LOW-DELAY CODE-EXCITED LINEAR-PREDICTIVE CODING OF WIDEBAND SPEECH AT 32 KBPS

Erik Ordentlich
Yair Shoham

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

ABSTRACT

The prospect of multi-channel/multi-user speech communication via the emerging ISDN has raised a lot of interest in advanced coding algorithms for 7KHz wideband speech. The capability of 32Kb/s wideband speech coding will support sustained high-quality stereo (or bilingual) voice transmission for audio-video teleconferencing over a basic rate 64Kb/s ISDN channel. In addition to the high-quality requirement, network applications are likely to dictate the use of algorithms with a very short coding delay. This paper reports on the use of the well known Codebook Excited Linear Predictive (CELP) [1-3] algorithm for 32Kb/s low-delay (LD-CELP) coding of wideband speech. The main problem associated with wideband coding, namely, spectral noise weighting, is discussed. The paper proposes an enhanced noise weighting technique and demonstrates its efficiency via subjective listening tests. In these tests, involving 20 listeners and 8 test sentences, the average rating for the proposed 32Kb/s LD-CELP was essentially equal to that of the 64Kb/s standard (G.722) CCITT wideband coder.

I. INTRODUCTION

The prospect of high-quality multi-channel/multi-user speech communication via the emerging ISDN has raised a lot of interest in advanced coding algorithms for wideband speech. In contrast to the standard telephony band of 200 to 3400 Hz, wideband speech is assigned the band 50 to 7000Hz and is sampled at a rate of 16000 Hz for subsequent digital processing. The added low frequencies increase the voice naturalness and enhance the sense of closeness whereas the added high frequencies make the speech sound crisper and more intelligible. The overall quality of wideband speech as defined above is sufficient for sustained commentary-grade voice communication as required, for example, in multi-user audio-video teleconferencing. Wideband speech is, however, harder to code since the data is highly unstructured at high frequencies and the spectral dynamic range is very high. In some network applications, there is also a requirement for a short coding delay which limits the size of the processing frame and reduces the efficiency of the coding algorithm. This adds another dimension to the difficulty of this coding problem.

The study reported in this paper investigates the use of the well known Codebook Excited Linear Predictive (CELP) [1-3] algorithm for wideband speech coding. The study is reported in detail in [4]. Here, a very brief report is given, which highlights the problem, the solution and the main results.

II. CELP AND LOW-DELAY CELP

The basic structure of conventional CELP [1-3] is shown in Figure 1. In the system, an excitation signal, drawn from an excitation codebook, excites an all-pole filter which is usually a cascade of an LPC-derived filter $1/A(z)$ and a so-called pitch filter $1/B(z)$. The LPC polynomial is given by $A(z) = \sum_{i=0}^{M} a_i z^{-i}$ and is obtained by a standard M^{th}-order LPC analysis of the speech signal. The pitch filter is determined by the polynomial $B(z) = \sum_{j=0}^{q} b_j z^{-j-P}$ where P is the current "pitch" lag - a value that best represents the current periodicity of the input and b_j's are the current pitch gains. Most often, the order of the pitch filter is $q=1$ and it is rarely more than 3. Both polynomial $A(z)$, $B(z)$ are monic.

The CELP algorithm implements a *closed-loop* (analysis-by-synthesis) search procedure for finding the best excitation. Each excitation vector is passed through the LPC and pitch filters in an effort to find the best match to the output, usually, in a weighted mean-squared error (WMSE) sense. As seen in Figure 1, the WMSE matching is accomplished via the use of a noise-weighting filter $W(z)$. The input speech $s(n)$ is first pre-filtered by $W(z)$ and the resulting signal $x(n)$ ($X(z)=S(z)W(z)$) serves as a reference signal in the closed-loop search. The quantized version of $x(n)$, denoted by $y(n)$, is a filtered excitation, closest to $x(n)$ in an MSE sense. The filter used in the search loop is the *weighted* synthesis filter $H(z) = W(z)/[B(z)A(z)]$.

The filter $W(z)$ is important for achieving a high perceptual quality in CELP systems and it plays a central role in the CELP-based wideband coder presented here, as will become evident later in this paper.

In general, the CELP transmitter codes and transmits the following five entities: the excitation vector, the excitation gain, the pitch lag, the pitch tap(s), and the LPC parameters. The overall transmission bit rate is determined by the sum of all the bits required for coding these entities.

