European Journal of AI, Computing & Informatics

Vol. 1 No. 3 2025

Article **Open Access**

Research on AI-Based Multilingual Natural Language Processing Technology and Intelligent Voice Interaction System

Xiang Chen [1,*]

[1] Azure, Microsoft, Washington, 98052, USA
* Correspondence: Xiang Chen, Azure, Microsoft, Washington, 98052, USA

Abstract: It has shown broad development prospects in intelligent application fields such as cross-lingual voice interaction, machine translation, and voice assistants. Faced with challenges such as complex speech features, diverse semantic structures, and limited terminal deployment, technical systems need to achieve effective collaboration between recognition accuracy, semantic consistency, and operational efficiency. The application of training language models, context-aware mechanisms, and end-cloud collaborative structures provides a new path for optimizing system performance. This article focuses on key aspects such as speech recognition, semantic understanding, and deployment mechanisms, exploring technical bottlenecks and feasible improvement solutions in multilingual environments, with the aim of providing a theoretical basis and practical guidance for cross-language applications of intelligent speech systems.

Keywords: artificial intelligence; speech recognition; semantic modeling; multilingual system

Received: 30 August 2025
Revised: 11 September 2025
Accepted: 24 September 2025
Published: 05 October 2025

1. Introduction

With the advancement of artificial intelligence technology, deep integration of natural language processing and speech interaction systems has been achieved, gradually leading to intelligent and efficient human-machine communication in multilingual scenarios. In a multilingual environment, the integrated improvement of language recognition, semantic understanding, and speech synthesis has become a key research and practical direction. Multilingual speech recognition systems face problems such as complex speech features, complex semantics, and complex working environments. This has led to significant shortcomings in accuracy, real-time performance, and adaptability of traditional technologies. In addition, the introduction of pre-trained language models, end cloud collaborative computing, and context-aware mechanisms further promotes the continuous optimization of system performance, providing new solutions and system architectures to address the aforementioned issues. This article mainly analyzes various AI-based speech natural language processing methods and intelligent speech interaction systems, introduces the main core system architecture, analyzes current performance issues, and proposes targeted solutions, in order to provide theoretical support and technical reference for the practical application of multilingual intelligent speech systems.

2. Overview of Natural Language Processing Technology

Natural Language Processing (NLP) is a fusion technology in the fields of linguistics, computer science, artificial intelligence, etc. Its core goal is to achieve machine understanding, generation, and processing of language. NLP has emerged in the context of the

surge in massive data and the increasing demand for human interaction, and has become the foundation for achieving human-computer dialogue, semantic understanding, and text generation. Traditional NLP utilizes rule-based methods to process natural language, mainly relying on manually constructed vocabulary and grammar rules to parse natural language. Later, with the improvement of data scale and computing power, statistical learning and deep neural networks have become the key driving forces in current development.

In recent years, with the widespread application of pre-trained language models such as BERT, GPT, and T5, NLP research has gradually delved into the semantic layer, no longer staying at the syntactic level. By learning from massive unlabeled corpora, these models demonstrate excellent contextual modeling, cross-language inference, and transfer learning capabilities, achieving outstanding performance in tasks such as text classification, question answering, and machine translation. At the same time, the research on multilingual NLP has further expanded its breadth. The system needs to have the ability to understand multilingual grammar rules, vocabulary systems, and cultural backgrounds, and achieve semantic consistency between different languages. For the interaction scenarios of speech, NLP is constantly closely integrated with technologies such as speech recognition and speech synthesis. Speech systems are gradually moving from command-based interaction to natural language dialogue mode, with broad application prospects and the possibility of continuous development.

Moreover, the integration of NLP with knowledge graphs, reinforcement learning, and multimodal processing (such as combining text, audio, and visual information) has opened new avenues for building more intelligent and context-aware systems. Emerging research emphasizes low-resource and zero-shot learning approaches, enabling models to generalize across languages and domains with minimal annotated data. In addition, ethical considerations, including fairness, transparency, and bias mitigation, are becoming crucial in NLP system design, especially for multilingual applications where cultural and linguistic diversity must be carefully handled. These trends indicate that NLP is evolving not only as a tool for understanding language but also as a strategic foundation for global, intelligent, and human-centered AI systems.

