International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

Overview of Code Excited Linear Predictive Coder

Minal Mulye[1], Sonal Jagtap[2]
[1]PG Student, [2]Assistant Professor, Department of E&TC, Smt. Kashibai Navale College of Engg, Pune, India

***Abstract*— Advances in speech coding technologies have enabled speech coders to achieve bit-rate reductions at a great extent while maintaining roughly the same speech quality. One of the most important driving forces behind this feat is the analysis-by-synthesis paradigm. Code Excited Linear Predictive coder (CELP) is the quite efficient closed loop analysis-by-synthesis method for narrow and medium band speech coding systems. CELP algorithm can produce low- rate coded speech comparable to that of medium- rate waveform coders thereby bridging the gap between waveform coders and Vocoders. This paper gives the general overview and conceptual literature of this highly proficient speech coder.**

***Keywords*— Analysis-by-synthesis, CELP, speech coder, Vocoders, waveform coders.**

I. INTRODUCTION

In telecommunications industry, speech coding plays a very important role. Over the years the capabilities of such techniques have developed significantly due to the rising demand of better performance. The fundamental objective of any speech coder is to represent the analog speech into a digital stream of bits so that it can be sent over the internet using minimum bandwidth. Hence we can say that modern telecommunications demand optimum bandwidth utilization with minimum delay and distortion. To accomplish this, now-a-days low bit rate coders are used in almost every telecom devices. Both LPC and CELP are such two techniques that also follow the ITU-E G.729 standard.

Among them Code Excited Linear Prediction (CELP) is the newest form of voice coder that is in actuality an enhancement of the LPC coder. It is a lossy compression algorithm which is used for low bit rate transmission. In conventional LPC, the excitation waveform is either a pulse train for voiced speech or a noise like waveform for unvoiced speech. This rigid classification also ignores the possibility of mixed forms of excitation and more general excitation patterns. However, in CELP the excitation waveform is obtained by optimizing the positions and amplitudes of a fixed number of pulses to minimize an objective measure of the performance. Here the objective measure is the frequency weighted mean square error correction. This frequency weighting reflects the properties of the human auditory perception reasonably accurately.

Another extension is the use of a codebook which contains all the excitation signals. These reduce the computational complexity as now only the excitation index is to be transmitted instead of the entire signal. All these points motivated for the elemental study of CELP coder which is done with the MATLAB software.

II. THE CELP CONCEPT

The basic principle that all speech coders exploit is the fact that speech signals are highly correlated waveforms. Speech can be represented using an autoregressive (AR) model:

$$x(m) = \sum_{k=1}^{p} a_k\, x(m-k) + e(m) \qquad \text{Eq.1}$$

Each sample is represented as a linear combination of the previous p samples plus a white noise. The weighting coefficients $a_1, a_2, \ldots, a_p$ are called Linear Prediction Coefficients (LPCs). We now describe how CELP uses this model to encode speech. The samples of the input speech are divided into blocks of *N* samples each, called frames. Each frame is typically 10-20 ms long. Each frame is divided into smaller blocks, of *l* samples (equal to the dimension of the VQ) each, called sub-frames. For each frame, we choose $a_1, a_2, \ldots, a_p$ so that the spectrum of $\{x_1, x_2, \ldots, x_M\}$, generated using the above model, closely matches the spectrum of the input speech frame. This is a standard spectral estimation problem and the LPCs $a_1, a_2, \ldots, a_p$ can be computed using the Levinson- Durbin algorithm.

Writing Eq. (1) in z-domain, gives

$$\frac{X(z)}{E(z)} = \frac{1}{1-(a_1 z^{-1} + a_2 z^{-2} + \ldots. + a_L z^{-L})} = \frac{1}{A(z)} \qquad \text{Eq.2}$$

From equations (1) and (2), we see that if we pass a 'white' sequence $e[n]$ through the filter $1/A(z)$, we can generate $X(z)$, a close reproduction of the input speech.

