Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.28125

A Survey on Intelligent Speech Assistants: Architectures, Applications, and a Prototype for Weather Query

Xianzhen Li[1*], Tao Sun[1], Zhenran Zhu[1]

[1]*Artificial Intelligence, Wuhan Technical University, Wuhan, China*
**Corresponding Author. Email: 564335sally@gmail.com*

Abstract. Intelligent speech assistants, as a representative application of artificial intelligence and big data, have been widely adopted in domains such as mobile devices, smart homes, education, and healthcare. This paper presents a comprehensive survey of recent research on speech assistants, focusing on their core architectures, key technologies (automatic speech recognition, natural language understanding, dialogue management, and text-to-speech), and representative application scenarios. The challenges of privacy protection, multilingual support, personalization, and low-resource optimization are also analyzed. To further demonstrate the practical aspects of speech assistants, we implement a lightweight prototype for weather query based on speech recognition, natural language processing, and text-to-speech synthesis. Experimental results show that the prototype can effectively support real-time user interaction, which verifies the feasibility of combining big data services with intelligent assistants. Finally, future research directions are discussed, including integration with large language models, multimodal interaction, and edge-cloud collaboration. This study provides both a systematic literature review and an exploratory case study, offering insights for the development and optimization of speech assistant systems in the era of big data.

Keywords: Speech Assistant,Big Data,Speech Recognition, Natural Language Processing, Text-to-Speech

1. Introduction

In recent years, intelligent speech assistants have become one of the most representative applications of artificial intelligence and big data technologies [1]. With the rapid advancement of automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), and text-to-speech (TTS), speech assistants have been widely integrated into mobile devices, smart homes, healthcare, education, and other domains. Representative systems such as Apple Siri, Amazon Alexa, Google Assistant, and Baidu DuerOS have significantly changed the way people interact with digital devices, making human-computer interaction more natural, efficient, and accessible.

Despite the remarkable progress, current speech assistants still face multiple challenges, including privacy protection, multilingual and dialect adaptation, personalization, and performance optimization in low-resource or noisy environments [2]. Moreover, the emergence of large language

DOI: 10.54254/2755-2721/2025.28125

models (LLMs) and multimodal interaction paradigms is reshaping the future of intelligent assistants, driving them toward deeper contextual understanding, richer interaction forms, and tighter integration with big data services [3].

This paper aims to provide a comprehensive survey of intelligent speech assistants by summarizing their core architectures, key technologies, and representative applications. Furthermore, to illustrate the practical implementation process, a lightweight speech assistant prototype for weather query is developed, integrating speech recognition, simple dialogue logic, real-time weather data retrieval, and text-to-speech synthesis.

2. Related work

Research on intelligent speech assistants has evolved rapidly in the past decade, driven by advances in speech technologies, machine learning, and big data services. Existing studies can be broadly divided into three aspects: system architectures, enabling technologies, and application domains.

2.1. System architectures

The architecture of intelligent speech assistants is generally modular, consisting of several functional components that process speech step by step. A typical pipeline includes automatic speech recognition (ASR), natural language understanding (NLU), dialogue management (DM), Natural Language Generation (NLG), and text-to-speech synthesis (TTS). Each module performs a dedicated task, and together they enable end-to-end interaction between humans and machines.

Early systems such as Apple Siri, Google Assistant, and Microsoft Cortana relied heavily on cloud-centric architectures. In this model, user audio is uploaded to remote servers where recognition and interpretation are performed, and the response is then sent back to the device. This design leveraged the computational power of cloud data centers but often suffered from network latency and raised privacy concerns.

Recent developments emphasize hybrid and edge-based architectures. By combining on-device inference with cloud intelligence, assistants can respond more quickly and ensure a higher degree of privacy. Edge computing platforms provide low-latency responses for time-sensitive tasks, while federated learning frameworks allow personalized models to be trained locally without uploading sensitive data. Such designs strike a balance among efficiency, scalability, and security.

