Language-Aware Routing in Mixtures of Experts for Low-Resource Automatic Speech Recognition

Jianan Chen*, Xiaoxue Gao†, Tatsuya Kawahara*, Nancy F. Chen†
* School of Informatics, Kyoto University, Kyoto, Japan
E-mail: chen.jianan.32d@st.kyoto-u.ac.jp, kawahara@i.kyoto-u.ac.jp
† Agency for Science, Technology and Research (A*STAR), Singapore
E-mail: {Gao_Xiaoxue, nancy_chen}@a-star.edu.sg

***Abstract*—Automatic speech recognition (ASR) for low-resource languages remains challenging due to limited annotated data and linguistic diversity. Unlike high-resource languages, low-resource languages often lack sufficient training data to train ASR models from scratch. Parameter-efficient fine-tuning of pre-trained speech models has become a widely adopted solution. Among them, Low-Rank Adaptation (LoRA) enables efficient adaptation by introducing only a small number of trainable parameters while preserving the knowledge in pre-trained models. In this work, we investigate the Mixtures of LoRA Experts (MoLE) framework on four Indic languages: Hindi, Bengali, Marathi, and Tamil. We compare traditional routing with language-guided and language-family-guided routing, and further combine the traditional and the model-predicted language-guided routing in a hybrid strategy. Experimental results show that top-1 routing generally outperforms top-2 routing and produces expert specialization towards individual languages. In addition, language-aware routing performs comparably to or better than traditional routing. The hybrid strategy achieves the best overall result, with a character error rate of 9.20%. Expert usage analysis further suggests that language-dependent specialization is more beneficial than uniform expert usage.**

***Index Terms*—speech recognition, mixture of experts, low-resource languages**

I. Introduction

Automatic speech recognition (ASR) for low-resource languages remains challenging due to limited labeled data and linguistic diversity [1], [2], [3], [4], [5]. Unlike high-resource languages, low-resource languages often lack sufficient labeled training data to effectively train ASR models from scratch. A widely adopted solution is to transfer knowledge encoded in multilingual pre-trained speech models to low-resource languages [6], [7], [8].

Recently, self-supervised speech representation learning has become one of the most successful approaches for multilingual ASR. Models such as wav2vec 2.0 learn universal acoustic representations from large-scale unlabeled speech corpora and have shown strong cross-lingual transfer capabilities [9], [10]. Nevertheless, adapting pre-trained multilingual models to low-resource target languages remains an important challenge. While multilingual training enables cross-lingual knowledge transfer, it can also lead to negative transfer, where the performance of some low-resource languages degrades [11].

Meanwhile, full-parameter fine-tuning of large pre-trained speech models is computationally intensive and often impractical in low-resource settings. As a result, parameter-efficient fine-tuning methods have attracted wide attention [12], [13], [14]. Among them, Low-Rank Adaptation (LoRA) [12] significantly reduces the number of trainable parameters, thus becoming one of the most popular methods. Previous studies have shown that LoRA can effectively adapt multilingual speech models to low-resource ASR tasks while maintaining competitive recognition performance [15], [16], [17].

However, a single shared LoRA adapter may not be sufficient to model the diverse acoustic and linguistic characteristics of multiple low-resource languages. Here, Mixture-of-Experts (MoE) architectures provide a promising alternative. MoE dynamically routes different inputs to different experts, allowing specialization among model components [18], [19], [20], [21]. Combining MoE with LoRA yields a Mixture-of-LoRA-Experts (MoLE) framework, in which multiple LoRA adapters act as experts and only some of the experts are activated by a routing mechanism (gating function) [22], [23].

Although MoLE has recently been applied to a variety of speech and language processing tasks [24], [25], [26], its behavior in multilingual low-resource ASR remains insufficiently understood. It is unclear whether expert routing should be learned entirely from data or guided by prior linguistic knowledge, such as language or language family.

In this work, we investigate the MoLE framework in multilingual low-resource ASR using XLSR-53 as the backbone model. We conduct a case study on four Indic languages: Hindi (hi), Bengali (bn), Marathi (mr), and Tamil (ta). Besides traditional routing, we explore language-guided, language-family-guided, and hybrid routing strategies to examine whether linguistic knowledge can improve expert specialization and recognition performance.

