International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Voice-Driven Query Recognition and Response System Using Mini Humanoid Robot

S. Mohan Roa[1], Nandineni Jagadeesh[2], L. Nandini[3], Muppur Gnaneswar[4], Manchireddy Thanmai[5], Machavaram Tejeswar Reddy[6]

[1,2,3,4,5,6]Department of ECE, Sri Venkateswara College of Engineering (Autonomous), Tirupati, AP, India

Abstract

Voice-based human–robot interaction has emerged as a significant research domain for enabling natural, intuitive, and real-time communication between humans and intelligent robotic systems. Conventional humanoid robots have predominantly relied on predefined commands, keyword-based matching, or external application interfaces, which is limit conversational flexibility and reduce the interaction intelligence. This paper presents the design, implementation, and evaluation of a voice- driven query to recognition and response system using the RoboMonk Mini humanoid robot.

The proposed system employs an onboard microphone to capture real-time user speech, which is the processed using speech recognition and the artificial intelligence-based language understanding the techniques. Recognized speech is converted into the textual queries and interpreted using an AI language model capable of generating context-aware and meaningful responses. The generated response is delivered through the text-to-speech module, while synchronized humanoid gestures are the executed using intelligent bus servo motors to enhance interaction of naturalness.

The system operates on an embedded Raspberry Pi processing platform, enabling autonomous and the real-time conversational interaction without the dependency on external control interfaces. Experimental results to demonstrate reliable of speech recognition performance, accurate AI-based response generation, smooth gesture synchronization, and improved user engagement. The proposed architecture is the modular and scalable, that supporting future enhancements such as multilingual interaction, emotion-based response adaptation, and autonomous mobility integration.

Keywords: Humanoid robot, speech recognition, conversational AI, text-to-speech, human–robot interaction, embedded systems, and voice-driven control.

INTRODUCTION

Human–robot interaction (HRI) has evolved the significantly with the increasing deployment of humanoid robots in educational, research, and service environments. The primary objective of an modern HRI systems is to be enable natural and intuitive communication between the humans and robotics platforms. Traditional humanoid robots typically operate using the predefined command sets, button- based interfaces, or mobile application control. Although such as systems simplify implementation, they limit interaction depth and fail to support natural conversation.

With advancements in speech has the recognition and artificial intelligence, voice-based interaction has become a promising alternative for enhancing the conversational intelligence in robotic systems. the Speech-driven communication enables hands-free operation and allows the users to interact with robots

in a manner similar to the human conversation. Integrating AI language to models further to enhances the robot's ability to interpret a open-ended questions, understand contextual meaning, and generate meaningful responses.

The Robomonk Mini humanoid robot provides a compact and programmable platform equipped with the intelligent bus servos and embedded processing capability. By the integrating speech recognition, AI-based language has the processing, text-to-speech synthesis, and synchronized gesture control, the robot can deliver expressive and the interactive conversational responses.

This paper has presented a voice-driven humanoid of a system designed to a enable real-time query recognition and response of generation. The proposed architecture has the integrates audio sensing, natural language understanding, the adaptive decision-making, and the humanoid motion control into a unified embedded framework. The system will operate the autonomously without any reliance on the predefined static responses, thereby enhancing interaction quality and user engagement.

The main contributions of this work are threefold. First, a real-time speech recognition framework is implemented on an embedded platform to capture and interpret user queries. Second, an AI-driven language processing module is a integrated to generate context-aware responses. Third, synchronized humanoid gesture execution is developed to the improve expressive communication and interaction naturalness.

RELATED WORK

Speech-based human–robot interaction has been extensively studied in the robotics and artificial intelligence research. In the Early a voice-controlled the robotic systems relied that primarily on keyword detection and the rule-based command execution. These systems were limited to be recognizing predefined phrases and that could not handle open-ended or conversational queries.

With the development of the natural language processing techniques, conversational AI systems began the incorporating machine learning algorithms to be improve contextual understanding. Statistical language models and neural of the network-based architectures significantly enhanced speech has recognition accuracy and response to the generation capabilities. However, it may be conversational AI implementations rely on cloud-based processing, introducing latency and reducing the system autonomy.

