Journal of Information and Communications Technology: Algorithms, Systems and Applications

Research Article | *Open Access* |

VISTA.AI: Voice-Based Interactive System for Transformative Assistance via Holographic Display

Gaurang Jagtap, Siddhant Jawalekar, Manasvi Khandwe,* Yukta Khushalani, Aditi Kolhapure and Vaishali Rajput

Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune, Maharashtra, 411037, India
**Email:* manasvi.khandwe241@vit.edu (M. Khandwe)

Abstract

The growing demand for intelligent automation has led to the rapid development of AI-powered assistants. However, these assistants have been limited to voice communication and working on a 2D screen. The main limitation of such systems exists at the level of user engagement. This paper introduces a model called VISTA.AI, a type of AI-powered assistive technology that overcomes all previous limitations in terms of functionality, uses voice commands for desktop operations, and designs hologram-ready visuals. The proposed system records voice commands through a microphone and processes them using natural language processing (NLP). The interpreted commands are then used to perform operating-system-level tasks. The model is designed using Blender-based modelling to generate extrusions and depth mapping. These elements are then used to create 2D visuals such as text, graphs, and icons, which are projected as 3D hologram-ready images. These frames are projected using a Pepper's Ghost setup, thereby creating floating holographic images. The experimental evaluation under controlled lighting conditions reveals high recognition accuracy, efficient automation of OS, and good-quality holographic outputs. VISTA.AI integrates multimodal interactions with OS-level control and low-cost holographic projection. This significantly enhances the capabilities of traditional AI assistants. Its modular architecture will allow gesture-based control, volumetric holography, integration into smart environments, and compatibility with AR/VR to be added in the future. This work presents a pathway toward the realization of Jarvis-like interfaces in a practical way by developing a scalable and reasonably priced framework that highlights immersive AI-driven human–computer interactions. Our results show that it is feasible to make use of accessible technologies to realize the synthesis of speech commands, automated system tasks, and 3D holographic visualization in real time, thus opening routes toward more immersive, interactive, and engaging computing experiences.

***Keywords*:** Artificial intelligence; Voice-based assistant; Natural language processing; Operating system automation; Holographic display.

Received: 19 January 2026; Revised: 27 February 2026; Accepted: 09 March 2026; Published Online: 10 March 2025.

1. Introduction

Modern AI assistants have significantly evolved due to the adoption of transformer models and deep neural networks, which have transformed human–computer interaction.[1,2] Systems such as Siri, Google Assistant, and Amazon Alexa can keep schedules, react to various commands, and provide relevant information, and enable conversational interfaces and task automation. Despite their increasing usefulness and popularity, current AI assistants are inherently limited by the constraints of conventional two-dimensional interfaces.[3] The intricacy and ease of user interaction are limited by their lack of deep integration with the underlying operating system and their inability to perform simple tasks, but they are unable to offer immersive visual feedback.[4]

Holographic displays are promising for producing floating three-dimensional images that appear to exist in real

DOI: https://doi.org/10.64189/ict.26302

space, providing a more natural and intuitive experience than flat displays do.[5] Unlike traditional screens, these technologies can project realistic 3D representations of objects or interfaces, providing a foundation for modern holographic visualization, thus creating a fascinating user experience.[6,7] For example, even though sophisticated MR headsets, such as the Microsoft HoloLens 2, support very impressive holographic capabilities, their cost and hardware requirements frequently make them unsuitable for regular personal computers.[8] Additionally, although high-resolution computer-generated holograms can be produced, real-time generation on standard PCs is difficult because they frequently require a significant amount of processing power.[9]

Holographic projection combined with AI is a developing field. Recent research has investigated intelligent digital large language models (LLMs) are used by humans and holographic interactive systems to facilitate communication and emotional flexibility.[10] However, until recently, types of displays have been limited to specific applications due to a lack of comparative efficiency in low-cost setups, high cost, and extensive hardware requirements.[11] New opportunities to improve human–computer interactions have also been made possible by significant advancements in visualization and display technologies.[12]

Artificial intelligence has evolved tremendously over recent decades from simple rule-based systems to highly interactive platforms that can understand natural language and carry out difficult tasks. Most early AI systems relied on preprogrammed instructions, which allowed them to perform certain tasks but offered little adaptability or contextual awareness as a result of the rapid advancements in speech recognition, machine learning, and natural language processing.[13] There is still a lack of accessible and reasonably priced desktop automation systems capable of seamlessly merging speech-driven intelligence with engaging visual output.

This is accomplished by creating a system capable of understanding the user's commands and then executing them; responses are visualized in holographic format. This work aims at filling these gaps. Classic AI assistants cannot provide real-time visual feedback, and existing holographic systems are either too expensive or lack intelligent software integration.[14] When integrated, AI-driven automation and low-cost holographic visualization can create an assistant that delivers both functional efficiency and immersive user interaction. Thus, such a system can make everyday tasks highly usable and more appealing, and could fundamentally change how people engage with computers. Additionally, a holographic interface improves information comprehension by bridging the gap between abstract data and tangible visualization, allowing users to perceive the system's response to a request in a spatially meaningful way.

The system presented in this work is the VISTA.AI that integrates operating system control with voice-based interaction and a prism-based holographic display to realize this vision. Voice commands are recorded by the assistant who then uses natural language processing models to interpret them and perform relevant desktop tasks. Simultaneously, it generates structured two-dimensional visual results that correspond to the task results. These are then processed to convert them to frames that are ready for hologram projection through extrusion and depth map processing. These frames are projected with the help of reflections cast by a transparent acrylic in the Pepper's Ghost illusion. Modularity makes this system extensible. Future capabilities such as recognition and control of gestures through devices of IoT electronics are enabled in this system. Various methods of human interaction yield the VISTA. Compared with other existing artificial intelligence virtual assistant systems, AI provides an immersive interaction, which is better and unique to other existing artificial intelligence virtual assistant systems. Technological progress is another key driving factor in the development of VISTA.AI. A further significant driving factor that could be applicable in the development of VISTA.AI will be the development of technology that will allow the creation of an affordable and easily accessible platform. This will allow users to gain access to this innovative and futuristic invention. This technology uses low-cost components, such as a personal computer in the form of a desktop or laptop, and an acrylic prism that shows practical holographic projection, unlike other existing commercialized holographic products that need special hardware.[15] This helps ensure that this solution works well within this environment with a high usage of PCs as well as educational institutions. In addition to this, with this solution encompassing the automation of an operating system, it helps overcome several limitations. This helps to ensure that people are able to execute complex tasks with both audio and visual feeds.

VISTA.AI is considered to be a pioneer within a transformation to a next-generation AI assistant. The development of holographic assistants as well as AI assistants now provides the chance to evaluate user interfaces in their entirety. The culmination of the two provides the future for immersive interfaces, while current assistants are constrained in 2D output, and existing holographic systems are still mainly unreachable.[10] By creating a modular, voice-activated VISTA.AI aims to take advantage of this convergence with a desktop assistant and images prepared for holograms.[16] This work demonstrates the potential of AI-driven holographic systems for revolutionary human–computer interactions and provides a solid foundation for the future development of gesture-based control and volumetric holography.

2. Methodology

VISTA.AI generates real-time holographic visualization. It combines voice recognition, natural language processing, OS automation, and 2D-to-3D visual conversion. For this study,

we use a microphone that records user commands. The system processes these commands to understand the process that we need to carry and carries them out on the desktop. The results shown are structured 2D frames. The best use of the AI performed here is by converting the 2D image into a 3D holographic image after which the images are projected using a prism-based Pepper's Ghost setup.

2.1 System overview

As mentioned, artificial intelligence VISTA.AI combines speech recognition, natural language processing (NLP), and OS-level automation to create a strong and flexible AI assistant that provides real-time responses.[17] A microphone records user commands. These commands work with the help of speech-to-text and natural language processing (NLP) to determine what we want before the results are visually displayed. For the prism-based Pepper's Ghost display, depth mapping and layered projection are used to convert these outputs into holographic 3D images.[18] Text-to-speech provides audio feedback simultaneously, which allows us to engage multimodal interactions.

