Virtual U: Robust Voice Assistant using Gaussian Mixture Model

Abhishek Malviya
Computer Science & Engineering
United Institute of Technology Prayagraj, India abhishekmalviya2050@gmail.com

Kritika Mishra
Computer Science & Engineering United Institute of Technology Prayagraj, India
mishrakritika2001@gmail.com

Amit Kumar Tiwari
Computer Science & Engineering United Institute of Technology
Prayagraj, India
kumartiwariamit@gmail.com

Adarsh Rai
Computer Science & Engineering United Institute of Technology Prayagraj, India adarsh.21rai@gmail.com

Aditya Srivastava
Computer Science & Engineering United Institute of Technology Prayagraj, India
adityasrivastava.united@gmail.com

Anamika Saini
Computer Science & Engineering United Institute of Technology Prayagraj, India
anamika08saini08@gmail.com

Abstract—Today the world is full of rush, and hustle. Desktop assistants are life-changing innovations in the world of Artificial Intelligence. These assistants help the user to use their desktops in touch free manner. Like Siri, Cortana, Google Assistant, Alexa, etc. there are other assistants in the market. Today's worlds need automation and fast work with the least manual input. The voice assistant aims to develop an intelligent virtual assistant that interacts with users through voice commands, providing a seamless and efficient user experience on desktop platforms. The assistant employs natural language processing (NLP), speech recognition, and machine learning techniques to understand and respond to user queries, perform tasks, and offer relevant information. A voice assistant is a tool that helps you to tackle this problem and control your system with your voice. When this assistant is combined with the security then it achieves the power of identifying the master among multiple speakers. This paper presents an overview of basic functions and common features of our Desktop Voice Assistant: Virtual U.

Keywords - Desktop Voice Assistant, Natural Language Pro- cessing(NLP), Gaussian Mixture Model, Mel Frequency Cepstral Coefficients (MFCC).

Introduction

The advent of voice assistants has revolutionized the way we interact with technology. With the rise of smart homes and the Internet of Things, voice assistants have become an integral part of our daily lives. However, with the rise of smart homes and the Internet of Things, there has been a growing demand for desktop voice assistants as well [1]. These intelligent assistants offer a convenient and hands- free way to interact with a computer, making tasks such as sending emails, scheduling appointments, and searching the web much easier. The Desktop Voice Assistant Virtual U is a revolutionary technology that utilizes voice recognition intelligence to process user input in the form of spoken commands. It can perform various actions and return outputs such as search results, all through the convenience of voice- based interaction. This technology is designed to enhance user experience

and streamline workflow, making it an essential tool in today's fast-paced digital landscape. The main objective of this project is to develop a voice assistant using Python that can perform operations such as copying and pasting files, sending messages to a user's mobile device, and ordering pizza using voice commands. This project is specifically designed for physically challenged individuals to aid in their daily tasks. [2]

I. Related Work

There are many existing solutions for voice assistants across various platforms, including IOT devices, mobile devices, and home automation systems. Some of the most popular voice assistants include Google Assistant, Siri, Alexa, and other voice assistants should be appreciated by everyone.

A. Google Assistant

This is a voice assistant developed by Google and is available on Google Home devices and other smart speakers. This Voice assistant is activated by a Hot word "Hey Google" or "Ok Google" and then providing a voice command. It allows users to get information, control their devices, and perform tasks using their voice. It supports multi-language recognition and can understand a wide range of accents.

B. Amazon Alexa

Amazon Alexa is a voice assistant developed by Ama- zon that is available on various platforms, including smart speakers, mobile devices, and home automation systems. It uses natural language processing and machine learning to understand voice commands and perform various tasks, such as setting reminders, playing music, answering questions, and controlling smart home devices.

C. Apple Siri

This is a voice assistant developed by Apple and is available on iPhone, iPad, and other Apple devices. This Voice assistant is activated by a Hot word "Hey Siri" and then providing a voice command. It allows users to perform tasks, get information, and control their devices using their voice.

Literature Review

The desktop assistant was found to be one of the prominent use cases. The old wish came to a reality with the advancement in desktop voice assistants. These voice assistants change the way a normal user interacts with the computer by offering a hands-free voice operating system.

As humans, we have long dreamed of the day when we could simply speak to our computers and have them execute our every command. With the advent of desktop voice assistants, that dream is becoming a reality. These virtual assistants are revolutionizing the way we interact with our desktop computers, offering a hands-free, voice-activated alternative to traditional input methods.