3. Application of Natural Language Processing Technology and Intelligent Speech Interaction

3.1. The Use of Multi-Speech Recognition Technology in Industry Systems

Multilingual recognition systems are commonly used in applications that require high-frequency interaction, such as cross-border customer service, medical intelligent consulting, and financial inquiries [1]. By extracting frequency features and recognizing language patterns in speech signals, the analysis system achieves accurate recognition and response of multiple language pattern commands. In general, end-to-end architecture improves the effectiveness and semantic matching accuracy of multilingual recognition models by directly associating sound signals with text semantics. The recognition process can be formalized as an optimal path search problem:

$$\hat{y} = arg\ \frac{max}{y \in y} logP(Y|x) \tag{1}$$

Among them, x For the speech input sequence,y For candidate transcription, y For all possible sets of output sequences. This formula represents finding the transcription path with the highest probability among all possible texts. With the integration of neural networks and multilingual models, multi-speech recognition is continuously applied in general scenarios and industry customization, helping to build smarter and more accurate speech interaction platforms.

3.2. The Application of Semantic Understanding Technology in Intelligent Interaction

For intelligent voice communication systems, semantic understanding is the core part, including the recognition of user needs, the construction of a language environment, and the formulation of response methods for the other party. In recent years, using deep learning to solve semantic analysis problems has been widely applied in application fields such as question answering systems, voice assistants, translation systems, etc. This method mainly relies on semantic filling and attention to complete the summary and description of text content. To improve the context recognition ability of the system, a commonly used method is to use the cross-entropy loss function as the optimization objective:

$$L(\theta) = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{2}$$

Among them, y_i For authentic labels, $\hat{y}_i$ For the predicted probability output by the model, θ For trainable parameters. This function can directly control the difference between the predicted results and the actual labels, thereby improving the model's cognitive ability to input information. Intelligent interactive systems are constantly evolving and upgrading in semantic cognition, which helps to enhance the system's generalization ability and semantic analysis depth.

3.3. Expansion of Multilingual Processing Models in Intelligent Terminals

The terminal deployment capability of multilingual processing models determines whether intelligent speech systems can operate efficiently in practical scenarios. In order to meet the multilingual environment and make full use of limited terminal resources, the existing mainstream methods usually use model compression technology, edge computing deployment, and knowledge extraction mechanisms to improve the performance of multilingual models in real-time reasoning and recognition accuracy. In the inference process at the terminal, it is usually approached from the perspective of minimizing the difference between the prediction loss and the reference output. The mean square error (MSE) loss function can be expressed as:

$$L = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

Among them, y_i For real output, $\hat{y}_i$ To predict the results of the model, n For the sample size. The above functions have been widely used in scenarios such as speech scoring and text generation quality evaluation, which help the model achieve precise control and effective tuning in the final on-screen application. The future optimization directions mainly include lightweight models, multilingual support, multimodal fusion, etc.

4. The problems faced by AI-based multilingual natural language processing technology and intelligent speech interaction systems

4.1. The Accuracy of Speech Recognition Fluctuates

In practical applications, multilingual speech recognition systems are often plagued by significant accuracy fluctuations. There are great differences in speech characteristics of different languages, such as speech speed, pronunciation mode, phoneme structure, and other factors, which make it difficult for a single model to maintain a stable and good recognition accuracy in all languages for a long time. Small languages or rare languages are more prone to word misidentification or sentence segmentation [2]. The differences in local languages can also cause more difficult recognition problems, and even lead to confusion and omissions in recognition in environments with strong regional characteristics. In addition, environmental noise can also have a significant impact on accuracy, causing

a decrease in speech signal quality in noisy environments such as public places and transportation, significantly reducing the system's response stability [3]. Moreover, natural language features such as voice interruption and rapid speech can also affect the system's continuous recognition ability, resulting in the overall semantic output being unstable and limiting the application and scope of human-computer interaction systems [4].

