Spoken Digit Compression: A Comparative Study between Discrete Wavelet Transforms and Linear Predictive Coding

Shijo M Joseph
School of Information
Science & Technology
Kannur University Kannur
Kerala India, 670 567

Firoz Shah A
School of Information
Science & Technology
Kannur University Kannur
Kerala India, 670 567

Babu Anto P
School of Information
Science & Technology
Kannur University Kannur
Kerala India, 670 567

ABSTRACT
In modern world communication service providers are continuously met with the challenge of accommodating more users with in a limited allocated bandwidth. Due to this motivation service providers and manufactures are continuously in search of low bit rate speech coders that deliver high quality speech.

This paper deals with spoken digit compression. Linear predictive coding and discrete wavelet transforms are used to materialise speech compression. We used Malayalam one of the south Indian language for this experiment. We could successfully compress and reconstruct Malayalam spoken words with perfect audibility using LPC and db4 wavelets**.** From the result we can see that the performance of wavelet transform is better than LPC.

General Terms

Signal Processing systems , Wavelet Representation.

Keywords

Speech compression, Linear Predictive coding, Discrete wavelets Transforms, Malayalam Spoken digits
.

1. INTRODUCTION
For the purpose of communication and storage, it is necessary to convert it into an electrical signal. This electrical representation of speech has certain properties such as [1].

1. It is a one-dimensional signal, with time as its independent variable.

2. It is random in nature.

3. It is non-stationary, i.e. the frequency spectrum is not constant in time.

4. Although human beings have an audible frequency range of 20Hz–20kHz, the human speech has significant frequency components only up to 4kHz, a property that is exploited in the compression of speech. Compression is the process of converting an input data stream into another data stream that has a smaller size and it is possible only because data is normally represented in the computer in a format that is longer than necessary i.e. the input data has some amount of redundancy associated with it. One of the main objective of speech compression systems is to eliminate this redundancy [2]. When compression is used to reduce storage requirements, overall program execution time may be reduced. This is because reduction in storage will result in the reduction of secondary memory access attempts. With respect to transmission of data, the data rate is reduced at the source by the compressor (coder), it is then passed through the communication channel and returned to the original rate by the expander (decoder) at the receiving end. The compression algorithms help to reduce the bandwidth requirements and also provide a level of security for the data being transmitted. In a mobile phone network, if speech compression is used, more users can be accommodated at a given time because of the lesser bandwidth [3]. Speech compression has many applications in teleconferencing and wireless communications. Study about speech compression is important. However, it is more important to ensure that compression retains the integrity of the speech. If the data is distorted in some way, it becomes difficult to understand. Thus, speech compression needs to be performed in a way which retains the key qualities of the data. This work is a comparison study between two flexible compression schemes that are based on daubechies wavelet and linear predictive coding. This paper is organized as follows: Section 2 explains the techniques used for feature extraction such as Discrete Wavelet Transform (DWT) and Linear Predictive Coding (LPC-10). Section 3 deals with the Malayalam spoken digit database used for the experiment. Section 4 explores the details of the experiments and result. With section 5 we conclude the paper.

2. TECHNIQUE USED FOR FEATURE EXTRACTION
This part of the paper explains two promising techniques used for the analysis / synthesis coding of speech signal. They are discrete wavelet transforms and linear predictive coding.

2.1 Discrete wavelet transform
The first phase of this experiment is handled with Discrete Wavelet Transforms. Discrete Wavelet Transform (DWT) is one of the most promising signal compression techniques which use multi resolution filter banks for the analysis. The Discrete Wavelet Transform is defined by the following equation [4].

$$W(J, K) = \Sigma_J \Sigma_K X(K)\, 2^{-J/2} \Psi(2^{-J}N-K) \quad (1)$$

Where ψ (t) si the basic analyzing function called the mother wavelet In DWT a time-scale representation of a signal is obtained by digital filtering techniques. Low frequency components of a signal are more significant than its high frequency components, since the low frequency components characterize a signal more than its high frequency components. The DWT is computed by successive low pass filtering and high pass filtering of the discrete time domain signal. This algorithm is called the Mallat algorithm [5]. At each level the decomposition of the input signal has two kinds of outputs. The low frequency components, known as the approximations a[n]and high frequency components, known as the details d[n]. At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. With this approach, the time resolution becomes good at high frequencies, while the frequency resolution becomes good at low frequencies [6]. The filtering and decimation process is continued until the desired level is reached. The DWT of the original signal is then obtained by concatenating all the coefficients, a[n] and d[n], starting from the last level of decomposition.

