



Efficient Multilingual ASR Finetuning via LoRA Language Experts

Jiahong Li^{*1,2}, Yiwen Shao², Jianheng Zhuo², Chenda Li¹, Liliang Tang², Dong Yu², Yanmin Qian^{†1}

¹Auditory Cognition and Computational Acoustics Lab
MoE Key Lab of Artificial Intelligence, AI Institute
School of Computer Science, Shanghai Jiao Tong University, China
²Tencent AI Lab, USA

{li-jiahong, yanminqian}@sjtu.edu.cn

Abstract

Recent advancements in deep learning have significantly enhanced multilingual automatic speech recognition (ASR) due to the development of advanced model architectures and available large-scale multilingual datasets. Despite that, multilingual ASR still suffers from the curse of multilinguality in that different languages tend to interfere with each other, making it difficult for the ASR model to identify multiple languages effectively while sharing model capacity across them. This paper proposes an efficient finetuning framework for customized multilingual ASR via prepared LoRA language experts based on Whisper. Through LoRA expert fusion or knowledge distillation, our approach achieves better recognition performance on target languages than standard fine-tuning methods. Experimental results demonstrate that the proposed models yield approximately 10% and 15% relative performance gains in language-aware and language-agnostic scenarios, respectively.

Index Terms: multilingual speech recognition, language experts, low-rank adaptation, knowledge distillation

1. Introduction

As a crucial component of human-machine interaction, automatic speech recognition (ASR) focuses on transcribing speech signals into their corresponding written text. With the advent of deep learning techniques, ASR has achieved remarkable success in various real-world scenarios[1, 2, 3, 4, 5]. Meanwhile, multilingual ASR, which necessitates a single system capable of recognizing multiple languages, has gained popularity due to the increasing demand for cross-lingual applications. Large-scale datasets[6, 7, 8] have enabled the development of massive multilingual ASR foundation models, such as Whisper [9], Google USM [10], and MMS [11], which facilitate efficient finetuning of customized multilingual ASR models.

Multilingual ASR models encounter two primary challenges in effectively dealing with multiple languages. Firstly, language interference is inevitable when the languages are belong to different linguistic families or the training data is imbalanced in terms of quantity and quality. Previous works attempt to mitigate this issue by language-specific modules[12, 13, 14, 15] or better sampling strategies[16, 17] for training data. This challenge can be more complex when we consider the *catastrophic forgetting* problem[18, 19], especially when language expansion[20, 21] is involved. Secondly, language identification is a hidden task inside multilingual ASR. While high performance is feasible with given language identity (LID), the performance can deteriorate dramatically when

LID is absent[22, 23]. This gap arises from the cascaded errors inherent in the implicit language identification and monolingual recognition pipeline.

Efficiency is another essential consideration in customized multilingual ASR. With the success of foundation models such as Whisper, fine-tuning has emerged as a popular method for developing downstream customized models. In this context, Low-Rank Adaptation (LoRA)[24] proves to be particularly effective regarding parameter scale and tuning time due to its low-rank modeling on the weight updates. Also, LoRA can be stored as separate residual modules, which avoids language interference when those languages are not finetuned. Knowledge distillation (KD)[25, 26] is another way to accelerate the finetuning process with guidance from the teacher model, usually a model with similar architecture but larger capacity.

To this end, we introduce an efficient multilingual ASR finetuning framework based on LoRA language experts, facilitating fast adaptation from the base Whisper model to downstream customized multilingual ASR models. A LoRA language expert refers to the language-specific LoRA parameters trained for one single language, which effectively captures the monolingual knowledge from sufficient training data and can be prepared in advance. Through simple linear combination, shallow layers of LoRA experts can be merged as the shared encoder and language router, producing a new customized LoRA mixture of language experts (MoLE) model to handle language-agnostic multilingual speech. Moreover, through layer-wise knowledge distillation, LoRA language experts can provide discriminative guidance for a multilingual LoRA student model, thereby efficiently enhancing the model performance. In summary, the contributions of this paper are as follows:

- We propose LoRA language experts to achieve improved language-aware recognition performance and apply the linear combination to merge different experts as LoRA MoLE for effective language-agnostic inference.
- Utilizing layer-wise knowledge distillation, a multilingual LoRA student model can achieve superior performance in language-agnostic scenarios with a faster finetuning process.

