

Electronics Letters

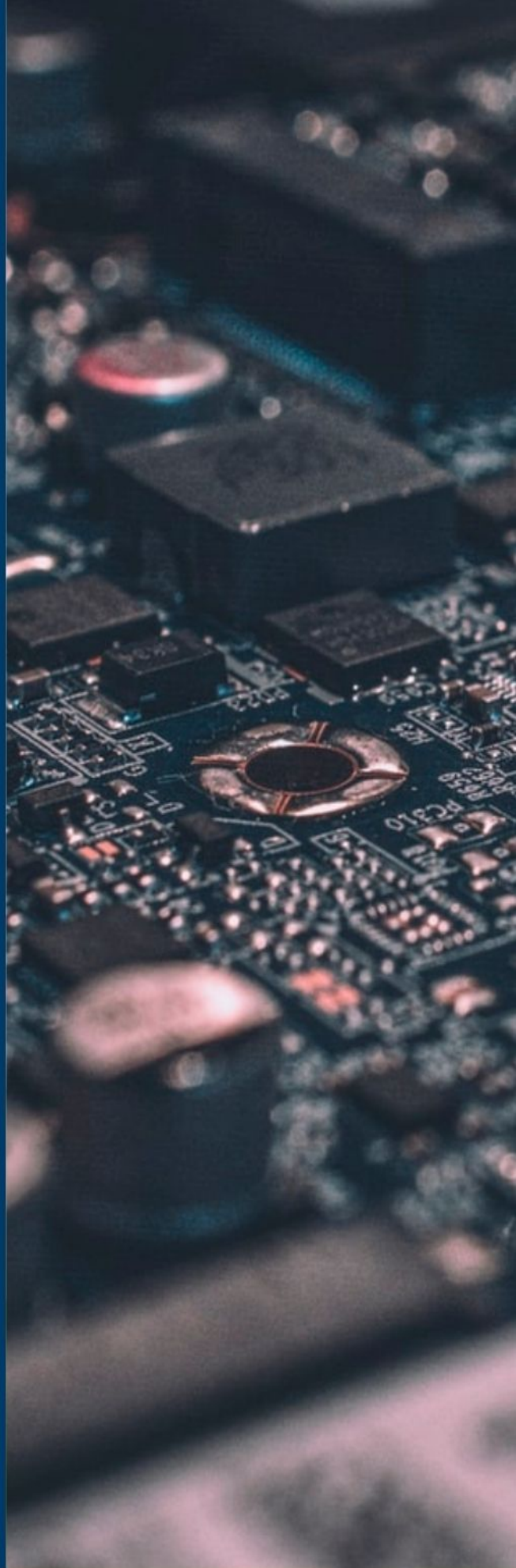
Special issue Call for Papers

**Be Seen. Be Cited.
Submit your work to a new
IET special issue**

Connect with researchers and experts in your field and share knowledge.


Be part of the latest research trends, faster.

[Read more](#)



The Institution of
Engineering and Technology

Wav2vec-MoE: An unsupervised pre-training and adaptation method for multi-accent ASR

Yuqin Lin,  Shiliang Zhang, Zhifu Gao, Longbiao Wang, Yanbing Yang, and Jianwu Dang
 Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
 E-mail: linyuqin@tju.edu.cn

In real life, either the subjective factors of speakers or the objective environment degrades the performance of automatic speech recognition (ASR). This study focuses on one of the subjective factors, accented speech, and attempts to build a multi-accent ASR system to solve the degradation caused by different accents, one of whose characteristic is the low resource. To deal with the challenge of the low-resource data and the different speech styles, a wav2vec-MoE (mixture of experts) is proposed to adapt the wav2vec 2.0 for multi-accent ASR. In the wav2vec-MoE, a domain MoE is developed by introducing pseudo-domain information in the pre-training stage, where the domain denotes a collection of speech varied by the same influence factors. The MoE is trained with two strategies according to the proposed domain mismatch assessment between unlabeled speech and target speech, without requiring any explicit domain information. Experiments show that the wav2vec-MoE achieves a 14.69% relative word error rate reduction (WERR) on the AESRC2020 accent dataset and an 8.79% relative WERR on the Common Voice English dataset.

