International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013)

Voice Activated Calculator

Suraj Jadhav[1], Shashank Kava[2], Sanket Khandare[3], Sayali Marawar[4], Savitha Upadhya[5]
Department of Electronics and Telecommunications Engineering, Fr.C.Rodrigues Institute of Technolog,Vashi,India.

***Abstract*- In the era of human machine interface, speech recognition is being looked upon as highly fascinating field to achieve human computer interaction. Several applications of speech recognition have emerged over the past years including voice dialling, voice query recognition for call routing and simple data entry. Speech recognition is the process of automatic determination of linguistic information conveyed by human speech using a computer and reconstructing the text of a spoken sentence from the continuous acoustic signal, overcoming the associated noise induced disturbances. The aim of this paper is to present an overview of the application of speech recognition system for mathematical computing using a basic computer without any additional hardware. The system developed would be able to take voice inputs from the user, match it with the database to recognise the digits, perform the required mathematical operation and display the output. This technology would make the calculation process easier for the user by supporting the concept of multitasking as well as eliminating human errors due to mistyping.**

***Keywords*- *H*uman computer interaction, speech recognition, mathematical computing, multitasking.**

I. Introduction

A recent survey conducted among people performing calculations on a daily basis shows that they prefer using a separate calculator instead of a calculator present in their work station computers to save time and achieve multitasking. They suggested that using the computer based calculator is time consuming as they have to switch screens and this leads to typing mistakes. Given an option, they would rather prefer a voice activated calculator running in the background on their computers, to which inputs can be given in the form of spoken digits and operations, and would display the result on the screen. This voice activated calculator can be implemented on a basic computer system with no additional hardware. The developed software would be able to take operands and the operation commands as voice inputs and perform the mathematical operation and display the output on screen. This system includes two main stages. First stage is the training phase which consists of feature extraction using Mel Frequency Cepstral Coefficients (MFCC) and storage of extracted features as training data in the form of reference templates. Second stage is the testing phase in which the features of real time voice inputs are extracted and compared with the reference template using Euclidean distance criterion to recognise the input digit.

After recognising the input digits, the required operation is preformed and the output is displayed in the form of text on the screen.

II. Block Diagram Of The System

The Figure 1 shows the basic block diagram of the Voice Activated Calculator. The speech is recorded using the software Audacity 1.2.6. The voice samples include digits from zero to nine and four basic mathematical operations, plus, minus, multiply and divide. All these samples are given as input to the feature extraction algorithm, their MFCCs are calculated and the feature vectors are stored as templates in a database.

The real time input speech is sampled at 16000 samples per second. The input speech is segmented using Energy Extraction algorithm and stored in buffers as and when the voice samples arrive so that the feature extraction of these voice samples takes place systematically. The obtained feature vectors from the real time input speech and the templates are matched using Euclidean Distance criterion. Once the match is found these recognized digits will be forwarded for the calculation and the result will be displayed in the form of text on the computer.

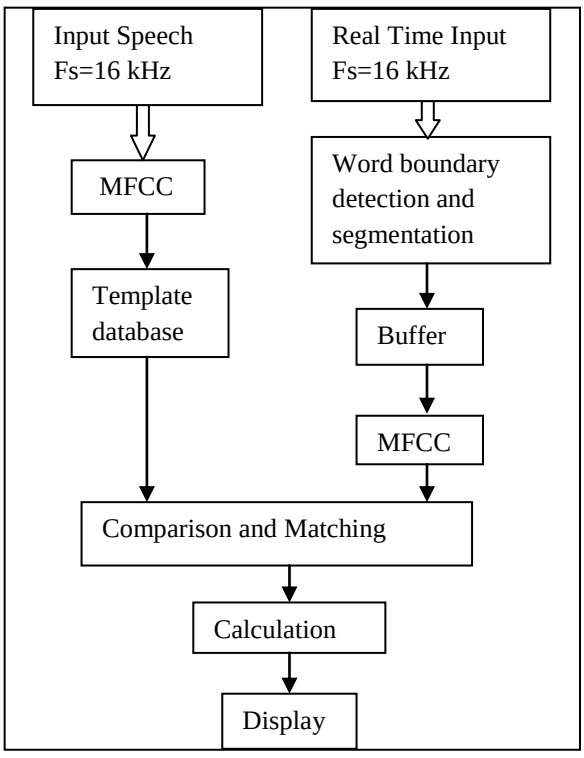

Figure 1. Block Diagram of the System

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013)

III. Mel Frequency Cepstral Coefficients

For speech/speaker recognition, the most commonly used acoustic features are Mel-scale Frequency Cepstral Coefficient (MFCC) [1], shown in Figure 2. MFCC takes human perception sensitivity with respect to frequencies into consideration, and therefore are best for speech/speaker recognition.