2. Background

2.1. Whisper

In this work, we adopt Whisper-medium[9] as the multilingual foundation model for convenience, yet other foundation models should have similar behaviors. Whisper[9] is a typical encoder-decoder Transformer model designed for multiple speech processing tasks, including multilingual speech recognition, speech translation, language identification, and voice activity detection. Whisper takes speech input \mathbf{X} in the form of an 80-dimensional

* This work was conducted during internship at Tencent.

† Corresponding author.

log-mel spectrum, with a fixed duration of 30 seconds after trimming or padding. The encoder module transforms the input speech features into hidden representation \mathbf{H} , capturing essential information for speech recognition. The consecutive decoder module generates in an auto-regressive manner the transcribed text tokens $\hat{\mathbf{Y}} = [\hat{y}_1, \dots, \hat{y}_L]$ where L denotes the total length of tokens. The current token is conditioned on the previous tokens together with the prompts \mathbf{P} . This process can be formulated as follows:

$$\mathbf{H} = \text{AudioEncoder}(\mathbf{X}) \quad (1)$$

$$\hat{y}_l = \text{TextDecoder}(\mathbf{P}, \hat{y}_{1:l-1}, \mathbf{H}) \quad (2)$$

2.2. LoRA

LoRA[24] is an advanced technique originally designed to enhance the parameter-efficient finetuning of large pretrained models in the context of natural language processing (NLP) tasks. Instead of updating all the model parameters during finetuning, LoRA introduces low-rank matrices as the updated components for target weight matrices inside the model. In practice, for a weight matrix $\mathbf{W}_i \in \mathbb{R}^{d_1 \times d_2}$, the new weight matrix $\hat{\mathbf{W}}_i$ after finetuning is modeled as

$$\hat{\mathbf{W}}_i = \mathbf{W}_i + \Delta \mathbf{W}_i \quad (3)$$

$$\Delta \mathbf{W}_i = \mathbf{B}_i \mathbf{A}_i \quad (4)$$

where $\mathbf{B}_i \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{A}_i \in \mathbb{R}^{r \times d_2}$ are trainable low-rank matrices after decomposition with rank $r \ll \min(d_1, d_2)$. LoRA significantly reduces the training parameters when the backbone weights are frozen. After finetuning, LoRA parameters can be merged back to the original weights to reduce storage, or kept isolated to serve as independent residual modules for specific tasks.

3. Methods

In this paper, we propose an efficient finetuning framework that leverages monolingual LoRA language experts to facilitate the finetuning process of customized multilingual ASR models.

3.1. Task description

Customized multilingual ASR models are designed to target a specific range of languages, encompassing both the base languages included during Whisper’s pretraining phase and additional languages for language expansion. Multilingual ASR can be decomposed into two primary tasks: language identification and monolingual speech recognition. While these two tasks are usually integrated to mitigate the risk of cascaded errors, it remains crucial to maintain balance between them. The Whisper decoder treats different language IDs as special tokens and decodes them prior to the sequence of text tokens.

3.2. LoRA language experts

LoRA is an effective approach for multilingual ASR finetuning, not only due to its parameter-efficiency but also because LoRA parameters can be isolated from the backbone Whisper model. Such isolation allows LoRA to naturally preserve recognition performance across all base languages as separate residual modules, thereby avoiding the *catastrophic forgetting* problem. We refer to one set of monolingual LoRA parameters including $\mathbf{A}_{i,j}$ and $\mathbf{B}_{i,j}$ as a LoRA expert for the j -th target language and an original Whisper weight matrix \mathbf{W}_i , which

