Reviewing Human-Machine Interaction through Speech Recognition approaches and Analyzing an approach for Designing an Efficient System

Krishan Kant Lavania
Associate Professor
Department of CS
AIET, RTU

Shachi Sharma
Research Student
Department of CS
AIET, RTU

Krishna Kumar Sharma
Assistant Professor
Department of CSE
Central University of Rajasthan
Kishangarh, Ajmer

ABSTRACT
Speech is most natural way of interaction for human. It has broad applications in the human-machine and human-computer interaction. This paper reviews the literature and the technological aspects of human-machine interaction through various speech recognition approaches. It also discusses the various techniques used in each step of a speech recognition process and attempts to analyze an approach for designing an efficient system for speech recognition. It also discusses that how this system works and its application in various areas.

Keywords
Speech recognition (SR);human-machine interaction;

1. INTRODUCTION
Speech interaction makes more interactive and easy interaction of human-machine interaction. Now-a-days it is used in application, but there is requirement of improvement in the recognition efficiency.
Some groups of society which are illiterate and nontechnical find technical gadgets, machines and computers less convenient and friendly to work with. So, in order to enhance this interaction with such machines and devices, speech interface is added as a new natural way for interaction, since most people find machines or computers which can speak and recognize speech more simple and easy to work with than the ones which can be operated only through some conventional mediums. Generally, Machine recognition of spoken words is carried out by matching the given speech signal (digitalized speech sample) against the sequence of words which best match the given speech sample [1].This paper presents different speech feature extraction techniques and their decision based recognition through artificial intelligence techniques as well as statistics techniques. And we present our comparatively results for these features.

2. GENERAL STRUCTURE OF A SPEECH RECOGNITION SYSTEM
In this system in order to recognize a voice the system is trained [3] such that it can recognize a person's voice. This is done by asking each person to speak out a word or any kind of utterance in the microphone.
After this the digitalization of the speech signal is followed by some signal processing. This creates a template for the speech pattern which is then kept saved in memory.
In order to recognize the speaker's voice a comparison is done by the system between the utterance and the template stored respectively for that utterance in the memory.

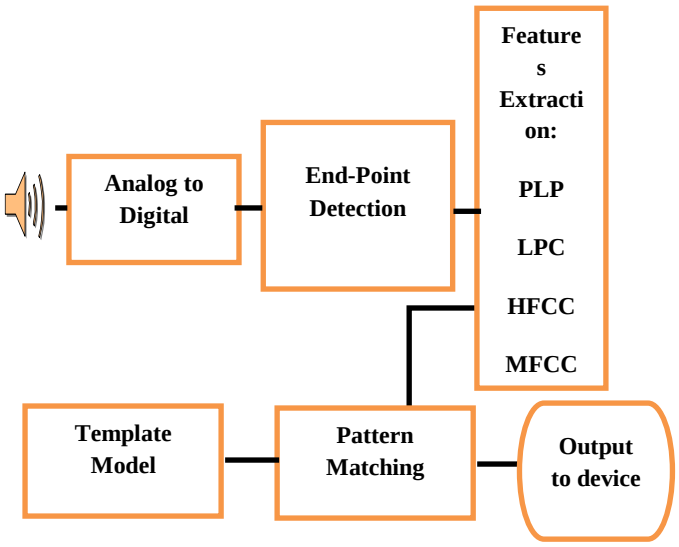

Fig. 1: Block diagram of the voice recognition system

3. SPEECH RECOGNITION APPROACHES
Basically speech recognition can be categorized under three methods or approaches [5], which are:

a) The acoustic phonetic approach
b) The pattern recognition method
c) The artificial intelligence technique

3.1 Acoustic Phonetic Method
Acoustic phonetic method is designed on the theory of acoustic phonetics that require distinctive and finite phonetic units in spoken language and that phonetic units are featured by a set of properties that are available in the signal, or its spectrum, over time.
Prime features of acoustic-phonetic approach are: Formants, Pitch, and Voiced/unvoiced Energy Nasality, Frication etc. Problems associated with the acoustic phonetics approach are requirement of extensive knowledge of acoustic properties; Choice of features is ad hoc; Not optimal classifier.

