Jarvis, Digital Life Assistant

ShrutiKa Khobragade
Department of Computer
Vishwakarma Institute of InformationTechnology Pune,INDIA

***Abstract*—"*Jarvis*" was main character of Tony's Stark's life assistant in Movies Iron Man. Unlike original comic in which Jarvis was Stark's human butler, the movie version of Jarvis is an intelligent computer that converses with stark, monitors his household and help to build and program his superhero suit.**

In this Project Jarvis is Digital Life Assistant which uses mainly human communication means such Twitter, instant message and voice to create two way connections between human and his apartment, controlling lights and appliances, assist in cooking, notify him of breaking news, Facebook's Notifications and many more. In our project we mainly use voice as communication means so the Jarvis is basically the Speech recognition application. The concept of speech technology really encompasses two technologies: Synthesizer and recognizer. A speech synthesizer takes as input and produces an audio stream as output. A speech recognizer on the 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 information. A direct analysisand 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. The mel-scaled frequency cepstral coefficients (MFCCs) derived from Fourier transform and filter bank analysis are perhaps the 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 in their daily life and also reduces their efforts. In our test we check all this functionality is working properly. We test this on 2 speakers(1 Female and 1 Male) for accuracy purpose.

Keywords: Feature extraction, MFCC.

I. INTRODUCTION

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 recognition 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/on 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

With the information presented so far one question comes naturally: how is speech recognition done? To get knowledge of how speech recognition problems can be approached today, a review of some research highlights will be presented. The earliest attempts to devise systems for automatic speech recognition by machine were made in the 1950's, when various researchers tried to exploit the fundamental ideas of acoustic-phonetics. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek built a system for isolated digit recognition for a single speaker [12]. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In 1959 another attempt was made by Forgie , constructed at MIT Lincoln Laboratories. Ten vowels embedded in a /b/-vowel-/t/ format were recognized in a speaker independent manner [13]. In the 1970's speech recognition research achieved a number of significant mile- stones. First the area of isolated word or discrete utterance recognition became a viable and usable technology based on the fundamental studies by Velichko and Zagoruyko in Russia , Sakoe and Chiba in Japan and Itakura in the United States . The Russian studies helped advance the use of pattern recognition ideas in speech recognition; the Japanese research showed how dynamic programming methods could be successfully applied; and Itakura's research showed how the ideas of linear predicting coding (LPC). At AT&T Bell Labs, began a series of experiments aimed at making speech recognition systems that were truly speaker independent. They used a wide range of sophisticated clustering algorithms to determine the number of distinct patterns required to represent all

variations of different words across a wide user population. In the 1980's a shift in technology from template-based approaches to statistical modeling methods, especially the hidden Markov model (HMM) approach [1].

The purpose of this paper is getting a deeper theoretical and practical understanding of a speech recognizer. The work started by examines a currently existing state of the art for feature extracting method MFCC. With this study from MFCC applying this knowledge in practical manner, the speech recognizer is implemented in .Net technology in C# language developed by Microsoft [11]. In our project we use The Speech Application Programming Interface or SAPI is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. Applications that use SAPI include Microsoft Office,Microsoft Agent and Microsoft Speech Server.. In general all API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages. In addition, it is possible for a 3rd-party company to produce their own Speech Recognition and Text-To-Speech engines or adapt existing engines to work with SAPI. Basically Speech platform consist of an application runtimes that provides speech functionality, an Application Program Interface (API) for managing the runtime and Runtime Languages that enable speech recognition and speech synthesis (text-to-speech or TTS) in specific languages.

II. SPEECH REPRESENTATION

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 short-time 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 non-stationary, 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.

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) - No speech is produced.
- 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 quasi-periodic.

Quasi-periodic means that the speech waveform can be seen as periodic over a short-time period (5-100 ms) during which it is stationary.