The CELP coder uses a long block of future samples mainly for the LPC analysis and coding. This implies a long coding delay which is unacceptable in many applications. Low-Delay CELP [5] is based on the same CELP principle but it avoids the excessive coding delay by performing the LPC analysis in a *backward-mode*. In this mode, LPC analysis is performed on recent past *quantized* speech, eliminating the need for a long look-ahead data block and no LPC information is transmitted. If the pitch loop is not used and the gain is processed in a fully backward mode [5], all

CH2977-7/91/0000-0009 $1.00 © 1991 IEEE

the transmission bits are assigned to the excitation, which facilitates the use of an extremely short data frame. This, in turn, results in a very short coding delay.

The LD-CELP proposed here for coding wideband speech at 32 Kb/s employs backward LPC. Two versions of the coder were studied. The first included pitch loop. The second did not, i.e., $B(z)=1$. The general structure of the coder is that of Figure 1, excluding the transmission of the LPC information.

III. PERCEPTUAL NOISE WEIGHTING ANALYSIS IN CELP SYSTEMS

In closed-loop MSE waveform coding the quantization the error signal between the output and the target is white. As a result, the signal-to-noise ratio (SNR) is not uniform across the frequency range and the quantization noise tends to mask the low-energy regions of the input spectrum. This problem has been recognized and addressed in the context of CELP coding of telephony-bandwidth speech [6]. The solution was in a form of a noise weighting filter, added to the CELP search loop in a way shown in Figure 1. The general form of this filter is:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} \quad ; \quad 1 \le \gamma_2 < \gamma_1 \le 1 \tag{1}$$

where $A(z)$ is the LPC polynomial. The effect of γ_1 or γ_2 is to move the roots of $A(z)$ towards the origin, de-emphasizing the spectral peaks of $1/A(z)$. With γ_1 and γ_2 as in (1), the response of $W(z)$ has valleys (anti-formants) at the formant locations and the inter-formant areas are emphasized. In addition, the amount of an overall spectral roll-off is reduced, compared to the speech spectral envelope as given by $1/A(z)$.

In the CELP system of Figure 1, the unweighted error signal $E(z) = Y(z) - X(z)$ is white since this is the signal that is actually minimized. The final error signal is

$$\hat{S}(z) - S(z) = E(z)W^{-1}(z) \tag{2}$$

and has the spectral shape of $W^{-1}(z)$. This means that the noise is now concentrated in the formant peaks and is attenuated in between the formants. The idea behind this noise shaping is to exploit the auditory masking effect. Noise is less audible if it shares the same spectral band with a high-level tone-like signal. Capitalizing on this effect, the filter $W(z)$ greatly enhances the perceptual quality of the CELP coder.

As will be shown shortly, incorporating $W(z)$ in the CELP coder *decreases the MSE efficiency of the coder*, that is, the minimum overall mean square error can not be achieved. This poses a particularly delicate problem in the case of wideband speech coding due to the high spectral dynamic range. Perceptually efficient spectral shaping is obtained at the price of an increased overall noise level. However, there is an apparent tradeoff of shaping vs. noise-level and an excessive noise shaping usually deteriorates the performance.

The performance of the CELP coder is related to predictability or the non-spectral-flatness of the speech input. This fact is not that obvious in the framework of a closed-loop analysis-by-synthesis system since prediction is not explicitly performed. The CELP system can be heuristically analyzed with the aid of an equivalent "DPCM" coder shown in Figure 2. In this figure, the transmitter of Figure 1. was redrawn in a different way. Virtual predictor and quantizer were added to make the circuit look like a standard differential coder. Note that the weighting filter is now represented by $\beta W(z)$ where $W(z)$ is assumed to be a rational *monic* filter, that is, the leading terms of both the numerator and denominator are strictly 1.0. The virtual predictor

$$P_r(z) = 1 - \frac{B(z)A(z)}{W(z)} \tag{3}$$

plays an important role in this analysis. It is easy to show that is guarantees the fundamental DPCM relation, namely,

$$\hat{R}(z) - R(z) = Y(z) - X(z) = E(z) \tag{4}$$

Note that for $P_r(z)$ to be a valid predictor the leading term of its impulse response must be strictly 0.0. This is guaranteed by $B(z),A(z),W(z)$ being all monic. Note, also, that if $W(z) = 1$, the predictor reduces to that of a standard DPCM system.

