A

Project Report

on

AI Assistant

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2023-24

in

COMPUTER SCIENCE AND ENGINEERING

By

Saksham Pandit (2000290100128)

Rishika Gupta (2000290100117)

Sajal Gupta (2000290100126)

Under the supervision of

Prof. Umang Rastogi

KIET Group of Institutions, Ghaziabad

Affiliated to

Dr. A.P.J. Abdul Kalam Technical University, Lucknow

(Formerly UPTU)

May, 2024

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:	Signature:
Name: Saksham Pandit	Name: Rishika Gupta
Roll No.: 2000290100128	Roll No.: 2000290100117
Date:	Date:

Signature:

Name: Sajal Gupta

Roll No.: 2000290100126

Date:

CERTIFICATE

This is to certify that Project Report entitled "**AI Assistant**" which is submitted by **Saksham Pandit, Rishika Gupta and Sajal Gupta** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, Lucknow is a record of the candidates own work carried out by them under my supervision. The matter embodied in this report is original and has not been submitted for the award of any other degree.

.

Prof. Umang Rastogi	**Dr. Vineet Sharma**
Assistant Professor	**(Head of Department)**

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to **Prof. Umang Rastogi**, Department of Computer Science & Engineering, KIET, Ghaziabad, for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. Vineet Sharma, Head of the Department of Computer Science & Engineering, KIET, Ghaziabad, for his full support and assistance during the development of the project. We also do not like to miss the opportunity to acknowledge the contribution of all the faculty members of the department for their kind assistance and cooperation during the development of our project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members, especially **Prof. Gaurav Parashar**, of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:

Name: Saksham Pandit

Roll No.: 2000290100128

Date:

Signature:

Name: Rishika Gupta

Roll No.: 2000290100117

Date:

Signature:

Name: Sajal Gupta

Roll No.: 2000290100126

Date:

ABSTRACT

Virtual voice assistants have transformed the way humans interact with computers, especially with mobile devices. This study examines the latest advancements in speech processing technologies to create a virtual voice assistant for desktop users. We examine the latest progress in speech recognition, natural language processing, and dialogue control. Important factors to consider are precision in quiet environments, compatibility with current desktop processes, and customization that suits user choices. We examine the possible advantages and obstacles of this approach, as well as identify areas for future research. This study proposes to connect mobile and desktop voice assistants using modern speech processing technologies, providing users with a smooth and effective method to control their computers through voice command.

Ensuring precision in quiet environments and compatibility with existing desktop processes are crucial considerations for seamless integration. Additionally, offering customization options tailored to user preferences can enhance user satisfaction and adoption.

Identifying potential advantages, such as hands-free operation and increased accessibility, along with obstacles like background noise interference and privacy concerns, will provide valuable insights into the feasibility and implementation of desktop voice assistants.

Connecting mobile and desktop voice assistants holds immense potential for streamlining tasks and enhancing user experiences across different computing platforms. I'd be interested to learn more about the specific methodologies and findings of your study.

TABLE OF CONTENTS

Page No.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ANN - Artificial Neural Network

CNN - Convolutional Neural Network

GMM - Gaussian Mixture Model

HMM - Hidden Markov Model

HTK - Hidden Markov Model Toolkit

MLP - Multilayer Perceptron

RBM - Restricted Boltzmann Machine

RNN - Recurrent Neural Network

SGD - Stochastic Gradient Descent

CHAPTER 1

INTRODUCTION

1.1 Introduction of the problem

Today as we are developing with technology and the notion of Artificial Intelligence, we are making our living simple and less complicated. One such clever use of this growing technology of artificial intelligence is voice assistant. Nowadays we are not even utilizing our fingers to type or touch. We merely talk about the tasks, and it gets done by the virtual assistant. Today we may talk to our devices exactly as we talk to any human to execute the duty. Software applications known as voice assistants are designed to recognize and understand spoken instructions in natural language and carry out the tasks that the user specifies. This virtual assistant program promotes user productivity by handling the everyday chores of the user and by presenting information from internet sources to the user. Today people are no longer surprised when they talk to their virtual assistant as it has become their part of lives and is always available with them on their mobile phones or other devices. When Apple Corporation debuted its Siri, it became everyone's focus of attention. This virtual assistant given by the Apple firm became a source of entertainment for many. Then the arrival of Google Assistant in Android phones allowed Android users an opportunity to experience the same. Sooner or later virtual assistants became a significant software and feature of smartphones. We may construct our own customized virtual assistant with the aid of Python programming language. Python includes several inherent libraries, modules, and packages, which help the building of the tailored and personalized virtual assistant with essential features and functions. With the rise in artificial intelligence which contains components like natural language processing and machine learning, virtual assistants are evolving out smarter than before and the work done by them is also growing better and more precise. It would not be inaccurate to state that it is all about allowing our Virtual assistant to work for us, select information, and produce a decent answer. The primary purpose of this project is to build software that will be able to serve individuals like a personal assistant. This program attempts to produce a virtual assistant for Windows-based platforms. The goal of this application software is to accomplish and execute the user’s duties for instructions, supplied in either voice or text. It will ease the task of the user. In this project, we have designed a virtual assistant ZEN which would save time for the consumers. It simplifies human life by enabling users to operate PCs or laptops using just voice instructions. Voice Assistants take up less time.

AI assistants are utilized across various domains and industries to streamline tasks, enhance productivity, and improve user experiences. In the realm of personal assistants, they assist users in managing schedules, setting reminders, making reservations, and providing personalized recommendations. In customer service, they facilitate efficient communication, offer automated support, and handle inquiries promptly. In healthcare, AI assistants aid in diagnosing ailments, monitoring patient progress, and providing medical information. Moreover, they assist in automating repetitive tasks in industries like finance, manufacturing, and transportation. Overall, AI assistants contribute to simplifying processes, increasing efficiency, and augmenting human capabilities in diverse fields.

1.2 Objective

The objectives of AI assistants encompass a multifaceted approach aimed at enhancing user experiences and solving a myriad of problems across diverse domains. These assistants strive to streamline tasks and processes, boosting efficiency by automating routine activities and freeing up human resources for more complex endeavors. Personalization is a key focus, with AI assistants leveraging data analysis and machine learning algorithms to deliver tailored content, recommendations, and interactions based on individual user preferences and context. They serve as knowledge repositories, providing users with access to vast amounts of information and resources, while also facilitating seamless communication through natural language processing and conversation capabilities. In customer service, AI assistants aim to improve support experiences by offering timely and accurate assistance, ultimately leading to higher satisfaction levels and increased brand loyalty. Moreover, they assist users in managing tasks, projects, and workflows, leveraging context awareness to anticipate needs and provide relevant recommendations. Continuous learning and improvement are fundamental, as AI assistants evolve over time through user interactions, feedback, and data analysis, ensuring they remain relevant and effective. Privacy and security are paramount, with AI assistants implementing robust measures to safeguard sensitive information and uphold user trust and confidence. Overall, the objectives of AI assistants revolve around enhancing productivity, personalization, and problem-solving across various domains through intelligent automation and continuous innovation.

1.3 Scope

The future potential of voice assistants is very promising and advancing quickly. Voice assistants have made substantial progress in smart homes, customer service, healthcare, education, and other areas and will soon become a crucial part of our lives.

There is a lot of scope for enhancing and upgrading this project There can be space for future scope for this project. We can implement a chatbot that can give the project a complete look. Also, so far this project is limited to desktop users so we can extend this project to smartphones or maybe some other gadgets as well. Also, ZEN is a customized virtual voice assistant that has limited functionalities so we can add on more functionalities to perform other tasks of users. After upgrading, enhancing, and extending this project, ZEN will become very efficient in making the lives of users comfortable as it will be performing their tasks effortlessly.

1.4 Project Description

Project ZEN is an innovative venture harnessing the power of artificial intelligence, particularly focusing on natural language processing and speech recognition, all implemented within the Python programming language. The project's core functionality revolves around interpreting user commands provided via voice input, processing them, and executing the desired actions.

At its heart, ZEN comprises several integral components:

Speech Recognition Module: This pivotal Python module facilitates the conversion of spoken words into text, enabling seamless interaction with the system. Leveraging the Recognizer class, the module listens to voice input and converts it into text format for further processing. Additionally, it utilizes a Microphone class to access the device's microphone, allowing for real-time audio input. The module is tailored to recognize Indian English, ensuring compatibility with diverse user bases.

Python Backend: Serving as the backbone of the project, the Python backend receives the text output generated by the speech recognition module. It processes this input, determining whether it necessitates a system call, an API call, or content extraction. The backend orchestrates the subsequent actions based on the user's request, orchestrating the flow of information within the system.

API Calls: ZEN leverages Application Programming Interfaces (APIs) to seamlessly connect with external applications or services. This enables the transmission of user requests to pertinent suppliers, subsequently retrieving and presenting the desired results to the user.

Content Extraction: Employing natural language processing techniques, content extraction entails extracting organized information from unstructured or semi-structured machine-readable resources. This pivotal activity involves processing human language documents to extract relevant data, enhancing ZEN's ability to provide accurate and contextually relevant responses.

System Call: Facilitating interaction between the program and the operating system's kernel during runtime, system calls enable ZEN to request essential services crucial for its operation.

Text-to-Speech Module: An essential component for user interaction, the text-to-speech module converts textual responses generated by ZEN into audible speech. Leveraging libraries like Pyttsx3, this module ensures a seamless user experience by rendering synthesized speech output. This component plays a vital role in conveying information and responses back to the user in a natural and comprehensible manner.

To augment its capabilities, ZEN integrates several foundational libraries and packages, including:

SpeechRecognition: Enabling the identification and translation of human speech into text format.

Pyttsx3: A versatile offline module facilitating text-to-speech conversion, enhancing ZEN's offline functionality.

Wikipedia: Empowering ZEN to access and manage Wikipedia requests, providing users with informative responses.

OS: Streamlining various operating system tasks, enhancing ZEN's automation capabilities.

