DialectMoE: An End-to-End Multi-Dialect Speech Recognition Model with Mixture-of-Experts

Anonymous ACL submission

Abstract

Dialect speech recognition has always been one of the challenges in Automatic Speech Recognition (ASR) systems. While lots of ASR systems perform well in Mandarin, their performance significantly drops when handling dialect speech. This is mainly due to the obvious differences between dialects and Mandarin in pronunciation and the data scarcity of dialect speech. In this paper, we propose DialectMoE, a Chinese multi-dialects speech recognition model based on Mixture-of-Experts (MoE) in a low-resource conditions. Specifically, DialectMoE assigns input sequences to a set of experts using a dynamic routing algorithm, with each expert potentially trained for a specific dialect. Subsequently, the outputs of these experts are combined to derive the final output. Due to the similarities among dialects, distinct experts may offer assistance in recognizing other dialects as well. Experimental results on the Ai-Datatang dialect public dataset show that, compared with the baseline model, DialectMoE reduces Character Error Rate(CER) for Sichuan, Yunnan, Hubei and Henan dialects by 23.6%, 32.6%, 39.2% and 35.09% respectively. The proposed DialectMoE model demonstrates outstanding performance in multi-dialects speech recognition.

1 Introduction

The application domains of speech recognition technology are extensive, encompassing diverse fields such as voice assistants, smart homes, and automotive voice interaction, among others. Thanks to the advancements in deep learning, Automatic Speech Recognition(ASR) systems have made remarkable strides in recognizing Mandarin speech(Malik et al., 2021; Wang et al., 2019; Alharbi et al., 2021). Dialect serves as a prevalent mode of everyday communication among the Chinese populace. However, the performance of ASR systems remains limited in dialect speech, posing a significant challenge in the field of speech recognition technology(Hinsvark et al., 2021; Alsharhan and Ramsay, 2020) due to the inherent variations and distinct characteristics in pronunciation among dialects and Mandarin. Therefore, improving the accuracy and adaptability of Chinese ASR systems is significant and meaningful for multi-dialect. Our study mainly focuses on Chinese dialects, the proposed method can also be generalized to other dialects.

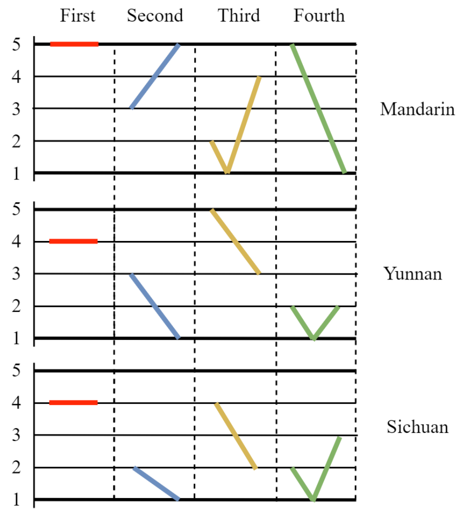


Figure 1: The tonal distinctions among Standard Mandarin, Yunnan dialect, and Sichuan dialect.

Chinese dialects are typically classified into ten main categories, each exhibiting notable differences in pronunciation, tone, vocabulary, and grammar(Ho, 2015). Chinese is a tonal language, where each character corresponds to a specific tone, a feature that is prevalent in most of its dialects as well. The pronunciation of a given Chinese character with different tones imparts markedly distinct meanings. This underscores the profound significance of tones in the comprehension of Chinese phonetics. (Ho, 2015; Sproat et al., 2004). Figure 1 depicts the tonal distinctions among Standard Mandarin, Yunnan dialect, and Sichuan dialect. It is evident that notable differences exist in tonal between Standard Mandarin and Yunnan dialect as well as Sichuan dialect in the second, third, and fourth tones. However, Yunnan dialect and Sichuan dialect exhibit a pronunciation similarity in specific tones. The change in tone reveal differences and similarities between Standard Mandarin and various dialects. Hence, considering both the differences and similarities in pronunciation among various dialects alongside Standard Mandarin becomes crucial for the advancement of Chinese speech recognition systems.

