ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

Embedded Speech Recognition System

Nikita Varshney[1], Sukhwinder Singh[2]

Student, Department of E&EC, PEC University of Technology, Chandigarh, India

Asst. Professor, Department of E&EC, PEC University of Technology, Chandigarh, India

Abstract-The purpose of this paper is to review the existing techniques and hence introduce suitable user interface for novice user and the test plan is to design the embedded system using technique according to the area of its application. This embedded system technique on further development can be used for deaf and dumb also thus providing them an easier method to communicate with the rest of the world.

KEYWORDS–Speech Recognition, LPC, Signal Processing, Covariance Method, Hamming Window, Windowing

I. INTRODUCTION

Speech recognition is the process of taking the spoken word as an input to a computer program. This process is important to virtual reality because it provides a fairly natural and intuitive way of controlling the simulation while allowing the user's hands to remain free. It is the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and these signals are transformed into coding patterns to which meaning has been assigned. Speech Recognition Systems that do not use training are called "speaker-independent" systems and that use training are called "speaker-dependent" systems. Combining speech recognition with network actuation can be used to control the actuator from a remote place. Voice controlled Embedded system (VCES) is a Semi -autonomous system whose actions can be controlled by the user by giving specific voice commands. The graphical user interface running along with the software provides a very convenient method for the users to first train the system and then run. The speech signal is captured through microphone and processed by software running on a PC. Speech characteristics can be extracted from sample by coding and is implemented through an Embedded System.

II. RELATED WORK

A speech recognition system makes human interaction with computers possible through a voice/speech to initiate an automated service or process. Controlling a machine by simply talking to it gives the advantage of hands-free, eyes-free interaction. Human computer interactions refers to the ways Users (humans) interact with the computers. Speech recognition systems help users who in one way or the other cannot be able to use the traditional Input and Output devices. For about four decades human beings have been dreaming of an “intelligent machine” which can master the natural speech. In its simplest form, this machine should consist of two subsystems, namely automatic speech recognition (ASR) and speech understanding (SU). ASR transcribes natural speech while SU is to understand the meaning of the transcription. A lot of speech aware applications are already there in the market developed by Dragon, IBM and Philips. Genie is an interactive speech recognition software developed by Microsoft. Various voice navigation applications, one developed by AT&T, allow users to control their computer by voice, like browsing the Internet by voice. Background noise is the worst part of a speech recognition process. It confuses the recognizer and makes it unable to hear what it is supposed to. Various capabilities of current speech recognizers in the field of telecommunications are in Voice Banking and Directory Assistance.The structure of a standard speech recognition system is illustrated along with description below:

Raw speech. Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time

Signal analysis. Raw speech should be initially transformed and compressed, in order to simplify subsequent

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

processing. There are many signal analysis techniques which can extract useful features and compress the data by a factor of ten without losing any important information. Fourier analysis (FFT), which is the most popular one, yields discrete frequencies over time, which can be interpreted visually.

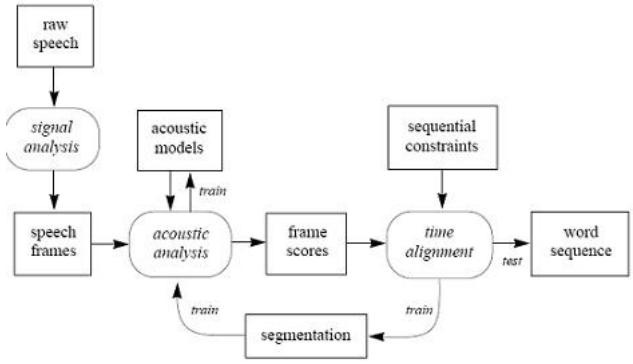

Fig.1 Structure of a Standard Speech Recognition System

Linear Predictive Coding (LPC) yields coefficients of a linear equation that approximate the recent history of the raw speech values. Cepstral analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the signal. Afterwards, procedures such as Linear Discriminant Analysis (LDA) may optionally be applied to further reduce the dimensionality of any representation, and to de-correlate the coefficients.

Speech frames. The result of signal analysis is a sequence of speech frames, typically at 10 msec intervals, with about 16 coefficients per frame. These frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics; this typically leads to improved performance. The speech frames are used for acoustic analysis.