Web browser: Facilitating seamless interaction with the system's default web browser for retrieving and presenting information.

Pywhatkit: Simplifying interactions with the browser, enriching ZEN's browsing capabilities.

Project ZEN represents a comprehensive integration of cutting-edge AI techniques and Python programming prowess, aiming to deliver a sophisticated and intuitive user experience through voice-driven interaction and intelligent task execution.

CHAPTER 2
LITERATURE REVIEW

Currently, a wide range of Smart Personal Digital Assistant applications are emerging in the market for various device plat-forms. And these new software applications perform significantly better than PDAs because they incorporate every feature of a smartphone. Additionally, VPAs are more dependable than personal assistants because they are portable and usable at any time. They have more information than any assistant since they are internet-connected [1].

The authors assert that although voice assistants currently have limited capabilities, this will soon change, as they make strides into space exploration and basic medicinal procedures. It transpires that the voice assistant's capabilities will expand beyond resolving basic user needs to encompass more intricate and financially burdensome duties, the execution of which will necessitate the ongoing education of artificial intelligence [2].

Adrian et al. proposed a solution that allows elderly individuals to track their daily physical activity using virtual voice assistants, IoT devices, and activity-monitoring smart bracelets. This enables the elder people to avoid developing sedentary habits by simply using their voice. Their endeavour, designated EMERITI sought to enhance the quality of life for the elderly by employing virtual assistants across various case studies [3].

Po-Sheng et al. [4] concentrated their study on the creation of a Deep Neural Network (DNN)-based campus virtual assistant. This research is presented in App format, which is economical and simple to use. The system offers a straightforward voice response interface, obviating the necessity for users to navigate intricate web pages or app menus in search of information.

The survey conducted by Peng et al. [5] delineates research domains characterized by a comparatively comprehensive understanding of the threat but a dearth of effective countermeasures, such as concealed vocal commands. It also addresses the work that examines the privacy implications, including research on the administration of consent recording. The purpose of this survey is to compile an all-encompassing study plan concerning the security and privacy of PVAs.

The authors of [6] discussed about voice assistant interface interaction with the BIM model from a distance has been enabled. Individuals with visual impairments can access and augment BIM models. BIM novices are capable of practicing BIM features and retrieving information with minimal skill.

According to a study by Atieh [7], voice engagement with a VA that combines sincerity, creativity, and intelligence allows users to assert control over their speech interactions with the Virtual Assistant, focus on the voice interaction, and engage in exploratory behaviour The exploratory behaviour of consumers results in their continued use of voice assistants.

A survey by Malik et al. [8] discussed and examined voice recognition methods. An ASR depends on three modules: feature extraction, classification, and language model, according to its fundamental design. Analysis of classification models shows HMM performed best. The addition of a language model may considerably affect ASR accuracy. Even when sub-optimal approaches are employed to develop language models, further research will enhance voice recognition.

S. Raju of [9] designed methods that use probability theory, pattern machine, and now deep learning. Because they function alone, these methods lack literal meaning and context. The context must accompany the translation. Time series context construction is difficult. Due to dynamic inputs, context may change. Audio corpus association with learned vectors is used for similarity, missing data, and prediction. Sequence modelling using diverse algorithms must enhance voice recognition. Language sequence modelling helps clarify ambiguous words. Additional emotions may improve context and processing.

Ali et al. [10] used data from 174 publications published between 2006 and 2018, this research conducted a statistical analysis of deep learning in voice applications. The bulk of publications explored voice recognition. The study’s databases were in English. Most research evaluated system efficiency using WER (word error rate). Most deep learning researchers extract voice features using MFCCs. HMM and GMM significantly utilized MFCCs. Many researchers considered Linear Predictive Coding for feature extraction in deep learning models. They found that Authors should utilize hybrid models since research shows that DNN models with HMM or GMM information perform better.

Singh et al. [11] surveyed that, HTK is the most popular toolkit used Indian language ASR. Researchers now often utilize Kaldi to create systems. An extensive literature review shows that not many Deep learning methods have been used to test much of ASR. The scarcity of big-voice corpora in multiple languages is the reason. The exploratory work on feature extraction is also confined to a few popular languages and most speech signal classification investigations employ HMM-GMM. Alam et al.

[12] reviewed cutting-edge DNN algorithms and architectures for vision and speech and found that RNN models dominate voice recognition systems, notably in NLP applications.

Al-Fraihat1 et al. [13], surveyed that hybrid DNN models are being used because they perform better than stand-alone models.

O'Shaughnessy [14] discussed that the MLP structure is the foundation of an ANN, while SGD search is the usual training method. Many enhancements to these fundamental approaches have taken use of additional computer power and large data volumes and thus ANN models remain opaque and hard to comprehend but proper structuring may enhance ANNs' power.

Abde et al. [15] proposed an innovative technique to use CNNs for voice recognition that directly accommodates some speech variability. In our hybrid CNN-HMM technique, the HMM handles temporal variability, while convolving along the frequency axis invariantly handles minor frequency changes caused by speaker variances in speech data. Song et al.

Song et al. [16] focused on the application of deep learning in acoustic features and speech attributes and proposed a deep speech-based English speech recognition algorithm that combines multiple features. The CNN–RBM–ASAT algorithm proposed in this paper has much higher accuracy than CNN and RBM algorithms, so combining the two can improve accuracy.

Bell et al. [17] meta-analysis shows that adaption techniques work for hybrid and E2E systems across corpora and classes. However, unsupervised and semi-supervised E2E system training utilizing uncertainty propagation techniques remains a major research problem. We have summarized the findings and limitations in the table I and summarized in figure 1.

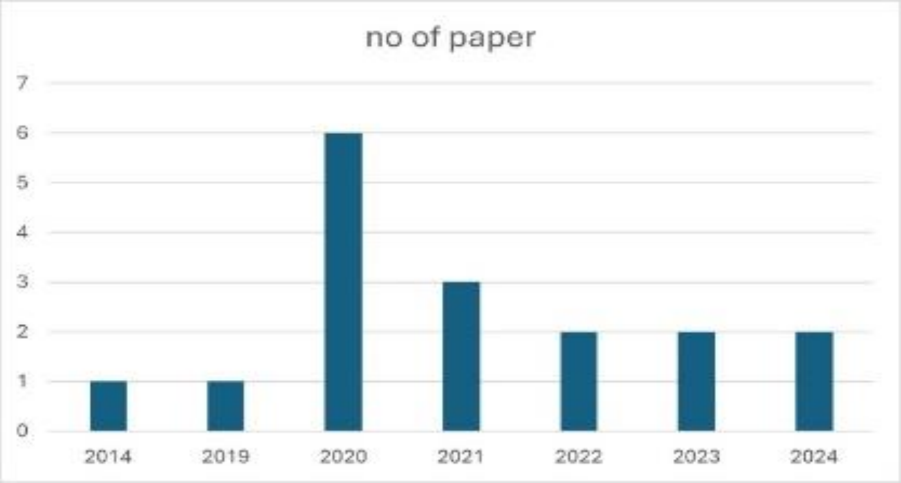


Fig 2.1. number of research papers referred according to the year distribution

Reference	Findings	Limitations
[8]	ASR is dependent on three modules: the feature extraction module, the classification module, and the language model. Among classification models, HMM performed the best	Language model helps to improve the accuracy of an ASR, but only sub-optimal methods are available.
[9]	Speech Recognition methods use probability theory, pattern machine and now deep learning	These techniques function independently and therefore they miss the literal meaning and context associated with it. Due to dynamic inputs, context may change. Audio corpus association with learned vectors is used for similarity, missing data, and prediction
[10]	Most articles employed WER (word error rate) to evaluate system efficiency. Most researchers employ MFCCs to extract voice features in deep learning models. Classifiers like HMM and GMM employed MFCCs extensively. When utilizing deep learning models	The power of RNN models, notably Long Short-Time Memory (LSTM), in voice recognition makes deep RNN highly accurate but Speech Recognition using RNNs is understudied
[11]	The most popular toolkit used	Deep learning methods have not been used

	for Indian language speech recognition is HTK (Hidden Markov Model Toolkit) whereas Kaldi is used for system building.	extensively in ASR systems for Indian languages. Because there is a scarcity of multilingual speech corpora. Feature extraction experiments are also confined to popular languages.
[12]	Reviewed current DNN algorithms and structures for vision and voice applications. For NLP applications RNN models were widely used in voice recognition systems.	Three fundamental model limitations: the risks of employing limited datasets mobile device hardware limits, and overoptimizing intelligent algorithms to substitute human experts.
[13]	Found that hybrid DNN models out-perform stand-alone models and are being used more.	Transformational models can parallelize, learn quicker, and perform better for low-resource languages, yet there is a research gap in voice recognition utilizing transformer model
[14]	Fundamental MLP is the fundamental ANN structure, while SGD search is the typical training method.	The methodology was rejected due to the failure of ES (expert system) to manage speech fluctuation and no mistake correction feedback.
[15]	Explained how to use CNNs for voice recognition in a unique approach that directly accommodates some speech variability and for this, they used a hybrid CNN-HMM technique	CNNs pre-trained with convolutional RBMs performed better in large vocabulary voice search but not phone identification. This disparity needs more study.
[16]	The CNN-RBM-ASAT algorithm proposed in this paper has much higher accuracy than CNN and RBM algorithm, so combining the two can improve the accuracy.	To generate a novel network feature extraction notion, a clustering technique before feature extraction and screening the features need to be included.
[17]	Used adaption techniques for hybrid and E2E systems across corpora and classes.	There is a need to use uncertainty propagation approaches for unsupervised and semi-supervised E2E system training.

