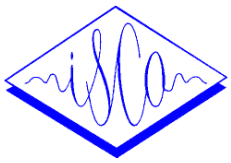

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

CONSTRAINED-STOCHASTIC EXCITATION CODING OF SPEECH AT 4.8 KB/S

Yair Shoham

Signal Processing Research Department
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974

ABSTRACT

This paper proposes a method for enhancing the performance of Codebook-Excited Linear Predictive (CELP) coders. It is based on the observation that the codebook-driven excitation in these coders is noisy and that the noisy component is not adequately filtered by the LPC filter. It is proposed to adaptively constrain the amount of the noisy excitation by linking its level to a performance index of the long-term (pitch-loop) sub-system. This operation reduces the noisy effects of the excitation, enhances the synthesized speech periodicity and hence, the perceptual quality of the coder. Listening test results are presented to demonstrate the subjective improvement of this coder over the basic CELP. The CSEC technique has been implemented in various AT&T coders at 4.8 to 8.0 Kbps, including low-delay CELP, with both stochastic and trained codebooks. Noticeable improvement in speech quality has been achieved. The technique has also been incorporated in the proposed federal standard PFS1016 4.8 Kbps coder.

I. INTRODUCTION

In the last few years, Code-Excited Linear Predictive (CELP) coding has emerged as the most prominent technique for digital speech communication at rates of 8 Kb/s and below, and it is now considered the best candidate coder for digital mobile telephony and secure speech communication. While the CELP coder is able to provide fairly good-quality speech at 8 Kb/s, its performance at 4.8 Kb/s is yet unsatisfactory for many applications. Considerable efforts have recently been made by many researchers towards improving the CELP coding technique at low bit rates.

CELP is an LPC-based coding technique in which speech is traditionally synthesized by passing an excitation signal through an all-pole LPC filter. The novelty of CELP is in the combination of vector quantization (VQ) with an analysis-by-synthesis (close-loop) approach to the optimization of the excitation signal. In CELP, this signal is drawn from an excitation codebook so as to minimize the error between the original and synthesized speech. The codebook may be either stochastic, i.e., pseudo-randomly populated, or pre-designed over some training data for minimum global distortion. In either case, the excitation contains a noisy component which does not contribute to the speech synthesis process and cannot be completely removed by the filter. It is a common opinion among many researchers that new forms of excitation need to be studied in order to improve the CELP performance at low bit rates.

This paper reports on one study in this direction. It is proposed in this study to adaptively constrain the amount of the noisy excitation by linking its level to a performance index of the long-term (pitch-loop) sub-system. This operation reduces the noisy effects of the excitation, enhances the synthesized speech periodicity and hence, the perceptual quality of the coder. The effect of the constrained excitation is more noticeable at lower bit rates and, in particular (but not exclusively), when a stochastic codebook is used. Hence, the proposed methods will be referred to as Constrained-Stochastic Excitation Coding (CSEC) of speech.

The CSEC technique has been implemented in various AT&T coders at 4.8 to 8.0 Kbps with noticeable improvement in speech quality. It has also been found to increase the coding quality of 8Kbps low-delay CELP coders that employed trained codebooks. The CSEC method has also been incorporated in the proposed federal standard PFS1016 4.8 Kbps coder [10].

The next section briefly reviews the basic CELP coder. Then, the concept of constrained excitation is introduced and the algorithm is discussed. Finally, listening test results are presented to demonstrate the subjective improvement of this coder over the basic CELP.

II. THE BASIC CODING SYSTEM

The coding system is based on the standard Codebook-Excited Linear Predictive (CELP) coder which employs the traditional excitation-filter model. A brief description of the system follows as a necessary introduction to the Constrained-Stochastic-Excitation Coding (CSEC) concept. More details on the CELP system can be found in numerous previous papers, e.g., [1]-[9].

The speech signal $s(n)$ is processed frame by frame and the frames are contiguous and equal in size. Throughout this paper, we use the convention that the current frame corresponds to the time window $[n = 0,..,N-1]$, N being the frame size.

