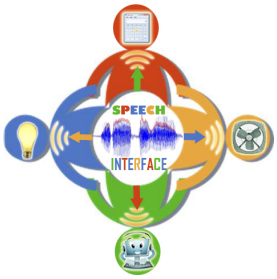

CHAPTER 3

Feature Extraction Techniques for Speech Recognition

3.1 Introduction

The goal of speech recognition area is to develop techniques and system for speech input based machine communication. Statistical, modeling of speech and automatic speech recognition has invented widespread application for human machine interface such as automatic call processing [1]. Since 1960s, computer scientists have been researching techniques for making computer able to record, interpret and understand human speech. Even the most rudimentary problem such as sampling voice was a huge challenge in the early years. The communication among the human being is dominated by spoken language. The researcher turn towards developing a system which communicate with computer in native language [2]. Machine recognition of speech involves real time application such as travel information and reservation, translators, natural language understanding and many more [3, 4].

The earliest speech recognition systems were first attempted in the early 1950s, at Bell Laboratories, this system was developed for isolated digit recognition system of a single speaker. The goal of automatic speech recognition is to analyse, extract characterize and recognize information about the speaker identity. Speech recognition system may be viewed as

working in a four stages. The techniques of speech recognition are classified in four classes they are Analysis, Feature extraction, Modeling and Testing techniques.

3.2 Speech Analysis Techniques

Speech data carries special type of information that express speaker identity. This includes speaker specific information such as information regarding vocal tract, excitation source and behaviour feature. Information about the behaviour feature was also embedded in signal and can be used for speaker recognition. The speech analysis method deals with suitable frame size for segmenting speech signal, which can be used for further analysis [5]. The speech analysis technique is classified in the following three techniques:

3.2.1 Segmentation Analysis

In this case speech is analysed using the frame size and shift in the range of 10-30 ms. to extract speaker information. This technique mainly provides vocal tract information for speech and speaker recognition.

3.2.2 Sub Segmental Analysis

In this technique speech is analysed using the frame size and shift in range 3 to 5 ms. This technique is used to analyse characteristic of the excitation state [6].

3.2.3 Supra Segmental Analysis

In this case, speech is analysed using the frame size. This technique is used to analyse behaviour characteristics of the speaker and the two supra segmental analysis techniques are discussed below:

3.2.3.1 Linear Predictive Coding (LPC)

In the area of speech recognition, the Linear Predictive Coding (LPC) is one of the robust and dynamic speech analysis techniques. It is also powerful for speech encoding using the lowest bit rate. This technique provides basic speech parameter and which can be used for efficient computation of

performance evaluation. The variation of LPC depends on intensity, frequency, pitch and formant. The number of LPC coefficient is executed from run source through filter on resulted coefficient of speech.

3.2.3.2 Rasta-PLP

The Perceptual linear prediction analysis depends on short term spectrum of the speech. For the improvement of a result in PLP the short term spectrum of the speech by different psychologically based transformation is used. The Linear Predictive speech analysis technique is based on short term spectrum of speech. It is one of the most powerful speech analysis technique and most useful methods for encoding good quality speech at a low bit rate. It provides extremely accurate estimation of speech parameters. The short-term spectral values are modified by the frequency response of communication, which makes this technique vulnerable. Rasta processing is an approach of feature extraction, enhancement and suppression for speech recognition. Rasta processing increases dependence of the data on its previous context. The Rasta processing works well in word model. The RASTA filter can be used either in the log spectral or Cepstral domains.

3.3 Feature Extraction Technique

In the classification problem the speech feature extraction is used for reducing the dimension of the input vector, while maintaining the perceptive power of the signal. Feature extraction is a special form of dataset and it results in extraction of specific features. These features carry the characteristics of the useful information regarding speech. Feature design and selection is the main challenging problem in the speech recognition system development for specific application. For speech identification and verification development, the number of training and testing vector are needed for the classification. The problem grows with the dimension of the given input. The techniques available in enrich literature for speech feature extraction is described in table 3.1, with their properties.

Table 3.1: The Feature Extraction technique with their comparative properties [7]

Sr.No.	Method	Property
1	Principal Component analysis (PCA)	• Eigenvector-based method. • Nonlinear feature extraction method • Supported to Linear map. • Faster than other technique. • It is good for Gaussian data.
2	Linear Discriminate Analysis(LDA)	• Linear feature extraction method • Supported to supervised linear map. • Faster than other technique, • Better than PCA for classification.
3	Independent Component Analysis (ICA)	• Blind course separation method • Support to Linear map • It is iterative in nature • It is good for non- Gaussian data.
4	Linear Predictive coding	• Static feature extraction Method. • It is used for feature Extraction at lower order coefficient.
5	Cepstral Analysis	• Static feature extraction method. • Power spectrum method. • Used to represent spectral envelope.
6	Mel-Frequency Scale Analysis	• Static feature extraction method. • Spectral analysis method. • Mel scale is calculated.
7	Filter Bank Analysis	• It required frequencies possible • Used for filter based feature extraction.
8	Mel-Frequency Cestrum Coefficient (MFFCs)	• Power spectrum is computed by performing Fourier Analysis, • Robust and dynamic method for speech feature extraction
9	Kernel Based Feature Extraction Method	• Nonlinear transformations method
10	Wavelet Technique	• Better time resolution than Fourier Transform, Real time factor is minimum
11	Dynamic Feature Extractions i)LPC	• Acceleration and delta coefficients • II and III order derivatives of

	ii)MFCCs	Normal LPC and MFCCs coefficients
12	Spectral Subtraction	• Robust Feature extraction method
13	Cepstral Mean Subtraction	• Robust Feature extraction method for small vocabulary based system
14	RASTA Filtering	• Used for Noisy speech recognition
15	Integrated Phoneme Subspace Method (Compound Method)	• A transformation based on PCA + LDA + ICA. • It gives Higher Accuracy than the existing Methods.

A new modification of Mel-Frequency Cepstral Coefficient (MFCC) feature has been proposed for extraction of speech features for speech recognition. The work uses multidimensional F-ratio for performance measure in speech recognition (SR) applications to compare discriminate ability of different multiple parameter methods [8]. There are many parameters that affect the accuracy of speech recognition system such as vocabulary size, speaker dependency, time for recognition, type of speech (continuous, isolated) and recognition environment. A speech recognition algorithm consists of several stages in which feature extraction and classification are most important. In feature extraction category best presented algorithms are

- Mel Frequency Cepstral Coefficients (MFCC),
- Linear discriminant analysis (LDA)
- Linear Prediction Cepstral Coefficient (LPCC)
- Linear Prediction Coefficients (LPC)
- Principal Component Analysis (PCA)

The enriched literature available on speech recognition hence reported that the MFCC is most popular and robust technique for feature extraction [9, 10].

3.3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) technique is the robust and dynamic technique for speech feature extraction [10]. Figure 3.1 shows the complete block diagram of the Mel Frequency Cepstral Coefficients. The

Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to extract important feature of sound file. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies [11].

