International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-3, February 2013

Advanced Speech Compression VIA Voice Excited Linear Predictive Coding Using Discrete Cosine Transform (DCT)

Nikhil Sharma, Niharika Mehta

Abstract: One of the most powerful speech analysis techniques is the method of linear predictive analysis. This method has become the predominant technique for representing speech for low bit rate transmission or storage. The importance of this method lies both in its ability to provide extremely accurate estimates of the speech parameters and in its relative speed of computation. The basic idea behind linear predictive analysis is that the speech sample can be approximated as a linear combination of past samples. The linear predictor model provides a robust, reliable and accurate method for estimating parameters that characterize the linear, time varying system. In this project, we implement a voice excited LPC vocoder for low bit rate speech compression.

Index Terms: Autocorrelation, Discrete Cosine Transform, Levinson Durbin Recursion, Linear predictive coding (LPC).

I. INTRODUCTION

Speech coding has been and still is a major issue in the area of digital speech processing in which speech compression is needed for storing digital voice and it requires fixed amount of available memory and compression makes it possible to store longer messages. Several techniques of speech coding such as Linear Predictive Coding (LPC), Waveform Coding and Sub band Coding exist. This is used to characterize the vocal track and inverse filter is used to describe the vocal source and therefore it is used as the input for the coding. The speech coder that will be developed is going to be analyzed using subjective analysis. Subjective analysis will consist of listening to the encoded speech signal and making judgments on its quality. The quality of the played back speech will be solely based on the opinion of the listener. The speech can possibly be rated by the listener either impossible to understand, intelligible or natural sounding. Even though this is a valid measure of quality, an objective analysis will be introduced to technically assess the speech quality and to minimize human bias.

II. BACKGROUND

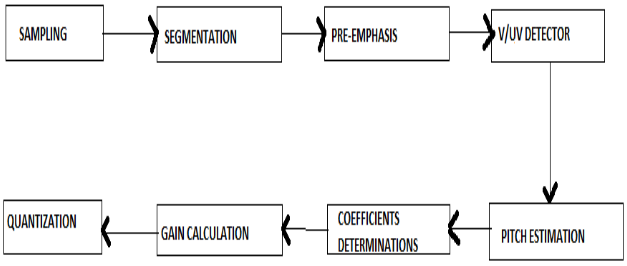
There are several different methods to successfully accomplish speech coding. Some main categories of speech coder are LPC Vocoders, Waveform and Sub band coders. The speech coding in this Project will be accomplished by using a modified version of LPC-10 technique.

Manuscript published on 28 February 2013.
*****Correspondence Author(s)
Nikhil Sharma, (Research Scholar) Nit Kurukshetra, India.
Niharika Mehta, (Research Scholar) Mriu Faridabad, India.

Linear Predictive Coding is one possible technique of analyzing and synthesizing human speech. The exact details of the analysis and synthesis of this technique that was used to solve our problem will be discussed in the methodology section. LPC makes coding at low bit rates possible. For LPC-10, the bit rate is about 2.4 kbps. Even though this method results in an artificial sounding speech, it is intelligible. This method has found extensive use in military applications, where a high quality speech is not as important as a low bit rate to allow for heavy encryptions of secret data. However, since a high quality sounding speech is required in the commercial market, engineers are faced with using other techniques that normally use higher bit rates and result in higher quality output. In LPC-10 vocal tract is represented as a time-varying filter and speech is windowed about every 30ms. For each frame, the gain and only 10 of the coefficients of a linear prediction filter are coded for analysis and decoded for synthesis. In 1996, LPC-10 was replaced by mixed-excitation linear prediction (MELP) coder to be the United States Federal Standard for coding at 2.4 kbps. This MELP coder is an improvement to the LPC method, with some additional features that have mixed excitation, aperiodic pulses, adaptive spectral enhancement and pulse dispersion filtering. Waveform coders on the other hand, are concerned with the production of a reconstructed signal whose waveform is as close as possible to the original signal, without any information about how the signal to be coded was generated. Therefore, in theory, this type of coders should be input signal independent and work for both speech and nonspeech input signals.

III. METHODOLOGY

A) LPC System Implementation



FIG(1):- LPC ENCODER BLOCK DIAGRAM

1-Sampling: First, the speech is sampled at a frequency appropriate to capture all of the necessary frequency components important for processing and recognition. According to the Nyquist theorem, the sampling frequency must be at least twice the bandwidth of the continuous-time signal in order to avoid aliasing. For voice transmission, 10 kHz is typically the sampling frequency of choice, though 8 kHz is not unusual. This is because, for almost all speakers, all significant speech energy is contained in those frequencies below 4 kHz (although some women and children violate this assumption).