The successive high pass and low pass filtering of the signal can be depicted by the following equations

$$Y_{high}[k]= \Sigma_n x[n]g[2k-n] \quad (2)$$

$$Y_{low}[k]= \Sigma_n x[n]h[2k-n] \quad (3)$$

Where Y_{high} and Y_{low} are the outputs of the high pass and low pass filters obtained by sub sampling by 2 [9] .

2.2 Linear predictive coding

LPC is used to estimate basic speech parameters like pitch formants and spectra. The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech and estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients. These predictor coefficients are normally estimated in every frame, size of 20 ms long. The predictor coefficients are represented by ak. Another important parameter is the gain (G). The transfer function of the time –varying digital filter is given by (4)

$$H(z) = \frac{G}{1-\sum a_k z^{-k}} \quad (4)$$

The summation is computed starting at k=1 up to p, which will be 10 for the LPC-10 algorithm. This means that only the first 10 coefficient are transmitted to the LPC synthesizer. The two most commonly used methods to compute the coefficients are the covariance method and the auto correlation formulation. For our implementation we will be using the auto correlation formulation. The reason is that this method is superior to the covariance method in the sense that the roots of the polynomial in the denominator of the above equation is always guaranteed to be inside the unit circle. Hence guaranteeing the stability of the system H (z). levinson – Durbin recursion will be utilized to compute the required parameters for the auto-correlation method [7].

The LPC analysis of each frame also involves the decision-making process of concluding if a sound is voiced or unvoiced. If a sound is decided to be voiced, an impulse train is used to represent it, with non zero taps occurring every pitch period. A pitch –detecting algorithm is employed to determine the correct pitch period/ frequency. We used the autocorrelation function to estimate the pitch period [8]. However if the frame is unvoiced, then white noise is used to represent it and a pitch period of T=0 is transmitted. Therefore, either white noise or impulse train becomes the excitation of the LPC synthesis filter. It is important to re-emphasize that the pitch, gain and coefficient parameters will be varying with time from one frame to another.

3. DATABASE USED FOR THE EXPERIMENT.

The speech signals are chosen from the Malayalam (one of the south Indian languages) spoken digit database. We have used a speaker independent database for our experiment. Our data base consists of 10 isolated Malayalam spoken digits from 0 to 9 by 10 different speakers. The speech samples are recorded from five male and five female speakers. The speakers were free from speech disabilities. The speech is recorded using high quality studio recording microphone, at a sampling rate of 8 KHz (4 KHz band limited). The recorded speech is processed, labeled and stored in the data base. The spoken digits in the data base and their IPA format are given in table1.

Table 1 speech data base and their IPA format

Number digit	Words in Malayalam	IPA format
0	പൂജ്യം	/pu: //ʤ *y*a/ /ṁ/
1	ഒന്ന്	/o//*nn*/
2	രണ്ട്	/ɾ a//*ṇṭ*/
3	മൂന്ന്	/mu: //*nn*/
4	നാല്	/n̪a: //l/
5	അഞ്ച്	/a//***ñc***/

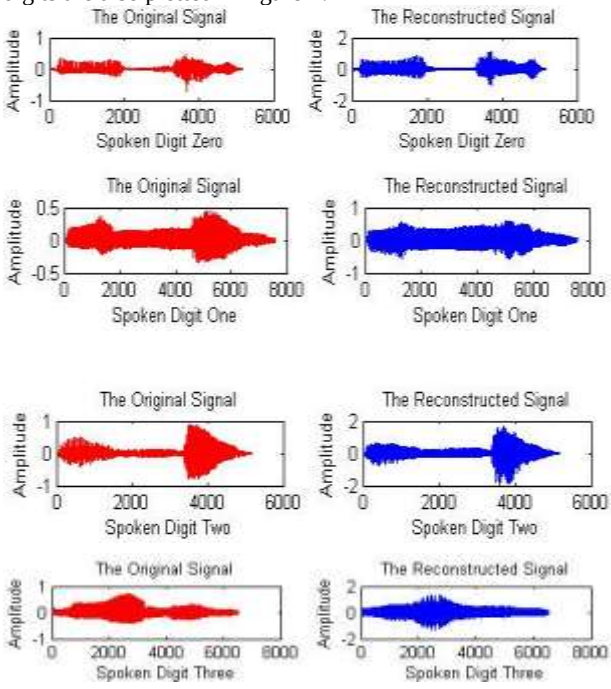
6	ആറ്	/a: //r/
7	ഏഴ്	/e: //ɻ /
8	എട്ട്	/e//ʈʈ/
9	ഒൻപത്	/o//*n*a//pa//t̪/

