International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No.9, December 2012 – www.ijais.org

A Natural Human-Machine Interaction via an Efficient Speech Recognition System

Shachi Sharma
Research Student
Department of CS
AIET, RTU

Krishna Kumar Sharma
Assistant professor
Deptt. of CS & informatics
University of kota ,Kota

Himanshu Arora
Associate Professor
Department of CSE
ACERC, RTU

ABSTRACT

This paper is motivated from non-technical users' problems in using technical interfaces of computer. In village areas, farmers face problems in using conventional ways to use computers, so in order to design a natural interaction way of human with computer, an efficient speech recognition system should be developed.

For this we designed a system application. User has to speak commands and the system performs according to commands. This is all tested in the mobile environment and with varying users. And from the results, conclusion has been derived that the hybrid feature set outperformed in the noisy environment as compared to individual feature set with their dynamic features. And the result was approximately 5% higher. When DHMM is implemented in the system, results increased.

General Terms

Speech Recognition, Pattern Recognition..

Keywords

Speech Recognition system, DHMM, hybrid feature set;

1. INTRODUCTION

Speech is most natural way of interaction for human. If it is being used by users for machine interaction (e.g., for interaction with computer, robot, mobile phone or various other technical gadgets) then human-machine interaction will become more interactive and easy [1]. Thus a robust speech recognition system has broad applications in the human-machine and human-computer interaction.

In today's world human-machine interaction has increased its scope in the social life and in almost every field [1], but still some groups of society which are illiterate and nontechnical find technical gadgets and devices less convenient and friendly to work with. Even some people find difficulty in using mobile phones also. So, in order to enhance this interaction with such machines there should be a natural and friendly interaction way, so that human can handle the machines efficiently.

Thus speech is added as a new natural way for interaction with these techie devices, as speech is the widely used interaction method for human [2]. When speech is the way of interaction, illiterate and nontechnical people can also easily command the computer and other such machines.

2. GENERAL STRUCTURE OF A SPEECH RECOGNITION SYSTEM

In order to design a natural interaction way of human with computer, an efficient speech recognition system should be developed. For this we designed a system application that can work in noisy environment and with changing users. Data from 35 different users has been collected, and each user speaks 11 times each word. We used a highly efficient head mounted Sennheiser microphone to collect data.

The design of the system has majorly two phases: 1) Training, and 2) Testing.

The process of extraction of features relevant for classification is common in both phases. During the training phase, the parameters of the classification model are estimated using a large number of class examples (Training Data). During the testing or recognition phase, the feature of test pattern (test speech data) is matched with the trained model of each and every class. The test pattern is declared to belong to that whose model matches the test pattern best.

The Training process involves several steps (i.e. the study implements the isolated word recognizer in six steps) as discussed below.

The first step performs the collection of speech samples to train system with possible all possible conditions. In the second step we preprocess data in order to make it ready to extract features. In the third step we detect end points of the speech samples. In the fourth step we extracted MFCC, dynamic features of MFCC, HFCC, and their dynamic features. A combined feature vector is also proposed of MFCC, HFCC, and their dynamic features, named as, INTEGRATED STATIC AND DYNAMIC CEPSTRAL COEFFICIENTS FEATURE VECTOR. In the fifth step we vector quantized data to remove data redundancy. And in the last step Discrete Hidden Markov Model is implemented to enhance recognition results.

As shown in the figure 2, speech recognition system is designed and features are extracted. This design is done in to two steps: training and testing. In the training above steps are used: Pre emphasis, end-point detection, frames blocking, windowing, FFT, Cepstral features extraction, VQ, and DHMM.

