LOW COMPLEXITY VQ FOR MULTI-TAP PITCH PREDICTOR CODING

Jayesh Patel
DSP Software Engineering, Inc.
Bedford, MA 01730
Email: jayesh@dspse.com

ABSTRACT

Pitch predictors are successfully used in Linear Prediction Analysis-by-Synthesis (LPAS) coders to model periodicity in speech. The various structures of pitch predictors are investigated and used in LPAS coders. In most of the low bit-rate LPAS coder design, single-tap or three-tap pitch are commonly used. Higher prediction gain can be achieved by using additional taps. 5-tap pitch predictor is rarely used in low bit-rate speech coder because of high complexity and bandwidth requirement in encoding additional tap gains. This paper describes a technique for reducing the complexity and bandwidth requirement for 5-tap pitch predictor.

1. INTRODUCTION

LPAS coders have given new dimension to the medium-rate and low-bit rate speech coding research. Various forms of LPAS coders are being used in applications like secure telephone, cellular phones, answering machines, voice mails, digital memo recorders etc. LPAS coders are based on a speech production model and fall into category between waveform coders and parametric coders (vocoders). It consists [1] of two cascaded time varying filters, short-term filter and long-term filter (pitch predictor or adaptive codebook) and stochastic codebook. A short-term predictor is used in removing sample-to-sample correlation in the speech. It estimates the short-time spectral envelope of the speech. A long-term predictor (pitch predictor) is used in removing long-term correlation exhibits during voiced speech. A stochastic codebook serves as a source excitation. The parameters of short-term filter, long-term filter and stochastic excitation are optimized sequentially using perceptually weighed mean squared error criterion. Most of the current LPAS coders differ in stochastic codebook implementation and the transfer function of pitch predictor.

The pitch predictor has played an important role in success of LPAS coder. Various structure of pitch predictor are studied and can be categorized into:

1. Multi-tap pitch predictor (MTPP)
2. Fractional pitch predictor (FPP)

The level of periodicity in speech is not consistent throughout the frequency spectrum. In most of the speech signal, higher periodicity is observed at lower frequencies than at higher frequencies. MTPP can provide this frequency dependent gain factor [2]. The transfer function of an odd-order MTPP with predictor coefficient's g_k center around delay M is given by:

$$P(z) = \sum_{k=0}^{p-1} g_k z^{-\left(M - \lfloor \frac{p}{2} \rfloor + k\right)}$$

In a conventional single-tap pitch predictor resolution of the delay M in not sufficient for high-pitched speakers (female and child) as it is dependent on the sampling frequency (8kHz). FPP can provide the fractional resolution for the delay. The transfer function of a FPP with integer delay M, fractional delay l, and interpolation factor D, and predictor coefficient g is

$$P(z) = g z^{-\left(M + \frac{l}{D}\right)}$$

Fractional delays in the FPP are implemented using polyphase filters [2]. For a single-tap pitch predictor $D = 0$ and $p = 1$ in MTPP and $l = 0$ in FPP. Use of additional taps and fractional delay in pitch predictor boosts the speech quality of the coder. The speech quality, complexity and bit-rate are a function of p in MTPP and l and D in FPP. The higher the values of p or (l and D), the higher the complexity, the bit-rate and better the speech quality.

The performance of multi-tap pitch predictor is close to fractional pitch predictor [2][3]. The disadvantage of FPP is it requires higher computational complexity at the encoder as well as at the decoder end due to an interpolation procedure. Single-tap or three-tap pitch predictors are widely used in LPAS based coder design based tradeoff between quality, complexity and bandwidth. Though higher-tap ($p = 5$) pitch predictors give better performance (objective and subjective) than 1-tap or 3-tap, they are rarely used in LPAS coder design due increase in complexity and bandwidth.

The parameter set of pitch predictor consist of delay M and predictor coefficients g_k. The accuracy of delay M and predictor coefficients depend on the predictor order [7]. Accuracy of delay M is more important in single-tap pitch predictor implementation while accuracy of predictor coefficients are more important in MTPP implementation. Hence in MTPP configuration delay M can updated at a slower rate than as compared to a single-tap pitch predictor [3]. Due to more predictor coefficients in MTPP than in single-tap pitch predictor, scalar quantization of predictor coefficients in MTPP requires extra bits. Also, higher complexity is required in estimating the predictor coefficients

in MTPP configuration. By Vector Quantizing (VQ) the predictor coefficients the overall bandwidth requirement of MTPP can match to bandwidth requirement of single-tap pitch predictor. Various issues in VQ design of predictor coefficients should be resolved for cost effective implementation of MTPP. This paper addresses the complexity (time and space) issue in VQ design for 5-tap pitch predictor coding.

