IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

# Voice-Controlled Navigation of Short Videos on Social Media Platforms Using Voice

**Swapnil Tawhare[1], Sampada Shinde[2], Prapti Unde[3], Prof. Bharti Patil[4]**
Computer Science Department
Dr. D. Y. Patil Arts, Commerce, Science College, Pimpri Pune

**Abstract:** *Short videos on social media are very popular today, but users mostly control them using touch. This paper presents a voice-controlled system that allows users to navigate short videos using simple voice commands. With this system, users can play, pause, skip, replay, like, or search videos without touching the screen. The main goal is to make video navigation easier and more convenient, especially for people who have difficulty using touch controls or when users are busy doing other tasks. A basic model of the system is created and tested to check how accurately it understands voice commands and how fast it responds.*

*The results show that voice control can improve user experience by making interaction faster and more natural. However, some challenges like background noise, different accents, and privacy issues still need to be solved. In the future, the system can be improved by supporting more languages and increasing accuracy..*

**Keywords:** Voice Control, Short Video Navigation, Speech Recognition, Social Media Platforms, Hands-Free Interaction, Natural Language Processing

## I. INTRODUCTION

In recent years, short-form videos have become one of the most popular types of content on social media platforms. Applications like Instagram Reels, YouTube Shorts, and TikTok allow users to quickly watch, share, and interact with videos in a continuous scrolling format. These platforms are mainly designed for touch-based interaction, where users tap or swipe to control video playback. However, touch-based navigation is not always convenient. Users may find it difficult to interact with videos while multitasking, such as cooking, driving, or exercising. Additionally, people with physical disabilities may face challenges while using touch controls. This creates a need for a more accessible and user-friendly method of interaction.

Voice-controlled technology has emerged as a powerful solution for hands-free operation. By using speech recognition and natural language processing, users can control devices and applications through simple voice commands. Integrating voice control into short video platforms can improve accessibility, increase convenience, and provide a more natural way of interaction. This research focuses on designing and implementing a voice-controlled navigation system for short videos on social media platforms. The study aims to evaluate its effectiveness in terms of accuracy, response time, and user experience, while also identifying challenges such as background noise and speech variations.

## II. OBJECTIVES

The primary objective of this research is to design and develop a voice-controlled navigation system that enables users to interact with short video content on social media platforms using spoken commands. The system aims to replace or support traditional touch-based controls by allowing users to perform actions such as play, pause, skip, replay, like, and search videos through voice input. Another key objective is to enhance accessibility by providing an alternative interaction method for users with physical disabilities or limitations that make touch-based navigation difficult. In

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

addition, the study focuses on improving user convenience by enabling hands-free interaction, which is especially useful in multitasking situations such as driving, cooking, or exercising.

The research also aims to evaluate the performance of the proposed system by analyzing factors such as speech recognition accuracy, response time, and overall user satisfaction. Furthermore, it seeks to identify and address challenges associated with voice-based systems, including background noise interference, variations in speech patterns and accents, and potential privacy concerns. Finally, the study aims to explore future improvements, such as multilingual support and enhanced command recognition, to make the system more efficient, inclusive, and widely applicable across diverse user groups.

## III. SCOPE

This research focuses on the development and implementation of a voice-controlled navigation system for short videos on social media platforms. The system is designed to support basic voice commands such as play, pause, skip, replay, like, and search, enabling users to interact with video content without relying on touch-based inputs. The scope of this study includes the use of speech recognition and natural language processing techniques to understand and process user commands. It also covers the design of a prototype model that demonstrates how voice interaction can be integrated into short video platforms. The performance of the system is evaluated based on parameters such as accuracy, response time, and ease of use.

This study is limited to controlled environments for testing, where factors such as background noise and internet connectivity are considered but not fully explored in real-world conditions. Additionally, the system focuses on a limited set of predefined voice commands and may not support complex or conversational inputs. The research does not cover full-scale deployment in commercial social media applications but provides a foundation for future development. Future scope includes expanding multilingual support, improving recognition accuracy in noisy environments, enhancing security and privacy features, and integrating the system into real-world platforms for broader usability.

