Real Time Isolated Word Speech Recognition System for Human Computer Interaction

Siva Prasad Nandyala
Department of Electronics and Communication Engineering,
National Institute of Technology
Warangal, India

Dr.T.Kishore Kumar
Associate Professor,
Department of Electronics and Communication Engineering,
National Institute of Technology
Warangal, India

ABSTRACT
This paper introduces a new approach to develop a real time isolated word speech recognition system for human computer interaction. The system is a speaker dependent system. The main task is to recognize list of words in which the speaker says through the microphone. The features used are the mel-frequecy cepstarl coefficients (MFCC) which gives the good discrimination of the speech signal. The Dynamic Programming algorithm is used in the system measures the similarity between the stored template and the test template for the speech recognition which gives the optimum distance. The recognition accuracy obtained for the system is 88.0%. We made a simple list of ten words of cities names in India in which for displaying the images of the cities when the word is spoken which can be used in tourism application. This work can be used in many areas after a little modification for the specific function for example to control the robot using simple commands and also in many applications.

General Terms
Signal Processing, Speech Processing, Pattern Recognition

Keywords
Isolated Word Recognition, Feature Extraction, Mel Frequency Cepstral Coefficients, Dynamic Programming.

1. INTRODUCTION

Speech is the basic medium of communication between the people and is now playing an important role in human computer interaction. Speech recognition is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. The recognized words can be the final results, as for applications such as commands & control, data entry, and document preparation. State-of-the-art speech recognition technology still lacks robustness with respect to environmental conditions and speaking style. ASR is still far from a solved problem [1]. Speech recognition is a broad term which means it can recognize almost anybody's speech, such as a call-centre system designed to recognize many voices. Speech has the potential to be a better interface than other computing devices used such as keyboard or mouse [2]. Speech recognition is a system trained to a particular user, where it recognizes their speech based on their unique vocal sound. Speech recognition is a difficult problem, largely because of the many sources of variability associated with the signal. It depends on many factors which can be stated as the first, the acoustic realizations of phonemes, the smallest sound units of which words are composed, are highly dependent on the context in which they appear. Second, acoustic variabilities can result from changes in the environment as well as in the position and characteristics of the transducer. Third, within speaker variabilities can result from changes in the speaker's physical and emotional state, speaking rate, or voice quality. Finally, differences in sociolinguistic background, dialect, and vocal tract size and shape can contribute to across speaker variabilities [3].

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, remote vehicle control [4], home appliance control(for washing machines, ovens),in military applications like training of air traffic controllers, in artificial integellegence(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.

2. CLASSIFICATION OF THE SPEECH RECOGNITION SYSTEMS

Speech recognition systems are generally classified as discrete or continuous systems that are speaker dependent, independent, or adaptive. Discrete systems maintain a separate acoustic model for each word, combination of words, or phrases and are referred to as isolated (word) speech recognition (ISR).Connected word recognition systems are similar to isolated word system but allows separate utterances to be 'run-together' with a minimal pause between words.Continuous speech recognition (CSR) systems, on the other hand, respond to a user who pronounces words, phrases, or sentences that are in a series or specific order and are dependent on each other, as if linked together.CSR are the most difficult systems to design when compared to ISR and Connected word systems.

A speaker-dependent system needs that the user record an example of the word, sentence, or phrase prior to its being recognized by the system; that is, the user "trains" the system.

Some speaker-dependent systems require only that the user record a subset of system vocabulary to make the entire vocabulary recognizable. A speaker-independent system does not require any recording prior to system use. A speaker independent system is developed to operate for any speaker of a particular type (e.g., American English). A speaker adaptive system is developed to adapt its operation to the characteristics of new speakers [5].

Basically there are three approaches to speech recognition. They are Acoustic Phonetic Approach, Pattern Recognition Approach and Artificial Intelligence Approach.

We have used the pattern recognition approach to develop the Real Time Isolated Word Speech Recognition system. The pattern-matching approach (Itakura 1975; Rabiner 1989; Rabiner and Juang 1993) involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm [6].