Table 2.1. Findings and limitations of papers

CHAPTER 3

PROPOSED METHODOLOGY

Our entire project ZEN is designed using several techniques of artificial intelligence like Natural language processing and speech recognition written in Python programming language. Python offers a plethora of specialized built-in libraries and packages to carry out the tasks that the user inputs. The user's voice, which is a list of actions that the user wants completed, is the data used in this project. Whenever the user gives the voice command as the input, the speech recognition module comes into effect it takes the voice as an input listen to the words of voice input identifies them with its capability, and converts spoken words into text which is further spoken by Zen, this spoken words becomes the output voice and the task is done by ZEN. The steps below show how the command is taken up by the assistant.

The library that we have named ”Recognizer” recognizes the command, and as a result, voice is converted to text. They should also be separated from the surrounding text by one space.

After the command is transformed into a query, it will identify the words in the sentence, look up the keywords that match its condition, and execute the function by the specified condition.

3.1. Speech Recognition Module

This Python module aids in the conversion of speech to text. This module automatically receives the voice input. The identical text is received and sent to the central processing unit. This enables us to transform audio into text for additional processing. We have imported a speech recognition module as “Sr”. The Recognizer class inside the speech recognition module lets us recognize the audio. The same module contains a Microphone class that offers access to the microphone of the device. So, using the microphone as the source, we attempt to listen to the audio using the listen () function in the Recognizer class. We have also set the pause threshold to 1, that is it will not complain even if we halt for one second while we talk. We have set the language to Indian English. It returns the transcript of the audio which is nothing but a string. We’ve put it in a variable named query. The speech input Texts from the distinct corpora arranged on the PC may be sent to users.

3.2. Python Backend

This aids in providing the user with the necessary output. The output generated by the voice recognition module is sent to the Python backend, which also determines if the output is a system call, an API call, or context extraction. Also, this component output is fed to the Text-to-speech module.

3.3. API Calls

API means Application Programming Interface. It helps to connect two applications and transmits user requests to the supplier, who then returns the result to the user.

3.4. Content Extraction

The process of extracting organized information from un-structured or semi-structured machine-readable resources is known as content extraction. This activity mostly involves using natural language processing (NLP) to process documents written in human languages. It all comes down to pulling pertinent and related data from the webpage.

3.5. System call

This facilitates the computer program’s request for a service from the operating system’s kernel while it is running.

3.6. Text-to-speech module

A text-to-speech engine is needed for a text-to-speech module. Written text can be converted into waveforms that aid in producing sound using a text-to-speech engine. And these are the next essential and fundamental libraries that support ZEN in doing the job.

Speech recognition. The foundation of this project is this library. This is used to identify human speech and translate it from input voice to text.

The assistant ZEN is composed of the components mentioned in Table II.

S.NO.	Components
1	Speech Recognition Module
2	Python Backend
3	API Calls
4	Content Extraction
5	System Calls
6	Text-to-Speech Module

Table 3.1. Table depicting the number of components used in ZEN.

These are the following libraries of python that play a key role in functioning of the Zen

3.7. Pyttsx3

This offline module is a significant resource. This module alone contains the run and wait functions as well. It specifies the time interval between inputs, or more precisely, how long the system will wait for the next input. This has the primary advantage of operating offline.

3.8. Wikipedia

It is an online Python library that needs to be connected to the internet to function. With the aid of this library, Zen may manage Wikipedia requests and provide answers.

3.9. OS

This library helps with a variety of operating system tasks that can be done automatically. This library offers functions for generating and erasing directories (folders), retrieving their contents, updating folders, and more.

3.10. Web browser

The platform that this library offers the system’s default web browser is helpful. Users must provide a filename or URL to operate with this library, and the output is then shown in the browser.

3.11. Pywhatkit

Utilizing this library is an absolute breeze, as it is designed to simplify any interaction with the browser.

3.12. DFD Level-1 Diagram

A Data Flow Diagram (DFD) Level 1 Diagram is a graphical representation of a system that illustrates the flow of data between external entities and processes within the system. At this level of abstraction, the diagram provides a high-level overview of the system's functionalities and interactions. It typically depicts the main processes or functions of the system as bubbles or circles, with arrows representing the flow of data between them. External entities, such as users or other systems, are shown interacting with the system through input and output data flows. Each process represents a transformation or manipulation of data within the system, and the data flows illustrate the movement of data between processes, external entities, and data stores. The DFD Level 1 Diagram helps stakeholders understand the overall structure of the system and its major components, laying the foundation for more detailed analysis and design at lower levels of abstraction.

A Data Flow Diagram (DFD) Level 1 Diagram consists of several key elements that together illustrate the flow of data within a system. At its core are external entities, depicted as squares or rectangles, which represent sources or destinations of data outside the system. Processes, represented as circles or ovals, denote the functions or activities performed within the system, transforming input data into output data. Data flows, depicted as arrows, signify the movement of data between processes, external entities, and data stores, while data stores, depicted as rectangles or cylinders, represent repositories of data within the system. Optionally, control flows, depicted as dashed lines or arrows, can indicate the sequence of operations or decision logic within the system. By visually representing these elements and their interactions, the DFD Level 1 Diagram provides a structured overview of the system's functionality and data flow, facilitating communication and comprehension among stakeholders.

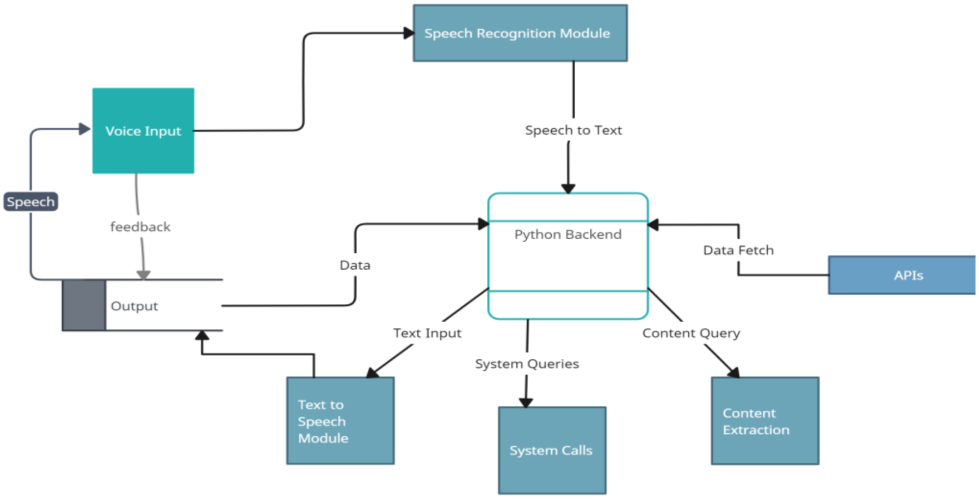


Figure.3.1 DFD level 1 diagram

3.13. Sequence Diagram

A sequence diagram is a visual representation in Unified Modelling Language (UML) that depicts the interactions between objects or components within a system over time. It provides a dynamic view of how these entities communicate and collaborate to accomplish specific tasks or functionalities. In a sequence diagram, objects are represented as vertical lifelines, and messages or method calls between them are depicted as arrows, illustrating the sequence of events as they occur. Activation bars show the periods during which objects are active, and optional elements like loops and conditions can be included to illustrate complex interactions. Sequence diagrams are valuable tools during the design phase of software development, helping developers and stakeholders understand the flow of control and communication within a system, identify potential issues, and refine system design to ensure efficient and effective implementation.

3.14. Algorithm

1. Receive Input: The AI assistant starts by receiving input from the user, typically in the form of text or voice commands.

2. Understand Input: It then analyses the input to understand what the user is asking or requesting. This involves natural language processing (NLP) to decipher the meaning of the text or speech.

3. Identify Intent: Once the input is understood, the AI assistant identifies the user's intent or what action needs to be taken based on the input. For example, if the user asks for the weather forecast, the intent is to provide weather information.

4. Retrieve Information: The assistant retrieves relevant information from its knowledge base or external sources. This could involve accessing databases, APIs, or the internet to gather the necessary data.

5. Process Information: After gathering the information, the AI assistant processes it to ensure relevance and accuracy. This might involve filtering out irrelevant data or performing calculations.

6. Generate Response: Using the processed information, the assistant generates a response to the user's query or request. This could be in the form of text, speech, or visual output, depending on the interface.

7. Deliver Response: Finally, the AI assistant delivers the response to the user through the appropriate channel, such as text on a screen or spoken words through a speaker.

CHAPTER 4

RESULTS AND DISCUSSION

When we start up the assistant ZEN it greets the user by saying “Hello Sir I am ZEN. How may I help you?”. And it waits for the user to command tasks. Following are the tasks that ZEN can perform when the user commands it.

4.1. WhatsApp Automation

We must state first that we are sending a “WhatsApp message” It will match the keyword and call the WhatsApp function. In WhatsApp automation, “name” will be taken from the take command function as described above. After conversion of the speech to the conversion of the name given by the user. After the AI matches the name of the user you want to send the message to it will proceed further. Otherwise, it will not proceed further and ask you to specify the phone number of the new user which is the name mentioned as a query by the user to the AI. But if matched then it will proceed to send the message by stating the message which you want to send then it will be taken by the command function, and it will be saved and then the ZEN will ask to state the time in hour, minute, or second format, and the message will be scheduled and sent at that time.

4.2. Open Application Automation

Automation can be delivered when we specify the application that we want to open. It can open only those messages that are mentioned in the project. Suppose you want to open an application ‘x’ which is mentioned in the project then it will open with the help of the “web browser” library.

4.3. YouTube Automation

In this we will open the YouTube application only when there is stated “YouTube search” in the speech. If it matches the condition, it will open the YouTube function and then it will show the user-selected result in the YouTube application. Besides these things, there are other functionalities in YouTube automation while playing the video some of the functions are:

- Pause
- Restart
- Mute
- Skip
- Back
- Fullscreen
- Film mode.

4.4. Dictionary

This is activated when you speak about “what is the meaning of”. It will convert it into the query and replace “what is the” with empty blanks and “meaning” with “”, then with the help of the Dictionary library it will find the meaning of the problem given and return the result.

