Implementation of Virtual Voice Assistant in collaboration of Emotion Detector

Siddharth Singh[1], Sonali Kumari[2], Sunny[3], Dr. Shivani Dubey[4], Dr. Ajay Kumar Sahu[5]
[1,2,3,4,5]*Department of Information Technology, Greater Noida Institution of Technology, Greater Noida, India*

***Abstract*— Realizing natural communication between humans and machines is one of the objectives of artificial intelligence. Many businesses have developed various types of virtual voice assistants based on their usage of conversation systems technology in various apps and domains, such as Microsoft's Cortana, Amazon Alexa, Apple Siri, Computer Assistant, and Facebook's M. The new model will make use of many technologies, including image/video recognition, gesture recognition and speech recognition, to improve interaction between people and robots. In order to sense and affect the environment, a voice assistant incorporates artificial intelligence, speech recognition, machine learning, natural language processing (NLP), numerous actuation techniques, and speech synthesis. The proposed project is designed to associate virtual voice assistant to guide the user by detecting their initial emotions, such as Happy, Sad, Angry, Confused, etc. In our research, we focuses on TTS (Text to Speech) module for the Virtual Voice Assistant and OpenCV for capturing the initial images for the detection of their emotions.**

Keywords—Virtual Voice Assistant, OpenCV, Emotion Detector.

I.INTRODUCTION

Intelligent agents known as spoken conversation systems can assist users in completing tasks more quickly through spoken interactions. Additionally, spoken conversation systems are being added to a variety of gadgets, including smart phones, smart TVs, and in-car navigation systems. A wide range of applications in business, education, government, healthcare, and entertainment can also be supported by dialogue systems or conversational systems. Personal assistants, sometimes called voice assistants, intelligent personal assistants, digital personal assistants, mobile assistants, or virtual personal assistants, go by many names. Numerous firms, including Microsoft's Cortana, Apple's Siri, Amazon Alexa, Google Assistant, Samsung S Voice, Nuance Dragon, and Facebook's M, have employed spoken dialogue systems in the development of their dialogue system devices. These businesses improved and designed their conversation systems using various strategies. Based on the application and the complexity of the VPAs, several design strategies are employed. As an illustration, Google has enhanced the Google Assistant by utilizing the Deep Neural Networks (DNN) [1]. technique, which emphasizes the key elements of conversation systems and the new deep learning architectures employed for these elements. Additionally, Microsoft enhanced the Cortana conversation system by combining the Microsoft Azure Machine Learning Studio with other Azure components.

Moreover. For developers to create applications with incredibly engaging user interfaces and realistic conversational interactions, Amazon offers advanced deep learning functionalities of automatic speech recognition for converting speech to text and natural language understanding to identify the intent of the text. Additionally, Facebook just unveiled Messenger M, its own personal assistant that aims to integrate contextual memory and machine learning. Facebook is using supervised learning to train Messenger's new virtual assistant. Supervised learning is a technique where computers learn by doing what their human instructors show them. All of these businesses are working to increase their competencies in a number of key technologies for their dialogue systems, including text-to-speech, dialogue management, synthetic talking faces, and automatic speech recognition. Additionally, some businesses and researchers have made an effort to enhance their applications by creating the Next-Generation of dialogue systems using the Multi-modal dialogue

technique. Two or more mixed user input methods, including as speech, pen, touch, manual gestures, gaze, and head and body movement, are processed by the multi-modal dialogue. For instance, this system, which includes a touch screen and a speech recognizer, is used in the Ford Model U Concept Vehicle to handle a number of non-essential vehicle functions, including the climate, entertainment, navigation, and telephone. As opposed to the conventional, speech-only, command-and-control interfaces used in some of the cars now on the market, the prototype employs a natural language spoken dialogue interface mixed with an intuitive graphical user interface [2].