3. DESIGN AND DEVELOPMENT OF THE ISOLATED WORD RECOGNITION SYSTEM

The main block diagram of the system is shown in Figure 1.

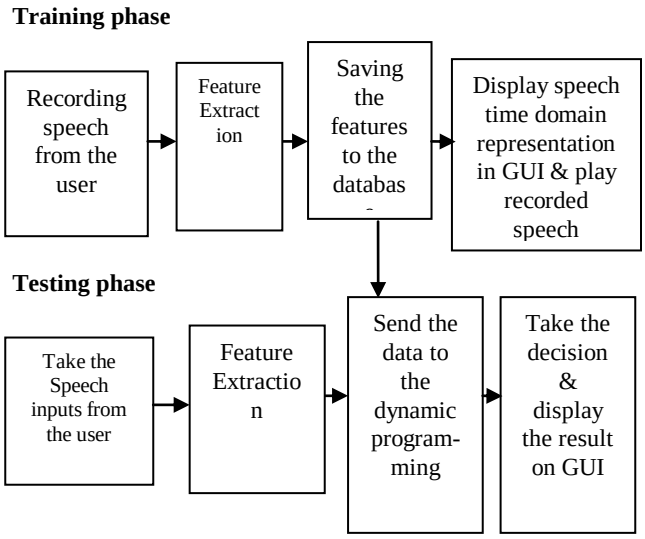


Figure 1: Real Time Isolate Word Recognition system

In our system the spoken words are instantly captured through microphone. The recognized word result image 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. One is the training part in which we record the speech of the user directly by using the laptop 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 and image is displayed on the GUI interface instantly.

The main steps in the designed system are explained as follows

3.1 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 2. When examined over a sufficiently short period of time, 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.

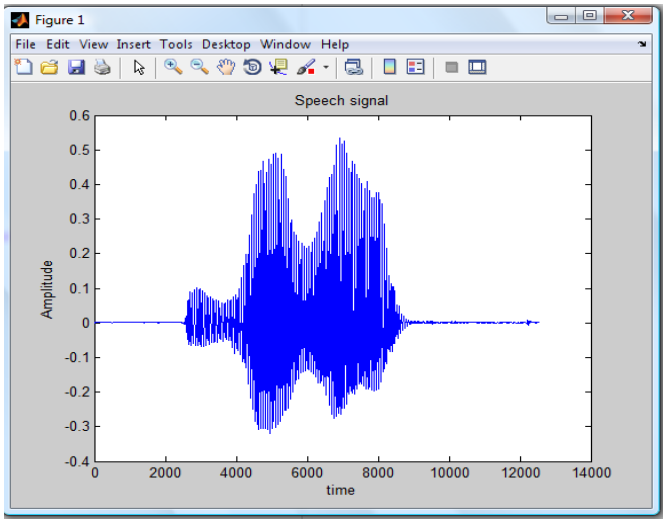


Figure 2: An example of speech signal

3.1.1 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. The following formula is used to compute the Mel's for a particular frequency:

$$Mel(f) = 2595*\log10(1+ f / 700) \quad (1)$$

A block diagram of the structure of an MFCC is given in Figure 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 is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to mentioned variations.

The main steps used for the MFCC are given as Pre-emphasizing, Frame Blocking, Windowing, Fast Fourier Transform, Mel-Frequency warping and Cepstrum

