Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 4.8 kbps

Ira A. Gerson and Mark A. Jasiuk
Chicago Corporate Research and Development Center
Motorola Inc.
1301 E. Algonquin Road, Schaumburg, IL 60196
Phone: (708) 576-3893
Fax: (708) 576-0541

ABSTRACT

Code Excited Linear Prediction (CELP) speech coders exhibit good performance at data rates as low as 4800 bps. The major drawback to CELP type coders is their large computational requirements. The Vector Sum Excited Linear Prediction (VSELP) speech coder utilizes a codebook with a structure which allows for a very efficient search procedure. Other advantages of the VSELP codebook structure will be discussed and a detailed description of a 4.8 kbps VSELP coder will be given. This coder is an improved version of the VSELP algorithm, which finished first in the NSA's evaluation of the 4.8 kbps speech coders [1]. The coder employs a subsample resolution single tap long term predictor, a single VSELP excitation codebook, a novel gain quantizer which is robust to channel errors, and a new adaptive pre/postfilter arrangement.

INTRODUCTION

Vector Sum Excited Linear Prediction falls into the class of speech coders known as Code Excited Linear Prediction (CELP) (also called Vector Excited or Stochastically Excited) [2,4,6]. The VSELP speech coder was designed to accomplish three goals:

1. Highest possible speech quality
2. Reasonable computational complexity
3. Robustness to channel errors

These three goals are essential for wide acceptance of low data rate (4.8 - 8 kbps) speech coding for telecommunications applications.

The VSELP speech coder achieves these goals through efficient utilization of a structured excitation codebook. The structured codebook reduces computational complexity and increases robustness to channel errors [1,3]. A single optimized VSELP excitation codebook is used to achieve high speech quality while maintaining reasonable complexity. A subsample resolution single tap long term predictor noticeably improves performance for high pitched speakers. A novel gain quantizer is employed which achieves high coding efficiency and robustness to channel errors. Finally, a new adaptive pre/post filter arrangement is used to enhance the quality of the reconstructed speech .

BASIC CODER STRUCTURE

Figure 1 is a block diagram of the VSELP speech decoder. The 4.8 kbps VSELP coder/decoder utilizes two excitation sources. The first source is the long term ("pitch") predictor state, or adaptive codebook [4]. The second is the VSELP excitation codebook. For the 4.8 kbps coder, the VSELP codebook contains the equivalent of 1024 codevectors. The excitation vectors, selected from the two excitation sources, are multiplied by their corresponding gain terms and summed, to become the combined excitation sequence ex(n). After each subframe, ex(n) is used to update the long term filter state (adaptive codebook). The synthesis filter is a direct form 10th order LPC all-pole filter. The LPC coefficients are coded once per 30 msec frame and updated in each 7.5 msec subframe through interpolation. The excitation parameters are also updated in each 7.5 msec subframe. The number of samples in a subframe, N, is 60 at an 8 kHz sampling rate. The "pitch" prefilter and spectral postfilter will be discussed below.

Table 1 shows the bit allocations for the 4.8 kbps VSELP coder. The 10 LPC coefficients are coded using scalar quantization of the reflection coefficients. An energy term, $R_q(0)$, which represents the average speech energy per frame is also coded once per frame. To accommodate noninteger values of the long term predictor delay, eight bits are used to code L. A polyphase

FIR interpolating filter generates the excitation vectors for noninteger delays [5]. The two excitation gains are vector quantized to 7 bits (GS-P0 code) per subframe.

