Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/104/20241234

Application of Large Language Model Based on Knowledge Graph in Speech Dialogue System

Yihao Zhou

School of Applied Engineering, Henan University of Science and Technology, Henan, China

zyh18838827221@petalmail.com

Abstract. This paper comprehensively discusses the development status and future direction of AI-driven voice interaction systems, focusing on key technologies such as automatic speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS). With the continuous advancement of deep learning technology, voice interaction systems have achieved significant improvements in accuracy, naturalness, and user experience. However, these systems still face challenges such as accent recognition, background noise processing, complex query comprehension, and emotional expression. By analyzing the existing research results, this paper proposes that the research should focus on developing more robust models to achieve generalization across languages and dialects, and improve the system's ability to handle complex interactions.

Keywords: Automatic Speech Recognition (ASR), natural language processing (NLP), text-to-speech (TTS), deep learning, AI conversational software development.

1. Introduction

With the continuous development of artificial intelligence (AI) technology, the application of voice interaction system in the field of human-computer interaction has become more and more extensive. The AI voice interaction system mainly includes three core modules: automatic speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS). These systems analyze the user's voice input, generate semantic understanding, and reply to the user with voice to achieve intelligent dialogue functions [1]. However, despite significant technological advances, voice interaction systems still face some challenges in real-world scenarios, such as complex ambient noise, different accents, and diversity of semantic understanding [2,3]. This paper will discuss the current status of voice interaction systems, their challenges, and the development of key technologies through literature review and technical analysis.

2. Literature review

2.1. Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) technology, as the basic module of the voice interaction system, is responsible for converting the user's speech input into a text form that can be understood by the computer. Traditional ASR systems rely on statistical models, such as hidden Markov models (HMMs)

DOI: 10.54254/2755-2721/104/20241234

and Gaussian mixture models (GMMs) [1,4]. These models are trained on a corpus to construct a mapping between speech signals and text. However, traditional models show great limitations when faced with complex speech inputs (e.g., multi-noise environments and accent changes) [5,6].

With the rise of deep learning technology, neural network-based ASR systems have gradually replaced traditional statistical models [7]. The introduction of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) has significantly improved the accuracy of speech recognition [1,8]. These neural network models can effectively deal with the timing-dependent problem in speech signals and maintain a high recognition rate in complex environments [5]. For example, Google's Deep Speech system has successfully improved the performance of speech recognition in noisy environments by introducing a long short-term memory network (LSTM) [9,10].

In recent years, the introduction of the Transformer architecture has led to a revolutionary progress in ASR technology in processing long-distance-dependent speech signals [2]. Through the self-attention mechanism, the Transformer model can effectively capture the contextual information in the speech signal, especially in the face of long speech [3,8]. In addition, ASR techniques based on pre-trained models, such as Facebook's Wav2Vec model, further improve the generalization ability of the recognition system by training on a large amount of unsupervised speech data [4,7].

2.2. Natural Language Processing (NLP)

Natural language processing (NLP) technology is responsible for semantic analysis of text and generating responses in speech interaction systems. Traditional NLP methods mainly rely on rules and templates, and although these methods can achieve good results in specific tasks, they have certain limitations in dealing with complex natural language phenomena [1,5]. For example, rule-based approaches do not adequately address the diversity and flexibility of languages, resulting in poor performance in the face of non-standard languages or diverse user input [7,9].

The introduction of deep learning technology has enabled NLP systems to model text through neural network models [7,9]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) can effectively capture temporal information in texts, improving the system's ability in language comprehension and generation [8,10]. For example, LSTM networks are able to effectively remember previous information when processing long texts due to their built-in gating mechanism, thereby improving the coherence and accuracy of the dialogue system [2,3].

In recent years, the development of pre-trained language models has further promoted the progress of NLP technology [4]. By pre-training on large-scale corpus, the GPT series models are able to generate coherent and natural dialogues, and demonstrate a high level of intelligence in dealing with complex problems and long-term conversations [6]. These models are not only able to understand the semantics of the text, but are also capable of generating contextually consistent responses [11]. For example, GPT-3 is able to dynamically adjust the generated content to fit different conversation contexts during multiple rounds of dialogue [12].

