ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

A Comparative Study of Feature Extraction Techniques for Speech Recognition System

Pratik K. Kurzekar [1], Ratnadeep R. Deshmukh [2], Vishal B. Waghmare [2], Pukhraj P. Shrishrimal [2]

M.Tech Student, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS), India

ABSTRACT: The automatic recognition of speech means enabling a natural and easy mode of communication between human and machine. Speech processing has vast applications in voice dialing, telephone communication, call routing, domestic appliances control, Speech to Text conversion, Text to Speech conversion, lip synchronization, automation systems etc. Here we have discussed some mostly used feature extraction techniques like Mel frequency Cepstral Co-efficient (MFCC), Linear Predictive Coding (LPC) Analysis, Dynamic Time Wrapping (DTW), Relative Spectra Processing (RASTA) and Zero Crossings with Peak Amplitudes (ZCPA).Some parameters like RASTA and MFCC considers the nature of speech while it extracts the features, while LPC predicts the future features based on previous features.

KEYWORDS: Speech Recognition, Feature Extraction, Linear Predictive Coding (LPC), Mel Frequency Cepstrum Coefficient (MFCC), Zero Crossings with Peak Amplitudes (ZCPA), Dynamic Time Wrapping (DTW), Relative Spectra Processing (RASTA).

I. INTRODUCTION

Speech is the most common form of communication among the human beings. There are various languages in the world that are spoken by human beings for communication [1]. Researchers are trying to develop the system which can analyze and classify the speech signal [2]. The computers system which can understand the spoken language can be very useful in various areas like agriculture, health care and government sectors etc. Speech recognition refers to the ability of listening spoken words and identifies various sounds present in it, and recognizes them as words of some known language [3]. Speech signals are quasi-stationary signals. When speech signals are examined over a short period of time (5-100 msec), its characteristics are stationary; but, for a longer period of time the signal characteristics changes; it reflects to the different speech sounds being spoken. Features are extracted from the speech signals on the basis of short term amplitude spectrum (phonemes). Feature extraction is the most important phase in speech recognition system. There are some problems which are faced during the feature extraction process because of the variability of the speakers [4].
This paper gives the comparative study of some of the mostly used feature extraction techniques for Speech Recognition system. The rest of the paper is divided as follows: section 2 describes what is meant by feature extraction, section 3 describes commonly used feature extraction techniques, and section 4 compares the feature extraction techniques. Conclusion is mentioned in Section 5.

II. RELATED WORK

Over the years a number of different methodologies have been proposed for isolated word and continuous speech recognition. These can usually be grouped in two classes: speaker-dependent and speaker-independent.

Speaker dependent methods usually involve training a system to recognize each of the vocabulary words uttered single or multiple times by a specific set of speakers [5, 6] while for speaker independent systems such training methods are generally not applicable and words are recognized by analyzing their inherent acoustical properties [7,8]. Various features have been used singly or in combination with others to model the speech signals, ranging from Linear

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

Predictive Coding (LPC), Dynamic Time Wrapping (DTW), Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing with Peak Amplitude and Relative Spectra Filtering (RASTA).

1. Linear Predictive Coding (LPC):

Based on a highly simplified model for speech production, the linear prediction coding (LPC) algorithm is one of the earliest standardized coders, which works at low bit-rate inspired by observations of the basic properties of speech signals and represents an attempt to mimic the human speech production mechanism.

Thiang et. al. Developed speech recognition system using LPC for controlling movement of mobile robot. He used LPC as a feature extracting method because it is powerful method for speech analysis; therefore, it is fast and simple, yet an effective method for estimating the main parameters of speech signal [9].

Omesh Wadhwani et.al. recognised the vernacular language speech for discrete word using LPC technique. He used this method for better interpretation of spoken words [10].

Urmila Shrawankar et. al. stated that LPC is the most powerful and useful method for encoding quality speech at low bit rate. She stated, in LPC, a specific speech sample at current time can be approximated as linear combination of past speech sample.

2. Dynamic Time Wrapping (DTW):

Maruti Limkar et.al. developed an ASR system for English letter (0-9) using DTW. He stated that DTW is used to detect the nearest recorded voice and to discriminate the speech data model into respective classes [11].

Ingyin Khaing is used DTW for developing continuous speech recognition system for Myanmar language. He used DTW for feature clustering [12].

3. Mel-Frequency Cepstral Co-efficient (MFCC):

MFCC is one of the most powerful speech feature extraction technique and works on human auditory perception system. Vibha Tiwari developed text based speaker recognition system using MFCC [13].

Lindasalwa Muda et. al. is developed voice recognition system. He used MFCC as feature extraction technique [14].

4. Relative Spectra processing (RASTA):

Hajer Rahali is used RASTA method to extract the relevant information from the audio signal. The main goal of the work is to improve the robustness of speech recognition system in additive noise and real time reverberant environment [15].