The contributions of this work include:

1) We investigate language-guided, language-family-guided, and hybrid routing strategies for MoLE-based multilingual low-resource ASR, demonstrating that incorporating linguistic information improves expert routing and recognition performance.
2) We analyze expert usage patterns and show that stronger expert specialization is associated with better ASR performance, while a hybrid routing strategy yields the best overall results.

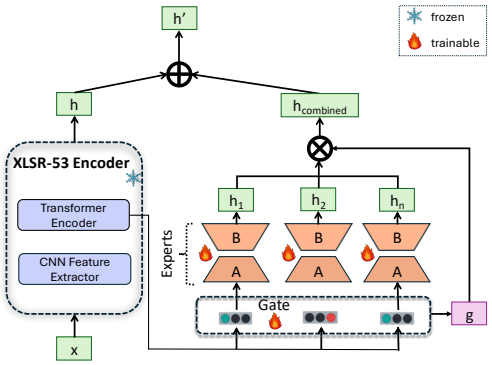


Fig. 1: Architecture of the Mixture of LoRA Expert (MoLE) framework. A pre-trained speech encoder (XLSR-53) is fine-tuned with multiple LoRA experts. A gate dynamically routes the input x to a set of experts. The weighted output $h_{combined}$ of the selected experts is then fused with the base model output h to produce the final representation h'.

II. Proposed Method

A. Mixture of LoRA Expert

To capture both language-specific and universal acoustic features across languages, we employ a Mixture of LoRA Expert (MoLE) framework as illustrated in Fig. 1 to fine-tune pre-trained speech encoders. Instead of fine-tuning all parameters, MoLE uses multiple lightweight LoRA adapters as experts, enabling language-aware adaptation while maintaining parameter efficiency.

LoRA is a fine-tuning technique that adapts large pre-trained models without updating all parameters. Instead of directly modifying the original weights, LoRA introduces two trainable low-rank matrices while keeping the pre-trained parameters frozen. Given an input representation x, let $h \in \mathbb{R}^d$ denote the hidden state produced by a layer of the pre-trained model. LoRA augments the hidden state with a low-rank update:

$$h' = h + \frac{\alpha}{r} BAx, \tag{1}$$

where $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ are trainable low-rank matrices, $r \ll d$ is the LoRA rank, and α is a scaling factor that controls the contribution of the LoRA update.

To further increase model capacity for multilingual ASR, multiple LoRA adapters are used as experts inside a sparse Mixture-of-Experts (MoE) mechanism. Instead of applying a single LoRA adapter to all languages, a gating network dynamically selects some of the experts based on the input representation or hidden states. Traditionally, the routing weight assigned to expert i is defined as:

$$g_i = \begin{cases} \dfrac{\exp(w_i^\top x)}{\sum_{j \in \mathcal{T}_k(w_j^\top x)} \exp(w_j^\top x)}, & i \in \mathcal{T}_k(w_i^\top x) \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathcal{T}_k(\cdot)$ denotes the set of top-k selected experts (activated experts) and w_i represents the routing parameters associated with expert i.

The final MoLE output is obtained by combining the outputs of all selected experts:

$$h' = h + \frac{\alpha}{r} \sum_{i=1}^{n} g_i B_i A_i x, \tag{3}$$

where n denotes the total number of experts.

This routing strategy is referred to as *traditional routing*, where expert assignments are learned directly without explicit linguistic supervision. Although effective, the resulting expert specialization is often difficult to interpret.

B. Ground-Truth Language-Guided Routing

We hypothesize that explicit linguistic knowledge can improve expert specialization in multilingual ASR. Instead of learning routing decisions solely from hidden representations, we investigate routing strategies based on ground-truth language information.

For language-guided routing, each language is assigned to a specific expert. Given an utterance belonging to language l, the routing weight is determined by a fixed language-to-expert mapping function $m(\cdot)$:

$$g_i = \begin{cases} 1, & i = m(l), \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

All utterances from the same language are then consistently routed to the same expert. This design explicitly encourages language-specific expert specialization and provides interpretable routing decisions.

We further investigate language-family-guided routing. Languages belonging to the same language family are grouped together and routed to a shared expert. Similarly, the routing weight is then determined by a language-family-to-expert mapping.

Compared with traditional learned routing, these strategies explicitly incorporate linguistic knowledge into the routing mechanism.

