Triple X: A LLM-Based Multilingual Speech Recognition System for the INTERSPEECH2025 MLC-SLM Challenge

Miaomiao Gao[13*], *Xiaoxiao Xiang*[2*], *Yiwen Guo*[4†]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences
[2]LIGHTSPEED
[3]University of Chinese Academy of Sciences
[4]Independent Researcher

xiangxiaoxiao18@mails.ucas.ac.cn, gaomiaomiao20@mails.ucas.ac.cn

Abstract

This paper describes our Triple X speech recognition system submitted to Task 1 of the Multi-Lingual Conversational Speech Language Modeling (MLC-SLM) Challenge. Our work focuses on optimizing speech recognition accuracy in multilingual conversational scenarios through an innovative encoder-adapter-LLM architecture. This framework harnesses the powerful reasoning capabilities of text-based large language models while incorporating domain-specific adaptations. To further enhance multilingual recognition performance, we adopted a meticulously designed multi-stage training strategy leveraging extensive multilingual audio datasets. Experimental results demonstrate that our approach achieves competitive Word Error Rate (WER) performance on both dev and test sets, obtaining second place in the challenge ranking.

Index Terms: Speech recognition, multilingual conversational environments, multi-stage training

1. Introduction

Speech recognition is the task of transcribing speech into text. It plays a critically important role in a wide range of applications, including human-computer interaction, voice assistants, real-time transcription, and content creation. High-accuracy speech recognition systems can significantly enhance user experience and accessibility, especially in multilingual and conversational contexts.

Classic end-to-end automatic speech recognition (ASR) frameworks have demonstrated great success in recent years. Representative methods include Paraformer [1], OWSM v3.1 [2], FireRedASR-AED [3], and Whisper [4]. These models typically follow an encoder-decoder paradigm and can be categorized into several mainstream modeling approaches: connectionist temporal classification (CTC) [5], recurrent neural network transducer (RNN-T) [6], Recurrent Neural Aligner (RNA) [7], and encoder-decoder approaches [8]. All of these methods aim to learn the complex mapping between sequences of acoustic features and sequences of textual tokens by leveraging large-scale paired speech and text datasets.

Recently, text-based large language models (LLMs) have demonstrated outstanding performance across a wide range of downstream tasks, including machine translation, question answering, and long-form text generation. Models such as DeepSeek [9], GPT [10], Qwen [11], and LLaMA [12] have emerged as foundational models for natural language understanding and generation due to their ability to capture rich linguistic and contextual knowledge from massive text corpora. Inspired by the success of pre-trained LLMs in textual domains, recent studies have explored integrating their reasoning and generation capabilities into ASR pipelines. Notable examples include Qwen-Audio [13] and FireRedASR-LLM [3], which improve performance in semantically complex or noisy conditions. However, despite these advances, existing LLM-augmented ASR systems have not sufficiently addressed the challenges posed by real-world multilingual conversational scenarios, which involve code-switching, speaker diversity, and informal speech patterns. This highlights the need for more robust architectures and training strategies specifically tailored to multilingual, conversational ASR.

Task 1 of the MLC-SLM Competition aims to develop an LLM-based ASR system that improves speech recognition accuracy in multilingual conversational scenarios. To this end, we adopt an Encoder–Adapter–LLM architecture that leverages the capabilities of large language models (LLMs). The Encoder extracts rich acoustic and semantic representations from speech, while the adapter bridges the encoder output to the semantic space of the LLM. The LLM then generates transcriptions by interpreting both the audio-derived features and a given task instruction. By utilizing LLMs, Triple X taps into their advanced text processing capabilities and reasoning potential, enabling more accurate speech-to-text conversion and better adaptation to diverse linguistic patterns and contexts.

Using this approach, we achieve word error rates of 9.73% and 9.67% on the validation and test sets, respectively, securing second place on the official leaderboard. These results demonstrate the effectiveness of our architecture in improving ASR performance in multilingual settings, showing competitive results compared to other state-of-the-art models.

2. Approach

In this section, we first introduce the network architecture, followed by a description of the dataset used in our experiments. Finally, we detail the experimental setup, including the training strategies, input features, and loss functions.

2.1. Network Architecture

Our proposed Triple X system adopts the widely used Encoder-Adapter-LLM architecture, as illustrated in Figure 1. Specifically, we employ the Whisper-large-v3 encoder to extract rich acoustic and semantic features from the input speech. This encoder follows a standard Transformer architecture. However, the output sequence of the encoder is longer than that of text,

[*] Equal contribution.
[†] Corresponding authors.