Text-to-speech technologies have also evolved from basic waveform concatenation methods to neural speech has synthesis techniques capable of generating natural-sounding voices. When integrated with the humanoid robots, speech synthesis improves the communication realism.

Recent research has explored the combining conversational AI with humanoid motion control to create expressive robotic systems. Synchronising gestures with speech output enhances user engagement and it improves perceived intelligence. Despite these advancements, challenges remain in the achieving real-time performance on embedded platforms while in maintaining accurate language understanding and smooth motion execution.

The proposed system builds upon these developments by implementing a fully embedded conversational humanoid robot capable of speech recognition, AI-based response generation, and the synchronised gesture execution without dependence on external computation infrastructure.

EXISTING SYSTEM AND LIMITATIONS

Before existing, conversational humanoid robots were developed, mostly robotic systems were designed to the operate through predefined voice on commands, rule-based responses, or external control

mechanisms. In the many conventional systems, robot interaction is initiated using the fixed keyword triggers, on mobile applications, or programmed response datasets stored within the control unit. Although these are the approaches simplify implementation and it reduce the system complexity, they significantly restrict the conversational flexibility and prevent the robot from engaging in dynamic, real-time dialogue with the users.

Several existing robotic platforms are employing basic speech recognition modules that rely on keyword matching rather than natural language understanding. While such systems has provide limited voice interaction capability, they are unable to interpret open-ended questions, contextual variations, or sentence-level semantics. As a result, the interaction remains the constrained to predefined phrases, making the communication mechanical and the repetitive. These are the systems fail to adapt responses based on the user intent, conversational context, or topic variation.

A. Limitations of Conventional Conversational Robots

Conventional conversational robotic systems suffer from several limitations that reduce their effectiveness in the naturally human–robot interaction scenarios. These include the dependency on predefined commands, inability to process open-ended or contextual questions, lack of semantic to understanding, and restricted conversational intelligence. Without AI-based language that processing the capability, robots cannot be interpret user intent beyond the basic keyword detection. that Similarly, without synchronized the motion control, verbal responses lack of expressive humanoid gestures, reducing the perceived intelligence and engagement quality.

Furthermore, static response has architectures limit adaptability and prevent the robot from learning or else generating the context-aware answers. As a result, user can experience becomes repetitive and less interactive, making such as systems unsuitable for advanced educational of demonstrations or AI research applications.

B. Computational and Deployment Challenges

Many modern conversational AI systems rely on the cloud-based processing to achieve a advanced natural language understanding. While the cloud-based models provide the high computational capability, they are introducing latency due to the network of communication and reduce the system autonomy. Dependency on the external servers also raises in the reliability and the privacy concerns.

Embedded humanoid platforms that wanted lightweight speech recognition and language processing frameworks that capable of operating in a original time under hardware constraints. Achieving low-latency of response while maintains accurate speech that interpretation and synchronized motion that execution remains a significant challenge. Environmental factors such as background noise, accent variations, and speech clarity further complicate reliable of deployment.

C. Summary of Existing System Drawbacks

In the summary, existing conversational humanoid systems face the challenges including dependency on a predefined command structure, limited to contextual understanding, the lack of expressive synchronization, reduced the autonomy, and reliance on external computation infrastructure. These limitations restrict the interaction intelligence and conversational depth.

The identified drawbacks motivate the development of an embedded AI-driven humanoid robot capable of real-time voice-based query recognition, contextual response generation, and synchronized the gesture execution, as proposed in this work.

