Integrated Method of Deep learning and Large Language Model in Speech Recognition

Bo Guan , Jin Cao , Bingjie Huang , Zhuoyue Wang , Xingqi Wang , Zixiang Wang *

Posted Date: 19 July 2024

doi: 10.20944/preprints202407.1520.v1

Keywords: deep learning; large language model; speech recognition; ensemble method

Article

Integrated Method of Deep learning and Large Language Model in Speech Recognition

Bo Guan [1], Jin Cao [2], Bingjie Huang [3], Zhuoyue Wang [4], Xingqi Wang [5] and Zixiang Wang [6,*]

[1] Virginia Tech, Blacksburg, VA, USA; jasonguan0107@gmail.com
[2] Independent Researcher, Dallas, TX, USA; caojinscholar@gmail.com
[3] Cornell University, Sunnyvale, CA, California; bingjiehuang1998@gmail.com
[4] University of California, Berkeley, Berkeley, CA, USA; zhuoyue_wang@berkeley.edu
[5] Johns Hopkins University, Baltimore, MD, USA; wxq19991001@gmail.com
[6] Syracuse University, Syracuse, NY, USA
* Correspondence: zwang161@syr.edu

Abstract: This research aims to explore the integration method of deep learning and large language models in speech recognition to improve the system's recognition accuracy and ability to handle complex contexts. Deep neural network (DNN), convolutional neural network (CNN), long short-term memory network (LSTM) and Transformer-based large language model are used to build an integrated acoustic and language model framework. Experiments on TIMIT, LibriSpeech and Common Voice datasets show that the ensemble model shows significant improvements in both word error rate (WER) and real-time factor (RTF) compared to traditional models. Especially in terms of adaptability to multiple languages and accent changes, the model shows superior performance. The conclusion shows that through technology integration, the performance of the speech recognition system in complex environments can be effectively improved, providing a new direction for the future development of speech recognition technology.

Keywords: deep learning; large language model; speech recognition; ensemble method

I. Introduction

In today's digital era, speech recognition technology has become one of the key technologies to improve the efficiency of human-computer interaction and is widely used in fields such as intelligent assistants, autonomous vehicle communication systems, and multilingual translation. With the rapid development of deep learning technology and large language models, their application potential in improving speech recognition accuracy and processing complex contexts has been gradually explored [1]. Deep learning technology, through its powerful feature extraction capabilities, has significantly improved the performance of acoustic models, while large language models optimize the speech-to-text conversion process by understanding context and semantics. However, how to effectively integrate these two technologies to fully utilize their respective advantages and improve the overall performance of the system is still an urgent challenge to be solved [2].

This article aims to explore the integration method of deep learning and large language models in speech recognition. Through in-depth analysis of existing technologies and experimental verification of integration strategies, a new solution is proposed in order to achieve higher recognition accuracy and real-time performance, providing theoretical basis and practical guidance for technological progress and application promotion in related fields.

II. Theoretical Overview

A. Basic Speech Recognition Technology

 doi:10.20944/preprints202407.1520.v1

As shown in Figure 1, basic speech recognition technology involves the process of converting speech signals into text and is a core component of human-computer interaction systems. This technology mainly consists of two major components: acoustic model and language model. The acoustic model is responsible for parsing and identifying the acoustic features in the speech signal and converting it into a sequence of phonemes or phonetic symbols, while the language model applies grammatical and semantic rules on this basis to infer the most likely word sequence [3] . Traditionally, this process relies on hidden Markov models (HMM) to handle acoustic modeling and time series prediction, while language models often use n-gram statistical models to predict the probability of word sequences.

With the development of technology, basic speech recognition technology has been able to support multiple languages and dialects, cope with various noise interferences, show high flexibility and accuracy, and provide users with a smoother and more natural interactive experience [4].

