IMPROVING VSELP CODING USING TRUNCATED PERCEPTUAL WEIGHTING FILTER AND IN-LOOP QUANTIZATION

C. F. Chan, W. H. Lau, S. P. Chui, and F. L. Hui

Department of Electronic Engineering
City Polytechnic of Hong Kong

Abstract

This paper proposes a method to reduce the complexity of VSELP coding and introduces a more efficient quantization scheme for the short-term predictor coefficients. It is shown that, by using a reduced-order perceptual weighting filter, the subjective speech quality was retained while the complexity in codebook searching is reduced tremendously. An efficient analysis/quantization scheme which quantizes the reflection coefficients inside the analysis loop was developed. The new scheme does not alter the coding format of the VSELP standard and obtains a lower quantization distortion than the original scalar quantization scheme.

Introduction

Vector Sum Excited Linear Predictive (VSELP) coding is a medium bit rate coding technique developed by Motorola in 1990[1]. It is based on the analysis-by-synthesis principle utilized in stochastic coding[2]. The novelty of VSELP coding is the excitation codebook which consists of codevectors constructed from a small set of basis vectors. The combination of basis vectors is governed by a special type of binary code so that sequential codebook search can be performed in great speed. Recently real-time implementation of the VSELP coder on Motorola DSP56156 has been demonstrated[3]. The objective of our research is to further improve the speed of codebook search so that real-time implementation of the VSELP coder on a lower speed DSP is possible. We also work on developing a better quantization scheme for quantizing the short-term filter coefficients without altering the coding format of the VSELP standard.

In analysis-by-synthesis speech coding, the objective of closed-loop minimization is to minimize the perceptually weighted mean squared error between the input speech $\boldsymbol{x}$ and the synthetic speech $\boldsymbol{s}$, i.e., $E = (\boldsymbol{x}-\boldsymbol{s})^T\boldsymbol{H}_w^T\boldsymbol{H}_w(\boldsymbol{x}-\boldsymbol{s})$, where $\boldsymbol{H}_w$ is the impulse response matrix of the weighting filter[2]. The perceptual weighting filter is necessary because human reception system tends to tolerate more error in the high signal energy regions in the speech spectrum. A widely used weighting filter is defined as $W(z) = \dfrac{A(z)}{A(z/\gamma)}$ where $A(z)$ is the inverse linear predictive filter, and γ is the weighting factor with a value between 0 and 1. Due to the memory hand-over in the synthesis filter, the synthetic speech is expressed as

$$\boldsymbol{s} = \boldsymbol{A}^{-1}(g\boldsymbol{u} - \boldsymbol{A}_o\boldsymbol{s}_o) \qquad (1)$$

where $A = \begin{bmatrix} 1 & 0\cdot & \cdot & \cdot & \cdot & 0 \\ a_1 & 1 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_p & \cdot & a_1 & 1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & a_p & \cdot & a_1 & 1 \end{bmatrix}$ is an $M \times M$ matrix, $A_o = \begin{bmatrix} a_p & a_{p-1} & \cdot & \cdot & a_1 \\ 0 & a_p & \cdot & \cdot & a_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 0 & a_p \\ \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$ is an $M \times P$ matrix, u and g are the excitation vector and the corresponding gain respectively, and s_o is the last P samples of the synthetic speech from the previous subframe. Similarly, the input speech can also be expressed as

$$x = A^{-1}(v - A_o x_o) \tag{2}$$

where v is the residual vector and x_o is the last P samples of the input speech from the previous subframe. In standard VSELP coder, there are 4 subframes in each analysis frame, the subframe size M is equal to 40 samples and the LPC order P is equal to 10. By using similar notation, the impulse response matrix of the perceptual weighting filter can be expressed as $H_w = \hat{A}^{-1} A$, where $\hat{A}$ is the impulse response matrix due to $A(z/\gamma)$. Now E can be expressed as

$$E = (v - gu - A_o x_o + A_o s_o)^T \hat{A}^{-T} \hat{A}^{-1} (v - gu - A_o x_o + A_o s_o) \tag{3}$$

Therefore, the analysis-by-synthesis structure can be re-organized to a different form as shown in Fig. 1. This analysis structure can be considered as the minimization of the weighted mean squared error between the residual signal and the excitation signal. The weighting filter in this case is $\dfrac{1}{A(z/\gamma)}$.