The relation (4) allows us to view the quantization operation as done by a virtual vector quantizer $Q(r)\beta$, (see Fig. 2) although, in reality, it is done in output domain by matching $Y(z)$ to $X(z)$. The performance of such a quantizer, in terms of MSE SNR, is given by the well known expression

$$\sigma_e = \sigma_r \varepsilon 2^{-b} \tag{5}$$

where σ_e is the variance of the quantization error, σ_r is the variance of the *virtual* residual $R(z)$, and b is the coding rate in bits per sample. Experimental results indicate that, for a low bit rate ($b<1$ as in the usual CELP case) and an uncorrelated input (such as $R(z)$), the quantizer constant ε is close to 1.0. Henceforth, we neglect this constant.

To proceed with this analysis we make the usual assumption that the output $Y(z)$ is very close to $X(z)$, i.e., the quantization error is small. So, using $Y(z) \simeq X(z)$, we get

$$\begin{aligned} R(z) &\simeq X(z) - [1 - B(z)A(z)W^{-1}(z)]X(z) \\ &= B(z)A(z)W^{-1}(z)X(z) = \beta S(z)B(z)A(z) \end{aligned} \tag{6}$$

The result $R(z) \simeq \beta S(z)B(z)A(z)$ is rather striking. It tells us that the virtual residual is approximately *independent of the weighting filter* up to some fixed gain. Moreover, it tells us that $R(z)$ is approximately a minimum-energy residual obtained by inverse-filtering that is characterized by maximum LPC prediction gain:

$$P_G = \frac{\sigma_s}{\sigma_r} \tag{7}$$

where σ_s is the variance of the input speech. Note that P_G characterizes the input signal only and it has nothing to do with the coding system.

Using the above, the variance of the error $E(z)$ is given by $\sigma_e = \beta(\sigma_s / P_G)2^{-b}$. Since $E(z)$ is statistically white, the average-magnitude-spectrum of the final error signal

$$D(z) = \hat{S}(z) - S(z) = E(z)\beta^{-1}W^{-1}(Z) \tag{8}$$

is given by

$$M_D(\omega) = \frac{\sigma_s}{P_G} 2^{-b} \mid W^{-1}(\omega) \mid \tag{9}$$

The MSE performance is, therefore, linearly related to the prediction gain of the signal. The output noise is affected only by the monic part of the weighting filter. Henceforth, we assume without loss of generality, that $W(z)$ is monic.

The power of the output error signal $D(z)$ is given by

$$\sigma_d^2 = \frac{\sigma_s^2}{P_G^2} 2^{-2b} \int_{-\pi}^{\pi} \mid W^{-2}(\omega) \mid d\omega/2\pi \tag{10}$$

Using the standard logarithmic inequality we get

$$\sigma_d^2 = \frac{\sigma_s^2}{P_G^2} 2^{-2b} \int_{-\pi}^{\pi} | W^{-2}(\omega) | d\omega/2\pi \geq$$

$$\frac{\sigma_s^2}{P_G^2} 2^{-2b} \left[\int_{-\pi}^{\pi} \log | W^{-2}(\omega) | d\omega/2\pi + 1 \right] \quad (11)$$

Since $W(z)$ is rational, monic and stable, the first term inside the brackets is zero, which leaves us with

$$\sigma_d^2 \geq \frac{\sigma_s^2}{P_G^2} 2^{-2b} \quad (12)$$

with equality if and only if $W(z) = 1$. Therefore, applying a weighting filter always increases the overall noise level.

A smoothed version of the speech input spectrum may be represented by

$$S_S(\omega) = \frac{\sigma_r}{B(\omega) A(\omega)} \quad (13)$$

Combining (9) and (13) we get an expression for a frequency-dependent signal to noise ratio in the log domain:

$$SNR(\omega) = \log \left| \frac{W(\omega)}{B(\omega) A(\omega)} \right| + b \log 2 \quad (14)$$

Averaging this expression over the frequency range shows that the only non-zero term left is $b \log 2$ (recall that all the filters are monic). Since $b \log 2$ is a small value, the SNR is bound to be negative in some frequency regions, mostly at high frequencies. It is also clear from (14) that the weighting filter $W(z)$ cannot improve this situation globally. It can only improve the SNR in some regions at the price of reducing the SNR in others.