In recent years, numerous researches have focused on tackling the challenge of poor performance in dialect speech recognition models(Li et al., 2018; Ren et al., 2019; Zhang et al., 2022b). The traditional way is based on different modeling methods to improve the effect of dialect speech recognition. Humphries et al. (1996) employed an adaptive method that utilizes a pronunciation vocabulary with dialect data to capture differences between standard and dialect pronunciations. Li et al. (2019) proposed a novel method for modeling Chinese characters based on radicals, effectively addressing the issue of dialect modeling difficulty. This method significantly reduces the required size of radical dictionaries compared to ordinary character dictionaries. Recently, multitask-based methods have been widely used in the task of dialect speech recognition. Compared with the traditional method, the multi-task learning method is more efficient. Elfeky et al. (2016) proposed constructing a dialect classification model and a separate speech recognition model for each dialect. The dialect classification model is used to select the corresponding dialect speech recognition model. Dan et al. (2022) proposed a multi-task training strategy that combines dialect classification with dialect speech recognition, bridging the substantial gap between Mandarin and dialect acoustic properties. However,, these investigations are contingent upon extensive dialectal datasets and do not examine the potential influence of commonalities among various dialects on model performance.

To construct a reliable dialect speech recognition model in low-resource conditions, Jiang (2023) introduced a transfer learning-based approach, it involves a model trained on Mandarin and fine-tunes with small-scale dialect data. However, relying solely on transfer learning may not adequately capture the distinctions between dialects and Mandarin. Wang et al. (2023) proposed Aformer model with multi-stage training strategy, which can capture diverse acoustic information in different training stage, enabling the model to effectively adapt to dialect data. The aforementioned studies focus on the training strategy and model expansion, they do not fully consider the differences and similarities between dialects and Mandarin.

In this paper, we present DialectMoE, a multi-dialect speech recognition model based on Mixture-of-Experts (MoE), aimed at improving the performance of multi-dialects speech recognition in low-resource conditions. DialectMoE is architecturally structured with dual encoders: a dialect encoder and a general encoder. The main contributions of the paper include:

- We propose a three-stage training methodology designed to enhance the model's adaptability in addressing low-resource multi-dialect scenarios through different stages. Detailed specifics will be expounded upon in Section 3.3.

- We introduce MoE layers and enhance the dynamic routing algorithm to enable the combination of acoustic features from both the input sequence and the dialect encoder during the expert selection process.

- The experiment results show that DialectMoE reduces Character Error Rate(CER) compared to the baseline model for Sichuan, Yunnan, Hubei and Henan dialects by 23.6%, 32.6%, 39.2% and 35.09%, respectively.

2 Related Work

2.1 Conformer-based ASR

The Conformer, a convolution-augmented Transformer introduced in (Gulati et al., 2020), has been widely acknowledged as the state-of-the-art end-to-end ASR technology owing to its exceptional performance in ASR tasks. In recent years, several researchers have proposed Conformer variants (Peng et al., 2022; Sehoon et al., 2023) to further enhance the capabilities of speech recognition. The Conformer module comprises two feed-forward modules, a multi-heads self-attention module, and a convolution module. The output y of one Conformer block for a given input x can be defined as follows:

$$\hat{x} = x + \frac{1}{2}\mathrm{FFN}_1(x) \tag{1}$$

$$\tilde{x} = \hat{x} + \mathrm{MHSA}(\hat{x}) \tag{2}$$

$$\bar{x} = \tilde{x} + \mathrm{Conv}(\tilde{x}) \tag{3}$$

$$y = \mathrm{LN}(\bar{x} + \frac{1}{2}\mathrm{FNN}_2(\bar{x})) \tag{4}$$

where FNN_1 denotes the first feedforward module, FNN_2 denotes the second feedforward network, MHSA denotes the multi-head self-attention module, Conv denotes the convolution module, and LN denotes layer normalization. For additional information regarding the Conformer ASR model, please refer to (Gulati et al., 2020).

During Conformer training, the Joint CTC-Attention loss function (Hori et al., 2017) is utilized. This loss function is commonly used in present-day speech recognition technology. In this paper, the joint CTC-Attention loss is incorporated into the total loss function. The loss function is outlined as follows:

$$\mathcal{L}_{all} = (1-\lambda)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc} \tag{5}$$

where $\mathcal{L}_{att}$ denotes the decoding loss of the Attention decoder, and $\mathcal{L}_{ctc}$ denotes the CTC loss., λ is a hyper parameter which denotes the weight of these two loss.

2.2 Mixture-of-Experts Based Speech Recognition

The MoE based methods offer a solution for more efficient training and inference by selectively activating different experts in the model based on different inputs (Jacobs et al., 1991; Shazeer et al., 2017). This enables the model to adapt to a wide range of inputs and scale to more parameters while maintaining a consistent computational cost. The MoE based models have demonstrated their effectiveness in natural language processing (Fedus et al., 2022; Du et al., 2022) and computer vision (Riquelme et al., 2021; Fan et al., 2022).