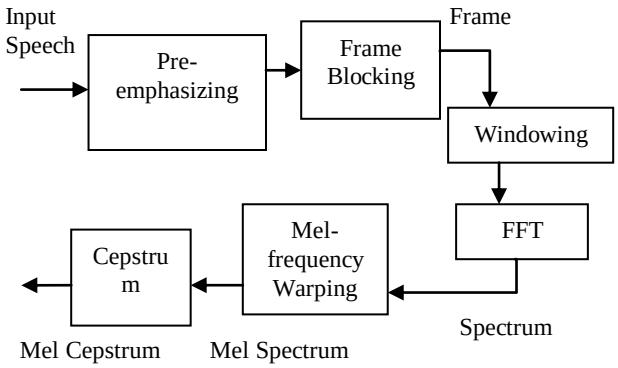
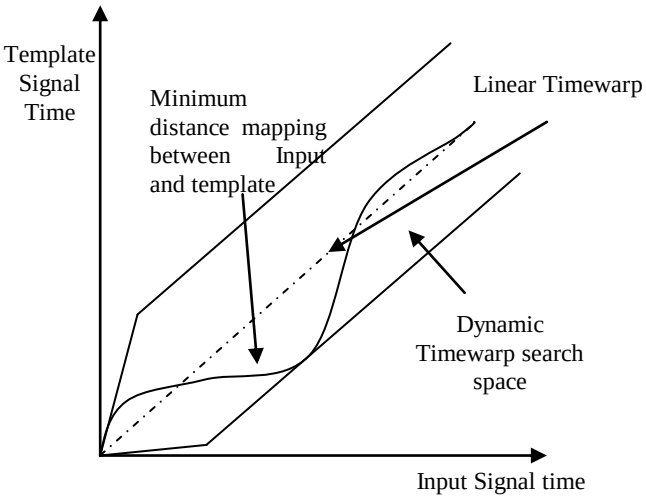


Figure 3: Block diagram of MFCC processing

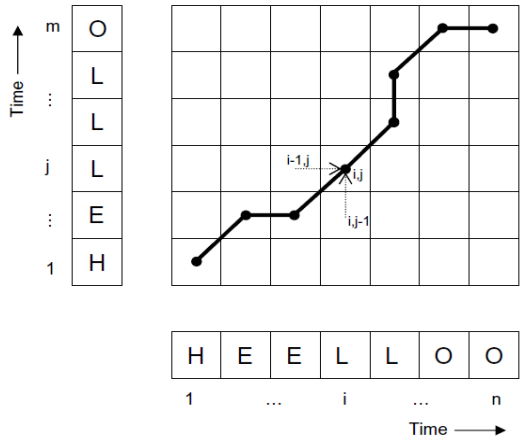
4. DYNAMIC PROGRAMMING

The Dynamic Time Warping algorithm (DTW) is a well-known algorithm in many areas. While first introduced in 60s and extensively explored in 70s by application to the speech recognition it is currently used in many areas: handwriting and online signature matching, sign language recognition and gestures recognition, data mining and time series clustering (time series databases search), computer vision and computer animation, surveillance, protein sequence alignment and chemical engineering, music and signal processing [7].

Dynamic Time Warping (DTW) is a much more robust distance measure for time series, allowing similar shapes to match even if they are out of phase in the time domain.



(a)



(b)

Figure 4: (a) Dynamic Time Warping, (b) Example word "Hello"

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" [8].

Figure 4(a) 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. DTW is an algorithm particularly suited to matching sequences with missing information, provided there

are long enough segments for matching to occur. The optimization process is performed using dynamic programming, hence the name is dynamic time warping.

In the DTW method the distance method is often used. In this method, the distance is designated to depict the greatest similarity between series by calculating the minimum distance between them, which is defined as follows

$$d(X,Y) = (\sum_{i=1}^{n} |x_i - y_i|)^{1/p} \quad (2)$$

4.1 Dynamic programming with modifications

In this method we used the DTW algorithm which has some modifications given below [9].

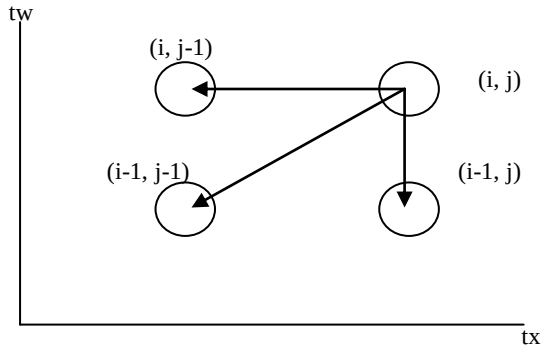


Figure5: Local path alternatives for a grid point

Let d (i, j) to be the distance d (wi, xj) between the two vectors wi and xi. Let δj(i) be the accumulated distance δ(i, j) at grid point (i, j) and δj−1(i) the accumulated distance δ(i, j − 1) at grid point (i, j − 1) in which the grid points are shown in Figure 5.

