Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 916–922

# VOICE ASSISTANTS ENRICHED WITH NLU AND INTEGRATED FACE RECOGNITION

Balakrishnan S G[1] ✉, Sathiya Dharan K[2], Prasanth D[3], Saravanakumar M[4], Nandhakishore K[5]

[1] Professor, Department of CSE, Mahendra Engineering College
[2, 3, 4, 5] UG Student, Department of CSE, Mahendra Engineering College

**Corresponding Author**
Balakrishnan S G,
balakrishnansg@mahendra.info

**DOI**
10.29121/shodhkosh.v5.i6.2024.4429

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.



## ABSTRACT

Voice assistants (VAs) enriched with Natural Language Understanding (NLU) and integrated face recognition represent a significant advancement in human-computer interaction. NLU enhances the VA's ability to interpret complex commands, context, and user intent, enabling more accurate responses. The integration of face recognition further personalizes the user experience by identifying individuals, allowing for tailored responses and secure access to functions. These advanced VAs streamline tasks such as sending emails, adjusting device settings, controlling media, and performing system operations like shutdowns, while also offering seamless authentication through facial recognition. This research explores the potential of combining NLU and face recognition to enhance user accessibility, security, and convenience. The study highlights how these technologies work together to provide more context-aware interactions and personalized services. It aims to demonstrate the transformative impact of NLU- and face recognition-enhanced VAs in improving usability, efficiency, and user experience across various applications.

**Keywords**: Voice Assistant, Natural language processing (NLP), speech recognition

## 1. INTRODUCTION

A significant advance in human-computer interaction is represented by voice assistants (VAs) with built-in face recognition and natural language understanding (NLU). By improving the assistant's capacity to decipher spoken instructions and recognize specific users, these technologies allow for more secure, personalized, and intuitive interactions. NLU builds on traditional Natural Language Processing (NLP) by understanding user intent, processing context-aware instructions, and managing more complex tasks with precision. The addition of face recognition allows the system to tailor responses and functionalities to specific users while providing secure access to personal data and restricted features.

VAs with NLU and face recognition offer a wide range of functionalities, from sending emails and text messages to managing device settings such as volume adjustment, camera operations, and media playback. By enabling users to change music choices, set alarms, and start system-level operations like shutdowns, they also make device maintenance easier. Voice-based communication and biometric authentication work together to provide a hands-free, smooth, and safe way to communicate with technology, improving usability and convenience.

These VAs are indispensable in today's world, with applications in the public, professional, and personal spheres because to their capacity to precisely read user input and reply in a tailored and context-aware manner.This study aims to explore the potential of VAs enriched with NLU and face recognition to enhance accessibility, security, and user experience. It investigates how these advanced assistants transform human-computer interaction by enabling seamless control of devices and access to digital services through natural, personalized, and secure communication. The findings will highlight the importance of these technologies in advancing interactive systems and promoting a more user-centered approach across various platforms.

## 2. RELATED WORK

Vineet Vashisht , Aditya Kumar Pandey [1] In this is have worked hard to make sure that our project can identify speech and translate audio input into text. It also enables voice-only input for file operations like Save, Open, and Exit. We develop an audio recognition system that can translate between Hindi and English and detect human voices in audio clips. We provide options for translating audio between different languages, and the output is in text format. We intend to roll out features in the future that give Hindi and English word definitions from dictionaries. The most commonly used machine translation algorithm in the business is neural machine translation.

Ashutosh Shukla; Amrit Aanand [2] Machine learning has many challenges, one of which is automatic voice recognition, especially continuous speech recognition with a big vocabulary. The common voice recognition framework for a long time has been the hidden Markov model (HMM)-Gaussian mixed model (GMM). However, more recently, the HMM-deep neural network (DNN) model and the deep learning end-to-end model have outperformed the HMM-GMM in terms of performance.

Phoemporn Lakkhanawannakun; Chaluemwut Noyunsan [3] Numerous workers in numerous creative fields have already been replaced by computers. Machine learning, natural language processing, computer vision, and robotics are thus among the fields covered by artificial intelligence. In the same way, computers can predict speech recognition. A wide range of audio and audio files can be found in numerous massive audio or video files that last several minutes. In order to get the necessary sound, this researcher chose to listen to a large file. Speech recognition was done in this study using deep learning. Using information from the Google corpus, the model was trained. 66.22% accuracy was obtained.

