A Review on Speech Recognition Technique

Santosh K.Gaikwad
Research Student
Department of CS& IT
Dr.Babasaheb Ambedkar Marathwada University Aurangabad

Bharti W.Gawali
Associate Professor
Department of CS& IT
Dr.Babasaheb Ambedkar Marathwada University Aurangabad

Pravin Yannawar
Assistant Professor
Department of CS& IT
Dr.Babasaheb Ambedkar Marathwada University Aurangabad

ABSTRACT
The Speech is most prominent & primary mode of Communication among of human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer .This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and also gives overview technique developed in each stage of speech recognition. This paper helps in choosing the technique along with their relative merits & demerits. A comparative study of different technique is done as per stages. This paper is concludes with the decision on feature direction for developing technique in human computer interface system using Marathi Language.

General Terms
Human computer Iterface, Modeling technique, speech processing, signal processing, Pattern Recognition.

Keywords
Analysis, feature extraction, Modeling, Testing, speech processing, HCI

1. INTRODUCTION
The speech is primary mode of communication among human being and also the most natural and efficient form of exchanging information among human in speech. So, it is only logical that the next technological development to be natural language speech recognition for HCI. Speech Recognition can be defined as the process of converting speech signal to a sequence of words by means Algorithm implemented as a computer program. Speech processing is one of the exciting areas of signal processing. The goal of speech recognition area is to developed technique and system to developed for speech input to macine.based on major advanced in statically modeling of speech ,automatic speech recognition today find widespread application in task that require human machine interface such as automatic call processing.[1]. Since the 1960s computer scientists have been researching ways and means to make computers able to record interpret and understand human speech. Throughout the decades this has been a daunting task. Even the most rudimentary problem such as digitalizing (sampling) voice was a huge challenge in the early years. It took until the 1980s before the first systems arrived which could actually decipher speech. Off course these early systems were very limited in scope and power. Communication among the human being is dominated by spoken language, therefore it is natural for people to expect speech interfaces with computer .computer which can speak and recognize speech in native languaguage [2]. Machine recognisation of speech involves generating a sequence of words best matches the given speech signal. Some of known applications include virtual reality, Multimedia searches, auto-attendants, travel Information and reservation, translators, natural language understanding and many more Applications (Scan soft, 2004 [3]; Robertson, 1998 [4])

1.1. Type of Speech
Speech recognition system can be separated in different classes by describing what type of ullerances they can recognize.

1.1.1 *Isolated Word*
Isolated word recognizes attain usually require each utterance to have quiet on both side of sample windows. It accepts single words or single utterances at a time .This is having "Listen and Non Listen state". Isolated utterance might be better name of this class [5].

1.1.2 *Connected Word*
Connected word system are similar to isolated words but allow separate utterance to be "run together minimum pause between them.

1.1.3 *Continuous speech*
Continuous speech recognizers allows user to speak almost naturally, while the computer determine the contnt.Recognizer with continues speech capabilities are some of the most difficult to create because they utilize special method to determine utterance boundaries.

1.1.4 *Spontaneous speech*
At a basic level, it can be thought of as speech that is natural sounding and not rehearsed .an ASR System with spontaneous speech ability should be able to handle a variety of natural speech feature such as words being run together.

1.2 ASR System classification
Speech Recognition is a special case of pattern recognition. There are two phase in supervised pattern recognition, viz., Training and Testing. The process of extraction of features relevant for classification is common in both phases. During the traning 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.

2. SPEECH RECOGNITION TECHNIQUES

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. The earliest speech recognition systems were first attempted in the early 1950s at Bell Laboratories, Davis, Biddulph and Balashek developed an isolated digit Recognition system for a single speaker [1]. The goal of automatic speaker recognisation is to analyze, extract characterize and recognize information about the speaker identity. The speaker recognisation system may be viewed as working in a four stages

1. Analysis
2. Feature extraction
3. Modeling
4. Testing

2.1 Speech analysis technique

Speech data contain different type of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis stage deals with stage with suitable frame size for segmenting speech signal for further analysis and extracting [7]. The speech analysis technique done with following three techniques

2.1.1 Segmentation analysis

In this case speech is analyzed using the frame size and shift in the range of 10-30 ms to extract speaker information.Studid made in used segmented analysis to extract vocal tract information of speaker recognition.

2.1.2 Sub segmental analysis

Speech analyzed using the frame size and shift in range 3-5 ms is known as Sub segmental analysis. This technique is used to mainly analyze and extract the characteristic of the excitation state. [8].

