IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

Jarvis: AI Virtual Assistant

Dr. V. K. Shandilya[1], Jaswantsingh Shekhawat[2], Prajwal Dakare[3], Sujay Dhawak[4], Anshul Gayakwad[5]
Head of Department, Department of Computer Science and Engineering[1]
Final Year B.E. Students, Department of Computer Science and Engineering[2,3,4,5]
SIPNA College of Engineering and Technology, Amravati, Maharashtra, India

Abstract: *"Jarvis" was the protagonist of Tony's Stark starring in the movie Iron Man. Unlike the first jokes when Jarvis was Stark's servant, Jarvis' movie version is a smart computer that communicates critical issues, monitors his family, and helps build and organize his hero's suit. In this project, Jarvis is a Digital Health Assistant who uses social media i.e., voice that creates two-way communication between people and his personnel Computer, performing various tasks such as calling and chatting on WhatsApp, taking notes, locating, etc. In our project, we use the word just as communication means that Jarvis is a Speech recognition program. The concept of speech technology combines two technologies: Synthesizer and vision. Speech synthesizer captures input and generates audio streams as output this is achieved using the pyttsx3 library accompanied by Speech function. Speech recognition. On the other hand, he did the opposite. It treats audio streaming as input and thus converts it into text writing and this output uses the Speech Recognition library accompanied by the TakeCommand function. The word is a symbol of endless knowledge. Direct analysis and integration of complex voice signal is due to the vast amount of information contained in the signal. Therefore, digital signal processes such as Feature Visibility and Feature Matching are introduced to represent the voice signal. For this project, we are directly using a speech engine that uses an output element such as Mel scaled frequency cepstral. The mel scale frequency cepstral coefficients (MFCCs) obtained from the fourier transform and filter bank analysis are probably the most widely used in high-level speech recognition systems. We aim to create more jobs that can help people in their daily lives and reduce their efforts as in this project we have introduced Chrome and Youtube Automation. In our test, we checked that all of this functionality works well. We test this on 2 speakers (1 female and 1 male) to find the exact motive for the speculation.*

Keywords: Chatbot, NLP, Artificial Intelligence, Infection, Important, Machine Learning

I. INTRODUCTION

A virtual assistant is a part of everyday life using objects that have escaped so far from the idea of being smart and technological devices. The concept of AI Assistant like Jarvis is a combination of intelligence and technology, Jarvis works on speech recognition that replaces the use of mice, keyboards, controls, and body language. It can be accessed through voice command. A speech impediment is usually a device that detects individual words and acts accordingly. A distinct feature of speech recognition is helping people with functional disabilities or other types of disabilities and mental retardation. Visual AI helper serves as a solution to many problems. It also serves as a new and attractive item that can be placed anywhere.

Important: how is speech recognition done? To determine how speech recognition problems can be addressed today, a review of some of the highlights of the study will be presented. Attempts to automate speech recognition were made in the 1950s when various researchers attempted to explore the basic ideas of acoustic phonetics. system recognition is based on the vowel point of each digit. Ten vowels embedded in the / b / -vowel- / t / format are recognized independently of the speaker, which is another attempt made in 1959 by Fororgie, developed at MIT Lincoln Laboratories. In the 1970's a speech recognition study gained many points of change. First, a separate word recognition center became a practical technology based on the basic studies of Velichko and Zagoruyko in Russia, Sakoe and Chiba in Japan, and Itakura in the United States. Russian studies are exploring the application of pattern recognition concepts in speech recognition; Japanese research has shown that powerful planning techniques can be used effectively. At AT&T Bell Labs, tests aimed at making speech recognition systems truly dominate the speaker. Use sophisticated integration algorithms to represent all Jarvis or Digital Life allies. In 1980 a shift in technology from model-based approaches to mathematical modeling methods, most notably the hidden method of the Markov model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

II. Speech Representation

All features of the speech signal can be represented on two different domains, background and time. If we look at the speech signal under long-term view this is not the case (approximately time t> 0.5 s). In this case, the features of the signal are not static, which means that it changes to reflect the different sounds spoken by the person in order to be able to use the speech signal and to interpret its features properly in a certain type of speech signal representation. preferred.

III. Three State Representation

How to distinguish events from speech representation of the three states. Here are some interesting facts

- Silence (S) - No speech produced.
- Unvoiced (U) - The vocal cords do not vibrate, resulting in random or intermittent speech stops.
- Voice (V) - The vocal cords tighten and vibrate occasionally, resulting in quasiperiodic vocal cords.

