

Voice and Hand Gesture-Based Multimodal Intelligent Assistant for Human-Computer Interaction

Y. Venkata Krishna, K. Chandu, K. Rohith Kumar, G. Mohan Manjunadh Reddy, M. Manoj Reddy

Department of Computer Science and Engineering

BTech Final Year Project Report

Abstract—The traditional modality used for Human-Computer Interaction (HCI) involves keyboard and mouse input, which limits the naturalness and accessibility of human interactions with computing systems. This paper describes the design and implementation of a multimodal intelligent assistant that allows hands-free communication with computers through voice interaction, hand gesture recognition, and artificial intelligence. The system consists of speech recognition, a Large Language Model (LLM) reasoning component for providing answers to natural-language queries, a Text-to-Speech (TTS) component for spoken answers, a voice dictation component for hands-free typing, screen awareness component with AI-vision capabilities for analyzing screen content, and a lightweight floating UI for showing the state of the assistant. Hand gesture recognition is implemented as an additional way to perform control actions along with natural-language voice commands. Multilingual switching between English and Telugu languages is implemented explicitly to avoid unreliable language detection. This paper presents the architecture and methodology of the proposed system and current status of its implementation, specifying implemented components, components in development, and proposed for future development. Instead of inventing performance metrics, The paper highlights qualitative findings based on tests carried out on the functioning prototype. The objective is to show how the combination of voice, gestures, and AI-based reasoning can bring human computer interaction to a more natural and inclusive model.

Keywords: Human Computer Interaction, Multimodal Interaction, Speech Recognition, Hand Gesture Recognition, Large Language Models, Desktop Automation, Text-to-Speech, Screen Understanding.

I. INTRODUCTION

The main aim of Human-Computer Interaction (HCI) is to make human-computer data exchange natural, efficient, and accessible. Traditional interaction paradigms relying on keyboard and mouse, despite their effectiveness in delivering precise input, fail to mimic the natural way in which people communicate – through voice and gestures. Speech recognition, computer vision technologies, and Large Language Models (LLMs) have paved the way to creating intelligent assistants that would interact with users using several complementary modalities.

This work suggests a multimodal intelligent assistant capable of voice interaction, hand gesture recognition, reasoning with the help of AI to facilitate natural, hands-free interaction with desktop computing system. The suggested solution aims at processing voice commands, interpreting hand gestures that would be captured by a camera, responding to natural language questions with the help of LLM, performing desktop automation tasks, supporting voice dictation, analyzing the screen content upon request, and notifying user of the state of the system via minimalistic floating interface. Voice and gesture are supposed to be

Section II provides literature survey. Section III provides the proposed methodology, and Section IV provides the system architecture. Section V provides the current status of implementation. Section VI provides qualitative results and discussion, which includes system benefits and constraints. Section VII provides future scope, and Section VIII provides the conclusion of the paper.

A. Problem Statement

Voice assistants and gesture recognition tools have thus far mostly been developed independently of each other and have remained as separate solutions in one particular modality. While voice assistants usually do not include advanced hands-free action mechanisms for executing quick commands, gesture recognition tools fail to incorporate natural language reasoning necessary for responding to open-ended questions and holding a conversation. Moreover, most traditional voice assistants lack the integration with the desktop environment needed for automating various processes, cannot recognize on-screen visual content in order to help users solve various context-dependent problems, such as an on-screen error message, and rely on inaccurate automatic language detection while incorporating

transparent about which capabilities are fully implemented and which remain under development.

A. Research Objectives

The main purposes of the project are as follows:

- Development of a multimodal HCI system which minimizes reliance on conventional keyboard and mouse operations.
- Implementation of voice-based interaction along with hand gesture recognition as another form of control input.
- Provision of intelligent question answering and dictation services with automatic grammatical correction by using AI.
- Hands-free control of the desktop, applications, and screenshot analysis to gain awareness of the screen.
- Implementation of multilingual switching between English and Telugu for avoiding detection errors.
- Development of an accessible floating interface for displaying feedback without interfering with work.

