

Case Study: Product Quantization in Digital Filter Realizations

Patrick O’Keefe, *Student Member, IEEE*

Abstract—Gain structures in digital filters are a source of product quantization noise due to finite wordlength. With certain assumptions, this gain structure’s noise can be modeled as a white noise source. This paper seeks to explore the effects of product quantization in the HVXC Decoder for the MPEG-4 Audio3 standard. The filter of interest is a fourth order high pass component in the post-processing stage of the decoder. Also, a scaling strategy will be employed to mitigate the effect of the quantization noise to reduce, if not completely eliminate, the chance of register overflow.

Index Terms—Product quantization, HVXC Decoder, L_p Norm, Scaling.

I. INTRODUCTION

DIGITAL filters are always subject to finite wordlength issues. Quantized coefficients can cause a pole that is very near to the unit circle to exceed that boundary, causing the filter to become unstable. Product quantization is another issue that arises. The coefficients of a digital filter are effectively gain structures for the signal path through which additional noise is accumulated. When two signals of length b bits are multiplied together, they can produce a result of length $2b$ bits. In order to keep a uniform bit length throughout the registers, truncation or rounding must be employed. Under certain conditions, this noise can be modeled as a white noise source with a magnitude dependent on the quantization resolution. Obviously, this additional noise has negative effects on the signal to noise ratio (SNR) of the system. This noise can easily cause the dynamic range of a finite length filter to be exceeded, in which case overflow occurs. [2]

Scaling is a technique that is used to prevent overflow. When scaled properly, a digital filter will avoid overflow even with the added noise from product quantization while maintaining a relatively high SNR. If a filter is over-scaled, the frequency response of the filter approaches the noise floor which directly reduces the filter’s SNR.

This paper analyzes a high pass filter from the MPEG-4 Audio3 standard [3]–[5]. The standard uses Harmonic Vector eXcitation Coding (HVXC) to achieve very low bit rates, particularly for speech signals. Our high pass filter of interest is one of the many filters that are contained within the HVXC, and it is used to eliminate unnecessary low frequencies. After introducing the filter, we will discuss the theory and effects of product quantization. To reduce or eliminate the chance of overflow, an appropriate scaling strategy will be developed that compensates for the product quantization.

Manuscript received October 19, 2009.

P. O’Keefe is with the Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL, 33146 USA e-mail:p.okeefe@umiami.edu.

II. THE HVXC DECODER

The HVXC Decoder contains a fourth order high pass filter that removes unnecessary low frequencies in the post-processing stage of the MPEG-4 Audio3 standard. The filter structure will be analyzed along with its performance in “infinite” wordlength conditions.

A. Filter Structure

As mentioned above, filters are sensitive to coefficient quantization under finite wordlength restrictions. A common way to reduce this sensitivity is to break up the filter into several cascaded second-order sections known as bi-quad sections. [1] These bi-quad sections independently realize their pole/zero pairs without disturbing any other sections. This results in a reduction in the propagation of quantization errors. A particular way of realizing a bi-quad section is known as Direct Form II and is represented by (1).

$$H(z) = \prod_k \frac{a_{0k} + a_{1k}z^{-1} + a_{2k}z^{-2}}{b_{0k} + b_{1k}z^{-1} + b_{2k}z^{-2}} \quad (1)$$

Direct Form II implementations are characterized by having the same number of delays as the order of the filter. A signal flow graph of one bi-quad section is shown in Figure 1.

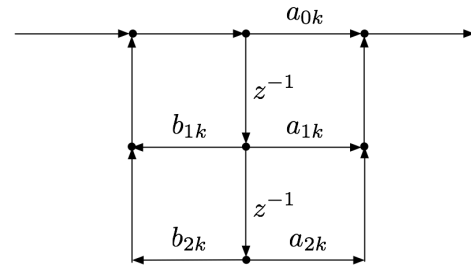


Fig. 1. Direct Form II Implementation of a second-order IIR

The fourth order high pass filter from the HVXC Decoder is given as two cascaded bi-quad sections implemented in Direct Form II as shown in (2) where the coefficients for the filter are given in Table I. This cascaded form is convenient because we would like to focus on the product quantization error and not the coefficient sensitivity.

$$H_{HPF}(z) = K_{HPF} \left(\frac{1 + a_{11}z^{-1} + a_{12}z^{-2}}{1 + b_{11}z^{-1} + b_{12}z^{-2}} \right) \dots \left(\frac{1 + a_{21}z^{-1} + a_{22}z^{-2}}{1 + b_{21}z^{-1} + b_{22}z^{-2}} \right) \quad (2)$$

TABLE I
HVXC COEFFICIENTS

Coefficient	Value
K_{HPF}	+1.100000000000000
a_{11}	-1.998066423746901
a_{12}	+1.000000000000000
b_{11}	-1.962822436245804
b_{12}	+0.9684991816600951
a_{21}	-1.999633313803449
a_{22}	+0.999999999999999
b_{21}	-1.858097918647416
b_{22}	+0.8654599838007603

