ICSV14
Cairns • Australia
9-12 July, 2007

14th International Congress on Sound & Vibration

APPLICATION OF CEPSTRUM ANALYSIS IN SPEECH CODING

Vahid Abolghasemi, Hossein Marvi

Faculty of Electrical & Robotic Engineering, Shahrood University of Technology
vahidabolghasemi@yahoo.com, marvi_hossein@yahoo.co.uk

Abstract

A simple Code Excited Linear Prediction Coder based on Cepstral-Analysis to decrease the bit-rate of coder has been proposed. The proposed method first applies Cepstral analysis to the input speech signal and then attempts to remove insignificant coefficients of the cepstrum. These modified coefficients rather than the input speech samples, are fed to our simple CELP coder to be encoded. At the decoder side, we compute inverse cepstrum of the decoded coefficients and obtain the reconstructed speech signal. Our experimental results approve the reduction of bit-rate.

1. INTRODUCTION

Speech coding is the compression of speech (into a code) for transmission with speech codecs that use audio signal processing and speech processing techniques. In addition, most speech applications require low coding delay, as long coding delays interfere with speech interaction [1].Speech coding algorithms can be classified into waveform coders, vocoders and hybrid coders.
In a waveform coder, in the encoder, a reduction of the signal dynamics can be achieved by a fixed or adaptive quantization. Better results are obtained if a (fixed or adaptive) prediction filtering, according to the correlation properties of the signal, is employed. Under certain presumptions, a prediction gain can be used to reduce the bit rate, if the prediction error (residual) signal is quantized instead of the original signal. The prediction filter parameters may be adapted using the reconstructed signal [1].
In vocoders, (or parametric coders) not the signal samples but the parameters of a source filter speech model are quantized and transmitted. This source-filter synthesis representation closely follows the model of speech production.

The time-varying synthesis filter corresponds to the vocal tract and may include a model of the acoustic tube and the lip radiation. In approximation, an all-pole model can be used. The usage of this filter corresponds to the principle of Linear Predictive Coding (LPC).

Pure vocoders are particularly used for low bit rate application (below 0.5 bits per sample).

The third type of speech (hybrid) coder is the intermediate class between waveform coder and vocoder. This type of coders is working for medium bit rates (0.5 ... 2 bits per sample) with relatively high quality. Similar to a parametric coder, it relies on a speech production model

during encoding. Additional parameters of the model are optimized in such a way that the decoded speech is as similar as possible to the original waveform.
The majority of modern hybrid speech coders is based on the principle of linear-predictive analysis-by-synthesis coding also known as CELP (Code-Excited Linear Prediction). [1][2].
In this paper we propose a novel approach to CELP design. The proposed method attempts to apply cepstral analysis to a simple CELP coder, to reduce bit rate. For this purpose the input speech signal first is divided into frames and then the cepstral coefficients of every frame is computed. After that, these coefficients are fed to the CELP coder, instead of the speech frames. Before feeding the frames to the CELP coder, the frame length has been decreased by removing some samples (cepstrum coeffs.) to reduce the bit rate. This idea is based on the fact that the most important information of the vocal tract is laid in low qu-frequencies and pitch information in high qu-frequencies, and the middle region of qu-frequency domain does not involve much speech information [3].
This paper is organized as follows. After introduction, the CELP concept and also the basic parts of a simple CELP coder is discussed in section 2. Section 3 describes our proposed method using Cepstral-Analysis. Experimental results are given in section 4 followed by a conclusion in section 5.

2. THE CELP (CODE EXCITED LINEAR PREDICTION) CONCEPT

CELP coding can be introduced by considering long-term and short-term linear prediction models. Figure.1 demonstrates a block diagram of the speech production model including an excitation codebook. As can be found from this figure, the extracted excitation is multiplied with a suitable gain at first. Then this scaled signal is passed through the cascade connection of pitch and formant synthesis filters to yield the synthetic speech. The advantage of using this pitch synthesis filter is that the periodicity of the speech signal is produced associated with the fundamental pitch frequency. Also the spectral envelope is generated using the formant synthesis filter [2].