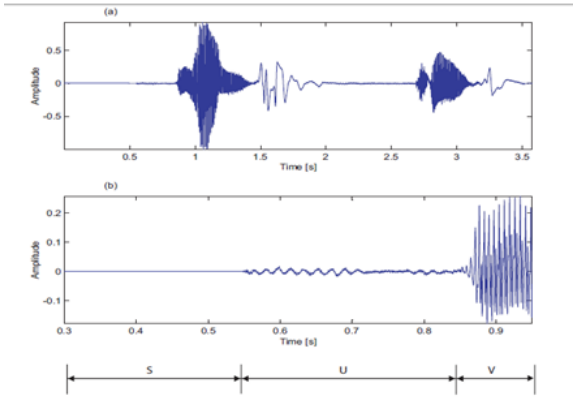

Fig1: Three State Representation

The upper plot (a) contains the whole speech sequence and in the middle plot (b) a part of the upper plot (a) is reproduced by zooming an area of the speech sequence. At the bottom of Fig. 1 the segmentation into a three-state representation, in relation to the different parts of the middle plot, is given. The segmentation of the speech waveform into well-defined states is not straight forward. But this difficulty is not as a big problem as one can think.

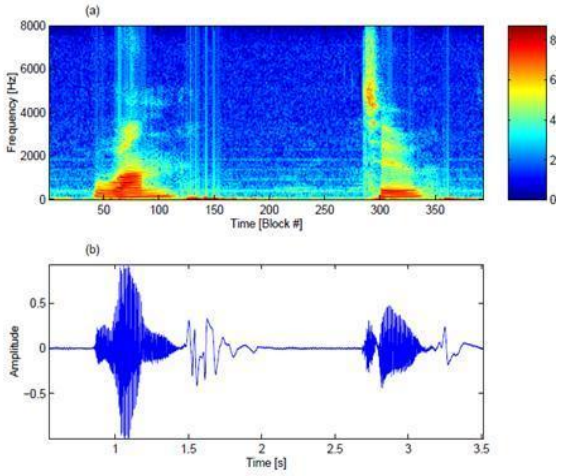

Fig 2 : Spectrogram using Welch's Method (a) and speech amplitude (b)

Here the darkest (dark blue) parts represents the parts of the speech waveform2 where no speech is produced and the lighter (red) parts represents intensity if speech is produced. speech waveform is given in the time domain. For the spectrogram Welch's method is used, which uses averaging modified periodograms [3]. Parameters used in this method are block size K= 320, window type Hamming with 62.5% overlap resulting in blocks of 20 ms with a distance of 6.25 ms between block.

2.PHONEMICS AND PHONETICS

The speech production begins in the humans mind, when he or she forms a thought that is to be produced and transferred to the listener. After having formed the desired thought, he or she constructs a phrase/sentence by choosing a collection of finite mutually exclusive sounds. The basic theoretical unit for describing how to bring linguistic meaning to the formed speech, in the mind, is called phonemes. Phonemes can be seen as a way of how to represent the different parts in a speech waveform, produced via the human vocal mechanism and divided into continuant (stationary) or non-continuant parts.

A phoneme is continuant if the speech sound is produced when the vocal tract is in a steady-state. In opposite of this state, the phoneme is non-continuant when the vocal tract changes its characteristics during the production of speech. For example if the area in the vocal tract changes by opening and closing the mouth or moving your tongue in different states, the phoneme describing the speech produced is non-continuant Phonemes can be grouped based on the properties of either the time waveform or frequency characteristics and classified in different sounds produced by the human vocal tract. The classification, may also be seen as a division of the sections in Fig 3

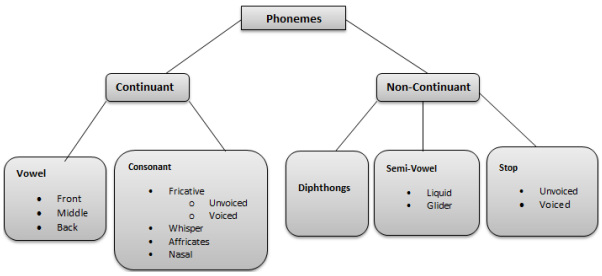

Fig3: Phoneme Classification

3 FEATURE EXTRACTION (MFCC)

The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior. MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency [8-10]. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. The overall process is shown in following Fig: 4.