As humans, we have long dreamed of the day when we could simply speak to our computers and have them execute our every command. With the advent of desktop voice assistants, that dream is becoming a reality. These virtual assistants are revolutionizing the way we interact with our desktop computers, offering a hands-free, voice-activated alternative to traditional input methods.

Recently many companies have invested in the field of voice assistants. Companies like Google, Amazon, IBM, Apple, and others are training people in the field of natural language processing, programming, and analytics which are the subfields of Artificial Intelligence (AI). Some of the revolutionary products can be seen in daily life. Google's voice assistant is provided for Android phones which we can access via the Internet. The assistant can make calls, news, and video recommendations. According to Nilsen Norman's research on the use of desktop assistants, the feedback from mixed native people was not satisfactory. They gave desktop assistants like Alexa and Google Assistants to a 12 people group for their usage. But problems were unease to explain the details, cannot understand fast-flow languages. Companies have recently improved their product.

Although these assistants are available easily in the market but not for Windows users. Microsoft has worked on desktop assistants and gave outstanding results which is popularly known as Cortana . It can be used for managing alarms, calendars, and opening applications on our software . But if it is then why don't we use it? It has played an important role in helping the user with daily activities, but it faces hard competition from Google Assistant, Siri, and Alexa. Some users have complained that noise cancelation and user voice understanding with clarification is still a problem in Cortana. Eavesdropping is also a threat to Cortana because it cannot identify who is speaking to it. Whether it is the real user or some other. And yes, hacking into it is also a big

problem [3]. Projects have been done on developing a personal assistant for PCs. Major problems have been focused on enhancing the understanding power of the desktop voice assistant. Different machine learning algorithms like HMM, GMM, CNN, and SVM are used in speech recognition which gives a drastic improvement in these voice assistants. One more focus is on the authentication capability of desktop voice assistants.

During our research on available projects in this domain, we came across the use of the gTTS library for speech 2 recognition and text output [4]. Also, authentication is still a problem. The voice assistants control our whole computer and so they might have access to the sensitive information in our system. The need for voice assistants for authentication is very important.

Proposed System

Voice Assistants are designed to assist the user in doing some simple tasks like alarm setting, browsing content on the web, and dictation. Assuming you are commanding our laptop/personal computer just our voice is fantastic. Virtual U is also a voice assistant that helps the master to do daily tasks automatically. It can talk to the master like a friend, browse the content of any topic on the web, YouTube video play, alarm management, and schedule messages for WhatsApp...Virtual U has the special feature of master identification also. This feature is related to the security purpose of this standalone application. The uses Python language and machine learning algorithm and Microsoft API for implementing its objective.

Methodology

Developing a Desktop voice assistant involves several key steps and methodologies. Here are some of the common steps that can be followed:

A. Machine Learning model

In our Desktop Voice Assistant: Virtual U we have used the Gaussian Mixture Model (GMM)` which is an unsupervised clustering algorithm. GMM model is a special use of (HMM) assuming independent frames [5]. It is a probabilistic approach for identifying the speaker. The feature extraction technique is Mel Frequency Cepstral Coefficients (MFCC) as it provides accurate feature extraction like human ears. [6]

1) Preprocessing of Voice data: The users must first speak to get entry into the system and take functions from virtual u. At first, the user's voice goes through testing for authorization. If the speaker passes this section, then only it gets authenticated to use the system functionalities. But the computer cannot directly process the audio signals from the microphone. It is stored in .mp3 files. This file is then used and converted into wav format for feature extraction. After entry into the system when we want to retrain the model the recording of the voice sample is done, and preprocessing is done. During the preprocessing, the Mono channel is used as it reduces the noise, the chunk size is 512, rate of sampling is 44000 for 20 seconds. By these configurations, the audio file becomes preprocessed and ready for feature extraction.