Additionally, deep neural network architecture was suggested by Waseda University's Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata that enables multimodal integrated learning of temporal sequences, including visual, audio, and motion. Two activities using a humanoid robot in a real-world setting were used to gauge how well their suggested framework performed. Furthermore, using techniques like adaptive conversational tactics and gradual voice creation, a multimodal conversation system was developed and put into use by Zhou Yu from Carnegie Mellon University to work with users' interest and attention while they were moving.

Three modalities—audio, video and text—are used in this work that leverage automatically extracted features. In this proposal, we suggest a method for developing the next generation of virtual personal assistants that will improve user-computer interaction by utilizing a multi-modal dialogue system that uses techniques like image/video recognition, gesture recognition, speech recognition, conversational knowledge base, sizable dialogue, and a general knowledge base. Additionally, our method will be used to a variety of activities, such as security access control, home automation, robots and vehicles, and solutions for the disabled. The method also incorporates some cutting-edge methods that make this device stand out, such as turning the device into a TV by using the data show or connecting it to a screen, watching TV and movies with translation language, chatting with anyone in any language, deciphering body language and gestures, and playing games with speech and gesture recognition. It may be used to interpret verbal and facial emotions. To change the general model of dialogue systems to multi-modal dialogue systems and to design the Next-Generation of Virtual Personal Assistants with high accuracy, we added some components, such as the gesture model, ASR model, graph model, user model, input model, output Model, Interaction model, Inference Engine, knowledge Base and cloud servers [3].

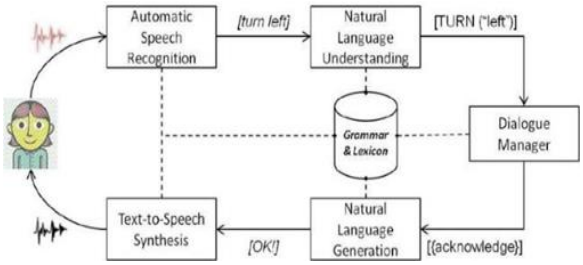

Fig 1. The Structure of General Dialogue System

II.LITERATURE REVIEW

Voice assistants have a lengthy history, and they have made a number of important advances throughout the years. Voice assistant for dictation, search, and voice commands has evolved into a regular feature on smartphones and wearable technologies. The study is based on an insufficient survey of the literature in order to provide broad knowledge (theory and concepts) on voice control, virtual assistants, sectors of application, and other issues. There are several real-world instances of intelligent software on the market today that can process natural language in a variety of contexts. The first voice recognition system, Audrey, was created by Bell Laboratories in 1952. Audrey had little knowledge of technology and could only interpret 10 digits uttered by select people (Pieraccini, 2012). IBM developed and introduced the Shoebox Machine around ten years later.

The device recognized and responded to 16 distinct spoken phrases, including all 10 numerals from "0" to "9," as well as computations like "plus" and "minus" (IBM, 2018). The Shoebox Machine understood and responded to 16 spoken words, including the 10 numbers from "0" to "9," but only when spoken in English by a designated speaker. These limitations later proven to be problematic [4], casting doubt on voice recognition. The HMM fundamentally altered the process of developing a functional speech recognition system. The possibility that sounds may be words was first calculated by speech recognition software using HMM. The idea of

being able to identify an endless number of words has now become near since the approach can raise the number of understandable words to a few thousand. Practically any type of data may be successfully represented due to the possibilities for observation distribution in each model step. When it debuted the Siri virtual personal assistant in 2011, Apple Inc. created the first voice command system that was generally used. 2013 (Bostic). Siri, an intelligent bot, is now a common feature and a significant part of Apple's mobile devices [5]. To act as a virtual assistant for computer users, Zabaware Inc. created the chatbot HAL in a manner similar to this. The bot also uses natural language processing algorithms to converse with the user and record what the user says in an effort to organize the data that has been presented to it [6].

