

SPEECH CODING WITH LINEAR PREDICTIVE CODING

G.CHENCHAMMA, P.L.CHOWDARY, K. SHALINI KATYAYANI
DEPARTMENT OF ELECTRONICS AND COMPUTERS ENGINEERING,
KLCE, GUNTUR, ANDHRA PRADESH, INDIA

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. In this paper we proposed a technique called LPC. 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 both subjective and objective 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. The objective analysis will be performed by computing Segmental Signal to Noise Ratio (SEGSNR) between the original and the coded speech signal.

Keywords: Waveform coding, Frequency bandwidth, speech coding, Segmental Signal to Noise Ratio

1. INTRODUCTION

A speech coder consists of two components: the encoder and the decoder. Speech is a time varying waveform. The analog speech signal $s(t)$ is first sampled at the rate $f_s \geq 2f_{max}$, where f_{max} is the maximum frequency content of $s(t)$. The sampled discrete time signal is denoted by $s(n)$. This signal is then encoded using one of several coding schemes such as PCM (pulse code modulation) or predictive coding. In predictive coding the encoder considers a group of samples at a time, extracts coefficients that can model those samples concisely, converts those coefficients to binary bits and transmits them. In this way the encoder encodes the speech signal in a compact form using fewer bits. The decoder reconstructs the speech signal from those transmitted parameters. The whole process is illustrated in Fig.1.

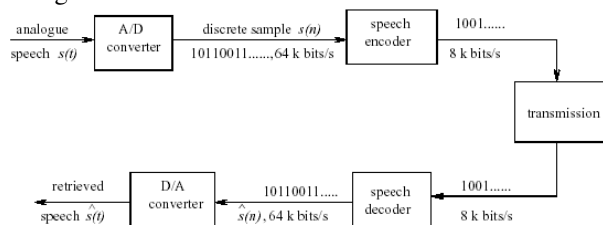


Figure 1: Speech Coding.

One of the major performance measures of speech coding is determined by how well the code speech is perceived. If the redundancies of the speech signal can be found adequately, and if the perceptual properties of

the ears are exploited properly, good audible performance can be achieved at low bit rates. The human hearing system acts like a filter bank and is most sensitive to the 200–5600 Hz frequency range in terms of perception. Important perception features, for instance voicing, are determined from a harmonic structure which is present at low frequencies (the harmonic structure does not go beyond 3 kHz). Voiced speech has a periodic or quasi periodic character. Poorly reproduced periodicity in the reconstructed voiced segment causes a major audible distortion. Perceptual aspects, such as the amplitude envelope, the amplitude and location of the first three formants and the spacing between the harmonics are found in the frequency domain. The first three formants are usually located below 3 kHz. The manner and place of articulation are other important perceptual features. The manner of articulation affects low frequencies. The place of articulation affects the second formant region, above 1 kHz. An unvoiced speech segment can be replaced by a noise-like signal with a similar spectral envelope, without significant auditory distortion.

2. 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 [1] technique. 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 40 ms. For each frame, the gain and only 18 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 non-speech signals.

3. METHODOLOGY

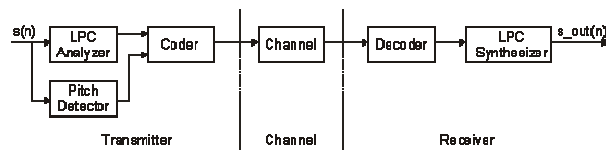


Figure 2: Block diagram of an LPC vocoder.

In this section an explanation of the LPC speech coding technique will be given. The specific modifications and additions done to improve this algorithm will also be covered. However, before jumping into the detailed methodology of our solution, it will be helpful to give a brief overview of speech production. Speech is produced when velum is lowered to make it acoustically

coupled with the vocal tract. Nasal sounds of speech are produced this way. Speech signals consist of several sequences of sounds. Each sound can be thought of as unique information. There are voiced and unvoiced types of speech sounds. The fundamental difference between these two types of speech sounds comes from the way they are produced. The vibrations of the vocal cords produce voiced sounds. The rate at which the vocal cords vibrate dictates the pitch of the sound. On the other hand, unvoiced sounds do not rely on the vibration of the vocal cords. The unvoiced sounds are created by the constriction of the vocal tract. The vocal cords remain open and the constrictions of the vocal tract force air out to produce the unvoiced sounds. LPC technique will be utilized in order to analyze and synthesize speech signals. This method is used to successfully estimate basic speech parameters like pitch, formants and spectra. A block diagram of an LPC vocoder can be seen in Fig.2. The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and the 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 every frame, which is normally 40 ms long. The predictor coefficients are represented by a_k . Another important parameter is the gain (G). The transfer function of the time-varying digital filter is given by

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

The summation is computed starting at $k=1$ up to p , which will be 10 for the LPC-10 algorithm, and 18 for the improved algorithm that is utilized. This means that only the first 18 coefficients are transmitted to the LPC synthesizer. The two most commonly used methods to compute the coefficients are, but not limited to, 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. The block diagram of simplified model for speech production can be seen in Fig.3.