Introduction: As one of the essential technologies of human-computer interaction, end-to-end automatic speech recognition (ASR) has achieved remarkable performance in recent research [1–3]. While in real life, either the subjective factors of speakers or the objective environment degrade the performance of ASR [4–8]. This study focuses on one of the subjective factors—accented speech. To deal with accented speech, one popular solution is to build an ASR for each accented speech [9–11]. This solution leads to a high-performance ASR for specific accents but requires huge labeled accented speech for training. Multi-accent ASR alleviates this problem by building a system to recognize various accented speech. However, the low resource of labeled data and the variety of speech styles between accents challenge the multi-accent ASR. Most of the approaches in recent advances work on the model training or fine-tuning stage, while the improvement is limited due to the low-resource labeled data [12–15].

Recently, unsupervised pre-training techniques (UPTs) are proven to be effective for low-resource tasks [16–19]. It learns speech representations as the initialization of the target model from a large amount of unlabeled speech, and then, the model is fine-tuned with labeled data for the downstream tasks. However, UPTs try to obtain a universal representation, which may lead to the lacking of some distinguishing speech features that are varied by subjective factors. Hence, directly using UPTs on multi-accent ASR could not deal with domain mismatch problem caused by the different speech styles.

This insight motivates us to enhance the recent popular UPT, wav2vec 2.0 [16], to address the two challenges of multi-accent ASR. We propose a wav2vec-MoE that introduces pseudo-domain information provided by a domain identification (DID) model to solve domain mismatch under an unsupervised framework. The domain refers to a collection of speech varied by the same influence factors, including but not limited to international accents. We use the DID instead of accent identification (AID) because it is difficult to extract the distinctive features of the accent from the input speech, especially when the data is extremely unbalanced on accent [12, 13]. In the wav2vec-MoE, we develop a domain MoE, which is a mixture of experts guided by pseudo-domain prior knowledge. We incorporate the domain MoE into the wav2vec 2.0 framework to make the speech representation robust to multiple domains. We propose a domain mismatch assessment (DMA) to evaluate the quality of pseudo-domain information and then train the domain MoE with two different strategies according to the judgment of DMA. The experiments were conducted on the AESRC2020 accent dataset [20] and the Common Voice English dataset [21] to evaluate the proposed method.

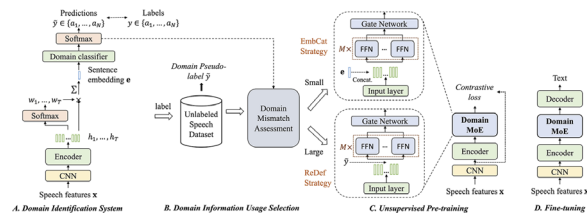


Fig. 1 The schematic diagram of the wav2vec-MoE.

Related work: Research on multi-accent ASR focus on dealing with the different speech style challenge and the low-resource challenge. One popular method is model adaptation, which makes the ASR system of standard speech adapt to accented speech. For example, integrating accent-related features from an accent identification (AID) model into acoustic features for adaptation [12–15]. However, these methods did not consider the quality of AID, and low-quality AID will bring noise to ASR, thus may affect the performance of ASR. Recently, the acoustic models based on the mixture of experts (MoE) achieved competitive performance. For example, You et al. [22, 23] introduced MoE in accented ASR and proposed speechMoE [22] and speechMoE2 [23], in which each expert is activated by route decisions. The models have large capacities but have large model sizes. Jain [15] et al. proposed a MixNet-based architecture to compensate for phonetic and accent variabilities by using MoE. Gong [8] et al. proposed a layer-wise adapter in which each expert learns the features of the inputs and finally merges the features from others. These methods demonstrate the potential of MoE in multi-accent ASR, but their improvement is limited by limited annotated speech data.

Wav2vec-MoE: This section describes the proposed wav2vec-MoE in detail. Figure 1 shows the schematic diagram of the method.