## IV. LITERATURE SURVEY

| SR.NO | TITLE | YEAR | AUTHER | SUMMARY |
|---|---|---|---|---|
| 1. | AI Voice in Online Video Platforms: A Multimodal Perspective on Content Creation and Consumption | 2025 | Xiaoke Zhang, Mi Zhou, Gene Moo Lee | This paper explains how AI-based voice technology is used in short video platforms. It shows that voice features make content creation easier and faster for users. However, it may reduce personal connection in videos. Overall, voice technology improves usability and plays an important role in modern social media platforms. |
| 2. | Hands-Free Video Player: Enhancing Accessibility with Voice-Controlled Navigation | 2025 | V. Karthika, A. Siva Ganesh | This paper presents a voice-controlled system for navigating short videos on smart platforms. It uses speech recognition and natural language processing to execute commands like play, pause, skip, and search. The system improves accessibility and provides a hands-free user experience, especially useful for disabled users and multitasking scenarios. It also highlights challenges such as noise interference and speech variation. |
| 3. | Voice-Controlled Autonomous Navigation for Smart Systems Using | 2025 | Walid Benayed, Mohamed | This paper describes a voice-controlled navigation system using deep learning-based speech recognition. It allows users to control systems through spoken |

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

| | | | | |
|---|---|---|---|---|
| | Deep Learning-Based Speech Recognition | | Slim Masmoudi | commands in real time. The study shows high accuracy in understanding voice inputs but notes that performance can be affected by noise and speech differences. |
| 4. | Voice-Controlled Navigation for Intelligent Human–Machine Interfaces Using Automatic Speech Recognition | 2025 | Ioannis Giachos, Vasilios-Stylianos Lefkelis, Evangelos Papakitsos, Petros Savvidis | This paper presents a voice-controlled system for human–machine interaction using Automatic Speech Recognition (ASR). The system allows users to control devices through spoken commands in real time. It focuses on improving accuracy, speed, and usability of voice-based navigation systems. The study shows that integrating ASR with intelligent interfaces makes interaction more natural and efficient, but challenges such as background noise and speech variation still affect performance. |
| 5. | Natural Language Processing in Voice-Controlled Systems | 2023 | Sharma and Gupta | This research discusses the role of natural language processing in understanding voice commands. It explains how NLP improves the accuracy of voice-controlled systems and enables better communication between users and devices. |

## V. METHODOLOGY

The proposed system for voice-controlled navigation of short videos on social media platforms is developed using a combination of speech recognition and natural language processing techniques. The main goal of the system is to allow users to control video playback through voice commands instead of traditional touch-based interactions.The process begins with voice input collection, where the user gives spoken commands such as “play”, “pause”, “skip”, “replay”, or “like”. These voice inputs are captured using a microphone. Before processing, the audio signal is pre-processed to reduce background noise and improve clarity, ensuring better recognition accuracy. After pre-processing, the system converts the speech input into text using an Automatic Speech Recognition (ASR) engine. This step plays an important role in understanding what the user has spoken. The converted text is then passed to the next stage for interpretation.

In the next step, Natural Language Processing (NLP) techniques are used to analyze the text and identify the user’s intent. The system is trained to recognize specific keywords and phrases that correspond to different video control actions. For example, if the system detects the word “pause”, it identifies the command and prepares to stop the video playback. Once the command is identified, the system executes the corresponding action on the video platform. This may include playing, pausing, skipping to the next video, replaying a video, or marking a video as liked. The response is designed to be quick and smooth to improve user experience.

Finally, the system is evaluated based on important performance factors such as accuracy of speech recognition, response time, and overall user satisfaction. Tests are conducted in different conditions to analyze how well the system performs in real-time usage.However, the system also considers challenges such as background noise, variations in accents, and differences in speech patterns, which may affect performance. These challenges are important for improving the system in future enhancements.

