Development Of Smart Homes Under the Integration of Artificial Intelligence

Zurui Yang *

Department of Japanese, Xiamen University Tan Kah Kee College, Zhangzhou, 363000, China

* Corresponding Author Email: merodei509@gmail.com

Abstract. With the development of the times, intelligence has become pervasive. As an important bridge connecting humans and machines, speech recognition technology is driving the intelligent development of people's lives. Among its applications, the use in the smart home sector stands as a major focus of this technology. It breaks the limitations imposed by traditional manual operations, bringing convenience to daily life in the simplest manner. This paper summarizes the current development status and principles of speech recognition technology. By integrating its application in smart homes, it introduces two core aspects, namely whole-house device control and personalized services. Additionally, taking the Xiaodu Smart Speaker as an example, the technology is elaborated upon. Meanwhile, three major bottlenecks currently facing speech recognition technology are pointed out, namely feature distortion caused by noise issues, inaccurate recognition, and vague commands resulting from dialects, and semantic misunderstandings arising from colloquial language. In response to these problems. This paper analyzes these issues and proposes targeted optimization solutions for the technology.

Keywords: Artificial Intelligence; Speech Recognition; Smart Home.

1. Introduction

In today's society, Artificial Intelligence (AI) has become a key focus of social attention. As one of the critical technologies in this field, speech recognition has brought significant changes to the era of intelligence. Voice, as the most natural medium for human-computer interaction, serves as a crucial hub connecting users with artificial intelligence. Empowered by advanced technologies such as AI and speech recognition, smart home systems are rapidly integrating into people's lives [1]. The development of smart homes not only facilitates daily life but also enables people with disabilities and the elderly to use them with ease [2]. According to predictions and analyses by the World Economic Forum (WEF), the market size of smart homes is projected to reach $13 trillion by 2030 [3].

With continuous breakthroughs in AI technology, speech recognition has been continuously endowed with new capabilities. As one of its core application scenarios, smart homes have achieved a significant transformation from single-device control to whole-house interaction, bringing intelligent experiences to people's lives. Enterprises like Xiaomi and Midea, through proprietary vertical-domain large models, enable their voice systems to accurately recognize multi-device linkage commands—such as turning on the living room air conditioner and adjusting the temperature to 26°C—thereby improving the accuracy of semantic understanding.

The deep integration of AI and speech recognition technology drives smart homes toward a more intelligent stage. Although speech recognition has reached a certain level of maturity, it still has shortcomings and faces many challenges, which can be specifically broken down into three key difficulties, namely noise interference, dialectal accents and semantic misunderstandings. By sorting out and integrating the current technological landscape, targeted optimization solutions are proposed here, providing references for research in this field and helping to promote the more user-friendly development of smart homes.

2. Principles of Speech Recognition

Speech recognition technology first emerged in the 1950s, when Bell Labs developed the Audrey system, capable of simple recognition of ten English digits [4]. After over 70 years of development, the core technology of speech recognition has transitioned from early statistical models like the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) to neural network technology driven by deep learning, achieving significant progress. Currently, overseas research on speech recognition mainly leans towards deep learning, end-to-end speech recognition, and enhancement, while domestic research focuses on end-to-end speech recognition, speech enhancement, and multi-language recognition [5]. At present, polymorphic models based on Transformer have become the mainstream technology in the field of speech recognition.

Essentially, the function of speech recognition is to convert the information contained in human language into specific commands that machines can understand and execute [6]. The steps of speech recognition can be broadly divided into four core stages, namely information processing, feature extraction, model matching and result output. As shown in Figure 1, this diagram illustrates the process of speech recognition technology. A human voice drives the diaphragm inside the microphone to vibrate, using principles such as electromagnetic induction to achieve the conversion of sound waves into analog electrical signals. After pre-processing steps like noise reduction and filtering, a relatively clean speech signal is obtained [6].