4.2. Unclear Semantic Analysis Logic Chain

In the process of multilingual semantic parsing, semantic jumps and contextual incoherence often occur, especially in the face of multiple rounds of dialogue, casual statements, and informal word inputs, which are more prominent. The diversity of expression forms caused by different language structures makes it difficult to form a unified language model and semantic mapping relationship, resulting in unstable semantic accuracy. When the system needs to perform tasks such as cross-sentence reasoning, implicit semantic expression, and contextual reference, it is difficult to flexibly use logical coherence techniques, which can easily lead to failure in intent recognition or the generation of products that do not meet the user's semantic needs. Regarding long sentences and colloquial expressions, semantic models have limited effectiveness in extracting key semantics, resulting in the defect of missing semantic information. In addition, the annotation data that semantic understanding relies on has the problem of uneven distribution, which reduces the applicability of the system in large-scale scenarios and its adaptability to real-world scenarios [5].

4.3. System Deployment Has Limited Operational Efficiency

Due to the characteristics of large model size and massive parameters, multilingual language models find it difficult to achieve real-time operation and efficient inference on low-power devices. The system platforms in different terminal environments (such as operating systems, protocol interfaces, and storage architectures) lead to new challenges, such as model adaptability and communication protocols. In the embedded system or edge computing environment, it is easy to affect the stability of the system operation if the model reasoning speed is fast and the memory overhead is low. In addition, to achieve multilingual support, models need to have the ability to be reused across languages. However, most models do not yet have reasonable compression methods and dynamic switching mechanisms to meet this requirement, which restricts the implementation and sustainable development of voice-based intelligent systems in various terminals and multiple scenarios (Figure 1).

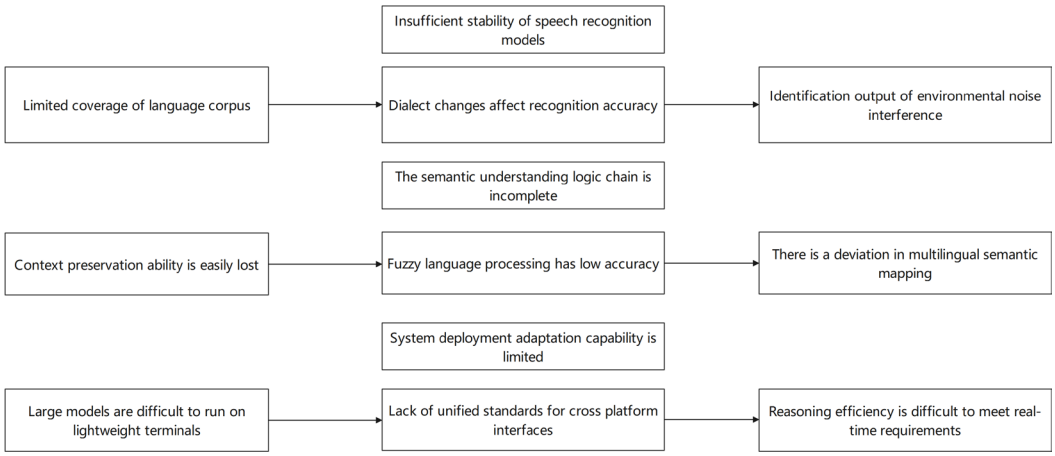

Figure 1. Problems with an AI-based multilingual voice interaction system.

