A HIGH QUALITY BI-CELP SPEECH CODER AT 8 KBIT/S AND BELOW

Soon Y. Kwon*, Hochong Park**, Hyokang Chang***

* Telogy Networks, Inc., 20250 Century Boulevard, Germantown MD, 20874
** Samsung Electronics Co., Ltd., 8-2 Karak-Dong, Songpa-Ku, Seoul Korea
*** Combasis Technology, Inc., 9400 Althea Ct., Rockville MD 20854

ABSTACT

This paper descibes "BI-CELP: baseline and implied CELP," which is a high quality speech coding method based on a code excited linear prediction (CELP) model employing excitation vectors combined from two codebooks, one from the baseline codebook and the other from the implied codebook. In this method the index of the baseline codebook is coded and transmitted to the receiver while the index of the implied codebook is extracted from the synthesized speech. This method has been applied to a lower rate voice coder at 8 Kbit/s to produce high quality voice comparable to that of the 16 Kbit/s G.728 LD-CELP. The performance of the 8 Kbit/s BI-CELP coder is measured in terms of SNRseg and MOS. The average SNRseg is 12.14 dB which is 0.6 dB higher than that of the 8 Kbit/s G.729 CS-ACELP. The MOS for the quiet input is 3.8 which is 0.02 higher than that of G.729 CS-ACELP. BI-CELP algorithm is implemented in real-time on a single TMS320C31 with 27 MIPS of CPU.

1. INTRODUCTION

CELP[1] has proven to be the most effective speech coding algorithm for the rates between 4 Kbit/s and 16 Kbit/s. The voice quality of the CELP coder has been improved by many researchers[2]-[5]. In particular, excitation codebooks have been extensively studied and developed for the CELP model. Random codebook of VSELP[2] and multi-pulse codebook of ACELP[3],[4] are successfully developed to characterize speech for the low rate applications. The addition of the pitch synchronous innovation(PSI)[5] to the CELP coder improves the perceptual voice quality significantly. Yet, the voice quality of the CELP coder operating at rates between 4 Kbit/s and 16 Kbit/s is not transparent because current codebooks for the CELP coder could specify only partial characteristics of the LPC residual signals, especially for the pitch harmonics in the synthesized speech. To generate high quality synthesized speech, the codebook for the CELP coder should be able to characterize the LPC residual spectrums of random noise source, energy concentrated pulse source, and mixture of both sources because of the speech characteristics.

Mixed excitation has been applied to the 2.4 Kbit/s LPC vocoder[6],[7] and has proven the effectiveness of the model. In the model of BI-CELP, an excitation vector is formed as a linear combination of two codevectors, one from the random codebook and the other from the multi-pulse codebook, to jointly characterize the properties of the excitation sources that may depend on the input speech and coding algorithms. The excitation vector thus formed is very effective in approximating the ideal excitation source in the frequency domain as shown in Fig. 4.

To reduce the data rate, only the index of the dominant codevector is transmitted while the index of the second codevector is derived from the synthesized speech delayed by the pitch period in the decoders. For this reason, the codebook that includes the dominant codevector is referred to as baseline codebook and the codebook that includes the second codevector as implied codebook. The gain of the implied codevector is transmitted for the decoder to recover the composite excitation vector. The selection of implied codevector using the delayed synthesized speech tends to maintain the pitch harmonic structure better in the synthesized speech than other CELP coders do. Previous models[3]-[5] to enhance pitch harmonics depend on the baseline codevector which may not be suitable for female speech whose residual signal is purely white.

BI-CELP model includes a delayed decision procedure to remove problems of discontinuity at the frame boundaries. The codebook parameters are jointly selected for the two consecutive half subframes to improve performance[5]. In this way, the frame boundary effects are greatly reduced without adopting a look-ahead procedure[8]. An efficient search algorithm for the real-time implementation has been developed to select the near optimum codebook parameters without significant performance degradation.