In real-world application, speech recognition systems need to handle various input conditions, including speaker variation, accent variation, and acoustic environment (Zilvan et al., 2021). However, conventional speech recognition models have a fixed computational cost and cannot adapt to the complexity of input instances. You et al. (2021, 2022) explore the MoE based model for speech recognition, named SpeechMoE, and propose a new router architecture which integrates additional global domain and various embedding into router input to promote adaptability. Additionally, a multilingual speech recognition network (MoLE) was introduced (Kwon and Chung, 2023) to analyze audio input data from multiple languages and identify expert networks suitable for each language. Simultaneously, a language-independent expert network was also introduced, and the selected expert network and the language-independent expert network collectively fulfill the language requirements necessary for effective speech recognition.

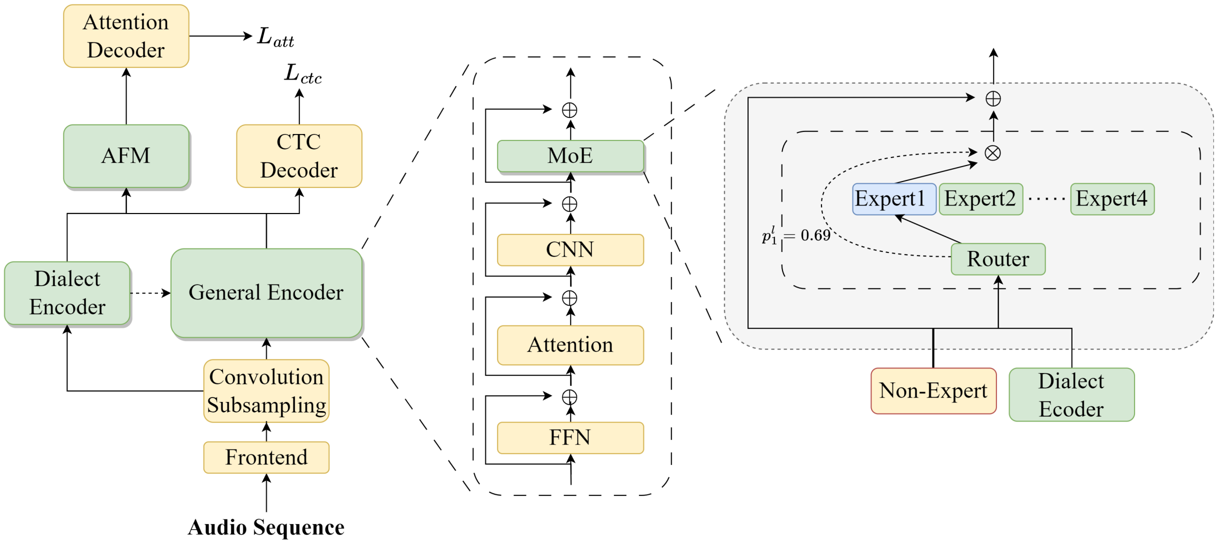
Employing the MoE mechanism to determine expert activation during the forward propagation process manifests a notable capacity for accommodating the inherent variability in multi-dialectal speech across different input sequences. Nevertheless, the conventional MoE paradigm relies solely on the input sequence for expert selection, and the information in the present input does not inherently ensure the optimal suitability of the selected experts. Therefore, the incorporation of supplementary dialectal information to facilitate expert selection stands forth as a judicious resolution, enhancing the precision and adaptability of the chosen experts to the distinctive intricacies characterizing the multi-dialectal speech context. Furthermore, the exploration of MOE-based methods in the domain of multi-dialect speech recognition remains limited.

3 DialectMoE

3.1 Overall Architecture of DialectMoE

The overall architecture of DialectMoE is shown in Figure 2a. The original audio sequence undergoes preprocessing by the frontend module to extract FBank features. Subsequently, the convolutional downsampling is applied to temporally downsample the audio feature sequence. The dialect encoder, consisting of 6 layers of vanilla Conformer encoder, captures dialect information from the feature sequences. On the other hand, the general encoder comprises 12 layers of DialectMoE encoder blocks, are responsible for capturing speech information in a normal manner. Both encoders share the same input but focus on different aspects of information.

The detailed structure of the DialectMoE block is presented in Figure 2b. With the DialectMoE block, the input sequence firstly passes through the Feed-Forward Network (FFN) layer, followed by Attention and Convolutional Neural Network



(a) DIalectMoE (b) DIalectMoE block

Figure 2: (a) DialectMoE overall architecture, where the general encoder consists of N DialcetMoE blocks. (b) Architecture of the DialectMoE encoder block module.

