*European Journal of Engineering and Technologies*

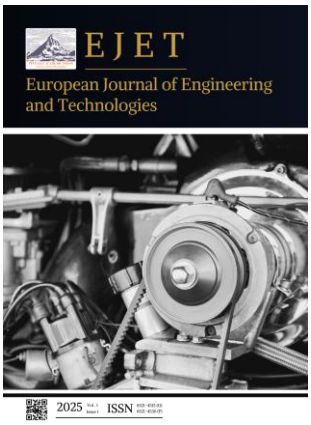
*Vol. 1 No. 3 2025*

*Article* **Open Access**

# Research on the Integration of Voice Control Technology and Natural Language Processing in Smart Home Systems

**Xiang Chen** [1,*]

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




Received: 19 August 2025
Revised: 02 September 2025
Accepted: 24 September 2025
Published: 01 October 2025



**Abstract:** This paper mainly focuses on the voice interaction system of smart homes, proposes an overall solution integrating voice control technology and natural language processing, and designs a modular framework integrating speech recognition, semantic understanding, multi-round dialogue, and feedback generation to enhance the system's understanding and response ability to user instructions. The experimental results show that this system outperforms traditional methods in terms of recognition accuracy, response speed, and dialogue experience, and has broad prospects and application value in terms of practicability and commercial promotion.

**Keywords:** smart home; voice control; natural language processing; speech recognition

## 1. Introduction

Voice control technology, due to its convenient operation and rapid response, is gradually becoming the main interaction method in smart home systems. However, the traditional voice interaction method triggered by instructions still has significant deficiencies in terms of comprehension ability, dialogue duration, and personalized response. In recent years, the development of natural language processing technology has provided a new breakthrough for improving the intelligence level of speech control technology. This paper establishes an intelligent voice interaction system by combining natural language processing and voice control technologies, so as to understand complex language information and conduct logical reasoning and effective judgment on it, thereby making human-computer dialogue more natural. This paper mainly studies the integration method of voice control technology and natural language processing, proposes a voice interaction system for smart homes based on semantic understanding, and conducts an in-depth analysis of the specific framework design and application effect of this system.

## 2. A Review of Voice Control Technology in Smart Home Systems

### *2.1. Fundamentals of Speech Recognition Technology*

For the voice control system of smart homes, speech recognition technology (ASR) is a key front-end module for achieving human-computer interaction. It converts users' voice information into processable text information, thereby providing a technical basis for semantic understanding and instruction execution at the back end [1]. This process includes voice collection, feature extraction, acoustic and language model matching, and decoding output. In recent years, speech recognition systems based on deep learning models such as CNN, LSTM, and Transformer have been widely applied, effectively en-

hancing the systems' capabilities in processing fluent speech, dealing with noise interference, and diverse pronunciations [2]. The CTC and attention mechanism structure has gradually replaced the traditional HMM-DNN architecture, improving the accuracy and response speed of the system. In the smart home scenario, the voice recognition system needs to have the ability to respond quickly while ensuring the accuracy of recognition.

### *2.2. Voice Wake-up and Voiceprint Recognition*

Voice wake-up and voiceprint recognition are the core technologies of the voice interaction system in smart homes. Voice wake-up is mainly used for system activation and user identity recognition. By using specific wake-up words, it can determine whether home appliances are in standby mode. Once awakened, it will switch to interaction mode and execute the next voice processing procedure. Voiceprint recognition is used to identify the identities of different users by analyzing the individual characteristics contained in the user's voice. This process combines the modeling of physiological and behavioral features in speech and can be used for the comparison and identity management of multi-person voices. Common voiceprint recognition technologies include i-vector, x-vector, and deep neural networks etc. This technology supports multi-user voiceprint comparison and identity management, enhancing the uniqueness and security of the system. The combination of the two can effectively enhance the stability and practicability of the system in complex home environments and is the basis for building an intelligent voice interaction platform [3].