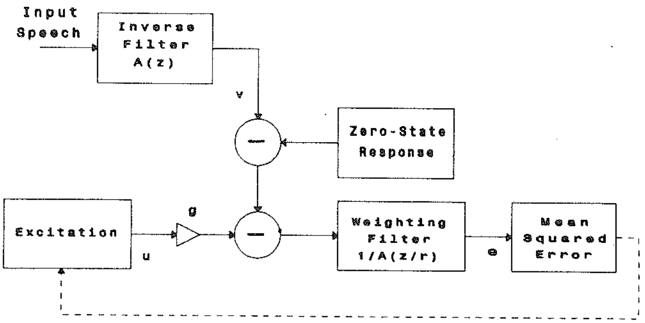

Fig. 1 Modified Analysis-by-Synthesis Structure

It is well known in CELP coding that the index to the optimum codevector in the codebook is computed as

$$i_{opt} = \underset{\forall u \,\in\, codebook}{argmax} \left[\frac{(u^T \hat{A}^{-T} \hat{A}^{-1} \bar{v})^2}{u^T \hat{A}^{-T} \hat{A}^{-1} u} \right] \tag{4}$$

where $\tilde{v} = v - A_0(x_0 - s_0)$. Practically, since $s_0 \approx x_0$, $\tilde{v} \approx v$. This assumption allows us to ignore the zero-state response and hence saves us from performing speech synthesis in the analyzer.

Truncated Perceptual Weighting Filter

The burden of high computational load in CELP coding is very much due to the weighting filter. If full weighting is applied $\gamma = 0$ and $\hat{A} = I$, and if no weighting is applied $\gamma = 1$ and $\hat{A} = A$. Obviously, for the case $\gamma = 0$, the computation will be reduced significantly. Conventional CELP coders set γ to a value between 0.8 and 0.9 which would not help to reduce the computation. The question we want to ask now is how can we modify $\hat{A}$ such that a lot of computations can be saved while the synthetic speech quality will not be affected subjectively. Since $\hat{A}$ has only $P+1$ non-zero diagonals, if the number of non-zero diagonals is reduced, a lot of computations can be saved. Obviously, this is achieved if a large part of coefficients of the weighting filter is truncated to zero, or, in other words, the order of the weighting filter is reduced. Since the idea of perceptual weighting is to increase the error weighting in high signal energy regions of speech spectrum, if the truncated filter can perform similar functions, the subjective quality of the synthetic speech will be retained. Fig. 2 shows the frequency responses of a typical LPC filter and the corresponding full weighting filter with $\gamma = 0.9$. The responses of two truncated filters with one and two coefficients, respectively, are also shown. We can see that the responses of the truncated filters actually resemble the trend of the response of the full perceptual weighting filter. We would expect that the subjective speech quality due to the truncation would not be affected significantly.

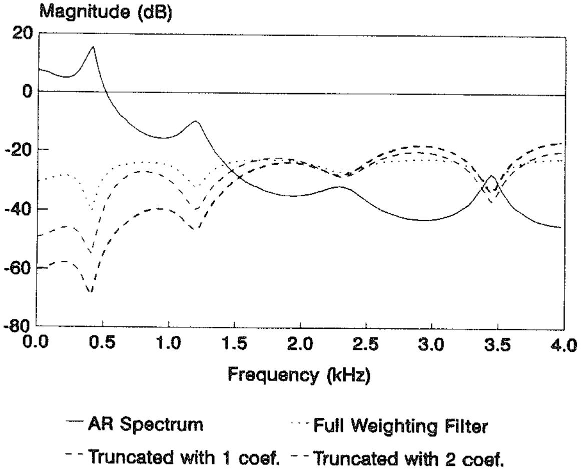

Fig. 2 Frequency Responses for Various Weighting Filters

In our work this hypothesis was tested objectively and subjectively using real speech signals.

Table I shows the average segmental SNR for cases of full weighting filter and truncated filters with one and two coefficients respectively. The results were achieved using a standard 8kbit/s VSELP coder adopted by TIA. We see that even using a truncated filter with one coefficient, the drop in SNR is merely 1.5 dB. Informal listening test has revealed that the degradation in speech quality as a result of this truncation is barely noticeable.

