Modified CELP Coder Using Root Cepstrum

Vahid Abolghasemi and Hossein Marvi
Shahrood University of Technology, Shahrood, Iran
vahidabolghasemi@yahoo.com, marvi_hossein@yahoo.co.uk

ABSTRACT

Due to increasing demand for speech communications, efficient techniques in low-rate speech coding are of interest. In this paper a new compression technique using root cepstral analysis has been proposed. Implementing the proposed method causes the coder to deal with root cepstrum coefficients instead of speech samples. The main idea in using root cepstrum analysis is that some of the trivial coefficients can be ignored to send toward the decoder. Although it leaves a little degradation in the quality of decoded speech signal, considerable reduction in the total bit-rate is achieved. Moreover it has the advantage of adjustability which can be used to optimize the coding procedure. The experimental results confirm the ability of the proposed method in speech coding problems.

1. INTRODUCTION

In general, speech coders aim to represent a digitized speech signal using as few bits as possible, maintaining simultaneously a reasonable level of speech quality.
The main goal of speech coding is either to maximize the perceived quality at a particular bit-rate, or to minimize the bit-rate for a particular perceptual quality [1].

There are mainly three types of coders classifying by coding techniques: waveform coders, parametric coders and hybrid coders.

In waveform coders a procedure to preserve the original model of the signal waveform has been proposed, hence the resultant coders can be applied to any signal source. These coders are better suited for high bit-rate coding, since performance lessens with decreasing bit-rate. In practice, these coders work best at a bit-rate of 32 kbps and higher [1].

In parametric coders, the speech signal is assumed to be generated from a model, which is controlled by some parameters. During encoding, parameters of the model are estimated from the input speech signal, with the parameters transmitted as the encoded bit-stream. Beside several proposed models in this class, the most successful one is based on linear prediction, such as linear prediction coders (LPC) and mixed excitation linear prediction coders (MELP) [2][1]. This class of coders works well for low bit-rates, typically in the range of 2 to 5 kbps.

The third type of speech coders is hybrid coders. As its name implies, a hybrid coder combines the strength of a waveform coder with that of a parametric coder. 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 closeness criterion often is measured by a perceptually weighted error signal. Similarity between the original signal and the decoded signal in time domain, such as in waveform coders, is desired. This class belongs to the medium bit-rate coders, which works well in the range of 5-15 kbps. Code-excited linear prediction (CELP) is one of the conspicuous of this kind [2][1].

Despite the fact cepstrum has been widely used in speech recognition, this paper intend to apply cepstrum in speech coding problems; the proposed algorithm aims to put cepstrum in a CELP coder to obtain a low-rate coder. The main purpose of the paper is to demonstrate and also compare the strength of Root rather than log cepstrum in low-rate coders. As description on the proposed method, a pre-processing step i.e. framing, windowing, etc, is applied to the input speech at first. Then the cepstral coefficients for each frame are computed. In the next step, in order to reduce the bit rate for transmitting the samples (cepstral coefficients), the frame length has to be decreased. We can do this by zeroing out some samples (cepstrum coeffs.) of each frame. The main idea of applying root cepstrum analysis and then zero out trivial coefficients is: cepstral analysis separates the vocal tract and pitch information. Data related to the vocal tract is located in low qu-frequencies, and the pitch information in upper part of qu-frequency axis. Hence the samples located in the central zone of qu-frequency domain carry insignificant speech information. This can be found as a suitable feature for reducing the bit-rates in speech coders [3].

The rest of the paper is as follows. Section 2 presents a brief overview of CELP concept. In section 3 we describe Root-Cepstral analysis. Modified CELP based on root cepstrum is proposed in section 4. Experimental results are given in section 5. Finally the paper is concluded in section 6.