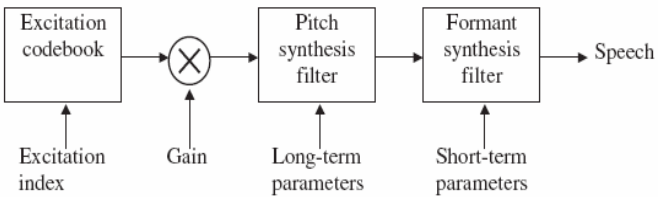


Figure1: The CELP Model of speech production

CELP coders have been classified into the analysis-by-synthesis category. Figure 2 shows a fundamental CELP diagram. As can be seen form the figure, a closed-loop search routine selects the excitation signal and feed it to the synthesis filters. The synthesized signal is then compared with the original speech frame followed by a distortion measurement. This process is repeated for all excitation codevectors stored in a codebook [1]. After this procedure the index of the best excitation sequence is sent to the decoder. Using this index, the excitation codevectors stored in a codebook is reproduced at the decoder. Figure 3 shows more details about the block diagram of the CELP. In the following these details are described briefly.

2.1 Linear Prediction (LP) model

The vocal-tract can be modelled as an all pole filter using linear prediction analysis. This filter generates the spectral envelope of the speech signal. The filter coefficients can be obtained by implementing a lattice filter acting both as a forward and backward error prediction filter [4].

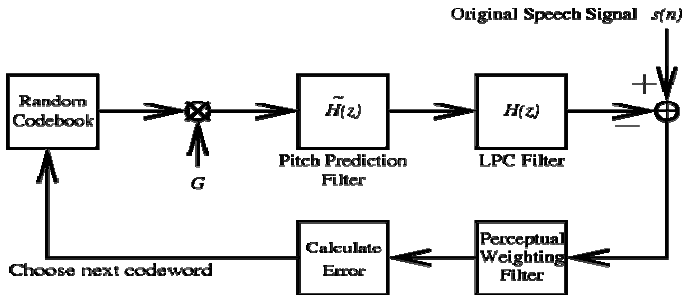


Figure 2: Basic CELP scheme, minimize error by selecting best codebook entry.

Eq. 1 shows the time domain equation that relates the previous samples to the current one. Take the advantage of Eq. 1, we can define frequency domain equation for the filter.

$$\hat{y}(n) = \sum_{i=1}^{p} a_i y(n-i) \tag{1}$$

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

$$H(z) = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \tag{2}$$

2.2 Perceptual Weighting Filter

The designed Linear Prediction filters output the synthetic speech frames. In order to evaluate the difference between the original and the synthesized speech as an error criterion, these two signals are subtracted [2]. The error sequence is passed through a perceptual error weighting filter with system function

$$w(n) = \frac{A(z)}{A(z/c)} = c^M \frac{(p_0 - z)(p_1 - z)...(p_{M-1} - z)}{(cp_0 - z)(cp_1 - z)...(cp_{M-1} - z)} \tag{3}$$

Where c is a parameter in the range $0 < c < 1$ that is used to control the noise spectrum weighting. The coefficients of the filter $A(z/c)$ are $a_i c^i$ which can be seen from

$$A(z/c) = 1 - a_1 (z/c)^{-1} - ... - a_M (z/c)^{-M} = 1 - (a_1 c) z^{-1} - ... - (a_M c^M) z^{-M} \tag{4}$$

2.3 Excitation sequence

The codebook is made up of vectors whose components are consecutive excitation samples. Each vector contains the same number of excitation samples as there are speech samples in a frame. The codebook is known to the encoder as well as the decoder [5][6].

The signal e(n) used to excite the LP synthesis filter 1/A(z) is determined every several milliseconds within the frame under analysis. An excitation sequence $d_k(n)$ is selected from a codebook of stored sequenced, where k is the index.

Excitation codebook search is the most computationally intensive part of CELP coding. Throughout the years, many ideas have been proposed to reduce this complexity issue [2].
In general, different types of codebooks can be found. The most widely used are adaptive codebooks in conjunction with the fixed codebook [2]. In our experiment the fixed codebook, which is a collection of Gaussian signals, is used.