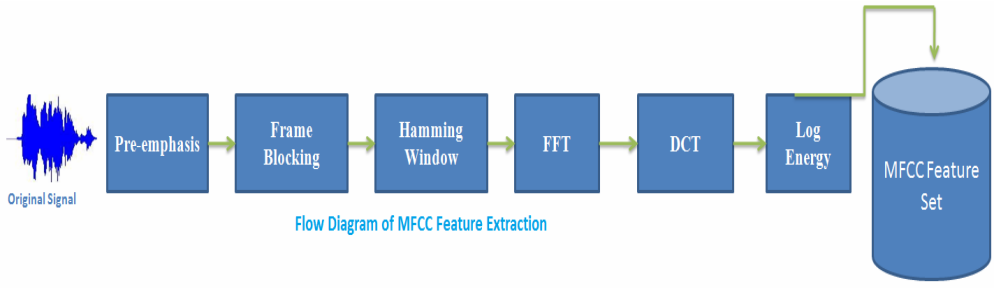

Figure 3.1: Block diagram of MFCC.

In this technique the logarithm of high frequencies used for capture the important characteristics of speech. From the literature it is observed that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels.

The following formula is used to compute the Mels for a particular frequency:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f / 700).$$

A step by step implementation of the MFCC is shown in Figure 3.2 [12]. In the Pre-emphasis each of the speech samples is sampled to 16000 Hz for analysis purpose. The sample speech signal was pre-emphasized with filter. In the pre-emphasized the signal is blocked onto the frame of N sample, with adjacent frame being separated by M. Finally, the log Mel spectrum was converted into time. The output is called Mel Frequency Cestrum Coefficients (MFCC).The MFCC is real numbers and it can be converted into time domain using the Discrete Cosine Transform (DCT). The MFCC

is used to discriminate the repetitions and prolongations in natural speech [13]. The researcher used MFCC with 12, 13, 26 and 39 variations as original feature and derivative of it.

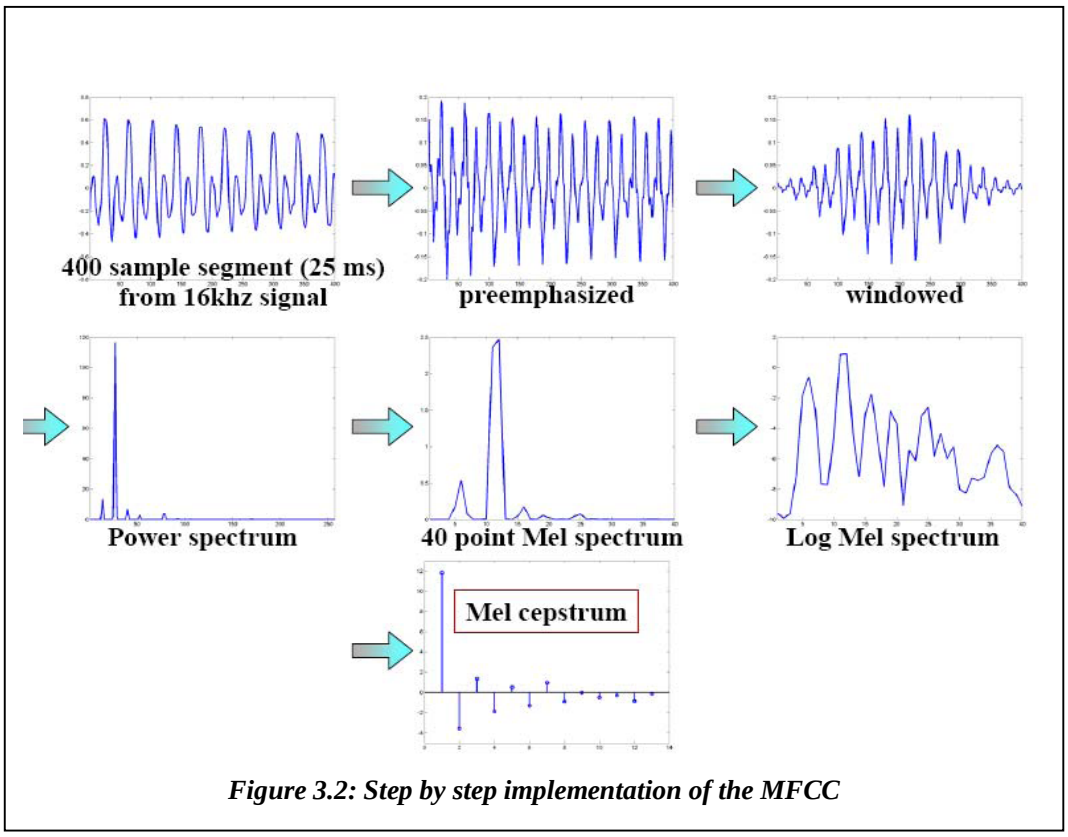

Figure 3.2: Step by step implementation of the MFCC

3.3.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is commonly used technique for data classification and dimensionality reduction. It easily handles the case where within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within class variance in any particular data set thereby guaranteeing maximal reparability. The use of Linear Discriminant Analysis for data classification is applied to classification problem of speech recognition [11]. LDA algorithm provides better classification compared to principal components analysis [14]. The figure 3.3 describes the LDA training after projection.

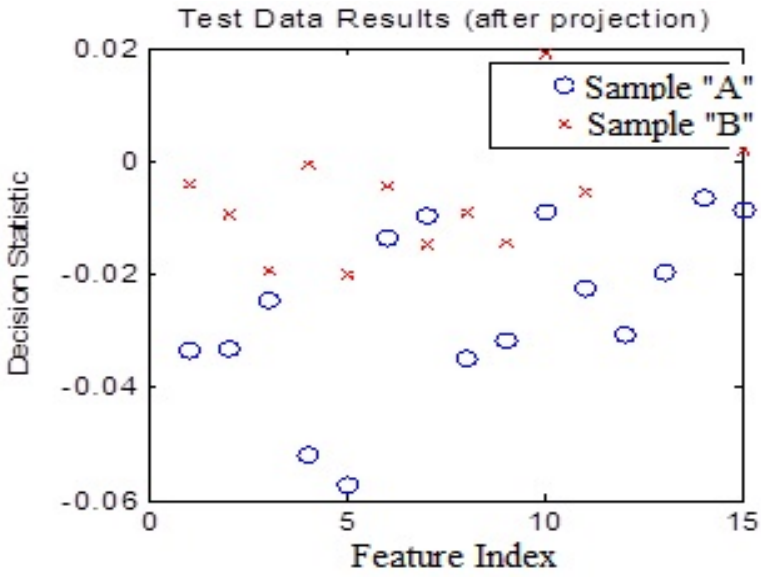

Figure 3.3: LDA Training after projection.

