HIPA-MoE: A Parameter-Efficient Fine-Tuning Architecture with Hierarchical Adapter-based Mixture-of-Experts for Multilingual ASR

Xun Lu*† , Xuyang Wang*‡ , Gaofeng Cheng*†, Lin Zheng*† , Pengyuan Zhang*†
* Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS, China
† University of Chinese Academy of Sciences, China
‡ Corresponding Author. E-mail: wangxuyang@hccl.ioa.ac.cn

***Abstract*—Multilingual automatic speech recognition (MASR) has advanced significantly with self-supervised pretraining (SSL). However, conventional fine-tuning remains constrained by data imbalance, especially for low-resource languages. Although Mixture-of-Experts(MoE) has provided a promising approach for multilingual automatic speech recognition, expanding feedforward networks (FFNs) into MoE layers often incurs significant parameter overhead. To solve these issues, we propose HIPA-MoE, a novel Hierarchical Inverted Pyramid Adapter Mixture-of-Experts (HIPA-MoE) architecture that uses lightweight adapters as experts with language-aware routing. Our model hierarchically organizes experts to capture both universal acoustic features and language-specific phonetic nuances. Specifically, shared adapters in the lower layers model cross-lingual patterns, while the upper layers deploy language-specific adapters for fine-grained specialization. Additionally, we incorporate an auxiliary phoneme-level loss via uroman transliterations to enhance cross-lingual speech representation. Experiments on ML-SUPERB 2.0 demonstrate that HIPA-MoE achieves state-of-the-art performance in low-resource and long-tail languages while maintaining high parameter efficiency and scalability.**

Index Terms—multilingual ASR, MoE, adapters

I. Introduction

Multilingual automatic speech recognition (MASR) has significantly benefited from advances in self-supervised pretraining (SSL). Models such as XEUS[1], USM[2], and MMS[3] are pre-trained on vast amounts of unlabeled speech data across numerous languages and demonstrating strong cross-lingual generalization. The practical integration of SSL models into downstream applications, such as automatic speech recognition (ASR), typically requires a specialized fine-tuning process. However, traditional fine-tuning methods are often impacted by data imbalance, leading to notable performance degradation for tail languages[4].

A promising approach involves the use of Mixture-of-Experts (MoE) architectures, which utilize specialized subnetworks to effectively process language-specific acoustic features. By dynamically routing to experts based on specific tasks and languages, MoE increases model multilingual capacity while maintaining inference efficiency. It has been widely applied in the fields of natural language processing[5] , and speech recognition[6]–[8]. Some MoE-based models have been proposed to tackle the multilingual ASR task using language-related router[6]. Nevertheless, conventional MoE configurations replace feed-forward network (FFN) layers with expert layers[9]. Given that FFNs constitute a considerable portion of the model's parameters, scaling MoE to accommodate a wide range of languages results in a rapid increase in the total number of parameters and incurs substantial training computational expenses.

Adapter-based methods provide a lightweight and modular alternative. Adapters are compact, trainable modules inserted into a frozen backbone network, enabling language-specific adaptation with minimal parameter overhead.

Additionally, another key yet often overlooked source of cross-lingual generalization lies in the shared phonetic features among languages. Many phonemes exhibit consistent articulatory and acoustic realizations across typologically diverse languages[10]. However, existing approaches typically do not explicitly model or leverage these shared phonological patterns, limiting the ability of ASR models to generalize pronunciation knowledge to low-resource or unseen languages.

To address these challenges, we propose hierarchical Inverted Pyramid Adapter Mixture-of-Experts (HIPA-MoE) multilingual ASR architecture that combines hierarchical representations with a modular, adapter-based MoE design. HIPA-MoE adopts an inverted pyramid structure, in which fewer shared experts are utilized in the lower encoder layers to capture language-universal acoustic features, while a larger number of language-specific adapter experts are introduced in the upper layers, with one expert per target language. This architecture allows the model to capture shared, generalizable features in the earlier layers while enabling fine-grained, language-specific specialization in the deeper layers, thereby optimizing both performance and scalability.

The key innovations of HIPA-MoE are as follows:

(1) Hierarchical Inverted Pyramid Structure: HIPA-MoE builds on a frozen self-supervised backbone and introduces an inverted pyramid hierarchical expert structure, enabling progressive adaptation from shared to language-specific experts for efficient multilingual speech recognition.