$s(n)$ is filtered by a pole-zero, noise-weighing linear filter to obtain $X(z) = S(z)A(z)/A'(z)$ where $x(n)$ is the *target signal* used in the coding process. $A(z)$ is the standard LPC polynomial corresponding to the current frame, with coefficients a_i , $i=0,..,M$. ($a_0=1.0$). $A'(z)$ is a modified polynomial, obtained from $A(z)$ by shifting the zeroes towards the origin in the z-plane, that is, by using the coefficients $a'_i = a_i \gamma^i$ with $0.<\gamma<1.$ (typical value: $\gamma=0.8$). This pre-filtering operation reduces the quantization noise in the coded speech spectral valleys and enhances the perceptual performance of the coder [6].

The LPC filter $A(z)$ is assumed to be a quantized version of an all-pole filter obtained by the standard autocorrelation-method LPC analysis. The LPC analysis and quantization processes are independent of the other parts of the CELP algorithm and will not be discussed here.

The coder attempts to synthesize a signal $y(n)$ which is as close to the target signal $x(n)$ as possible, usually, in a mean-square-error (MSE) sense. The synthesis algorithm is based on the following simple equations

$$\sum_{i=0}^{M} a'_i \, y(n-i) = r(n) \qquad (1)$$

$$r(n) = \beta \, r'(n,P) + g \, c(n) \qquad (2)$$

$$r'(n,P) = \begin{cases} r(n-P) & , \ n < P \\ r'(n-P,P) & , \ n \geq P \end{cases} \qquad (3)$$

β and P are the so-called pitch tap and pitch lag respectively. g is the excitation gain and $c(n)$ is an excitation signal. Each of the entities β, P, g, $c(n)$ takes values from a predetermined finite table. In particular, the table for the excitation sequence $c(n)$ (the excitation codebook) holds a set of N-dimensional codevectors.

The task of the coder is to find a good (if not the best) selection of entries from these tables so as to minimize the distance between the target and the synthesized signals. The sizes of the tables determine the number of bits available to the system for synthesizing the coded signal $y(n)$.

Notice that Eq. (2) and (3) represent a 1st-order pitch-loop (with periodic extension [7]). Higher-order pitch loops could also be used. However, spreading the limited number of bits for transmitting parameters of more than one pitch loop has not been found to yield higher performance.

The actual output signal, denoted by $z(n)$ ($Z(z)$ in the z-domain), is obtained by using the inverse of the noise-weighting filter. This is accomplished simply by computing $Z(z) = R(z)(1/A(z))$ where $R(z)$ is the z-domain counterpart of $r(n)$. Note that, in general, minimizing the MSE distance between $x(n)$ and $y(n)$ *does not* imply the minimization of the MSE between the input $s(n)$ and the output $z(n)$. Nevertheless, the noise-weighting filtering has been found to significantly enhance the perceptual performance the CELP coder.

A key issue in CELP coding is the strategy of selecting a good set of parameters from the various codebooks. A global exhaustive search, although possible in principle, is prohibitively complex. Therefore, sub-optimal procedures are used. A common and sensible strategy is to separate the pitch parameters P and β from the excitation parameters g and $c(n)$ and to select the two groups independently. P and β are found first and then, for a fixed such selection, the best g and $c(n)$ are found. $y(n)$ can be expressed in the form

$$y(n) = y_0(n) + \beta r'(n,P)*h(n) + g\,c(n)*h(n) \qquad (4)$$

where $y_0(n)$ is the response to the filter initial state without any input and $h(n)$ is the impulse response of $1/A'(z)$ in the range $[0,..,N-1]$. The notation * denotes the convolution operation. The best P and β are given by

$$P^*, \hat{\beta} = \underset{P,\beta}{\operatorname{argmin}} \; \| \; x(n) - y_0(n) - \beta r'(n,P)*h(n) \; \| \qquad (5)$$

where the search is done over all the entries in the tables for β and P. The notation $\| \ . \ \|$ indicates the Euclidean norm of the corresponding time-sequence. The values for P are typically in the integer range $[20,..,147]$ (7 bits). The table for β typically contains 8 discrete values (3 bits) in the approximate range $[0.4,..,1.5]$. Numerous low-complexity methods have been used for minimizing (5). A common sub-optimal method is to first minimize (5) for P with an *unquantized* β and, then, to quantize β that corresponds to the best P [3].

