MOSA: Mixtures of Simple Adapters Outperform Monolithic Approaches in LLM-based Multilingual ASR

Junjie Li[1,2], Jing Peng[1,2], Yangui Fang[4], Shuai Wang[3], Kai Yu[1,2]†
[1] *X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University*
[2] *MoE Key Lab of Artificial Intelligence, Jiangsu Key Lab of Language Computing*
[3] *School of Intelligence Science and Technology, Nanjing University*
[4] *School of Electronic Information and Communications, Huazhong University of Science and Technology*
{junjieli, jing.peng, kai.yu}@sjtu.edu.cn, shuaiwang@nju.edu.cn

***Abstract*—End-to-end multilingual ASR aims to transcribe speech from different languages into corresponding text, but is often limited by scarce multilingual data. LLM-based ASR aligns speech encoder outputs with LLM input space via a projector and has achieved notable success. However, prior work mainly improves performance by increasing data, with little focus on cross-lingual knowledge sharing. Moreover, a single complex projector struggles to capture both shared and language-specific features effectively. In this work, we propose MOSA (Mixture of Simple Adapters), leveraging a Mixture-of-Experts mechanism to combine lightweight adapters that learn shared and language-specific knowledge. This enables better utilization of high-resource language data to support low-resource languages, mitigating data scarcity issues. Experimental results show that MOSA-Base achieves a 15.4% relative reduction in average WER compared to the Baseline-Base and consistently outperforms it across all languages. Remarkably, MOSA-Base surpasses the Baseline-Base even when trained with only 60% of its parameters. Similarly, MOSA-Large outperforms the Baseline-Large in average WER and demonstrates greater robustness to data imbalance. Ablation studies further indicate that MOSA is more effective at handling individual languages and learning both language-specific and shared linguistic knowledge. These findings support that, in LLM-based ASR, a mixture of simple adapters is more effective than a single, complex adapter design.**

***Index Terms*—Multilingual ASR, LLM-Based ASR, Mixture of Experts.**

I. Introduction

End-to-end multilingual automatic speech recognition (ASR) requires a single model to accurately transcribe speech from multiple languages into corresponding text. This technology plays a vital role in facilitating cross-lingual communication, such as real-time translation, and in preserving linguistic diversity. However, it also faces significant challenges, including data scarcity and the complexity of dialectal variations. In recent years, large language models (LLMs) have demonstrated remarkable capabilities across various domains [1]–[7]. The strong performance of LLMs has sparked growing interest in a new ASR paradigm: aligning the speech input space with LLM input spaces through an adapter module. This alignment enables direct utilization of the LLMs' extensive commonsense knowledge and contextual understanding learned from large-scale text corpora.

The paradigm of aligning speech encoders with LLM input spaces via adapters has been extensively explored in ASR. For instance, SLAM-ASR [8] achieves state-of-the-art performance on the English LibriSpeech 960h [9] dataset by training a simple linear adapter layer. [10] systematically compares the effects of various combinations of speech encoders, LLMs, and adapter architectures on Chinese ASR performance. FireRedASR [11] employs a more powerful speech encoder to enhance Mandarin speech recognition performance. In multilingual ASR, [12] investigates the impact of frame stacking to reduce the sequence length fed into LLMs, exploring how different frame rates influence recognition accuracy. Qwen-Audio [13] leverages a carefully crafted multi-task training framework to enable multilingual speech recognition, while Qwen2-Audio [14] further enhances understanding by replacing task-specific labels with natural language descriptions, thus better utilizing the reasoning capabilities of LLMs. Seed-ASR [15] introduces a multi-stage training strategy inspired by LLMs, which enables both multilingual recognition and improved context-guided understanding. Ideal-LLM [16] proposes a dual-encoder architecture that incorporates a language classifier to assign different speech encoder weights for different languages, thereby improving multilingual adaptability.

While previous work has improved ASR performance by enhancing the speech encoder or even fine-tuning LLMs [11], [13]–[15], relatively little attention has been paid to the design of adapter modules [16], [17]. On one hand, even if these adapters are complex and include many transformer layers, a single adapter may still fail to effectively align speech representations from all languages with the LLM input space. On the other hand, most prior approaches focus on increasing the amount of training data [11], [13]–[15], without fully exploiting the potential for knowledge sharing across languages, such as syntactic or phonetic similarities [18], which could benefit low-resource languages by transferring knowledge from high-resource ones. Meanwhile, each expert in the Mixture-of-Experts (MoE) [19]–[23] model specializes in different domains, making it a highly effective architectural design for multilingual and multi-domain settings. HDMoLE [17] addresses multi-accent ASR using a hierarchical MoE [19] framework, where the LoRA [24] expert weights assigned by a global pre-trained accent recognition model and local models are used to handle different accents.

