IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Authenticated Voice Assistant for Desktop Applications

[1]Dr. Yogita Bhise, [2]Evangeline Rebello, [3]Ruchika Kahane, [4]Samruddhi Badgujar, [5]Kranti Sonawane

[1]Asst. Professor, Department of Computer Engineering

[2,3,4,5] Student, Department of Computer Engineering

[1,2,3,4,5]K. K. Wagh Institute of Engineering Education and Research, Nashik, India

***Abstract:*** In today's exponentially growing data-centric industry security is the most important factor. Considering that, authentication that is based on biometrics traits is more secure for many applications. The aim is to build a model based on voice authentication which would recognize the verbal commands of the respective user to perform tasks on desktop applications, for example – sending emails and WhatsApp messages, performing file operations, etc. just like present systems, but there would be an additional layer of security in which it will only recognize the specific voice and thus follow the commands of the authenticated user. As the existing systems have the drawback of not authenticating the user's voice and directly functioning on any input voice, so to add a security aspect and maintain confidentiality we have proposed this model which would authenticate the user's voice and also address the issues where the existing voice-based systems fail such as authenticating pre-recorded voice of the authenticated user. The main focus would be on increasing the security level and accuracy of biometric authentication. The phases involved in this system are Voice authentication and Desktop voice assistance. The implementation will consist of feature extraction, model building, and model evaluation. The next part is the Voice Assistant on Desktop Applications. This includes listening to spoken words and identifying them. The system takes voice and converts spoken words into text. This text is then fed as a query or the result is obtained and the system generates a reply. It uses a speech recognition library for performing speech recognition. This project would be the first step leading toward an era of secured voice-based assistance where confidentiality is a major requirement.

***Index Terms*** **- Machine Learning, Communication, Voice Recognition, Speech Recognition, Security and Biometric Authentication**

## I. INTRODUCTION

Verbal communication is one of the most preferred forms of communication for humans due to its convenience. The aim is to build a model based on voice authentication and then recognize the verbal commands of the respective user to perform tasks on desktop applications. This would be the first step leading toward an era of secure voice-based transactions in fields where confidentiality is a major requirement. This will also overcome the drawbacks of existing voice assistants. Nowadays, the market is filled with voice-based assistants which follow every user's command. This somewhere hampers the security of users in terms of confidential tasks. Smart Devices are becoming popular in the world today, these are devices that can effectively interact with humans. Human Interaction with technology is the basis of research and study in the field of biometrics. The concept of biometrics refers to verifying if the human is indeed the authorized identity. There are several techniques used for biometric authentication. This field has been used in various applications for a very long time but the most recent advancements are using face, fingerprint, iris, hand geometry, and voice for recognizing the identity of the user. Face and Iris scanning are popular but not convenient as the person needs to constantly be in front of the camera, whereas iris and hand geometry require the user to hold the device physically. The voice does not have such limitations and is more convenient. Voice is the most used source of communication with machines and authentication of users due to convenience. Through this project, the motive is to provide an additional layer of security in which it will only recognize the specific voice and thus follow the commands of the authenticated user. It also focuses on preventing people from using recorded audio files of authorized users' voice for gaining authorization. The system has been trained on different pitches and accents of a particular user. Thus, building an AI and ML-based voice assistant using the required algorithms and engines. The main focus would be on increasing the security level and accuracy of biometric authentication.

## II. LITERATURE REVIEW

The system is based on two parts namely voice authentication and voice assistant. Different pre-processing, feature extraction, and machine learning techniques using MLP and present are needed to evaluate their efficiency. The noise is an important feature that continues to exist. Complete or almost 'close to complete' elimination of noise can greatly improve speaker recognition [1]. One of the major issues concerned while working with human audio is the dialect variation that happens in the same language at the time of pronunciation. The system needs to focus on identifying voice characteristics to uniquely recognize a speaker, independent of what has been said. [2] The voice assistant will take input of the user's voice, and then using speech-recognition module will convert the audio into text in English using GTTS. But, GTTS has a drawback that it doesn't work without an active internet

