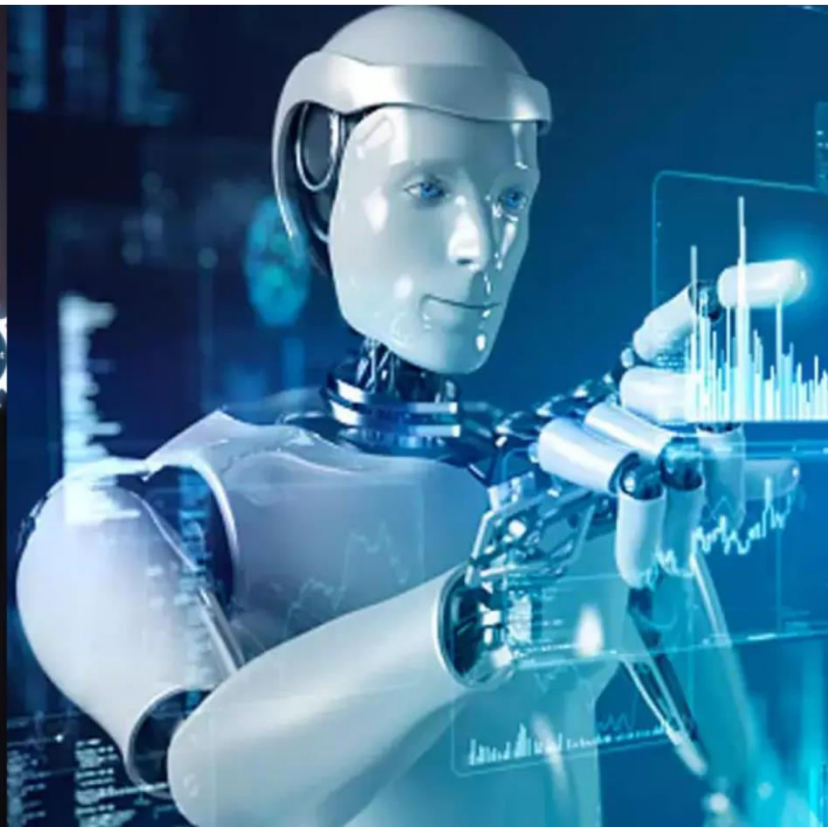
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 6, June 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506041|

SpeakEasy – AI Voice Assistant

Prof. A. N. Jalgeri, Ms. Sana Qasim Shaikh

Guide, Department of CSE, Vidya Vikas Pratishthan Institute of Engineering and Technology, Solapur, Maharashtra, India

Student, Department of CSE, Vidya Vikas Pratishthan Institute of Engineering and Technology, Solapur, Maharashtra, India

ABSTRACT: SpeakEasy AI is an advanced multimodal intelligent assistant designed to bridge the gap between conversational artificial intelligence and real-time system-level automation. Traditional AI assistants are limited to text-based interaction and lack awareness of the user's environment and local system control. SpeakEasy AI addresses this limitation by integrating natural language processing, computer vision, and desktop automation into a unified platform.

The system leverages Google Gemini Flash for contextual understanding, YOLOv8 for real-time object detection, and DeepFace for facial recognition. It enables users to perform complex tasks such as application control, file operations, voice-based communication, and visual recognition using natural language commands. The system is implemented using a hybrid architecture consisting of a React-based frontend, Node.js backend, and Python AI engine. Experimental results demonstrate improved productivity, reduced manual effort, and seamless human-computer interaction through voice and vision-based control.

KEYWORDS: Multimodal AI, Desktop Automation, YOLOv8, Gemini Flash, Computer Vision, Voice Assistant, Human-Computer Interaction.

I. INTRODUCTION

Artificial Intelligence has evolved significantly in recent years, especially with the emergence of Large Language Models (LLMs). However, most AI assistants remain confined to chat interfaces and lack the ability to interact with the real world or operating systems. Users often perform repetitive tasks such as opening applications, sending messages, or managing files manually. Existing automation tools require predefined scripts and lack flexibility, while AI chatbots lack execution capabilities.
SpeakEasy AI introduces a unified system that combines:

- Natural Language Understanding
- Computer Vision
- System Automation

This allows users to control their computer and environment using simple voice or text commands, making interaction more intuitive and efficient.

II. LITERATURE SURVEY

Traditional OCR systems such as Tesseract focus only on text extraction and lack contextual understanding. Similarly, AI assistants like chatbots provide conversational abilities but cannot interact with local systems.

Recent advancements include:

- YOLO (You Only Look Once): Real-time object detection
- Gemini AI: Multimodal reasoning and language understanding
- Automation tools: Script-based execution systems

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506041|

However, existing solutions suffer from:

- Lack of integration between vision and automation
- No real-time contextual awareness
- Limited system-level control

SpeakEasy AI overcomes these limitations by combining vision, NLP, and automation into a single intelligent system.