A. Pre-emphasis

The speech signal s(n) is sent to a high-pass filter:

$$s_2(n) = s(n) - \alpha s(n-1) \quad (1)$$

Where $s_2(n)$ is the output signal and the value of α is usually between 0.9 and 1.0.

The z-transform of the filter is,

$$H(z) = 1 - \alpha z^{-1} \quad (2)$$

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans to amplify the importance of high-frequency formants.

B. Frame Blocking

The input speech signal is segmented into frames of 10~20 ms with overlap of 50% of the frame size. . Since the sample rate is 16 kHz and the frame size is 256 sample points, the frame duration is 256/16000 = 0.016 sec = 16 ms. Additionally, for 50% overlap meaning 128 points, then the frame rate is 16000/(256-128) = 125 frames per second.

C. Windowing using Hamming window

Since speech signal is quasi-stationary, it is analysed for a short duration of time. Hence windowing technique is used. Hamming window is used as it has less noisy spectrum in comparison with rectangular window. Length of hamming window is 256 samples. Hamming window defined by:

$$w(n) = 0.54 - 0.46\cos(2\pi n/(N-1)) \quad (3)$$
$$\text{for } 0 \leqq n \leqq N-1.$$

D. Fast Fourier Transform (FFT)

Spectral analysis of speech signal indicates that different timbres in speech signals corresponds to different energy distribution over frequencies. Therefore FFT is performed to obtain the magnitude frequency response of each frame.

E. Mel Frequency Filter Bank

The magnitude frequency response is multiplied by a set of 40 triangular bandpass filters that are designed according to the Mel scale which is related to the human auditory system [2].

The Mel scale has 13 filters which are linearly spaced upto 1 kHz and remaining 27 filters which are logarithmically spaced upto 8kHz. Frequency response of Mel filter bank is shown in Figure 3. The relationship between Mel frequency and natural frequency is given by,