4. EXPERIMENT AND RESULTS

The speech compression is materialized by using linear predictive coding and daubechies wavelet transformation.

4.1 DWT experiment

The speech compression is materialized by using discrete wavelet transformation using daubechies wavelets. The samples in the data base are compressed using db4 wavelets. In each level of compression the speech signal is compressed without losing its audibility by splitting it into high frequency and low frequency components. In each level of compression the numbers of samples are down sampled by a factor of 2. The original form of speech signal for the digit zero (/pu: / /ʤya/ /ṁ//) is given in the figure 1 with the reconstructed one. The compressed signal is chosen from the third level of the decomposition using db4 wavelet. The reconstructed signal from the compressed one ensures perfect audibility. Second digit one (//o//*nn*/) is plotted in original, and in reconstructed form. Similarly remaining eight digits are also plotted in figure 1.



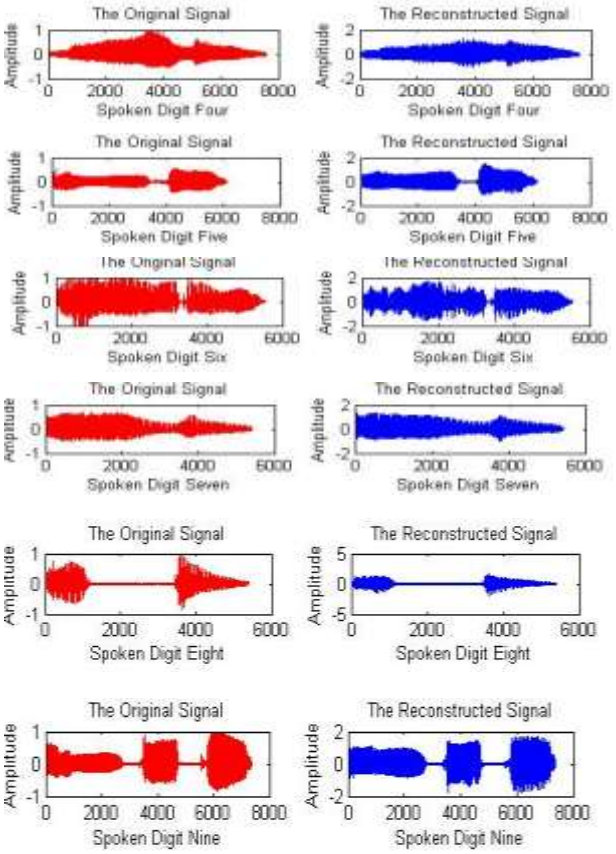


Fig.1

4.2 LPC experiment

In LPC analysis the input signal is broken up in to frames or blocks. The 8000 samples in each second of speech signal are broken in to 240 sample segments which means that each frame represents 30 milliseconds of the speech signal. This frame size gives an intelligible speech with good compression. The signal is passed through a low pass filter with bandwidth 1 KHz to split up the signal in to voiced and unvoiced sound. The voiced sounds have very high amplitude since they have average energy level. The voiced sounds have distinct resonant or formant frequencies. Unvoiced sounds have less energy and therefore smaller amplitudes. Pitch period of the voiced sound is found by applying Average Magnitude Difference Function (AMDF). Since pitch period (P) for humans is limited, the AMDF is evaluated for a limited range of the possible pitch period values.