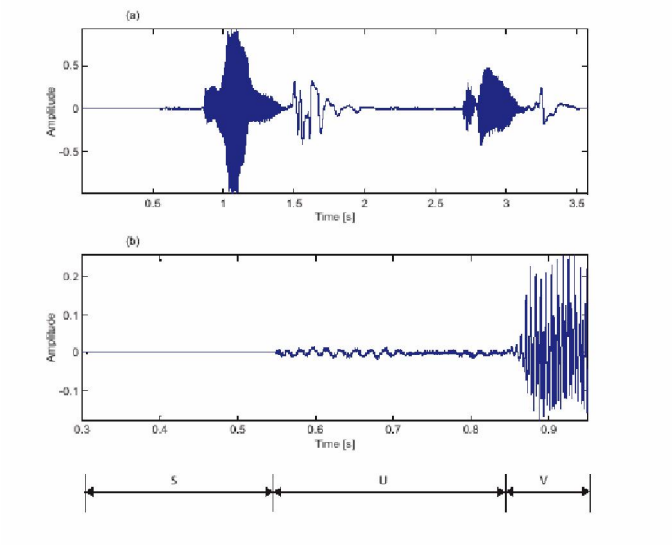

Figure 1: Three State Representation

The upper structure (a) contains the whole speech sequence and in the middle structure (b) part of the upper structure (a), the speech sequence area is reproduced by zoom. Below Figure: 1 a division into the representation of the three states, in respect of the different parts of the central structure is given. The division of the speech wave format into well-defined contexts is not straightforward. But this issue is not as big as one might think.

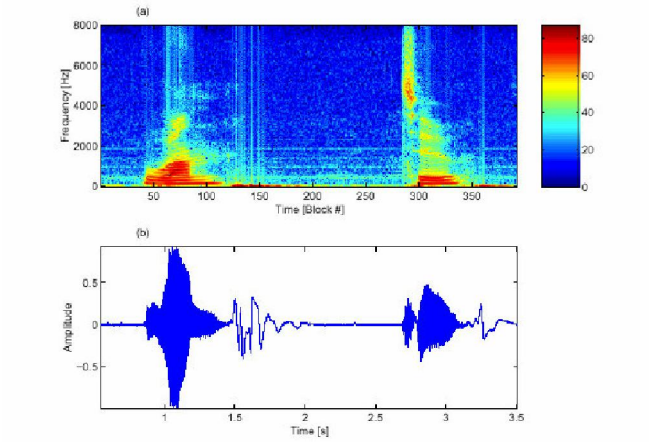

Figure 2: Spectrogram using Welch's Method (a) and speech amplitude (b)

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

In this illustration, the dark blue parts represent parts of the speech structure, in which speech is not produced, and the simple red parts represent the intensity of speech when speech is produced. In the background the waveform expression is given. The Welch method is used for spectrogram, which uses periodically modified periodograms. K = 320 block size parameters used in this method, a Hamming window type with a spacing of 62.5% leading to 20 ms blocks and a distance of 6.50 ms between blocks.

IV. PHONEMICS AND PHONETICS

Speech is produced in the human brain when a person develops an idea that must be produced and transmitted to the listener. After making the thought he wants, one creates a sentence / sentence / episode by selecting a set of different limited sounds.

to select a collection of special corresponding sounds. A unit of basic theory to explain how to deliver linguistically, phonemes are known as mind-built speech. Representing different parts in the speech format is a way of expressing phonemes, produced by a person's pronunciation, and divided into writing (continuous) or immersive (non-continuous) parts. When the voice is in motion the phonograph continues when the sound of speech is produced. In contrast, the phoneme does not move when the voice changes its characteristics when producing speech. For example, when the tone of a voice changes by closing and opening the mouth or by moving your tongue in different situations, the phonograph that describes the speech produced does not develop a time waveform and is separated by different sounds produced by the human voice. Separation can also be seen as the division of the categories in the picture: 3.

