Available Online: 14 May 2025 DOI: 10.54254/2977-3903/2025.23018

Response delay in voice-controlled robots: the combined effect of sound propagation speed and circuit processing time

Zhaozhang Huang

Ulink High School of Suzhou Industrial Park, Suzhou, China

changchang819@126.com

Abstract. With the rapid advancement of voice recognition and artificial intelligence technologies, voice-controlled robots are increasingly utilized in human-computer interaction. However, the issue of response delay during the voice control process has become a critical bottleneck affecting user experience and system performance. This paper systematically analyzes the causes and characteristics of response delays in voice-controlled robots from two key dimensions: the speed of sound propagation in air and the internal circuit processing time of the robot. Through an examination of representative cases (such as smart speakers, surgical robots, and industrial control systems), this study reveals the relative contributions of physical and technological factors to total response time across different application scenarios and evaluates their impact on user perception and interaction efficiency. The findings indicate that in short-range scenarios, speech recognition processing is the primary bottleneck, while in long-range and high real-time scenarios, sound propagation delay also becomes a significant factor.

Keywords: human-computer interaction, speech processing latency, edge computing

1. Introduction

With ongoing progress in voice recognition and artificial intelligence technologies, voice control is being increasingly integrated into the field of robotics, allowing humans to interact with machines through natural language commands. However, in voice-controlled robotic systems, the time lag between the issuance of a command and the robot's response has emerged as a key factor impacting both user experience and system performance. Excessive response delays can lead to user frustration or distrust, reduce the naturalness of interaction, and even pose safety risks in critical situations. Thus, it is essential to thoroughly analyze the sources and influencing factors of response delays in voice-controlled robots, in order to provide a theoretical basis for reducing latency and enhancing system responsiveness [1].

The factors influencing the response time of voice-controlled robots are multifaceted. Among them, the two primary components are the delay in sound propagation through the air and the processing delay within the robot's internal circuitry. These two types of delays combine to determine the system's overall response time. The delay caused by sound propagation is rooted in physical factors—since the speed of sound is finite, a sound wave requires a certain amount of time to travel from the user to the robot's microphone when there is a physical distance. On the other hand, circuit processing delay is a technological factor—it arises from the robot's need to parse, recognize, and execute the received voice command using its hardware and algorithmic models, which consumes time. This paper focuses on these two major aspects, exploring their individual characteristics, their contributions to total response delay, and how they can be jointly considered to optimize the response speed of voice-controlled robots [2].

In current research, many scholars and engineers have recognized the problem of latency in voice interaction and have proposed corresponding improvements. For example, studies testing the response times of mainstream voice assistants on the market have revealed significant differences across systems, with delays ranging from less than one second to several seconds. In specialized fields such as surgical robotics, research on voice control has similarly identified recognition accuracy and response speed as major barriers to broader application. Therefore, a systematic review and analysis of the response delay in voice-controlled robots is not only of academic significance but also of practical value.

2. Literature review

In recent years, research on response delays in voice interaction systems—both in China and internationally—has primarily focused on three areas: (1) user experience and perception, (2) analysis and modeling of delay sources, and (3) technological and algorithmic approaches to delay reduction.

2.1. User experience and delay tolerance

Response delay has a significant impact on user experience. According to classic principles in the field of human-computer interaction, if a system responds within 0.1 seconds, users perceive it as instantaneous; if within approximately 1 second, users notice a brief wait but are unlikely to lose their train of thought; however, delays of several seconds or more may cause users to lose focus or feel frustrated. In the domain of service robots, a study by Kang et al. (2024) further confirmed the influence of delay on users' subjective evaluations, and explored strategies for mitigating negative perceptions through robotic feedback. Their experiments showed that when a robot cannot respond immediately, appropriate feedback—such as progress indicators or auditory prompts—can alleviate users' anxiety caused by waiting, thereby enhancing the overall interaction experience. These studies highlight the critical importance of minimizing latency in voice systems and suggest that, when delay is unavoidable, optimization of user perception must also be considered [3].

2.2. Analysis of delay sources

The sources of delay in voice interaction systems span several stages, including audio signal propagation, processing, and network transmission. For voice-controlled robots specifically, the primary factors are the delay caused by sound propagation in the medium (typically air) and internal system processing delay. In standard conditions, the speed of sound in air is approximately 340 meters per second. This means that for every additional meter between the user and the robot's microphone, the propagation time of the voice signal increases by about 3 milliseconds. While this physical delay typically remains within a few dozen milliseconds during short-range indoor interactions—well below the threshold of human perception—in larger spaces or long-distance voice control scenarios (such as operating a remote robot), the propagation delay may reach the scale of several hundred milliseconds, which cannot be ignored [4].