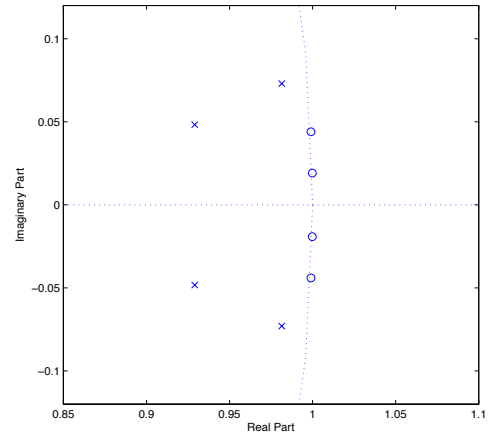
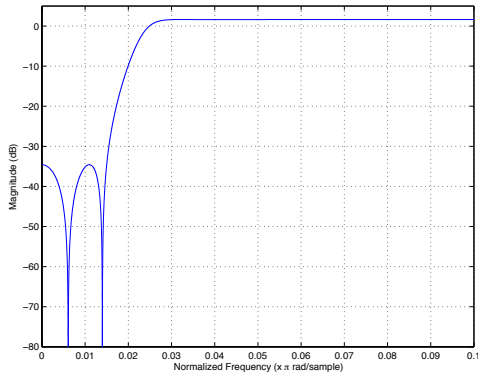


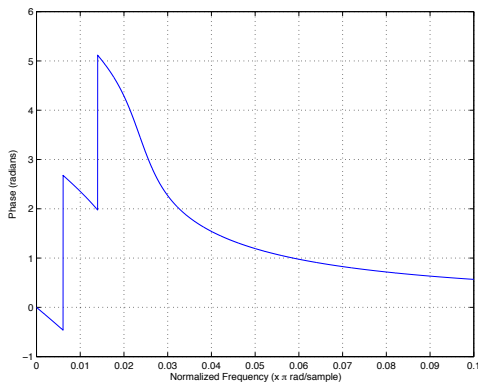
Fig. 3. Pole-Zero plot of the fourth order high pass filter

B. Filter Performance

Because MATLAB uses double-precision floating point values, we can assume that our initial analysis of the filter is using “infinite” wordlength. The magnitude and phase response of the filter can be seen in Figure 2 and the pole-zero plot is shown in Figure 3. The axes are adjusted to show the relevant information. Note that the the passband magnitude is slightly above 0 dB in Figure 2a. This can be attributed to the K_{HPF} term before the cascaded stages and will be taken into account when developing a scaling strategy.



(a)



(b)

Fig. 2. (a) Magnitude and (b) Phase response of the fourth order high pass filter from the HVXC Decoder

III. PRODUCT QUANTIZATION EFFECTS

As mentioned above, truncation or rounding must be employed after multiplication to maintain a uniform bit length. This introduces quantization noise into the system. If the quantization level q is small and the signal moves between several quantization levels from sample to sample, we can assume that the noise introduced can be modeled as a white noise source. [2] Note that $q = 2^{-B}$ where B is the number of bits used. Let us focus our discussion sign magnitude rounding (SMR) because of its zero mean. The following holds for SMR processes:

$$\begin{aligned} \text{Range} &= \left[-\frac{q}{2}, +\frac{q}{2} \right] \\ \mu_e &= 0 \\ \sigma_e^2(m) &= \frac{q^2}{12} \delta(m) \\ \phi_{ee}(m) &= \frac{q^2}{12} \delta(m) \end{aligned} \quad (3)$$

The fact that the sources are uncorrelated means that they can be summed to a common node. Our high pass filter of interest contains nine coefficients for multiplication. However, notice that in our first bi-quad section coefficient a_{12} is one. This means that we have three gain structures in the first bi-quad, four in the second, and one before either section. The covariance of a noise source is a function of the number of zero multipliers, M , and the number of pole multipliers, N and is given by (4).

$$\sigma_{ee}^2 = (M + N + 1) \frac{q^2}{12} \quad (4)$$

Therefore, for 8-bit fixed point SMR, our first bi-quad has $\sigma_{ee}^2 = 5.0863 \times 10^{-11}$ and our second bi-quad has $\sigma_{ee}^2 = 6.3578 \times 10^{-11}$. When these noise sources are added into the signal path, a new magnitude spectrum arises and can be seen in Figure 4.

With fewer bits, product quantization is more severe and the signal to noise ratio is even worse. In the next section,

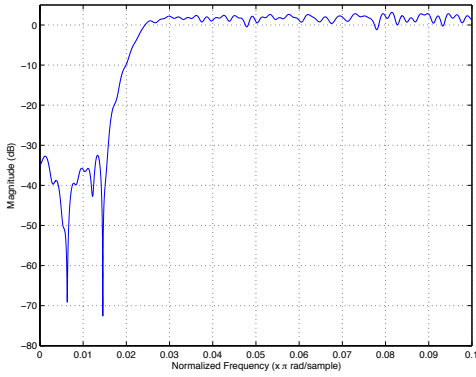


Fig. 4. Magnitude spectrum of fourth order high pass filter with 8-bit product quantization noise

a scaling strategy will be developed to compensate for this noise.