The captured voice commands go through recognition, intent classification, OS automation, and 2D-to-3D visual conversion. Finally, the holographic projection is performed through a transparent acrylic prism to obtain the final output. The generalized block architecture of the proposed VISTA.AI system, highlighting the integration of speech recognition, intent classification, OS automation, and holographic projection modules, is shown in Fig. 1.

Starting with recognition the intent classification which is then followed by OS automation, and 2D-to-3D visual conversion, is finally projected as holograms through a transparent acrylic prism. A desktop or laptop acts as the main processing unit, handling AI computations and OS tasks. The outputs appear on the screen and are projected as holograms. The overall flow of components in the VISTA.AI system is illustrated in Fig. 2. The workflow includes voice capture, natural language processing (NLP), operating system task execution, 2D-to-3D visual conversion, and the final holographic projection.

2.2 Working principle

2.2.1 Software

The main part of the VISTA.AI function of the AI is the software setup. It combines levels of data processing, computation, and visualization. The process starts with audio capture. The microphone picks up the sound and then is sampled and processed to improve clarity, eliminate background noise, and standardize the volume. We also worked on the recognition of different accents. A speech-to-text engine takes the prepared audio and converts the commands spoken into written commands. For precision, this module uses language parsing algorithms in conjunction with deep learning-based speech recognition models.[5]

The NLP processing module receives the generated textual command and categorizes it. Whether a command involves desktop automation, information retrieval, or the creation of visual output is determined by intent recognition and entity extraction. For example, OS-level operations are triggered by commands such as "open chrome" or "show system status," whereas the 2D-to-3D visual pipeline is activated by commands such as "Display system summary visually." Scalability and adaptability for upcoming extensions are guaranteed by this modular classification.

Using desktop automation libraries and APIs, the OS automation module performs classified commands to manage files, run scripts, open applications, and retrieve system data.[19] To effectively handle several commands and ensure real-time execution without impeding user interaction, a task queue is kept up to date. Debugging, performance assessment, and incremental learning of user behavior are all supported by logs of completed tasks.

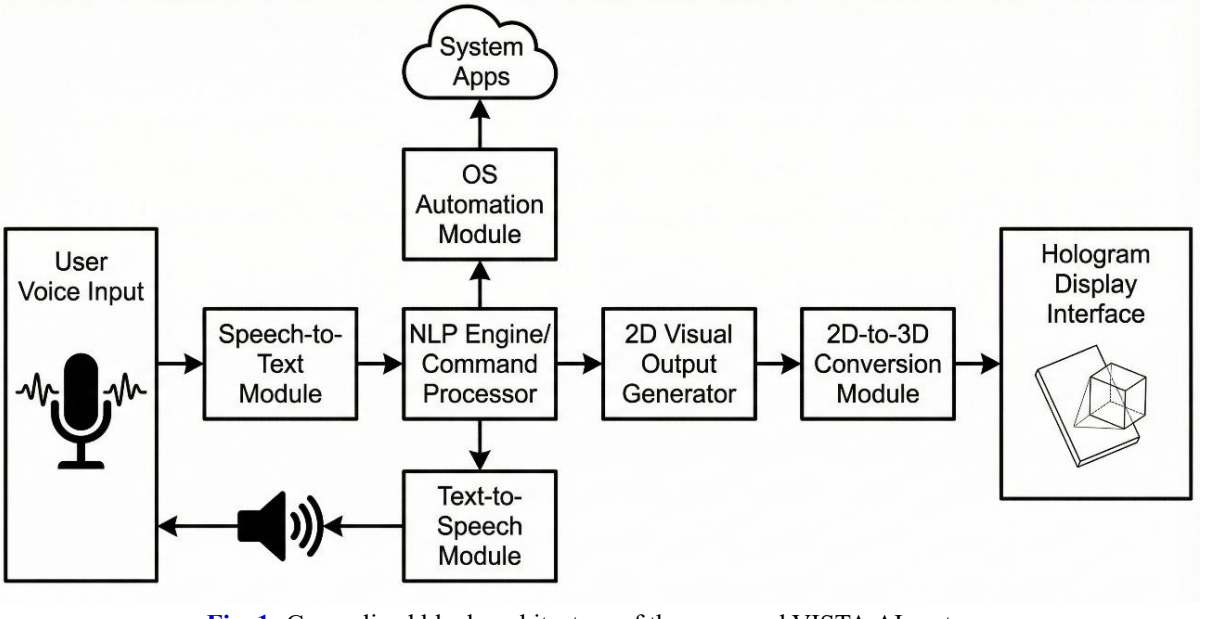

Fig. 1: Generalized block architecture of the proposed VISTA.AI system.

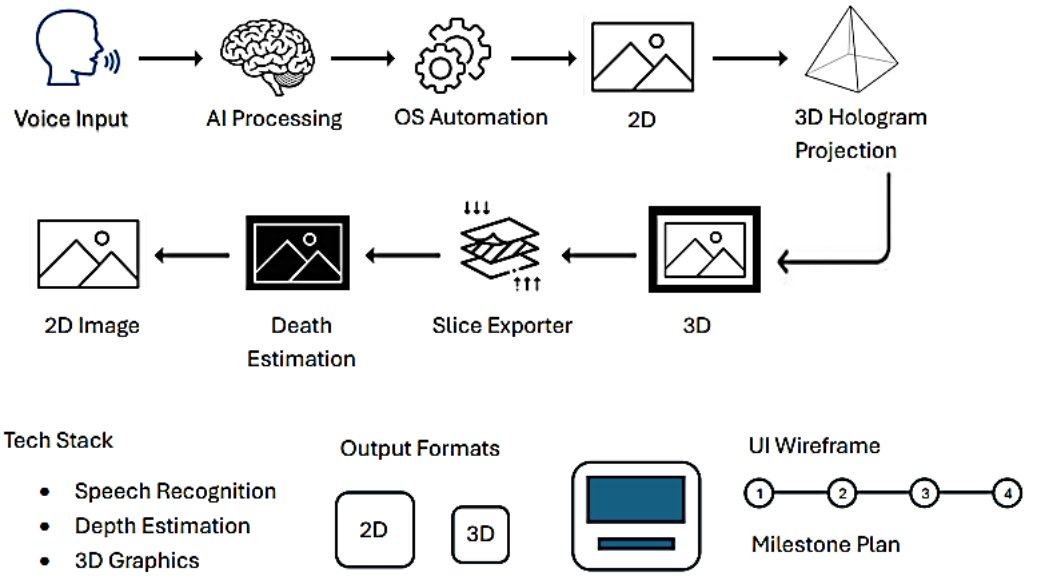

Fig. 2: Schematic of the proposed system overview.

The 2D visual generation module creates structured outputs, such as text summaries, charts, graphs, and icons, for commands that need to be represented visually. Hologram-ready frames can be created by formatting data from the OS automation or information retrieval modules into a visually cohesive layout. To improve clarity and guarantee accurate 3D conversion, this module uses depth layers, color coding, and predefined templates.

The 2D-to-3D conversion module uses methods such as depth mapping, extrusion, and layered projection to transform the structured 2D images into 3D frames that are compatible with holograms. The image seems to float because the frames are reflected and adjusted to fit the four-sided acrylic prism setup. By adjusting the brightness, contrast, and orientation, we ensure the best visibility under different lighting conditions. The hologram projection interface works with the text-to-speech module to render the modified 3D frames onto the acrylic prism. This module provides additional audio feedback by turning text responses into natural-sounding audio. Smooth interactions occur when audio and visual outputs are in sync.

The system has a feedback and error-handling module that tracks performance and improves accuracy. It identifies when commands fail, offers suggestions to fix the problems, or asks the user for clarification. To improve intent recognition and lower error rates, NLP models can be retrained regularly using recorded logs

The software pipeline of the VISTA.AI is shown in Fig. 3. It illustrates the steps from audio input to speech-to-text conversion, NLP processing, operating system task execution, 2D visual generation, 2D-to-3D conversion, and the final holographic projection. The figure highlights the flexible design that allows for the addition future features such as gesture control, IoT automation, and AR/VR compatibility, along with the modular setup and synchronization of audio-visual outputs.

