A Review on Different Feature Recognition Techniques for Speech Process in Automatic Speech Recognition.

Ashwini P, Dr. Bharathi S.H . Ananya.K.Nayaka

ABSTRACT- Speech is the fundamental form of communication in human being because of which speech processing has evolved and exists as an everlasting limb of speech processing. Automatic speech recognition is one of the recent research areas in speech processing field. This paper presents a broader review about different speech recognition techniques. Few of the discussed techniques are MFCC, LPC, PLP, PLDA, and RASTA. Among all theses feature extraction techniques MFCC and LPC are more widely used because of their nearness to the original speech signal.

KEYWORDS- Automatic Speech Recognition (ASR) , Feature Extraction, MFCCs, LPC, RASTA, PLDA and PLP.

I. INTRODUCTION

Speech is the important way of communication. Speech processing is one of the most rousing research areas under signal processing. The signals are generally processed in digital domain; hence speech processing can also be distinctively called as digital signal processing appertained to speech signal. Automatic Speech Recognition (ASR) is a computer speech recognition system. It is a course of action of converting speech signal into series of words and other lingual units with help of algorithms which could be implemented as computer programs. The predominant objective of ASR is to develop different techniques and also system to enable the computers to identify speech signals which are fed as input. Speech recognition and its applications have evolved from past few decades. In any of the speech recognition system the speech signal is converted into text form out of which, the text form will be the output from ASR and this text will be almost equivalent to the speech fed as input. This recognition has its procesecution in voice search, voice dialling, robotics etc. most of the speech recognition systems are working on HMMs-Hidden Markov Models. The important aspect of HMM being extensively used in HMM is the parameters which can easily and automatically being erudite and trained. It is also computationally practical to use. Though many forge on have been made in field of ASR still we are impotent to develop machine which can understand all kinds of human speech in any environment. In this paper we discuss about different feature recognition techniques which will help in ASR.[1-3]

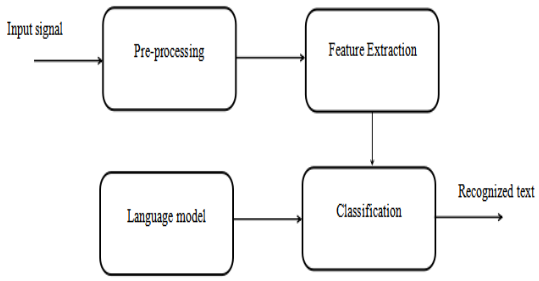

Fig.1 Speech Recognition Stages

II. SPEECH-RECOGNITION TECHNIQUES:

The manifest of an ASR is to have an ability to follow the speech and then function on spoken words. A speech recognition system .embrace for stages which can be systematized as displayed in depiction

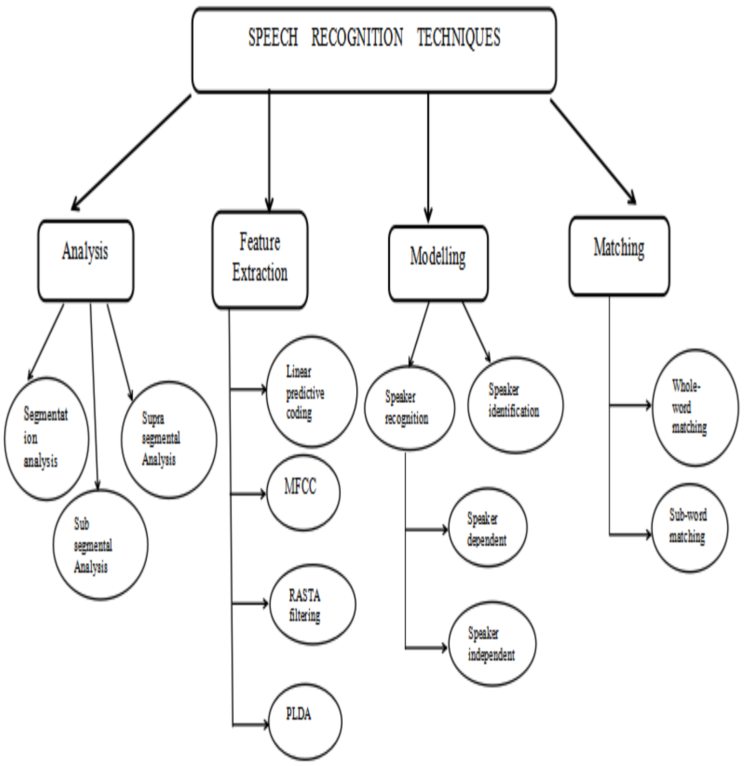

Fig 2: Speech Recognition Technique [2]

