A 4 KBPS ADAPTIVE FIXED CODE-EXCITED LINEAR PREDICTION SPEECH CODER

Hong Kook Kim[1] , Mi Suk Lee[2], and Hwang Soo Lee[2,3]

[1] currently at AT&T Labs Research, 180 Park Avenue, Florham Park NJ 07932, USA
work performed while author was at MMC Technology, Inc., Seoul, Korea
[2] Dept. of Electrical Eng., Korea Advanced Institute of Science and Technology, Taejon 305-701, Korea
[3] Central Research Laboratory, SK Telecom, 58-4 Hwaam-dong, Yusong-gu,Taejon 305-348, Korea
Email :[1] hkkim@research.att.com, [2] lms@spectra.kaist.ac.kr, [3] hslee@sktelecom.re.kr

ABSTRACT

In this paper, we propose an adaptive fixed code-excited linear prediction (AF-CELP) speech coder operating at 4 kbps. By exploiting the fact that a fixed codebook contribution to speech signal is also periodic as the corresponding adaptive codebook contribution, the adaptive fixed codebook model efficiently represents excitation signals. In order to overcome the quality degradation caused by the coarse quantization of excitation, a paired pulse algebraic codebook structure is also applied to the excitation model. Additionally, a pitch prefiltering, a noise spreading, and a harmonic enhancement technique are adopted in the decoding process. The spectrogram reading and informal listening tests proved that the AF-CELP reproduces high quality speech.

1. INTRODUCTION

In order to develop a CELP structure coder working at 4 kbps, the number of bits assigned to a spectral envelope, pitch, and an excitation signal should be reduced as much as possible. Spectral vector quantization (VQ) should provide the transparent performance in a view of spectral distortion [1]. However, the current speech coders employ a spectral VQ that performs somewhat lower than the transparent condition, which reaches about 20 bits/frame. As a way to further reduce the quantization bits, the predictive and/or classified VQ have been proposed by using speech specific information [2]. The relaxation CELP [3], which is standardized as IS127 - enhanced variable rate speech coder (EVRC) - in digital mobile communications, estimates and transmits a pitch value once a frame and modifies the input signal so that the resulting signal has the estimated period in a frame. The EVRC modifies the signal through pitch interpolation [4]. A successful way to reduce excitation bits was proposed by an extension of ITU-T G.729 [5], which reduces the number of bits assigned to the algebraic codebook from 34 bits of the 8 kbps CS-ACELP [6] to 22 bits. To compensate for the quality degradation due to a sparse algebraic codebook, the coder employed an additional postfiltering technique.

In this paper, we propose a 4 kbps CELP coder, which is based on the adaptive fixed codebook model [7]. Four additional techniques - a paired pulse algebraic excitation model, a pitch prefiltering, a noise spreading, and a harmonic enhancing technique - are introduced to overcome the artifacts occurring from modeling the fixed codebook coarsely.

2. OVERVIEW OF AF-CELP CODER

Figures 1 and 2 show the block diagram of the AF-CELP encoder and decoder, respectively. The frame size of the coder is 20 ms, which corresponds to 160 speech samples. Each frame is divided into three subframes. The length of each subframe is 53, 54, and 53 samples. The linear prediction coefficients of order 10 are obtained once every frame and then converted into line spectrum pairs (LSPs). The LSPs are quantized by using an 18 bits/frame multi-stage split vector quantizer. The quantized LSPs are used to construct a perceptual weighting filter as well as a synthesis filter for an analysis-by-synthesis loop. Input signal is first filtered by the perceptual weighting filter, and used as a target signal to find an optimal excitation signal. An initial pitch is obtained by using an open-loop search procedure so that the correlation of the filtered signal is maximized. The AF-CELP has three kinds of excitation codebooks: adaptive codebook, algebraic structured fixed codebook, and adaptive fixed codebook. For all subframes, the conventional method of adaptive codebook search is used. At the first subframe, the initial pitch is used as a starting value to search a final lag of adaptive codebook. 8 bits are used to quantize this lag with a resolution of 0.5. For the second and third subframes, the open-loop pitch is replaced with the integral part of the lag of the first frame and the search range is restricted from $-3\frac{1}{3}$ to $3\frac{1}{3}$ at a center of the integer pitch. This results in assigning 9 bits to the differential pitches of the second and the third subframes.