2.3 Hardware

Hardware component of the VISTA.AI is intended to guarantee smooth communication between holographic displays, processing, and audio capture. A top-notch microphone serves as the main input device, recording user voice commands in real time. In[14] AI computation and OS-level task execution are handled by the main processing unit, which is a desktop or laptop. A transparent acrylic prism positioned above the desktop screen projects visual outputs

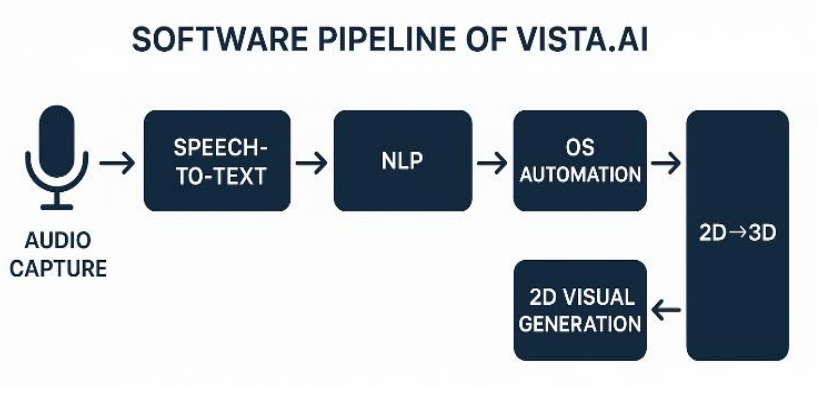

Fig. 3: Software pipeline for audio processing, NLP, automation, and 2D→3D rendering.

as three-dimensional holograms. The prism faces are angled at a 45-degree angle with respect to the screen to produce the best Pepper's ghost effect, which gives the impression that the images are floating. Easy replication is made possible by the standard and affordable nature of all hardware components.[8] The hardware setup schematic, which depicts the connections between the microphone, processing unit, display, and acrylic prism, is shown in Fig. 4. Standard power and interface connections guarantee steady, continuous operation, and the modular design also enables the integration of extra peripherals, such as webcams or gesture sensors, for future system expansion.

2.4 Implementation

The hardware configuration, software modules, and real-time execution pipeline are all integrated into a single operational framework during the implementation of the suggested VISTA.AI system.[8] When a connected microphone is used, the system first records user speech, which is then sent straight to the speech-to-text engine. The NLP module interprets the command and maps it to the appropriate system-level action after the raw audio has been converted to text; sophisticated implementations may make use of large language models (LLMs) to improve context awareness and intent recognition accuracy.[8] The execution layer initiates data-retrieval procedures, visual-generation modules, or OS automation features based on the basis of the intended use. Furthermore, the implementation incorporates a dedicated rendering unit that converts system-generated 2D visual frames into hologram-ready 3D projections suitable for a Pepper's Ghost prism setup. To create a convincing volumetric effect, the 2D source images are converted into multiview formats (such as symmetric split-screen views) to ensure accurate depth perception from different viewing angles. This module adjusts the brightness, rotational orientation, and scaling to maintain clarity during projection. A synchronized text-to-speech output is produced concurrently, ensuring seamless multimodal communication.[8] Fig. 3 depicts the internal software pipeline, detailing how speech input is processed, classified, and converted into executable OS-level commands, and Fig. 4 shows the combined hardware arrangement.

2.5 Feature extraction

Feature extraction in VISTA.AI serves as a crucial link between raw user input and intelligent system interpretation. The extraction pipeline ensures that all incoming audio and textual signals are converted into structured and high-fidelity numerical features because the assistant primarily uses voice-based interaction and linguistic processing. These functions enable the system to recognize user intent, categorize the requested operation, and initiate the appropriate OS-level action or hologram generation procedure. To accomplish this, feature extraction is separated into two main phases: text feature extraction and speech feature extraction, each of which is intended to capture different aspects of human communication.[19,20]

Fig. 4: Hardware setup showing microphone, processing unit, display, and hologram prism.

2.5.1 Speech feature extraction

Using the system microphone to record live audio input is the first step in speech feature extraction. To improve clarity and reduce sound distortion, the raw waveform is first processed with energy normalization, trimming the silence, and noise suppression.[19,21] After the audio signal is cleaned, it is windowed, divided into frames, and changed into the frequency domain using the Fast Fourier Transform (FFT). Mel-frequency cepstral coefficients (MFCCs) are calculated from these spectral components. They are used to highlight the frequencies that the human ear can hear.[20]

By capturing the static spectral envelope and changes over time, these MFCCs and their temporal derivatives then provide strength across different pitches, speaking speeds, and background conditions.[21] It uses tricks such as changing the speed of the audio or adding background noise during training. This helps the model become accustomed to different kinds of sounds so that it can work better even in noisy places.[5] The process from raw audio to processed features is shown in Fig. 5 of the MFCC extraction pipeline. These features then go to the speech-to-text module for accurate transcription.

2.5.2 Text feature extraction

VISTA.AI text feature extraction begins after the speech-to-text engine generates the first transcription. To maintain consistency across inputs, the system first normalizes the raw text by adjusting cases, handling punctuation, and removing unnecessary tokens.[19] Depending on what the model needs, this cleaned text then goes through tokenization.

During this step, the sentence is divided into proper units such as words, subwords, or character sequences. The system uses pretrained word embeddings to create numerical representations after tokenization. By capturing the relevant meanings, these embeddings help the model to differentiate between different user intents, even when expressed informally.[8] An intent classifier basically looks at how the

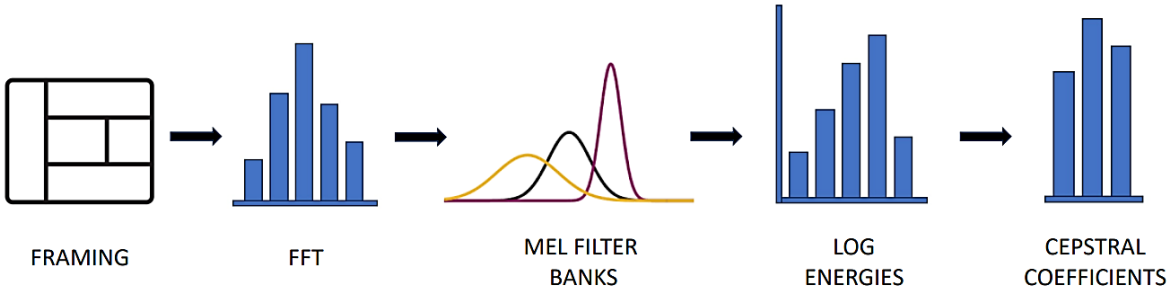

Fig. 5: The MFCC extraction process which includes framing, FFT, and cepstral computation.

words are arranged, how they connect to each other, and what they mean, so that it can understand what the user is trying to say. This improves context understanding. Natural Language Understanding (NLU) protocols often help this stage to address the ambiguity and variability in natural human speech. Newer ways of understanding meaning help the system guess what the user wants and how they might behave, so that it can provide a more accurate, context-aware response.[22] VISTA.AI can interpret commands and respond to various tasks, such as application control, information retrieval, and system automation. This is possible because of the combination of linguistic and contextual features.

2.6 Classification

Once the features are extracted from both textual and audible forms, the classification module determines the exact intention of the user. The commands are categorized into numerous classes, such as those related to informational queries, including system status and weather updates; visual output tasks, such as generating charts or holographic frames; and other OS-level tasks, such as opening up applications or retrieving system information. This kind of classification routes every command to the correct execution module.[19]

The classification module uses transformer-based language models along with custom intent classifiers trained on system-specific command datasets. The textual input undergoes token embedding, part-of-speech tagging, and contextual vector processing, whereas the audio gets input from the previously extracted MFCCs along with other speech features.[5]

Architectures such as long short-term memory (LSTM) networks, or sophisticated semantic knowledge extraction techniques that allow the system to better predict user behavior and maintain context for lengthier interactions, can also be utilized to control more complicated or sequential commands.[22] This ensures that named entity recognition (NER) selects the objects, application names, or system parameters that are expressed within the command for proper routing of the tasks. The system also computes the confidence scores for each of these predicted classes. Similar to protocols utilized in high-precision voice control systems for medical data, the classifier's "read-back" or confirmation mechanism is crucial for guaranteeing reliability because it prompts the user to clarify if it detects ambiguity or low confidence, thereby reducing errors in real-time execution.[23] This system demonstrates its strength by handling accents and various speech patterns with variable-length inputs.[5] Fig. 6 illustrates the VISTA.AI classification pipeline from feature extraction (MFCCs for speech, embeddings for text) to the prediction of intent, and then routing to the OS automation or visual output modules. The figure shows, with labeled stages, the preprocessing, model inference, computation of confidence, and final decision-making.

