003, pp. 536–540.

[10] L. Arslan and J. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, pp. 28–40, 1999.

[11] L. Mayfield and A. Waibel, "Adaptation methods for non-native speech," in *Multilinguality in Spoken Language Processing*, 2002.

[12] Y. Huang, D. Yu, and C. Liu, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Conference of the International Speech Communication Association (Interspeech)*, 2014, pp. 2977–2981.

[13] J. Shor, D. Emanuel, and O. Lang, "Personalizing asr for dysarthric and accented speech with limited data," in *Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 784–788.

[14] T. Tan, Y. Lu, and R. Ma, "Aispeech-sjtu asr system for the accented english speech recognition challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6413–6417.

[15] Z. Jicheng, P. Yizhou, and P. Tung, "E2e-based multi-task learning approach to joint speech and accent recognition," in *Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 1519–1523.

[16] K. Deng and S. Cao, "Improving accent identification and accented speech recognition under a framework of self-supervised learning," in *Conference of the International Speech Communication Association (Interspeech)*, 2021, pp. 1504–1508.

[17] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[18] B. Li, T. N. Sainath, and K. C. Sim, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4749–4753.

[19] S. Li, B. Ouyang, and D. Liao, "End-to-end multi-accent speech recognition with unsupervised accent modelling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6418–6422.

[20] A. Gulati, J. Qin, and C. Chiu, "Conformer: Convolution-augmented transformer for speech recognition," in *Conference of the International Speech Communication Association (Interspeech)*, 2020, pp. 5036–5040.

[21] P. Guo, F. Boyer, and X. Chang, "Recent developments on espnet toolkit boosted by conformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5874–5878.

[22] M. Zeineldeen, J. Xu, and C. Lüscher, "Conformer-based hybrid asr system for switchboard dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7437–7441.

[23] J. Deng, X. Xie, and T. Wang, "Confidence score based conformer speaker adaptation for speech recognition," in *Conference of the International Speech Communication Association (Interspeech)*, 2022, pp. 2623–2627.

[24] L. T. Xie Xurong, Liu Xunying, "Blhuc: Bayesian learning of hidden unit contributions for deep neural network speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5711–5715.

[25] V. Panayotov, G. Chen, and D. Povey, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] R. Ardila, M. Branson, and K. Davis, "Common voice: A massively-multilingual speech corpus," in *Conference on Language Resources and Evaluation (LREC)*, 2020, pp. 4218–4222.

[27] H. Bu, J. Du, and X. Na, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, 2017, pp. 1–5.

[28] S. Watanabe, T. Hori, and S. Karita, "Espnet: End-to-end speech processing toolkit," in *Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 2207–2211.

[29] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[30] A. D. Gotmare, N. S. Keskar, and C. Xiong, "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," in *International conference on learning representations*, 2019.

[31] C. Szegedy, V. Vanhoucke, and S. Ioffe, "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[32] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *The Association for Computational Linguistics (ACL)*, 2015, pp. 1715–1725.

[33] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 66–71.