II. RELATED WORK

This part examines the existing literature related to intelligent personal assistants, speech recognition in human-computer interaction, multimodal speech-gesture fusion and language model-based voice interaction, as well as contextualizes the proposed system in the existing literature.

In their paper, Silva et al. [1] performed a systematic review of the literature on Intelligent Personal Assistants (IPA). The authors suggest an IPA taxonomy and identify usability, security, and privacy as critical determinants of user trust in these systems. The review demonstrates that most studies devoted to IPAs focus on voice-controlled natural language interfaces, and not voice in combination with other modalities like gestures.

Rebman, Aiken, and Cegielski [2] studied speech recognition in the human-computer interaction domain through an experiment on text entry using speech and typing. Their experiment proved that, even though speech recognition enabled text creation faster, the errors caused by inaccurate speech recognition, and the need for correcting those errors decreased the efficiency and the acceptance of the process. This result is particularly useful in designing the voice dictation feature of our system, which uses artificial intelligence for text correction to ease the correction process.

Ambadar and Martinez [3] tested a multimodal fusion approach in three different environmental situations, which were normal, moderate, and extreme conditions. As a result, they were able to develop a new multimodal configuration and environment detection model and discovered that the multimodal system outperforms the unimodal system,

providing further support to the reasons behind the integration of voice and gestures in the suggested system.

Deng, Fang, and Huang [4] have come up with a multimodal human-robot interaction system which integrates speech (using corpus-based natural language understanding algorithm) with gestures which are captured by the Leap Motion sensor and filtered by Interval Kalman Filter to minimize sensor noise. They showed that integration of both speech and gestures leads to more intuitive and convenient manipulation compared to each of these modes used separately; however, their system was meant for robot command execution rather than desktop tasks.

In a relatively recent research by Mahmood et al. [5], the effects of Large Language Models on user interaction styles with voice assistants were studied through an exploratory study of a ChatGPT-powered voice assistant in the context of medical self-diagnosis, creative planning, and discussion. The results suggest that LLM-powered assistants enable richer interaction styles and greatly minimize intent recognition errors as opposed to traditional voice assistants, but at the same time create new interaction difficulties which need to be overcome with proper design.

As can be seen from the review of literature above, current systems typically consider only one of the voice interaction and hand gesture recognition, or sometimes both of them at a very restricted level, without taking into account AI/LLM-based reasoning, desktop automation, and vision-based screen awareness capabilities. The main purpose of the proposed system is to combine the above components in a single multimodal assistant system. Table I provides a comparison of representative existing work against the proposed system.

TABLE I
COMPARISON OF EXISTING RESEARCH WITH THE PROPOSED SYSTEM

RESEARCH	MODALITIES	LLM	CONTROL	AWARENESS	LIMITATIONS / COMMENTS
Silva et al. [1]	Voice (general survey)	Partial	No	No	Reviews IPAs broadly; does not propose an integrated gesture + desktop + screen system.
Rebman et al. [2]	Voice only	No	No (dictation only)	No	Focused solely on speech-driven text entry; no gesture or AI reasoning layer.
Ambadar & Martinez [3]	Voice + Gesture	No	No	No	Improves fusion reliability across environments but does not include LLM reasoning or screen/desktop layers.
Deng et al. [4]	Voice + Gesture	No	Robot commands	No	Designed for robot control, not general-purpose desktop HCI.
Mahmood et al. [5]	Voice (LLM-based)	Yes	No	No	Studies LLM-voice interaction patterns; no gesture, desktop automation, or screen awareness.
Proposed System	Voice + Hand Gesture	Yes	Yes	Yes (vision-based)	Unifies all modalities; gesture set and some modules are still under active development.

III. PROPOSED METHODOLOGY

The proposed methodology adopts a modular architecture wherein each interaction ability is built as a separate module managed by a command processing layer. The components that form part of this methodology includes.

A. Voice Input and Speech Recognition

The system uses an audio input captured by a microphone and translates the spoken word into text with the help of a speech recognizer engine. Text that is obtained is then sent to the command processor unit and analyzed. It is determined if the input corresponds to a system command such as Stop, Sleep, Start Dictation, a question addressed to the AI component, or a need to analyze the screen. Certain commands such as Stop and Sleep have high priority and are executed first..