Domain identification training: The DID model is to identify the domain where the speech comes from. The domain refers to a collection of speech varied by the same influence factors, for example, international accents, and regional accents. The model consists of an encoder and a linear layer for the domain classifier. To extract sentence-level domain embeddings, we apply attention pooling to the outputs of the encoder of the DID model. Given the input speech features $\mathbf{x} = \{x_1, \dots, x_L\}$ with length L , corresponding domain label $y \in \{a_1, \dots, a_N\}$ with N domains, the outputs of the encoder refer to the latent speech features $\mathbf{h} = \{h_1, \dots, h_L\}$. The sentence-level domain embeddings \mathbf{e} pass through a linear layer and a softmax layer to predict the maximum probability of the domain where the speech is from. The DID model is optimized by minimizing the cross-entropy loss.

Domain information usage selection: Previous work shown that high-quality domain features help the ASR. It performs well when the quality of DID is high enough to help the ASR. In a real situation, however, low-quality domain features may harm the performance of ASR. The quality depends on the mismatch between the source data and the target data. The greater the mismatch, the worse the quality. Hence, we need to assess the mismatch between the unlabeled speech and the target speech.

We compute the proportion of the uncertain label as the measure of mismatch. A high proportion indicates a large mismatch. Since a large amount of data, we randomly sampled a subset from the unlabeled dataset for assessment. Denote α as the confidential coefficient that the sentence with speech features \mathbf{x} is identified as being from the domain y by the DID model. The predicted domain \hat{y} can be expressed as

$$\hat{y} = \begin{cases} a_i, i = \text{argmax}(\mathbf{I}), & \text{if } \max(\mathbf{I}) > \alpha \\ \text{uncertain}, & \text{else,} \end{cases} \quad (1)$$

where \mathbf{I} is the probability that the speech comes from each of the domains, and $\text{argmax}(\cdot)$ is the function that outputs the index of the highest-probable domain. In this paper, we set $\alpha = (N + 1)/2N$, which varies with the difficulty of the DID task. After labeling the data, we compute the proportion p of the *uncertain* label in the sampled subset to assess the mismatch between the unlabeled data and the target data. We

regard $p > \beta$ as the case of a large mismatch, and vice versa is the case of a small mismatch. β is the experience value. We carried out experiments with different β values to optimize the hyper-parameter and found that when the unlabeled speech is large (over 1k), setting β to 0.5–0.6 gives the best and most stable ASR performance. When the unlabeled is small, set β to 0.1–0.2 is the best choice. We adopted $\beta = 0.5$ for large unlabeled data and $\beta = 0.2$ for small unlabeled data.

Unsupervised pre-training: To make speech representations based on wav2vec2.0 [16] robust to multiple target domains, we develop a domain MoE and incorporate it into wav2vec 2.0 framework, as shown in Figure 1, part C. The domain MoE consists of a mixture of expert networks using feedforward networks (FFNs) and a gate network using a recurrent neural network (RNN). Each expert maps the encoded features to different subspaces with pseudo-domain information and then aggregates them through the gate mechanism. It improves the ability of the model to extract representation. The gate mechanism is realized by the gate network followed by a softmax. To make a fair comparison of the two training strategies introduced below, this paper set the number of expert as $N + 1$. Finally, the domain-robust speech representations $\mathbf{r} = \{r_1, \dots, r_L\}$ are derived from a weighted sum of the weights δ and the experts outputs $\mathbf{E}_i(\mathbf{c})$.

Since there is a mismatch between the source (unlabeled) speech and the target (accented) speech in real life, we propose two training strategies (EmbCat, ReDef) for unsupervised pre-training according to the quality assessments describe in Section 1.

EmbCat strategy: When the mismatch is small, we use domain embeddings as one of the inputs in domain MoE because the embeddings contain richer domain information than pseudo labels. Specifically, the sentence-level domain embeddings \mathbf{e} are extracted from the DID model and concatenated with the outputs of the encoder. After a linear layer, the concatenated features are fed into the domain MoE. Every sample updates all the experts.

ReDef strategy: In the case that mismatch is large, we use domain information implicitly, inspired by [24]. The domain MoE is divided into one common expert and specific experts. During training, every expert provides the outputs, while only the common expert and the corresponding specific expert are updated.

Fine-tuning: The fine-tuning stage is related to the training strategy in the pre-training. The model pre-trained with the EmbCat strategy is fine-tuned with the EmbCat strategy. For the model pre-trained with the ReDef strategy, the experts are updated according to the pseudo-label provided by the DID model. The features extracted by each expert are from the subspace of the pseudo-domain. The subspaces of the pseudo-domain are different from the target domain. If the model is fine-tuned with the ReDef strategy same as in the pre-training stage, the mapping space of the experts will be redefined; this will destroy the pre-trained feature structures. Therefore, we directly fine-tuned the model pre-trained with the ReDef strategy.