A. Speech Recognition

Various classes may include speech, Continuous Speech, Linked words, Isolated Word, Spontaneous speech Each utterance is normally silent on both sides of the sample frame when isolated words are recognized. It will accept a single word or single speech once. The state of "Listen and Non-Listen" is this. For this class, isolated utterance might be a better phrase. Similar to isolated words, connected word systems enable various statements to be "run together" with a minimum of pause. Continuous speech recognizers let the user speak almost naturally while the computer figures out what they are saying. Because they use a unique technique to identify utterance boundaries, continuous speech recognizers are among the hardest to develop. It might be conceptualized as speech that sounds natural and is unrehearsed at its most fundamental level [7].

Word runs are one example of a natural speech characteristic that an ASR system with spontaneous speech capabilities should be able to manage. Python has a number of voice recognition programs at your disposal. APIAI, Assembly AI, Google Cloud Speech, Pocket Sphinx, Wit and Watson developer cloud. When it comes to usability, the Speech Recognition package excels. Speech should be given as an audio input for speech recognition. Speech Recognition eliminates the need to spend hours creating custom scripts to use microphones and handle audio files. Instead, you can get started using it right away. As a wrapper for several popular speech APIs, the Speech recognition library is extremely scalable.

The Google Web Speech API, which supports all of these, has a hard coded default API key in the speech recognition library. This guarantees that you can begin moving without needing to register for a facility. The Speech Recognition package is a fantastic option for any Python project because of how straightforward and user-friendly it is. However, it cannot be assured that it will support all of the features of the API functions it encapsulates [8].

Finding out how Speech Recognition will function in your particular circumstance will require some time spent researching the different options. The following is a basic example of how these tasks are carried out in Figure 2.

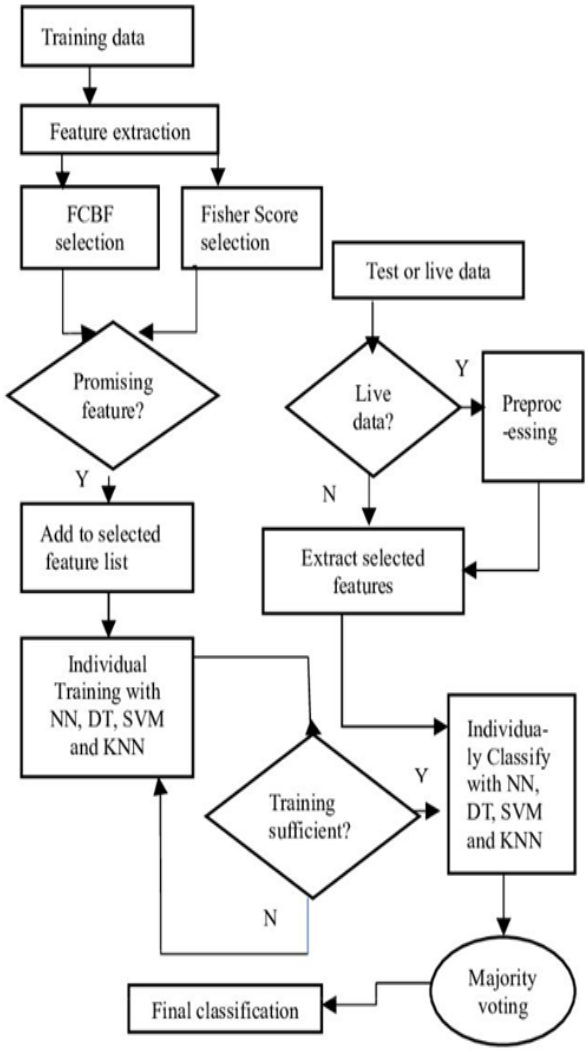

Fig 2. General flow of speech recognition task

B. Image Recognition

As the user converses with the Chatbot, the facial analysis module recognizes his emotions. Once the question is posed, the emotion is recorded and kept in a user profile in the database. It recognizes a variety of emotions, including rage, grief, and happiness. This module's implementation made advantage of the IBM Watson Studio service. The facial analysis aids in identifying the outward symptoms of depression. The face analysis is carried out on live footage that is recorded as the user communicates with the chatbot.