## 3. The Role of Natural Language Processing Technology in Speech Interaction

### *3.1. Intent Recognition and Slot Extraction*

Intention recognition and slot extraction in the voice interaction system are important links in the voice interaction system of a smart home, directly affecting the intention reasoning effect of the system. Intent recognition is used to determine the operation goals expressed in user statements. For example, "Turn on the living room light" corresponds to the intent of light control, and "Adjust 26 degrees" corresponds to the intent of temperature control. Slot extraction extracts useful information such as "living room", "air conditioner", and "26 degrees" from it, forming structured instructions. Through this process, the system can not only understand the user's intention but also accurately match the corresponding devices and control parameters, achieving precise control of different devices and scenarios. The synergy of the two enhances the accuracy of semantic understanding and the contextual coherence of conversations, meeting the interaction requirements of smart homes in multi-device and multi-scenario environments [4].

### *3.2. Multi-Round Dialogue Management*

As an important component of the smart home voice interaction system, multi-round dialogue management is responsible for understanding the context and maintaining the dialogue status, enabling users to experience a natural interaction process that is semantically coherent and context-dependent. The main function is to track the user's historical instructions, analyze the omitted or substituted parts in the statements, and better respond to fuzzy instructions such as "a little darker". Through dynamic modeling of the context, the system can not only semantically understand complex instructions but also automatically clarify, complete, or confirm operations, further improving the task completion rate and error correction capability. In scenarios involving multi-device control and multi-parameter adjustment, multi-round dialogue management helps to plan the dialogue structure, thereby ensuring the consistency of semantic instructions during the transmission and execution process [5]. This mechanism significantly enhances the intelligence and flexibility of human-computer interaction, promoting the transformation of voice control from the traditional imperative input to the natural language dialogue mode

based on semantic understanding, thereby improving the smoothness of user operation and the corresponding personalization level of the system.

### *3.3. Natural Language Generation and Response*

In the voice interaction of smart homes, natural language generation (NLG) and response technologies mainly convert the processing results of the system into natural language that is easy for users to understand, and construct a human-machine feedback channel. Its functions are not limited to providing feedback on the execution results of tasks or operations, such as "The light has been turned off" or "the temperature has been set to 26°C", but also include various forms of information interaction, such as responding to user inquiries, warning reminders, and scene suggestions. By using template generation, semantic filling, or neural network generation methods, more natural and emotionally rich language expressions can be generated to enhance the naturalness and affinity of human-computer dialogue. NLG should also have the ability to maintain consistency in conversation topics, respond smoothly and coherently in multiple rounds of interaction, and enhance the naturalness and humanized experience of the overall conversation.

## 4. The Integrated Design of Voice Control Technology and Natural Language Processing in Smart Home Systems

### *4.1. Overall Architecture Design*

To achieve more natural and effective human-computer voice interaction in smart homes, voice control is deeply integrated with natural language processing technology to construct an overall system framework. The speech interaction system proposed in this paper is divided into five modules, namely speech acquisition and preprocessing, speech recognition (ASR), Natural language understanding (NLU), instruction generation and device control, and natural language generation and speech feedback (NLG). The voice acquisition module is mainly responsible for picking up sounds and performing noise reduction and echo cancellation. The ASR module converts speech into text and serves as the data source of the NLU module. The NLU module can accomplish intent recognition, slot extraction, and context management, and achieve semantic understanding. The control module can convert semantic information into structured instructions to drive the corresponding devices to perform specific operations. The NLG module generates feedback content based on the execution results and converts the text information into speech form through speech synthesis technology to provide voice feedback to the user. This system adopts a modular and decoupled design, supports local and remote collaborative execution, is compatible with a variety of intelligent devices, and is suitable for human-computer interaction in various complex environments. It has good scalability and compatibility, providing strong support and guarantee for realizing intelligent and personalized voice control systems (Figure 1).