3.3.3 Principal Component Analysis

PCA is a well-established technique for feature extraction and dimensionality reduction. It is based on the assumption that most information about classes is contained in the directions along where the variations are the large. The most common derivation of PCA is in terms of a standardized linear projection which maximizes the variance in the projected space. Principal components analysis (PCA) is a method of identifying patterns in data, and highlights their similarities and differences. It is powerful tool for analyzing data. The main advantage of PCA is that once the patterns in the data, and from data were found then compression may be done i.e. dimension may be reduced. Figure 3.4 describes the working of PCA.

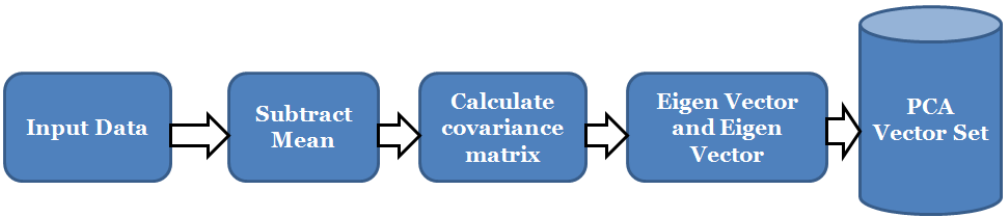

Figure 3.4: working of PCA

3.3.4 Discrete Wavelet Transformation

The Wavelet series is just a sampled version of continuous wavelet transformation and its computation may consume significant amount of time and resources, depending on the resolution required. The discrete wavelet transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. The DWT level 1 was used for the experiment.

The procedure starts with passing this signal (sequence) through a half band digital low pass filter with impulse response h[n]. Filtering a signal corresponds to the mathematical operation of convolution of the signal with the impulse response of the filter. The figure 3.5 described the basic structure of wavelet [15].

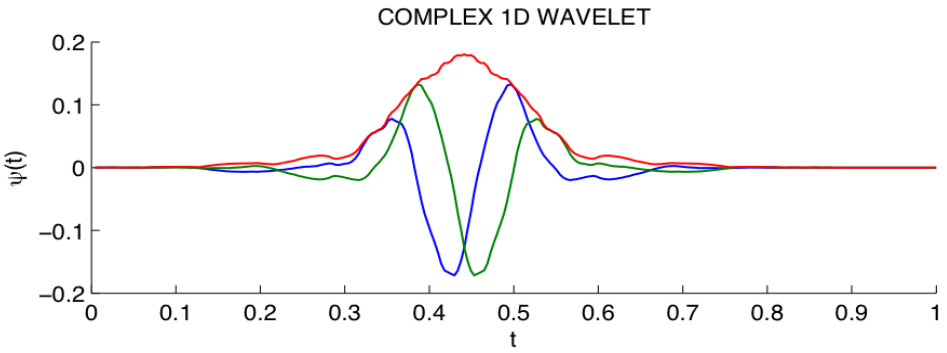

Figure 3.5: Basic structure of wavelet

3.4 Modeling Technique

The objective of modeling technique is to generate models using speaker specific feature vector. This modeling technique is divided into two classifications such as speech identification and recognition. The speech identification technique automatically identifies speech which is in the trained database. The speech recognition is also divided into two parts that means speaker dependent 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 other hand, in case of speaker dependent recognition machine should extract speaker characteristics in the acoustic signal [16]. The main aim of speech identification is comparing a speech signal from an unknown speech to a database of known speech. The system can recognize the speech, which has been trained with a number of speakers [17]. The modeling approach for speech recognition process is described as below.

3.4.1 The Acoustic Phonetic Approach

This method is definitely viable and has been studied in great depth for more than 60 years. This approach was based upon theory of acoustic phonetics and postulates [18]. The earliest approaches to speech recognition were based on acoustic phonetic approach. It is assumed in the acoustic-phonetic approach that rules governing the variability are straightforward and can be readily learned by a machine [19]. Formal evaluations conducted by the national institute of science and technology (NIST) in 1996 which demonstrated that automatic language identification (LID) used the phonetic feature of speech signal and discriminate among set of languages [20]. For speech recognition system the phone based approach demonstrates good performance [21, 22]. Phone recognition, Gaussian mixture modeling, and support vector machine classification are two techniques have been applied to the language identification. The acoustic phonetic approach has not been widely used in most commercial applications [23].

3.4.1.1 Fundamental Frequency

Voice signals can be considered as quasi-periodic. The fundamental frequency is called the pitch. The average pitch period, time pattern, gain, and fluctuation change from one individual speaker to another speaker.

3.4.1.2 Vocal Tract Resonance:

The vocal tract filter frequency response is nothing but the shape and gain of envelop of the signal with values of formants and bandwidth. The following feature are used for the acoustic phonetic approach

A) Energy

B) Formants

C) Pitch

A) Energy

The energy also affects the performance of acoustic model in the speech recognition. The energy of speech is basic and independent parameter. Energy of each frame is calculated by equation 1 given below. The mathematical formulation of energy described in equation1.

$$E_t = \int_t^{t+\tau} |X(t)|\, dt \qquad \text{Equation 1}$$

Energy of all the frames is ordered and the top ones are selected for the following process to obtain the pitch feature. The voiced frame was determined by calculating the energy contained within certain bandwidths [52]. No doubt, the computation complexity is greatly reduced. The following experimental results indicate that such a simple energy calculation is able to yield speech frames which contain relatively strong pitch feature. The amplitude of unvoiced segments is noticeably lower than that of the voiced segments. The short-time energy of speech signals reflects the amplitude variation.

B) Formant

Formant was outlined by Gunnar Fant (1960). The spectral peaks of the spectrum |P (f)| are referred to as formant. This definition is generally utilized in acoustic analysis and trade. The literature shows various definitions of formants described in [24], the peaks that are determined within the spectrum envelope are referred to as formant. In elements of the speech analysis community, however, formant has come back to possess different meanings. In the process of formant [25], it defines resonance frequencies of the vocal tract in terms of a gain operate T (f) of the vocal tract: The frequency location of a maximum in |T (f)|. The resonance and formant are so conceptually distinct. The acoustics of the vocal tract are usually sculptured employing a

mathematical model of a filter [26]. The frequencies of the poles of this filter model fall near those of the formant.

C) Pitch

The voiced regions look like periodic signal in the time domain representation. Pitch is defined as the fundamental frequency of the excitation source. Hence an efficient pitch extractor and an accurate pitch estimate calculated can be used in an algorithm of gender identification. The pitch feature allows people to communicate verbally. This unique feature can help the improvement the performance of emotion recognition. Everyone has a distinctive voice, different from all others. One's voice is unique it can act as an identifier. The human voice is composed by the magnitude of different components which makes each voice different such as pitch, attitude, and sampling rate. The periodicity associated with such segments is defined is pitch period.