2.1.3 Supra segmental analysis

In this case, speech is analyzed using the frame size this technique is technique is used mainly to analyze and characteristic due to behavior character of the speaker.

2.1.4 Performance of System

The performance of speaker recognisation system depends on the technique employed in the various stages of speaker recognisation system. The state of art of speaker recognisation system mainly used segmental analysis, Mel frequency Spectral coefficients (MFFCs), Gaussian mixture model (GMM) and feature extraction, modeling and testing stage. There are practical issues in the speaker recognisation field other technique may also have to be used for resulting a good speaker recognisation performance some of practical issues are a s follows ….

2.1.4.1. Nonacoustic sensor provide an exciting opportunity for multimodal speech processing with application to areas such as speech enhancement and coding .this sensor provide measurement of function of the glottal excitation and can supplement acoustic waveform[1].

2.1.4.2. A universal background Model (UBM) is a model used in a speaker verification system to represent general person independent the feature characteristics to be compared against a model of person specific feature characteristics when making an accept or reject decision[2].

2.1.4.3. A Multimodel person recognisation architecture has been developed for the purpose of improving overall recognition performance and for addressing channel specific performance. This multimodal architecture includes the fusion of speech recognisation system with the MIT/LL GMM/UBM speaker recognition architecture [3].

2.1.4.4. Many powerful for speaker recognition have introduced in high level features, novel classifiers and channel compression methods [4].

2.1.4.5. SVMs have become a popular and powerful tool in text independent speaker verification at the core of any SVM type system give a choice of feature expansion.[5]

2.1.4.6. A recent areas of significant progress in speaker recognition is the use of high level features-idiolect, phonetic relations, prosody. A speaker not only has distinctive acoustic sound but uses language in a characteristic manner. [6]

2.2. Feature Extraction Technique

The speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need feature extraction of speech signal. Following are some feature extraction.

Table1: List of technique with their properties For Feature extraction

Sr.No	Method	Property	Procedure for Implementation
1	Principal Component analysis (PCA)	Non linear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data
2	Linear Discriminate Analysis(LDA)	Non linear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification[9]
3	Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non- Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower order coefficient,	It is used for feature Extraction at lower order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope[9]
6	Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency scale i.e. Mel-frequency Scale.
7	Filter bank analysis	Filters tuned required frequencies	
8	Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	This method is used for find our features
9	Kernel based feature extraction method	Non linear transformations	Dimensionality reduction leads to better classification and it is used to redundant features, and improvement in classification error.[11]
10	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e. II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or runtime Feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram[4]
13	Cepstral mean subtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter
14	RASTA filtering	For Noisy speech	It is find out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	A transformation based on PCA+LDA+ICA	Higher Accuracy than the existing Methods[14]

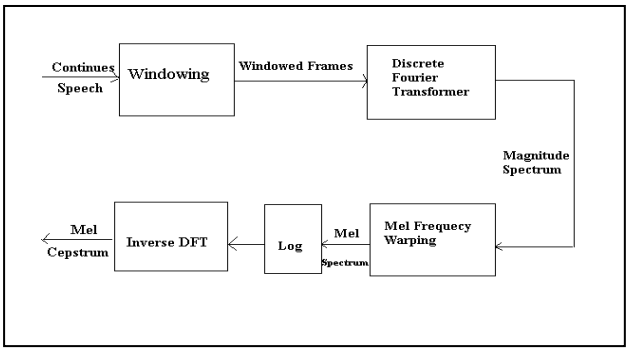

Fig.1: Feature Extraction diagram

A new modification of Mel-Frequency Cepstral Coefficient (MFCC) feature has been proposed for extraction of speech features for Speaker Verification (SV) application. This is compared with original MFCC based feature extraction method and also on one of the recent modification. The work uses multi-dimensional F-ratio as performance measure in Speaker Recognition (SR) applications to compare discriminative ability of different multi parameter methods [10]. A number of problems with the standard method of hidden Markov models (HMMs) and features derived from fixed, frame-based spectra (e.g. MFCCs) are discussed. Based on these problems, a set of desirable properties of an improved acoustic model are proposed, and we present a "parts-based" framework as an alternative. The parts-based model (PBM), based on previous work in machine vision, uses graphical models to represent speech with a deformable template of spectrum-temporally localized "parts", as opposed to modeling speech as a sequence of fixed spectral profiles. We discuss the proposed model's relationship to HMMs and segment-based recognizers, and describe how they can be viewed as special cases of the PBM [11]. Each person voice is different thus the quran sound which had been recited by person by person that means using mfcc we can calculate a verses of sound in that mfcc consist of framing,windowing.dft,mel filter bank and inverse dft[12].