Experiments and discussions

Datasets: All experiments are conducted on the AESRC2020 accent dataset [20], the Mozilla Common Voice English dataset (version 7) [21], Librispeech [25] corpus (1k hours of reading English speech), and a private general English dataset (7k hours of unlabeled speech). The AESRC2020 accent dataset contains eight accented English in England (UK), America (US), China (CHN), Japan (JPN), Russia (RU), India (IND), Portugal (PT), and Korea (KR). Each accent has 20 h of speech for training. The validation set contains eight accents, while the test set contains another two out-set accented English in Canada (CAN) and Spain (ES). The Common Voice English dataset contains sixteen accented English in Africa (AF), Australia (AU), Bermuda (BU), Canada (CAN), England (UK), Hongkong (HK), India (IND), Ireland (IR), Malaysia (MY), New Zealand (NZ), Philippines (PH), Scotland (SC), Singapore (SG), South Atlantic (SA). The duration of the accented speech is unbalanced, ranging from about 1 h to 500 h. We select the eight lowest resource accented speech data as an out-set test set. The other 8 accented speech data are divided into the training set, validation

Table 1. The DID accuracy and the mismatch assessment (MA) on the Common Voice Dataset and the AESRC2020 Dataset.

Dataset	DID accuracy (%)		MA	
	Validation	Test	Validation	Unlabeled speech
Common Voice	81.36	64.96	0.29*	0.62*
AESRC2020	82.09	77.16	0.09	0.48

* indicates the quality of domain embeddings from the DID system is not good enough.

set, and in-set test set with the ratio of 8:1:1. The training, validation, and in-set test set lasted 431.5, 52.7, and 53.8 h, respectively. The private general English dataset contains 7k hours of unlabeled speech data with a 16 kHz sampling rate. Recording files are from the Ali Industrial application, mainly obtained through user-authorized data collection, including different fields, such as including live broadcasting, input software, customer service, finance, and so on.

Experimental setups

Domain identification system: The inputs of the domain identification (DID) model are 80-dimension log Mel-filterbank domain features, computed with a 25 ms sliding window shifted by 10 ms each time step. We use the memory-equipped self-attention model (SAN-M) [26] with 45 encoder blocks and 12 decoder blocks. Each layer has a dimension of 320 and 6 heads. For the AESRC2020 dataset, we train the DID model for identifying accents between countries, while for the Common Voice dataset, we train the DID model for identifying accents between continents because of the extremely unbalanced data on accents between countries. Due to the limited speech data, we use a pre-trained ASR system to initialize the DID model. Table 1 shows the DID accuracy and mismatch assessment values on the Common Voice dataset and the AESRC2020 dataset.

Automatic speech recognition: We use the SAN-M model as the backbone, which has 12 encoder blocks and 6 decoder blocks. The settings of feature extraction are the same as the DID system. The attentions in the baseline/baseline-Large/wav2vec-MoE have dimensions of 256/320/256 with 4 heads, respectively. There are 9 experts for AESRC dataset and 5 experts for Common Voice dataset. The hidden dimensions of each expert network are set to 1280/2048 for AESRC2020/CommonVoice dataset. During training, we adopt an Adam optimizer [27] with a warm-up learning rate [28]. For pre-training, the models are trained in 350k iterations with a batch size of 320k tokens. The learning rate is set to 4.0. For fine-tuning, the outputs are the 1024 English Byte Pair encoding (BPE) [29] subword units. The models for the AESRC2020/CommonVoice dataset are trained 350k/100k iteration. The learning rate is set to 2.0/3.0 on the AESRC2020/CommonVoice dataset. Besides, SpecAugment [30] is applied for data augmentation. For evaluation, all models are decoded using a beam search with a beam width of 10.