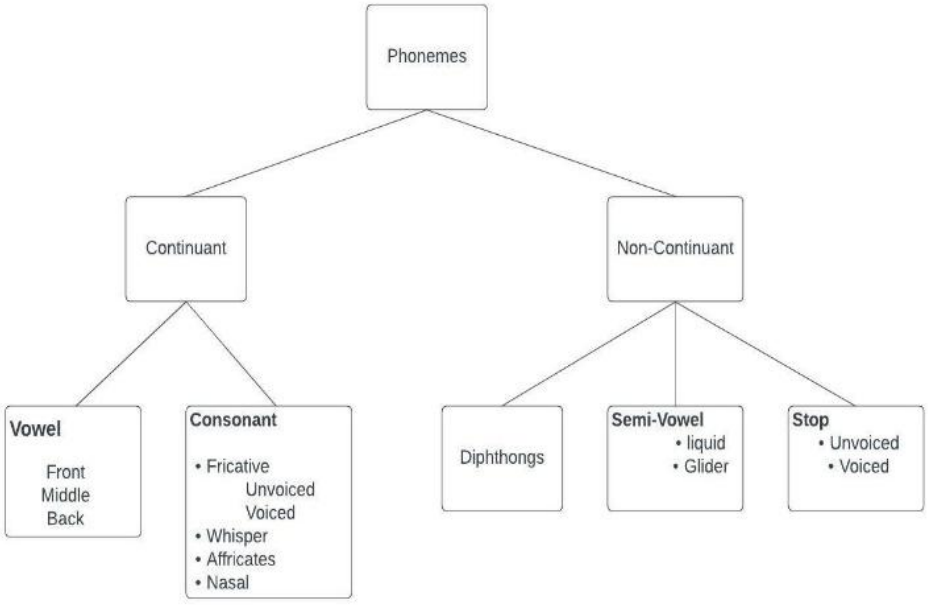

Figure 3: Phoneme Classification

V. FEATURE EXTRACTION (MFCC)

To get the best performance of the detection the best representation of the acoustic signal parameter is an important task. The efficiency of this category is important in the next section as we touch on its operation. The Mel-Frequency Cepstral Coefficients (MFCC) approach is the most effective way to extract speech quality and current research aims to identify performance enhancements, in the MFCC based on knowing the differences in the sensitive ear band frequency and frequency [7 - 10]. The MFCC has two types of filters which is a logarithmic space above 1000 Hz and is separated by a line at frequencies below 1000 Hz. The whole process is shown in the following diagram: 4.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

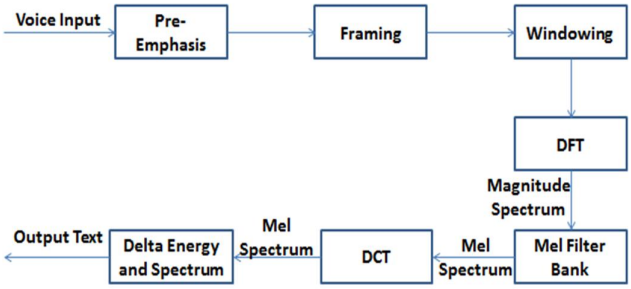

Figure 4: MFCC Block Diagram

VI. METHODOLOGIES

As mentioned, voice recognition works in the postulate that the human voice reflects different features on a different speaker. The signal during the test and the training session can vary due to many factors such as changes in the user's voice, health status (e.g., the speaker has a fever), frequency of speaking and echoing in the audio and frequency of microphone recording. The table below provides detailed information on recording and training time and Figure 5 shows the voice process chart.

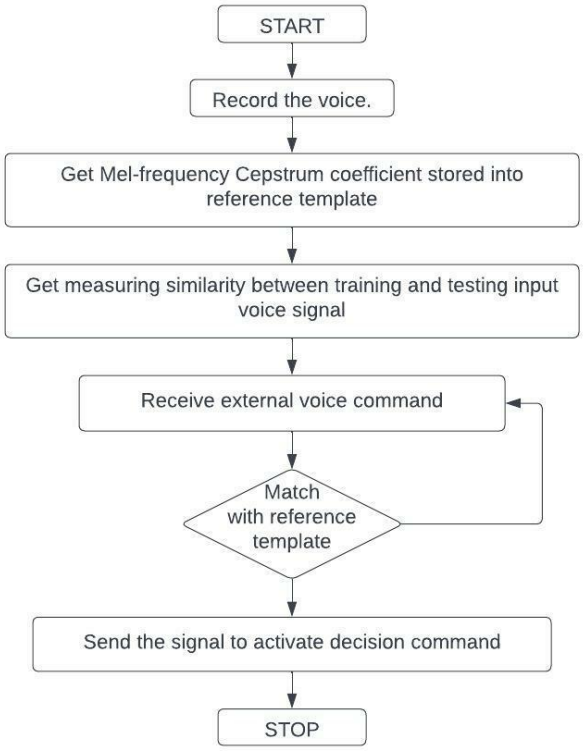

Figure 5: Flowchart for Voice Flow Algorithm

Process	Description
Speech	2Female (age=20, age53) 2Male (age=22, age=45)
Tool	Mono Microphone Microsoft Speech Software
Environment	College Campus
Utterance	Twice each of the following word