Acoustic models. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties. The simplest acoustic technique isa template, which is just a stored sample of the unit of speech to be modeled, e.g.,a recording of a word. An unknown word can be recognized by simply comparingit against all known templates, and finding the closest match. Templates have twomajor drawbacks: (1) they cannot model acoustic variability, except in a coarseway by assigning multiple templates to each word; and (2) they are limited to whole-word models, because it's hard to record or segment a sample shorterthan a word – so templates are useful only in small systems which can afford theluxury of using whole-word models.

III. HUMAN SPEECH PRODUCTION

Speech is a natural form of communication for human beings. The process of speech production in humans can be summarized as air being pushed from the lungs, through the vocal tract, and out through the mouth to generate speech. The lungs are the source of the sound and the vocal tract act as a filter that produces the various types of sounds that make up speech. Thus, all the humans produce sound by the same process but the sound produced is different by different people. We have to devise such a system by which we can analyse different types of sound because a Speech controlled Embedded System can have multiple users to universalize VCES. A limited set of individual sounds is termed as phonemes which are of two types, voiced and unvoiced. Voiced sounds are usually vowels, have high average energy levels and are very distinct resonant frequencies. They are generated by air from the lungs which is forced over the vocal cords. The rate at which the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

vocal cords vibrate determines the pitch of the sound produced. Unvoiced sounds are usually consonants and generally have less energy and higher frequencies than voiced sounds. The production of unvoiced sound involves air being forced through the vocal tract in a turbulent flow from the lungs. During this process the vocal cords do not vibrate, instead they stay open until the sound is produced. The amount of air in the lungs also affects the production of sound in humans. The air from the lungs is the source of air for the vocal tract which acts as a filter for the air from the lungs by taking in the source and producing speech. Higher the volume of air, louder the sound is.Speech signal can be considered to be of quasi-stationary nature. Quasi-stationary means that speech can be treated as a stationary signal for short intervals of time. This allows us to use techniques which are generally used for stationary signals for the processing of speech signals. We can also use speech compression as raw audio data can take up a great deal of memory. During compression, the data is compressed so that it will occupy less space. This compressed speech can also be encrypted for security purposes. Speech compression becomes important with teleconferencing and other applications; sending data is expensive, and anything which reduces the volume of data which needs to be sent can help to cut costs. The amplitude of speech signal varies slowly with time, which is another characteristic that is commonly exploited for the speech compression purpose.

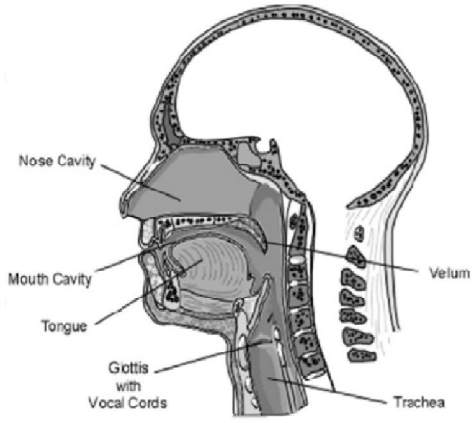

Fig. 2 Human anatomy for Speech Production

IV. WHAT IS VOICED AND UNVOICED SPEECH?

Voiced sounds, e.g., a, b, are essentially due to vibrations of the vocal cords, and are oscillatory in nature. Therefore, over short periods of time, they can be appropriately by sums of sinusoids. Unvoiced sounds such as s, sh are more noise-like. We need to distinguish between voiced and unvoiced speech depending upon our application. There are two methods for doing it: Short-time power function and Zero Crossing rate. Typically power of voiced signals is greater than that of unvoiced signals. As unvoiced signals oscillate much faster, they have a much higher rate of zero- crossing than voiced signals.