Once $\hat{\beta}$ and P^* are found, the coder attempts to find a best match to the resulting error signal $d(n) = x(n) - y_0(n) - \beta r'(n,P^*)*h(n)$ by finding

$$\hat{g}, \hat{c}(n) = \underset{g,c(n)}{\arg\min} \; \| \; d(n) - g\,c(n)*h(n) \; \| \qquad (6)$$

where the search is performed over all entries of the gain table and the excitation codebook. As for the pitch loop, the search for g, $c(n)$ can be performed sub-optimally by first searching for the best excitation with an unconstrained (unquantized) gain and, then, quantizing that gain.

The CSEC system departs from the basic CELP described above at the stage of selecting g and $c(n)$. In the CSEC system, these parameters are selected in such a way as to constrain the level of the excitation and make it adaptive to the performance of the long-term subsystem. The concept behind this approach is discussed next.

III. THE CONCEPT OF CONSTRAINED EXCITATION

The CELP coding approach is based on the observation that the LP residual is essentially white and can be characterized as a Gaussian process. The CELP coder attempts to replace the ideal LP residual by external pseudo-random Gaussian excitation sequences, held in a finite-size codebook, with the hope of obtaining a reasonable match to the source spectrum. Since these artificial excitation signals have nothing to do with the source, they poorly represent the *perceptually relevant* component of the residual and contain a significant amount of irrelevant noise.

There is no known explicit way of identifying or reducing the noisy components in the excitation codebook, based on the local source characteristics. Therefore, our philosophy is to treat the excitation as mainly a noise signal and to restrict its use in order to reduce the amount of noise injected into the system.

The two components of $y(n)$ in Eq. (4) which carry new information about the source are the "pitch" signal $p(n) = \beta r'(n,P)*h(n)$ and the filtered excitation $e(n) = gc(n)*h(n)$. $p(n)$ is the result of attempting to utilize the periodicity of the source. There is no additive noisy component in it and the new information is introduced by modifying the delay P and the scale factor β. It is therefore expected to be perceptually more appealing than the excitation noisy component $e(n)$. Fortunately, in voiced (periodic) regions, $p(n)$ is the dominant component. Figure 1 shows the RMS ratio (dB) of $p(n)$ to $e(n)$ as a function of time for a typical speech segment (shown at the top of the figure) composed of voiced and unvoiced regions. The RMS values were calculated over frames of 50 samples. In voiced regions, the RMS of $p(n)$ is about 15 dB higher than that of $e(n)$. Our aim will be to make $p(n)$ even more dominant in voiced regions by appropriately suppressing the signal $e(n)$.

We propose to reduce the level of the noisy excitation and to impose a heavier reconstruction burden on the pitch signal $p(n)$. However, since $p(n)$ is not always efficient in reconstructing the output, particularly in unvoiced and transitional regions, the amount of excitation reduction should depend on the efficiency of $p(n)$. The efficiency of $p(n)$ should reflect its closeness to $x(n)$ and may be defined in various ways. In this work we use the (signal to noise) ratio