B. Hand Gesture Input

The computer vision module uses a camera to track the gestures of the hand, such as open-palm gestures for activating, closed fists for stopping, swiping actions for scrolling, and pointing gestures for navigating or selecting. The use of hand gestures is meant to supplement the use of voice commands for quick, non-verbal commands to perform actions that are frequently used, while the voice can be used as the main command for complicated actions.

C. Multimodal Command Processing

The Multimodal Input and Command Fusion Module receives recognized speech and detected gestures and reconciles them into a single actionable command. The Intent and Command Processing Layer then classifies the fused input to determine which downstream module—AI/LLM reasoning, desktop automation, or screen awareness—should handle the request.

D. AI/LLM Processing

Natural-language questions, such as requests for explanations or general knowledge queries, are routed to an AI module built around a Large Language Model. The LLM generates a contextually relevant response, which can be displayed on the floating interface and converted into speech. The AI component is also used to support intelligent writing assistance, including text improvement during dictation.

E. Desktop Automation

Commands related to desktop and application control—such as typing text, keyboard and mouse interaction, browser control, scrolling, and media control—are executed through a desktop automation module that is invoked by the command-processing layer.

F. Screen Awareness and Vision AI

When the user requests screen analysis (for example, by saying "Analyze my screen" or "What is this error?"), the system captures a screenshot and forwards it to a vision-capable AI model. The model analyzes the visual content—including error messages, on-screen questions, and general textual or application content—and generates a response that is delivered through both text and speech.

G. Text-to-Speech Output

Responses generated by the AI module or the screen awareness pipeline are converted into speech through a Text-to-Speech module, allowing the assistant to communicate conversationally. The user may interrupt an ongoing spoken response with the voice command "Stop," which halts only the current utterance without disabling future responses.

H. Voice Dictation

The dictation module allows the user to place the text cursor in a supported application and dictate content using voice. Transcription of spoken input into the targeted application, where the use of AI in correcting missing spaces,

capital letters, punctuation marks, and simple grammar, without changing the meaning of what the user has spoken. Commands like "Stop Dictation" are interpreted and not recorded as is.

I. Multilingual Interaction

The system allows the use of both English and Telugu through explicit language selection and not automatic language detection. The assistant starts with English by default and only changes to Telugu when an explicit request like "Switch to Telugu" is made. Input language and output language are considered independent options such that the use of one word from the other language does not affect the output language.

J. Floating User Interface

A lightweight, movable floating interface displays the assistant's current state—Listening, Thinking, Speaking, or Processing—without obstructing the user's primary work area. The interface supports being hidden or shown on demand and automatically fades after a period of inactivity, while the assistant continues to operate in the background.

IV. SYSTEM ARCHITECTURE

The proposed architecture organizes the system into four conceptual layers: an input layer, a processing layer, an intelligence layer, and an output layer.