3.2 Pattern Recognition Method
Speech recognition is one in which the speech pattern are required directly without explicit feature determination and segmentation. Most pattern recognition methods have two steps-namely, training of data, and recognition of pattern via

pattern comparison. Data can be speech samples, image files, etc.

In pattern recognition method, features will be output of the filter bank, Discrete Fourier Transform (DFT), and linear predictive coding. Problems associated with the pattern recognition approach are: System's performance is directly dependent over the training data provided. Reference data are sensitive to the environment. Computational load for pattern trained and classification proportional to number of patterns being trained.

3.3 Artificial Intelligence (AI) Method

Sources of knowledge are: Acoustic knowledge; Lexical knowledge; Syntactic knowledge; Semantic knowledge; Pragmatic knowledge. In AI method, there are different techniques which can be brought into use to solve the problem as given below:

• Single/Multilayer perceptrons

• Hopfield or recurrent networks

• Kohonen or self-organizing network

Advantages associated with artificial intelligence method are: Parallel computation is possible; Knowledge can acquire from knowledge sources; Fault tolerant.

4. FEATURE EXTRACTION TECHNIQUES

This technique is basically used for analyzing a given speech signal.

It can be categorized mainly as: a) temporal analysis technique, and b) Spectral analysis techniques.

The basic difference between both the techniques is that, that in temporal analysis technique, analysis is carried out by the speech waveform only, whereas for spectral analysis, analysis is performed by using the spectral representation of the speech signal.

Fig.2: General feature extraction process

4.1 Spectral Analysis Techniques

Spectral analysis techniques are mainly required to recognize a time domain signal when it is in its frequency domain representation. This is basically done by performing a fourier transform over it. Few prominently used techniques are discussed below [4]:

4.1.1 Cepstral Analysis

This is an important analysis technique by which excitation and vocal tract can be set apart, the speech signal is given as

$$s(n) = g(n) \times v(n) \quad (1)$$

Where $v(n)$., is the vocal tract impulse response and $g(n)$ is the excitation signal

Also the frequency domain is represented as

$$S(f) = G(f).V(f) \quad (2)$$

Logarithmically,

$$\log(S(f)) = \log(G(f)) + \log(V(f)) \quad (3)$$

Thus we see that by excitation and vocal tract could be set apart from each other and can also be superimposed if logarithm is taken in the given frequency domain.

4.1.2 Mel Cepstrum Analysis

Mel Cepstrum is an analysis technique which consists of a cepstrum along a frequency axis. It also consists a*mel* scale.

Mel-frequency cepstrum provides a better and closer response to the human auditory system than an ordinary cepstrum because the frequency bands [6] in the Mel-frequency cepstrum are placed logarithmically over the *mel* scale. This helps in providing a closer response of human auditory system than the linearly spaced frequency bands which are derived from FFT (Fast Fourier Transform) and DCT [7] (Discrete Cosine Transform). Thus a mel frequency cepstrum results in more accurate processing of data. But MFCCs still has one limitation that it does not consist an outer ear model due to which it cannot represent perceived loudness precisely.

The block for computing MFC coefficients is given in Fig.3:

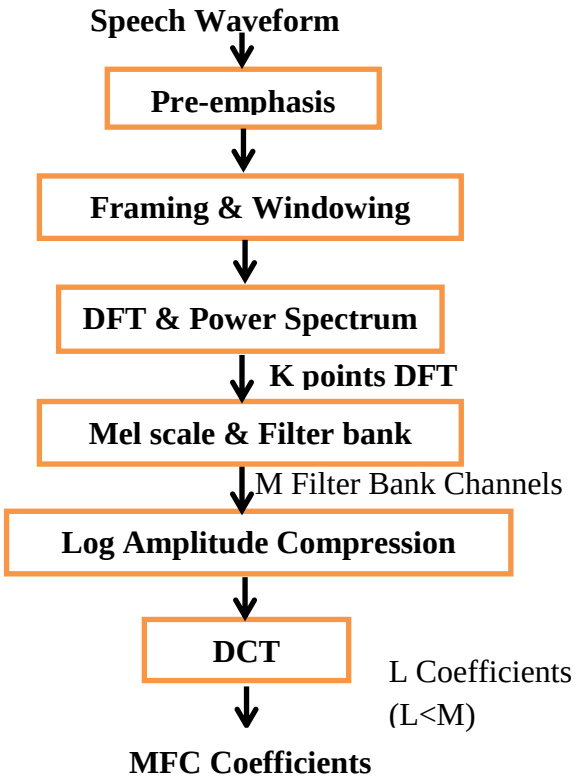

Fig. 3: MFCC extraction Process

4.1.3 Human Factor Cepstrum Analysis

Human factor cepstrum coefficients are closer to human auditory perception than MFCC because it uses HFCC filter. Its extraction technique is similar to MFCC feature extraction instead of filter.

4.2 LPC Analysis

The fundamental concept of this analysis technique is that a speech sample derived from a signal can be represented by a linear combination [6] of all other previous speech samples. We can derive a set of coefficients by reducing the total squared differences along a finite range between the derived speech samples and the linearly predicted samples.

LPC analysis states that a given speech sample for a signal at time *n*, $s(n)$.can be represented as a linear combination of all the previous *p* speech sample as given below:

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \cdots + a_n s(n-n).$$

Where, the predictor coefficients $a_1, a_2, \ldots . a_n$ are assumed to be constant over the speech analysis domain. The block diagram for computing LPC coefficients are given in Fig. 4.

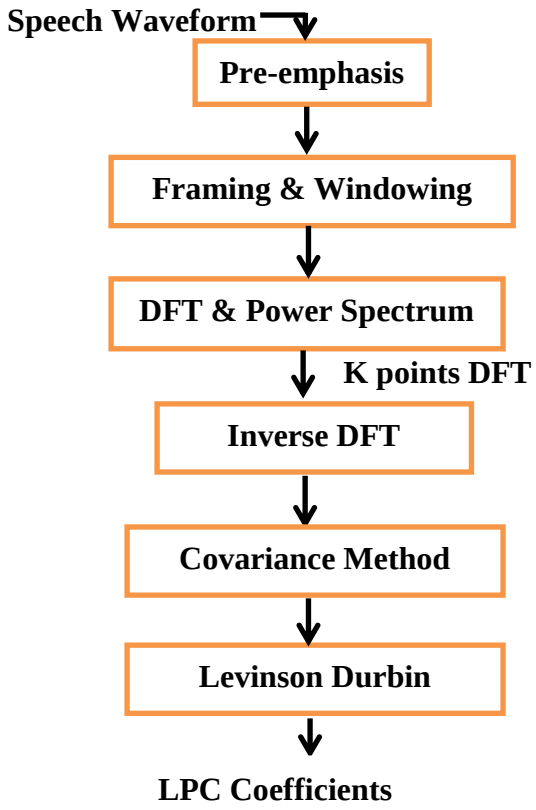

Fig. 4: LPC extraction process

4.3 PLP based analysis

PLP analysis models perceptually motivated auditory spectrum by a low order all pole function, using the autocorrelation LP technique

PLP analysis technique is basically based over the following three important factors derived from the mechanism ofhuman auditory response to an approximation of the hearing spectrum: (1) the critical-band derived for spectralresolution, (2) the intensity-loudness energy concept. And (3) equal-loudness curve,

PLP analysis technique is more efficient in autocorrelation response with human auditory system than the linear predictive analysis technique, conventionally.

PLP analysis technique has a higher computational efficiency and provides a low one-dimensional representation of speech samples.

An automatic speech recognition system takes the maximum advantage of these characteristic for speaker-independent systems.

