International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

Implementation of a Hybrid Voice and Gesture Recognition System for Intelligent Robotic Vehicles

Divyansh Singh[1], Adityaraj Deokar[2], Shubham Dhavane[3], Tanuja Abhang[4], Ankush Kadu[5]

[1,2,3,4] UG Student, Department of Electronics and Telecommunication Engineering, Dr. D. Y. Patil School of Engineering, Pune, Maharashtra, India

[5]Professor, Department of Electronics and Telecommunication Engineering, Dr. D. Y. Patil School of Engineering and Technology, Pune, Maharashtra, India

---

Abstract - *Recent advancements in embedded systems and human–machine interaction technologies have increased the demand for intuitive robotic control systems. This paper presents the development of a smart robotic vehicle operated through both speech commands and hand gesture recognition. The proposed system combines an ESP32-based robotic platform with a Raspberry Pi-enabled control interface to achieve wireless real-time navigation. Voice commands are interpreted using a predefined speech recognition mechanism for directional control, while hand movements are detected through an MPU6050 accelerometer and gyroscope sensor for continuous gesture-based operation. Wi-Fi communication is utilized for reliable data transfer between the controller and robotic vehicle. The implemented dual-mode architecture enhances operational flexibility and allows users to switch between speech and gesture control according to environmental conditions. Experimental testing indicates that the gesture recognition module provides stable and accurate performance, whereas speech recognition efficiency varies depending on surrounding noise levels and voice clarity. The developed system offers a cost-effective and user-friendly solution suitable for robotic assistance, surveillance applications, educational projects, and smart automation systems.*

Key Words: **Speech Recognition, Gesture Recognition, Human–Machine Interaction, ESP32, MPU6050, Robotic Vehicle, Embedded Systems, Multimodal Control.**

1. INTRODUCTION

Robotic systems have become an important part of modern technology in applications such as industrial automation, surveillance, healthcare assistance, defense systems, disaster management, and intelligent transportation. As robotic platforms continue to evolve, the demand for efficient and intuitive Human-Machine Interaction (HMI) techniques has increased significantly. Conventional robotic control systems based on joysticks, switches, and remote controllers require continuous manual operation and often provide limited flexibility in dynamic environments. To overcome these limitations, researchers have increasingly focused on natural interaction methods such as speech recognition and gesture recognition for robotic control applications [10], [14].

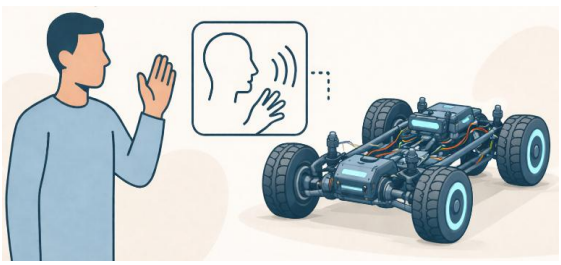

Fig -1: Multimodal framework for robotic vehicle

Speech recognition enables hands-free communication between humans and robotic systems through verbal commands. Several studies have demonstrated the feasibility of speech-based robotic control using embedded platforms and wireless communication technologies [3], [8], [13], [16]. Although voice-controlled systems provide ease of operation and accessibility, their performance is often affected by environmental noise, speaker variability, and limited offline recognition capability [20]. These challenges reduce system reliability in real-world operating conditions.

Gesture recognition has emerged as another effective interaction technique for robotic navigation and control. Vision-based and sensor-based gesture systems have been widely explored for interpreting human hand movements and translating them into robotic commands [1], [7], [12], [15], [17]. Recent developments involving accelerometer, IMU, and depth-camera technologies have improved gesture recognition accuracy and response time [4], [5], [9]. However, vision-based approaches may suffer from lighting variations, computational complexity, and hardware cost limitations.

Recent research indicates that combining speech and gesture recognition within a multimodal framework improves operational reliability, flexibility, and user interaction quality [2], [6], [11], [18], [19]. Hybrid interaction architecture enables complementary operation between both input modes, allowing the robotic system to maintain functionality even when one control method becomes unreliable due to environmental conditions.

The proposed work presents the implementation of a hybrid voice and gesture recognition system for intelligent robotic vehicles using low-cost embedded hardware components.

The developed framework integrates an ESP32 microcontroller, MPU6050 inertial sensor, offline voice recognition module, wireless communication interface, and motor driver circuitry to achieve real-time robotic navigation. Gesture commands are detected using accelerometer and gyroscope sensing, while speech commands are processed using predefined offline voice patterns. Wireless transmission enables efficient communication between the control unit and robotic vehicle. The proposed system aims to provide a compact, cost-effective, and scalable solution suitable for surveillance systems, assistive robotics, educational platforms, industrial automation, and intelligent mobility applications.

