VOICE ASSISTANT IN PYTHON: STREAMLINING HUMAN-DEVICE INTERACTION AND REDUCING DEPENDENCY ON INPUT DEVICES

Pravesh Kumar Bansal[1*], Rakesh Kumar Vishram[2]

Abstract:
Today, technological evolution is accelerating. It may not be worth it if people still attempt to use multiple input devices during this evolutionary period. As a result, we developed Voice Assistant (VA) in Python to lower the barriers between end users and VA operating systems and allow users to perform any command on their devices without using a keyboard. The primary goal of VA is to reduce the need for input devices such as keyboards and microphones; by reducing the number of input devices, space and expense are saved. VA is simple to use on the most recent devices. Millions of devices now have them in their homes and other places. The work presented in this paper describes the operation of a VA and its primary process, which is to issue a voice command (VC) and receive output in text format or on the screen.

Keywords: Voice Assistant, Python, Personal Desktop Virtual Assistant.

[1*] Department of Computer Science and Engineering, University of Engineering and Management Jaipur, Rajasthan -303807, India, Email: Bansal086@gmail.com
[2] Department of Computer Science and Engineering University of Engineering and Management Jaipur, Rajasthan -303807, India, Email: Rakeshvishram12@gmail.com

***Corresponding Author:**
* Department of Computer Science and Engineering, University of Engineering and Management Jaipur, Rajasthan -303807, India, Email: Bansal086@gmail.com

DOI: 10.48047/ecb/2023.12.si10.00470

1. INTRODUCTION

In the modern era, VA cannot sustain users because accessing apps and services on various devices is more accessible. Microsoft Cortana, Google Assistant, and Amazon Alexa are well-known and found on laptops and mobile platforms. The goal of the Virtual Assistant (VLA), which serves as a socio-conversational agent in virtual worlds, is to simulate how a human would behave in a conversation [1]. The primary motivation for creating a VLA is to provide helpful information such as weather, facts from Wikipedia, setting the alarm, watching movies on television streaming from Netflix, making a shopping list, or scheduling your day's work properly. By using only your simple VC.

VA is also used in a dialogue system to help with practical things like customer service, getting information, and navigating [2].

VA has a lengthy history of substantial additions with multiple waves. Virtual assistants (VA) for dictation, search, and VC have become a regular function on smartphones and wearable devices [3].

Complement or replace customer service with humans; that is, instead of hiring humans and wasting extra resources on them, their cost, and training on them. We can replace them with a VA who will do the same job with better speed and work without extra cost or resources. VA has several exciting capabilities, such as:

- Send Messages.
- Play music from streaming music services.
- Answer questions asked by users.
- Provide weather information.

The capabilities of VA are increasing day by day. Amazon and Google have provided platforms for developers in structure to extend their assistant's capabilities [4].

The structure of this Research is as follows: Some earlier research employed in VA is discussed in Section 2. The proposed model and methods for enhancing recognition are shown in Section 3. The proposed algorithm is then further explained in Section 4. In screenshots taken from our system without a keyboard or mouse, Section 6 shows the VA outcome. The system's conclusion and future scope are shown in Section 7.

RELATED WORK

There are many voices assistant is which can run on many platforms or for many devices, like Android (Alexa), Apple (Siri), Microsoft (Cortona), Google Home, and Amazon Alexa. Every firm that develops a VA has its own unique set of development processes and strategies. However, their objective remains the same: to offer specific assistants that can be integrated into other platforms, such as intelligent speakers, and are virtually available at a lower cost [5].

These companies may have their second motive for creating these VA is to make a Socio-Conversational (SC) agent in the virtual world or as a conversational partner; they conduct a conversation via auditory or textual methods (Chatbot). During their Research, they think that making these types of VA as an SC agent can be used as a dialog system for various practical purposes like customer service, information acquisition, and how a human would behave as a conversational partner [6]. They clarified that speaking is the most common method of communication used by humans in general. Because this is the most effective method through which people would be able to communicate with robots using voice detection, such as the Hidden Markov Model (HMM) and Dynamic Time Wrapping (DTW) [7].