2. MAIN POINT OF CELP CODERS

Description of a CELP coder can be done with regard to the long-term and short-term linear prediction models. Figure.1 shows the block diagram of the speech production model, where an excitation sequence is extracted from the codebook through an index. As can be seen from this figure, the extracted excitation is scaled to the appropriate level and filtered by the cascade connection of pitch synthesis filter and formant synthesis filter to yield the synthetic speech. The pitch filter creates periodicity in the signal associated with the fundamental pitch frequency, and the formant filter generates the spectral envelope [2].
CELP is an analysis-by-synthesis method. Figure 2 shows a basic CELP scheme, where the excitation signal is selected by a closed-loop search procedure and applied to the synthesis filters. The synthesized waveform is compared to the original speech segment, the distortion is measured, and the process is repeated for all excitation codevectors stored in a codebook [1]. The index of the best excitation sequence is transmitted to the decoder, which retrieves the excitation codevectors from a codebook identical to that at the encoder.

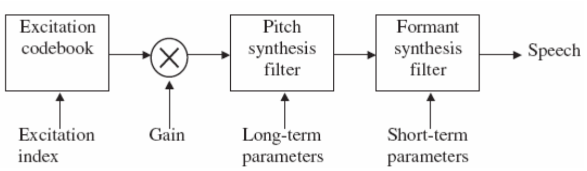

Figure1- The CELP Model of speech production

In a brief expression we can summarize the CELP algorithm based on four main ideas: Using the source-filter model of speech production through linear prediction (LP); Using an adaptive or a fixed codebook as the input (excitation) of the LP model; Performing a search in closed-loop in a perceptually weighted domain; Applying a type of quantization specially vector quantization (VQ). Figure 5 demonstrates more details about the block diagram of a simple CELP.

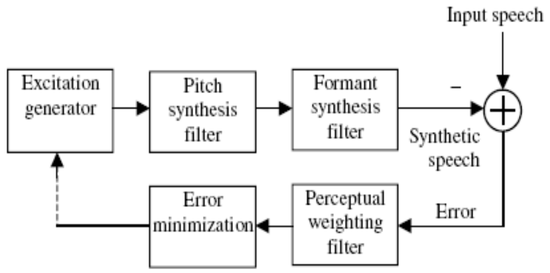

Figure 2 - Analysis-by-synthesis loop of a CELP encoder

3. ROOT CEPSTRUM ANALYSIS

In general a speech frame can be modeled as a convolution of a vocal tract filter and excitation sequence, consisting of periodic pulses for voiced speech and noise for unvoiced speech. Transfer function of the vocal tract can be modeled by an all pole filter [4]. In order to analyze the excitation and vocal tract information independently we may use an operator to map the convolution onto a linear domain and separate the vocal tract and excitation information. This is done by using a homomorphic transformation such as a “Log” operator.

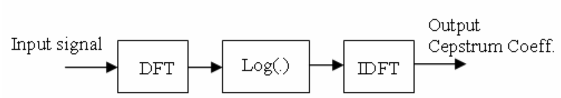

Figure 3- Practical implementation of system for obtaining the log cepstrum

Figure 3 shows a scheme to obtain separated information using log operator. It should be noted that the log cepstrum itself is obtained in two ways. The real cepstrum, which is only the Log magnitude and complex cepstrum, which is achieved by considering both real and imaginary parts of the Fourier Transform coefficients [4]. The log-cepstrum representation of speech signal is attractive in speech processing; however it presents one major problem. Since, as x tends to zero, log(x) tends to mines infinity, the function is very sensitive to small values of x. In the cepstrum this means that there is most sensitivity to those parts with lower power, i.e. to those parts where the SNR is normally worse. One well technique for dealing with this problem is to replace log(x) function with root function x^{γ}, where $-1<\gamma<1$. Lim [5] approximated the logarithmic deconvolution scheme with a root function, where $(.)^{\gamma}$ and $(.)^{1/\gamma}$ are used instead of log and exponential operations. Later, the Log-cepstrum is covered as a special case under generalized Root-cepstrum [4].

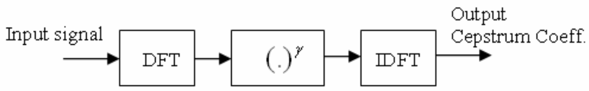

Figure 4 - block diagram of a Root-Cepstrum Analyzer

As stated earlier, it is known that the Log-cepstral analysis is sensitive to noise [4]. On the other hand, Root-cepstrum is more immune to noise, exhibiting small deviations from the clean cepstrum. The average deviation of noisy Log-cepstra from clean Log-Cepstra is 34.2%, whereas noisy Root-cepstra deviation from clean Root-cepstra is only 23.5% [4]. Figure 4 shows the block diagram for obtaining the RCCs (Root Cepstrum Coefficients).