The different feature extraction technique describe as follows

- Spectral feature like band energies, formats, spectrum and Cepstral coefficient mainly speaker specific information due to vocal tract.
- Excitation source feature like pitch and variation in pitch.
- Long term feature like duration ,information energy due to behavior feature

2.3 Modeling Technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classification speaker recognition and speaker identification. The speaker identification technique automatically identify who is speaking on basis of individual information integrated in speech signal The speaker recognisation is also divided into two parts that means speaker dependant and speaker independent. In the speaker independent mode of the speech recognisation the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message .on the other hand in case of speaker recognisation machine should extract speaker characteristics in the acoustic signal [13]. The main aim of speaker identification is comparing a speech signal from an unknown speaker to a database of known speaker .The system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be divide into two methods, text- dependent and text independent methods. In text dependent method the speaker say key words or sentences having the same text for both training and recognition trials. Whereas text independent does not rely on a specific texts being spoken [14]. Following are the modeling which can be used in speech recognition process:

I. The acoustic-phonetic approach

This method is indeed viable and has been studied in great depth for more than 40 years. This approach is based upon theory of acoustic phonetics and postulates [15]. The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach (Hemdal and Hughes 1967). Which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time? Even though, the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (the so-called co articulation effect), it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned by a machine [16]. Formal evaluations conducted by the National Institute of Science and Technology (NIST) in 1996 demonstrated that the most successful approach to automatic language identification (LID) uses the phonotactic content of a speech signal to discriminate among a set of languages[18]. Phone-based systems, such as those described in [19] and [20]. There are three techniques that have been applied to the language identification. Problem phone recognition, Gaussian mixture modeling, and support vector machine classification. [22][23].Using IPA Methods we can find similarities for probabilities of content dependant acoustic model for new language.[24]. The acoustic phonetic approach has not been widely used in most commercial applications [25].

II.Pattern Recognition approach

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. A pattern recognition has been developed over two decade received much attention and applied widely too many practical pattern recognition problem [25].A speech pattern representation can be in the form of a speech template or a statistical model (e.g., a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness

of match of the patterns. The pattern-matching approach has become the predominant method for speech recognition in the last six decades ([24] pg.87).

III.Template based approaches

Template based approaches matching (Rabiner *et al.*, 1979) Unknown speech is compared against a set of pre-recorded words (templates) in order to find the best Match. This has the advantage of using perfectly accurate word models. Template based approach [26] to speech recognition have provided a family of techniques that have advanced the field considerably during the last six decades. The underlying idea is simple. A collection of prototypical speech patterns are stored as reference patterns representing the dictionary of candidate s words. Recognition is then carried out by matching an unknown spoken utterance with each of these reference templates and selecting the category of the best matching pattern. Usually templates for entire words are constructed. This has the advantage that, errors due to segmentation or classification of smaller acoustically more variable units such as phonemes can be avoided. In turn, each word must have its own full reference template; template preparation and matching become prohibitively expensive or impractical as vocabulary size increases beyond a few hundred words. One key idea in template method is to derive typical sequences of speech frames for a pattern (a word) via some averaging procedure, and to rely on the use of local spectral distance measures to compare patterns. Another key idea is to use some form of dynamic programming to temporarily align patterns to account for differences in speaking rates across talkers as well as across repetitions of the word by the same talker. But it also has the disadvantage that pre-recorded templates are fixed, so variations in speech Can only be modeled by using many templates per word, which eventually becomes Impractical [27].