Let ψ (i, j) be the predecessor grid point (k, l) chosen during the optimization step at grid point (i, j). There are mainly four steps present in this approach which gives the best optimal path for the algorithm.

4.1.1 Initialization

Grid point (0, 0):

Ψ (0, 0) = (− 1, −1)

δj (0) = d (0, 0) (eq.12)

Initialize first column (only vertical path possible):

For i = 1 to TW − 1

{

δj (i) = d(i, 0) + δj(i − 1) (3)

ψ (i, 0) = (i − 1, 0)

}

4.1.2 Iteration

compute all columns:

for j = 1 to TX − 1

{

swap arrays δj− 1(.) and δj(.)

first point (i = 0, only horizontal path possible):

δj(0) = d(0, j) + δj− 1(0) (4)

ψ(0, j) = (0, j − 1)

compute column j:

for i = 1 to TW − 1

{

optimization step:

$$\delta_j(i) = \min \begin{cases} \delta_{j-1}(i) + d(i,j) \\ \delta_{j-1}(i-1) + 2 \cdot d(i,j) \\ \delta_j(i-1) + d(i,j) \end{cases} \quad (5)$$

Tracking of path decisions

$$\psi(i,j) = \underset{(k,l)\in \begin{cases} (i,j-1), \\ (i-1,j-1), \\ (i-1,j) \end{cases}}{\arg\min} \begin{cases} \delta_{j-1}(i) + d(i,j) \\ \delta_{j-1}(i-1) + 2 \cdot d(i,j) \\ \delta_j(i-1) + d(i,j) \end{cases} \quad (6)$$

}

}

4.1.3 Termination

D (TW − 1, TX − 1) = δj (TW − 1, TX − 1)

(7)

4.1.4 Backtracking

Initialization:

i = TW − 1, j = TX – 1

(8)

while ψ(i, j) = − 1

{

get predecessor point:

i, j = ψ(i, j)

(9)

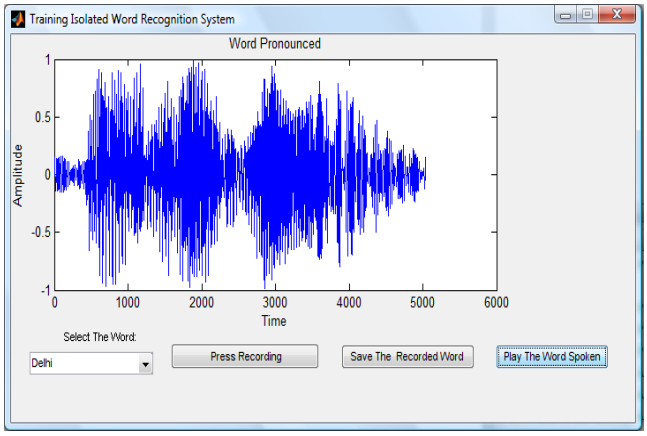
}

The algorithm above will find the optimum path by considering all path hypotheses through the matrix of grid points.

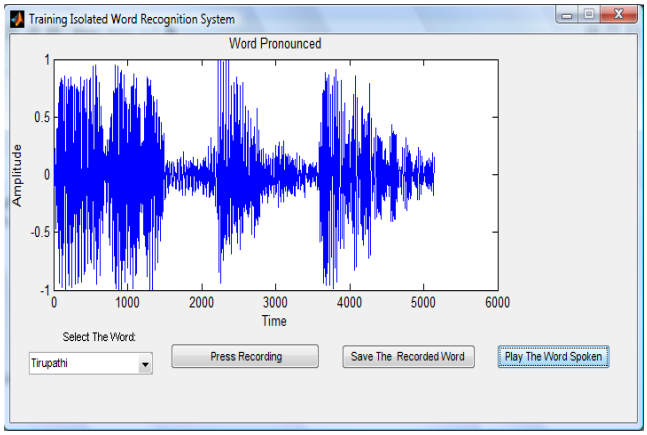
Although there are other advanced techniques in speech recognition such as the hidden Markov modeling (HMM), artificial neural network (ANN) techniques and Support Vector Machines(SVM), the DTW is widely used in the small-scale embedded-speech recognition systems such as those embedded in cell phones. The reason for this is owing to the simplicity of the hardware implementation of the DTW, which makes it suitable for many mobile devices. Additionally, the training procedure in DTW is very simple and fast, as compared with the HMM, ANN and SVM rivals.