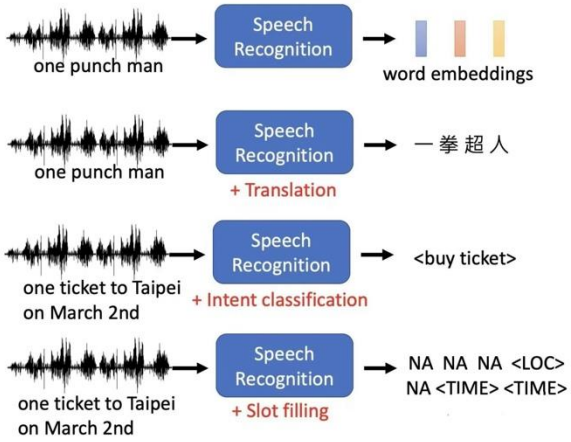

Figure 1. Principle of Speech Recognition Technology.

B. Application of Deep Learning Technology in Speech Recognition

Deep learning techniques have revolutionized the field of speech recognition, providing a powerful method to capture complex patterns and features in speech data. By using multi-layer neural networks, deep learning models such as deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN) can effectively learn the temporal and spatial dependencies of speech signals [5].

In particular, recurrent neural networks are widely used in acoustic models due to their advantages in processing time series data, allowing the system to better understand and predict the dynamic characteristics of speech streams. In addition, deep learning models can automatically extract meaningful features from large amounts of data without the need for complex manual feature design, which greatly improves the efficiency and accuracy of speech recognition systems. For example, in the CNN-LSTM-DNN (CLDNN) framework, the CNN layer first processes the input features to extract local dependencies, then processes the time series information through the long short-term memory network (LSTM) layer, and finally performs feature classification through the DNN layer.

C. The Role of Large Language Models in Speech Recognition

The role of large language models in speech recognition is mainly reflected in their ability to process and understand complex language structures, thereby improving the accuracy and naturalness of speech-to-text conversion. These models, including Transformer, BERT, and GPT, use deep network architectures to capture long-distance dependencies, allowing the models to not only understand individual words, but also grasp the context of entire sentences and even paragraphs [10–12].

 doi:10.20944/preprints202407.1520.v1

By pre-training and fine-tuning these models, they can effectively adapt to specific speech recognition tasks and handle the diversity and complexity of various linguistic expressions [9]. In addition, the use of large language models significantly improves the ability to recognize nuances in language. For example, the processing of homonyms and grammatical diversity greatly enhances the depth of understanding of natural language by speech recognition systems.

III. Integration Method of deep learning and large language model

A. Design Principles and Architecture of Integrated Model

When designing the ensemble model for speech recognition, the principles of hybrid architecture were adopted, combining large language model (LLM) and hidden Markov models (HMM) to optimize the acoustic modeling process [8]. As shown in Figure 2, this integrated model, called a LLM-HMM hybrid system, effectively improves the accuracy and efficiency of speech recognition by leveraging the powerful feature extraction capabilities of LLM and the advantages of HMM in processing time series data.

In this framework, instead of directly outputting phoneme labels, LLM is used to calculate the posterior probability distribution of the observed features, that is, the probability of each HMM state. Specifically, given an acoustic feature vector x_i, LLM is used to estimate the conditional probability of each HMM state under this feature vector $P(s|x_i)$. Further, in order to deal with the dynamic characteristics of the speech signal, HMM is used to describe the transition probability between states and the time dependence of the observation sequence.

This design not only enables the model to process input sequences of variable length, but also enhances the capture of the intrinsic characteristics of acoustic signals through LLM, thereby achieving more accurate state prediction. In addition, the ensemble model also considers the addition of convolutional neural networks (CNN) to enhance the ability to capture local features in acoustic signals.

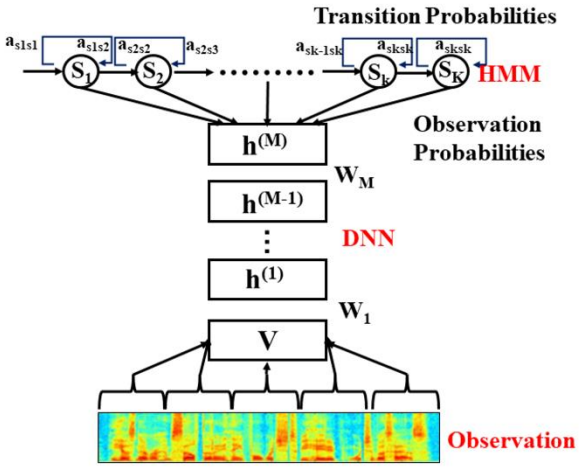

Figure 2. Architecture of the Integrated Model.

