JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

JARVIS: THE DIGITAL LIFE ASSISTANT

Priyanjali kumari

Galgotias University Department of Computer

Science and Engineering,

Greater Noida, Uttar Pradesh, INDIA

ABSTRACT

The objective of this project is to create a personal assistant for Linux-based Software. Visual assistants such as Cortana for Windows and Siri for iOS serve as inspiration for Jarvis. It's intended to give a simple interface for carrying out a variety of actions based on exact instructions. Users can interact with the assistant by using voice commands or by using keyboard input.

As a personal assistant, Jarvis assists the end user with daily activities such as regular human chat, search queries on google, Bing or yahoo, video search, retrieval images, live weather, word descriptions, search for tree description. Signalling health-based recommendations and reminding the user about planned events and activities. Machine learning is used to assess user statements or instructions in order to produce the best answer.

Keywords: Linux Systems, Personal Assistant, Automation and Machine Learning

I. INTRODUCTION

Speech complements or replaces the use of mouse, keyboards, controllers, and body language as an effective and natural way for individuals to engage with programmes. A hands-free, yet intuitive way to communicate with apps, speech allows people to produce and stay informed in a variety of situations where other visual connectors may not. Speech recognition is a very useful topic in many programs and environments in our daily lives. Often a speech impediment is a machine that understands people and their voice spoken in a certain way and can do something afterward. A different aspect of speech recognition is helping people with disabilities or other types of disabilities. To make their daily routine easier, voice control may be helpful. With their voice they could turn on / off the light switch or use other household appliances. This leads to a discussion about smart homes where these services can be made available to ordinary people and people with disabilities.

With the information presented so far one question arises automatically: how is speech recognition performed? For information on how speech recognition problems can be addressed today, a review of some of the highlights of the study will be presented. Various researchers attempted to explore the core ideas of acoustic-phonetics in the 1950s, which led to the first attempts to create automatic speech recognition systems. In 1952, at Bell Laboratories, Davis, Biddulph, and Balashek developed a single-digit single-system recognition system [12]. The system relies heavily on measuring spectral sound during each digital vowel region. Fororgie, which was created at MIT Lincoln Laboratories, attempted a second attempt in 1959. Ten vowels embedded in / b / -vowel- / t / are recognized independently of the platform [13]. In the 1970's a speech recognition study acquired a number of gemstones. Initially the location of the distinctive name recognition or incomplete pronunciation became a practical and useful technology based on the basic research of Velichko and Zagoruyko in Russia, Sakoe and Itakura in the United States, as well as Chiba in Japan. Russian studies helped to improve the use of pattern recognition concepts in speech recognition; Japanese research has shown how effective planning methods can be used effectively; and Itakura research has shown that the ideas for predicting linear code (LPC). At AT&T Bell Labs, a series of experiments aimed at making speech recognition systems truly independent of speakers have begun. They used complex integration algorithms to determine the number of different patterns needed to represent all different types of words for multiple users. In the 1980s the technological shift from model-based approaches to mathematical modeling methods, most notably the hidden Markov model (HMM) method [1].

The purpose of this paper is to gain a deeper understanding of the theory and practice of the speech therapist. Work initiated by examining the current state of the MFCC feature. With this MFCC study using this information in a practical way, the speech therapist is using Net # technology in the C # language developed by Microsoft [11]. In The Speech Application Programming Interface, or SAPI, is an API that we employ in our project. developed by Microsoft to allow the use of speech recognition and speech integration within Windows applications. Programs that use SAPI include Microsoft Office, Microsoft Agent and Microsoft Speech Server. Normally all APIs are designed in such a way that a software developer can write an application to perform speech recognition and integration using a standard set of interactive, multi-accessibility areas of programming languages. In addition, it is possible for a third party company to produce its own Speech and Text-ToSpeech engines or modify existing engines to work with SAPI. Basically the Speech field contains app launch times that provide speech functionality, Application Interface (API) for running time management and runtime languages that allow speech recognition and speech integration (text-to-speech or TTS) in specific languages.

II.SPEECH REPRESENTATION

Speech signal and all its components can be represented in two different domains, time and frequency frequency Speech signal is a dynamic time-varying signal in the sense that, when tested in a short period of time (between 5 and 100 ms), its features are suspended for some while. This is not the case if we look at the speech signal under long-term perspective (approximately T> 0.5 s). In this case the features of the symbols are not fixed, which means that they change to produce different sounds spoken by the speaker.