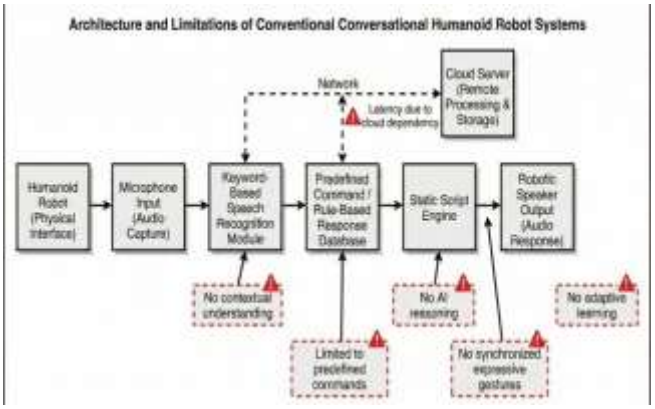


FIG 1. BLOCK DIAGRAM OF EXISTING SYSTEM AND LIMITATIONS VOICE_BASED HUMANOID ROBOT ARCHITECTURE

PROPOSED SYSTEM ARCHITECTURE

The proposed Voice-Driven Query as Recognition and Response System Using Robomonk Mini Humanoid Robot is the developed to enable natural, hands-free, and the intelligent human–robot interaction through speech-based communication. The system is designed to overcome the limitations of conventional rule-based conversational robots by eliminating dependency on predefined command structures and static response datasets. Instead, the robot autonomously captures spoken input, interprets user intent using the artificial intelligence, and generates the context- aware verbal and gestural responses in real time. The proposed architecture integrates multiple functional layers to ensure efficient audio sensing, speech recognition, language understanding, adaptive decision-making, and synchronized humanoid motion execution. This system operates as a closed-loop conversational framework, where the continuous speech inputs are processed by the embedded intelligence modules to produce coordinated verbal and physical responses.

A. Audio Input and Sensing Layer

The input layer has the consists of an onboard microphone module that continuously captures real-time audio signals from the user. This voice-based is the sensing mechanism serves as the primary interaction interface, enabling natural communication without requiring physical buttons, remote of controllers, or mobile applications. Continuous the audio acquisition ensures low-latency speech detection and the real-time responsiveness during the conversational interaction.

B. Speech Recognition and Preprocessing Layer Captured audio signals are forwarded to the speech recognition module, where the preprocessing operations such as noise filtering, signals normalization, and voice activity detection are performed. These are preprocessing steps enhance recognition reliability under the varying indoor acoustic conditions. The processed has audio is then converted into the textual format using speech-to-text algorithms, transforming spoken of queries into a machine- interpretable data suitable for language has analysis.

C. AI Language Processing and Intent Interpretation Module

The recognized textual query is transmitted to the AI language processing module, which functions as the core intelligence unit of the system. This module analyzes sentence structure, contextual semantics, and user intent using artificial intelligence-based language modeling techniques. Unlike conventional keyword-based systems, the AI processing the framework generates meaningful and context-aware responses capable of handling open-ended conversational queries.

The language model can dynamically interpret user input and formulates appropriate responses based on

the contextual as reasoning rather than the static rule mapping. This capability as significantly enhances to the conversational flexibility and interaction intelligence.

D. Text-to-Speech Synthesis Layer

Once the AI module generates the responses text, the output is forwarded to the text-to-speech (TTS) synthesis engine. The TTS module of converts the generated textual response into a natural-sounding audio signals. Speech synthesis parameters are the configured to ensure clarity, intelligibility, and appropriate pacing for human–robot communication. The synthesized speech is delivered through the robot's onboard speaker system, enabling the verbal interaction.

E. Adaptive Gesture Mapping and Motion Control Layer

To enhance the communication naturalness, conversational responses are synchronized with the expressive humanoid gestures. The adaptive gesture mapping module at associates conversational states—such as explanation, greeting, confirmation, or emphasis—with corresponding predefined servo motion sequences. Priority of handling mechanisms ensure that gesture execution remains the synchronized with speech output without interrupting verbal delivery.

The motion control module generates coordinated servo trajectories and transmits synchronized control signals to intelligent bus servos. This ensures smooth and expressive head, arm, and body movements aligned with the spoken response.

F. Output and System Support Layer

The output layer as represents the combined with execution of verbal response and it synchronized humanoid gestures. A regulated power management and the embedded support layer to ensures stable operation of the microphone, processor, speaker system, and servo motors during the continuous interaction sessions. The modular architecture to allows scalability and supports future enhancements such as multilingual speech and processing, emotion-aware response adaptation, and cloud-based knowledge integration.