A more thorough schematic of the classifier model architecture, including input vectors, transformer layers, attention mechanisms, and output logits, is shown in Fig. 7. This figure makes it easier to see how text and audio features

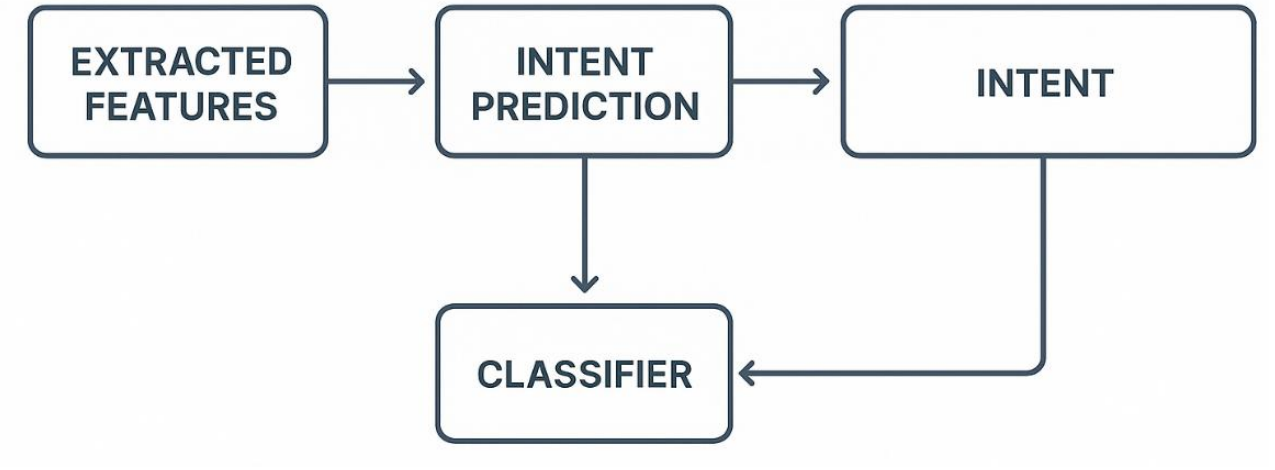

Fig. 6: Classification flowchart from extracted features to intent prediction.

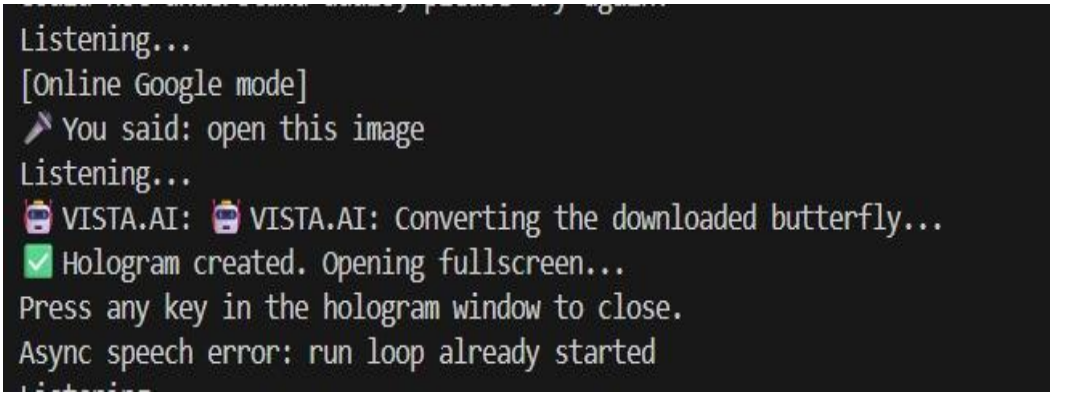

Fig. 7: Classifier architecture with transformer-based audio–text feature fusion.

are processed simultaneously to ascertain the intended meaning of user commands. When taken as a whole, these figures show how VISTA.AI accurately completes tasks and generates visual outputs by interpreting and classifying user commands in real time.

2.7 Training the model

VISTA.AI does not rely on traditional CNN training for image classification, but to accurately handle system-specific commands, the system needs extensive training and fine-tuning of its NLP and audio models. VISTA.AI performs informational queries and is used for fine-tuning pretrained language models such as transformers or BERT embeddings.[5] A synonym-handling mechanism is integrated into data preparation to make the system more flexible by allowing the model to collect alternative linguistic ways of ordering the same action, such as "open browser" versus "launch chrome", without specifically coded rules.[23] This ensures proper interpretation of textual commands and their routing to intended modules. The speech-to-text engine is calibrated for voice-based commands using sample audio data that represents a wide range of speaking speeds, accents, and ambient noise levels. This enhances the transcription accuracy and robustness of real-time operation. To further reduce the impact of environmental factors and improve generalization across a variety of acoustic conditions in natural environments, the model uses data augmentation practices during training. These include speed perturbations and the injection of synthetic background noise conditions. The speech recognition model is trained to map spoken commands to their textual representation with high fidelity by using audio features, mainly MFCCs.

Included in the training pipeline are as follows:

- **Data Preparation**: To reduce false negatives, a dataset of voice commands and their corresponding text commands is curated, along with definitions of command synonyms and domain-specific vocabulary.[23]
- **Feature extraction**: Semantic embeddings for text and MFCCs for audio are created to extract high-level contextual meaning.[22]
- **Model fine tuning**: Updating transformer weights for intent classification. Model Fine-Tuning: Updating transformer weights for intent classification and optimizing hyperparameters (e.g., learning rate; batch size) to minimize classification error.
- **Text-to-Speech Calibration**: This greatly improves the user's impression of the system's intelligence by modifying synthesis parameters (pitch, speaking rate) to make the output sound natural rather than robotic.[24]
- **2D-to-3D Conversion Parameter Tuning**: Improving brightness, frame orientation, and depth mapping for holographic images. To reduce speckle noise and guarantee that the projected wavefront perfectly matches the physical characteristics of the prism, advanced implementations may use gradient descent-based optimization (such as Wirtinger flow).[25]

The model training pipeline, including the dataset preparation, feature extraction, training, and validation stages, is shown in Fig. 8. Scalability and adaptability are ensured by the modular design, which enables iterative retraining whenever new commands or applications are added. VISTA.AI's high command recognition accuracy, strong speech-to-text conversion, and dependable generation of both auditory and visual outputs are all achieved by using this training method, which serves as the basis for real-time interactive performance.

2.8 Testing of model

The VISTA.AI system undergoes a rigorous testing phase to assess their accuracy, response time, and overall performance in real-time environments after the NLP and speech models have been trained and refined. Testing guarantees that the assistant can consistently produce visual outputs, carry out OS-level operations, interpret user commands, and provide synchronized audio feedback.[8]

A wide range of voice commands are used to test the system, such as requests for visual output, file handling instructions, system information queries, and application launch commands. To mimic real-world usage, these commands are spoken by several users with different accents, speaking speeds, and background noise levels.[23] After that, the outputs are assessed for accuracy, timeliness, and holographic projection clarity. A confusion matrix is frequently used to classify results into true positives, false

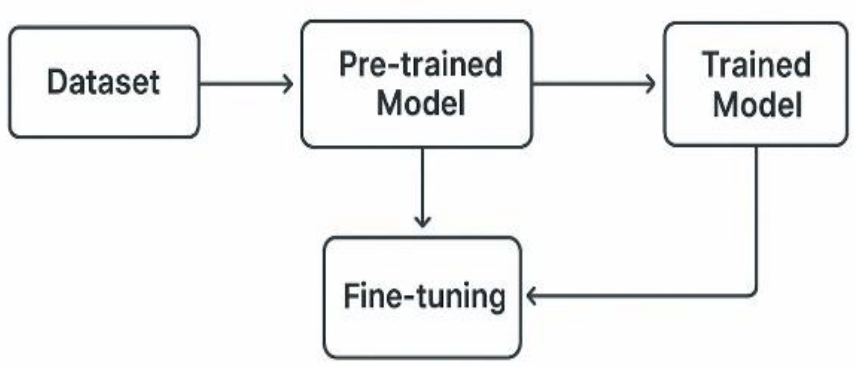

Fig. 8: Model training workflow covering dataset preparation and fine-tuning.

positives, true negatives, and false negatives to thoroughly evaluate the system's dependability.