In this work, we posit that there exists both shared and language-specific knowledge across different languages. A single complex adapter struggles to effectively learn and map these diverse types of knowledge. To address this, we propose MOSA (Mixture of Simple Adapters), a novel approach based on the MoE mechanism. Each adapter expert is responsible for learning either shared or language-specific knowledge, while a router predicts expert weights for each speech input, enabling dynamic expert mixture. For the sake of simplicity, we use only a single encoder, and each adapter consists of just two linear layers. Unlike previous methods that employ a variety of loss functions, we compute only the cross-entropy loss on transcriptions. This design allows each language to better align with the input space of large language models (LLMs), leveraging high-resource language information to enhance the performance of low-resource languages. Consequently, it helps mitigate the issue of data scarcity in low-resource languages. Experimental results show

†Kai Yu is the corresponding author.

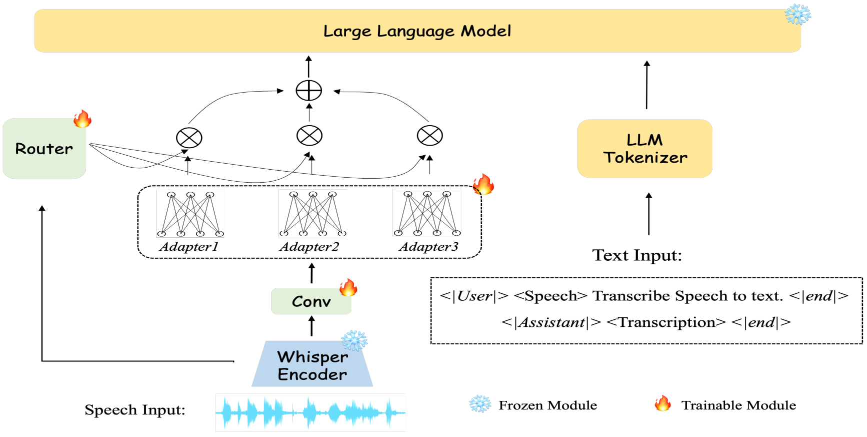

Fig. 1. Model Architecture. The speech encoder extracts features from speech. Adapters map these features into the LLM input space, guided by a Router that dynamically weights their outputs. The LLM then performs speech recognition based on the aligned representation and instruction.

that MOSA-Base achieves a 15.4% relative reduction in average WER compared to the Baseline-Base and consistently outperforms it across all languages. Remarkably, MOSA-Base surpasses the Baseline-Base even when trained with only 60% of its parameters. Similarly, MOSA-Large outperforms the Baseline-Large in average WER and demonstrates greater robustness to data imbalance. Ablation studies further indicate that MOSA is more effective at handling individual languages and learning both language-specific and shared linguistic knowledge. These findings support that, in LLM-based ASR, a mixture of simple adapters is more effective than a single, complex adapter design.

II. Method

A. Mixture of Experts

The Mixture-of-Experts (MOE) [19]–[23] architecture consists of two main components: a set of N experts $\{E_i\}_{i=1}^{N}$, where each expert is expected to learn specialized or shared knowledge, and a gating network or router G, which is responsible for dynamically determining the contribution or weight of each expert based on the input. Given an input x, the weights for the experts are computed as:

$$w = \text{SoftMax}(G(x)) \tag{1}$$

The final output is then obtained by computing a weighted sum of the expert outputs. In this work, we use softmax gating instead of a sparse MOE based on top-k selection:

$$y = \sum_{i=1}^{N} w_i E_i(x) \tag{2}$$

where w_i denotes the weight assigned to the i-th expert E_i.

B. Model Architecture

As shown in Figure 1, MOSA consists of a speech encoder, a pre-trained large language model, and a projector module incorporating a MoE mechanism.