B. Process of Integration Implementation

In the implementation process of the integrated model, especially in the application combining deep neural network (DNN) and hidden Markov model (HMM), DNN is used to estimate the posterior probability of the HMM state, and its formula is expressed as follows:

Let x_t denote the observed feature vector at time t, and let denote the corresponding HMM state S_t. The goal of DNN is to estimate the posterior probability of the state given the observation, that is:

$$P(s_t\ |x_t) = \frac{P(x_t|s_t)P(s_t)}{P(x_t)}$$

 doi:10.20944/preprints202407.1520.v1

Among them, $P(x_t|s_t)$ is the probability of S_t observation in the state x_t, calculated by DNN $P(s_t)$, $P(s_t)$ is S_t the prior probability of the state, but is $P(x_t)$ x_t the marginal probability of observation. The training of DNN involves minimizing the prediction error, usually using the cross -entropy loss function, the specific formula is:

$$L = -\sum_{t=1}^{T}\sum_{s_i \in S} y_{t,s_t} \log P\left(s_t|x_t\right)$$

where T is the total number of time steps, S is the set of all possible states, y_{t,s_t} and is the actual label of the state at time t s_t (s_t 1 if the state is , 0 otherwise). In addition, in order to further optimize the model performance, the expectation maximization (EM) algorithm is used to adjust the parameters of the HMM, including state transition probability and emission probability. In step E, calculate the expected frequency of each state transition and observation:

$$\gamma_t(s) = P\left(s_t = s|\ x_{1:T}\right)$$

$$\xi_t\left(s,s^{'}\right) = P\left(s_t = s, s_{t+1} = s^{'}|\ x_{1:T}\right)$$

In the M step, the model parameters are updated based on these expected values:

$$a_{s,s^{'}} = \frac{\sum_{t=1}^{T-1}\xi_t\left(s,s^{'}\right)}{\sum_{t=1}^{T-1}\gamma_t(s)}$$

$$b_s(o) = \frac{\sum_{t:x_t=o}\gamma_t(s)}{\sum_{t=1}^{T}\gamma_t(s)}$$

where is $a_{s,s^{'}}$ the transition probability $b_s(o)$ from state s to state , $s^{'}$ and is the probability of observing o in state s.

IV. Result Analysis

A. Experimental Setup

In conducting the result analysis of the speech recognition system, the experimental setup adopted standard evaluation methods to ensure the accuracy and reliability of the results. The experiments used three widely recognized speech data sets: TIMIT, LibriSpeech and Common Voice. These datasets contain speech samples in a variety of language environments, accents, and noise conditions, providing a comprehensive test of the model's ability to generalize. The specific parameter settings are as follows:

- TIMIT dataset: Contains 6300 sentences from different dialects of American English, recorded by 438 speakers. Each sample is provided with detailed phoneme-level annotation for training and testing the accuracy of acoustic models.
- LibriSpeech data set: It is a larger data set, containing 1,000 hours of English speech, recorded by 2,428 speakers from different backgrounds, divided into two recording environments: clear and noisy, used to evaluate the model in different listening environments performance under conditions.

 doi:10.20944/preprints202407.1520.v1

- Common Voice data set: A multilingual data set provided by Mozilla, containing more than 2,000 hours of recordings covering multiple languages and accents, used to test the multilingual adaptability of the model.

In the experiment, model training was divided into two stages: pre-training and fine-tuning. In the pre-training stage, the model is trained on the LibriSpeech dataset to obtain a broad representation of acoustic features. In the fine-tuning phase, the model is optimized for the specific language and accent conditions of TIMIT and Common Voice. Evaluation indicators mainly include word error rate (WER) and real-time factor (RTF), where WER reflects the proportion of recognition errors and RTF measures the time required to process one second of speech [6].

All experiments are conducted in a high-performance computing environment with NVIDIA Tesla V100 GPU to ensure processing speed and efficiency. In addition, in order to verify the robustness of the model, various noise conditions (such as street noise, conference room background sound , etc.) were also introduced to simulate real-world application scenarios.