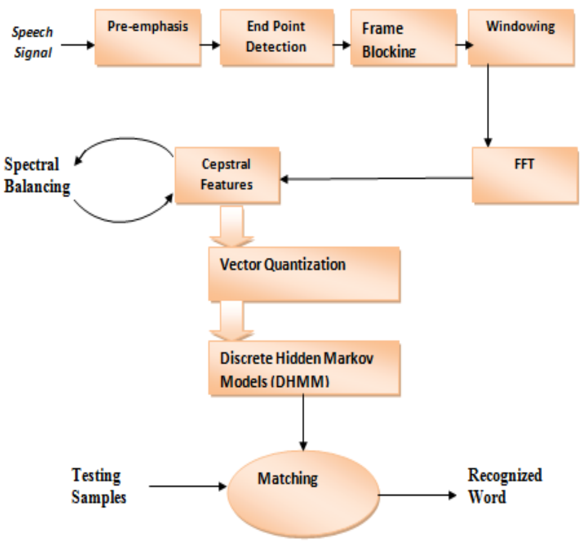


Fig. 2.1: Block diagram of a Speech Recognition System

3. PROPOSED METHOD

Human auditory systems can intelligently recognize spoken word and can perform actions accordingly it. To resemble human auditory system mel frequency cepstral coefficient (MFCC) and human factor cepstral coefficient (HFCC) parameters are used. Obtained parameters are static and to make them robust for varying features, dynamic features of MFCC and HFCC parameters are used. MFCC and HFCC with their dynamic parameters are immune to noises in the signal.

For the testing purpose an efficient head mounted Sennheiser microphone is used in varying acoustic ambience.Hidden Markov Model (HMM) [3] is used to design models of each word, HMM enhance the recognition rate of the spoken words. After successfully extracting features from MFCC coefficient [4] and its dynamic coefficient are used as features. Similarly HFCC coefficient and its dynamic coefficient are also used as feature vector. As shown in the Figure 3.1, the key difference between these two parameters is in the design of the filter bank that is described in the ASR system. To increase system efficiency we proposed possible combination of the features, which include both filter's characteristics, in the following steps:

STEP 1: we obtain MFCC and HFCC features and their dynamic parameters also.

STEP 2: according to proposed method, we make combined feature vector of both parameter and vector quantized the parameter, named as, INTEGRATED STATIC AND DYNAMIC CEPSTRSL COEFFICIENTS FEATURE VECTOR.

STEP 3: and from generated codebook after vector quantization, we develop HMM model of the each word.

STEP 4: system is trained from step 3, now to use this system we find out testing samples maximum likelihood from the HMM models using Viterbi algorithm

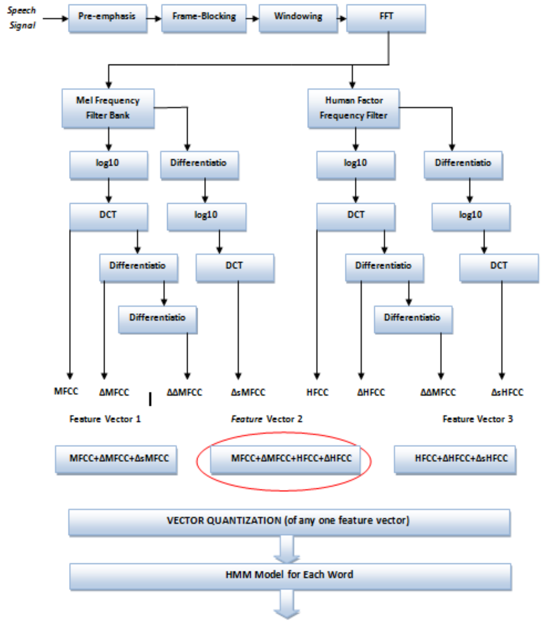


Fig. 3.1: Block Diagram from the feature extraction to design HMM model.

The main aim of this system is to enhance the recognition result in the mobile environment, in varying acoustic conditions, and to work efficiently in odd situation for the small vocabulary consists of fifteen words: {Up, Down, Forward, Backward, Left, Right, Start, Stop, Hold, Krishna, Save, One, Two, Hello, Move}. And this speech database is collected from 35 different speakers of varying age and sex, and each speaker speaks 11 times each words. These words are collected by an efficient head mounted Sennheiser microphone by Sonarca sound recorder free software at 16 KHz and 16 bits. Discrete Hidden Markov Model technique is used to design recognizer. DHMM is preferred because speech samples are of short length. We used methodology that is proposed by Sorensen [5] (as shown in the block diagram 3.2). As previously described we used following steps to design this system.