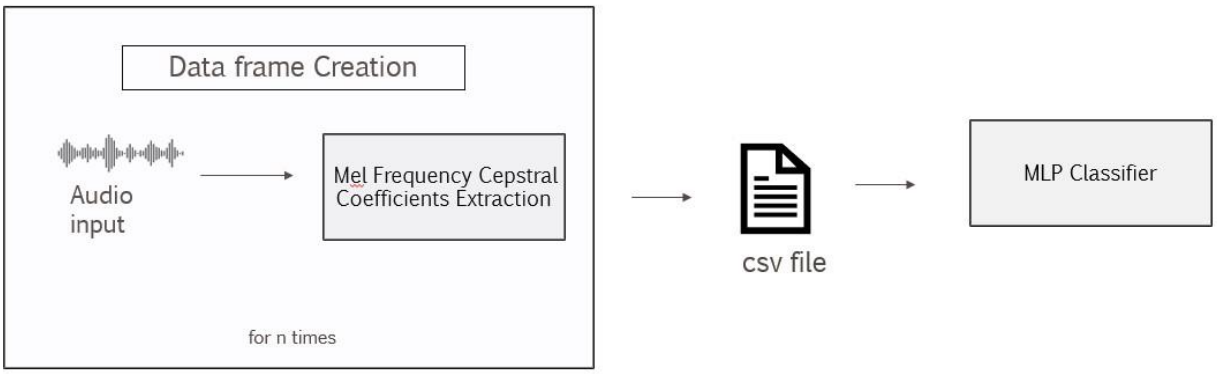
connection and doesn't support multiple tts-engine. [3] Voice assistants are the systems which can be integrated into bigger systems and we need to understand the various aspects along with flaws and limitations of various techniques to implement them. Some are desktop voice assistants while some are personal voice assistants. These use speech recognition, python backend and API calls. Multiple engines for the conversion of audio to text are available like pyttx3, gttx, etc. [4] There are different conditions to record an audio, in different conditions including noise, and thus would need different extraction techniques. Speaker authentication is recognizing the right user and giving access to that particular user only. The speaker could be at any place where noise plays a major factor for authentication. It is necessary to deal with audio recorded in an unconstrained environment containing a lot of unwanted noise. [5]

## III. METHODOLOGY

The user's voice is taken as input by the system, after which the system extracts the audio function vectors. These vectors are used to train the machine learning algorithms.

### 3.1 Dataset

The audio formats used extensively are WAV, MP3, etc. MP3 files are compressed files and do not contain the entire information. Whereas, WAV files cover the entire range of frequencies that are audible to the human ear. Hence, WAV are usually used in audio research studies. In the system database, the sample voice of the authenticated user is stored in WAV format. Various attributes like frequency and spectral range of the audio input are extracted and added to the csv file as shown in figure 1. Audio has three features - Rhythmic, Temporal and Spectral Features. The objective of the system is to clear out the noise and extract the range and features of the audio into a dataset using an appropriate algorithm and predefined coefficients. [6] For creating the dataset we have used recordings in WAV format which is then converted to a csv file containing the MFCC features extracted from various types of voices.




Dataset Creation

### 3.2 Phase 1 – Voice Authentication

The training data phase involved creating a dataset such that the classifier is well trained to detect minor changes in voice of an user so that it can differentiate accurately as authorised and non-authorized. The aspects involved are pitch, accent, notations, tones, distance, mood, volume, rhythm, environment and other variations that can affect the voice of the same user. For this purpose, the dataset was trained using voices of multiple people with vast differences in these aspects. Also, if there was a false positive prediction, we trained similar voices of different users so that the classifier can detect accurately among similar voices. The dataset also involved different tones and languages, background voices and noise so that it doesn't affect the actual prediction. The classifier also differentiates between pre-recorded and instantaneous voice of an authorised user.

#### 3.2.1 Training the data

The training data phase involved creating a dataset such that the classifier is well trained to detect minor changes in voice of an user so that it can differentiate accurately as authorised and non-authorized. The aspects involved are pitch, accent, notations, tones, distance, mood, volume, rhythm, environment and other variations that can affect the voice of the same user. For this purpose, the dataset was trained using voices of multiple people with vast differences in these aspects. Also if there was a false positive prediction, we trained similar voices of different users so that the classifier can detect accurately among similar voices. The dataset also involved different tones and languages, background voices and noise so that it doesn't affect the actual prediction. The classifier also differentiates between pre-recorded and instantaneous voice of an authorised user.

#### 3.2.2 Feature Extraction

The audio input file can be recorded in varying conditions. Classification of audio based on its features can be Rhythmic, Temporal and Spectral. Rhythmic features are musical notes. Temporal features describe an audio signal over a sampled period. Spectral features are based on the frequency of audio waves.

Human ear receives temporal signals, then converts them to their frequency domains resulting in corresponding vibrations giving us the ability to hear and comprehend an audio signal. Spectral features are very similar to how the human ear perceives audio signals and are widely used in speech and speaker recognition.

Several methods for generating frequency domains are : Linear Predictive Coding (LPC), Rastafilter, and Mel Frequency Cepstral Coefficients (MFCC).

We will be using MFCC as according to research they represent the closest relation to the human earing model and are becoming increasingly popular in speech recognition.