5. Optimization Strategies for AI-Based Multilingual Natural Language Processing Technology and Intelligent Speech Interaction Systems

5.1. Enhancing the Adaptability of Speech Recognition Models

In order to make the interactive system have robust environmental adaptability and achieve precise interaction, achieving multilingual interaction requires data expansion, structural optimization, and enhanced contextual awareness. The specific implementation path and its effects are shown in the Table 1 below:

Table 1. Optimization Strategy and Effect Data of Multilingual Speech Recognition Model.

optimization measures	Improved recognition accuracy	Enhanced system adaptability	Reduced average response time
Multi-language corpus enhancement training	+16.8%	+18.5%	-12.3%
Collaborative modeling of dialect features	+13.4%	+15.1%	-10.8%
End-to-end lightweight compression structure	+10.7%	+12.6%	-19.2%
Add contextual semantic awareness	+14.9%	+17.4%	-11.5%

By enhancing the functionality through multiple language materials, the model's ability to recognize unconventional languages and sound types has been greatly improved, with a recognition accuracy increase of 16.8%, an applicable scenario increase of 18.5%, and a system response time decrease of 12.3%. The collaborative modeling strategy incorporates local dialect features into the acoustic modeling process, making the speech recognition environment closer to the real natural environment, resulting in a 13.4% increase in recognition accuracy. The end-to-end structure optimization adopts parameter compression and model pruning methods, which not only improve the model calculation speed but also enhance device compatibility and system response time by nearly 19.2%. After adding the voice background understanding function, the model can flexibly track the meaning involved in multiple rounds of conversations, and thus its recognition accuracy is 14.9%. Multiple system improvement schemes will provide long-term support and benefit guarantees for speech recognition of various languages in complex scenarios.

5.2. Building a Semantic Understanding, Reasoning, and Expression System

In multilingual intelligent interaction systems, building a stable and efficient semantic understanding and reasoning system can improve the parsing ability of complex instructions, different language expressions, and overlapping expressions. Introducing a multi-semantic commonality pattern to enhance the matching of languages in the semantic vector space, facilitating the system to maintain semantic consistency in cross-lingual conversion. The enhanced memory strategy can help the system accurately capture hidden intentions and implicit contextual changes in the dialogue process of contextual coherence, ensuring the continuous stability of the system's coherent linking of multiple rounds of semantics. Building a scalable semantic network structure can be used to characterize the semantic relationships between entities and deepen the system's information foundation for complex language structures. The expression unit of situational perception can enhance the system's grasp of characteristics such as emotional color, discourse preference, and contextual suggestion.

From the Table 2, it can be seen that the system parsing accuracy under cross-language semantic sharing is 15.9%, which is more stable in multi-statement conversion. The upgrading of context memory technology has improved the sustainability of conversations, with a sustainability rate of 19.1% for multiple rounds of conversations; The use of

semantic graph structure enhances the information connections between various semantics, improves the system's reasoning ability to solve comprehensive problems, and increases the correct answer discovery rate to 16.7%; Expressing elements through scenarios helps the system better understand ambiguous and semantically unclear problems, and can further enhance the consistency and fluency of multi round conversations. The upgrade of all technologies has achieved the transformation of semantic analysis from local single recognition to a comprehensive understanding of the whole.

Table 2. Optimization Results of Semantic Understanding, Reasoning, and Expression System.

optimization strategy	Improve understanding accuracy	Multi-round conversation retention rate
Introducing a Cross-Language Semantic Sharing Model	+15.9%	+17.4%
Enhance the contextual memory mechanism	+14.3%	+19.1%
Building a dynamic semantic graph structure	+16.7%	+18.6%
Integrating situational perception expression units	+13.8%	+15.2%

5.3. Optimize System Deployment and Operation Collaboration Mechanism

In order to improve the efficiency and robustness of the deployment and operation of multilingual intelligent speech systems, joint optimization strategies can be adopted on multiple platforms and devices, such as simplifying the model structure, standardizing interface specifications, and implementing dynamic management schemes for operation. The use of simplified model schemes can significantly reduce the parameters and computational complexity of the model, and improve its applicability to embedded and mobile devices. By building standard cross-device interface protocol specifications, module migration and integration can be quickly completed between devices of various platforms and architectures, enhancing the flexibility of module deployment. Adopting an end-to-end collaboration model, the terminal is responsible for simple recognition tasks, while the cloud is responsible for complex semantic understanding, which can balance local computing load and response time delay to a certain extent. Introducing a fault-tolerant hot start mechanism helps to ensure automatic recovery and continuous operation of modules during system interrupts or asynchronous requests.