4.5. Screenshot

In the screenshot, it will take the screenshot of the currently opened screen and then Zen will ask you the name which you want to give to the file. The screenshot is done with the help of “pyautogui” which consists of a function screenshot( ), and it will save it and open it with the help of “os”.

4.6. Temperature

“What’s the temperature in Ghaziabad” This speech will tell you the temperature in Ghaziabad by simply converting it into a query and getting the knowledge of the temperature of your area.

4.7. Speed Test

In the speed test, we will check two speeds upload speed and download speed. Here we will use the library known as “Speed test” which is used to calculate the speed and after finding the speed it is converted and displayed by the AI accordingly.

- Introducing Zen (ask zen to introduce)
- Some Human Resource Questions (how are you?)
- Search on Wikipedia (how to make pani puri)
- Jokes by AI (ask him for a joke)
- Repeat my Word (say to him that to “repeat my words”)
- Current Location (webbrowser library will open your current location)
- Play music (with the help of “playonyt” in pywhatkit, it plays the query)
- Video Downloader (Windows will open where the link of the video needs to be provided by the user and then video will be downloaded by AI).

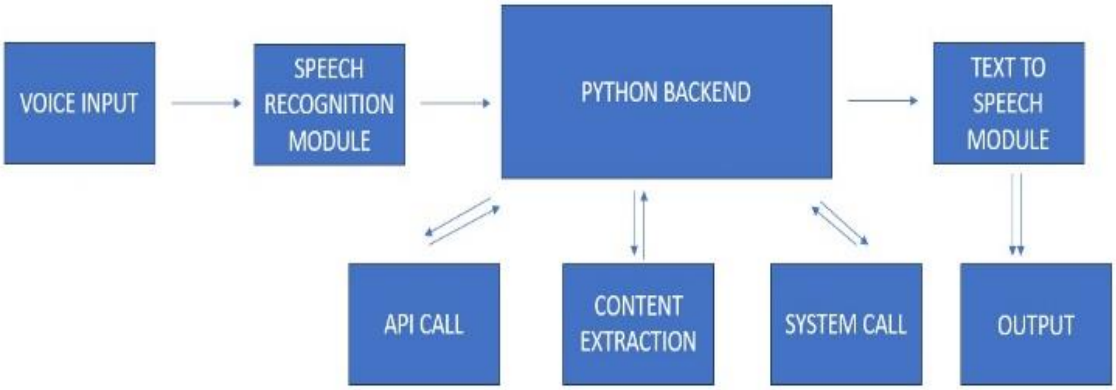


Fig 4.1. Working of the voice assistant ZEN.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1. Conclusion

Virtual Assistants on the desktop are a highly efficient method to arrange your schedule and execute numerous activities. They aid users in appropriately managing and arranging their time. When the user gives the voice assistant an instruction, they will carry it out. We can create a Python voice assistant that works with all Windows versions, like Alexa, Siri, or Google Assistant. Voice assistants are more straightforward and user-friendly, and they will do routine tasks on client request. They are beneficial in a range of sectors, including day-to-day usage, home appliances, etc. Those who are illiterate can access any information simply by conversing with the assistant, which is particularly useful. Voice assistant integration into daily life is increasing. The majority of the user's activities—including sending WhatsApp messages, utilizing Chrome, YouTube, and, conducting query searches, and more are automated. Over the past few years, voice assistants have undergone a significant period of development. This project utilizes Artificial Intelligence and Python to create a desktop assistant, ZEN, that can handle automated tasks in daily life. The assistant consists of three automations: OS automation, Chrome automation, and YouTube automation. OS automation allows users to open programs, software, and settings using voice commands. Chrome automation allows users to perform various tasks on Chrome without physical effort. Virtual personal assistants offer numerous benefits, such as being more trustworthy, portable, and providing more information than personal assistants. They are also linked to the internet, making them accessible at any time.

5.2. Future Scope

The future potential of voice assistants is very promising and advancing quickly. Voice assistants have made substantial progress in smart homes, customer service, healthcare, education, and other areas and will soon become a crucial part of our lives.

There is a lot of scope for enhancing and upgrading this project There can be space for future scope for this project. We can implement a chatbot that can give the project a complete look. Also, so far this project is limited to desktop users so we can extend this project to smartphones or maybe some other gadgets as well. Also, ZEN is a customized virtual voice assistant that has limited functionalities so we can add on more functionalities to perform other tasks of users. After upgrading, enhancing, and extending this project, ZEN will become very efficient in making the lives of users comfortable as it will be performing their tasks effortlessly.

REFERENCES

1. D. Lahiri, P. C. P. Kandimalla, and A. Jeysekar, “Hybrid multi purpose voice assistant,” in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEEE, 2023, pp. 816– 822.
2. I. Shazhaev, D. Mikhaylov, A. Shafeeg, A. Tularov, and I. Shazhaev, “Personal voice assistant: from inception to everyday application,” Indonesian Journal of Data and Science, vol. 4, no. 2, pp. 64–70, 2023.
3. A. Valera Roman,´ D. Pato Mart´ınez, A. Lozano Murciego, D. M. Jimenez´-Bravo, and J. F. de Paz, “Voice assistant application for avoiding sedentarism in elderly people based on iot technologies,” Electronics, vol. 10, no. 8, p. 980, 2021.
4. P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. Lee, “Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus,” IEEE Access, vol. 8, pp. 62 032–62 041, 2020.
5. P. Cheng and U. Roedig, “Personal voice assistant security and privacy—a survey,” Proceedings of the IEEE, vol. 110, no. 4, pp. 476–507, 2022.
6. F. Elghaish, J. K. Chauhan, S. Matarneh, F. P. Rahimian, and M. R. Hosseini, “Artificial intelligence-based voice assistant for bim data management,” Automation in Construction, vol. 140, p. 104320, 2022.
7. A. Poushneh, “Humanizing voice assistant: The impact of voice assistant personality on consumers’ attitudes and behaviors,” Journal of Retailing and Consumer Services, vol. 58, p. 102283, 2021.
8. M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” Multimedia Tools and Applications, vol. 80, pp. 9411–9457, 2021.
9. S. Raju, V. Jagtap, P. Kulkarni, M. Ravikanth, and M. Rafeeq, “Speech recognition to build context: A survey,” in 2020 international conference on computer science, engineering and applications (ICCSEA). IEEE, 2020, pp. 1–7.
10. A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” IEEE access, vol. 7, pp. 19 143–19 165, 2019.
11. A. Singh, V. Kadyan, M. Kumar, and N. Bassan, “Asroil: a comprehensive survey for automatic speech recognition of indian languages,” Artificial Intelligence Review, vol. 53, pp. 3673–3704, 2020.
12. M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, “Survey on deep neural networks in speech and vision systems,” Neurocomputing, vol. 417, pp. 302–321, 2020.
13. D. Al-Fraihat, Y. Sharrab, F. Alzyoud, A. Qahmash, M. Tarawneh, and A. Maaita, “Speech recognition utilizing deep learning: A systematic review of the latest developments,” HUMAN-CENTRIC COMPUTING AND INFORMATION SCIENCES, vol. 14, 2024.

14. D. O’Shaughnessy, “Trends and developments in automatic speech recognition research,” Comput. Speech Lang., vol. 83, no. C, jan 2024. [Online]. Available: https://doi.org/10.1016/j.csl.2023.101538

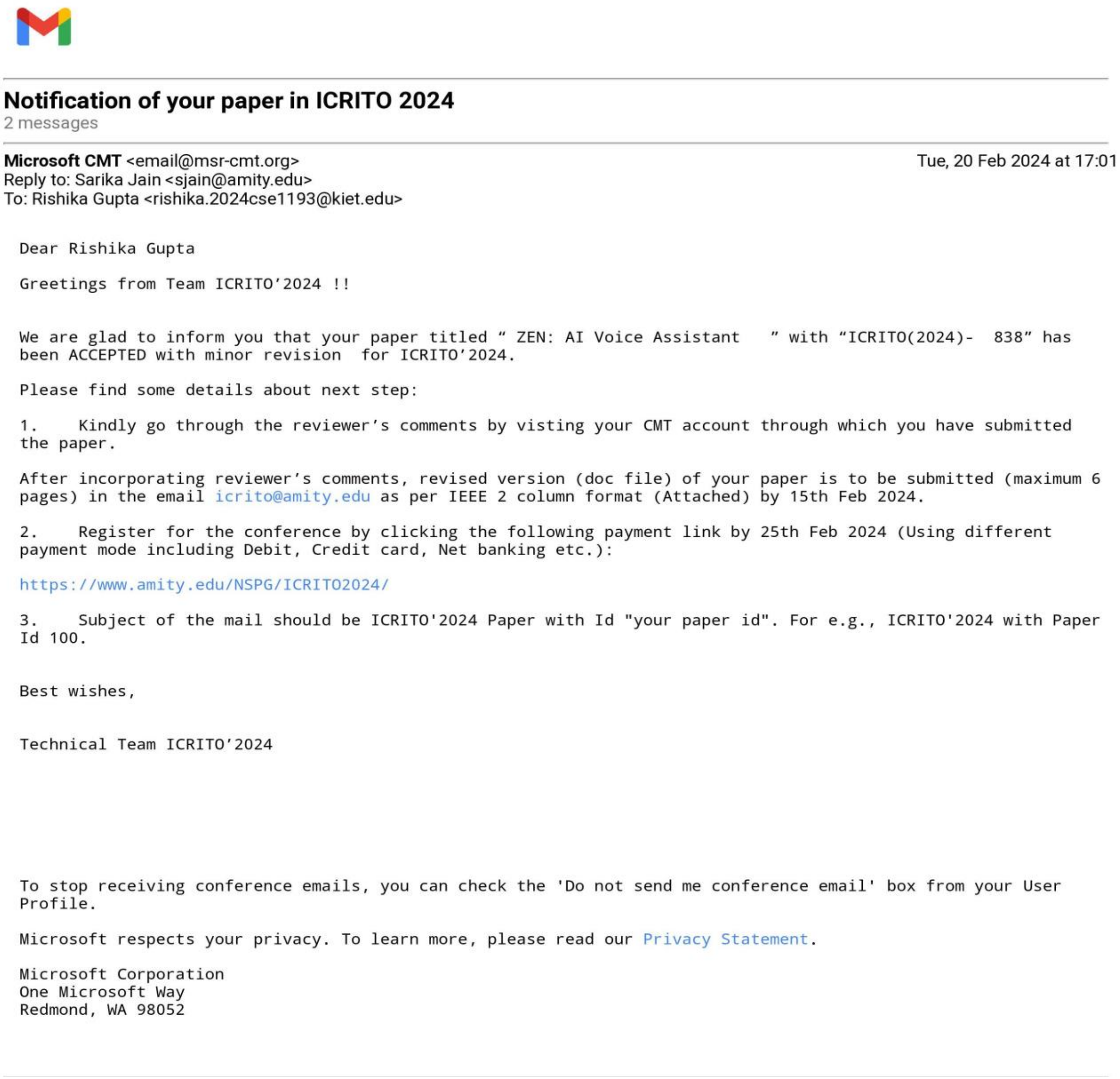
15. O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Transactions on audio, speech, and language processing, vol. 22, no. 10, pp. 1533–1545, 2014.