V.SOURCE-FILTER MODEL OF SPEECH PRODUCTION

Sound is variations in pressure of air. The creation of sound is the process of setting the air in rapid vibration. A speech production model has two major components:

Excitation (for Voiced sounds) Periodic air pulses pass through vibrating vocal chords**.(for Unvoiced sounds)** Air is forced through a constriction in vocal tract thus producing turbulence.This model has a feature extractor too which has three modules like sound recording, pre- emphasis filter and speech coding. Here the feature extractor uses a standard LPC

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

Cepstrum coder, which converts the incoming speech signal into LPC Cepstrum feature space. A vector of 12 Linear Predicting coding Cepstrum coefficients is calculated from each data block using Durbin's method and the recursive expressions developed by Furi.The period T is called the pitch period, and 1/T is called the pitch frequency.
On average:

Male: T = 8ms - pitch125Hz Female: T = 4ms - pitch250Hz

Vocal tract: Different voiced sounds are produced by changing the shape of the vocal tract.This system is time-slowly varying. Changes occur slowly compared to the pitch period i.e. each sound is approximately periodic, but different sounds are different periodic signals. This implies that we can model the vocal tract as an LTI filter over short time intervals. When a wave propagates in a cavity, there is a set of frequencies called natural frequencies of the resonator which get amplified and depend on the shape and size of the resonator. Therefore, the magnitude response of the vocal tract for one voiced sound (phoneme) can be modelled. The waveform for this particular phoneme will then be the convolution of the driving periodic pulse train x(t) with the impulse response v(t) and the magnitude of its spectrum |S(f)| will be the product of X(f) and the magnitude response |V (f)|. The maxima of |S(f)| are called the formant frequencies of the phoneme. Locations are dictated by the poles of the transfer function. Roll-off is such that the first 3-4 formants (range: up to 3.5 kHz) are enough for reasonable reconstruction. Thus, sampling at 3.52 kHz = 7 kHz is typically enough. Depending on the application, the sampling rate is usually 7- 20 kHz.
Poles of H(z) near the unit circle correspond to large values of H|ej|. So, we can design an all-pole filter, with poles which are close to the unit circle, corresponding to formant frequencies. The larger the magnitude response at the formant frequency, the closer the corresponding pole(s) to the unit circle.

VI. TYPES OF SPEECH RECOGNITIONSYSTEMS

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as the following:

- Isolated Words: Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses).
- Connected Words: Connected word systems or more correctly 'connected utterances' are similar to isolated words, but allows separate utterances to be 'run-together' with a minimal pause between them.
- Continuous Speech: Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content.
- Spontaneous Speech: It is a speech that is natural sounding and not rehearsed.

VII. LINEAR PREDICTIVE CODING

Compression of digital audio signals was started in the 1960s by telephone companies who were concerned with the cost of transmission and width of the transmitted signal. Linear Predictive Coding's origins begin in the 1970s with the development of the first LPC algorithm. The idea of using LPC for speech compression came up in 1966. LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds (sibilants and plosive sounds). Although apparently crude, this model is actually a close approximation of the reality of speech production. The glottis (the space between the vocal folds) produces the buzz, which is characterized by its intensity and frequency. The vocal tract forms the tube, which is characterized by its resonances, which give rise to formants, or enhanced frequency bands in the sound produced. Hisses and pops are generated by the action of the tongue, lips and throat during sibilants and plosives. LPC analyzes the speech signal by estimating the

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue. The numbers which describe the intensity and frequency of the buzz, the formants, and the residue signal, can be stored or transmitted somewhere else. LPC synthesizes the speech signal by reversing the process: use the buzz parameters and the residue to create a source signal, use the formants to create a filter (which represents the tube), and run the source through the filter, resulting in speech. Now as speech signals vary with time, this process is done on short chunks of the speech signal, which are called frames; generally 30 to 50 frames per second give intelligible speech with good compression.For voiced speech, each sample is known to be highly correlated with the corresponding sample that occurred one pitch period earlier. In addition, each sample is correlated with the immediately preceding samples because the resonances of the vocal tract. Therefore short durations of speech show an appreciable correlation.

VIII. THROUGH LINEAR PREDICTION

The possibility that a signal can be predicted from its past samples depends upon the autocorrelation function, or the bandwidth and the power spectrum of the signal. A predictable signal has a smooth and correlated fluctuation in the time domain and in the frequency domain, the energy of a predictable signal is concentrated in the narrow band of frequencies. In contrast, the energy of a non-predictable signal, such as white noise, is spread over a wide band of frequencies and not in the main lobe mainly. For a signal to have a capacity to convey information it must be of variable nature. These signals can be modelled as the output of a filter excited by an uncorrelated input. The random input models the unpredictable part of the signal, whereas the filter models the predictable structure of the signal.

