Sanjib Das / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.2071-2087

Speech Recognition Technique: A Review

Sanjib Das
Department of Computer Science, Sukanta Mahavidyalaya, (University of North Bengal), India

ABSTRACT

Speech is the primary, and the most convenient means of communication between people. 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 and demerits. A comparative study of different technique is done as per stages. This paper concludes with the decision on feature direction for developing technique in human computer interface system in different mother tongue and 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. The objective of this review paper is to 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 **– Analysis, ASR, Feature Extraction, Modeling, Testing**

I. Introduction

Speech Recognition is also known as Automatic Speech Recognition (ASR), or computer speech recognition 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]. Generally, machine recognition of spoken words is carried out by matching the given speech signal against the sequence of words which best matches the given speech sample [2]. The main goal of speech recognition area is to develop techniques and systems for speech input to machine. Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to desire to automate simple tasks necessitate human machine interactions. The research in ASR by machines has attracted a great deal of attention for about sixty years [3] and ASR today finds widespread application in tasks that require human machine interface, such as automatic call processing [4]. India is a linguistically rich area which has 18 constitutional languages written in 10 different scripts [5]. Hence there is a special need for the ASR system to develop in different native languages [6].

1.1 ASR System Classification

Speech Recognition is a special case of pattern recognition. There are two phases in supervised pattern recognition, viz., Training and Testing. The process of extraction of features relevant for classification is common in both phases. During the training phase, the parameters of classification model are estimated by 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.

1.2 Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize.

1.2.1 Isolated Word

Isolated word-recognizers usually require each utterance to exit 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 having "Listen and Non-listen state". 'Isolated Utterance' might be better name for this class [7]. 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 are the major advantages of this type. The disadvantage of this type in choosing different boundaries affects the results.

1.2.2 Connected Word

Connected word systems are similar to isolated words but allow separate utterance to be 'run-together' with a minimal pause between them.

1.2.3 Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. Basically, it is computer dictation [8]. Recognizers with continuous speech capabilities are some of the most

difficult job to create because they utilize special methods to determine utterance boundaries. As vocabulary grows larger, confusability between different word sequences grows.

1.2.4 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, "ums" and "ahs" and even slight stutters [8]. Spontaneous or unrehearsed speech may include mispronunciations, false-starts, and non-words.

1.3 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 main categories based on speaker models, namely, speaker dependent and speaker independent.

1.3.1 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 others speakers. This systems are usually easier to develop, cheaper and more accurate, but not as flexible as speaker adaptive or speaker independent systems.

1.3.2. Speaker independent models

Speaker independent system 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.

1.4 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. direction machines). In ASR systems the types of vocabularies can be classified as follows.

- ✓ Small vocabulary - ten of words
- ✓ Medium vocabulary - hundreds of words
- ✓ Large vocabulary – thousands of words
- ✓ Very-large vocabulary – tens of thousands of words
- ✓ Out-of-Vocabulary – Mapping a word from the vocabulary into the unknown word

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

1.5. Basic Principles of ASR

All ASR systems operate in two phases. First, a training phase, during which the system learns the reference patterns representing the different speech sounds (e.g. phrases, words, phones) that constitute the vocabulary of the application. Each reference is learned from spoken examples and stored either in the form of templates obtained by some averaging method or models that characterize the statistical properties of pattern. Second, a recognizing phase, during which an unknown input pattern, is identified by considering the set of references.
The Speak-recognizer process is shown below (Fig: 1).

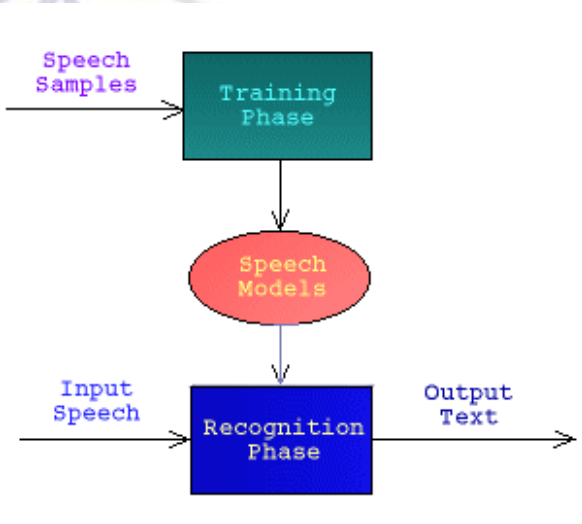

Fig: 1 Basic Principle of Speak-recognizer

Most ASR systems consist of three major modules i.e. signal processing front-end, acoustic modeling and language modeling. The signal processing front-end transforms the speech signal into a sequence of feature vectors to be used for classification. Generally, this representation has a considerably lower information rate than the original speech waveform.

1.5.1 Growth of ASR Systems

Recent years have seen a substantial growth in the deployment of practical systems for 'automatic speech recognition' (ASR). These ongoing commercial successes are a direct result of a significant increase in the capabilities of ASR devices over the past thirty years driven by both improvements in the underlying ASR algorithms and the relentless increase in available computer power. 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. Today, there are various systems which incorporate these combinations. However, with these numerous

technological barriers in developing ASR system, still it has reached the highest growth. 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.

1.5 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 $X = X1,X2\ldots X_n$, the goal is to find out the corresponding word sequence $W = W1,W2\ldots W_n$ that has the maximum posterior probability P(W|X) expressed using Bayes theorem as shown in equation (1). The following figure 1: shows the overview of ASR system.

