Volume 3, Issue 5, May 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Survey On: Voice Command Recognition Technique

Om Prakash Prabhakar
Research Student
Department of ET&T
CSIT Durg (CG) India.

Navneet Kumar Sahu
Assistant Professor
Department of ET&T
CSIT Durg (CG) India.

Abstract-- The Speech is most essential & 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. Today, speech technologies are commercially available for an unlimited but interesting range of tasks. These technologies enable machines to respond correctly and reliably to human voices, and provide useful and valuable services. This paper gives an overview of major technological perspective and appreciation of the fundamental progress of speech recognition and gives overview technique developed in each stage of speech recognition and also summarize and compare different speech recognition systems and identify research topics and applications which are at the forefront of this exciting and challenging field.

Keywords – Speech Recognition; Feature Extraction; MFCC; LPC; HMM; DTW; Modeling; Testing.

I. Introduction

Speech is the most basic, common and efficient form of communication method for people to interact with each other. People are comfortable with speech therefore persons would also like to interact with computers via speech, rather than using primitive interfaces such as keyboards and pointing devices. This can be accomplished by developing an Automatic Speech Recognition (ASR) system. Which is the process of converting a speech signal to a sequence of words by means of an algorithm implemented as a computer program. It has the potential of being an important mode of interaction between humans and computers [1]. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. The research in ASR by machines has attracted a great deal of attention for about sixty years [2] and ASR today finds widespread application in tasks that require human machine interface, such as automatic call processing [3], and also computer which can speak and recognize speech in native language [4].

II. Classification Of Speech Recognition Systems

Speech recognition systems can be separated in several different classes by describing the type of speech utterance, type of speaker model, type of channel and the type of vocabulary that they have the ability to recognize. Speech recognition is becoming more complex and a challenging task because of this variability in the signal. These challenges are briefly explained below.

A. Types of Speech Utterance

An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences. The types of speech utterance are:

1) Isolated Words: Isolated word recognizers usually require each utterance to have quiet on both sides of the sample window. It doesn't mean that it accepts single words, but does require a single utterance at a time. This is fine for situations where the user is required to give only one word responses or commands, but is very unnatural for multiple word inputs. It is comparatively simple and easiest to implement because word boundaries are obvious and the words tend to be clearly pronounced which is the major advantage of this type. The disadvantage of this type is choosing different boundaries affects the results.

2) Connected Words: Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allow separate utterances to be 'run-together' with a minimal pause between them.

3) Continuous Speech: Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it's computer dictation. It includes a great deal of "co articulation", where adjacent words run together without pauses or any other apparent division between words. Continuous speech recognition systems are most difficult to create because they must utilize special methods to determine utterance boundaries. As vocabulary grows larger, confusability between different word sequences grows.

4) Spontaneous Speech: This type of speech is natural and not rehearsed. An ASR system with spontaneous speech should be able to handle a variety of natural speech features such as words being run together and even slight stutters. Spontaneous (unrehearsed) speech may include mispronunciations, false-starts, and non-words.

B. Types of Speaker Model

All speakers have their special voices, due to their unique physical body and personality. Speech recognition system is broadly classified into two main categories based onspeaker models namely speaker dependent and speaker independent.

1) Speaker dependent models: Speaker dependent systems are designed for a specific speaker. They are generally more accurate for the particular speaker, but much less accurate for other speakers. These systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems.

2) Speaker independent models: Speaker independent systems are designed for variety of speakers. It recognizes the speech patterns of a large group of people. This system is most difficult to develop, most expensive and offers less accuracy than speaker dependent systems. However, they are more flexible.

C. Types of Vocabulary

The size of vocabulary of a speech recognition system affects the complexity, processing requirements and the accuracy of the system. Some applications only require a few words (e.g. numbers only), others require very large dictionaries (e.g. dictation machines). In ASR systems the types of vocabularies can be classified as follows.

i. Small vocabulary - tens of words
ii. Medium vocabulary - hundreds of words
iii. Large vocabulary - thousands of words
iv. Very-large vocabulary - tens of thousands of words
v. Out-of-Vocabulary- Mapping a word from the vocabulary into the unknown word

Apart from the above characteristics, the environment variability, channel variability, speaking style, sex, age, speed of speech also makes the ASR system more complex. But the efficient ASR systems must cope with the variability in the signal.

III. Growth Of Asr Systems

Building a speech recognition system becomes very much complex because of the criterion mentioned in the previous section. Even though speech recognition technology has advanced to the point where it is used by millions of individuals for using variety of applications. The research is now focusing on ASR systems that incorporate three features: large vocabularies, continuous speech capabilities, and speaker independence. The milestone of ASR system is given in the following table 1.