2. LITERATURE SURVEY

Human–Machine Interaction (HMI) has gained increasing importance in robotic systems where intuitive and reliable control is essential. Traditional control mechanisms based on physical interfaces offer limited flexibility and fail to support natural communication between humans and robots. To overcome these limitations, recent research has focused on multimodal interaction techniques that integrate speech and gesture inputs to enhance usability and control efficiency [10], [14].

Speech-based robotic control has been widely explored due to its natural and hands-free communication capability. Early and recent studies have demonstrated the feasibility of controlling robotic platforms using embedded speech recognition modules and microcontroller-based systems [3], [8], [13]. However, speech-only systems are highly sensitive to background noise, speaker variability, and limited offline vocabulary support, which restrict their performance in real-world environments [6], [20]. These challenges highlight the need for complementary interaction modalities.

Gesture recognition has been proposed as an effective alternative to voice commands, particularly in scenarios where verbal communication is unreliable. Vision-based gesture recognition approaches using depth cameras or pose estimation techniques have achieved high recognition accuracy under controlled conditions [1], [7]. Nevertheless, such systems often suffer from performance degradation due to lighting variations, occlusions, and high computational requirements. To address these issues, sensor-based gesture recognition using accelerometers and inertial measurement units has been widely adopted for robotic control applications [9], [12], [17].

Table -1: Comparative Analysis of Existing Systems

Author(s)	Year	Findings
Salinas-Martínez et al.	2024	Fusion improved accuracy and reduced errors.
Baksh et al.	2024	Increased usability and learning outcomes.
Garcia et al.	2024	Contextual cues improved interaction.
Rautiainen et al.	2022	Boosted efficiency in noisy settings.
Feng	2021	Achieved real-time control with limitations.
Chmurski et al.	2021	Good accuracy with low compute load.
Saad et al.	2018	Enabled low-cost voice-based control.
Yang et al.	2018	Allowed intuitive gesture interaction.

Recent studies emphasize that combining speech and gesture recognition within a single framework significantly improves interaction robustness and command accuracy. Multimodal systems exploit complementary input channels to reduce ambiguity and enhance reliability during human–robot interaction [2], [6]. Experimental evaluations reported in the literature indicate that fusion-based approaches outperform single-modality systems, particularly in noisy or dynamic environments [11], [18].

With the growing demand for portable and cost-effective robotic solutions, researchers have increasingly focused on implementing multimodal interaction systems on low-cost embedded platforms. Several works have demonstrated real-time robotic control using Arduino, Raspberry Pi, and ESP32 microcontrollers without reliance on cloud-based processing [4], [12], [16]. These systems enable offline operation, reduced latency, and improved reliability, making them suitable for practical deployment in assistive robotics, education, and automation applications.

Several robotic control systems have also explored wireless communication between sensing and actuation units. The transceiver is frequently used for data exchange between the hand unit and robotic unit due to its long-range, low-power, and high-reliability features. In this context, the ESP32 microcontroller acts as a gesture and voice data processor in the hand unit, transmitting control signals wirelessly to the Raspberry Pi 4, this functions as the main controller on the robot side. The TB6612FNG motor driver then interprets these signals to operate the 4WD robotic chassis, ensuring smooth movement.

2.1 Gap Analysis and Observations

The gap analysis of existing research on speech and gesture-based human–machine interaction for robotic vehicles reveals several critical limitations. Most studies focus on either speech or gesture control individually, which restricts the system's reliability in real-world environments. Speech-only systems often face issues with background noise and limited vocabulary, while gesture-only systems are highly sensitive to lighting conditions or require costly hardware such as vision or radar modules.

Table -2: Gap Identified and Analysis

Author(s)	Gap Identified	Analysis
Salinas-Martínez et al.	System limited to industrial applications.	Adapting to Raspberry Pi would make it low-cost and portable.
Baksh et al.	Tested in controlled environment only.	Requires validation in noisy or dynamic conditions.
Yang et al.	Gesture-only system.	Adding speech input would increase reliability.
Rautiainen et al.	High sensor cost.	Cost-effective sensors needed for education and embedded systems.
Garcia et al.	Emotion-based focus, not real-time control.	Real-time robotic vehicle control should be integrated.
Saad et al.	Small command set and noise sensitivity.	Offline ASR needed for robustness.
Feng	Vision-only system affected by lighting.	Adding IMU improves low-light performance.
Chmurski et al.	Not ported to Raspberry Pi.	Optimization for embedded platforms required.
Ismail et al.	Domain-specific healthcare system.	Expand for general robotic applications.
Qu et al.	Requires cloud connectivity.	Offline alternatives increase reliability.