Next, we find an excitation signal that matches best the residual signal subtracted from the adaptive codebook contribution. For the first and the third subframes, we adopt an algebraic codebook by using three paired pulses in three tracks of 8 positions. Therefore, we can quantize the algebraic codebook with 12 bits (3 bits/track $\times$ 3 tracks + 3 signs). For the second subframe, an adaptive fixed codebook is applied to the excitation signal, but it does not transmit any information to the decoder. Finally, we compute two gains: one corresponds to an adaptive codebook gain and the other to an algebraic fixed codebook gain or an adaptive fixed codebook gain according to the subframe index. These gains are quantized every subframe by using a 7-bit, 2-dimensional VQ.

Table 1 shows the bit allocation of the 4 kbps AF-CELP. From the table, we know that the number of bits assigned to the fixed codebook occupies at most 30% of the total bits.

3. EXCITATION MODELING

3.1 Paired Pulse Algebraic Codebook

A paired pulse algebraic codebook structure is proposed for the AF-CELP. In this codebook, each codevector contains three signed pulses, and each pulse is accompanied by a smaller sized pulse, which is located at the next position of the pulse. Consequently, the excitation has six pulses in a

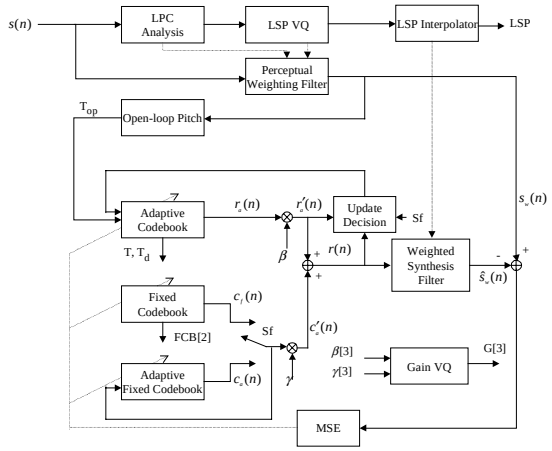

Figure 1. Block diagram of the 4 kbps AF-CELP encoder.

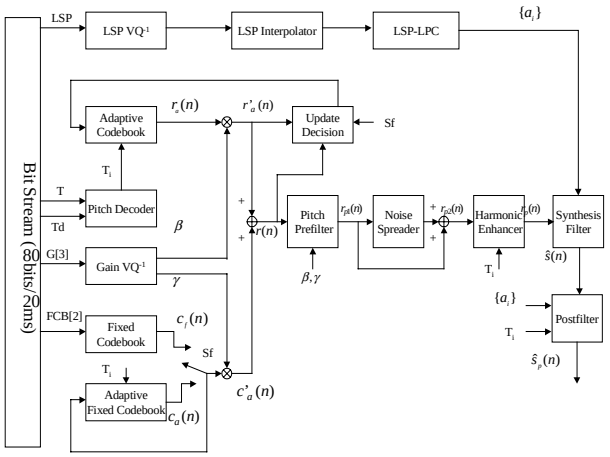

Figure 2. Block diagram of the 4 kbps AF-CELP decoder.

subframe as $c_f(n) = \sum_{i=0}^{2}\left(s_i\delta(n-p_i) + \Delta s_i\delta(n-p_i+1)\right)$, where s_i represents an *i*-th pulse amplitude, $\Delta s_i = c \cdot s_i$, and p_i is the *i*-th pulse position. We set c to 0.5 in the AF-CELP. Table 2 shows a track assignment for the paired pulse algebraic codebook structure.