5. TRAINING

Because of the real time characteristic of system, matlab is enabled to interface with the sound card of our laptop. The data acquisition tool box is used in this work to transfer the read voice to the database. For saving the voice, the user enters the word and the feature vectors are calculated and loaded to our database with its unique ID, that is previously set into an excel sheet, with using excel sheet for easy user editing without the aid of any programming algorithms. The training was done for the ten words of the cities names 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. The time domain representation of some of the spoken words is given in the below Figure 6.



(a)



(b)

Figure 6: (a) & (b) Time domain representation of the spoken words "Delhi" and "Tirupathi"

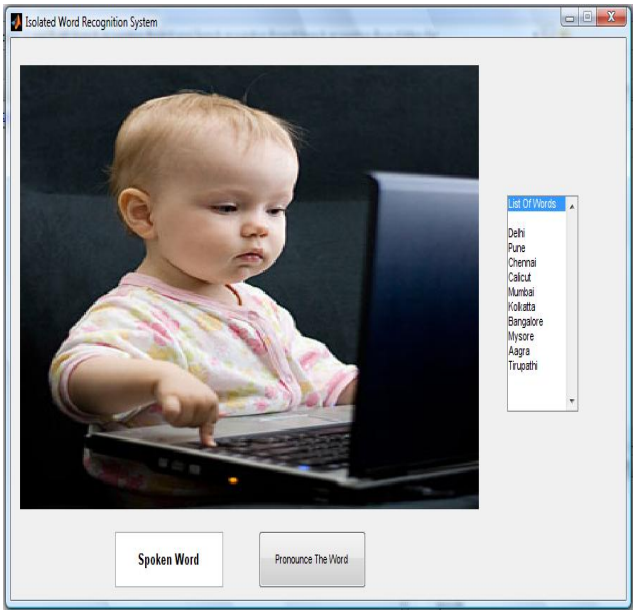
6. TESTING & EXPERIMENTAL RESULTS

In the experiment, our database consists of 10 different names of cities and the algorithm is tested for the percentage of accuracy. We did the testing of 25 utterances of each 10 words. The accuracy obtained using DTW using dynamic programming for each word is as shown in the following table:

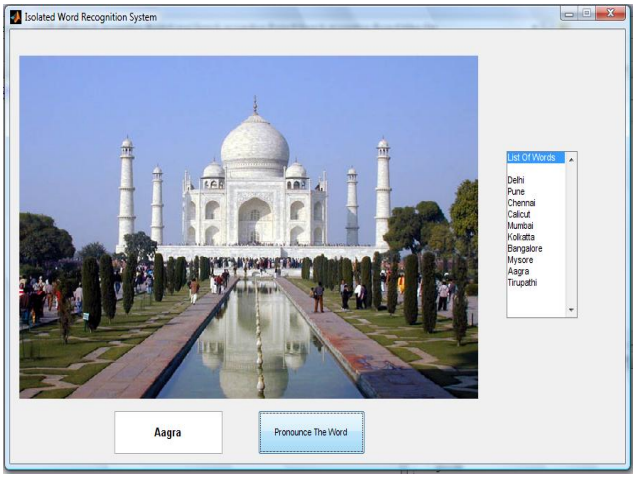
Table 1: Recognition accuracy results for DTW using dynamic programming

Word to be recognized	Accuracy
Delhi	96
Pune	92
Chennai	84
Calicut	92
Mumbai	80
Kolkata	80
Bangalore	80
Mysore	92
Aagra	92
Tirupathi	84

