2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Mar 14-15, 2024

2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 979-8-3503-5035-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRITO61523.2024.10522178

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 andfunctions. 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 studyon 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 SGDsearch 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.

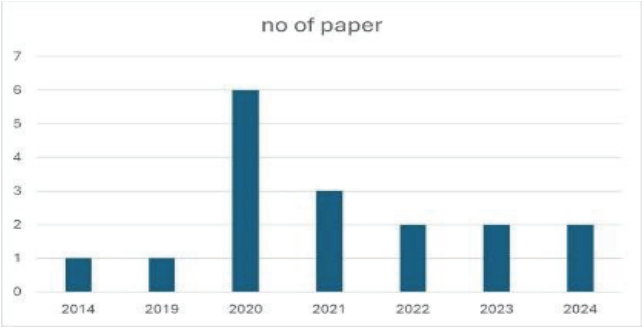

Fig. 1. Working of the voice assistant ZEN.

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 classification models, HMM performed the best	Language model helps to improve theaccuracy of an ASR,but only sub-optimal methods are available.
[9]	Speech Recognition methods use probability theory, pattern machineand now deep learning	These techniques function independently and thereforethey 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), invoice 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	To generate a novel network feature extraction notion, a clustering technique beforefeature extraction and screening the features
	accuracy.	need to be included.
[17]	Used adaption techniques for hybrid and E2Esystems 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 thespecified 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 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.

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 givenby 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 providedby the user and then video will be downloaded by AI).

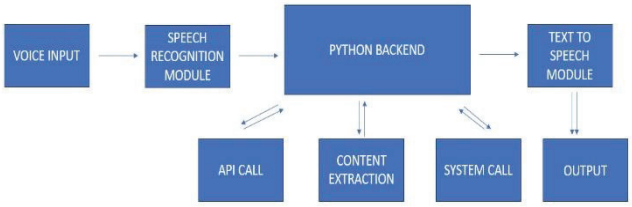

Fig. 2. Working of the voice assistant ZEN.

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 physicaleffort. 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 andU. 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.