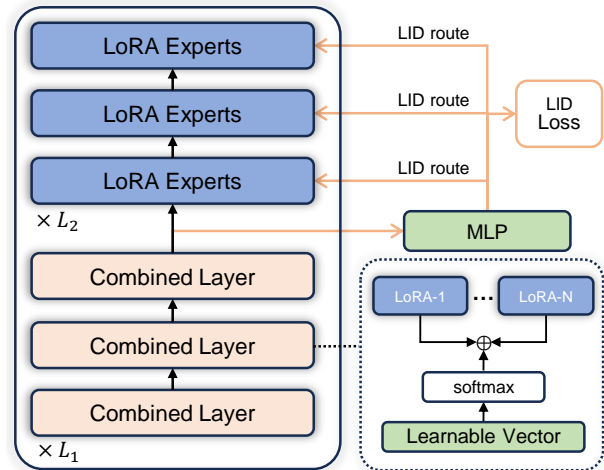


Figure 1: Paradigm of the MoLE encoder with combined LoRA language experts.

is finetuned on sufficient monolingual ASR data until convergence. LoRA language experts perform better than a single multilingual model due to the curse of multilinguality, where different languages can interfere with each other, causing performance degradation. Information from various languages in the database can be effectively stored within the corresponding LoRA language experts, facilitating the tuning process of new customized multilingual ASR models.

3.3. Language-agnostic LoRA MoLE

Although LoRA language experts can achieve nearly optimal performance when LID is provided, they cannot perform language-agnostic inference. To handle that, a language identification module is required to provide routing decisions among different experts. Rather than training an independent classifier, we propose to combine shallow encoding layers of the prepared LoRA language experts to form shared multilingual layers. This leads to a LoRA mixture of language experts (MoLE) system, inspired from the MoE[27, 28] architecture.

As shown in Figure 1, we take L_1 encoding layers for linear combination, where all the weights of LoRA language experts are summed using learnable weights. For N LoRA experts associated with a specific weight matrix \mathbf{W}_i , a learnable vector $\mathbf{V}_i \in \mathbb{R}^N$ followed by a softmax layer is trained to generate the linear combination weights. The combination process can be formulated as follows:

$$\alpha_{i,j} = \frac{\exp(\mathbf{V}_{i,j})}{\sum_k \exp(\mathbf{V}_{i,k})} \quad (5)$$

$$\hat{\mathbf{A}}_i = \sum_j \alpha_{i,j} \mathbf{A}_{i,j}, \quad \hat{\mathbf{B}}_i = \sum_j \alpha_{i,j} \mathbf{B}_{i,j} \quad (6)$$

$$\hat{\mathbf{W}}_i = \mathbf{W}_i + \hat{\mathbf{B}}_i \hat{\mathbf{A}}_i \quad (7)$$

In this manner, multiple LoRA experts within those layers can be merged to reduce the total number of parameters and retain knowledge of all languages to produce LID routes with a simple multi-layer perceptron (MLP). LoRA language experts and the Whisper backbone are kept frozen during finetuning of LoRA MoLE, with only the learnable vectors and the MLP updated based on the averaged LID loss and ASR loss.

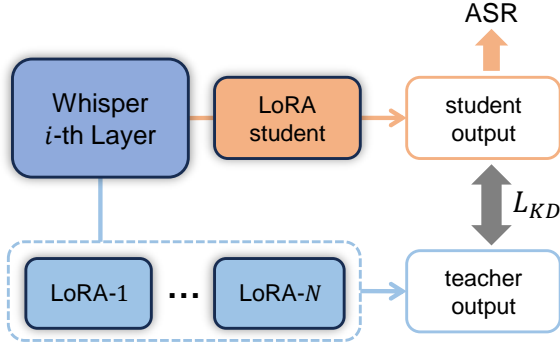


Figure 2: Process of layer-wise distillation from LoRA language experts to a multilingual LoRA student.

3.4. LoRA knowledge distillation

Although LoRA MoLE is effective, it contained more model parameters due to the remaining expert modules as they cannot be merged to the backbone weight matrices. To obtain a simpler model, we propose to distill knowledge from the LoRA language experts into a single multilingual LoRA student. Nonetheless, standard distillation on output ASR logits provides limited assistance on the fully-supervised finetuning process. Considering the residual format of LoRA, we assume the latent representations from different LoRA experts exhibit only minor differences. Consequently, a single multilingual LoRA student with a larger rank can align with multiple LoRA experts at the layer-wise granularity, and thus mimics the discriminative hidden distribution produced by the experts, which enhances the model’s ability to distinguish languages.