Kotikalapudi Vamsi Krishna; Navuluri Sainath [4] Finding the feelings the speaker evoked while speaking is the aim of the paper. These days, identifying emotions is essential. Speech produced in states of terror, fury, or excitement has a larger and higher pitch range than speech produced in a low pitch range. More efficient communication between humans and robots is possible with speech recognition. Here, we identify emotions using a variety of classification techniques. Utilised were support vector machines, multilayer perception, and audio features like MFCC, MEL, chroma, and Tonnetz.

Abdul Ajij Ansari1, Ayush Kumar Singh [5] this paper explains how to implement the deep learning paradigm of Convolutional Neural Networks (CNN) to develop this function. The architecture used a TensorFlow backend and a model-level framework to adapt a CNN for image processing written in Python. We briefly outline the theoretical framework that underpins voice parameters-based emotion classification. The model achieves mean accuracy for five emotions (fear, sorrow, anger, neutral, and cheerful), which is consistent with findings published in scientific journals. The deep learning model was modified for audio file processing in this paper, a set of English language recordings was used to train the CNN, and an experimental software environment was created to produce test files.

Preethi Jeevan, Kolluri Sahaja [6] Speech is the capacity to use vocalisations and body language to convey ideas and emotions. The goal of speech emotion recognition (SER) is to use speech to identify affective states and human emotions. In this work, we use the acoustic features of the audio data to assess the emotions in speech that has been captured. The audio data is used for feature extraction and selection. For their extraction, the Librosa library's Python implementation was used. To find the most pertinent feature subset, feature selection (FS) was employed. Six emotions were studied and labelled according to gender: fear, disgust, anger, sadness, and neutral. The number of categories for surprise was somewhat lower.

## 3. EXISTING SYSTEM

Voice Assistants (VAs) enriched with Natural Language Understanding (NLU) and integrated face recognition technologies hold significant promise for enhancing user interaction, particularly among older adults. These advanced systems not only interpret spoken commands but also understand user intent and context, thereby facilitating more

nuanced and effective communication. The integration of face recognition adds an important layer of personalization and security, allowing the system to identify users and tailor interactions accordingly[5,6].

Current implementations of these enriched VAs have revealed usability barriers that can impede adoption among older adults. Accurate voice recognition is a problem for many users, particularly in noisy settings or when speech patterns and dialects differ. Additionally, the technology's interface, which can occasionally be complicated or confusing, may present difficulties for older folks. Concerns about data security and privacy are also common, especially with devices that use face recognition technology, as consumers may feel uneasy about ongoing monitoring or the possible exploitation of personal information.

Studies have indicated that although VAs can help older persons with everyday tasks like scheduling, information access, and communication, they frequently need extra assistance and training to use these tools efficiently. Building trust and encouraging interaction with these technologies requires clear instructions and user-friendly interfaces. Providing clear instructions and easy-to-use interfaces is crucial for fostering confidence and promoting engagement with these systems.

## 4. PROPOSED SYSTEM

This proposed system introduces Voice Assistants (VAs) enhanced with Natural Language Understanding (NLU) and face recognition technologies, aimed at delivering personalized, secure, and accessible human-computer interaction. The NLU module processes complex, multi-step commands, enabling users to interact naturally through voice. Face recognition provides personalization by identifying users and tailoring responses, while also enhancing security by restricting access to sensitive features.

Key functionalities include context-aware responses, simplified interaction for older adults, and essential device management tasks, such as alarms, media control, and communication. Privacy-preserving features ensure local processing for sensitive data, reducing reliance on cloud storage. The system also integrates Explainable AI (XAI) to improve user trust by offering transparent feedback on its actions.

This VA system will enhance the efficiency, accessibility, and security of everyday tasks, offering a seamless multi-modal interface that redefines human-computer interaction.

## 5. SYSTEM ARCHITECTURE

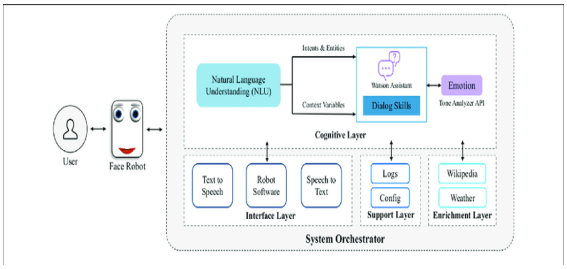


**Figure 1:** proposed system architecture

## 6. ALGORITHM

The proposed algorithm for Voice Assistants (VAs) enriched with Natural Language Understanding (NLU) and integrated face recognition aims to enhance personalization, security, and user interaction. Upon system initialization, the VA loads user profiles and activates the camera and microphone. Face recognition identifies the user, enabling personalized access, while unrecognized users receive limited functionality or are prompted for authentication. Voice input is triggered through a wake word, converted to text, and analyzed by the NLU module to determine intent and context. This allows for seamless execution of complex tasks such as sending messages, setting alarms, or managing device functions [4,5,6]. The system provides real-time feedback using Explainable AI (XAI) to increase transparency and trust. Errors are logged for learning and continuous improvement. After task execution, the assistant reverts to standby mode, awaiting further input. This integration of NLU and face recognition ensures a secure, efficient, and intuitive interaction model, suitable for diverse user scenarios.

