Speech Recognition System Using AI

Sneha Sabale[1], Madhavi Vora[2], Prof.Abhijit Shinde[3], Prof.Sachin Pandare[4]
[1, 2] Dept of Computer Science and Engineering
[3]Assistant Professor, Dept of Computer Science and Engineering
[1, 2, 3] Sahakar Maharshi Shankarrao Mohite Patil Institute of Technology and Research, Akluj, Solapur ,Maharashtra,(India)

Abstract- *This paper presents a desktop application which allows manipulation of voice to command. The user will be able to handle computer by using voice, and can open any file in the computer using this application, user can manipulate and control computer system using voice commands. These manipulations will occur in real-time. Furthermore, the user will be able to load additional text editors at the same time, and manipulate them. The interaction between user and program will occur through a GUI (Graphical User Interface) which is developed in LabVIEW. The GUI will attempt to replicate similar controls found on professional voice recognition software's. Jarvisis Just Rather Virtual Assistant System which make easiest our life. In this project Jarvis is digital life assistant which uses only human communication means such twitter, instant message and voice to create two way connection between human and his apartment, controlling light and appliances assist in booking , notify him of breaking news, Facebook notification and many more. In our project we mainly use voice as a communication means so the Jarvis is basically speech recognition system. The concept of speech technology really encompasses two technologies: Synthesizer and Recognizer. Speech synthesizer takes as input and produce and audio stream as output. The voice is a signal of infinite command. The direct analysis and synthesizing the complex voice due to too much information contained in the signal.*

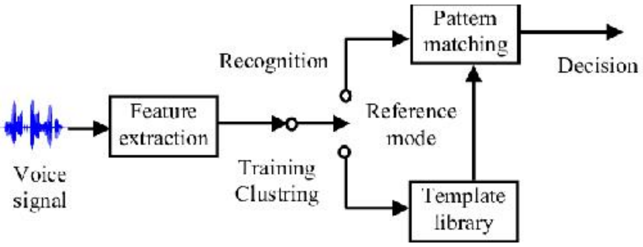
Keywords- Feature Extraction, Speech recognizer, Feature Machine

I. INTRODUCTION

Jarvis is Just Rather Virtual Assistant System which make easiest our life. In this project Jarvis is digital life assistant which uses only human communication means such twitter, instant message and voice to create two-way connection between human and his apartment, controlling light and appliances assist in booking, notify him of breaking news, facebook notification and many more. In our project we mainly use voice as a communication means so the Jarvis is basically speech recognition system. The concept of speech technology really encompasses two technologies: Synthesizer and Recognizer. Speech synthesizer takes as input and produce and audio stream as output. Speech recognizer on other hand does opposite. It takes an audio stream as input and thus turns it into text transcription. The voice is a signal of infinite command. The direct analysis and synthesizing the complex voice due to too much information contained in the signal. Therefore the digital signal process such as Feature Extraction and Feature Matching are introduced to represent the voice signal. In this project we directly use speech engine which use Feature Extraction technique as Mel scaled frequency cepstral. The mel scaled frequency centrals coefficients (MFCCs) derived from most widely used front-ends in state-of-the-art speech recognition systems. Our aim to create more and more functionalities which can help human to assist their daily life and reduce their effort. In our test we check all these functionalities is working properly. It takes an audio stream as input and thus turns it into text transcription.

The voice is a signal of infinite information. A direct analysis and synthesizing the complex voice signal is due to too much information contained in the signal. Therefore, the digital signal processes such as Feature Extraction and Feature Matching are introduced to represent the voice signal. In this project we directly use speech engine which use Feature extraction technique as Mel scaled frequency cepstral. Now a days voice recognition is highly influenced like google having function to open any site using voice. The objective of this project is to develop voice recognition with filtration of voice so that user can open any application using voice without disturbance of noise. Computer Jarvis is special software which is useful to blind people to operate system easily. Speech is an effective and natural way for people to interact with applications, complementing or even replacing the use of mice, keyboards, controllers, and gestures. A hands-free, yet accurate way to communicate with applications, speech lets people be productive and stay informed in a variety of situations where other interfaces will not.

Speech Rerecognition is a topic that is very useful in many applications and environments in our daily life. Generally, speech recognizer is a machine which understands humans and their spoken word in some way and can act thereafter. A different aspect of speech recognition is to facilitate for people with functional disability or other kinds of handicap. To make their daily chores easier voice control could be helpful. With their voice they could operate the light switch turn off/own or operate some other domestic appliances. This leads to the discussion about intelligent homes where these operations can be made available for the common man as well as for handicapped read using the function away read in MATLAB tool. MFCC method is used for detecting emotion from voice signal. Proposed work is based on feature extraction using MFCC and decision making using standard deviation.



The speech signal made to undergo Farming, after which it is passed through Hamming window for windowing process. It is most usable application to people for less hardwork.