2- Segmentation: The speech is then segmented into blocks for processing. Properties of speech signals change with time. To process them effectively it is necessary to work on a frame-by-frame basis, where a frame consists of a certain number of samples .The actual duration of the frame is known as length. Typically, length is selected between 10 and 30 ms or 80 and 240 samples. Within this short interval, properties of the signal remain roughly constant. Simple LPC analysis uses equal length blocks of between 10 and 30ms. Less than 10ms does not encompass a full period of some low frequency voiced sounds for male speakers. For certain frames with male speech sounded synthetic at 10ms sample windows, pitch detection became impossible. More than 30ms violates the basic principle of stationarity.

3- Pre-emphasis: The typical spectral envelope of the speech signal has a high frequency roll-off due to radiation effects of the sound from the lips. Hence, high-frequency components have relatively low amplitude, which increases the dynamic range of the speech spectrum. As a result, LP analysis requires high computational precision to capture the features at the high end of the spectrum. One simple solution is to process the speech signal using the filter with system function

$H(z) = 1-\alpha z^{-1}$…………………(1)

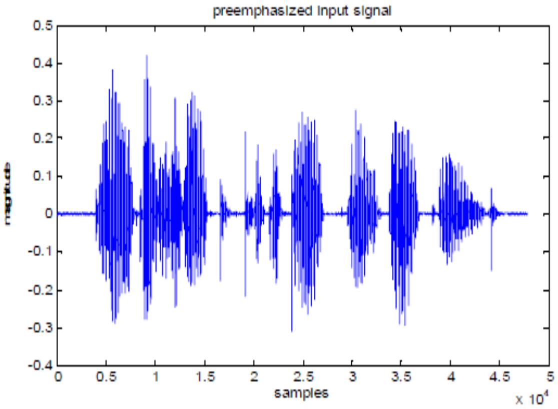
This is high pass in nature. The purpose is to augment the energy of the high frequency spectrum. The effect of the filter can also be thought of as a flattening process, where the spectrum is ‘‘whitened’’. Denoting x[n] as the input to the filter and y[n] as the output, the following difference equation applies:

$Y[n]=x[n]-\alpha x[n]$………………(2)

The filter described in (1) is known as the pre-emphasis filter. By pre-emphasizing, the dynamic range of the power spectrum is reduced. This process substantially reduces numerical problems during LP analysis, especially for low precision devices. A value of α near 0.9 is usually selected. It is common to find in a typical speech coding scheme that the input speech is first pre-emphasized using (1). To keep a similar spectral shape for the synthetic speech, it is filtered by the de-emphasis filter with system function at the decoder side, which is the inverse filter with respect to pre-emphasis.

$G(z)=1/(1-az^{-1})$……………….(3)

The main goal of the pre-emphasis filter is to boost the higher frequencies in order to flatten the spectrum. This pre-emphasis leads to a better result for the calculation of the coefficients using LPC. There are higher peaks visible for higher frequencies in the LPC spectrum. Clearly the coefficients corresponding to higher frequencies can be better estimated.



4- Voicing detector: The purpose of the voicing detector is to classify a given frame as voiced or unvoiced. In many instances, voiced/unvoiced classification can easily be accomplished by observing the waveform; a frame with clear periodicity is designated as voiced, and a frame with noise-like appearance is labeled as unvoiced. In other instances, however, the boundary between voiced and unvoiced is unclear; this happens for transition frames, where the signal goes from voiced to unvoiced or vice versa. The necessity to perform a strict voiced/unvoiced classification is indeed one of the fundamental limitations of the LPC model. In this section we discuss some measurements that a voicing detector relies on to accomplish its task. For reliable operation, the detector must take into account as many parameters as possible so as to achieve a high degree of robustness. These parameters are input to a linear classifier having binary output. The voicing detector is one of the most critical components of the LPC coder, since misclassification of voicing states can have disastrous consequences on the quality of the synthetic speech. These parameters are discussed bellow.

a- Energy: This is the most obvious and simple indicator of voicedness. Typically, voiced sounds are several orders of magnitude higher in energy than unvoiced signals. For the frame (of length N) ending at instant m, the energy is given by