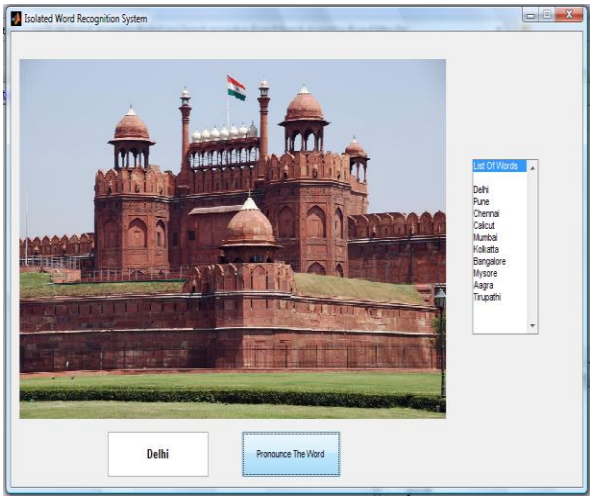
The results of the System are shown in Figure.7.


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysore
Aagra
Tirupathi
Spoken Word
Pronounce The Word

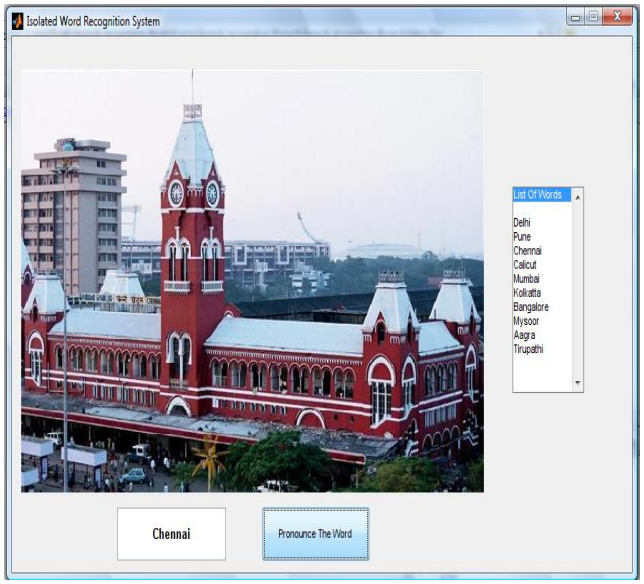
(a)


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysore
Aagra
Tirupathi
Aagra
Pronounce The Word

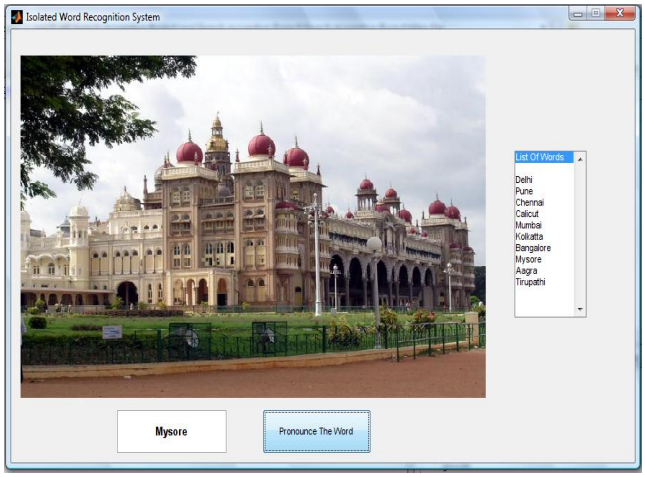
(b)


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysore
Aagra
Tirupathi
Delhi
Pronounce The Word

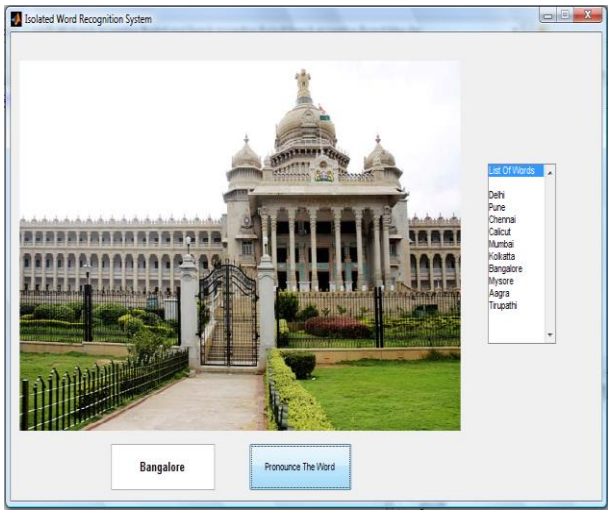
(c)