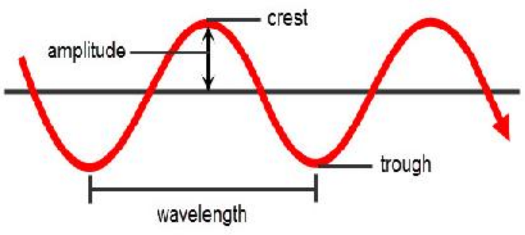
II. LITERATUREREVIEW

In named "Speech based Human Emotion Recognition Using MFCC",IEEE WiSPNET year of 2017 conference proposed by author name as M.S. Likitha, Sri Raksha R.Gupta. A database consist of voices of 60 people with different emotions. Speech signal of speakers read using the function awav read in MATLAB tool.MFCC method is used for detecting emotion from voice signal. Proposed work is based on feature extraction using MFCC and decision making using standard deviation. The speech signal made to undergo Farming, after which it is passed through Hamming window for windowing process performed on the input signal. After which the Mel Frequency Cepstral Coefficients were obtained. The standard deviation for the mean value was found, and this value was passed through as if else statement, where the obtained standard deviation of that particular emotion is compared with the optimized value of standard deviation for different emotion and corresponding emotion were displayed. It can predict some basic emotion such as happy, sad , angry from MFCC waves. In named "Emotion Recognition from Speech using Convolutional Neural Network with Recurrent Neural Network Architecture", IEEEWiSPNET year of 2017 conference proposed by authors named as Saikat Basu, Jaybrata Chakraborty. For the last two decades, several intelligent system are proposed by researches. These different system also differ by nature of features used for classification of speech signals. There are widely used spectral feature are Mel- frequency cepstrum coefficients (MFCC) and linear predictive cepstral. Speech understanding is a major component of human machine interaction and its quality affect the user experience. coefficient(LPCC). Only pitch and MFCC features are used for recognition of emotion. In[3] named "Emotion Recognition from Speech using Emotional Statistical Parametric Speech Synthesis Using LSTM- RNNs" IEEE WiSPNET of 2017 conference year of 12-15 December 2017,proposed by authors name as Shumin An, Zhenhua Ling and Lirong Da. Two modeling approaches, emotion dependent modeling and unified modeling with emotion codes,are implemented and compared by experiments. LSTM-RNN-based acoustic model are built separately for each emotion type. By using two or four – dimensional emotion space gives better result In named "Training Spoken Language Understanding System With non-parallel speech and Text", IEEE WiSPNET of 2020 conference proposed by author named Leda Sari, Samual Thomas. End Spoken Language understanding(SLU) systems are typically trained on large amount of data. Kanchan Patil, Avinash Kharat2, Pratik Chaudhary3, Shrikant Bider4, Rushikesh Gavane 5 Department of Information and Technology, SREsS Sanjivani College of Engineering Koregoan, India M.S. Likitha,Sri Raksha Gupta "Speech based Human Emotion Recognition Using MFCC",IEEE WiSPNET year of 2017 Leda Sary1 ,Samuel Thomas2, Mark Hasegawa johnson1 Department of computer engineering,Universities of Illinois at Urbana-chamaign. Shrutika Khobragade Department of computer Engineering.

III. PROPOSEDSYSTEM

The speech signal and all its characteristics can be represented in two different domains, the time and the frequency domain A speech signal is a slowly time varying signal in the sense that, when examined over a short period of time (between 5 and 100 ms), its characteristics are shorttime stationary. This is not the case if we look at a speech signal under a longer time perspective (approximately time T>0.5 s). In this case the signals characteristics are nonstationary, meaning that it changes to reflect the different sounds spoken by the talker To be able to use a speech signal and interpret its characteristics in a proper manner some kind of representation of the speech signal are preferred.

THREE STATE REPRESENTATIONS



he three-state representation is one way to classify events in speech. The events of interest for the three-state representation are • Silence(S) - No speech is produced. • Unvoiced Unvoiced (U) - Vocal cords are not vibrating, resulting in an aperiodic or random speech waveform. • Voiced (V) - Vocal cords are tensed and vibrating periodically, resulting in a speech waveform that is quareious period Block Diagram of Proposed System With it deep learning. Proposed model gives better accurancy for dataset .For real time imaginary large dataset is needed .The acoustic characterstics of the speech signal is Feature. extraction is a small amount of a data from the speech signal is extracted. In speech analysis,it is common to use two types of features : acoustic, which have physical sense and representation characterstics that correspond to values calaculated over the to analyze the signal without disturbing its acoustic properties. .

METHODOLOGY USED IN PROPOSED SYSTEM

voice recognition works based on the premise that a person voice exhibits characteristics are unique to different speaker. The signal during training and testing session can be greatly different due to many factors such as people voice change with time, health condition (e.g. the speaker has a cold), speaking rate and also acoustical noise and variation recording environment via microphone. Table II gives detail information of recording and training session, whilst Figure shows the flowchart for overall voice recognition process.