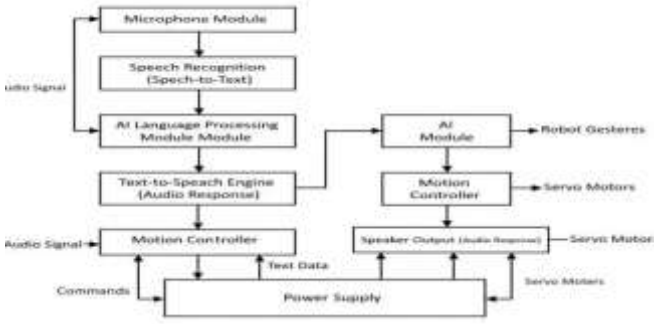


Fig 2. Block diagram of the proposed Voice_Based humanoid robot architecture.

TABLE I. Functional Layers Of The Proposed System Architecture

Layer	Module	Description
Input Layer	Microphone Module	Captures real-time voice input from the user for conversational interaction
Processing Layer	Speech Recognition & Preprocessing Module	Performs noise filtering, voice activity detection, and converts speech to text

Recognition Layer	AI Language Processing Module	Interprets user intent and generates context-aware textual responses
Decision Layer	Conversational Response Logic	Selects appropriate response behavior and gesture mapping
Control Layer	Text-to-Speech (TTS) Engine	Converts AI-generated text into natural-sounding speech output
Output Layer	Verbal and Humanoid Motion Execution	Delivers spoken response through speaker and executes expressive gestures
Support Layer	Power and Embedded System Management	Provides regulated power and stable operation for processing and actuation modules

A. Architecture Summary

In summary, the proposed system architecture has integrated speech sensing, artificial intelligence-based on language understanding, adaptive of decision-making, and synchronized at humanoid motion control into a unified embedded conversational framework. By enabling the real- time voice-driven interaction, the system as achieves autonomous and the expressive humanoid of communication suitable for educational, research, and intelligent service applications.

HARDWARE ARCHITECTURE

The hardware architecture of the proposed voice-driven query of recognition and response system is designed to the support real-time speech processing, embedded artificial intelligence execution, and the synchronized humanoid motion control. The system has integrated an audio sensing unit, an embedded processing platform, the intelligent servo actuation, communication modules, and regulated power management to ensure stable and autonomous conversational performance. The hardware configuration is optimized for the indoor interactive environments where low-latency response, clear speech capture, and the smooth humanoid gestures are essential.

A. Humanoid Robotic Platform

The Robomonk Mini humanoid robot serves as the primary mechanical platform for expressive interaction. The robot structure provides that as multiple degrees of freedom through articulated joints, enabling coordinated movements of the head, arms, and upper body. This is anthropomorphic configuration allows the robot to perform synchronized gestures such as head turns, arm raises, greeting the motions, and that expressive posture adjustments aligned with the verbal responses.

The mechanical frame as an ensures stability during the continuous conversational demonstrations while its maintaining compact and portable design suitable for educational and the research environments.

B. Audio Sensing Unit

A built-in microphone module is employed as the primary input device for capturing user speech. The microphone continuously acquires audio signals from the surrounding environment, enabling the natural and hands- free communication. Proper placement and the sensitivity configuration ensure reliable voice of detection under typical indoor acoustic conditions. The audio sensing unit forms the foundational of layer for speech-based human–robot interaction, eliminating the need for physical control buttons or an external triggering mechanism.

C. Embedded Computing Module

A Raspberry Pi 5 functions as the central processing unit of the system. The embedded processor performs speech recognition, AI-based language processing, response generation, and motion control command execution. The processor is selected to provide sufficient computational capability for real-time conversational at inference while it is maintaining low power consumption and autonomous operation.

By executing all major computational tasks locally, the system has reduced dependency on cloud infrastructure and the ensures responsive interaction is performance.