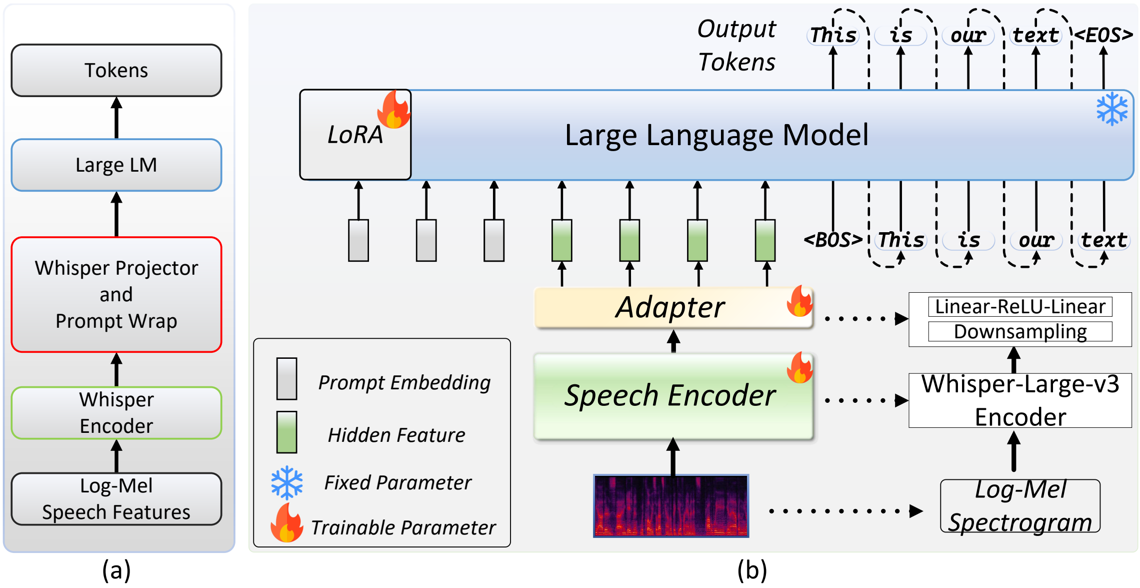

Figure 1: *The schematic diagram of the Triple X architecture, which includes an encoder, an adapter and an text-based LLM.*

which can negatively impact the processing efficiency of the LLM. To reduce the sequence length and align the output dimension of the audio encoder with the input embedding dimension of the pre-trained text-based LLM, our adapter first applies a downsampling module to reduce the sequence length, followed by a Linear–ReLU–Linear transformation to map the output semantic information from of the encoder into the semantic space of the LLM. Notably, we use the simplest frame splicing for the downsampling module, as we have found that different downsampling methods yield similar results. For the LLM component in Triple X, we initialize it with pretrained weights from Qwen-3B, a high-quality open-source LLM developed by Alibaba. As shown in Figure. 1, the input to the LLM includes the output features of the encoder and the user prompt.

2.2. Datasets

In our experiments, we use two types of training sets. The first training set consists of approximately 1,500 hours of multilingual conversational speech data provided by the competition organizers. It covers around 11 languages, including English, French, German, Italian, Portuguese, Spanish, Japanese, Korean, Russian, Thai, and Vietnamese. The English portion includes about 500 hours of recordings from diverse regions, including British, American, Australian, Indian, and Philippine English, while each of the other languages contributes approximately 100 hours. We apply oracle segmentation and use speaker labels for each conversation to segment long utterances into shorter ones. The second training set is constructed from publicly available datasets, including GigaSpeech2 [14], KsponSpeech [15], Reazonspeech [16], and Multilingual LibriSpeech [17]. We have selected 30,000 hours of audio data from these datasets, and statistical information regarding the amount and language of data are shown in Figure. 2.

To evaluate the model, we use the development and evaluation sets provided by the competition organizers, with the evaluation set containing 4 hours of recordings for each language.

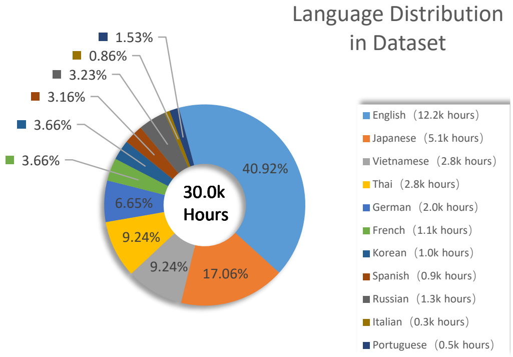

Figure 2: *Illustration of the dataset distribution across different languages and their respective data volumes.*