The block diagram of a CELP encoder is shown in Fig.1. There is a codebook of size *M* and dimension *l*, available to both the encoder and the decoder.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

The code vectors have components that are all independently chosen from *N* (0, 1) distribution so that each code vector has an approximately 'white' spectrum. For each sub frame of input speech (*l* samples), the processing is done as follows: Each of the code vectors is filtered through the two filters (labeled 1/*A*(*z*) and 1/ *B*(*z*)) and the output y_l is compared to the speech samples. The code vector whose output best matches the input speech (least MSE) is chosen to represent the sub frame.

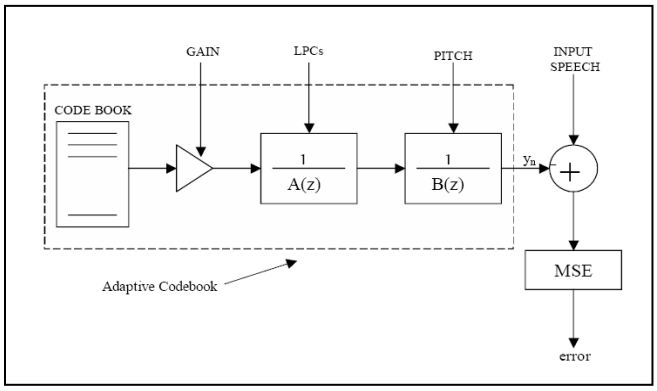

Fig.1: Basic CELP scheme

The first of the filters, 1/ *A*(*z*) , is described by Eq.(2). It shapes the 'white' spectrum of the Code vector to resemble the spectrum of the input speech. Equivalently, in time-domain, the filter incorporates short-term correlations (correlation with P previous samples) in the white sequence. Besides the short-term correlations, it is known that regions of voiced speech exhibit long term periodicity. This period, known as pitch, is introduced into the synthesized spectrum by the pitch filter 1/ *B*(*z*) . The time domain behavior of this filter can be expressed as:

$$y[n] = x[n] + y[n\text{-}T]$$

Where *x*[*n*] is the input, *y*[*n*] is the output and *T* is the pitch.

The speech synthesized by the filtering is scaled by an appropriate gain to make the energy equal to the energy of the input speech. To summarize, for every frame of speech, we compute the LPCs and pitch and update the filters. For every sub-frame of speech (*l* samples), the code vector that produces the 'best' filtered output is chosen to represent the sub-frame.

The decoder receives the index of the chosen code vectors and the quantized value of gain for each sub-frame. The LPCs and the pitch values also have to be quantized and sent every frame for reconstructing the filters at the decoder. The speech signal is reconstructed at the decoder by passing the chosen code vectors through the filters.

An interesting interpretation of the CELP encoder is that of a forward adaptive VQ. The filters are updated every *N* samples and so we have a new set of code vectors y_l every frame. Thus, the dashed block in Fig.1 can be considered a forward adaptive codebook because it is 'designed' according to the current frame of speech.

III. Analysis Of Celp

A block diagram of CELP analysis-by-synthesis coder is shown in the Fig.2. It is called analysis by synthesis because we encode and then decode the speech at the encoder and then find the parameters that minimize the energy of the error signal. First LP analysis is used to estimate the vocal system impulse response in each frame. Then the synthesized speech is generated at the encoder by exciting the vocal system filter. The difference between the synthetic speech and the original speech signal constitutes an error signal, which is spectrally weighted to emphasize perceptual important frequencies and then minimized by optimizing the excitation signal. Optimal excitation sequences are computed over four blocks within the frame duration, meaning that the excitation is updated more frequently than the vocal system filter. In our implementation frame duration of 20ms is used for the vocal-tract analysis (160 samples of an 8 kHz sampling rate) and 5ms block duration (40 samples) for determining the excitation.

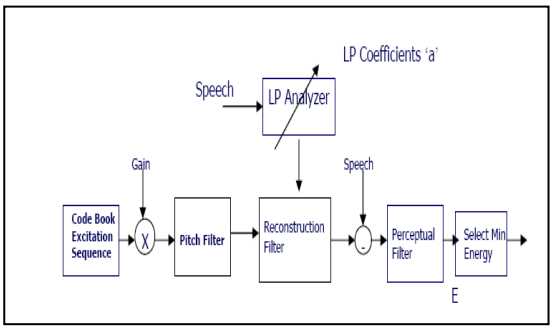

Fig.2: Block diagram of CELP

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

A. Required parameters

Looking at the encoder diagram, we see that we need to transmit five pieces of information to the decoder side for proper functioning.