The three services used are Watson Studio (to build the CNN model), Machine Learning (to train the models), and Cloud Object Storage (to store the data sets, trained models, and training results). Convolution Layer (Conv 2D), Image data, ReLU, and Max Pooling (Pool 2D) are the layers of a convolution Neural Network constructed with model networks from IBM Watson Studio [9]. The IBM Cloud Storage instance connected to the facial analysis module is represented by this node. The testing dataset, training dataset, and validation dataset are the three buckets that make up Cloud Object Storage. Images from FERC-2013 and contributions from volunteers from our college make up the training data set. The network's kernel is a crucial component. The information encoded in the pixels is changed by applying this to the entire image. Activation maps, which identify the activated regions, are then obtained by convolving the kernel with the input value.

In essence, the kernel computes the dot product between the kernel matrix and the patches selected from the image with size equal to the kernel. The dot product's resultant terms are added up to create a single entry in the activation matrix.

Stride is the quantity of pixels that move across the input matrix. When the patch is selected and moved (to the right or downward until the matrices border is reached), this value is utilized. Up until the full input picture has been processed, the procedure is repeated [10]. The kernel occasionally does not fit the supplied picture correctly. Therefore, one may either eliminate the area of the picture where the kernel does not fit or pad it with zeros. The first layer to utilize a mathematical operation that requires an image matrix and a kernel as inputs is the convolution layer. This layer retains the connection between pixels and is used to extract characteristics from the image.

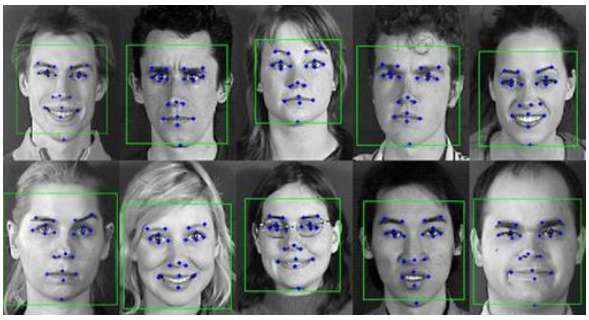

Fig 3. The dataset containing different facial expressions

III.METHODOLOGY

The Natural Language Processing algorithm is used by the Virtual Voice Assistant. The system's ability to operate offline, or without an internet connection, is its most crucial feature [11]. Due to the fact that it only accepts voice instructions as input, this software will assist users in saving time.

The system basically works in following phases:

- Image Recognition
- Emotion Detection
- Speech Input
- Speech Preprocessing
- Task Execution
- Text to Speech

Voice commands from the user serve as input. Waves carrying the speaker's speech are picked up by the listener. There is definitely unwanted background noise and room reverberation in the vocal input. We therefore process the spoken input to obtain the user's context in order to eliminate this noise. Speech preprocessing is crucial in removing unrelated causes of variance. In the end, it increases speech recognition's precision [12]. The preprocessed speech input is used to extract features from it.

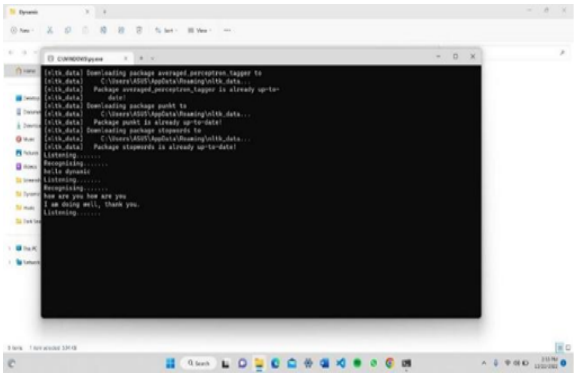

Fig 4. Sample of execution in backend

The basic UI is being designed using Tkinter. Tkinter is the standard Python GUI library. When paired with Tkinter, Python offers a quick and simple way to build GUI applications. Tkinter offers a versatile object-oriented interface for the Tk GUI toolkit [13].