IV. PERCEPTUAL NOISE WEIGHTING IN CELP CODING OF WIDEBAND SPEECH

The spectral dynamic range of wideband speech is considerably higher then that of telephony speech where the added high-frequency region of 3400 to 6000 Hz is usually near the bottom of this range. Expression (14) in the previous analysis tells us that, in this case, the unweighted SNR tends to be highly negative at high frequencies. While coding of the low-frequency region seems to be easier, coding of the high-frequency region poses a severe problem. The auditory system is quite sensitive in this region and the quantization distortions are clearly audible in a form of crackling and hiss. Noise weighting is, therefore, more crucial, in wideband CELP. The balance of low to high frequency coding is more delicate. The major effort in this study was towards finding a good weighting filter that would allow a better control of this balance.

The starting point of this investigation was the weighting filter of the conventional CELP as in (1). The goal was to find a good set (γ_1, γ_2) for best perceptual performance. It was found that, similar to the narrow-band case, the values γ_1=0.9 , γ_2=0.4 produced reasonable results. However the performance was not yet satisfactory. It was found that the filter $W(z)$ as in (1) has an inherent limitation in modeling the format structure and the required spectral tilt concurrently. The spectral tilt is more or less controlled by the difference $\gamma_1 - \gamma_2$. The tilt is global in nature and it is not possible to emphasize it separately at high frequencies. Also, changing the tilt affects the shape of the formants of $W(z)$. A pronounced tilt is obtained along with higher and wider formants, which puts too much noise at low frequencies and in between the formants. The conclusion was that the formant and tilt problems ought to be *decoupled.* The approach taken was to use $W(z)$ only for formant modeling and to add another section for controlling the tilt only. The general form of the new filter is

$$W'(z) = W(z) P(z) \quad (15)$$

where $P(z)$ is responsible for the tilt only. Various forms of $P(z)$ were studied. A detailed discussion of these forms can be found in [4]. In this paper we focus on a solution that has been found, after careful listening to coded speech, to work the best. In the proposed system, the weighting filter (15) incorporates two-pole section

$$P(z) = \frac{1}{1 + \sum_{i=1}^{2} p_i \delta^i z^{-i}} \quad (16)$$

The coefficients p_i are found by applying the standard LPC algorithm to the first three correlation coefficients of the current-frame LPC inverse filter ($A(z)$) sequence a_i. The parameter δ is used to adjust the spectral tilt of $P(z)$. The values $\delta = 0.7$, γ_1=0.95 and γ_2=0.8 were found to yield the best perceptual performance.

Figure 3 demonstrates the effect of the enhanced noise-shaping filter. The dashed curve in the figure shows a typical spectrum of a conventional inverse filter $W^{-1}(z)$. The solid curve is the spectrum of an enhanced inverse filter $W^{-1}(z) P^{-1}(z)$ for the same underlying LPC filter. As seen, for the same general tilt, the enhanced filter has less pronounced formants especially at lowest and highest formants. Compared to the old filter, the enhanced filter attenuates the noise by about 5 dB at the lowest and highest formant, while achieving the right overall spectral tilt.

The CELP coder with the noise weighting as in (15, 16) was implemented with various pitch loops, i.e., various orders for $B(z)$ and various number of bits for the pitch taps. Interestingly, the best system - in terms of both perceptual and MSE performance - was the one *without a pitch loop,* i.e., $B(z) = 1$.

V. SUBJECTIVE PERFORMANCE TEST

The performance of the wideband LD-CELP was assessed by comparing it to the G.722 CCITT standard wideband coder [7]. This coder operates at 64 Kb/s, twice the rate of LD-CELP. The LD-CELP algorithm employed an adaptive two-pole tilt filter with δ=0.7, in combination with a formant filter $W(z)$ defined by γ_1=0.98 and γ_2=0.8. Backward LPC analysis was used, of order 32. The coder did not use a pitch loop. The coding block size was 5 and the codebook size was 1024.

The test material included 4 male and four female utterances. Each utterance was coded by the G.722 and by the LD-CELP to form a pair of utterances. The order in a pair was set at random. However, each pair was played in both orders to eliminate biased decisions. Twenty listeners took part in the test. The listener was asked to vote for the better sounding utterance in his judgement, or, to split his vote equally, if no preference could be made. The final scores were defined as the percent number of votes for each system per given condition.