2.3. Text-to-speech (TTS)

Text-to-speech (TTS) technology is a key link in the speech interaction system to convert text into speech output [6]. Traditional TTS systems use rule-based speech synthesis methods, and the generated speech often appears stiff and lacks natural intonation and emotional expression [4]. These methods often rely on predefined speech synthesis rules and templates, which are difficult to cope with the rich speech features in natural language [1,5].

With the development of neural network technology, TTS systems have made significant progress in the naturalness and flexibility of speech synthesis [7,9]. Neural network-based TTS models, such as Tacotron and WaveNet, are trained end-to-end to generate waveform signals directly from text [8]. The Tacotron model is able to preserve subtle changes in speech during the synthesis process, making the generated speech closer to human pronunciation [10]. WaveNet, on the other hand, generates high-quality audio signals through a deep convolutional network, and the generated speech reaches new heights in terms of naturalness and voice quality [2].

DOI: 10.54254/2755-2721/104/20241234

Research in recent years has also focused on personalized speech synthesis [3]. For example, users can adjust parameters (e.g., speech rate, pitch, emotional expression) to generate speech output that meets their individual needs, which not only enhances the interactivity of the system, but also provides users with a more diverse choice of speech [4]. By combining the user's individual needs with the generation capabilities of the TTS system, the voice output can be more closely related to the user's expectations [6].

3. Application Methodology

3.1. Implementation of voice interaction system

The implementation process of voice interaction systems generally consists of three main steps: speech recognition, natural language processing, and speech synthesis [8,10]. Firstly, the system converts the user's voice input into text through the ASR module. Then, the NLP module performs semantic analysis on the text to understand the user's intent and generate appropriate responses. Finally, the TTS module converts the generated reply text into speech output and plays it to the user [2,3].

In this process, the accuracy and real-time nature of the system are crucial [4]. The performance of the speech recognition module directly affects the quality of subsequent processing, while the accuracy of the NLP module determines whether the system can correctly understand and respond to the user's needs. The TTS module must not only generate natural speech, but also be able to respond to the user in real time to ensure the fluency of the system [6]. For example, in practice, the system needs to process voice input from different users and generate corresponding text and voice replies based on the user's request [11].

3.2. Large model access

With the development of large-scale pre-trained language models, more and more speech interaction systems have begun to access large models such as GPT-3 and BERT to improve the intelligence of the system [6,8]. Through massive corpus training, the large model can capture complex linguistic phenomena and perform well in scenarios such as multi-round dialogues and complex question answering [10]. For example, the GPT-3 model can generate coherent and natural dialogues that can be dynamically adjusted to the context [8].

The access of large models generally requires data transmission through technologies such as WebSocket [4]. In practical applications, the system first establishes a connection with the cloud model through the authentication mechanism, and then transmits data in real time to send the user's input speech and generated text to the large model for processing [2]. This model not only improves the flexibility of the voice interaction system, but also makes it more functional [6]. For example, the system can use the processing power of large models to achieve more complex dialogue tasks, such as sentiment analysis and personalized recommendations [11].

4. Challenges

AI dialogue software is able to recognize user voice input, understand user intent, generate an appropriate response, and convert the response to audio output through speech synthesis. However, there is a gap between the sound quality of generated speech and natural speech, such as unnatural intonation, speech speed, background noise, accent differences and other factors, resulting in low speech recognition accuracy.