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

	1) Volume Up 2) Volume Down 3) "Jarvis there" 4) Introduce yourself 5) Show data
Sampling Frequency	16000 KHz
Feature Computational	39 double delta MFCC coefficient

VII. Result and Discussion

Below figure shows the input voice signals of two different speakers.

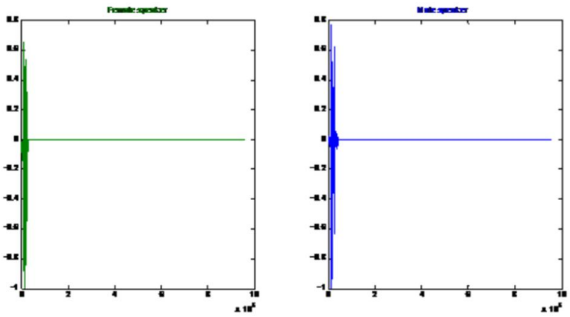

Figure 6: Example voice signal input of two difference speakers.

The above figure is used for carrying voice analysis performance evaluation using MFFC. A MFCC cepstral is a matrix, the problem with this approach is that, if constant window spacing is used and the lengths of the input and stored sequences is unlikely to be the same. Moreover, there will be variation in the length of individual phonemes as discussed before, Example the word Volume Up might be uttered with a long and short or with a short and long.

VIII. Conclusion

This paper represents voice recognition algorithms that are important in improving voice recognition functionality. The method was able to authenticate a specific speaker based on personal information embedded in the voice signal. The results show that these methods can be used effectively for voice purposes. Several other strategies such as Liner Predictive Coding (LPC), Dynamic Time Wrapping (DTW), and Artificial Neural Network (ANN) are currently under investigation. The findings will be presented in future publications.

References

[1]. https://docs.python.org/3/

[2]. Ashish Jain, Hohn Harris, Speaker identification using MFCC and HMM based techniques, university Of Florida, April 25,2004.

[3]. http://www.microsoft.com/MSDN/speech.html, downloaded on 2Oct 2012.

[4]. Young Steve, A Review of Large-vocabulary Continuous-speech Recognition, IEEE SP Magazine, 13:45- 57, 1996, ISSN 1053-5888.

[5]. https://www.udemy.com/?deal_code=&utm_term=Homepage&utm_content=Textlink&utm_campaign=Rakuten-default&ranMID=39197&ranEAID=%2F68Yt01SgtI&ranSiteID=_68Yt01SgtI-VvNIQiohSWcLJRXuQasY bg&LSNPUBID=%2F68Yt01SgtI&utm_source=aff-campaign&utm_medium=udemyads

[6]. Mammone Richard J., Zhang Xiaoyu, Ramachandran Ravi P., Robust Speaker Recognition, IEEE SP Magazine, 13:58-71, 1996, ISSN 1053-5888.

[7]. http://web.science.mq.edu.au/~cassidy/comp449/html/ch11s 02.html, downloaded on 2 Oct 2012.

[8]. Rabiner Lawrence, Juang Bing-Hwang. Fundamentals of Speech Recognition Prentice Hall , New Jersey, 1993, ISBN 0-13-015157-2

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 4, May 2022
Impact Factor: 6.252

[9]. Deller John R., Jr., Hansen John J.L., Proakis John G. ,Discrete-Time Processing of Speech Signals, IEEE Press, ISBN 0-7803-5386-2

[10]. http://www.cse.unsw.edu.au/~waleed/phd/html/node38.html , downloaded on 2 Oct 2012.

[11]. https://rapidapi.com/#

[12]. https://www.analyticsvidhya.com/blog/2022/02/full-guide-on-whatsapp-automation-using-python/#:~:text=various%20other%20features.-,What%20is%20Whatsapp%20Automation%20Using%20Python%3F,web%20and%20send%20the%20message.

[13]. https://medium.com/future-vision/google-maps-in-python-part-2-393f96196eaf