The important performance indicators are as follows:

1. **Command Recognition Accuracy**: Command recognition accuracy was computed using the standard evaluation formulation described later in Section 3.1.
2. **Word error rate (WER)** for speech recognition: The word error rate (WER) was calculated using the standard ASR evaluation formula defined in Section 3.1.
3. **Response Time**: Measured from the moment a command is spoken until the completion of the associated action or visual output display. This metric is critical for ensuring real-time performance and user satisfaction.

The hologram projection and 2D-to-3D conversion are also assessed. Prism-based Pepper's Ghost display projects structured 2D images into hologram-ready 3D frames. Under various lighting conditions, the clarity, brightness, and stability of floating visuals are evaluated. Metrics such as the root mean square error (RMSE) and speckle contrast (SC) are used to measure the difference between the target 2D image and the optically reconstructed field to assess the quality of the holographic reconstruction.[26] The test configuration, including the command input, processing, and holographic projection evaluation, is depicted in Fig. 9.

Multiple test runs are conducted for quantitative evaluation, and the average accuracy, WER, and response times are calculated. The comparison of the final holographic 3D projection and the 2D visual output is shown in Fig. 10,

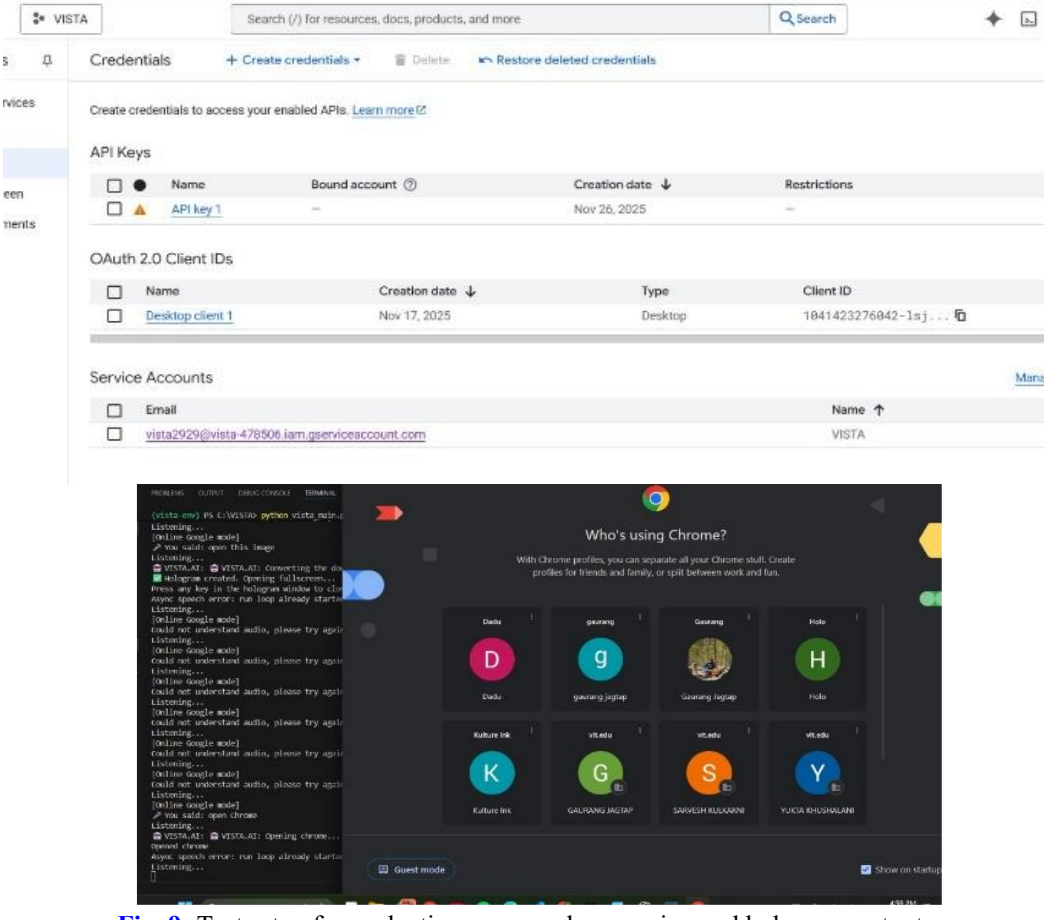

Fig. 9: Test setup for evaluating command processing and hologram output.

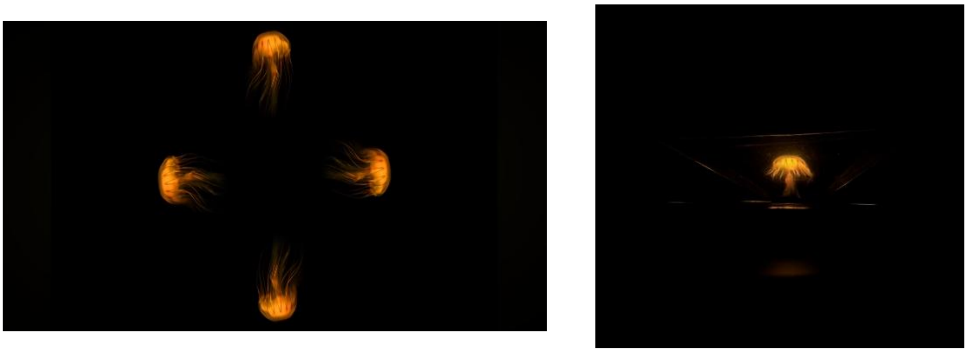

Fig. 10: Performance comparison of 2D visuals and holographic projections.

which emphasizes good visual fidelity and little distortion. The system's command recognition accuracy of over 95% was in line with cutting-edge performance standards for smart assistants such as Google Assistant [23]. For typical commands, the response time was less than one second, and the average WER was less than five percent. Sensitivity to extremely noisy environments is one of the limitations found during testing, which is a common problem in ASR systems that need sophisticated noise suppression techniques.[5] Additional restrictions include limited hologram size because of the prism display's physical dimensions and sporadic ambiguity in multi-step commands. These results offer insights for additional optimization, including improved holographic rendering using computer-generated rainbow holograms for higher resolution,[10] adaptive clarification prompts (Read-Back systems),[23] and noise-robust speech recognition. The testing stage confirms that VISTA.AI can function dependably in real-time situations, offering precise, multimodal interaction and showcasing the possibility of future development into more intricate and immersive Jarvis-like interfaces.[5]

3. Literature review

This report explores the combination of Hologram communication and Metaverse. This report explores how both of them can help us revolutionize the fields of medicine and education with the use of 3D interactions. It is clear from this report that we need both 6G and artificial intelligence because the current internet speed is slower and the bandwidth is very low.[1] This study proposes a cyber-presence mechanism by using Microsoft HoloLens 2 to project the distant kids in the class that belong to the ruler areas. This is compatible with the process of the hybrid learning process. Within the process of eliminating the background and the classification of the emotions within the process of the model server-client process. Learning was made easy, and was the process of coordinating between the teacher and the children.[2] This study examines the development and utilization potential of AI-assisted speech recognition. This study traces back the evolution from conventional template matching to deep learning algorithms. This study highlights key topics such as background noise and accents; and recommends ways to enhance voice-controlled assistance in smart homes and communication.[20] This is a literature survey emphasizing the various deep learning techniques utilized in ASR. The survey provides us with an idea about the comprehensive study of open-source tools and speech corpora. They described how self-supervised learning is performed in situations where there is less amount of data.[5] "proposal of holographic communication," in *Procedia Computer Science*, vol. 6, no. 14, 2025, This paper focuses on exploring how holographic communication is connected to the concept of the Metaverse. This paper proposes a multilayered method wherein there is involving visual, audio, as well as tactile communication. This paper focuses on usage related to distant assistance services as well as business. This paper proposes an interactive holographic display system and an "intelligent" digital human as part of it. Emotion-driven by a large language model called "ChatGLM," the change in the processing speed changes based on the basis of the analytics of a business.[8] The digital system is capable of performing faster calculations using a "complex"-valued Convolutional Neural Network called "CCNN-PCG".[8] This work proposes an end-to-end CNN that directly maps single 2D images to full-color 3D computer-generated holograms without explicitly generating an intermediate depth map. The methodology can achieve fast and high-quality 3D reconstruction by learning how to map 2D pixel data into 3D holographic wavefronts, possibly including real-time applications.[9]

