Dynamic Language Routing Mixture of Experts Model for Multilingual Speech Recognition

Junchen Wang, Yonghe Wang†, Feilong Bao, Guanglai Gao

College of Computer Science, Inner Mongolia University, Hohhot, China
Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian

Abstract

Multilingual Automatic Speech Recognition (MASR) systems have inherent difficulties in balancing computational efficiency and modeling phonetic similarities between different languages. This study proposes a MASR framework based on a Mixture-of-Experts (MoE) architecture, which incorporates a dynamic Top-k expert routing strategy. This method dynamically selects the optimal expert subnetwork based on the input speech features, thereby achieving more efficient and accurate language modeling. To address the ambiguity in expert selection caused by phonetic similarities between different languages, we further proposes a *Token-Level Realignment* (TLR) method that accurately aligns language representations with expert groups at the framework level, effectively alleviating the cross-language interference problem. The experiments achieved excellent recognition performance in four languages (English, Chinese, Khalkha Mongolian, Chahar Mongolian), with WERs/CERs of 12.88%, 5.49%, 7.64%, and 19.30%, respectively.

Keywords: multilingual automatic speech recognition; mixture-of-experts; dynamic routing; token-level realignment; mongolian language

1. Introduction

Multilingual Automatic Speech Recognition (MASR) has garnered significant research interest in recent years, driven by increasing demands for cross-lingual communication. While traditional MASR systems require separate models for each language, modern end-to-end (E2E) approaches enable unified multilingual modeling. Benefiting from self-supervised learning and large amounts of training data, representative Automatic Speech Recognition (ASR) models, like Whisper [1], Google USM [2], and MMS [3], demonstrate remarkable multilingual capabilities spanning over a thousand languages while maintaining competitive performance metrics. However, the practical deployment of such models poses challenges, particularly in optimizing the trade-off between computational efficiency and cross-lingual interference mitigation.

To address these limitations, several approaches have been investigated to enhance recognition performance in multilingual settings, such as multi-task learning [4, 5], language information integration [6, 7], and Mixture-of-Experts (MoE) [8, 9]. The MoE paradigm fundamentally

†Corresponding author: Yonghe Wang (Email: cswyh@imu.edu.cn; ORCID:0000-0003-1647-1539)

differs from conventional monolithic architectures that process all linguistic inputs through a unified shared parameter. MoE employs a dynamic routing mechanism to distribute input features to language-specific processing paths to reduce cross-lingual interference and maintain computational efficiency. For example, the LR-MoE framework [10] introduced language-specific routing through MoE-based Feed-Forward Networks (FFNs) to mitigate cross-lingual interference, and activate only one expert at a time during training and inference for computational efficiency. BLR-MoE [11] advances LR-MoE architectures by extending MoE to both FFN and self-attention layers to reduce language confusion, while enhancing router robustness through expert pruning and augmented language identification (LID) classification.

However, both architectures lack the hierarchical granularity of intra-language semantic modeling. To address this issue, DLG-MoE [12] introduces a hierarchical routing mechanism that combines explicit language modeling with implicit attribute learning, which has been empirically validated to achieve superior performance in code-switching speech recognition (CS-ASR) tasks. The DLG-MoE framework first uses language routers to explicitly identify inputs and route them to the corresponding language expert groups, and then uses unsupervised routers within each expert group to implicitly capture finer-grained language differences such as dialects, accents, and domains, enabling more nuanced processing of code-switching speech.

Inspired by the above, based on DLG-MoE [12], we propose a multilingual speech recognition modeling approach based on the hierarchical routing mechanism. which comprises three fundamental components: (1) a language identification router for multilingual discrimination; (2) a lower-level expert routing network that integrates feedforward neural networks with local attention mechanisms to model the acoustic features of each language; and (3) an upper-level expert routing network based on multi-head attention mechanisms to capture common structural information across languages. Our main contributions are as follows:

- To more effectively learn the acoustic features and common language structures of different languages, we adopted a hierarchical expert routing network and introduced a *Token-Level Realignment* mechanism to achieve frame-level precise alignment, effectively reducing cross-language interference and alleviating language confusion issues.
- To achieve the optimal balance between computational efficiency and model performance in multilingual speech recognition, we introduced a dynamic top-k routing strategy in the expert selection process.
- We systematically evaluated the impact of expert count and inference strategies on model performance, validating robust recognition performance in mixed speech scenarios and multilingual streaming speech recognition tasks.

2. Related Work And Motivation

2.1. Mixture-of-Experts based MASR

In recent years, the MoE mechanism has made significant progress in MASR. For example, [13, 14, 15] research utilizing shared embedding networks and hierarchical MoE representations has improved the expert routing mechanism. Although these methods have enhanced the performance of multilingual systems, they still face challenges in handling cross-lingual interference and optimizing computational efficiency.

To further improve the cross-lingual performance of MASR, language-specific expert routing strategies have gained attention. MoLE [16] proposes activating language-specific experts

and aggregating them with language-agnostic experts, demonstrating its applicability for low-resource languages. M-MoE [17] introduces a dual-layer MoE structure, with routing mechanisms designed for both known and unknown languages, supporting a wide range of languages. These approaches [18, 19] can be seen as information-driven expert models that rely on language information to select the corresponding experts. [20] improved the performance of the language model by introducing prompt information. Additionally, due to the scarcity of Mongolian language resources, [21, 22] provides a standardized lexical framework for low-resource languages. With the growing demand for streaming speech recognition tasks, MoE-Conformer [23] combines the advantages of streaming ASR and MoE by activating a fixed number of experts, thereby improving the efficiency of multilingual streaming ASR. Building on the work of these researchers, we extend the adaptability of MoE-based models to low-resource languages, such as Khalkha Mongolian (kMN) and Chahar Mongolian (cMN).