Hynek Hermansky et. al. discussed relationship with human auditory perception system and extend the original metho to the combination of additive noise and convolution noise. He used band pass filter of time trajectories of logarithmic parameter of speech [16].

5. Zero Crossing with Peak Amplitude (ZCPA):

Young-Giu Jung et. al. is developed an optimal feature extraction for throat signal analysis using throat microphone. He used this technique to improve recognition rate of the system. This microphone is used to minimize the impact of environment noise [17].

Ying Sun developed an emotion speech recognition system using ZCPA. He proposed ZCPA for feature extraction by studying the length of frame and human auditory characteristics [18].

III. FEATURE EXTRACTION

Feature extraction is a basic and fundamental pre processing step to pattern recognition and machine learning problem. It's a special form of dimensionality reduction technique which is used to reduce the data which is very large to be processed by an algorithm. In feature extraction the provided input data is transformed into a set of features which provides the relevant information for performing a desired task without the need of the full size data but using the reduced set. The speech recognition technique is having a background of DSP i.e. Digital signal processing. DSP is the centre of progress in speech processing during the complete development of the speech processing and speech recognition systems [19].

Feature Extraction is not only used in speech analysis, synthesis, coding, recognition and enhancement but also in voice modification, speaker recognition and language identification. Theoretically it is possible to recognize speech directly from the digital waveform of the speech. However, as speech is time varying the idea to perform some form of feature extraction came into existence which is used to reduce the variability of speech signal. In the context of Automatic speech recognition feature extraction is the process of retaining the useful information from the speech signal while the

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

unnecessary and unwanted information is removed which involves the speech signal analysis. However, while removing the unwanted information from the speech signal some useful information may also lose.

The main objective of feature extraction is to untangle the speech signal into the different acoustically identifiable components and to obtain the set of feature with low rate of change in order to keep the computation feasible. The feature extraction for speech recognition can be divided in spectral analysis, parametric transformation and statistical modelling. This phenomenon is shown in the following fig.1.

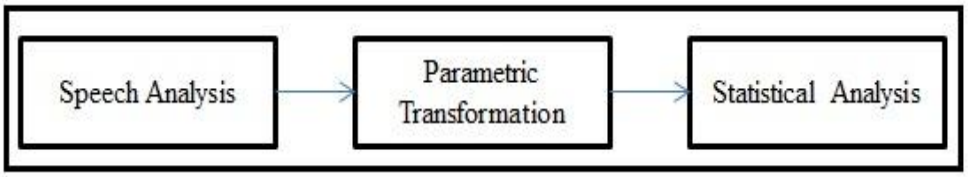

Fig.1 Feature Extraction Process [20]

When speech is produced in the sense of time varying signal, its characteristics can be represented via parameterization of the spectral activity. There are six major classes of spectral analysis algorithms i.e. Digital filter bank (Power estimation), Fourier Transform (FT Derived Filter Bank Amplitudes, FT Derived Cepstral Coefficients), Linear Prediction (LP, LP Derived Filer Bank Amplitudes, LP Derived Cepstral Coefficients) used in speech recognition system**.** Signal parameters are generated from signal measurements through two fundamental operations: differentiation and concatenation. The output of this stage of processing is a parameter vector containing our raw estimates of the signal.

The third step of the feature extraction process is Statistical Modeling. Here, it assumes that the signal parameters are generated from some underlying multivariate random process. To learn the nature of this process, it impose a model on the data, optimize (or run) the model, and then measure the quality of the approximation. The only information about the process is its observed outputs of the signal parameters that have been computed. For this reason, the parameter vector output from this stage of processing is often called the signal observations. A statistical analysis is to be performed on the vectors to determine if they are part of a spoken word or phrase or whether they are merely noise.

IV. COMMONLY USED FEATURE EXTRACTION TECHNIQUES FOR SPEECH

In speech recognition, the main goal of the feature extraction step is to compute a sequence of feature vectors providing a compact representation of the given input signal. Commonly LPC, MFCC, ZCPA, DTW and RASTA are used as feature extraction techniques for speech recognition system.

1. LINEAR PREDICTIVE CODING (LPC)

The basic idea behind the Linear Predictive Coding (LPC) analysis is that a speech sample can be approximated as linear combination of past speech samples.LPC is a frame based analysis of the speech signal which is performed to provide observation vectors of speech.LPC feature extraction process can be explained from following figure. The input speech signal digitized spectrally flatten speech signal is put through a low order digital system to make it less susceptible to finite precision effects later in the signal processing [21].

To compute LPC features, initially the speech signal is blocked into frames of N samples. The output of the pre-emphasizer network is related to the input to the network. After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. Typical window is the Hamming window.