2.3. Technological and algorithmic improvements to reduce delay

To address the issue of speech processing delays, the introduction of high-performance computing and efficient algorithms has been a primary strategy. On the hardware side, there has been a notable shift from reliance on cloud-based servers in early systems to the current emphasis on edge computing and dedicated acceleration chips, allowing speech recognition to be completed locally and in real-time, thereby avoiding the additional delays associated with network transmission. On the software side, advancements in algorithm design have also significantly reduced speech recognition latency.

3. Case studies

This section analyzes several representative application scenarios of voice-controlled robots to examine the structure of response delays across different contexts. For instance, we compare the delay composition between short-range indoor interaction (e.g., smart speaker robots) and long-range outdoor interaction (e.g., voice-remote control of large service robots in open spaces). Additionally, we cite recent studies on surgical and industrial robot voice control to validate the theoretical analysis with empirical data.

3.1. Case 1: voice command response of an indoor smart speaker robot

Assume a user issues the command "play music" from a distance of 2 meters in a living room to a smart speaker-style robot. In this scenario:

Speech Processing Delay: Smart speakers typically adopt a hybrid architecture of local wake-word detection and cloud-based speech recognition. The wake word (e.g., "Hey, Xiaoyi") is detected locally, usually within tens of milliseconds. Subsequently, the voice command is uploaded to the cloud for recognition and parsing. Reports indicate that mainstream voice assistants respond to user queries in approximately 0.6 to 1 second. Of this, around 0.5 seconds is attributed to speech recognition and comprehension, while the remainder is due to network transmission and the initiation of music playback [5].

Other Delays: The command "play music" is relatively straightforward, and the execution phase—starting the music—occurs almost instantly.

Summary: The total response delay is approximately 0.6 to just over 1 second, with sound propagation contributing less than 0.01 seconds and thus being negligible. The primary delay originates from cloud-based processing and network round-trips. This

explains why, in daily use, users generally do not perceive delay variations due to distance as long as they are within the same room. In such scenarios, system processing capability is the main factor affecting response speed. Users are typically tolerant of delays up to about one second without irritation. If the processing delay is further reduced—for example, by adopting edge computing to shorten recognition time to 0.2 seconds—the total delay may drop to around 0.3 seconds, rendering the interaction nearly real-time.

3.2. Case 2: long-range outdoor voice command to a robot

Consider an outdoor security robot that is voice-controlled by a user shouting "stop" from 50 meters away, with no other communication method available:

Speech Processing Delay: If the robot possesses local recognition capabilities, it can immediately process the audio. In cases where it uses a simplified local recognition system (e.g., limited command vocabulary), processing and issuing a stop command can be completed in about 0.1 seconds. If the system has lower performance, processing may take up to 0.5 seconds.

Other Delays: Executing the stop command requires braking, which involves a mechanical delay estimated at 0.1 seconds.

Summary: The total response delay ranges from approximately 0.35 seconds (in ideal conditions) to 0.75 seconds (under slower processing conditions). Here, sound propagation accounts for 20% or more of the total delay. As the distance increases—for example, to 100 meters—sound propagation delay alone can reach 0.3 seconds, becoming the dominant bottleneck. To mitigate delay in such scenarios, two approaches can be considered: Use of Wireless Communication Devices: If the user transmits commands via a radio device (e.g., a walkie-talkie connected to the robot), electromagnetic waves travel at the speed of light, taking only about 0.00017 seconds to cover 50 meters—effectively eliminating sound propagation delay. Directional Microphone Arrays: Equipping the robot with long-range directional microphone arrays can enhance far-field voice capture while minimizing the additional processing overhead caused by ambient noise.

3.3. Case 3: voice control system of a surgical robot

Taking the experiment conducted by Davila et al. (2024) as an example, the surgeon uses voice commands to control the surgical assistant arm. Common commands include fixed phrases such as "zoom camera" and "open forceps." Due to the high demands for precision and reliability in surgical environments, the system runs a high-performance speech recognition model (OpenAI Whisper) on a local server. The measured average response time is approximately 1.7 seconds. The breakdown is as follows [6]:

Sound Propagation Delay: The surgeon wears a headset microphone, with the distance to the receiver being negligible (<0.5 meters), resulting in a delay of only about 1–2 milliseconds.