They are working on the next generation of virtual assistants, which can multitask and simultaneously interpret input types, including audio, images, videos, touch sensors, etc. In addition, specific approaches are attempting to apply VA in creating self-driving automobiles [8]. These individuals study in the medical area to see how they can best serve Veterans Affairs patients. According to the report, 66% of Americans use some prescription medication, yet, around 50% of those customers do not take their medication as prescribed. Patients may schedule timely reminders to ensure they don't forget to take their medication at the specified time. By incorporating VLA into healthcare systems, patients can converse with their digital assistants to either make an appointment with their favorite doctor or discuss their concerns about their health with that physician [9]. These individuals conducted studies and descriptions of the language, which explained how voice is the primary interface for the language and how it is the best form of communication. Speech signal was translated into analog and digital wave structure so that a machine could understand it [10]. It provided the interface between humans and machines. They discuss a genetic algorithm and NLP solicitations of Indian languages at work [11].

PROPOSED MODEL

We all have our part of the creative; therefore, it makes sense that any VA firm's R&D team would use its unique techniques and perspectives while creating its products. Some helpers can execute various activities correctly and as the user desires, whereas others can only accomplish a subset of those jobs as accurately as possible.

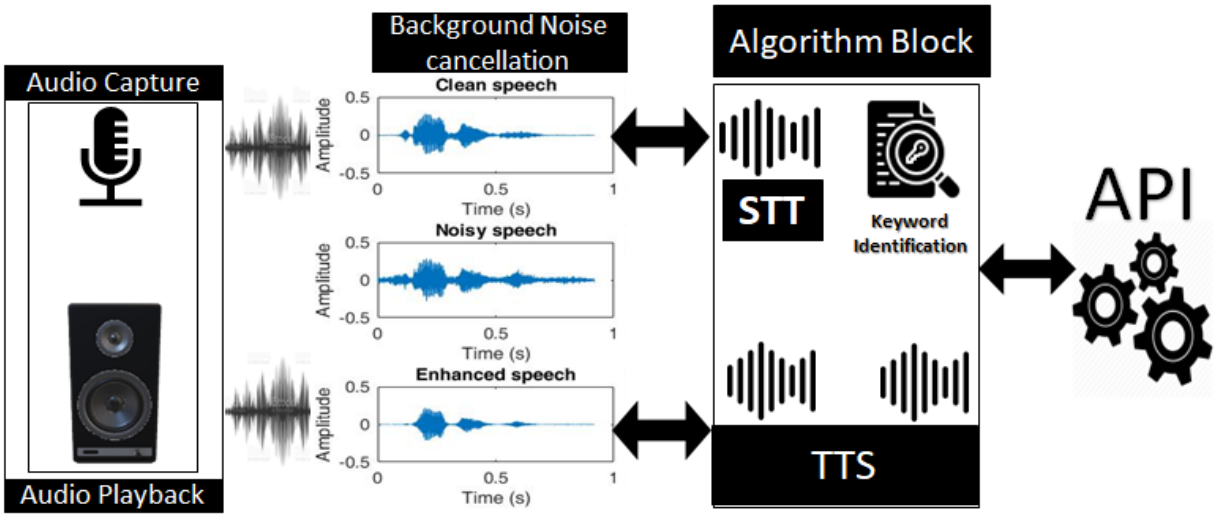

Fig. 1. Process of converting Speech to Text (STT) and Text to Speech(TTS) and noise cancellation process [12].

Fig. 1 represents the process of VA. This diagram shows how a VA system can capture the audio through users, cancel the background noise, implement a voice recognition algorithm, and use an API with Python's help.

A. Audio Capture

The User's Words or sentences act as input spoken to the system (VA). After taking the input in words and sentences, it will try to clear the background noise with the help of Noise Cancellation (NC) or Wiener Filtering method. This way, our system can capture the audio and filter that VC using the wiener filtering method. Then, after converting into text via an online voice recognition technology. The User who wants to give their input as their VC into the system may get a text from a specified bulk arranged on the network server located in the data canter by accessing the microphone and reading the text temporarily stored in the system.