The next step is to auto correlate each frame of windowed signal Where the highest autocorrelation value is the order of the LPC analysis. The next processing step is the LPC analysis, which converts each frame of autocorrelations into LPC parameter set by using Durbin's method.LPC cepstral coefficients, is a very important LPC parameter set, which can be derived directly from the LPC coefficient set.

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

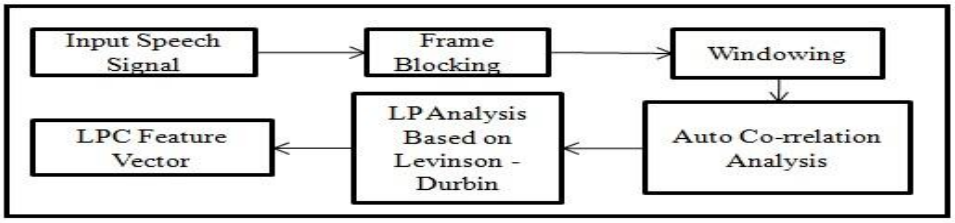

Fig.2: LPC Feature Extraction Process

Linear prediction is a mathematical operation which provides an estimation of the current sample of a discrete signal as a linear combination of several previous samples. The prediction error i.e. the difference between the predicted and actual value is called the residual. If the current sample x_i of the audio signal be predicted by the past p samples and x_i' is the predicted value then we have:

$$x_i' = -a_2 x_{i-1} - a_3 x_{i-2} - \cdots - a_{p+1} x_{i-p}$$

Here $\{1, a_2, \ldots, a_{p+1}\}$ are the (p+1) filter coefficients. In this case the signal is passed through an LPC filter which generates a element feature vector and a scalar which represents the variance of the predicted signal [22].

2. MEL-FREQUENCY CEPSTRUM CO-EFFICIENT (MFCC)

The Mel-frequency Cepstrum Coefficient (MFCC) technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. The signal is divided into overlapping frames to compute MFCC coefficients. Let each frame consist of N samples and let adjacent frames be separated by M samples where M<N. Each frame is multiplied by a Hamming window where the Hamming window equation is given by:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$$

In the third step, the signal is converted from time domain to frequency domain by subjecting it to Fourier Transform. The Discrete Fourier Transform (DFT) of a signal is defined by the following:

$$X_k = \sum_{i=0}^{N-1} x_i e^{-\frac{j2\pi ki}{N-1}}$$

In the next step the frequency domain signal is converted to Mel frequency scale, which is more appropriate for human hearing and perceptions. This is done by a set of triangular filters that are used to compute a weighted sum of spectral components so that the output of the process approximates a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. The following equation is used to calculate the Mel for a given frequency:

$$M = 2595 log_{10}\left(1 + \frac{f}{700}\right)$$

In the next step the log Mel scale spectrum is converted to time domain using Discrete Cosine Transform (DCT). DCT is defined by the following, where is a constant dependent on N:

$$X_k = \alpha \sum_{i=0}^{N-1} x_i \cos\left\{\frac{(2i+1)\pi k}{2N}\right\}$$

The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficients is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vectors.

A block diagram of the MFCC processes is shown in Figure. Block diagram of MFCC The speech waveform is cropped to remove silence or acoustical interference that may be present in the beginning or end of the sound file. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts each frame from the time domain to the frequency domain. In the Mel-frequency wrapping block, the signal is plotted against the Mel spectrum to mimic human hearing. In the final step, the Cepstrum, the Mel-spectrum scale is converted back to standard frequency scale. This spectrum provides a good representation of the spectral properties of the signal which is key for representing and recognizing characteristics of the speaker.

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

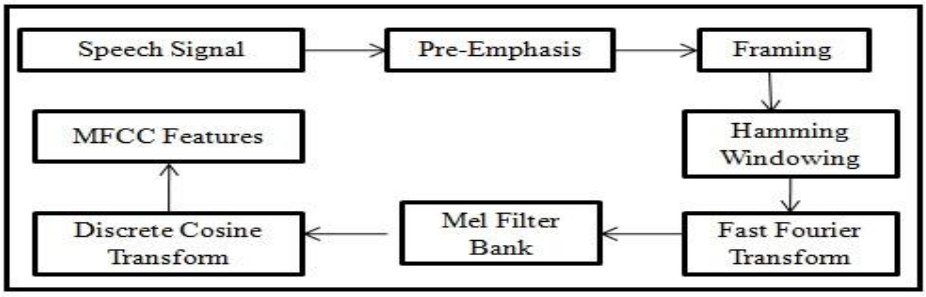

Fig.3 Block Diagram of MFCC Feature Extraction Techniques [23]

3. ZERO CROSSINGS WITH PEAK AMPLITUDES (ZCPA)

This feature extraction technique is based on Human Auditory System. It uses zero-crossing interval to represent signal frequency information and amplitude value to represent intensity information, finally frequency information and amplitude information is combined to form the complete feature output. The following figure shows ZCPA principle diagram for feature extraction. Block diagram consist of Band Pass Filters (BPF) block, zero-crossing detection block, peak detection block, non-linear amplitude compression block and the frequency receiver block [24].