2. BI-CELP CODEC DESCRIPTION

The block diagrams of the BI-CELP encoder and decoder are shown in Fig. 1 and Fig. 2, respectively. The basic parameters of the speech codec are as follows:

Sampling Rate	8 KHz
Frame Length	80 samples(10 ms)
Subframe Length	40 samples(5 ms)
Half Subframe	20 samples(2.5 ms)
Data Rate	8 Kbit/s
Short-term Predictor Order	10
Long-term Predictor Order	1

In the BI-CELP encoder, input speech samples are high-pass filtered to remove dc-bias and unwanted low frequency background noise and buffered to form a frame of speech samples for the calculation of short-term (LPC) filter parameters, long-term (pitch filter) parameters, and codebook parameters. The parameters, shown in Table 1, are encoded and transmitted to the decoder in the receiver. There are 80 bits per frame and the bit allocation is shown in Table 1.

Table 1 BI-CELP Parameters and Bit Allocation

Transmission Parameters	Sub-frame 1	Sub-frame 2	Sub-frame 1&2
LSP low (lsp_1 to lsp_4)		9 bits	9
LSP high (lsp_5 to lsp_{10})		9	9
Pitch Period	8	5	13
Pitch Gain	3	2	5
Codebook Index/Half Subframe	6 6	6 6	24
Codebook Gains	10	10	20
Total number of bits	33	47	80

As we note from the block diagrams of Fig. 1 and Fig. 2, implied codevector or implied codebook index is calculated from the LPC output samples delayed by the pitch period at both encoder and decoder.

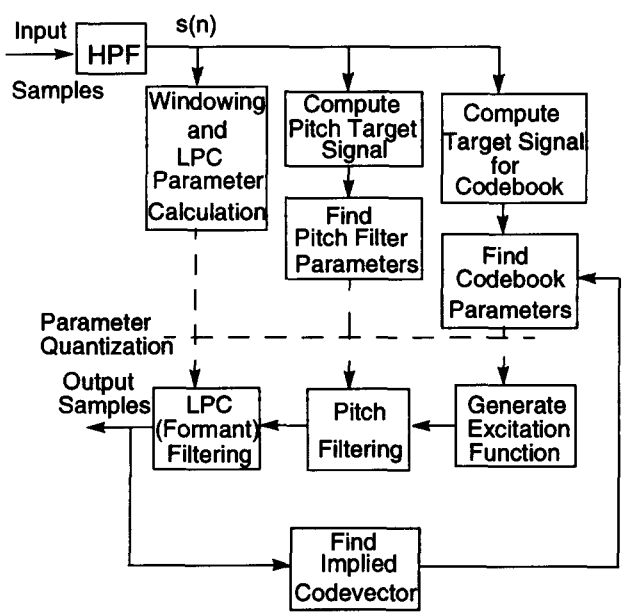

Fig. 1 BI-CELP Encoder

In the BI-CELP decoder, pitch filter input is calculated from decoded codebook parameters along with the implied codebook index calculated from the LPC filter output. The output of the pitch filter is filtered by the formant(LPC) filter to produce a speech signal which is filtered again by the adaptive post filter to enhance the perceptual voice quality.

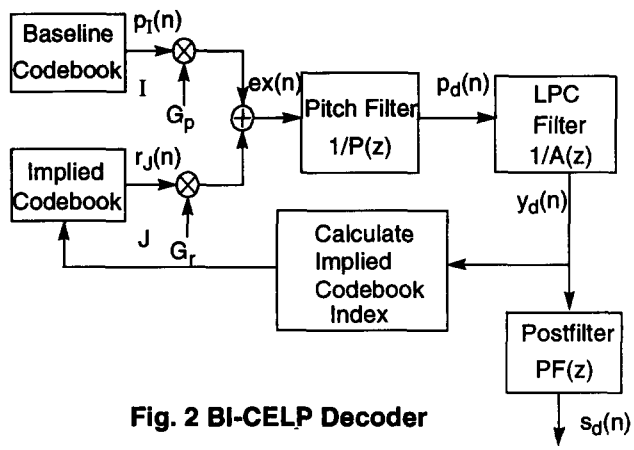

Fig. 2 BI-CELP Decoder

2.1 Windowing and Short-Term (LPC) Filter

The short-term(LPC) filter parameters are calculated at every frame from the autocorrelation functions computed from the Hamming windowed (160 samples including look ahead 40 samples, current frame 80 samples, and overlap 40 samples) input speech. The LPC prediction coefficients are scaled to perform bandwidth expansion and they are transformed into LSPs (line spectrum Pairs). The LSP quantizer encodes only for the second codebook subframe and the LSPs for the first frame are obtained by linearly interpolating the LSPs of current frame and previous frame. As the LSPs of adjacent frames are highly correlated, high coding efficiency of LSPs can be obtained by the moving average(MA) prediction as follows:

- Ten LSPs are split into low 4 LSPs and high 6 LSPs, i. e.,(w_1, w_2, w_3, w_4) and (w_5, w_6,..., w_{10})
- Mean values are removed, i. e., $f_i = w_i - Bias_i$
- LSP residuals are calculated via MA(Moving Average) prediction as

$$\delta_i(n) = f_i - \sum_{k=1}^{M} \alpha_k^{(i)} \hat{\delta}_i(n-k) \quad 1 \le i \le 10 \qquad (1)$$

where $\alpha_k^{(i)}$ = predictor coefficients, $\hat{\delta}_i(n)$ = quantized residuals for frame n and M=4=predictor order. The MA prediction coefficients are trained from a large data base to minimize the mean squared error of $\delta_i(n)$-$\hat{\delta}_i(n)$. Two vector quantization tables are developed for the vectors of δ_l = (δ_1, δ_2, δ_3, δ_4) and δ_h = (δ_5, δ_6, δ_7, δ_8, δ_9, δ_{10}). Quantized LSPs are obtained from the quantized LSP residuals of eq. (1), stability of the filter ensured, and converted to LPC prediction coefficients.

2.2 Long-Term (Pitch) Filter

The pitch filter parameters, pitch lag and pitch gain, are calculated from the perceptually weighted input signal x(n) with no filter of P(z) shown in Fig. 3 and the past pitch filter output by analysis-by-synthesis closed loop. The transfer function of the perceptual weighting filter is

$$W(z) = \frac{A(z)}{A(z/\zeta)} \qquad (2)$$

where A(z) is the LPC prediction filter and the perceptual weighting filter constant $\zeta = 0.89$. The transfer function of the pitch filter is

$$P(z) = \frac{1}{1 - \beta z^{-\left\lfloor \frac{n+L}{L} \right\rfloor L}}, \quad 0 \le n \le N_s - 1 \qquad (3)$$

where N_s =40 and $\lfloor x \rfloor$ is the floor function of x which provides the largest integer that is equal or less than x, β is the pitch gain, and n is the sample number in the subframe. The pitch lag is represented by 8 bits ranging from 19 to 147 and includes fractional numbers below 90. The pitch period for the first subframe is quantized with 8 bits while it is quantized with 5 bits for the second subframe. The pitch gain β is quantized with 3 and 2 bits for the first and second subframe, respectively.

The search routine for the pitch parameters of the entire pitch lag including fractional pitch periods involves substantial amount of calculations, and optimal long-term lags are usually fluctuating around the actual pitch period. In order to reduce the computations for the search of pitch parameters, an open loop pitch period(integer pitch period) is searched using the 30 ms window. The actual search for the pitch parameters is limited to ±20 lags around the open loop pitch period.

The open loop pitch may be extracted from the input speech signals directly or it may be extracted from the LPC prediction error signals. Pitch extraction from the LPC residual signals[6] is preferred to the one from the speech signals directly, since the pitch excitation sources are shaped by the vocal tract in the process of human speech production.

2.3 Codebook Parameters

BI-CELP uses two codevectors, one from the baseline codebook and the other from the implied codebook as shown in Fig. 3. There are two codebook subframes (5 ms) in a frame (10 ms). Each codebook subframe consists of two half codebook subframes (2.5 ms). Two codevectors (baseline codevector and implied codevector) are selected during each half codebook subframe. First, two implied codevectors, one from the random codebook and the other from the multi-pulse codebook are selected for each half codebook subframe based on the minimum mean square error (MMSE) between the target signal and the weighted LPC output due to the excitation function from the implied codebook.