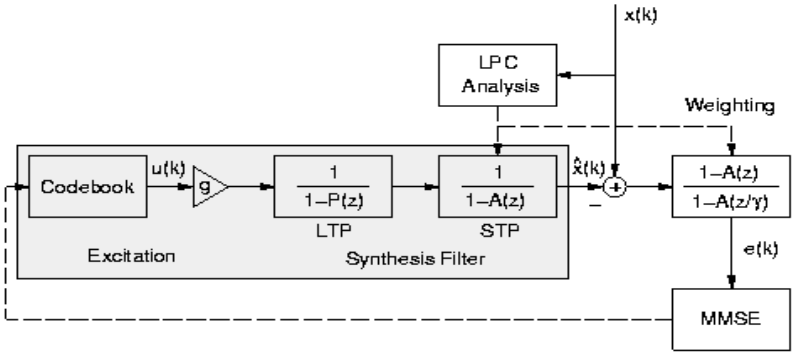


Figure 3: Block diagram of CELP encoder

2.4 The Pitch Synthesis Filter

As for voiced speech frames, the excitation sequence demonstrates a significant correlation from one pitch period to the next, a long-delay correlation filter is used to generate the pitch periodicity in voiced speech [2][3]. Therefore the filter with system function

$$J(z) = \frac{1}{1 - bz^{-P}} \tag{5}$$

describing the effect of the long-term predictor in synthesis, is known as the long term synthesis filter or pitch synthesis filter.

2.5 Error Minimization

As mentioned above, 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

$$e(n) = Gd_k(n) + be(n - P) \tag{6}$$

The excitation is applied to the vocal-tract response 1/A(z) to produce a synthetic speech sequence given by
Let,

$$F(z) = \frac{1}{A(z)}, \quad \hat{s} = e(n) * f(n) = Gd_k(n) * f(n) + be(n - P) * f(n) \tag{7}$$

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.

$$E(n) = w(n) * (s(n) - \hat{s}(n)) \tag{8}$$

Let

$$I(z) = F(z)W(z) \tag{9}$$

Then the error signal can be written as

$$E(n) = w(n) * s(n) - Gd_k(n) * I(n) - be(n - P) * I(n) = E_0(n) - GE_1(n,k) - bE_2(n,P) \tag{10}$$

Where

$$E_0(n) = w(n) * s(n) \tag{11}$$

$$E_1(n,k) = dk(n) * I(n) \tag{12}$$

$$E_2(n,P) = e(n-P) * I(n) \tag{13}$$

Note here that since P can be greater than sub frame length, we need to buffer previous values 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

$$Y_2(P,b) = \sum_n [E_0(n) - bE_2(n,P)]^2 \tag{14}$$

Thus, for a given value to P, the optimum value of b is given by differentiating the equation with respect to b and putting equal to zero. We get the result

$$\hat{b}(P) = \frac{\sum_n E_0(n)E_2(n,P)}{\sum_n E_2^2(n,P)} \tag{15}$$

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

$$Y_2(P,\hat{b}) = \sum_n E_0^2(n) - \frac{\left[\sum_n E_0(n)E_2(n,P)\right]^2}{\sum E_2^2(n,P)} \tag{16}$$

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

$$E_3(n) = E_0(n) - \hat{b}E_2(n,\hat{P}) \; and \; GE_1(n,k) \tag{17}$$

Thus P and k are chosen by a complete search of the Gaussian codebook to minimize

$$Y_1(k,G) = \sum_n \left[E_3(n) - GE_1(n,k)\right]^2 \tag{18}$$

which is solved in a similar manner as above. Note, that 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. And so we need to store final conditions of the filters, the previous values of b and e(n) to be used in the later frames.

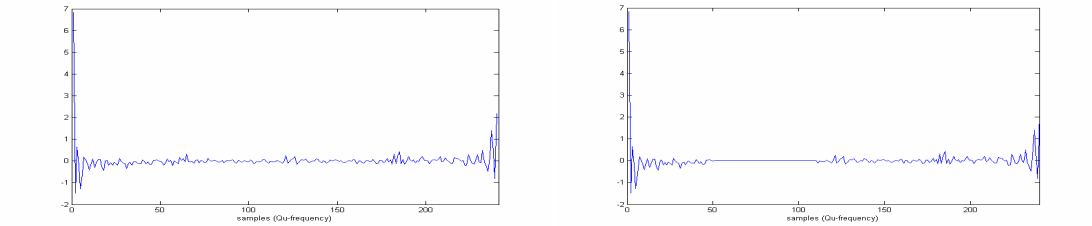
2.6 Quantization

In our experiment for quantization we use uniform scalar quantization for the four parameters i.e. Gain, Pitch Delay, Pitch Filter coefficients, Linear Predict coefficients and Excitation Sequence Index [1][2].