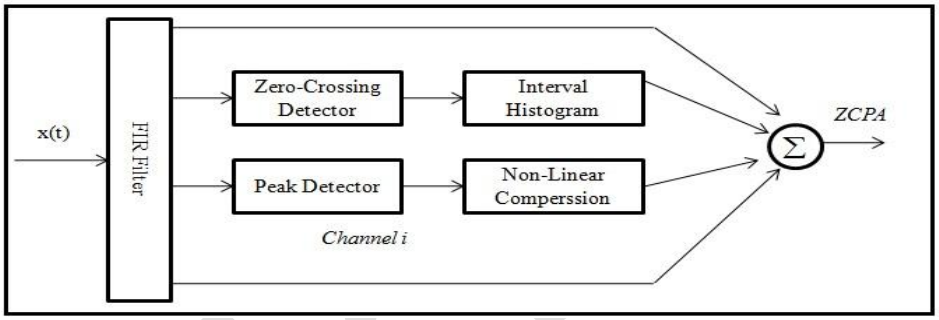

Fig.4 Principle diagram of ZCPA for Feature Extraction Techniques

The filters block is based on perception of human auditory system, containing 16 BPF, and covers the frequency range of 200-4000Hz. The original speech data pass the filter bank and are transformed to 16different data processing paths. Speech signal is divided into frames and each frame undergoes the upward going zero-crossing interval detection and peaks detections in each interval. The non-linear compression of peak value uses the equation. The equation is a monotony function; x represents a peak in the upward going zero-crossing interval. After it is compressed logarithmically the result is g(x). This process is to mimic the biological relationship between the hearing nerves stimulating intensity and the lock-up intensity.

$$g(x) = \log(1.0 + 20x)$$

Frequency receiving block is to divide the frequency band into several sub-bands, each band is called a frequency bin. The 16 path outputs formed together the ZCPA feature.

The frequency information and the intensity information are combined by the frequency receiver. Separate the frequency band of speech signal into several sub-bands by ERB-rate coordinate, and each sub-band is called a frequency bin. This paper uses 16 frequency bins, the frequency range is from 156Hz to 4055Hz. Convert the upward zero-crossing into the right frequency of frequency bins, and the frequency information can be known by looking up the number of the frequency bins[25].

In normalization of the time and amplitude section, it is necessary to normalize the time and the amplitude to satisfy the training and testing of Support Vector Machine (SVM) algorithm. The output feature vector can represent the pitch and intensity features of speech. Because pitch and intensity features are typical features of the expression in speech.

4. DYNAMIC TIME WARPING (DTW)

DTW algorithm is based on Dynamic Programming. This algorithm is used for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be “warped” non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

similarity between the two time series. To align two sequences using DTW, an n*m matrix where the (ith, jth) element of the matrix contains the distance d (qi, cj) between the two points qi and cj is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation as shown in equation:

$$d(qi, cj) = 2*(qi - cj)$$

Each matrix element (i, j) corresponds to the alignment between the points qi and cj. Then, accumulated distance is measured by equation.

$$D(i, j) = \min [D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i, j)$$

This algorithm performs a piece wise linear mapping of the time axis to align both the signals. The best match or alignment between these two sequences is the path through the grid, which minimizes the total distance between them, which is termed as Global distance. The overall distance (Global distance) is calculated by finding and going through all the possible routes through the grid, each one compute the overall distance. The global distance is the minimum of the sum of the distances (Euclidean distance) between the individual elements on the path divided by the sum of the weighting function. For any considerably long sequences the number of possible paths through the grid will be very large. Global distance measure is obtained using a recursive formula [26].

$$GDxy = LDxy + \min (GDx-1\ y-1, GDx-1\ y, GDx\ y-1)$$

Here, GD = Global Distance (overall distance),

LD = Local Distance (Euclidean distance).

5. RELATIVE SPECTRAL PROCESSING (RASTA)

Relative spectral processing [RASTA] based speech enhancement involves linear filtering of the trajectory of the short-term power spectrum of noisy speech signal, as shown in Figure. The spectral values of input speech signal are compressed by a nonlinear compression rule (a=2/3) before performing the filtering operation and expanded after filtering (b=3/2).