D. Intelligent Servo Actuation System

The humanoid robot utilizes high-voltage intelligent bus of servos to achieve precise joint-level motion control. These servos provide the accurate position feedback and the enable synchronized multi-joint movement during the gesture execution. So, Servo coordination ensures smooth and expressive humanoid gestures aligned with speech output. The intelligent servo architecture supports controlled acceleration and deceleration, minimizing abrupt movements and enhancing interaction naturalness.

E. Communication and Expansion Interface

A dedicated expansion and control board manages communication between the embedded processor and the intelligent servo motors and communication interface will ensure reliable data transmission for synchronized motion execution. Additionally, Wi-Fi or Ethernet connectivity enables optional external communication, software updates and if required add functionality.

F. Power Supply and Regulation System The system is powered by an 11.1V lithium battery pack along with appropriate voltage regulation circuitry. The power management system supplies stable voltage to the Raspberry Pi, microphone module, speaker unit, servo motors, and control electronics. Proper regulation is essential because dynamic servo loads can introduce voltage fluctuations during expressive motion execution. The integrated power management framework will ensure the continuous and stable operation during extended interactive sessions.

G. Hardware Summary

The proposed hardware architecture combines real-time audio sensing, embedded AI processing and synchronized humanoid motion control into a unified autonomous robotic platform. The combination of microphone-based speech input, AI-driven embedded processing, and intelligent servo actuation enables natural conversational interaction without reliance on physical control interfaces or predefined static responses.

TABLE II. HARDWARE COMPONENTS USED IN THE PROPOSED SYSTEM

Component	Function
Microphone Module	Captures real-time user speech for conversational interaction
Speaker Output Unit	Delivers synthesized verbal responses to the user
High-Voltage Intelligent Bus Servos	Provides smooth joint actuation
Servo Control Board	Enables smooth and coordinated joint-level actuation
Servo Control / Expansion Board	Synchronizes motion execution between processor and servo motors
Wi-Fi / Ethernet Communication Module	Supports network connectivity and optional external integration

Lithium Battery Pack (11.1V, 2000mAh)	Supports network connectivity and optional external integration

SOFTWARE IMPLEMENTATION AND SYSTEM WORKFLOW

The software implementation of the proposed voice-driven query recognition and response system is designed to achieve real-time speech perception, intelligent conversational processing ad synchronized humanoid motion execution. This software architecture integrates speech acquisition, audio preprocessing, understanding of natural language, response generation, speech synthesis and servo motion control within an embedded processing framework. The main objective of the software system is enabling the hands-free conversational interaction by continuously interpreting spoken queries, generating meaningful physical and verbal responses.

The software pipeline begins with continuous audio acquisition from the onboard microphone module. The captured audio stream will undergo preprocessing operations like noise filtering, signal normalization and voice activity detection to improve recognition reliability under varying indoor acoustic conditions. These preprocessing techniques reduce background noise and will make sure to enhance speech clarity prior to recognition.

The processed audio signal is forwarded to the speech recognition module, where spoken language is converted into textual format using speech-to-text algorithms. The generated text represents the user query in a machine- readable form and the textual input transmitted to the AI language processing module for the semantic interpretation and response generation.

The AI language processing module functions as core intelligence component of the system. It will analyze the contextual meaning of the query, interprets user intent and will generate a coherent and context-aware response. Unlike conventional keyword-based systems the AI language processing module will support open-ended conversational input and dynamically formulates responses without relying on statistic datasets.

Once the response text is generated, it will be shared to the text-to-speech (TTS) engine. The TTS module will synthesize natural-sounding speech which will be forwarded through the robot's speaker system. The speech parameters will be optimized to maintain clarity, proper pacing, and standard tone.

Simultaneously, the conversational state is transmitted to the gesture synchronization module. This module maps response categories—such as greeting, explanation, confirmation, or emphasis—to predefined servo motion patterns. The motion control interface generates synchronized control signals for intelligent bus servos, enabling coordinated head and arm movements aligned with speech output.