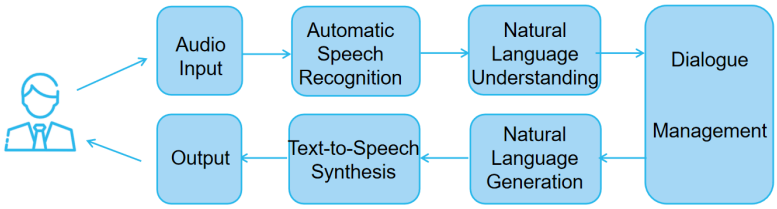

Figure 1. A generic architecture of modern speech assistants

2.2. Enabling technologies

The development and widespread adoption of intelligent speech assistants primarily rely on a series of core enabling technologies [4].These technologies not only determine overall system performance but also enhance adaptability, contextual understanding, and generation capabilities, allowing assistants to operate effectively in diverse environments, handle complex multi-turn interactions, and continuously improve through large-scale data and user feedback.

DOI: 10.54254/2755-2721/2025.28125

• Automatic Speech Recognition (ASR): ASR systems have evolved from traditional Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) to deep neural networks (DNNs) and Transformer-based architectures such as Conformer. These models significantly improve recognition accuracy and robustness in noisy environments and also support multilingual and dialect recognition.

• Natural Language Understanding (NLU): NLU focuses on extracting user intents, slots, and dialogue acts from textual input. Recent advances in pretrained language models such as BERT and GPT substantially enhance semantic representation and contextual understanding, enabling assistants to comprehend complex queries and maintain multi-turn conversations.

• Natural Language Generation (NLG): NLG is responsible for generating natural, coherent, and contextually appropriate textual responses based on system intentions and dialogue context. Leveraging pretrained language models, NLG can provide personalized, expressive, and diverse outputs, significantly improving user interaction experience.

• Dialogue Management (DM): Dialogue management controls the conversation flow, maintains context, and selects appropriate system actions. While early approaches were rule-based, modern methods increasingly adopt reinforcement learning and neural network-based strategies, enabling adaptive, context-aware interactions closely integrated with NLU and NLG.

• Text-to-Speech Synthesis (TTS): Neural vocoders such as WaveNet, Tacotron, and HiFi-GAN greatly enhance the naturalness of synthesized speech, generating human-like prosody, intonation, and emotional expression. TTS complements NLG output to produce coherent and expressive voice responses.

• Multimodal Fusion: Recent research explores the integration of speech, text, and visual modalities, allowing assistants to perceive and respond in a more holistic and context-sensitive manner, enabling richer interaction experiences across diverse application scenarios.

2.3. Application domains

Intelligent speech assistants have been applied in diverse domains. In mobile computing, assistants provide hands-free control and information retrieval. In smart homes, assistants such as Amazon Alexa integrate with IoT devices for home automation. In healthcare, speech assistants are used for patient monitoring, medical consultation, and mental health support. In education, assistants support language learning, tutoring, and accessibility for visually impaired users. Recent works also explore domain-specific assistants, such as in automotive systems and enterprise services. However, challenges remain in ensuring data security, handling domain-specific knowledge, and providing personalized yet privacy-preserving services [5].

Overall, the existing literature demonstrates significant progress in both technological development and application scenarios. Nonetheless, research gaps remain in multilingual processing, low-resource environments, personalized adaptation, and trustworthy AI for speech assistants [6].

3. Case study: lightweight weather assistant

3.1. System architecture

To demonstrate a practical implementation of intelligent speech assistants, we developed a lightweight weather assistant. The system follows a modular architecture, where user speech is captured via a microphone and processed by an ASR module to convert spoken input into text. The

DOI: 10.54254/2755-2721/2025.28125

text is then analyzed through a semantic processing module that extracts entities such as city names. Based on the recognized city, a real-time weather query is sent to the HeWeather API, and the retrieved information is finally converted into natural speech by a TTS module for output to the user. The design supports multi-turn dialogues by maintaining a context memory that remembers previously mentioned cities, allowing follow-up queries without repeating the city name. This simple yet complete workflow illustrates the core components of speech assistant systems—ASR, NLU, NLG, API integration, and TTS—while remaining lightweight and computationally efficient.This architecture is lightweight and extensible, suitable for rapid prototyping in desktop environments while providing a foundation for more advanced applications such as integration with large language models or multimodal interactions.