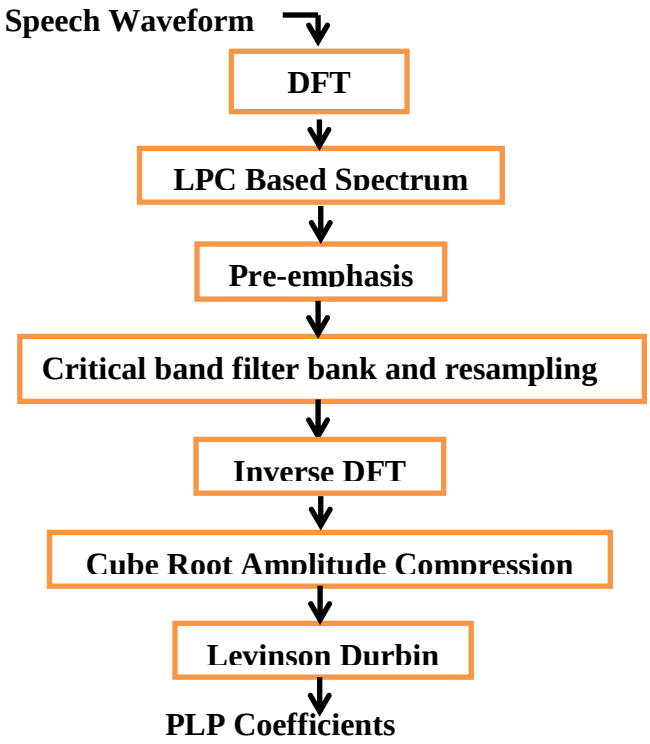

Fig. 5: PLP Extraction Process

4.4 Temporal Analysis

It involves processing of the waveform of speech signal directly. It involves less computation compared to spectral analysis but limited to simple speech parameters, e.g. power and periodicity.

4.4.1 Power Estimation

Power is rather simple to compute. It is computed on frame by frame basis as [1]

$$P(n) = \left({}^{1}/_{N_s}\right) \sum_{m=0}^{N_s} \left(w(m) s(n - \frac{N_s}{2} + m) \right)$$

Where N_s symbolises the sample numbers used to derive energy, $s(n)$denotes the signal, $w(m)$denotes the window function, and n denotes the sample index of center of the window in most speech recognition system Hamming window is almost exclusively used.

The major significance of $P(n)$is that it provides basis for distinguishing voiced speech segments from unvoiced speech segments.

The values of $P(n)$for the unvoiced segments are significantly smaller than for voiced segments.

5. PATTERN MATCHING TECHNIQUES

The models for pattern matching [5] techniques can be classified in two ways: (1) The Stochastic models, and (2) The Template models.

For a given stochastic model, pattern matching results in conditional probability, or a measure of analogy, of the observation, which implies that the pattern matching is probabilistic for a given model.

For a given template model, it is presumed that the observation is not a perfect copy of the original template and the alignment of the observed frames is chosen in such way

that it minimizes the distance measure 'd', this implies that the pattern matching is deterministic for a given model.

5.1 Template Models

In template based matching in order to evaluate the best matching pattern an unknown speech is compared with a set of pre-recorded words or templates.

5.2 Dynamic time warping

Dynamic Time warping is a template based system and it is one of the most common and majorly used procedures and is used to recompense speaking-rate inconsistency. Basically, Dynamic Time warping is used in automatic speech recognition to differentiate between various patterns of speech samples.

5.2.1 Concepts of DWT

Dynamic Time Warpingis an algorithm for pattern matching and it also has a non-linear time normalization effect [8]. The basic concept of DTW is derived from Bellman's principle for optimality. Bellman's principle states that for a given optimal path 'W', with starting point 'A' , ending point 'B' and having a point 'C' placed randomly somewhere over the optimal path, the path segment AC is the optimal path from A to C and the path segment CB is said to be optimal from C to B.

The DTW algorithm establishes an alignment (as shown in fig. 6) for two sequences of feature vectors, viz, $(T_1, T_2, \dots, T_N)$and $(S_1, S_2, \dots, S_N)$. A distance, say $d(i,j)$, is known as local distance if it can be calculated for any given two arbitrary feature vectors, say, T_i and S_j .

In DTW, for any two arbitrary feature vectors, say, T_i andS_j, we can evaluate the global distance, say (i,j) , between them by recursively summing its local distance $D(i,j)$, with the global distance which has been already calculated for the best predecessor.