(2) Adapter-Based MoE Experts: Adapters serve as MoE

979-8-3315-7206-8/25/$31.00

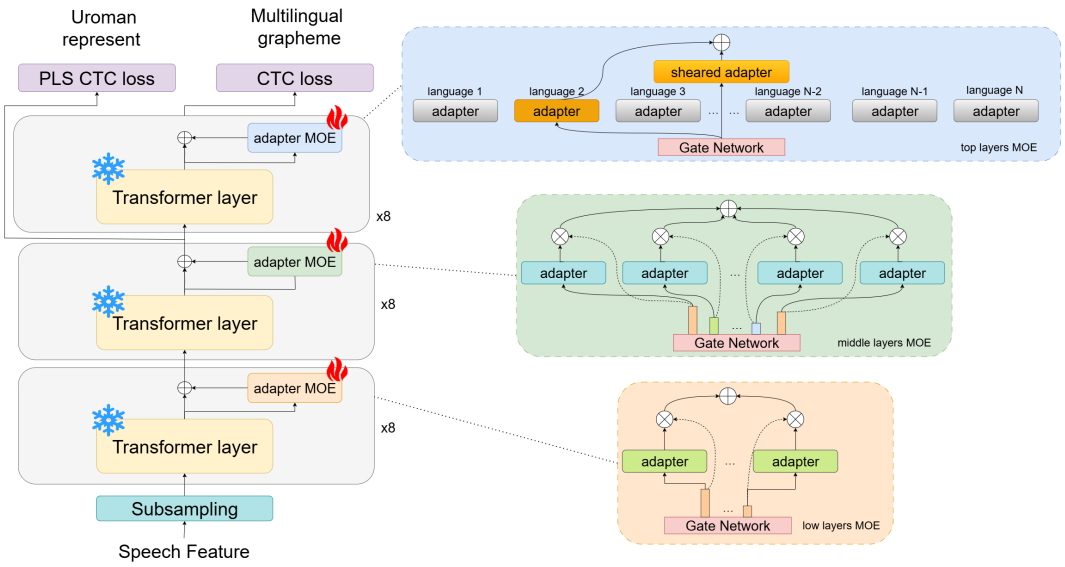

Fig. 1. Structure of HIPA-MOE
The left part shows the overall structure, and the right part shows the moe structure of different layers.

experts, offering a compact and scalable parameter expansion mechanism that avoids the overhead of full-model expansion while dealing with a large number of languages.

(3) Multilevel Language-Aware Expert Routing: A multi-level routing strategy gradually integrates language information at different layers, leveraging more language cues at higher levels to achieve precise and flexible expert selection.

(4) Cross-Lingual Phonological Supervision: A phoneme-level auxiliary loss is introduced at intermediate layers using uroman[11] transliterations, guiding the model to capture shared phonological features and promoting better generalization across languages.

By integrating hierarchical model structure with multi-level supervision, HIPA-MoE addresses the limitations of previous approaches in both parameter efficiency and cross-lingual generalization, offering a scalable and effective framework for multilingual ASR.

II. Related Work

A. Mixture-of-Experts (MoE) in ASR

Several prior works [6], [12], [13] have explored the application of MoE architectures in ASR. In [12], the FFN of the Conformer model is replaced with MoE layers to enable streaming multilingual ASR via dynamic top-2 expert activation. LR-MoE[13] introduces language-specific experts, while M-MoE[6] combines both language-specific and language-unkown experts: inputs with language IDs are routed to the corresponding experts, and those without IDs are handled by the language-agnostic MoE. BLR-MoE[7] extends the MoE mechanism to the attention modules and adopts a hierarchical routing strategy, with shared encoders at lower layers and MoE layers at the top. However, these studies mainly focus on designing novel MoE architectures, which can be computationally expensive when training from scratch. Recent work such as UME[9] proposes upcycling the FFN layers of pretrained ASR models into MoE layers, enabling the transformation into larger MoE-based architectures.

