Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST36285

Lightweight LLM-based Speech Recognition via KAN Adapters

Yuxi Li[1*], Yan Wang[2]

[1]College of Arts and Sciences, The Ohio State University, Columbus, USA
[2]Glasgow college, University of Electronic Science and Technology of China, Chengdu, China
**Corresponding Author. Email: li.16334@osu.edu*

Abstract. In recent years, the combination of large language model (LLM) and pre-trained voice encoder has shown great potential in the field of automatic speech recognition (ASR). However, bridging the modal communication between acoustic characterization and language embedding often requires a large number of training parameters, which makes it difficult for them to apply in environments with limited resources. This study proposes to use the Kolmogorov-Arnold network (KANs) as a simplified adapter for the automatic speech recognition (ASR) system based on the Large Language Model (LLM). And by introducing a KAN adapter between the pre-trained voice encoder and TinyLlama-1.1B, the system improves the correspondence between acoustic characterization and language characterization with very few training parameters. The experimental results show stable optimization characteristics, with a word error rate (WER) of 16.79% and a character error rate (CER) of 10.46%. These results highlight the potential of KAN-based adapters in ASR systems with limited resources and parameters. The KAN-based adapter provides a promising and parameter-efficient solution for matching acoustic and language scenarios. In another words, in the resource-limited automatic speech recognition (ASR) scenario, which is crucial to computing efficiency and training stability, it shows significant advantages.

Keywords: Automatic Speech Recognition (ASR), Kolmogorov-Arnold Networks (KAN), Large Language Models (LLMs), Parameter-Efficient Learning

1. Introduction

Automatic speech recognition (ASR) technology has greatly benefited from the rapid development of large language models (LLMs), which provide powerful insights into language structure and improve spelling in a variety of language tasks. However, most LLM-based ASRs rely on large backbone models and computational complexity, making them difficult to implement in resource-limited environments. As natural language hardware and inference optimization applications gain acceptance, optimizing sound models with simple LLMs, which while retaining satisfactory insights, has become a major challenge. In order to solve this problem, we have studied the application of Kolmogorov-Arnold network (KAN) as fast adapters in the automatic speech recognition (ASR) framework based on large language model (LLM). Unlike the traditional multi-layer perception machine (MLP) that relies on a set of functions, the Kolmogorov-Arnold network

DOI: 10.54254/2755-2721/2026.AST36285

(KAN) uses discrete functions; and these functions show stronger characterization ability and require less training data. This feature makes them very suitable for integrating speech language processing systems into high-quality speech recognition systems. For this study, we developed an automatic voice recognition (ASR) system by combining KAN-based adapters with pre-trained voice encoders and TinyLlama-1.1B. We evaluated the proposed method on the voice recognition task, and evaluated its accuracy and error rate. Experimental results show that under minimalist conditions, the KAN adapter shows excellent connectivity and efficiency, highlighting its potential to achieve good connectivity in ASR systems based on small language mapping (SLM).

2. Related work

Recent advances in augmented speech recognition (ASR) have enabled the development of the more sophisticated speech and language systems described earlier. HuBERT has emerged as a popular self-supervised speech encoder due to its ability to learn complex harmonic concepts from large amounts of unstructured speech [1]. In terms of natural language processing, LLaMA images show excellent natural language comprehension and reproduction, making them suitable for advances in natural language processing tasks [2]. To build a large pre-trained model, LoRA introduces a low-parameter architecture that significantly reduces the number of trainable parameters while maintaining good performance [3]. Accordingly, some studies have focused on the combination of speech encoding and speech representation for meaning comprehension. Hono and colleagues. developed a system that specifically combines pre-trained speech and vocabulary, providing efficiency through a simple learning module [4]. My mother and her friends. also showed that a simple but well-designed topic-LLM system can improve the performance of ASR using simple architectural problems [5]. Furthermore, Whisper builds a robust image from large-scale ASR and demonstrates the benefits of using deep voice recognition in speech extraction [6]. In recent years, Kolmogorov–Arnold networks (KANs) have emerged as an alternative to the classical multiple neural networks (MLPs), which use learnable nonlinear functions instead of fixed functions [7]. Due to its high spatial resolution and good parameter control, KAN has attracted considerable attention in speech recognition applications. There are others. investigated a KAN-based architecture and showed comparable performance to neural layers for spoken language recognition [8]. The city. The KAN method has improved in terms of speech accuracy, showing advantages in representing complex speech patterns [9]. Based on these results, we investigated whether pitch synchronization could be achieved in a simple LLM-based ASR system using the KAN adapter. Unlike previous studies that have focused on speech or language processing, this study investigates the potential of KAN adapters for speech recognition tasks using the robust TinyLlama system in a complex learning context.