III. METHODOLOGY

1) System Workflow

1. User provides input via voice/text or activates camera
2. Input is processed using NLP (local + Gemini)
3. Vision module analyzes real-time frames
4. Intent is extracted and mapped to actions
5. System executes commands using automation engine
6. Results are displayed in the frontend

2) AI Processing

- **Input:** Voice/Text/Image
- **Models Used:**
  - Gemini Flash (NLP & reasoning)
  - YOLOv8 (object detection)
  - DeepFace (face recognition)
- **Output:** Structured command execution or response

3) Automation Engine

- Uses PyAutoGUI and Win32 APIs
- Controls applications like:
  - WhatsApp
  - Spotify
  - MS Word
  - Browser

4) System Architecture

- **Frontend:** React (HUD Interface)
- **Backend:** Node.js (API & orchestration)
- **AI Core:** Python (Vision + Automation)
- **Database:** Firebase Firestore
- **Flow:** User → Frontend → Backend/Python → AI Models → Execution Layer → Response

5) Features

- Voice-controlled AI assistant
- Real-time object & Face detection (YOLOv8)
- Multi-language support (English, Hindi, Hinglish)
- Open apps (Spotify, WhatsApp, Word)
- Send messages via voice & Create files/folders
- System controls (shutdown, restart, volume)
- Code generation (Python, Java, React, Android)

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506041|

IV. EXPERIMENTAL RESULTS

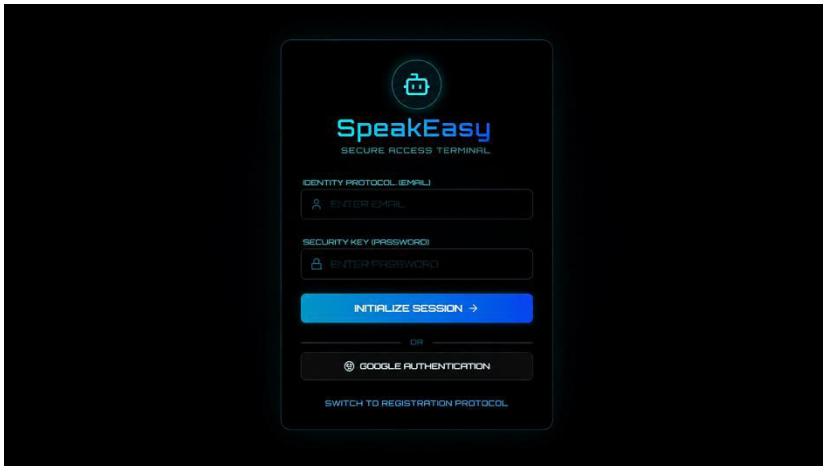


Fig (a) Shows the login interface of SpeakEasy AI with secure authentication options including email login and Google authentication.

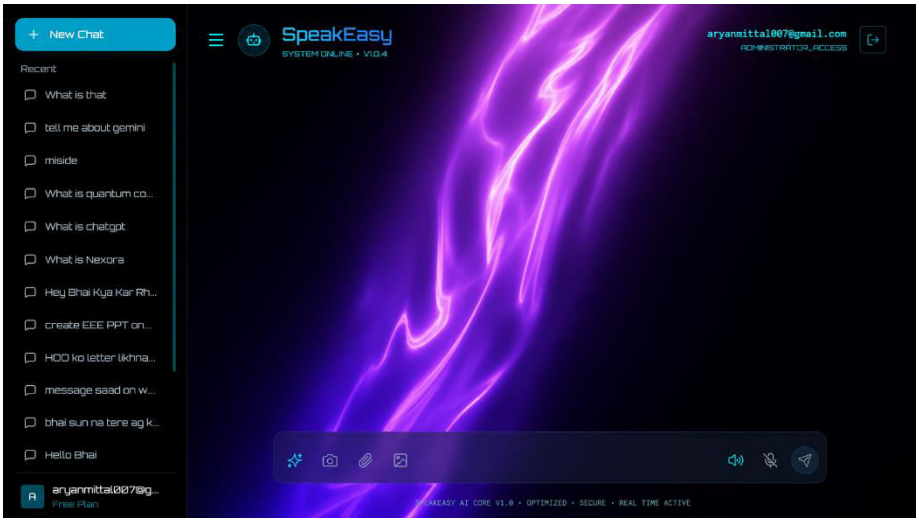


Fig (b) Displays the main dashboard (HUD) of SpeakEasy AI featuring chat interface, recent activities, and system status.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506041|