Next comes the feature extraction stage, where key parameters representing the essence of speech are extracted. In this process, Mel Frequency Cepstral Coefficients (MFCCs) are the most widely used classic method. Then follows model matching, the core link of speech recognition. With the collaboration of acoustic models (e.g., DNN-HMM, end-to-end models) and language models, machines achieve the transition from hearing to understanding. The main task of the acoustic model is to establish probabilistic correlations between speech feature vectors and basic units of speech (i.e., phonemes). Current mainstream approaches include Deep Neural Network-Hidden Markov Model (DNN-HMM) and end-to-end models. The language model, in turn, optimizes the rationality and accuracy of recognition results, ultimately generating preliminary text outcomes in this stage. Result output involves converting the text into device commands and finally executing them.

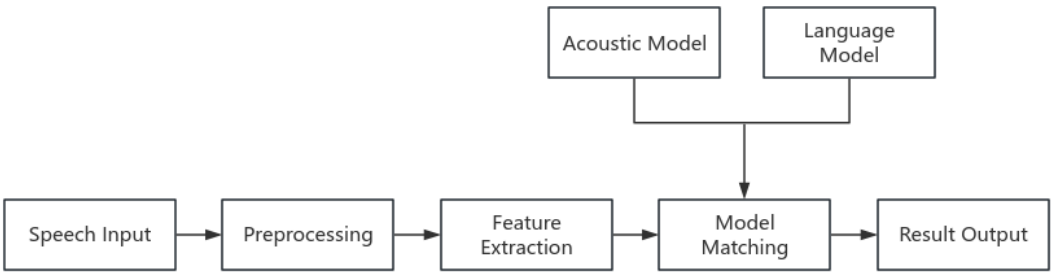

Fig. 1 Speech Recognition Flow Chart

3. Applications of Speech Recognition in Home Life

With the development of speech recognition technology, traditional manual interactions are gradually being replaced. Daily life at home is one of the most widely applied fields for this technology. Speech recognition addresses the pain points of illiteracy and complex operation among children and the elderly, while also making life more convenient by freeing hands. Users only need to convey commands through simple verbal instructions, with the core purpose of simplifying operational processes and adapting to diverse scenarios. As a link connecting users with smart home systems, speech recognition not only enables faster control of various household facilities but also allows for scheduling daily tasks [7]. With the advancement of future technologies, people's daily lives will become increasingly simplified and intelligent.

The accuracy rate of speech recognition is the core of its integration into the smart home field. Taking the IFLYTEK Input Method as an example, in standard scenarios of daily conversations, its accuracy rate for Mandarin speech recognition can reach 99.4% [8]. Additionally, the IFLYTEK Spark X1.5 and a series of AI software-hardware integrated solutions released by IFLYTEK have further driven the advancement of speech recognition technology. In a factory noise environment of 90 decibels, the recognition accuracy rate can reach 98.69%, while in scenarios like subways and bus stations, the accuracy rate is as high as 97.1%. Moreover, with just one recording, any timbre can be replicated, and sounds of any style can be created with a single command. These have brought users an extremely high-quality experience.

3.1. Whole-Home Device Control and Personalized Services

The core of speech recognition technology in smart homes lies in whole-house device control and personalized services. Whole-house device control effectively addresses efficiency issues during multi-device collaborative work, while personalized services provide more convenient and intelligent experiences to meet the needs of different types of users.

In whole-house device control, with a voice interaction hub (such as smart speakers) as the core, various smart devices in the home are connected via communication protocols (like WiFi, ZigBee, etc.) to enable centralized management of different devices. This is further divided into single-device precise control and multi-device collaborative linkage. Single-device precise control achieves directional operations of specific commands for a single device through speech recognition technology. In contrast, multi-device collaborative control triggers simultaneous responses from multiple devices via a single-sentence command, shifting from device intelligence to scenario intelligence.

Personalized services not only enhance the convenience of use but also bring exclusive experiences. Based on voice recognition, they better integrate context and proactively optimize response methods by learning users' living habits [6]. It places greater emphasis on making smart homes no longer mere combinations of technology, but entities that align more closely with life needs and provide practical assistance to people's lives.