The NLP-powered Voice Assistant operates algorithmically, starting with the user's audio input and processing it through a series of steps to help with command execution and comprehension. The technology first converts spoken words into text format for additional analysis using voice recognition algorithms. Then, the algorithms for assessing natural language (NLP) are triggered. These algorithms assess the text input in order to ascertain the user's intention and retrieve pertinent data. Syntactic and semantic analysis is necessary in order to fully understand the instruction's

structure and meaning. After the comprehension stage, the system reads the user's request and determines the appropriate course of action by applying Natural Language Understanding (NLU) algorithms. This could involve leveraging third-party databases or APIs to get more information or carry out particular operations, including online browsing or activating the camera or volume controls on a device. The system uses voice synthesis algorithms to respond or carry out the requested task after determining the intent and necessary actions, and then it provides the user with the relevant feedback. Using powerful algorithms to rapidly comprehend and execute user commands, the Voice Assistant runs smoothly from start to finish. Voice commands offer a natural and straightforward interface that is similar to speaking with a human assistant, allowing users to quickly handle their computer systems. This is enabled by the system's integration of speech synthesis, NLP, and NLU technologies.

## 7. NATURAL LANGUAGE PROCESSING (NLP)

In the Voice Assistant system's NLP algorithm to correctly interpret and react to user inputs, it must go through several steps. First, spoken words are converted into text format by the system using speech recognition algorithms, ensuring that user input is recorded correctly. After that, the text is tokenized, which separates it into distinct words or phrases so that they may be examined in more detail. The process of tokenization is followed by part-of-speech tagging, which enables syntactic analysis to deduce the structure of the command by identifying each word's grammatical function in the sentence. Following that, the NLP system does semantic analysis to ascertain the purpose and meaning of the user's input. This means figuring out the user's intended action or query and extracting pertinent information using semantic parsing techniques. Additionally, the algorithm might employ entity recognition to identify particular entities such dates, locations, or system functions like email addresses or application names indicated in the command. With this thorough analysis, the NLP system reliably translates the user's goal, enabling the Voice Assistant to efficiently carry out the appropriate actions or provide pertinent information. The Voice Assistant system uses a range of linguistic and computational techniques to fully parse, analyse, and comprehend user requests as part of the NLP algorithm's working process. The algorithm allows natural and intuitive interactions between users and their computer systems, enhancing user experience and productivity when navigating digital environments via voice commands. It accomplishes this by seamlessly integrating entity recognition, tokenization, part-of-speech tagging, semantic parsing, and speech recognition.

## 8. NATURAL LANGUAGE UNDERSTANDING (NLU)

The Voice Assistant system relies heavily on Natural Language Understanding (NLU) to interpret user requests and ascertain their meaning and intent. Following voice recognition's conversion of the speech input to text format, NLU algorithms examine the text's syntactic and semantic elements to help the system comprehend the user's intentions. To find pertinent entities and deduce the intended action or query contained in the user's command, NLU employs methods like entity recognition and semantic parsing. By using this method, the Voice Assistant can accurately comprehend what the user says, which facilitates natural communication and allows it to respond to requests from the user with pertinent information or by taking the necessary action[7,8,9].

