

## Real Time Isolated Word Recognition using Adaptive Algorithm

Siva Prasad Nandyala<sup>1+</sup>, T. Kishore Kumar<sup>2</sup>

<sup>1,2</sup> Department of Electronics and Communication Engineering  
National Institute of Technology, Warangal, Andhra Pradesh, 506004, India

**Abstract.** In this paper, we implemented the speech recognition with including the speech enhancement before the feature extraction to remove the noise source mixed with the speech signal. The main task is to recognize list of words which the speaker says through the microphone, and export the word one by one on the application immediately. The features used are the mel-frequency cepstral coefficients (MFCC) which gives the good discrimination of the speech signal. The Dynamic Programming algorithm used in the system measures the similarity between the stored template and the test template for the speech recognition. Noise removal is very important in speech recognition as it degrades the performance of the recognition rate. The Least Mean Square (LMS) algorithm is used in this work to remove the noise present in the speech signal. The recognition rate obtained with speech enhancement is 93% compared to the 82% without enhancement.

**Keywords:** Isolated Word Recognition, Speech Enhancement, Feature Extraction, Mel Frequency Cepstral Coefficients, Dynamic Time Warping, Least Mean Square (LMS) algorithm

### 1. Introduction

Speech recognition (also known as automatic speech recognition or computer speech recognition) converts spoken words to text or some electrical format. The term “voice recognition” is sometimes used to refer to recognition systems that must be trained to a particular speaker, as is the case for most desktop recognition software. Recognizing the speaker can simplify the task of translating speech. The speech recognition has applications in so many areas like Telephone Conversation (with out the assistance of operator for searching telephone directory), Education system (for teaching the foreign students in correct pronunciation), in playing video games and toys control, home appliance control(for washing machines, ovens),in military applications like training of air traffic controllers, in artificial intelligence (for robotics), data entry, preparation of documents for specific application like in dictation for lawyers and doctors, for assisting people with disabilities and in translation of one language from another language between people of different nations.

Speech enhancement is also the most important field of speech processing .The existence of background noise may result in a significant degradation in the quality and intelligibility of the speech signal in many applications, for example, speech recognition system, hearing aids, teleconferencing systems and mobile communications [1].

### 2. Design and Development of the Isolated Word Recognition System

In our system the spoken words are instantly captured through Microphone. The recognized word result is displayed immediately on the graphical user interface. That is the reason we say the system is the real time. The design and development of our work can be divided into two parts.

---

<sup>+</sup> Corresponding author.  
E-mail address: speech.nitw@gmail.com.

One is the training part in which we record the speech of the user directly by using the laptop microphone and get its features and save the original speech signal and its features to the database. In the testing part first we record the speaker's speech and its MFCC feature is calculated and dynamic programming is applied to calculate the distance between this speech and the saved speech to take the decision. The result of the spoken word is displayed on the GUI interface instantly. The basic block diagram of the system is shown in the below figure 1.

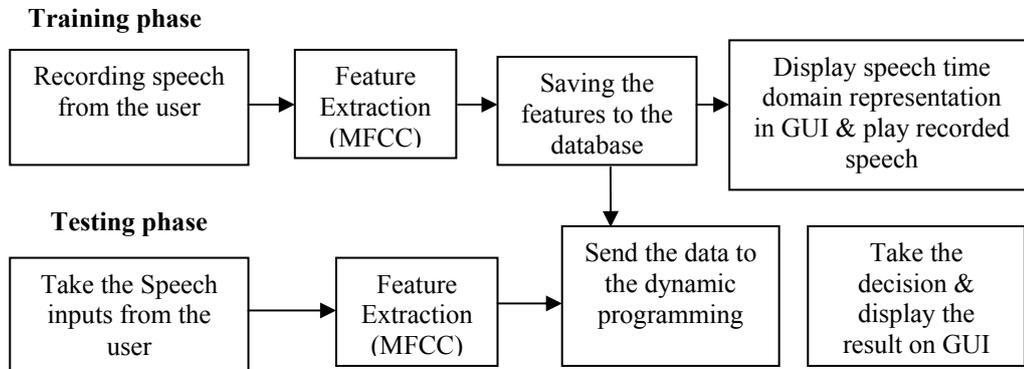


Figure 1: Real Time Isolate Word Recognition system

The LMS algorithm is used for the removal of noise in speech signal before the feature extraction block.