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysoor
Aagra
Tirupathi
Chennai
Pronounce The Word

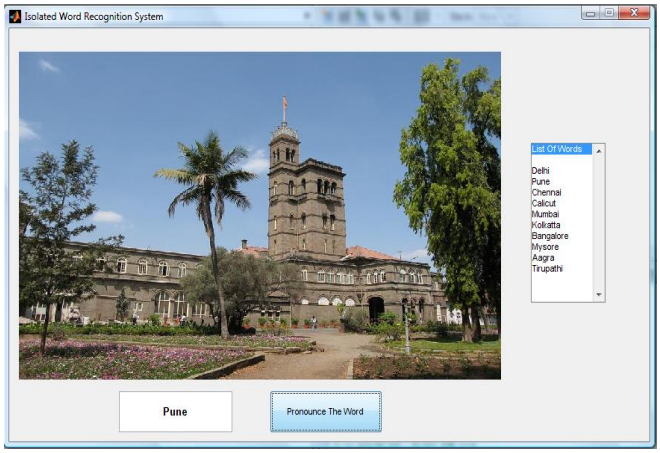
(d)


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysore
Aagra
Tirupathi
Mysore
Pronounce The Word

(e)


Isolated Word Recognition System
List Of Words
Delhi
Pune
Chennai
Calicut
Mumbai
Kolkatta
Bangalore
Mysore
Aagra
Tirupathi
Bangalore
Pronounce The Word

(f)



(h)

Figure 7:(a) Main GUI of the system, (b to h) :Results of the pronounced words "Aagra","Delhi","Chennai","Mysore" "Bangalore"and "Pune"

7. CONCLUSIONS

In this paper we proposed a new approach for real time isolated word speech recognition system for human computer interaction. The system is able to recognize the words at a recognition accuracy of 88.0 % which is relatively high for real time recognition. For the future work, adaptive filtering techniques can be used to remove the noise from speech signal so that accuracy can be further improved.

8. REFERENCES

[1] S.A.R. Al-Haddad, S.A. Samad, A. Hussain, K.A. Ishak and A.O.A. Noor "Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering" American Journal of Applied Sciences 6 (2): 290-295, 2009

[2] Jurafsky, D., and Martin, J. H. "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition", 2nd Edition. (2007).

[3] Ron Cole, Joseph Mariani Hans, UszkoreitGiovanni, Batista Varile, Annie Zaenen Antonio Zampolli, Victor Zue " Survey of the state of the art in human language technology"Cambridge University Press and Giardini 1997

[4] Shi-Huang Chen, YuRu Wei,"A Study on Speech-Controlled Real-TimeRemote Vehicle On-Board Diagnostic System"Proceeding of the International multiconference on Engineers and Computer Scientists 2010 vol I,IMECS 2010,March 17-19,2010,Hong Kong.

[5] Fadhilah Rosdi, Raja N. Ainon "Isolated Malay Speech Recognition Using Hidden Markov Models", Proceedings of the International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 2008 May 13-15.

[6] M.A.Anusuya, S.K.Katti,"Speech Recognition by Machine: A Review",(IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009

[7] Pavel Senin, "Dynamic Time Warping Algorithm Review ", Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA.

[8] R. Solera-Urena, J. Padrell-Sendra, D. Martın-Iglesias, A. Gallardo-Antolın, C. Pelaez-Moreno, and F. Dıaz-de-Marıa " SVMs for Automatic Speech Recognition: A Survey " Progress in nonlinear speech processing Pages: 190-216, 2007.

[9] B. Plannerer "An Introduction to Speech Recognition "March28,2005.