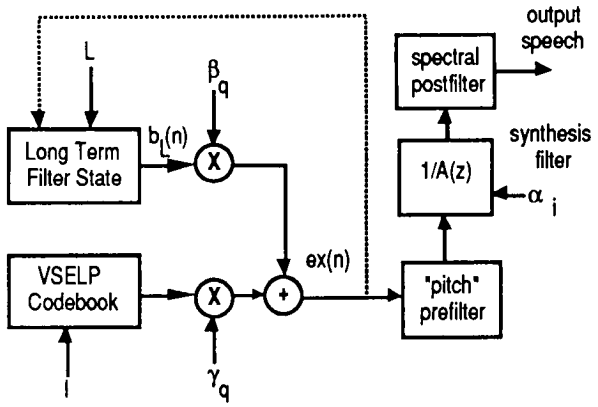

Figure 1 – VSELP Speech Decoder

PARAMETER	BITS/SUBFRAME	BITS/FRAME
LPC coefficients		37
energy - $R_q(0)$		5
excitation code - I	10	40
lag - L	8	32
GS-P0 code	7	28
<synch, parity>		2
TOTAL	25	144

Table 1 – Bit Allocations for 4.8 kbps

VSELP CODEBOOK STRUCTURE

The coder uses a single VSELP excitation codebook, which contains 2^M codevectors. These are constructed from a set of M basis vectors, where M = 10 for the 4.8 kbps coder. Defining $v_m(n)$ as the m^{th} basis vector and $u_i(n)$ as the i^{th} codevector, each from the VSELP codebook, then:

$$u_i(n) = \sum_{m=1}^{M} \theta_{im} v_m(n) \qquad (1)$$

where $0 \le i \le 2^M-1$ and $0 \le n \le N-1$.

In other words, each codevector in the codebook is constructed as a linear combination of the M basis vectors. The linear combinations are defined by the θ parameters. θ_{im} is defined as:

$\theta_{im} = +1$ if bit m of codeword i = 1

$\theta_{im} = -1$ if bit m of codeword i = 0

Note that if we complement all the bits in codeword i, the corresponding codevector is the negative of codevector i. Therefore, for every codevector, its negative is also a codevector in the codebook. These pairs are called complementary codevectors since the corresponding codewords are complements of each other.

The excitation codewords for the VSELP coder are more robust to bit errors than the excitation codewords for random codebooks. A single bit error in a VSELP codeword changes the sign of only one of the basis vectors. The resulting codevector is still similar to the desired codevector.

SELECTION OF EXCITATION VECTORS

Figure 2 is a block diagram which shows the process used to select the two codebook indices L and I. These excitation parameters are computed every subframe.

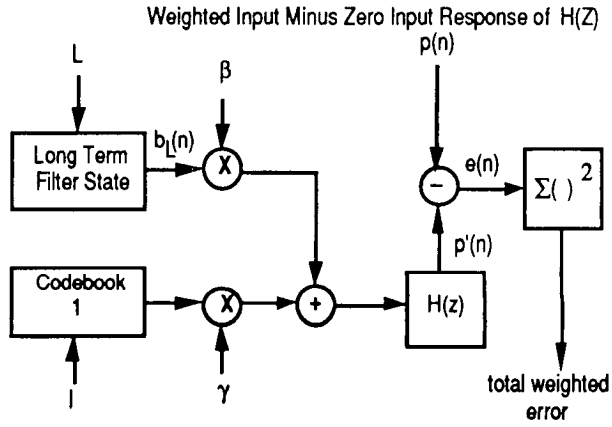

Figure 2 – Excitation Codeword Selection

H(z) is the bandwidth expanded synthesis filter, $H(z) = 1/A(z/\lambda)$, where λ is the noise weighting factor. Signal p(n) is the perceptually weighted (with noise weighting factor λ) input speech for the subframe with the zero input response of bandwidth expanded synthesis filter (H(z)) subtracted out [6].

The two excitation vectors are selected sequentially, one from each of the two excitation codebooks (adaptive codebook and VSELP

codebook). Each codebook search attempts to find the codevector which minimizes the total weighted error.