3.1 Word Recording

Sound is recorded through a head mounted Sennheiser microphone. This is done to make communication more natural and free talk with the machine. But, we put certain constrains on the range to communicate with robot. We collect 375 speech samples of each word spoken by 35 speakers 11 time (375=35*11) each word.

3.2 A/D Converter-

After recording we get the analog signal. This analog signal is hard to process, so we need to convert it in to digital form. Signal is digitized at 16 KHz at 16 bit quantization scheme.

3.3 Pre-emphasis-

This digitized signal now passed through pre-emphasis steps and get in to the windowed form. It consists of framing of signal, windowing of signal, and FFT operations.

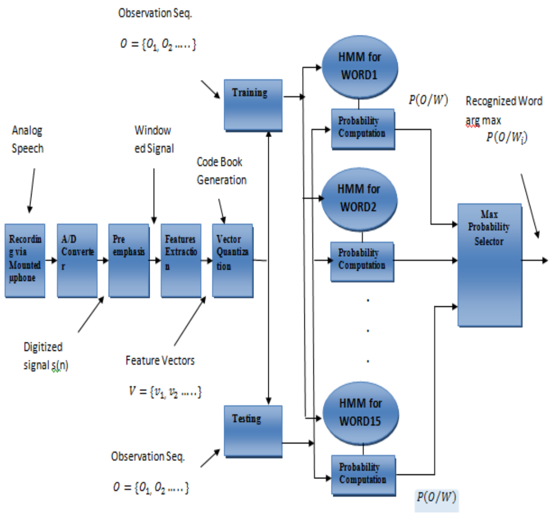


Fig. 3.2: System Block Diagram DHMM Recognizer ([5])

3.4 Feature extraction

Cepstral coefficients are obtained from speech frames. Cepstral coefficients are obtained by passing through MFCC filter banks and HFCC filter banks. A filter bank of 24 triangular filters was used for the mel-scale conversion in the MFCC computation and for HFCC computation also. First 12 coefficients are used to make feature vector of each frame. Delta features of the MFCC and HFCC are also measured to get dynamic features.

3.5 Vector quantization

Four fifth of the database was selected randomly to form training set, while reminder will be used as the testing samples. Feature vectors are generated from the samples with the K-means algorithms. Size of 128 symbols a codebook is generated from the samples. Note that a stochastically generated codebook was also an effort to decrease the computational load by eliminating the K-means algorithm training phase.

3.6 Discrete hidden markov model

A discrete HMM is used with eight states for each word in as shown in the Figure 3.3. The training of the samples is preceded by vector quantization and quantized symbols are used to train each model and Baum Welch algorithm is used to find out parameters that increase the likelihood of the training set.

3.7 Hidden markov model recognition step

Recall the all previous steps Word recording, Pre emphasis, Feature extraction, Vector Quantization, and DHMM for the testing samples. After calculating the parameters we need to find out a model which maximizes the posteriori probability, i.e. $\arg max_i\ [P(O/W_i)]\ for\ i = 1,2, \ldots \ldots .15$.

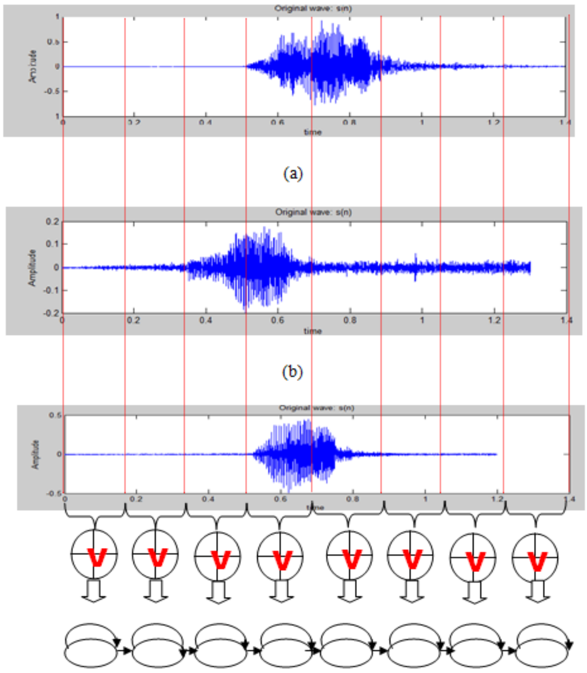


Figure 3.3: Representation of Word "right" and discrete HMM observation classifier [6].