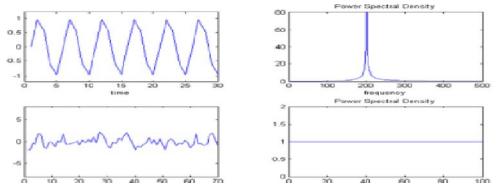

Fig.1 Time and Power Spectral Density Representation

The aim of linear prediction is to model the mechanism that introduces the correlation in a signal.

$$E_n = \sum_0^{N-1} e_n{}^2(m)$$

where $\Phi_n(i,k)$ is defined as

$$\Phi_n(i,k) = \sum_{m=0}^{N-1} x_n(m-1) * x_n(m-k) \begin{cases} 1 \le i \le p \\ 0 \le k \le p \end{cases} \qquad (4)$$

IX. THE AUTOCORRELATION METHOD

In this method, the speech segment is assumed to be zero outside the interval 0 < m < N -1. Thus the Speech Sample can be expressed as

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

$$x_n(m) = \begin{cases} x(n+m) * w(m), 0 \le m \le N-1 \\ 0, \quad \text{otherwise} \end{cases} \quad (1)$$

Another least square technique called covariance method windows the error signal, instead of the actual speech signal. In the autocorrelation method the filter is guaranteed to give best, but problems of the parameter accuracy can arise because of the necessity of the windowing (truncating) the rime signal. This is usually a problem if the signal is a portion of an impulse response. For example, if the impulse response of an all-pole filter is analysed by covariance method, the filter parameters can be computed accurately from only a finite number of samples of the signal. Using the autocorrelation method, one cannot obtain the exact parameters values unless the whole infinite impulse response is used in the analysis. However, in practice, very good approximations can be obtained by truncating the impulse response at a point where most of the decay of the response already occurred.

X. BY CAPTURING WORDS

The signal coming from the microphone is processedonly when we speak something. The program waits until the sample value exceeds some threshold value (which can be adjusted by the user). When the program is triggered by a significant sample, a number of following samples are captured to process. After that to determine the actual boundaries of the word spoken, 'edge detection' is performed. Here the center of gravity of the energy distribution of the signal is calculated and then from that point intervals where the amplitude level lies below a threshold level are removed. Finally we can have a set of voice samples corresponding to a particular word free of silent periods. In this process the noises introduced in the signals because of disturbances in the surrounding also get eliminated and also the software initially measures the noise present in the environment and subtracts this threshold value from the signals to be recorded.

XI. USE OF WINDOWING TECHNIQUE

Windowing of a simple waveform, like cos(ωt) or sin(ωt) causes its Fourier transform to develop non- zero values (commonly called spectral leakage) at frequencies other than ω i.e. to avoid Gibbs Phenomenon which is due to slow conversion of fourier series at the points of discontinuities. The leakage tends to be worst (highest) near ω i.e. the main lobe and least at frequencies farthest from ω i.e. at the side lobes. Thus, we need to minimize the width of side lobe provided the maximum part of energy is under the main lobe. If the signal under analysis is composed of two sinusoids of different frequencies, leakage can interfere with the ability to distinguish them spectrally. If their frequencies are dissimilar and one component is weaker, then leakage from the larger component can obscure the wearer's presence. But if the frequencies are similar, leakage can render them irresolvable even when the sinusoids are of equal strength. The rectangular window has excellent resolution characteristics for signals of comparable strength, but it is poor choice for signals of disparate amplitudes. This characteristic is sometimes described as low-dynamic-range. At the other extreme of dynamic range are the windows with the poorest resolution. These high- dynamic-range low-resolution windows are also poorest in terms of sensitivity; this is, if the input waveform contains random noise close to the signal frequency, the response to noise, compared to the sinusoid, will be higher than with a higher-resolution window. High-dynamic-range windows are probably most often justified in wideband applications, where the spectrum being analysed is expected to contain many different signals of various strengths. In between the extremes are moderate windows, such as Hamming and Hann. They are commonly used in narrowband applications, such as the spectrum of a telephone channel.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

