Jarvis- AI Virtual Voice Assistant

Punam D. Dhangar, Jitentra A. Gaikwad
Department of Instrumentation Engineering Vishwakarma Institute of Technology Pune, India

Abstract - JARVIS AI Voice Assistance is an advanced system designed for seamless voice interactions to manage tasks, control smart devices, and provide real-time assistance. Inspired by Iron Man's fictional AI, JARVIS combines natural language processing and machine learning to understand and respond to user commands. Its capabilities include managing household systems, scheduling, retrieving information, and assisting in complex tasks.

This project focuses on JARVIS as a digital assistant that connects to a user's home via Twitter, instant messaging, and voice commands, enabling two-way control over lights, appliances, cooking help, news alerts, and more. Built as a speech recognition application, it uses a synthesizer to convert text to speech and a recognizer to turn spoken words into text, enabling smooth, voice-based communication.

Keywords: Voice Assistant, NLP, Neural Network, Google Search.

I. INTRODUCTION

Speech is one of the most natural and effective ways for people to engage with technology, offering a hands-free, seamless alternative to traditional input methods like keyboards and mice. By simply speaking, users can interact with applications, stay productive, and access information in scenarios where using their hands isn't feasible—whether they're driving, cooking, or multitasking. Speech not only enhances convenience, but it also adds a layer of accessibility that many other interfaces can't match.

At its core, speech recognition technology listens, understands, and responds to spoken commands. This ability to transform voice into action is becoming increasingly vital in daily life, from personal assistants on smartphones to voice-controlled smart home devices. Beyond convenience, speech recognition holds transformative potential for individuals with disabilities. For those who struggle with mobility or dexterity, voice control can significantly simplify daily tasks. A person could, with just their voice, turn lights on or off, adjust the thermostat, or operate various household appliances, giving them greater independence.

This brings us to the exciting world of intelligent homes, where voice-activated systems can be used not only by the common person but also to greatly assist those with disabilities. Imagine controlling lights, appliances, or even security systems with just your voice, offering both convenience and independence.

But how exactly is speech recognition achieved? To understand modern advancements, it's important to look back at its origins. The journey of automatic speech recognition began in the 1950s, when researchers first attempted to use machines to understand human speech. These early systems were based on acoustic-phonetics, the study of the physical properties of speech sounds. One of the pioneering efforts came in 1952 at Bell Laboratories, where Davis, Biddulph, and Balashek developed a system that could recognize isolated digits spoken by a single speaker. This system marked the first step towards the sophisticated voice-controlled systems we have today, laying the foundation for decades of innovation in speech recognition technology.

As research progressed, these early concepts evolved into the highly advanced, multi-speaker systems we now see integrated into everyday technology, making speech recognition a key part of intelligent homes and beyond.

The early speech recognition system focused on measuring spectral resonances during the vowel sounds of each digit. In 1959, Forgie at MIT Lincoln Laboratories made another important advancement by recognizing ten vowels in a /b/-vowel-/t/ format without being tied to a specific speaker.

During the 1970s, speech recognition research made significant progress, particularly in isolated word recognition, which became a usable technology. This progress was influenced by studies from Velichko and Zagoruyko in Russia, Sakoe and Chiba in Japan, and Itakura in the United States. The Russian research helped develop pattern recognition techniques, the Japanese research applied dynamic programming methods, and Itakura introduced linear predictive coding (LPC).

At AT&T Bell Labs, researchers started experiments to create speech recognition systems that could

understand speech from any speaker. They used various clustering algorithms to identify the different patterns needed to represent variations in speech across a wide range of users.

In the 1980s, there was a shift in technology from template-based methods to statistical modeling approaches, particularly the hidden Markov model (HMM), which greatly improved speech recognition accuracy and efficiency [1].

The purpose of this paper is to explore and deepen our theoretical and practical understanding of speech recognition technology. We begin by examining the state-of-the-art feature extraction method known as Mel-Frequency Cepstral Coefficients (MFCC). By studying MFCC, we aim to apply this knowledge practically, leading to the implementation of a speech recognizer using .NET technology in C#, developed by Microsoft.

For our project, we utilize the Speech Application Programming Interface (SAPI), a robust API created by Microsoft that enables developers to integrate speech recognition and speech synthesis capabilities into Windows applications. This integration allows us to leverage advanced speech processing techniques while creating a user-friendly interface for interaction, showcasing the potential of speech recognition in real-world applications.

Applications that use the Speech Application Programming Interface (SAPI) include Microsoft Office, Microsoft Agent, and Microsoft Speech Server. Generally, all APIs are designed so that software developers can create applications for speech recognition and synthesis using a standard set of interfaces. These interfaces are accessible from various programming languages, making it easier to integrate speech capabilities into different applications. In addition, third-party companies can create their own Speech Recognition and Text-to-Speech (TTS) engines or modify existing ones to work with SAPI. Essentially, the speech platform includes application runtimes that provide speech functionality, an Application Programming Interface (API) for managing the runtime, and runtime languages that enable speech recognition and TTS in specific languages.