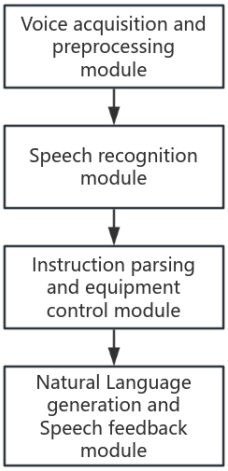



**Figure 1.** Architecture of the integrated system of voice control and natural language processing in a smart home.

*4.2. Module Function Division and Collaboration Mode*

In the voice interaction system of a smart home, the selection of functional modules should be reasonable and scientific. The smart home voice interaction system is composed of five modules: the voice acquisition module, the speech recognition module (ASR), the natural language understanding module (NLU), the command execution module, and the voice feedback module. Efficient connection and collaborative operation are achieved among various modules through a unified data structure and message mechanism.

The speech recognition module uses the acoustic feature vector $x = (x_1, x_2, \dots, x_T)$. Taking $x_T$ as the input, a deep neural network (such as LSTM or Transformer) is adopted to model the speech probability distribution $P(y_i|X)$, and the output text sequence $y = (y_1, y_2, \dots, y_N)$. The common training objective function is the cross-entropy loss:

$$L_{ASR} = -\sum_{i=1}^{N} log\ P\ (y_i|X) \quad (1)$$

After the text is recognized, it is passed into the NLU module for intent recognition and slot extraction. Intention recognition is usually modeled as a multi-class classification task, and its output probability is:

$$P(c|y) = soft\ max(W_c \cdot h + b_c) \quad (2)$$

Among them, h is the encoded semantic vector, and $W_c$ and $b_c$ are the parameters of the intent classifier. Slot extraction adopts A sequence labeling model, such as BiLSTM-CRF, to maximize the conditional probability of label sequence $s = (s_1, \dots, s_N)$:

$$L_{slot} = log\ P\ (s|y) \quad (3)$$

The above-mentioned structured semantic information is transformed into device control commands by the instruction mapping module and sent to the target device via the MQTT or HTTP protocol. The final result feedback is generated by the natural language generation module to produce response statements, and the language generation probability is calculated using the language model:

$$P(r|m) = \prod_{t=1}^{T} P(r_t|r_{<t}, m) \quad (4)$$

*4.3. Module Interface and Data Flow Design*

Clear module interfaces and efficient data flow design are important prerequisites for the stable operation and coordinated work of the smart home voice interaction system. The smart home voice interaction system adopts a modular structure. Each functional module uses an independent communication mode based on standard interfaces. The overall modular communication structure covers six major modules: voice collection,

voiceprint recognition, semantic understanding, command generation, device control, and voice feedback. Communication among various modules is completed through RESTful API or MQTT protocol to enhance the scalability and maintainability of the system.

In terms of the data flow, the system collects and preprocesses the user's voice information through the voice acquisition module, and transmits the processed voice information to the voice recognition module to convert it into text information. The natural language understanding module extracts the intent and slot information in the text, generates structured semantic data, and, based on the semantic data and scene logic, generates control commands and transmits them to the corresponding devices. After the task execution is completed, the voice feedback module uses natural language generation technology to synthesize sentences and broadcasts them to the user through voice synthesis, achieving a complete human-computer interaction closed loop. The output data format of the natural language understanding module in the system is as follows (Figure 2):


```
{
  "intent": "control_device",
  "slots": {
    "device": "light",
    "location": "bedroom",
    "action": "turn_on"
  }
}
```


**Figure 2.** Semantic structured data format output by the NLU module.

This structure is transmitted between modules through a unified interface, which is not only simple and clear but also easy to expand and debug.

