Error Spectrum Shaping and Vector Quantization

Jon Dattorro
Christine Law

in partial fulfillment of the requirements for EE392c
Stanford University
Autumn 1997

0. Introduction

We view truncation noise and quantization noise in basically the same way, arising from similar phenomena, regardless of bit rate. Truncation error feedback has been employed successfully in the scalar quantization of audio signals for about 20 years. The fundamental idea is to gain control over the spectral shape (the color) of the truncation error noise which is defined as the difference between the original signal and its truncated, rather, quantized version. The motivation can be both perceptual and quantitative; that is, in some circumstances it is advantageous to have colored truncation noise because of a certain perceptual insensitivity, while in other circumstances the MSE (mean square error), in what is considered to be the spectral baseband, can actually be reduced by the coloration.[1]

In this report, we demonstrate and conclude that truncation error feedback is *not* successful in the quantization of still images. In the images we examine, we do not see a preponderance of quantization error energy localized in any particular spatial frequency region. As we will see, the monic FIR noise feedback filters that we choose have the unfortunate side-effect of boosting truncation noise in part of the spectral range. Hence, truncation error feedback is not effective in reducing bit rate. This may help to explain why we have not seen this idea appearing in the literature. (See [2] for a good survey up to 1991.)

[1]Two examples of this are: 1) error feedback in a system known to have spectral error localized in a particular frequency region, and 2) error feedback in over-sampled systems. The former might occur in digital filtering applications while the latter might occur in D/A conversion applications [1].

1. Truncation Error Feedback in 1 Dimension

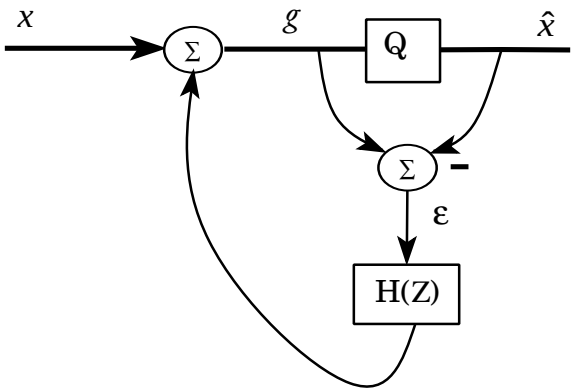


Figure 1. Truncation error feedback in a quantization application.

Figure 1 shows the basic concept for which we now develop fundamental equations. H(z) is the digital filter applied only to the truncation error.

$$\begin{aligned} \varepsilon &\equiv g - \hat{x} \\ g &= x + \varepsilon * h \end{aligned} \qquad (1)$$

Where $*$ denotes the convolution operator. These two time-domain equations are read directly from Figure 1. We combine them to get

$$\hat{x} = x + \varepsilon * h - \varepsilon \qquad (2)$$

Now since these signals are all deterministic, they have Fourier spectra. Hence we can say

$$\hat{X} = X - E(1 - H) \qquad (3)$$

which is a frequency domain expression where $E \leftrightarrow \varepsilon$. We see how the quantized output signal can be expressed in terms of the original input signal in Equations (2) and (3). Equation (3) says that the truncation error spectrum E is modified by the feedback filter H(z). In the absence of the feedback filter (in the case that H=0), Equation (3) would say that the quantized output signal spectrum can be conceived in terms of the original input signal spectrum less the error spectrum; that is,

$$\hat{X} = X - E \qquad ; H = 0 \qquad (3a)$$

So we see that the impact of the error feedback filter H is to somehow change the spectrum of the quantization error. That is our intended goal; to gain control over the error spectrum.