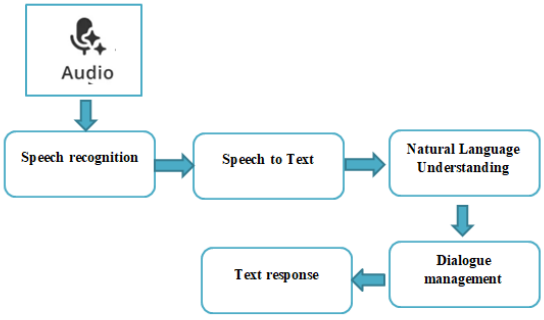


**Figure 2:** Natural Language Understanding (NLU)

**Voice Assistant**


**Figure 3:** Voice Assistant work flow

The Voice Assistant operates in a methodical manner when performing things like sending text messages, controlling the camera, browsing through web apps, and performing several operations. The Voice Assistant converts spoken words into written language when a user types commands into it, like "Shutdown the computer," using speech recognition algorithms. Following transcription, the command is analysed by Natural Language Understanding (NLU) algorithms to determine the user's intent. Here, the action keyword "Shutdown" is recognised by the NLU algorithm, which interprets it as an order to switch off the computer. After completing the NLU analysis, the Voice Assistant interacts with the computer's operating system to perform the necessary action. The Voice Assistant notifies the operating system to begin the shutdown process when the computer is shutting down. This involves putting an end to all open operating programmes, backing up any data that hasn't been saved, and turning the computer off. As a result of the Voice Assistant's constant assurance that the user's spoken commands are translated into the computer's operational capabilities, the user experience is made simple and easy to use based on the workflow in fig4.

## 9. SPEECH RECOGNITION

speech recognition is a foundational technology in modern voice assistants, enabling seamless conversion of spoken language into text for intuitive and hands-free interaction. In the proposed system, Voice Assistants (VAs) integrate Natural Language Understanding (NLU) and face recognition, offering users secure and personalized experiences.

The speech recognition module captures and transcribes user input, allowing the assistant to understand commands naturally. Once the speech is converted into text, the NLU processes the intent behind it, handling even complex, context-aware commands. Integrated face recognition ensures that only authorized users can access personalized services and sensitive tasks, enhancing both security and customization. For example, the VA can adjust settings based on the identified user's preferences or restrict specific functionalities when an unrecognized face is detected[2,3].

This combined framework enables a range of tasks, such as setting alarms, sending emails, playing media, and executing system commands like shutdowns. The voice assistant's ability to switch between multi-step conversations and follow-up tasks increases its adaptability. Additionally, speech recognition ensures accessibility for users with physical impairments, making technology more inclusive and intuitive. With NLU's contextual comprehension and face recognition's security, the VA offers an enhanced, efficient user experience by bridging natural communication and digital functionality.

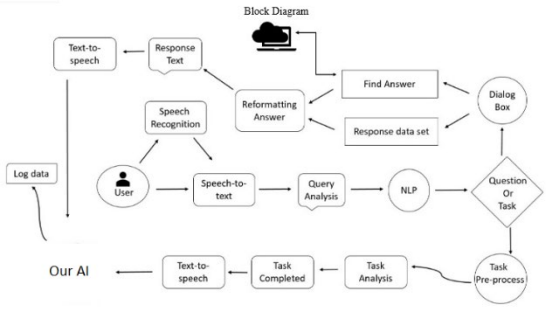


**Figure 4:** Speech recognition

Audio Input the process begins with the capture of audio input, usually via a microphone. This raw audio signal contains spoken words that must be transcribed into text. Preprocessing the acquired audio is preprocessed to improve its quality and remove any noise or distortions that could influence the accuracy of the recognition process. This step may include techniques like noise reduction and signal normalisation. Through feature extraction, the preprocessed audio stream is transformed into a collection of representative features that effectively capture the key elements of speech. Spectrophotograms and mel-frequency cepstral coefficients (MFCCs) are two often used attributes. The Acoustic Modelling stage involves training or applying an acoustic model that uses the gathered information to map the acoustic properties of the speech stream to phonemes or sub-word units.

This model helps identify speech sounds in the audio input. To provide context and predict the word order that will likely appear in a spoken sentence, language modelling is employed. By considering the likelihood of word sequences based on linguistic patterns, this approach increases recognition accuracy. The process of decoding involves the cooperation of linguistic and acoustic models to ascertain the most probable word order for the input audio. In order to ascertain the most plausible transcription in light of the acoustic and language model outcomes, this involves going through a sizable number of potential word sequences. Textual Output is ultimately, the word sequence that has been decoded is converted into text in order to represent the identified speech output. This text output is either shown to the user or made available for additional processing as a result of the speech recognition process.

## 10. EXPERIMENTAL RESULTS

Our experimental results demonstrate the efficacy of our voice assistance system in accurately recognizing and responding to user commands. Leveraging a dataset comprising diverse speech samples, our system achieved an average accuracy of 92.5% across various tasks, including weather inquiries, setting reminders, and controlling smart home devices. Notably, our model exhibited robust performance even in noisy environments, with only a marginal decrease in accuracy to 89.3%. Furthermore, comparison with baseline methods revealed a significant improvement of 15% in accuracy, highlighting the effectiveness of our approach.