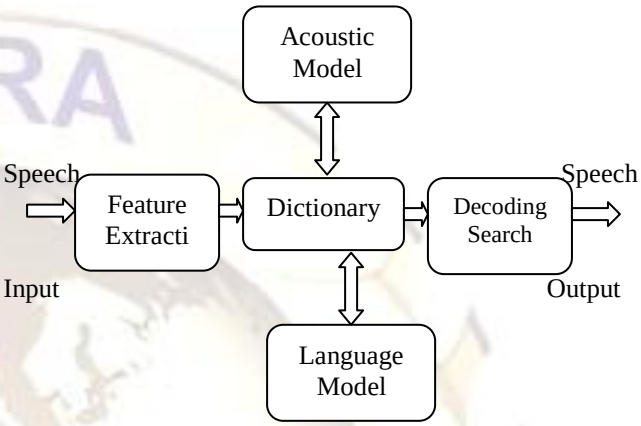

Figure 1: Overview of ASR system

1. 7 Basic Model of Speech Recognition:

Research in speech processing and communication for the most part, was motivated by people's desire to build mechanical models to emulate human verbal communication capabilities. Speech is the most natural form of human communication, and the speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages. Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human machine interface, such as automatic call processing in telephone networks, and query based information systems that provide updated travel information, stock price quotations, weather reports, data entry, voice dictation, access to information: travel, banking, Commands, Avoinics, Automobile portal, speech transcription, handicapped people (blind people) supermarket, railway reservations etc. Speech recognition technology was increasingly used within telephone networks to automate as well as to enhance the operator services. This report reviews major highlights during the last six decades in the research and development of automatic speech recognition, so as to provide a technological perspective. Although many technological progresses have been made, still there remains many research issues that need to be tackled.

Fig.2 shows a mathematical representation of speech recognition system in simple equations which contain front

end unit, model unit, language model unit, and search unit. The recognition process is shown below (Fig: 2).

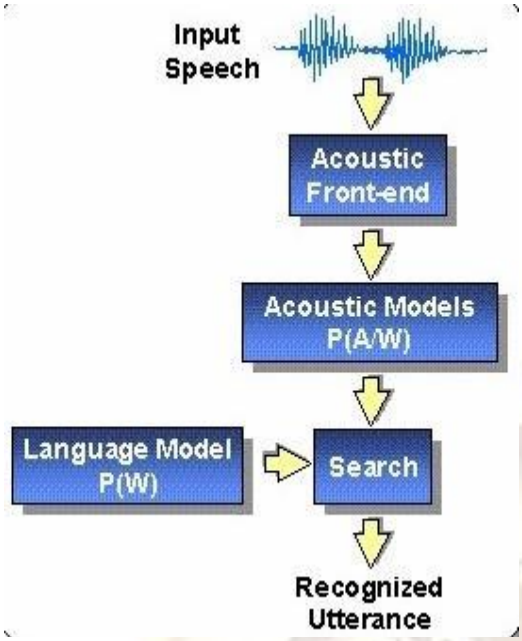

Fig.: 2 Basic model of speech recognition

The standard approach to large vocabulary continuous speech recognition is to assume a simple probabilistic model of speech production whereby a specified word sequence, W, produces an acoustic observation sequence Y, with probability P(W,Y). The goal is then to decode the word string based on the acoustic observation sequence, so that the decoded string has the maximum *a posteriori* (MAP) probability.

$$\hat{P}(W/A) = \arg\max_{W} P(W/A) \quad \text{.....(1)}$$

Using Baye's rule, equation (1) can be written as

$$P(W/A) = \frac{P(A/W)P(W)}{P(A)} \quad \text{.....(2)}$$

Since P(A) is independent of W, the MAP decoding rule of equation(1) is

$$\hat{W} = \arg\max_{W} P(A/W)P(W) \quad \text{..... .. (3)}$$

The first term in equation (3) P(A/W) is generally called the acoustic model, as it estimates the probability of a sequence of acoustic observations, conditioned on the word string. Hence P(A/W) is computed. For large vocabulary speech recognition systems, it is necessary to build statistical models for sub word speech units and to build up word models from these sub-word speech unit models (using a lexicon to describe the composition of words), and then postulate word sequences and evaluate the acoustic model probabilities via standard concatenation methods. The second term in equation (3) P(W), is called the language model. It describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints of the language and the recognition task.

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. The goal of automatic speaker recognition 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

- ✓ Analysis
- ✓ Feature extraction
- ✓ Modeling
- ✓ Testing

2.1 Speech analysis

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. This uniqueness is embedded in the speech signal during speech production and can be used for speaker used for speaker recognition. The behavioral tracts as to how the vocal tract and excitation source are controlled during speech production are also unique for each user. The information about behavioral tracts is also embedded in the speech signal and can be used for speaker recognition. The information about the behavior feature also embedded in signal and that can be used for speaker recognition. The speech analysis deals with stages with suitable frame size for segmenting speech signal for further analysis and extracting [9]. The speech analysis is 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. Studies have been made in using 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 mainly to analyze and extract the characteristic of the

excitation state. [10]. 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 [11-15].

2.1.3 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. The behavioral tracts vary restively slowly compares to the vocal tract information, which is the reason for the choice of large frame size and shift [11, 16-18].

2.1.4 Performance of System

The performance of speaker recognition system depends on the technique employed in the various stages of speaker recognisation system. The state of the art speaker recognition system mainly used in 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 recognition field. Other techniques may also have to be used for resulting a good speaker recognition performance. Some of practical issues are as follows:

2.1.4.1. Non-acoustic sensor provides an exciting opportunity for multimodal speech processing with application to areas, such as speech enhancement and coding. This sensor provides measurement of function of the glottal excitation and can supplement acoustic waveform.