Speech Processing Delay: The vast majority of the 1.7-second delay is due to speech recognition and command parsing.

Other Delays: The delay for the robotic arm to begin moving is minimal.

This case illustrates that even under conditions prioritizing high recognition accuracy, processing delay can remain substantial (exceeding one second). Although a 1.7-second delay would be considered relatively long in typical conversational contexts, it is acceptable in surgical settings where operations proceed at a slower pace and a delay of one to two seconds is generally tolerable. To improve response speed, stream-based recognition could be employed to reduce the wait time for complete utterances, or custom lightweight models could be developed to accelerate processing. However, in such critical applications, the trade-off between accuracy and speed must be carefully balanced.

The above case studies demonstrate that the relative contribution and significance of sound propagation delay and circuit processing delay vary depending on the application scenario. In close-range situations involving simple commands, processing delay often dominates while propagation delay is nearly negligible. In long-distance or high real-time requirement scenarios, propagation delay becomes increasingly significant. In complex recognition tasks, processing delay remains the primary bottleneck, though advanced technologies offer potential for reduction. Overall, the two delays have a cumulative effect, and which one constitutes the main bottleneck depends on the specific application context. Understanding this interplay is crucial for implementing targeted strategies to optimize system performance.

4. Conclusion and future prospects

This paper systematically analyzes the two main types of delays experienced by voice-controlled robots from the reception of a spoken command to the generation of a response, and discusses their implications for user experience and system design. Through the introduction and literature review, we identified the significance of the delay issue and the current state of research. Combined with the results and case studies, we examined the differences in delay composition across various application scenarios and the performance levels achievable by state-of-the-art systems. In the discussion, we synthesized the following insights: physical and technical factors jointly determine the lower bound of response delay; users' perceptual thresholds provide reference points for optimization; application-specific requirements influence optimization strategies; and advances in algorithms and system design will continue to drive delay reduction.

Looking ahead, as artificial intelligence continues to evolve, the response speed of voice-controlled robots is expected to approach real-time. Several promising directions include: More efficient speech recognition models and chips, potentially leveraging neuromorphic computing hardware to achieve millisecond-level processing of complex voice commands; Smarter interaction strategies, such as predictive processing where the robot anticipates the user's intent before the command is fully spoken, enabling immediate action once the utterance is complete; Multimodal human-machine collaboration, where voice input is complemented by visual cues, gestures, and other signals to infer user intent in advance and reduce reaction time. At the same time, researchers will continue exploring how to enhance user experience in situations where delay is inevitable. For instance, robots may respond in a human-like manner by saying "Please wait a moment" to alleviate user anxiety during pauses.

In conclusion, response delay in voice-controlled robots is a multidisciplinary systems issue, requiring the integration of knowledge from physical acoustics, electronic engineering, computer algorithms, and human-computer interaction. Through continuous technological innovation and design optimization, we have every reason to believe that future voice-controlled robots will offer faster and more natural responses, revolutionizing voice interaction across diverse application scenarios.

References

[1] Jiang, S.-Y., Xie, N.-J., & Yin, D.-N. (2025). Virtual-physical interaction system for mobile robots based on digital twin. *Industrial Instrumentation & Automation Equipment,*(02), 69–75. https://doi.org/10.19950/j.cnki.CN61-1121/TH.2025.02.013

[2] Ma, F., & Yang, L. (2025). "First humanoid robot stock" UBTECH: From Spring Festival Gala "star" to factory "rookie". *Southern Daily,* A05.

[3] Zhang, W., & Wu, C.-X. (2025). Human cognition and simulation modeling and its application in voice interaction design of smart homes. *Packaging Engineering, 46*(04), 226–236. https://doi.org/10.19554/j.cnki.1001-3563.2025.04.019

[4] Wang, J.-X., Wang, H., & Xie, Y.-Y. (2023). Relay transmission optimization algorithm in mobile delay-tolerant sensor networks based on a trust mechanism. *Journal of Sensor Technology, 36*(08), 1281–1289.

[5] Jiang, Z.-W. (2023). Construction of an influencing factor model for perceived privacy risk in human-computer interaction: An empirical study based on smart speaker users. *Journal of Journalism,*(08), 83–96. https://doi.org/10.15897/j.cnki.cn51-1046/g2.20230905.001

[6] Xu, W.-K. (2025). The dilemma and response of liability determination in AI medical robot accidents. *Journal of Panzhihua University,* 1–9. Advance online publication. http://kns.cnki.net/kcms/detail/51.1637.Z.20250407.1033.002.html