The predecessor which provides the minimum global distance, say $D(i,j)$, (i.e. at row i and coloumn j) is considered as the best predecessor, as given below:

$$D(i,j) = \min_{m \le i, k \le j} [D(m,k)] + d(i,j)$$

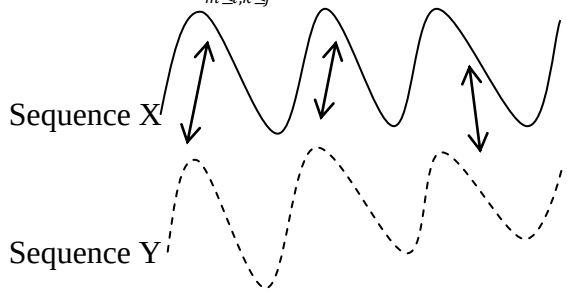

Fig. 6: Dynamic Time Warping

5.3 Vector Quantization

A VQ code book is a collection of code-words and it is typically designed by a clustering procedure. For every speaker, who is enrolled for speech recognition, a code book is developed with the help of his training data. This is generally done on the basis of how a specific text is read. A pattern match score can be formed as the distance between an input vector $xj$and the minimum distance code-word $x$in the claimant's VQ code book C.

This match score for L frames of speech is

$$Z = \sum_{j=1}^{L} \min_{x \in C} d(xj, \bar{x})$$

Vector Quantization (VQ) is often applied to ASR. The goal of this system is the data compression. Different VQ techniques are as follows:

5.3.1 K-means Algorithm

In this algorithm clustered the vectors based on attributes into k partitions. The main goal of this algorithm is to reduce the entire intra-cluster variance [9], V, to the least possible. $V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$

Here we have taken *k* clusters *Si, i* = 1,2... *K* and have kept μ*i*as the centroid or mean point of all these points, given, $x_j \in S_i$

The process of k-means algorithm uses:

a) Least-squares partitioning method to divide the input vectors into k initial sets.
b) Next it evaluates the mean point, or the centroid, of every individual set separately. It then builds a new partition by joining each point with the closest centroid.
c) After that the re-evaluation of all the centroids are performed for all the possible new clusters.
d) Algorithm is iterated till the time vectors stop switching clusters or else centroids are not changed again.

The *K*-means algorithm has also been named after Linde, Buzo and Gray as the *generalized LBG algorithm* in speech processing literature

5.3.2 Distortion Measure

The quantized code vector is selected which is approximated to be the closest to the input feature vector for a given speech sample in terms of Euclidean distance. The *Euclidean distance* is defined by:

$$d(x, y_i) = \sqrt{\sum_{i=1}^{L} (x^i - y_j^i)^2}$$

Where $x^i$is the i^{th}*component* of the input speech feature vector, and$y_j^i$is the i^{th}component of the code-wordy_i. Here the unknown speaker is recognized to be the one which has the least distortion distance.

5.3.3 Nearest Neighbors

Nearest Neighbors (NN) is a methodology of integrating the best features of DTW and VQ techniques into one.). Contrary to the vector quantization method it forms a very simple code book [10] without creating the clusters of training data which was enrolled. In fact, it maintains the database of all the training data and thus it can also make use of temporal information

5.4 Stochastic Models

With the help of a stochastic model we can formulate the pattern-matching problem as one measuring the likelihood of a particular observation (a feature vector of a cluster of vectors).

5.4.1 Hidden Markov Model

In an HMM, a given model behaves as a doubly embedded stochastic procedure [11] in which stochastic method which is underlying is not clearly noticeable for observation (it lies hidden). Here, the observations are actually a probabilistic function of the state.

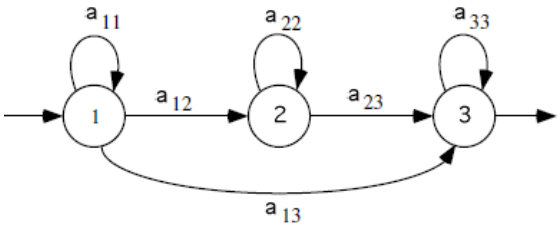

Fig. 7: an example of a three-state HMM

Basically, we can observe the HMM only through some other set of stochastic procedure which can produce the series of observations. The HMM can be considered as a finite-state machine, in which a probability density function (or feature vector stochastic model $\left({}^{x}/_{s_i}\right)$) is added with every state s_i(i.e. underlying main model). All the states are associated with each other through a transition network, in such a model, the state transition probabilities are represented as,$a_{ij} = p\left({}^{s_i}/_{s_j}\right)$.