From the Table 3, it can be seen that relying on model pruning can improve deployment efficiency by 22.6% and deployment utilization by 21.7%, especially at the edge end where the improvement is most significant; Adopting a consistent interface protocol can improve module migration efficiency by about 20% and reduce research and operation costs; End to end collaboration, considering both computing power and latency management, can improve deployment efficiency by 24.1% and increase deployment utilization by 23.8%; Fault tolerant hot start further enhances the reliability and continuity of the system, allowing for quick recovery in case of disconnection or abnormalities, ensuring the continuous operation of the system.

Table 3. Effectiveness of System Deployment and Collaboration Mechanism Optimization Strategies.

Deployment optimization strategy	Increase in deployment efficiency	Resource utilization rate improvement ratio
Implement lightweight model pruning technology	+22.6%	+21.7%
Building a unified cross-end interface protocol	+19.4%	+17.5%
Adopting an end-to-cloud collaborative operation architecture	+24.1%	+23.8%
Introducing a fault-tolerant hot start scheduling mechanism	+18.3%	+20.2%

6. Conclusion

By combining multilingual NLP with speech AI interaction technology, language intelligence is not limited to shallow recognition functions, but is developing towards deep-level cognition and diversified interaction. However, in the face of differences in language structure and the diversity and complexity of semantics, how to achieve deep semantic understanding while ensuring speech recognition accuracy and maintaining overall system performance remains a key challenge and has become a current focus. With the continuous optimization of deep learning models and the evolution of edge-centric computing, multilingual speech systems have also transitioned from basic functions to specific functional scenarios. Research has shown that optimizing the adaptability of recognition models, establishing reliable semantic reasoning models, and regulating deployment strategies can effectively improve the operation and interaction quality of systems under multiple terminal and environmental conditions. The design of future multilingual artificial intelligence will focus on building efficient system architectures with high semantic consistency, strong cross-language transfer capabilities, and low computational resource consumption. And committed to building multilingual AI systems with lower computing power consumption, laying a solid foundation for cross-cultural intelligent communication.

References

1. A. Lastrucci, Y. Wandael, A. Barra, R. Ricci, A. Pirrera, G. Lepri, and D. Giansanti, "Revolutionizing radiology with natural language processing and chatbot technologies: a narrative umbrella review on current trends and future directions," *Journal of Clinical Medicine*, vol. 13, no. 23, p. 7337, 2024, doi: 10.3390/jcm13237337.
2. H. L. Ellis, and J. T. Teo, "The influence of AI in medicine," *Medicine*, vol. 52, no. 12, pp. 811-815, 2024.
3. K. V. Raja, R. Siddharth, S. Yuvaraj, and K. R. Kumar, "An Artificial Intelligence based automated case-based reasoning (CBR) system for severity investigation and root-cause analysis of road accidents-Comparative analysis with the predictions of ChatGPT," *Journal of Engineering Research*, vol. 12, no. 4, pp. 895-903, 2024.
4. A. Mamillapalli, B. Ogunleye, S. Timoteo Inacio, and O. Shobayo, "GRUvader: Sentiment-Informed Stock Market Prediction," *Mathematics*, vol. 12, no. 23, p. 3801, 2024, doi: 10.3390/math12233801.
5. D. Scharp, J. Song, M. Hobensack, M. H. Palmer, V. Barcelona, and M. Topaz, "Applying natural language processing to understand symptoms among older adult home healthcare patients with urinary incontinence," *Journal of Nursing Scholarship*, vol. 57, no. 1, pp. 152-164, 2025, doi: 10.1111/jnu.13038.