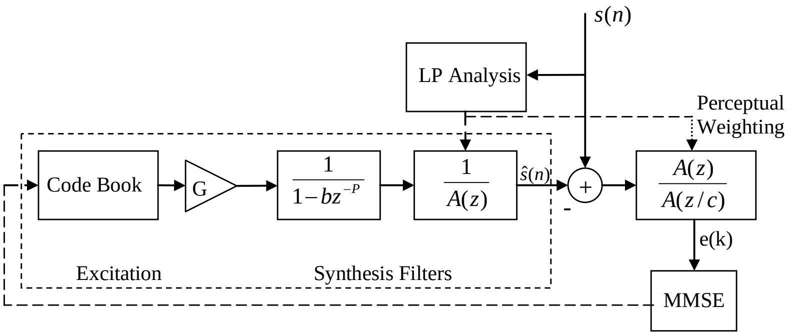

Figure 5- Block diagram of CELP encoder

4. MODIFIED CELP CODER

Root cepstrum analysis has been widely used in speech recognition [4][5] and has been achieved very promising results. Moreover the two-dimensional and also modified two-dimensional root cepstrum is used in speech and speaker recognition problems [6] [7], while applying in speech coding is less considered yet. The proposed modified coder takes the advantage of Root cepstrum in speech coding.

One can be realized at a first looking at the speech spectrum is that a speech segment is consist of two major parts. A smooth varying portion issued from the vocal tract model, denoted by dashed curve in the Figure 6(a), and a rapidly varying fine structure which is related to the periodic excitation of the speech signal [3]. This has been shown clearly in Figure 6(a).

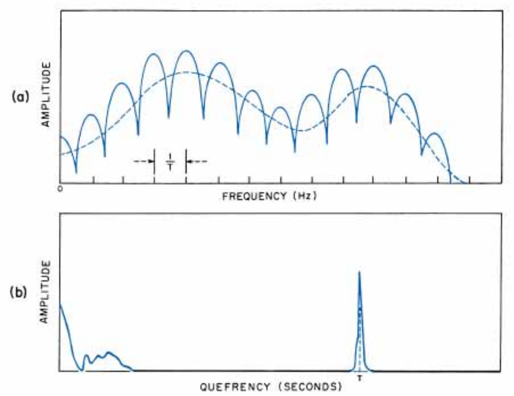

Figure 6- (a) Spectrum of a speech frame (b) the computed cepstrum

Modeling a speech frame as a convolution of a vocal tract filter and excitation sequence will result in product of the vocal tract envelope and the pitch harmonics in the frequency domain. The cepstrum which is computed as the inverse Fourier transform of the root (or logarithm) of the magnitude are in the time domain but they are differ from the input speech samples. Since the vocal tract envelope varies smoothly, it contains low frequency components, while the fine structure varies more rapidly and contains high frequency components. Of course after transformation, the low and high frequency components correspond to low time (qu-frequency) and high time (qu-frequency). Figure 6(b) shows these two properties. Note that the periodic pitch component is transformed to a high time (qu-frequency) peak [3].

From Figure 6(b) it can be seen 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 segments which doesn't include significant information. These components which are laid in medium qu-frequencies correspond to unvoiced information which in comparison with vocal tract and pitch components, have low (nearly zero) amplitude. Thus, it is expected that by removing a reasonable number of these components the total bit-rate will decrease. This assumption has to be proved through the experiments. It is also clear that the quality of reconstructed waveform will decrease, too.

The most important priority of root rather than log cestrum is raised when considering the quality of reconstructed speech signal. As it has known that root cepstrum is much less sensitive to noise we assumed that the quality of reconstructed speech signal might be better than that when using log operator. On the other hand using Root cepstrum analyzer, the best quality of the reconstructed speech signal can be achieved by adjusting an optimum value for γ. Thus both two operators have been applied in the experiments and the results confirmed our assumption.