The complete workflow forms a closed-loop conversational cycle:

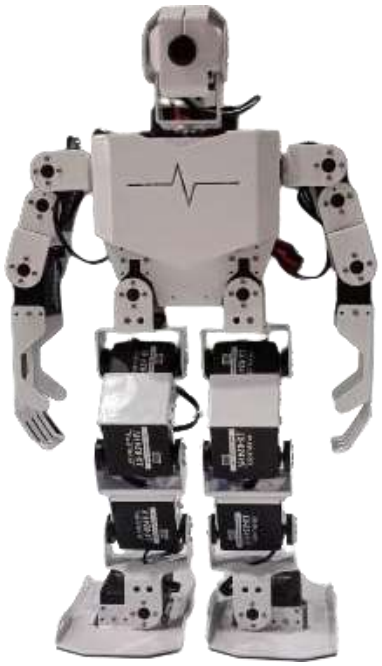
1. Audio capture
2. Speech preprocessing
3. Speech-to-text conversion
4. AI-based response generation
5. Text-to-speech synthesis
6. Gesture synchronization and motion execution

The modular design of the software architecture allows scalability and supports future enhancements such as multilingual processing, contextual memory retention, emotion-aware responses, and cloud-assisted knowledge expansion.

TABLE III. SOFTWARE TOOLS AND LIBRARIES USED

Software Component	Purpose
Programming Environment (Python)	Implementation of speech processing, AI response generation, and motion control algorithms
Speech Recognition Library	Converts spoken audio input into textual format
Audio Preprocessing Module	Performs noise filtering, normalization, and voice activity detection
AI Language Processing Model	Interprets user intent and generates context- aware responses
Text-to-Speech (TTS) Engine	Converts generated text into natural- sounding speech output
Servo Control Library	Generates synchronized servo commands for humanoid gesture execution
Embedded Operating System (Raspberry Pi OS)	Provides real-time system coordination and hardware resource management

Fig 3: Developed Humaniod Robot



SPEECH RECOGNITION AND CONVERSATIONAL RESPONSE METHODOLOGY

The proposed methodology integrates real-time speech recognition, artificial intelligence-based language understanding, and synchronized humanoid motion control to enable natural and expressive conversational interaction. The methodology is designed as a continuous perception interpretation response pipeline, where spoken input from the user is analyzed, processed and converted into coordinated verbal and gestural outputs. By combining speech understanding with expressive motion execution. The system achieves improved conversational intelligence, interaction naturalness.

A. Audio Data Acquisition and Preprocessing

The methodology starts with continuous acquisition of audio signals from the onboard microphone module. The captured speech will be digitized and subjected to preprocessing operations including noise

filtering, amplitude normalization, and voice activity detection. These preprocessing steps will improve speech clarity and reduce background noise and will reliable recognition under varying indoor acoustic conditions. The processed audio stream is segmented into some meaningful speech frames which are suitable for further linguistic analysis.

B. Speech Recognition Process

The Speech recognition will be performed using embedded speech-to-text algorithms which will convert spoken language into textual representation.

Acoustic features are extracted from the audio signal and matched against trained language models to identify words and sentence structure. Result of this stage is a textual query representing the spoken input of user.

Temporal validation mechanisms will ensure stable recognition by filtering out incomplete phrases, noise artifacts and unintended speech fragments. This stage forms the foundation for intelligent conversational interpretation.

C. AI-Based Language Understanding and Response Generation

The recognized textual query is transmitted to the AI language processing module for semantic interpretation. This module analyzes sentence structure, contextual meaning, and user intent using artificial intelligence-based language modeling techniques.

Contrast to the conventional keyword-based systems, the AI module can support open-ended conversational input and can dynamically generates context-aware responses. The output is not only limited to predefined datasets, but is formulated based on some contextual reasoning and linguistic understanding. This will be improving standard flexibility and will allow the robot to respond to diverse topics and question formats

D. Text-to-Speech Synthesis and Expressive Output Once the response text is generated, it will be shared to the text-to-speech (TTS) engine. The TTS module will synthesize natural-sounding speech which will be forwarded through the robot's speaker system. The speech parameters will be optimized to maintain clarity, proper pacing, and standard tone.