2.2. Motivation

Many MoE-based architectures have achieved success in handling multilingual recognition through shared language embedding networks and multi-layer MoE designs. However, these methods still face challenges in mitigating cross-lingual interference and optimizing computational efficiency. To address these issues, researchers in recent years have proposed language-specific expert routing mechanisms and multi-language expert selection strategies. These methods have significantly improved the recognition ability of multilingual speech and enhanced the model's performance in multilingual tasks. In particular, MoE-based expert networks dynamically select and activate the most relevant experts by assigning appropriate sub-networks to different languages, thus reducing cross-lingual interference and improving cross-lingual adaptability.

However, existing MoE architectures still face the issue of complex hierarchical structures, especially when dealing with acoustic similarities between different languages. For instance, when handling mixed-language speech, the model may misrecognize speech of similar languages as belonging to another language. To solve this problem, we propose a Token-Level Realignment (TLR) method, which filters candidate tokens at each frame to ensure that only tokens from the target language are selected, effectively reducing cross-lingual interference.

Based on this, we propose a TLR method for multilingual speech recognition, optimizing adaptability for low-resource languages. Through an improved dynamic routing strategy and language identity recognition, we aim to enhance the model's computational efficiency and cross-lingual adaptability, better handling multilingual speech recognition tasks, particularly for low-resource languages.

3. The Proposed Approach

In the overall architectural design, we introduce a fundamental model variant: the Byte Pair Encoding based Mixture-of-Experts model, as illustrated in Figure 1. This model adopts a unified encoder-decoder framework. The model is built upon a standard sequence-to-sequence architecture, where the encoder consists of multiple stacked Conformer blocks. We integrate a MoE structure within each Conformer layer composed of multiple FFNs. A multilingual speech sharing router (MSSR) mechanism allows the model to automatically detect the language of the speech input and dynamically route it to the most appropriate expert subnetworks. This design aids in the grouped and multi-scale modeling of language-specific acoustic features. Dynamic

expert selection not only enhances the model's representational capacity but also significantly improves its adaptability in cross-lingual transfer tasks. Given the acoustic similarities across languages, we have designed the TLR method to effectively eliminate language confusion, thereby enhancing the distinguishability of speech.

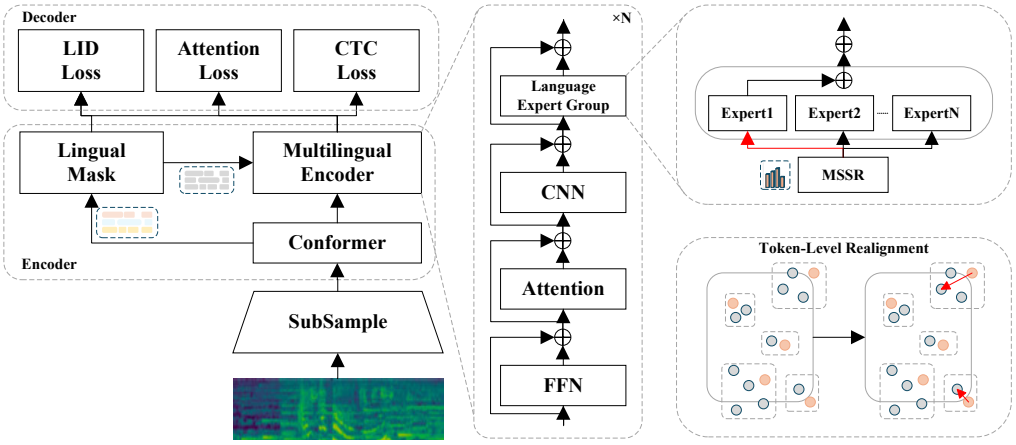


Figure 1: Overview of the model architecture, which includes a shared language route in the encoder, and a dynamic language expert route for each language. The **Token-Level Realignment** refers to a schematic for realigning non-target language tokens during the decoding process.

3.1. Multilingual Speech Sharing Router

To improve the linguistic discrimination of speech in the multilingual speech recognition model, we design a Multilingual Speech Sharing Router (MSSR), which achieves a fine-grained grouping and routing of speech frames by introducing a language category supervisory signal, thus improving the overall system's ability to model mixed multilingual inputs.

Specifically, the input speech features are first modeled using a conformer-based speech encoder, and the hidden states obtained from the last layer of the conformer encoder for the input speech features are denoted as $x = (x_1, x_2, \ldots, x_T)$, $x_t \in \mathbb{R}^D$, where x_t represents the t-th frame, and D is the dimensionality of the hidden feature vector. To train MSSR, we assign each frame a language token sequence $y = (y_1, y_2, \ldots, y_L)$ corresponding to its textual transcription. Each token is replaced with its language label, such that $y_i \in \{1, 2, \ldots, K\}$, where L is the length of the label sequence and K is the number of language categories (e.g., 4 classes: English (EN), Chinese (ZH), Khalkha Mongolian (kMN), and Chahar Mongolian (cMN)). Each frame representation x_t is then fed into a language classifier for prediction: $P = \text{Softmax}(Wx_t + b)$, $P \in \mathbb{R}^{T\times K}$, where $W \in \mathbb{R}^{K\times D}$ and $b \in \mathbb{R}^K$ are learnable parameters. To train the language classifier, we adopt the CTC loss to align the predicted sequence with the ground-truth label sequence. Suppose the batch contains B samples, and the b-th sample has a predicted distribution $y^{(b)}$ with ground-truth label sequence $y^{*(b)}$, then the total loss is defined as (1):