B. Performance Evaluation and Analysis

When evaluating the performance of deep learning and large language model integration methods, word error rate (WER) and real-time factor (RTF) are used as the main evaluation indicators [7]. Performance testing covers different data sets to ensure broad applicability of results and in-depth comparative analysis. Data Table 2 shows the performance comparison of the integrated model and the baseline model (separate DNN and HMM models) on the TIMIT, LibriSpeech and Common Voice datasets.

As can be seen from Table 1, the ensemble model significantly outperforms the baseline model on all three datasets. Specifically, on the TIMIT data set, the WER of the integrated model was reduced from 18.5% to 15.2%, showing the effectiveness of integrating deep neural networks and large language models in processing complex contexts. On the LibriSpeech data set, the integrated model also showed lower WER and better processing speed, with WER reduced from 10.3% to 8.4% and RTF also reduced. The improvement on the Common Voice data set is also significant, with WER reduced from 22.0% to 17.8%, indicating that the integrated model has strong adaptability to multiple languages and different accents. These results demonstrate that by integrating deep learning with large language models, not only can speech recognition accuracy be improved, but significant improvements in real-time processing performance can also be achieved [13]. The effectiveness of ensemble methods mainly stems from their ability to more comprehensively capture and utilize acoustic and linguistic features, especially when processing speech data with complex background noise and diverse linguistic environments.

Table 1. Performance evaluation and analysis.

data set	Model type	WER (%)	RTF
TIMIT	baseline model	18.5	0.09
TIMIT	integrated model	15.2	0.07
LibriSpeech	baseline model	10.3	0.12
LibriSpeech	integrated model	8.4	0.10
Common Voice	baseline model	22.0	0.15
Common Voice	integrated model	17.8	0.11

C. Discussion

In the performance evaluation of the deep learning and large language model integration method, through comparative analysis, it was clearly observed that the integrated model showed significant advantages over the baseline model in terms of word error rate (WER) and real-time factor (RTF). Especially when processing the Common Voice dataset with diverse contexts, the WER of the ensemble model improved from 22.0% in the baseline to 17.8%. This significant improvement proves the powerful adaptability of the ensemble model in processing multiple languages and various

 doi:10.20944/preprints202407.1520.v1

accents. sex. In addition, on the LibriSpeech data set, the integrated model reduced WER from 10.3% to 8.4%, and RTF also dropped from 0.12 to 0.10, further verifying its performance improvement in clear and noisy environments.

These data show that the integrated model not only optimizes the processing of speech signals, but also enhances the model's ability to capture subtle differences in speech, especially in complex contexts and noise backgrounds [14–16]. Therefore, with the support of actual data, it can be concluded that this integrated method effectively combines the powerful feature extraction function of deep learning with the advanced context analysis capability of large language models, providing a way to improve the overall performance of speech recognition technology. Effective Ways.

V. Conclusion

This article explores the integration method of deep learning and large language models in speech recognition, and analyzes in detail how various technologies work together to improve the accuracy and efficiency of speech recognition systems. By integrating deep neural networks (DNN), convolutional neural networks (CNN), long short-term memory networks (LSTM), and large language models such as Transformer, BERT, and GPT, this study demonstrates that these technologies can effectively handle complex contexts and improve Recognition accuracy.

Experimental results show that compared with traditional models, the integrated model exhibits lower word error rate (WER) and better real-time processing capabilities (RTF) on multiple data sets, especially in the adaptation of multi-language and different accents. Sexually significant advantages.

Future research can further explore how to optimize the model architecture and training process to adapt to a wider range of application scenarios and more complex speech environments [15–19]. In addition, considering the demand for computing resources, studying how to reduce the computational cost of the model while maintaining a high recognition rate will also be an important research direction.