2.1.4.2. A Universal Background Model (UBM) is a model used in a speaker verification system to represent general person independent of the feature characteristics to be compared against a model of person specific feature characteristics while accepting or rejecting a decision.

2.1.4.3. A Multi-model person recognition 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 [19].

2.1.4.4. Many powerful models for speaker recognition have been introduced in high level features, novel classifiers and channel compression methods [20].

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.

2.1.4.6. A recent area of significant progress in speaker recognition is the use of high level features-idiolect, phonetic relations, prosody. A speaker possesses distinctive acoustic sound and also uses language in a characteristic manner. [21]

2.2 Feature Extraction Technique

Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other. Because every speech has different individual characteristics embedded in utterances. These characteristics 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:

- ✓ Easy to measure extracted speech features
- ✓ It should not be susceptible to mimicry
- ✓ It should show little fluctuation from one speaking environment to another
- ✓ It should be stable over time
- ✓ It should occur frequently and naturally in speech

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. The purpose of feature extraction stage is to extract the speaker-specific information in the form of feature vectors. The feature vectors represent the speaker-specific information due to one or more of the following: Vocal tract, excitation source and behavioral tracts. A good feature set should have representation due to all of the components of speaker information. Just as a good feature set is required for a speaker, it is necessary to understand the different feature extraction techniques. This section describes the same. Spoken digit recognition conducted by P Denes in 1960 suggested that inter-speaker differences exist in the spectral patterns of speakers [22]. S Pruzansky, motivated from this study, conducted the first speaker identification study in 1963. In his study, spectral energy patterns were used as the features. It was shown that the spectral energy patterns yielded good performance, confirming the usefulness for the speaker recognition [23]. Further, he reported a study using the analysis of variance in 1964 [24]. In this work, a subset of features was selected from the analysis of variance using F ratio test defined as the ratio of the variance of the speaker means to average within speaker variance [24]. It was reported that the subset of features provided equal performance, thus significantly reducing the number of computations. Speaker verification study was first conducted by Li in 1966 using adaptive linear threshold elements [25]. This study used spectral representation of the input speech, obtained from the bank of 15 band pass filters spanning the frequency range 300-4000Hz. Two stages of adaptive linear threshold elements operate on the rectified and

smoothed filter outputs. These elements are trained with speech utterances. The training process results in a set of weights that characterize the speaker. This study demonstrated that the spectral band energies as feature contain speaker information. A study by Glenn in 1967 suggested that acoustic parameters produced during nasal phonation are highly effective for speaker recognition [26]. In this study, average power spectral of nasal phonation was used as the features for the speaker recognition. In 1969, Fast Fourier Transform (FFT) based cepstral coefficients were used in speaker verification study. In this work, a 34-dimensional vector was extracted from speech data. The first 16 components were from FFT spectrum, the next 16 were from log magnitude FFT spectrum and the last two components were related to pitch and duration. Such a 34-dimensional vector seems to provide a good representation of speaker.

In 1972, Atal demonstrated the use of variations in pitch as a feature for speaker recognition. In addition to the variation in pitch, other acoustic parameters, such as glottal source spectrum slope, word duration and voice onset were proposed as features for speaker recognition by Wolf in 1971 [27]. The concept of linear prediction for speaker recognition was introduced by Atal in 1974 [28]. In this work, it was demonstrated that Linear Prediction Cepstral Coefficients (LPCCs) were better than the Linear Prediction Coefficients (LPCs) and other features, such as pitch and intensity.

Earlier studies neglected the features, such as formant bandwidth, glottal source poles and higher formant frequencies, due to non-availability of measurement techniques. The studies introduced after the linear prediction analysis explored the speaker specific potential of these features for speaker recognition [29]. A study carried by Rosenberg and Sambur suggested that adjacent cepstral coefficients are highly correlated and hence all coefficients may not be necessary for speaker recognition [30]. In 1976, Smbur proposed to use orthogonal linear prediction coefficients as feature in speaker identification [31]. In this work, he pointed out that for a speech feature to be effective, it should reflect the unique properties of the speaker's vocal tract and contain little or no information about linguistic content of the speech. In 1977, long term parameter averaging, which includes pitch, gain and reflection coefficients for speaker recognition was studied [32]. In this study, it was shown that reflection coefficients are informative and effective for speaker recognition. In 1981 Furui introduced the concept of dynamic features to track the temporal variability in feature vector in order to improve the speaker recognition performance [33, 34]. A study made by G R Doddington in 1985 converts the speech directly into pitch, intensity and formant frequency, all sampled 100 times per second. These features were also demonstrated to provide good performance.

A study by Reynolds in 1994 compared the different features like Mel Frequency Cepstral Coefficients (MFCCs), Linear Prediction Cepstral Coefficients (LPCCs), LPCCs and Perceptual Linear Prediction Cepstral Coefficients (PLPCCs) for speaker recognition [35]. He reported that among these features, MFCCs and LPCCs gave better performance than other features. In 1995 P. The Venaz and H Hugli [36] reported that Linear Prediction (LP) residual also contains speaker-specific information that can be used for speaker recognition. Also, it has been reported that though the energy of LP residual alone gives less performance, combining it with LPCC improves the performance as compared to that of LPCC alone. Similarly, several studies reported that though the energy of LP residual alone gives less performance, combining it with MFCC improves the performance as compared to that of MFCC alone. In 1996, Plumpe developed a technique for estimating and modeling the glottal flow derivative waveform from speech for speaker recognition [37]. In this study, the glottal flow estimate was modeled as coarse and fine glottal features, which were captured using different techniques. Also it was shown that combined course and fine structure parameters gave better performance than the individual parameter alone. In 1996, M J Carey, E S Paris carried out a study on the significance of long term pitch and energy information for speaker recognition [38]. In 1998, M K Sonmez, E Sriberg carried out a study on pitch-tracks and local dynamics for speaker verification [39].