In the following, section 2 describes conventional VQ design of predictor coefficients and joint optimization of delay and VQ index search in closed-loop. In section 3, we describe the low complexity VQ design and section 4 and 5 shows the comparative performance of the proposed design as compared to conventional VQ design.

2. VECTOR QUANTIZATION OF $\mathbf{G_K}$

Vector quantization [3] of the multiple coefficient's $\mathbf{g_k}$ in MTPP is necessary to reduce the bandwidth requirement. Predictor coefficients can be optimized using open-loop fashion or closed-loop fashion. A closed-loop optimization performs better than open-loop, but it is more computationally expensive than open-loop optimization. The computational complexity of closed-loop optimization can be reduced using Restricted Pitch Deviation Coding (RPDC) [6] approach. In this method pitch analysis is done in two stages

- First, an open-loop pitch M_{olp} is estimated.
- Than closed-loop optimization is done on restricted deviation around M_{olp}.

Better performance can be achived by joint optimization of delay M and predictor coefficients. Using RPDC approach computational requirement of joint optimization can be reduced. The weighted mean squared error (WMSE) [4] is used as a distortion measure in optimization process. Next section describes the procedure.

2.1. Joint Optimization of Parameters M and $\mathbf{g_k}$

Let $r(n)$ be contribution from the adaptive codebook or pitch predictor and let $s_{tv}(n)$ be the target vector and $h(n)$ be the impulse response of the weighted synthesis filter. The error between synthesized speech and target assuming zero contribution from stochastic codebook and 5-tap pitch predictor, is given as.

$$e(n) = s_{tv}(n) - \sum_{j=0}^{j=n} h(n-j) \sum_{k=0}^{k=4} g_k r\,(n-(M-2+k))$$

In matrix notation, with vector length equal to subframe size.

$$\mathbf{e} = \mathbf{s_{tv}} - g_0\mathbf{Hr_0} - g_1\mathbf{Hr_1} - g_2\mathbf{Hr_2} - g_3\mathbf{Hr_3} - g_4\mathbf{Hr_4}$$

Where $\mathbf{H}$ is the impulse response matrix of the weighted synthesis filter. Then total mean squared error is given by

$$E = \mathbf{e^T e} = \mathbf{s_{tv}^T s_{tv}} - 2\mathbf{c_M}^{\mathbf{T}}\mathbf{g}$$

Where $\mathbf{g} = [\ g_0, g_1, g_2, g_3, g_4, \mathbf{g'}\ ]$

$\mathbf{g'} = [\ -0.5{g_0}^2, -0.5{g_1}^2, -0.5{g_2}^2, -0.5{g_3}^2, -0.5{g_4}^2,$
$-g_0g_1, -g_0g_2, -g_0g_3, -g_0g_4, -g_1g_2,$
$-g_1g_3, -g_1g_4, -g_2g_3, -g_2g_4, -g_3g_4\ ]$

$$\begin{aligned}\mathbf{c_M} = [\ & s_{tv}^T\mathbf{Hr_0}, s_{tv}^T\mathbf{Hr_1}, s_{tv}^T\mathbf{Hr_2}, s_{tv}^T\mathbf{Hr_3}, s_{tv}^T\mathbf{Hr_4},\\ & \mathbf{r_0}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_0}, \mathbf{r_1}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_1}, \mathbf{r_2}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_2},\\ & \mathbf{r_3}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_3}, \mathbf{r_4}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_4}, {g_1}^2\mathbf{r_0}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_1},\\ & {g_2}^2\mathbf{r_0}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_2}, {g_3}^2\mathbf{r_0}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_3}, {g_4}^2\mathbf{r_0}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_4}\\ & {g_2}^2\mathbf{r_1}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_2}, {g_3}^2\mathbf{r_1}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_3}, {g_4}^2\mathbf{r_1}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_4}\\ & {g_3}^2\mathbf{r_2}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_3}, {g_4}^2\mathbf{r_2}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_4}, {g_4}^2\mathbf{r_3}^{\mathbf{T}}\mathbf{H}^{\mathbf{T}}\mathbf{Hr_4}\ ]\end{aligned}$$