Weighting Filters	Full: 10 Coefficients	Truncated: 2 Coefficients	Truncated: 1 Coefficient
Average Seg. SNR (dB)	13.2	12.1	11.7

Table I

Complexity Reduction Due to Truncation

In the VSELP coder there are one adaptive codebook due to the long-term predictor and two fixed codebooks. The optimum codevectors from all three codebooks can be determined sequentially by repeated applications of Eqn. 4 with the orthogonized codevector. There are 128 codevectors from each fixed codebook and they are actually determined from 7 basis vectors as $\boldsymbol{u} = \boldsymbol{Qb}$ where $\boldsymbol{Q} = [\boldsymbol{q}_1 \boldsymbol{q}_2 \cdot\cdot \boldsymbol{q}_7]$ with its column vectors being the stochastic basis vectors, and $\boldsymbol{b}$ is a binary vector with its element being either 1 or -1. This codebook can be searched very efficiently by sequencing $\boldsymbol{b}$ in Gray code pattern, i.e., only 1 bit is changed in consecutive search. In order to calculate the optimum codeword index, the term that needed to be maximized now becomes $\frac{(\boldsymbol{b}^T\boldsymbol{d})^2}{\boldsymbol{b}^T\boldsymbol{R}\boldsymbol{b}}$ where $\boldsymbol{d} = \boldsymbol{Q}^T\hat{\boldsymbol{A}}^{-T}\hat{\boldsymbol{A}}^{-1}\boldsymbol{v}$, and $\boldsymbol{R} = \boldsymbol{Q}^T\hat{\boldsymbol{A}}^{-T}\hat{\boldsymbol{A}}^{-1}\boldsymbol{Q}$ is a 7x7 symmetric matrix. Note that most part of the computation can be done outside the codebook searching loop. Nevertheless, even using Gray code searching strategy, the energy term in the denominator still requires 7 mult/add operations for each codevector search, and the cross-correlation term in the numerator requires 1 addition operation; provided that $\boldsymbol{d}$ has been calculated before the codebook search. If a truncated filter with only one coefficient is used, the complexity in calculating $\boldsymbol{d}$ off line for each subframe is $O(9M)$. Note that to calculate $\hat{\boldsymbol{A}}^{-T}\hat{\boldsymbol{A}}^{-1}\boldsymbol{v}$ requires only $2M$ mult/add operations by using backward and forward filtering. Meanwhile, in the VSELP coding standard, the short-term predictor is represented by 10 reflection coefficients, i.e., k_1 to k_{10}, and each coefficient is individually scalar quantized. Because k_1 is quantized to 64 levels, and as the result of this quantization and truncation process the energy term can be pre-calculated and stored in a table with size 64x64 (note that Gray code produces complementary codewords and only half of the energy terms are needed to be stored). Therefore, only 2 mult/add operations are absolutely needed for each codeword search; this is a significant improvement over standard VSELP algorithm.

In-Loop Quantization of Reflection Coefficients

In our work a method to quantize the reflection coefficients inside the computation loop of the lattice algorithm is used. Define the auto- and cross-correlation matrices as

$$\alpha_m(i,j) = \sum_{n=P}^{N-1} f_m(n-i) f_m(n-j) \qquad 0 \le i,j \le P \tag{5}$$

$$\beta_m(i,j) = \sum_{n=P}^{N-1} b_m(n-i) b_m(n-j) \qquad 0 \le i,j \le P \tag{6}$$

$$\phi_m(i,j) = \sum_{n=P}^{N-1} f_m(n-i) b_m(n-j) \qquad 0 \le i,j \le P \tag{7}$$

where $f_m(n)$ and $b_m(n)$ are the forward and backward prediction errors in the m^{th} lattice stage, respectively. By using the lattice recursive structure, we have

$$\alpha_{m+1}(i,j) = \alpha_m(i,j) + k_{m+1}^2 \beta_m(i+1,j+1) - k_{m+1}[\phi_m(i,j+1) + \phi_m(j,i+1)] \tag{8}$$

$$\beta_{m+1}(i,j) = \beta_m(i+1,j+1) + k_{m+1}^2 \alpha_m(i,j) - k_{m+1}[\phi_m(i,j+1) + \phi_m(j,i+1)] \tag{9}$$

$$\phi_{m+1}(i,j) = \phi_m(i,j+1) + k_{m+1}^2 \phi_m(j,i+1) - k_{m+1}[\alpha_m(i,j) + \beta_m(i+1,j+1)] \tag{10}$$