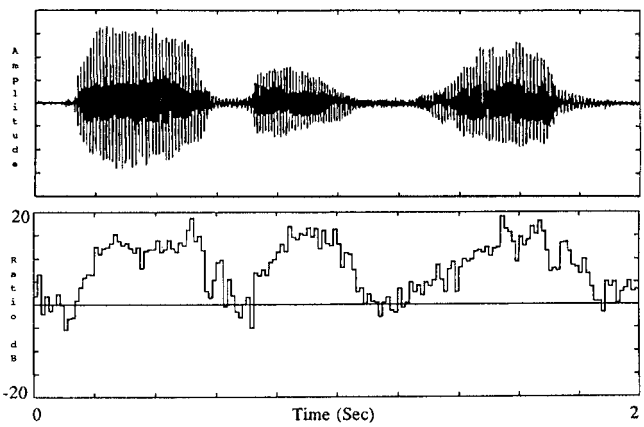

Figure 1. RMS Ratio (dB) of $p(n)$ to $e(n)$ in a Typical Speech Segment.

$$S_p = \frac{\| x(n) \|}{\| x(n) - y_0(n) - p(n) \|} \quad (7)$$

The quantity S_p is used in controlling the level of the excitation. Recalling that the excitation is perceived as essentially a noisy component, we define the signal-to-noisy-excitation ratio

$$S_e = \frac{\| x(n) \|}{\| e(n) \|} \quad (8)$$

The basic requirement now is that S_e be higher than some monotone-nondecreasing threshold function $T(S_p)$:

$$S_e \geq T(S_p) \quad (9)$$

Figure 2 shows the empirical function $T(S_p)$ on a dB scale, used in this work. It consists of a linear slope followed by a flat region. When S_p is high namely, $p(n)$ is capable of efficiently reconstructing the output, S_e is forced to be high and $e(n)$ contributes very little to the output. As S_p goes down, the constraint on $e(n)$ is relaxed and it gradually takes over, since $p(n)$ becomes inefficient. $T(S_p)$ is shaped by a slope factor α and a saturation level f. Based on limited listening to coded speech, we use the preliminary parameters $\alpha = 6.0$ and $f = 24.0$ dB. However, these parameters should be a subject of more careful optimization by intensive listening to coded speech.

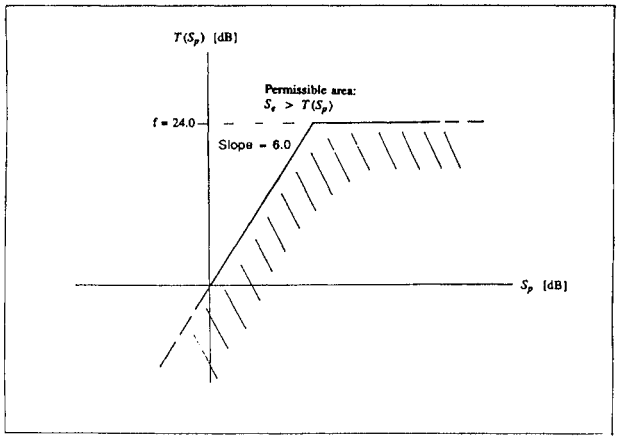

Figure 2. The Threshold Function $T(S_p)$

The procedure for constraining the excitation, whose details are discussed next, is quite simple: the system measures S_p for the current frame, determines the threshold using $T(.)$ and selects the best excitation $\hat{c}(n)$ and the best gain $\hat{g}$ subject to the constraint of Eq. (9).

IV. CONSTRAINED-EXCITATION SEARCH ALGORITHM

The objective is to find the best gain and excitation vector from the corresponding codebooks, under the constraint of Eq. (9). Defining the unscaled excitation response $c_h(n) = c(n)*h(n)$, the minimization problem is, therefore, stated as:

$$\hat{g}\ ,\ \hat{c}(n) = \underset{g\ ,\ c(n)}{\text{argmin}}\{-2g < d(n)\ ,\ c_h(n) > + g^2 \| c_h(n) \|^2\} \quad (10)$$

subject to:

$$| g | \ \| c_h(n) \| \leq \frac{\| x(n) \|}{T(S_p)} \quad (11)$$

where $< .,. >$ denotes the inner product of the arguments. The minimization range is the set of all the entries of the gain and excitation codebooks. It is clear from the quadratic form of the problem that for a fixed excitation $c(n)$ the best quantized gain is obtained by quantizing the unconstrained optimal gain, given by

$$g^* = \frac{< d(n)\ ,\ c_h(n) >}{\| c_h(n) \|^2} \quad (12)$$

Thus, for a given $c(n)$ the best quantized gain is:

$$\hat{g} = \underset{g}{\text{argmin}} \| g - g^* \| \quad (13)$$

subject to Eq. (11).