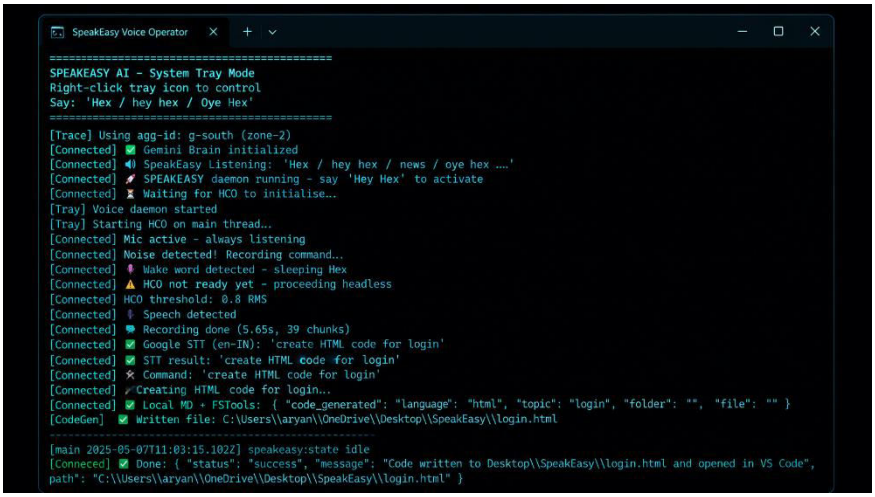


Fig (c) Shows the backend voice daemon running in terminal, processing voice commands and executing AI-driven tasks.

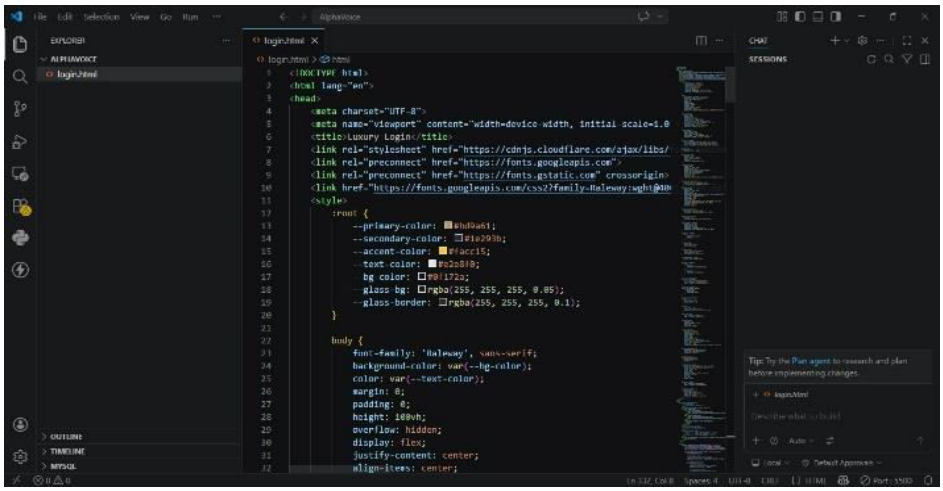


Fig (d) Illustrates the code generation output where the system creates HTML login page code based on voice input.

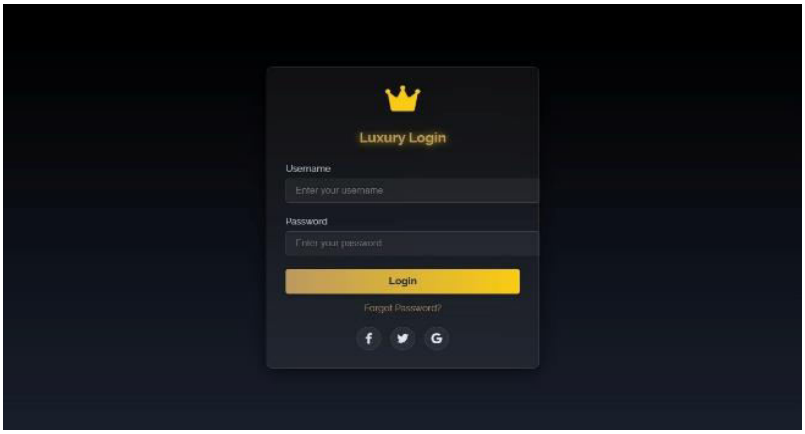


Fig (e) Shows the generated luxury login page created automatically by SpeakEasy AI using voice command.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 6, June 2026

|DOI: 10.15680/IJIRSET.2026.1506041|

V. CONCLUSION

SpeakEasy AI presents a next-generation intelligent assistant that goes beyond traditional chat-based AI systems. By integrating multimodal AI technologies, the system enables seamless interaction between users and machines. It significantly enhances productivity, reduces manual effort, and demonstrates the potential of combining computer vision, natural language processing, and automation into a unified platform.

REFERENCES

[1] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," CVPR, 2016.
[2] G. Jocher et al., "YOLOv8," Ultralytics, 2023.
[3] Google, "Gemini AI Documentation," 2024.
[4] OpenAI, "GPT Models and Multimodal Systems," 2023.
[5] OpenCV Library Documentation, 2024.
[6] PyAutoGUI Documentation, 2024.
[7] Firebase Documentation, Google, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details