IV. Dynamic time warping

Dynamic time warping is an algorithm for measuring similarity between two sequences which may vary in time or speed. For instance, similarities in walking patterns would be detected, even if in one video, the person was walking slowly and if in another, he or she were walking more quickly, or even if there were accelerations and decelerations during the course of one observation. DTW has been applied to video, audio, and graphics indeed, any data which can be turned into a linear representation can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. In general, DTW is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in the context of hidden Markov models. Dynamic time warping (DTW) is such a typical approach for a template based approach matching for speech recognition and also DTW stretches and compresses various sections of utterance so as to find alignment That results in best possible match between template and utterance on frame-by frame basis .By "frame" we mean short segment (10-30ms) of speech signal which is basis of parameter vector computation, and "match" defined as sum of frame-by frame distances between template and input utterance. Template with closest match defined in manner chosen as recognized word .To absorbed acoustic variations; statistical methods can be integrated into DTW approaches.DTW quite efficient for isolated word recognition and can be adapted to connected word recognition. One example of the restrictions imposed on the matching of the sequences is on the monotonicity of the mapping in the time dimension. Continuity is less important in DTW than in other pattern matching algorithms; 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.

V. Knowledge based approaches

An expert knowledge about variations in speech is hand coded into a system. This has the advantage of explicit modeling variations in speech; but unfortunately such expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and automatic learning procedure was sought instead. Vector Quantization (VQ)[28] is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. Since transmission rate is not a major issue for ASR, the utility of VQ here lies in the efficiency of using compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. For IWR, each vocabulary word gets its own VQ codebook, based on training sequence of several repetitions of the word. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure [29].

VI.Statistical based approaches

In which variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM. The approaches represent the current state of the art. The main disadvantage of statistical models is that they must take priori modeling assumptions which are answerable to be inaccurate, handicapping the system performance. In recent years, a new approach to the challenging problem of conversational speech recognition has emerged, holding a promise to overcome some fundamental limitations of the conventional Hidden Markov Model (HMM) approach (Bridle et al., 1998 [8]; Ma and Deng, 2004 [9]).This new approach is a radical departure from the current HMM-based statistical modeling approaches. For text independents speaker recognition use left-right HMM for identifying the speaker from simple data and also HMM having advantages based on Neural Network and Vector Quantization.

The HMM is popular statistical tool for modeling a wide range of time series data .In Speech recognition area HMM have been applied with great success to problem such as part of speech classification[30].

A weighted hidden markov model HMM algorithm and a subspace projection algorithm are proposed in[31], to address the discrimination and robustness issues for HMM based speech recognition. Word models were constructed for combining phonetic and fenonic models [31] A new hybrid algorithm based on combination of HMM and learning vector were proposed in [30]. Learning Vector Quantization[31] (LVQ) method showed an important contribution in producing highly discriminative reference vectors for classifying static patterns. The ML estimation of the parameters via FB algorithm was an inefficient method for estimating the parameters values of HMM. To overcome this problem paper [32] proposed a corrective training

method that minimized the number of errors of parameter estimation. A novel approach [33] for a hybrid connectionist HMM speech recognition system based on the use of a Neural Network as a vector quantize. Showed the important innovations in training the Neural Network. Next the Vector Quantization approach showed much of its significance in the reduction of Word error rate. MVA[33] method obtained from modified Maximum Mutual Information(MMI) is shown in this paper. Nam So Kim et.al., have presented various methods for estimating a robust output probability distribution(PD) in speech recognition based on the discrete Hidden Markov Model(HMM) in their paper[34].An extension of the viterbi algorithm[35] made the second order HMM computationally efficient when compared with the existing viterbi algorithm. In paper[36] a general stochastic model that encompasses most of the models proposed in the literature, pointing out similarities of the models in terms of correlation and parameter time assumptions, and drawing analogies between segment models and HMMs have been described. An alternative VQ [37] method in which the phoneme is treated as a cluster in the speech space and a Gaussian model was estimated for each phoneme. The results showed that the phoneme-based Gaussian modeling vector quantization classifies the speech space more effectively and significant improvements in the performance of the DHMM system have been achieved [38]. The trajectory folding phenomenon in HMM model is overcome by using Continuous Density HMM which significantly reduced the Word Error Rate over continuous speech signal as been demonstrated by [39]. A new hidden Markov model [37] showed the integration of the generalized dynamic feature parameters into the model structure was developed and evaluated using maximum-likelihood (ML) and minimum-classification-error (MCE) pattern recognition approaches. The authors have designed the loss function for minimizing error rate specifically for the new model, and derived an analytical form of the gradient of the loss function.

The K-means algorithm is also used for statistical and clustering algorithm of speech Based on the attribute of data .The K in K-means represents the number of clusters the algorithm should return in the end. As the algorithm starts K points known as cancroids are added to the data space. The **K-means algorithm** is a way to cluster the training vectors to get feature vectors. In this algorithm clustered the vectors based on attributes into k partitions. It uses the *k* means of data generated from Gaussian distributions to cluster the vectors. The objective of the k-means is to minimize total intra-cluster variance [40].