1 THREE STATE REPRESENTATIONS

Representation of the three kingdoms is one way to distinguish events from speech. These are exciting developments in the representation of the three kingdoms

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

Quasi-periodic means that the speech wave can be seen as a period in a short period of time (5-100 ms) when stopped.

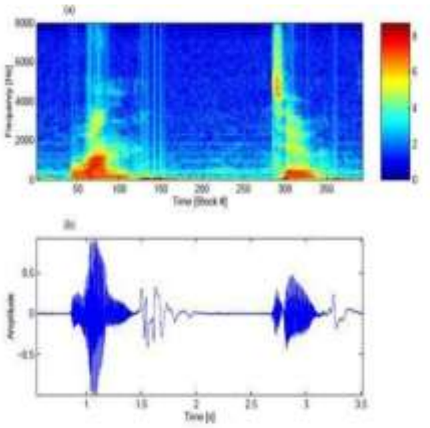

Fig 2 : Spectrogram using Welch's Method (a) and speech amplitude (b)

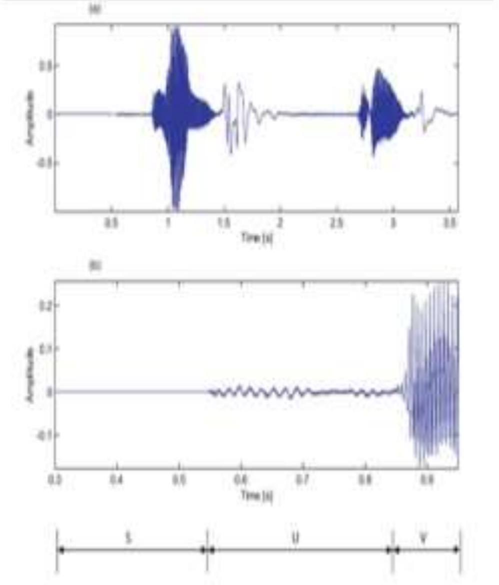

Fig1: Three State Representation

The above structure (a) contains the whole speech sequence and in the middle section (b) part of the upper structure (a) is redone by bringing the speech sequence closer. Under Figure 1 the division into representation of the three regions, with respect to the different parts of the central structure, is given. The division of the waveform format into well-defined contexts does not mean going forward. But this difficulty is not as big a problem as one

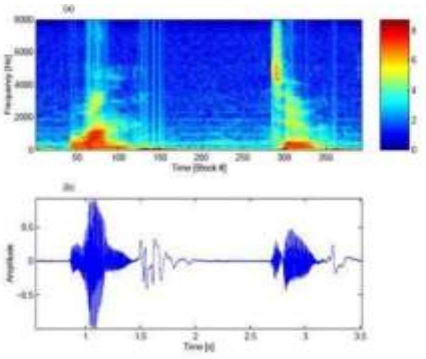

Fig 2 : Spectrogram using Welch's Method (a) and speech amplitude (b)

The lightest regions (red) indicate the intensity when speech is generated, whereas the darkest parts (blue) indicate the parts of the speech structure2 where speech is not created. Speech waveform is given in time zone. The Welch spectrogram method is used, which uses periodically modified periodograms [3]. The parameters used in this method are block size K = 320, a Hamming window type with a spacing of 62.5% that leads to 20 ms blocks and a distance of 6.25 ms between blocks.

2. PHONICAL AND PHONICAL

Speech production begins in the human mind, when he makes up an idea that must be produced and passed on to the listener. After constructing the idea you want, form a sentence / sentence by selecting a set of different sound effects. The basic unit of theory to explain how to convey the meaning of language in a built-in speech, in mind, is called phonemes. Phonemes can be seen as a means of representing different parts of the speech wave system, produced by the human voice machine and divided into continuous (vertical) or non-continuous area.