At the same time the conversational state is mapped to corresponding humanoid gestures using gesture synchronization module. For example, greeting responses may trigger arm movements, explanatory responses may start head nodding and emphasis outputs may involve in coordinated posture adjustments. Intelligent bus servos execute these movements in synchronization with speech playback, ensuring expressive and natural interaction.

E. Integrated Conversational Response Execution

The final stage of the methodology will be involved in synchronized verbal and physical response execution. Speech output and servo activation are temporally aligned to prevent desynchronization between motion and audio. The closed-loop conversational framework will make sure that each spoken query results in a coordinated and contextually appropriate humanoid response.

This integrated methodology will enhance perceived intelligence, improve user engagement, and provide a scalable platform for advanced conversational robotics applications.

EXPERIMENTAL RESULTS AND DISCUSSION

The proposed voice-driven query recognition and response system was experimentally assessed in an indoor conversational environment to evaluate its real- time performance, speech recognition reliability, AI response quality and humanoid gesture synchronization stability. The evaluation focused on

measuring speech recognition accuracy, response generation latency, synchronization between verbal output and gestures, and overall interaction adaptability under varying user conditions

A. Experimental Setup

The system was tested using real-time conversational interaction, where users asked open-ended and topic-based questions within the robot's auditory range. Experiments were conducted under various indoor acoustic conditions, which includes varying background noise levels and different speaker distances, to estimate robustness.

The Raspberry Pi 5 has executed all speech recognition, AI language processing, text-to-speech synthesis, and motion control tasks locally. The robot has operated on its own and no relying on manual triggering or predefined command datasets. Users interacted naturally without the constraints on phrasing, unlocking evaluation of contextual understanding capability.

B. Performance Metrics

The primary evaluation metrics included:

- Speech Recognition Accuracy – Percentage of the correctly interpreted spoken queries
- Response Generation Latency – Time interval between the speech input completion and verbal response initiation
- Conversational Relevance – Contextual correctness of the AI-generated responses
- Gesture Synchronization Quality – Alignment between the speech output and the servo-based motion execution
- System Stability – Continuous operational reliability during the extended interaction sessions

The speech recognition accuracy will be evaluated by comparing the recognized text output with queries which are actually spoken. Response latency was measured as the elapsed time from end-of-speech detection to speech playback initiation

C. Results and Observations

Experimental outputs demonstrated that the system achieved reliable real-time speech recognition performance in clear indoor acoustic conditions. Recognition accuracy remained high when background interruptions were minimal. Minor performance degradation was observed under moderate noise conditions, though overall conversational reliability remained acceptable.

AI-generated responses were relevant and demonstrated the ability to handle open-ended questions in multiple topics. contrast from keyword-based systems, the robot successfully interpreted the sentence variations and generated meaningful explanations rather than the static replies.

Gesture synchronization performance was stable, with servo-based humanoid movements occurring in

coordination with speech output. Head nodding, arm gestures and posture adjustments enhanced the expressiveness and improved perceived interaction intelligence.

The embedded processing framework maintained the responsive conversational performance with low latency, enabling smooth and engaging user interaction.

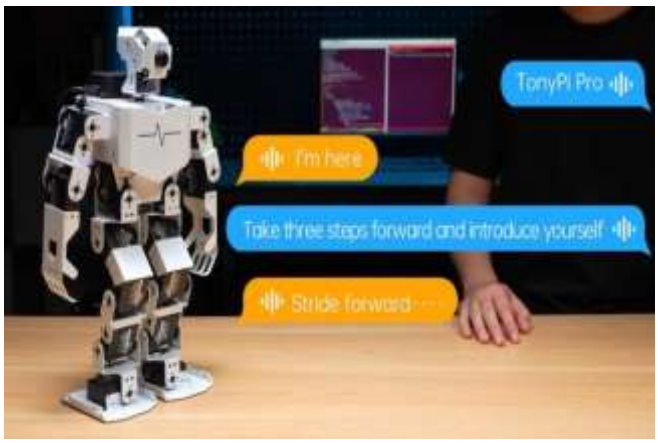


Fig 4: Voice interaction with robot and how it responds

TABLE IV. PERFORMANCE EVALUATION RESULTS

Metric	Observed Performance
Speech Recognition Accuracy	Approximately 90–94% under indoor conditions
Average Response Latency	Approximately 1–2 seconds
Conversational Relevance	Contextually appropriate and topic- adaptive
Gesture Synchronization Stability	Smooth and synchronized with speech
System Operational Stability	Reliable during continuous interaction

