A 4800 bps CELP Vocoder with an Improved Excitation

Hisham Hassanein, André Brind'Amour and Karen Bryden
Department of Communications
Communications Research Centre
3701 Carling Ave., P.O.Box 11490, Station H
Ottawa, Ont., CANADA. K2H 8S2.
Phone : 613-998-2462
Fax : 613-990-7987

ABSTRACT

The Stochastic or Code-Excited Linear Predictive Coder (CELP) is among the promising candidates for producing good quality speech at low bit rates. However, the speech quality produced suffers from perceived roughness. Many researchers have used pole-zero postfilters to mask the roughness at the output of the synthesis filter. Although the postfilters are effective in masking the noise at low bit rates, they produce spectral distortions. In this paper[1], it is proposed to improve the speech quality by introducing two modifications to the fixed stochastic codebook. In the first modification, the stochastic codebook is used only when the long-term corrrelations are low. Otherwise a pulse-like codebook is selected. In the second modification, the selected codebook output is weighted using an adaptive spectral shaping procedure. These two modifications have been incorporated in a 4800 bps CELP coder and have resulted in a perceptually improved vocoded speech.

1. INTRODUCTION

Code-Excited Linear Predictive Coding (CELP) was first introduced by Atal and Schroeder in 1984 [1]. This algorithm represented a breakthrough for achieving good quality speech at rates below 4800 bps. Its major drawback was its computational complexity, which was prohibitive for real-time applications. Since then, many speech researchers, recognizing the tremendous potential of this algorithm, experimented with different methods for simplifying the algorithm. At the same time, the DSP chips became more powerful and floating point processors were introduced. These activities have resulted in the implementation of CELP (or similar algorithms developed later like Vector Adaptive Predictive Coders (VAPC) [2] or Vector Sum Excited Linear Predictive Coder (VSELP) [3]) using a single DSP chip. In this paper, a brief description of a *reference* CELP coder is described, followed by two modifications proposed for the stochastic codebook. The modified CELP algorithm has been implemented on a single TMS320C25 DSP chip and operates in full-duplex at the rate of 4800 bps.

2. REFERENCE CODER DESCRIPTION

The reference coder's synthesizer is depicted in Fig. 1. The excitation to the synthesizer e(n) given by (1) is the linear combination of a vector from the stochastic codebook x(K+n) and a vector from the adaptive codebook e(n-L)

$$e(n) = Gx(K+n) + \beta e(n-L), \quad n=0,59 \qquad (1)$$

where K and L are the optimal indices of the stochastic and adaptive codebooks respectively, and G and β the gains of the vectors from the respective codebooks. During voiced sounds, the lag coefficient β is close to unity. In this case most of the contribution to the excitation comes from the adaptive codebook (which represents the excitation history) and the stochastic codebook entry acts only as a correcting term. For lower bit rates the stochastic codebook contribution may cause noisy synthetic speech for two reasons. First, the LPC model with a limited number of coefficients (10) fails to entirely remove all the short-term correlations from the speech input signal. The result is an intelligible residual with a non-flat spectrum. This may suggest that the stochastic codebook may be inadequate, particularly for a

[1]This project was sponsered by DCEM/DRDCS, code number 0417U.

small codebook size (60 entries in our implementation) giving rise to rough synthetic speech. The second reason for the synthetic speech roughness may be attributed to the use of the stochastic codebook during voiced frames. In order to mask the noise, the reference coder uses a formant and lag postfilters [7] at the output of the synthesizer, in order to compensate for the two defficiencies listed above. The postfilters were reported to be effective in masking the noise but have the disadvantage of creating spectral distortions, particularly when the vocoders are used in tandem. In order to improve the synthesized speech quality without introducing the inherent spectral distortions produced by the postfilters, two modifications to the stochastic codebook are introduced, namely the use of a pulse-like codebook instead of the stochastic codebook during segments with long-term correlation coefficients larger than a certain threshold, and the spectral weighting of the output of the codebooks. These two modifications are described in the following sections.