3. Method

3.1. Overall architecture

The system follows an automatic speech recognition (ASR)-based large-scale language model (LLM), which consists of three stages: pretrained speech segmentation, KAN-based speech segmentation, and speech segmentation method. As shown in Figure 1, the embedded speech is first processed by a pre-trained HuBERT embedder to extract voice features; these samples were transformed with a low-resistance KAN adapter and imported into TinyLlama-1.1B, which was used

DOI: 10.54254/2755-2721/2026.AST36285

as a transcription factor. In order to improve the transfer and reduce the computational cost, this project uses LoRA in its fine-tuning stage.

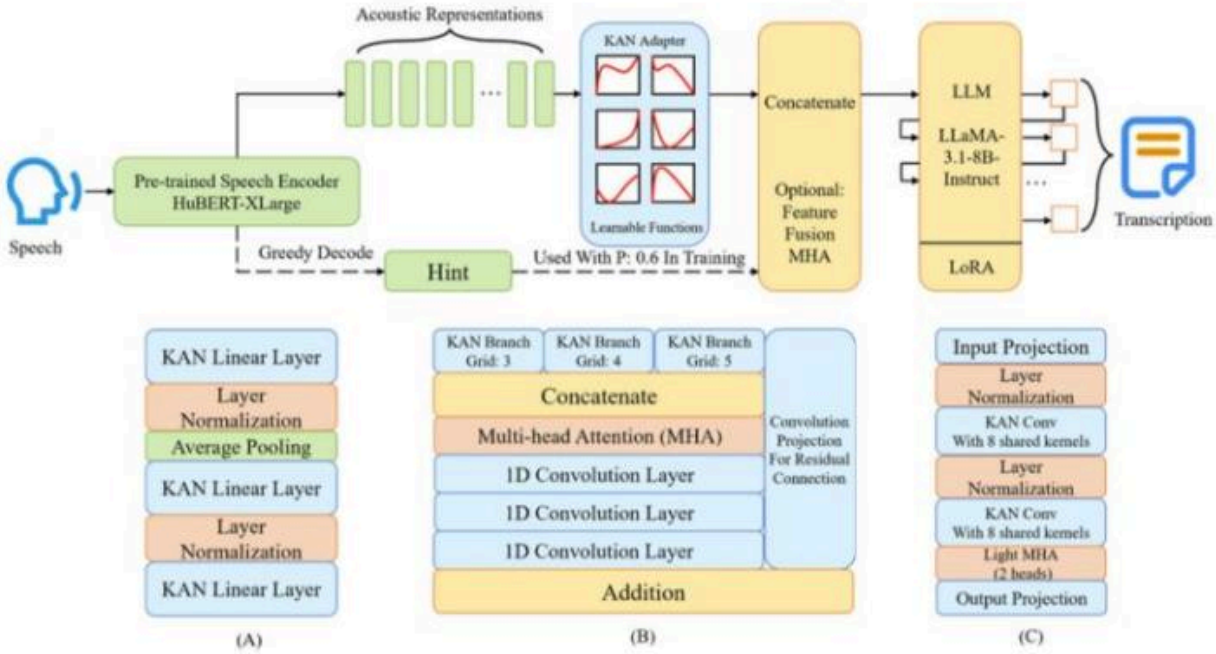

Figure 1. Model's general architecture and design of 3 KAN adapters

3.2. KAN-based adapter

This adapter acts as a bridge between the acoustic representation generated by the speech encoder and the token input required by the language model. Traditional automatic speech recognition (ASR) systems based on large-scale language models (LLM) usually use multi-layer perception machines (MLP) to perform this mapping operation. In contrast, we replace the MLP mechanism with an adapter based on the Kolmogorov-Arnold network (KAN). KAN uses the learned nonlinear functions to model complex transformations and shows excellent performance in various voice-related tasks [7-9]. In this study, the adapter is made up of a simple KAN convolutional layer combined with standard matching functions and operations. We choose to share the core design to reduce the amount of parameter calculations while maintaining the ability to identify local acoustic patterns. The adapter achieves efficient matching between the acoustic domain and the language domain without significantly increasing the number of parameters to be trained.