4. DHMM RECOGNITION RESULT

Now, sixty four tests are performed for each word to evaluate overall performance of the system with the DHMM based recognizer for two times for different-different data sets. We test it for varying parameter sets obtained from HFCC and MFCC filter banks. Initially, we tested it for individual parameter sets and then we find out that we need to test system on the hybrid parameters set that contain static features, dynamic logarithmic features, and logarithmic dynamic features. An Isolated Word Speech Recognition System designed in Matlab 7.10.0.499 (R2010a) 64-bit is shown below in fig 4.1.

Combinations that are used to classify the spoken words are shown below:

Sample Frequency:	16 KHz
Frame length:	400 samples.
Shift Interval:	160 samples
Mel-scale filters Bank:	24 filters.
HFCC filter Bank:	24 filters.

Speech Features:

1) $MFCC\ + \Delta_s MFCC + \ \Delta MFCC$

2) $HFCC\ + \Delta_s HFCC + \ \Delta HFCC$

3) $MFCC\ + HFCC + \Delta MFCC + \ \Delta HFCC$

Codebook size:	128 symbols.
Size of HMM:	8 states.

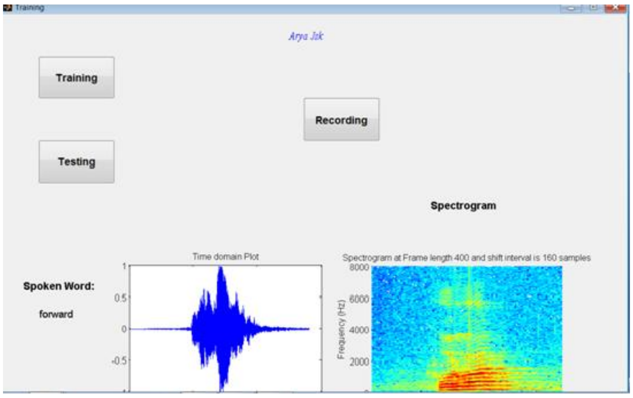


Fig. 4.1: Isolated Word Speech Recognition System designed in Matlab 7.10.0.499 (R2010a) 64-bit.

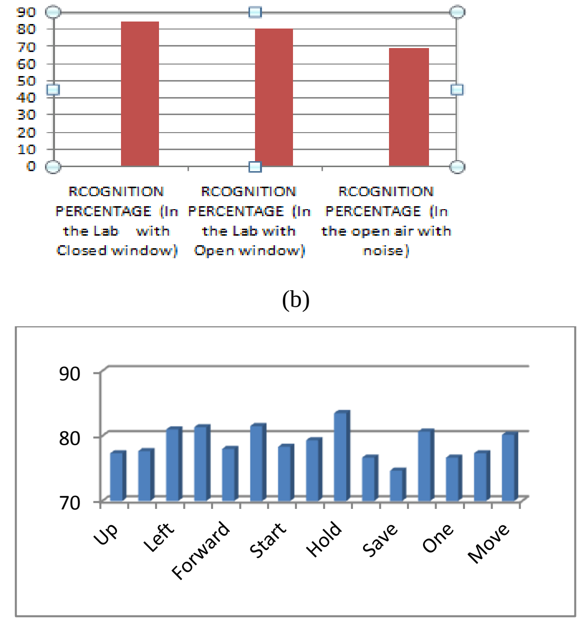
4.1 Speech feature $MFCC + \Delta_s MFCC + \Delta MFCC$ based result

Results obtained from $MFCC + \Delta_s MFCC + \Delta MFCC$ features sets are presented in the following table. Each word is tested for the 64 times in varying conditions (i.e, in lab with windows closed, in lab with windows kept open and in open air with noise). Results obtained are discussed for the training data as given in the table 4.2(a) with the performance graphs also (shown in figure 4.2(b), 4.2(c).