As shown in Figure 2, we apply knowledge distillation to layer-wise output of the language-specific LoRA language experts and the language-agnostic LoRA student. To alleviate the cascaded mismatch in the forwarding path, the student output is interpolated with the teacher output before being passed to the next layer with a probability of 0.5. The layer-wise KD losses from different layers, including the output ASR logits, are averaged and combined with the ASR loss for backpropagation during training. We use cosine similarity to compute the KD loss on the layer-wise outputs and Jensen divergence on the final ASR logits. The process of knowledge distillation is follows:

$$\mathbf{Y}_i = \text{Layer}_i(\mathbf{Y}_{i-1}) \quad (8)$$

$$\hat{\mathbf{Y}}_i = \text{Layer}_i(\tilde{\mathbf{Y}}_{i-1}) \quad (9)$$

$$\mathcal{L}_{\text{KD}_i} = 1 - \frac{\mathbf{Y}_i \cdot \hat{\mathbf{Y}}_i}{\|\mathbf{Y}_i\| \cdot \|\hat{\mathbf{Y}}_i\|} \quad (10)$$

$$\tilde{\mathbf{Y}}_i = \frac{1}{2}(\hat{\mathbf{Y}}_i + \mathbf{Y}_i) \quad (11)$$

$$\mathcal{L} = \mathcal{L}_{\text{ASR}} + \alpha \frac{1}{M} \sum_{i=0}^M \mathcal{L}_{\text{KD}_i} \quad (12)$$

Here, α is the balancing weight, M is the total number of layers, and \mathbf{Y}_i , $\hat{\mathbf{Y}}_i$, $\tilde{\mathbf{Y}}_i$ represent the teacher output, student output, and interpolated student output of the i -th layer, respectively. During training, the Whisper backbone and LoRA language experts are kept frozen, and only the LoRA student will be updated, which is initialized from the average value of LoRA experts.

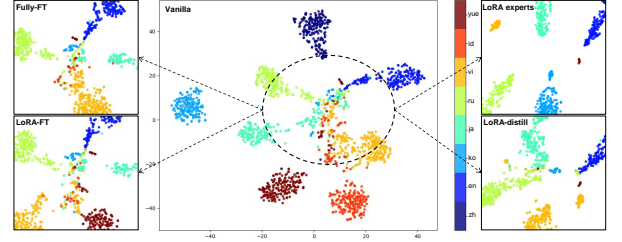


Figure 3: Visualization of the encoded hidden representations by t-SNE for different models.

4. Experiments

4.1. Setup

The backbone model is Whisper-medium which is well-pretrained to produce reasonable transcriptions for a wide range of languages. We select Zh (Chinese), En (English), Ko (Korean), Ja (Japanese), Ru (Russian), Vi (Vietnamese), and Id (Indonesian) as base languages and Yue (Cantonese) as language expansion. Experiments are conducted on an internal multilingual database with over 10,000 hours of data for Zh and En and around 5000 hours for other languages, which is sufficient for finetuning. For testing, we adopt AISHELL[29] and LibriSpeech[30] for Zh and En, and CommonVoice[8] for the others.

We apply LoRA to the attention modules, including the querying, key and value matrices, and fully-connected layers across all layers in both the encoder and decoder, with a rank of 64 for monolingual LoRA experts and 256 for multilingual LoRA baseline and teacher model. During training, we use AdamW[31] optimizer with a peak learning rate of $1e-4$ for LoRA models and $1e-5$ for fully-finetuned models, and the dynamic batch size is set to 160 seconds. The prepared monolingual LoRA language experts are trained to converge and stored for later use. The maximum training steps are set to 100,000 for other multilingual models and 20,000 for the LoRA MoLE model. All models are trained with 8 NVIDIA V100 32GB GPUs. During training, we remove Whisper’s restriction of 30-second input to make it more efficient. We apply a uniform sampling strategy for multilingual training to balance different languages. During inference, we use greedy search to decode test data and apply no external language model. The performance is evaluated using the word error rate (WER) for each language.