3.3. Training strategy

The HuBERT encoder is used to extract audio features, while TinyLlama-1.1B is large language model. During this training process, most of the pre-training parameters remain frozen, and only the KAN adapter and LoRA parameters are adjusted, so as to fine-tune with fewer computing resources. And the model is optimized by AdamW optimizer and cosine learning rate algorithm. For the training process, the training begins with a short warm-up stage, followed by a gradual decrease in the learning rate; this method ensures the stability of the optimization process and helps the model to

DOI: 10.54254/2755-2721/2026.AST36285

converge smoothly throughout the training process. In the process of reasoning, the language model generates the final output directly based on the processed phonetic characterization.

4. Experiments

4.1. Dataset

This experiments were conducted using the LibriSpeech corpus, which is a large dataset widely used in speech recognition [10]. And the dataset consists of approximately 1,000 hours of English speech from audio recordings, sampled at 16 kHz. To evaluate the performance of the proposed framework on non-users, the "train-clean-100" dataset which contains 100 seconds of recorded data and also is used for training; the "dev-clean" version is used for validation, and the "test-clean" version is used for decision-making. This framework allows us to study the benefits of KAN-based adaptation with limited training resources.

4.2. Experimental setup

The system is implemented based on PyTorch and trained on a single NVIDIA GPU. The HuBERT model is used as a voice encoder, and TinyLlama-1.1B is used as the basic language model; a KAN-based adapter is constructed between the two to align acoustic characterization and language characterization. In order to improve parameter efficiency, most of the pre-training parameters are frozen, and the LoRA method is adopted in the subsequent training (fine-tuning) process. The optimization process adopts AdamW algorithm combined with cosine learning rate scheduling. The effective batch size is set to 4, and the training process contains about 170,000 optimization steps. And the periodic verification is carried out during the training process to monitor the convergence of the model. The model performance is evaluated by standard speech recognition indicators, which are word error rate (WER) and character error rate (CER).

4.3. Training dynamics

Figure 2 summarizes the training dynamics of the proposed system in 15 epoch. The training error rate gradually decreased, which is from about 2.7 to 0.3, while the verification error rate decreased from about 1.9 to less than 0.2. This continuous downward trend shows that the model has effectively learned from the data and maintained a stable convergence during the training process. The word error rate (WER) also decreased significantly, from more than 0.9 in the initial training stage to about 0.17 at the end of the training; at the same time, the character error rate (CER) also decreased from about 0.75 to 0.10. These trends show that with the training, the recognition accuracy of the model at the level of words and characters has gradually improved. The learning rate curve shows a steady downward trend, ensuring the optimization and stability of the whole training process. In general, the training curve shows that the proposed adapter based on the KAN architecture can effectively align the acoustic characterization with the "TinyLlama" language model, even under the condition of limited resources.