The Liner Prediction Coefficients, 'a'

The Gain, 'G'
The Pitch Filter, 'b'
The Pitch Delay 'P'
The Codebook Index, 'k'

Following is an explanation of all the blocks and how we find these parameters.

B. LP Analysis

The linear prediction analysis estimates the all-pole (vocal-tract) filter in each frame, used to generate the spectral envelope of the speech signal. The filter typically has 10-12 coefficients. In our implementation it has 12 coefficients. MATLAB's 'lpc' function is used to obtain these coefficients however they can be obtained by implementing a lattice filter which acts both as a forward and backward error prediction filter. It gives us reflection coefficients which can be converted to filter coefficients. Levinson-Durbin method can be used effectively to reduce complexity of the filter.

$$\hat{x}(m) = \sum_{k=1}^{p} a_k\, x(m-k) \qquad \text{Eq.3}$$

So we define H(z) as the IIR reconstruction filter used to reproduce speech.

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{k=1}^{p} a_k\, z^{-k}} \qquad \text{Eq.4}$$

C. Perceptual weighting Filter

The output of the LP filter is the synthetic speech frame, which is subtracted from the original speech frame to form error signal. The error sequence is passed through a perceptual error weighting filter with system function

$$w(n) = \frac{A(z)}{A(z/c)} = c^{M}\, \frac{(p_0 - z)(p_1 - z)\ldots.(p_{M-1} - z)}{(cp_0 - z)(cp_1 - z)\ldots.(cp_{M-1} - z)} \qquad \text{Eq.5}$$

Where c is a parameter in the range 0 < c < 1 that is used to control the noise spectrum weighting. In practice, the range 0.7 < c < 0.9 has proved effective.

D. Excitation Sequence

The codebook contains a number of Gaussian signals which are used as the excitation signals for the filter. In our implementation we generated a codebook of 512 sequences each of length 5ms i.e. 40 samples. The codebook is known to the encoder as well as the decoder. The signal *e(n)* used to excite the LP synthesis filter is determined every 5 milliseconds within the frame under analysis. An excitation sequence is selected from a Gaussian codebook of stored sequenced, where k is the index. If the sampling frequency is 8 kHz and the excitation selection is performed every 5ms, then the codebook word size is 40 samples. A codebook of 512 sequences has been found to be sufficiently large to yield good-quality speech, and requires 9 bits to send the index.

E. Pitch Filter

Human voices have pitch in a few hundred hertz. For 8 kHz signal these frequencies correspond to pitch delay of 16 to 160 samples. For voiced speech, the excitation sequence shows a significant correlation from one pitch period to the next. Therefore, a long-delay correlation filter is used to generate the pitch periodicity in voiced speech. This typically has the form given by

$$J(z) = \frac{1}{1 - bz^{-P}} \qquad \text{Eq.6}$$

Where 0<b<1.4 and P is an estimate of the number of samples in the pitch period which lies in the interval [16, 160].

F. Energy Minimization

The excitation sequence e(n) is modeled as a sum of a Gaussian codebook sequence $d_k(n)$ and a sequence from an interval of past excitation, that is

$$e(n) = G\, d_k(n) + b\, e(n-p) \qquad \text{Eq.7}$$

The excitation is applied to vocal tract filter response to produce a synthetic speech sequence given by

Let

$$F(z) = \frac{1}{A(z)}$$

$$\hat{S}(n) = e(n) * f(n)$$

$$\hat{S}(n) = G\, d_k(n) * f(n) + be(n-p) * f(n)$$

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

Where the parameters G, k, b and P are selected to minimize the energy of the perceptually weighted error between the speech S(n) and the synthetic speech over small block of time i.e.

$$E(n) = w(n) * (s(n) - \hat{s}(n)) \quad \text{Eq.8}$$

Let

$$I(z) = F(z)W(z) \quad \text{Eq.9}$$

Then the error signal can be written as

$$\begin{aligned} E(n) &= w(n) * s(n) - Gd_k(n) * I(n) - be(n-P) * I(n) \\ &= E_0(n) - GE_1(n,k) - bE_2(n,P) \end{aligned} \quad \text{Eq.10}$$