4.2. Results and analysis

In Table 1, we present the ASR performance of different multilingual ASR systems across eight languages. The vanilla Whisper model performs well on base languages but poorly on expanded Cantonese speech (using Zh for LID). After fully-finetuning, the *Fully-FT* model shows improved performance on Zh, Id, Vi and Yue, but performs worse on other languages, indicating interference between different languages. The *Curated Fully-FT* model is an enhanced version that uses four times the training steps and a curated language sampling strategy with weights of $\{1.0, 1.0, 0.25, 0.25, 0.25, 0.25, 0.25, 0.5\}$ for those languages, as concluded from experiments. This model significantly improves performance on language-aware tests, particularly for En, Vi, and Yue, although the performance gain is not consistent in language-agnostic scenarios.

Table 1: Performance Comparison of Different Multilingual ASR Systems on Eight Languages in WER(%).

Model	#Train Param	Language-Aware Test									Language-Agnostic Test								
		Zh	En	Ko	Ja	Id	Ru	Vi	Yue	avg	Zh	En	Ko	Ja	Id	Ru	Vi	Yue	avg
Vanilla	0	9.20	3.65	5.12	15.75	12.89	9.78	18.95	38.90	14.28	9.20	3.65	5.24	16.95	17.91	10.60	23.12	50.32	17.12
Fully FT	727M	5.24	4.41	5.34	17.21	9.88	9.74	16.68	9.68	9.77	5.24	5.02	5.69	16.83	12.20	9.93	20.90	9.68	10.69
Curated Fully FT		4.34	3.03	5.55	15.86	9.87	9.16	13.87	7.74	8.68	4.34	5.02	6.19	16.46	13.19	9.82	19.20	<u>7.74</u>	10.25
Mono. LoRA LE	66M*8	<u>3.59</u>	<u>2.44</u>	<u>4.54</u>	<u>14.31</u>	<u>8.93</u>	8.79	<u>11.75</u>	<u>7.48</u>	7.73	-								
LoRA MoLE	1M	4.07	2.54	4.65	14.39	10.51	<u>8.77</u>	14.79	8.88	8.58	<u>4.07</u>	<u>2.54</u>	<u>5.07</u>	<u>14.92</u>	12.88	<u>8.93</u>	17.25	8.88	9.32
Multi. LoRA	264M	5.34	2.93	5.37	15.58	10.28	10.25	18.37	9.64	9.72	5.34	4.88	5.73	16.22	13.10	10.59	26.54	9.68	11.51
Multi. LoRA KD		4.85	2.84	4.77	15.31	9.74	9.50	14.63	8.25	8.74	4.85	4.67	4.78	16.29	<u>11.68</u>	9.80	<u>17.22</u>	8.33	9.70

The *Monolingual LoRA Experts* achieves the best performance with LID input due to their language independence, while it cannot perform language-agnostic inference with lack of language identification ability. The proposed *LoRA MoLE* model with 20 layers merged can remedy that drawback with the sacrifice of the monolingual performance and minimum training parameters. The average WER rises to 8.68%, but this model provides the best average performance when LID input is unavailable. Notably, the performance of MoLE is closely tied to language classification accuracy, which is adversely affected by language imbalance. Consequently, performance on relatively high-resource languages, such as Zh and En, aligns well with language-aware inference, while performance deteriorates for other languages, particularly Vi.

In the last block, the finetuned *Multilingual LoRA* model performs worse than the fully-finetuned models, except for En, indicating a weaker modeling ability. The proposed *Multilingual LoRA KD* model exhibits approximately 10% and 15% relative performance gains on language-aware and language-agnostic tests, respectively, due to the layer-wise alignment with LoRA language experts, demonstrating effective knowledge transfer.

In Figure 3, we visualize the hidden representations of different languages from five models using t-SNE[32]. It is evident that both the fully fine-tuned model and the LoRA model struggle to separate multilingual samples near the decision boundary, whereas the distilled LoRA student aligns well with the LoRA language experts, producing distinctly separable results.