Baum-Welch decoding [11] can be used to deduce the probability that a series of speech frames was created with the help of this model. The score for L frames of a given input speech frame is the likelihood of this model. This can be represented as follows:

$$P(x(1;L)|model) = \sum_{\substack{all\ state \\ Sequence}}^{L} {}_{i=1}\pi p(x_i|s_i)\, p(s_i|s_{i-1})$$

5.5 Artificial Neural Networks (ANN)

ANN is used to classify speech samples in the intelligent ways as shown in the figure 5.5.

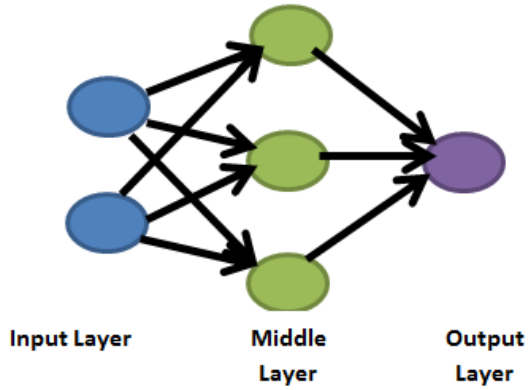

Fig. 8: Simplified view of an artificial neural network

The basic and main feature of ANN is its capability of learning by gaining strength and properties of inter-neuron connections (also called as synapses).

In the approach of Artificial Intelligence to speech recognition various sources of knowledge [2] are required to be set up. Thus, artificial intelligence is classified in two processes broadly: a) Automatic knowledge acquisitions learning and b) Adaptation. Neural networks have many similarities with Markov models. Both are statistical models which are represented as graphs.

Fig. 8: Simplified view of an artificial neural network

Where Markov models use probabilities for state transitions, neural networks use connection strengths and functions. A key difference is that neural networks are fundamentally parallel while Markov chains are serial.

Frequencies in speech occur in parallel, while syllable series and words are essentially serial. This means that both techniques are very powerful in a different context.

5.6 Hybrid Model (HMM/NN)

In many speech recognition systems, both techniques are implemented together and work in a symbiotic relationship [2].Neural networks perform very well at learning phoneme probability from highly parallel audio input, while Markov models can use the phoneme observation probabilities that neural networks provide to produce the likeliest phoneme sequence or word. This is at the core of a hybrid approach to natural language understanding.

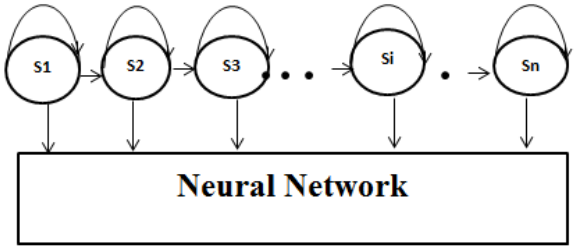

Fig. 9: n-state Hybrid HMM Model

6. EXPERIMENTAL ANALYSIS

A database of 100 speakers is created. Each speaker speaks a word 10 number of times. Totally, 10000 samples are collected from all the speakers. These words are collected by a laptop mounted microphone by using sonarca sound recorder software. The silence is removed from the all the samples through end point detection and they are stored as speech samples in wave format files with 16KHz sampling rate and 16 bits. Experiments are conducted on 50 speech samples of each word in different environmental conditions. Table 1 lists the words which are spoken by all 100 speakers and stored in the database.

Table 1: Dictionary of spoken words

Speaker number	Word
1	Hello
2	Shachi
3	AIET
4	MTech
5	December
6	Krishna
7	Diwali
8	Happy
9	Yellow
10	Google

The experiments are performed on several pattern matching techniques. This is done by applying various feature extraction techniques over them for word error recognition, as shown in fig. 10. Each word is recognized independently. We establish a recognition model from the training set for every

word. Technical results are described in the tables below: The results in table 2 shows that features extracted from MFCC are more efficient than the PLP, LPC and HFCC and the WER reached is 94.8%.We remark that among the entire pattern matching techniques, extraction features based on MFCC are the most promising one with the maximum word recognition rate reaching to 94.8 %(highest among all the feature extraction techniques).