(CNN) layer to extract global and local information, respectively. Then the appropriate expert within the MoE layer is selected based on the dynamic routing. The output of experts are multiplied by the weight assigned by the router layer.

Compared to the widely used vanilla Conformer block(Gulati et al., 2020), our DialectMoE block incorporates MoE layers to address complex and variable scenarios encountered in real-world situations. The dialect information captured by the dialect encoder is weighted by the router layer, which enables the router layer to choose more appropriate experts based on both dialect features and general features obtained from two encoders. This dynamic routing mechanism proves more effective in intricate speech scenarios, especially those involving multiple dialects.

3.2 Dialect Adaptive Dynamic Expert Routing

In the context of multi-dialect speech recognition, effectively addressing the diversity of dialectal variations is crucial. We present a novel dynamic routing algorithm aimed at enhancing the adaptability and generalization of the model to diverse dialects speech input sequences. The proposed algorithm leverages the input sequences from the current MoE layer and the dialect information provided by the dialect encoder to select appropriate experts. To evaluate the impact of different dialect embedding on routing, we consider the following three strategies: input sequence **concat** dialect embedding,

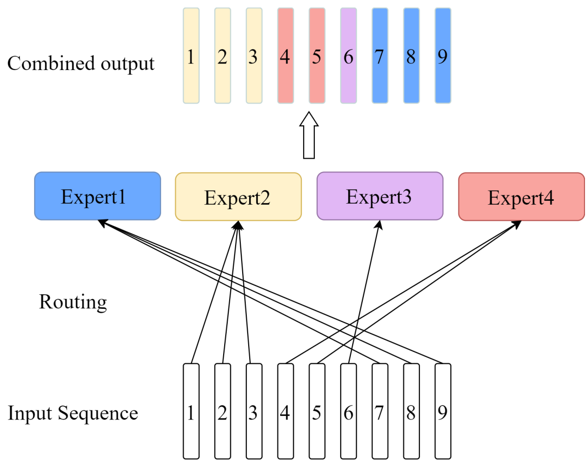


Figure 3: Illustration of dynamic routing algorithm.

input sequence **add** dialect embedding, and only use embedding(**embed**) of dialects. The output of the dialect encoder is denoted as $\mathbf{X}^D_{encoder} \subseteq \mathbb{R}^{T\times d}$, where T represents the sequence length and d denotes the feature dimension. Assuming that there are N experts, the output $r \subseteq \mathbb{R}^{T\times N}$ of the routing layer can be defined as follows. :

$$r = W_r \cdot \mathrm{Concat}(\bar{x}, \mathbf{X}^D_{ecndoer}) \tag{6}$$

$$r = W_r \cdot \mathrm{Add}(\bar{x}, \mathbf{X}^D_{ecndoer}) \tag{7}$$

$$r = W_r \cdot \mathbf{X}^D_{ecndoer} \tag{8}$$

where W_r represents the weight parameter of the router layer, and $\bar{x}$ denotes the output of the convolution module. These three dynamic routing strategies are the ones we consider employing. It is worth noting that while the general router layer selects experts based on the input sequence $\bar{x}$, the algorithm we designed intuitively makes more sense as it incorporates the output of the dialect encoder to select the most suitable expert.

The router layer selects the expert with the highest probability through dynamic routing, which is based on the input sequence r. The dynamic routing probability is then defined as follows:

$$p_i = \frac{exp^{r_i}}{\sum_{j=1}^{N} exp^{r_j}} \tag{9}$$

where $p_i \subseteq \mathbb{R}^{T \times N}$ is the probability that the i expert is selected, the output $\mathbf{O}_{moe} \subseteq \mathbb{R}^{T \times d}$ of the MoE layer can be formally defined as follows:

$$\mathbf{O}_{moe} = p_i \cdot E_i(\bar{x}) \tag{10}$$

where E_i is the output of the i expert selected. Figure3 illustrates the process of dynamic routing.

In order to incorporate dialect information into the decoder, DialectMoE incorporates an information fusion step by combining the outputs of two separate encoders. This fusion process, illustrated as the Acoustic Fusion Module(**AFM**) in Figure 2(a), occurs prior to transmitting the results to the decoder. The fusion process is defined as follows:

$$\mathbf{X}^{A}_{encoder} = \text{Concat}(\mathbf{X}^{G}_{encoder}, \mathbf{X}^{D}_{encoder}) \tag{11}$$

where $\mathbf{X}^{A}_{encoder}$ denotes the result of fusion of information output by two two different encoders, and $\mathbf{X}^{G}_{encoder}$ denotes the result output by a general encoder.