### 3.2.3 Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients known as MFCCs are the set of spectral audio features that are used to uniquely identify one audio sample from other samples. These coefficients form the dataset where the classifier will learn the pattern of the coefficients and then gives authority to an user.

### 3.2.4 Machine Learning Classifier

The system uses a supervised classification machine learning algorithm, Multilayer Perceptron. MLP is a type of Artificial Neural Network where there are two or more hidden layers in addition to input layer and output layer. It also facilitates back propagation which minimises the error rate by updating the weights of input. This system uses MLP Classifier as it is one of the most efficient and also gives accurate results. The dataset created in this project by using MFCC is used to train the MLP classifier.

### 3.2.5 Speech Recognition of randomly generated sentences

To overcome the issue of someone playing the pre-recorded voice of the authorized user to gain authorization, we overcome it by generating a random sentence of five words using the random module in Python. And then verify the input sentence of the user against the randomly generated one, by using the speech-recognition library which converts the audio into theken text.

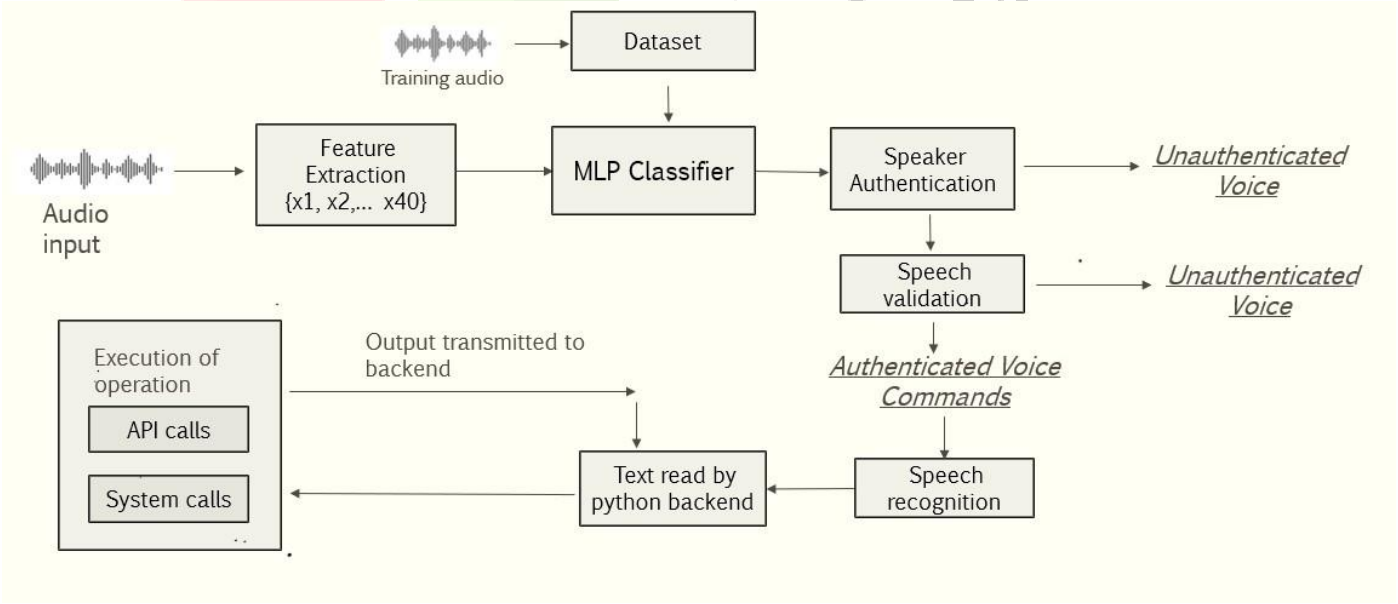
## 3.3 Phase II – Voice Assistant

Virtual assistance is a very user-friendly software that allows tasks to be done easily. A virtual assistant is software that performs the tasks which are assigned by the user. Virtual assistants use Speech Recognition Library to recognize the user's voice input and compare them with executable commands. The Virtual Assistant helps to build a system that totally works on the user's commands.