$\mathrm{En,g}[m] = \sum_{n=m-N+1}^{m} y^2\,[n]$…………..(4)

For simplicity, the magnitude sum function defined by

$\mathrm{MSF}[m] = \sum_{n=m-N+1}^{m} |y[n]|$…………(5)

Serves a similar purpose. Since voiced speech has energy concentrated in the lowfrequency region, due to the relatively low value of the pitch frequency, better discrimination can be obtained by low pass filtering the speech signal prior to energy calculation. That is, only energy of low-frequency components is taken into account. A bandwidth of 800 Hz is adequate for the purpose since the highest pitch frequency is around 500Hz.[4]

b- Zero Crossing Rate: The zero crossing rate of the frame ending at time instant m is defined by

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)
© Copyright: All rights reserved.

$$\mathrm{ZC}[m] = \frac{1}{2\sum_{n=m-N+1}^{m} |\mathrm{sgn}(y[n]) - \mathrm{sgn}(y[n-1])|} \ldots\ldots(6)$$

With sgn(.) the sign function returning ±1 depending on the sign of the operand. Equation (3.6) computes the zero crossing rates by checking the samples in pairs to determine where the zero crossings occur. Note that a zero crossing is said to occur if successive samples have different signs. For voiced speech, the zero crossing rate is relatively low due to the presence of the pitch frequency component (of low frequency nature), whereas for unvoiced speech, the zero crossing rate is high due to the noise-like appearance of the signal with a large portion of energy located in the high frequency region.

c- Pitch period: Since voiced speech concentrated in the low-frequency region, as a consequence, its pitch period has higher values than the unvoiced.

Voicing Detector Design: A voicing detector can rely on the parameters discussed so far (energy, zero crossing rate, and pitch period) to make the proper decision. A simple detector can be implemented by using just one parameter as input. For instance, the zero crossing rate can be used for voicing detection in the following manner: if the rate is lower than a certain threshold, the frame is declared voiced; otherwise, it is unvoiced. The design problem is therefore to find the proper threshold so that a voicing decision can be accomplished reliably. By analyzing a large amount of speech signals, it is possible to come up with a reasonable value of a decision threshold so as to minimize the total classification error. Relying on just one parameter, however, limits the robustness of the system. For the voicing detector using the zero crossing rates alone, noise contamination can increase the rate in such a way that voiced frames are classified as unvoiced frames. Thus, using more parameters of the frame is necessary to improve the reliability in voicing detection.

5- Pitch period estimation: One of the most important parameters in speech analysis, synthesis, and coding applications is the fundamental frequency, or pitch, of voiced speech. Pitch frequency is directly related to the speaker and sets the unique characteristic of a person. Voicing is generated when the airflow from the lungs is periodically interrupted by movements of the vocal cords. The time between successive vocal cord openings is called the fundamental period, or pitch period. For men, the possible pitch frequency range is usually found somewhere between 50 and 250 Hz, while for women the range usually falls between 120 and 500 Hz. In terms of period, the range for a male is 4 to 20 ms, while for a female it is 2 to 8ms. Pitch period must be estimated at every frame. By comparing a frame with past samples, it is possible to identify the period in which the signal repeats itself, resulting in an estimate of the actual pitch period. Note that the estimation procedure makes sense only for voiced frames. Meaningless results are obtained for unvoiced frames due to their random nature. Design of a pitch period estimation algorithm is a complex undertaking due to lack of perfect periodicity, interference with formants of the vocal tract, uncertainty of the starting instance of a voiced segment, and other real world elements such as noise and echo. In practice, pitch period estimation is implemented as a trade-off between computational complexity and performance. Many techniques have been proposed for the estimation of pitch period and only one is included here.[4]

a- The Autocorrelation Method: The pitch period could be estimated by taking the average separation between peaks. The overall peaks and troughs in the spectrum are referred to as the formant structure (where the formants are the frequencies where resonances occur). The autocorrelation of a stationary sequence x (n) is defined as

$$R_z(\tau) = x(n) * x(n+\tau) = 1/N \sum_{n=0}^{N-1} x(n)x(n+\tau) ..(7)$$

Where τ is termed the lag. Auto means self or from one signal, and correlation means relation between two samples. An autocorrelation is the average correlation between two samples from one signal that are separated by τ samples. It should be noted that the upper limit in the summation will be less than N−1 when τ is positive, and the lower limit will be greater than 0 when τ is negative. Thus, the autocorrelation can be rewritten as

$$R_z(\tau) = 1/N \sum_{n=0}^{N-1-|\tau|} x(n)\, x(n+|\tau|) \ldots\ldots\ldots\ldots(8)$$

6- Coefficients determination: The coefficients of the difference equation (the prediction coefficients) characterize the formants, so the LPC system needs to estimate these coefficients. The estimate is done as mentioned above by minimizing the mean-square error between the predicted signal and the actual signal. This is a straight forward problem, in principle. In practice, it involves (1) the computation of a matrix of coefficient values, and (2) the solution of a set of linear equations. An efficient algorithm known as the Levinson- Durbin algorithm is used to estimate the linear prediction coefficients from a given speech waveform.