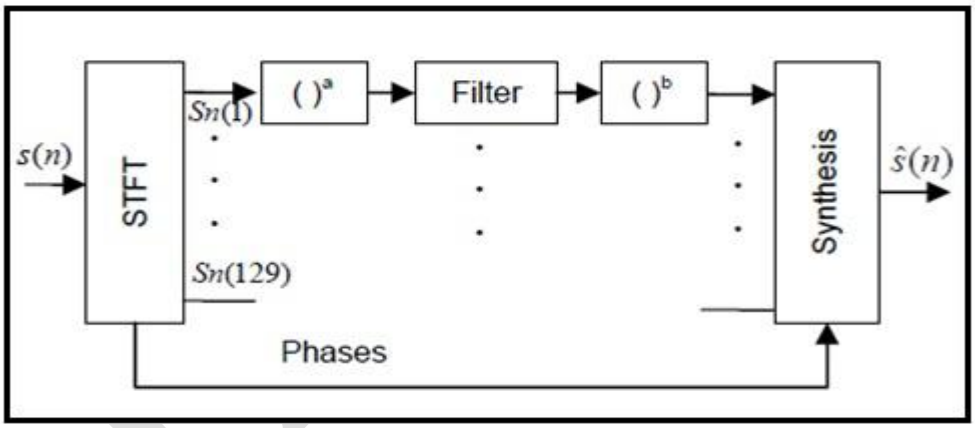

Fig.5 Block Diagram of RASTA Method

To obtain better noise suppression for communication systems the fixed RASTA filters were replaced by abank of non-causal FIR Wiener-like filters. For 256point FFT, 129 unique filters are required. The output ofeach filter is given as

$$S_i(k) = \sum_{j=-M}^{M} w_i(j) Y_i(k-j)$$

Here, $S_i(k)$is estimate of clean speech in frequency bin "i" and frame-index "k", $Y_i(k)$is noisy speech spectrum,$w_i(j)$are the weights of the filter and M is order of the filter. In this method the weights$w_i(j)$ are obtained suchthat $S_i(k)$is least square estimate of clean speech $S_i(k)$or each frequency bin *i*. The order M = 10 corresponds to 21 tap non-causal filters. The filters were designed based on optimization on 2 minutes of speech of a male speaker recorded at 8 kHz sampling over public analog cellular line from a relatively quiet library [13]. The published response of the filter corresponding to bins in the frequency range 300Hz to 2300 Hz is a band-pass filter, emphasizing modulation frequency around 6-8 Hz. Filters corresponding to the 150-250 Hz and 2700- 4000 Hz regions are low gain,

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

low-pass filters with cut off frequency of 6 Hz. For very low frequency bins (0-100 Hz) the filters have flat frequency response with 0 dB gain.

V. COMPARISON OF FEATURE EXTRACTION TECHNIQUES

The following table shows the above discussed techniques used for the analysis of the developed speech databases in different languages and accuracy achieved by the developed Automatic Speech Recognition (ASR) System.

Sr. No.	Technique	Language	Database Used	Recognition Rate	Application
1	Linear Predictive Analysis (LPC)	English [27]	Developed during research	91.4%	Controlling the movement of mobile robot
		Vernacular	Developed during research	App. 90%	To overcome the limitations on software associated with speech recognition
		English Numerals (0-9)	Developed during research	94%	To recognize English words corresponding to digits Zero to Nine
		English	Developed during research	97.3%	To reduce effects of variations in the frequency response of telephone connections
		English	Developed during research	97.5%	Controlling the movement of mobile robot
		Devnagari [28]	Developed during research	82.3%	To find out better technique for continuous speech using PCA
2	Mel-Frequency Cepstral Coefficient (MFCC)	English Numerals (0-9) [29]	Developed during research	90.5%	Isolated digit recognition
		English Numerals (0-9)	AURORA 2	92.93%	New algorithm for extracting MFCC to make hardware implementation more efficient
		Urdu [30]	Developed during research	86.67%	Voice recognition system for security purpose
		Arabic Numerals (0-9) [31]	Developed during research	97.66%	Telephone Dialling, Airline reservation
		Devnagari	Developed during research	85.3%	To find out better technique for continuous speech using PCA
		English [32]	TIDIGITS	99.9%	Front-end for automatic speech recognition system
3	Zero-Crossing with Peak Amplitude (ZCPA)	English [33]	Developed during research	96.67%	Speaker identification and isolated letter recognition
		Korean [34]	Developed during research	97.14%	Throat signal analysis
		Chinese [35]	Developed during research	91.1%	Robust speech recognition system
		Korean [36]	Developed during research	87%	Front-end for automatic speech recognition system
		English	TIDIGITS	95.4%	Front-end for automatic speech recognition system
		English[37]	Developed during	90.9%	Robust speech recognition