TABLE 1. GROWTH OF ASR SYSTEM

Year	Progress of ASR System
1952	Digit Recognizer
1976	1000 word connected recognizer with constrained grammar
1980	1000 word LSM recognizer (separate words w/o grammar)
1988	Phonetic typewriter
1993	Read texts (WSJ news)
1998	Broadcast news, telephone conversations
1998	Speech retrieval from broadcast news
2002	Rich transcription of meetings, Very Large Vocabulary, Limited Tasks, Controlled Environment
2004	Finnish online dictation, almost unlimited vocabulary based on morphemes
2006	Machine translation of broadcast speech
2008	Very Large Vocabulary, Limited Tasks, Arbitrary Environment
2009	Quick adaptation of synthesized voice by speech recognition (in a project where TKK participates in)
2011	Unlimited Vocabulary, Unlimited Tasks, Many Languages, Multilingual Systems for Multimodal Speech Enabled Devices
Future Direction	Real time recognition with 100% accuracy, all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent.

IV. Overview Of Automatic Speech Recognition (Asr) System

The task of ASR is to take an acoustic waveform as an input and produce output as a string of words. Basically, the problem of speech recognition can be stated as follows. When given with acoustic observation X = X1,X2…Xn, the goal is to find out the corresponding word sequence W = W1,W2…Wm that has the maximum posterior probability P(W|X) expressed using Bayes theorem as shown in equation (1). The following figure 2 shows the overview of ASR system.

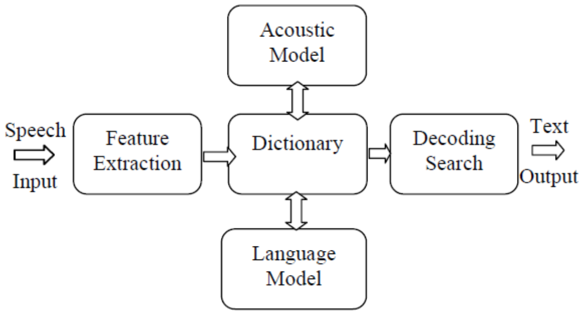

Figure 2.Overview of ASR system

$$W = \arg\max_{w} P(W / X) = \arg\max_{w} \frac{P(W)P(X / W)}{P(X)} \quad (1)$$

In order to recognize speech, the system usually consists of two phases. They are called pre-processing and post-processing. Pre-processing involves feature extraction and the post-processing stage comprises of building a speech recognition engine. Speech recognition engine usually consists of knowledge about building an acoustic model, dictionary and grammar. Once all these details are given correctly, the recognition engine identifies the most likely match for the given input, and it returns the recognized word.

V. 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.The speaker recognition system may be viewed as working in a four stages.

i. Analysis
ii. Feature extraction
iii. Modeling
iv. Testing/Matching techniques

A. SPEECH ANALYSIS

In speech analysis technique Speech data contains different types of information that shows a speaker identity. This includes speaker specific information due to vocal tract, excitation source and behavior feature. The physical structure and dimension of vocal tract as well as excitation source are unique for each speaker. The speech analysis deals with stages with suitable frame size for segmenting speech signal for further analysis and extracting [5]. The speech analysis is done with following three techniques.

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. Studies have been made in using segmented analysis to extract vocal tract information of speaker recognition.

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 mainly to analyze and extract the characteristic of the excitation state. [6]. The excitation source information is relatively fast varying compared to vocal tract information, so small frame size and shift are required to best capture the speaker-specific information [7].

3) Supra-segmental Analysis: In this case, speech is analyzed by using the frame size and shift of 100-300 ms to extract speaker information mainly due to behavioral tract and here speech is analyzed using the frame size. This technique is used mainly to analyze and characteristic due to behavior character of the speaker. These include word duration, intonation, speaker rate, accent etc.

B. SPEECH FEATURE EXTRACTION TECHNIQUES

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. The utterance can be extracted from a wide range of feature extraction techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:

i. Easy to measure extracted speech features
ii. It should not be susceptible to mimicry
iii. It should show little fluctuation from one speaking environment to another
iv. It should be stable over time
v. It should occur frequently and naturally in speech

The most widely used feature extraction techniques are explained below.