## VI. SYSTEM ARCHITECTURE

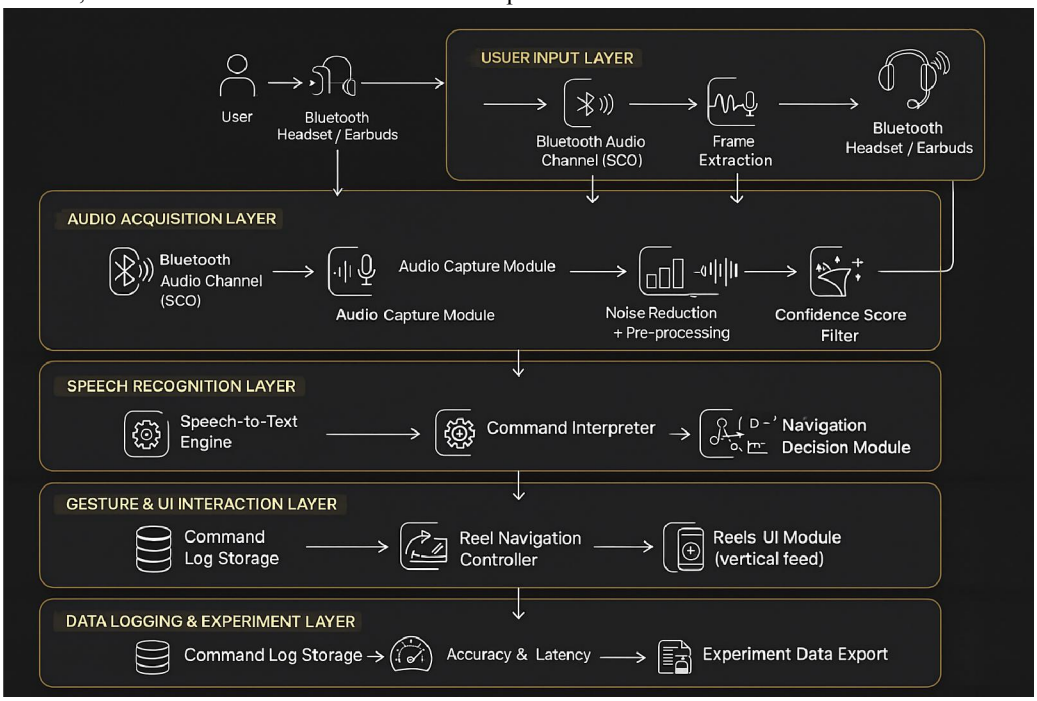
### 6.1 Overview of System Architecture

The system follows a layered architecture that enables users to control short videos using voice commands. It consists of multiple layers, including User Input, Audio Processing, Speech Recognition, UI Interaction, and Data Logging.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

The process starts when the user gives a voice command through a microphone or Bluetooth device. The audio is then captured and pre-processed to remove noise. Next, a Speech-to-Text (ASR) engine converts the voice into text, and NLP identifies the user's intent.

Based on the command, the system performs actions like play, pause, skip, or like on the video interface. Finally, the system logs data such as accuracy and response time for performance analysis.Overall, the architecture ensures real-time, hands-free, and efficient interaction with short video platforms.




### 6.2 Architectural Components

**Hardware Components**

Microphone – Captures user voice commands clearly
Bluetooth Headset / Earbuds – Provides wireless, hands-free voice input
Mobile Device / Computer – Acts as the main processing unit for running the system
Display Screen – Shows short videos and user interface (Reels/Shorts)
Speakers – Outputs audio of the video and system feedback
Network Connectivity (Wi-Fi/Mobile Data) – Supports cloud-based processing

**Software Components**

Operating System – Android, iOS, Windows, or Linux to run the application
Programming Languages – Python, JavaScript for development
Speech Recognition Engine (ASR) – Converts voice input into text
Natural Language Processing (NLP) Module – Understands user intent from text
Audio Processing Library – Handles noise reduction and signal processing
Application Framework – React, Android SDK, or similar for UI development
Video Player Module – Controls playback (play, pause, skip, etc.)
Database / Storage – Stores logs, command history, and performance data

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

### 6.3 Architectural Models

The system architecture shown in the diagram mainly follows a layered and pipeline model. It is divided into multiple layers such as User Input, Audio Acquisition, Speech Recognition, UI Interaction, and Data Logging, where each layer performs a specific function.