In 2003, B Peskin, J Navratil reported that combination of prosodic features like long-term pitch with spectral features provided significant improvement as comapared to only pitch features [40]. A study by L Mary, K S Rao, B Yegnanarayana in 2004 were carried out on supra-segmental features like duration and intonation capyurd using neural network for speaker recognition. In 2005, B Yegnanarayana, S R M Prasanna demonstrated the use of features such as long term pitch and duration information obtained using Dynamic Time Warping (DTW), along with source and spectral features for text-dependent speaker recognition. In 2008, M Girmaldi, F Cummins carried out a study on Amplitude Modulation (AM)-Frequency Modulation (FM)-based parameter of speech for speaker recognition. In this study, it was demonstrated that using different instantaneous frequencies due to the presence of formants and harmonics in speech signal, it is possible to discriminate speakers [41].

In 2007, Min-Seok Kim and Ha-Jin Yu introduced a new feature transformation method based on rotation for speaker identification [42]. In this study, they have proposed a new feature transformation method that is optimized for diagonal covariance Gaussian mixture models [43] which is used for a speaker identification system. They first have defined an object function as the distances between the Gaussian mixture components and rotate each plane in the feature space to maximize the object function. The optimal degrees of the rotations are found using the Particle Swarm Optimization [44] algorithm. In 2008, Min-Seok Kim, IL-Ho Yung and Ha-Jin Yu have proposed a feature transformation method to maximize the distance between the Gaussian mixture models for speaker verification using PSO [45].

The different feature extraction techniques described above may be summarized as follows:

- ✓ Special features like band energies, formants, spectrum and cepstral coefficients representing mainly the speaker-specific information due to the vocal tract.

- ✓ Excitation source features like pitch, variations in pitch, information from LP residual and glottal source parameters.
- ✓ Long-term features like duration, intonation, energy, AM and FM components representing mainly the speaker-specific information due to the behavioral traits.

Among these the most commonly used cepstral coefficients are MFCCs and LPCCs, because of less intra-speaker variability and also availability of spectral analysis tools. However, the speaker-specific information due to excitation source and behavioral tract represents different aspects of speaker information. The main limitation for the use of excitation source and behavioral tract is non – availability of suitable feature extraction tools.

2.3 Speaker Modeling Technique

The objective of modeling technique is to generate speaker models using speaker specific feature vector. The speaker modeling technique divided into two classifications: 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 recognition is also divided into two parts that means speaker dependant and speaker independent. In the speaker independent mode of the speech recognition the computer should ignore the speaker specific characteristics of the speech signal and extract the intended message, on the one hand. On the other, in case of speaker recognisation machine should extract speaker characteristics in the acoustic signal [46]. 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 divided into two methods, text-dependent and text-independent methods. In text-dependent method, the speaker says 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 [47]. Following are the modeling which can be used in speech recognition process:

2.3.1 The acoustic-phonetic approach

The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. 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 [48]. 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 [49]. 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[50]. Phone-based systems are described in [51] and [52]. There are three techniques that have been applied to the language identification: Problem phone recognition, Gaussian mixture modeling, and support vector machine classification. [53][54]. Using IPA Methods we can find similarities for probabilities of content dependant acoustic model for new language.[55]. The acoustic phonetic approach has not been widely used in most commercial applications [56].

2.3.2 Pattern Recognition Approach

Speech recognition is one in which the speech patterns 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: Systems' 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. A block schematic diagram of pattern recognition is presented in fig. 3: below. In this, there exist two methods, namely, template approach and stochastic approach.

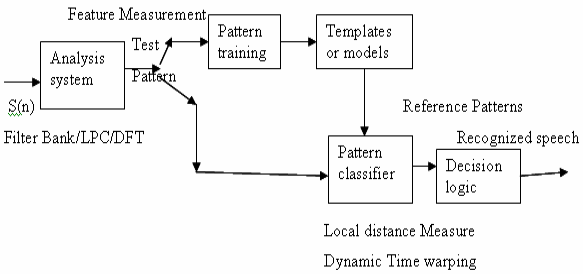

Fig 3. Block diagram of Pattern recognition speech recognizer

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 decades received much attention and applied widely to many practical pattern recognition problems [56]. 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 ([57] p. 87).

2.3.3 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. Template based approach [58][59] 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 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. 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 [60].

2.3.4. 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 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. 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, and hence the name.

Dynamic Time Warping is an algorithm for measuring similarity between two sequences which may vary in time or speed [61]. In general, it is a method that allows a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions, i.e. the sequences are "warped" non-linearly to match each other. This sequence alignment method is often used in the context of HMM. This technique is quite efficient for isolated word recognition and can be modified to recognize connected word also [61].

2.3.5 The Artificial Intelligence Approach

The Artificial Intelligence approach [62] 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. Expert system is used widely in this approach (Mori et al., 1987) [63] [64]. 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, and thereby making error- analysis and knowledge-based system enhancement difficult. A large body of linguistic and phonetic literature provided insights and understanding to human speech processing [65]. 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 but 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.