$$\mathcal{L}_{\text{LID}} = -\frac{1}{B}\sum_{b=1}^{B}\log \sum_{\pi\in\mathcal{B}^{-1}(y^{*(b)})} P(\pi|x^{(b)}), \tag{1}$$

where $\mathcal{B}^{-1}(y^{*(b)})$ denotes all valid CTC alignment paths corresponding to $y^{*(b)}$. By exploiting the CTC loss, the model learns to assign the most appropriate language labels to the speech features along the temporal dimension, thus aligning the speech frames with the corresponding language expert groups for multilingual speech recognition and routing. Meanwhile, in each language expert group, the predicted linguistic information is introduced into the MoE routing mechanism as a priori knowledge, which allows the in the language expert group can be freed from the heavy task of speech differentiation to capture the fine-grained features of acoustic features (e.g., accent, dialect) in an unsupervised manner, which not only enhances the model's adaptability to different languages, but also promotes the dynamic collaboration among the experts and improves the overall recognition performance. We use the joint loss function proposed by [24], which integrates the alignment capability of CTC and the modeling capability of Attention, and the total loss is as follows (2):

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

3.2. Multi-scale MoE

In MASR tasks, there are significant differences between languages in terms of speech morphology, rhythmic features, and pronunciation patterns. In order to capture these differences more efficiently and improve the cross-lingual modeling capability, we introduce a multi-layer MoE module in the encoder to achieve multi-scale and multi-granularity modeling of speech features in different languages.

When the input speech signal is encoded into an intermediate representation x, the system calculates the matching probabilities with each language expert through a router based on the language attributes of each frame feature, denoted as $P(x) = (p_1, p_2, \ldots, p_x)$. The top-$k$ function is then applied to select the experts with the highest probabilities to participate in the computation. Each frame feature is assigned to the most appropriate group of experts, while the remaining experts are masked. The final weight vector of all experts is obtained by applying $G(x) = \text{Softmax}(\text{Top}_k(P(x), k))$, and the output of the MoE layer is obtained by weighting the representations of different experts $E_i(x)$ using the weight $G_i(x)$ of the i-th expert (3):

$$y = \sum_{i=1}^{n} G_i(x)E_i(x) \tag{3}$$

where top-k function is given by (4):

$$\text{Top}_k(v_j, k) = \begin{cases} v_j, & v_j \text{ in top } k \text{ elements of } v \\ -\infty, & \text{otherwise} \end{cases} \tag{4}$$

During training, the maximum number of experts is set to $K_{\max}$, which represents a trade-off between model performance and computational efficiency. At each forward pass, an integer k is randomly sampled from a discrete uniform distribution $U[1, K_{\max}]$ to determine the number of active experts for that computation. This strategy encourages each expert to become more independent and robust.

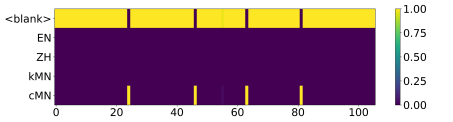
3.3. Token-Level Realignment

Although the Multilingual Speech Sharing Router (MSSR) is able to route speech to the correct language group, as shown in Figure 2b, the model still suffers from frequent lexical-element-level language confusions when dealing with languages with highly similar acoustic

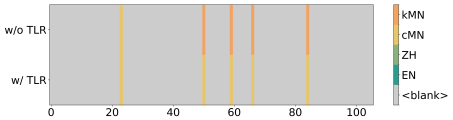
features such as Khalkha Mongolian (kMN) and Chahar Mongolian (cMN), as shown in Figure 2a. To solve this problem, we propose the Token-Level Realignment method, which does not require frame-level linguistic annotations for the input audio at the decoding stage. The acoustic features are mapped to probability distributions over a shared vocabulary, and the beam search strategy is constrained by predefined language boundaries to minimize interference from non-target languages. The recognition errors are effectively corrected and the model's ability to discriminate between similar languages is significantly improved, as shown in Figure 2c.

Text	ᠬᠦᠬᠡᠳ ᠦᠭᠡᠢ ᠡᠮᠦ ᠪᠠᠶᠢᠬᠤ ᠳᠠᠭᠠᠨ ᠲᠠᠭᠠᠷᠠᠨ᠎ᠠ
Reference	heuhed -u'd eimu baihv -d'agan tagaran_a
w/o TLR	heuhed ийм байх даа та
w/ TLR	heuhed -u'd eimu baihv da da

(a) Recognition example



(b) Language router probabilities



(c) TLR correction visualization

Figure 2: Illustration of the language confusion problem and the effect of the proposed Token-Level Realignment (TLR) method.

Specifically, V denotes the shared vocabulary across all languages, and L is the set of supported languages. For each language $l \in L$, we define a language-specific subvocabulary $\mathcal{V}_l \subseteq \mathcal{V}$. To enforce language boundary filtering, we introduce an indicator function as follows (5):

$$\mathbb{I}_{\mathcal{V}_l}(v) = \begin{cases} 0 & \text{if } v \in \mathcal{V}_l \\ -1 & \text{otherwise} \end{cases} \tag{5}$$

At each frame t, instead of directly selecting the top-k tokens with the highest posterior probability from the entire vocabulary, we apply a language-aware score adjustment using a penalty term λ (6):

$$\tilde{v}_t = \arg\max_{v \in \mathcal{V}} \log p_t(v|X) + \lambda \cdot \mathbb{I}_{\mathcal{V}_l}(v) \tag{6}$$

where X represents the input feature sequence and λ is used to control the penalty strength for out-of-language tokens, which is defined over the interval $[0, +\infty)$. In the extreme case $\lambda \to +\infty$, any out-of-language token receives an infinitely large negative score and is thus completely excluded from the beam search.