The data flows in a sequential pipeline, starting from voice input, then audio processing, followed by speech-to-text conversion, command interpretation, and finally action execution on the video interface.

Additionally, the system works on an event-driven approach, where voice commands act as triggers to perform actions like play, pause, or skip. The inclusion of a data logging layer helps in analyzing performance through accuracy and response time.

### 6.4 Case Studies or Examples

The proposed voice-controlled navigation system can be understood better by comparing it with existing real-world applications that use similar concepts. Popular short video platforms such as YouTube Shorts, Instagram Reels, and TikTok are widely used for consuming video content, but they primarily depend on touch-based interactions like tapping and swiping. The proposed system enhances these platforms by introducing voice-based commands, allowing users to control video playback without physical interaction, thereby improving convenience and accessibility.

A similar concept is already implemented in voice assistants such as Google Assistant, Alexa, and Siri. These systems allow users to control media playback using simple voice commands like “play,” “pause,” or “next.” This demonstrates that voice-based interaction is both practical and effective. The proposed system applies this idea specifically to short video navigation, making the interaction more seamless and user-friendly.

In addition, smart TVs and OTT platforms such as YouTube TV and Netflix support voice-enabled controls through remote devices. Users can search for content and control playback using speech, which highlights the growing adoption of voice technology in entertainment systems. Similarly, modern automotive infotainment systems allow drivers to control media using voice commands, ensuring safety and hands-free operation while driving.

These real-world examples clearly show that voice-controlled systems are already being successfully used in different domains. The proposed system builds upon these existing technologies and adapts them to short video platforms, making user interaction faster, easier, and more accessible.

### 6.5 Future Trends

The future of voice-controlled short video navigation systems is expected to advance significantly with the growth of artificial intelligence and machine learning technologies. One major trend is the development of more accurate speech recognition systems that can better understand different accents, languages, and speaking styles, even in noisy environments. This will make voice interaction more reliable and user-friendly.

Another important trend is the introduction of multilingual support, allowing users from different regions to interact with the system in their preferred language. This will make the system more inclusive and globally accessible. Additionally, improvements in natural language processing (NLP) will enable the system to understand more complex and conversational commands instead of only predefined keywords.

Future systems are also likely to integrate with advanced devices such as smart glasses, AR/VR platforms, and wearable technology, enabling a more immersive and hands-free experience. Personalized voice recognition is another emerging trend, where the system can identify individual users and adapt to their preferences, providing a more customized experience.

Furthermore, there will be a stronger focus on privacy and security, ensuring that user voice data is protected through secure processing and storage methods. Cloud computing and edge computing will also play a key role in improving processing speed and reducing latency.

Overall, these advancements will make voice-controlled systems more intelligent, efficient, secure, and widely adopted across various applications, including social media platforms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

## VII. FINDINGS

The study of the voice-controlled short video navigation system shows that voice-based interaction provides a more convenient and hands-free way to control video content compared to traditional touch-based methods. The system was able to successfully recognize basic commands such as play, pause, skip, and replay with a good level of accuracy under normal conditions. This demonstrates that integrating speech recognition and natural language processing can significantly improve user experience. The findings also indicate that the system performs efficiently in terms of response time, providing quick execution of commands, which is essential for real-time applications. Users found the system useful, especially in multitasking situations such as cooking or exercising, where touch interaction is difficult.

However, some limitations were observed during testing. The system's accuracy decreases in environments with high background noise, and variations in accents or speech patterns can affect command recognition. Additionally, the system currently supports a limited set of predefined commands, which restricts flexibility. Overall, the findings confirm that voice-controlled navigation is a practical and effective solution for improving accessibility and convenience in short video platforms, while also highlighting areas for improvement such as noise handling, accuracy, and command expansion.