The samples in the data base are compressed by LPC. In LPC compression the speech signal is compressed without losing its audibility by breaking the speech in to segments and then sending the voiced/ unvoiced information, the pitch period and the coefficient for the filter. The original form of speech signal for the digit zero (/pu: / /ʤya/ /ṁ//) is given in the figure 2 with the reconstructed one. Second digit one (//o//*nn*/) is plotted in original, and in reconstructed form. Similarly remaining eight digits are also plotted in figure 2.

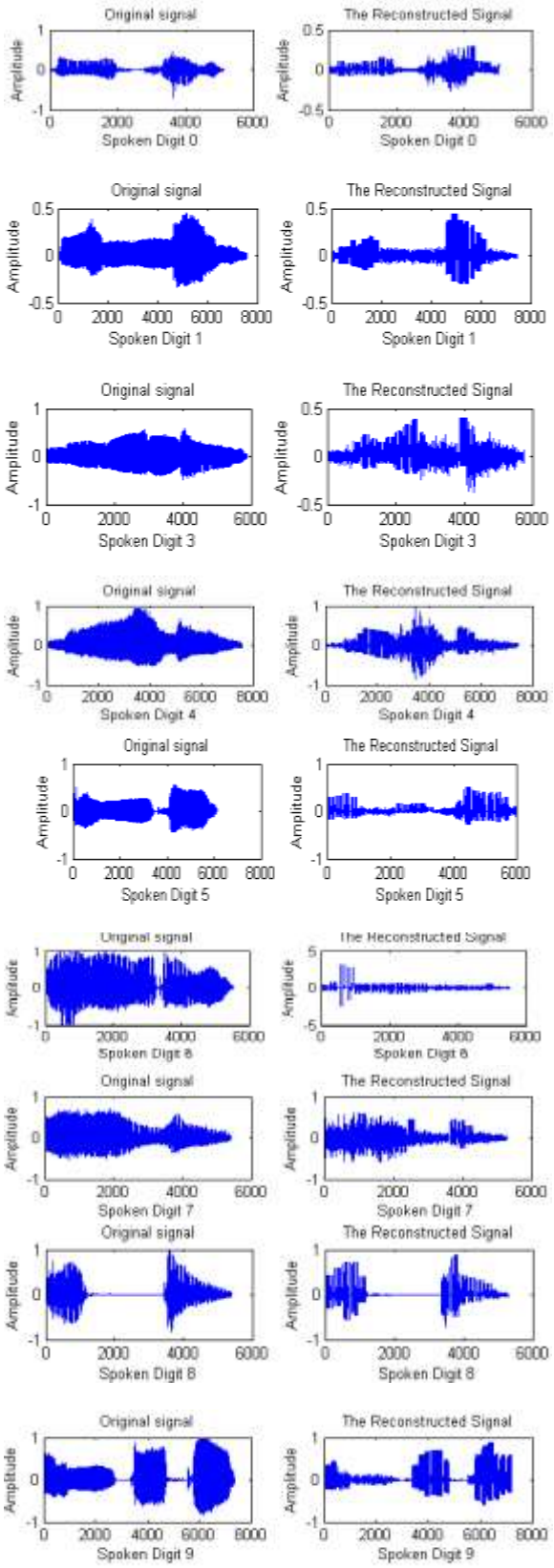


Fig. 2

4.3. Performance measures.

The distortion or error caused in the recovered signal by the signal compression process is measured in several ways. The standard distortion measures are [10]

4.3.1 *Signal to noise ratio* (SNR)

$$SNR = 10\log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \quad (5)$$

σ_x^2 is the mean square of the speech signal and σ_e^2 is the mean square difference between the original and reconstructed signals.

4.3.2 *Peak signal to noise ratio (PSNR)*

$$PSNR = 10\log_{10}\frac{NX^2}{||x-r||^2} \quad (6)$$

Where N is the length of the reconstructed signal, X is the maximum absolute square value of the signal x and $||x-r||^2$ is the energy of the difference between the original and reconstructed signals.

4.*3.3 Normalised root mean square error (NRMSE)*

$$NRMSE = \sqrt{\frac{x\ \ x(n)-r(n)^2}{x(n)-\mu_x(n)^2}} \quad (7)$$

Where x(n) is the speech signal, r(n) is the reconstructed signal, and $\mu_x(n)$ is the mean of the speech signal.