II. SPEECH REPRESENTATION

The speech signal and all of its features can be represented in two separate domains: time and frequency domain. A speech signal is a slowly fluctuating signal in the sense that when studied over a short period of time (between 5 and 100 ms), its features are short-term stationary. This is not the case when we examine a voice signal over a longer time period (roughly time T>0.5 s). In this scenario, the signal characteristics are non-stationary, which means they shift to represent the diverse sounds spoken by the speaker. A voice signal representation is preferred in order to use and comprehend its properties correctly.

1 THREE STATE REPRESENTATION

The three-state representation is one way to classify events in speech. The events of interest for the three-state representation are

- Silence (S) - In this state, No speech is produced.

- Unvoiced (U) - This state occurs when the vocal cords are not vibrating, leading to an aperiodic or random speech waveform.

- Voiced (V) - In this state, the vocal cords are tensed and vibrating periodically, resulting in a quasi-periodic speech waveform.

Quasi-periodic means that the speech waveform appears periodic over a short time period (5-100 ms) when it remains stable.

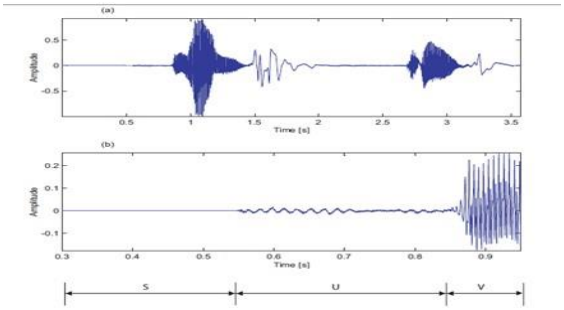

Fig1: Three State Representation

The upper plot (a) illustrates the complete speech sequence, while the middle plot (b) focuses on a specific section by zooming in on a part of the upper plot (a). At the bottom of Fig. 1, the segmentation into a three-state representation is displayed, highlighting the correlation with different segments of the middle plot. Although segmenting the speech waveform into clearly defined states can be complex, this challenge is often less daunting than it might initially appear. With careful analysis, the underlying patterns can be discerned, making the task manageable.

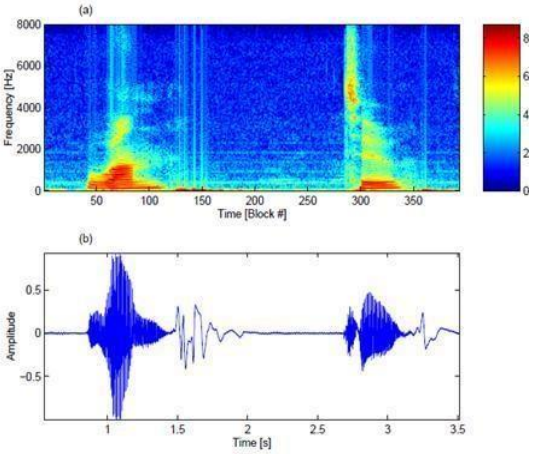

Fig 2 : Spectrogram using Welch's Method (a) and speech amplitude (b)

In this Fig2 representation, the darkest areas (dark blue) indicate segments of the speech waveform where no speech is produced, while the lighter areas (red) signify periods of vocalization with varying intensity. The speech waveform is presented in the time domain, illustrating the dynamic nature of speech over time. For the spectrogram, Welch's method is employed, utilizing averaged modified periodograms to provide a clearer view of the frequency content within the speech signal [3]. This method enhances the analysis of speech by revealing how its spectral characteristics evolve, allowing for a deeper understanding of vocal patterns. The parameters used in this method include a block size of \(K = 320 \), a Hamming window with 62.5% overlap, resulting in blocks of 20 ms with a 6.25 ms distance between each block.

2. PHONEMICS AND PHONETICS

The Speech production begins in the human mind when a thought is formed and prepared for communication to a listener. After developing the desired thought, the speaker constructs a phrase or sentence by selecting from a collection of finite, mutually exclusive sounds. The basic theoretical unit for conveying linguistic meaning in this mental framework is called a phoneme. Phonemes represent the various components of a speech waveform, produced through the human vocal mechanism, and are divided into two categories: continuant (stationary) parts and non-continuant parts.