Speech Encoder. For the speech encoder, we use the encoder component of Whisper-large-v3[1]. Whisper [25] is a multi-task, multilingual model based on an encoder-decoder architecture, which has achieved remarkable performance in both ASR and speech translation tasks through multi-task training strategies. Although Whisper is primarily trained for ASR and translation, its encoder representations capture rich information [26], and can even reconstruct original waveforms [27].

Projector. The Projector module aligns the speech input space with the LLM text input space. We apply the MoE mechanism introduced in Section II-A to this module. Specifically, the Projector consists of a Router and N Adapters, where each Adapter is capable of learning shared or language-specific information across different languages. Given a speech input x, the speech encoder E_a first extracts a hidden representation h_a. This representation is then fed into both the Router and the Adapter modules. The Router computes a weight vector w based on h_a, representing the contribution of each Adapter. Before being passed to the Adapters, h_a is downsampled through two convolutional layers with a stride of 2, resulting in h_{conv}. This operation allows each element in the h_{conv} sequence to incorporate information from neighboring frames. Each Adapter then maps h_{conv} into the input space of the LLM, yielding h_{adapt_i}. Finally, the aligned representation h_{adapt} is obtained by taking a weighted sum of h_{adapt_i} using the weights w.

Text Decoder. For the pretrained large language model, following the Ideal-LLM [16] framework, we adopt the Phi-3-mini-4k-instruct [28] model[2], which contains 3.8 billion parameters. This model demonstrates strong performance in language understanding, reasoning, and multilingual tasks. The aligned speech representation h_{adapt} is embedded into the user input section, forming the following structure: `<|user|>` $<h_{\text{adapt}}>$ `Instruction <|end|> <|assistant|> Transcription <|end|>`. The text labels and instruction prompt are tokenized using the LLM tokenizer and converted into input embeddings. Finally, the LLM processes the aligned representation to understand the audio input and, conditioned on the instruction, performs speech recognition. The cross-entropy loss is applied exclusively to the transcription tokens, ensuring that the model focuses on accurate speech recognition.

III. Experiments

A. Datasets

Following Ideal-LLM [16], we adopt the Multilingual LibriSpeech [29] dataset as our multilingual ASR benchmark. This dataset con-

[1] https://huggingface.co/openai/whisper-large-v3

[2] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

TABLE I
WER(%) RESULTS ON MULTILINGUAL LIBRISPEECH OF DIFFERENT MODELS. LOWER IS BETTER. IDEAL-LLM AND MOSA SHARE THE SAME LLM.

Model	Trainable Params (B)	en	de	nl	fr	es	it	pt	pl	Avg
LLaMA with ASR [12]	0.240	6.20	6.70	11.30	5.50	5.20	10.80	16.20	15.90	9.73
Ideal-LLM Base [16]	0.172	7.44	8.25	12.47	6.71	5.47	11.84	10.87	9.38	9.05
Ideal-LLM Large	0.303	6.15	7.12	11.23	5.40	4.26	9.93	12.41	**6.02**	7.81
MOSA-Base	0.155	5.91	**6.30**	**11.05**	5.27	4.47	9.99	9.68	8.61	7.66
MOSA-Large	0.287	**5.25**	6.33	11.50	**5.05**	**4.17**	**9.84**	**9.03**	8.84	**7.50**

tains approximately 50,000 hours of speech across eight languages: English (en), German (de), Dutch (nl), French (fr), Spanish (es), Italian (it), Portuguese (pt), and Polish (pl). Although the dataset is large in total volume, the distribution of hours among languages is highly imbalanced. Specifically, English accounts for 44,658.74 hours, German for 1,966.51 hours, Dutch for 1,554.24 hours, French for 1,076.58 hours, Spanish for 917.68 hours, Italian for 247.38 hours, Portuguese for 160.96 hours, and Polish for 103.65 hours. Before computing the WER (Word Error Rate), we apply Whisper's multilingual normalization[3], which replaces any non-standard markers, symbols, or punctuation with spaces while preserving diacritics.

B. Implementation Details

In MOSA, the speech encoder is based on the encoder of Whisper-large-v3. The output of the Whisper encoder, denoted as h_a, corresponds to 20 ms segments of the original audio. The `conv` module shown in Figure 1 consists of two convolutional layers, each with a kernel size of 3 and a stride of 2. Consequently, h_a is further downsampled to obtain h_{conv} with a temporal resolution of 12.5 Hz. Following prior work, we also apply the SpecAugment [30] data augmentation technique. For the large language model, we adopt Phi-3-mini-4k-instruct. The instruction prompt is: *Transcribe speech to text*. Following the Ideal-LLM [16] framework, both the speech encoder and the LLM are kept frozen during training.