7- Gain Calculation: Power of the prediction-error sequence is calculated next, which is different for voiced and unvoiced frames. Denoting the prediction-error sequence as , with N being the length of the frame, we have for the unvoiced case

$$p = 1/N \sum_{n=0}^{N-1} e^2\,[n] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

For the voiced case,

$$p = 1/[N/T]T \sum_{n=0}^{[N/T]T-1} e^2\,[n] \ldots\ldots\ldots\ldots\ldots(10)$$

power is calculated using an integer number of pitch periods: It is assumed that N > T, and hence use of the floor function ensures that the summation is always performed within the frame’s boundaries. Gain computation is performed as follows. For the unvoiced case, denoting the gain by g, we have

$$g = \sqrt{p} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(11)$$

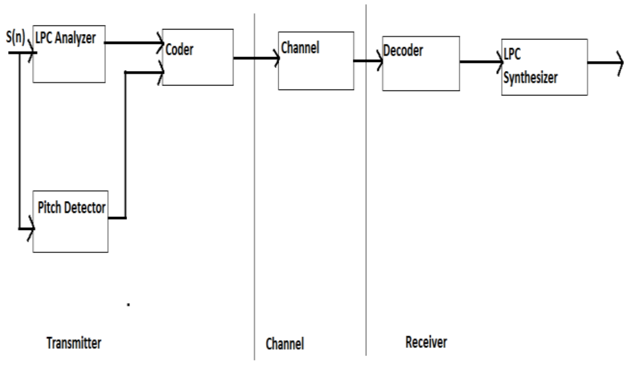
Since the white noise generator has unit-variance output. For the voiced case, the power of the impulse train having an amplitude of g and a period of T, measured over an interval of length [N/T]T, must equal p.

8- Quantization: Usually direct quantization of the predictor coefficients is not considered. To ensure stability of the coefficients (the poles must lie within the unit circle in the z-plane) a relatively high accuracy (8-10 bits per coefficients) is required. This comes from the effect that small changes in the predictor coefficients lead to relatively large changes in the pole positions.

Quantizing intermediate values is less problematic than quantifying the predictor coefficients directly. These intermediate values are called Line spectral frequency coefficients (LSFs) .Line spectral frequency coefficients (LSFs) were first introduced by Itakura (1975) as an alternative representation of LPCs (LSFs are mathematically equivalent (one-to-one) to LPCs). Due to many desirable properties, the LSF has received widespread acceptance in speech coding applications. Line spectral frequency, possesses several desirable features that make it attractive as an alternative LPC representation. The values of the LSFs directly control the property of the signal in the frequency domain, and changes of one parameter have a local effect on the spectrum. Also, the LSFs are bounded, they are located inside the (0, π) interval and ordered ( ). LSF are more amenable to quantization. LSFs are more correlated from one frame to the next than LPCs. For frame size of 20 msec. There are 50 frames/sec. 2400 bps is equivalent to 48 bits/frame.

B) *Voice-excited LPC Vocoder*

As the test of the sound quality of a plain LPC-10 vocoder showed, the weakest part in this methodology is the voice excitation. It is know from the literature that one solution to improve the qualityof the sound is the use of voice-excited LPC vocoders . Systems of this type have been studied by Atal et al. and Weinstein. Fig.3. shows a block diagram of a voice-excited LPC vocoder. The main difference to a plain LPC-10 vocoder, as shown in Fig.3, is the excitation detector, which will be explained in the sequel.