A phoneme is considered continuant when the speech sound is produced while the vocal tract remains in a steady state. In contrast, a phoneme is classified as non-continuant when the vocal tract undergoes changes in its characteristics during speech production. For instance, if the shape of the vocal tract alters due to actions like opening and closing the mouth or moving the tongue, the resulting phoneme is non-continuant. Phonemes can be grouped based on properties of the time waveform or frequency characteristics, leading to various classifications of sounds produced by the human vocal tract. This distinction highlights the dynamic nature of speech and the intricate mechanisms involved in phoneme production.The classification, may also be seen as a division of the sections in Fig 3

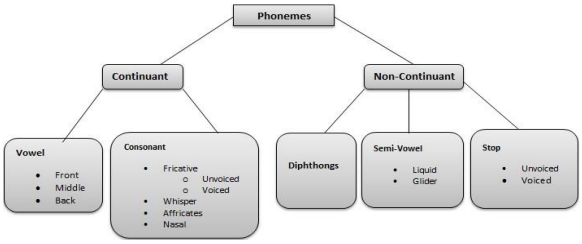

Fig3: Phoneme Classification

3. FEATURE EXTRACTION (MFCC)

The extraction of optimal parametric representations of acoustic signals is crucial for achieving superior recognition performance in speech processing. This efficiency directly impacts the subsequent phases of recognition, as it influences how well the system can interpret and act on the input signals. Mel-Frequency Cepstral Coefficients (MFCC) play a key role in this process, as they are designed to align with human auditory perceptions, particularly the fact that the human ear cannot perceive frequencies above 1 kHz. Essentially, MFCC utilizes established variations in the critical bandwidth of human hearing with frequency to enhance the accuracy of speech recognition systems. MFCC utilizes two types of filters: one that is spaced linearly for frequencies below 1000 Hz and another that uses logarithmic spacing for frequencies above 1000 Hz. The Mel Frequency Scale incorporates a subjective pitch to capture important phonetic characteristics in speech. The overall process is shown in following Fig: 4.

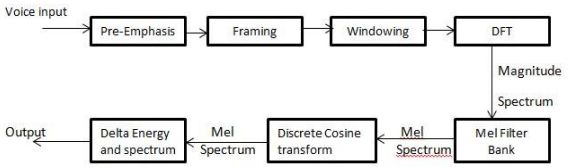

Fig4: MFCC Block Diagram

As shown in Figure 4, the Mel Frequency Cepstral Coefficients (MFCC) extraction process consists of

seven computational steps. Each step serves a specific function and employs various mathematical approaches, as briefly discussed below:

STEP 1: PRE- EMPHASIS

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$Y[n]=X[n]-0.95X[n-1] \quad (1)$$

Let's consider a = 0.95, which make 95% of any one sample is presumed to originate from previous sample.

STEP 2: FRAMING

The process of segmenting the speech samples obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M<N).Typical values used are M
= 100 and N= 256

STEP 3: HAMMING WINDOWING

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window equation is given as: If the window is defined as W (n), $0 \le n \le N-1$ where

N = number of samples in each frame

Y[n] = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is

Shown below:

$$Y(n)=X(n)\times W(n) \quad (2)$$

$$w(n)=0.54-0.46\cos[[2\pi n/(N-1)] \quad 0\le n\le N-1 \quad (3)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

STEP 4: FAST FOURIER TRANSFORM

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$y(w)=FFT[h(t)*X(t)]=H(w)*X(w) \quad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

STEP 5: MEL FILTER BANK PROCESSING

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 5 is then performed.

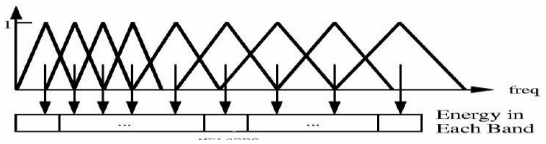

Fig. 5. Mel scale filter bank, from (young et al, 1997)

This figure shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decrease linearly to zero at center frequency of two adjacent filters [7, 8]. Then, each filter output is the sum of its filtered spectral components. After that the following equation issued to compute the Mel for given frequency f in HZ.

$$F(Mel)=[2595*\log_{10}[1+f/700]] \quad (5)$$

STEP6: DISCRETE COSINE TRANSFORM

This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cestrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

STEP 7: DELTA ENERGY AND DELTA SPECTRUM

The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time . 13 delta or velocity

features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t1 to time sample t2, is represented at the equation below:

$$Energy=\sum X^2[t] \quad (6)$$

Each of the 13 delta features represents the change between frames corresponding to cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t)=[c(t+1)-c(t-1)]/2 \quad (7)$$

III. METHODOLOGIES

As highlighted in [11], voice recognition operates on the principle that each person's voice has unique characteristics. During training and testing sessions, the audio signals can vary significantly due to several factors. For instance, a speaker's voice may change over time, be affected by health conditions (such as having a cold), and vary with the speaking rate. Additionally, external factors like background noise and differences in the recording environment can influence the audio captured by the microphone.