The comprehensive loss function for speech recognition comprises the combined CTC-Attention loss(Hori et al., 2017), as explained previously, along with the supplementary balance loss(Fedus et al., 2022). The complete formulation of the loss function is as follows:

$$\mathcal{L}_{all} = \lambda \mathcal{L}_{ctc} + (1 - \lambda)\mathcal{L}_{att} + \alpha \mathcal{L}_b \tag{12}$$

where α is the weight of the balance loss ($\alpha = 0.1$) and λ is the weight of the speech recognition loss ($\lambda = 0.3$), $\mathcal{L}_b$ denotes the balance loss.

3.3 Training Strategies

Considering the significant disparity in the quantities of Mandarin and dialect data, low-resource dialect speech recognition scenarios commonly exhibit limited labeled dialect speech data, typically ranging from a few to tens of hours. This insufficiency hampers the development of a reliable speech recognition model. To address this issue, this study introduces a multi-stage training strategy. The training process encompasses the following sequential steps:

1. Pre-training: The Conformer model is used as a general encoder for DialectMoE to implement pre-training on Mandarin datasets. The pre-training step allows the model to capture various common speech features, thus reducing the complexity of learning for the dialect recognition task.

2. Training Dialect Encoder: A Conformer Encoder is initialized as a dialect encoder and is trained on the dialect classification task using both dialect and Mandarin data. The objective of this step is to enable the dialect encoder to learn the acoustic differences between Mandarin and dialects, assisting the general encoder in dialect speech recognition tasks.

3. Training DialectMoE: The parameters of the dialect encoder are frozen, and the second feedforward network layer in the pre-trained Conformer model is initialized with N experts. Use only low-resource dialect training data to train the final DialectMoE model.

By pre-training phase, the initial model acquires a substantial set of effective parameters, thereby conferring notable advantages for last training stages. In the second phase, the dialect encoder is trained on a dialect identification task, enabling it to focus on differences between multi-dialects and Mandarin. In the last stage, only multi-dialect data is used for training, This strategic approach enhances DialectMoE's capability to adeptly capture shared acoustic characteristics across diverse dialects. This approach enhances DialectMoE's capability to adeptly capture shared acoustic characteristics across multi-dialects. The proposed method is evaluated based on extensive comparison and ablation experiments, which are comprehensively detailed in Section 4.

4 Experiments

4.1 Datasets

The Aishell dataset (Bu et al., 2017) serves as the Mandarin speech corpus in this study. This extensive collection of Mandarin Chinese speech data,

Dataset	Train(h)	Test(h)
Aishell	164	10
Sichuan(SC)	28.5	1.5
Yunnan(YN)	28.5	1.5
Henan(HN)	0	1.5
Hubei(HB)	0	1.5

Table 1: Details of both Dialect and Mandarin datasets.

encompasses diverse acoustic scenarios such as reading and dialogue.

For the Chinese dialect dataset, an open-source dataset provided by AiDatatang[1] is utilized in this study. It comprises a training set of 30 hours of Sichuan and Yunnan dialects and a test set of 1.5 hours featuring Henan and Hubei dialects. Within this study, Sichuan(SC) and Yunnan(YN) dialects are used to test the adaptability of the model to multi-dialect data, and Henan(HN) and Hubei(HB) dialects are used to test the generalization of the model to multi-dialect data. More details are shown in Table1.

4.2 Experiment Setup

All experiments were conducted using the Wenet(Zhang et al., 2022a) end-to-end speech toolkit. Our methodology involved extracting an 80-dimensional log-Mel filter bank (Fbank) as the acoustic input feature, with a window size of 25 ms and a step size of 10 ms. To ensure feature normalization, we applied cepstrum mean variance normalization (CMVN) calculated from the training set on Fbank. To augment the low-resource dialect data, we employed speed perturbation and SpecAugment(Park et al., 2019) techniques. No additional language models were incorporated into the experiments.

For the pre-training model, we utilized a Conformer encoder trained on the Mandarin dataset. The general encoder of DialectMoE consists of 12 Conformer encoder layers with a feed-forward dimension of 2048 and an attention dimension of 256, employing 4 self-attention heads. This model was trained using the Adam optimizer (Kingma and Ba, 2014). Furthermore, we adopted the warmup learning schedule (Gotmare et al., 2018) for the initial 25K training iterations and set the label smoothing (Szegedy et al., 2016) weight and dropout to 0.1 for model regularization. The decoder consists of a 6-layer Transformer, while the dialect encoder comprises a 6-layer Conformer encoder. The loss function for the dialect classification task applies cross-entropy loss to all training datasets.