16. Z. Song, “English speech recognition based on deep learning with multiple features,” Computing, vol. 102, no. 3, pp. 663–682, 2020.

17. P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, “Adaptation algorithms for neural network-based speech recognition: An overview,” IEEE Open Journal of Signal Processing, vol. 2, pp. 33–66, 2020.

APPENDIX-I

Paper Acceptance Form



Notification of your paper in ICRITO 2024
2 messages

Microsoft CMT <email@msr-cmt.org> Tue, 20 Feb 2024 at 17:01
Reply to: Sarika Jain <sjain@amity.edu>
To: Rishika Gupta <rishika.2024cse1193@kiet.edu>

Dear Rishika Gupta

Greetings from Team ICRITO’2024 !!

We are glad to inform you that your paper titled “ ZEN: AI Voice Assistant ” with “ICRITO(2024)- 838” has been ACCEPTED with minor revision for ICRITO’2024.

Please find some details about next step:

1. Kindly go through the reviewer’s comments by visting your CMT account through which you have submitted the paper.

After incorporating reviewer’s comments, revised version (doc file) of your paper is to be submitted (maximum 6 pages) in the email icrito@amity.edu as per IEEE 2 column format (Attached) by 15th Feb 2024.

2. Register for the conference by clicking the following payment link by 25th Feb 2024 (Using different payment mode including Debit, Credit card, Net banking etc.):

https://www.amity.edu/NSPG/ICRITO2024/

3. Subject of the mail should be ICRITO'2024 Paper with Id "your paper id". For e.g., ICRITO'2024 with Paper Id 100.

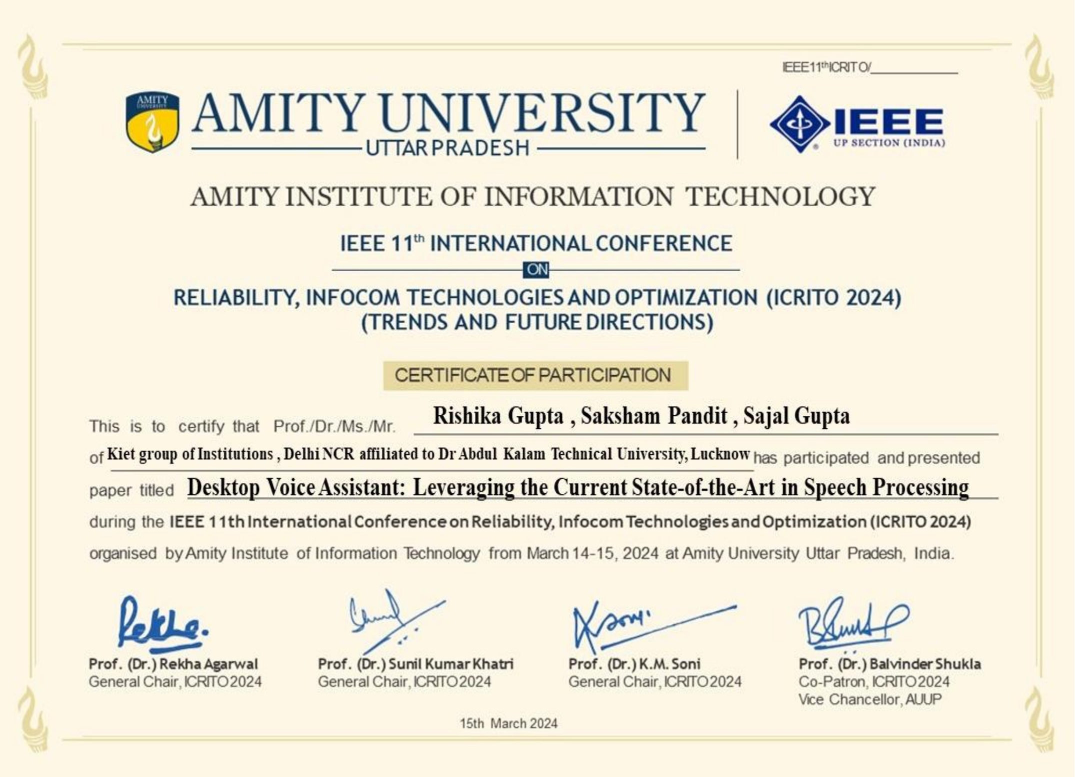
Best wishes,

Technical Team ICRITO’2024

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our Privacy Statement.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052



IEEE11thICRITO/____________

AMITY UNIVERSITY
UTTAR PRADESH

IEEE UP SECTION (INDIA)

AMITY INSTITUTE OF INFORMATION TECHNOLOGY

IEEE 11th INTERNATIONAL CONFERENCE
ON
RELIABILITY, INFOCOM TECHNOLOGIES AND OPTIMIZATION (ICRITO 2024)
(TRENDS AND FUTURE DIRECTIONS)

CERTIFICATE OF PARTICIPATION

This is to certify that Prof./Dr./Ms./Mr. **Rishika Gupta , Saksham Pandit , Sajal Gupta** of **Kiet group of Institutions , Delhi NCR affiliated to Dr Abdul Kalam Technical University, Lucknow** has participated and presented paper titled **Desktop Voice Assistant: Leveraging the Current State-of-the-Art in Speech Processing** during the **IEEE 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2024)** organised by Amity Institute of Information Technology from March 14-15, 2024 at Amity University Uttar Pradesh, India.

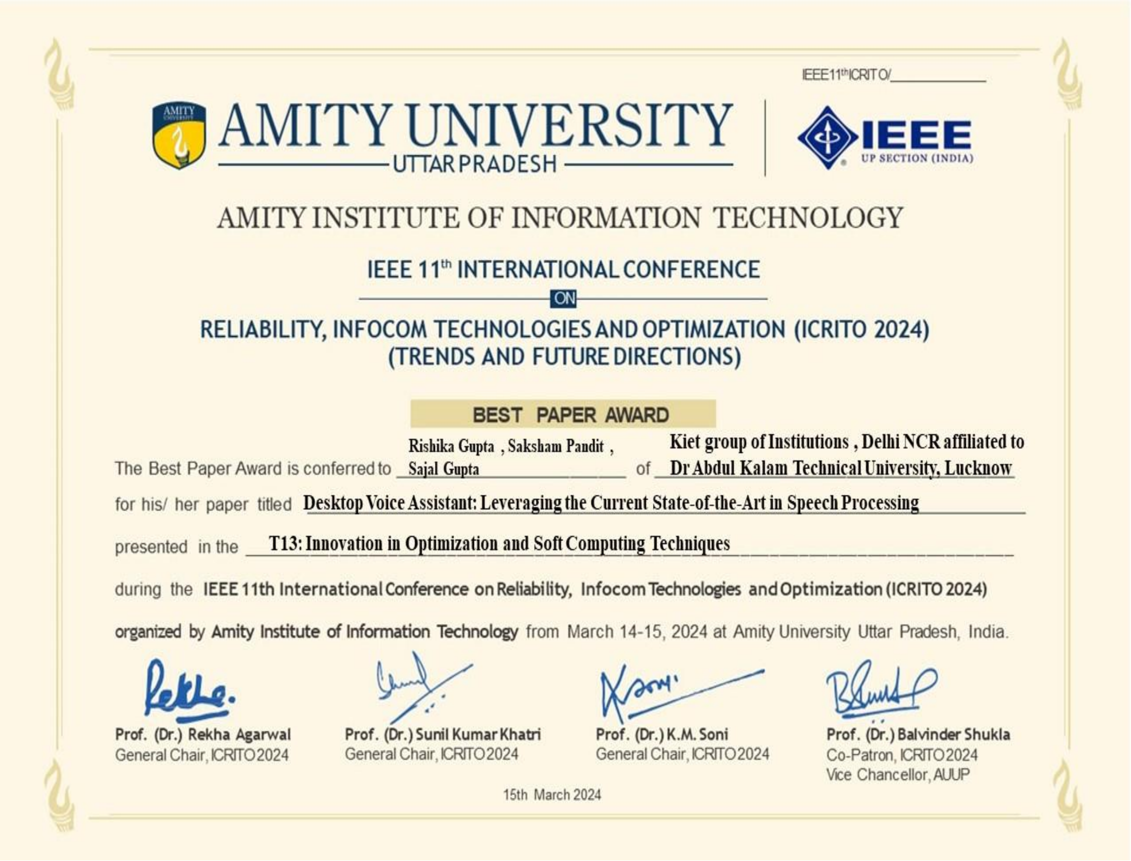
Rekha.
Prof. (Dr.) Rekha Agarwal
General Chair, ICRITO 2024

Prof. (Dr.) Sunil Kumar Khatri
General Chair, ICRITO 2024

Ksoni
Prof. (Dr.) K.M. Soni
General Chair, ICRITO 2024

Prof. (Dr.) Balvinder Shukla
Co-Patron, ICRITO 2024
Vice Chancellor, AUUP

15th March 2024



IEEE11thICRITO/____________

AMITY UNIVERSITY
UTTAR PRADESH

IEEE
UP SECTION (INDIA)

AMITY INSTITUTE OF INFORMATION TECHNOLOGY

IEEE 11th INTERNATIONAL CONFERENCE
ON
RELIABILITY, INFOCOM TECHNOLOGIES AND OPTIMIZATION (ICRITO 2024)
(TRENDS AND FUTURE DIRECTIONS)

BEST PAPER AWARD

The Best Paper Award is conferred to Rishika Gupta , Saksham Pandit , Sajal Gupta of Kiet group of Institutions , Delhi NCR affiliated to Dr Abdul Kalam Technical University, Lucknow

for his/ her paper titled **Desktop Voice Assistant: Leveraging the Current State-of-the-Art in Speech Processing**

presented in the **T13: Innovation in Optimization and Soft Computing Techniques**

during the **IEEE 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2024)** organized by **Amity Institute of Information Technology** from March 14-15, 2024 at Amity University Uttar Pradesh, India.