If the number of valid candidates is less than k, we augment the candidate set by selecting additional high-confidence tokens from the target sub-vocabulary $\mathcal{V}_l$, ensuring the beam size remains unchanged (7):

$$T_t^* = \arg\max_{S \subseteq \mathcal{V}_l,\ |S|=k} \sum_{v \in S} \log p_t(v|X) \tag{7}$$

Finally, decoding is performed within the constrained search space defined by the target language (8):

$$\mathcal{Y}_l := \left\{ Y = (v_1, \ldots, v_T) \in \mathcal{V}_l^T \right\}, \quad \hat{Y} = \arg\max_{Y \in \mathcal{Y}_l} P(Y|X) \tag{8}$$

This strategy ensures that the output sequence strictly adheres to the target language vocabulary, improving recognition robustness in multilingual scenarios involving closely related languages.

4. Experiments and Analysis

4.1. Datasets

Our experiments utilize a multilingual speech recognition dataset comprising four distinct languages: English (EN), Chinese (ZH), Khalkha Mongolian (kMN), and Chahar Mongolian (cMN).

English: The English consists of the 100-hour training subset from LibriSpeech [25], a widely adopted benchmark in speech recognition research. This dataset contains read speech derived from audiobooks, featuring clean recordings with high-quality transcriptions. The development and test data sets are come with its own and are split into "clean" and "other" subsets.

Chinese: For Chinese data, we employ the AISHELL-1 [26] corpus, which offers approximately 178 hours of Mandarin speech recordings. Collected in quiet environments by 400 native speakers, this dataset covers a broad vocabulary and various speaking styles. We used about 150 hours of speech for training, about 18 hours of speech for development, and about 10 hours of speech for test.

Khalkha Mongolian: The Khalkha Mongolian corpus contains of about 290 hours of speech data, a total of 192,711 Khalkha audio, with a sampling frequency of 16 kHz, including 416 speakers from different parts of Mongolia. In our experiments, the dataset was divided into three parts: training set, validation set, and test set, and was not duplicated. The training set is about 278 hours, the validation set is about 8 hours, and the test set is about 4 hours.

Chahar Mongolian: The Chahar Mongolian data consists of about 345 hours of speech sampled at 16 kHz that were collected from 10 domains, including trending, news, education, tourism, and so on [27]. The training set contains 325 hours of speech and involves 889 speakers. The development set contains 4,665 utterances ($\sim$ 8 hours). The test set contains 7,643 utterances ($\sim$ 12 hours).

4.2. Experimental Setup

Basic settings : We conducted multilingual speech recognition experiments based on the WeNet [28] framework, utilizing a hybrid decoding architecture that combines CTC and attention mechanisms. The input acoustic features are 80-dimensional log-Mel filterbanks (FBank), extracted with a frame shift of 10 ms and a window length of 25 ms. We use audio speed-Perturbation [29] and SpecAugment [30] for data augmentation. The audio speed-Perturbation is changing the speed of the audio signal, producing 3 versions of the original signal with speed factors of 0.9, 1.0 and 1.1. The SpecAugment method sets the frequency mask width to 27 and the time mask width to 100. Both frequency masking and time masking are used twice.

The model architecture consists of a 12-layer encoder, where the first 6 layers are standard Conformer blocks and the last 6 layers are MoE layers enhanced with expert routing mechanisms. Each MoE layer includes multiple feedforward sub-networks with a hidden dimension of 2048, supporting both top-2 and dynamic top-k expert selection strategies. We designed two variants

of the model: **Expert4** and **Expert6**. The encoder is configured with a model dimension of $d_{\text{model}} = 256$, number of attention heads $n_{\text{heads}} = 4$, and feedforward dimension $d_{\text{ff}} = 2048$. The activation function used is swish, and the encoder integrates relative positional encoding as well as convolutional modules with a kernel size of 31. The decoder adopts a bidirectional Transformer structure, consisting of 3 forward and 3 backward layers, with 4 attention heads in each direction.

Training and Inference settings : we use Adam algorithm [31] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and Noam learning rate schedule to optimize the models [32]. The maximum learning rate is set to 0.001, with a linear warm-up strategy applied over the first 25,000 steps. A static batch strategy is used during training, with batch_size set to 8. Gradient accumulation is employed with a step size of 8, and gradient clipping with a threshold of 5.0 is applied to improve training stability.

During inference, we employ two inference approaches: (1) a simple greedy approach to report the 1st pass results directly, and (2) a two-stage attention rescoring method that first generates N-best candidate sequences via CTC prefix beam search, and then applies attention-based decoding to utilize more extensive contextual information for more accurate sequence-level scoring. We obtained the recognition error with SCLITE toolkit [33]. The results are performed in terms of character error rate (CER in %) for ZH task, and word error rate (WER in %) for EN, kMN, and cMN tasks.

Table 1: Comparison of MASR performance under different Expert configurations and decoding methods. ✕ indicates that Language Gating inference is not used. ✓ indicates the use of Language Gating inference. All experiments were conducted using top-2 inference settings.