Where

$$E_0(n) = w(n) * s(n)$$
$$E_1(n,k) = dk(n) * I(n)$$
$$E_2(n,P) = e(n-P) * I(n)$$

Since P can be greater than subframe length of 40 samples, we need to buffer previous samples of e(n) to use at this point. To simplify the optimization process, the minimization of the energy of error is performed in two steps. First, b and P are determined to minimize the error energy.

$$Y_2(P,b) = \sum_n [E_0(n) - bE_2(n,P)]^2 \quad \text{Eq.11}$$

Thus, for a given value to P, the optimum value of b is given by differentiating the equation with respect to b and equating with zero.

$$\hat{b}(P) = \frac{\sum_n E_0(n)E_2(n,P)}{\sum_n E_2^2(n,P)} \quad \text{Eq.12}$$

Which can be substituted for b in the equation for Y_2(P, b) that is

$$Y_2(P,\hat{b}) = \sum_n E_0^2(n) - \frac{[\sum_n E_0(n)E_2(n,P)]^2}{\sum_n E_2^2(n,P)} \quad \text{Eq.13}$$

Hence the value of P minimizes Y_2(P) or, equivalently, maximizes the second term in the above equation. The optimization of P is performed by exhaustive search, which could be restricted to a small range around the initial value obtained from the LP analysis.

Once these two parameters are determined, the optimum choices of gain G and codebook index k are made based on the minimization of the error energy between

$$E_3(n) = E_0(n) - \hat{b}E_2(n,\hat{P}) \text{ and } GE_1(n,k). \quad \text{Eq.14}$$

Thus P and k are chosen by an exhaustive search of the Gaussian codebook to minimize

$$Y_1(k,G) = \sum_n [E_3(n) - GE_1(n,k)]^2 \quad \text{Eq.15}$$

Which is solved in a similar manner as above. As the output of the filters because of the memory hangover (i.e. the output as a result of the initial filter state, with zero input) of previous intervals, must be incorporated into the estimation process. Hence we need to store final conditions of the filters, the previous values of b and e(n) to be used in the later frames.

IV. Results

The quality of a synthesized speech is determined by observing how a synthesized signal is approximated according to the original signal. This approximation mainly depends on how the synthesized signal copies the envelope or the pattern of the original signal. The more is the replication, the better is the quality. Therefore, as observed from the graphs, the quality of a speech signal is well maintained in CELP, since it has better envelope replication of the original signal.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