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

			research		system
		Devnagari	Developed during research	38.5%	To find out better technique for continuous speech using PCA
4	Dynamic Time Wrapping (DTW)	English	Developed during research	90.5%	Isolated digit recognition
		Arabic [38]	Developed during research	98.5%	Arabic speech recognition for isolated words
		English Numeral (1-5) [39]	Developed during research	91.1%	Voice recognition system
		English [40]	Developed during research	75.19%	Recognizing voice command for robot
		Arabic [41]	Developed during research	77%	Isolated digit recognition
		English [42]	Developed during research	83%	Template matching for continuous speech recognition system
5	Relative Spectral (RASTA) Filtering	English	Developed during research	95.92%	Isolated digit recognition
		English [43]	AURORA	94.27%	To improve the robustness of speech recognition systems in additive noise and real-time reverberant environments
		Spanish [44]	Developed during research	93.90%	Robust automatic speech recognition system
		English [45]	TIDIGIT, AURORA	90%	Impulsive noisy speech recognition by relative spectral analysis
		English [46]	AURORA-2, AURORA-3	78.8%	Automatic speech recognition for auditory processing

Table 1--Table for Different Speech Feature Extraction Techniques for Different Languages with Recognition Rate

The above table shows developed speech recognition systems using different techniques with recognition rate for different languages. From table we can conclude that database developed for English language has more recognition rate compared to other language. The recognition rate for Automatic Speech Recognition System for English language is above 90% while for other language is in between 80-90%. During the study we observed that maximum work is done for isolated word recognition system while negligible amount of work is done for continuous speech recognition system.

Linear predictive coding (LPC) is one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters. LPC is generally used for speech analysis and synthesis. We observed from the study that LPC technique is commonly used in electrical and musical companies for making mobile robots, in telephone companies, tonal analysis of violins and other string musical instruments etc.

Mel Frequency Cepstral Coefficient's (MFCC) is most commonly used features extraction technique in speech recognition systems. The reason for MFCC being most commonly used for extracting features is that it is most nearest to the actual human auditory speech perception. MFCC is used to recognize numbers automatically spoken into a telephone, airline reservation, voice recognition system for security purpose etc. Some researchers have proposed modifications to the basic MFCC algorithm to improve robustness, such as by raising the log-mel-amplitudes to a suitable power (around 2 or 3) before taking the DCT, which reduces the influence of low-energy components.

Zero Crossing Peak Amplitude (ZCPA) is similar to MFCC and mostly used for development of automatic speech recognition in noisy environments, speaker identification, throat signal analysis, development of noise robust speech recognition system etc. ZCPA provides more accurate and clear description to speech signal, and it does not exhibit the information redundancy.

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

Dynamic Time Wrapping (DTW) has been applied to temporal sequences of video, audio, and graphic. DTW is commonly used for measuring similarity between two temporal sequences which may vary in time or speed. It is used to cope with different speaking speed, speaker recognition, online signature recognition, and generally used to solve time alignment problem, template matching etc. Also it is observed that it can be used in partial shape matching application. DTW is a method that can calculate an optimal match between two given sequences with certain restrictions. DTW is used to align the utterances properly and calculating the minimum distance between two utterances or samples. It is an efficient method to solve the time alignment problem.

Relative Spectra Filtering (RASTA) was originally developed to reduce the sensitivity of recognizers to frequency characteristics of an operating environment. RASTA method is generally used for speech analysis in which speech signals are enhanced, to develop noise robust speech recognition system and etc.

VI. CONCLUSION

In this paper we have studied few of the features extraction techniques used for the development of Automatic Speech Recognition System. During the study it was observed that technique has some limitation, sometimes reliability and performance of the system is affected. It was also observed that the maximum work in the area of Automatic Speech Recognition is for English language. Less work has been carried out for Indian languages. Even the recognition rate for the ASR is higher for English language than any other language. The recognition rate for the Indian language is very low due to the phonetic nature of Indian language. It was also observed that the researchers have tried single techniques for the recognition; there is a need to develop hybrid approach which may give better performance for the development of robust speech recognition area. This paper provides the comparative study of the commonly used feature extraction techniques for speech recognition. This paper will help the researchers willing to work in the area of speech recognition know the basic difference between the discussed feature extraction techniques.

VII. ACKNOWLEDGEMENT

The authors would like to thank the University Authorities for providing the infrastructure to carry out the research.