D. Discussion

The experimental results confirm that integrating speech recognition, AI-based language processing, and synchronized humanoid motion significantly enhances conversational interaction quality. The system successfully eliminates dependency on predefined responses and manual control mechanisms, enabling adaptive and natural communication.

The synchronization of verbal output with expressive gestures improves user engagement and perceived intelligence compared to purely speech-based robotic systems. Although environmental noise and accent variations can influence recognition accuracy, the overall system performance demonstrates the feasibility of embedded conversational humanoid robotics.

Future optimization may focus on improved noise robustness, multilingual speech handling, and reduced response latency through computational optimization techniques.

CONCLUSION AND FUTURE SCOPE

This paper has presented a voice-driven query recognition and response system implemented on the Robomonk Mini humanoid robot platform. The proposed system addresses the limitations of standard rule-based conversational robots by enabling natural, hands-free, and context-aware human– robot interactions. By integrating real-time speech recognition and artificial intelligence-based language processing, text-to-speech synthesis, and synchronized humanoid motion control, the robot will achieve

expressive and adaptive conversational behavior.

The developed architecture establishes a unified perception– interpretation–response framework in which continuous audio input is processed through speech recognition and AI language understanding modules to generate meaningful verbal outputs. At the same time gesture synchronization mechanisms will ensure coordinated servo-based movements aligned with speech delivery. Experimental evaluation demonstrated reliable speech recognition accuracy, AI-generated responses which is contextually irrelevant low-latency interaction and stable gesture execution under indoor conditions.

The results confirm that integrating conversational AI with humanoid motion control will enhance the user engagement and interaction naturalness. contrast static or keyword-based systems, the proposed framework supports open-ended queries and dynamic response generation, thereby improving conversational intelligence and adaptability.

Future enhancements of the proposed system may include multilingual speech recognition support to enable interaction Future enhancements of the proposed system may include multilingual speech recognition support to enable interaction

across diverse language groups. Emotion-aware conversational modeling can be integrated to adapt tone and gestures based on user sentiment. Advanced contextual memory retention mechanisms may further improve dialogue continuity and personalization. Additionally, integration with cloud-based knowledge databases and IoT- enabled environments can expand the robot's functional capability. Autonomous navigation and mobility integration may also allow the humanoid platform to operate in larger interactive spaces.

Overall, the proposed voice-driven humanoid interaction system that provides a scalable and the modular foundation for conversational of robotics applications in education, research, public has exhibitions, and intelligent service of environments.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the Department of Electronics and the Communication Engineering for providing the necessary facilities and the technical resources to carry out this work. The authors also extend their appreciation to the project guide and faculty members for their continuous encouragement, valuable suggestions, and expert of guidance throughout the development and the evaluation of the voice-driven query recognition and the response system using the Robomonk Mini humanoid robot.

Special thanks are conveyed to all team members and the laboratory staff for their support in system integration, testing, and experimental validation. Their contributions played a significant role in the successful implementation of the proposed conversational humanoid platform.

REFERENCES

1. D. Jurafsky and J. H. Martin, "Speech and language processing," Pearson, Upper Saddle River, NJ, USA, 2009.
2. L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," Prentice Hall, Englewood Cliffs, NJ, USA, 1993.
3. A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., Vancouver, BC, Canada, pp. 6645–6649, May 2013.

4. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," unpublished.
5. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre- training of deep bidirectional transformers for language understanding," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Minneapolis, MN, USA, pp. 4171–4186, June 2019.
6. H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," Speech Commun., vol. 51, no. 11, pp. 1039–1064, Nov. 2009.
7. S. Russell and P. Norvig, "Artificial intelligence: A modern approach," Pearson, London, U.K., 2016.