Some multimodal approaches exist but are designed for specific industrial or academic applications and rely on expensive sensors or cloud-based processing, making them unsuitable for low-cost embedded platforms. These gaps highlight the need for a lightweight, offline, and multimodal framework that integrates both speech and gesture recognition using affordable components like ESP32, MPU6050, and Raspberry Pi. Such a system can improve accuracy, reduce cost, and enhance usability across a wide rangeofpracticalapplications.

The pie chart highlights the usage of different components across 20 journals. It shows that wireless modules and motor drivers are the most widely used, forming the core of robotic vehicle systems.

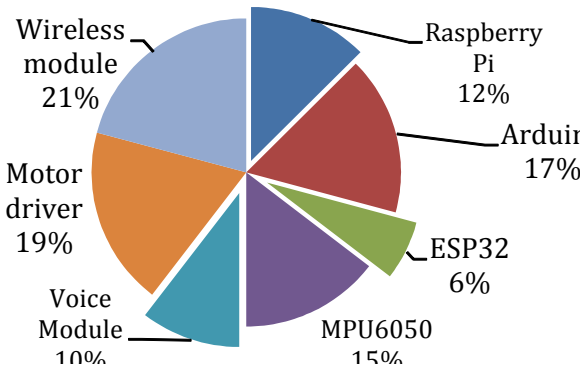

Chart -1: Component usage across 20 journals

Arduino is used more often in simpler setups, while Raspberry Pi appears in projects needing higher processing power. Gesture recognition with MPU6050 is common, but voice modules are less frequently used. Overall, the chart suggests that combining gesture and speech control is still less explored, making our project unique and relevant.

3. METHODOLOGY

The proposed methodology presents a multimodal human–machine interaction framework that combines speech and gesture inputs for real-time robotic vehicle control. At system startup, all peripheral modules including the microcontroller, inertial sensor, voice recognition unit, and wireless transceiver are initialized to ensure synchronized operation. Voice commands are acquired through the embedded recognition module, while hand motion data is captured using the IMU sensor. The controller processes both input streams and maps them to predefined motion instructions using rule-based decision logic.

The encoded control data are transmitted wirelessly to the robotic unit through a low-power RF link. Upon reception, the embedded controller decodes the command and generates appropriate drive signals for the motor driver circuitry. The motor driver regulates the speed and direction of the DC motors to achieve the intended navigation behavior.

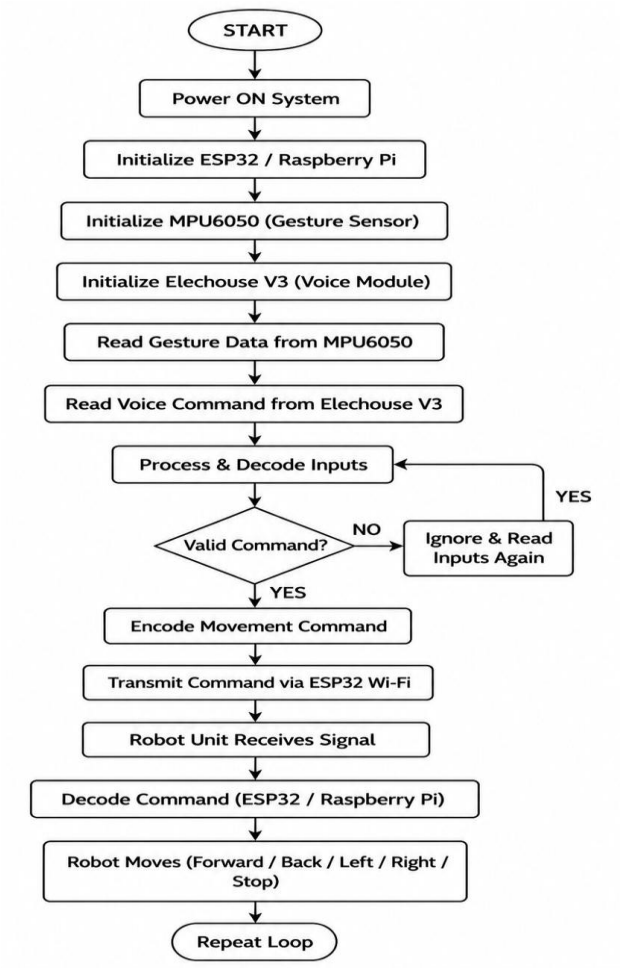

Fig -2: Flowchart of the proposed system

In the methodology flowchart shows how project integrates speech and gesture controls for a robotic vehicle using an ESP32 microcontroller. Voice commands from a microphone and gestures from an IMU/accelerometer are processed to generate control signals for the motor driver. Prototyping is done on a breadboard, followed by PCB integration, ensuring accurate and low-latency responses for smooth vehicle navigation.