When the $\mathbf{g}$ vector comes from the stored VQ codebook and distortion measure is minimized over restricted pitch range $[M_{olp}-1, M_{olp}+2]$, then above equation is a function of M and VQ index i.

$$E(N,i) = \mathbf{e^T e} = \mathbf{s_{tv}^T s_{tv}} - 2\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}$$

$$M \in [M_{olp}-1, M_{olp}+2], i \in [0, N]$$

Where N is the size of the codebook. Minimizing $E(M,i)$ is equivalent to maximize $\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}$, inner product of two 20 dimensional vectors. The best combination of M and i which maximize $\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}$ is selected. Mathematically,

$$\underset{(M,i)}{MAX} \quad \{\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}\}$$

$$M \in [M_{olp}-1, M_{olp}+2], i \in [0, N]$$

$\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}$ is used as distortion measure for joint optimization of delay M and predictor coefficients $\mathbf{g_k}$. Since the distortion measure is evaluated for each VQ entry times restricted pitch range, the time complexity for $\mathbf{c_M}^{\mathbf{T}}\mathbf{g_i}$ computation of such design is high. Also, additional time complexity comes from computation of $\mathbf{g_i'}$ vector for each $\mathbf{g_i}$. This additional time complexity can be reduced at an expense of increase in auxiliary table $\mathbf{g_i'}$ storage for each $\mathbf{g_i}$. Such design is practical when there is a leverage on either time or space complexity in the design. But for application like answering machine and digital memo recorders where time and space complexity both are crucial, such method can not be employed.

3. PRODUCT CODE VECTOR QUANTIZATION

Product code Vector Quantization (PCVQ) [5] is a structured VQ method by which an excellent performance-complexity tradeoff can be achieved. It is also known as split-VQ. In this method, the original vector is divided into two or more feature vectors. Each feature vector can be quantized jointly or separately. At the decoder, each feature vector is decoded first and than concatenated to form an approximation to the original vector. Product Code VQ techniques are already applied to spectral (LSF) quantization [6].

In order to resolve the time and space complexity issue, we have employed the Product Code VQ (PCVQ) method for quantizing the 5-tap predictor gains g_k. The $\mathbf{g}$ vector is divided into two feature vector $\mathbf{g_1}$ and $\mathbf{g_2}$. The $\mathbf{g_1}$ and $\mathbf{g_2}$ are quantized using two separate codebooks $\underline{C_1}$ and $\underline{C_2}$.

3.1. Full Search PCVQ

For optimum performance of PCVQ, each possible combination of $\mathbf{g_1}$ and $\mathbf{g_2}$ to make $\mathbf{g}$ is searched in closed-loop configuration. Mathematically,

$$\mathbf{g_{i,j}} = g_{1_i} + g_{2_j} + g_{12_{i,j}}$$

$$\underset{(M,i,j)}{MAX} \quad \{\mathbf{c_M}^{\mathbf{T}} \mathbf{g_{i,j}}\}$$

$$M \in [M_{olp} - 1, M_{olp} + 2], i \in [0, N_1], j \in [0, N_2]$$

$N = N_1 * N_2$. N_1 and N_2 depend on the bits allocated to quantize $\mathbf{g_1}$ and $\mathbf{g_2}$ feature vector.

In our design, $\mathbf{g_1}$ feature vector contains center three elements g_1, g_2 and g_3 of $\mathbf{g}$

$$\mathbf{g_1} = [0, g_1, g_2, g_3, 0, 0, -0.5{g_1}^2, -0.5{g_2}^2, -0.5{g_3}^2, 0, 0, 0, 0, -g_1 g_2, -g_1 g_3, 0, -g_2 g_3, 0, 0]$$

And $\mathbf{g_2}$ feature vector contains outer two elements g_0 and g_4 of $\mathbf{g}$.

$$\mathbf{g_2} = [g_0, 0, 0, 0, g_4, -0.5{g_0}^2, 0, 0, 0, -0.5{g_4}^2, 0, 0, 0, -g_0 g_4, 0, 0, 0, 0, 0, 0]$$

And $\mathbf{g_{12}}$ vector is computed to form $\mathbf{g}$ vector.

$$\mathbf{g_{12}} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -g_0 g_1, -g_0 g_2, -g_0 g_3, 0, 0, 0, -g_1 g_4, 0, -g_2 g_4, -g_3 g_4]$$

A huge savings in space complexity is obtained using this method, but there is a slight increase in time complexity as it requires additional computation for the $\mathbf{g_{12}}$ vector. Hence, there is no savings in time complexity when full search product code VQ is used.