DOI: 10.54254/2755-2721/2026.AST36285

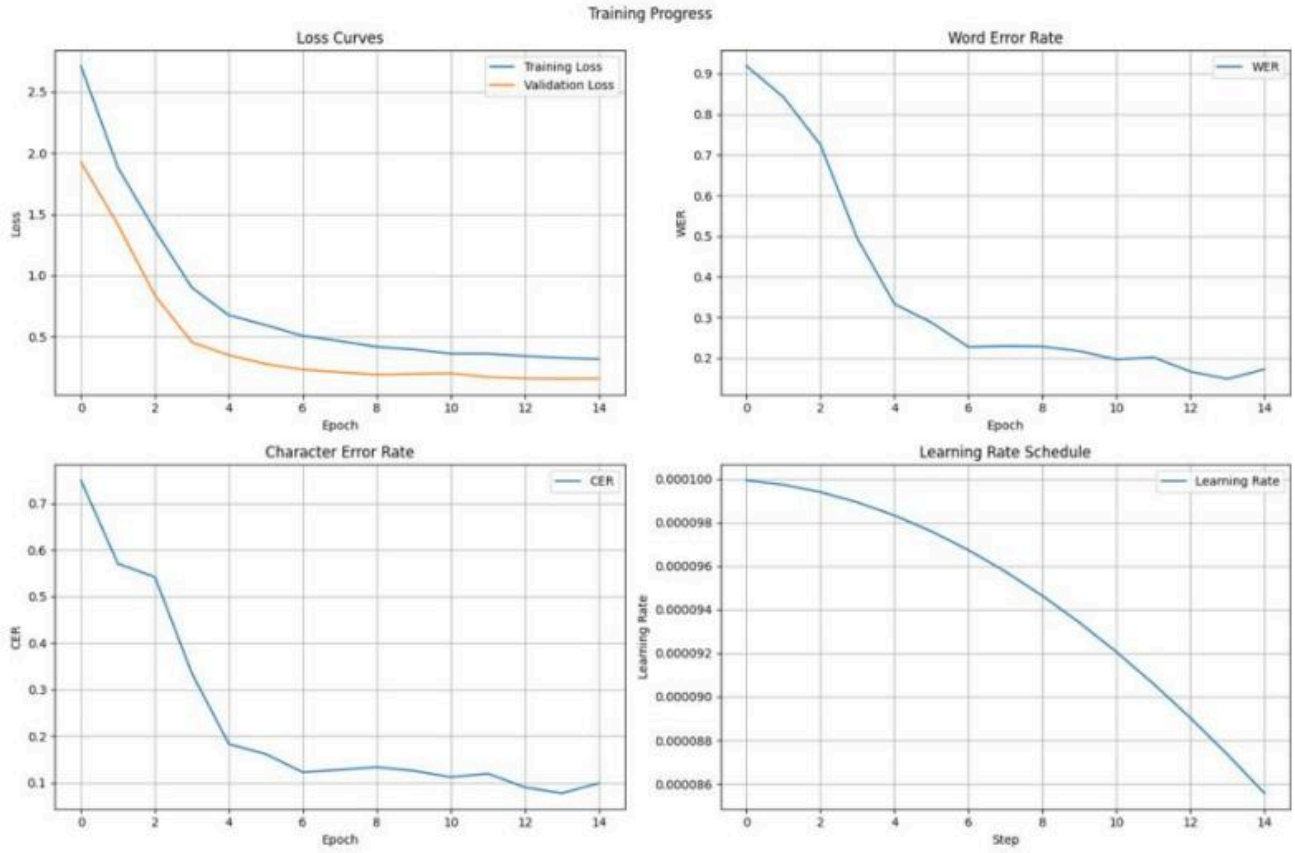

Figure 2. The proposed training dynamics of the KAN-ASR framework

Notes: This figure shows the loss, word error rate (WER), character error rate (CER) and learning rate scheduling changes with the training round (epoch) in the process of training and verification. The loss, WER and CER curves all show a trend of continuous improvement, indicating that the model has achieved stable convergence.

4.4. Recognition performance

The final identification performance of the proposed KAN-ASR framework is shown in Table 1.

Table 1. Recognition performance on the LibriSpeech test set

Metric	Value
Test Loss	0.187
WER (%)	16.79
CER (%)	10.46

The loss function value of the developed system on the test set is 0.187, the word error rate (WER) is 16.79%, and the character error rate (CER) is 10.46%. Because the system adopts the TinyLlama-1.1B language model with a large number of parameters and only uses the train-clean-100 subset for training, these results verify the parameter-efficient language recognition performance based on the large language model (LLM) and combined with the KAN adapter. As shown in the learning curve shown in Figure 2, the model shows strong convergence throughout the optimization process. The gradual decrease in the values of the loss function, WER, and CER

DOI: 10.54254/2755-2721/2026.AST36285

indicates that the KAN adapter can establish an efficient mapping between the acoustic representation produced by the speech encoder and the local space of the speech pattern.