The target signal for the implied codevector is the LPC filter output delayed by the pitch period. Therefore, implied codevector controls the pitch harmonic structure of the synthesized speech depending on the gain of the implied codevector.This gain is the new mechanism to control the pitch harmonic structures of the synthesized speech regardless of the selected baseline codevector. The approaches of PSI excitation [5] and pre-pitch filter excitation [3],[4] can control the pitch harmonic structure depending on the selected codevector which may not be adequate for some speech.

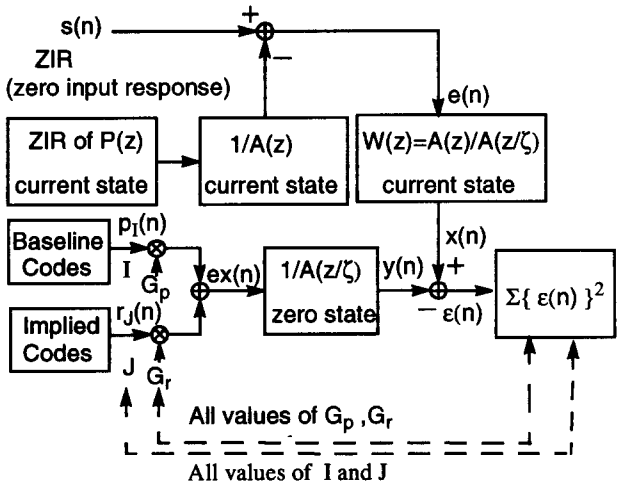

Fig. 3 Search for BI-CELP Codebiook Parameters

Second, the baseline codevectors are selected jointly with the candidate implied codevectors based on the weighted MMSE criterion with the target signal x(n) shown in Fig. 3. For the implied codevector from the multi-pulse codebook, the baseline codevector is selected from the random codebook, and for the implied codevector from the random codebook, the baseline codevector is selected from the multi-pulse codebook. In this way, the excitation functions always consist of multi-pulse and random codevectors as follows:

$$ex(n) = \begin{cases} G_{p1} p_{i1}(n) + G_{r1}(n) r_{j1}(n) & \text{for } 0 \le n \le 19 \\ G_{p2} p_{i2}(n) + G_{r2}(n) r_{j2}(n) & \text{for } 20 \le n \le 39, \end{cases} \qquad (4)$$

where $p_{i1}(n)$ and $r_{j1}(n)$ are the i1-th baseline codevector and j1-th implied codevector, respectively, for the first half codebook subframe, and $p_{i2}(n)$ and $r_{j2}(n)$ are the i2-th baseline codevector and j2-th implied codevector, respectively, for the second half codebook subframe. $G_{p1}, G_{r1}, G_{p2}, G_{r2}$ are the gains for the corresponding codevectors.

For the 8 Kbit/s BI-CELP, twelve bits are allocated for indices i1 and i2, and ten bits are allocated for the gain vector $\{G_{p1}, G_{r1}, G_{p2}, G_{r2}\}$. The gain vector and codebook indices are selected by a perceptually weighted MMSE criterion for all the baseline indices and gain vectors. In order to reduce the CPU load, the codebook parameters are searched and selected in three steps:

(1) Implied codevectors are identified for the first and second half codebook subframes.
(2) K and L sets of codebook index only are selected for the

first and second codebook subframes, respectively.

(3) One set of codebook parameters(codebook indices and corresponding gains) is selected from the KxL candidates of (2).

BI-CELP codebook parameters are jointly selected for the two consecutive half subframes from KxL=3x2=6 candidates to improve performance. In this way, the frame boundary effects are greatly reduced without adopting a look-ahead procedure[8]. The power spectrums of BI-CELP output(dotted line) and CS-ACELP output(step line) are plotted for a typical female input speech (solid line) in Fig. 4 to indicate the well produced pitch harmonic structure.