XII. USE OF PRE-EMPHASIS

Pre-emphasis refers to a system process which is designed to increase (within a frequency band) the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies to improve the overall signal-to-noise ratio by the minimization of the adverse effects of phenomena such as attenuation, distortion or saturation of recording media in subsequent parts of the system. The whole system is called emphasis. The frequency curve is decided by special time constants. The cut off frequency is to be calculated from that value. It is commonly used in telecommunications, digital audio recording, record cutting, in FM broadcasting transmissions, and in displaying the spectrograms of speech signals. In high speed digital transmission, we use pre-emphasis to improve the signal quality at the output of a data transmission. In transmitting signals at high data rates, the transmission medium may introduce distortions, so we use pre-emphasis to distort the transmitted signal to correct for this distortion. When done properly a received signal is produced which more closely resembles the original or desired signal, allowing the use of higher frequencies or producing fewer bit errors. This operation is necessary for removing DC and low frequency components of the incoming speech signal. It also flattens the signal spectrum. In Pre-Speech Recognition Based Embedded Control System Emphasis is done by using a first order FIR filter which can be described

XIII. PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other main advantage of PCA is that once you have found these patterns in the data, and you can compress the data, i.e. by reducing the number of dimensions to two, without much loss of information.

XIV. IMPLEMENTATION

The proposed system will be advantageous due to the following reasons.
a) The proposed model is quite portable. b) Differently-abled persons can easily be trained.c) The component and implementation cost are quite low. d) The proposed model is affordable to all masses of the society.e) Translation can be made to any language.f) Wide range applications.g) Upgraded easily.
After processing the speech, the necessary command instructions are sent to the Embedded System via Parallel Port Interface or by using a special function IC embedded in the system. The software is microphone dependent. The special feature of the application is the ability of the software to train itself for the above voice commands for a particular user. An ANN was used to classify the feature vectors of new speech. The outputs of network are the triphones obtained from baseline system whose identity often depends not on the spectral features at one point in time and on how the features change over time, so the inputs to the network consist of the features for the frame to be classified, as well as the features for frames at -60, -30, 30, and 60 millisecond relative to the frame to be classified (for a total of 195 input values). It can also be trained for its use by multiple users. The graphical user interface (GUI) running along with the software provides a very convenient method for the users to train. It also provides many other facilities in operating the Embedded System. The flowchart of the algorithm implemented is as shown in fig. below.
The algorithm is as follows:
• During the training phase, the user inputs his sound files into the software.
• This file is then passed through a signal conditioning block where noise cancellation takes place, by adjusting the mic noise level.
• After the voice is recorded i.e. word capturing takes place, the process of word filtering take place.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

• For better filtering we make use of "Hamming window" function out of all the other window functions for windowing as the voice captured lies in the narrow band frequency.
• Linear predictive coding (LPC) is performed on the captured word for getting the cepstral coefficient.
• The comparison is done by comparing the training mode word captured and running mode word captured.
• Interfacing between embedded system and pc is done via parallel port connection and the programming of the microcontroller of the circuit thus takes place in the manner.

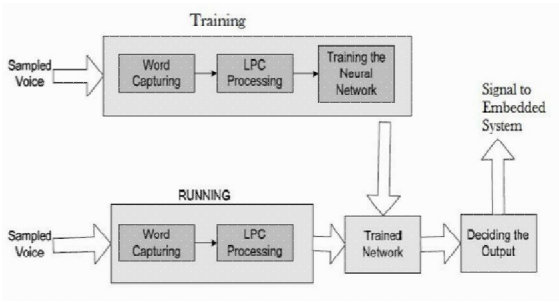

Fig.4 Flowchart of Algorithm

Process
• Return a frequency domain data of the word captured.
• Save waveform of the signal.
• Create a System object to read from a multimedia file.
• Create an FIR digital filter System object used for pre- emphasis.
• Create a Hamming window object.
• Create an autocorrelation System object which lags in the range [0:12] scaled by the length of input.
• Create a System object to compute the reflection coefficients.
Setup plots for visualization:
• Read audio input.
• Pre-emphasis.
• Buffer and Window
• Autocorrelation
• Compare the stored data and return the difference value
For deaf and dumb, the device will be beneficial in teaching and training them by further technical modifications, with basics of any language depending upon the preference of language in their region. For multipurpose applications voice enabled home automation system using wireless RF transmission and reception techniques can be used in addition with the above proposed system which will be handy and useful for deaf and dumb persons or physically challenged and also for elderly persons.