Despite the limitations of quantitative metrics compared to large-scale ASR systems， which are rained on large data sets and large speech models, and this approach provides computationally efficient solutions to speech recognition problems under resource-limited conditions.

4.5. Error analysis

In order to better understand the limitations of the experimental data, we analyzed the overall identification errors on the test set. Table 2 provides some examples of model output.

Table 2. Representative transcription errors from the test set

Model Output	Reference	Error Type
countled	counselled	Rare word substitution
saver	Xavier	Proper noun recognition
saint francis savior	saint francis xavier	Proper noun recognition
equation	requisition	Semantic substitution
cold loose indifference range	cold lucid indifference reigned	Multi-word substitution

The first mistake is to confuse unfamiliar words with familiar words with similar meanings, resulting in misunderstandings. For example, the word "recommended" may be mistaken for a word or expression with a similar meaning; This phenomenon highlights the complexity of high-level vocabulary classification for compact sentence models. The second common mistake involves miscognizing proper nouns and other nouns. For example, "Xavier" may be mistaken for "the savior", and "St. Francis Xavier" may be mistakenly transcribed as "St." Such errors are often caused by the limited ability of speech recognition systems to accurately model and identify unfamiliar words and proper nouns. The third type of error involves semantic substitution, which means that although the highlighted words seem reasonable, they do not match the actual words. For example, translating "request" as "equation" shows that the language model sometimes prioritizes common words when decoding. Finally, when several adjacent words in the sentence are misunderstood, errors will occur; this means that acoustic deviations may accumulate in distant fragments, resulting in distortion of spatial resolution.

In summary, the error analysis clearly shows that most of the knowledge involved contains synonymous combinations of words, adjectives and adverbs. Future research can further strengthen this work by introducing language models, training data or decoding methods that provide phonetic information.

5. Conclusion

This reaserch discusses the application of Kolmogorov-Arnold network (KAN) in automatic speech recognition (ASR) systems based on small-scale large-scale language models (LLM). By integrating a KAN-based adapter between the pre-trained HuBERT model and TinyLlama-1.1B, the system achieves efficient alignment between voice and language while reducing computing costs. The experimental results on the LibriSpeech data set show stable performance, achieving a word error rate of 16.79% (WER) and a character error rate of 10.46% (CER). And the error analysis shows that

DOI: 10.54254/2755-2721/2026.AST36285

identification failure is mainly related to virtual words, proper nouns and word sequences with similar phonetic characteristics. All these research results show that in an automatic speech recognition (ASR) system based on a large language model (LLM), the KAN adapter provides an alternative to be more efficient than the traditional projection layer. Future work will explore more powerful language models and refined alignment strategies to further improve the recognition accuracy.

References

[1] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 3451–3460.

[2] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The Llama 3 herd of models. *arXiv:*2407.21783.

[3] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv:*2106.09685.

[4] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov-Arnold networks. *arXiv:*2404.19756.

[5] Hono, Y., Mitsuda, K., Zhao, T., Mitsui, K., Wakatsuki, T., & Sawada, K. (2024). Integrating pre-trained speech and language models for end-to-end speech recognition. *arXiv:*2312.03668.

[6] Ma, Z., Yang, G., Yang, Y., Gao, Z., Wang, J., Du, Z., Yu, F., Chen, Q., Zheng, S., Zhang, S., & Chen, X. (2024). An embarrassingly simple approach for LLM with strong ASR capacity. *arXiv:* 2402.08846.

[7] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv:* 2212.04356.

[8] Koudounas, A., La Quatra, M., Pastor, E., Siniscalchi, S. M., & Baralis, E. (2025). "KAN you hear me?" Exploring Kolmogorov-Arnold networks for spoken language understanding. *arXiv:* 2505.20176.

[9] Li, H., Hu, Y., Chen, C., Siniscalchi, S. M., Liu, S., & Chng, E. S. (2025). From KAN to GR-KAN: Advancing speech enhancement with KAN-based methodology. *arXiv:* 2412.17778.

[10] Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). LibriSpeech: An ASR corpus based on public domain audio books. In *Proceedings of ICASSP 2015*, 5206–5210. doi: 10.1109/ICASSP.2015.7178964