The feature extraction is explained below

## 2.1. Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly timed varying signal (it is called quasi-stationary). An example of speech signal is shown in below figure. When examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. The sample speech signal is shown in the below figure 2.

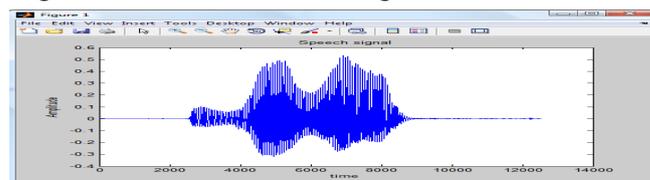


Figure 2: An example of speech signal

## 2.2. Mel-Frequency Cepstral Coefficients

The features we used are the Mel-frequency Cepstrum Coefficients (MFCC) which has been the dominant features for recognition from a long time. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. Studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency,  $f$ , measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 Mel's[3].

The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. As been discussed previously, the main purpose of the MFCC processor is to mimic the behaviour of the human

ears. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to mentioned variations[3].

The main steps used for the MFCC are as follows

1) Pre-emphasizing 2) Frame Blocking 3)Windowing 4)Discrete Fourier Transform 5)Mel-Frequency warping 6) Cepstrum

### 1) Pre-emphasizing

The range of human hearing is about 20-20,000 Hz. Frequencies below 1000 Hz are stated as low frequencies and frequencies above 1000 Hz are termed as high frequencies. In a voice waveform major of the information is found at high frequency compared to low frequencies. So initially the high frequency components of the speech input are boosted 20 dB/decade to improve the overall SNR ratio. This process helps the voice to resume its major information effectively thereby lowering the spectral distortion in frequency domain.

### 2) Frame Blocking

In this step the continuous speech signal is blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame, and overlaps it by N - M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N - 2M samples. This process continues until all the speech is accounted for within one or more frames. Typical values for N and M are N = 256 (which is equivalent to ~ 30 msec windowing) and M = 100.

### 3) Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. If we define the window as

$$w(n), 0 \leq n \leq N - 1 \quad (\text{eq.1})$$

Where N is the number of samples in each frame, then the result of windowing is the signal.

$$y_l(n) = x_l(n)w(n), \quad 0 \leq n \leq N - 1 \quad (\text{eq.2})$$

The window we used is the Hamming window, which has the form:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (\text{eq.3})$$

### 4) Discrete Fourier Transform

The next processing step is the Discrete Fourier Transform, which converts each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of N samples {xn}, as follow:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn / N}, \quad k = 0,1,2,\dots, N - 1 \quad (\text{eq.4})$$

The result after this step is often referred to as spectrum or periodogram

### 5) Mel-Frequency Wrapping

As mentioned above, psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The Mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000

Mels. Therefore we can use the following approximate formula to compute the Mels for a given frequency  $f$  in Hz:

$$\text{Mel}(f) = 2595 \cdot \log_{10}(1 + f / 700) \quad (\text{eq.5})$$

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the Mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel frequency interval. The modified spectrum of  $S(\omega)$  thus consists of the output power of these filters when  $S(\omega)$  is the input. The number of Mel spectrum coefficients,  $K$ , is typically chosen as 20. Note that this filter bank is applied in the frequency domain. A useful way of thinking about this Mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

## 6) Cepstrum

In this final step, we convert the log Mel spectrum back to time. The result is called the Mel frequency Cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those Mel power spectrum coefficients that are the result of the last step are

$$\tilde{S}_0, k = 0, 2, \dots, K - 1$$

We can calculate the MFCC's,  $\tilde{c}_n$ , as

$$\tilde{c}_n = \sum_{k=1}^K (\log \tilde{S}_k) \cos \left[ n \left( k - \frac{1}{2} \right) \frac{\pi}{K} \right], \quad n = 0, 1, \dots, K-1 \quad (\text{eq.6})$$

Note that we exclude the first component,  $\tilde{c}_0$ , from the DCT since it represents the mean value of the input signal which carried little speaker specific information.