3.2 Adaptive Fixed Codebook Model

The adaptive fixed codebook (AFC) modeling [7] was proposed by observing the periodicity of the fixed codebook excitation. The codebook contribution, $c'_a(n)$, is represented as $c'_a(n) = \gamma c_a(n) = \gamma\{\beta_a c_a(n-T_a)\}$. We have to determine γ as well as T_a and β_a. T_a and β_a can be replaced with the pitch lag T and the adaptive codebook gain β. This is because the target signal of the fixed codebook has high amplitudes near the pitch onset of the residual or the adaptive codebook signal. Also, γ is determined in a sense of minimum mean square as

$$\gamma = \frac{\sum_{n=0}^{N-1}(s_w(n) - s_0(n) - h_w(n) * r'_a(n))(h_w(n) * c_a(n))}{\sum_{n=0}^{N-1}(h_w(n) * c_a(n))^2}, \qquad (1)$$

where $s_w(n)$ is a perceptually weighted signal, $s_0(n)$ is a zero input response of the weighted synthesis filter whose impulse response is $h_w(n)$, and $r'_a(n)$ is previously determined from the result of the conventional adaptive codebook search. Finally, an AFC excitation scaled by γ is added to the adaptive codebook excitation and thus the resultant excitation is finally passed through the weighted synthesis filter to obtain the decoded speech signal.

Table 1. Bit allocation of the 4 kbps AF-CELP.

Parameters	Assigned Bits			Frame
	Sf 1	Sf 2	Sf 3	
LSP	-			18
Pitch	8	4.5	4.5	17
Fixed CB	12	0	12	24
Gain	7	7	7	21
Total				80

Table 2. Algebraic codebook structure of the 4 kbps AF-CELP.

Pulse	Sign	Positions
p_0	± 1	0, 7, 14, 21, 28, 35, 41, 47
p_1	± 1	2, 9, 16, 23, 30, 37, 43, 49
p_2	± 1	4, 11, 18, 25, 32, 39, 45, 51

The adaptive codebook is updated differently according to the subframe index (Sf) and the correlation of the adaptive codebook excitation. When the algebraic codebook is employed for the first or the third subframe, the memory of the adaptive codebook for the corresponding subframe is updated in the same way as a conventional adaptive codebook. At the second subframe, the adaptive codebook memory is updated by the rule of the following update decision logic. We first compute g_r and g'_r that are the correlation coefficients with a lag T of $r(n)$ and $r'_a(n)$, respectively.

$$g_r = \frac{\sum_{n=0}^{N-1} r(n)r(n-T)}{\sum_{n=0}^{N-1} r(n-T)r(n-T)} \quad \text{and}$$

$$g'_r = \frac{\sum_{n=0}^{N-1} r'_a(n)r'_a(n-T)}{\sum_{n=0}^{N-1} r'_a(n-T)r'_a(n-T)}, \qquad (2)$$

where N is the subframe size. When $g_r > g'_r$, we update the adaptive codebook memory with the current excitation. Otherwise, the adaptive fixed codebook excitation does not contribute to updating the memory of adaptive codebook. The numerator of g_r can be expanded in terms of $r'_a(n)$ and $c'_a(n)$. In transient regions, $\sum_n (r'_a(n)c'_a(n-T) + r'_a(n-T)c'_a(n))$ becomes small or negative and thus g_r decreases. In case the decision logic is not considered, using the conventional updating method increases a periodicity even though the original residual is not periodic.

4. EXCITATION ENHANCEMENT

4.1 Pitch Prefilter Technique

The all-pole pitch prefilter is used as

$$H_p(z) = \frac{1}{1 - g_p z^{-T}} \ , \qquad (3)$$

where g_p and T are the gain of pitch filter and the estimated period of decoded excitation, respectively. The excitation signal is reconstructed by the contribution of adaptive and fixed codebooks as

$$r(n) = \beta r'_a(n) + \gamma c'_a(n) \ . \qquad (4)$$

In the proposed technique, the excitation of (4) is filtered by (3). This is expressed as $r_{p1}(n) = \alpha(n)(r(n) + g_p Fr(r(n-T)))$, where $g_p = \max\{0.5\beta,\ 0.4\}$ and $g_p = \max\{0.4\beta,\ 0.4\}$ for the second subframe and the others, respectively, and $Fr(r(n-T))$ is a delayed excitation signal with a fractional lag of T. We use the same procedure to find the adaptive codebook search of the encoder. In order to match the gain of the prefiltered excitation with that of the unfiltered excitation, we perform a gain matching procedure. The gain scale value, $\alpha(n)$, can be approximated as