REFERENCES

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers.html,Viewed 1st June 2014.
[2] Pukhraj P. Shrishrimal, Ratnadeep R. Deshmukh, and Vishal B. Waghmare, "Indian Language Speech Database: A Review", International Journal of Computer Applications (0975 – 888) Volume 47– No.5, pp. 17-21, June 2012.
[3] Bharti W. Gawali, Santosh K. Gaikwad, Pravin Yannawar, and Suresh C. Mehrotra, "Marathi Isolated Word Recognition System using MFCC and DTW Features", ACEEE International Journal on Information Technology, Vol. 01, No. 01, pp.21-24, Mar 2011.
[4] Urmila Shrawankar, and Dr. Vilas Thakare, "Techniques for Feature Extraction in Speech Recognition System: A Comparative Study", International Journal of Computer Applications in Engineering, Technology and Sciences (IJCAETS), ISSN 0974-3596, pp 412-418, 2010.
[5] M. B. Herscher, and R. B. Cox, "An adaptive isolated word speech recognition system", Proceding Conference on Speech Communication and Processing, Newton, MA, pp.89-92, 1972.
[6] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Transaction on Acoustics, Speech and Signal Processing (ASSP), Vol.23,pp.67-72, 1975.
[7] V. N. Gupta, J. K. Bryan, and J. N. Gowdy, A speaker-independent speech recognition system based on linear prediction, IEEE Transactions on Acoustics, Speech, Signal Processing (ASSP), Vol. 26, pp. 27-33, 1978 27-33.
[8] L. R. Rabiner, J. G. Wilpon, Speaker independent isolated word recognition for a moderate size vocabulary, IEEE Transactions on Acoustics, Speech, Signal Processing (ASSP),Vol. 27, pp. 583-587, 1979.
[9] Thiang, and Wanto, "Speech Recognition Using LPC And HMM Applied for Controlling Movement of Mobile Robot" Seminar National Teknologi Informasi, pp. 67-72, 2010.
[10] Omesh Wadhwani, and Amit Kolhe, "Recognition of Vernacular Language Speech for Discrete Words Using LPC Technique", Journal of Global Research in Computer Science, Vol.2, No. 9, pp.25-27, September 2011.
[11] Maruti Limkar, Rama Rao, and Vidya Sagvekar, "Isolated Digit Recognition Using MFCC and DTW", International Journal on Advanced Electrical and Electronics Engineering, (IJAEEE), ISSN (Print): 2278-8948, Vol.1, Issue-1, pp.59-64, 2012.
[12] Ingyin Khaing, "Myanmar Continuous Speech Recognition System Based on DTW and HMM", International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 Issue 1, pp.78-83, February 2013.
[13] Vibha Tiwari, "MFCC and Its Applications in Speaker Recognition", International Journal on Emerging Technologies, Vol.1 Issue 1, pp. 19-22, 2010.