3. VOICED/UNVOICED CODEBOOKS

It was observed that the synthetic speech produced by Pitch-Excited LPC coders sounds smoother (although less natural) than the reference coder. This may be attributed to the inadequacy of the stochastic codebook during voiced sounds. In order to smooth the CELP synthetic speech, we tried to imitate the excitation model of the Pitch-Excited coder. Thus, instead of using the stochastic codebook during voiced subframes, a pulse-like codebook is used. The number of pulses in the codebook is a function of the lag (which has been determined earlier in the analysis process) and the subframe size. The number of the codebook indices is chosen to be equal to the subframe size. For lag values larger than the subframe size, each codebook vector contains only one non-zero pulse positioned at a location equal to the codebook index. For lag values less than the subframe size, the first S-L+1 vectors, where S is the subframe size and L is the lag, contain two pulses each (with equal amplitudes) separated by the lag value and the location of the first pulse is equal to the codebook index. The remaining vectors contain only one non-zero pulse. The dual codebook is shown in Fig. 2, where if $|\beta|$ is larger than a certain threshold T, (T>0), the speech is considered voiced, and consequently the pulse-like codebook is chosen. It is to be noted that the threshold is valid for values of $|\beta|$ close to unity, where speech is voiced, as well as large values of $|\beta|$ which indicates voicing onset. Otherwise, the stochastic codebook is used. The output of either codebook is frequency weighted as described in the following section.

4. WEIGHTING OF THE DUAL CODEBOOK

As mentioned earlier, the residual signal exhibits some short-term correlations which have not been successfully removed by LPC analysis. Intuitively, it makes sense to spectrally weight the dual codebook adaptively so that its spectrum will be as close as possible to that of the residual (after removing the long-term correlations). Several researchers have tried to apply spectral shaping to the excitation. In [4], an all-pole spectral shaping was applied to the excitation of a Pitch-Excited Linear Predictive Coder. In [5] a pole-zero weighting filter was used to shape the excitation of a CELP coder. The result was a perceived quality comparable to that obtained with the postfilter without the inherent distortion introduced by the latter. We have chosen a filter similar to [4]. The dual codebook is adaptively weighted every frame by

$$W(z) = \frac{1}{1+F\sum_{i=1}^{10} a_i z^{-i}} \qquad (2)$$

where the a_i's are the LPC filter coefficients and F is a modulus reduction factor given by

$$F = \alpha\prod_{i=1}^{10}(1 - k_i^2), \quad 0<=\alpha<=1 \qquad (3)$$

In (3), the k_i's are the reflection coefficients and α is a scaling factor. When the LPC prediction is efficient as is the case for front vowels, murmurs and nasals, the residual has a relatively flat spectrum and F is small. In this case W(z) acts as an all-pass filter and the dual codebook is minimally weighted. For speech sounds where the LPC prediction is not as good, F is relatively large and W(z) gives the dual codebook a spectral shape similar to that of the residual.

4. EXPERIMENTAL RESULTS

The two modifications described above have been implemented in the TMS320C25 code. The objective measure chosen to quantify the results is the segmental Signal to Weighted Noise ratio (SWNR) defined as

$$SWNR = \frac{\sum_{n=0}^{N-1} s^2(n)}{\sum_{n=0}^{N-1} e_w{}^2(n)} \qquad (4)$$

where s(n) is the original input speech signal, N is the frame size in samples, and $e_w(n)$ is the weighted error signal. The latter is obtained by frequency weighting the difference between the input and synthesized speech signals, by the conventional CELP weighting filter CW(z) defined by

$$CW(z) = \frac{1 + \sum_{i=1}^{10} a_i z^{-i}}{1 + \sum_{i=1}^{10} \gamma^i a_i z^{-i}}, \ \gamma = 0.75 \qquad (5)$$