Model	Train	Language Gating	ID	Decoding Method	EN	ZH	kMN	cMN
Expert4	Top2	✕	(1)	CTC	17.31	6.82	12.67	25.46
			(2)	Attention Rescoring	14.67	6.01	9.81	22.49
		✓	(3)	CTC	17.30	6.81	12.48	23.16
			(4)	Attention Rescoring	14.67	6.00	9.60	20.16
	Dynamic	✕	(5)	CTC	14.89	6.38	11.29	23.48
			(6)	Attention Rescoring	12.91	5.75	8.78	20.91
		✓	(7)	CTC	14.94	6.40	11.13	22.37
			(8)	Attention Rescoring	12.93	5.77	8.61	19.75
Expert6	Top2	✕	(9)	CTC	15.34	6.60	11.24	23.93
			(10)	Attention Rescoring	13.21	5.88	8.59	21.15
		✓	(11)	CTC	15.34	6.39	11.23	23.07
			(12)	Attention Rescoring	13.21	5.81	8.58	20.15
	Dynamic	✕	(13)	CTC	14.86	6.02	9.98	23.32
			(14)	Attention Rescoring	12.88	5.49	7.76	20.81
		✓	(15)	CTC	14.90	6.02	9.87	22.25
			(16)	Attention Rescoring	**12.88**	**5.49**	**7.64**	**19.61**

4.3. Experimental Results and Analysis

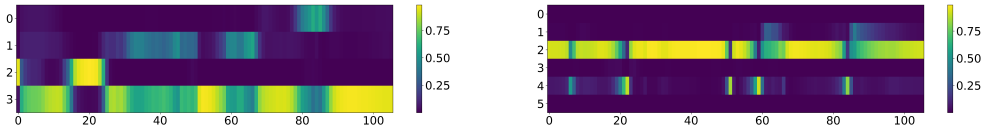
4.3.1. Analysis of Evaluation Results under Different MoE Configurations

Table 1 presents the performance of different MoE configurations on MASR tasks across four languages: EN, ZH, kMN, and cMN.

Effect of Top-k Strategy : Using top-2 inference with the same number of experts and model structure, the dynamic Top-k routing strategy demonstrates superior performance compared to the fixed Top-2 approach (e.g., 1—4 and 5—8, or 9—12 and 13—16). In the Expert4 architecture, (5) achieves results of 14.89%, 6.38%, 11.29%, and 23.48% for EN, ZH, kMN, and cMN. Compared to (1), the recognition accuracy was improved by 13.98%, 6.45%, 10.89%, and 7.78%, respectively. This performance gain benefits from the flexibility of dynamic routing, which can adaptively determine the number and combination of activated experts, thereby enhancing the model's ability to capture complex speech patterns and better accommodate cross-lingual variations. To further analyze computational trade-offs, we additionally compared top-1 and top-2 inference on the Expert6-Dynamic model. As summarized in Table 2, we report GFLOPs, real-time factor (RTF) and latency: the increased computational overhead of the top-2 strategy relative to top-1 underscores the inherent trade-off between achieving higher performance and maintaining computational efficiency.

Effect of the Number of Experts : Introducing a larger number of experts (e.g., 1—8 and 9—16) significantly improves the model's ability to capture multilingual acoustic details. By increasing the number of expert sub-networks, the Expert6 architecture outperforms Expert4 in both language modeling and recognition accuracy. In the comparison between (4) and (12), the overall result decreases from (EN: 14.67%, ZH: 6.00%, kMN: 9.60%, and cMN: 20.16%) to (EN: 13.21%, ZH: 5.81%, kMN: 8.58%, and cMN: 20.15%).

We visualized the internal routing behavior of the model for different numbers of experts for the example in Figure 2a. In the **Expert4** model (shown in Figure 3a), we observe that its routing pattern is relatively centralized. A small number of experts (e.g., expert3) are continuously activated with high probability, and it is not only responsible for recognizing a small number of real characters, but also takes on the task of outputting a large number of `<blank>`. In contrast, the routing behavior of the **Expert6** model (shown in Figure 3b) exhibits a clearer division of specialization, where expert2 is responsible for outputting high-probability `<blank>`, and expert4 is activated at different time steps to focus on recognizing and distinguishing more challenging real speech characters. A larger number of experts can accurately capture and model more complex acoustic details, improving the model's generalization ability and cross-linguistic adaptability.



(a) Expert4 (b) Expert6

Figure 3: Effect of the number of experts

Effect of Decoding Method : Across all experimental settings, the Attention Rescoring method consistently outperforms the traditional CTC decoding approach. Comparing the results of (15) (EN: 14.90%, ZH: 6.02%, kMN: 9.87%, cMN: 22.25%) with those of (16) **(EN: 12.88%, ZH: 5.49%, kMN: 7.64%, cMN: 19.61%)**, we observe a significant improvement in recognition

performance across all four languages(**13.55%, 8.80%, 22.59%, 11.87%**), achieving the best capability of the model. This is primarily attributed to the Attention mechanism's ability to incorporate contextual information and perform global alignment, which compensates for CTC's limitations in local modeling and effectively reduces acoustic confusion and language boundary ambiguity.

Table 2: Comparison of computational efficiency for the Expert6-Dynamic model under different top-*k* inference settings, reporting GFLOPs, real-time factor (RTF), and latency.