B. Parameter-efficient Fine-tuning

In the ASR domain, there has been growing interest in parameter-efficient fine-tuning approaches. Low-Rank Adaptation (LoRA) [14] introduces trainable low-rank matrices for iterative adaptation of ASR models. Inspired by MoE, researchers have begun to treat LoRA modules as domain experts to address the challenges of real-world multi-domain applications. HDMoLE [8] integrates LoRA with MoE, using dynamic thresholds to adaptively activate a variable number of experts in LLM-based ASR. However, LoRA is primarily optimized for large language models. In ASR, simpler residual adapters have proven more parameter-efficient[15]. For instance, [4] introduces language-aware adapters to mitigate the long-tail problem. Hyper-Adapter[16] uses a hyper-network to generate adapter weights, achieving a balance between efficiency and performance in MASR. Drawing from these insights, we propose a hierarchical MoE design where adapters serve as experts.

C. Hierarchical Representation and Cross-lingual Supervision

To improve multilingual and cross-lingual generalization, Whistle[17] uses weak speech supervision to enhance data-efficient multilingual recognition, validating the transferability of speech features in low-resource settings. [18] enhances cross-lingual generalization in end-to-end multilingual phoneme recognition via phoneme set unification. [10] introduces IPA symbols and phonetic features as inductive biases,

training models to predict IPA targets and general pronunciation features. LUPE[19] integrates language identification (LID), phonetic features, language-specific processing modules, and cross-lingual self-supervised speech representations to progressively encode multi-granular language and acoustic information across layers. However, many previous approaches rely on IPA as a phonetic representation, which is academically precise but overly complex. Inspired by MMS[3], we adopt uroman[11] to convert language-specific graphemes into a unified and more readable Romanized format.

III. Proposed Method

To effectively tackle the challenges of data imbalance, parameter scalability, and insufficient cross-lingual generalization in multilingual ASR, we propose a **H**ierarchical **I**nverted **P**yramid **A**dapter **M**ixture-**o**f-**E**xperts (HIPA-MoE) architecture. Our design combines the efficiency of adapter modules with the representational power of MoE and phoneme-level supervision, delivering a lightweight yet expressive multilingual ASR framework. The overall architecture is illustrated in Figure 1.

A. Hierarchical Inverted Pyramid Structure

HIPA-MoE is built upon a pre-trained SSL encoder MMS, which is frozen during fine-tuning to retain general-purpose acoustic representations. On top of this frozen backbone, we introduce a hierarchical structure composed of shared and language-specific experts distributed asymmetrically across the encoder layers:

- **Lower Layers – Shared Adapter Experts:** The bottom layers (Layer 1–8) contain a small number of *shared adapter experts*. These experts capture generalizable acoustic and phonological patterns that are common across languages. Their reduced count and position ensure low parameter cost and promote cross-lingual transfer.
- **Middle Layers – Experts with language information:** The moderate layers (Layer 9–16) contain a moderate number of *shared adapter experts*. Unlike the lower layers, the middle layers introduce language information as a part of the routing, enabling the model to capture acoustic information related to language.
- **Upper Layers – Language-Specific Adapter Experts:** In contrast, the upper encoder layers (Layer 16–24) are equipped with *language-specific adapter experts*, each dedicated to a target language. In order to maintain cross-lingual capture capability, a shared expert is employed in addition to language-specific experts.

This inverted pyramid design ensures efficient parameter utilization, with greater specialization occurring in the top layers, while retaining generalization capability in the lower layers.

B. Adapter-Based MoE Experts

Instead of replacing full feed-forward networks with large expert modules, we leverage **adapters** as MoE experts. Each adapter consists of an architecture with a down-projection, non-linearity, and up-projection:

$$\text{Adapter}(x) = W_{\text{up}} \cdot f(W_{\text{down}} \cdot x) \tag{1}$$

where $W_{\text{down}} \in \mathbb{R}^{d \times r}$, $W_{\text{up}} \in \mathbb{R}^{r \times d}$, and $f(\cdot)$ is a non-linear activation function such as ReLU. This compact structure allows scalable expansion of experts without substantially increasing the model's footprint. All adapters are inserted in a residual fashion after each Transformer layer's feed-forward block, maintaining the backbone's original flow while enabling task- and language-specific adaptation.