Informal listening tests using the real-time hardware were performed on a variety of input speech samples. Both of the modifications introduced above produced synthetic speech which may be described as cleaner and fuller than that produced by the reference coder. However, this improvement was not translated into a significant improvement in the average segmental SWNR. Over a limited database of 999 frames, the improvement was only 0.27 dB in the case of the weighted codebook and 0.24 dB in the case of the dual codebook. However, the improvement seems to be localized and went as high as 5 dB in the case of the dual codebook and 7 dB in the case of the weighted codebook. The combined use of the two modifications resulted in increases as high as 10 dB in some frames. In Fig. 3a, 3 seconds of input speech from a male speaker are shown and the segmental SWNR of the corresponding synthesized speech is shown in Fig. 3b. The difference in segmental SWNR between the synthesized speech for each modification and the reference signal are shown in Figs. 3c and 3d. In can be noted from Fig. 3d that the largest increases in SWNR between the coder using the dual codebook and the reference coder tend to occur at the beginning of transitions.

5. CONCLUSIONS

In this paper, two modifications to the stochastic codebook were introduced, namely the adaptive frequency weighting of the stochastic codebook and the use of a dual codebook. This has resulted in an improved speech quality of the vocoder. Although these modifications resulted in an increase of up to 10 dB, the improvements seem to be localized and were not translated into a significant increase in the average segmental SWNR.

REFERENCES

1. Atal B.S. and M.R. Schroeder 1984. Stochastic Coding of Speech at Very Low Bit Rates. *Proc. of ICC, Amsterdam*, pp. 1610-1613.

2. Chen J.H. and A. Gersho 1987. Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, p. 51.3.

3. Gerson I.A. and M.A. Jasiuk 1990. Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 461-464.

4. Kang G.S. and S.S. Everett 1985. Improvement of the Excitation Source in the Narrow-Band Linear Prediction Vocoder. *IEEE Trans. Acoust., Speech and Signal Processing*, ASSP-33(2).

5. Kroon P. and B.S. Atal 1988. Strategies for Improving the Performance of CELP Coders at Low Bit Rates. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 151-154.

6. Kroon P. and E.F. Deprettere Feb 1988. A Class of Analysis-by-Synthesis Predictive Coders or High Quality Speech Coding at Rates between 4.8 and 16 kb/s. *IEEE J. on Selected Areas in Communications*, SAC-5.

7. Kroon P. and B.S. Atal 1987. Quantization Procedures for the Excitation in CELP coders. *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 1649-1652.