The phoneme is continuous when speech sounds are produced when the voice level is constant. In contrast, the phoneme does not continue when the voice changes its characteristics when producing speech. For example, if the position in the voice region changes by opening and closing the mouth or by moving your tongue in different contexts, the phonograph describing the speech produced does not proceed and is separated by different sounds produced by the human voice. Separation, it can also be seen as the division into sections in Fig. 3

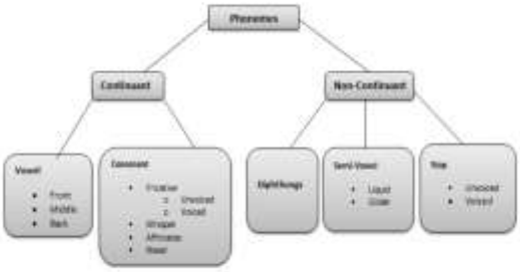

Fig3: Phoneme Classification

3 FEATURE EXTRACTION (MFCC)

The output of the best representation of the acoustic signal parameter is an important function to produce better recognition performance. The effectiveness of this category is important in the next phase as it affects its behavior. MFCC is based on human perceptions that can detect more than 1 Khz waves. In other words, in the MFCC it is based on the known differences in the critical ear bandwidth and frequency [8-10]. The MFCC has two consecutive types of filters that are subdivided at low levels below 1000 Hz and logarithmic space above 1000Hz. A specific tone is present in the Mel Frequency Scale to capture the important phonetic aspect of speech. The whole process is shown in the following figure: 4.

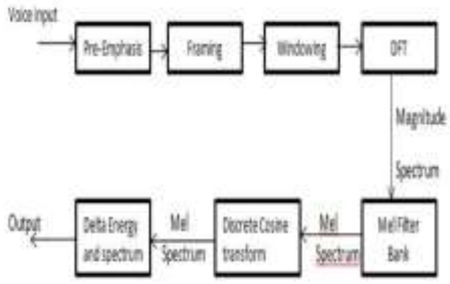

Fig4: MFCC Block Diagram

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

STEP 1: PRE-EMPHASIS

This step detects signal passing through a filter that emphasizes high waves. This process will increase signal strength at higher frequencies.

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

Let us consider = 0.95, which makes 95% of any single sample considered to be from the previous sample.

STEP 2: FRAMING

The process of separating speech samples obtained from analog to digital (ADC) translation into a small frame with a length within 20 to 40 msec. The voice signal is divided into N-frame frames. The adjacent frames are divided by M (M <N) .The standard values used are M = 100 and N = 256.

STEP 3: HAMMING WINDOWING

The Haming window is used as a window shape by considering the following block in the feature processing series and covers all adjacent frequency lines. The equation of the Haming window is given as: If the window is defined as W (n), 0 ≤ n ≤ N-1 where

N = number of samples per frame

Y [n] = Exit signal

X (n) = input signal

W (n) = Haming window, then the window signal is

Shown below:

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

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

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

STEP 4: FAST FOURIER TRANSFORM

Convert each frame of N samples from a time zone to a frequency domain. Fourier Transform is the transformation of the glottal pulse U [n] and vocal tract impulse response H [n] over time. This statement supports the figure below:

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

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

STEP 5: MELFILTER BANK PROCESSING

The range of frequencies in the FFT spectrum is very wide and the voice signal does not follow the line scale. The filter bank according to Mel's scale as shown in Figure 5 is made.

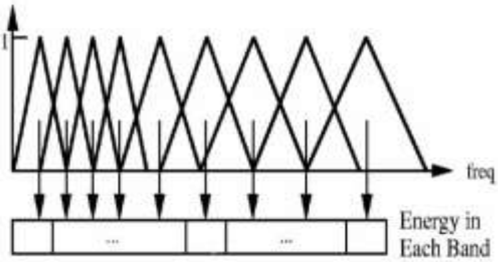

Fig. 5. Mel scale filter bank, from (young et al, 1997)

This figure shows the set of triangular filters used to calculate the total weight of the spectral components of the filter so that the process output reaches the Mel scale. The frequency response of each filter size is triangular and equal to one in the middle frequencies and descends along the zero to the average frequencies of the two adjacent filters [7, 8]. Then, the output of each filter is the sum of its spectral filtered components. After that next number extracted to calculate Mel at a given frequency in HZ.

$F(Mel) = [2595 * \log\_10 [1 + f / 700]]$ (5)

STEP 6: Formal COSINE TRANSFORMATION

This is the process of converting the Mel spectrum log into a time zone using Discrete Cosine Transform (DCT). The effect of the modification is called the Mel Frequency Cestrum Coefficient. A set of coefficient is called acoustic vectors. Therefore, each input word is converted into an acoustic vector sequence.

STEP 7: DELTA ENERGY & DELTA SPECTRUM

Voice signal and frames change, as does the format's inclination to change. Therefore, there is a need to add changes related to cepstral features over time.