The estimation of pitch is one of the important issues in speech processing. There are a large set of methods that have been developed for the estimation of pitch. Among them three frequently used methods are auto correlation, cestrum analysis and single inverse technique (SIFT). SIFT is successful due to the involvement of simple steps for the estimation of pitch. Even though auto correlation method is of theoretical interest, it produce a frame work for SIFT methods [27].

The basic period is called the pitch period. The average pitch frequency (in short, the pitch), time pattern, gain, and fluctuation change from one individual speaker to another. A well-known method for pitch detection is given in [28]. The pitch detector's algorithm can be given by equations 2 and 3. Figure 3.6 describes a vocal phoneme, in which the pitch marks are denoted in red.

$$\langle x, y \rangle = \int_{t_0}^{t_0+\tau} x(t) \cdot y(t) dt \quad ; \quad y(t) = x(t-\tau) \quad \text{Equation 2}$$

Where, $T_0 = \arg\max(\rho_\tau)$;

$$\rho_\tau = \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|} \quad ; \quad \|x\| = (\langle x, x \rangle)^{1/2} \quad \text{Equation 3}$$

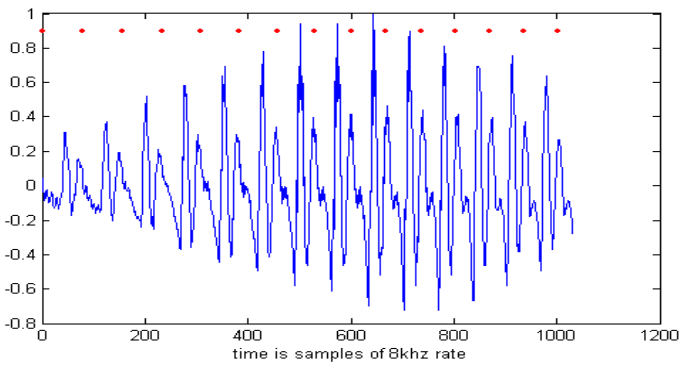

Figure 3.6: A phoneme with its pitch cycle marks (in red).

i. Pitch Synchronous Overlap and Add Technique

The PSOLA (Pitch Synchronous Overlap and Add) technique is the dynamic technique for pitch estimation. In the basic TD-PSOLA (Time Domain PSOLA) system, prosodic modifications are made directly on the speech waveform. The same approach can also be applied on the error detection in signal resulting from the LPC analysis. The Pitch marks are calculated and denoted on the error signal, which is then divided into a stream of short-term signals, synchronized to the local pitch cycle [29].

The second stage is to warp the reference sentence in order to align it with the target sentence. Once this is done, the sequence of short-term signals of the reference speaker is converted into a modified stream of synthesized short-term signals synchronized on a new set of time instants. This new set of time instants is determined in order to comply with the desired modifications. A new error signal is then obtained by overlapping and adding the new stream of synthesized short-term signals. The last step is to synthesize the synchronized signal to the new pitch marks.

ii) Spectrum Parameter Modification Algorithm

In the formant analysis the moving frequency around the circle is result in loss of voice quality individually. Along with bandwidth and gain changes,

can be achieved through direct changes to the filter. This technique is also known as speech morphing technique, which changes in spectrum parameter and fundamental frequency [30]. In this algorithm the first stage is to find the pitch marks of each speaker's utterance and to create a correspondence table to match the source and the target. The values are set for the amount of modification desired is targeted in second step. The third step is to make the necessary spectrum modifications.

iii) Cross-Correlation Technique

Cross-correlation is calculated between two consecutive pitch cycles. The cross-correlation values between pitch cycles are higher (close to 1) in voiced speech than in unvoiced speech.

iv) Filter Bank Analysis Technique

In signal processing, a filter bank is an array of pass filters that separates the input signal into multiple components, each one carrying a single frequency sub of the original signal. One application of a filter bank is a graphic equalizer, which can attenuate the components differently and recombine them into a modified version of the original signal. The process of decomposition performed by the filter bank is called analysis (meaning analysis of the signal in terms of its components in each sub-band); the output of analysis is referred to as a sub band signal with as many sub bands as there are filters in the filter bank. The speech synthesis means the reconstruction of completer signal resulted from filtering process. The vocodor uses a filter bank to determine the amplitude information of the sub-bands for a modulator signal (such as a voice). It is used to control the amplitude of the sub-bands of a carrier signal (such as the output of a guitar or synthesizer), thus it is commanding the dynamic characteristics of the modulator on the carrier [31].

3.4.2. Pattern Recognition Approach

Pattern recognition is almost synonymous with machine learning. This branch of artificial intelligence focuses on the recognition of patterns and regularities in data. In many cases, these patterns are learned from labelled training data (supervised learning), but when no labelled data are available

other algorithms can be used to discover previously unknown patterns (unsupervised learning) [32]. A pattern recognition has been developed over four decades has received much attention and applied widely too many practical pattern recognition problem [23]. The pattern matching approach includes two essential phases namely, pattern training and pattern comparison. The essential advantages of this method are it uses a well formulated mathematical framework. Speech pattern representation was in the form of a speech template or a statistical model such as Hidden Markov Model. In Pattern comparison method the direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage. The pattern matching approach has become the leading method for speech recognition in the last six decades [33].

3.4.2.1 K- Nearest Neighbours Technique(KNN)

The KNN Algorithm is a widely applied method for classification or regression in pattern recognition and machine learning. As a lazy learning, KNN Algorithm is instance-based and used in many Applications in the field of statistical pattern recognition, Data mining, Image processing and many applications. The KNN Algorithm is simple but computationally intensive. When the size of train data set and test data set are both very Large, the execution time may be bottleneck of the application [34].

3.4.3 Template Based Approach

In the template based approach the unknown speech is compared against set of pre-recorded words (templates) in order to find the best Match. This approach has the advantage of using perfectly accurate word models. Template based approach of speech recognition have provided family of techniques that have advanced the field considerably during the last six decades [35]. One key idea in template method is to derive typical sequences of speech frames for a pattern (word) via some averaging procedure, and to rely on the use of local spectral distance measures for compare patterns [36].

For the template matching dynamic time warping, support vector machine techniques are used.

3.4.3.1 Support Vector Machine

The support vector machine (SVM) is the robust and dynamic clustering technique. The foundations of Support Vector Machines (SVM) have been developed by Vapnik [15] and gained reputation due to many promising features such as better empirical performance. It belongs to a family of generalized linear classifiers. It can be defined as systems which use hypothesis space of linear functions in a high dimensional feature space, also trained with a learning algorithm from optimization theory. Support vector machine was originally popular with the NIPS community and now is an active part of the machine learning research domain around the world. It becomes eminent when, using pixel maps as input. It gives comparable accuracy to sophisticated neural networks with elaborated features [37]. It is used especially for pattern classification and regression based applications. The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior [38] to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks. SVMs were developed for solving the classification problem, but currently it has been extended to solve regression problems [40]. Support Vector Machines acts as one of the excellent approach to data modeling. The kernel mapping provides a common base for most of the commonly employed model architectures, enabling comparisons to be performed [39]. The minimization of the weight vector can be used as a criterion in regression problems, with a modified loss function. This technique also used for choosing the kernel function, additional capacity control and development of kernels with invariance [41].