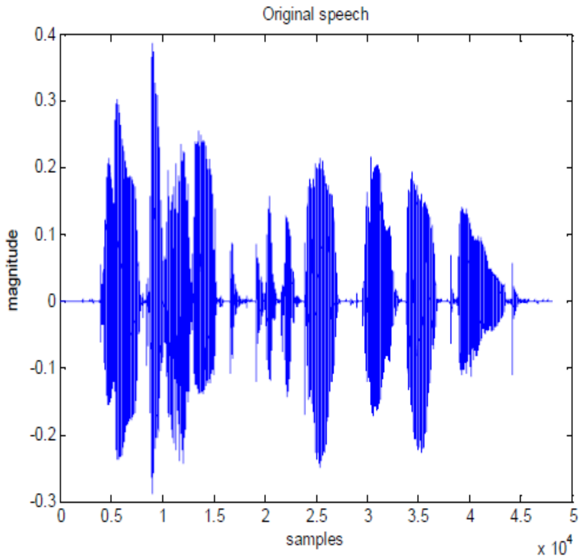


The main idea behind the voice-excitation is to avoid the imprecise detection of the pitch and the use of an impulse train while synthesizing the speech. One should rather try to come up with a better estimate of th excitation signal. Thus the input speech signal in each frame is filtered with the estimated transfer function of LPC analyzer. This filtered signal is called the residual. If this signal is transmitted to the receiver one can achieve a very good quality. To achieve a high compression rate ,the discrete cosine transform (DCT) of the residual signal could be employed. The DCT concentrates most of the energy of the signal in the first few coefficients. Thus one way to compress the signal is to transfer only the coefficients, which contain most of the energy. The tradeoff, however, is paid by a higher bit rate, although there is no longer a need to transfer the pitch frequency and the voiced /unvoiced information. We therefore looked for a solution to reduce the bit rate to 16 kbits/sec.

IV. IMPLEMENTATION

The project has been implemented in MatlabR2009a. It has been divided into 3 parts namely basic LPC vocoder, Voice excited LPC model compressed using DCT, Voice excited LPC model compressed without using DCT. The waveform generated by each of these techniques have been plotted and analysed.

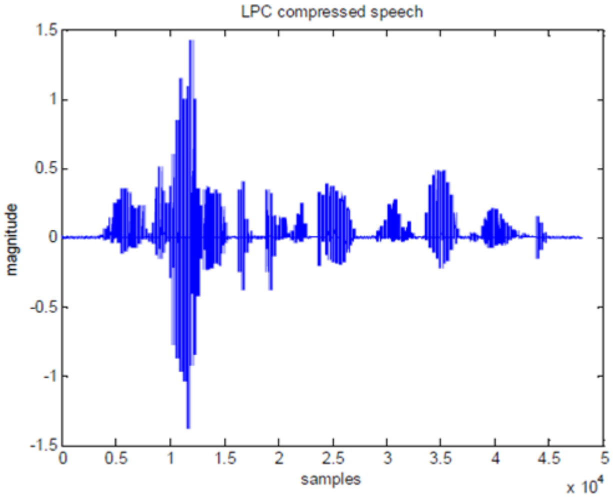
A) *Original Speech*



B) *Basic LPC Vocoder:*

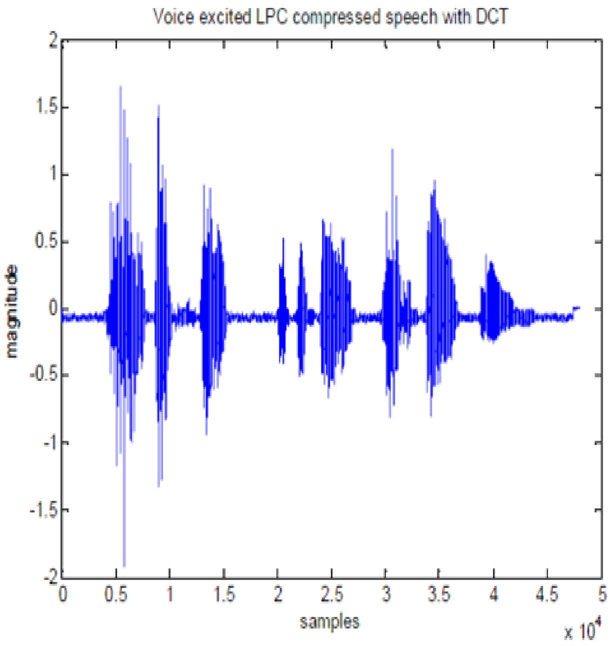
For implementing the basic LPC Vocoder, the pitch period is assumed to be 7.5ms. The filter coefficients have been evaluated using the Levinson-Durbin recursion algorithm. The original speech is recovered from the coefficients by passing it though a train of impulses which models the voiced sections of the speech.