2.3.6 Knowledge Based Approach

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. 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)[66] 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 [67]. 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 enable 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.

2.3.7 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. 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 nowadays. 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.

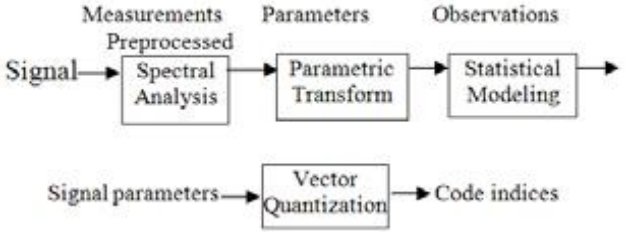

Fig. 4: Statistical Models in Speech Recognition

In which variations in speech are modeled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models, or HMM. These approaches represent the current state of the art. The main disadvantage of statistical models is that they must take *a 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 [68]; Ma and Deng, 2004 [69]).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 [70].

A weighted hidden markov model HMM algorithm and a subspace projection algorithm are proposed in [71], to address the discrimination and robustness issues for HMM based speech recognition. Word models were constructed for combining phonetic and fenonic models [71] A new hybrid algorithm based on combination of HMM and learning vector were proposed in [70]. Learning Vector Quantization [71] (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 [72] proposed a corrective training method that minimized the number of errors of parameter estimation. A novel approach [73] 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[73] 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[74].An extension of the viterbi algorithm[75] made the second order HMM computationally efficient when compared with the existing viterbi algorithm. In this paper[76] 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 model VQ [77] in which the phoneme is treated as a cluster in the speech space and Gaussian Model were 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 [78]. 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 has been demonstrated by [79]. A new hidden Markov model [77] 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 algorithm 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 [80].

The process of K-means algorithm uses:

- ✓ Least-squares partitioning method to divide the input vectors into k initial sets.
- ✓ 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.
- ✓ After that the re-evaluation of all the centroids are performed for all the possible new clusters.
- ✓ 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. The most well-known codebook generation algorithm is the K-means algorithm. In 1985, Soong et al. [81] used the LBG algorithm for generating speaker-based vector quantization (VQ) codebooks for speaker recognition. It is demonstrated that larger codebook and test data give good recognition performance. Also, the study suggested that VQ codebook can be updated from time to time to alleviate the performance degradation due to different recording conditions and intra-speaker variations [81]. The disadvantages of the VQ classification is that it ignores the possibility that a specific training vector may also belong to another cluster. As an alternative to this, fuzzy vector quantization (FVQ) using the well-known fuzzy C-means method was introduced by Dunn, and its final form was developed by Bezdek [82] [83]. In [84] and [85], FVQ was used a classifier for speaker recognition. It was demonstrated that FVQ gives better performance than the traditional *K-means* algorithm. This is because the working principle of FVQ is different from VQ, in the sense that the soft decision- making process is used while designing the codebooks in FVQ [82]; whereas in VQ, the hard decision process is used. Moreover, in VQ each feature vector has an association with only one of the clusters, there is relatively more number of feature vectors for each cluster; and hence the representative vectors, viz., *code-vectors*, may be more reliable than VQ. Therefore, clustering may be better performance compared to VQ.

In order to model the statistical variations, the Hidden Markov Model (HMM) for text-dependent parameters are observation symbols. Observation symbols are created by VQ codebook levels. Continuous probability measures are created using Gaussian Mixtures Models (GMMs). The main assumption of HMM is that the current state depends on the previous state. In training phase, state transition probability distribution, observation symbol probability distribution and initial state probabilities are estimated for each speaker as a speaker model. The probability of observations for a given speaker model is calculated for speaker recognition. Kimbal *et al.* studied the use of HMM for text-dependent speaker recognition under the constraint of limited data and mismatched channel conditions [86-89]. In this study the MFCC feature was extracted for each speaker and then models were built using the Board Phonetic Category (BPC) and the HMM-based Maximum Likelihood Leaner Regression (MLLR) adaptation technique. The BPC modeling is based on identification of phonetic categories in an utterance and modeling them separately. In HMM-MLLR, first, speaker independent (SI) model is created using HMM, and then MLLR technique is used to adapt SI model to each speaker. It was shown that the speaker model built using the adaptation technique gave better performance than the BPC and GMM for cross-channel conditions.

The capability of neural networks to discriminate between patterns of different classes is exploited for speaker recognition [90][91][92]. Neural network has an input layer, one or more hidden layers and an output layer. Each layer consists of processing units, where each unit represents model of an artificial neuron, and the interconnection between the two units as a weight associated with it. The concept of multi-layer perception (MLP) was used for speaker recognition in [93]. In this study, it was demonstrated that one- hidden layer network with 128 hidden nodes gave the same performance as that achieved with the 64 codebook VQ approach. The disadvantage of MLP is that it takes more time for training network. The problem was alleviated using the radial basis

function (RBF) network took lesser time than the MLP and outperformed both VQ and MLP.