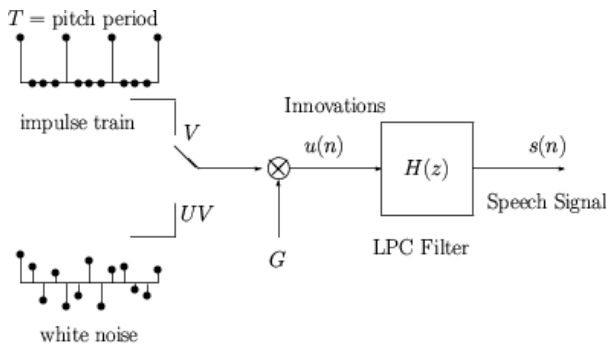


Figure 3: Simplified model for speech production.

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 nonzero taps occurring every pitch period. A pitch-detecting algorithm is employed to determine to correct pitch period / frequency. We used the autocorrelation function to estimate the pitch period. 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.

3.1 Pre-emphasis Filter

From the speech production model it is known that the speech undergoes a spectral tilt of -6dB/oct . To counteract this fact a pre-emphasis filter of the following form is used:

$$y(n) = x(n) - \alpha \cdot x(n-1) \dots\dots\dots(2)$$

The reconstruction of the speech signal and is as follows:

$$y(n) = x(n) + \alpha \cdot x(n-1) \dots\dots\dots(3)$$

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 as

can be seen in Fig.4. Clearly the coefficients corresponding to higher frequencies can be better estimated.

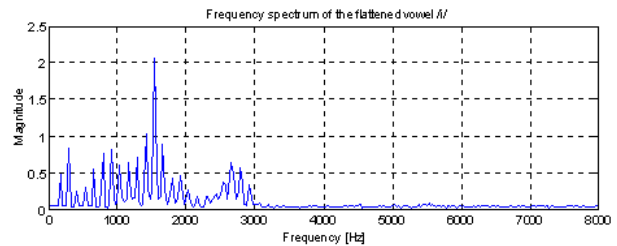


Figure 4: Frequency spectrum of the vowel /i/ in the word nine.

3.2 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 known from the literature that one solution to improve the quality of the sound is the use of voice-excited LPC vocoders. Systems of this type have been studied by Atal et al. and Weinstein. Fig.5. shows a block diagram of a voice-excited LPC vocoder. The main difference to a plain LPC-10 vocoder, as shown in Fig.6, is the excitation detector, which will be explained in the sequel.

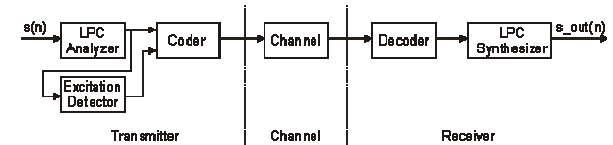


Figure 5: Block diagram of a voice-excited LPC vocoder.

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 the 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. 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, which is described in the following section.

4. PERFORMANCE ANALYSIS

4.1 Bit Rates

In the sequel the necessary bit rates of the two solutions are computed. The bit rate for a plain LPC vocoder is shown in Table 1 and the bit rate for a voice-excited LPC vocoder with DCT is printed in Table 2. The following parameters were fixed for the calculation:

- Speech signal bandwidth $B = 8$ kHz
- Sampling rate $F_s = 16000$ Hz (or samples/sec.)
- Window length (frame): 20 ms
which results in 320 samples per frame by the given sampling rate F_s
- Overlapping: 10 ms (overlapping is needed for perfect reconstruction)
hence the actual window length is 30ms or consists of 480 samples
- There are 50 frames per second
- Number of predictor coefficients of the LPC model = 18 (see calculation in)

Table 1: Bit rate for plain LPC vocoder

	Number of bits per frame
Predictor coefficients	$18 * 8 = 144$
Gain	5
Pitch period	6
Voiced/unvoiced switch	1
Total	156
Overall bit rate	$50 * 156 = 7800$ bits / second

Table 2: Bit rate for voice-excited LPC vocoder with DCT

	Number of bits per frame
Predictor coefficients	$18 * 8 = 144$
Gain	5
DCT coefficients	$40 * 4 = 160$
Total	309
Overall bit rate	$50 * 309 = 15450$ bits / second

4.2 Objective Performance Evaluation

We measured the segmental signal to noise ratio (SEGSNR) of the original speech file compared to the coded and reconstructed speech file using the provided

Matlab-function "segsnr". The obtained results are as follows:

- 1) A Male speaker saying: "The students of ECM in KLU."
- 2) A Female speaker saying: "The students of ECM in KLU."