C. Multilevel Language-Aware Expert Routing

To accommodate the varying needs of different hierarchical layers in the adapter structure, we design a **layer-wise expert routing strategy** with progressively increasing usage of language information. In the lower and middle layers, we adopt a soft routing strategy based on Top-2 gating. For the lower layers, routing is purely data-driven, relying on local token-level representations:

$$\alpha_i = \text{Top2Softmax}(W_g x) \tag{2}$$

where α_i denotes the activation for expert i, W_g is a learnable gating projection, and only the top-2 experts receive non-zero weights per token.

In the middle layers, we enhance the routing mechanism by incorporating language identity. Specifically, a language embedding l is concatenated with the token hidden state x before gating. The language embedding l is obtained from a learnable word embedding table, where the language ID is used as the input index and the output dimension matches the hidden state dimension.:

$$\alpha_i = \text{Top2Softmax}(W_g[x; l]) \tag{3}$$

This progressive design allows the model to gradually transition from language-agnostic to language-aware expert selection, improving generalization in the lower layers and adaptability in the middle layers.

In the upper layers, we apply a static, language-specific routing policy. Each input is routed to a fixed expert corresponding to its language, alongside a globally shared expert to promote robustness and mitigate under-utilization:

$$f_{\text{top}}(x, l) = x + \text{Adapter}_l(x) + \text{Adapter}_{\text{shared}}(x) \tag{4}$$

where Adapter_l is the language-specific expert, and $\text{Adapter}_{\text{shared}}$ is a globally shared adapter added to improve robustness and expert utilization. This deterministic routing enhances specialization and ensures accurate modeling of fine-grained language-specific patterns.

D. Cross-Lingual Phonological Supervision

To enhance phoneme-level generalization, we incorporate an **auxiliary uroman loss** at the middle encoder layers. All transcriptions are converted into uroman[11] transliterations, which normalize language-specific scripts into standard Latin

alphabet phoneme-like representation. We then apply a CTC loss over these targets. This auxiliary supervision encourages shared layers to model articulatory-consistent units across languages, improving transferability to low-resource or unseen languages. In the final encoder layers, language-specific experts are activated to produce output representations tailored to the target language. Final predictions are generated via grapheme-level CTC decoding.

The final object loss function is shown in Eq. 5:

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

where λ denotes the weight of the grapheme-level target loss.

IV. Experimental

A. Experimental Data

We conduct our experiments on the ML-SUPERB 2.0 benchmark[20], a large-scale multilingual speech corpus designed to assess MASR models across a wide range of languages and resource conditions. The dataset includes labeled speech data from 142 languages, spanning the spectrum from high-resource to endangered languages. We used the one-hour training subset of ML-SUPERB2.0 [20] in our experiment.

Each utterance is paired with language-specific grapheme-level transcriptions, which serve as the primary decoding targets. We use the grapheme set is characterized by a size of 6417. To promote phonological generalization across languages, we also employ uroman tools[11] to convert text to the standard Latin alphabet as auxiliary phoneme-like labels. We use this standard target to capture similar phonetic characteristics between different languages in a simple way, rather than using more complex and professional international phonetic symbols.

B. Experimental Design

Our experimental framework is designed to evaluate the effectiveness, efficiency, and scalability of the proposed HIPA-MoE architecture under multilingual low-resource conditions. All models are fine-tuned based on the MMS-300M[3], a 24-layer Transformer model pretrained on over 490,000 hours of multilingual speech. All of our experiments were conducted under the fairseq[21] framework.

We use XLSR[22] and MMS Partial fine-tuning [3] models as benchmark references, with XLSR fully fine-tuned and MMS fine-tuned only on layers 9–12[20]. As baselines, we include a fully fine-tuned MMS model with end-to-end encoder updates, an adapter-based model that inserts lightweight residual adapters while keeping the backbone frozen, and a fully fine-tuned model enhanced with an auxiliary phoneme-level CTC loss (PLS) derived from uroman transliterations.

We then explore several MoE architectures. We introduce adapter-based MoE (A-MoE), where transformer layers are augmented with sparse adapter experts instead of full feed-forward MoE layers. Each adapter has a hidden dimension of 16 and applies ReLU as the activation function. Based on A-MoE, we examine different routing strategies. Static routing assigns each language to a fixed expert, with or without additional shared experts. Top-2 dynamic routing activates the two most relevant experts per input token. To prevent expert collapse and promote balanced expert usage, we also introduce a variant with a load-balancing loss (BLS), configured aligning with GShard[23]. In addition, we implement a top-2 FFN-wide MoE model in which the standard FFN layers are replaced with large-scale FFN experts, which have 8 experts, yielding up to 1.7B parameters.