$$\alpha(n) = SCALE_\alpha \ \alpha(n-1) + (1 - SCALE_\alpha) G_p \, , \qquad (5)$$

where $SCALE_\alpha$ is a slowly varying weighting factor and set to be 0.925. A scale vale for a subframe, G_p, in (5) is computed as follows: we can denote $r_{p1}(n) = G_p\,(r(n) + g_p Fr(r(n-T)))$ for $G_p > 0$. From the equality of $\sum_{n=0}^{N_s-1} r_{p1}(n)^2 = \sum_{n=0}^{N_s-1} r(n)^2$ and the assumption of $Fr(r(n-T)) \approx r(n)/\beta$, we obtain the approximation of $G_p \approx 1/(1 + g_p/\beta)$.

4.2 Noise Spreading Technique

As with the excitation prefiltering, we add a random fluctuation into the prefiltered excitation, which provides enhanced perceptual quality of decoded speech. The resultant excitation is given by

$$r_{p2}(n) = \lambda(n)(r_{p1}(n) + g_c \ rand(n)) \ , \qquad (6)$$

where $g_c = \max\{0.3\gamma,\ 0.3\}$ and $rand(n)$ is a random number whose pdf has uniform distribution with zero mean. We also perform a gain matching procedure. Fortunately, we obtain that there is no need to match a gain. That is, $\lambda(n) \approx 1$. The simple proof is as follows: The power of the input signal is equal to that of the output signal: $\sum_{n=0}^{N_s-1} r_{p2}(n)^2 = \sum_{n=0}^{N_s-1} r_{p1}(n)^2$. Let the scale value for the current subframe be $\lambda(n) = \lambda$. From (6) and $\sigma_r^2 = \mathrm{var}(rand(n)) = 1/6$, we obtain $\lambda^2 (\sum_{n=0}^{N_s-1} r_{p1}(n)^2 + g_c \sigma_r^2 N_s) = \sum_{n=0}^{N_s-1} r_{p1}(n)^2$. Also. $\lambda^2 = 1/(1 + g_c \sigma_r^2 N_s / \sum_{n=0}^{N_s-1} r_{p1}(n)^2)$. Therefore, we can obtain an approximation of $\lambda \approx 1$ since $g_c \sigma_r^2 N_s \le 0.3/6 \cdot 54 = 2.7$ and $\sum_{n=0}^{N_s-1} r_{p1}(n)^2 >> 2.7$. Finally, $r_{p2}(n)$ is put into the harmonic enhancer of the decoder.

4.3 Harmonic Enhancement

A harmonic enhancer generates high frequency components of excitation signals and thus improves the quality of synthesized speech. The output of the harmonic enhancer is

$$r_p(n) = r_{p2}(n) + r_h(n) \ . \qquad (7)$$

In order to generate high frequency components, the following model is proposed.

$$r_h(n) = \sum_{m=M_l(T_0)}^{M_h(T_0)} A_m c_{s,m}(n; T_0, n_{0,m}) \ , \qquad (8)$$

$$c_{s,m}(n; T_0, n_{0,m}) = \cos((n + n_{0,m}) \frac{2\pi m}{T_0}), \ \ 0 \le n \le N_s - 1 \, , \qquad (9)$$

where T_0 is the lag of the adaptive codebook, $n_{0,m}$ and A_m are the estimated starting phase and amplitude of the m-th sinusoid for the current subframe, respectively. Additionally, $M_l(T_0) = BW_l \left\lfloor T_0 / 2 \right\rfloor$ and $M_h(T_0) = BW_h \left\lfloor T_0 / 2 \right\rfloor$ corresponding to the low and high cutoff harmonic bands, respectively.