Model	Inference	ID	Decoding Method	GFLOPs	RTF	Latency
Expert6-Dynamic	top-1	(1)	CTC	3.16	0.0052	29.9
		(2)	Attention Rescoring	3.16	0.0294	167.21
	top-2	(3)	CTC	3.46	0.0159	89.91
		(4)	Attention Rescoring	3.46	0.0541	307.46

4.3.2. Analysis of Cross-Lingual Stability and Streaming Capability

Cross-lingual Stability and Robustness Evaluation : We randomly sampled 500, 1000, and 2000 utterances for each of the four languages to construct mixed speech inputs, and compared the performance of two MoE models under different inference strategies. As shown in Table 3, increasing the number of sampled utterances per language from 500 to 2000 significantly improved overall recognition performance, especially under the Attention Rescoring strategy. In the Expert6-Dynamic model, the overall results decreased from 11.99% (8) to 11.74% (12), indicating that richer training data enables the model to better learn cross-lingual acoustic features, thereby enhancing its multilingual modeling capabilities. Furthermore, with the same amount of data, case (12) with 11.74% outperformed case (6) with 11.95%, further confirming that increasing the number of experts positively contributes to improving recognition accuracy in multilingual scenarios.

Table 3: MASR performance in different random sampling conditions

Model	Mixed Speech	ID	Decoding Methond	EN	ZH	kMN	cMN	Overall
Expert4-Dynamic	shuffle 500	(1)	CTC	14.86	6.55	11.90	23.27	14.16
		(2)	Attention Rescoring	12.58	5.92	9.37	20.70	12.18
	shuffle 1000	(3)	CTC	14.91	6.38	10.96	23.87	14.07
		(4)	Attention Rescoring	12.93	5.75	8.46	21.26	12.20
	shuffle 2000	(5)	CTC	14.78	6.07	11.32	23.02	13.79
		(6)	Attention Rescoring	12.82	5.54	8.80	20.43	11.95
Expert6-Dynamic	shuffle 500	(7)	CTC	14.73	6.23	10.65	22.94	13.74
		(8)	Attention Rescoring	12.73	5.75	8.52	20.46	**11.99**
	shuffle 1000	(9)	CTC	14.96	5.97	9.43	23.09	13.54
		(10)	Attention Rescoring	13.02	5.48	7.41	20.53	**11.81**
	shuffle 2000	(11)	CTC	14.81	5.68	10.36	22.30	13.37
		(12)	Attention Rescoring	13.01	5.21	8.20	20.05	**11.74**

Multilingual Streaming ASR Capability Analysis : Table 4 presents the performance of the Expert6-Dynamic model under different chunk sizes in MASR tasks, comparing streaming and non-streaming inference modes. In the same model structure, the non-streaming configuration consistently achieved the best recognition performance. When the chunk size was set to 16 (2), the performance dropped to (EN: 14.31%, ZH: 6.06%, kMN: 9.10%, and cMN: 21.56%). This suggests that the streaming mode, due to segmenting inputs into limited windows, leads to context truncation, thereby affecting recognition accuracy. As the chunk size was further reduced to 8 (3), performance worsened to (EN: 15.02%, ZH: 6.44%, kMN: 9.65%, and cMN: 22.48%), which marks the worst result among all tested configurations. These findings indicate that in streaming ASR tasks, the chunk size setting has a significant impact on recognition performance, and overly small windows severely limit the model's capacity to model context.

Table 4: The performance of different chunk sizes in streaming speech recognition across various languages.

Model	ID	Chunk Size	EN	ZH	kMN	cMN
Expert6-Dynamic	(1)	Non-Streaming	12.88	5.49	7.64	19.61
	(2)	16	14.31	6.06	9.10	21.56
	(3)	8	15.02	6.44	9.65	22.48

4.4. Discussion

Due to the high acoustic similarity between cMN and kMN, the model tends to confuse the two languages when recognizing cMN, resulting in a decline in recognition performance. Figure 4 presents the reference and recognized transcripts under different model configurations. We first compare the performance of the model without language expert routing (w/o Expert) and with language-specific experts (w/ Expert). In Cases 1, 2, and 3, the model mistakenly recognizes cMN as kMN, while the inclusion of specific language experts effectively resolves this issue. However, in Cases 4, 5, and 6, recognition errors still occur even with language expert routing, indicating that expert selection alone is insufficient to fully eliminate language confusion.

To address this problem, we design an ablation study that incorporates both the language-specific expert module and a TLR strategy. As shown in Table 5, the introduction of language-specific experts significantly enhances the model's discriminative capability, reducing the WER to 19.61%(2). Furthermore, by integrating the TLR method, the model achieves the best recognition performance on Chahar Mongolian, reducing the WER to **19.30%**(3) and an improvement of **7.26%** over 20.81%(1). These results demonstrate that fine-grained TLR is effective in mitigating recognition errors caused by phonetic overlap across languages.

Table 5: Ablation study of Expert6-Dynamic with expert selection and TLR on cMN.

Configuration	ID	CTC	Attention Rescoring
Expert6-Dynamic(w/o Expert)	(1)	23.32	20.81
w/ Expert	(2)	22.25	19.61
w/ Token-Level Realignment	(3)	21.81	**19.30**