C. Model-Predicted Language Routing

Although ground-truth language labels provide an upper bound for language-aware routing, such information is generally unavailable during inference. To address this problem, we further investigate routing strategies based on automatically predicted language information, as illustrated in Fig. 2.

Given an input speech utterance x, we first estimate the posterior probabilities over the target languages:

$$p^{lang} = LID(x) \tag{5}$$

where $p^{lang} \in \mathbb{R}^N$ denotes the predicted language probability distribution, and N is the number of target languages.

The predicted language probabilities are then converted into routing weights similar to Eq. (2):

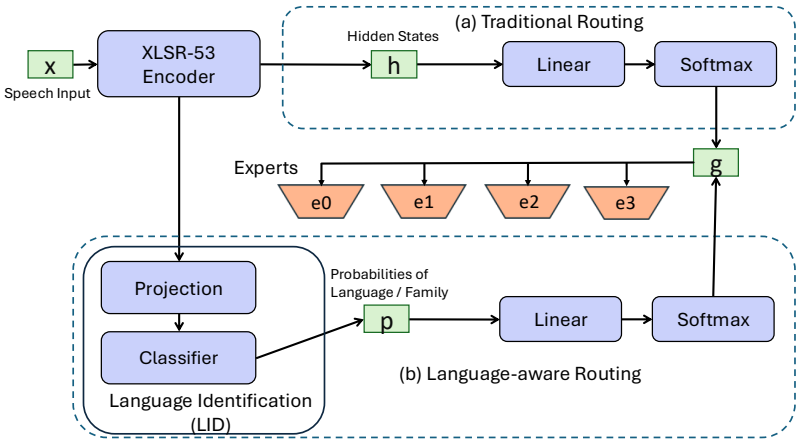


Fig. 2: Comparison of the routing strategies. (a) Traditional routing learns directly from the hidden states of the XLSR-53 encoder. (b) Language-aware routing incorporates language information to guide expert selection.

$$g_i = \begin{cases} \dfrac{\exp(w_i^\top p_i^{lang})}{\sum_{j \in \mathcal{T}_k(w_j^\top x)} \exp(w_j^\top p_j^{lang})}, & i \in \mathcal{T}_k(w_i^\top p_i^{lang}) \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

For family-guided routing, the probability of a language family can be calculated by summing the probabilities of the languages belonging to the family. Finally, g_i is normalized with the number of activated experts k.

We further investigate a hybrid routing strategy that combines the traditional gating network with the language probabilities generated by the LID model. Let $\mathbf{g}^{\text{trad}}$ denote the routing weights produced by the traditional gating network and $\mathbf{g}^{\text{LID}}$ denote the routing weights derived from the LID module. The final hybrid routing weights are computed as

$$\mathbf{g}^{\text{hybrid}} = \lambda \mathbf{g}^{\text{LID}} + (1 - \lambda)\mathbf{g}^{\text{trad}} \tag{7}$$

where $\lambda \in [0, 1]$ controls the relative contribution of the two routing mechanisms.

III. Experimental Setups

A. Dataset

We conduct experiments on four Indic languages from two language families: Indo-Aryan (Hindi, Bengali, and Marathi) and Dravidian (Tamil). The selected languages exhibit both shared and distinct linguistic characteristics. Hindi and Marathi use the Devanagari script, but Bengali and Tamil use different scripts. This setting evaluates whether language-aware and language-family-aware expert routing can facilitate multilingual adaptation in low-resource ASR.

We use the CommonVoice 16.1 corpus [27]. The `train` split is used for training, the `dev` split for development, and the `test` split for evaluation. Detailed statistics of the four languages are shown in Table I.

TABLE I: Durations in hours of train, dev, and test splits of four languages.

Family	Language	L.ID.	Train	Dev	Test
Indo-Aryan	Hindi	hi	10.0	3.1	4.4
	Bengali	bn	16.7	16.2	16.6
	Marathi	mr	7.5	1.7	2.5
Dravidian	Tamil	ta	14.9	16.3	12.7

B. Backbone Model

The backbone model used in this study is XLSR-53 [10], a multilingual self-supervised speech representation model pre-trained on 53 languages. A CTC prediction head is appended for the ASR task.