Results obtained from the data set:

$MFCC + \Delta_s MFCC + \Delta MFCC$ feature vector					
			Successful word recognition rate (PERCENTAGE)		
S. No.	Spoken Words (to recognize	No. of times word spoken by each person	In the lab with windo w closed	In the lab with window open	In open air noise
1	Up	64	85	77	72
2	Down	64	87.5	80	70
3	Left	64	85	83	67
4	Right	64	85	82	72
5	Forward	64	90	81	75
6	Backward	64	85	80	73
7	Start	64	83	81	72
8	Stop	64	85	82	67
9	Hold	64	83	80	70
10	Krishna	64	85	79	65
11	Save	64	84	80	69
12	Hello	64	83	80	65
13	One	64	85	80	70
14	Two	64	87.5	78	65
15	Move	64	87.5	87	68

(a)



(b)

(c)

Fig. 4.2(a): Percentage (%) Recognition Results from the $MFCC + \Delta_s MFCC + \Delta MFCC$ feature vector.

Fig. 4.2(b): average recognition in Percentage (%) of the three conditions.

Fig. 4.2(c): word wise recognition percentage (%).

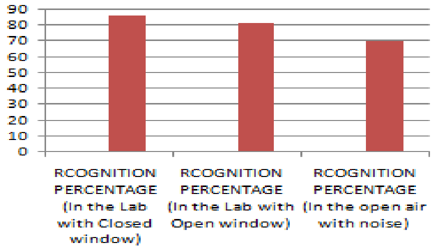
From the data set we can conclude that in the closed window environment, average recognition rate vary is 84.80%. 'Backward' word has highest recognition rate and 'Krishna' has lowest recognition rate. In the open window environment it declines to approximately 4% and average recognition rate found is 80.67% and in the open air noisy environment the performances degrade to approx. 10%. It gets average recognition of 69.33%. In order to increase performance HFCC and its dynamic features are used for recognition.

4.2. Speech feature $HFCC + \Delta_s HFCC + \Delta HFCC$ based result

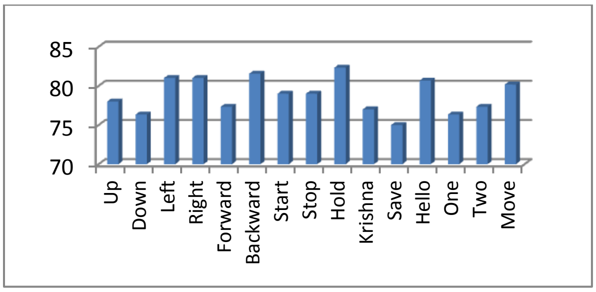
Results obtained from $HFCC + \Delta_s HFCC + \Delta HFCC$ features sets are presented in the following table. Each word is tested for the 64 times in varying conditions. (i.e, in lab with windows closed, in lab with windows kept open and in open air with noise). Results obtained are discussed for the training data set, as given in the table 4.3(a) with the performance graphs (as shown in fig. 4.3(b) and 4.3(c)).

$HFCC\ +\Delta_s HFCC +\ \Delta HFCC$ Feature vector.					
			Successful word recognition rate (PERCENTAGE)		
S. No.	Spoken Words (for recognition)	No. of times word spoken by person	In the lab with window closed	In lab with window open	In open air noise
1	Up	64	86	81	67
2	Down	64	83	77	69
3	Left	64	91	84	68
4	Right	64	88	85	70
5	Forward	64	83	77	72
6	Backward	64	87.5	78	79
7	Start	64	83	83	71
8	Stop	64	85	82	70
9	Hold	64	87	85	75
10	Krishna	64	83	80	68
11	Save	64	83	77	65
12	One	64	85	83	74
13	Two	64	84	80	65
14	Hello	64	86	84	62
15	Move	64	87.5	81	72

(a)



(b)



(c)

Fig: 4.3(a): Percentage (%) Recognition Results from the $HFCC\ +\Delta_s HFCC + \Delta HFCC$ feature vector.