The search procedure is to obtain the best gain for each excitation vector as in (13), record the resulting distortion and to select the pair $\hat{g}\ ,\ \hat{c}(n)$ corresponding to the lowest distortion.

Notice that the constraint (11) is "soft" in the sense that if it is satisfied for the truly optimal gain no restriction takes place. In other words, the standard best excitation is used. This happens when the quantized version of the optimal gain g^* falls within the permissible range for the gain. This situation occurs mainly when $T(S_p)$ is low (unvoiced and transitional segments) which essentially defaults the system to a regular CELP. It may also happen in (voiced) regions of a very high LPC prediction gain, that is, very low excitation power. Note, however, that a pair $\hat{g}\ ,\ \hat{c}(n)$ for which the constraint does not apply is not necessarily the best choice. There may be another pair with a constrained gain that actually yields a lower distortion.

There may arise a situation in which the constraint (11) can not be satisfied for any pair of gain and excitation vector. This could be easily remedied by including a zero-valued gain in the table. However, such a solution would be inefficient from a coding efficiency standpoint since such a gain would rarely be used. Another practical solution to this problem is to somewhat relax the constraint by applying it to the optimal (unquantized) gain rather than to the quantized one. In this approach a modified optimal gain is defined as

$$g^{**} = \min\{\ | g^* |\ ,\ \frac{\| x(n) \|}{\| c_h(n) \|\ T(S_p)}\ \} sign(g^*) \quad (14)$$

This gain is computed and quantized for each excitation codevector and the pair $\hat{g}^{**}\ ,\ \hat{c}(n)$ minimizing the distortion, is selected.

Figure 3 shows the ratio in dB of a regular gain (basic CELP) to a constrained gain (CSEC) as a function of time, for the same speech segment. As shown, the gain reduction is high in voiced regions (up to 15 dB) and it is around zero in unvoiced regions.

Since the gain of the excitation is constrained, the highest possible SNR cannot, in general, be achieved. Therefore, it is sensible to use some gain-independent performance criterion. The correlation between the target $x(n)$ and the reconstructed signal $y(n)$ is a natural measure to think of. This measure turns out to be particularly simple if the inequality in the constraint (11) is replaced by an equality. In this case maximization of the correlation amounts to

$$\hat{c}(n) = \underset{c(n)}{\text{argmax}}\ \hat{g} < x(n)\ ,\ c_h(n) > \quad (15)$$

with

$$g = \frac{\| x(n) \|}{\| c_h(n) \| \; T(S_p)} \; sign(< x(n) \; , \; c_h(n) >) \qquad (16)$$

and $\hat{g}$, the quantized version of g. Based on a limited comparative listening test we obtained the impression that the correlation method performed slightly better than the MSE method.

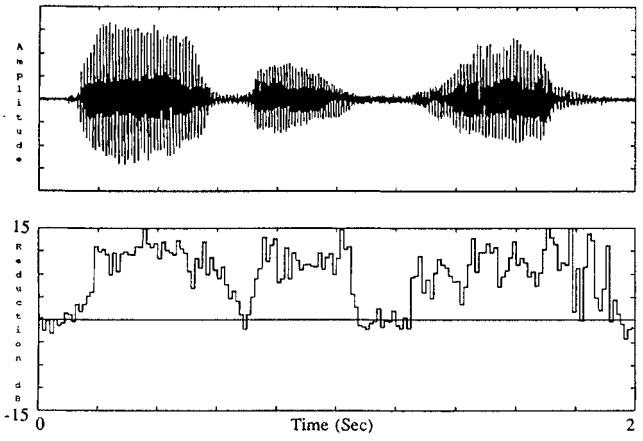

Figure 3. Gain Reduction (dB) vs. Time, as Determined by the CSEC System, for a Typical Speech Segment.