1) Linear Predictive Coding (LPC)

One of the most powerful signal analysis techniques is the method of linear prediction. LPC [8] [9] of speech has become the predominant technique for estimating the basic parameters of speech. It provides both an accurate estimate of the speech parameters and it is also an efficient computational model of speech. The basic idea behind LPC is that a speech sample can be approximated as a linear combination of past speech samples. Through minimizing the sum of squared differences (over a finite interval) between the actual speech samples and predicted values, a unique set of parameters or predictor coefficients can be determined. These coefficients form the basis for LPC of speech [10]. The predictor coefficients are therefore transformed to a more robust set of parameters known as cepstral coefficients. The following figure3 shows the steps involved in LPC feature extraction.

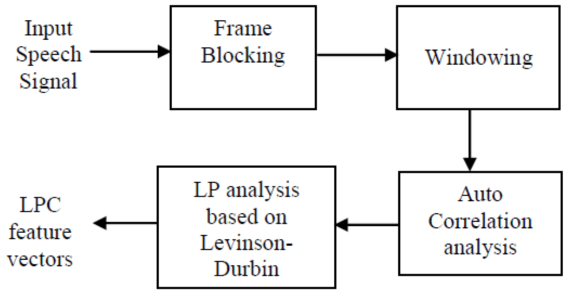
Input Speech Signal
Frame Blocking
Windowing
Auto Correlation analysis
LP analysis based on Levinson-Durbin
LPC feature vectors

Figure3. Steps involved in LPC Feature extraction

2) Mel Frequency Cepstral Coefficients (MFCC)

The MFCC [8] [9] is the most evident example of a feature set that is extensively used in speech recognition. As the frequency bands are positioned logarithmically in MFCC [11], it approximates the human system response more closely than any other system. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. In order to extract the coefficients the speech sample is taken as the input and hamming window is applied to minimize the discontinuities of a signal. Then DFT will be used to generate the Mel filter bank. According to Mel frequency warping, the width of the triangular filters varies and so the log total energy in a critical band around the center frequency is included. After warping the numbers of coefficients are obtained. Finally the Inverse Discrete Fourier Transformer is used for the cepstral coefficients calculation [8] [9]. It transforms the log of the quefrench domain coefficients to the frequency domain where N is the length of the DFT. MFCC can be computed by using the formula (2).

$$\text{Mel}(f) = 2595*\log10(1+f/700) \quad (2)$$

The following figure 4 shows the steps involved in MFCC feature extraction.

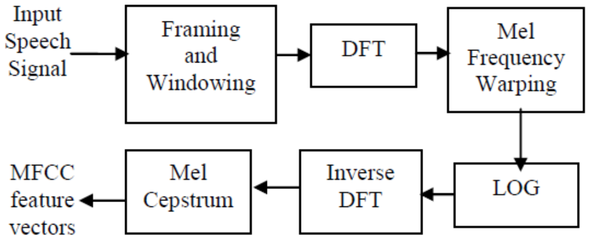

Figure4. Steps involved in MFCC feature extraction

C. 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 system can recognize the speaker, which has been trained with a number of speakers. Speaker recognition can also be dividing 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. Where as text independent does not rely on a specific texts being spoken [12]. Following are the modeling which can be used in speech recognition process:

1) 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 [13]. 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 [14].There are three techniques that have been applied to the language identification. Problem phone recognition, Gaussian mixture modeling, and support vector machine classification. [15][16]. The acoustic phonetic approach has not been widely used in most commercial applications [17].

2) 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 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[18].

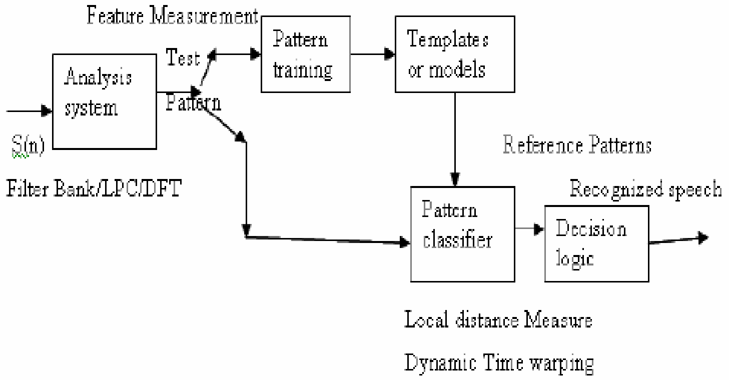

Figure5. Block diagram of Pattern recognition

3) 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. Recognition is 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. One key idea in template method is to derive a typical sequence 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. [19].

4) Dynamic Time Warping (DTW)

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 was 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. 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. Continuity is less important in DTW than in other pattern matching algorithms. This technique is quite efficient for isolated word recognition and can be modified to recognize connected word also[20].