Process	Description

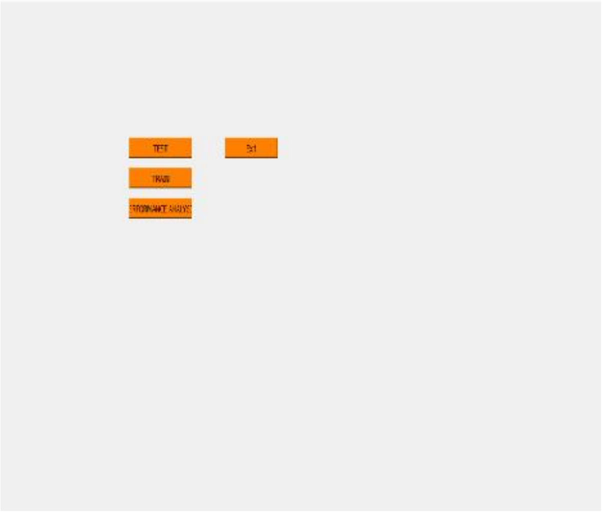
1 Speech	2Female(age=20,age=53)
	2Male(age=22,age=45)

2 Tool	Mono Microphone Microsoft Speech System
3 Environment	College

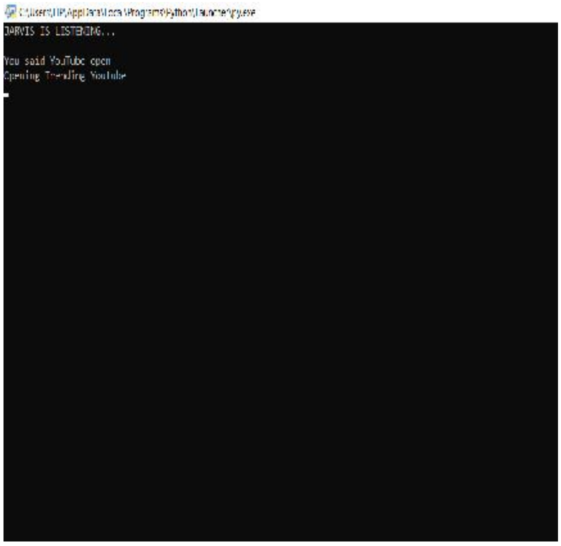
In this project we directly use speech engine which use feature extraction technique as Mel Scaled frequency cepstral coefficient (MFCCs) derived from fourier transform representation of speech signal, and those that in general does not correspond to any physical sense and filter bank analysis are perhaps the most widely used front-ends in state-of-the-art speech recognition systems. In our test we check all this functionality is working properly.

IV. OUTPUT

1 Voice Test



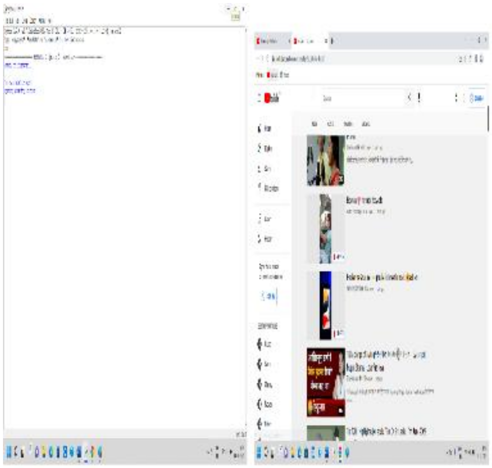
2 Voice is recognized and working



```
C:\Users\HP\AppData\Local\Programs\Python\Launcher\py.exe
JARVIS IS LISTENING...

You said YouTube open
Opening Trending Youtube
```

3 Run Successfully



V. CONCLUSION

This paper has discussed voice recognition algorithms which are important in improving the voice recognition performance. The technique was able to authenticate the particular speaker based on the individual information that was included in the voice signal. The results show that these techniques could use effectively for voice recognition purposes. Several other techniques such as Liner Predictive Coding (LPC), Dynamic Time Wrapping (DTW), and Artificial Neural Network (ANN) are currently being investigated. The findings will be presented in future publicationsIn this paper, we described the initial steps we have taken to build Jarvis, an intelligent system that can help people learn physical tasks via apprenticeship training. The use of augmented reality provides an opportunity to provide instructional support within the context of task performance.

REFERENCES

[1] Kanchan Patil, Avinashi Kharat2, Pratik Chaudhary3, Shrikant Bider4, Rushikesh Gavane5 Department of Information and Technology, SRESs Sanjivani College of Engineering Kopargoan,India.

[2] M.S. Litika’s Raksha Gupta “Speech based Human Emotion Recognition Using MFCC”, IEEE WiSPNET year of 2017

[3] Leda Sary1, Samuel Thomas2, Mark Hasegawa johnson1 Department of computer engineering, Universities of Illinois at Urbana-chamaign. Shrutika Khobragade Department of computer Engineering.