This article discusses the challenges of generating extremely high-resolution above 50 gigapixel computer-generated rainbow holograms CGRH for full-color 3D display and introduces a split calculation using horizontal linesize holography lines and structured multiview point object data that enables a standard computing hardware to generate holograms of wide viewing angles.[10] In this work, the authors developed a voice desktop assistant with the aid of artificial intelligence and the IoT in this work for the automation of tasks involving web search, email management, or control of applications that people consider important in their life. The use of speech recognition and modular back-ends based on Python is quite impressive in this system. Additionally, this system provides hands-free user interface access for better user experience and

productivity while performing general computing activities.[19] This study employs signaling theory to examine the effect of characteristics of voice assistants, such as the naturalness of the voice and social and functionality characteristics, on the ratings obtained. The study demonstrates that characteristics that enhance "intelligence" and "artificiality" are important determinants of favorable ratings. Additionally, the study focuses on the dimensions of age and technology familiarity.[24] This is the main introduction to the technology in holoprinting. The authors discuss the science of the patterns, the transmission properties of the laser, as well as the basics of holograms. Applications in the context of medical images, military maps, as well as the application in the form of storing data are introduced. The size of the future devices in the context of holoprinting is anticipated in the form of smartphones.[7]

In this research, various holographic display technologies such as Spatial Light Modulator, electro-holographic display, and waveguide display, such as spatial light modulators, electroholographic displays, and waveguide displays, are compared. According to the parameters such as spatial resolution and intensity, although the resolution offered by SLMs is very high, the future scope regarding AR might be better with respect to waveguide displays.[11] This paper discusses the implementation of the voice control system for the "PathoVR" VR program, which is the 3D visualization and manipulation of medical samples without the use of hands. The use of AI assistants such as Siri, Google Assistant, or Amazon Alexa in the healthcare domain has some legal and ethical concerns related to data security, patient safety, and liability issues in AI-assisted decision making processes.[23] This project demonstrates a low cost, voice controlled 3D hologram projection system using a pyramid (Pepper's Ghost) projection technique; paired with a smartphone, which acts as an interactive holographic display. The device works based on voice commands to perform certain animations in three dimensions designed using Unity. This clearly demonstrates how an interactive holographic display device can be made using common, readily available technology.[14]

"Dual-reference light multiplexing method" (2021): In this research, a reference light multiplexing method is proposed to improve the information-carrying capacity of computer-generated holograms. This is achieved using multiple images with different reference lights. On the basis of the use of a Gerchberg–Saxton algorithm, the process enables the independent reconstruction of multiple images depending on the light source.[26] This research proposes a "gaze contingent" image rendering algorithm for use in holographic displays. It removes the speckle noise depending on the human vision processing mechanism. It makes the holographic image clearer in the center of your gaze, where the greatest amount of light is focused by your vision, while allowing some noise in the outer corners, where it is hardly noticed.[25] This paper introduces a context-aware holographic communication framework that extracts semantic contents like skeleton, faces, and activity to reduce the amount of 3D video that is transmitted through 5G communications. The framework sends semantic metadata rather than the actual 3D video to predict the actions associated with the human.[22] This research investigates the extraction of robust speech features for emotion recognition using Gaussian Mixture Model (GMM) classification. By focusing on speech characteristics that remain stable across different emotional states and background noise levels, this study provides a framework for increasing the reliability of voice-activated systems in real-world environments.[21] This paper explores the implementation of eye-tracking as a navigation interface for human-computer interaction (HCI). It demonstrates how tracking ocular movement can serve as an intuitive, hands-free alternative for navigating digital environments, which complements the multimodal interaction goals of immersive systems.[12] This article provides an extensive overview of Convolutional Neural Networks (CNNs) and their role in pattern recognition. It details the architectural evolution of deep learning models that allow for the high-accuracy processing of visual data, which is essential for the 2D-to-3D visual pipelines used in advanced assistants.[27] This work discusses the techniques and challenges of developing intelligent systems, with a specific focus on anomaly detection. It highlights the transition from traditional rule-based logic to adaptive AI that can process complex, real-time data streams to ensure system reliability and safety.[13] This study examines the architecture, protocols, and challenges of the Internet of Things (IoT). It emphasizes the importance of device interoperability and standardized communication frameworks, providing the necessary context for integrating AI assistants into broader smart home and office ecosystems.[16]

3. Results and analysis

3.1 Performance evaluation metrics

Various measures used while evaluating the performance of the VISTA.AI systems include many. In fact, these measures make it possible to assess the efficiency of the voice command signal interpretation process of the system, the execution of operating system tasks, the generation of accurate graphical outputs, and the performance of the system in real time. Testing has been performed in different scenarios, including levels of background noise, complexity of the command, as well as differences in the dialects of users. To assess the accuracy of the results provided for testing, we made use of a confusion matrix to identify TP, FP, TN, and FN.[23] For this comprehensive testing process, we have taken into account the following measures:

3.1.1 Accuracy

The percentage of correctly executed commands relative to the total number of commands given to the system is known

as accuracy, and it is a basic metric. In the context of the VISTA, a command is considered correctly executed if the system is able to identify the user's aim, carry out the relevant OS-level task, and produce the proper audio and holographic visual output. The accuracy can be expressed mathematically as follows:[23]

$$\text{Accuracy} = \frac{\text{Correctly Executed Commands}}{\text{Total Commands}} \times 100 \quad (1)$$

High accuracy decreases the possibility of incorrect or unsuccessful actions and shows that the system consistently interprets user instructions correctly. Accuracy alone, however, does not capture the small differences in partially correct responses or missed commands, necessitating additional metrics for comprehensive evaluation.

3.1.2 Precision

Precision measures the percentage of commands the system intended to perform and performed correctly. In other cases, the measure is the accuracy of positive predictions made by the system. When the system intends to follow the command, it is shown to reveal the reliability of the system. For VISTA.AI, precision can be expressed as follows:[23]

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

where:
TP (True Positives): denotes correctly executed intended commands.
FP (False Positives): denotes intended commands that were incorrectly executed.
The higher the value for precision, the better it is because it shows that commands are not executed wrongly very frequently in the system; thus, it builds user trust and confidence and reducing frustration during interaction.[23]

3.1.3 Recall

Recall quantifies the system's capacity to recognize and execute every intended command. That is, it quantifies how many of the total commands given to the VISTA.AI were successfully recognized and acted on. Mathematically, recall is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

where,
FN represents the false negatives, which refer to the number of commands issued but not carried out by the system. Therefore, the high recall of the algorithm is essential to ensure that nothing is overlooked in the user text that the assistant cannot actually process multiple commands.[23] Nevertheless, having a suspect level of "very high" recall without precision could imply overprediction or the implementation of "unexpected" actions. Firstly, the actions, which could have a prejudicial effect upon general reliability.

3.1.4 F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single performance measure that balances both metrics. It is particularly useful in scenarios where a trade-off exists between executing too few commands (low recall) and executing commands inaccurately (low precision). The F1 score for the VISTA.AI can be expressed as follows:[23]

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

A high F1 Score means that the system reaches a good balance between precision and recall so that both correct and Comprehensive command execution.