The proposed DialectMoE is initialized with pre-trained general encoder, dialect encoder, and decoder. The second feedforward layer in each Conformer layer of the general encoder is initialized as an N expert ($N = 4$), with the expert parameters being the pre-trained feedforward network parameters. Training employed the same Adam optimizer, and the number of warm-up steps in the pre-training learning plan was adjusted to 10000, with an initial learning rate of 0.001.

4.3 Main Results

In this paper, we meticulously design comparison experiments with other speech recognition models to showcase the effectiveness of our proposed method. The experimental results presented in this study were reproduced using the open-source speech processing toolkit Wenet (Zhang et al., 2022a). Table2 illustrates the performance of each ASR model in dialect speech recognition under low-resource conditions. The evaluation metric employed is the Character Error Rate (CER).

M1 represents the Conformer model that was exclusively pre-trained on the Aishell Mandarin dataset, consisting of 178 hours of data.

M2 denotes the model fine-tuned from M1 using the low-resource dialect dataset.

M3 corresponds to the model trained directly on the combined dataset of both dialect and Mandarin speech.

M4 and **M5** refer to the multi-tasking models trained on the combined dataset, with a distinction that M4 predicts the dialect category in the encoder while the decoder focuses on recognizing the speech text, whereas M5 predicts both the dialect category and the speech text in the decoder.

M6 represents the multi-pass model proposed in (Wang et al., 2023) for the training of the Aformer.

M7 signifies the DialectMoE model proposed in this paper.

The results obtained from the M1 model demonstrate a notably poor performance in recognizing dialectal speech within the Mandarin speech recognition model. However, by fine-tuning the M1 model with dialect data, the CER of the M2 model for Sichuan and Yunnan dialects is significantly reduced, although further optimization is still required. To address this, our paper proposes the Di-

[1] https://www.datatang.com

ID	Model	Params(M)	SC	YN	HN	HB
M1	Conformer (Mandarin)	46.1 M	82.75	81.42	82.26	87.47
M2	FT-Conformer	46.1 M	15.60	14.06	54.91	57.18
M3	Conformer (Mandarin+Dialect)	46.1 M	13.86	12.02	41.28	48.37
M4	MT-Conformer(DID+ASR)	47.2 M	17.79	15.64	47.78	56.55
M5	MT-Conformer(DID&ASR)	46.2 M	16.09	14.05	41.28	48.37
M6	Aformer	68.3 M	13.21	12.76	35.32	39.89
M7	DialectMoE	93.8 M	**11.91**	**9.84**	**33.38**	**37.11**

Table 2: CER(%) on Chinese Dialect ASR task. **FT** represents the fine-tuning step and **MT** represents the multitask-based approach.

Strategy	SC	YN	HN	HB
normal	13.43	11.83	35.22	38.67
embed	12.53	10.20	34.95	38.65
concat	12.18	9.97	33.55	**37.09**
add	12.41	10.36	34.19	37.23
normal+fusion	13.92	12.19	35.27	38.72
embed+fusion	12.23	10.21	33.50	37.55
concat+fusion	**11.91**	**9.84**	**33.38**	37.11
add+fusion	12.49	10.13	33.66	37.65

Table 3: Ablation of different routing strategy.

Model	Params(M)	SC	YN	HN	HB
MoE-2e	68.5M	12.39	10.21	33.84	38.37
MoE-4e	93.8M	**11.91**	**9.84**	**33.38**	**37.11**
MoE-8e	134.5M	12.94	10.44	34.09	38.26

Table 4: Ablation of experts number.

alectMoE model, which surpasses existing studies and baselines in terms of performance. In comparison to the fine-tuned model of M2, the DialectMoE model exhibits a reduction in CER of 23.6% and 32.6% for the Sichuan and Yunnan dialects, respectively. Additionally, it achieves a reduction of 39.2% and 35.09% for the Henan and Hubei dialects, respectively.