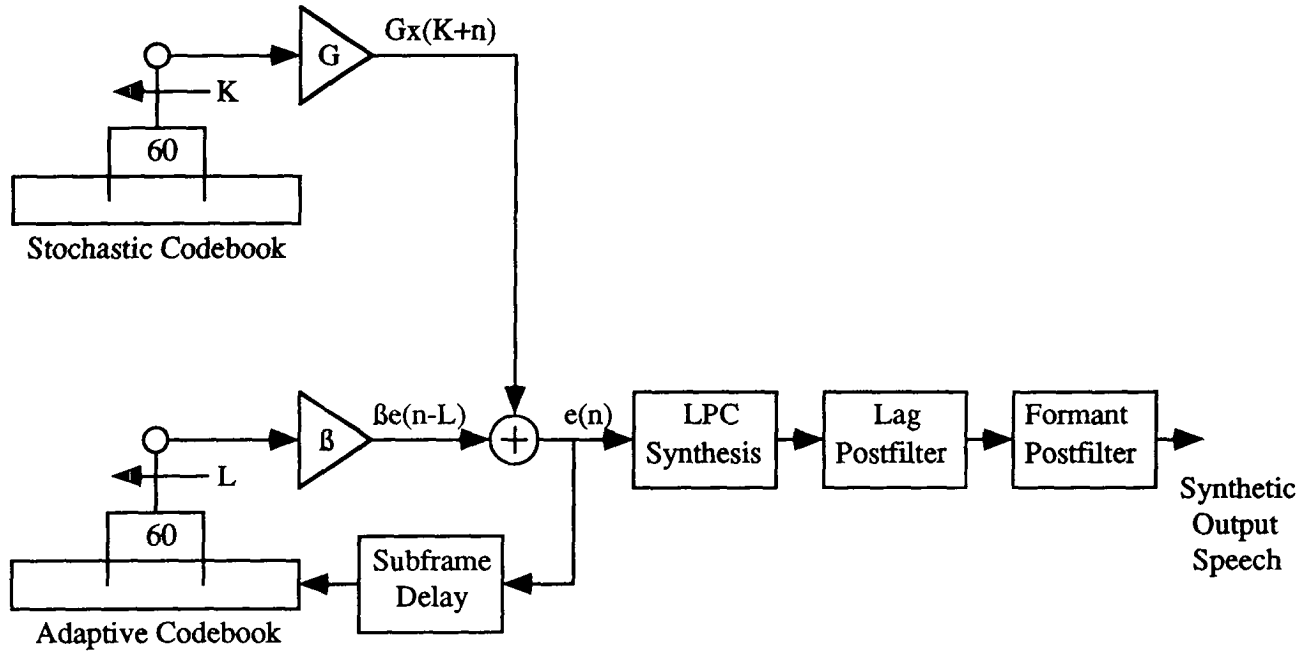

Figure 1. Synthesis structure of the reference coder.

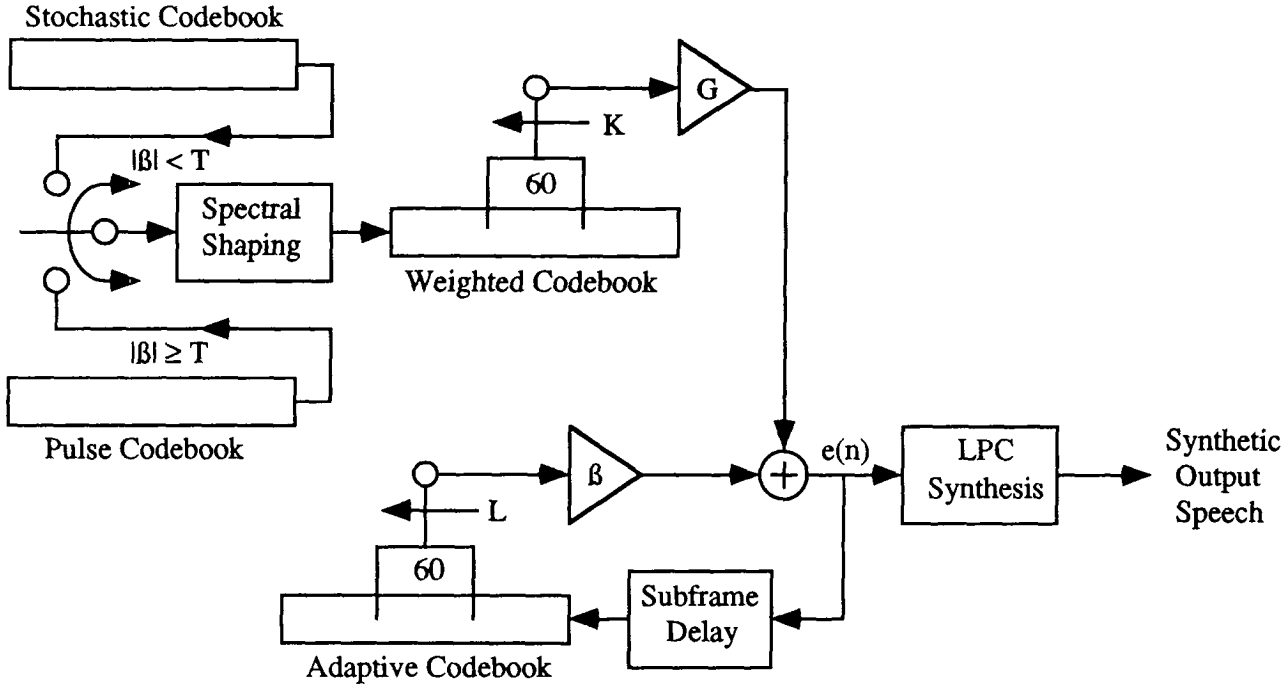

Figure 2. Modified synthesis structure.