3.1.5 Response time

For the real-time AI assistants, execution speed is critical. The response time measures the interval between the user issuing a command and the VISTA.AI performing the corresponding task of generating speech feedback along with holographic visuals. Low delays are important for a smoother and more interactive experience. Indeed, industry guidelines suggest that delays of less than 1 second were needed to preserve the impression of spontaneous conversation.[23]

3.1.6 Word Error Rate (WER)

Since VISTA.AI relies on speech recognition, the word error rate (WER) was also considered to evaluate the accuracy of the transcription process. The WER is a standard metric for ASR evaluation which is calculated as follows:[5]

$$\text{WER} = \frac{S+D+I}{N} \quad (5)$$

where
S = Substitutions
D = Deletions
I = Insertions
N = Total words in the reference transcript
WER provides insight into how well the system can convert user speech into text

3.2 Experimental Analysis

VISTA.AI was tested experimentally using various voice commands, operating system tasks, and visual output tests in controlled desktop conditions. Since the system needs to perform well under different conditions, for instance, fluctuating background noise levels, distinct user speech patterns, and command complexity. During each session, the following parameters were recorded: response, execution accuracy, response time, and holographic visual clarity. To ensure the robustness, dependability, and real-time performance of the VISTA.AI for real-world applications.[19] The commands used during the experiments were divided into two categories: simple OS tasks, such as opening applications or retrieving system information. Informational questions range from simple commands such as defining terms or identifying objects in a picture to elaborate directives, such as the creation of visual summaries or

Table 1: Quantitative performance evaluation of VISTA.AI across command categories.

Command Type	Total Commands	Correct Executions	Incorrect Executions	Accuracy (%)	Avg. Response Time (s)
Simple OS Tasks	50	49	1	98	0.9
Complex Commands	50	44	6	88	1.5
Visual Output Commands	30	27	3	90	1.8
Overall	130	120	10	92	1.2

multistep automation in an operating system. Complex commands with the exception of processing the visual output for hologram conversion, the system successfully managed to perform basic tasks with near-perfect accuracy. The flow of a sample experimental session is shown in Fig. 11, which then shows the user's voice input to NLP processing, the execution of OS tasks, and, finally, hologram projection.

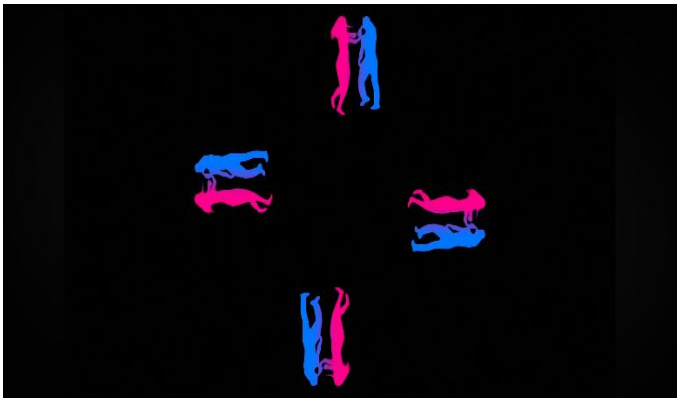

Fig. 11: Sample 2D visual frame generated for hologram conversion.

3.3 Results

The performance metrics for the VISTA.AI used in the quantification included the accuracy, precision, recall, and F1 score, which are described in Section 3.1 and the response times are summarized in Table 1.

On the basis of the test results, the VISTA.AI achieved a 92% F1-score with an overall accuracy of 92%, precision of 94%, and recall of 90%. The average response time for all the commands was 1.2 seconds, which reflects real-time performance. Compared with the 2D-to-3D conversion process, complex commands that needed the generation of holograms had response times that were somewhat longer; However, the output still appeared responsive and visually clear. Sample outputs of some command inputs are as follows: Categories include those shown in Fig. 12, highlighting hologram projection and execution of commands.

3.4 Discussion

It is evident from the results that the VISTA.AI combines automated OS, voice recognition, natural language processing, and hologram projection perfectly in a unified modulated system. The high accuracy and precision levels ensure that the system is reliable for executing the user's command accurately. High precision and accuracy demonstrate the system's ability to correctly interpret and execute user instructions. The recall value checks whether most of the commands are indeed identified and performed; however, errors were primarily noted when there was a lot of background noise or significant accent fluctuations.

The addition of a holographic visual output loads the interaction with a multimodal, immersive quality. Pepper's Ghost projection provides a clear and floating 3D visual that enhances user interaction and system perception. Although not true volumetric holography, the content of the source display greatly influences the brightness of the hologram; in order to avoid "ghosting" artifacts and guarantee high-contrast visuals, maximizing the brightness of the source

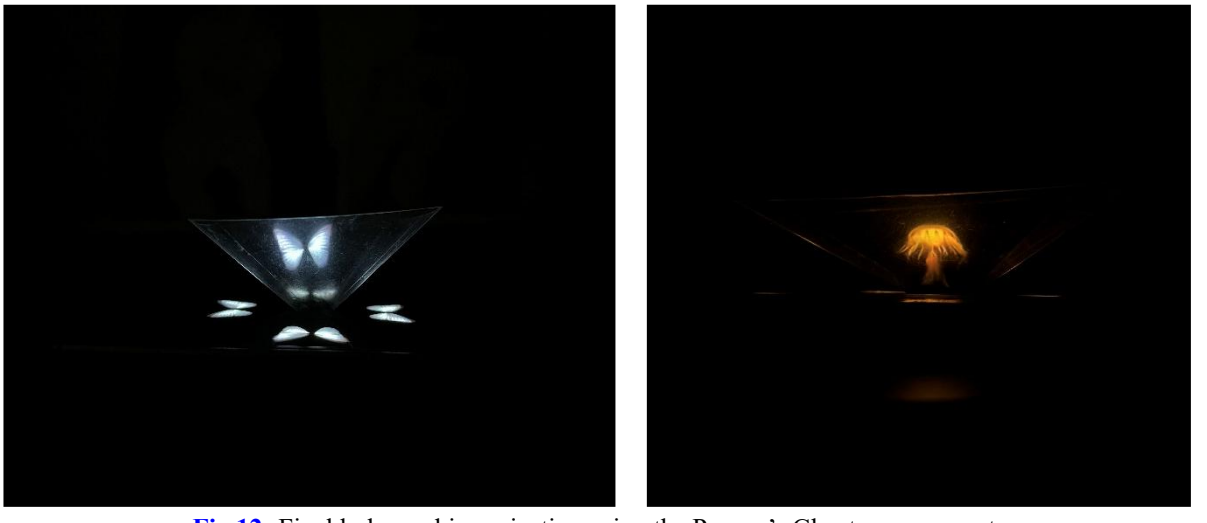

Fig 12: Final holographic projection using the Pepper's Ghost arrangement.

and minimizing ambient light is essential.[8] Future work can optimize the conversion pipeline using higher performance hardware to reduce minor limitations, such as a slight lag during complex visual command execution. Overall, through experimental analysis, with real-time operating system control, precise speech-based command execution can be obtained. In terms of execution and interactive 3D holographic feedback, AI is a competitive low-cost substitute for current AI assistants, bringing the system closer to a Jarvis-like experience.

4. Conclusion

The present work introduces the VISTA.AI, a low-cost, modular AI assistant offering holographic visual output combined with natural language processing, operating system automation, and voice recognition. It successfully uses a microphone to record user commands, interprets them using natural language processing algorithms, performs OS-level tasks, and utilizes text-to-speech. The method of synthesis coupled with the use of the prism-based Pepper's Ghost projection method provides both visual and audio feedback. In the experimental findings, the F1 measure and response time were observed to be 92% on average and 1.2 seconds, respectively. High accuracy, precision, and recall in command execution indicate real-time performance. In contrast to a typical AI assistant, the addition of 3D images with a focus on being hologram compliant has increased interactive interaction. The system does manage to complete the tasks and maintains visual clarity in low light, despite some shortfalls in performance drops under high background noise or complicated command sequences. Future versions of VISTA.AI can support more features, sensors, and cutting-edge display technologies because of its modular architecture. Overall, this work creates the foundation for immersive Jarvis-like AI assistants by fusing OS automation, multimodal interaction, and affordable holographic visualization. It opens the door for future developments in home automation. Virtual assistants and human–computer interactions can be identified by showing how AI-driven systems can go beyond voice interaction to visually interactive environments.

CRediT Author Contribution Statement

Gaurang Jagtap: Methodology; Formal analysis; Investigation; Writing – review & editing. **Siddhant Jawalekar:** Writing – review & editing. **Manasvi Khandwe**: Conceptualization; Writing – original draft. **Yukta Khushalani:** Methodology; Formal analysis; Investigation; Writing – review & editing. **Aditi Kolhapure:** Conceptualization; Writing – review & editing. **Vaishali Rajput:** Writing – review & editing, Validation, Visualization, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The experimental data generated during system evaluation and testing are available from the corresponding author upon reasonable request.