Rekha.

Prof. (Dr.) Rekha Agarwal
General Chair, ICRITO 2024

Prof. (Dr.) Sunil Kumar Khatri
General Chair, ICRITO 2024

Prof. (Dr.) K.M. Soni
General Chair, ICRITO 2024

Prof. (Dr.) Balvinder Shukla
Co-Patron, ICRITO 2024
Vice Chancellor, AUUP

15th March 2024

Desktop Voice Assistant: Leveraging the Current State-of-the-Art in Speech Processing

Saksham Pandit
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
saksham.2024cse1022@kiet.edu

Rishika Gupta
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
rishika.2024cse1193@kiet.edu

Sajal Gupta
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
sajal.2024cse1046@kiet.edu

Shivali Tyagi
Computer Science and Engineering Department
KIET Group of Institutions
Ghaziabad, India
shivali.tyagi@kiet.edu

***Abstract*— Virtual voice assistants have transformed the way humans interact with computers, especially with mobile devices. This study examines the latest advancements in speech processing technologies to create a virtual voice assistant for desktop users. We examine the latest progress in speech recognition, natural language processing, and dialogue control. Important factors to consider are precision in quiet environments, compatibility with current desktop processes, and customization that suits user choices. We examine the possible advantages and obstacles of this approach, as well as identify areas for future research. This study proposes to connect mobile and desktop voice assistants using modern speech processing technologies, providing users with a smooth and effective method to control their computers through voice commands.**

Keywords*— *Voice Assistant, Speech Recognition

I. Introduction

Today as we are developing with technology and the notion of Artificial Intelligence, we are making our living simple and less complicated. One such clever use of this growing tech-nology of artificial intelligence is voice assistant. Nowadays we are not even utilizing our fingers to type or touch. We merely talk about the tasks, and it gets done by the virtual assistant. Today we may talk to our devices exactly as we talk to any human to execute the duty. Software applications known as voice assistants are designed to recognize and understand spoken instructions in natural language and carry out the tasks that the user specifies. This virtual assistant program promotes user productivity by handling the everyday chores of the user and by presenting information from internet sources to the user. Today people are no longer surprised when they talk to their virtual assistant as it has become their part of lives and is always available with them on their mobile phones or other devices. When Apple Corporation debuted its Siri, it became everyone's focus of attention. This virtual assistant given by the Apple firm became a source of entertainment for many. Then the arrival of Google Assistant in Android phones allowed Android users an opportunity to experience the same. Sooner or later virtual assistants became a significant software and feature of smartphones. We may construct our own customized virtual assistant with the aid of Python programming language. Python includes several inherent libraries, modules, and packages, which help the building of the tailored and personalized virtual assistant with essential features and functions. With the rise in artificial intelligence which contains components like natural language processing and machine learning, virtual assistants are evolving out smarter than before and the work done by them is also growing better and more precise. It would not be inaccurate to state that it is all about allowing our Virtual assistant to work for us, select information, and produce a decent answer. The primary purpose of this project is to build software that will be able to serve individuals like a personal assistant. This program attempts to produce a virtual assistant for Windows-based platforms. The goal of this application software is to accomplish and execute the user's duties for instructions, supplied in either voice or text. It will ease the task of the user. In this project, we have designed a virtual assistant ZEN which would save time for the consumers. It simplifies human life by enabling users to operate PCs or laptops using just voice instructions. Voice Assistants take up less time.

II. Literature Review

Currently, a wide range of Smart Personal Digital Assistant applications are emerging in the market for various device plat-forms. And these new software applications perform significantly better than PDAs because they incorporate every feature of a smartphone. Additionally, VPAs are more dependable than personal assistants because they are portable and usable at any time. They have more information than any assistant since they are internet-connected [1]. The authors assert that although voice assistants currently have limited capabilities, this will soon change, as they make strides into space exploration and basic medicinal procedures. It transpires that the voice assistant's capabilities will expand beyond resolving basic user needs to encompass more intricate and financially burdensome duties, the execution of which will necessitate the ongoing education of artificial intelligence [2].

Adrian et al. proposed a solution that allows elderly individuals to track their daily physical activity using virtual voice assistants, IoT devices, and activity-monitoring smart bracelets. This enables the elder people to avoid developing sedentary habits by simply using their voice. Their endeavor, designated EMERITI sought to enhance the quality of life for the elderly by employing virtual assistants across various case studies [3]. Po-Sheng et al.[4] concentrated their study on the creation of a Deep Neural Network (DNN)-based campus virtual assistant. This research is presented in App format, which is economical and simple to use. The system offers a straightforward voice response interface, obviating the necessity for users to navigate intricate web pages or app menus in search of information. The survey conducted by Peng et al. [5] delineates research domains characterized by a comparatively comprehensive understanding of the threat but a dearth of effective countermeasures, such as concealed vocal commands. It also addresses the work that examines the privacy implications, including research on the administration of consent recording. The purpose of this survey is to compile an all-encompassing study plan concerning the security and privacy of PVAs. The authors of [6] discussed about voice assistant interface interaction with the BIM model from a distance has been enabled. Individuals with visual impairments can access and augment BIM models. BIM novices are capable of practicing BIM features and retrieving information with minimal skill. According to a study by Atieh [7], voice engagement with a VA that combines sincerity, creativity, and intelligence allows users to assert control over their speech interactions with the Virtual Assistant, focus on the voice interaction, and engage in exploratory behavior The exploratory behavior of consumers results in their continued use of voice assistants.

A survey by Malik et al. [8] discussed and examined voice recognition methods. An ASR depends on three modules: fea-ture extraction, classification, and language model, according to its fundamental design. Analysis of classification models shows HMM performed best. The addition of a language model may considerably affect ASR accuracy. Even when sub-optimal approaches are employed to develop language models, further research will enhance voice recognition. Authors of [9] designed methods that use probability theory, pattern machine, and now deep learning. Because they function alone, these methods lack literal meaning and context. The context must accompany the translation. Time series context construction is difficult. Due to dynamic inputs, context may change. Audio corpus association with learned vectors is used for similar-ity, missing data, and prediction. Sequence modeling using diverse algorithms must enhance voice recognition. Language sequence modeling helps clarify ambiguous words. Additional emotions may improve context and processing.

Ali et al. [10] used data from 174 publications published between 2006 and 2018, this research conducted a statistical analysis of deep learning in voice applications. The bulk of publications explored voice recognition. The study's databases were in English. Most research evaluated system efficiency using WER (word error rate). Most deep learning researchers extract voice features using MFCCs. HMM and GMM signifi-cantly utilized MFCCs. Many researchers considered Linear Predictive Coding for feature extraction in deep learning models. They found that Authors should utilize hybrid models since research shows that DNN models with HMM or GMM information perform better. Singh et al. [11] surveyed that, HTK is the most popular toolkit used Indian language ASR.

Researchers now often utilize Kaldi to create systems. An extensive literature review shows that not many Deep learning methods have been used to test much of ASR. The scarcity of big-voice corpora in multiple languages is the reason. The exploratory work on feature extraction is also confined to a few popular languages and most speech signal classification investigations employ HMM-GMM. Alam et al. [12] reviewed cutting-edge DNN algorithms and architectures for vision and speech and found that RNN models dominate voice recognition systems, notably in NLP applications. Al-Fraihat1 et al. [13], surveyed that hybrid DNN models are being used because they perform better than stand-alone models. O'Shaughnessy [14] discussed that the MLP structure is the foundation of an ANN, while SGD search is the usual training method. Many enhancements to these fundamental approaches have taken use of additional computer power and large data volumes and thus ANN models remain opaque and hard to comprehend but proper structuring may enhance ANNs' power. Abde et al. [15] proposed an innovative technique to use CNNs for voice recognition that directly accommodates some speech variability. In our hybrid CNN-HMM technique, the HMM handles temporal variability, while convolving along the frequency axis invariantly handles minor frequency changes caused by speaker variances in speech data. Song et al. Song et al. [16] focused on the application of deep learning in acoustic features and speech attributes and proposed a deep speech-based English speech recognition algorithm that combines multiple features. The CNN–RBM–ASAT algorithm proposed in this paper has much higher accuracy than CNN and RBM algorithms, so combining the two can improve accuracy. Bell et al. [17] meta-analysis shows that adaption techniques work for hybrid and E2E systems across corpora and classes. However, unsupervised and semi-supervised E2E system training utilizing uncertainty propagation techniques remains a major research problem. We have summarized the findings and limitations in the table I and summarized in figure 1.