We adopt the PEFT framework [28] and inject LoRA adapters into both the attention and feed-forward layers of XLSR-53. To determine an effective LoRA configuration for multilingual low-resource ASR, we perform a grid search over the LoRA rank $r \in 16, 32, 64, 128$ and scaling factor α, with α/r ranging from 0.125 to 4. We select $r = 32$ and $\alpha = 128$ for all subsequent MoLE experiments, as these are the best parameters for constructing the LoRA-only baseline.

Character Error Rate (CER) and Word Error Rate (WER) are used as evaluation metrics, where lower values indicate better recognition performance. The overall CER and WER are computed by concatenating the test sets of all languages and calculating the error rate over the merged hypotheses and references.

C. Routing Strategies

To investigate whether linguistic knowledge can improve expert routing, we compare four routing strategies: 1) **Traditional Routing** (T1-T2): Expert routing is determined by a learnable gating network in Section II-A. 2) **Ground-Truth Routing** (L1, L3, F1, F3): Expert routing is determined using manually provided language or language-family labels, as described in Section II-B. 3) **Predicted Routing** (L2, L4, F2): Expert routing is determined using language information

TABLE II: Comparison of different routing strategies under different expert configurations.

ID	Routing	n	k	hi		bn		mr		ta		Overall Avg.	
				WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓	WER↓	CER↓
Traditional Routing													
T0	LoRA-only	-	-	62.23	20.80	69.85	21.36	75.24	21.27	85.85	25.52	72.30	22.27
T1	Traditional	4	2	22.27	8.16	29.36	8.39	27.09	6.39	58.02	18.16	32.37	10.19
T2	Traditional	4	1	22.22	8.06	29.25	8.41	27.35	6.48	57.18	18.01	32.24	10.16
Language-Guided Routing													
L1	Ground Truth LID	4	2	22.10	8.07	29.06	8.37	27.35	6.47	57.73	18.01	32.26	10.15
L2	Predicted LID	4	2	22.43	8.13	28.91	8.35	27.47	6.42	57.57	18.07	32.30	10.16
L3	Ground Truth LID	4	1	20.89	7.43	28.22	8.11	25.40	5.97	**56.44**	**17.58**	30.95	9.71
L4	Predicted LID	4	1	21.52	7.60	28.95	8.25	24.97	5.82	57.41	17.69	31.41	9.77
Language-Family-Guided Routing													
F1	Ground Truth Family	4	2	21.75	7.77	28.45	8.12	26.71	6.36	57.12	17.85	31.71	9.95
F2	Predicted Family	4	2	22.24	8.07	29.16	8.41	27.97	6.57	57.73	18.00	32.48	10.18
F3	Ground Truth Family	2	1	20.28	7.20	28.35	8.02	25.09	5.98	57.28	17.91	30.90	9.72
F4	Predicted Family	2	1	20.33	7.25	28.21	7.99	25.63	6.15	57.68	18.03	30.99	9.77
Hybrid Strategies ($\lambda = 0.1$)													
C1	Predicted LID + Traditional Gate	4	2	21.12	7.57	28.73	8.17	26.21	6.12	57.16	17.65	31.52	9.81
C2	Predicted LID + Traditional Gate	4	1	**18.84**	**6.66**	**27.66**	**7.75**	**21.45**	**5.00**	57.00	17.63	**29.35**	**9.20**

predicted by a language identification (LID) model in Section II-C. 4) **Hybrid Routing** (C1, C2): The routing weights generated by the traditional gating network are combined with the language probabilities predicted by the LID model.

For language-guided routing, we evaluate both top-1 ($k = 1$) and top-2 ($k = 2$) expert selection. Under top-1 routing, Hindi, Bengali, Marathi, and Tamil are assigned to experts 0, 1, 2, and 3. Under top-2 routing, each language is assigned to two experts. Hindi, Bengali, Marathi, and Tamil are assigned to experts $\{0, 3\}$, $\{1, 3\}$, $\{0, 2\}$, and $\{1, 2\}$. These overlapping assignments are designed to encourage linguistic knowledge sharing among languages. For family-guided routing, experts are assigned by the Indo-Aryan and Dravidian language families. Under top-1 routing, we use two family-specialized experts. Each language family is assigned to a dedicated expert. Under top-2 routing, we use four experts. Each family is assigned to two experts without overlapping.