Conflict of Interest

There is no conflict of interest.

Artificial Intelligence (AI) Use Disclosure

The authors confirm that there was no use of artificial intelligence (AI)-assisted technology for assisting in the writing or editing of the manuscript and no images were manipulated using AI.

Supporting Information

Not applicable

References

[1] A. V. Shvetsov, S. H. Alsamhi, When holographic communication meets metaverse: Applications, challenges, and future trends, *IEEE Access*, 2024, **12**, 197488-197515, doi: 10.1109/ACCESS.2024.3514576.

[2] B. Dominguez-Dager, F. Gomez-Donoso, R. Roig-Vila, F. Escalona, M. Cazorla, Holograms for seamless integration of remote students in the classroom, *Virtual Reality*, 2024, **28**, 24, doi: 10.1007/s10055-023-00924-7.

[3] B. Shneiderman, Human-centered artificial intelligence: reliable, safe & trustworthy, *International Journal of Human–Computer Interaction*, 2020, **36**, 495–504, doi: 10.1080/10447318.2020.1741118.

[4] E. Luger, A. Sellen, Like having a really bad PA: The gulf between user expectation and experience of conversational agents, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 2016, 5286–5297, doi: 10.1145/2858036.2858288.

[5] H. Ahlawat, N. Aggarwal, D. Gupta, Automatic speech recognition: A survey of deep learning techniques and approaches, *International Journal of Cognitive Computing in Engineering*, 2025, **6**, 201-237, doi: 10.1016/j.ijcce.2024.12.007.

[6] S. H. Alsamhi, F. Nashwan, A. V. Shvetsov, transforming digital interaction: Integrating immersive holographic communication and metaverse for enhanced immersive experiences, *Computers in Human Behavior Reports*, 2025, **18**, 100605, doi: 10.1016/j.chbr.2025.100605.

[7] N. Sharma, K. Ali, A Review Paper on Holographic Technology Three-Dimensional Visualization, *International Journal of Engineering Research & Technology*, 2016, **4**, 1-3.

[8] Y. Zhao, Z. Xu, T.-Y. Zhang, M. Xie, B. Han, Y. Liu, Interactive holographic display system based on emotional adaptability and CCNN-PCG, *Electronics,* 2025, **14**, 2981, doi: 10.3390/electronics14152981.

[9] C. Chang, C. Zhao, B. Dai, Q. Wang, J. Xia, S. Zhuang, D. Zhang, Conversion of 2D picture to color 3D holography using end-to-end convolutional neural network, *PhotoniX*, 2025, **6**, 30, doi: 10.1186/s43074-025-00186-3.

[10] T. Yamaguchi, H. Yoshikawa, High resolution computer-generated rainbow hologram, *Applied Sciences*, 2018, **8**, 1955, doi: 10.3390/app8101955.

[11] A. Kumar, V. M. B, Advancements in holographic display technology: A comparative Analysis, *International Journal of Engineering Research & Technology*, 2023, 11.

[12] M. Lakshmi Pavani, A. V. Bhanu Prakash, M. S. Shwetha Koushik, J. Amudha, C. Jyotsna, Navigation through eye-tracking for human–computer interface, Information and Communication Technology for Intelligent Systems, Smart Innovation, Systems and Technologies, 2019, **107**, 2019, doi: 10.1007/978-981-13-1747-7_56.

[13] D. K. Saini, D. Ahir, A. Ganatra, Techniques and challenges in building intelligent systems: Anomaly detection in camera surveillance, Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, Smart Innovation, Systems and Technologies, 2016, **51**, doi: https://doi.org/10.1007/978-3-319-30927-9_2.

[14] T. A. Syed, M. S. Siddiqui, H. B. Abdullah, S. Jan, A. Namoun, A. Alzahrani, A. Nadeem, A. B. Alkhodre, A. B. In-depth review of augmented reality: tracking technologies, development tools, AR displays, collaborative AR, and security concerns, *Sensors*, 2023, **23**, 146, doi: 10.3390/s23010146.

[15] B. Chao, M. Gopakumar, S. Choi, J. Kim, L. Shi, G. Wetzstein, Large Étendue 3D Holographic Display with Content-adaptive Dynamic Fourier Modulation. In SIGGRAPH Asia 2024 Conference Papers (SA '24). Association for Computing Machinery, New York, NY, USA, 2024, **26**, 1–12, doi: 10.1145/3680528.3687600.

[16] P. Aswale, A. Shukla, P. Bharati, S. Bharambe, S. Palve, An Overview of Internet of Things: Architecture, Protocols and Challenges, Information and Communication Technology for Intelligent Systems, Smart Innovation, Systems and Technologies, 2019, **106**, 2019, doi: 10.1007/978-981-13-1742-2_29.

[17] V. Këpuska, G. Bohouta, Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home), 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2018, 99-103, doi: 10.1109/CCWC.2018.8301638.

[18] M. Thiyagesan, J. Govardhan, E. Dinesh Raj, K. Dhanush Kumar, N. Ahamed Basha, T. Magesh, Interactive voice-controlled 3D holographic display, *Turkish Online Journal of Qualitative Inquiry*, 2021, **12**, 2343-2352.

[19] H. Tyagi, V. Kumar, M. Danish, G. Agarwal, P. Mishra, Speech Recognition Intelligence System for Desktop voice Assistant by using AI & IoT, *International Journal of Intelligent Systems and Applications in Engineering*, 2023, **11**, 266-272.

[20] T. Jiang, Application and development prospect of AI speech recognition technology, Proceedings of the 3rd International Conference on Signal Processing and Machine Learning, 2023, 198-202, doi: 10.54254/2755-2721/6/20230766.

[21] M. Navyasri, R. RajeswarRao, A. DaveeduRaju, M. Ramakrishnamurthy, Robust Features for Emotion Recognition from Speech by Using Gaussian Mixture Model Classification, Information and Communication Technology for Intelligent Systems (ICTIS 2017), Smart Innovation, Systems and Technologies, 2018, **84**, 2018, doi: 10.1007/978-3-319-63645-0_50.

[22] A. Manolova, K. Tonchev, V. Poulkov, S. Dixit, P. Lindgren, Context-aware holographic communication based on semantic knowledge extraction, *Wireless Personal Communications*, 2021, **120**, 2307–2319, doi: https://doi.org/10.1007/s11277-021-08560-7.

[23] M. Vincze, B. Molnar, M. Kozlovszky, The use of voice control in 3D medical data visualization implementation, legal, and ethical issues, *Information*, 2025, **16**, 12, doi: 10.3390/info16010012.

[24] A. Guha, T. Bressgott, D. Grewal, D. Mahr, M. Wetzels, E. Schweiger, How artificiality and intelligence affect voice assistant evaluations, *Journal of the Academy of Marketing Science,* 2022, **51**, 312–330, doi: 10.1007/s11747-022-00874-7.

[25] P. Chakravarthula, Z. Zhang, O. Tursun, P. Didyk, Q. Sun, H. Fuchs, Gaze-contingent retinal speckle suppression for perceptually-matched foveated holographic displays, arXiv preprint arXiv:2108.06192, 2021.

[26] D. Pi, J. Liu, Computer-generated hologram based on reference light multiplexing for holographic display, *Applied Sciences,* 2021, **11**, 7199, doi: 10.3390/app11167199.

[27] A. Patil, M. Rane, Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition, Information and Communication Technology for Intelligent Systems (ICTIS 2020), Smart Innovation, Systems and Technologies, 2021, 195, doi: 10.1007/978-981-15-7078-0_3.

the ideas, methods, or products mentioned. GR Scholastic remains neutral regarding jurisdictional claims in published maps and institutional affiliations.

Citation

G. Jagtap, S. Jawalekar, M. Khandwe, Y. Khushalani, A. Kolhapure, V. Rajput, VISTA.AI: Voice-based interactive system for transformative assistance via holographic display, *Journal of Information and Communications Technology: Algorithms, Systems and Applications*, 2026, 2(1), 26302, doi: 10.64189/ict.26302.