2. Choice of Error Feedback Filter *H*(z)

The *filter factor* multiplying E in Equation (3) is (1 - H). H(z) must be a polynomial having no zero-order terms if a delay-free loop is to be avoided in Figure 1. The simplest choice is

$$H(z) = z^{-1} \qquad (5)$$

This choice is well known to produce limit cycle tones in the circuit of Figure 1. [3] [4] Gray [3,ch.6.5] discusses the resultant tones in terms of the spectrum of the error signal ε for a restricted class of input signals. Gray also makes the connection to truncation error feedback [3,Fig.6.2,Eq.6.5.2,Eq.6.6.4]. Gray explains that for sinusoidal input signals, the truncation error spectrum E will *not* be white for feedback filter orders less than 3. [5] He says [3,ch.6.6]

> "For third order and higher, however, the noise is white for sinusoids and finite sums of sinusoids." (proof is in [4])

From [3] and [4] we learn that the best choices for H(z) are all of the form:

$$H(z) = 1 - (1 - z^{-1})^p \qquad ; p=1,2,3,\ldots \qquad (5a)$$

These choices reduce the likelihood of discrete-frequency limit cycles.

We have found in our work on this low rate application that the only useful p for coding of images is p=1 as in (5). All other choices lead to instability because of gain in the filter factor as shown in Figure 2 through Figure 4. Empirically, we find that instability can be ameliorated by increasing the size of the codebook.[2]

3. Sense of the Error Filter Factor

By inverting the error filter H, we can change the sense of the filter factor acting on the error spectrum from high to lowpass. Equation (3) becomes modified;

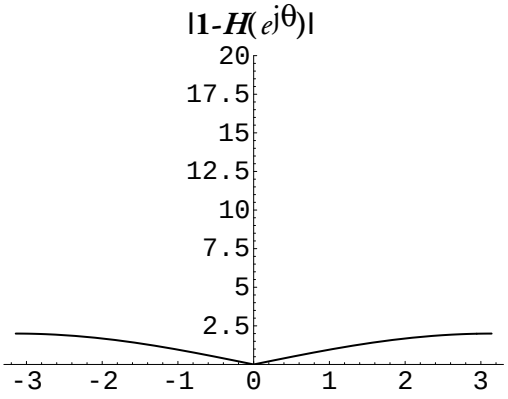
$$\hat{X} = X - E(1 + H) \qquad (3b)$$

In one dimension the choices of filter factor up to order 3 would appear as in Figure 2 through Figure 4. Theoretically, one would select the filter factor sense based upon some a priori knowledge of the quantization error spectrum. For example, were we to know in advance that the error spectrum were predominantly lowpass, then we would choose the highpass sense of the error filter factor so as to obliterate low frequency errors during the encoding process.

In two dimensions (2D), the filter factors become surfaces [6] as we shall see in the examples of actual error spectrum shaping of images in the attachments.

---

[2]We eliminated instability by increasing codebook size, but we found that the **SQNR** remained inferior for error feedback processed images.



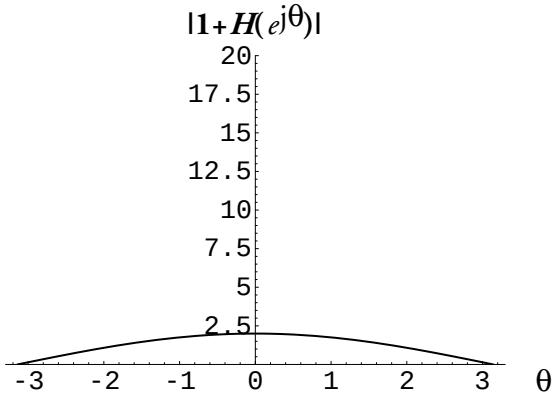


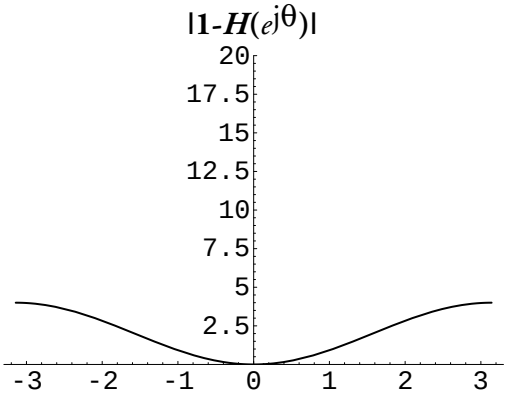
Figure 2. 1st order error feedback.