D. Training Details

All experiments are conducted on NVIDIA RTX A6000 GPUs with 48 GB of memory. During MoLE training, the parameters of the XLSR-53 backbone are frozen, while the LoRA experts, routing modules, LID modules, and CTC prediction head are trainable. We set LoRA rank $r = 32$, and LoRA scaling factor $\alpha = 128$. LoRA adapters are inserted into both attention and feed-forward layers, and no model quantization is used.

Models are trained for 10 epochs with fp16 mixed precision, a learning rate of 1×10^{-4}, a per-device batch size of 4, and gradient accumulation of 4. For language- and family-aware routing, routing features are extracted from the 10th layer of XLSR-53, which is the best layer among $\{4, 7, 10\}$.

TABLE III: Overall CER (%) for different hybrid routing weights λ on the dev split.

λ	0.1	0.3	0.5	0.7	0.9
$k = 1$	**10.69**	10.85	10.88	10.82	10.90
$k = 2$	**10.95**	11.03	11.10	11.02	11.06

IV. Results and Analysis

A. Effects of Language-Guided Routing

Table II compares traditional, language-guided, and language-family-guided routing.

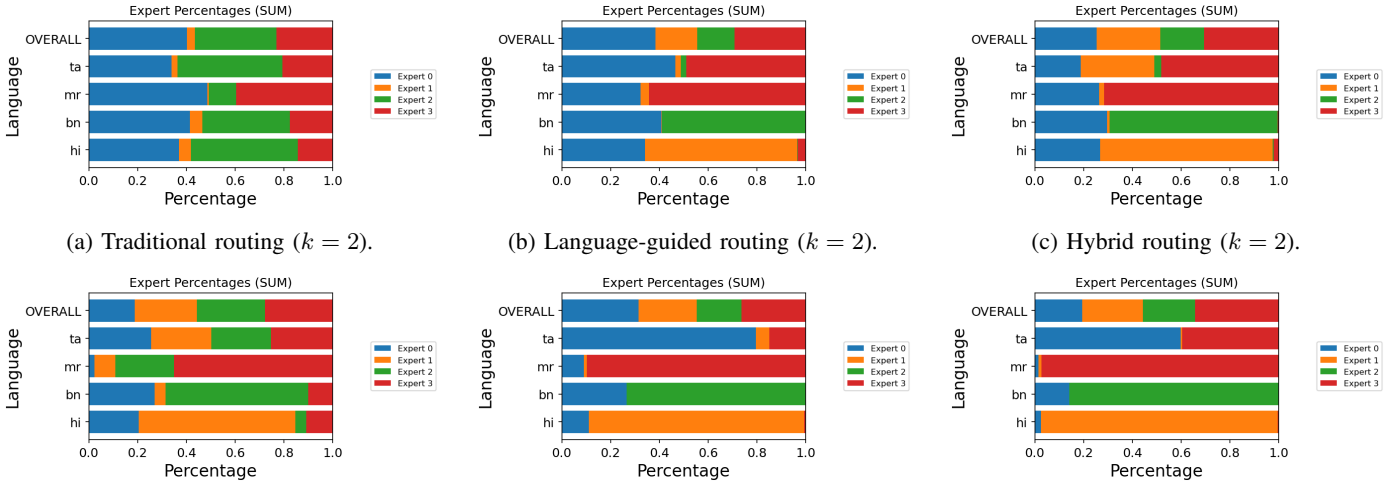
Under top-2 routing ((k=2)), ground-truth family routing (F1) achieves an overall CER of 9.95%, compared with 10.19% for traditional routing (T1), while ground-truth language routing (L1) obtains 10.15%. Under top-1 routing ((k=1)), L3 and F3 achieve CERs of 9.71% and 9.72%, respectively.

These results suggest that incorporating linguistic information into expert routing can benefit multilingual low-resource ASR under the XLSR-53 backbone. However, the relatively small differences among the routing strategies indicate that the effectiveness of MoLE depends more on expert specialization itself than on the specific form of linguistic information.

B. Effects of Top-k Expert Selection

Table II also shows a consistent performance difference between top-2 and top-1 routing. From the results, top-1 routing generally achieves a lower error rate than top-2 routing even though top-2 routing activates twice parameters than top-1 routing.

For traditional routing, reducing the number of activated experts from two to one slightly improves the overall CER from 10.19% (T1) to 10.16% (T2). Similar improvements



(a) Traditional routing ($k = 2$). (b) Language-guided routing ($k = 2$). (c) Hybrid routing ($k = 2$).