3.2. Sequential Search PCVQ

Sequential search method is used to reduce time complexity PCVQ, In this method each feature vector is quantized sequentially. The procedure is explained below

Stage 1: For all candidates of M, the best index $i = I[M]$ from codebook $\underline{C_1}$ is determined using the same distortion criterion.

$$I[M] = arg \ \underset{i}{MAX} \quad \{\mathbf{c_M}^{\mathbf{T}} \mathbf{g_{1_i}}\}$$

$$M \in [M_{olp} - 1, M_{olp} + 2], i \in [0, N_1]$$

Stage 2: Best combination M, $I[M]$ and index j from codebook $\underline{C_2}$ is selected using the same distortion criterion.

$$\mathbf{g_{I[M],j}} = g_{1_{I[M]}} + g_{2_j} + g_{12_{I[M]j}}$$

$$\underset{(M,I[M],j)}{MAX} \quad \{\mathbf{c_M}^{\mathbf{T}} \mathbf{g_{I[M],j}}\}$$

$$M \in [M_{olp} - 1, M_{olp} + 2], j \in [0, N_2]$$

With sequential PCVQ approach, the distortion measure is evaluated only $(N_1 + N_2)$ times as compared to $(N_1 * N_2)$ times in full search PCVQ and conventional VQ. Also the complexity of distortion measure evaluated N_1 times is less in sequential PCVQ as compared to full search PCVQ. Hence there is considerable saving in time complexity as well as space complexity using sequential PCVQ approach. Next section shows the comparative performance of each method described above.

VQ Method	Total Multiplication	Storage Requirement
Fast Conventional VQ	$N * R * D$	$N * D$
Low Memory Conventional VQ	$N * R * (D + T_x)$	$N * T$
Full Search PCVQ	$N * R * (D + D_{12})$	$N_1 * D_1 + N_2 * D_2$
Sequential Search PCVQ	$N_1 * R * (D_1 + T_{1_x}) + N_2 * R * (D_2 + T_{2_x})$	$N_1 * T_1 + N_2 * T_2$

Table 1. Complexity of VQ design for MTPP

4. COMPARISONS

A comparison based on total multiplication (time complexity) required in calculating the distortion measure and storage requirement (space complexity) among all the four vector quantization techniques for predictor coefficients of MTPP

1. Fast Conventional VQ
2. Low Memory (LM) Conventional VQ
3. Full Search PCVQ
4. Sequential Search PCVQ

is shown in the Table 1. Where,

T =Taps of pitch predictor
D =Length of g vector $= T + T_x$
$T_x = \frac{T(T+1)}{2}$ Length of extra vector
N = size of g vector VQ
D_1 = Length of $\mathbf{g_1}$ vector $= T_1 + T_{1_x}$
$T_{1_x} = \frac{T1(T1+1)}{2}$
N_1 = size of $\mathbf{g_1}$ vector VQ
D_2 = Length of $\mathbf{g_2}$ vector $= T_2 + T_{2_x}$
$T_{2_x} = \frac{T2(T2+1)}{2}$
N_2 = size of $\mathbf{g_2}$ vector VQ
D_{12} = size of $\mathbf{g_{12}}$ vector $= T_x - T_{1_x} - T_{2_x}$
R = Pitch search range
$N = N_1 * N_2, T = T_1 + T_2$

The time complexity is function of search range R and size N of VQ, while space complexity is a function of size N of VQ.