1. **Analysis**: The primary stage in speech recognition is analysis. When a presenter is made to speak, different types of information will be contributed in identification of speaker. The information varies because of various reasons like. Vocal track, origin of excitation and also behavioural features. Further speech analysis can be categorized in 3 analyses: Segment analysis, Sub-segmental Analysis, Super-segmental Analysis[3]

2. Feature extraction techniques

Feature Extraction is the principal sector of the ASR. It is contemplated as cardinal of the speech recognition system. The task of this technique is to extricate the properties from the speech signal fed as input. Which the computer to recognize the expounder. Feature extraction abbreviates the magnitude of the input (vector) without destructing the power of the input speech. Signal there are different feature extraction approaches for speech recognition system which will be discussed below in the consecutive sections.[2]

2.1. Mel-Frequency cepstral co-efficient (MFCC)

Mel Frequency cepstral co-efficient is most prevalent technique and to bring out spectral features. MFCC employed in speech recognition is rooted on frequency domain utilizing Mel scale found on the human ear scale. It is one of the acquired techniques for feature extraction. MFCCs which works on frequency domain is in fact more pinpoint compared to time domain features using techniques [2]

Human speech is corollary of the frequencies is not linear in character as a result the pitch of an acoustic Speech signal of a particular frequency is mapped into a “Mel” scale[4][5]. Any frequencies spacing under 1khz is linear and the frequencies beyond 1khz will be logarithms hertz frequencies Mel frequencies can be calculated with the help of equation(1) from [6]

$$\text{fmel} = 2596*\log\left(1+\frac{f}{700}\right) \qquad (1)$$

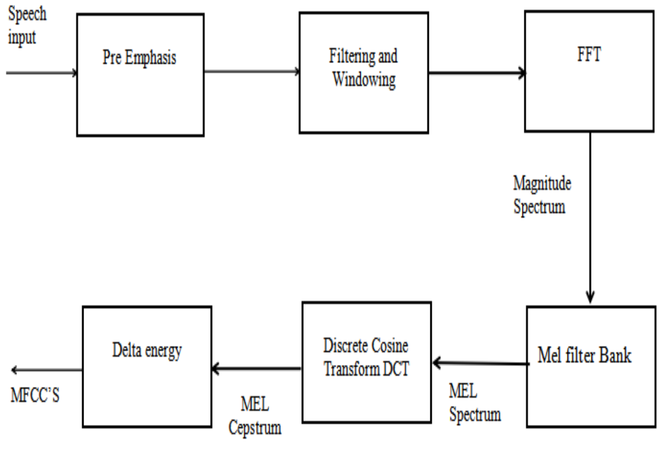

Fig. 3 Block Diagram for MFCC Computation [1]

The above block diagram shown different stages in MFCCs computation is depicted. At pre-emphasis stage speech signals are recorded which are processing at sampling rate of 16 kHz. Each voice or audio samples are stored into distinct audio files. This step performs lifting of energy by pre emphasis at high frequencies. The pre emphasis is filter’s difference equation is given below in equation 2

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0+b_1z^{-l}}{1} = 1\text{-}0.97z^{-l} \qquad (2)$$

The next step in MFCC is framing and windowing where non stationary speech signal is divided into little chunk. Which are over lapping with one another, this process is called as framing Followed by framing windowing will be performed to eliminate the disruption at boundaries of frame Most frequently used window is hamming window whose expression is as below in equation 3

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] 0 \le n \le N-1 \\ 0 \qquad \text{otherwise} \end{cases} \qquad (3)$$

Where N is summed quantity of sample in a single frame.

To the same signal out of windowing fast Fourier transform is applied to calculated discrete Fourier transform (DFT) because of which the signal is converted into frequency domain Equation 3 gives FFT [7].

$$x[k] = \sum_{n=0}^{N-1} x(n)e^{-j2\frac{\pi}{N}kn} \qquad (4)$$

In the next step i.e. when signal is passed through Mel filter Bank which is made up of triangle shaped overlapping filter the Hertz signal will be converted into Mel scale [8]. The DCT of the output from Mel filter bank is extracted. The output obtained out of DCT will be progress through Delta energy converter where in base 10 logarithms of the output will be obtained. The main reason behind thi9s conversion is because human ear can acknowledge to acoustic speech signal is not linear. The human ear is merely sensitive to variations in amplitude higher frequencies. Energy consumption is given by

$$E = \sum_{t=t1}^{t=t2} x^2(t) \qquad (5)$$

Few of the advantages of MFCC are the identification accuracy is high which leads to greater performance. It can apprehend important features with less complexity. But MFCC fails to give more accurate output in the presence of noise. As filter banks number varies performance get altered.