Kohonen developed self organization map (SOM) as an unsupervised learning classifier. SOM is a special class of neural network based on competitive learning [94][95]. Thus the performance of SOM depends on the parameters, such as neighborhood, learning rate and number of iterations. These parameters are to be fine-tuned for good performance. The SOM and associative memory model were used together as a hybrid model for speaker identification in [96]. It was shown that the hybrid model gave better recognition performance than the MLP. A text-independent speaker recognition system based on SOM neural networks also suited in [97]. The disadvantage of SOM is that it does not use class information while modeling speakers, resulting in a poor speaker model that leads to degradation in the performance. This can be alleviated by using Kohonen learning vector quantization (LVQ) [65]. LVQ is a supervised learning technique that uses class information to optimize the positions of code vectors obtained by SOM, so as to improve the quality of the classifier-design regions. An input vector is picked at random from the input space. If the class lebel of the input vector and the code-vector agree, then the code-vector is moved away from the input vector. Due to this fine-tuning, there may be improved recognition rate compared to SOM. LVQ was proposed for speaker recognition in [98]. Speaker recognition using VQ, LVQ and GVQ (GroupVector Quantization) was demonstrated for YOHO database in [99]. The experimental results show that LVQ gives better performance when the data is small, as compared to the traditional VQ and proposed GVQ; but GVQ yields better recognition performance when the size is large.

2.3.8 Artificial Neural Networks (ANN)

ANN is used to classify speech samples in the intelligent ways as shown in the figure 6.

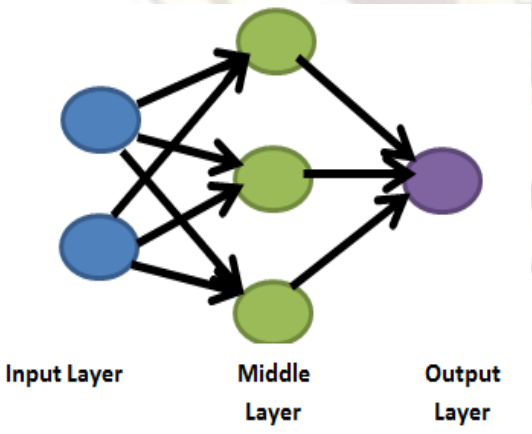

Fig. 6: Simplified view of an artificial neural network

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 [100] 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. 6: 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. 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). The Artificial Intelligence approach 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. 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.

2.3.9 Hybrid Model (HMM/NN)

In many speech recognition systems, both techniques are implemented together and work in a symbiotic relationship [101]. 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.

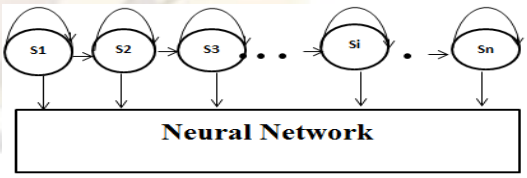

Fig. 7: n-state Hybrid HMM Model

2.3.10. 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 and they can be learned automatically through emulations or evolutionary process.

2.3.11. Matching Techniques

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

2.3.11.1 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 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 [102].

2.3.11.2. 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.[103] [104] discuss that research in the area of automatic speech recognition had been pursued for the last three decades.

3. Pronunciation modeling techniques

As mentioned in the introduction of Section 2, some speech variations, like foreign accent or spontaneous speech, affect the acoustic realization to the point that their effect may be better described by substitutions and deletion of phonemes with respect to canonical (dictionary) transcriptions. As a complementary principle to multiple acoustic modeling approaches reviewed in Section 4.2.2, multiple pronunciations are generally used for the vocabulary words. Hidden model sequences offer a possible way of handling multiple realizations of phonemes (Hain and Woodland, 1999) possibly depending on phone context. For handling hyper articulated speech where pauses may be inserted between syllables, ad hoc variants are necessary (Matsuda et al., 2004). And adding more variants is usually required for handling foreign accents. Modern approaches attempt to build in rules underlying pronunciation variation, using representations frameworks such as FSTs (Hazen et al., 2005; Seneff and Wang, 2005), based on phonological knowledge, data and recent studies on the syllabic structure of speech, for instance in English (Greenberg and Chang, 2000) or French (Adda-Decker et al., 2005).

4. 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)[105].

5. 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 [106][107]. 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.

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

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 2 lists the words which are spoken by all 100 speakers and stored in the database.

Table 2: 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. 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 3 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 3: 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 and Future Works

This paper gathers important references to literature related to the endogenous variations of the speech signal and their importance in automatic speech recognition. Important references addressing specific individual speech variation sources are first surveyed. This covers accent, speaking style, speaker physiology, age, emotions. General methods for diagnosing weaknesses in speech recognition approaches are then highlighted. Finally, the paper proposes an overview of general and specific techniques for better handling of variation sources in ASR, mostly tackling the speech analysis and acoustic modeling aspects. In this review, I have discussed the technique developed in each stage of speech recognition system. I also presented the list of techniques 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 are best among all modeling techniques. I have also 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. I hope this paper would bring about understanding and inspiration amongst the research communities of ASR.

8. Acknowledgements

The author remains thankful to Prof. Manoranjan Singha and Dr. Benulal Dhar, University of North Bengal, for their useful discussions and suggestions during the preparation of this technical paper.

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] 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.
[3] Dat Tat Tran, "Fuzzy Approaches to Speech and Speaker Recognition", A thesis submitted for the degree of PhD of the university of Canberra.
[4] R.Klevansand R.Rodman, "voice Recognition, Artech House, Boston, London 1997.