4.3.1 *Phase Estimation*

$n^*_{0,m}$ is the optimal phase of the m-th sinusoid that satisfies the following equation:

$$n^*_{0,m} = \arg\max_{0 \le n_{0,m} < \lceil T_0/m \rceil} \sum_{n=0}^{N_s-1} r_{p2}(n) c_{s,m}(n; T_0, n_{0,m}) \, , \qquad (10)$$

where N_s is the subframe length. In order to find $n^*_{0,m}$ for all m, (10) is computed ($M_h(T_0) - M_l(T_0) + 1$) times.

4.3.2 *Amplitude Estimation*

We estimate the amplitudes of harmonics on the basis of the minimum mean-square error criterion. In other words, we obtain $\{A_m\}$ which minimizes the following equation: $\sum_{n=0}^{N_s-1} (r_{p2}(n) - r_h(n))^2 = \sum_{n=0}^{N_s-1} (r_{p2}(n) - \sum_{m=M_l(T_0)}^{M_h(T_0)} A_m c_{s,m}(n; T_0, n^*_{0,m}))^2$

By differentiating the above equation with respect to A_k, we obtain

$$A_k = \frac{\sum_{n=0}^{N_T-1} r_{p2}(n) c_{s,k}(n; T_0, n^*_{0,k})}{\sum_{n=0}^{N_T-1} c_{s,k}(n; T_0, n^*_{0,k})^2} \ , \ \ M_l(T_0) \le k \le M_h(T_0) \, , \qquad (11)$$

where $N_T (\le N_s)$ is an integral number whose integer multiple is T_0. In (11), we simplify the equation by using the orthogonal property of cosine functions: $\sum_{n=0}^{N_T-1} c_{s,m}(n; T_0, n^*_{0,m}) c_{s,k}(n; T_0, n^*_{0,k}) = \delta(m-k), \ 1 \le k \le \lfloor T_0/2 \rfloor$.

In addition, we can enhance the amplitudes as

$$A_k = (1 + C \cdot \frac{k}{M(T_0)}) A_k \, , \ \ M_l(T_0) \le k \le M_h(T_0) \, , \qquad (12)$$

where C corresponds to an enhancing factor.

4.3.3 *Harmonic Generation*

The enhanced excitation signal, $r_h(n)$, in (7) is generated as

$$r_h(n) = \sum_{m=M_l(T_0)}^{M_h(T_0)} A_m(n)\cos\Theta_m(n), \quad 0 \le n \le N_s - 1. \qquad (13)$$

In (13), the amplitude and the phase are interpolated in a manner similar to the method proposed in [8]:

$$A_m(n) = A_m(-1) + \frac{n+1}{N_s}(A_m - A_m(-1)), \quad 0 \le n \le N_s - 1, \qquad (14)$$

$$\Theta_m(n) = \frac{2\pi m\, n^*_{0,m}}{T_0} + \frac{2\pi mn}{T_0(-1)} + \left(\frac{2\pi}{T_0} - \frac{2\pi}{T_0(-1)}\right)\frac{mn^2}{2N_s}, \quad 0 \le n \le N_s - 1, \qquad (15)$$

where $A_m(-1)$ and $T_0(-1)$ are the amplitude of the m-th harmonic and the pitch period of the previous subframe, respectively. In the AF-CELP, we empirically set $BW_l = 0.7$ and $BW_h = 0.9$, which correspond to 2800 Hz and 3600 Hz. Additionally, C is set to 0.5.

5. PERFORMANCE EVALUATION

In order to verify the performance of the 4 kbps AF-CELP, we compared the spectrogram of the AF-CELP with that of the 8 kbps CS-ACELP as shown in Figure 3. In low-frequency bands, the AF-CELP regenerates harmonic structures faithfully and there is no difference with the harmonic structures of the CS-ACELP. For the AF-CELP, the power of a harmonic frequency is lowered at high-frequency bands. However, the harmonic components are still appeared and thus the AF-CELP provides a comparable performance to that of the CS-ACELP. The subjective listening test is going on and we expect that the AF-CELP will have a high MOS score.