2.2. Linear predictive coding

Linear predictive coding is an arithmetical computation operation which is linear blending of a great many former samples. It is method of obtaining equal roughly fundamental parameters of language. [8][9] It provides exact approximation of language parameters. The primary concept behind LPC is that a language sample could be evaluated as a linear mixture of former language samples. One set of variable co-efficient can be found by reducing the summing of squared variances amid real language samples and contemplated values. The co-efficient acquired will be the basis for LPC [11]. Figure for LPC.

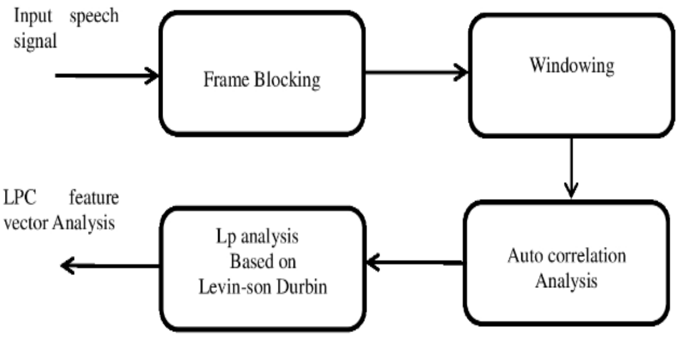
Input speech signal
Frame Blocking
Windowing
Auto correlation Analysis
Lp analysis Based on Levin-son Durbin
LPC feature vector Analysis

Fig.4 Block diagram of Linear Predictive Coding[1]

LPC will also have different stages as like MFCCs. Different steps involved an

- pre-emphasis,
- frame blocking,
- windowing;
- auto correction analysis,
- LPC analysis and LPC parameters conversion to cepstral co-efficient [10,12]

Few of the advantage of LPC includes reliability, higher computational speed, encodes speech at a lower bit rate. It has disadvantages of generating residual error in the output [2].

2.3. Perceptual Linear Prediction

Perceptual Linear Prediction In the year 1990 Herman sky introduced the Perceptual Linear Prediction (PLP) model. The main objective of this model is to obscure psychophysics of particular persons hearing; specifically it concentrates on feature extraction process. PLP MODEL shows unacceptability towards grabage information in the language thus improves rate of speech recognition.

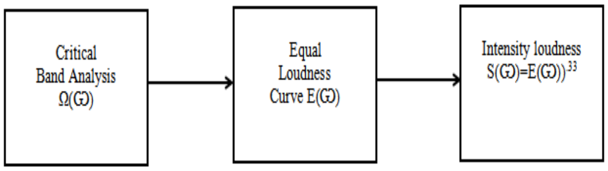
Critical Band Analysis Ω(Ꞡ)
Equal Loudness Curve E(Ꞡ)
Intensity loudness S(Ꞡ)=E(Ꞡ))^33

Fig 5. Block diagram of PLP Processing [1]

Bark scaling is applied whose subjective is to measure loudness, where the frequency is converted to the bark, which is an more effective representation of signal (individual hearing) in terms of frequency resolution. More elobrated steps of PLP computation is depicted below.

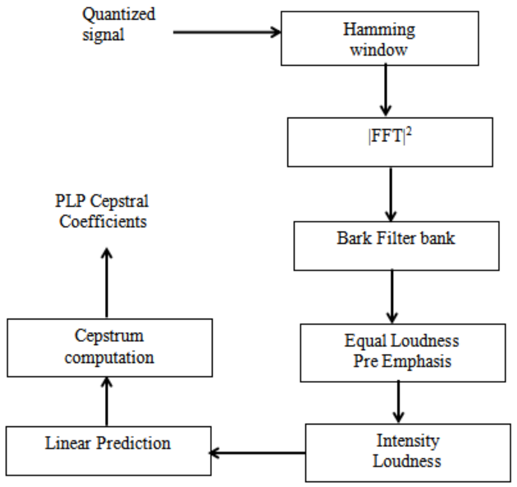

FIG. 6. PLP PARAMETER COMPUTION [1]

The bark frequency that corresponds to an audio frequency is given by equation (6) from [1]

$$\Omega(\omega) = 6\ln\S\frac{\omega}{1200\pi}\S x\left(\frac{\omega}{1200\pi}\right)^{2} + 1^{..0.5..} \qquad (6)$$