Fig. 4.3(b): average recognition in Percentage (%) of the three conditions.

Fig. 4.3(c): word wise recognition percentage (%).

From the data set we can conclude that in the closed window Environment average recognition rate obtained is 85.47%. 'Hold' word has highest recognition rate and 'Save' has lowest recognition rate. In the open window environment it decline to approximately 4% and average recognition rate found is 81.13%. And in the open air noisy environment the performance degrade to approx. 11%. It gets avg. recognition rate of 69.8%.

4.2 Speech feature $MFCC + HFCC + \Delta MFCC + \Delta HFCC$ based result

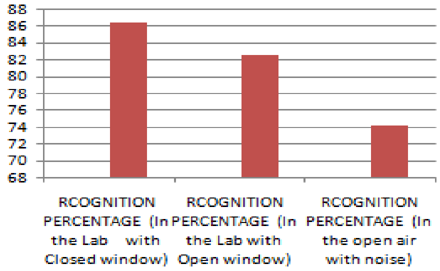
In order to increase performance MFCC and HFCC and their dynamic features are used for recognition. This INTEGRATED STATIC AND DYNAMIC CEPSTRAL COEFFICIENTS FEATURE VECTOR results in visible performance improvement (fig.4.4 (a)), and performance graphs (fig. 4.4 (b) and (c)).

Results obtained from the data set:

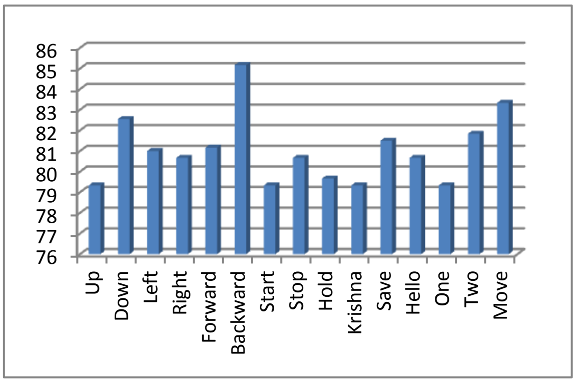
$MFCC + HFCC + \Delta MFCC + \Delta HFCC$ Feature vector.					
			Successful word recognition rate (PERCENTAGE)		
S. No.	Spoken Words (to be recognize)	No. of times word spoken by person	In the lab with windows closed	In lab with window open	In open air noise
1	Up	64	85	81	72
2	Down	64	87.5	85	75
3	Left	64	85	83	75
4	Right	64	85	83	74
5	Forward	64	87.5	83	73
6	Backward	64	90	87.5	78

7	Start	64	83	82	73
8	Stop	64	86	83	73
9	Hold	64	84	80	75
10	Krishna	64	85	80	73
11	Save	64	87.5	82	75
12	Hello	64	85	82	75
13	One	64	85	80	73
14	Two	64	87.5	83	75
15	Move	64	93	83	74

(a)



(b)



(C)

Fig: 4.4 (a): Percentage (%) Recognition Results from the $MFCC + HFCC + \Delta HFCC + \Delta MFCC$ Feature vector.
Fig. 4.4 (b): average recognition in Percentage (%) of the three conditions.

Fig. 4.4 (c): word wise recognition percentage (%).

From the above results (as shown in figure 4.4(a), (b) and (c) we can say that recognition rate increases considerably when we use hybrid data features. 'Backward' word has highest recognition rate and 'Up' has lowest recognition rate. In the closed window environment average recognition rate comes out to be 86.37%, in the open window environment it is 82.53% and in the noisy environment rate is 74.13%.

More precisely, we can state that the combined feature set outperformed in recognition percentage as compared to individual feature set (with their dynamic features) with the accuracy of 86.375% in the lab environment with windows closed, 82.53% in lab environment with windows kept open, and 74.13% in open air noisy environment.