Results: Table 2 compares the performance of the proposed wav2vec-MoE on the Common Voice dataset and the AESRC2020 dataset. \mathcal{P}_d and \mathcal{F}_d denote pre-training and fine-tuning the proposed domain MoE model, respectively. $\mathcal{P}_d + \mathcal{F}_d$ is the proposed wav2vec-MoE. From the table, the wav2vec-MoE gains the best performance comparing other methods. Compared with the baseline on the AESRC2020 dataset, the wav2vec-MoE achieves about 13%~15% relative word error rate reduction (WERR) on two test sets. Compared with the baseline on the Common Voice dataset, the wav2vec-MoE achieves about 8% WERR on two test sets. Comparing \mathcal{S}_2 and \mathcal{S}_4 , the amount of unlabeled speech is larger, and the domain MoE improves the ASR performance more obviously. The wav2vec-MoE is more effective than the wav2vec 2.0 on multi-accent ASR.

Ablation study

The effectiveness of the domain mismatch assessment: Table 3 compares the performance of ASR with/without the domain mismatch

Table 2. The performance (WER%) of the proposed wav2vec-MoE on the Common Voice dataset and the AESRC2020 dataset. ‘ \mathcal{P}_d ’/‘ \mathcal{F}_d ’ Denotes Pre-training/Fine-tuning the proposed Domain MoE Model. $\mathcal{P}_d + \mathcal{F}_d$ is the Proposed Wav2vec-MoE.

System	Method	Pretraining data	AESRC2020				Common voice			
			Param.	In-set test	Out-set test	Avg.	Param.	In-set test	Out-set test	Avg.
S_0	Baseline	-	25.7M	12.16	11.60	12.05	25.7M	19.52	22.56	20.58
S_1	Baseline-Large	-	33.6M	11.50	11.01	11.41	33.6M	18.99	21.93	20.01
S_2	Wav2vec 2.0	1kh	33.6M	11.08	10.11	10.89	33.6M	18.82	21.41	19.92
S_3	Wav2vec 2.0 + \mathcal{F}_d	1kh	32.8M	10.55	10.71	10.58	33.0M	18.39	21.28	19.39
S_4	Wav2vec 2.0	7kh	33.6M	10.88	10.28	10.76	33.6M	18.50	21.28	19.45
S_5	Wav2vec 2.0 + \mathcal{F}_d	7kh	32.8M	10.51	10.46	10.72	32.8M	18.28	20.89	19.19
S_6	$\mathcal{P}_d + \mathcal{F}_d$	7kh	32.8M	10.33	10.10	10.28	33.0M	17.81	20.57	18.77

Table 3. The effectiveness of Domain Mismatch Assessment (DMA) on the Test Set of Common Voice.

ID	Method	Param.	In-set	Out-set	Avg.
1	Wav2vec 2.0	33.6M	18.50	21.28	19.45
2	Wav2vec-MoE	33.0M	17.81	20.57	18.77
3	-,DMA (FT+EmbCat)	33.0M	18.22	20.99	19.18
4	-,DMA (FT+ReDef)	33.0M	18.55	21.23	19.48
5	-,DMA (PT)	33.0M	18.45	21.21	19.41

Table 4. The impact of domain information richness on the AESRC2020.

ID	Method	Param.	In-set	Out-set	Avg.
1	Wav2vec 2.0	33.6M	10.88	10.28	10.76
2	PT+EmbCat, FT+EmbCat	32.8M	10.33	10.10	10.28
3	PT+ReDef, FT+ReDef	32.6M	10.66	10.39	10.61
4	PT+ReDef, FT+Direct	32.6M	10.79	10.23	10.68

assessment (DMA) on the Common Voice dataset. The baseline is wav2vec 2.0 (ID-1). Compared with the wav2vec-MoE (ID-2), we first remove DMA in the fine-tuning stage, that is, we fine-tune the model with the EmbCat strategy (ID-3) or ReDef strategy (ID-4). Fine-tuning with the EmbCat strategy degrades the performance because the quality of domain embeddings is poor. As expected, fine-tuning with the ReDef harms the performance because it will destroy the feature structure as described in Section 1. Finally, we further remove DMA in the pre-training stage (ID-5), that is, we pre-train and fine-tune the model with the EmbCat strategy. The results suggest the necessity of DMA.