VII.Learning based approaches

To overcome the disadvantage of the HMMs machine learning methods could be introduced such as neural networks and genetic algorithm programming. In those machine learning models explicit rules or other domain expert knowledge) do not need to be given they a can be learned automatically through emulations or evolutionary process.

VIII. The artificial intelligence approach

The artificial intelligence approach attempts to mechanize the recognition procedure According to the way a person applies its intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. Expert system is used widely in this approach (Mori *et al.*, 1987) [41]). The Artificial Intelligence approach [26] is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic phonetic and pattern recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Some speech researchers developed recognition system that used acoustic phonetic knowledge to develop classification rules for speech sounds. While template based approaches have been very effective in the design of a variety of speech recognition systems; they provided little insight about human speech processing, thereby making error analysis and knowledge-based system enhancement difficult. On the other hand, a large body of linguistic and phonetic literature provided insights and understanding to human speech processing [42]. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert s speech knowledge into a recognition system. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. Pure knowledge engineering was also motivated by the interest and research in expert systems. However, this approach had only limited success, largely due to the difficulty in quantifying expert knowledge. Another difficult problem is the integration of many levels of human knowledge phonetics, phonotactics, lexical access, syntax, semantics and pragmatics. Alternatively, combining independent and asynchronous knowledge sources optimally remains an unsolved problem. In more indirect forms, knowledge has also been used to guide the design of the models and algorithms of other techniques such as template matching and stochastic modeling. This form of knowledge application makes an important distinction between knowledge and algorithms Algorithms enable us to solve problems. Knowledge enables the algorithms to work better. This form of knowledge based system enhancement has contributed considerably to the design of all successful strategies reported. It plays an important role in the selection of a suitable input representation, the definition of units of speech, or the design of the recognition algorithm itself.

IX. Stochastic Approach

Stochastic modeling [43] entails the use of probabilistic models to deal with uncertain or incomplete information. In speech recognition, uncertainty and incompleteness arise from many sources; for example, confusable sounds, speaker variability s, contextual effects, and homophones words. Thus, stochastic models are particularly suitable approach to speech recognition. The most popular stochastic approach today is hidden Markov modeling. A hidden Markov model is characterized by a finite state markov model and a set of output distributions. The transition parameters in the Markov chain models, temporal variability's, while the parameters in the output distribution model, spectral variability's. These two types of variability's are the essence of speech recognition. Compared to template based approach, hidden Markov modeling is more general and has a firmer mathematical foundation. A template based model is simply a continuous [44].

2.4 Matching Techniques

Speech-recognition engines match a detected word to a known word using one of the following techniques (Svendsen *et al.*, 1989 [12]).

I. **Whole-word matching**. The engine compares the incoming digital-audio signal against a prerecorded template of the word. This technique takes much less processing than sub-word matching, but it requires that the user (or someone) prerecord every word that will be recognized - sometimes several hundred thousand words. Whole-word templates also require large amounts of storage (between 50 and 512 bytes per word) and are practical only if the recognition vocabulary is known when the application is developed [45].

II. Sub-word matching. The engine looks for sub-words - usually phonemes and then performs further pattern recognition on those. This technique takes more processing than whole-word matching, but it requires much less storage (between 5 and 20 bytes per word). In addition, the pronunciation of the word can be guessed from English text without requiring the user to speak the word beforehand.[46] [47] discuss that research in the area of automatic speech recognition had been pursued for the last three decades.

3. PERFORMANCE OF SYSTEMS

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with word error rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR)[48].

3.1 Word Error Rate (WER)

Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [50][51]. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as

$$WER = \frac{S + D + I}{N}$$

Where

- *S* is the number of substitutions,
- *D* is the number of the deletions,
- *I* is the number of the insertions,

N is the number of words in the reference

When reporting the performance of a speech recognition system, sometimes word recognition rate (WRR) is used instead.

$$WRR = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N}$$

Where

- *H* is N-(S+D), the number of correctly recognized words

4. CONCLUSION AND FUTURE WORKS

In this review, we have discussed the technique developed in each stage of speech recognition system. We also presented the list of technique with their properties for Feature extraction. Through this review it is found that MFCC is used widely for feature extraction of speech and GHM and HMM is best among all modeling technique.