5) Knowledge Based Approach Knowledge

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. Vector Quantization (VQ) [21] is often applied to ASR. It is useful for speech coders, i.e., efficient data reduction. 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. The test speech is evaluated by all codebooks and ASR chooses the word whose codebook yields the lowest distance measure [22]. 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. 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.

6) The Artificial Intelligence Approach

The Artificial Intelligence approach [23] 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. 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. A large body of linguistic and phonetic literature provided insights and understanding to human speech processing [24]. This knowledge is usually derived from careful study of spectrograms and is incorporated using rules or procedures. 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

7) Statistical Based Approach

In this approach, variations in speech are modeled statistically (e.g., HMM), using automatic learning procedures. This approach represents the current state of the art [25]. Modern general-purpose speech recognition systems are based on statistical acoustic and language models. Effective acoustic and language models for ASR in unrestricted domain require large amount of acoustic and linguistic data for parameter estimation. Processing of large amounts of training data is a key element in the development of an effective ASR technology now a days. The main disadvantage of statistical models is that they must make *a priori* modeling assumptions, which are liable to be inaccurate, handicapping the system's performance.

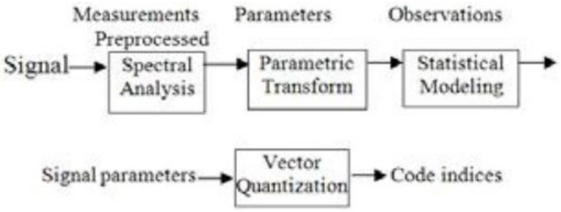

Figure6. Statistical models in speech recognition

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 as part of speech classification [26].

The K-means algorithm is also used for statistical and clustering algorithm of speech Based on the attribute of data .The K in Kmeans 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 [27].

8) Stochastic Approach

Stochastic modeling [28] 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 [29].

D. Matching Techniques

Speech-recognition engines match a detected word to a known word using one of the following techniques (Svendsen et al., 1989[30]).

1) 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 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 [31].

2) 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 [32] [33].

VI. PERFORMANCE OF SYSTEM

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) [34].

VII. 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 [35] [36]. 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.

The speed of a speech recognition system is commonly measured in terms of Real Time Factor (RTF). It takes time *P* to process an input of duration *I*. It is defined by the following formula.

$$RTF = \frac{P}{I}$$

The comparison of the various speech recognition research based on the dataset, feature vectors, and speech recognition technique adopted for the particular language are given in the table 2.

TABLE 2.COMPARISION OF VARIOUS SPEECH RECOGNITION APPLICATIONS BASED ON DATASET, FEATURE EXTRACTION AND RECOGNITION APPROACH

Author	**Year**	**Research Work**	**Nature of the Data**	**Feature Extraction Technique**	**Recognition Technique**	**Language**	**Accuracy**
Meysam Mohamad pour, Fardad Farokhi		Spoken digit recognition	Isolated Digit	Discrete Wavelet Transform (DWT)	Multilayer Perceptron + UTA algorithm	English	98%
Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda	2009	Automatic Speech Recognition for Bangia Digits	Small vocabulary Speaker independent Isolated digit	Mel-Frequency Cepstral Coefficients (MFCCs)	Hidden Markov Model (HMM)	Bangia	more than 95% for digits (0-5) and less than 90%

							for digits (6-9)
Corneliu Octavian Dumitru, Inge Gavat	2006	A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language	Large vocabulary Speaker independent Continuous speech	PLP, MFCC, LPC	Hidden Markov Models (HMM)	Romanian	MFCC-90,41% , LPC-63,55% . and PLP 75,78%
Douglas O'shaughnessy		Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis		LPC	HMM	English	Good acuuracy
Sid-Ahmed Selouani, Yousef Ajami Alotaibi	2003	Investigating Automatic Recognition of Non-Native Arabic Speech	Large vocabulary Speaker independent Phonetic/word	MFCC	HMM	Arabic	New words makes less accuracy for non-native speakers
Vimal Krishnan V.R Athulya Jayakumar Babu Anto.P	2008	Speech Recognition of Isolated Malayalam Words Using Wavelet Features and Artificial Neural Networ	Small vocabulary Speaker independent Isolated word	Discrete Wavelet Transform	Artificial Neural Network (ANN)	Malayalam	89%
Zhao Lishuang , Han Zhiyan	2010	Speech Recognition System Based on Integrating feature and HMM	Large vocabulary Speaker independent vowels	MFCC	Genetic Algorithm + HMM	Chinese	effective and high speed and accuracy
Bassam A. Q. Al-Qatab , Raja N. Ainon		Arabic Speech Recognition Using Hidden Markov Model Toolkit (HTK)		MFCC	HMM	Arabic	97.99%
Javed Ashraf , Dr Naveed Iqbal, Naveed Sarfraz Khattak, Ather Mohsin Zaidi		Speaker Independent Urdu Speech Recognition Using HMM	Small vocabulary Speaker independent Isolated word	MFCC	Hidden Markov Model	Urdu	Little variation in WER for new speakers