For example, in an article co-authored by Jiao Limin, Qu Zongfeng, Li Hongwei, Liu Zechao, and Hu Yaxin, a coupling solution of speech interaction and large language models (similar to GPT) was proposed to address the pain point of smart homes not understanding human language in daily life. By integrating factors such as context and environment, this solution effectively improves such issues [9]. Additionally, Haier Smart Home launched a new whole-home smart solution named 1+3+5+N, which also promotes the development of whole-home intelligence and integration. Leveraging the whole-home neural network unit, it gains insights into users' various needs to meet their diverse requirements. For instance, it creates a romantic dining atmosphere during meals or provides soft lighting when sleeping. Furthermore, on August 1, 2025, Midea Group released a new whole-home smart solution, which mainly consists of three parts namely AI-assisted functions, a linkage system between smart homes and appliances, and advanced technologies that strengthen interconnections among people, vehicles, and houses [10]. This not only makes life smarter and more convenient, simplifies operations, but also optimizes the issue of coherent device usage across different scenarios. Through this solution, as long as a vehicle model cooperating with Midea is driven within the home range, the lights will automatically turn on and curtains will open inside, etc.

3.2. Typical Cases of Speech Recognition

Taking the Xiaodu smart speaker as an example, when on standby, the device remains in a low-power listening state. It uses a microphone array to capture ambient sounds, and upon detecting the user's wake-up command, it awakens and processes the request accordingly.

For smart home appliances like smart TVs and air conditioners, users must first bind them to the Xiaodu app. Once paired successfully, this information is synchronized to a cloud database, serving

as the core basis for subsequent recognition. Traditional appliances, on the other hand, can be controlled by certain Xiaodu models equipped with infrared functionality.

The Xiaodu smart speaker receives the text information decoded by the speech cloud, performs semantic analysis to understand the user's intent and convert the command [11]. For example, when the remote control for an air conditioner or TV is misplaced, one can simply wake up Xiaodu and issue a voice comman to control the device directly.

4. Existing Bottlenecks in the Development of Smart Homes Driven by Speech Recognition Technology

Despite the remarkable progress achieved in the field of smart homes empowered by speech recognition technology, it still faces numerous challenges. For instance, under the influence of noisy environmental factors, the recognition accuracy drops significantly. The accuracy is notably low when recognizing dialects, accents, or colloquial expressions. Additionally, a single sentence may sometimes carry implicit meanings. These are all major challenges in the development of speech recognition technology.

Regarding the issue of noise, feature extraction stands as a core step in the speech recognition process. To better analyze the information conveyed by a speech signal, the system performs a Fast Fourier Transform (FFT) on each frame of the signal to obtain a spectrogram. By arranging these spectrograms in chronological order, a complete speech spectrogram is formed, which serves as the core basis for subsequent steps. However, noise interference directly impacts the spectrogram at the spectral level, thereby causing deviations in the subsequent calculation of MFCC (Mel-Frequency Cepstral Coefficients) features.

The dialect system is highly complex. The uniqueness of pronunciation rules in different regions, significant differences from Mandarin, and the frequent use of dialect vocabulary mixed with Mandarin sentence structures by middle-aged and elderly users all pose enormous challenges for speech recognition technology. Although current speech recognition systems have made some progress in identifying mainstream dialects such as Cantonese and Sichuan dialect, they still cannot support less widely spoken dialects, making it difficult to meet the interaction needs of multi-regional families.

The language people use in daily communication tends to be more colloquial, which is also one reason why machines struggle to understand human speech. The colloquialism and ambiguity of voice commands can lead to repeated or incorrect execution of instructions. The core reasons for this can be roughly summarized as insufficient corpus coverage, lack of contextual association ability and unclear user intent. Currently, the technology is still in the stage of merely processing received voice commands and is not yet proficient in integrating the context of the speaker’s situation and inferring the speaker’s intent.