2) Feature Extraction: Feature extraction is a technique to know the various parameters of the audio files which are then analyzed to determine the correct speaker. Mel-Frequency Cepstral Coefficient is a popular feature extraction technique from audio files. Mfc stands for Mel Frequency Cepstrum which is the short-term power spectrum of sound. Mel scale is a scale of pitches being converted into a mel spectrum of sound. In this scale sound above 40db is considered as the listener's threshold [7]. MFCC coefficients are a set of DCT decorrelated parameters, which are calculated by transformation of the logarithmically compressed filtered output energies, derived through a perceptually spaced triangular filter bank that processes the Discrete Fourier Transformed (DFT) speech signal. MFCC which maps the signal onto a non-linear Mel- Scale that mimics human hearing [8]. MFCC extraction is done by a common function provided by python speech features library of python. Some of the parameters used in Virtual u are signal(audio signal), samplerate ,winlen (25 milliseconds), winstep (10 milliseconds), nfft. [9]

Algorithm for feature extraction:

1) Sample the speech at 44,100 Hz, successive window overlaps as 0.01.
2) Take the length of each window as 0.025 seconds and fast Fourier components as 1200.
3) Windowing is done to divide the large continuous voice into samples. Since adjacent voice samples can have similarities overlapping windows are

chosen.

4) Now discrete Fourier transformation is applied on time series voice signal to convert it into frequency series data as these are easy to extract features.
5) Apply the mel filter banks. Here we apply the mel scale to map the frequency to that one which human perceive by formula [10]: mel(f) = 1127log(1+f/700) where f is frequency of voice signal in each window.
6) Now you have a feature vector of MFCC which are nothing but first and second order derivatives which can be used in model training.

3) Model Training: Gaussian Mixture Models (GMM) are used as underlying technique, because this mode has a good recognition capability . GMMs are commonly used as a parametric model of the probability distribution of continuous features like vocal tract related spectral features. The MFCC feature map is given as input to GMM model for parameter estimation. According to we used delta mfcc feature map for gmm model which give 90.68efficiency in identification. Its parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum. Python language provides a powerful library called sklearn.mixture which has model Gaussian Mixture

. The model is fitted by max iter number of times till significant changes are seen in parameter values. During model training we give the number of gaussian components, maximum number of iterations and covariance type. Now the trained model i.e., the model of each speaker is saved in folder.

Expectation-Maximization Algorithm

1) For GMM models let there be n training vectors x1. . . ..xn. So now the GMM likelihood is given as - where is GMM parameters and x is input features [11].

Since GMM likelihood is

$$L(\lambda/X) = Pr(X/\lambda) = \prod_{k=1}^{k=n} Pr(x_k/\lambda) \quad (1)$$

$$= Pr(x_1/\lambda)Pr(x_2/\lambda)...Pr(x_n/\lambda) \quad (2)$$

non-linear function of lamda so direct

- **wave** :- wave is a python library which is used to store voice samples in wav format. The problem of working the voice in digital manageable formats is solved by the wave module. [12]

maximization is not possible.

2) E-step: Here the initial value of lamda is estimated. Using this the likelihood is calculated.
3) M-step: Now the data of the estimation step is used to update the parameters of gmm likelihood.
4) Repeat steps 3 and 4 till convergence is reached. This is the state where maximum likelihood is reached.

B. Python Backend Model

In this part, the libraries and methods are used. The Desktop Voice Assistant is implemented using the Python language.

- **pyttsx3** :- pyttsx3 is a text-to-speech conversion library in Python. It can work offline and is supported by python 2 as well as 3. Virtual U invokes the init() to get a reference to an Engine instance. It is a very easy to use tool which converts the entered text into speech. The pyttsx3 module gives index-0 voice as female voice and index-1 is male which is provided by “sapi5” for windows. [12]
- **Datetime** :- Datetime is a python module which is used to work with dates and time. This module is preinstalled and provides classes and functions to work with date and time. [12]
- **Wikipedia** :- Wikipedia is a Python library. It is for accessing and parsing data from Wikipedia, searching Wikipedia, getting article summaries, getting data like links and images from a page, etc. Wikipedia uses the MediaWiki API. [12]
- **webbrowser** :- The webbrowser module is an easy web browser controller. It provides an appropriate interface which gives output of the Web documents to users. It accepts a URL as the argument. [12]
- **OS** :- OS module in Python provides functions for coordinating with the operating system. OS is included in standard utility modules. This module provides a way of using operating system functionality. [12]
- **pywhatkit** :- pywhatkit is a Python library to ease our work. Among them for sending whatsapp messages to know contacts, browsing youtube videos, doing google searches, and getting information on a particular topic from the internet. [12]

- **pickle** :- pickle module uses binary rules and regulations for serializing and de-serializing a Python object struc- ture. The process where a Python object hierarchy is con- verted into a byte stream is pickling, and is unpickling is the inverse

operation, whereby a byte stream is converted back into an object hierarchy. [12]