Although the codevectors are chosen sequentially, the gain of excitation vector chosen from the adaptive codebook is left "floating". The adaptive codebook is searched first, assuming gain γ is zero. The VSELP codebook is searched next, with optimal γ and β used for each codevector being evaluated. This joint optimization can be achieved by orthogonalizing each weighted (filtered) codevector to the weighted excitation vector selected from the adaptive codebook, prior to the VSELP codebook search. While this task seems impractical in general, for VSELP codebook it reduces to orthogonalizing only the weighted basis vectors.

The adaptive codebook is searched first for an index L which minimizes:

$$E'_L = \sum_{n=0}^{N-1} \left(p(n) - \beta' \, b'_L(n)\right)^2 \qquad (2)$$

where $b'_L(n)$ is the zero state response of H(z) to $b_L(n)$ and where β' is optimal for each codebook index L.

To search the VSELP codebook, the zero state response of each codevector to H(z) must be computed. From the definition of the VSELP codebook (1), filtered codevector $f_i(n)$ can be expressed as:

$$f_i(n) = \sum_{m=1}^{M} \theta_{im} \, q_m(n) \qquad (3)$$

where $q_m(n)$ is the zero state response of H(z) to basis vector $v_m(n)$, for $0 \le n \le N-1$.

The orthogonalized filtered codevectors can now be expressed as:

$$f'_i(n) = \sum_{m=1}^{M} \theta_{im} \, q'_m(n) \qquad (4)$$

for $0 \le i \le 2^M-1$ and $0 \le n \le N-1$. Thus $q'_m(n)$ is $q_m(n)$ with the component correlated to $b'_L(n)$ removed. The codebook search procedure now finds the codeword i which minimizes:

$$E'_i = \sum_{n=0}^{N-1} \left(p(n) - \gamma' \, f'_i(n)\right)^2 \qquad (5)$$

where γ' is optimal for each codevector i. Once we have filtered and orthogonalized the basis vectors, the VSELP codebook search is initiated. Defining:

$$C_i = \sum_{n=0}^{N-1} f'_i(n) \, p(n) \qquad (6)$$

and

$$G_i = \sum_{n=0}^{N-1} \left(f'_i(n)\right)^2 \qquad (7)$$

then the codevector which maximizes:

$$\frac{(C_i)^2}{G_i} \qquad (8)$$

is chosen. The search procedure evaluates (8) for each codevector. Using properties of the VSELP codebook structure, the computations required for computing C_i and G_i can be greatly simplified. Defining:

$$R_m = 2 \sum_{n=0}^{N-1} q'_m(n) \, p(n) \qquad 1 \le m \le M \qquad (9)$$

and

$$D_{mj} = 4 \sum_{n=0}^{N-1} q'_m(n) \, q'_j(n) \qquad 1 \le m \le j \le M \qquad (10)$$

C_i can be expressed as:

$$C_i = \frac{1}{2} \sum_{m=1}^{M} \theta_{im} R_m \qquad (11)$$

and G_i is given by:

$$G_i = \frac{1}{2} \sum_{j=2}^{M} \sum_{m=1}^{j-1} \theta_{im} \, \theta_{ij} \, D_{mj} + \frac{1}{4} \sum_{j=1}^{M} D_{jj} \qquad (12)$$

Assuming that codeword u differs from codeword i in only one bit position, say position v such that $\theta_{uv} = -\theta_{iv}$ and $\theta_{um} = \theta_{im}$ for $m \ne v$ then:

$$C_u = C_i + \theta_{uv} \, R_v \qquad (13)$$

and

$$G_u = G_i + \sum_{j=1}^{v-1} \theta_{uj} \, \theta_{uv} \, D_{jv} + \sum_{j=v+1}^{M} \theta_{uj} \, \theta_{uv} \, D_{vj} \qquad (14)$$

If the codebook search is structured such that each successive codeword evaluated differs from the previous codeword in only one bit position, then (13) and (14) can be used to update C_i and

G_i in a very efficient manner. Sequencing of the codewords in this manner is accomplished using a binary Gray code.