Table II provides detailed information regarding the recording and training sessions, while Figure 7 illustrates the flowchart of the overall voice recognition process. This process encompasses various stages, from capturing the voice input to recognizing and interpreting the spoken words, ultimately leading to accurate speech recognition despite the inherent variability in voice characteristics.

Process	Description
1) Speech	2Female(age=20,age=53) 2 Male(age=22,age=45)
2) Tool	Mono Microphone Microsoft Speech software
3) Environment	College Campus
4) Utterance	Twice each of the following word 1) Volume Up 2) Volume Down 3) "Jarvis there" 4) Introduce yourself 5) Show date.
5) Sampling Frequency	16000 KHz
6) Feature Computational	39 double delta MFCC coefficient

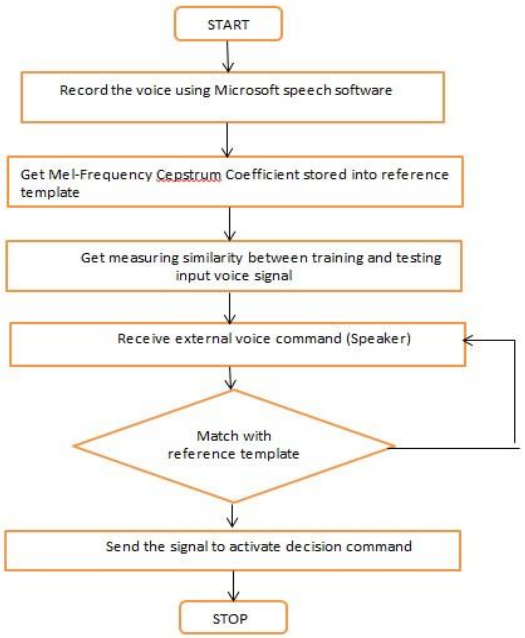

Fig7: Flowchart for Voice Flow Algorithm

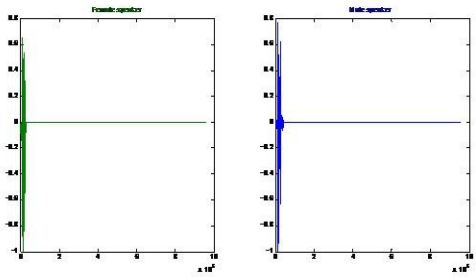

Fig 8: . Example voice signal input of two difference speakers

Figure 8 depicts how the speech analysis performance evaluation is carried out utilizing MFFC. A MFCC cepstral is a matrix; the disadvantage of this approach is that if constant window spacing is utilized, the lengths of the input and stored sequences are unlikely to be the same. Furthermore, as previously noted, the length of individual phonemes within a word will vary. For example, the word Volume Up may be said with a long /O/ and a short final /U/ or with a short /O/ and a long /U/.

The input voice signals of two different speakers are shown in Figure 8.

IV.RESULT AND DISCUSSION

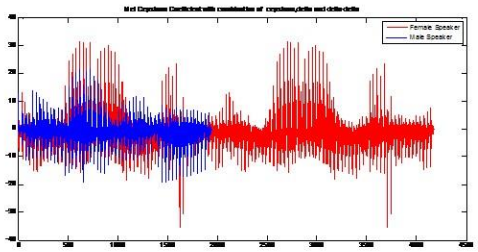

Fig.9. Mel Frequency Cepstrum Coefficients (MFCC) of one Female and Male speaker

Figure 9 shows the MFCC output of two different speakers.

The matching process needs to compensate for length differences and take account of the non-linear nature of the length differences within the words.

V. CONCLUSIONS

This paper has explored various voice recognition algorithms that are crucial for improving voice recognition performance. The techniques discussed effectively authenticated individual speakers by analyzing unique characteristics in their voice signals. The results demonstrate that these methods can be reliably employed for voice recognition applications. Furthermore, several additional techniques, such as Linear Predictive Coding (LPC), Dynamic Time Warping (DTW), and Artificial Neural Networks (ANN), are currently being researched. The insights gained from this investigation will be presented in forthcoming publications, contributing to the ongoing advancements in the field of voice recognition technology.

VI. REFERENCES

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

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

[3] Hayes H. Monson,Statistical Digital Signal Processing and Modeling, John Wiley & Sons Inc. , Toronto, 1996, ISBN 0-471-59431-8

[4] Proakis John G., Manolakis Dimitris G.,Digital Signal Processing, principles, algorithms, and applications, Third Edition, Prentice Hall , New Jersey, 1996, ISBN 0-13-394338-9

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

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

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

[8] Hiroaki Sakoe and Seibi Chiba, Dynamic Programming algorithm Optimization for spoken word Recognition, IEEE transaction on Acoustic speech and Signal Processing, Fecruary 1978.

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

[10] Davis K. H., Biddulph R. and Balashek S.,Automatic Recognition of Spoken Digits, J. Acoust. Soc. Am., 24 (6):637-642, 1952

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