- **warnings** :- warnings are used to handle the errors during the training and testing of the Machine Learning Model.
- NumPy, SciPy, Sklearn and sklearn.mixtures are some machine learning libraries where the voice verification is done.
- **Gaussian Mixture Model** - The Gaussian Mixture Model is used in virtual u. This model uses the Mel-Frequency Cepstral Coefficients of the voice to compare the voice in a pretrained model. This Model is a probabilistic clus- tering model. The model is trained using unsupervised learning and classes of each person's voice are created. During testing of any new user, the mfcc extraction of voice is done and given a Gaussian mixture trained model and probability of authenticity is calculated. The model with highest probability is given as a result of prediction. Hence if an authentic user speaks then the virtual u will enable itself for further operations otherwise it will not let the user use it.

Result & Analysis

The vital modules and packages of Python programming language have been installed and the code was implemented using Visual Studio Code IDE. Below are some outputs of the Desktop Voice Assistant: Virtual U.

A. *Opening/Closing of an Application*

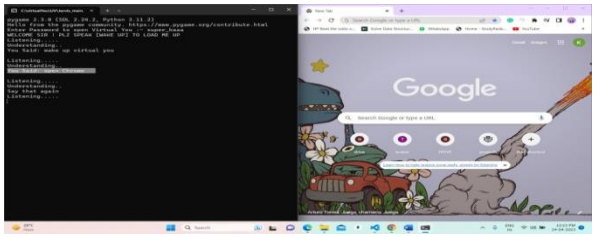

Fig. 1. Opening Google Chrome by Voice Command

B. *Setting an Alarm*

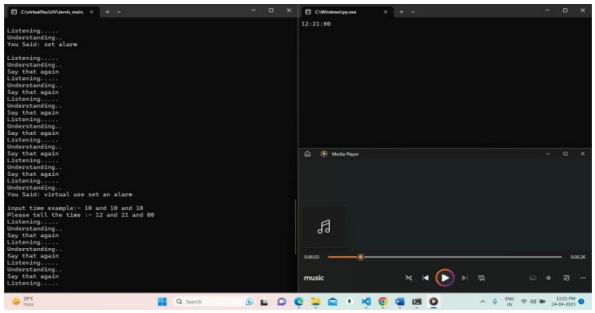

Fig. 2. Setting up Alarm by Voice Command

C. *Searching through Browser*

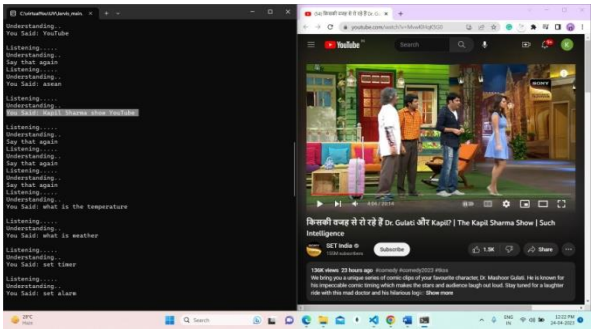

Fig. 3. Searching through YouTube in Browser by Voice Command

Comparative Result

Desktop Voice Assistants are becoming increasingly popular among users for their ability to perform various tasks using voice commands. These assistants can be divided into two categories: those that do not include authentication through voice and those that do include authentication through voice. There are a number of Desktop voice assistants that don't include authentication through voice and are designed to do tasks, such as scheduling appointments, browsing, weather reporting and playing music, etc. Examples of these voice assistants are Siri, Google Assistant, and amazon Alexa. These assistants are widely used by people and businesses for their ease of use and ability to streamline tasks. On the other hand, Virtual U (Desktop Voice Assistant) includes authentication through voice. It is designed to provide one more layer of security to users by verifying their identity using voice recognition technology. Therefore, Virtual U can play an important role for businesses and individuals who do transactions with sensitive information. Our virtual u is fit for both entertainment and business purposes.

Conclusion

With the advancement of technology in today's era, AI- powered voice assistants are becoming more frequent in our daily lives. A Voice Assistant can improve our efficiency, overall quality of life, and productivity. This paper introduces a Voice-based Virtual Assistant designed for Windows OS. In this system, we have integrated both Authenticated speech recognition and speaker recognition technology. Virtual U uses voice communication mode to interact with people. This Desktop Voice Assistant aimed to create an integrated version of authentication and provide users with personalized access.