4.3.4 Compression ratio(CR)

$$C = \frac{Length(X(n))}{Length(cWC)} \quad (8)$$

cWC is the length of the compressed wavelet transform vector.

4.3.5 M*ean opinion score (MOS).*

Subjective evaluation by listeners is still a method commonly used in measuring quality of reconstructed speech. MOS provides a numerical indication of the perceived quality of received media after compression and / or transmission. The MOS is expressed as a single number in the range 1 to 5. Where 1 is the highest perceived quality and 5 is the lowest perceived quality. When taking subjective test, listeners focus on the difference between the original and reconstructed signal and rateing it.

The MOS is generalized by averaging the result of a set of standard, subjective tests, where a number of listeners rate the reconstructed speech signals. In our experiment the MOS rate is 1.02 in Discrete wavelet transforms and 1.16 in LPC coding. The result of these analyses is given in table 2.

Table. 2

ANALYSIS USING LPC & DWT FOR MALAYALAM SPOKEN DIGITS

Digit	SNR db		PSNR db		NRMSE	
	LPC-10	Db4	LPC-10	Db4	LPC-10	Db4
0	5.8367	25.8254	86.6749	82.2533	2.7815	0.0140
1	5.6609	19.9395	86.7188	72.4854	2.9621	0.0789
2	5.8036	24.5111	86.2418	59.9067	2.7361	0.1127
3	5.7907	15.5768	86.3815	56.0216	2.9652	0.1108
4	5.8120	23.4511	86.64	72.5787	2.8393	0.096
5	5.7146	29.352	86.8529	62.7256	2.9698	0.1098
6	5.7694	19.2491	86.7827	68.4288	2.7755	0.1036
7	5.8377	25.6482	86.3023	68.3727	2.8901	0.0784
8	5.7533	23.0929	86.0007	76.6237	3.0364	0.0635
9	5.7113	15.2427	86.4251	66.034	3.0072	0.2252