3. PROPOSED METHOD BASED ON CEPSTRAL-ANALYSIS

Homomorphic analysis is based on the cepstrum, which is the inverse Fourier transformation of the logarithm of the speech spectrum magnitude [7].

By considering the speech spectrum it can be realized that a speech segment is consist of two major parts. One is a smooth varying portion which corresponds to the vocal tract, and the second part is a rapidly varying fine structure which is related to the periodic excitation of the speech signal [3].
If we represent the spectrum as the product of the vocal tract envelope and the pitch harmonics, then the cepstrum is computed as the inverse Fourier transform of the logarithm of the magnitude. Since the vocal tract envelope varies smoothly, it contains low frequency components, while the fine structure varies more rapidly and contains high frequency components [7]. Of course after transformation, the low and high frequency components correspond to low time (qu-frequency) and high time (qu-frequency) as shown in Figure 4 (a) Note that the periodic pitch component is transformed to a high time (qu-frequency) peak [3].
From Figure 4 (a) it can be concluded that Cepstrum, compresses and separates the vocal tract information in low qu-frequencies and the pitch information in high qu-frequencies, so there is a gap between these two parts which doesn't include significant information. Of course these components which are laid in medium qu-frequencies correspond to unvoiced information which in comparison with vocal tract and pitch components, have low amplitude. Thus, it is expected that by removing a reasonable number of these components the total bit-rate will decrease, which is a promotion in speech coding. It is also clear that the quality of reconstructed waveform will decrease, too.
Using this characteristic to reduce the frame length, in our proposed method, after computing the cepstral coefficients of the input speech frames (Figure 4 (a)), we decrease the frame length by removing the coefficients involved the gap (Figure 4(b)). Then we feed the retained coefficients, instead of the speech envelopes, to the CELP coder which described in the previous section. Figure 5 shows block diagram of this modified CELP which we call it "CEPS-CELP".



(a) (b)

Figure 4: (a)Real Cepstrum for one frame of the input speech signal. (b) Real Cepstrum for one frame of the input speech signal after removing insignificant samples

We can select a wider gap to remove more coefficients, but as mentioned before, the more coefficients have been removed; the poorer quality of reconstructed speech is achieved. So there is always a tradeoff between the quality of decoded speech and number of rejected coefficients.
At the decoder side, after synthesis of these coefficients from the codebook, we compute inverse-cepstrum for every frame and obtain the reconstructed speech signal. At the decoder side, it is observed that the bit-rate reduced, but a bit degradation of speech quality also is produced.
By applying log-cepstrum for every frame we have:

$$c(n) = \mathrm{Re}\left(F^{-1}\left\{\log\left|F\{s(n)\}\right|\right\}\right) \tag{19}$$

Where s(n) is the input speech signal and F is the Fourier operator.
At the decoder, the reconstructed speech signal is obtained by applying

$$s_r = \mathrm{Re}\left(F^{-1}\{\exp(F\{c_r(n)\}\}\right) \quad (20)$$

Where $c_r(n)$ is the reconstructed Cepstral coefficients for every frame.

4. EXPERIMENTAL RESULTS

Experiments have been conducted to evaluate and approve the proposed method.
In our implementation vocal-tract filter has L = 15 coefficients and the fidelity criterion is MSE
The speech signal is sampled at 8 kHz, the frame size is 30ms (240 samples), the block duration for the excitation sequence selection is 5 ms (40 samples). Furthermore, the codebook has 1024 sequences which require 10 bit to send the index k. and the lag of the pitch filter, P, is searched in the range 16 to 160 (equivalent to 50Hz to 500Hz) which require 8 bit to represent. It should be noted that using uniform quantization at least 4 bits are required to represent every sample.
In CEPS-CELP Since we have made zero coefficients located in the interval (50 110) for 240 frame length, non-zero samples per frame would be 160. These zeroed samples are not sent