(d) Traditional routing ($k = 1$). (e) Language-guided routing ($k = 1$). (f) Hybrid routing ($k = 1$).

Fig. 3: Expert usage distributions across languages under different routing strategies.

are observed for both language-guided and language-family-guided routing. While linguistic information provides useful guidance for expert routing, the benefit becomes more evident when each utterance is routed to a single expert rather than being distributed across multiple experts.

C. Overall Comparison of Routing Strategies

Table II shows the ASR performance of different MoLE routing strategies. The LoRA-only baseline (T0) obtains an overall CER of 22.27%. All MoLE configurations outperform this baseline, indicating that the introduction of LoRA experts consistently improves multilingual ASR performance. The comparable r=128 LoRA showed no further gains.

The hybrid routing strategy (C2) achieves the best performance by combining the traditional gate with the model-predicted gate in Eq. (7). To determine the best hybrid routing weight λ, which controls the contribution of g_{LID}, we search different weights from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ on the dev split. Based on the results in Table III, we set $\lambda = 0.1$ for all hybrid routing experiments. Compared with the traditional routing baselines (T1 and T2), C1 and C2 achieve relative CER reductions of 3.7% and 9.4%. C2 achieves an overall CER of 9.20%, outperforming all other routing configurations.

These results suggest that incorporating linguistic information into expert routing can benefit multilingual low-resource ASR under the XLSR-53 backbone. Furthermore, language-guided routing can be viewed as a lightweight linguistic bias to traditional routing. While its effect on the routing is relatively small, the additional linguistic information leads to more effective expert utilization. As a result, the hybrid routing strategy achieves the best overall ASR performance.

D. Expert Usage Analysis

Fig. 3 illustrates the distributions of activated experts under different routing strategies. For the traditional routing strategy with top-2 routing (Fig. 3a) and top-1 routing (Fig. 3d) , the activations are relatively evenly distributed across the four experts. A similar pattern can be observed in language-guided top-2 routing (Fig. 3b) and hybrid routing (Fig. 3c). Since each utterance is processed by two experts, the activations are shared among multiple experts, resulting in more balanced utilization.

In contrast, stronger expert specialization is shown in language-guided top-1 routing experiments. As shown in Fig. 3e and Fig. 3f, different languages are routed to different experts. For example, Bengali is mainly processed by expert 2, while Marathi relies heavily on expert 3. This behavior indicates that expert tends to learn language-dependent features.

Another pattern can be observed in the hybrid strategy Fig. 3f. While the overall activations remain distributed across all experts, Tamil utterances are distributed more evenly two experts. This suggests that combining both routing strategies encourages experts to focus on different languages while retaining a degree of feature sharing across languages.

V. Conclusions

In this work, we investigated traditional, language-guided, and hybrid MoLE routing for multilingual low-resource ASR using XLSR-53 as the backbone. Experimental results show that language-aware routing generally performs comparably to or better than traditional routing. In addition, top-1 routing achieves better performance than top-2 routing by encouraging stronger expert specialization. The best results are obtained by a hybrid routing strategy that combines a traditional gate with language probabilities from a language identification model. Expert usage analysis further suggests that effective expert specialization is associated with better ASR performance. In future work, we will investigate the correlation between the degree of expert specialization and the language-specific ASR results, and extend to other multilingual speech encoders.

Acknowledgment

This research/project is supported by the National Research Foundation, Singapore under its National Large Language

Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is also supported by A*STAR under its Japan-Singapore Joint Call: Japan Science and Technology Agency (JST NEXUS 251043539) and A*STAR 2024 (R24I6IR136). This work is further supported by JST BOOST JPMJBS2407.

References

[1] P. Arisaputra, A. T. Handoyo, and A. Zahra, "Xls-r deep learning model for multilingual asr on low-resource languages: Indonesian, javanese, and sundanese," *ICIC Express Letters, Part B: Applications*, vol. 15, no. 6, pp. 551–559, 2024.

[2] S. Bandarupalli, B. Akkiraju, S. C. D, V. R. S. Narasinga, and A. Vuppala, "Efficient asr for low-resource languages: Leveraging cross-lingual unlabeled data," in *Proc. IJCNLP-AACL*, 2025, pp. 1611–1617.