3.4.3.2 Dynamic Time Warping

Dynamic time warping is an algorithm for measuring similarity between two sequences based on time or speed. It has been applied to video, audio, and graphics indeed, any data, which can be turned into a linear representation. A

well-known application has been automatic speech recognition, to manage with different speaking speeds. In general, DTW is methods that allow a computer to find an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are warped non-linear manner for time dimension to determine the measure of their similarity independent of certain non-linear variations. It stretches and compresses various sections of utterances to find alignment those results in the best possible match between template and utterance frame by frame basis. This method is quite efficient for isolated word recognition and can be adapted to connected word recognition [42].

3.4.4. Knowledge Based Approach

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 is judged to be impractical and automatic learning procedure was sought instead. Vector Quantization [43] 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 [44].

3.4.4.1 Vector Quantization

The Vector Quantization (VQ) is the fundamental and most successful technique used in speech coding, recognition, and synthesis [45]. This technique is applied in the speech analysis where the mapping of large vector space is divided into a finite number of regions in same space. The VQ technique is commonly applied to develop discrete or semi-continuous HMM

based speech recognition system. The VQ encoder encodes a given set of k-dimensional data vectors with a much smaller subset.

In VQ, an ordered set of signal samples or parameters can be efficiently coded by matching the input vector to a similar pattern [24]. The VQ techniques are also known as data clustering methods in various disciplines. It is an unsupervised learning procedure widely used in many applications domain. Basically, the data clustering methods are classified as hard and soft clustering methods. These are centroid-based parametric clustering techniques based on a large class of distortion functions known as Bregman divergences [25]. In the hard clustering, each data point belongs to exactly one of the partitions in obtaining the disjoint partitioning of the data whereas each data point has a certain probability of belonging to each of the partitions in soft clustering. The parametric clustering algorithms are very popular due to its simplicity and scalability. The commonly used vector quantization is based on nearest neighbour called Verona or nearest neighbour vector quantization.

3.4.5 Statistical Based Approach

In statistical based approach variations are modelled statistically, using automatic, statistical learning procedure, typically the Hidden Markov Models. The main disadvantage of statistical models is that they must take prior modeling assumptions which are answerable to be inaccurate, handicapping the system performance. For speaker independents speech recognition use left-right HMM for identifying the speaker from simple data. The HMM is popular statistical tool for modeling wide range of time series data. In Speech recognition HMM has been applied with great success to problem such as speech classification [46]. A weighted Hidden Markov Models algorithm and a subspace projection algorithm are proposed to address the discrimination and robustness issues for HMM based speech recognition. Word models were constructed for combining phonetic and fenonic models. Learning Vector Quantization (LVQ) method showed an important contribution in producing highly discriminate reference vectors for classifying static patterns [47]. The Machine learning estimation of parameter

via Forward Backward algorithm was an inefficient method for estimating the parameter value of HMM. An alternative of VQ method in which the phoneme is treated as a cluster for speech space and a Gaussian model was estimated for each phoneme [48]. 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 [49].

3.4.5.1 Hidden Markov Models (HMM)

A hidden Markov model (HMM) is a statistical Markov model for speech recognition in which the system being modelled. It is assumed to be a Markov process with unobserved (hidden) states. An HMM can be presented as the simplest dynamic Bayesian network. The basic theory behind the Hidden Markov Models (HMM) dates back to the late 1900s, when Russian statistician Andrej Markov first presented Markov chains. Baum and his colleagues introduced the Hidden Markov Model as an extension to the first-order stochastic Markov process and developed an efficient method for optimizing the HMM parameter estimation in the late 1960s and early 1970s [50]. The Markov process itself cannot be observed, and only the sequence of labelled balls can be observed, thus this arrangement is called a "hidden Markov process", it is described in figure 3.7.

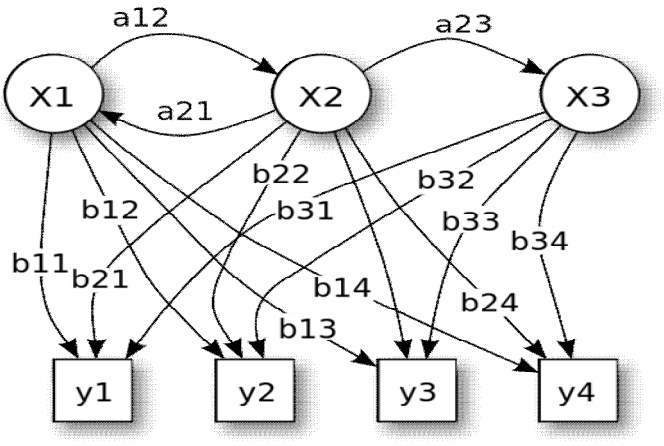

Figure 3.7: Probabilistic parameters of a hidden Markov model [49]

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech recognition [51].

The technique of HMM has been broadly accepted in today's modern state or the art ASR systems mainly for two reasons: its capability to model the non-linear dependencies of each speech unit on the adjacent units and a powerful set of analytical approaches provided for estimating model parameters [52,53]. The Hidden Markov Model (HMM) is a variant of a finite state machine having a set of hidden states Q, an output alphabet (observations) O, transition probabilities A, output (emission) probabilities B, and initial state probabilities II. The current state is the state of not observable. In its place, each state produces an output with a certain probability (B). Typically the states Q, and outputs O, are unspoken, so an HMM is said to be a triple (A, B, II).

3.4.6 The Artificial Intelligence Approach

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way person applies. Its intelligence is visualizing, analysing, and finally making a decision on the measured acoustic features. Expert system is used widely in this approach [54]. The Artificial Intelligence approach is hybrid of the acoustic phonetic approach and pattern recognition approach. It exploits the ideas and concepts of Acoustic phonetic and pattern recognition approach. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. While template based approaches have been very effective in the design of a variety of

speech recognition systems. In human speech processing the error analysis and knowledge based system enhancement is little difficult. In its pure form, knowledge engineering design involves the direct and explicit incorporation of expert's speech knowledge into a recognition system. 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.

Artificial neural networks were first developed in the early nineteen forties when a neurophysiologist, Warren McCulloch and a mathematician, Walter Pitts, wrote a paper on how neutrons might work by modeling a simple electrical circuit to describe the process. The idea with this model was to investigate the activity of neurons in the thinking process.

The artificial neural network is a network of a number of interconnected units with each unit having input or output characteristics that implements a local computation or function. Typical neural networks operate in parallel nodes whose function is determined by the network structure, the connection strengths and the function in each node. Neural networks have the unit ability to "learn". In other words, the human does not necessarily have to be able to explain the "problem" to the system [55]. The figure 3.8 described the structure of neural network.