Table 1 shows the experimental results. As shown, the overall scores are 51.72 and 48.28 for the LD-CELP and the G.722, respectively. This means that, on the average, the two systems performed alike, which is extremely encouraging, recalling that the LD-CELP rate is half that of the G.722. Also, the coding delay of the G.722 coder is 1.5

msec. whereas that of the LD-CELP is only 0.94 msec. In the ISDN environment, this means stereo, multi-user or bilingual communication over a basic-rate channel without loss of quality. When the scores are broken up in terms of utterance gender, one can notice an asymmetry in quality: the LD-CELP does better on males (62.81% preference) whereas the G.722 does better on females (59.38% preference). The reason for this is not yet understood.

The results show that 32Kb/s LD-CELP an an excellent candidate for wideband speech coding in high-quality communication networks. Although LD-CELP is more complex than the G.722, future advances in DSP technology will render this disadvantage insignificant.

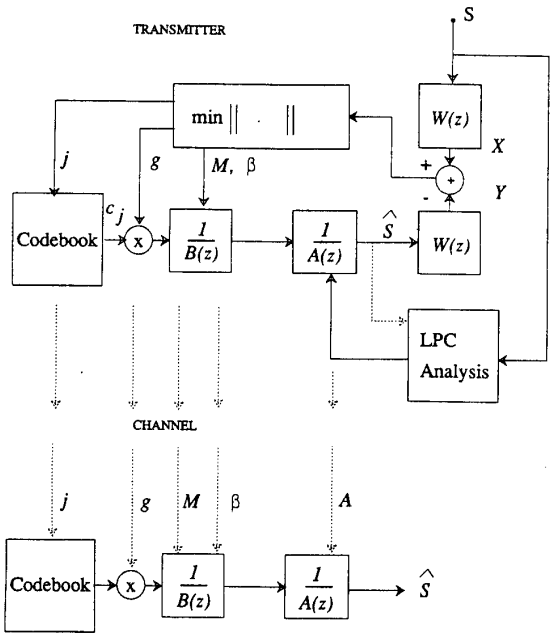

Figure 1. CELP Coder

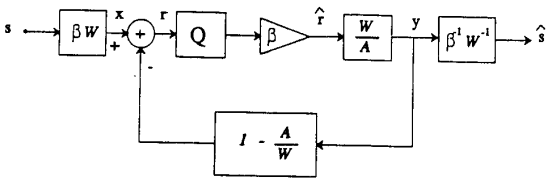

Figure 2. "DPCM" Equivalent of a CELP Coder

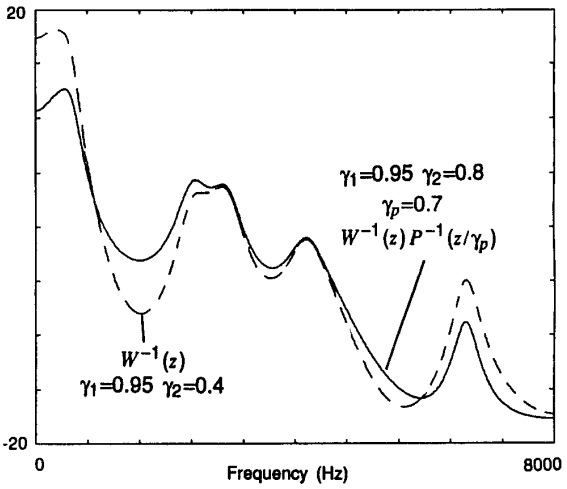

Figure 3. Effect of the enchanced noise-shaping filter.

Dashed line: Spectrum of a conventional inverse filter $W^{-1}(z)$.
Solid line: Spectrum of an enhanced inverse filter $W^{-1}(z)P^{-1}(z)$.

Utterances	LD-CELP (%)	G.722 (%)
All	51.72	48.28
Male	62.81	37.19
Female	40.62	59.38

Table 1. Listening test results for the proposed LD-CELP and the CCITT-G.722 wideband coders.

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] E. Ordentlich, "Low Delay Code Excited Linear Predictive (LD-CELP) Coding of Wide Band Speech at 32Kbit/sec.", MS Thesis, EE Dept., MIT, March 1990.

[5] J.H. Chen, "A Robust Low-Delay CELP Speech Coder at 16 Kb/s", Proc. GLOBECOM-89, Vol. 2, Nov. 89, pp. 1237-40.

[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] P. Mermelstein, "G.722, a New CCITT Coding Standard for Digital Transmission of Wideband Audio Signals", IEEE Comm. Mag., pp. 8-15, Jan. 1988.