For the voice assistant, the Speech Recognition library provides a wide range of in-built functions, where the user commands will be executed and sent back to the user in voice cite [7]. Using a Virtual Assistant is time-saving since it is very fast. Another major benefit of using a virtual assistant is that the user can use it at any time and it can adapt to changes easily. A virtual assistant performs particular tasks with the help of a Python backend by interpreting the voice commands using natural language processing. [8] Speech Recognition is a library in Python that is used to process human speech into a text format. Speech Recognition is used in many applications such as the Internet of Things, Artificial Intelligence, etc. The system uses the speech-recognition library. Using the speech Recognition library, it will convert the speech input to text using the recognizer function. Next the text is read by the python backend and the operation is performed (e.g. – a message is sent to the preferred contact) with the help of API / system calls. Further this output is then fed to the Python backend. Lastly, the written text is converted to a sound by the pyttsx3 engine. Various libraries like WolframAlpha, speech recognition, smptlib, etc. are used for the assistant work. [9] pyttsx3 is a text-to-speech conversion library used in Python for conversion of text into speech. pyttsx3 library is used by the system as a response after the execution of a task. It also allows working offline and is compatible with both versions of Python (Python 2 and 3). An application invokes the pyttsx3.init() factory function to get a reference to pyttsx3. It is a tool which converts the entered text into speech with ease. The pyttsx library supports two voices as output: A Female voice and A Male voice. Tasks done by virtual assistant

Multiple tasks done by the virtual assistant are as follows -

1. Task manager - user can command verbally to manage their day-to-day activities, set reminders, schedule meetings and add events.
2. Restart, shut down and hibernate - the system will automatically perform these activities on getting the user commands
3. Manage calls - user can manage calling over verbal commands
4. Capture photos




Architecture of Authenticated Virtual Assistant

## IV. CONCLUSION

This project is broadly divided into two parts – Voice Authentication and Voice Assistant. Audio pre-processing is an extremely crucial part in such kinds of projects. The two most important things to focus on for pre-processing are reducing the ambient noise and highlighting the human vocals. This will be achieved using shelf filters and MFCCs for noise reduction and vocal enhancements respectively. Feature extraction again is important as this is the heart of the classification. Converting the raw audio datasets with meaningful vectors would directly impact how the classification algorithms work on this dataset. Hence, the project uses Mel Frequency Cepstral Coefficients (MFCCs) for this phase using the Multilayer Perceptron.

Additionally, most speaker recognition studies calculate MFCC deltas up to the second order (differentials and accelerations) to improve the accuracy of the model. While this is an effective choice, it works best only on audio with no external intervention of sound. Lastly, choosing an appropriate machine learning classification technique is important too as those techniques determine how the model interprets the data. Thus, leveraging existing research and applying the knowledge, the project contains a Multilayer Perceptron classifier. For voice assistance, the project will use a speech recognition library for converting speech to text and perform the respective functions on the application with the help of System and API calls. The overall work is explained in the architecture.

## V. ACKNOWLEDGMENT

First and foremost, we would like to thank our Project guide, Prof. Dr. Y. D. Bhise, for her guidance and support. Her valuable insights, feedback, and assistance were crucial in the completion of this study. Through our discussions, she helped us to form and solidify ideas. With a deep sense of gratitude, we wish to express our sincere thanks to, Prof. Dr. S. S. Sane, Head of Department, Computer Department for his immense help in planning and executing the work on time. Our grateful thanks to the departmental staff members for their support.

## REFERENCES

**[1]** M.Subba Rao, P.V.S.Lokeswari, K.sireesha, A.Purushotham, “Automatic Speaker Recognition Using MultiLayer Perceptron Algorithm”, Volume XII, Issue IV, April 2020

**[2]** Saritha Kinkiri and Simeon Keates, “Speaker Identification:Variations of a Human voice”, IEEE Explore, 2020

**[3]** Subhash S, Prajwal N Srivatsa, Siddesh S, Ullas A, Santhosh B “Artificial Intelligence based Voice Assistant”, IEEE Explore, 2020

**[4]** Ayush Chinchane, Aryan Bhushan, Ayush Helonde, Prof. Kiran Bidua, “SARA: A Voice Assistant Using Python”, IRASET June 2022

**[5]** Abhishek Manoj Sharma “Speaker Recognition Using Machine Learning Techniques”, San Jose State University, Spring 2019

**[6]** Shaojin Ding, Tianlong Chen2, Xinyu Gong Weiwei Zha, Zhangyang Wang, “AutoSpeech: Neural Architecture Search for Speaker Recognition”, August 2020

**[7]** V. Geetha, C.K.Gomathy, Kottamasu Manasa Sri Vardhan, Nukala Pavan Kumar, “The Voice Enabled Personal Assistant for Pc using Python”, International Journal of Engineering and Advanced Technology, April 2021

**[8]** Kumar, Lalit, "Desktop Voice Assistant Using Natural Language Processing (NLP)", International Journal for Modern Trends in Science and Technology, 2020

**[9]** Vishal Kumar Dhanraj, Lokeshkriplani, Semal Mahajan, “Research Paper onDesktop Voice Assistant”, February 2022