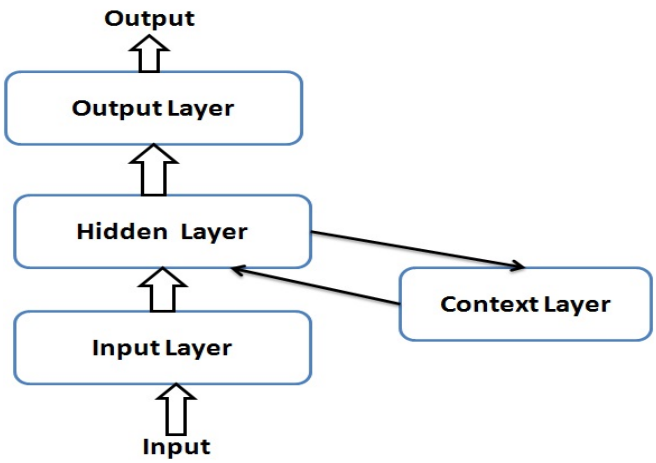

Figure 3.8: Showing structure of recurrent neural network.

Neural nets can be conveniently described as black-box computational methods for addressing basic Stimuli-Response processes (S-R). On each side of the black-box (ANN) is a known set of inputs corresponding to their respective output set hence any distortion in the input of the system would employ algorithms and codes within the black-box to produce a unique output for that stimulus. It is through this process that the "new" output is added to the already existing set of standard neural network responses for known stimuli. It is important to note that the standard S-R pairs encoded into the artificial neural network ought to represent the stable states of the system during normal operation.

3. 4.7 Stochastic Approach

Stochastic modeling requires the use of probabilistic models to deal with uncertain or incomplete information [6]. In speech recognition, uncertainty and incompleteness arise from many sources, such as confuse sounds, speaker variability's, contextual effects, and homophones words. Stochastic models are particularly suitable approach to speech recognition.

Hidden Markov Model is most popular stochastic approach and characterized by a finite state Markov Model. The transition parameters in the Markov chain models, temporal variability's, while the parameters in the output distribution model, spectral variability's. Hidden Markov modeling is more general and has a stronger mathematical foundation. A template based model is simply used for continuous speech recognition [56].

3.4.8 Proposed Fusion Approach

The fusion approach is nothing but the combination of different features extraction techniques.

3.4.8.1 Fusion Approach with MFCC

The Fusion approach means combination of different techniques. Total 13 MFCC features were extracted and feature vector was formed. The formed

feature vector was passed to fusion technique as an input. Figure 3.9 describes result of fusion MFCC feature extraction using LDA.

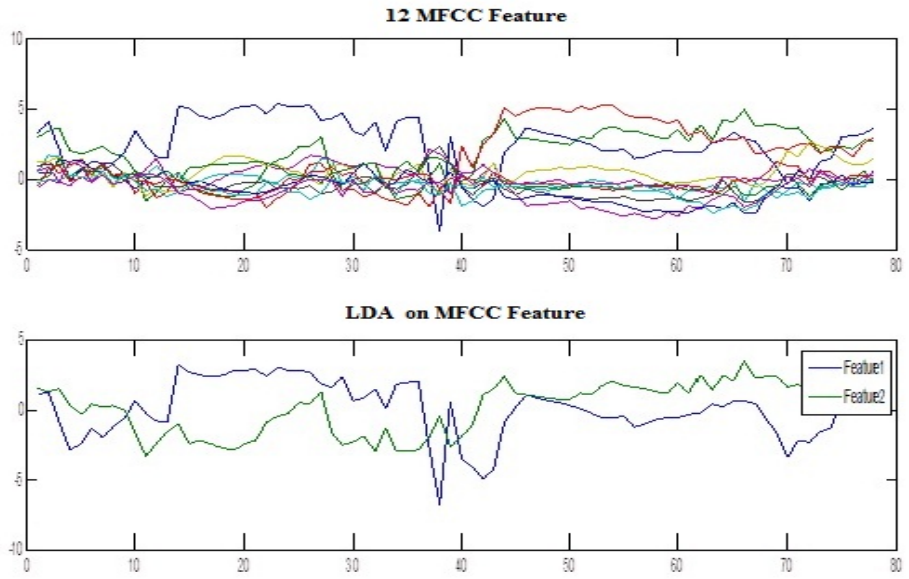

Figure 3.9: Fusion MFCC feature extraction

The detail fusion approach of different techniques with MFCC and their properties are explained in table 3. 2. From the table 3.2, we tested the fusion approach of MFCC with the LDA, PCA and DWT. The fusion approach is used for dimension reduction and for reduction of the time complexity. From the input the 13 MFCC feature are passed as input of the fused technique we got a 02 feature vector in minimum time. The performance of this fusion approach technique will improve.

Table 3.2: Fusion approach with MFCC and their properties

Sr.No	***Name of Fusion Technique***	***Combination of Techniques***	***Input feature vector***	***Output feature vector***	***Properties***
1	MFLDA	Fusion of MFCC and LDA	13	02	It is used for dimension reduction and classification without loss of information.
2	MFPCA	Fusion of MFCC and PCA	13	02	It is used for dimension reduction

3	MFDWT	Fusion of MFCC and DWT	13	01	It is used for dimension reduction. The speed is fast as compare to other techniques.
4	MFPDWT	Fusion of MFCC,PCA and DWT	13	01	It is used to reduce time complexity.
5	MFLDWT	Fusion of MFCC,LDA and DWT	13	01	It is used for dimension reduction and classification. It is also used to reduce time complexity.

3.4.8.2 Proposed approach

- ***Wavelet Decomposed Cepstral Coefficient (WDCC)***

The wavelet packet method is a generalization of wavelet decomposition that offers a richest improvement in the performance of the recognition system. The wavelet packets feature is indexed by three naturally interpreted parameter scale, position and frequency. In the wavelet packet analysis, each detail coefficient vector is also decomposed into two different parts using same approach as used in approximation vector splitting. In the splitting between two approximation coefficient vector and successive detail coefficient vectors never reanalysed. This strategy of decomposition offers richest analysis of signal [57, 58, 59, 60, 61]. From the decomposition process the complete binary tree is produced. The wavelet packet coefficient used for filter analysis. The wavelet packet was proposed by Coffman as a collection of bases in a hierarchical tree structure. The packet coefficient offers different time frequency representation qualities and consequently potential, for adaptation of the time series phenomenon [62, 63, 64, 65, 66]. In the proposed wavelet decomposed Cepstral coefficient, the original speech signal is decomposed second level. The approximation and detail coefficient is a distinguished output from decomposition steps. The figure 3.10 described the decomposition of wavelet.