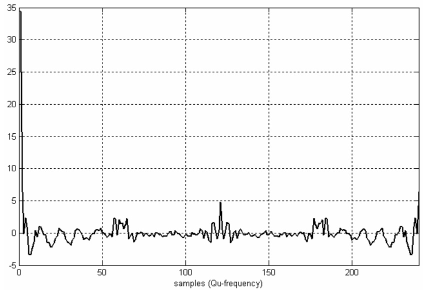

Figure 7- Root Cepstrum for one frame of the input speech signal (selected $\gamma = 0.5$)

Using this characteristic to reduce the frame length, in our proposed method, after computing the cepstral coefficients of the input speech frames (Figure 7), we decrease the frame length by removing the coefficients involved the gap (Figure 8). Then we feed the retained coefficients, instead of the speech samples, to the CELP coder. Figure 9 shows block diagram of this modified CELP coder. The pre-processing block set contains the framing and windowing and other extra required operation.

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 obtained. So there is always a tradeoff between the quality of decoded speech and number of rejected coefficients.

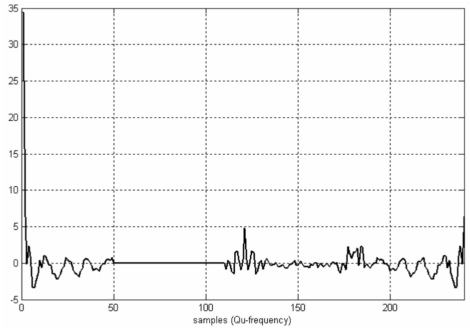

Figure 8- Speech frame (figure 7) after removing insignificant samples

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 root-cepstrum for every frame we have:

$$R(n) = \mathrm{Re}\left(F^{-1}\left\{\left(F\left\{s(n)\right\}\right)^{\gamma}\right\}\right) \qquad (1)$$

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

$$s_r = \mathrm{Re}\left(F^{-1}\left\{\left(F\left\{R(n)\right\}\right)^{1/\gamma}\right\}\right) \qquad (2)$$

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

In the equation 1 and 2 the parameter γ can inherit different values in the range $-1<\gamma<1$. Furthermore the value of γ can be adjusted to enhance performance.

5. EXPERIMENTAL RESULTS

To investigate the effect of the proposed method on bit-rate reduction and also give a comparison between Log and Root Cepstrum the proposed method implemented and applied to our database for both English and Persian speakers. Different utterances using both isolated words and continuous sentences have been used and the related results are given.

All the speech signals are 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.

The amount of bit-rate reduction depends on the number of zeroed samples of Cepstrum coefficients. After conducting many experiment on different utterances it has been empirically derived that zeroing out more than 60 samples (for the frame length of 240) will result a defective quality of reconstructed speech. We ignored sending the coefficients located in the interval (50, 110), so the non-zero samples per frame would be 160. These zeroed samples are not sent toward the decoder. Hence the total bit-rate decreases.

At the decoder with the knowledge about the number and location of zeroed samples, we replace 60 zeroed samples in the appropriate interval to obtain the reconstructed cepstrum coefficients. It leads to resize the frame length, up to 240.

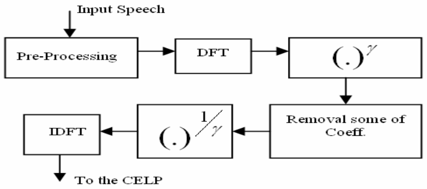

Figure 9- Block diagram of the proposed coder

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.

The bit-rate computed before applying Root cepstrum analyzer is 6.5 Kbps, while it decreases to less than 5 Kbps when using the proposed method.

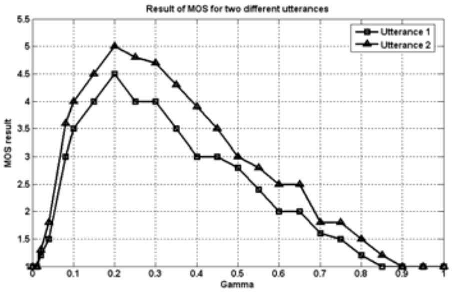

Figure 10 - MOS Average result for different values of γ

Utterance1: is in Persian

Utterance2: mentioned in the text

Vector Quantization (VQ) as a suitable way of optimizing the quantization procedure is very useful in speech coders. A simple CELP according to the VQ instead of scalar quantization achieves a more acceptable bit-rate. We observed that our CELP coder with VQ inside yields 4.2 Kbps of bit-rate.