Table 2: Comparative result analysis of features

Patten Matching techniques	LPC	PLP	HFCC	MFCC
DTW	76.4	85.6	85.7	90.4
VQ	65.8	78.5	74.6	96.5
HMM	80.5	77.6	80.4	86.2
Hybrid HMM	79.6	90.4	89.6	93.6
Average	77.6	85.7	88.7	94.8

In the next experiment we compare various pattern matching techniques (the HMM, VQ, Hybrid HMM/ANN, DTW) and tested for maximum word recognition efficiency in different environmental conditions (i.e. i.e. in closed room, in class room, in a car, in a seminar-hall, in open-air), as shown in figure 11 and results in table 3.
The results show that pattern matching based on HMM or VQ yield better results in different environmental conditions. DTW though is also closely promising one but it is visible from results that it gives less good accuracy.
The results in Table 2 also show that the two techniques (viz HMM and hybrid) are comparable but the HMM one provides slightly best results.
We remark that for the pattern matching based on Hybrid HMM , the efficiency of performances are better than all others with word recognition rate reaching up to (93.7%) longer need a human operator for much help and the service provider no longer need a bigger staff. But still security concerns require more research and development in some areas to make the speech recognition technology more dependent.

7. CONCLUSION

We have discussed various techniques for speech recognition that include processes for the feature extraction and pattern matching. From the above presented results we can conclude results regarding these techniques. In overall test MFCC with hybrid HMM technique. MFCC behave its characteristics like human auditory perception and hybrid HMM involves Neural net in its processing and shown maximum results as compare to other techniques.

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.

8. REFERENCES

[1] M. Cowling, R. Sitte, Analysis of Speech Recognition Techniques for use in a Non-Speech Sound Recognition System, Member, IEEE, Griffith University, Gold Coast, Qld, Australia.

[2] W. Gevaert, G. Tsenov, Senior Member, IEEE "Neural Networks used for Speech Recognition" Journal of Automatic Control, Belgrade, VOL. 20:1-7, 2010.

[3] S. K.Gaikwad, B.W.Gawali, "A Review on Speech Recognition Technique" International Journal of Computer Applications (0975 – 8887)Volume 10– No.3, November 2010.

[4] M. P. Kesarkar,"Feature Extraction for Speech Recognition", Electronic Systems, EE. Dept., IIT Bombay, November, 2003

[5] M AAnusuya, "Classification Techniques used in Speech Recognition Applications: A Review" International Journal Computer Technology Application, Vol. 2 (4), 910-954.

[6] K Sharma, H.P.Sinha "Comparative Study Of Speech recognition System using various feature extraction techniques" Int. J. IT and Knowledge ManagementJuly-Dec 2010, Volume 3, No. 2, pp. 695-698

[7] Mporas, T.Ganchev," Comparison of Speech Features on the Speech Recognition Task", Journal of Computer Science 3 (8): 608-616, 2007

[8] N. Meseguer, "Speech analysis for automatic speech recognition" Nowegian University of science and Technology.

[9] M. Gill, R. Kaur, "Vector Quantization based Speaker Identification", Int. Journal of computer applications",Vol 4 – No.2, July 2010

[10] S.Vimala, "Convergence Analysis of Codebook Generation Techniques for Vector Quantization using K-Means Clustering Technique", International Journal of Computer Applications Vol. 21– No.8, May 2011

[11] S.Melnikoff, S.Quigley, "Implementing a Hidden Markov Model SpeechRecognition System" 11th International Conference on Field Programmable Logic and Applications, FPL 2001.