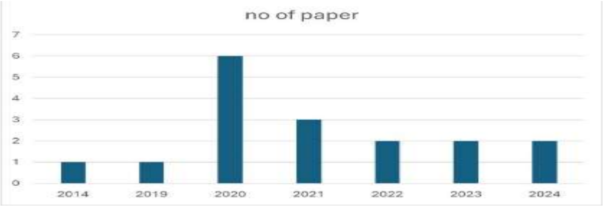


Fig. 1. Working of the voice assistant ZEN.

TABLE I: Findings and Limitations of authors

Reference	Findings	Limitations
[8]	ASR is dependent on three modules: the feature extraction module, the classification module, and the language model. Among	Language model helps to improve the accuracy of an ASR, but only sub-optimal methods are available.

	classification models, HMM performed the best	
[9]	Speech Recognition methods use probability theory, pattern machine and now deep learning	These techniques function independently and therefore they miss the literal meaning and context associated with it. Due to dynamic inputs, context may change. Audio corpus association with learned vectors is used for similarity, missing data, and prediction
[10]	Most articles employed WER (word error rate) to evaluate system efficiency. Most researchers employ MFCCs to extract voice features in deep learning models. Classifiers like HMM and GMM employed MFCCs extensively. When utilizing deep learning models	The power of RNN models, notably Long Short-Time Memory (LSTM), in voice recognition makes deep RNN highly accurate but Speech Recognition using RNNs is understudied
[11]	The most popular toolkit used for Indian language speech recognition is HTK (Hidden Markov Model Toolkit) whereas Kaldi is used for system building.	Deep learning methods have not been used extensively in ASR systems for Indian languages. Because there is a scarcity of multilingual speech corpora. Feature extraction experiments are also confined to popular languages.
[12]	Reviewed current DNN algorithms and structures for vision and voice applications. For NLP applications RNN models were widely used in voice recognition systems.	Three fundamental model limitations: the risks of employing limited datasets mobile device hardware limits, and overoptimizing intelligent algorithms to substitute human experts.
[13]	Found that hybrid DNN models out-perform stand-alone models and are being used more.	Transformational models can parallelize, learn quicker, and perform better for low-resource languages, yet there is a research gap in voice recognition utilizing transformer model
[14]	Fundamental MLP is the fundamental ANN structure, while SGD search is the typical training method.	The methodology was rejected due to the failure of ES (expert system) to manage speech fluctuation and no mistake correction feedback.
[15]	Explained how to use CNNs for voice recognition in a unique approach that directly accommodates some speech variability and for this, they used a hybrid CNN-HMM technique	CNNs pre-trained with convolutional RBMs performed better in large vocabulary voice search but not phone identification. This disparity needs more study.
[16]	The CNN-RBM-ASAT algorithm proposed in this paper has much higher accuracy than CNN and RBM algorithm, so combining the two can improve the accuracy.	To generate a novel network feature extraction notion, a clustering technique before feature extraction and screening the features need to be included.
[17]	Used adaption techniques for hybrid and E2E systems across corpora and classes.	There is a need to use uncertainty propagation approaches for unsupervised and semi-supervised

		E2E system training.

III. RESEARCH METHODOLOGY

Our entire project ZEN is designed using several techniques of artificial intelligence like Natural language processing and speech recognition written in Python programming language. Python offers a plethora of specialized built-in libraries and packages to carry out the tasks that the user inputs. The user’s voice, which is a list of actions that the user wants completed, is the data used in this project. Whenever the user gives the voice command as the input, the speech recognition module comes into effect it takes the voice as an input listen to the words of voice input identifies them with its capability, and converts spoken words into text which is further spoken by Zen, this spoken words becomes the output voice and the task is done by ZEN. The steps below show how the command is taken up by the assistant:

The library that we have named ”Recognizer” recognizes the command, and as a result, voice is converted to text. They should also be separated from the surrounding text by one space.

After the command is transformed into a query, it will identify the words in the sentence, look up the keywords that match its condition, and execute the function by the specified condition.

The assistant ZEN is composed of the components mentioned in Table II.

TABLE II
Table depicting the number of components used in Zen.

S.No.	Component
1	Speech Recognition Module
2	Python Backend
3	API Calls
4	Content Extraction
5	System call
6	Text-to-speech module

A. Speech Recognition Module

This Python module aids in the conversion of speech to text. This module automatically receives the voice input. The identical text is received and sent to the central processing unit. This enables us to transform audio into text for additional processing. We have imported a speech recognition module as “Sr”. The Recognizer class inside the speech recognition module lets us recognize the audio. The same module contains a Microphone class that offers access to the microphone of the device. So, using the microphone as the source, we attempt to listen to the audio using the listen () function in the Recognizer class. We have also set the pause threshold to 1, that is it will not complain even if we halt for one second while we talk. We have set the language to Indian English. It returns the transcript of the audio which is nothing but a string. We’ve put it in a variable named query. The speech input Texts from the distinct corpora arranged on the PC may be sent to users.

B. Python Backend

This aids in providing the user with the necessary output. The output generated by the voice recognition module is sent to the Python backend, which also determines if the output is a system call, an API call, or context extraction. Also, this component output is fed to the Text-to-speech module.

C. API Calls

API means Application Programming Interface. It helps to connect two applications and transmits user requests to the supplier, who then returns the result to the user.

D. Content Extraction

The process of extracting organized information from unstructured or semi-structured machine-readable resources is known as content extraction. This activity mostly involves using natural language processing (NLP) to process documents written in human languages. It all comes down to pulling pertinent and related data from the webpage.

E. System call

This facilitates the computer program’s request for a service from the operating system’s kernel while it is running.

F. Text-to-speech module

A text-to-speech engine is needed for a text-to-speech module. Written text can be converted into waveforms that aid in producing sound using a text-to-speech engine. And these are the next essential and fundamental libraries that support ZEN in doing the job.

Speech recognition. The foundation of this project is this library. This is used to identify human speech and translate it from input voice to text.

Pyttsx3. This offline module is a significant resource. This module alone contains the run and wait functions as well. It specifies the time interval between inputs, or more precisely, how long the system will wait for the next input. This has the primary advantage of operating offline.

Wikipedia. It is an online Python library that needs to be connected to the internet to function. With the aid of this library, Zen may manage Wikipedia requests and provide answers.

OS. This library helps with a variety of operating system tasks that can be done automatically. This library offers functions for generating and erasing directories (folders), retrieving their contents, updating folders, and more.

Web browser. The platform that this library offers the system’s default web browser is helpful. Users must provide a filename or URL to operate with this library, and the output is then shown in the browser.

Pywhatkit. Utilizing this library is an absolute breeze, as it is designed to simplify any interaction with the browser.

IV. Result Analysis

When we start up the assistant ZEN it greets the user by saying “Hello Sir I am ZEN. How may I help you?”. And it waits for the user to command tasks. Following are the tasks that ZEN can perform when the user commands it.

A. WhatsApp Automation

We must state first that we are sending a “WhatsApp message” It will match the keyword and call the WhatsApp function. In WhatsApp automation, “name” will be taken from the take command function as described above. After conversion of the speech to the conversion of the name given by the user. After the AI matches the name of the user you want to send the message to it will proceed further. Otherwise, it will not proceed further and ask you to specify the phone number of the new user which is the name mentioned as a query by the user to the AI. But if matched then it will proceed to send the message by stating the message which you want to send then it will be taken by the command function, and it will be saved and then the ZEN will ask to state the time in hour, minute, or second format, and the message will be scheduled and sent at that time.

B. Open Application Automation

Automation can be delivered when we specify the application that we want to open. It can open only those messages that are mentioned in the project. Suppose you want to open an application ‘x’ which is mentioned in the project then it will open with the help of the “web browser” library.

C. YouTube Automation

In this we will open the YouTube application only when there is stated “YouTube search” in the speech. If it matches the condition, it will open the YouTube function and then it will show the user-selected result in the YouTube application. Besides these things, there are other functionalities in YouTube automation while playing the video some of the functions are:

- Pause
- Restart
- Mute
- Skip
- Back
- Fullscreen
- Film mode.

D. Dictionary

This is activated when you speak about “what is the meaning of”. It will convert it into the query and replace “what is the” with empty blanks and “meaning” with “”, then with the help of the Dictionary library it will find the meaning of the problem given and return the result.

E. Screenshot

In the screenshot, it will take the screenshot of the currently opened screen and then Zen will ask you the name which you want to give to the file. The screenshot is done with the help of “pyautogui” which consists of a function screenshot (), and it will save it and open it with the help of “os.”

F. Temperature

“What’s the temperature in Ghaziabad” This speech will tell you the temperature in Ghaziabad by simply converting it into a query and getting the knowledge of the temperature of your area.

G. Speed Test

In the speed test, we will check two speeds upload speed and download speed. Here we will use the library known as “Speed test” which is used to calculate the speed and after finding the speed it is converted and displayed by the AI accordingly.

- Introducing Zen (ask zen to introduce)
- Some Human Resource Questions (how are you?)
- Search on Wikipedia (how to make pani puri)
- Jokes by AI (ask him for a joke)
- Repeat my Word (say to him that to “repeat my words”)
- Current Location (webbrowser library will open your current location)
- Play music (with the help of “playonyt” in pywhatkit, it plays the query)
- Video Downloader (Windows will open where the link of the video needs to be provided by the user and then video will be downloaded by AI).

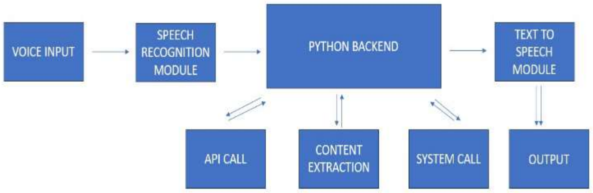


Fig. 2. Working of the voice assistant ZEN.