One of the most effective ways to evaluate the quality of reconstructed speech is the MOS standard test [1]. MOS (Mean Opinion Score) standard test has been shown in Table.1. Table 2 introduces a comparison between three types of implementation i.e. simple CELP, Log-cepstrum and Root cepstrum. As can be seen from the table with selecting the optimum γ, Root Cepstrum has best performance among the others. The utterances which have been used in this experiment are "Hello", "are you ok", "where is your mam", which has been obtained from twelve listeners. The results confirm the strength of the proposed method.

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

Table 1: The MOS scale

Although the quality of decoded speech signal is a bit lower than that of the simple CELP, the bit-rate has been decreased considerably. Moreover the table shows the outperformance of Root operator against the log operator in cepstral analysis.

	Simple CELP	Root cepstrum Optimum γ =0.2	Log cepstrum
Average MOS Score	4.83	4.54	4.33

Table 2: MOS test for utterance "Hello", "Are you ok?", "Where is your mam?".

Another experiment has been conducted to search for the optimum value of γ. The result has been demonstrated graphically in figure 10 we observed that the best quality of decoded speech signal will achieve for γs close to 0.2. Choosing $\gamma = 1$ which is called Pseudo cepstrum doesn't have a reasonable performance. Similarly choosing $\gamma \le 0$ does not have an acceptable result.

Interestingly it has been observed that the MOS result for Persian speeches are lower. The reason may be related to the differences between types of sounds (vowels, consonants, etc) of two languages. More precise reasons about this result have to be explored.

Also we computed the mean squared error for different speaker with different utterances as a criterion to measure the deviation of the output from the original signal. This measurement has been depicted in figure 11.

6. CONCLUSION

The paper stated application of root-cepstrum analysis in speech coding. The aim is to design low-rate speech coders. The proposed idea on influence of root-cepstrum analysis in reducing the total bit-rate has been clarified. Then a CELP coder based on the root cepstrum has been presented. In the experiments the performance of the system has been evaluated and the promising results have been presented. A comparison of the simple CELP and the log-cepstrum CELP with the proposed CELP has been presented, too.

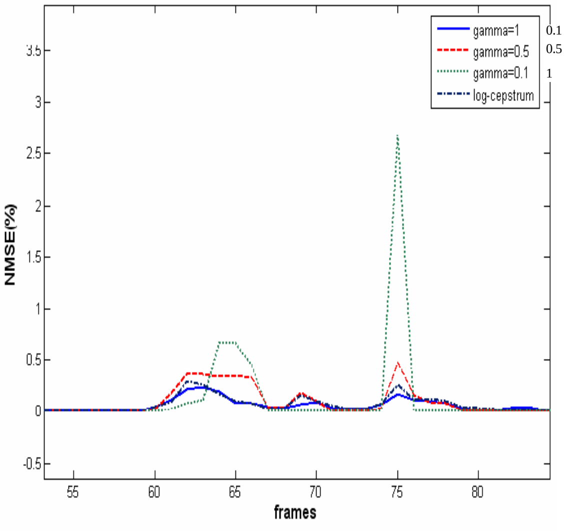

Figure 11- Mean Square Error for 4 different tests

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] Ruhi Sarilaya, John H.L. Hansen. "Analysis of the Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition, " *EUROSPEECH-2001 Aalborg, Denmark, Sept. 3-7, 2001.*

[5] J.S. Lim, "Spectral Root Homomorphic Deconvolution System," *IEEE Trans. ASSP*, vol. 27, no 3, pp. 223-233, 1979.

[6] ARIKI, Y., MIZUTA S., NAGATA, M., and SAKAI, T.: "Spoken word recognition using dynamic feature analysed by two-dimentional cepstrum, " IEEE Proc. I, Commin. Speech Vis., 1989 , 136,pp. 133-140

[7] E. Chilton, H.Marvi. "Two-dimensional root cepstrum as feature extraction method for speech recognition, " *IEE Electronic letters, Speech Processing.* Vol.39 No.10,15th May 2003. pp. 815-81