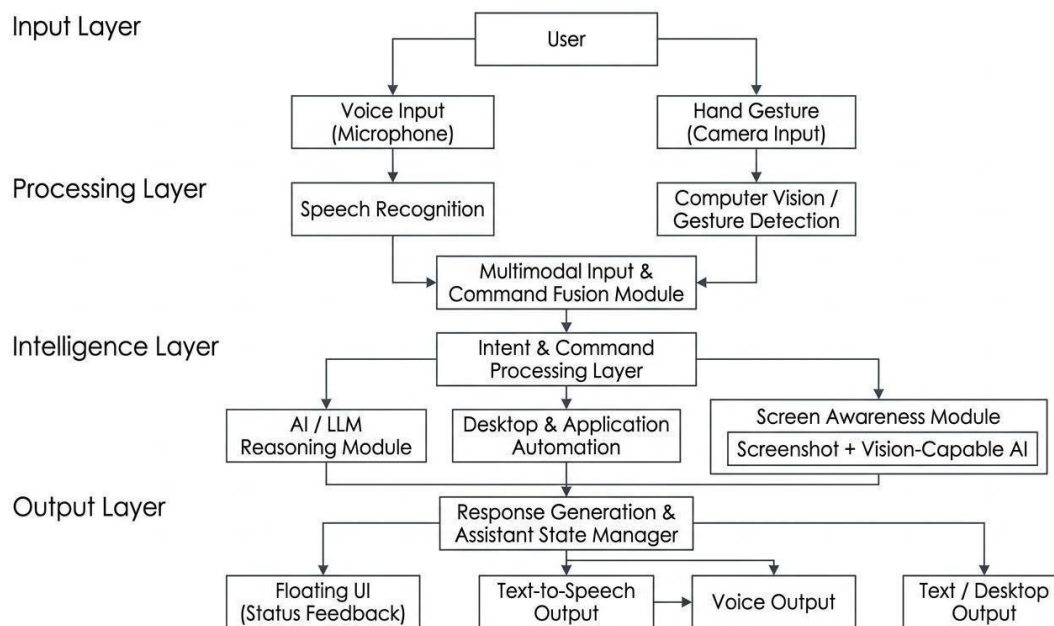


Fig. 1. Proposed Architecture of the Voice and Hand Gesture-Based Multimodal Intelligent Assistant.

The workflow of the architecture proceeds as follows. The user provides input through voice or hand gestures. Voice input is converted into text through the speech recognition module, while hand gestures are detected through the camera-based computer vision module. Both recognized inputs are passed to the Multimodal Input and Command Fusion Module, which forwards a unified command to the Intent and Command Processing Layer. This layer determines whether the request is a system command, a gesture-triggered action, an AI question, a dictation request, a screen-awareness request, or another supported operation, and routes it accordingly to the AI/LLM Reasoning Module, the Desktop and Application Automation module, or the Screen Awareness Module. Screen-related requests additionally trigger screenshot capture followed by analysis through the vision-capable AI model. All resulting outputs converge at the Response Generation and Assistant State Manager, which coordinates the final output through the Floating UI, Text-to-Speech, and text/desktop channels, with the Text-to-Speech output further rendered as

V. IMPLEMENTATION

The implementation status of the proposed system is organized into three categories to ensure an accurate and transparent representation of project progress.

A. Currently Implemented Modules

- Voice command capture and speech-to-text conversion.
- Command handling logic which favors high priority commands like Stop and Sleep.
- Question-answering by AI/LLM in response to general natural language questions.
- Text to speech conversion of AI responses with capability of interrupting mid-conversion through "Stop" command.
- Voice dictation with text entry in the focused application and basic corrections made through AI assistance.
- English to Telugu language switch.

- A floating user interface indicating assistant status (Listening, Thinking, Speaking, Processing).

B. Modules Currently Under Development

- Pipeline incorporating screen capture with vision-capable AI to facilitate the complete context-based analysis of content shown on the screen.
- Broadened the scope of desktop automation to perform a wider array of application/browser control commands.
- Fine-tuning the multimodal fusion mechanism to better accommodate voice + gesture inputs at once.

C. Proposed / Future Enhancements

- Fully developed camera-based hand gesture recognition module for the entire range of hand gestures, which includes activating, deactivating, scrolling, and pointing.
- Supported languages other than English and Telugu.
- Better offline capability of speech recognition and artificial intelligence reasoning.

One must realize that the system of hand gesture recognition is still in the concept stage and not yet developed into a functional module. This paper only states it as one of the possible modes of interaction in the multimodal framework as explained above.

VI. RESULTS AND DISCUSSION

A. Qualitative Performance Observations

The functional testing of the modules that have been developed was done informally via qualitative usage of the prototype instead of a quantitative study. During the testing process, it became clear that voice commands were identified and routed to the correct module (query for AI, dictation, or system commands) under normal and quiet indoor conditions. The AI/LLM module generated appropriate answers to the general knowledge queries while the Text-to-Speech module generated audible answers which could be interrupted by using the "Stop" command without the ability to answer other commands. The dictation module generated transcribed text into the current application, and the process of the AI-assisted correction added proper spacing and capitalization of the transcripts. Switching the languages explicitly worked as intended and did not involve any of the mistakes associated with switching languages automatically when an isolated word of a second language was used. The floating interface correctly demonstrated the current state of the assistant and did not interfere with the main work area.