To enable a fair comparison with Ideal-LLM, we implement both `base` and `large` variants of our model. The Adapter architecture is consistent across both variants: it consists of two linear layers with a ReLU [31] activation in between, projecting the hidden dimension from 3072 to 4096 and back to 3072. Compared to Transformer [32] or Q-former [33] based adapters commonly used in previous work, this is a much simpler design. The main difference between the two variants lies in the number of Adapters used: the `base` version includes 4 Adapters, whereas the `large` version uses 8. For the Router module, the `base` version is composed of two linear layers with a ReLU activation in between, mapping from 1280 to 512 and finally to 4 output dimensions. In contrast, the `large` version employs a deeper architecture with five linear layers and intermediate ReLU activations, with dimensions: $1280 \rightarrow 2560 \rightarrow 5120 \rightarrow 2560 \rightarrow 1280 \rightarrow 8$.

For training, we use the AdamW [34] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The maximum learning rate is set to $5\text{e}{-4}$ for the `base` version and $2\text{e}{-4}$ for the `large` version. We apply 2000 warmup steps followed by a linear decay learning rate scheduler. Training is conducted on 8 GPUs. Each device uses a batch size of 8 with gradient accumulation steps set to 16, processing approximately 1280 seconds of audio per device. The `base` version is trained for 11k steps, while the `large` version is trained for 36k steps. Following Ideal-LLM [16], both the `base` and `large` versions use the complete training data for the other seven languages. However, the base version only uses 10k hours of English data, while the large version uses the full 44k hours of English data. Finally, due to the extreme imbalance among languages in the dataset, we follow [35] to construct multilingual batches.

C. Main Results

Table I presents the WER (%) results of our proposed model MOSA and two recent baselines on the test sets of 8 languages from the Multilingual LibriSpeech dataset. The metric used is Word Error Rate (WER), where lower values indicate better performance. Among the baselines[4], Ideal-LLM [16] adapts to multilingual settings by assigning weights to a dual-encoder architecture using a weight selector, allowing different encoders to contribute differently depending on the input. LLaMA with ASR [12] integrates a CTC-trained encoder with LLaMA for speech recognition and explores several factors affecting performance, such as encoder size and stride, as well as LoRA-based fine-tuning of LLaMA. MOSA-Base achieves an average WER reduction of 21.3% and 15.4% compared to LLaMA with ASR and Ideal-LLM Base, respectively. It also outperforms both baselines across all 8 languages. Furthermore, MOSA requires fewer training parameters, making it a more efficient architecture that achieves state-of-the-art results. These results suggest that a single, large adapter may struggle to align multilingual representations with the LLM space. In contrast, combining multiple lightweight adapters, each responsible for either language-specific or shared knowledge, enables a simpler architecture to achieve superior performance. For MOSA-Large, the average WER is further improved. MOSA demonstrates a better ability to handle data imbalance. On the two languages with the least data, Polish and Portuguese, although its performance on Polish is slightly worse than that of Ideal-LLM Large, it significantly outperforms Ideal-LLM Large on Portuguese. Moreover, the degradation on Polish is marginal compared to MOSA-Base. And MOSA-Large still achieves better or comparable results than Ideal-LLM Large across other languages.

D. Ablation Study

Since the only difference in training data between the `base` and `large` versions lies in the English portion, the impact on other languages is minimal. Furthermore, the Large version requires significantly more computational resources, so we conduct the ablation studies on the `base` version.

A. Impact of the Number of Adapters. Table II presents the impact of varying the number of adapters on ASR performance. Regardless of whether 2, 3, 4, or 5 adapters are used, the average WER and monolingual WER consistently outperform the Ideal-LLM Base model. In other words, even with just two adapters (60% training

[3] https://github.com/openai/whisper/blob/main/whisper/normalizers/basic.py

[4] These results are taken from the Ideal-LLM paper.

TABLE II
WER(%) RESULTS ON MULTILINGUAL LIBRISPEECH OF DIFFERENT NUMBER OF ADAPTERS. LOWER IS BETTER.