XIV. RESULTS

There are various Speech Recognition Techniques like Linear Predictive Coding, Linear Prediction, Windowing, Pre-emphasis, Word Capturing, Embedded Model, Auto-Correlation Method etc and they all have their advantages and disadvantages. Following inferences can be drawn on the basis of the study of these techniques:

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

Sr No	Technique	Advantages	Disadvantages
1.	Linear Predictive Coding	Next sample can be easily determined from the previous ones, attempts to model human sound production instead of an estimate of sound.	Lossy Form of Compression Technique
2.	Linear Prediction	All-pole linear prediction is a highly useful technique, Acceptable for most voiced speech sounds	Not appropriate for nasal and fricative sounds, works only when frequency response consists of poles only and not zeros, the prediction order should be twice the number of formants present in the signal bandwidth
3.	Auto-Correlation	minimises the error signal over all time,all samples outside the interval of interest are taken to be 0, greater redundancies so easier to compute	Used only for fricative sounds and not periodic speech sounds
4.	Word Capturing	Easy and direct method	More prone to atmospheric noise
5.	Pre-emphasis	Improves the overall signal-to-noise ratio of recording media, improves signal quality at the output of a data transmission, fewer bit errors	SMT (Surface Mount) resistors and capacitors used in this technique occupy considerable portion, discrete resistors and capacitors can introduce undesirable EM interference, adverse yield impact by assembling more components, added cost on inventory and part management
6.	Principal Component Analysis	Simplest of the true eigenvector based multivariate analysis, can concentrate much of the signal into the first few principal components, optimal orthogonal transformation for keeping subspace having largest variance, reduces the no. of parameters	Saves the signal from distortion as later principal components may be dominated by noise, greater computational requirements, sensitive to variable scaling, Mean subtraction necessary
7.	Windowing	Relatively simpler, discontinuities become transition bands, eliminates ringing effect at the band-edge	Resulting filter never optimal, very little design flexibility, limited use in speech recognition

ACKNOWLEDGMENT

The author is grateful to Assistant Prof.Sukhwinder Singh, Electronics& Electrical Communication Department, PEC University of Technology, Chandigarh, India for his constant support and encouragement. Without his guidance and approval, the review on this technique would not have been possible as his valuable suggestions and deep knowledge about the technique played an important part in nailing out this review paper.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2014

REFERENCES

[1] .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., pp 5-7, September 1966.

[2] D.B.Fry,"Theoritical Aspects of Mechanical speech Recognition" , andP.Denes," The design and Operation of the Mechanical Speech Recognizer",Universtiy College London, J.British Inst. Radio Engr., 19:4,211-299,1959.

[3] WiqasGhai and NavdeepSingh,"Literature Review on Automatic Speech Recognition", International Journal ofComputer Applications vol. 41– no.8, pp. 42-50, March 2012.

[4] Rabiner L.R., S.E.L.evinson: (1981)"Isolated and connected word recognition – Theory and selected applications", IEEE Trans. COM-29, pp.621-629

[5] Chenghui Yang, "Based on Artificial Neural Networks for voice recognition word segment", May 2011

[6] S.Katagiri," Speech pattern recognition using neural networks", CRC Press, pp113-153, 2003

[7] Bo Lu, "A speech recognition system based on multiple neural networks", pp 2-3, Aug 2010

[8] AipingNing," A Speech Recognition System Based on Fuzzy Neural Network Optimized by Time Variant PSO", pp 4-5, Sept 2010

[9] W.Chou and B.H.Juang (Eds.) ,"Pattern Recognition in Speech and Language Processing", CRC Press, pp.115-147,2003.

[10] RathinaveluChengalvarayan, "Use of Generalized Dynamic Feature Parameters for Speech Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 5,No. 3, May 1997.

[11] Jing Zhang, "A speech recognition method based clustering neural network integration", pp7-9, April 2011

[12] P.A.Devijver and J.Kittler, "Pattern Recognition: A Statistical Approach" , London, Prentice Hall, 1982

[13] M.A.Anusuya and S.K.Katti, "Speech Recognition by Machine: A Review", (IJCSIS) International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.

[14] Rajesh Kumar Aggarwal and M. Dave, "Acoustic modeling problem for automatic speech recognition system: advances and refinements Part (Part II)", Int J Speech Technol, pp. 309– 320, 2011