Note that complementary codewords will have equivalent values for (8). Therefore only half of the codevectors need to be evaluated. Once the codevector which maximizes (8) is found, the sign of the corresponding C_i will determine whether the selected codevector or its negative will yield a positive gain. If C_i is positive then i is the selected codeword; if C_i is negative then the one's complement of i is selected as the codeword.

QUANTIZATION OF EXCITATION GAINS

The quantization of the excitation gains consists of two stages. The first stage codes the average speech energy once per frame. The quantized value of this energy, $R_q(0)$, is specified with five bits, using 2 dB quantization steps for 64 dB of dynamic range. In the second stage, a GS-P0 code is selected every subframe. This code, when taken in conjunction with $R_q(0)$ and the state of the speech decoder, determines the excitation gains for the subframe. The selection of the GS-P0 code takes place after the two excitation vectors, L and I, have been chosen.

The following definitions are used to determine the GS-P0 code. The combined excitation function, ex(n), is given by:

$$ex(n) = \beta\, c_0(n) + \gamma\, c_1(n) \qquad 0 \le n \le N-1 \quad (15)$$

where:

$c_0(n)$ is the long term prediction vector, $b_L(n)$

$c_1(n)$ is the codevector selected from the VSELP codebook, $u_I(n)$

The energy in each excitation vector is given by:

$$R_x(k) = \sum_{n=0}^{N-1} c^2_k(n) \qquad k = 0,1 \qquad (16)$$

Let RS be the approximate residual energy at a given subframe. RS is a function of N, $R_q(0)$, and the normalized prediction gain of the LPC filter. It is defined by:

$$RS = N\, R_q(0) \prod_{i=1}^{N_p} (1-r_i^2) \qquad (17)$$

where r_i is the i^{th} reflection coefficient corresponding to the set of direct form filter coefficients (α_i's) for the subframe. GS, the energy offset , is a coded parameter which refines the estimated value of RS. R, the approximate total subframe excitation energy, is defined as:

$$R = GS\ RS \qquad (18)$$

P0, the approximate energy contribution of the long term prediction vector as a fraction of the total excitation energy at a subframe, is defined to be:

$$P0 = \frac{\beta^2 R_x(0)}{R} \qquad \text{where } 0 \le P0 \le 1 \quad (19)$$

Thus β and γ are replaced by two new parameters: GS and P0. The transformations relating β and γ to GS and P0 are given by:

$$\beta = \sqrt{\frac{RS\ GS\ P0}{R_x(0)}} \qquad (20)$$

$$\gamma = \sqrt{\frac{RS\ GS\ (1-P0)}{R_x(1)}} \qquad (21)$$

The GS-P0 pair is vector quantized using a codebook of 128 vectors. The codebook was designed using the LBG algorithm [7], using the normalized weighted error as the distortion criterion. Figure 3 shows the distribution of the GS-P0 codebook vectors.

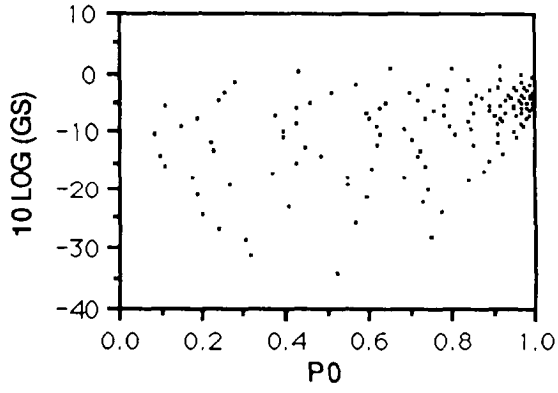

Figure 3 – P0 vs GS in dB for gain codebook

The vector, which minimizes the total weighted error energy for the subframe, is chosen from the GS-P0 codebook. The codebook search procedure requires only five multiply-accumulates per vector evaluation.