1	Text	ᠨᠢᠭᠡ ᠪᠤᠳᠤᠭᠤᠨ ᠪᠢᠪᠠ ᠬᠣᠭᠶᠠᠷ ᠲᠤᠭᠠᠯ ᠬᠤᠳᠠᠯᠳᠤᠨ ᠠ᠋
	Reference	nige budugun bvh_a hqyar tvgvl -tai hvdaldvn_a
	w/o Expert	nige budugun bvh_a хоёр тугалтай худалдана
	w/ Expert	nige budugun bvh_a hqyar tvgvl -tai hvdaldvn_a
2	Text	ᠪᠢ ᠪᠠᠷᠢᠮᠵᠢᠶ᠋ᠠ ᠪᠠᠷ ᠶᠠᠷᠢᠵᠤ ᠴᠢᠳᠠᠬᠤ ᠦᠭᠡᠢ ᠪᠠᠢᠨ᠋ᠠ ᠃
	Reference	bi barimjiy_a -bar yarijv cidahv ugei bain_a
	w/o Expert	bi баримжаа_а -bar yarijv cidahv ugei bain_a
	w/ Expert	bi barimjiy_a -bar yarijv cidahv ugei bain_a
3	Text	ᠨᠢᠭᠡ ᠪᠤᠳᠤᠭᠤᠨ ᠪᠢᠪᠠ ᠬᠣᠭᠶᠠᠷ ᠲᠤᠭᠠᠯ ᠬᠤᠳᠠᠯᠳᠤᠨ ᠠ᠋
	Reference	nige budugun bvh_a hqyar tvgvl -tai hvdaldvn_a
	w/o Expert	nige budugun bvh_a хоёр тугалтай худалдана
	w/ Expert	nige budugun bvh_a hqyar tvgvl -tai hvdaldvn_a
4	Text	ᠪᠢ ᠰᠠᠶᠢ ᠬᠪᠰᠢᠭᠪᠨ᠋ ᠳᠡᠭᠡᠷ᠋ᠡ ᠢᠷᠡᠭᠰᠡᠨ ᠪᠠᠢᠨ᠋ᠠ ᠃
	Reference	bi sayi hvsigvn deger_e iregsen bain_a
	w/o Expert	би sayi hvsigvnyyн deger_e iregsen байна_а
	w/ Expert	bi sayi hvsigvn deger_e iregsen байна_а
	w/ Token-Level Realignment	bi sayi hvsigvn deger_e iregsen bain_a
5	Text	ᠵᠠ ᠠ᠋ ᠨᠠᠷ ᠬᠣᠳᠤ ᠶᠠᠭᠤ ᠬᠢᠵᠤ ᠪᠠᠢᠨ᠋ᠠ ᠃ ᠪᠢ ᠬᠣᠳᠤ ᠵᠤᠮᠤᠭᠠᠷ ᠬᠤᠭᠤᠷᠠᠭ᠋ ᠪᠣᠯᠴᠢᠬᠠᠭᠰᠠᠨ ᠰᠢᠦ
	Reference	ja ta nar qdq yagv hiju bain_a bi qdq yqsqtai mvndag bqlcihagsan siu
	w/o Expert	ja та нар одоо юу hiju байна_а bi одоо ёстой мундаг bqlcihagsan шүү
	w/ Expert	ja ta нар одоо юу hiju bain_a bi qdq yqsqtai mvndar bqlcihagsan siu
	w/ Token-Level Realignment	ja ta nar qdq yagv hiju bain_a bi qdq yqsqtai mvndag bqlcihagsan siu
6	Text	ᠶᠠᠭᠤ ᠤᠤ ᠳᠠᠰᠢᠯ ᠬᠢᠬᠦ ᠦᠭᠡᠢ ᠤᠤ
	Reference	gvw_a -tai dashal hihu ugei uu
	w/o Expert	гоотай dashal хийх ugei uu
	w/ Expert	гоо_а -tai dashal hihu ugei uu
	w/ Token-Level Realignment	gvw_a -tai dashal hihu ugei uu

Figure 4: The impact of different inference methods on the recognition accuracy of cMN.

5. Conclusion and Future Work

In this work, we introduce a flexible multilingual speech recognition MoE model, which incorporates a dynamic top-*k* expert routing strategy within the MoE framework, aiming to achieve a better trade-off between computational efficiency and recognition performance. We leverage a multilingual speech sharing router to facilitate routing across multiple languages and thoroughly investigate the impact of varying expert counts and top-*k* strategies on the model's performance. However, issues such as language confusion due to phonetic similarities between languages remain a challenge. To address this, we propose a *Token-Level Realignment* method, which precisely aligns input frames with the corresponding language expert modules. The integration of language-specific experts has significantly improved recognition accuracy, and the proposed

Token-Level Realignment strategy further enhances the system's ability to distinguish between similar languages. In the future, we plan to explore more fine-grained modeling techniques for low-resource languages to improve speech recognition capabilities.

Acknowledge

This work is supported by the National Natural Science Foundation of China (No.62366037), Natural Science Foundation of Inner Mongolia (2025MS06001), Science and Technology Program of Inner Mongolia Autonomous Region (2025KYPT0041, 2025KYPT0064).

References

[1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[2] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

[3] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

[4] AFM Saif, Lisha Chen, Xiaodong Cui, Songtao Lu, Brian Kingsbury, and Tianyi Chen. M2asr: Multilingual multi-task automatic speech recognition via multi-objective optimization. In *Interspeech*, volume 2024, pages 1240–1244, 2024.

[5] Thomas Palmeira Ferraz, Marcely Zanon Boito, Caroline Brun, and Vassilina Nikoulina. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10716–10720, 2024.

[6] Yosuke Kashiwagi, Hayato Futami, Emiru Tsunoo, Siddhant Arora, and Shinji Watanabe. Rapid language adaptation for multilingual e2e speech recognition using encoder prompting. In *Proc. Interspeech 2024*, pages 2900–2904, 2024.

[7] Wei Liu, Jingyong Hou, Dong Yang, Muyong Cao, and Tan Lee. A parameter-efficient language extension framework for multilingual asr. In *Proc. Interspeech 2024*, pages 3929–3933, 2024.

[8] Peikun Chen, Fan Yu, Yuhao Liang, Hongfei Xue, Xucheng Wan, Naijun Zheng, Huan Zhou, and Lei Xie. Bamoe: Boundary-aware mixture-of-experts adapter for code-switching speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE, 2023.