Fig 5. GUI based output screen

Using arrays, the speech will be divided into its unique, pertinent words. There would be a specific dataset with keywords for each command category (e.g.: - open, play). Utilizing loops will be used to match. The necessary work will be completed if a match is discovered. opening a document, for instance and if not, using the Selenium Web Driver, such a keyword does not match present the first 3–4 most relevant results from a Google search as web pages.

The finished product will be made available via the cloud, possibly Google Cloud [14].

This module's implementation made advantage of the IBM Watson Studio service. The facial analysis aids in identifying the outward symptoms of depression. Real-time video that is recorded when the user interacts is used for the facial analysis.

IV. OVERVIEW OF THE SYSTEM

The steps of the overall system design are as follows:

1. Speech-based data collecting.
2. Text conversion and voice analysis.
3. Processing and storing of data.
4. Producing voice from the result of the text processing.
5. Recognize the initial facial expression.
6. Works according to the emotion detected.

Speech data is gathered in the first step and saved as an input for processing in the second phase. In the second step, speech-to-text technology is used to continually process and transform the input voice to text. The transformed text is then examined using Python data script and NLP techniques to determine the appropriate action to execute in response to the command. Finally, output is produced from a straightforward text to speech conversion utilizing text to speech after the response has been recognized. Even though speech recognition offers a number of advantages, it also has many drawbacks [15].

Implicitly, the creation of speech recognition software likewise carries these constraints. The current voice assistants rely on Python pattern recognition techniques, which have limitations in terms of context, accuracy, and misinterpretations, as well as time, expenses, and productivity. They only function in online mode. They keep the data on database servers, which increases the complexity of time and space. Some of them keep the data in the cloud, which raises security concerns. Interference from background noise is yet another difficult issue with speech recognition software.

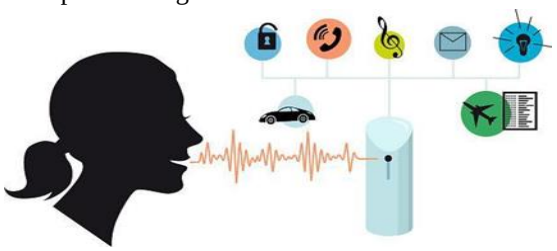

Fig 6. Proposed system diagram

When a user registers, a user profile is generated. The results of each question, together with any emotions that were discovered, are stored in each user profile. Base64 format is used for the captured image. Following the conversion, this is sent to the facial analysis module in PNG format. The database then contains the emotion that was discovered. The two main requirements for using this program are familiarity with the application's instructions and fundamental English communication abilities (High School). Outcomes of tests employing face analysis on a few samples. Following the full screening procedure, the patients receive suggestions depending on their score. The algorithm classifies emotions into three categories—angry, joyful, and sad—using 697 photos as the training and testing dataset. 15% of the 697 photos were utilized for testing, while 85% were used for training. After removing specific face traits, the facial analysis module determines the user's emotional state. For each of the three emotions it notices, it assigns a score in accordance with that [16]. The emotion with the

highest rating is taken into account. There were 100 test photos utilized. For the assessment of joyful, sad, and angry emotions, respectively, 40, 30, and 30 photographs each were employed. A mistaken emotion classification was made for 16 photos. For the 100 training photos employed, the accuracy of the Facial Analysis Model is determined to be 86%.