5. EXPERIMENTAL RESULTS

To asses subjective and objective performance of the different methods we assign 8-bits for quantizing the predictor

VQ Method	Total Multi.	Storage Space	Avg. Pred. Gain dB	Avg. Seg. SNR dB
Fast Conventional VQ	20480	5120	6.51	9.17
LM Conventional VQ	20480 +15360	1280	6.51	9.17
Full Search PCVQ	20480 +6144	288 +40	6.36	9.13
Sequential Search PCVQ	1920 +256	96 +16	6.17	9.09

Table 2. 5-Tap Pitch Predictor Complexity and Performance

coefficients. The codebook for 8-bit 5-dimensional VQ is designed using K-means algorithm. For PCVQ design, 5-bits were allocated for quantizing $\mathbf{g_1}$ vector, and 3-bit were allocated for quantizing $\mathbf{g_2}$ vector. The codebooks for both the feature vectors were designed using K-means algorithm.

With this specification, $T = 5, N = 256, T_1 = 3, T_2 = 2, N_1 = 32, N_2 = 8, R = 4, D = 20, D_1 = 9, D_2 = 5, D_{12} = 6, T_x = 15, T_{1_x} = 6, T_{2_x} = 3$.

All four methods were incorporated in the CELP coder for objective and subjective evaluation. Since we are evaluating the performance of the MTPP, the spectral parameters of short-term predictor and stochastic codebook excitation parameters are not quantized. The frame size selected is 224 samples with 4 subframes. MTPP analysis is done every subframe. The objective performance of all the methods is shown in Table 2. Column 4 is the average prediction gain (gain between the target vector and target vector minus contribution from the adaptive codebook). Column 5 is the average segmental SNR of synthesized speech using the specified method. The objective performance of sequential PCVQ is very close to conventional 5-dimensional VQ and full search PCVQ. An informal listening was done for subjective evaluation of each method. It was found that speech quality between conventional VQ and full search PCVQ was same. While speech quality using sequential search PCVQ method was very close to full search PCVQ.

Quality performance of sequential PCVQ can be improved by increasing the fanout [5] for the feature vectors. Instead of quantizing the feature vector using one codebook, multiple codebooks can be used to quantize the feature vector. There is a slight increase in space complexity using this approach. By increasing search range R, improvement in objective and subjective performance of VQ can be achived. Since it plays major role in defining the time complexity of the VQ, smallest value of R should be selected. In our simulation, with $R = 4$, gave good speech quality while maintaining the computational complexity of VQ design to desired level.

6. CONCLUSION

A low complexity VQ design of 5-tap pitch predictor is presented. This method is very useful where quality and complexity (time and space) are key factors in speech coder design. Tradeoff between complexity and performance for PCVQ design can be done by properly selecting $R, N1$ and $N2$ based on the DSP resources available for real time implementation of coder. Also auxiliary table storage allows tradeoff between time and space complexity of PCVQ design. Such a 5-tap pitch predictor PCVQ has application in low cost, high quality low bit rate speech coder design for digital telephone answering machine and digital memo recorders.

7. ACKNOWLEDGMENTS

Author would like to thank John DellaMorte, Doug Kolb and Jim DellaMorte of DSP Software Engineering, Inc. for their helpful advice and discussion.

REFERENCES

[1] 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 Acoust., Speech and Signal Processing*, pp. 937-940, 1985.

[2] Peter Kroon and Bishnu Atal, "On improving the performance of pitch predictors in speech coding systems," in *Advances in Speech Coding*, pp. 321-327, Kluwner Academic Publisher, Boston

[3] D. Veeneman and B. Manzor, "Efficient Multi-Tap Pitch Predictor for Stochastic coding," in *Speech and Audio Coding for wireless and network applications*, pp. 225-229, Kluwner Academic Publisher, Boston, 1993

[4] J-H Chen, "Toll-Quality 16kbps CELP speech coding with very low complexity," *Proc. ICASSP*, pp. 9-12, 1995

[5] Wai-Yip Chan, "The Design of Generalized Product-Code Vector Quantizer," in *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing*, pp. 389-392, 1992 Kluwner Academic Publisher, Boston, 1993

[6] Shihua Wang, Erdal Paksoy, and Allen Gersho, "Product code vector quantization of LPC parameters," in *Speech and Audio Coding for wireless and network applications*, pp. 251-258, Kluwner Academic Publisher, Boston, 1993

[7] M. Young, and Allen Gersho, "Efficient encoding of the long-term predictor in vector excitation coders," in *Advances in Speech Coding*, pp. 329-338, Kluwner Academic Publisher, Boston, 1991