The impact of domain information richness: Table 4 explores the impact of domain information richness on the performance of ASR. ‘PT+x’/‘FT+x’ denotes we use the x strategy in the pre-training/fine-tuning stage. Among the methods in the table, the richness of the domain information decreases from ID-2 to ID-4. The richer domain information contributes to the ASR when the quality of embedding is good.

Analysis: Figure 2 demonstrates the improvements of the wav2vec 2.0 and the wav2vec-MoE on each accent. One can see that wav2vec 2.0 degrades the ASR performance of some accents, which indicates that wav2vec 2.0 is not robust to the multi-domain scenario. Our wav2vec-MoE improves the ASR performance on every accent. The possible reason for the effectiveness of the wav2vec-MoE is that the proposed method alleviates the domain sensitivities by introducing pseudo-domain information, which reflects the mismatch between unlabeled speech and target speech.

Conclusion: This paper proposed a wav2vec-MoE, which is an unsupervised pre-training and adaptation method for multi-accent speech

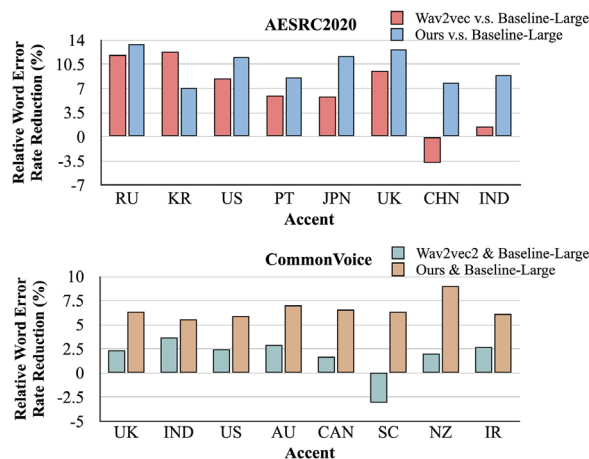


Fig. 2 Relative word error rate reduction of the Wav2vec 2.0 and our Wav2vec-MoE compared to the baseline on each accent in the AESRC2020 and CommonVoice datasets.

recognition to solve the challenge of low-resource and the challenge of different speech styles. The method introduced pseudo-domain information to solve domain mismatch under the unsupervised framework. A domain mismatch assessment (DMA) was proposed to evaluate the quality of pseudo-domain information. The wav2vec-MoE was trained with two proposed strategies according to the judgment of the DMA, without requiring any explicit domain information. Experiments on the Common Voice English dataset and the AESRC2020 accent dataset validated the effectiveness of the proposed method in alleviating the two challenges.

Author contributions: Yuqin Lin: Conceptualization; Investigation; Methodology; Software; Writing – Original – draft; Shiliang Zhang: Methodology; Resources; Writing – Review – Editing; Zhifu Gao: Software; Writing – Review – Editing; Longbiao Wang: Resources; Supervision; Validation; Writing – Review – Editing; Yanbing Yang: Investigation; Jianwu Dang: Resources; Supervision; Validation; Writing – Review – Editing.

Acknowledgements: This work was supported in part by Alibaba Group through Alibaba Innovative Research Program and the National Natural Science Foundation of China under Grant 62176182. (corresponding authors: L. Wang and J. Dang)

Conflict of interest statement: The authors declare no conflict of interest.

Data availability statement: The AESRC2020 dataset that supports the findings of this study is openly available at https://github.com/R1ckShi/AESRC2020/blob/master/README_en.md at <https://doi.org/10.1109/icassp39728.2021.941338>, reference number [20]. The CommonVoice dataset that supports the findings of this study is openly available

at <https://commonvoice.mozilla.org/en/datasets> at <https://arxiv.org/pdf/1912.06670.pdf>, reference number [21].