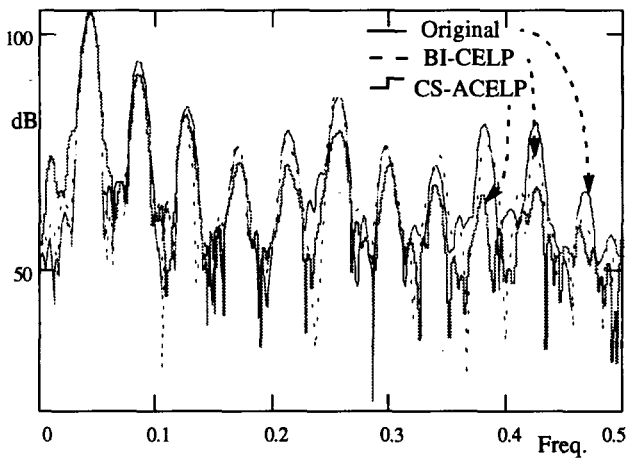

Fig. 4 Output Spectrums of BI-CELP and CS-ACELP

3. IMPLEMENTATION AND PERFORMANCE OF BI-CELP

The performance of the 8 Kbit/s BI-CELP speech codec is measured in terms of SNRseg and MOS. The objective voice quality in terms of segmental signal-to-noise ratio is measured from a large data base of various types of speech and tabulated in Table 2 along with the SNRseg of 8 Kbit/s G.729 CS-ACELP. The average SNRseg is 12.14 dB which is 0.59 dB higher than that of G.729 CS-ACELP

Table 2: SNRseg (dB) for 8 Kbit/s BI-CELP and CS-ACELP

File ID	Frame #	CS-ACELP	BI-CELP
1	2200	11.25	11.65
2	33670	10.52	11.45
3	37065	12.50	12.80
AVG	16207	11.55	12.14

The subjective voice quality is measured in terms of MOS by Dynastat in Austin Texas. The MOS for the clean input speech is 3.8 which is 0.02 higher than that of CS-ACELP.

Table 3: MOS for 8 Kbit/s BI-CELP, CS-ACELP, and VSELP

Input	Source	BI-CELP	CS-ACELP	VSELP
Clean	4.08	3.80	3.78	3.44

The BI-CELP algorithm is implemented in real-time on a single TMS320C31 DSP with 27 MIPS of CPU. Real-time communication tests in the office environment indicate that most listeners are comfortable about the voice quality of 8 Kbit/s BI-CELP. The effects of channel error, frame erasure, and acoustic background noise are now under investigation.

4. CONCLUSION

We have described a high quality speech coding method based on the BI-CELP model. In this method, mixed excitation functions are generated from the random codebook and multi-pulse codebook to specify the characteristics of the LPC residual signals. Implied codebook is used to reduce bit rate and delay decision is used to improve the performance while reducing the effects of frame boundaries. Real-time implementation indicates that 8 Kbit/s BI-CELP coder provides near toll-quality voice communication.

REFERENCES

[1] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction: High quality speech at very low rates," Proc. ICASSP-85, pp.937-940, 1985.

[2] I. Gerson and M. Jansiuk, "Vector sum excited linear prediction (VSELP) speech coding at 8 kb/s." Proc. ICASSP-90, pp. 461-464, 1990.

[3] R. Salami et. al., "Description of the proposed ITU-8 kb/s speech coding standard," Proc. IEEE Workshop on Speech Coding for Telecommunications, pp. 3-4, 1995.

[4] "Draft Recommendation G.729: Coding of speech at 8 Kbit/s using Conjugate-Structure Algebraic-Code-Excited Linear Predictive (CS-ACELP) Coding," ITU-T Study Group 15 Contribution - Q.12/15, June 1995.

[5] K. Mano et. al., "Design of a pitch synchronous innovation CELP coder for mobile communications," IEEE J. Sel. Areas Commun. vol. 13 pp. 31-41, Jan. 1995.

[6] S. Y. Kwon and A. J. Goldberg, "An enhanced LPC vocoder with no voiced/unvoiced switch," IEEE Trans. Acoust. Speech and Signal Processing, vol. ASSP-32, pp. 851-858, Aug. 1984.

[7] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," IEEE Trans. Speech and Audio Processing, vol. 3, pp.242-250, July 1995.

[8] S. Cucchi, M. Fratti, and M. Ronchi, "On improving performance of analysis by synthesis speech coders," IEEE Trans. Speech and Audio Processing, vol. 4, pp.243-247, May 1996.