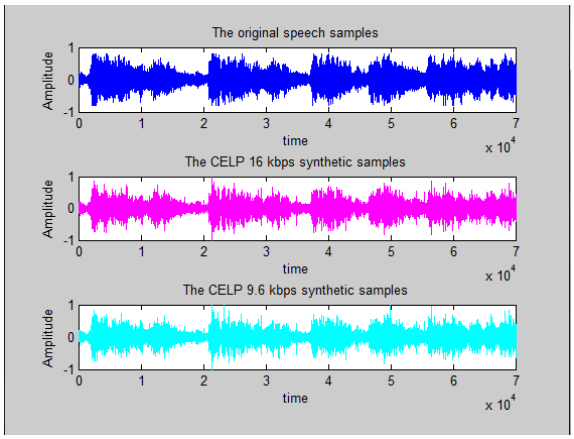

Fig.3: Comparison of original with CELP coders

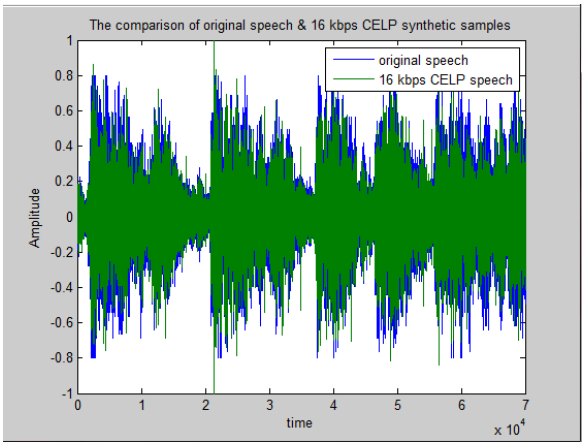

Fig.4: Original speech and 16kbps CELP synthesized speech

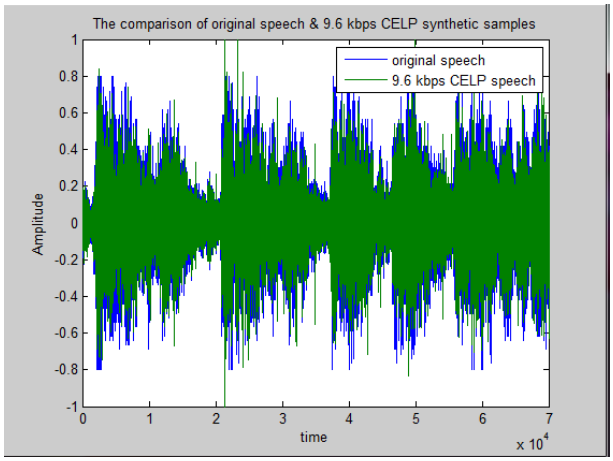

Fig.5: Original speech and 9.6kbps CELP synthesized speech

TABLE I
PARAMETERS USED IN ANALYSIS

Sr. No.	Parameter Name	Value
1	Frame length (N)	160
2	Sub frame length (L)	40
3	Order of LP Analysis (M)	12
4	Constant parameter for perceptual weighted filter (c)	0.85
5	Estimate of number of samples in the pitch period (Pidx)	[16, 160]

TABLE II
BIT ALLOCATION FOR 16 KBPS CELP

Parameter	Bits/ Parameter	Bits/frame
Codebook index, k	10	40
12 LPC coefficients	12	144
Gain	13	52
Pitch filter coefficient, b	13	52
Lag of pitch filter, P	8	32

Length of bit rate frame after quantization $\sum 320$

TABLE III
BIT ALLOCATION FOR 9.6 KBPS CELP

Parameter	Bits/ Parameter	Bits/frame
Codebook index, k	10	40
12 LPC coefficients	6	60
Gain	7	28
Pitch filter coefficient, b	8	32
Lag of pitch filter, P	8	32

Length of bit rate frame after quantization $\sum 192$

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 3, Issue 1, July 2014)

In our case, we have used a simplest variable bit rate Vocoder having a codebook containing 1024 sequences of length 40 and operated in two modes:

- High bit rate (16 Kbps) CELP.
- Low bit rate (9.6 Kbps) CELP.

Tables 2 and 3, show bit allocation for the specific bit rate.

V. Conclusions

The CELP coder exploits the fact that after removing the short and long term prediction from the speech signal, the residual signal has little correlation with itself. It also gives an approach to reduce the number of bits per sample. As CELP can preserve some phase information from the original signal, so it is capable of replicating the original envelope more precisely. Hence, for speech synthesis purposes, CELP is undeniably of best use.

Acknowledgement

I sincerely thank Prof. S.K Jagtap, Assistant professor, Smt. Kashibai Navale College of Engineering, Pune for her valuable guidance for all my study endeavors.

REFERENCES

[1] Kamboh, A., Lawrence, K., Thomas, A., Tsai, P. 2005 Design of a CELP coder and analysis of various quantization techniques.

[2] Devalapalli, S., Rangarajan, R., Venkatramanan, R. Design of a CELP coder and study of complexity vs quality trade-offs for different codebooks.

[3] Saha, N. K., Sarkar, R. N., Rahman, M. 2011 Comparison of musical pitch analysis between LPC and CELP.

[4] Prokopov, V., Chyrkov, O. 2011 Eavesdropping on encrypted VoIP conversations: phrase spotting attack and defense approaches.

[5] Kabal, P. 2011 The equivalence of ADPCM and CELP coding.

[6] Kabal, P. 2009 ITU-T G.723.1 Speech coder- MATLAB Implementation.

[7] Dutoit, T., Moreau, N., Kroon, P. Speech processed in a cell phone conversation.