$$mel(f) = 1125 * ln(1 + f/700) \quad (4)$$

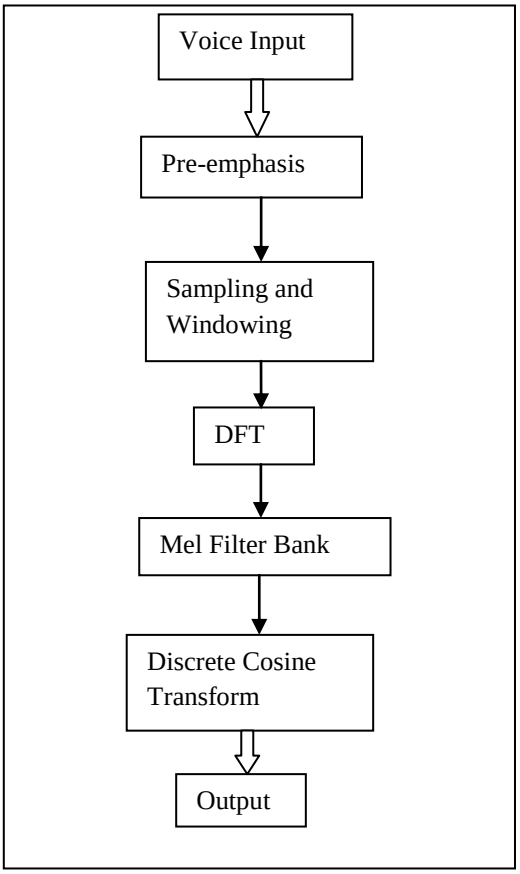

Figure 2. MFCC Block Diagram

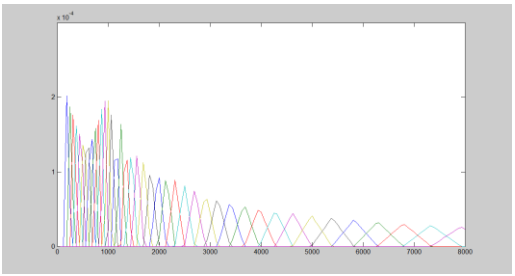

Figure 3. Mel Filter Bank

F. Discrete Cosine Transform (DCT)

Here DCT is applied to the output of Mel filter bank to get 'L' MFCCs. The equation for DCT is,

$$Cm = \sum_1^N \cos\left[m * (k - 0.5) * \frac{\pi}{N}\right] * E_k; \text{ m=1,2…L} \quad (5)$$

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013)

Where N is the number of triangular bandpass filters, L is the number of MFCCs.

IV. Endpoint Point Detection

An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the endpoint detection problem. The accurate detection of a word's start and end points means that subsequent processing of the data can be kept to a minimum. The start and end of the voice input is detected by analysing its energy profile. When the energy of the voice signal rises above a certain set threshold value, it marks the presence of voice input and when the energy level falls below the threshold it is considered as the end of the signal. The same concept is used to achieve isolation of digits in a continuous speech signal.

V. Speech Matching

Voice samples of digits zero to nine and the mathematical operators of 'plus', 'minus', 'into', 'by' are passed through the MFCC algorithm and their respective Mel Frequency Cepstral Coefficients are extracted and a reference template is obtained. Now the output of end point detection algorithm that takes the real time voice input and isolates the digits and mathematical operators, is also processed and their MFCCs are obtained.

The MFCCs of the template voice sample and real time voice input for the digit "1" is shown in the figure 4. The MFCCs of the two samples are found to be relatively similar and thus speech recognition is achieved.

To achieve speech matching, Euclidean distance [3] is calculated between the MFCCs of real time input and the MFCCs of all the signals in the template. The signal corresponding to the minimum Euclidean distance is the desired digit or mathematical operator.

For two discrete signals x and y of equal length N the Euclidean distance is given by,

$$d(x,y)=\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \quad (6)$$

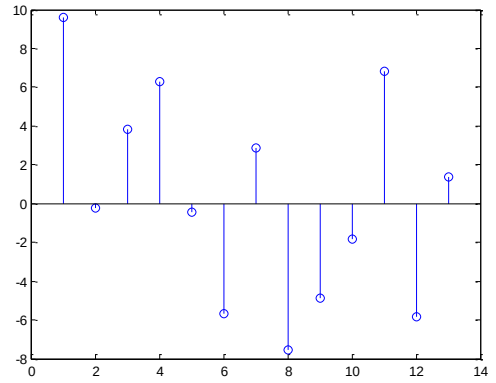

Figure 4.1. Digit '1' Template

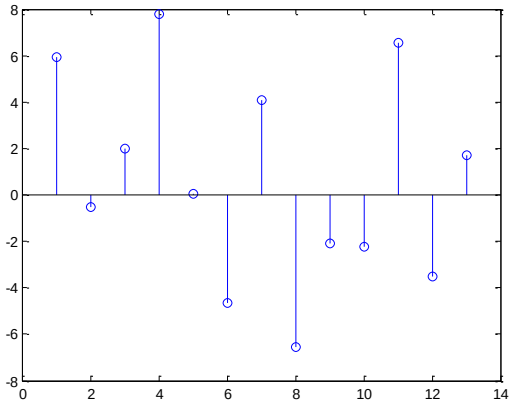

Figure 4.2. Digit '1' Real-time input

VI. Implementation And Results

Audio files of digits spoken by the speaker are recorded using the Audacity software. Figure 5 shows typical sound signal recorded using Audacity software. The recorded speech signals are sampled and stored. The sampling is done at a rate of 16000 samples per second. Each speech signal is divided into windows of 16ms each and hence, 256 samples each.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 2, February 2013)