5. ACKNOWLEDGMENTS

The authors would like to thank the university authorities for Providing infrastructure to carry out the experiments. This work is supported by DST.

6. REFERENCES

[1] R.Klevansand R.Rodman, "Voice Recognition, Artech House, Boston, London 1997.

[2] Samudravijaya K. Speech and Speaker recognition tutorial TIFR Mumbai 400005.

[3] Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn "An Evaluation Of Audio-Visual person Recognition on the XM2VTS corpus using the Lausanne protocol" MIT Lincoln Laboratory, 244 Wood St., Lexington MA

[4] W. M. Campbell_, D. E. Sturim W. Shen D. A. Reynolds_, J. Navr´atily "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition" MIT Lincoln Laboratory IBM 2006.

[5] Zahi N.Karam,William M.Campbell "A new Kernel for SVM MIIR based Speaker recognition "MIT Lincoln Laboratory, Lexington, MA, USA.

[6] Asghar .Taheri ,Mohammad Reza Trihi et.al,Fuzzy Hidden Markov Models for speech recognition on based FEM Algorithm, Transaction on engineering Computing and Technology V4 February 2005,IISN,1305-5313

[7] GIN-DER WU AND YING LEI " A Register Array based Low power FFT Processor for speech recognition" Department of Electrical engineering national Chi Nan university Puli ,545 Taiwan

[8] Nicolás Morales1, John H. L. Hansen2 and Doorstep T. Toledano1 "MFCC Compensation for improved recognition filtered and band limited speech" Center for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA

[9] M.A.Anusuya , S.K.Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security 2009.

[10] Goutam Saha, Ulla S. Yadhunandan " Modifield Mel-Frequency Cepstral coefficient Department of Electronics and Electrical communication Engineering India Institute of

Technology ,Kharagpur Kharagpur-721302 West Bengal,India. .

[11] Kenneth Thomas Schutte "Parts-based Models and Local Features for Automatic Speech Recognition" B.S., University of Illinois at Urbana-Champaign (2001) S.M., Massachusetts Institute of Technology (2003).

[12] Zaidi Razak, Noor Jamaliah Ibrahim, Emran Mohd Tamil, Mohd Yamani Idna Idris "Quarnic Verse recitation feature extraction using Mel-Frequency Cepstral Coefficient(MFCC)" Department of Al-Quran & Al-Hadith, AcademyOf Islamic Studies, University of Malaya .

[13] Samudravijay K "Speech and Speaker recognition report" source: http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html

Viewed on 23 Feb. 2010.

[14] Sannella, M Speaker recognition Project Report report"

From http://cs.joensuu.fi/pages/tkinnu/research/index.html

Viewed 23 Feb. 2010.

[15] IBM (2010) online IBM Research

Source:-http://www.research.ibm.com/Viewed 12 Jan 2010.

[16] P.satyanarayana "short segment analysis of speech for enhancement" institute of IIT Madras feb.2009

[17] David, E., and Selfridge, O., Eyes and ears for computers, Proc.IRE 50:1093.

[18] SadokiFuruki,Tomohisa Ichiba et.al,Cluster-based Modeling for Ubiquitous Speech Recognition, Department of Computer Science Tokyo Institute of Technology Interspeech 2005.

[19] Spector, Simon Kinga and Joe Frankel, Recognition ,Speech production knowledge in automatic speech recognition , Journal of Acoustic Society of America,2006

[20] M.A Zissman,"Predicting,diagonosing and improving automatic Language identification performance" ,Proc.Eurospeech97,Sept.1997 vol.1,pp.51-54 1989.

[21] Y.Yan and E.Bernard ,"An apporch to automatic language identification basedon language depandant phone recognition ",ICASSP'95,vol.5,May.1995 p.3511

[22] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds ,"Accoustic ,phonetic and discriminative approach to automic Language Idantification".

[23] Viet Bac Le, Laurent Besacier, and Tanja Schultz, Acoustic-phonetic unit similarities for context dependant acoustic model portability Carnegie Mellon University, Pittsburgh, PA, USA

[24] C.S.Myers and L.R.Rabiner, A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition , IEEE Trans. Acoustics, Speech Signal Proc.,ASSP-29:284-297,April 1981.

[25] D.R.reddy,An Approach to Computer speech Recognition by direct analysis of the speech wave,Tech.Report No.C549,Computer Science Department ,Stanford University,sept.1996

[26] Tavel R.K.Moore,Twenty things we still don't know about speech proc.CRIM/FORWISS Workshop on Progress and Prospects of speech Research and Technology 1994.