#Adapter	Trainable Params (B)	en	de	nl	fr	es	it	pt	pl	Avg
1	0.079	6.32	6.82	11.86	5.46	4.55	10.64	9.38	10.51	8.19
1*	0.079	6.17	7.00	11.91	5.47	4.54	10.34	10.16	10.53	8.27*
2	0.104	6.12	6.46	11.46	5.45	4.53	10.55	8.95	9.26	7.85
3	0.130	6.32	6.36	11.32	5.29	4.64	9.99	9.04	9.01	7.75
4	0.155	5.91	6.30	11.05	5.27	4.47	9.99	9.68	8.61	7.66
5	0.180	5.97	6.23	11.14	5.10	4.43	9.94	9.95	9.16	7.74

parameters of Ideal-LLM Base), the performance surpasses that of Ideal-LLM Base. As the number of adapters increases, the average WER gradually decreases, and the performance across most individual languages also improves. However, using five adapters does not yield further improvement, likely due to the model becoming too large relative to the available training data. Furthermore, we investigated the performance of using a single adapter without incorporating a router. The results indicate that, across multiple languages, the single-adapter setup consistently underperforms compared to the multi-adapter configuration. This is particularly evident in low-resource languages such as Polish. The relatively strong performance on Portuguese may be attributed to the adapter being optimized for that specific language at checkpointing. To further examine this, we fine-tuned the single-adapter model individually on the data of each language, resulting in eight distinct systems (denoted as 1^*). However, the improvements observed were still marginal across all languages. These findings suggest that multiple adapters are capable of capturing language-specific or shared knowledge, thereby alleviating the burden on a single adapter and leading to better overall results.

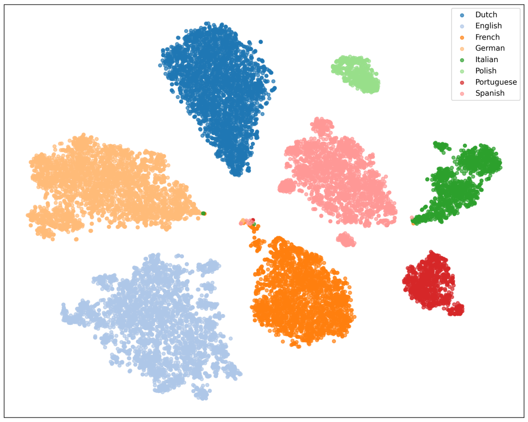

Fig. 2. T-SNE visualization of aligned speech embeddings.

B. Analysis of Speech Embeddings Across Languages. We conducted a distributional analysis of the speech embeddings generated by the projector of MOSA-Base. Specifically, we applied t-SNE [36] visualization on the test set of the MLS dataset, as shown in Figure 2. The results reveal that the embeddings from all eight languages in MLS form clearly separable clusters, with only a few outlier utterances. This indicates that the model is capable of treating each language similarly to a monolingual system. These findings further demonstrate that mixing multiple adapters can better accommodate multilingual settings.

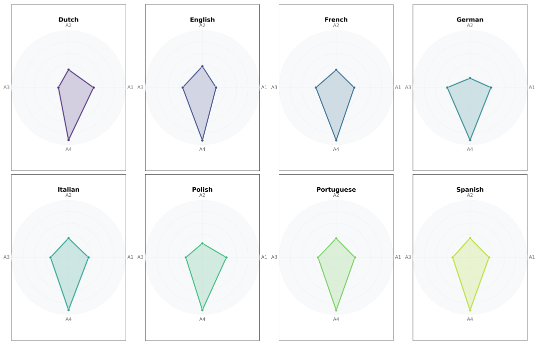

Fig. 3. Adapter weight distribution across languages.

C. Analysis of Adapter Weights Across Languages. We further analyzed the adapter weight distribution of MOSA-Base across different languages, reflecting each language's preference for specific adapters. The results, illustrated in Figure 3, show that all languages heavily utilize Adapter 4, while the remaining three adapters receive varying weights depending on the language. This suggests that there exists shared knowledge across languages, which is captured by the commonly used adapter. Meanwhile, the other adapters are likely responsible for learning language-specific knowledge, functioning as a mechanism to adjust or refine the shared representations.