B. Noise Removal Method

Noise is loud, unwelcome, or interferes with hearing. It can also be used to describe an obnoxious hissing electronic noise signal. Any extraneous sound audible while hearing or observing other sounds is called background or ambient noise. Drilling machines, mixers, and electronic equipment are just a few noise sources—unwanted conversations with those nearby, animal noises, and obnoxious noises from the building site.

Most noise removal algorithms subtract those bonds from the original signal. Many noise removal methods use static filters such as lowpass, high pass, and bandpass filters. These filters are designed for a specific parameter to isolate the dominant signal. Sometimes, these filters are ineffective in some conditions, like when the properties of the background noise overlap with the clean signal to be isolated. We have an isolated indication of both noisy speech and background noise. However, we can't subtract background noise for a precise address because of the macroscopic distance between the microphones; both have different environments. Another reason is various factors, including electrostatic charges within hardware components and some vibrations in the atmosphere, because that noise. Wiener filter uses properties of two signals to get a clear speech, and it checks an error which is the mean squared error that is calculated and minimized to produce clean speech.

To get the filtered speech or clean speech, we use the wiener filtering equation, which is

$$X(f) = S(f) + A(f). \quad (1)$$

Where.

$X(f)$ It is a noisy speech signal.
$S(f)$ It is an original speech signal.
$A(f)$ It is an additive speech signal.

$S(f)$ It can be extracted from the noisy signal by multiplying the loud speech signal $X(f)$. With the wiener filter function.

$$S(f) = W(f) * X(f) \quad (2)$$

Here. W(f) represents the winner filter in the frequency domain..

$$W(f) = |S(f)|^2 |S(f)| + |A(f)|^2 \quad (3)$$

In our VA, we remove the background noise using the non-gradient method, which we can see in Fig. 2. The noisy speech part in which it shows the unwanted frequency signal from the User's surrounding.

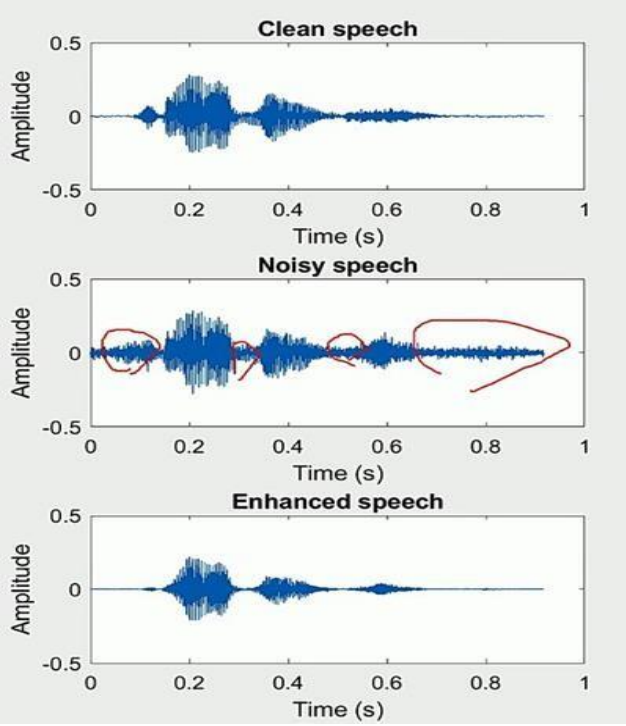

Fig. 2. Checking background noise frequency.

The diagram above depicts the background noise surrounding the system or the User. Noise comes from the User's immediate surroundings, as illustrated graphically. Enhanced speech procedure, shown in the diagram, ensures that we can give input to VA without considering background noise.