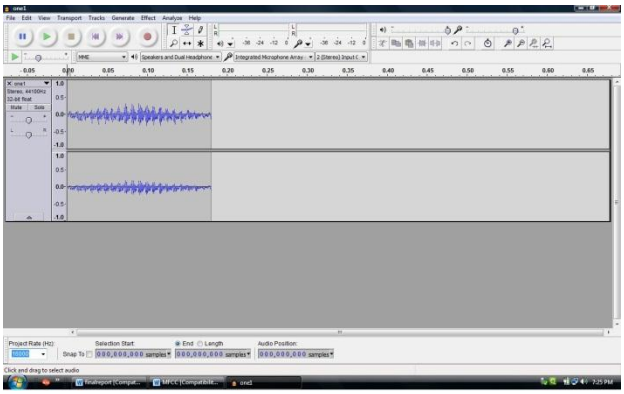

Figure 5. Voice Sample of Digit '1' using Audacity software

The 13 MFCCs of the recorded speech signal are extracted using the extraction algorithm and stored in a template database.

The real time input is a continuous speech signal for example 'one plus two', recorded using Audacity software. This continuous signal is isolated into individual digits and mathematical operator using Endpoint detection algorithm. The threshold is set experimentally after analysing various speech signals to be 0.0005 in the energy spectrum for an individual frame. The MFCCs of this signal are calculated. The recorded speech and the isolation of the signal are shown in Figure 6.

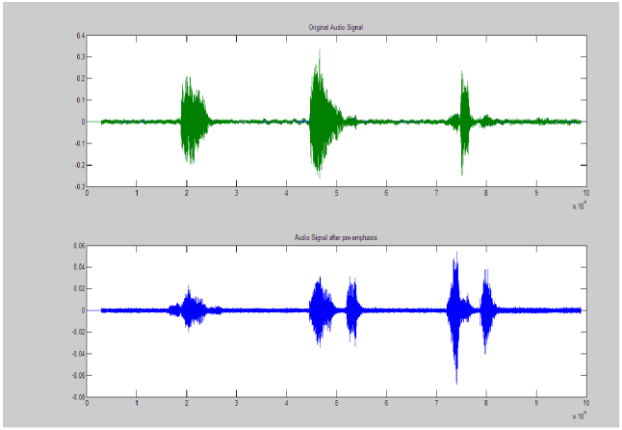

Figure 6.1 Real time speech input signal

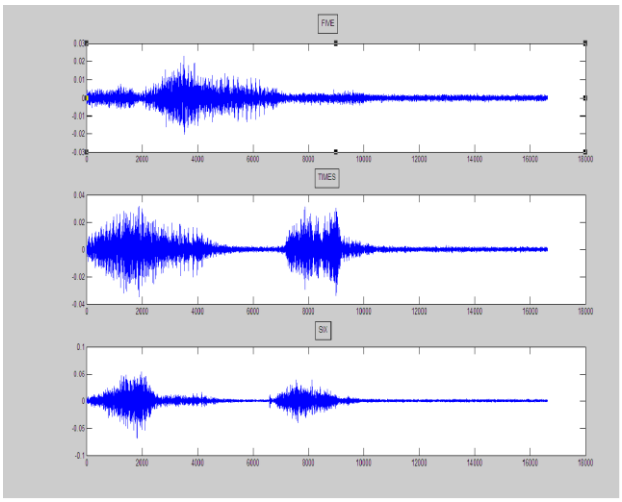

Figure 6.2. Isolated Digits

These MFCCs are compared with the template database and word matching is obtained using Euclidean distance algorithm. After the input speech is matched and the requested digit and operation is identified, it is given as the input to the calculation algorithm. The output is calculated and displayed on the screen.

REFRENCES

[1] S. Molau, M. Pitz R. Schliitel and H. Ney, "Computing Mel frequency cepstral coefficients on the power spectrum", Proc. IEEE International conference on Acoustis, Speech and Signal processing, 2001 (ICASSP 2001), vol.1, pp. 73-76, 2001.

[2] L.R.Rabiner and R.W. Schafer, (Third Edition), Digital Processing of Speech Signals, Pearson Education, Delhi, 2009.

[3] Douglas O'Shaughnessy, (Second Edition),Speech Communications Humans and Machines, Universities Press, 2001.