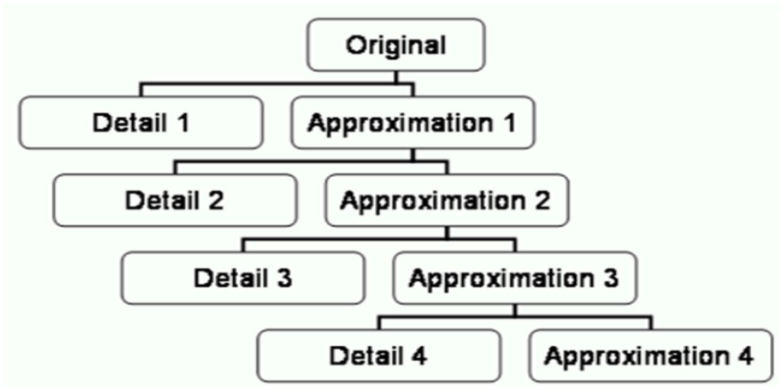

Figure 3.10: Flow chart of wavelet packet decomposition level.

The FFT, DCT and Log operation are performed on approximation level coefficient. In a parallel manner horizontal coefficient are also highlighted from the extracted detail coefficient. The DCT operation performed on horizontal coefficient, fused with basic acoustic coefficient are derived to first and second derivation where we got 18, 36, 39 WDCC coefficients. The flow diagram of WDCC is described in figure 3.11.

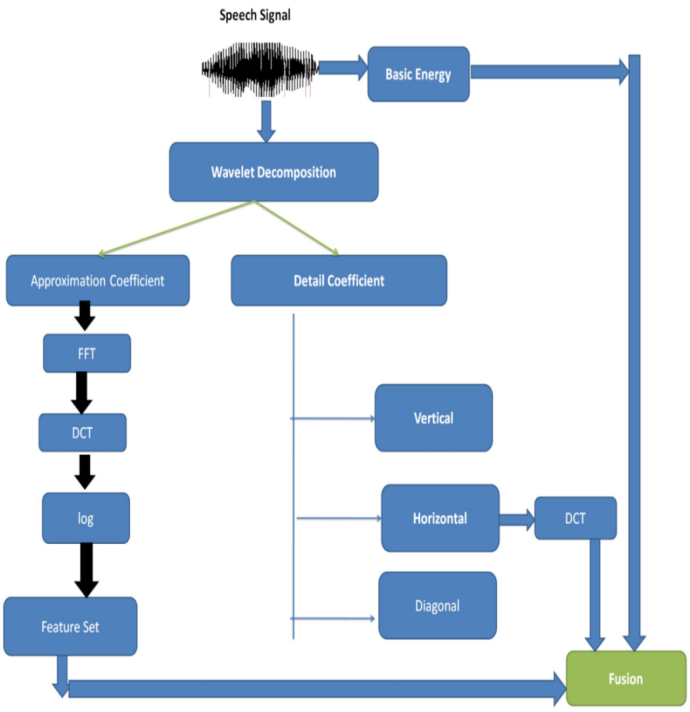

Figure 3.11: The proposed flow diagram for WDCC

3.5 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) [67]. For the performance of the system the researcher of current era also used the Real Time Factor value.

3.5.1 Word Error Rate (WER)

Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level [68, 69]. 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.

3.5.2 Real Tim Factor (RTF)

The real time factor (RTF) is a common metric for computing the speed of an automatic speech recognition system. It can also be used in other frameworks

where an audio or video signal is processed (usually automatically) at approximately constant rate [68]. If it takes time **P** to process an input of duration **I**, the real time factor is defined as

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

The performance of the system is not only depending on accuracy but also highly dependent of RTF value.

$$Performance = Accuracy * RTF$$

The accuracy of a speech recognition system, on the other hand, is measured with the word error rate.

3.6 Conclusion

This chapter deals with the available methodology used in order to analyse speech patterns for speech recognition system. The methodology clearly classified in speech analysis, feature extraction, modeling and testing technique. From the study of all technique, it is observed that MFCC is robust and dynamic technique for speech recognition domain. The fusion approach of the MFCC with other technique will improve the performance of the system. For the development of speech interface system the respond time and accuracy of the system is very important so, we proposed the Wavelet based feature extraction and tested the system. This chapter focused on an informative approach towards fusion based approach and proposed feature extraction technique for speech interface system. The details of the experiments designed as per the methodology will be discussed in chapter 04.

References

1. R. Klevansand, R. Rodman, "Voice Recognition", Artech House, Boston, London 1997.
2. Samudravijaya K., "Speech and Speaker Recognition Tutorial", TIFR Mumbai 400005.
3. Kevin Brady, Michael Brandstein, Thomas Quatieri, Bob Dunn, "An Evaluation of Audio-Visual person Recognition on the XM2VTS corpus using the Lausanne protocol", MIT Lincoln Laboratory, 244 Wood St., Lexington MA.
4. W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, J. Navratily, "The MIT- LL/IBM Speaker recognition System using High performance reduced Complexity recognition", MIT Lincoln Laboratory IBM 2006.
5. 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
6. Nicolás Morales, John H. L. Hansen, Doorstep T. Toledano, "MFCC Compensation for improved recognition filtered and band limited speech" Centre for Spoken Language Research, University of Colorado at Boulder, Boulder (CO), USA
7. M. A. Anusuya, S. K. Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009
8. Goutam Saha, Ulla S. Yadhunandan, "Modifield Mel- Frequency Cepstral Coefficient", Department of Electronics and Electrical communication Engineering India Institute of Technology, Kharagpur Kharagpur-721302 West Bengal, India.
9. Bharti W. Gawali, Santosh Gaikwad, Pravin Yannawar, Suresh C. Mehrotra, "Marathi Isolated Word Recognition System using MFCC and DTW Features" ACEEE 2010.
10. M. Kesarkar, "Feature Extraction for Speech Recognition" Indian Institute of Technology, Bombay, 2003.

11. Kashyap Patel, R. K. Prasad, "Speech Recognition and Verification Using MFCC & VQ ", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013

12. Santosh Gaikwad, Bharti Gawali, Suresh Mehrotra, "MFCC and TW Approach for Accent Identification", IEEE's International Conferences for Convergence of Technology, Pune, India.5th Apr 2014.

13. Lim Sin Chee, Ooi Chia Ai, M. Hariharan, Sazali Yaacob, "MFCC based Recognition of Repetition and Prolongations in Stuttered speech using K-NN and LDA", Proceedings of 2009 IEEE Student Conferences on Research and Development (SCOReD 2009), 16-18 Nov, 2009, UPM Serdang, Malaysia

14. Oh-Wook Known, Kwokleung Chan, Te-Won Lee, "Speech Feature Analysis Using Variational Bayesian PCA", in IEEE Signal Processing letters, Vol. 10, pp.137 – 140, May 2003

15. Santosh Gaikwad, Bharti Gawali, Suresh C. Mehrotra, "*Novel Approach Based Feature Extraction For Marathi Continuous Speech Recognition*", Published in ACM Digital Library, ACM New York, NY, USA ©2012 Pages 795-804. ISBN: 978-1-4503-1196-0 /2012,

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

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

18. 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.