C. NLP (Natural language processing)

A technique called NLP enables VA to communicate with people via language. Applied computational linguistics is another name for it. It has something to do with knowledge representation, reasoning, and search. Machines must be able to understand and comprehend human languages to get through the barrier preventing machine-human communication. Information extraction, machine translation, document classification, and information representation are only a few of the many subdomains that it encompasses. The use of NLP applications is required for processing text data. In this case, the text data indicate that the User's speech has been converted to text.

The two appearances of NLP domains that are used in making VA are

Classification. The condition of conversational systems is to find the intent of the spoken Sentence or loosely the action that must be performed.

Information extraction or tagging. It is used to find helpful information from within the Sentence, which is needed to complete the task.

The steps used by NLP in VA are:

1) *Setup the environment:*

First of all, we install these libraries:

Speech Recognition, nltk, gtts, and many more.

2) *In the second step*: we use SpeechRecognition to convert speech into text. It gives the ability to listen and understand the User's voice by transforming the audio signal into text.
3) *After converting speech into text:* VA uses different steps of NLP for further processes to get desired output that the User wants. The first stage is :

a) Segmentation :

In this process, the Sentence is divided into

component sentences, usually along a particular mark.
Example: writing: write

b) Tokenization *:*
In this method, NLP breaks the Sentence or text into its constituent words. Those constituent words are known as tokens.
Example: The sole meaning of life is to serve humanity:
['The,' 'sole,' 'meaning,' 'of,' 'life,' 'is, "too,' 'serve,' 'humanity']

c) Stemming :
It is the process by which tokens are converted back to their original form.
Example: Knows: Know
Knowing: Know

d) *Lemmatization* :
This step works the same as stemming to convert tokens into their base form, but sometimes containing cannot give accurate results. That's why we use lemmatization for more precise results.
Example: He was riding.
He was taking the ride.

e) POS tag :
In this step, the POS tag will check that the base word has what type of parts of speech.
Example: Book the flight.
[("Book," "Verb"), ("the," "Det"), ("flight," "Noun"].

f) Name Entity Recognition *:*
Sometimes one word has two or more POS tags. For this disadvantage, we use Name Entity Recognition to clarify the meaning of that word or for what purpose that is used—some words are used for Company, Location, Name, etc.

4) The above three steps came under automatic speech recognition. After that, we use Natural Language Understanding (NLU). This is the same as Name Entity Recognition. Here, VA tries to understand what the User wants to command. Like, if a user command to open a YouTube channel of some specific person, VA will attempt to know that you are supervising to open a YouTube channel, not that person's profile.

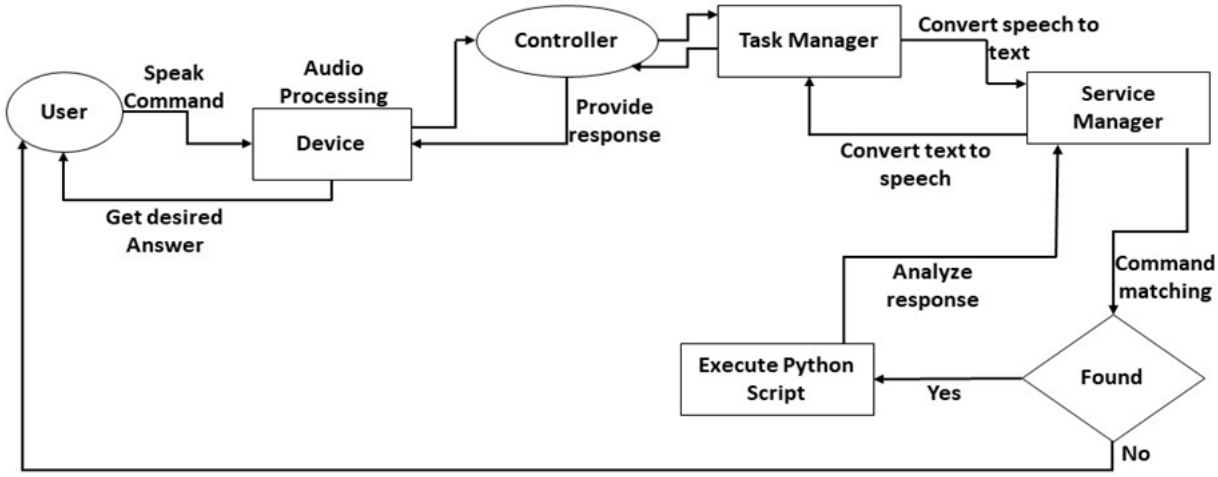

Fig. 3. It shows the detailed process of the VA [12].