The plot of the generated waveform is shown in



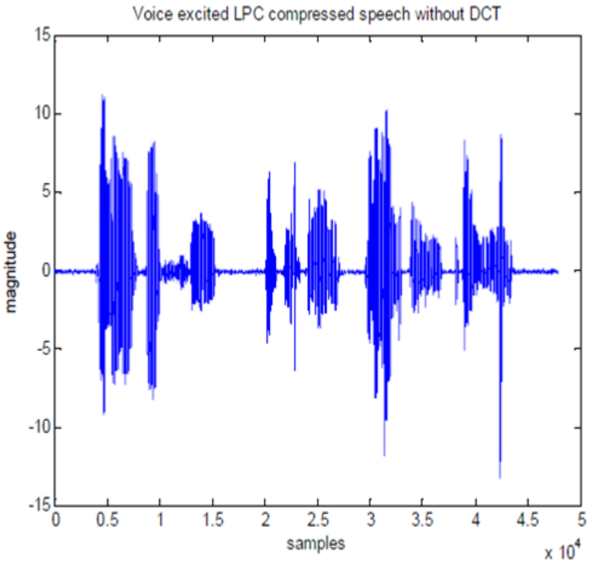
C) *Voice excited LPC Model with Discrete Cosine transform:*

In this implementation, the residual signal is compressed by using Discrete Cosine Transform (DCT). Since most of the energy of the signal is concentrated in the first few coefficients, the first 50 coefficients of DCT are considered in our implementation to compress the speech. These coefficients are the encoded using a scalar quantizer with each coefficient taking up 8 bits. The recovered signal after decoding is shown in



D) *Voice excited LPC Model without Discrete Cosine transform:*

The quality of the compressed speech can be improved but at the cost of a higher bit rate. This is achieved by transmitting the encoding the residual signal as a whole without using DCT. This will help us achieve better reconstruction of the transmitted signal. The recovered signal from this method is shown in



The original speech signal is then compared with the above mentioned results to check the effectiveness of each implementation.

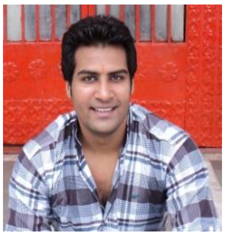
V. CONCLUSIONS

The results achieved from the voice excited LPC are intelligible. On the other hand, the plain LPC results are much poorer and barely intelligible. This first implementation gives an idea on how a vocoder works, but the result is far below what can be achieved using other techniques. Nonetheless the voice-excited LPC used gives understandable results and is not optimized. The tradeoffs between quality on one side and bandwidth and complexity on the other side clearly appear here. If we want a better quality, the complexity of the system should be increased or a larger bandwidth has to be used. Since the voice-excited LPC gives pretty good results with all the required limitations of this project, we could try to improve it. A major improvement could come from the compression of the errors. If we can send them in a loss-less manner to the synthesizer, the reconstruction would be perfect. An idea could be the Use of Huffman code for the DCT.

REFERENCES

1. L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Prentice- Hall, Englewood Cliffs, NJ.
2. B. S. Atal, M. R. Schroeder, and V. Stover, "Voice- Excited Predictive Coding Systetm for Low Bit-Rate Transmission of Speech", Proc. ICC, pp.30-37 to 30-40
3. http://www.data-compression.com/speech.html
4. C. J. Weinstein, "A Linear Predictive Vocoder with Voice Excitation", Proc. Eascon, September.
5. Speech coding. a tutorial review by Andres s. Spanias member IEEE.
6. Proakis John G. 'Digital Communications' .New York: Macmillan Pub. Co,
7. Haykin Simon. 'Digital Communication'. New
8. M. H Johnson and A. Alwan, " Speech Coding: Fundamentals and Applications", to appear as a chapter in the encyclopedia of telecommunications, Wiley, December 2002.
9. Orsak, G.C rt al, "Collaborative SP education using the internet and MATLAB" IEEE Signal Processing Magazine, Nov, 2009 vol 12, no6, pp 23-32.

FIRST AUTHOR: NIKHIL SHARMA



(RESEARCH SCHOLAR) NIT KURUKSHETRA
EMAIL ID- niks4u31@gmail.com

SECOND AUTHOR: NIHARIKA MEHTA

(RESEARCH SCHOLAR) MRIU FARIDABAD
EMAIL ID- niharika.mehta18@yahoo.com