4.3. Ablation Study

LoRA MoLE In LoRA MoLE, the number of merged layers L_1 is a trade-off between the accuracy of language identification and the performance of remaining LoRA expert layers. We evaluate the performance variation with $L_1 \in \{8, 12, 16, 20, 24\}$, and present the results in Figure 4. When $L_1 = 8$, the combination is too shallow to generate useful features for language identification. As more expert layers are merged, the ASR performance of the remaining layers declines when LID is known, while the identification accuracy peaks with 16 or 20 merged layers. The best overall language-agnostic performance is achieved at $L_1 = 20$, which also significantly reduces the number of expert parameters within the encoder.

LoRA KD For LoRA KD, we compare the average WER with or without LID against other configurations, as shown in Table 2. As we reduce the student rank from 256 to 128, the WER increases correspondingly but remains better than the

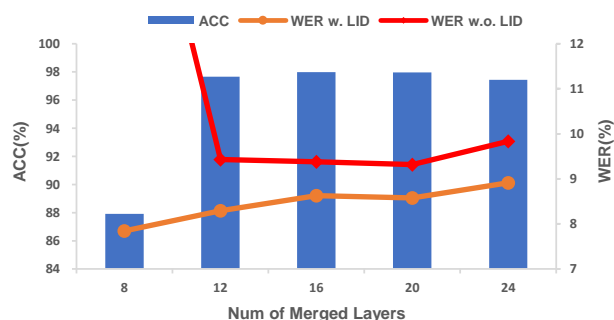


Figure 4: Variation of language routing accuracy and recognition WER regarding the number of merged layers in LoRA MoLE. Whisper encoder has 24 layers in total.

Table 2: Ablation Study on LoRA Knowledge Distillation in Averaged WER(%) among Eight Languages.

Models	Language-Aware	Language-Agnostic
Multi. LoRA	9.72	11.51
LoRA-KD	8.74	9.7
LoRA-KD-rank128	9.02	10.13
LoRA-KD-logits-only	9.39	10.32

basic multilingual LoRA. Additionally, we apply the standard output-level KD on the ASR logits, which resulted in only minor performance enhancements.

4.4. Limitation

The total number of parameters for LoRA language experts increases as the number of target languages grows. This can lead to increased memory costs when training models with more languages.

5. Conclusion

In this paper, we introduce an efficient finetuning framework for customized multilingual ASR based on foundation models such as Whisper. By leveraging language-specific LoRA language experts, we can train a language-agnostic LoRA MoLE model at minimal cost, and achieve superior performance on both testing scenarios with a multilingual LoRA student through layer-wise knowledge distillation.

6. Acknowledgement

This work was supported in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Commission Project under Grant 2021SHZDZX0102, and in part by Tencent AI Lab Rhino-Bird Focused Research Program.