5) ***Natural Language Generation(NLG)***
It produces a human language text response based on user VC. In this, NLG can give output to the User in terms of human language or text format; the User can change the voice form to hear the production, like if he wants a male voice or female voice, they can change it by their need.

D. API(Application Programming Interface)
For Application Programming Interface, it is an acronym (API). A piece of software called an API makes it possible for two apps to communicate. In basic terms, it functions as a messenger, delivering your request to the source you are seeking it from and then returning the result to you.

- VA is expected to provide insights into the future of Virtual Personal Assistants for a specific region or locality. The system will keep listening to your commands, and by giving your orders, it will recognize them.
- In Fig.3, you can see the detailed working of VA. If the VA cannot identify or acquire information from a user command, it will continue to inquire until the need for the ideal solution is satisfied.
- For user suitability, VA can have both male and female voices.

- Playing music, sending emails and messages, searching on Wikipedia, chatting on WhatsApp by voice command, building lists, accessing system-installed programs, opening anything on the web browser, etc., are some of the services that our VLA enables.

In Fig.3, the controller processes the command we give to VA to clear the noise. The task manager for converting speech to text, and then the service manager checks whether the command matches. If it matches, then VA executes the Python script file. Else it will move to its starting point. After performing the Python script file, it will provide our optimal output.

PROPOSED ALGORITHM

Algorithm: Human helping hand proposed model

Input: Voice Command
Output: Response

BEGIN
Step-1: Listen to what the User wants to command and then recognize it.
Step-2: Convert speech to text method.
Step-3: A non-gradient algorithm can purify the given input voice command by clearing the background noise.
Step-4: Store the command in text format and analyze the keyword VA needs to execute.
Step-5: Then, the task manager manages the input command. According to the service manager, it can match whether that command matches the preferable output.
Step-6: If the command is not matching, it moves towards the starting state again.
Where VA can take input and say your command does not recognize.
Step-7: Finally, when the command is matched, it will show the User the desired output in voice command, text command, or screen form. It depends upon the User what type of command they will give.
END

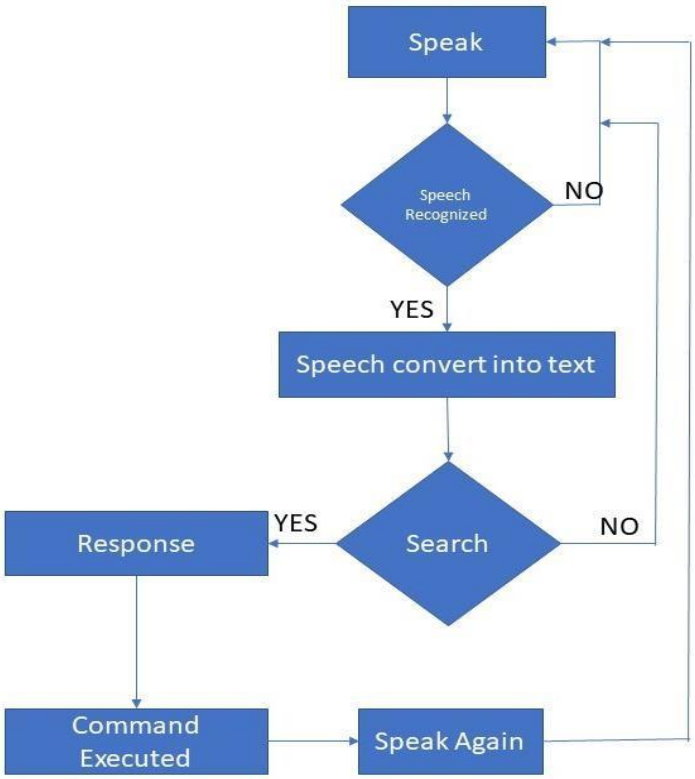

Fig. 4. Detail description of VA using flow chart representation

Fig.4. Shows the Flow chart of VA and the entire process of taking input, recognizing process, and action taken to generatethe output.