5. Countermeasures for Speech Recognition Technology to Address Challenges

First, regarding the approach to addressing the noise issue, noise in home environments is characterized by randomness and diversity. Therefore, spectral subtraction can be used to filter out some stationary noise. In addition, beamforming technology of microphone arrays can be leveraged to focus on the direction of the user's voice input, effectively reducing interference from non-stationary noise. Since noise can cause shifts in the Mel-Frequency Cepstral Coefficients (MFCC) features of speech, Cepstral Mean Subtraction (CMS) can also be applied—by statistically calculating the average cepstrum of noisy speech over a period of time and subtracting it, the overall interference from noise can be significantly reduced, restoring the true distribution of speech features [12].

In optimizing dialect recognition, systematically collect command corpora of dialects in home control scenarios and construct corresponding corpora. A basic dialect recognition model can be built based on the Transformer architecture, where its self-attention mechanism captures dependencies

between different positions in the speech sequence. Parallel computing is utilized to improve training efficiency and shorten the training cycle for dialect corpora, and combined with transfer learning to enhance recognition accuracy.

For semantic understanding, a semantic intent library dedicated to smart homes can be constructed to clarify the true intents and key parameters corresponding to ambiguous commands. Contextual modeling techniques such as Transformer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) can be introduced, and combined with Dialogue State Tracking (DST) to alleviate issues like semantic disambiguation and multi-turn intent coherence to a certain extent.

6. Conclusion

With the deep integration of artificial intelligence and speech recognition technology in smart homes, people's daily lives have undergone leapfrog improvements. Language interaction between humans and smart homes not only effectively lowers the operational threshold for users but also marks a monumental shift from manual to automatic operation. Although current speech recognition technology has achieved certain milestones, many difficulties still stand in the way of its development. At the same time, these challenges continuously drive the ongoing advancement of speech recognition technology. Looking ahead, with the continuous progress of technology, the connection between AI, speech recognition, and smart homes will only grow closer. Smart homes may further achieve deeper penetration into whole-home scenarios, refining diverse application contexts and catering to the needs of different user groups through iterative technological upgrades. Empowered by both enhanced intelligence and simplified usability, they will continue to elevate the quality of people's lives.

References

[1] Zhang H Y. Potential application risks and protection methods of voice assistants in smart homes. Secrecy Science and Technology, 2024, (7): 64-70.

[2] Chakraborty A, Islam M, Shahriyar F, et al. Smart home system: a comprehensive review. Journal of Electrical and Computer Engineering, 2023, 2023(1): 7616683.

[3] World Economic Forum. Digital Transformation Initiative. Cologny, Switzerland: World Economic Forum, 2018.

[4] Zhu Y. Continuous speech recognition based on deep learning. Guilin University of Electronic Technology, 2024.

[5] Qin K X, Wang W X, Wang Y S. Research on robot speech recognition method based on improved MFCC feature extraction and DNN network. Computer Measurement & Control, 2025, 33(2): 246-253.

[6] Li C. Application of speech recognition technology in smart home control systems. Audio Engineering, 2025, 49(6): 49-51.

[7] Zhang J. Application of artificial intelligence in speech recognition. Computer Knowledge and Technology, 2024, 20(17): 46-48.

[8] Cao C, Wang G. Evaluation of intelligent speech technology in epidemic prevention: Take iflytek input software in Chinese and Japanese recognition as an example. In: Journal of Physics: Conference Series. IOP Publishing, 2020, 1631: 012047.

[9] Jiao L M, Qu Z F, Li H W, et al. Research on the construction of smart home voice interaction based on the ChatGPT mechanism. China Standardization, 2023, (11): 88-92.

[10] Chen J B. Midea and Huawei jointly build the "second growth curve" of the industry smart home unlocks a trillion-yuan market. China Business Journal, 2025-10-13(B12).

[11] Wang Q H, Lin J P. Application of voice cloud platform in smart homes. Information and Computers (Theoretical Edition), 2024, 36(9): 118-120.

[12] O'Shaughnessy D. Automatic speech recognition: History, methods and challenges. Pattern Recognition, 2008, 41(10): 2965-2979.