This technique of quantizing the gains has a number of advantages. First, the coding is efficient. The coding of the energy once per

frame solves the dynamic range issue. The gain quantization performs equally well at all signal levels within the range of the $R_q(0)$ quantizer. With the average energy factored out, the two gains can be vector quantized efficiently. In minimizing the weighted error, the vector quantizer takes into account the correlation between the two weighted excitation vectors. Second, the values of GS and P0 are well behaved as can be seen in Figure 3. Whereas the optimal value for β, the adaptive codebook gain, can occasionally get very large, P0 is bounded by 0 and 1. Error propagation effects are also greatly reduced by this quantization scheme. Since the energies in the excitation vectors are used to normalize the excitation gains, previous channel errors affecting the energy in the adaptive codebook vector have very little effect on the decoded speech energy. Channel errors in the LPC coefficients are also automatically compensated for at the decoder in calculating the excitation gains. In fact as long as the code for the average frame energy, $R_q(0)$, is received correctly, the speech energy at the decoder will not be much greater than the desired energy (see Figure 3 for range of GS) and no "blasting" will occur.

OPTIMIZATION OF BASIS VECTORS

The basis vectors defining the VSELP codebook are optimized over a training database. The optimization criterion is the minimization of the total normalized weighted error. The normalized weighted error for each subframe can be expressed as a function of individual samples of each of the 10 basis vectors from the VSELP excitation codebook, given I, $b_L(n)$, p(n), the excitation gains, and the impulse response of H(z) for each subframe of the training data. The optimal basis vectors are computed by solving the 600 (10 basis vectors, 60 samples per vector) simultaneous equations which result from taking the partial derivatives of the total normalized weighted error function with respect to each sample of each basis vector and setting them equal to zero. Since the coder subframes are not independent, this procedure is iterated in a closed loop fashion. Figure 4 shows the improvement in weighted segmental SNR for each iteration.

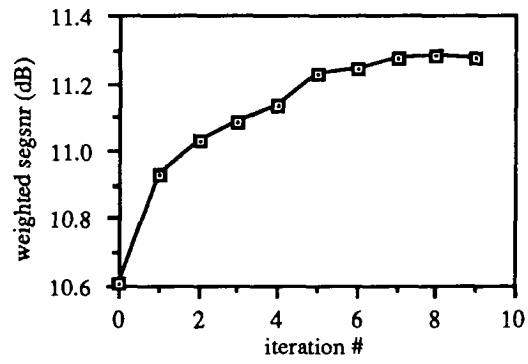

Figure 4 – Basis Vector Optimization

Initially the basis vectors are populated with random Gaussian sequences (iteration 0) which yields a weighted segmental SNR of 10.61 dB. The weighted segmental SNR increases to 11.28 dB after nine iterations. The subjective quality improvement due to the optimization of the basis vectors is significant. The objective as well as subjective improvements are retained for speech data outside the training data base.

ADAPTIVE PRE AND POSTFILTERING

The speech decoder creates the combined excitation signal, ex(n), from the long term filter state and the VSELP excitation codebook. The combined excitation is then processed by an adaptive "pitch" prefilter to enhance the periodicity of the excitation signal (see Figure 1). Following the adaptive pitch prefilter, the prefiltered excitation is applied to the LPC synthesis filter. After reconstructing the speech signal with the synthesis filter, an adaptive spectral postfilter is applied to further enhance the quality of the reconstructed speech. The pitch prefilter transfer function used is given by:

$$H_p(z) = \frac{1}{1 - \xi z^{-L}} \qquad (22)$$

where $\xi = \varepsilon \, \mathrm{Min}[\, \beta, \sqrt{P0}\,]$ and $\varepsilon = 0.4$ (23)

Note that the periodicity enhancement is performed on the synthetic residual in contrast to pitch postfiltering which performs the enhancement on the synthesized speech waveform [8]. This significantly reduces artifacts in the reconstructed speech due to waveform discontinuities which pitch postfiltering sometimes introduces. For

noninteger values of L, a polyphase FIR filter is used to compute the fractionally delayed excitation samples. Finally to ensure unity power gain between the input and the output of the pitch prefilter, a gain scale factor is calculated to scale the pitch prefiltered excitation prior to applying it the LPC synthesis filter.