19. P. Satyanarayana, "Short segment analysis of speech for enhancement", Institute of IIT Madras February 2009.

20. Sadoki Furuki, Tomohisa Ichiba et.al, "Cluster-based Modeling for Ubiquitous Speech Recognition", Department of Computer Science Tokyo Institute of Technology Inter speech 2005.
21. Spector, Simon King, Joe Frankel, "Recognition, Speech production knowledge in automatic speech recognition", Journal of Acoustic Society of America, 2006.
22. M. A. Zissman, "Predicting, diagnosing and improving automatic Language identification performance", Proceeding of Eurospeech97, Sept.1997 Vol.1, pp.51-54 1989.
23. D. R. Reddy, "An Approach to Computer speech Recognition by direct analysis of the speech wave", Technical Report No.C549,Computer Science Department ,Stanford University,Sept.1996
24. Tzu-Chuen Lu, Ching-Yun Chang [2010], "A Survey of VQ Codebook Generation", Journal of Information Hiding and Multimedia Signal Processing, Ubiquitous International, Vol. 1, No. 3, pp. 190-203
25. Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Joydeep Ghosh [2005], "Clustering with Bregman Divergences", Journal of Machine Learning Research, Vol. 6, pp. 1705–1749
26. Benade A. H., (1976), "Fundamentals of musical acoustics", Oxford University Press, London.
27. Fant G., (1960). "Acoustic Theory of Speech Production". Mouton & Co, The Hague, Netherlands.
28. Yoav Meden, Eyal Yair and Dan Chazan, "Super Resolution Pitch Determination of Speech Signals", IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 39, NO 1, JANUARY 1991
29. H. Valbret, E. Moulines, J. P. Tubach, "Voice Transformation using PSOLA technique", Speech Communication 11 (1992) 175-187 North Holland.
30. Masanobu Abe, "Speech Morphing by Gradually Changing Spectrum Parameters and Fundamental Frequency", NTT Human Interface Laboratories.

31. Filter Bank [online] http://en.wikipedia.org/wiki/Filter_bank
32. Pattern Recognition [online] http://en.wikipedia.org/wiki/Pattern_recognition
33. C. S. Myers, 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
34. Ravi Prasad., Nagarjuna D, Ali Moulaei Nejad, "K -Nearest Neighbors Algorithm", Department of Study in Computer Science University of Mysore, Mysore, India 2009©
35. Travel, 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.
36. H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", IEEE Trans. Acoustics, Speech, Signal Proc.,ASSP-26(1).1978
37. Tutorial slides by Andrew Moore. Http://www.cs.cmu.edu/~awm
38. Burges C., "A tutorial on support vector machines for pattern recognition", In "Data Mining and Knowledge Discovery". Kluwer Academic Publishers, Boston, 1998, (Volume 2).
39. Nello Cristianini, John Shawe-Taylor, "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods", Cambridge University Press, 2000.
40. V. Vapnik, S. Golowich, A. Smola, "Support vector method for function approximation, regression estimation, and signal processing". Advances in Neural Information Processing Systems 9, pages 281– 287, Cambridge, MA, 1997. MIT Press
41. Vapnik V., "Estimation of Dependencies Based on Empirical Data", Empirical Inference Science: Afterword of 2006, Springer, 2006
42. M. A. Anusuya, S. K. Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 3, 2009

43. Keh-Yih Su, et.al, “Speech Recognition using weighted HMM and subspace”, IEEE Transactions on Audio, Speech and Language.
44. 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
45. S. Furui, [1986], “Speaker-independent isolated word recognition using dynamic features of speech spectrum”, IEEE Transactions on Acoustic, Speech, Signal Processing, Vol. 34, No. 1, pp. 52-59.
46. H. Sakoe, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transaction on Acoustics, Speech, Signal Proc.,ASSP-26(1).1978
47. Keh-Yih Su et.al, “Speech Recognition using weighted HMM and subspace”, IEEE Transactions on Audio, Speech and Language.
48. 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
49. 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
50. Baum L. E., Petrie, T., (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". The Annals of Mathematical Statistics 37 (6): 1554–1563. doi:10.1214/aoms/1177699147. Retrieved 28 November 2011
51. Rabiner L., Juang, B.H. (1986), “An Introduction to Hidden Markov Models”, IEEE ASSP Magazine, Vol. 3, No.1, Part 1, pp. 4-16
52. Flaherty, M. J. and Sidney T., (1994), “Real Time implementation of HMM speech recognition for telecommunication applications”, proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP), Vol. 6, pp. 145-148.
53. Picone J., (1990), “Continues Speech Recognition using Hidden Markov Models”, IEEE ASSP Magazine, Vol. 7, Issue 3, pp. 26-41.

54. 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.
55. Schalkoff Robert, "Artificial Neural Networks" Clemson University, McGraw-Hill, 1997
56. John Butzberger, "Spontaneous Speech Effect in Large Vocabulary speech recognition application", SRI International Speech Research and Technology program Menlo Park, CA94025
57. H Hermansky, N. Morgan, "RASTA processing of speech". IEEE Trans Speech Audio Process. 2, 578–589 (1994). doi:10.1109/89.326616
58. A.E Rosenberg, CH Lee, FK Soong, Cepstral channel normalization techniques for hmm-based speaker verification, in Proceeding of ICSLP, Yokohama, Japan, 1835–1838 (1994)
59. M. J. F. Gales, S. J. Young, "Robust speech recognition using parallel model combination", IEEE Trans Speech Audio Process. 4, 352–359 (1996). doi:10.1109/89.536929.
60. S. Mallat, "A wavelet tour of signal processing", Academic Press, 1998.
61. Y. T. Chan, "Wavelet Basics", Kulwer Academic Publications, ©1995.
62. J. S. Walker, "Wavelets and their Scientific Applications", Chamman and Hall/CRC, © 1999.
63. Daubechies, "Ten lectures on wavelets," society for industrial and Applied mathematics, 1992.
64. Nikhil Rao, "Speech compression using wavelets", ELEC 4801 THESIS PROGEC, School of Information Technology and Electrical Engineering, The university Of Queensland, October 2001.
65. Santosh K. Gaikwad, Bharti W. Gawali, Pravin Yannawar "A Review on Speech Recognition Technique", International Journal of Computer Applications (0975 – 8887), Volume 10– No.3, November 2010.

66. Silva J., Narayanan S., August 2007, "Minimum probability of error signal representation", International IEEE Workshop Machine Learning for Signal Processing
67. K. Nagata, Y. Kato, S. Chiba, "Spoken Digit Recognizer for Japanese Language", NEC Res. Develop., No.6,1963
68. 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.
69. Lawrence Rabiner, Biing Hwang Juang, "Fundamental of Speech Recognition", Copyright 1999by AT&T.