## 2.3. DYNAMIC PROGRAMMING

Dynamic Time Warping is one of the pioneer approaches to speech recognition. Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed.



Figure 3: Dynamic Time Warping

DTW operates by storing a prototypical version of each word in the vocabulary into the database, then compares incoming speech signals with each word and then takes the closest match. But this poses a problem because it is unlikely that the incoming signals will fall into the constant window spacing defined by the host. For example, the password to a verification system is "HELLLO". When a user utter "HEELLOO", the simple linear squeezing of this longer password will not match the one in the database. This is due to the mismatch spacing window of the speech "HELLLO" [2].

Figure 3 above shows the graph on Dynamic Time Warping, where the horizontal axis represents the time sequence of the input stream, and the vertical axis represents the time sequence of the template stream. The path shown results in the minimum distance between the input and template streams. The shaded in area represents the search space for the input time to template time mapping function.

Let  $X(x_1, x_2, \dots, x_n)$  and  $Y(y_1, y_2, \dots, y_m)$  be two series with the length of  $n$  and  $m$ , respectively, and an  $n \times m$  matrix  $M$  can be defined to represent the point-to-point correspondence relationship between  $X$  and  $Y$ , where the element  $M_{ij}$  indicates the distance  $d(x_i, y_j)$  between  $x_i$  and  $y_j$  [4]. Then the point-to-point

alignment and matching relationship between X and Y can be represented by a time warping path  $W = \langle w_1, w_2, \dots, w_K \rangle$ ,  $\max(m, n) \leq K < m + n - 1$ , where the element  $w_k(i, j)$  indicates the alignment and matching relationship between  $x_i$  and  $y_j$ . If a path is the lowest cost path between two series, the corresponding dynamic time warping distance is required to meet

$$DTW(X, Y) = \min_W \left\{ \sum_{k=1}^K d_k, W = \langle w_1, w_2, \dots, w_k \rangle \right\} \quad (\text{eq.7})$$

Where  $d_k = d(x_i, y_j)$  indicates the distance represented as  $w_k = (i, j)$  on the path W. Then the formal definition of dynamic time warping distance between two series is described as

$$DTW(\langle \rangle, \langle \rangle) = 0$$

$$DTW(X, \langle \rangle) = DTW(\langle \rangle, Y) = \infty \quad (\text{eq.8})$$

$$DTW(X, Y) = d(x_i, y_i) + \min \left\{ \begin{array}{l} DTW(X, Y[2:-]), \\ DTW(X[2:-], Y), \\ DTW(X[2:-], Y[2:-]) \end{array} \right.$$

where  $\langle \rangle$  indicates empty series,  $[2 : -]$  indicates a subarray whose elements include the second element to the final element in an one-dimension array,  $d(x_i, y_j)$  indicates the distance between points  $x_i$  and  $y_j$  which can be represented by the different distance measurements, for example, Euclidean Distance. The DTW distance of two-time series can be calculated by the dynamic programming method based on accumulated distance matrix, whose algorithm mainly is to construct an accumulated distance matrix

$$r(i, j) = d(x_i, y_j) + \min \{ r(i-1, j), r(i, j-1), r(i-1, j-1) \} \quad (\text{eq.9})$$

Any element  $r(i, j)$  in the accumulated matrix indicates the dynamic time warping distance between series  $X[1:i]$  and  $Y[1:j]$ . Series with high similar complexity can be effectively identified because the best alignment and matching relationship between two series is defined by the dynamic time distance. The distance threshold used is 200 in this work.