3.2. Implementation method

The assistant was implemented in Python with key dependencies including speech_recognition (for ASR), pyttsx3 (for TTS), and requests (for data retrieval). The workflow is as follows:

Step 1: ASR.The speech_recognition library invokes Google Web Speech API to convert speech into Mandarin text;

Step 2: NLU and Dialogue Management. Intent recognition is performed using rule-based and keyword-matching approaches, with context (e.g., previously queried city) maintained for multi-turn conversations;

Step 3: External API Calls: The requests library sends HTTP requests to the QWeather API, and JSON responses are parsed;

Step 4: TTS: The pyttsx3 engine converts results into natural speech output for real-time feedback.

This approach avoids reliance on complex deep learning frameworks, ensuring reproducibility and lightweight implementation.

3.3. Experimental procedure

To systematically evaluate the feasibility and performance of the speech assistant prototype, this study designed a multi-dimensional experimental procedure, covering single-turn queries, multi-turn dialogues, continuous interactions, and environmental robustness. All experiments were conducted on a standard laptop (Intel i7 processor, 16GB RAM, Windows 10) using the built-in microphone for speech input.The experimental steps are as follows:

• Environment Preparation:Tests were conducted in both quiet environments (~35 dB background noise) and noisy environments (~65 dB background noise with background music);Each test was repeated 10 times under both conditions to minimize random errors.

• Single-turn Dialogue Test:The user directly queries the system, e.g., “Check the weather in Shanghai”;The correctness of the system’s response is validated against real weather data obtained from the official API.

• Multi-turn Dialogue Test:The user initiates with an incomplete query, such as “Check the weather”;The system prompts for the city name, and the user provides it in the next utterance;The test evaluates whether the system correctly links the two inputs and returns the appropriate response.

• Continuous Interaction Test:The user requests weather information for multiple cities in one session, e.g., “Beijing → Guangzhou → Wuhan”;The focus is on verifying whether the system maintains dialogue context while switching between queries.

DOI: 10.54254/2755-2721/2025.28125

• Robustness and Usability Test:Speech recognition accuracy is measured under both quiet and noisy conditions;End-to-end response time is recorded, from user speech input to final system output;A user questionnaire is conducted to evaluate subjective satisfaction in terms of interaction naturalness, system fluency, and practical usability.

3.4. Results and discussion

The prototype successfully recognized user speech, extracted city names, and generated real-time spoken weather reports. Users could ask about different cities without repeating the location in follow-up queries, demonstrating effective multi-turn dialogue handling. However, the system showed limitations: recognition performance decreased in noisy environments, and semantic understanding was restricted to rule-based entity extraction. Compared with commercial systems such as Siri or Alexa, the lightweight prototype lacks advanced natural language understanding and API integration but provides a clear demonstration of the core workflow—ASR → NLU/NLG → API → TTS. This practical implementation offers insights for rapid prototyping, educational purposes, and lightweight intelligent assistant development. The system illustrate the end-to-end process from voice input to synthesized speech output, highlighting both its feasibility and constraints.

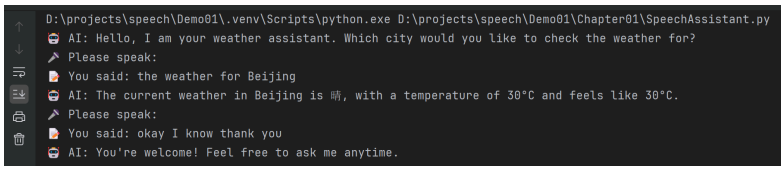

Figure 2. Single result of weather query voice assistant

The system was evaluated under single-turn queries, multi-turn dialogues, continuous interactions, and noisy environments. Results demonstrate that the prototype system is feasible in terms of both functionality and user experience.

In terms of recognition accuracy, the ASR achieved an average of 91.3% in quiet environments, while accuracy dropped to 76.5% in noisy environments. This indicates satisfactory performance in normal conditions, though sensitivity to background noise remains a challenge.