4. Hardware Requirements

4.1 ESP32: Central microcontroller for processing inputs from sensors and controlling the robotic vehicle.

4.2 MPU6050: IMU sensors are used to detect gestures and motion for gesture-based control.

4.3 ElecHouse V3: Captures voice commands for speech-based interaction with the vehicle.

4.4 Raspberry Pi 4: Acts as a secondary processing unit.

4.5TB6612FNG Motor Driver: Controls the DC motors on the robotic chassis based on signals from the ESP32.

4.6 DC-DC Buck Converter: Steps down the higher battery voltage (e.g., 12V) to a stable 5V output to safely power the ESP32.

4.7 2WD Robotic Chassis: Provides locomotion platform for the robot, including wheels and mounting structure for components.

4.8Ultrasonic Sensor: Detects distance and obstacles using ultrasonic waves.

4.9Logic Level Converter: Converts 3.3V and 5V signals for safe communication between modules.

5. SOFTWARE REQUIREMENTS

5.1 Arduino IDE: Used for programming and uploading code to the ESP32.

5.2 Embedded C/C++: Used for developing the embedded control system.

5.3 Raspberry Pi OS: Used for Raspberry Pi operations.

6. SYSTEM DESIGN

The system is divided into a transmitter and a receiver, facilitating a wireless control loop for a robotic vehicle.

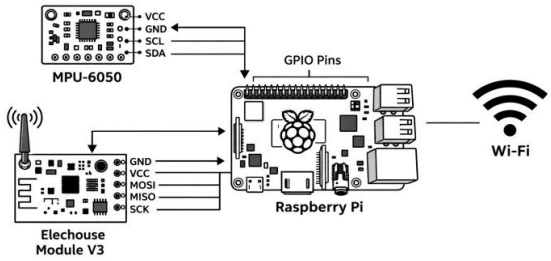

Fig -3: Transmitter of the proposed framework

At the Transmitter, a Raspberry Pi acts as the central processor, aggregating movement data from an MPU6050 IMU/accelerometer and vocal inputs from a voice recognition module before broadcasting commands via a wireless link.

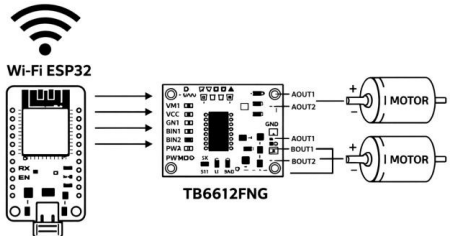

Fig -4: Receiver of the proposed framework

The Receiver features an ESP32 microcontroller that captures these signals and translates them into logic for the TB6612FNGmotordriver, which regulates power to two motors for precise navigation.

7. RESULTS

The developed prototype of the proposed multimodal robotic vehicle system is shown in the figures below. The complete hardware setup consists of two main units: the Hand Unit (Transmitter) and the Robot Unit (Receiver). The system was assembled using low-cost embedded components and tested under laboratory conditions to validate real-time operation and communication reliability.

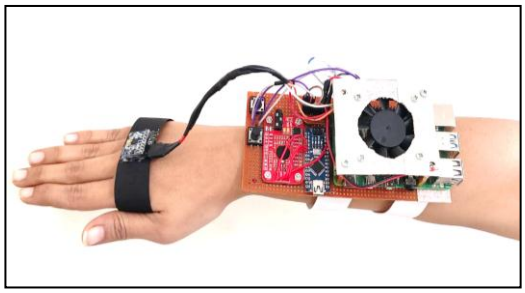

Fig -5: Hand Unit (Transmitter)

Fig.5. Illustrates the hardware implementation of the transmitter unit. The Hand Unit integrates the MPU6050 inertial sensor for gesture input and the Elechouse V3 voice recognition module for speech input. Both input signals are processed by the ESP32 microcontroller and transmitted wirelessly to the Robot Unit.