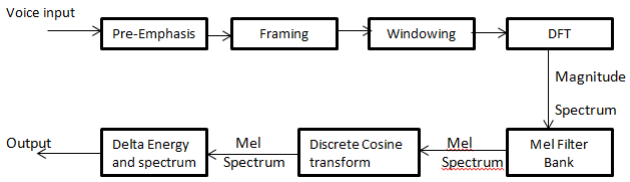

Fig4: MFCC Block Diagram

As shown in Figure 4, MFCC consists of seven computational steps. Each step has its function and mathematical approaches as discussed briefly in the following:

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.

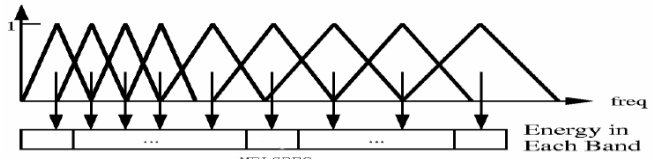

Fig. 5. Mel scale filter bank, from (young et al, 1997)

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.

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

STEP 6: 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.

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

4 METHODOLOGIES

As mentioned in [12], 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 7 shows the flowchart for overall voice recognition process.

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

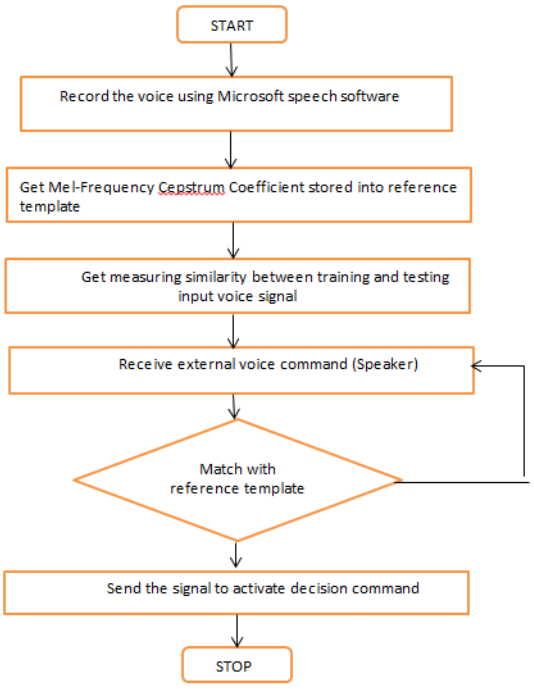

Fig7: Flowchart for Voice Flow Algorithm

5 RESULT AND DISCUSSION

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

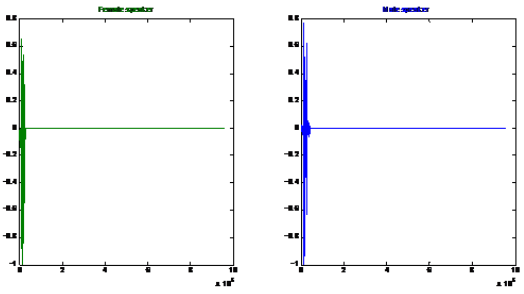

Fig 8: . Example voice signal input of two difference speakers

Figure 8 is used for carrying the 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, the lengths of the input and stored sequences is unlikely to be the same. Moreover, within a word, there will be variation in the length of individual phonemes as discussed before, Example the word Volume Up might be uttered with a long /O/ and short final /U/ or with a short /O/ and long/U/

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.

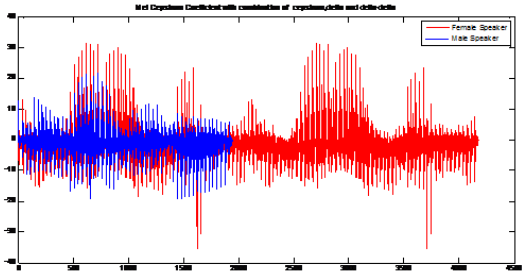

Fig.9. Mel Frequency Cepstrum Coefficients (MFCC) of one Female and Male speaker

III. CONCLUSIONS

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 publications.

IV. 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.

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

[8] http://web.science.mq.edu.au/~cassidy/comp449/html/ch11s02.html, downloaded on 2 Oct 2012.

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

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

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

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

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