However, since the hand gesture detection and screen detection modules are in the process of development, there have been no results from these two modules yet; their analysis will be done in the future when they are implemented. There have been no accuracy percentages, dataset sizes, or user study

the results of these experiments have been presented, since no controlled experiment has been performed till date. The above discussion is purely qualitative in nature and is meant to explain the functionality of the prototype.

B. Key Advantages

- Multimodal assistant is advantageous compared to existing one-modality assistants in several ways:
- Helps to become independent from the keyboard and mouse by providing voice and, in future, gesture control.
- Bridges together the best parts of voice control (natural, expressive commands) and gestures (fast and low latency actions).
- Integrates AI/LLM reasoning that answers open-ended questions instead of supporting only command sets.
- Provides accessibility by offering an alternative way of interaction to those who cannot use regular input devices.
- Gives transparent and layered feedback via floating interface without covering the screen of the user.
- Language selection via explicit language switch helps to avoid unreliable automatic language detection.

C. System Limitations

While having a number of benefits, the present prototype has some disadvantages which are currently being resolved:

- The hand gesture recognition and screen awareness modules are currently being developed and have not been completed.
- Accuracy of speech recognition may be compromised due to ambient noise, accents, and quality of microphones.
- The current design requires an Internet connection in order to perform AI/LLM processing and speech recognition and Text-to-Speech in the vast majority of cases.
- Multilingual is currently only supported in English and Telugu.

VII. FUTURE SCOPE

Future research on the system will aim at improving its functionalities and testing its usability:

- Advanced gesture recognition for hands with an extended vocabulary of gestures.
- Support for more gesture commands for additional actions with the desktop.
- Off-line speech recognition to minimize the dependence on the Internet connection.
- More languages supported.
- The ability of the system to change its AI behavior according to the preferences of the user.
- Better screen understanding.
- Mobile implementation of the assistant.
- Accessibility support.

VIII. CONCLUSION

This research paper has discussed the design and current development of a multimodal intelligent assistant, which integrates the features of voice interaction, hand gesture recognition, AI/LLM reasoning, desktop automation, voice dictation, multilingual capability, and screen awareness into a single framework of Human-Computer Interaction. The currently developed modules, such as voice command processing, AI-based answering questions, Text-to-Speech, voice dictation, explicit multilingual capability, and floating status bar, have been discussed, together with those which are under development, including full screen awareness and complete hand gesture recognition. By taking the voice and gesture interaction as complementary modes of interaction and basing the responses of AI on the structured command-processing pipeline, the proposed system tries to reach a hands-free way of computer interaction.

REFERENCES

- [1] A. de Barcelos Silva, M. M. Gomes, C. A. da Costa, R. da Rosa Righi, J. L. V. Barbosa, G. Pessin, G. De Doncker, and G. Federizzi, "Intelligent personal assistants: A systematic literature review," *Expert Systems with Applications*, vol. 147, p. 113193, 2020, doi: 10.1016/j.eswa.2020.113193.
- [2] C. M. Rebman, M. W. Aiken, and C. G. Cegielski, "Speech recognition in the human-computer interface," *Information & Management*, vol. 40, no. 6, pp. 509–519, 2003, doi: 10.1016/S0378-7206(02)00067-8.
- [3] F. F. Ambadar and J. J. L. Martinez, "Improving the performance of speech-gesture multimodal interface in non-ideal environments," *Procedia Computer Science*, vol. 216, pp. 587–596, 2023, doi: 10.1016/j.procs.2022.12.173.
- [4] Y. Deng, L. Fang, and X. Huang, "Research on multimodal human-robot interaction based on speech and gesture," *Computers & Electrical Engineering*, vol. 72, pp. 443–454, 2018, doi: 10.1016/j.compeleceng.2018.09.014.
- [5] A. Mahmood, J. Wang, B. Yao, D. Wang, and C.-M. Huang, "LLM-powered conversational voice assistants: Interaction patterns, opportunities, challenges, and design guidelines," *arXiv preprint arXiv:2309.13879*, 2023.