Furture Work

Voice assistants integration with haptics which is the study of perceptions in through sense can give the way to develop voice tutors which can be a robust system to create assistant tutors. Virtual U is wide and promising, especially in areas where security is of the most importance. Voice assistants acquiring visual information of the user and its surroundings gives way to more functionalities like fall detection, dangerous activities detection. Banks and financial institutions are to be the primary users of this technology. Users can authenticate themselves by their own voice as their identity. Thus, reducing the requirement for PINs, passwords, or other forms of iden- tification like biometrics, retina scan, etc. This not only saves time for an individual but also removes the risk of fraud, as voice recognition technology is highly accurate and difficult to copy.The virtual u is a basic voice assistant in which we attempt to give authentication by voice. The more scope can be added to it by give it some other features like mood analysis or sentiment analysis which is used in chat bots to provide better user experience.

References

[1] Gaurav Agarwal, Harsh Gupta, Divyanshu Jain , Chinmay Jain , Prof. Ronak Jain , Department of Information Technology, A.I.T.R, Indore, Madhya Pradesh, India. Assistant Professor, Department of Information Technology, A.I.T.R, Indore, Madhya Pradesh, India, "Desktop Voice Assistant" IJRNETS, Volume 2, Issue 5, May 2020

[2] Smita Srivastava, Dr. Deves Katiyar, Mr. Gaurav Goel, Department of Computer Applications, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India, "Desktop Virtual Assistant" IJRASET 2022

[3] Patjoshi, Rajesh. (2020). Desktop Assistant Based on Voice Recognition and Face Detection. International Journal of Grid and Utility Computing.
13. 2020-2030.

[4] Subhash S, Prajwal N, Siddhesh S, Ullas A, Santosh B Department of Telecommunication Engineering Dayananda Sagar College of Engi- neering Bengaluru, India "Artificial Intelligence-based Voice Assistant", 2020 IEEE.

[5] Hamidia, Mahfoud and Zenati, Nadia and Belghit,Hayet and Guetiteni, Kamila and Achour, Nouara(2015). Voice interaction using Gaus- sian Mixture Models for Augmented Reality applications. 10.1109/IN- TEE.2015.7416773.

[6] Hossan, Md and Memon, Sheeraz and Gregory, Mark. (2011). A novel approach for MFCC feature extraction. 1 - 5. 10.1109/IC- SPCS.2010.5709752.

[7] G. Iannizzotto, L. L. Bello, A. Nucita and G. M. Grasso, "A Vision and Speech Enabled, Customizable, Virtual Assistant for Smart En- vironments," 2018 11th International Conference on Human System Interaction (HSI), Gdansk, Poland, 2018, pp. 50-56.

[8] O. Portillo-Rodriguez, C. A. Avizzano, A. Chavez-Aguilar, M. Raspolli,
S. Marcheschi and M. Bergamasco, "Haptic Desktop: The Virtual Assistant Designer," 2006

2nd IEEE/ASME International Conference on Mechatronics and Embedded Systems and Applications, Beijing, China, 2006, pp. 1-6.

[9] v, Geetha and Gomathy, C K and Kottamasu, Manasa and Kumar, Nukala. (2021). The Voice Enabled Personal Assistant for Pc using Python. International Journal of Engineering and Advanced Technology.

[10] Vrushali Kolte1 , Kalyani Kasar1 , Samidha Jadhav1 , Sunil Rathod, Department of Computer Engineering, Dr. D. Y. Patil School of Engi- neering, Pune, Maharashtra, India "IVA:An Intelligent Virtual Assistant System Implementation and Speaker Recognition", International Journal of Scientific Research in Computer Science Engineering.

[11] Kapoor, Ayush and Hemani, Harsh and Sakthivel, N. and Chaturvedi, S.. (2015). MPI Implementation of Expectation Maximization Algorithm for Gaussian Mixture Models. Advances in Intelligent Systems and Computing.

[12] Vishal Kumar, Dhanraj Lokeshkriplani, Semal Mahajan, Dr.Akhilesh Das Gupta Department of Information Technology Institute of Tech- nology And Management "Research Paper onDesktop Voice Assistant", International Journal of Research in Engineering and Science (IJRES) volume 10, issue 2