During practical implementation, the system successfully responded to predefined motion commands such as Forward, Backward, Left, Right, and Stop. Gesture-based control demonstrated smooth directional transitions with minimal delay, while voice-based control enabled hands-free operation.

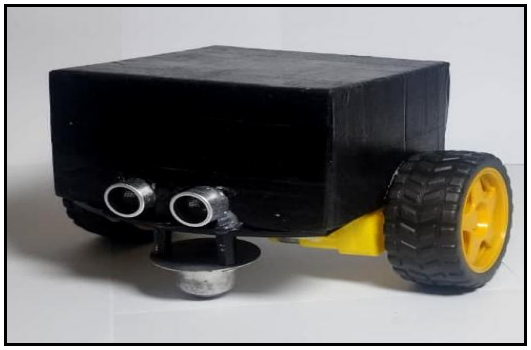

Fig -6: Robot Unit (Receiver)

Fig.6. Illustrates the hardware implementation of the receiver unit. The Robot Unit receives the encoded commands and decodes them using the controller, which generates appropriate control signals for the TB6612FNG motor driver. The motor driver then regulates the speed and direction of the DC motors mounted on the 2WD robotic chassis.

The physical prototype validates the feasibility of integrating speech and gesture recognition into a single embedded framework. The modular hardware design ensures easy expansion for additional sensors or control features in future development.

8. CONCLUSION

The proposed multimodal human–machine interaction system successfully integrates gesture and speech recognition for real-time robotic vehicle control. Experimental results demonstrate that the gesture recognition module achieved an average accuracy of approximately 95%, while the speech recognition module achieved around 70% accuracy under controlled conditions. These results confirm that gesture-based control provides higher reliability, whereas voice commands offer convenient hands-free interaction.

Future work will focus on enhancing the functionality and reliability of the proposed system. Collision avoidance sensors such as ultrasonic or infrared sensors can be integrated into the robotic vehicle to detect obstacles and prevent accidents during navigation. In addition, improving the wireless communication range will allow the robot to operate effectively over larger distances and in more complex environments. Further improvements may also include optimizing the speech recognition module for better performance.

9. REFERENCES

[1] Jianan Xie, Zhen Xu, “Depth-MediaPipe & Semantic-Pose System for Gesture Control of Quadruped Robot,” 2025.
[2] Paul, Nicolescu, “Combined Vision Gestures and NLP-Based Speech Processing,” 2024.
[3] Adekunle T. Oyelami et al., “Arduino 4-DOF Robotic Arm with HM2007 Speech Module,” 2023.
[4] Dhanashree Wadaye et al., “ESP32-CAM, MPU6050 and Gesture-Controlled Surveillance Robot,” 2023.
[5] Ing-Jr Ding, Ya-Cheng Juang, “sEMG + IMU Gesture System with ROS & SLAM,” 2023.
[6] Ryumin, Ivanko & Ryumina, “Multimodal Recognition Using Mobile Sensors and ML,” 2023.
[7] Moysiadis et al., “Hand Gesture Recognition in ROS with Depth Camera,” 2022.
[8] Surjeet, Nishu Gupta, “MSP432 Microcontroller with WiFi-Based Speech and Video,” 2021.

[9] E. E. Atimati et al., "Bluetooth Glove with Accelerometer-Based Gesture Input," 2021.
[10] Laura-Bianca Bilius, Radu-Daniel Vatavu, "Survey and User Study on Input Preferences," 2020.
[11] Zunaid Kazi et al., "Speech and Gesture-Based Remote Teleoperation Interface," 1998.
[12] Amudhan Rajarajan, Sakthivel Murugesan, "Arduino-Based Robot with Gesture Control via Accelerometer," 2018.
[13] Surjeet, Nishu Gupta, "Voice-Based Robot Control Using MSP432 and WiFi," 2021.
[14] Hongyi Liu, Lihui Wang, "Review on Gesture Recognition for Human-Robot Collaboration," 2017.
[15] Hasan U. Zaman et al., "Hand Gesture Robot Control Using Ultrasonic Sensors," 2017.
[16] Ahmet Vatansever, Hilmi Kuscu, "Voice-Controlled Robotic Vehicle with Direction Mapping," 2019.
[17] Juber Salgar et al., "Gesture Control Robot Using Accelerometer and Arduino," 2020.
[18] T. Thivagar, A. Sriram, "Voice + Gesture Smart Vehicle with Arduino," 2020.
[19] Lijuan Liu et al., "Interactive System Using Gestures for Water Play Robot," 2019.
[20] Jie He et al., "Attention-Based Command Detection for Voice Systems," 2019.