Fig 3. Performance Analysis based on SNR

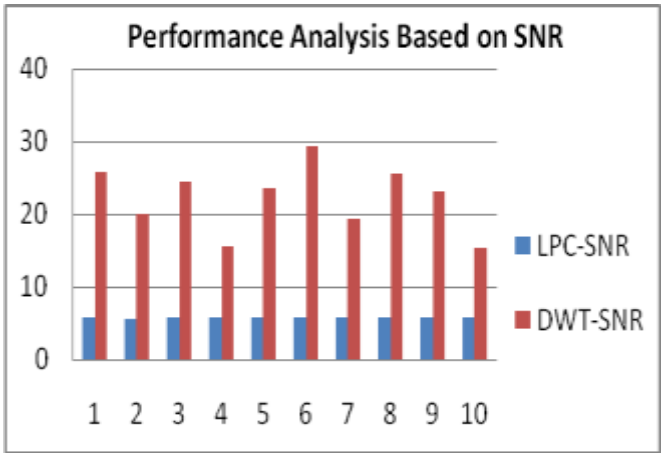


Fig-3

Fig .4. Performance Analysis based on PSNR

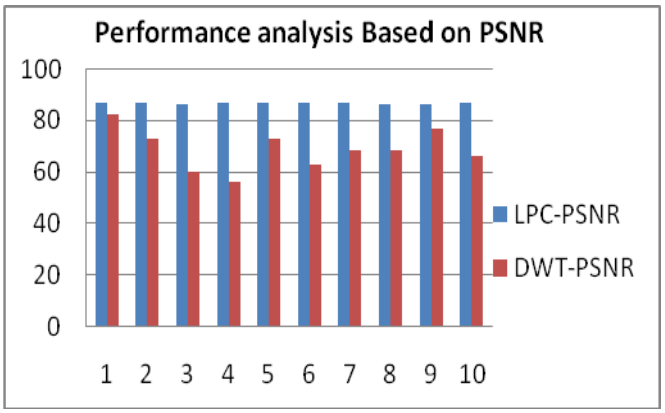


Fig-4

Fig 5. Performance Analysis based on NRMSE

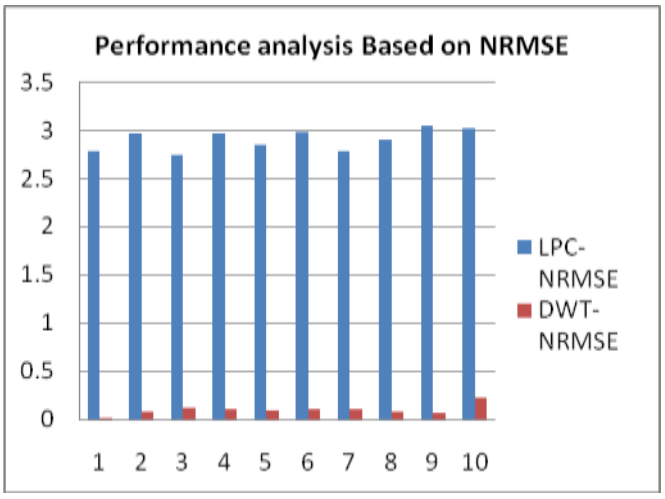


Fig-5

5. CONCLUSION

The performance of two promising compression techniques such as Discrete Wavelet Transforms and LPC-10 on spoken digits was tested and the following points were observed. High compression ratios were attained with acceptable SNR. In wavelet experiment further enhancement beyond the level 3 decomposition affect the intelligibility of the spoken digits. In LPC-10 experiment SNR ,PSNR and NRMSE remained almost same for all experiments with insignificant changes. But in DWT these parameters were varied.

In this work the compressed signals can be reconstructed back to its initial form with full audibility. In general a good reconstructed signal is the one with low NRMSE and high PSNR. That means that the signal has low error and high signal fidelity. In this work we have obtained average NRMSE as 0.09927 and average PSNR 67.0916 in DWT experiment. In LPC-10 experiment we get average NRMSE is 2.8963 and PSNR is 86.5020. This shows that both the techniques are effective. However LPC is not a very effective technique for representing quasi-stationary signals like speech because LPC assumes speech signals are stationary within a

given time and frame. If frame size is larger LPC technique is not able to analyze localized events accurately. The Wavelet transforms proved to be a useful tool for analysis of non-stationary signals. It uses short windows for high frequencies and long window at low frequencies. This results in multi-resolutions analysis by which the signal is analyzed with different resolutions at different frequencies.

The obtained results are tabulated in table2 and comparative analyses of different parameters are plotted in figure 3, figure 4 and figure 5. From the plotted figure 3 we can say that the signal to noise ratio of the Wavelet has high value comparable with LPC-10 which produces a good quality speech. Figure 4 shows that PSNR is also comparatively good in Wavelet based experiment that means high signal fidelity. Figure 5 prove that normalized root mean square error is comparatively very less in Wavelet decomposition. The measuring quality of reconstructed speech through Mean Opinion Score (MOS) is higher in wavelets. From the stated results we can conclude that wavelet is more ideal for speech compression applications.

REFERENCES

[1] Lawrence R Rabiner 1978 "Theory and application of Digital Signal Processing", Prentice Hall Publication 2nd edition .

[2] Kalid Sayood 2005 "Introduction to data Compression" ,Morgan Kaufmann Publishers 2edition.

[3] Jalal Karam, 2007 'Various Speech Processing Techniques for Speech Compression and Recognition", Proceeding of world academy of science, engineering and technology volume 26 December ISSN1307- 6884,.

[4] Y.T Chan 1995 "Wavelet Basics", Kulwer Academic Publications.

[5] Daubechies,1992 "Ten lectures on wavelets", society for industrial and applied mathematics.

[6] J.S Walker 1999 "Wavelets and their Scientific Applications", Chamman and Hall/CRC.

[7] Amol R Madane, Zalak Shah and Raina Shah 2009 "Speech compression using Linear predictive coding", proceedings International workshop on Machine Intelligence Research MIR labs.

[8] Mahmoud A. Osman, Nasser AI, Hussein M. Magboub and S.A Alfandi 2010 " Speech Compression Using LPC and Wavelet" 2nd International Conference on Computer Engineering and Technology.

[9] Jalal Karam 2006 "The effect of Different Compression Schemes on Speech Signals" World Academy of Science , Engineering and Technology.

[10] http://www.otolith.com/pub/u/howitt/lpc.tutorial.html