Furthermore, the instruction generation module will generate intermediate control commands based on this structure, for example (Figure 3):


```
{
  "device_id": "light_bedroom_01",
  "command": "ON",
  "timestamp": "2025-06-20T10:45:32Z"
}
```


**Figure 3.** shows the execution instruction structure generated by the equipment control module.

Such control commands can be directly issued to the central control system or connected to the protocol gateway to achieve fine control of specific intelligent terminals.

*4.4. Scene Modeling and Intent Mapping*

In the voice control system of a smart home, users express their operation intentions through natural language, but these languages have problems such as ambiguity, diversity, and context dependence. To achieve precise control, it is necessary to construct a scene modeling and intention mapping mechanism. Scene modeling involves semantic abstraction of the home environment, device categories, and personal behaviors to form a unified information representation form, enabling the system to identify the specific directions of words such as "bedroom", "curtains", and "turn on the light", and then perform further operations and bind the devices with scene semantics.

Based on the results of scene modeling, the intent mapping technology converts natural language instructions into standard command syntax structures that the system can execute. This process is accomplished by the natural language parsing module and usually includes two stages: intention recognition and slot extraction. Intention recognition is used to determine the target of the user's operation, such as "playing music" or "adjusting

temperature". Slot position extraction is used to extract key information such as equipment names, spatial positions, and numerical parameters, which is provided for the backend to generate and execute control instructions. To achieve stable transmission and system compatibility, user intentions and their parameters are usually expressed in a unified data format.

## 5. Experimental Design and Result Analysis

### *5.1. System Setup and Experimental Environment Description*

This research builds a smart home experimental system integrating speech recognition and natural language processing based on a modular speech interaction framework. The system can be divided into four modules, namely speech recognition, natural language understanding, instruction mapping, and speech feedback. To further enhance the effect, a hybrid architecture of on-premises deployment and cloud collaboration is adopted. The Speech recognition part is based on the Deep Speech engine and improves the accuracy of Mandarin speech recognition through noise suppression algorithms. Semantic processing adopts the BERT model for intent recognition and slot extraction, and realizes dialogue process management through the Rasa platform. The control module introduces the MQTT protocol to convert semantic information into smart home control instructions, which are used to control various devices such as lights, air conditioners, and sockets.

### *5.2. Test Indicators and Experimental Methods*

For the smart home control system integrating voice control and natural language processing (NLP) technology proposed in this paper, to comprehensively evaluate its performance, four evaluation indicators centered on recognition accuracy, semantic understanding ability, interaction response efficiency, and user experience are designed, and an overall evaluation framework is constructed, supplemented by testing methods. During the experiment, the performance of the speech recognition module was mainly measured and analyzed through the Speech Recognition Accuracy Rate (WAR). Its calculation formula is:

$$WAR = \left(1 - \frac{S+C+I}{N}\right) \times 100\% \tag{5}$$

This indicator can effectively reflect the system's ability to accurately restore users' voice instructions.

The semantic understanding ability is evaluated through the accuracy rate of intent recognition and the accuracy rate of slot extraction, which respectively reflect the system's ability to determine user behavior goals and the ability to extract parameters. Their calculation methods are respectively:

$$\text{Intent Accuracy} = \frac{C_{\text{intent}}}{N_{samples}} \times 100\% \tag{6}$$

$$\text{Slot Accuracy} = \frac{C_{\text{slots}}}{T_{slots}} \times 100\% \tag{7}$$

The system response efficiency is measured by the Mean Response Time (ART), which is defined as the average time consumed by the system from receiving a voice instruction to feeding back the execution result:

$$\text{ART} = \frac{1}{n}\sum_{i=1}^{1} t_i \tag{8}$$

In complex interaction scenarios, to further evaluate the context processing ability and conversation coherence of the system, this paper introduces the multi-round dialogue completion rate index:

$$DRR = \frac{C_{\text{dialongue}}}{T_{\text{dialongue}}} \times 100\% \tag{9}$$

*5.3. Experimental Results and Comparative Analysis*

To verify the advantages of the designed system in speech recognition, natural language understanding, and overall interaction performance, a set of comparative experiments was carried out in this paper, and the performance of this system was evaluated and tested with the traditional system based on rule matching and keyword extraction. The experiment was conducted in a unified experimental environment, focusing on six indicators: the accuracy rate of speech recognition, the accuracy rate of intention recognition, the accuracy rate of slot extraction, the average response time, the completion rate of multi-round conversations, and user satisfaction. The empirical results show that in terms of the accuracy rate of speech recognition, the system designed in this paper has reached 93.8%, which is 8.2% higher than the 85.6% of the traditional design. In terms of the accuracy rates of intention recognition and slot extraction, this system has reached 92.5% and 90.2% respectively, which is more than 6% higher than that of traditional systems. This indicates that modeling prior language patterns and semantic scenarios is helpful to enhance the semantic understanding ability of the system. In terms of response time, the average response time of this system remains within the range of 1.32 seconds, while traditional systems have obvious response delays, with an average response time of more than 2.05 seconds, which can have a negative impact on the user experience. In terms of the completion rate of multi-round dialogues, this system supports context management and intention continuity modeling, with a completion rate of 88.0%, while that of traditional systems is only 71.0%. The subjective satisfaction score of users of the smart home system is also significantly higher than that of the traditional system, with the average score increasing by 3.9 points. The smart home system, after integrating voice control and natural language processing technology, performs better in terms of recognition accuracy, interaction smoothness, and user experience, showing strong market promotion potential (Table 1).

**Table 1.** Results of the system performance comparison experiment.

| Test indicators | The result of this system | Traditional system results |
|---|---|---|
| Speech Recognition Accuracy (WAR) | 93.8% | 88.2% |
| Accuracy rate of intention recognition | 92.5% | 86.1% |
| Slot position extraction accuracy rate | 90.2% | 80.4% |
| Average response time (seconds) | 1.32 | 2.05 |
| Completion rate of multiple rounds of dialogues | 88.0% | 71.0% |
| User satisfaction (out of 5 points) | 4.6 | 3.9 |

## 6. Conclusion

This paper researches the integration of voice control technology and natural language processing in the smart home system, and constructs an integrated system architecture. Through experiments, it is proven that compared with the traditional solution, this system has greatly improved in terms of recognition accuracy, response speed, and user experience, further verifying the effectiveness of the deep integration of voice and semantics in improving the overall intelligence level of the system.

## References

1. Z. Li, "Research and implementation of low resource voice awakening technology in smart home scene," Acad. J. Comput. Inf. Sci, vol. 7, no. 11, pp. 96-101, 2024, doi: 10.25236/AJCIS.2024.071113.
2. M. Pilz, S. Zimmermann, J. Friedrichs, E. Wördehoff, U. Ronellenfitsch, M. Kieser, and J. A. Vey, "Semi-automated title-abstract screening using natural language processing and machine learning," Systematic Reviews, vol. 13, no. 1, p. 274, 2024, doi: 10.1186/s13643-024-02688-w.

3. U. F. Dimlo, V. Rupesh, and Y. Raju, "The dynamics of natural language processing and text mining under emerging artificial intelligence techniques," International Journal of System Assurance Engineering and Management, vol. 15, no. 9, pp. 4512-4526, 2024, doi: 10.1007/s13198-024-02468-8.
4. N. K. Baqer, Y. J. Harbi, and H. A. J. Al-Asady, "Impact of Delay in ZigBee WSNs for Smart Home Applications," Journal of Engineering Research and Reports, vol. 26, no. 8, pp. 372-380, 2024, doi: 10.9734/jerr/2024/v26i81252.
5. Y. Fan, "Research on the Application of Artificial Intelligence in Smart Home Systems and Its Impact on Privacy Protection," The Frontiers of Society, Science and Technology, vol. 6, no. 6, 2024, doi: 10.25236/FSST.2024.060620.