IV. CONCLUSION

In this work, we observe that a single, complex adapter struggles to effectively adapt across multiple languages. To address this, we propose MOSA (Mixture of Simple Adapters), a method based on the MoE mechanism that combines multiple lightweight adapters. This design allows different experts to learn either language-specific or shared knowledge, enabling high-resource languages to benefit low-resource ones and alleviating the data scarcity problem in multilingual settings. Moreover, this approach simplifies the overall model architecture. Experimental results indicate that MOSA-Base achieves a 15.4% relative improvement in average WER over Baseline-Base and consistently outperforms it across all evaluated languages. Notably, MOSA-Base maintains superior performance even when trained with only 60% of the parameters used by Baseline-Base. Likewise, MOSA-Large surpasses Baseline-Large in average WER and exhibits enhanced robustness to data imbalance. Ablation studies further reveal that MOSA is more effective at handling individual languages and at learning both language-specific and shared linguistic features. These results suggest that, in LLM-based ASR, using a mixture of simple adapters is more effective than relying on a single, complex adapter architecture.

References

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[3] R. Larenz, "On overcoming the alleged alienation between christianity and physics," 2023.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[5] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[7] J. Peng, Y. Wang, Y. Fang, Y. Xi, X. Li, X. Zhang, and K. Yu, "A survey on speech large language models," *arXiv preprint arXiv:2410.18908*, 2024.

[8] Z. Ma, G. Yang, Y. Yang, Z. Gao, J. Wang, Z. Du, F. Yu, Q. Chen, S. Zheng, S. Zhang *et al.*, "An embarrassingly simple approach for llm with strong asr capacity," *arXiv preprint arXiv:2402.08846*, 2024.

[9] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[10] X. Geng, T. Xu, K. Wei, B. Mu, H. Xue, H. Wang, Y. Li, P. Guo, Y. Dai, L. Li *et al.*, "Unveiling the potential of llm-based asr on chinese open-source datasets," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 26–30.

[11] K.-T. Xu, F.-L. Xie, X. Tang, and Y. Hu, "Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration," *arXiv preprint arXiv:2501.14350*, 2025.

[12] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli *et al.*, "Prompting large language models with speech recognition abilities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 13 351–13 355.

[13] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[14] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[15] Y. Bai, J. Chen, J. Chen, W. Chen, Z. Chen, C. Ding, L. Dong, Q. Dong, Y. Du, K. Gao *et al.*, "Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition," *arXiv preprint arXiv:2407.04675*, 2024.

[16] H. Xue, W. Ren, X. Geng, K. Wei, L. Li, Q. Shao, L. Yang, K. Diao, and L. Xie, "Ideal-llm: Integrating dual encoders and language-adapted llm for multilingual speech-to-text," *arXiv preprint arXiv:2409.11214*, 2024.

[17] 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," *arXiv preprint arXiv:2409.19878*, 2024.

[18] Z. Yu, Y. Zhang, K. Qian, C. Wan, Y. Fu, Y. Zhang, and Y. C. Lin, "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.

[19] 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," *arXiv preprint arXiv:1701.06538*, 2017.

[20] W. Fedus, J. Dean, and B. Zoph, "A review of sparse expert models in deep learning," *arXiv preprint arXiv:2209.01667*, 2022.

[21] B. Zoph, I. Bello, S. Kumar, N. Du, Y. Huang, J. Dean, N. Shazeer, and W. Fedus, "St-moe: Designing stable and transferable sparse expert models," *arXiv preprint arXiv:2202.08906*, 2022.

[22] Y. Xie, S. Huang, T. Chen, and F. Wei, "Moec: Mixture of expert clusters," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 807–13 815.

[23] X. Song, D. Wu, B. Zhang, D. Zhou, Z. Peng, B. Dang, F. Pan, and C. Yang, "U2++ moe: Scaling 4.7 x parameters with minimal impact on rtf," *arXiv preprint arXiv:2404.16407*, 2024.

[24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[26] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," *arXiv preprint arXiv:2307.03183*, 2023.

[27] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[28] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[29] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[30] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[31] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[33] S. Kim, A. Lee, J. Park, A. Chung, J. Oh, and J.-Y. Lee, "Towards efficient visual-language alignment of the q-former for visual reasoning tasks," *arXiv preprint arXiv:2410.09489*, 2024.

[34] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[35] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.