Vocoder Type	SNR 1	SNR 2
Plain LPC	-20.00 dB	-20.00 dB
Voice-excited LPC	0.4426 dB	0.6553 dB

5. QUALITY

5.1 Subjective quality

A comparison of the original speech sentences against the LPC reconstructed speech and the voice-excited LPC methods were studied. In both cases, the reconstructed speech has a lower quality than the input speech sentences. Both of the reconstructed signals sound mechanized and noisy with the output of plain LPC vocoder being nearly unintelligible. The LPC reconstructed speech sounds guttural with a lower pitch than the original sound. The sound seems to be whispered. The noisy feeling is very strong. The voice-excited LPC reconstructed file sounds more spoken and less whispered. The guttural feeling is also less and the words are much easier to understand. Overall the speech that was reconstructed using voice-excited LPC sounded better, but still sounded muffled. The waveforms in Fig.5 give the same idea. The voice-excited waveform looks closer to the original sound than the plain LPC reconstructed one.

5.2 Segmental signal to noise ratio

Looking at the segmental SNR, computed in section 4.2, it is obvious that the first sound is very noisy, having a negative SNR. The noise in this file is even stronger than the actual signal. The voice-excited LPC encoded sound sounds far better, and its SNR, although barely, is in the positive side. However, even the speech coded with the improved voice-excited LPC does not sound exactly like the original signal. It is noticeable that

both the plain LPC and the voice-excited vocoders are not sensitive to the input sentence, the result is the same for a sentence with many voiced sounds and for a sentence with many fricatives or other unvoiced sounds. The good point is that any spoken sentence can be transmitted with the same overall results. The disadvantage is that we cannot focus on a specific aspect of the vocoder that would give much poorer results. To improve the quality, the overall system has to be

improved. We cannot just improve the unvoiced sounds production to make the vocoder sound perfect.

5.3 Quality-performance tradeoffs

The LPC method to transmit speech sounds has some very good aspects, as well as some drawbacks. The huge advantage of vocoders is a very low bit rate compared to what is achieved for sound transmission. On the other hand, the speech quality achieved is quite poor.

6. DISCUSSION OF RESULTS

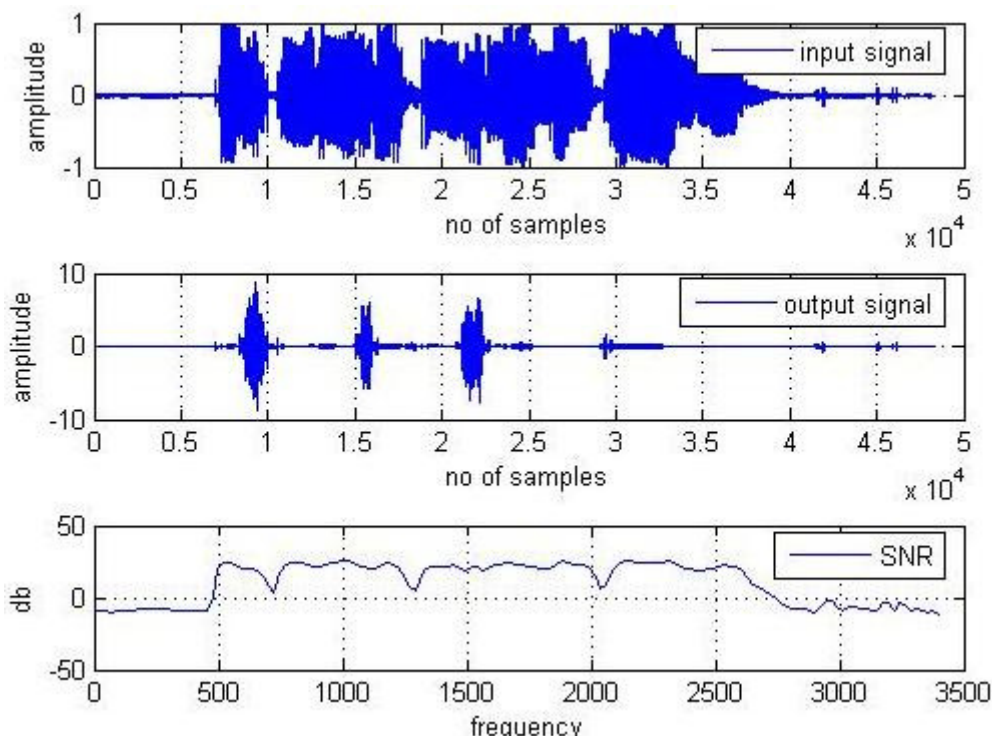


Figure 6: a)Waveform of the sentence “THE STUDENTS OF ECM IN KLU” b)LPC reconstructed speech signal c)SNR of the LPC reconstructed speech signal.

7. 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

coefficient. Many simulations have to be done to get the right code book.

REFERENCES

1. T.Tremain. The government standard linear predictive coding algorithm LPC-10 In speech Technology magazine, pages 40-49, April 1982.
2. L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals", Prentice- Hall, Englewood Cliffs, NJ, 1978.
3. B. S. Atal, M. R. Schroeder, and V. Stover, "Voice-Excited Predictive Coding System for Low Bit-Rate Transmission of Speech", Proc. ICC, pp.30-37 to 30-40, 1975
4. C. J. Weinstein, "A Linear Predictive Vocoder with Voice Excitation", Proc. Eascon, September 1975.