4.4 Ablation Studies

4.4.1 Ablation of dynamic routing strategy

This paper incorporates ablation experiments to investigate the effectiveness of the proposed dynamic routing algorithm and model design. Table 3 presents the impact of utilizing different dynamic routing algorithms and the merging of two encoder outputs before the decoder. In "normal", the dynamic routing algorithm proposed in this paper is not employed, and the experts are directly selected based on the input sequence, similar to the approach in (Fedus et al., 2022). The strategy column in Table3 indicates the usage of different dynamic routing algorithms: "**embed**" signifies the utilization of only the dialect encoder outputs, "**concat**" denotes the concatenation of the dialect encoder outputs with the input sequence, and "**add**" indicates the summation of the dialect encoder outputs with the input sequence. The "**fusion**" entry indicates whether or not the two encoder outputs should be fused before reaching the decoder., whether they go through AFM. The experiments employing the ”**concat+fusion**“ strategy along with the fusion of the two encoder outputs demonstrate optimal results across the four different dialect test sets.

4.4.2 Ablation of experts number

To investigate the impact of initializing a different number of experts in DialectMoE on the overall model performance, we conducted an experiment with varying numbers of experts, specifically 2, 4, and 8. The experimental results, as shown in Table4, highlight that the model size increases with an increasing number of experts. However, the common notion that a larger number of model participants leads to improved performance does not hold true under low-resource conditions. The results indicate that, for low-resource dialect data, an excessive number of experts does not enhance model performance; in fact, it diminishes it. Experimental evidence supports the conclusion that setting the number of experts to 4 is more appropriate in this context.It is noteworthy that when the number of experts is set to 2, the number of model parameters matches the number of Aformer(Wang et al., 2023) parameters. However, despite this similarity, our results outperform the baseline. This finding further validates the efficacy and correctness of our proposed method.

Top-k	Time(s)	SC	YN	HN	HB
4	1.46s	12.01	10.03	**33.31**	**37.08**
2	0.94s	11.93	**9.81**	33.57	37.24
1	**0.68s**	**11.91**	9.84	33.38	37.11

Table 5: Ablation of the number of experts selected.

4.4.3 Ablation of the number of experts selected

In the vanilla MoE, a top-k approach is employed to select a combination of k experts for routing the input sequence. However, in this paper, a Softmax approach, specifically top-1, is utilized. To further investigate the effectiveness of the proposed dynamic routing algorithm, experiments were conducted to explore the impact of the number of selected experts. As presented in Table5, when the number of selected experts is set to 4, there is an improvement in performance for dialects that are not part of the training dataset (Henan and Hubei dialects). This suggests that increasing the number of selected experts can enhance the model's generalization to external data. The model's performance remains similar when the number of selected experts is 2 or 1. However, it is worth noting that the decoding time for a single speech increases by approximately 53% when the number of selected experts is 4 compared to when it is 1. This indicates that the number of selected experts has a minimal impact on the model's performance but significantly affects decoding efficiency, which is crucial for a robust speech recognition system.

4.5 Layer-Wise Analysis of Experts

In Figure 4, we present a visualization of the expert weights applied to the test sets corresponding to Sichuan and Kunming dialects. We can observe certain patterns in the weights. Across the initial three layers of the model, both dialects manifest a heightened degree of distinctiveness in expert selection, indicative of specific groups of experts concentrating exclusively on dialect-specific information. Within the intermediate layers of the model, expert weights display a diminished prominence, yet discernible differences persist in the expert weights associated with the two dialects. This observation suggests that varying combinations of experts implicitly encapsulate distinctive information pertaining to dialectal variations. In the concluding three layers of the model, the deployed experts exhibit near-identical characteristics, thereby indirectly affirming the model's proficiency in capturing shared features between Sichuan and Kunming dialects.

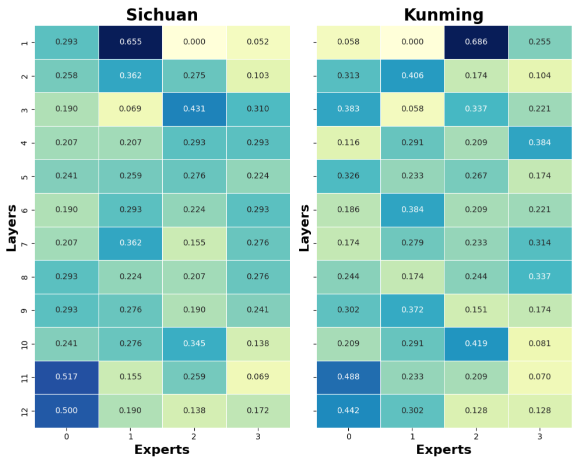


Figure 4: The expert weights are visualized on Sichuan dialect and Kunming dialect.