From the graph shown below in fig.4.5 we can conclude that in the noisy environment, combined feature set (i.e. $MFCC + HFCC + \Delta MFCC + \Delta HFCC$) results performed 4.33% and 4.80% higher as compared to single filtered feature set (i.e. $MFCC + \Delta_s MFCC + \Delta MFCC$ and $HFCC + \Delta_s HFCC + \Delta HFCC$ respectively). In lab environment when windows are kept closed, it $(MFCC + HFCC + \Delta MFCC + \Delta HFCC)$ performed 0.835% and 1.3% higher than single filtered feature sets (i.e. $MFCC + \Delta_s MFCC + \Delta MFCC$ and $HFCC + \Delta_s HFCC + \Delta HFCC$ respectively). And, in lab environment with windows kept open, the combined feature the combined feature set (i.e. $MFCC\ HFCC + \Delta MFCC + \Delta HFCC$) performed 1.285% and 1.78% higher than single filtered feature sets (i.e. $MFCC + \Delta_s MFCC + \Delta MFCC$ and $HFCC + \Delta_s HFCC + \Delta HFCC$ respectively).

The comparison of all the three speech feature vectors on the basis of their word recognition rate is shown in figure 4.6.

5. CONCLUSION

This paper sets its goal in the starting to design an isolated word speech recognition system. In order to design this system we collect speech samples from 35 different persons and each person speaks each word 11 times with the head mounted Sennheiser microphone. And this is successfully executed in the Matlab programming. We obtained high recognition rate with our proposed model, INTEGRATED STATIC AND DYNAMIC CEPSTRAL COEFFICIENTS FEATURE VECTOR by using MFCC, HFCC, and their dynamic coefficients as compared to individual feature sets. This model for the speech recognition was tested in all odd situations as well as in even situation like noisy, varying speakers, and system independent.

6. REFERENCES

[1] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A survey of socially interactive robots, robotics and Autonomous Systems", ISBN *978-3-902613-13-4*, Elsevier Publications Ltd., vol. 4*2*, pp. 143-166, 2003.

[2] M. A. Goodrich, A. Schultz, "Human Robot Interaction: A Survery," Foundations and Trends in Human-Computer Interaction, *ISBN 978-1-60198-092-2*, Goodrich's Publications Ltd., vol. 1, pp. 203-275, 2007.

[3] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, Feb1989.

[4] N. Zheng, Xia Li, Houwei Cao, Tan Lee, P. C. Ching, "Deriving MFCC parameters from the dynamic spectrum for robust speech recognition", *ISCLP'08. 6th International Symposium on Chinese Spoken Language Processing*, 2008.

[5] Sorensen and M. Swanholm, Speech coding and recognition course notes, [http://www.itu.dk./courses/TKG/E2002], last accessed February 15, 2006.

[6] A.A.M. Abushariah, T. S. Gunawan, O.O. Khalifa, "English Digits Speech Recognition System Based on Hidden Markov Models", *International Conference on Computer and Communication Engineering (ICCCE 2010)*, May 2010.

[7] K. K. Lavania , S. Sharma, K. K. Sharma, "Reviewing Human-Machine Interaction through Speech Recognition approaches and Analyzing an approach for Designing an Efficient System", *Proc. of Int. Journal of Computer Applications*, January 2012. Vol 38, No. 3, pp. 466-677.