N.Uma Maheswari, A.P.Kabilan, R.Venkatesh		A Hybrid model of Neural Network Approach for Speaker independent Word Recognition		LPC	Hybrid model of Radial Basis Function and the Pattern Matching method	English	91%
Raji Sukumar.A Firoz Shah.A Babu Anto.P	2010	Isolated question words Recognition from speech queries by Using artificial neural networks	Medium vocabulary Speaker dependent Isolated word	D Γ	ANN	Malayalam	80%
R. Thangarajan, A.M. Natarajan and M. Selvam		Phoneme Based Approach in Medium Vocabulary Continuous Speech Recognition in Tamil language	Medium vocabulary Speaker independent Continuous Speech	MFCC	Hidden Markov Model (HMM)	Tamil	good word accuracy for trained and test sentences read by trained and new speakers
A.Rathinavelu, G.Anupriya, A.S.Muthananthamurugavel	2007	Speech Recognition Model for Tamil Stops	Small vocabulary Speaker independent phonems	first five formant values	Feed forward neural networks	amil	81%
M. Chandrasekar, and M.Ponnavaikko	2008	Tamil speech recognition: a complete model	Medium vocabulary Speaker dependent Isolated Speech	MFCC	Back Propagation Network	Tamil	80.95%
A.P.Henry Charles1 & G.Devaraj2	2004	Alaigal-A Tamil Speech Recognition	Speaker independent Continuous speech	MFCC	HMM	Tamil	Offers High Performance

VIII. CONCLUSION

In this review, the fundamentals of speech recognition are discussed and its recent progress is investigated. The various approaches available for developing an ASR system are clearly explained with its merits and demerits. The performance of the ASR system based on the adopted feature extraction technique and the speech recognition approach for the particular language is compared in this paper. In recent years, the need for speech recognition research based on large vocabulary speaker independent continuous speech has highly increased. Based on the review, the potent advantage of HMM approach along with MFCC features is more suitable for these requirements and offers good recognition result. These techniques will enable us to create increasingly powerful systems, deployable on a worldwide basis in future.

REFERENCES

[1] Bassam A. Q. Al-Qatab, Raja N. Ainon, “Arabic Speech Recognition Using Hidden Markov Model Toolkit(HTK)”, 978-1- 4244-6716-7110/$26.00©2010 IEEE.

[2] Dat Tat Tran, “Fuzzy Approaches to Speech and Speaker Recognition”, a thesis submitted for the degree of PhD of the University of Canberra.

[3] R.Klevansand R.Rodman, “voice Recognition, Artech House, Boston, London 1997.

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

[5] 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

[6] 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

[7] B. Yegnanarayana, S.R.M. Prasanna, J. M. Zachariah, and C.S. Gupta, "Combining evidence from source, suprasegmental and spectral features for a fixed- text specker verification system," IEEE Trans. Speech Audio Process., vol. 13(4), pp. 575-82, July 2005.

[8] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June 2006, Zadar, Croatia.

[9] DOUGLAS O'SHAUGHNESSY, "Interacting With Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, VOL. 91, NO. 9, September 2003, 0018-9219/03$17.00 © 2003 IEEE.

[10] N.Uma Maheswari, A.P.Kabilan, R.Venkatesh, "A Hybrid model of Neural Network Approach for Speaker independent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6, December, 2010 1793-8201.

[11] A.P.Henry Charles & G.Devaraj, "Alaigal-A Tamil Speech Recognition", Tamil Internet 2004, Singapore.

[12] Sannella, M Speaker recognition Project Report report" From http://cs.joensuu.fi/pages/tkinnu/research/index.html Viewed 23 Feb. 2010.

[13] IBM (2010) online IBM Research Source:-http://www.research.ibm.com/Viewed 12 Jan 2010.

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

[15] 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".

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

[17] 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

[18] 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.

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

[20] Santosh K.Gaikwad, Bharti W.Gawali and Pravin Yannawar, "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887) Volume 10– No.3, November 2010

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

[22] 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

[23] R.K.Moore, Twenty things we still don t know about speech, Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research an Technology, 1994.

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

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

[26] 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

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

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

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

[30] 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 .

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

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

[33] 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.

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

[35] 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.

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