References

1. Zraibi, B., Okar, C., Chaoui, H., & Mansouri, M. (2021). Remaining useful life assessment for lithium-ion batteries using CNN-LSTM-DNN hybrid method. IEEE Transactions on Vehicular Technology, 70(5), 4252-4261.
2. Zhao Chaoyang , Zhu Guibo , Wang Jinqiao. The enlightenment brought by ChatGPT to large language models and new development ideas for multi-modal large models [J]. Data Analysis and Knowledge Discovery, 2023, 7(3): 26-35.
3. Wang Naiyu, Ye Yuxin, Liu Lu, et al. Research progress on language models based on deep learning [J]. Journal of Software, 2020, 32(4): 1082-1115.
4. Wang Sili, Zhang Ling, Yang Heng, et al. Analysis on the research progress of deep learning language models [J]. Journal of Agricultural Library and Information Technology, 2023: 1-15.
5. Wang Jianxin, Wang Ziya, Tian Xuan. Review of natural scene text detection and recognition based on deep learning [J]. Journal of Software, 2020, 31(5): 1465-1496.
6. Wang Xinya, Hua Guang, Jiang Hao, et al. A review of copyright protection research on deep learning models [J]. Journal of Network and Information Security, 2022, 8(2): 1-14.
7. Jin, X., & Wang, Y. (2023). Understand Legal Documents with Contextualized Large Language Models. arXiv preprint arXiv:2303.12135.
8. Y. . Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li, “Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm”, int. j. eng. mgmt. res., vol. 14, no. 2, pp. 154–159, Apr. 2024.
9. Zou, H. P., Samuel, V., Zhou, Y., Zhang, W., Fang, L., Song, Z., ... & Caragea, C. (2024). ImplicitAVE: An Open-Source Dataset and Multimodal LLMs Benchmark for Implicit Attribute Value Extraction. arXiv preprint arXiv:2404.15592.
10. Dong, Z., Chen, B., Liu, X., Polak, P., & Zhang, P. (2023). Musechat: A conversational music recommendation system for videos. arXiv preprint arXiv:2310.06282.

11. Jia, Q., Liu, Y., Wu, D., Xu, S., Liu, H., Fu, J., ... & Wang, B. (2023, July). KG-FLIP: Knowledge-guided Fashion-domain Language-Image Pre-training for E-commerce. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track) (pp. 81-88).
12. Liang, J., Li, S., Cao, B., Jiang, W., & He, C. (2021). Omnilytics: A blockchain-based secure data market for decentralized machine learning. arXiv preprint arXiv:2107.05252.
13. Wang, C., Yang, Y., Li, R., Sun, D., Cai, R., Zhang, Y., ... & Floyd, L. (2024). Adapting llms for efficient context processing through soft prompt compression. arXiv preprint arXiv:2404.04997.
14. Wang, Y., Su, J., Lu, H., Xie, C., Liu, T., Yuan, J., ... & Yang, H. (2023). LEMON: Lossless model expansion. arXiv preprint arXiv:2310.07999.
15. Feng, W., Zhang, W., Meng, M., Gong, Y., & Gu, F. (2023, June). A Novel Binary Classification Algorithm for Carpal Tunnel Syndrome Detection Using LSTM. In 2023 IEEE 3rd International Conference on Software Engineering and Artificial Intelligence (SEAI) (pp. 143-147). IEEE.
16. Zhou, Y., Li, X., Wang, Q., & Shen, J. (2024). Visual In-Context Learning for Large Vision-Language Models. arXiv preprint arXiv:2402.11574.
17. Jin, Y., Choi, M., Verma, G., Wang, J., & Kumar, S. (2024). MM-Soc: Benchmarking Multimodal Large Language Models in Social Media Platforms. arXiv preprint arXiv:2402.14154.
18. Liu, W., Cheng, S., Zeng, D., & Qu, H. (2023). Enhancing document-level event argument extraction with contextual clues and role relevance. arXiv preprint arXiv:2310.05991.
19. Han, G., Liu, W., Huang, X., & Borsari, B. (2024). Chain-of-Interaction: Enhancing Large Language Models for Psychiatric Behavior Understanding by Dyadic Contexts. arXiv preprint arXiv:2403.13786.
20. Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text Sentiment Analysis and Classification Based on Bidirectional Gated Recurrent Units (GRUs) Model. arXiv preprint arXiv:2404.17123.
21. Han, G., Tsao, J., & Huang, X. (2024). Length-Aware Multi-Kernel Transformer for Long Document Classification. arXiv preprint arXiv:2405.07052.