There are 13 delta or speed features (12 cepstral features as well as power), and 39 has a double delta or acceleration feature. The strength of the x-frame frame in the window from the t1 time sample to the t2 time sample, is represented in the figures below:

$E = \sum X^2 [t]$ - (6)

Each of the 13 delta elements represents a change between frames corresponding to the cepstral or force factor, while each of the 39 double delta elements represents a change between frames on the corresponding delta elements.

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

4. METHODOLOGY

Process	Description

1) Speech	2women(age=20, age=53)
	2Male(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	16000KHz
6) Feature Computational	39 double delta MFCCcoefficient

As mentioned in [12], voice recognition works based on the premise that individual traits are different from different speakers. The signal during training and the test session can vary greatly due to many factors such as changes in human voice over time, health status (e.g. fluorescent speaker), speech quality and acoustical sound and location of various microphones. Table II provides details of the recording details and training session, while Figure 7 shows the flow chart of the entire voice recognition process.

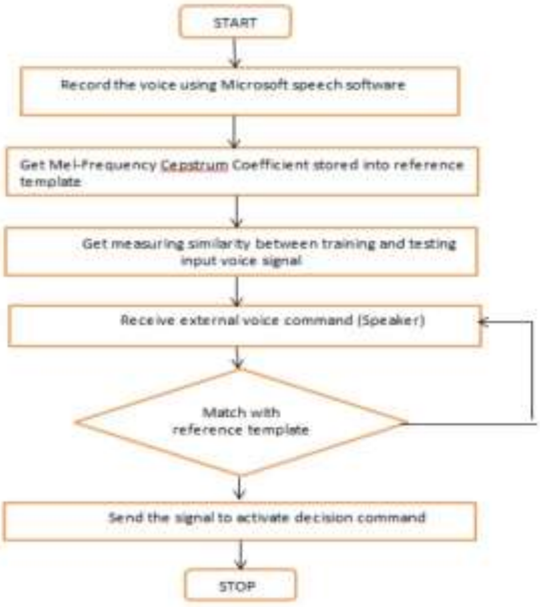
START
Record the voice using Microsoft speech software
Get Mel-Frequency Cepstrum Coefficient stored into reference template
Get measuring similarity between training and testing input voice signal
Receive external voice command (Speaker)
Match with reference template
Send the signal to activate decision command
STOP

Fig7: Flowchart for Voice Flow Algorithm

5 RESULT AND CONCLUSION

The included voice features of two different speakers are shown in Figure

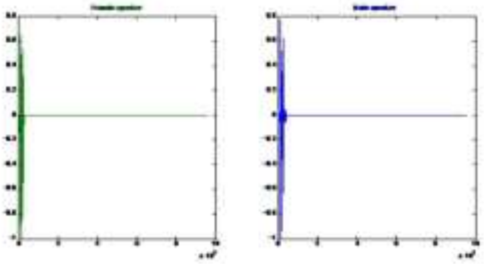

Fig 8: . Example voice signal input of two difference speakers

Figure 8 is used to manage the performance analysis of voice analysis using MFFC. The MFCC cepstral is a matrix, the problem with this is that if continuous window partitions are used, the length of the input and the stored sequence are unlikely to be the same. In addition, within a single word, there will be variations in the length of each of the phonemes as discussed earlier, for example the word Volume Up may be pronounced with / O / and short / U / or / O / and length / U /

The MFCC output of two separate speakers is shown in Figure 9. The matching process needs to compensate for the length difference and consider the indirect type of length difference within the words.

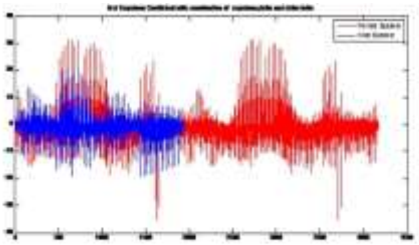

Fig.9. Mel Frequency Cepstrum Coefficients (MFCC) of one Female and Male speaker

III. CONCLUSIONS

This paper discusses voice recognition algorithms that are important in improving voice recognition functionality. The method was able to authorize a specific speaker based on personal information embedded in the voice signal. The results show that these techniques 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 outcomes will be published in the near future.

IV. REFERENCES:

[1] Rabiner Lawrence, Juang Bing-Hwang. Highlights 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., Timely Analysis of Speech Symbols, IEEE Press, ISBN 0-7803-5386-2

[3] Statistical Digital Signal Processing & Modeling, Hayes H. Monson, John Wiley and 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 ownership using MFCC and HMM-based strategies, University of Florida, April 25,2004.

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