### 3. Training

In the training phase we have trained the ten digits from zero to nine by the specific user as the system is speaker dependent. The main advantage of our system is that very little training effort is required when compared to the other methods like HMM, ANN and SVM's need a lot of training.

### 4. Testing

In the recognition part, dynamic programming is has been used to find the best match for the spoken word. We tested our system with different speakers, 100 Times the test was conducted (Randomly each digit 10 times) and the results found satisfactory with the recognition rate of 93.0 %.The below figure 4 shows the some of the recognition results of matlab.

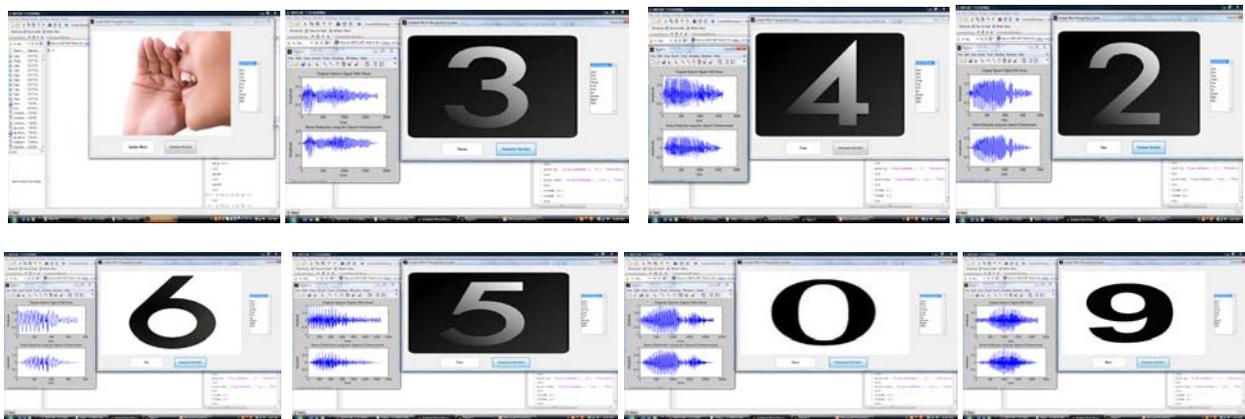


Figure 4: Main GUI interface, Result for the digits "three", "four", "two", "six", "five", "zero" and "nine"

### 4.1. Experimental Results

This section provides the experimental results in recognizing the isolated words. In the experiment, our database consists of digit names. The recognition rate using dynamic programming with speech enhancement for each word is as shown in the following table:

$$\text{Recognition Rate} = (\text{Correct Recognition} / \text{Total Number of Samples for Each word}) * 100$$

Digit	Recognition Rate (Without Enhancement)	Recognition Rate (With Enhancement)
One	60	100
Two	80	90
Three	90	90
Four	80	100
Five	80	100
Six	90	100
Seven	90	90
Eight	70	70
Nine	80	90

## 5. Conclusion

In this paper we proposed a new approach for real time isolated word speech recognition system with adaptive filter. The system is able to recognize the digits at a recognition rate of 93.00 % which is relatively high for real time recognition. From the results it is observed that LMS algorithm used for the speech enhancement which has improved the performance of the recognition rate.

## 6. References

- [1] Oliver Gauci, Carl J. Debono, Paul Micallef, "A Reproducing Kernel Hilbert Space Approach for Speech Enhancement" *ISCCSP 2008*, Malta, 12-14 March 2008
- [2] R. Solera-Urena, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Pelaez-Moreno, and F. Diaz-de-Maria, "SVMs for Automatic Speech Recognition: A Survey", *Progress in nonlinear speech processing* Pages: 190-216, 2007.
- [3] Minh N. Do, "An Automatic Speaker Recognition System", Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1999.
- [4] B. Plannerer "An Introduction to Speech Recognition", pp.27-47, March 28, 2005