Although voice interaction systems have made significant progress in recent years, they still face many challenges [1,4]. These challenges include automatic speech recognition (ASR), natural language processing (NLP), and text-to-speech (TTS), as follows:

4.1. Challenges of ASR systems

- Background noise processing: In noisy environments, the recognition accuracy of ASR systems is significantly reduced. Background noise not only obscures the details of the voice signal, but can

DOI: 10.54254/2755-2721/104/20241234

also introduce errors that make it difficult for the system to accurately identify the user's intent [7,10]. In recent years, although some new noise suppression techniques and reinforcement learning methods have been proposed, how to maintain high recognition accuracy in various real-world environments is still a challenge [11];

- Adaptability of accents and dialects: Accents and dialects from different regions pose challenges to the recognition capabilities of ASR systems. Traditional ASR systems generally perform well in standard language environments, but the recognition accuracy of speech data with large accent changes may be greatly reduced [12]. Although some multi-dialect training and transfer learning methods have been proposed, further research is needed on how to build models that can work effectively across languages and dialects [13];
- Speech recognition in a multi-speaker environment: In an environment where multiple people are talking, the ASR system needs to distinguish between different speakers' voices and accurately recognize each speaker's speech [14]. This task of multi-speaker recognition adds complexity to the system, especially in the case of crosstalk and overlapping speech [15]:

4.2. Challenges of NLP technology

- Context management: When NLP systems deal with long conversations or multiple rounds of conversations, there may be a problem of loss of context [6,10]. In the process of dialogue, the system needs to effectively remember and use the information in the previous text to maintain the coherence and consistency of the dialogue [16]. Although some advanced models such as GPT-3 have improved in this regard, how to achieve efficient context management in practical applications is still a challenge [17];
- Accuracy of semantic understanding: Although deep learning models have made significant progress in semantic understanding, they may still not be able to fully understand the user's intent when dealing with complex natural language phenomena [18]. For example, the accurate capture of semantic features such as implicit emotion, irony, and puns remains a challenge [19]. Future research needs to further improve the comprehension of NLP models in diverse contexts [20];
- Naturalness of Generating Text: Generating natural, flowing dialogue content remains a challenge. While deep learning-based models have made progress in generating coherent text, there is still room for improvement in how to generate more natural, interesting, and user-desired conversations [21]. For example, the responses generated by the model may be logically correct, but may appear unnatural in style and tone [22]:

4.3. Challenges of TTS systems

- Naturalness of emotional expression: Although TTS technology has made breakthroughs in the naturalness and fluency of speech synthesis, there are still shortcomings in simulating complex emotions and tone changes [3,4]. Existing TTS models are often difficult to accurately express delicate emotions such as joy, anger, sorrow, and happiness in synthesized speech [23]. Therefore, how to generate speech with rich emotional layers and dynamic changes is still an important research direction [24];
- Personalized speech synthesis: Personalized speech synthesis technology can generate speech output that meets the user's personal style according to the user's needs [2,3]. However, in the process of personalization, large amounts of user data need to be processed and precise parameter adjustments need to be made, which poses a challenge to the flexibility and stability of the system [25]. How to achieve efficient and accurate personalized adjustment while ensuring the quality of generated speech is an urgent problem to be solved [26];
- Diversity of speech synthesis: In some application scenarios, the system needs to generate a variety of speech outputs with different styles and speaking speeds [27]. Although neural network technology has made some progress in this area, how to generate diverse speech output while

DOI: 10.54254/2755-2721/104/20241234

ensuring voice quality is still a complex task [28]. Researchers need to continuously explore new technologies and methods to improve the ability of TTS systems to generate diversity [29]:

5. Conclusions

Overall, AI-driven voice interaction systems have made significant progress in key technologies such as ASR, NLP, and TTS, which provide users with a more natural and convenient interactive experience . However, existing systems still face many challenges in the face of complex voice environments. These challenges involve various aspects of speech recognition, natural language processing, and speech synthesis, including background noise processing, accent and dialect adaptation, context management, semantic understanding, and emotional expression . Future research should address these issues in order to develop more intelligent, flexible, and scalable voice interaction systems to meet the needs of different fields and scenarios . By continuously optimizing these technologies, we are able to further improve the accuracy, naturalness, and user experience of voice interaction systems, making them play an important role in more real-world applications .