The objective of lattice analysis is to minimize the sum of the forward and backward prediction energy, this is achieved if

$$k_{m+1} = \frac{2\phi_m(0,1)}{\alpha_m(0,0) + \beta_m(1,1)} \tag{11}$$

It is well known that if k_{m+1} is calculated using Eqn. 11, the prediction error vector and the input speech vector are orthogonal and k_{m+1} can be fixed for the analysis of subsequent stages. Practically, this is not the case because k_{m+1} has to be coarsely quantized to low bit rates. In standard VSELP coder, 10 reflection coefficients are firstly calculated using Eqn 11 and then each reflection coefficient is individually scalar quantized. The total number of bits required in each analysis frame is 38 bits.

We are proposing an in-loop quantization scheme which, at each stage of lattice analysis, a trained quantization table is searched and a codeword that achieves the lowest prediction error is selected, i.e.,

$$i_{m+1} = \underset{k_{m+1} \in C_{m+1}}{argmin} \left\{ [1 + k_{m+1}^2][\alpha_m(0,0) + \beta_m(1,1)] - 4k_{m+1}\phi_m(0,1) \right\} \tag{12}$$

where C_{m+1} is the quantization codebook for the $(m+1)^{th}$ stage reflection coefficient. Because the quantization error in the present stage will affect the determination of reflection coefficients in the following stages, the optimum set of quantized coefficients should be determined by exhaustively testing all combinations of codewords in the quantization tables. However, this will be too computationally expensive. In our work, two suboptimum searching schemes under the context of in-loop quantization are examined. In the first scheme, each reflection coefficient is quantized independently. The best codeword is obtained by searching the codebook using Eqn. 12, then from Eqns. 8, 9, and 10 the auto- and cross-correlations are updated using the quantized coefficient. These steps are repeated until all reflection coefficients have been fixed. Since division is no longer needed, the complexity of this in-loop scheme is actually lower than the off-loop scalar quantization scheme. In the second scheme, two reflection coefficients are quantized in a single step by searching two consecutive tables simultaneously. However, the complexity of this scheme is much higher. Note that the standard coding format of the VSELP coder will not be altered by using these in-loop quantization schemes, however, the distortion level is lower. Fig. 3 illustrates plots of the averaged likelihood ratio distortion against coding rates for the scalar and the two in-loop quantization schemes.

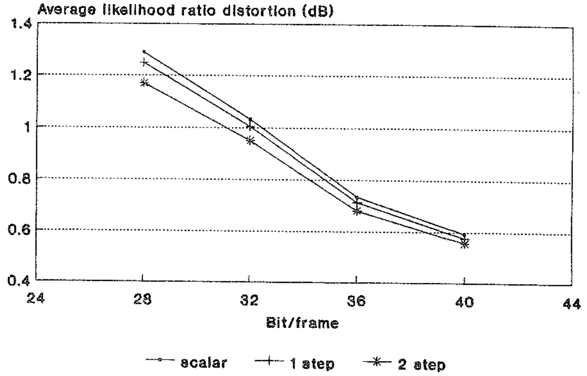

Fig. 3 Plots of Quantization Distortion Against Coding Rates

We see that the improvement on quantization distortion is quite significant. Experimentally, by applying the one-step in-loop quantization scheme on the 8 kbit/s VSELP coder we are able to achieve, in average, about 0.85 dB increase in segmental SNR. In future works more improvements could be achieved by using trellis search codebooks.

Conclusion

It is demonstrated that a truncated perceptual weighting filter can be used in VSELP coding without lost of subjective speech quality while codebook searching can be speed up significantly. It is also demonstrated that a simple in-loop scheme for the quantization of short-term predictor coefficients can be used to reduce the quantization distortion with no increase in complexity.

References

1. I. A. Gerson and M. A. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps," IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp.461-464, 1990.
2. M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction Coding (CELPC); High Quality Speech at Very Low Bit Rates," IEEE Int. Conf. on Acoustic, Speech, and Signal Processing, pp.25.1.1-25.1.4, 1985.
3. M. H. Sunwoo and S. Park, "Principles of Vector-Sum Excited Linear Predictive (VSELP) Speech Coder and Its Implementation on the DSP56156," Motorola Digital Signal Processors, 1991.