The form of the adaptive spectral postfilter used is:

$$H_S(z) = \frac{1 - \sum_{i=1}^{10} \eta_i z^{-i}}{1 - \sum_{i=1}^{10} \nu^i \alpha_i z^{-i}} \qquad 0 \le \nu \le 1 \quad (24)$$

where the α_i's are the coefficients of the synthesis filter. To derive the numerator, the $\nu^i \alpha_i$ coefficients are converted to the autocorrelation domain (the autocorrelation of the impulse response of the all pole filter corresponding to the denominator of (24) is calculated for lags 0 through 10). A binomial window is then applied to the autocorrelation sequence [9] and the numerator polynomial coefficients are calculated from the modified autocorrelation sequence via the Levinson recursion. This postfilter is similar to that proposed by Gersho and Chen [10]. However, the use of the autocorrelation domain windowing results in a frequency response for the numerator that tracks the general shape and slope of the denominator's frequency response more closely. To increase postfiltered speech "brightness", an additional first order filter is used of the form:

$$H_B(z) = 1 - u z^{-1} \qquad (25)$$

The following postfilter parameter values are used: ν=.8, B_{eq}=1200 Hz, u=0.4. Note that B_{eq} is the bandwidth expansion factor which specifies the degree of smoothing which is performed on the denominator to generate the numerator.

As in the case of the pitch prefilter, a method of automatic gain control is needed to ensure unity gain through the spectral postfilter. A scale factor is computed for the subframe in the same manner as was done for the pitch prefilter. In the case of the spectral postfilter, this scale factor is not used directly. To avoid discontinuities in the output waveform, the scale factor is passed through a first order low pass filter before being applied to the postfilter output.

CONCLUSIONS

A high quality 4.8 kbps speech coder has been described. The complexity of the coder is low enough so that it can be implemented on a single DSP device such as the Motorola DSP56000. In addition to its very high speech quality, the parameters are inherently robust to channel errors and can be error protected very efficiently.

REFERENCES

[1] D. P. Kemp, R. A. Sueda and T. E. Tremain, "An Evaluation of 4800 bps Voice Coders", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, May 1989.

[2] M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): High Quality Speech at Very Low Bit Rates", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 937-940, March 1985.

[3] I. Gerson and M. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP)", *IEEE Workshop on Speech Coding for Telecommunications*, pp. 66-68, September 1989.

[4] W. B. Kleijn, D. J. Krasinski and R. H. Ketchun, "Improved Speech Quality and Efficient Vector Quantization in SELP", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 155-158, April 1988.

[5] P. Kroon and B.S. Atal, "Pitch Predictors with High Temporal Resolution", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, April 1990.

[6] G. Davidson and A. Gersho, "Complexity Reduction Methods for Vector Excitation Coding", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 3055-3058, May 1986.

[7] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design", *IEEE Trans. Comm.*, vol. COM-28, pp. 84-95, Jan. 1980.

[8] P. Kroon and E. F. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s", *IEEE J. Select. Areas Commun.*, vol. SAC-6, No. 2, February 1988.

[9] Y. Tohkura, F. Itakura and S. Hashimoto, "Spectral Smoothing Technique in PARCOR Speech Analysis-Synthesis", *IEEE Trans. Acoustics Speech and Signal Processing*, vol. ASSP-26, pp. 587-596, Dec. 1978.

[10] J. Chen and A. Gersho, "Real-Time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering", *Proc. IEEE Int. Conf. on Acoustics Speech and Signal Processing*, pp. 51.3.1-51.3.4, April 1987.