V. Conclusion

Virtual Assistants on the desktop are a highly efficient method to arrange your schedule and execute numerous activities. They aid users in appropriately managing and arranging their time. When the user gives the voice assistant an instruction, they will carry it out. We can create a Python voice assistant that works with all Windows versions, like Alexa, Siri, or Google Assistant. Voice assistants are more straightforward and user-friendly, and they will do routine tasks on client request. They are beneficial in a range of sectors, including day-to-day usage, home appliances, etc. Those who are illiterate can access any information simply by conversing with the assistant, which is particularly useful. Voice assistant integration into daily life is increasing. The majority of the user’s activities—including sending WhatsApp messages, utilizing Chrome, YouTube, and, conducting query searches, and more are automated. Over the past few years, voice assistants have undergone a significant period of development. This project utilizes Artificial Intelligence and Python to create a desktop assistant, ZEN, that can handle automated tasks in daily life. The assistant consists of three automations: OS automation, Chrome automation, and YouTube automation. OS automation allows users to open programs, software, and settings using voice commands. Chrome automation allows users to perform various tasks on Chrome without physical effort. Virtual personal assistants offer numerous benefits, such as being more trustworthy, portable, and providing more information than personal assistants. They are also linked to the internet, making them accessible at any time. The future potential of voice assistants is very promising and advancing quickly. Voice assistants have made substantial progress in smart homes, customer service, healthcare, education, and other areas and will soon become a crucial part of our lives.

REFERENCES

[1] D. Lahiri, P. C. P. Kandimalla, and A. Jeysekar, "Hybrid multipurpose voice assistant," in 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). IEE,2023, pp. 816-822.

[2] I. Shazhaev, D. Mikhaylov, A. Shafeeg, A. Tularov, and I. Shazhaev, "Personal voice assistant: from inception to everyday application," Indonesian Journal of Data and Science, vol. 4, no.2, pp. 64-70,2023.

[3] A. Valera Roman, 'D. Pato Martínez, A. Lozano Murciego, D.M. Jimenez'-Bravo, and J.F. de Paz, "Voice assistant application for avoiding sedentarism in elderly people based on iot technologies," Electronics, vol.10, no.8, p. 980,2021.

[4] P.-S. Chiu, J.-W. Chang, M.-C. Lee, C.-H. Chen, and D.-S. Lee, "Enabling intelligent environment by the design of emotionally aware virtual assistant: A case of smart campus," IEEE Access, vol. 8, pp. 62 032-62 041,2020.

[5] P. Cheng and U. Roedig, "Personal voice assistant security and privacy – a survey, "Proceedings of the IEEE, vol. 110, no.4, pp.476-507, 2022.

[6] F. Elghaish, J. K. Chauhan, S. Matarneh, F. P. Rahimian, and M. R. Hosseini, "Artificial intelligence-based voice assistant for bim data management," Automation in Construction, vol. 140, p. 104320, 2022.

[7] A. Poushneh, "Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors," Journal of Retailing and Consumer Services, vol. 58, p. 102283, 2021.

[8] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," Multimedia Tools and Applications, vol. 80, pp. 9411-9457, 2021.

[9] S. Raju, V. Jagtap, P. Kulkarni, M. Ravikanth, and M. Rafeeq, "Speech recognition to build context: A survey," in 2020 international conference on computer science, engineering, and applications (ICCSEA). IEEE 2020, pp. 1-7.

[10] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," IEEE access, vol. 7, pp. 19 143-19 165, 2019.

[11] A. Singh, V. Kadyan, M. Kumar, and N. Bassan, "Asroil: a comprehensive survey for automatic speech recognition of indian languages," Artificial Intelligence Review, vol. 53, pp. 3673-3704, 2020.

[12] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, "Survey on deep neural networks in speech and vision systems," Neurocomputing, vol. 417, pp. 302-321, 2020.

[13] D. Al-Fraihat, Y. Sharrab, F. Alzyoud, A. Qahmash, M. Tarawneh, and A. Maaita, "Speech recognition utilizing deep learning: A systematic review of the latest developments," HUMAN-CENTRIC COMPUTING AND INFORMATION SCIENCES, vol. 14, 2024.

[14] D. O'Shaughnessy, "Trends and developments in automatic speech recognition research," Comput. Speech Lang., vol. 83, no. C, jan 2024. [Online]. Available: https://doi.org/10.1016/j.csl.2023.101538.

[15] O. Abdel-Hamid, A. -r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," IEEE/ACM Transactions on audio, speech, and language processing, vol.22, no. 10, pp. 1533-1545, 2014.

[16] Z. Song, "English speech recognition based on deep learning with multiple features," Computing, vol. 102, no. 3, pp. 663-682, 2020.

[17] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation algorithms for neural network-based speech recognition: An overview," IEEE Open Journal of Signal Processing, vol. 2, pp. 33-66, 2020.

desktop

ORIGINALITY REPORT

3%	1%	2%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Zhaojuan Song. "English speech recognition based on deep learning with multiple features", Computing, 2019 Publication	1%
2	Submitted to KIET Group of Institutions, Ghaziabad Student Paper	1%
3	Submitted to The NorthCap University, Gurugram Student Paper	<1%
4	Atieh Poushneh. "Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors", Journal of Retailing and Consumer Services, 2021 Publication	<1%
5	www.ijitee.org Internet Source	<1%
6	cse.cvrce.edu.in Internet Source	<1%

desktop

PAGE 1

Verb This verb may be incorrect. Proofread the sentence to make sure you have used the correct form of the verb.

Run-on This sentence may be a run-on sentence.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Run-on This sentence may be a run-on sentence.

Verb This verb may be incorrect. Proofread the sentence to make sure you have used the correct form of the verb.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

PAGE 2

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

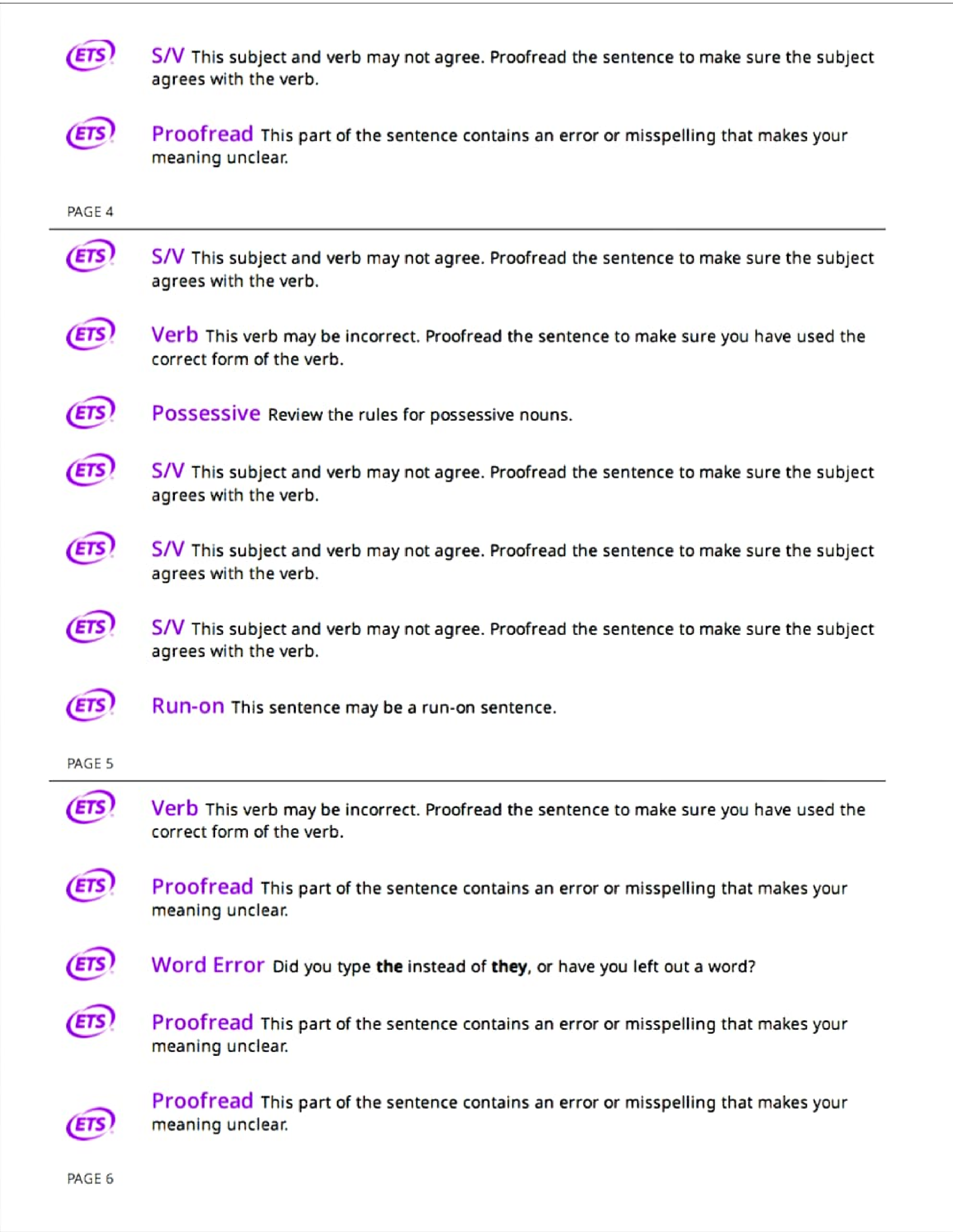
S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

PAGE 3

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.



S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

PAGE 4

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Verb This verb may be incorrect. Proofread the sentence to make sure you have used the correct form of the verb.

Possessive Review the rules for possessive nouns.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Run-on This sentence may be a run-on sentence.

PAGE 5

Verb This verb may be incorrect. Proofread the sentence to make sure you have used the correct form of the verb.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

Word Error Did you type **the** instead of **they**, or have you left out a word?

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

PAGE 6

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Pronoun This pronoun may be incorrect.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Verb This verb may be incorrect. Proofread the sentence to make sure you have used the correct form of the verb.

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

PAGE 7

S/V This subject and verb may not agree. Proofread the sentence to make sure the subject agrees with the verb.

Proofread This part of the sentence contains an error or misspelling that makes your meaning unclear.

PAGE 8