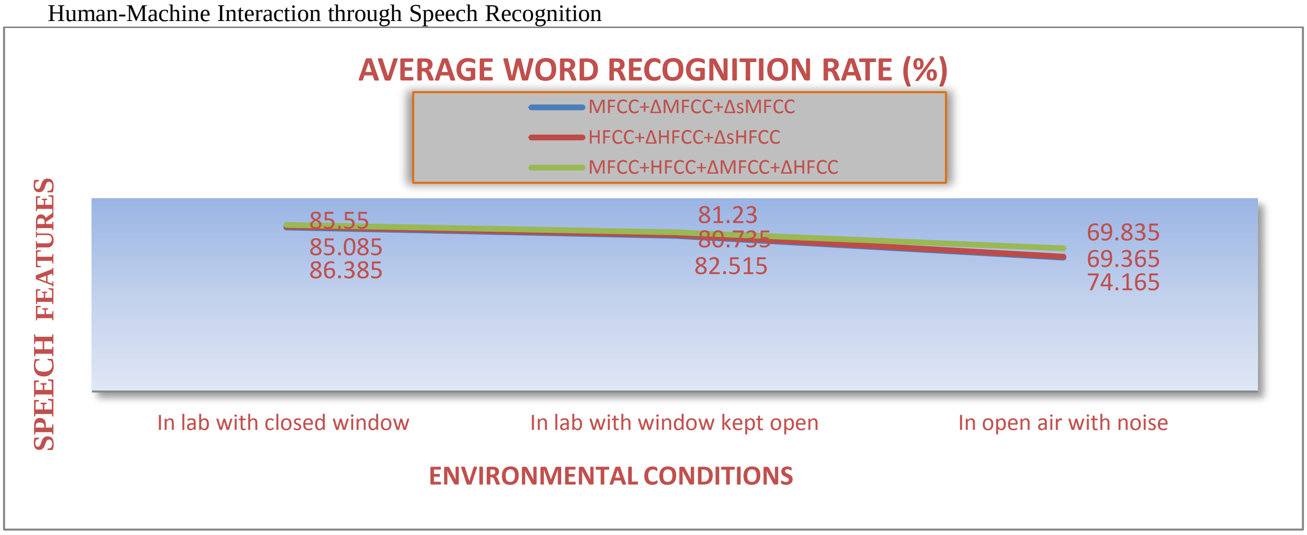
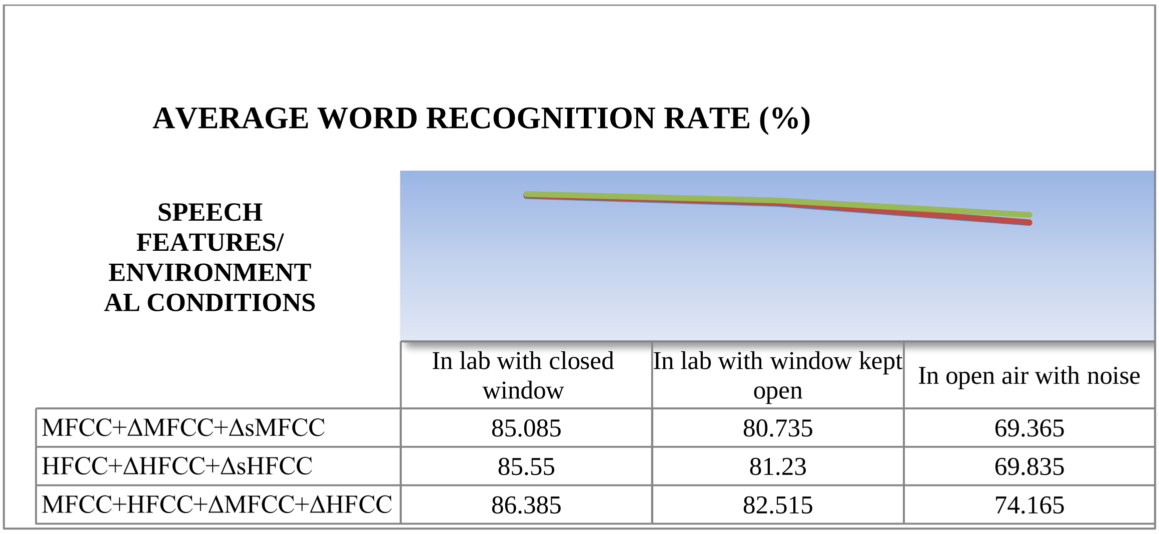


Fig 4.5: Average word recognition rate of three different speech features in different environmental conditions.



SPEECH FEATURES/ ENVIRONMENTAL CONDITIONS	In lab with closed window	In lab with window kept open	In open air with noise
MFCC+ΔMFCC+ΔsMFCC	85.085	80.735	69.365
HFCC+ΔHFCC+ΔsHFCC	85.55	81.23	69.835
MFCC+HFCC+ΔMFCC+ΔHFCC	86.385	82.515	74.165

Fig. 4.6: Word recognition result for three different speech feature vectors.