2.4. RASTA filtering

RASTA is a short from derived from RelAtive SpecTrAl. It is a procedure used for boosting signal which have been recorded in noisy atmosphere. It adopts band pass filtering in processing signals in log spectral domain ,earlier it was solely used to lessen the effect of noise in the signal lately it can also be used to boost the signal.[13,14]

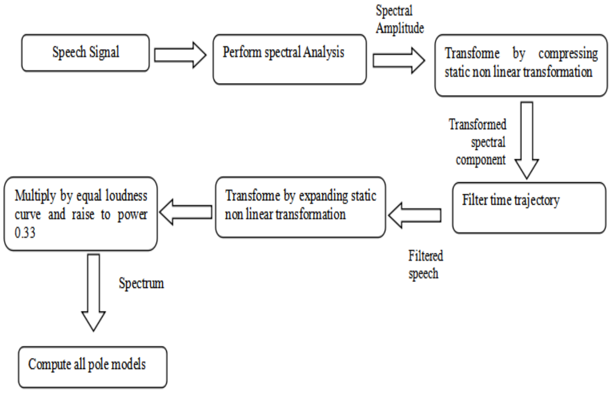
Speech Signal
Perform spectral Analysis
Spectral Amplitude
Transforme by compressing static non linear transformation
Transformed spectral component
Filter time trajectory
Filtered speech
Transforme by expanding static non linear transformation
Multiply by equal loudness curve and raise to power 0.33
Spectrum
Compute all pole models

Fig. 7: Process of RASTA Technique [16]

It removes the variation in environment; it's efficient because system does not depend on type of microphone and its position

Minor damaging is caused in performance of cleaning information yet it cut downs the error into half for filtered cases. We could yield a better performance when RASTA is merged with PLP [2].

2.5. Probabilistic Linear Discriminate Analysis (PLDA):

It is basically a dimensional reduction method where in it is used in recognition of object for feature extraction; earlier it was use for face recognition recently now it has also been used for voice recognition. The state contingent variable of HMM has been used by this technique, generative model has been used in order to construct PLDA. It is basically based on i- vector extraction The variables from Hidden Markov Model have been used and the variable is state dependent. PLDA has been constructed by generative model where data distribution is learnt under unsupervised way. It's a ductile acoustic model this factor being the favourable part of the model It makes use of different number of input frames which are in co-ordinate with each other without the requirement of convenience. [2, 15]

3. **Modelling Techniques**: In this they extract the speaker model by inspection of extracted features and vectors.As reveled in figure beside its been grouped into speaker recognisation & identification. Speaker identification procedure allows the system to recognizeorator with respect to source drained data from speech signal.speaker recognisation has beside grouped in speaker dependent and independent and hear succeeding modeling approaches could be utilizedapproaches:

3.1. Acoustic-Phonetic approach: the primary rule which this technique follow is to identify the speech and mark to signal with pertinent labels. This method of modeling proposes that in a language limited number of phonemes would exist, which can be expounded by the acoustic properties.[1,19]

3.2. Pattern recognition approaches: This approach demands two steps firstly pattern comparison and patter training. The predominant attribute is that it makes use of a mathematical model and builds uniform speech pattern for definitive patter comparison. Its further representation is of speech template or statistical models form like HMM. In patter comparison stage the unknown speech and compared with pattern already learned through training to find the identity of the unknown signal.[1,19]

3.3. Dynamic Time Warping (DTW): DWT is an algorithm with is used to identify or compute the resemblance among two sequences which might be varying with respect to time or speed. As an efficient ASR the system should be able to cope with different orator's speeds. The DWT is more efficient if it has to recognize isolated words and can be modified for connected word identification. DWT is much apt algorithm to match sequences with lost information , on a condition that length of the segment is long enough to match [1,19]

3.4. Artificial Intelligence Approach (AI): This approach is a blend of acoustic phonetic and patter approach. This method tries to automate recognition process, such a way that it seems to be as similar like an human being applying his intelligence to imagine , investigate and then give a verdict on measured parameter with respect to speech signal.