V. PERFORMANCE

The subjective performance of the proposed coder was measured by a so-called informal A-B comparison listening test and by a formal Mean Opinion Score (MOS) test. In the A-B test, 10 speech sentences were processed by CELP and CSEC coders. 10 listeners took part in this test and voted for the better coder in their judgement. The CSEC was similar to the CELP in all respects, except for the excitation search. The intent was to show the improvement obtained by the constrained-excitation principle with the rest of the system parameters unchanged. Both coders ran at 4.8 Kb/s. The speech sentences were taken from the "Phoneme-Specific" database [8]. The test scores, defined as an average percent votes in favor of a given system, are given in Table 1. As shown, the total average scores are 85% in favor of the CSEC system in contrast to 15% in favor of the basic CELP.

Listener	1	2	3	4	5	6	7	8	9	10	Average
CELP	0	20	20	20	30	10	10	30	10	0	15
CSEC	100	80	80	80	70	90	90	70	90	100	85

Table 1. Listening Test Scores (%) of a Comparison Between CELP and CSEC at 4.8 Kb/s.

In a formal MOS test, a 6.6 Kb/s CSEC coder was compared to two other 6.6 Kb/s CELP systems. The first one was the AT&T version of CELP implemented in hardware (real-time), which scored 3.45. The second was the software version of the first coder, with a mu-law input. It scored 3.44. The CSEC scored in this test 3.73, noticeably higher than the two other coders.

The above results shows that the CSEC coder performs distinctly better the CELP coder. The complexity of the CSEC coder is essentially the same as that of the CELP since the same type and amount codebook-search arithmetic is needed in both coders. Also, most of the complexity-reducing "tricks" that have been proposed for the CELP algorithm can be combined with the CSEC method. Therefore, the CSEC method Is essentially a no-cost improvement of the CELP algorithm.

VI. REFERENCES

[1] B.S. Atal, M.R. Schroeder , "Stochastic Coding of Speech Signals at Very Low Bit rates", Proc. IEEE Int. Conf. Comm., May 1984, P. 48.1

[2] M.R. Schroeder, B.S. Atal, "Code-Excited Linear Predictive (CELP): High Quality Speech at Very Low Bit Rates", Proc. IEEE Int. Conf. ASSP., 1985, pp. 937-940.

[3] P. Kroon, E.F. Deprettere "A Class of Analysis-by-Synthesis Predictive Coders for High-Quality Speech Coding at Rate Between 4.8 and 16 Kb/s.", IEEE J. on Sel. Area in Comm. SAC-6(2), Feb. 1988, pp. 353-363.

[4] P. Kroon, B.S. Atal, "Quantization Procedures for 4.8 Kb/s CELP Coders", Proc. IEEE Int. Conf. ASSP 1987 pp. 1650-1654.

[6] B.S. Atal, M.R. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria", IEEE Tr. ASSP, Vol. ASSP-27, No. 3, June 1979, pp. 247-254.

[7] W.B. Kleijn, D.J. Krasinski, R.H. Ketchum, "Improved Speech Quality and Efficient Vector Quantization in SELP", Proc. IEEE Int. Conf. ASSP, 1988, pp. 155-159.

[8] A.W.F. Huggins, R.S. Nickerson, "Speech Quality Evaluation Using Phoneme-Specific Sentences", J.Acoust. Soc.Am. 77(5) pp. 1896-1906, May 1985.

[9] G. Davidson, A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding", Proc. Int. Conf. ASSP 1986 pp. 3055-58

[10] J.P. Campbell Jr., T.E. Tremain, V.C. Welch, "The Proposed Federal Standard 1016 4800 bps Voice Coder: CELP", SPEECH TECHNOLOGY, Apr./May 1990, pp. 58-64.