Our proposed HIPA-MoE architecture with an inverted pyramid structure employs 16 adapter experts in the lower layers (1-8) with token-level top-2 routing, 64 experts in the middle layers (9-16) using top-2 routing combined with language embeddings, and 142 experts in the upper layers (17-24) adopting language-based static routing along with a shared expert. We evaluate four variants: the base model, a BLS-regularized version, a version augmented with an auxiliary phoneme-level CTC loss using uroman targets, and the final version that incorporates both BLS regularization and the auxiliary phoneme-level supervision.

All models are trained using the Adam optimizer with linear warm-up and cosine learning rate decay. The learning rate is set based on model architecture and training dynamics: 3e-5 for full fine-tuning, and 3e-4 when freezing the backbone to fine-tune adapters. Training is conducted on 4 NVIDIA A100 GPUs with 40GB of memory, using a batch size of approximately 1.6M tokens per GPU with gradient accumulation 2 steps. For models with auxiliary phoneme-level losses, we apply a loss weighting λ of 0.3 relative to the primary grapheme-level objective.

V. Results

We evaluate model performance using three metrics: Validation CER, Test CER, and Worst-30 CER, which reflects the average performance on the 30 lowest-performing languages. This last metric is particularly important for assessing fairness and robustness in multilingual ASR systems, highlighting the ability of a model to generalize beyond high-resource languages into the long-tail of typologically diverse speech communities.

A. Benefits of Dynamic Routing and Expert Specialization

The fully fine-tuned CTC baseline achieves a test CER of 14.5% with a Worst-30 CER of 38.37%. This highlights the difficulty of multilingual recognition under data imbalance, particularly for low-resource languages. By using the forward-layer MoE approach (FFN MoE), the number of trainable parameters reached 1.7B, but the performance was not as good as that obtained through direct fine-tuning. The fine-tuning method based on adapters achieves lightweight parameter updates (only 65M parameters need to be trained) by freezing the main network and training only the inserted adapter modules. But this method results in a significant performance decline, with a test CER of 37.2% and a Worst-30 CER of 69.15%.

TABLE I
COMPARISON OF THE PERFORMANCE OF THE INVERTED PYRAMID STRUCTURE WITH OTHER STRUCTURES, SHOWING THE METHOD, NUMBER OF MODEL TUNABLE PARAMETERS, AND CERS. * MEANS JOIN MULTILINGUAL ASR AND LID

Method	Tunable Params	Valid CER	Test CER	Worst-30 CER
XLSR* [20]	322M	–	15.8	–
Partial fine-tuning* [20]	91M	–	15.5	–
Full Fine-Tuning	322M	14.92	14.55	38.37
FFN MoE	1734M	21.36	19.84	37.13
Adapter Only	65M	39.85	37.23	69.15
A-MoE /w Static	125M	27.31	28.99	63.63
+ /w Shared Expert	126M	21.23	21.31	47.80
A-MoE /w Top-2&BLS	132M	22.08	21.81	51.48
HIPA-MoE base(Ours)	70.5M	14.75	14.24	35.59
+ /w BLS	70.5M	**14.18**	**13.70**	**33.35**

Early MoE models with static, language-specific expert routing perform poorly, with the adapter-based static MoE reaching a test CER of 28.99% and a Worst-30 CER of 63.63%, despite over 120M trainable parameters. Adding shared experts yields only modest gains, revealing the limitations of rigid expert assignment. In contrast, dynamic top-2 token-level routing with 64 experts and a balanced load score significantly improves performance, reducing the test CER to 21.81%. Our proposed HIPA-MoE achieves the best results, reaching a test CER of 14.24% and a Worst-30 CER of 35.59% with just 70.5M trainable parameters, surpassing the full fine-tuning baseline. Further adding load balancing loss to the lower layers reduces the test CER to 13.70%.