Sl.no	Image	Expected Emotion	Actual Emotion
1.		Angry	Angry
2.		Happy	Happy
3.		Happy	Angry
4.		Happy	Sad
5.		Sad	Sad

Fig 7. Emotion Detection Test Results

CONCLUSION & FUTURE SCOPE

This proposal describes the next generation of virtual personal assistant coherent structure and new virtual voice assistant technology created to communicate with people. This virtual voice assistant system uses speech, video, photos, gestures, and other forms of communication for input and output. Through the use of speech recognition, gesture recognition, image/video recognition, and the Knowledge Base, the VPAs system will also be utilized to enhance user-computer interaction. Although by using the huge discourse knowledge base, this system may hold a lengthy discussion with users. This system can be utilized for a variety of functions including security access 102 control, home automation, robotics and vehicles, and aid with impairments. It can also be a good solution for applications like customer service, education, training, transaction facilitation, travel information, online shopping, counselling, tutoring systems, remote banking, ticket booking, travel reservation, stock transactions, information inquiry, taxi bookings, path planning, etc. The future enhancement of the system is to improve the number of emotions detected. People with disabilities may find this kind of program useful for making daily tasks easier. They can move forward with tasks that they might find challenging to do on their own with the help of basic orders in their native tongue. According to the study, we suggest developing an application that meets the demands of diverse users. The user primarily wants to use the voice assistant to make their lives easier, thus by incorporating the features stated below, the user may be assisted.

1. Creating content for various tongues and accents.
2. Versatility in any setting.
3. To increase security, voice authentication technology can be used.
4. Implementing a chatbot requires a corpus.
5. Dialog flow requires neuronal stacking
6. Use Flask or Django to deploy to the web
7. Cloud deployment using Heroku and Amazon EC2.
8. NLP features including topic modelling and entity detection.
9.

REFERENCE

[1] DOUGLAS O'SHAUGHNESSY, SENIOR MEMBER, IEEE, "Interacting with Computers by Voice: Automatic Speech Recognition and Synthesis" proceedings of THE IEEE, VOL. 91, NO. 9, SEPTEMBER 2003

[2] Nil Goksel-Canbek2 Mehmet EminMutlu, "On the track of Artificial Intelligence: Learning with Intelligent Personal Assistant" International Journal of Human Sciences, 13(1), 592-601. doi:10.14687/ijhs. v13i1.3549.

[3] Easwara Moorthy, A., Vu, K.-P.L.: Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. International Journal of Human-Computer Interaction 31, 307–335 (2015)

[4] Tsiao, J.C.-S., Tong, P.P., Chao, D.Y.: Natural Language Voice-Activated Personal Assistant, United States Patent (10), Patent No.: US 7,216,080 B2 (45), 8 May 2007

[5] S. Arora, K. Batra, and S. Singh. Dialogue System: A Brief Review. Punjab Technical University.

[6] R. Mead. 2017.Semio: Developing a Cloud-based Platform for Multimodal Conversational AI in

Social Robotics. 2017 IEEE International Conference on Consumer Electronics (ICCE).
[7] K. Noda, H. Arie, Y. Suga, and T. Ogata. 2014.Multimodal integration learning of robot behavior using deep neural networks. Elsevier: Robotics and Autonomous Systems.
[8] R. Pieraccini, K. Dayanidhi, J. Bloom, J. Dahan, M.l Phillips. 2003. A Multimodal Conversational Interface for a Concept Vehicle. Euro speech 2003.
[9] "Word stream." https://www.wordstream.com/google-now.
[10] T. Guide, "Tom Guide." https://www.toms guide. com/us/amazon-alexa-faq,review- 4016.html.
[11] "The Ultimate Guide to Speech Recognition with Python." https://www.selenium.dev/ documentation/en/webdriver/ understand ing_the_ components/.
[12] "Understanding the components." https://www.selenium.dev/documentation/en/webd river/ understand ing_the_components/.
[13] "What is Microsoft Azure Platform-as-a-Service (PaaS)?" https://www.sherweb.com/blog/cloud-server/what-is-azure-paas/.
[14] Veton Kepuska and Gamal Bohouta, "Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa, and Google Home)", 2018 IEEE.
[15] Pubudu M. Dias and Kithsiri Jayakody "Virtual Assistant in Native Language", 2021 IEEE.
[16] Nadja Damij and Suman Bhattacharya, "The Role of AI Chatbots in Mental Health Related Public Services in a (Post) Pandemic World: A Review and Future Research Agenda", 2022 IEEE.