## VIII. DISCUSSION

The implementation of a voice-controlled navigation system for short videos demonstrates a significant shift from traditional touch-based interaction to a more natural and hands-free approach. The system effectively integrates speech recognition and natural language processing to allow users to control video playback using simple voice commands. This improves usability, especially in situations where touch interaction is not convenient. The discussion highlights that the system performs well in controlled environments, offering quick response times and accurate command recognition for basic operations such as play, pause, and skip. It enhances accessibility for users with physical limitations and supports multitasking scenarios, making it a practical solution for modern applications.

However, the system also faces certain challenges. Background noise can interfere with speech recognition accuracy, and variations in accents and speech patterns may lead to misinterpretation of commands. Additionally, the current system is limited to predefined commands and does not fully support complex or conversational inputs. These limitations suggest the need for further improvements in noise filtering, advanced NLP techniques, and adaptive learning models. Overall, the discussion indicates that while the system is effective and promising, continuous enhancements are required to make it more robust, accurate, and adaptable for real-world usage.

## IX. CONCLUSION

In conclusion, the proposed voice-controlled navigation system for short videos provides an effective and innovative solution to overcome the limitations of traditional touch-based interaction. By integrating speech recognition and natural language processing, the system enables users to control video playback using simple voice commands, making interaction more natural, efficient, and hands-free. The system successfully demonstrates improved accessibility and convenience, especially for users who are multitasking or have physical limitations. It also shows promising performance in terms of response time and command recognition under normal conditions.

However, certain challenges such as background noise, variations in speech patterns, and limited command support need to be addressed for real-world implementation. Future improvements focusing on enhanced accuracy, multilingual support, and better noise handling can further strengthen the system. Overall, the study confirms that voice-controlled navigation has strong potential to enhance user experience in short video platforms and represents an important step toward more intelligent and user-friendly human-computer interaction systems.

## X. REFERENCES

[1]. H. Ahlawat, "Automatic Speech Recognition: A Survey of Deep Learning Approaches," *ScienceDirect*, 2025.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, May 2026
Impact Factor: 8.2

[2]. Md. Nayeem et al., "Automatic Speech Recognition in the Modern Era: Architectures, Training, and Evaluation," *arXiv*, 2025.
[3]. Sharma et al., "Enhancing Voice Assistant Systems through Advanced AI and NLP Techniques," 2026.
[4]. H. Ji, "Application of Intelligent Speech Recognition Technology," *ScienceDirect*, 2025.
[5]. Dinh et al., "Benchmarking Speech-to-Text and NLP Pipeline Systems," *MDPI Computers*, 2025.
[6]. Gupta et al., "Speech Recognition-Based Wireless Control System," *MDPI*, 2025.
[7]. Nguyen et al., "AI-Powered Voice Recognition for Learning Systems," *CALL-EJ Journal*, 2025.
[8]. "Advancements in Speech Recognition Using Transformer Models," *Research Survey*, 2026.
[9]. "AI-Enhanced Speech and Voice Recognition Tools," *ResearchGate*, 2025.
[10]. "Applications of Speech Recognition in 2025," *OpenCV Blog*, 2025.
[11]. Radford, A., et al., "Robust Speech Recognition via Large-Scale Weak Supervision (Whisper)," *OpenAI*, 2024.
[12]. Zhang, Y., Chen, G., "Transformer-Based Speech Recognition: Recent Advances and Challenges," *IEEE Access*, 2024.
[13]. Baevski, A., et al., "wav2vec 2.0: Self-Supervised Learning of Speech Representations," *IEEE Transactions on Audio, Speech, and Language Processing*, 2024 (updated studies).
[14]. Pratap, V., et al., "Massively Multilingual Speech Recognition," *Meta AI Research*, 2024.
[15]. Google Research, "Advances in End-to-End Speech Recognition Systems," 2025.