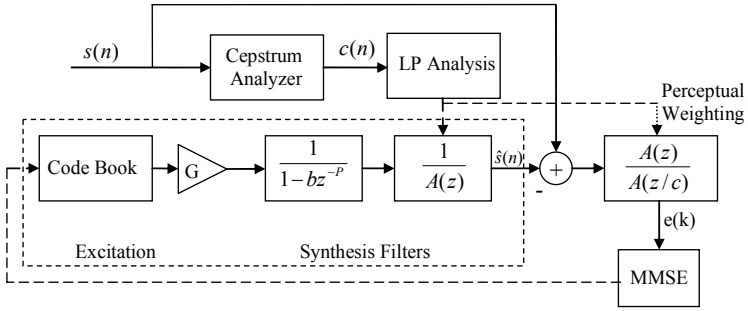


Figure 5: Block diagram of the proposed coder: CEPS-CELP

toward the decoder. Therefore just 160 non-zero samples per frame are sent. At the decoder with the knowledge about the number and location of zeroed samples, we rearrange them into each decoded frame. It leads to resize the frame length, up to 240. In fact this procedure is a lossy compression.
Thus the bit allocation for above parameters changes as follows:
-Vocal tract filter L = 10 coefficients
- Pitch filter coeff. = 5
- Gain = 5
- Pitch Delay = 8

MOS scale	Speech quality
1	Bad
2	Poor
3	Fair
4	good
5	Excellent

Table 1: The MOS scale

We applied both Persian and English speeches of male and female to the proposed coder and found that, quality of the reconstructed speech signal for English speakers is better, compared to Persian speakers.
Another experiment has been done using MOS standard test. MOS (Mean Opinion Score) scale has been shown in Table.1. The utterances which are used in this experiment are "Exercise one", "conversation", "listen and practice", which has been obtained from twelve

listeners The results has been summarized in Table.2. From Table 2, it can be realized that the quality of reconstructed speech signal in CELP is comparable with that of in CEPS-CELP which confirms the capability of the proposed method.
A comparison of bit rate between CELP and CEPS-CELP is shown in Table 3. As can be seen from this table, the bit rate is considerably reduced in CEPS-CELP. But the MOS test is decreased too.

	CELP	CEPS-CELP
Average MOS Score	4.83	4.42

Table 2: MOS test for utterance "Exercise one", "Conversation", "Listen and practice".

	Symbol	Bit allocation CELP	Bit allocation CEPS-CELP
Codebook index	k	10x4	10x4
Pitch delay	P	12x4	8x4
Pitch filter coeffs	b	7x4	5x4
Gain	G	7x4	5x4
LP coeffs	a	15x5	10x4
Total bit rate		7.3 kbps	5.1 kbps

Table 3: Comparison of two type of CELP

5. CONCLUSION

In this paper, the CELP concept has been presented. A basic CELP coder has been designed and implemented in MATLAB. Then, an improved approach to reduce bit rate based on cepstral analysis proposed.
We compared the performance of both basic CELP and our coder "CEPS-CELP". Experimental results show outperformance of suggested method to original CELP.

REFERENCES

[1] Andreas S. Spanias. “Speech Coding: A Tutorial Review, ” *Proceeding of the IEEE* Vol. 82, No. 10, pp. 1541-1582, October 1994
[2] WAI C. CHU. *Speech Coding Algorithms Foundation and Evolution of Standardized Coder*. John-Wiley, , pp. 299-325, 2003
[3] Panos E.Papamichalis, *Practical Approaches To Speech Coding*, , Prentice-Hall, , pp. 92-174, 1987
[4] Lawrence R.Rabiner and Ronald W.Schafer, *Digital Processing Of Speech Signals*. Prentice-Hall, Inc. pp 355-388, 1978
[5] K. Koishida, K. Tokuda, T. Kobayashi and S. Imai, “CELP coding system based on mel-generalized cepstral analysis,” *Proc. ICSLP’96*, pp.318–321, 1996.
[6] K. Koishida, G. Hirabayashi, K. Tokuda and T. Kobayashi, “A wideband CELP speech coder at 16 kbit/s based on melgeneralized cepstral analysis,” *Proc. ICASSP’98,* pp.157–160, 1998.
[7] CLIFFORD J. WEINSTEIN**,** ALAN V. OPPENHEIM, “Predictive Coding in a Homomorphic Vocoder, ” *IEEE Trans. Audio And Electro acoustics.* Vol. AU-19, NO. 3, 1971