B. Effect of Uroman Auxiliary Supervision

TABLE II
PERFORMANCE COMPARISON OF ADDING INTER-UROMAN LOSS, SHOWING THE METHOD, NUMBER OF MODEL TUNABLE PARAMETERS, AND CERS. * MEANS JOIN MULTILINGUAL ASR AND LID

Method	Tunable Params	Valid CER	Test CER	Worst-30 CER
XLSR* [20]	322M	–	15.8	–
Partial Fine-tuning* [20]	91M	–	15.5	–
Full Fine-tuning	322M	14.92	14.5	38.37
+ PLS	322M	**13.66**	**13.36**	36.08
A-MoE /w Top-2 & BLS	132M	22.08	21.82	51.48
+ /w PLS	132M	20.71	20.27	48.96
HIPA-MoE base(Ours)	71M	14.75	14.24	35.59
+ /w PLS	71M	14.22	13.72	33.46
+ /w PLS & BLS	71M	14.03	13.53	**32.76**

Table II evaluates the effect of adding auxiliary CTC loss on *uroman*-transliterated phoneme targets. For the full model fine-tuning baseline, this improves validation/test CERs from 14.92%/14.5% to 13.66%/13.36%, highlighting enhanced phonetic alignment across languages.

When applied to MoE-based systems, phoneme-level supervision consistently improves convergence. For example, the top-2 expert model improves from 22.08%/21.82% to 20.71%20.27% in validation/test CERs. Applying this strategy to HIPA-MoE leads to a validation/test CER of 14.22%/13.72%, while adding load balance loss, it further achieves validation/test CERs to 14.03%/13.53%, further demonstrating the synergy of hierarchical structure, language-aware routing, and phonetic supervision. On the Worst-30 languages, HIPA-MoE achieves the best overall performance, surpassing all baseline models, and notably outperforms the globally fine-tuned model with intermediate phoneme-level loss, which achieves the lowest overall CER.

While the phoneme-augmented CTC baseline achieves the lowest overall CER, it requires full model fine-tuning and lacks structural modularity. In contrast, our approach balances performance, scalability, and interpretability, and is better suited for extension to massively multilingual settings.

VI. CONCLUSIONS

In this paper, we introduce HIPA-MoE, a hierarchical and parameter-efficient Mixture-of-Experts architecture for MASR. By leveraging a frozen SSL backbone with an inverted pyramid of adapter experts, HIPA-MoE enables effective modeling of both language-universal and language-specific features. Our language-aware routing mechanism ensures precise expert selection, while auxiliary phoneme-level supervision via uroman transliterations enhances cross-lingual phonological generalization. Experiments on ML-SUPERB 2.0 show HIPA-MoE outperforms traditional fine-tuning and static MoE methods in accuracy, robustness, and scalability, using only 22% of trainable parameters.

Moving forward, HIPA-MoE provides a strong foundation for scalable MASR systems. Future work will extend this architecture to code-switching, streaming ASR, and integration with multilingual large language models, aiming to further improve performance in complex and dynamic multilingual environments.

ACKNOWLEDGMENT

This work is partially supported by the National Key Research and Development Program of China(No.2024YFF0907304), the National Natural Science Foundation of China(No.62401560), the Youth Innovation Promotion Association Chinese Academy of Sciences, the Basic and Frontier Exploration Project Independently Deployed by Institute of Acoustics, Chinese Academy of Sciences(No.JCQY202411).

REFERENCES

[1] W. Chen, W. Zhang, Y. Peng, *et al.*, "Towards robust speech representation learning for thousands of languages," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 10 205–10 224.

[2] Y. Zhang, W. Han, J. Qin, *et al.* "Google USM: Scaling automatic speech recognition beyond 100 languages." arXiv: 2303.01037. (Sep. 25, 2023), [Online]. Available: http://arxiv.org/abs/2303.01037 (visited on 10/14/2024), pre-published.

[3] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *J. Mach. Learn. Res.*, vol. 25, no. 1, 97:4798–97:4849, Jan. 1, 2024, ISSN: 1532-4435.

[4] J. Bai, B. Li, Q. Li, T. N. Sainath, and T. Strohman, "Efficient adapter finetuning for tail languages in streaming multilingual ASR," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024.

[5] H. Zhou, Z. Wang, S. Huang, *et al.*, "MoE-LPR: Multilingual extension of large language models through mixture-of-experts with language priors routing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 24, pp. 26 092–26 100, 24 Apr. 11, 2025, ISSN: 2374-3468.