RESULT AND DISCUSSION

Fig. 5. Open Calculator Using VA

In Fig.5. The picture shows that we tested our VA by giving the command to open the calculator. This is not limited to a calculator; VA can open any application installedon a computer or desktop.

Fig. 6. Open WhatsApp using VA, where you can text anyone using a simple voice command. Calculator Using VA

In Fig.6. VA is testing for opening WhatsApp. In WhatsApp, you can send messages using your voice command, so we don't have to use our keyboard to type the letters.

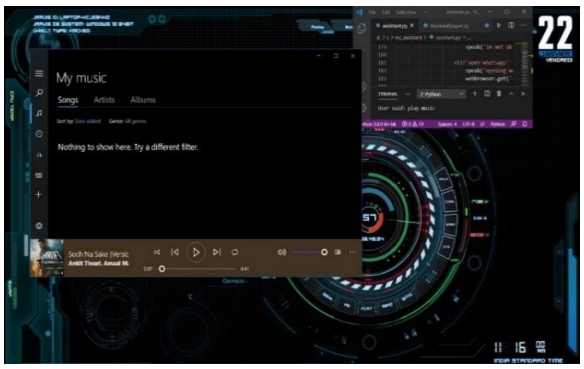

Fig. 7. Open Music Player on the request of the User and Play the music that the User wants or loves more

In Fig.7. it is clear from the above figure that the User gives a command to VA to open the music player. Plays that music in the playlist or search for that song with their name or singer's name to play. You can also interrupt VA while music plays to change the theme.

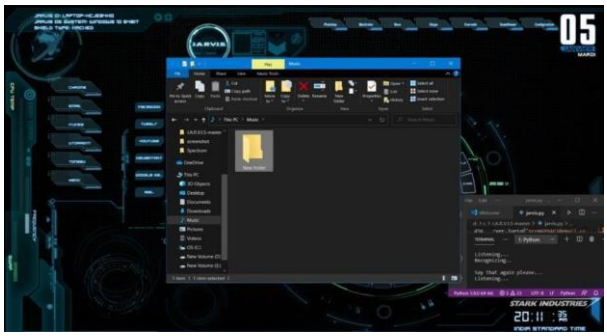

Fig. 8. The process of creating a new folder using VA.

Figure 8 shows the testing of a straightforward voice command to create a new folder. For instance, if a user requests to create a new folder while on their desktop, it will do so and allow them to delete any existing folders as well.

TABLE I . TIME IS TAKEN BY FOLLOWING COMMANDS.

TableNo	CONNECTION TIME TAKEN		
	Commands	***ConnectionLAN***	***Time Taken(In second)***
1	Open Calculator	LAN-1	5 Sec
		LAN-2	9 Sec
2	Open WhatsApp	LAN-1	10 Sec
		LAN-2	14 Sec
3	Open New Search Tab	LAN-1	10 Sec
		LAN-2	12 Sec
4	Created New Folder	LAN-1	4 Sec
		LAN-2	7 Sec
5	Open YouTube	LAN-1	9 Sec
		LAN-2	12 Sec
6	Open iemcrp Test	LAN-1	10 Sec
		LAN-2	15 Sec
7	Remove Current Tab	LAN-1	5 Sec
		LAN-2	5 Sec

In Table I, we check how long a VA can take to open these commands using different LAN connections. For opening the calculator, it takes time 5sec to connect LAN-1. 9sec from connection LAN-2. As it is, we apply the same method for different commands.