[27] H.Sakoe and S.Chiba, Dynamic programming algorithm optimization for spoken word recognition ,IEEE Trans. Acoustics, Speech, Signal Proc.,ASSP-26(1).1978

[28] Keh-Yih Su et.al., Speech Recognition using weighted HMM and subspace IEEE Transactions on Audio, Speech and Language.

[29] L.R.Bahl et.al, A method of Construction of acoustic Markov Model for words, IEEE Transaction on Audio ,speech and Language Processing ,Vol.1,1993

[30] Shigeru Katagiri et.al., A New hybrid algorithm for speech recognition based on HMM segmentation and learning Vector quantization , IEEE Transactions on Audio Speech and Language processing Vol.1,No.4

[31] G. 2003 Lalit R .Bahl et.al.,Estimating Hidden Markov Model Parameters so as to maximize speech recognition Accuracy,IEEE Transaction on Audio, Speech and Language Processing Vol.1 No.1 , Jan.1993.

[32] Gerhard Rogoll,Maximum Mutual Information Neural Networks for hybrid connectionist-HMM speech Recognition systems ,IEEE Transaction on Audio, speech and Language Processing Vol.2 ,No.1,Part II,Jan.1994.

[33] Antonio M. Peinado et.al, discriminative codebook design using Multiple Vector quatization in HMM based speech recognizers,IEEE Transaction on Audio,Speech and language Processing Vol.4 No.2 March.1996

[34] Nam Soo kim et.al,On estimating robust Probability Distribution in HMM in HMM based Speech Recognition ,IEEE Transaction on Audio, Speech and Language Processing Vol.3,No.4 ,July 1995.

[35] Jean Francois, Automatic word Recognition Based on Second Order hidden Markov Models.IEEE Transaction on Audio, Speech and Language ProcessingVol.5, No.1, Jan.1997.

[36] Mari ostendorf et.al. from HMM to segment Models: a Unified View stochastic Modeling for speech Recognition ,IEEE Transaction on audio, speech and Language Processing Vol.4,No.5,September 1996.

[37] John butzberger ,Spontaneous speech effects In Large Vocabulary Speech Recognition application,SRI International Speech Research and Technology Program Menlo Park,CA 94025

[38] Dannis Norris, "Merging Information in Speech Recognition" feedback is never Necessary workshop.1995

[39] Yifan gong, stochastic trajectory Modeling and Sentence searching for continuous Speech Recognition,IEEE Transaction on Speech and Audio Processing,1997.

[40] Alex weibel and Kai-Fu Lee, reading in Speech recognition ,Morgan Kaufman Publisher,Inc.San Mateo,California,1990.

[41] John Butzberger, Spontanious Speech Effect in Large Vocublary speech recognition application, SRI International Speech Research and Technology program Menlo Park, CA94025.

[42] M.J.F.Gales and S.J young, Parallel Model combination for Speech Recognition in Noise technical Report, CUED/FINEFENG/TRI135, 1993.

[43] A.P.Varga and R.K.Moore, "Hidden Markov Model Decomposition of Speech and Noise, Proc.ICASSp, pp.845-848, 1990.

[44] M.Weintraub et.al, linguistic constraints in hidden markov Model based speech recognition, Proc.ICASSP, pp.699-702, 1989.

[45] S.katagiri, Speech Pattern recognition using Neural Networks.

[46] L.R.Rabiner and B.H.jaung ," Fundamentles of Speech Recognition Prentice-Hall, Englewood Cliff, New Jersy, 1993

[47] D.R.Reddy, An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave , Tech.Report No.C549, Computer Science Dept., Stanford Univ., September 1966

[48] K.Nagata, Y.Kato, and S.Chiba, Spoken Digit Recognizer for Japanese Language , NEC Res.Develop., No.6,1963

[49] D.B.Fry, Theoretical Aspects of Mechanical speech Recognition , and P.Denes, The design and Operation of the Mechanical Speech Recognizer at University College London, J.British Inst. Radio Engr., 19:4,211-299,1959.

[50] Dat Tat Tran, Fuzzy Approaches to Speech and Speaker Recognition , A thesis submitted for the degree of Doctor of Philosophy of the university of Canberra

[51] Lawrence Rabiner, Biing Hwang Juang, Fundamental of Speech Recognition, Copyright 1999by AT&T.