[5] M. Chandrasekhar, M. Ponnavaikko, "Tamil speech recognition: a complete model", electronic Journal << Technical acoustics>> 2008, 20.
[6] Samudravijaya K. Speech and speaker recognition tutorial TIFR Mumbai 400005.
[7] Zahi N.Karam, William M.Campbell "A new Kernel for SVM MIIR based Speaker recognition" MIT Lincoin Laboratory, Lexington, MA, USA.
[8] Fundamentals of Speech Recognition by Lawrence Rabiner & Bing-Hwang Juang.
[9] 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
[10] 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
[11] 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.
[12] K.S.R. Murthy, and B. Yegnanarayana, "Combining evidence from residue phase and MFCC features for specker recognition ," IEEE Trans. Signal Process. Lett. , vol. 13(1), pp. 52-6, Jan. 2006.
[13] B. Yegnanarayana, K. Sharat Reddy, and S.P. Kishore, "Source and system features for specker recognition using AANN models ," in proc. Int. Conf. Acoust., Speech, Signal Process. , Utah, USA, Apr. 2001.
[14] K. Sharat Reddy, "Source and system features for specker recognition ," Master's thesis, Indian Institute of Technology Madras, Dept. of computer Science anf Engg., Chennai, India, 2003.
[15] C.S. Gupta, " Significance of source features for specker recognition ," Master's thesis, Indian Institute of Technology Madras, Dept. of computer Science anf Engg., Chennai, India, 2003.
[16] B.S. Atal, "Automatic specker recognition based on pitch contours ," J. Acoust. Soc. Amer., vol. 52, no. 6(part 2), pp. 1687-97, 1972.
[17] L. Mary, K.S. Rao ,S.V. Gangashetty, and B. Yegnanarayana, "Neural networls model for capturing duration anf intination knowlwdge for language and speaker identification," in Proc. Int.Conf. Cognitive Neural Syatem,Boston,Massachusetts, May 2004.
[18] F.Farahani, P.G. Georgiou, and S.S. Narayanan, "Speaker identification using suprasegmental pitch patterns dynamics," in Proc. Int.Conf. Acoust. , Speech Signal Process., Montreal, Canada, May 2004, pp.89- 92.
[19] 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
[20] 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.
[21] 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.
[22] P. Denes, and M.V. Mathews, "Spoken digit recognition using time-frequency pattern matching," J. Acoust. Soc. Amer., vol. 32(11), pp. 1450-5, Nov. 1960.
[23] S.Pruzansky, "Pattern-matching procedure for automatic talker recognition," J. Acoust. Soc. Amer., vol. 35(3), pp. 354-8, Mar. 1963.
[24] S.Pruzansky and M.V. Mathews, "Talker-recognition procedure based on analysis of varuance," J. Acoust. Soc. Amer., vol. 36(11), pp. 2041-7-8, Nov. 1964.
[25] K.P Li, J. E. Dammann, and W.D. Chapman, "Experimental studies in speaker verification using an adaptive system," J. Acoust. Soc. Amer., vol. 40(5), pp. 966-78, Nov. 1966.
[26] J.W. Glenn, and N. Kleiner, "Speaker identification based on nasal phonation," J. Acoust. Soc. Amer., vol. 43(2), pp. 368-72, June 1967.
[27] J.J. Wolf, "Efficient acoustic parameters for speaker recognition," J. Acoust. Soc. Amer., vol. 51, no.6(part2) , pp. 2044-56, 1971.
[28] B.S. Atal, "Effectness of linear prediction characteristics of the speech wave for Automatic specker identification and verification ," J. Acoust. Soc. Amer., vol. 55, pp. 1304-12, 1974.
[29] M.R.Sambur, "Selection of acoustic features for specker identification ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-23 (2), pp. 176-82, Apr. 1975.
[30] A.E. Rosenberg , and M.R.Sambur, "New techniques for automatic specker verification ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-23 (2), pp. 169-76, Apr. 1975.
[31] M.R.Sambur, "Specker recognition using orthogonal linear prediction ," IEEE Trans. Acoust., Speech Signal Process., vol. ASSP-24 (4), pp. 283-9, Aug. 1976.
[32] J.D. Markel, B.T. Oshika, and A.H. Grey, Jr., "Long-term feature avarging for specker recognition," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-25 (4), pp. 330-7, Aug. 1977.
[33] S. Furui., "Speaker-independent isolated word recognition using dynamic features of speech spectrum ," IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34, pp. 52-9, Feb. 1983.
[34] Sasaoki Furui., "Spectral analysis technique for automatic speaker verification ," IEEE Trans.

Acoust., Speech, Signal Process., vol. 29 (2), pp. 254-72, Apr. 1981.

[35] D.A. Reynolds, "Experimaental evalution of features for robust speaker identification," IEEE Trans. Speech Audio Process., vol. 2 (4), pp. 639-43, Oct. 1994.

[36] P. Thevenaz and H Hugli, "Usefullness of LPC-residue in text -independent speaker verification," Speech communication., vol. 17, pp. 145-57, 1995.

[37] M.D. Plumpe, T.F. Quatieri, and D.A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification ," IEEE Trans. Speech Audio Process., vol. 7 (5), pp. 569-85, 1999.

[38] M J Carey, E S Paris,H. Lloyd-Thomas, and S.Bennett, "Robust prosotic features for speaker identification ," in proc. Int. Spoken Language Process., Philadelphis,PA,USA,Oct 1996.

[39] M K Sonmez, E Sriberg,L.Heck, amd M. Weintraub, "Modeling dynamic prosodic variation for speaker verification ," in proc. Int. Spoken Language Process., Sydney ,NSW, Australia Nov – Dec. 1998.

[40] B Peskin, J Navratil , J. Abramson, D.Jones, D. Klusacek , D.A. Reynolds,and B.Xiang, "Using prosodic and conservational feature for high-performance speaker recognition ," in Int. Conf Acoust., Speech ,Signal Process., vol. IV,Hong Kong pp. 784-7, Apr. 2003.