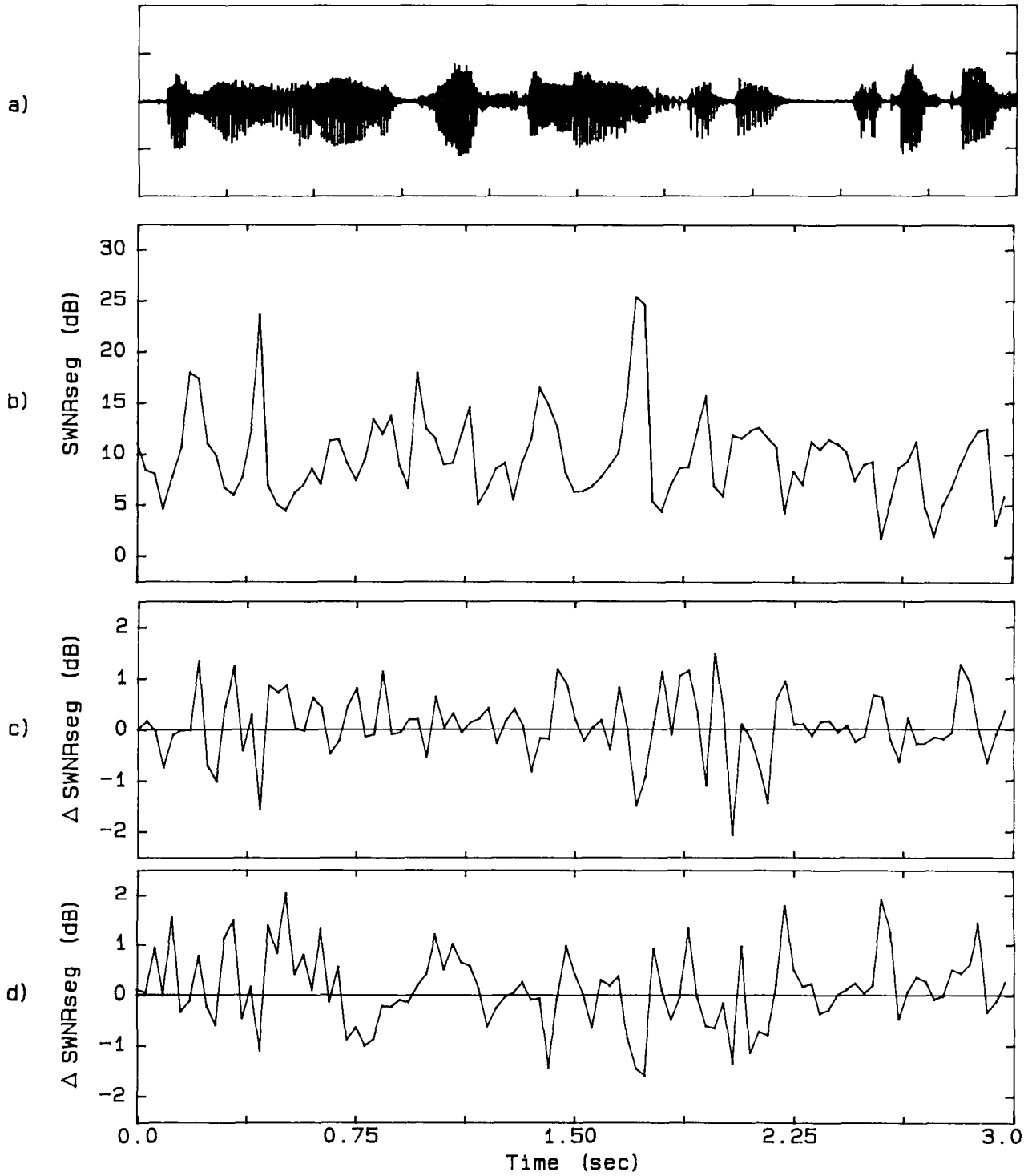

Figure 3. Illustration of SWNR improvements. a) Input speech signal, b) SWNR of the reference coder, c) difference in SWNR between the coder with a weighted codebook and the reference coder, d) difference in SWNR between the coder with a dual codebook and the reference coder.