IV. SCALING

Scaling is the attenuation of a signal before it enters a stage or filter. An efficient method of scaling that utilizes norms is referred to as L_p Scaling. [1] To understand L_p Scaling, it is important to know what is meant by a norm. The L_p norm of a transfer function $H(\omega)$ is given by (5). [2]

$$\|H\|_p = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |H(\omega)|^p d\omega \right]^{1/p} \quad (5)$$

provided that

$$\int_{-\pi}^{\pi} |H(\omega)|^p d\omega < \infty \quad (6)$$

It can be shown that the L_∞ norm is the maximum value of the transfer function as in (7).

$$\|H(\omega)\|_\infty \equiv \lim_{p \rightarrow \infty} \|H(\omega)\|_p = \max_{\omega \in [-\pi, +\pi]} |H(\omega)| \quad (7)$$

Given the maximum input, u_{\max} , and overflow bound, M_{overflow} , a scaling factor of λ is necessary to remain bounded. This relationship is given by the following. [2]

$$\lambda u_{\max} \leq \frac{M_{\text{overflow}}}{\max_k \|H_k\|_p} \quad (8)$$

This indeed works, but the problem of scaling is complicated when cascading sections. The scale factor λ can be placed at the beginning of the sections, but this has a negative effect on the SNR at the filter output. The scale factor can be placed at the end, but then there is a risk of overflow in the initial stages. Therefore, one solution to lessen either effect is to distribute the scaling parameter λ at each cascaded section. [2] In a filter with L cascaded sections, we need to find the scaling factors K_k where $k = 1, 2, 3, \dots, L$. Using (8) as a guide, we can choose our distributed scaling factors such that

$$\prod_{l=1}^k K_l u_{\max} \leq \frac{M_{\text{overflow}}}{\|\prod_{l=1}^k H_l\|_p} \quad (9)$$

In this case, both M_{overflow} and u_{\max} are defined to be one. The scaling factors need to be assigned to the sections in order of most sensitive to product quantization to least sensitive. Our first section is the most sensitive, so its scaling parameter will simply take the inverse of the norm for the Direct Form II transfer function. Typically, either L_2 or L_∞ norms are used. Using the L_∞ norm guarantees that there will be no chance of overflow at the expense of SNR. Using the L_2 norm reduces the chance of overflow without guaranteeing its occurrence, but it yields a higher SNR. Here, the L_∞ norm is employed to minimize overflow and also to compensate for the K_{HPF} scaling inherent in the filter design that occurs before either stage. The result of the new scaling strategy can be seen in Figure 5. It is apparent that the SNR is lower than in Figure 4, but the passband magnitude has been reduced. Therefore, the chance of overflow has also been reduced.

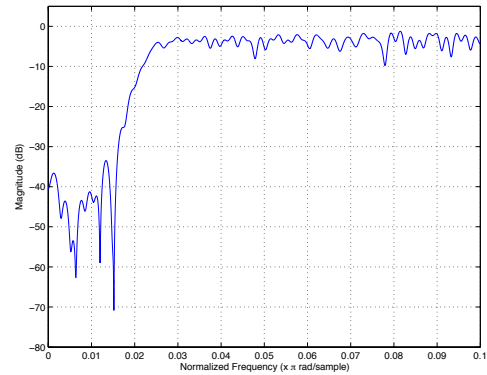


Fig. 5. Magnitude spectrum of fourth order high pass filter with 8-bit product quantization noise and a scaling strategy based upon the L_∞ norm

V. CONCLUSION

Product quantization effects can cause register overflow, and therefore need to be compensated with a scaling strategy. The benefits of using L_∞ norm scaling constants was shown above on the fourth order high pass filter in the post-processing stage of the HVXC Decoder in the MPEG-4 Audio3 standard.

REFERENCES

- [1] S.K. Mitra, *Digital Signal Processing: A Computer Based Approach*, ser. McGraw-Hill Series in Electrical and Computer Engineering, S. W. Director, Ed. New York, NY: McGraw-Hill 2006.
- [2] K. Premaratne, *DSP Algorithm Implementation: Finite Wordlength Issues*, EEN536 Class Notes. Department of Electrical and Computer Engineering, University of Miami, 2009.
- [3] S. Battista, F. Casalino, and C. Lande, "MPEG-4: a multimedia standard for the third millennium 1," *IEEE Multimedia*, vol. 6, no.4, pp. 74-83, Oct./Dec. 1999.
- [4] S. Battista, F. Casalino, and C. Lande, "MPEG-4: a multimedia standard for the third millennium 2," *IEEE Multimedia*, vol. 7, no.1, pp. 76-84, Jan./Mar. 2000.
- [5] *Information Technology Very Low Bitrate Audio-Visual Coding, Part 3 Audio*, MPEG Working Group, International Standards Organization/International Electrotechnical Commission (ISO/IEC) Std. ISO/IEC FCD 14 496-3 Subpart 1, May 1998.



Patrick O'Keefe is an undergraduate pursuing a Bachelor of Science in Electrical Engineering in the Audio Engineering Program at the University of Miami, Coral Gables, FL USA. His research interests include human computer interaction, low-latency audio processing on embedded platforms, computer vision, and music information retrieval.