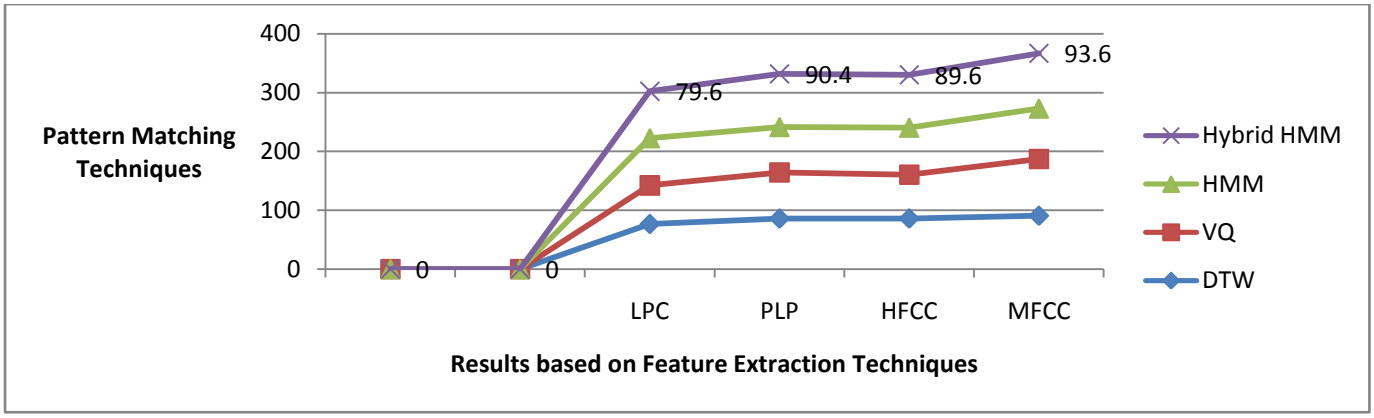

Fig. 10: Results based on different pattern matching techniques

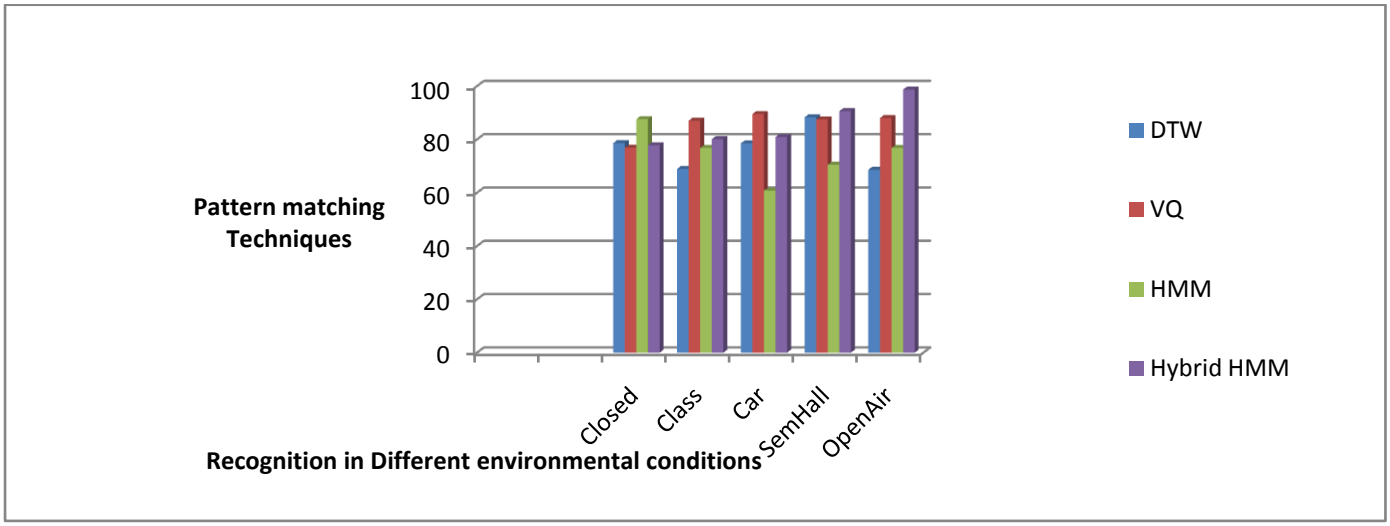

Fig. 11: Recognition results in the different environmental conditions

Table 3: Recognition results Table in the different environmental conditions.

Pattern Matching Technique	Closed	Class	Car	SemHall	OpenAir	Average
DTW	78.6	68.9	78.5	88.3	68.6	78.7
VQ	76.9	87	89.5	87.5	88	83.4
HMM	87.6	76.8	60.8	70.5	76.8	75.9
Hybrid HMM	77.8	80.1	80.9	90.6	98.7	93.7