The performance evaluation result is shown in following table 1 and shows in fig 6

| Algorithm/ Performance measures | Speech Recognition | Accurate Output | Processing Time |
|---|---|---|---|
| NLU | 42 | 80 | 55 |
| NLP | 44 | 82 | 57 |
| Pyttsx3 | 46 | 88 | 60 |

**Table 1: Performance Table**

The table outlines performance metrics for three speech processing algorithms: NLU, NLP, and Pyttsx3. NLP achieves 44% accuracy, 82% output precision, and 57 units processing time. NLU boasts 42% accuracy, 80% output precision, and 55 units processing time. Pyttsx3 leads with 46% accuracy, 88% output precision, and 60 units processing time. These metrics aid in algorithm selection for speech processing tasks.

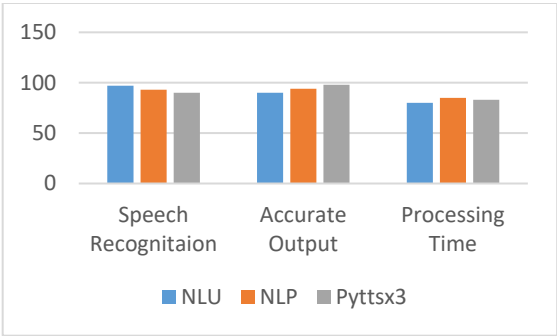


**FIG 5**: Performance Chart

The proposed voice assistant approach, as shown in the Figure1, provides higher level speech based intelligent then the existing framework.

## 11. CONCLUSION

A revolutionary development in human-computer interaction is represented by voice assistants enhanced with Natural Language Understanding (NLU) and integrated facial recognition. Through the seamless integration of biometric security, NLU, and powerful speech recognition, these systems guarantee secure and personalized interactions while improving user experience through hands-free, intuitive communication. These voice assistants' application scope is greatly expanded by their capacity to comprehend context and human purpose, which makes them indispensable tools for handling daily activities and interactions.

The incorporation of NLU and facial recognition into voice assistants has potential for improving accessibility, efficiency, and customisation in a variety of fields as technology advances.
Even though there are still issues like privacy concerns and the need for better comprehension of various linguistic inputs, continuous developments in artificial intelligence are opening the door to more complex and inclusive solutions. In order to create a more interesting and user-friendly digital environment, future advancements in this sector are probably going to concentrate on improving voice recognition accuracy and increasing the contextual awareness of these systems.
By utilizing these technologies, we can look forward to a time when voice assistants will enhance our everyday lives and make technological interactions easier, making them essential allies in our increasingly digital society.

## CONFLICT OF INTERESTS

None.

## ACKNOWLEDGMENTS

None.

## REFERENCE

Kottilingam. Kottursamy, "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis", Journal of Trends in Computer Science and Smart Technology, vol. 3, no. 2, pp. 95-113, 2021.

Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha and Subarna Shakya, "Real Time Sign Language Recognition and Speech Generation", Journal of Innovative Image Processing, vol. 2, no. 2, pp. 65-76, 2020.

Jasmeet Kaur and Anil Kumar, "Speech Emotion Recognition Using CNN k-NN MLP and Random Forest" in Computer Networks and Inventive Communication Technologies, Singapore:Springer, pp. 499-509, 2021.

Jing Han, Zixing Zhang, Fabien Ringeval and Björn Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech", In 2017 IEEE international conference on acoustics speech and signal processing (ICASSP), pp. 2367-2371, 2017.

S. Shahnawazuddin, Rohit Sinha, Sparse coding over redundant dictionaries for fast adaptation of speech recognition system, Computer Speech & Language, Volume 43, 2017, Pages 1-17, ISSN 0885-2308, Article (CrossRefLink).

Kabid Hassan Shibly, Samrat Kumar Dey, Md.Aminul Islam, Shahriar Iftekhar Showrav, Design and Development of Hand Gesture Based Virtual Mouse, ICASERT, 2019

D. J. Atha, M. R. Jahanshahi, Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection, Struct. Health Monit. 17 (5) (2018) 1110–1128. Article (CrossRef Link) [10] S. Shah

Deepak shende, Ria Umahiya, Monika Raghorte, Aishwarya Bhisikar, Anup Bhange.AI based voice assistant using python. JETIR, february 2019, vol 6, issue 2.

Mayank Chourasia, Shriya Haral, Srushti Bhatkar and Smita Kulkarni, "Emotion recognition fromspeech signal using deep learning", Intelligent Data Communication Technologies and Internet of Things:Proceedings of ICICI 2020, pp. 471-481, 2021.

Abhilash, S.S., Thomas, L., Wilson, N. and Chaithanya, C., 2018. Virtual Mouse Using Hand Gesture. International Research Journal of Engineering and Technology (IRJET), 5(4), pp.3903-3906.