CONCLUSION AND FUTURE WORK

This paper aims to reduce the barrier between human and machine interaction. Our proposed algorithm used analog audio to convert speech into English Text. Then it used NLP to convert the user command as an instruction, and our system performs an operation on this instruction and shows some output on the screen. It uses Python, which has many libraries, which are accessible in making this system, and it has many advantages. Therefore, with our paperwork, we mostly try to reduce the communication barrier between humans and machines.

It also uses a non-gradient algorithm to hear the User's pure voice or actual command. By reading this paper, you already know that our system depends entirely on the Internet and only works on laptops or desktops. For future work, our approach is to make it available for mobile phones, i.e., Android or Apple OS, and try to make it offline.

REFERENCES

1. B. Ravenet, M. Ochs, and C. Pelachaud, "Architecture of a socio-conversational agent in Virtual Worlds," 2014 IEEE International Conference on Image Processing (ICIP), 2014.
2. D. J. Patil, A. Shewale, E. Bhushan, A. Fernandes, and R. Khartadkar, "A voice-based assistant using Google Dialogflow and machine learning," International Journal of Scientific Research in Science and Technology, pp. 06–17, 2021.
3. Dr. K V. Kulhalli, Dr.K Sirbi, Mr. A J. Patankar.: Personal Assistant with Voice Recognition Intelligence. International Journal of Engineering Research and Technology. ISSN 0974-3154 Volume 10, Number 1 (2017). International Research Publication House.
4. G. TERZOPOULOS and M. SATRATZEMI, "Voice assistants and smart speakers in Everyday Life and Education," Informatics in Education, pp. 473–490, 2020.
5. V. Kepuska and G. Bohouta, "Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa, and Google Home)," 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), 2018.
6. H. K, J. Zohu, and M. Yang, "Clustering of thread posts in online discussion forums," Extracting Chatbot Knowledge from Online Discussion Forums, vol. 7, no. 1, 2007. (Vol. 7, (pp. 423-428).
7. v Radha and c vimala, "Speech recognition for agglutinative languages," A review on speech recognition challenges and approaches, vol. 2, no. 1, pp. 423–428, 2012.
8. G. Lugano, "Virtual assistants and self-driving cars," 2017 15th International Conference on ITS Telecommunications (ITST), 2017., Warsaw, Poland.
9. v Radha and c vimala, "Speech recognition for agglutinative languages," A review on speech recognition challenges and approaches, vol. 2, no. 1, pp. 423–428, 2012.
10. k v kulhalli, a j patankar, and k sirbi, "Desktop Voice assistant with speech recognition intelligence (DVAbot)," International Journal for Research in Engineering Application & Management, vol. 10, no. 1, pp. 301–305, 2020.
11. S. Srivastava and S. Prakash, "Security enhancement of IOT-based smart home using a hybrid technique," Communications in Computer and Information Science, vol. 6, pp. 543–558, 2020.
12. Ravikumar N R, Prateek C, Sathvik Bhadhar, Rahul Kumar, Mayura D Tapkire. "Virtual Voice Assistant" International Research Journal of Engineering and Technology (IRJET) Vol. 07 Issue: 04 Apr 2020.
13. R. Sarikaya, "The technology behind Personal Digital assistants: An overview of the system architecture and Key Components," IEEE Signal Processing Magazine, vol. 34, no. 1, pp. 67–81, 2017.
14. S. K and P. A. J, "Personal assistant with voice recognition intelligence," 2019 IEEE International Conference on Robotics, Automation, Artificial-intelligence and Internet-of-Things (RAAICON), vol. 10, no. 1, pp. 416–419, 2017.
15. A. S. Tulshan and S. N. Dhage, "Survey on virtual assistant: Google Assistant, Siri, Cortana, Alexa," Communications in Computer and Information Science, pp. 190–201, 2019.
16. M. Chkroun and A. Azaria, "Lia: A virtual assistant that can be taught new commands by speech," International Journal of Human–Computer Interaction, vol. 35, no. 17, pp. 1596–1607, 2019.