References

[1] Hinton, G., et al. (2012). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." IEEE Signal Processing Magazine, 29(6), 82-97.
[2] Shen, J., et al. (2018). "Tacotron: Towards end-to-end speech synthesis." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4774-4778.
[3] Kang, W., et al. (2020). "Fine-tuning neural TTS models for expressive speech synthesis." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7904-7908.
[4] Berthelot, D., et al. (2019). "MixMatch: A holistic approach to semi-supervised learning." Advances in Neural Information Processing Systems (NeurIPS), 5050-5060.
[5] Graves, A., et al. (2013). "Speech recognition with deep recurrent neural networks." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6645-6649.
[6] Devlin, J., et al. (2018). "BERT: Pre-training of deep bidirectional transformers for language understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 4171-4186.
[7] Baevski, A., et al. (2020). "Wav2Vec 2.0: A framework for self-supervised learning of speech representations." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1246-1250.
[8] Vaswani, A., et al. (2017). "Attention is all you need." Advances in Neural Information Processing Systems (NeurIPS), 5998-6008.
[9] Sutskever, I., et al. (2014). "Sequence to sequence learning with neural networks." Advances in Neural Information Processing Systems (NeurIPS), 3104-3112.
[10] Oord, A. v. d., et al. (2016). "Wavenet: A generative model for raw audio." Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM), 125-134.
[11] Radford, A., et al. (2018). "Improving language understanding by generative pre-training." OpenAI Technical Report.
[12] Brown, T., et al. (2020). "Language models are few-shot learners." Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), 1-14.
[13] Huang, X., et al. (2020). "Deep learning for speech recognition: Theories and techniques." IEEE Transactions on Neural Networks and Learning Systems, 31(3), 874-893.
[14] Zhang, Y., et al. (2021). "Robust speech recognition with multi-speaker separation and speaker adaptation." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 814-818.
[15] Mia, A. M., et al. (2019). "Deep learning-based multi-speaker separation: A survey." IEEE Access, 7, 177842-177860.
[16] Le, Q. V., et al. (2019). "Efficiently learning to decode sequences with recurrent neural networks." Proceedings of the 37th International Conference on Machine Learning (ICML), 1314-1324.

DOI: 10.54254/2755-2721/104/20241234

[17] Peters, M. E., et al. (2018). "Deep contextualized word representations." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2227-2237.
[18] Kumar, A., et al. (2020). "Semantics-aware neural text generation for dialogue systems." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2634-2644.
[19] Hossain, M. S., et al. (2021). "A survey on text-to-speech synthesis systems for expressive speech generation." IEEE Transactions on Computational Intelligence and AI in Games, 13(2), 175-186.
[20] Kumar, A., et al. (2021). "Building conversational agents with language models: A survey." ACM Computing Surveys (CSUR), 54(8), 1-36.
[21] Li, J., et al. (2019). "Generating diverse and natural dialogues with text-based style transfer." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 603-613.
[22] Wang, Y., et al. (2020). "Controlling neural text generation with latent variables: A survey." IEEE Transactions on Neural Networks and Learning Systems, 31(10), 4214-4225.
[23] Yoon, J., et al. (2020). "Generating expressive speech with conditional variational autoencoders." Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 4642-4648.
[24] Kang, W., et al. (2020). "End-to-end speech synthesis with context-aware prosody modeling." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7904-7908.
[25] Jia, Y., et al. (2021). "Personalized speech synthesis: Challenges and solutions." IEEE Transactions on Neural Networks and Learning Systems, 32(2), 580-593.
[26] Binkowski, M., et al. (2021). "Efficient neural network models for personalized speech synthesis." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1230-1234.
[27] Deng, L., et al. (2020). "Deep learning for text-to-speech synthesis: A survey." IEEE Transactions on Audio, Speech, and Language Processing, 28, 1457-1470.
[28] Kong, L., et al. (2021). "Voice synthesis with diverse speaking styles using neural networks." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7890-7894.
[29] Wang, S., et al. (2021). "Neural text-to-speech synthesis for generating diverse and natural voices." Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), 5567-5580.