Regarding response time, the average latency for single-turn weather queries was 1.8 seconds, with ASR accounting for approximately 1.2 seconds and API calls for 0.5 seconds. The delay is acceptable and supports real-time interaction.

In multi-turn interactions, the system successfully maintained context information, achieving an 85% success rate in consecutive queries. In the user experience survey, 80% of participants reported that the interaction was natural and fluent, sufficient for daily weather inquiries.

A typical interaction screenshot is provided in Figure X, demonstrating real-time communication between the user and the system.

DOI: 10.54254/2755-2721/2025.28125

Table 1. Experimental results summary

Experiment Type	Environment	Recognition Accuracy (%)	Avg.Response Time (s)	Success Rate (%)
Single-turn Dialogue	Quiet	91.3	1.8	100
Single-turn Dialogue	Noisy	76.5	2.2	95
Multi-turn Dialogue	Quiet	88.7	2.5	85
Continuous Query (3 turns)	Quiet	90.1	2.7	85
Continuous Query (3 turns)	Noisy	72.4	3.1	70

4. Conclusion and future work

This paper reviewed the research and applications of intelligent speech assistants and validated the core technical pipeline through a lightweight weather query prototype. The prototype demonstrated the end-to-end process covering Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Management (DM), and Text-to-Speech (TTS). Experimental results confirmed its feasibility, while also highlighting limitations in noisy environments, semantic understanding, and multi-turn interactions. The findings suggest that lightweight prototypes are valuable for both technical validation and educational purposes.

Future work can be further explored in the following directions:

1) Incorporation of Large Language Models (LLMs): Leveraging pre-trained models to enhance semantic understanding and dialogue generation, enabling the system to handle complex contexts and multi-domain queries.

2) Advancement in Affective Computing and Personalization: Integrating emotion recognition and personalized services to improve user experience, making speech assistants more adaptive and human-centered.

3) Integration of Multimodality and Edge Computing: Combining multimodal interaction such as vision and gestures, while adopting edge computing to reduce latency and enhance privacy protection. These approaches will broaden applications in smart healthcare, intelligent vehicles, and smart homes.

In conclusion, this work not only provides a comprehensive overview of speech assistants but also validates the feasibility of core technologies through a prototype system, laying a foundation for future academic research and industrial applications.

Funding projects

2024 National Higher Vocational College Information Technology Classroom Teaching Reform Research Project (Online and Offline Deep Integration of Big Data Technology Course Teaching Reform Practice);Hubei Provincial Department of Education Science and Technology Research Program Guidance Project (Research on Zero-Carbon Campus Monitoring and Analysis Platform)

References

[1] Chen Yan, Xiaoyu Ji, Kai Wang, Qinhong Jiang, Zizhi Jin, and Wenyuan Xu. 2022. A Survey on Voice Assistant Security: Attacks and Countermeasures. ACM Comput. Surv. 55, 4, Article 84 (April 2023), 1-36.

[2] Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In

DOI: 10.54254/2755-2721/2025.28125

Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 109, 1–14.

[3] Xiao Zhan, Noura Abdi, William Seymour, and Jose Such. 2024. Healthcare Voice AI Assistants: Factors Influencing Trust and Intention to Use. Proc. ACM Hum.-Comput. Interact. 8, CSCW1, Article 62 (April 2024), 37 pages.

[4] Bradley Rey, Yumiko Sakamoto, Jaisie Sin, and Pourang Irani. 2024. Understanding User Preferences of Voice Assistant Answer Structures for Personal Health Data Queries. In Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI '24). Association for Computing Machinery, New York, NY, USA, Article 17, 1–15.

[5] Luca Lazzaroni, Francesco Bellotti, and Riccardo Berta. 2024. An embedded end-to-end voice assistant. Eng. Appl. Artif. Intell. 136, PB (Oct 2024).

[6] Jingjin Li, Chao Chen, Mostafa Rahimi Azghadi, Hossein Ghodosi, Lei Pan, and Jun Zhang. 2023. Security and privacy problems in voice assistant applications: A survey. Comput. Secur. 134, C (Nov 2023).