As we all know it is no more covert that the science of speech recognition has extensively developed. As the technology has progress, speech recognition has turn out to be gradually more embedded in our on a daily basis lives with voice-motivated applications. It can be expounded as "human intelligence manifested by machine ".Fore most it was used investigate and rapidly figure data; Artificial intelligence enables computer to take over the task that humans were skilled of. Machine learning a division of (AI) are self taught systems. It includes making a computer to recognize patterns, rather than programming it with set of rules. The training procedure demands feeding huge amounts of data to the algorithm and enabling it to learn from that data and distinguish patterns. Near the beginning days, programmers would have to write code for every object they wanted to identify but then now one system can recognize all objects by showing it with many examples of each. Hence, these systems continue to get smarter over time without human intervention. There are many different techniques and approaches to machine learning. Some other common applications of artificial intelligence in the present days are object recognition, translation, speech recognition, and natural language processing [20].

4. **Matching Techniques**: The word that has been perceived is employed by the engine of speech recognizer to a word that is already well-known by deploying the following approaches

4.1. Sub word matching: The generally phonemes has being surveyed by an engine and achieve additional pattern recognition on those sub words. This method takes a lot of processing compared to whole-word matching, but advantage over whole-word matching is that it constrained storage(5 to 20 bytes for a word). Added to this pronunciation of sub-word can be hypothesized from English text without making the user to speak.[1,16,17]

4.2. Whole word matching: In this matching incoming digitized audio signal is compared with a guide with is prerecorded. This method had advantage of consuming much lesser processing, over sub-word matching. The disadvantage is that it requires all the words to be recognized to be prerecorded, which intern leads to requirement of heavy storage capacity(50 to 512 bytes/word).[1,18]

5. CONCLUSION:

This paper has given an overview of five different feature recognition techniques. Out of all available feature extraction techniques MFCC and LPC are more predominantly used techniques because of their outstanding advantage where in the output will be almost same as speech signal fed as input.

6. REFERENCES:

[1] Jolad, B., & Khanai, R., Dr. (n.d.). Different Feature Extraction Techniques For Automatic Speech Recognition: A Review. International Journal Of Engineering Sciences & Research Technology, 7(2), 181-188.

[2] Shreya Narang, Ms. Divya Gupta, Speech Feature Extraction Techniques: A Review, International Journal of Computer Science and Mobile Computing, Vol. 4, Issue. 3, March 2015, pg.107 – 114.

[3] Anjivani S. Bhabad, Gajanan K. Kharate, An Overview of Technical Progress in Speech Recognition, International Journal of Advanced Research in Computer Science and Software Engineering, March 2013,.Volume 3, Issue 3.

[4] Lei Xie, Zhi-Qiang Liu, "A Comparative Study of Audio Features For Audio to Visual Cobversion in MPEG-4 Facial Animation," Proc. of ICMLC, Dalian, 13-16 Aug-2006.

[5] Alfie Tan Kok Leong, "A Music Identification System Based on Audio Content Similarity," Thesis of Bachelor of Engineering, Division of Electrical Engineering, The School of Information Technology and Electrical Engineering,The University of Queensland, Queensland, Oct-2003.

[6] S "DWT and MFCC Based Human Emotional Speech Classification Using LDA" International Conference on Biomedical Engineering (ICoBE), Penang, 27-28 February 2012, pp. 203-206.

[7] Alfie Tan Kok Leong, "A Music Identification System Based on Audio Content Similarity," Thesis of Bachelor of Engineering, Division of Electrical Engineering, The School of Information Technology and Electrical Engineering,The University of Queensland, Queensland, Oct-2003.

[8] Corneliu Octavian DUMITRU, Inge GAVAT, "A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language", 48th International Symposium ELMAR-2006, 07-09 June2006, Zadar, Croatia

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

[10] Lawrence Rabiner, and Biing Hwang Juang, Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993.

[11] Iindependent Word Recognition", International Journal of Computer Theory and Engineering, Vol.2, No.6,December, 2010 1793-8201

[12] Ethnicity Group. "Cepstrum Method". 1998 http://www.owlnet.rice.edu/~elec532/PROJECTS98/speech/cepstrum/cepstrum.html.

[13] H. Hermansky and N. Morgan, Rasta processing of speech, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 578{589, 1994.

[14] Hynek Hermansky , Eric A. Wan, and Carlos Avendano, Oregon Graduate Institute of Science & Technology Department of Electrical Engineering and Applied Physics, Speech enhancement based on temporal processing.

[15] Liang Lu, Member, IEEE and Steve Renals, Fellow, IEEE, IEEE SIGNAL PROCESSING LETTERS, Probabilistic Linear Discriminant Analysis for Acoustic Modelling, VOL. X, NO. X, 2014

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

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

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

[19]]Santosh K.Gaikwad and Pravin Yannawar, A Review, International Journal of Computer Applications A Review on Speech Recognition Technique Volume 10– No.3, November 2010.

[20] www.temi.com