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

[14] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, ISSUE 3, pp. 138-143,MARCH 2010.
[15] Hajer Rahali, Zied Hajaiej, and Noureddine Ellouze, "Robust Features for Impulsive Noisy Speech Recognition Using Relative Spectral Analysis", International Journal of Computer, Information, Systems and Control Engineering Vol.8 No.9, pp. 1421-1426, 2014.
[16] Hynek Hermansky, and Nelson Morgan, "RASTAA Processing of Speech", IEEE Transactions On Audio, Speech & Language Processing, Vol 2, No. 4, pp. 578-589, October 1994.
[17] Young-Giu Jung, Mun-Sung Han, and Sang Jo Lee, "Development of an Optimized Feature Extraction Algorithm for Throat Signal Analysis", Electonics and Telecommunication research Institute Journal, Volume 29, Number 3, pp.292-299, June 2007.
[18] Ying Sun, Jiemin Yin, and Xueying Zhang, "Study for Classification of Emotional Speech by using Optimized Frame Zero Crossing with Peak Amplitudes Feature Extraction Algorithm", Journal of Computational Information Systems Vol.7 Issue 10, pp.3508-3515, 2011.
[19] Bhupinder singh, Rupinder Kaur, Nidhi Devgun, and Ramandeep Kaur, "The process of feature extraction in automatic speech recognition system for computer Machine interaction with humans: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 2, Feb 2012.
[20] Kashyap Patel, and 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, pp. 33-37, May 2013.
[21] Eslam Mansour mohammed, Mohammed Sharaf Sayed, Abdallaa Mohammed Moselhy, and Abdelaziz Alsayed Abdelnaiem, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 6, No. 3,pp. 55-66, June, 2013.
[22] Bishnu Prasad Das, and Ranjan Parekh, "Recognition of Isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", International Journal of Modern Engineering Research (IJMER) Vol.2, Issue.3, pp-854-858 ISSN: 2249-6645,pp. 854-858, May-June 2012.
[23] E. Chandra, K. Manikandan, and M. Sivasankar, "A Proportional Study on Feature Extraction Method in Automatic Speech Recognition System", International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering, Vol. 2, Issue 1, pp. 772-775, January 2014.
[24] Handy K. Elminir, Mohamed Abu El-Soud, and L. M. Abou El-Maged, "Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition", International Journal of Science and Technology, ISSN 2224-3577, Volume 2 No.10, October 2012.
[25] Shivanker Dev Dhingra, Geeta Nijhawan, Poonam Pandit, "Isolated Speech Recognition Using MFCC and DTW", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Issue 8, pp. 4085-4092, August 2013.
[26] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition", International Journal for Advance Research in Engineering and Technology, Volume 1, Issue VI, pp. 1-5, July 2013.
[27] Ram Singh, and Preeti Rao, "Spectral Subtraction Speech Enhancement with RASTA Filtering", Proceding of National Conference on Communications (NCC), Kanpur, India, 2007.
[28] Thiang, and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot", International Conference on Information and Electronics Engineering IPCSIT vol.6, IACSIT Press, pp.179-183, Singapore, 2011.
[29] Fumitada Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition", IEEE Transaction on Acoustic,Speech and Signal Processing, pp.854-858, February 1975.
[30] Anuradha S. Nigade, and J. S. Chitode, "Throat Microphone Signals for Isolated Word Recognition Using LPC", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 8, pp.906-913, August 2012.
[31] Garima Vyas, and Barkha Kumari, "Speaker Recognition System Based On MFCC and DCT", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Vol.2, Issue-5, pp. 145-148, June 2013.
[32] Shumaila Iqbal, Tahira Mahboob, and Malik Sikandar Hayat Khiyal, "Voice Recognition using HMM with MFCC for Secure ATM", International Journal of Computer Science Issues (IJCSI), Vol. 8, Issue 6, No 3, pp.297-303, November 2011.
[33] Santosh Gaikwad, Bharti Gawali, and Pravin Yannawar, "Performance Analysis of MFCC & DTW for Isolated Arabic Digit", International Journal of Advanced Research in Computer Science, Vol. 2 (1), pp. 513-518 Jan. –Feb, 2011.
[34] Serajul Haque, Roberto Togneri, and Anthony Zaknich, "Zero-Crossings with Adaptation for Automatic Speech Recognition", Proceeding of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson.University of Auckland, New Zealand. December 6-8, pp. 199-204, 2006.
[35] J. Manikandan, and B. Venkataramani, "Design of a real time automatic speech recognition system using Modified One Against All SVM classifier" Microprocessors and Microsystems, ELSEVIER, Vol.35, pp. 568–578, 2011.
[36] Xueying Zhang, and Wuzhou Liang, "A Robust Speech Recognition Based on the Feature of Weighting Combination ZCPA", Proceeding of the First International Conference on Innovative Computing, Information and Control (ICICIC'06) 0-7695-2616-0/06, pp. 2006.
[37] Doh-Suk-Kim, Jae-HoonJeong, Soo-Young Lee, and Rhee M. Kil, "Comparative Evaluations of Several Front Ends for Robust Speech Recognition", Advance Institute of Science and Technology Korea, 1997.
[38] Serajul Haque, Roberto Togneri, and Anthony Zaknich, "Zero Crossings with Peak Amplitudes and Perceptual Features for Robust Speech Recognition", School of Electrical, Electronic and Computer Engineering University of Western Australia, 2012.
[39] Khalid A. Darabkh, Ala F. Khalifeh, Baraa A. Bathech, and Saed W. Sabah, "Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language", International Scholarly and Scientific Research & Innovation, pp. 106-113, Vol: 7 2013-05-25.
[40] Sahil Verma, Tarun Gulati, and Rohit Lamba, "Recognizing Voice For Numeric's Using MFCC and DTW", International Journal of Application or Innovation in Engineering & Management, Vol.2, Issue 5, pp.127-130, May 2013.
[41] Nidhi Desai, Kinnal Dhameliya, and Vijayendra Desai, "Recognizing voice commands for robot using MFCC and DTW", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 5, pp. 6456-6459, May 2014.
[42] Z. Hachkar1, A. Farchi, B.Mounir, and J. El Abbadi, "A Comparison of DHMM and DTW for Isolated Digits Recognition System of Arabic Language", International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3 No. 3, pp.1002-1008, Mar 2011.
[43] Mathias De Wachter, Mike Matton, Kris Demuynck, Patrick Wambacq, Member, Ronald Cools, and Dirk Van Compernolle, "Template Based Continuous Speech Recognition", IEEE Transactions On Audio, Speech & Language Processing, pp. 898-901, 2007.

DOI: 10.15680/IJIRSET.2014.0312034

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2014

[44] Hajer Rahali, Zied Hajaiej, and Noureddine Ellouze, "Robust Features for Speech Recognition using Temporal Filtering Technique in the presence of Impulsive Noise", I.J. Image, Graphics and Signal Processing, Vol.11, pp.17-24, 2014.

[45] Pere Pujol Marsal, Susagna Pol Font, Astrid Hagen, H. Bourlard, and C. Nadeu, "Comparison And Combination Of Rasta-Plp And Ff Features In A Hybrid Hmm/Mlp Speech Recognition System", Speech and Audio Processing, IEEE Transactions on Vol.13, Issue: 1, 20 December 2004.

[46] Marcus Holmberg, David Gelbart, Werner Hemmert, "Automatic Speech Recognition with an Adaptation Model Motivated by Auditory Processing", Audio, Speech, and Language Processing, IEEE Transactions on Vol. 14, Issue. 1, 19 December 2005.

DOI: 10.15680/IJIRSET.2014.0312034