[41] M Girmaldi, F Cummins, "Speaker identification using instataneous frequencies ," IEEE Trans. Audio, Speech, Language Process., vol. 16(6), pp. 1097-111, Aug. 2008.

[42] Min-Seok Kim and Ha-Jin Yu, "A new feature transformation method based on rotation for speaker identification," 19th IEEE Int. Conf. on Tools with Artificial Intelligence , pp. 68-73, 2007.

[43] M. Inal, and Y.S. Fatihoglu, "Self organizing map and associative memory model hybrid classifier for speaker recognition," in proc. Neu., Net., App., Elec., Engg. (NEUREL'02), Belgrade, Yugoslavia., pp. 71-4, Sep. 2002.

[44] J. Kennedy and R. Eberhart, "Particle Swarm Optimization", Proceedings of IEEE International Conference on Neural Networks (ICNN'95), Vol. IV, pp.1942-1948, Perth, Australia, 1995.

[45] Min-Seok Kim, IL-Ho Yung and Ha-Jin Yu, "Maximize the distance between the Gaussian mixture models for speaker verification using PSO," 4th IEEE Int. Conf. on Natural Computation , pp. 175-78, 2008.

[46] Samudravijay K "Speech and Speaker recognition report"source:http://cs.jounsuu.fi/pages/tkinnu/reaserch/index.html Viewed on 23 Feb. 2010.

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

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

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

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

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

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

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

[54] 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".

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

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

[64] 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.

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

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

[67] 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

[68] 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

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

[70] 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

[71] 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.

[72] 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.

[73] 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

[74] 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.

[75] 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.

[76] 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.

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

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

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

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

[81] F. K. Soong, A. E Rosenberg, L. R. Rabiner, B. H. Jung, "Vector quantization approach to speaker recognition", in proc, IEEE Int. conf. Acoust., Speech, Signal process., Vol 10, Detroit, Michingon, Apr. !985, pp. 387-90

[82] J.C. Bezdek, and J.D Harris, "Fuzzy portions and relations; an axiomatic basis for clustering, "Fuzzy Sets and Systems, vol. 1, pp. 111-27, 1978.

[83] H. J. Zimmermann, Fuzzy set theory and its applications, 1st ed. Kluwer academic, 1996.

[84] L. Lin, and S. Wang, "A Kernel method for speaker recognition with little data", in Int. Conf. Signal Process., Budapest, Hungery, May, 2006.

[85] V. Chatzis, A.G. Bors, and I. Pitas, "Multimodal decision-level fusion for person authentication". IEEE Trans. Man Cybernetics Part A: Systems and Humans, vol. 29, pp. 674-81, Nov. 1999.

[86] A. E. Rosenberg , and S.Parthasarathy, " Speaker background models for connected digit password speaker verification, " in proc. Int. Conf. Acoust ., Speech ,Signal Process., Atlanta Georgia , May 1996, pp. 81-4.

[87] J.M.Naik, L.P. Nestch , and G.R. Doddington , " Speaker verification using long distance telephone lines , "in proc. Int. Conf. Acoust., Speech , Signal Process., Glasgow UK May 1989,pp.524-7.

[88] T. Matsu, and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and Discrete/continuous HMMs," IEEE Trans. Speech Auto Process., vol. 2(3), pp. 456-9, July 1994.

[89] O. Kimball, M. Schmidt, H. Gish, and J. Waterman, "Speaker verification with limited enrollment data," in proc. European Conf. speech commun. And Tech.(EUROSPEECH '97), Rhodes, Greecee, Sep. 1997, PP. 967-70.

[90] R. P. Lipmann, "An introduction to computing with neural nets," IEEE Trans. Acoust., Speech Signal Process., vol. 4, pp. 4-22, Apr. 1989

[91] G. Bannani, and P Gallinari, "Nural networks for discrimination and modelization of speakers," "Speech Communication., vol. 17, pp.159-75, 1995.

[92] B. Yegnanarayana, Artificial Neural Networks. New Delhi: Prentice Hall, 1999.

[93] J. Oglesby, and J. S. Mason, "Optimization of neural models for speaker identification," in proc. Int. Conf. Acoust., Speech, signal Process., Albuquerque, NM, May 1990, pp. 261-4.

[94] "Radial basis function for speaker recognition ," in proc. Int. Conf. Acoust., Speech, Signal Process., Toronto , Canada , May 1991, pp.393-6.

[95] T. Kohonen , " The self-organizing map ," Proce. IEEE, vol. 78(9), pp. 1464-80, Sep. 1990.

[96] M. Inal, and Y.S.Fatihoglu, "Self organizing map and associative memory model hybrid classifier for speaker recognition , " in proc. Neu., Net., App., Elec., Engg. (NEUREL'02) , Belgrade , Yugoslavia, Sep. 2002 pp. 71-4.

[97] A.T. Mafra, and M.G.Simoes, " Text independent automatic speaker recognition using self-organizing maps," in proc. Ind. App. Society conf., vol. 3 Victoria , British Columbia , Oct.2004,pp. 1503-10.

{98] G.Binnani, F. Fogelman, P. Gallinari, " A connectionist approach for speaker identification ," in proc. Conf. Acoust., Speech, Signal Process., Albuquerque, NM , May 1990 , pp. 265-8.

[99] J. He , L. Liu, G. Palm , "A discriminative training algorithm for VQ-based speaker identification, "IEEE Trans. Speech Audio Process., vol. 7, pp. 353-6 , May 1999.

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

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

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

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

[104] 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.

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

[106] 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.

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