[9] Yihan Wu, Yifan Peng, Yichen Lu, Xuankai Chang, Ruihua Song, and Shinji Watanabe. Robust audiovisual speech recognition models with mixture-of-experts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 43–48. IEEE, 2024.

[10] Wenxuan Wang, Guodong Ma, Yuke Li, and Binbin Du. Language-routing mixture of experts for multilingual and code-switching speech recognition. In *Proc. Interspeech 2023*, pages 1389–1393, 2023.

[11] Guodong Ma, Wenxuan Wang, Lifeng Zhou, Yuting Yang, Yuke Li, and Binbin Du. Blr-moe: Boosted language-routing mixture of experts for domain-robust multilingual e2e asr. *arXiv preprint arXiv:2501.12602*, 2025.

[12] Hukai Huang, Shenghui Lu, Yahui Shan, He Qu, Fengrun Zhang, Wenhao Guan, Qingyang Hong, and Lin Li. Dynamic language group-based moe: Enhancing code-switching speech recognition with hierarchical routing. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[13] Zhao You, Shulin Feng, Dan Su, and Dong Yu. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. In *Proc. Interspeech 2021*, pages 2077–2081, 2021.

[14] Zhao You, Shulin Feng, Dan Su, and Dong Yu. 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, 2022.

[15] Zhao You, Shulin Feng, Dan Su, and Dong Yu. 3m: Multi-loss, multi-path and multi-level neural networks for speech recognition. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 170–174. IEEE, 2022.

[16] Yoohwan Kwon and Soo-Whan Chung. 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, 2023.

[17] Songjun Cao, Xiong Wang, Yike Zhang, Xiaoming Zhang, and Long Ma. M-moe: Mixture of mixture-of-expert model for ctc-based streaming multilingual asr. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[18] Anjuli Kannan, Arindrima Datta, Tara N Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. Large-scale multilingual speech recognition with a streaming end-to-end model. In *Proc. Interspeech 2019*, pages 2130–2134, 2019.

[19] Bo Li, Dongseong Hwang, Zhouyuan Huo, Junwen Bai, Guru Prakash, Tara N Sainath, Khe Chai Sim, Yu Zhang, Wei Han, Trevor Strohman, et al. Efficient domain adaptation for speech foundation models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[20] Ping Feng, Xin Zhang, Jian Zhao, Yingying Wang, and Biao Huang. Relation extraction based on prompt information and feature reuse. *Data Intelligence*, 5(3):817–833, 2023.

[21] Meirong Bao and Qinggeletu De. Construction of mongolian near-synonymous compound qualitative adjective sets. *Data Intelligence*, 7(1):221–236, 2025.

[22] Yingli Shen, Lingzhi Yu, Xiaoke Qi, and Xiaobing Zhao. Chinese-mongolian bilingual dictionary dataset. *Data Intelligence*, 2025.

[23] Ke Hu, Bo Li, Tara Sainath, Yu Zhang, and Françoise Beaufays. Mixture-of-expert conformer for streaming multilingual asr. In *Proc. Interspeech 2023*, pages 3327–3331, 2023.

[24] Takaaki Hori, Shinji Watanabe, and John R Hershey. 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, 2017.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[26] He Bu, Jun Du, Xingyu Na, Jiayu Wu, and Ming Li Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *Proceedings of the 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, 2017.

[27] Yihao Wu, Yonghe Wang, Hui Zhang, Feilong Bao, and Guanglai Gao. Mnasr: A free speech corpus for mongolian speech recognition and accompanied baselines. In *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6. IEEE, 2022.

[28] Binbin Zhang, Di Wu, Zhendong Peng, Xingchen Song, Zhuoyuan Yao, Hang Lv, Lei Xie, Chao Yang, Fuping Pan, and Jianwei Niu. Wenet 2.0: More productive end-to-end speech recognition toolkit. In *Proc. Interspeech 2022*, pages 1661–1665, 2022.

[29] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *Interspeech*, pages 3586–3589, 2015.

[30] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*, pages 2613–2617, 2019.

[31] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *ICONIP*, pages 5998–6008, 2017.

[33] Jon Fiscus. Sclite scoring package version 1.5. *US National Institute of Standard Technology (NIST), URL http://www. itl. nist. gov/iaui/894.01/tools*, 1998.

Author Biographies

Junchen Wang received his B.E. degree from Inner Mongolia University of Technology in 2024, and enrolled in the School of Computer Science at Inner Mongolia University in 2024, where he is currently working on his M.S. degree in Computer Technology, with research interests that include Speech Recognition.
ORCID: 0009-0002-7752-2037, Email: 32409144@mail.imu.edu.cn

Yonghe Wang received the Ph.D. degree from Inner Mongolia University (China) in 2022. He is currently a lecturer at the College of Computer Science, Inner Mongolia University. His research interests include speech recognition, speech translation, speech signal processing, natural language processing, and Mongolian intelligent information processing.
ORCID: 0000-0003-1647-1539, Email: cswyh@imu.edu.cn

Feilong Bao received the Ph.D. in Engineering and is a professor and Ph.D. supervisor at the School of Computer Science, Inner Mongolia University. His research work centers on Mongolian information technology, computational linguistics, and Mongolian information systems.
ORCID: 0000-0001-7312-1629, Email: csfeilong@imu.edu.cn

Guanglai Gao is a professor and Ph.D. supervisor at the College of Computer Science, Inner Mongolia University. His research work centers on artificial intelligence, pattern recognition, and intelligent information processing in Mongolian.
ORCID: 0009-0005-5513-1192, Email: csggl@imu.edu.cn