[6] S. Cao, X. Wang, Y. Zhang, X. Zhang, and L. 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)*, Apr. 2025, pp. 1–5.

[7] G. Ma, W. Wang, L. Zhou, Y. Yang, Y. Li, and B. Du, "BLR-MoE: Boosted language-routing mixture of experts for domain-robust multilingual E2E ASR," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[8] B. Mu, K. Wei, Q. Shao, Y. Xu, and L. Xie, "HDMoLE: Mixture of LoRA experts with hierarchical routing and dynamic thresholds for fine-tuning LLM-based ASR models," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[9] L. Fu, S. Yu, S. Li, L. Fan, Y. Wu, and X. He, "UME: Upcycling mixture-of-experts for scalable and efficient automatic speech recognition," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[10] J. Lee, M. Mimura, and T. Kawahara, "Leveraging IPA and articulatory features as effective inductive biases for multilingual ASR training," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[11] U. Hermjakob, J. May, and K. Knight, "Out-of-the-box universal romanization tool uroman," in *Proceedings of ACL 2018, System Demonstrations*, F. Liu and T. Solorio, Eds., Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 13–18.

[12] K. Hu, B. Li, T. Sainath, Y. Zhang, and F. Beaufays, "Mixture-of-expert conformer for streaming multilingual ASR," in *INTERSPEECH 2023*, ISCA, Aug. 20, 2023, pp. 3327–3331.

[13] W. Wang, G. Ma, Y. Li, and B. Du, "Language-routing mixture of experts for multilingual and code-switching speech recognition," presented at the Proc. Interspeech 2023, 2023, pp. 1389–1393.

[14] E. J. Hu, Y. Shen, P. Wallis, *et al.*, "LoRA: Low-rank adaptation of large language models," presented at the International Conference on Learning Representations, Oct. 6, 2021.

[15] K. C. Sim, Z. Huo, T. Munkhdalai, *et al.*, "A comparison of parameter-efficient ASR domain adaptation methods for universal speech and language models," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 6900–6904.

[16] Z. Hou, D. Garcia-Romero, and K. J. Han, "Hyper-adapter for parameter-efficient multilingual ASR adaptation," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5.

[17] S. Yusuyin, T. Ma, H. Huang, W. Zhao, and Z. Ou. "Whistle: Data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision." arXiv: 2406.02166. (Jun. 4, 2024), [Online]. Available: http://arxiv.org/abs/2406.02166 (visited on 11/27/2024), pre-published.

[18] H. Yen, S. M. Siniscalchi, and C.-H. Lee, "Boosting end-to-end multilingual phoneme recognition through exploiting universal speech attributes constraints," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of: IEEE, Apr. 14, 2024, pp. 11 876–11 880, ISBN: 979-8-3503-4485-1.

[19] W. Liu, J. Hou, D. Yang, M. Cao, and T. Lee, "LUPET: Incorporating hierarchical information path into multilingual ASR," presented at the Proc. Interspeech 2024, 2024, pp. 3979–3983.

[20] J. Shi, S.-H. Wang, W. Chen, *et al.* "ML-SUPERB 2.0: Benchmarking multilingual speech models across modeling constraints, languages, and datasets." arXiv: 2406.08641 [cs]. (Jun. 12, 2024), [Online]. Available: http://arxiv.org/abs/2406.08641 (visited on 06/04/2025), pre-published.

[21] M. Ott, S. Edunov, A. Baevski, *et al.* "Fairseq: A Fast, Extensible Toolkit for Sequence Modeling." arXiv: 1904.01038 [cs]. (Apr. 1, 2019), [Online]. Available: http://arxiv.org/abs/1904.01038 (visited on 09/03/2024), pre-published.

[22] A. Babu, C. Wang, A. Tjandra, *et al.* "XLS-R: Self-supervised cross-lingual speech representation learning at scale." arXiv: 2111.09296 [cs]. (Dec. 16, 2021), [Online]. Available: http://arxiv.org/abs/2111.09296 (visited on 06/12/2025), pre-published.

[23] D. Lepikhin, H. Lee, Y. Xu, *et al.*, "Gshard: Scaling giant models with conditional computation and automatic sharding," *ArXiv*, vol. abs/2006.16668, 2020.