6. CONCLUSION

We developed a 4 kbps CELP coder called the AF-CELP. The coder is designed based on the combination of an adaptive encoding technique and a modification of an algebraic structure. The former is exploited by the fact that a fixed codebook contribution is periodic to some extent in voiced regions. The latter, called paired pulse algebraic codebook structure, is also applied to reduce the number of bits for an algebraic codebook while keeping their effect on the excitation signals. Additionally, a pitch prefilter technique, a noise spreading technique, and a harmonic enhancement of excitation are introduced in the decoder side of the coder. Informal listening tests showed that the 4 kbps AF-CELP provides high quality of decoded speech.

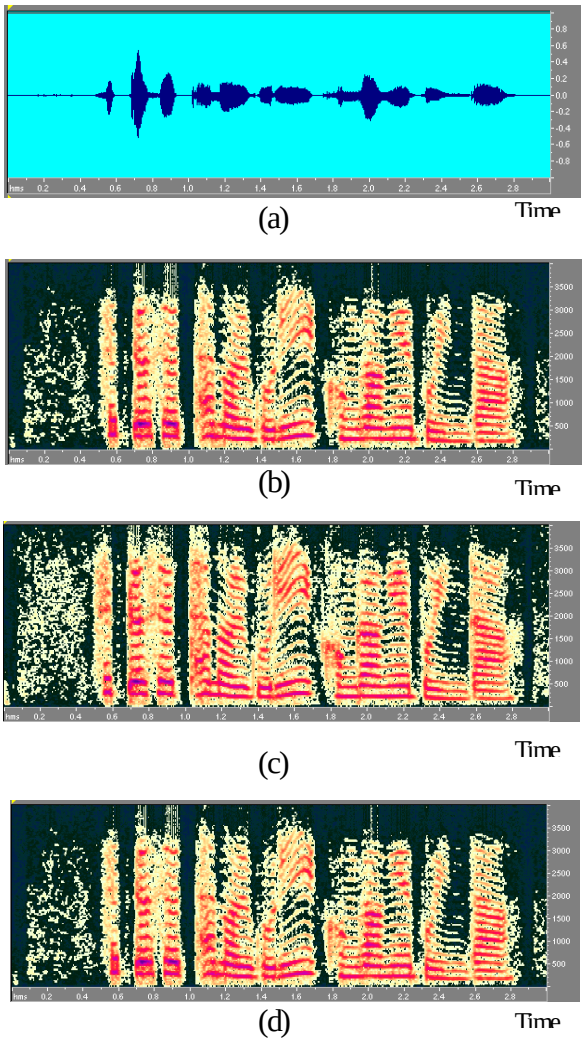

Figure 3. For (a) 3-second speech signals, comparisons of the spectrograms of (b) the original speech, (c) the decoded speech from the 4 kbps AF-CELP, and (d) the decoded speech from the 8 kbps CS-ACELP.

7. REFERENCES

[1] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3-14, Feb. 1993.

[2] M. Y. Kim, H. K. Kim, Y. D. Cho, and S. R. Kim, "Spectral envelope quantization with noise robustness," in *Proc. of 1997 IEEE Workshop on Speech Coding*, Pacono Manor, PA, pp. 77-78, Sept. 1997.

[3] W. B. Kleijn, P. Kroon, and D. Nahumi, "The RCELP speech coding algorithm," *European Trans. Telecomm.*, vol. 4, no. 5, pp. 573-582, 1994.

[4] W. B. Kleijn, R. P. Ramachandran, and P. Kroon, "Interpolation of the pitch-predictor parameters in analysis-by-synthesis speech coders," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 1, pp. 42-54, Jan. 1994.

[5] NTT and Ericsson, "High level description of the optimized 6.4 kbps extension to ITU-T Rec. G.729 (DRAFT)," ITU-T Contribution, Nov. 1997.

[6] R. Salami, C. Laflamme, J-P. Adoul, and D. Massaloux, "A toll quality 8 kbps speech codec for the personal communications system (PCS)," *IEEE Trans. Veh. Technol.*, vol. 43, no. 3, pp. 808-816, Aug. 1994.

[7] H. K. Kim, "Adaptive encoding of fixed codebook in CELP coders," in *Proc. of ICASSP*, Seattle, WA, pp. SP 5.4.1-4, May 1998.

[8] D.W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 36, no. 8, pp.1223-1235, Aug. 1988.