2.3. Experiment Setup

We adopt a carefully designed three-stage training strategy to improve multilingual speech recognition accuracy. First, we fine-tune Whisper-large-v3 and use the resulting encoder weights to initialize the encoder of Triple X. This enhances the encoder's speech feature representation capabilities and facilitates faster convergence in subsequent training stages. Next, we freeze the encoder and train the adapter parameters to align the semantic information embedded in the encoded representations with the semantic space of the LLM. Finally, we apply trainable Low-Rank Adaptation (LoRA) to fine-tune the LLM while keeping its core parameters fixed. This approach strikes a balance between adaptability and preservation of pre-trained knowledge. For input speech, similar to conventional end-to-end ASR systems, we apply SpecAug [18] and speed perturbation [19] for data augmentation. We extract 128-dimensional log-Mel spectrograms as input feature of the encoder with a window of 25ms, a hop length of 10ms, without applying global mean and variance normalization. During training, cross-entropy loss is used and computed only at positions cor-

Table 1: *Detailed WER (%) results of different LLM backbone on the official evaluation set for various beam sizes. The **BOLD** values show the best results.*

System	beam size	WER (%) ↓
Qwen3-8B	2	10.83
Qwen3-8B	3	10.75
Qwen3-8B	5	10.71
Qwen3-8B	8	10.70
Qwen3-8B-Base	2	10.56
Qwen3-8B-Base	3	10.47
Qwen3-8B-Base	5	10.43
Qwen3-8B-Base	8	**10.41**
Qwen3-8B-Base	10	10.42

Table 2: *WER(%) results on interspeech 2025 MLC-SLM Task 1 evaluation set. The **BOLD** values show the best results.*

System	WER (%) ↓
MLC-SLM Baseline	20.17
Triple X	**9.67**

responding to the text transcription.

3. Evaluation

We initially optimized the model using the training set provided by the competition organizers to facilitate rapid model selection and performance validation. Table.1 presents the results of Qwen3-8B and Qwen3-8B-Base on the evaluation set, revealing several key insights. First, Qwen3-8B-Base consistently achieved higher speech recognition accuracy than Qwen3-8B, as evidenced by lower WER scores across various beam settings. This suggests that the base version may serve as a more effective backbone for speech recognition tasks. Second, increasing the beam size initially improved recognition accuracy but later led to a decline, with the best performance (lowest WER) observed at a beam size of 8. Therefore, to balance computational efficiency and recognition accuracy, we adopt a beam size of 8 as the optimal setting in subsequent experiments.

Even without incorporating additional data, the aforementioned models already achieved impressive performance, surpassing 80% of the participants. To further enhance our results, we collected a substantial amount of publicly available datasets to better map the semantic information from the encoded representation into the semantic space of the LLM. The distribution of these datasets is illustrated in Figure 2. After pre-training, we fine-tuned both the adapter and LoRA modules using the official training set with a reduced learning rate. As shown in Table 2, the proposed method achieved a WER of 9.67% on the official evaluation set, corresponding to a recognition accuracy of 90.33%. This represents a 13.15% improvement over the baseline accuracy of 79.83%. Overall, our model achieved WERs of 9.73% and 9.67% on the validation and evaluation sets, respectively, securing second place on the competition leaderboard.

4. Conclusions

We have developed a multilingual speech recognition system named Triple X, which leverages a LLM. By employing a multi-stage training strategy, our system achieved a WER of 9.67 on the MLC-SLM evaluation set, securing second place on the leaderboard. For future endeavors, we plan to collect more extensive multilingual conversational datasets to further enhance recognition accuracy in multilingual dialogue settings. Additionally, we aim to extend our current ASR model to support both speech recognition and response generation within a unified framework.

5. References

[1] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[2] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan, Y. Sudo, M. Shakeel, K. Choi, J. Shi, X. Chang *et al.*, "Owsm v3. 1: Better and faster open whisper-style speech models based on e-branchformer," *arXiv preprint arXiv:2401.16658*, 2024.

[3] 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.

[4] 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.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[6] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[7] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping." in *Interspeech*, vol. 8, 2017, pp. 1298–1302.

[8] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[9] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[11] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[12] 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.

[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. Yang, Z. Song, J. Zhuo, M. Cui, J. Li, B. Yang, Y. Du, Z. Ma, X. Liu, Z. Wang *et al.*, "Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement," *arXiv preprint arXiv:2406.11546*, 2024.

[15] J.-U. Bang, S. Yun, S.-H. Kim, M.-Y. Choi, M.-K. Lee, Y.-J. Kim, D.-H. Kim, J. Park, Y.-J. Lee, and S.-H. Kim, "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, vol. 10, no. 19, p. 6936, 2020.

[16] Y. Y. D. M. S. Fujimoto, "Reazonspeech: A free and massive corpus for japanese asr," 2016.

[17] 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.

[18] 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.

[19] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition." in *Interspeech*, vol. 2015, 2015, p. 3586.