© 2023 The Authors. *Electronics Letters* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made. Received: 20 March 2023 Accepted: 30 April 2023
doi: 10.1049/ell2.12823

References

- 1 Li, J., et al.: Recent advances in end-to-end automatic speech recognition. *APSIPA Trans. Signal Inf. Process.* **11**, 1–64 (2022)
- 2 Nassif, A.B., et al.: Speech recognition using deep neural networks: A systematic review. *Access* **7**, 19143–19165 (2019)
- 3 Miao, H., Cheng, G., Zhang, P.: Low-latency transformer model for streaming automatic speech recognition. *Electron. Lett.* **58**, 44–46 (2022)
- 4 Chen, M., et al.: Improving deep neural networks based multi-accent Mandarin speech recognition using i-vectors and accent-specific top layer. In: Proc. Interspeech, pp. 3620–3624 (2015)
- 5 Li, B., et al.: Multi-dialect speech recognition with a single sequence-to-sequence model. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4749–4753. IEEE, Piscataway (2018)
- 6 Jain, A., Upreti, M., Jyothi, P.: Improved accented speech recognition using accent embeddings and multi-task learning. In: Proceedings of Interspeech, pp. 2454–2458 (2018)
- 7 Jain, A., Singh, V.P., Rath, S.P.: A multi-accent acoustic model using mixture of xperts for speech recognition. In: Proceedings of Interspeech, pp. 779–783 (2019)
- 8 Gong, X., et al.: Layer-wise fast adaptation for end-to-end multi-accent speech recognition. In: Proceedings of Interspeech, pp. 1274–1278 (2021)
- 9 Zheng, Y., et al.: Accent detection and speech recognition for Shanghai-accented Mandarin. In: Proceedings of Interspeech, pp. 217–220 (2005)
- 10 Najafian, M., et al.: Unsupervised model selection for recognition of regional accented speech. In: Proceedings of Interspeech, pp. 2967–2971 (2014)
- 11 Winata, G.I., et al.: Learning fast adaptation on cross-accented speech recognition. In: Proceedings of Interspeech, pp. 1276–1280 (2020)
- 12 Li, S., et al.: End-to-end multi-accent speech recognition with unsupervised accent modelling. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6418–6422. IEEE, Piscataway (2021)
- 13 Gao, Q., et al.: An end-to-end speech accent recognition method based on hybrid CTC/attention transformer ASR. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7253–7257. IEEE, Piscataway (2021)
- 14 Turan, M.A.T., Vincent, E., Jouvét, D.: Achieving multi-accent ASR via unsupervised acoustic model adaptation. In: Proceedings of Interspeech, pp. 1286–1290 (2020)
- 15 Jain, A., Singh, V.P., Rath, S.P.: A multi-accent acoustic model using mixture of experts for speech recognition. In: Proceedings of Interspeech, pp. 779–783 (2019)
- 16 Baevski, A., et al.: Wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **33**, 12449–12460 (2020)
- 17 Yi, C., et al.: Applying wav2vec2.0 to speech recognition in various low-resource languages. CoRR abs/2012.12121 (2020). <https://arxiv.org/abs/2012.12121>
- 18 Schneider, S., et al.: Wav2vec: Unsupervised pre-training for speech recognition. In: Proceedings of Interspeech 2019, pp. 3465–3469 (2019)
- 19 Fan, Z., et al.: Exploring wav2vec 2.0 on Speaker Verification and Language Identification. In: Proceedings of Interspeech, pp. 1509–1513 (2021)
- 20 Shi, X., et al.: The accented English speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6918–6922. IEEE, Piscataway (2021)
- 21 Ardila, R., et al.: Common Voice: A massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4218–4222 (2020)
- 22 You, Z., et al.: SpeechMoE: Scaling to large acoustic models with dynamic routing mixture of experts. In: Proc. Interspeech, pp. 2077–2081 (2021)
- 23 You, Z., et al.: SpeechMoE2: Mixture-of-experts model with improved routing. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7217–7221. European Language Resources Association, Paris (2022)
- 24 Gaur, N., et al.: Mixture of informed experts for multilingual speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6234–6238. IEEE, Piscataway (2021)
- 25 Panayotov, V., et al.: LibriSpeech: an ASR corpus based on public domain audio books. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. IEEE, Piscataway (2015)
- 26 Gao, Z., et al.: SAN-M: Memory equipped self-attention for end-to-end speech recognition. Proceedings of Interspeech, pp. 6–10 (2020)
- 27 Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
- 28 Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. **30**, pp. 5998–6008. MIT Press, Cambridge (2017)
- 29 Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 66–75. Association for Computational Linguistics, Stroudsburg, PA (2018)
- 30 Park, D.S., et al.: SpecAugment: A simple data augmentation method for automatic speech recognition. In: Proceedings of Interspeech, pp. 2613–2617 (2019)