5 Conclusion

In this manuscript, we present a multi-dialectal speech recognition model based on MoE termed DialectMoE. Structurally, it incorporates a dual-encoder architecture, wherein the general encoder is dedicated to acquiring general acoustic representations, and the dialect encoder is specialized for acquiring acoustic representations across various dialects. A refinement in the dynamic routing strategy within the MoE layer of the universal encoder has been introduced to enable the selection of appropriate experts based on the acoustic information specific to the dialect in the input sequence. Furthermore, we propose a three-stage training methodology to facilitate DialectMoE in learning distinct tasks at different phases, thereby enhancing its adaptability and performance across varying aspects of the multi-dialectal speech recognition task. Experimental results demonstrate that the proposed DialectMoE model achieves remarkable performance in multi-dialects speech recognition tasks.

6 Limitations

While the MoE-based approach can effectively enhance model performance, it inherently results in an increase in the number of model parameters. This increase in parameters can lead to higher training costs and longer inference times, which are inevitable consequences. Therefore, it is imperative to conduct further research on model compression techniques to mitigate these issues.

References

Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876.

Eiman Alsharhan and Allan Ramsay. 2020. Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition. *Language Resources and Evaluation*, 54:975–998.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*, 24(10):1429.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Mohamed Elfeky, Meysam Bastani, Xavier Velez, Pedro Moreno, and Austin Waters. 2016. Towards acoustic model unification across dialects. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 624–628. IEEE.

Zhiwen Fan, Rishov Sarkar, Ziyu Jiang, Tianlong Chen, Kai Zou, Yu Cheng, Cong Hao, Zhangyang Wang, et al. 2022. M^3vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35:28441–28457.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

Dah-an Ho. 2015. Chinese dialects. *The Oxford handbook of Chinese linguistics*, pages 149–159.

Takaaki Hori, Shinji Watanabe, and John R Hershey. 2017. Joint ctc/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529.

Jason J Humphries, Philip C Woodland, and D Pearce. 1996. Using accent-specific pronunciation modelling for robust speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2324–2327. IEEE.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Rui Jiang. 2023. Chinese dialect recognition based on transfer learning. *INTERSPEECH 2023*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yoohwan Kwon and Soo-Whan Chung. 2023. Mole: Mixture of language experts for multi-lingual automatic speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4749–4753. IEEE.

Sheng Li, Xugang Lu, Chenchen Ding, Peng Shen, Tatsuya Kawahara, and Hisashi Kawai. 2019. Investigating radical-based end-to-end speech recognition systems for chinese dialects and japanese. In *INTERSPEECH*, pages 2200–2204.

Mishaim Malik, Muhammad Kamran Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. In *International Conference on Machine Learning*, pages 17627–17643. PMLR.

Zongze Ren, Guofu Yang, and Shugong Xu. 2019. Two-stage training for chinese dialect recognition. *arXiv preprint arXiv:1908.02284*.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Kim Sehoon, Gholami Amir, Shaw Albert, Lee Nicholas, Mangalam Karttikeya, Keutzer Kurt, et al. 2023. Squeezeformer: An efficient transformer for automatic speech recognition.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Richard Sproat, Liang Gu, Jing Li, Yanli Zheng, Yi Su, Haolang Zhou, Philip Bramsen, MIT David Kirsch, Izhak Shafran, Stavros Tsakalidis, et al. 2004. Dialectal chinese speech recognition. In *CLSP Summer Workshop*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Dong Wang, Xiaodong Wang, and Shaohe Lv. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Xuefei Wang, Yanhua Long, Yijie Li, and Haoran Wei. 2023. Multi-pass training and cross-information fusion for low-resource end-to-end accented speech recognition. *arXiv preprint arXiv:2306.11309*.

Zhao You, Shulin Feng, Dan Su, and Dong Yu. 2021. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. *arXiv preprint arXiv:2105.03036*.

Zhao You, Shulin Feng, Dan Su, and Dong Yu. 2022. Speechmoe2: Mixture-of-experts model with improved routing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7217–7221. IEEE.

Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu. 2022a. Wenet 2.0: More productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455*.

Fengrun Zhang, Xiang Xie, and Xinyue Quan. 2022b. Chinese dialect speech recognition based on end-to-end machine learning. In *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*, pages 14–18. IEEE.

Vicky Zilvan, Ana Heryana, Asri Rizki Yuliani, Dikdik Krisnandi, R Sandra Yuwana, and Hilman F Pardede. 2021. Front-end based robust speech recognition methods: A review. In *Proceedings of the 2021 International Conference on Computer, Control, Informatics and Its Applications*, pages 136–140.