7. References

- [1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. ICML*, 2006, pp. 369–376.
- [2] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. IEEE ICASSP*, 2013, pp. 6645–6649.
- [3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proc. IEEE ICASSP*, 2016, pp. 4960–4964.
- [4] S. Kim, T. Hori, and S. Watanabe, “Joint CTC-attention based end-to-end speech recognition using multi-task learning,” in *Proc. IEEE ICASSP*. IEEE, 2017, pp. 4835–4839.
- [5] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Trans. on Signal and Information Processing*, vol. 11, no. 1, 2022.
- [6] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, “MLS: A large-scale multilingual dataset for speech research,” in *Proc. ISCA Interspeech*, 2020, pp. 2757–2761.
- [7] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *Proc. IEEE SLT*, 2023, pp. 798–805.
- [8] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, “Common Voice: A massively-multilingual speech corpus,” in *Proc. ACL*, 2020, pp. 4218–4222.
- [9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. ICML*, 2023, pp. 28 492–28 518.
- [10] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, “Google USM: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.
- [11] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, “Scaling speech technology to 1,000+ languages,” *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.
- [12] N. Gaur, B. Farris, P. Haghani, I. Leal, P. J. Moreno, M. Prasad, B. Ramabhadran, and Y. Zhu, “Mixture of informed experts for multilingual speech recognition,” in *Proc. IEEE ICASSP*, 2021, pp. 6234–6238.
- [13] L. Zhou, J. Li, E. Sun, and S. Liu, “A configurable multilingual model is all you need to recognize all languages,” in *Proc. IEEE ICASSP*, 2022, pp. 6422–6426.
- [14] W. Wang, G. Ma, Y. Li, and B. Du, “Language-routing mixture of experts for multilingual and code-switching speech recognition,” in *Proc. ISCA Interspeech*, 2023, pp. 1389–1393.
- [15] S. Li, Y. You, X. Wang, K. Ding, and G. Wan, “Enhancing multilingual speech recognition through language prompt tuning and frame-level language adapter,” in *Proc. IEEE ICASSP*, 2024, pp. 10 941–10 945.
- [16] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, “Large-scale multilingual speech recognition with a streaming end-to-end model,” in *Proc. ISCA Interspeech*, 2019, pp. 2130–2134.
- [17] A. Piñeiro-Martín, C. García-Mateo, L. Docio-Fernandez, M. del Carmen López-Pérez, and G. Rehm, “Weighted cross-entropy for low-resource languages in multilingual speech recognition,” in *Proc. ISCA Interspeech*, 2024, pp. 1235–1239.
- [18] T. Xu, K. Huang, P. Guo, Y. Zhou, L. Huang, H. Xue, and L. Xie, “Towards rehearsal-free multilingual ASR: A LoRA-based case study on Whisper,” in *Proc. ISCA Interspeech*, 2024, pp. 2534–2538.
- [19] C. Y. Kwok, J. Q. Yip, and E. S. Chng, “Continual learning optimizations for auto-regressive decoder of multilingual ASR systems,” in *Proc. ISCA Interspeech*, 2024, pp. 1225–1229.
- [20] Y. Khassanov, Z. Chen, T. Chen, T. Y. Chong, W. Li, L. Lu, and Z. Ma, “Extending multilingual ASR to new languages using supplementary encoder and decoder components,” in *Proc. IEEE ICASSP*, 2024, pp. 10 586–10 590.
- [21] Z. Song, J. Zhuo, Y. Yang, Z. Ma, S. Zhang, and X. Chen, “LoRA-Whisper: Parameter-efficient and extensible multilingual ASR,” in *Proc. ISCA Interspeech*, 2024, pp. 3934–3938.
- [22] W. Liu, J. Hou, D. Yang, M. Cao, and T. Lee, “A parameter-efficient language extension framework for multilingual ASR,” in *Proc. ISCA Interspeech*, 2024, pp. 3929–3933.
- [23] B. Yan, V. Pratap, S. Watanabe, and M. Auli, “Improving multilingual ASR in the wild using simple N-best re-ranking,” *arXiv preprint arXiv:2409.18428*, 2024.
- [24] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” in *Proc. ICLR*, 2022.
- [25] F. C. Salinas, K. Kumatani, R. Gmyr, L. Liu, and Y. Shi, “Knowledge distillation for mixture of experts models in speech recognition,” Microsoft Tech. Report, MSR-TR-2022-6, May 2022, Tech. Rep., 2022.
- [26] T. P. Ferraz, M. Z. Boito, C. Brun, and V. Nikoulina, “Multilingual Distilwhisper: Efficient distillation of multi-task speech models via language-specific experts,” in *Proc. IEEE ICASSP*, 2024, pp. 10 716–10 720.
- [27] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in *Proc. ICLR*, 2017.
- [28] S. Dou, E. Zhou, Y. Liu, S. Gao, W. Shen, L. Xiong, Y. Zhou, X. Wang, Z. Xi, X. Fan *et al.*, “LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin,” in *Proc. ACL*, 2024, pp. 1932–1945.
- [29] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in *Proc. O-COCOSDA*, 2017, pp. 1–5.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE ICASSP*, 2015, pp. 5206–5210.
- [31] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proc. ICLR*, 2019.
- [32] L. Van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.