[3] Z. Li and J. Niehues, "Enhance contextual learning in ASR for endangered low-resource languages," in *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, Association for Computational Linguistics, May 2025, pp. 1–7.

[4] J. Lee, M. Mimura, and T. Kawahara, "Leveraging ipa and articulatory features as effective inductive biases for multilingual asr training," in *ICASSP 2025*, pp. 1–5.

[5] Y. Fang et al., "Low-resource domain adaptation for speech llms via text-only fine-tuning," *arXiv preprint arXiv:2506.05671*, 2025.

[6] T. Luo et al., *Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models*, 2024. arXiv: 2402.12851.

[7] J. M. Janeiro et al., "Mixture of languages: Improved multilingual encoders through language grouping," in *Proceedings of the 2025 Conference on EMNLP*, 2025, pp. 29 695–29 710.

[8] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, and S. Watanabe, "Improving massively multilingual asr with auxiliary ctc objectives," in *ICASSP 2023*, IEEE, 2023, pp. 1–5.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, *Wav2vec 2.0: A framework for self-supervised learning of speech representations*, 2020. arXiv: 2006.11477.

[10] A. Babu et al., "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[11] J. Chen, C. Chu, S. Li, and T. Kawahara, "Data selection using spoken language identification for low-resource and zero-resource speech recognition," in *2024 APSIPA ASC*, 2024, pp. 1–6.

[12] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[13] Y. Zeng and K. Lee, "The expressive power of low-rank adaptation," *arXiv preprint arXiv:2310.17513*, 2023.

[14] C. Poth, H. Sterz, I. Paul, and et al., "Adapters: A unified library for parameter-efficient and modular transfer learning," in *Proc. EMNLP Demo*, 2023, pp. 149–160.

[15] J. Mainzinger and G.-A. Levow, "Fine-tuning asr models for very low-resource languages: A study on mvskoke," in *Proc. ACL SRW*, 2024, pp. 76–82.

[16] Z. Y. et al., "Master-asr: Achieving multilingual scalability and low-resource adaptation in asr with modular learning," in *International Conference on Machine Learning*, PMLR, 2023, pp. 40 475–40 487.

[17] X. Y. et al., "Adapting pre-trained self-supervised learning model for speech recognition with light-weight adapters," *Electronics*, vol. 13, no. 1, 2024.

[18] Y. Zhou et al., "Mixture-of-experts with expert choice routing," *Advances in Neural Information Processing Systems*, vol. 35, pp. 7103–7114, 2022.

[19] N. Shazeer et al., *Outrageously large neural networks: The sparsely-gated mixture-of-experts layer*, 2017. arXiv: 1701.06538 [cs.LG].

[20] K. Hu, B. Li, T. Sainath, Y. Zhang, and F. Beaufays, "Mixture-of-Expert Conformer for Streaming Multilingual ASR," in *Interspeech 2023*, 2023, pp. 3327–3331.

[21] Y. Liang, F. Meng, and J. Zhou, "Thor-moe: Hierarchical task-guided and context-responsive routing for neural machine translation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 21 433–21 445.

[22] L. Bandarkar, C. Yang, M. Fayyaz, J. Hu, and N. Peng, "Multilingual routing in mixture-of-experts," *arXiv preprint arXiv:2510.04694*, 2025.

[23] A. Gupta, A. Parulekar, S. Chattopadhyay, and P. Jyothi, "Parameter-efficient adaptation of multilingual multimodal models for low-resource asr," in *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, 2024, pp. 175–185.

[24] X. Zhao, X. Chen, Y. Cheng, and T. Chen, "Sparse moe with language guided routing for multilingual machine translation," in *The Twelfth International Conference on Learning Representations*, 2024.

[25] Y. Yu, D. Zhang, Y. Ren, and et al., "Progressive lora for multimodal continual instruction tuning," in *Findings of ACL*, 2025, pp. 2779–2796.

[26] R. Bagat, I. Illina, and E. Vincent, "Mixture of lora experts for low-resourced multi-accent automatic speech recognition," in *Interspeech 2025*, ISCA, Aug. 2025, pp. 1143–1147.

[27] R. Ardila, M. Branson, K. Davis, and et al., "Common voice: A massively-multilingual speech corpus," in *Proc. LREC*, 2020, pp. 4218–4222.

[28] S. Mangrulkar et al., *PEFT: State-of-the-art parameter-efficient fine-tuning methods*, 2022.