(a) highpass sense

$\hat{x}_n = x_n - (\varepsilon_n - \varepsilon_{n-1})$

(b) lowpass sense

$\hat{x}_n = x_n - (\varepsilon_n + \varepsilon_{n-1})$



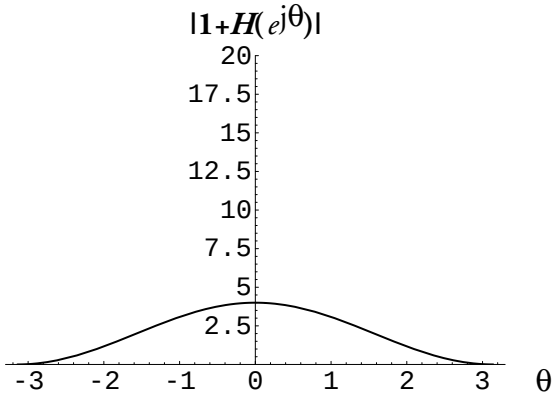


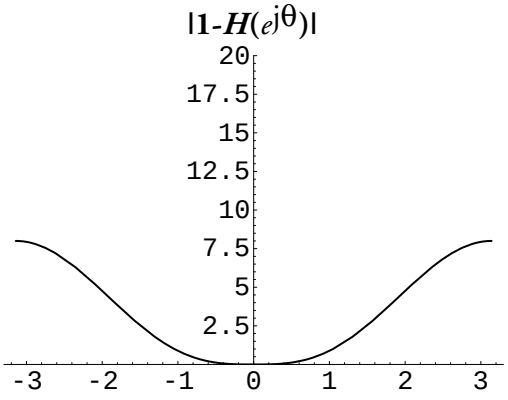
Figure 3. 2nd order error feedback.

(a) highpass sense

$\hat{x}_n = x_n - (\varepsilon_n - 2\varepsilon_{n-1} + \varepsilon_{n-2})$

(b) lowpass sense

$\hat{x}_n = x_n - (\varepsilon_n + 2\varepsilon_{n-1} + \varepsilon_{n-2})$



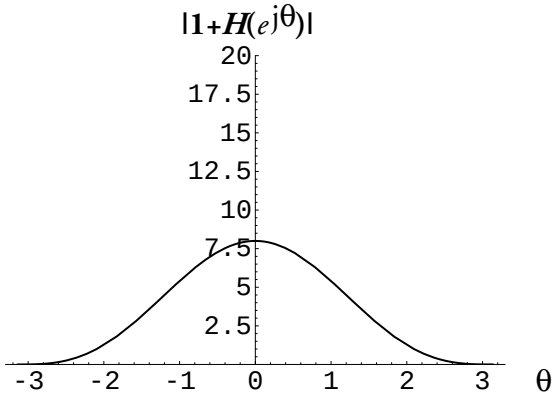


Figure 4. 3rd order error feedback.

(a) highpass sense

$\hat{x}_n = x_n - (\varepsilon_n - 3\varepsilon_{n-1} + 3\varepsilon_{n-2} - \varepsilon_{n-3})$

(b) lowpass sense

$\hat{x}_n = x_n - (\varepsilon_n + 3\varepsilon_{n-1} + 3\varepsilon_{n-2} + \varepsilon_{n-3})$

4. Delay-Free Loop when applied to Audio

The first plan was shape the error spectrum of one audio stream subject to vector quantization. Under those circumstances, the vector quantizer schematic takes the form shown in Figure 5, where $H(z)$ is as defined in Equation (5a). It is important to realize that Figure 5, less the error feedback circuit, represents the classical interpretation of vector quantization when applied to one dimensional data streams in time or space; Figure 5 is *not* our own concoction, rather, it is the correct interpretation of vector quantization.

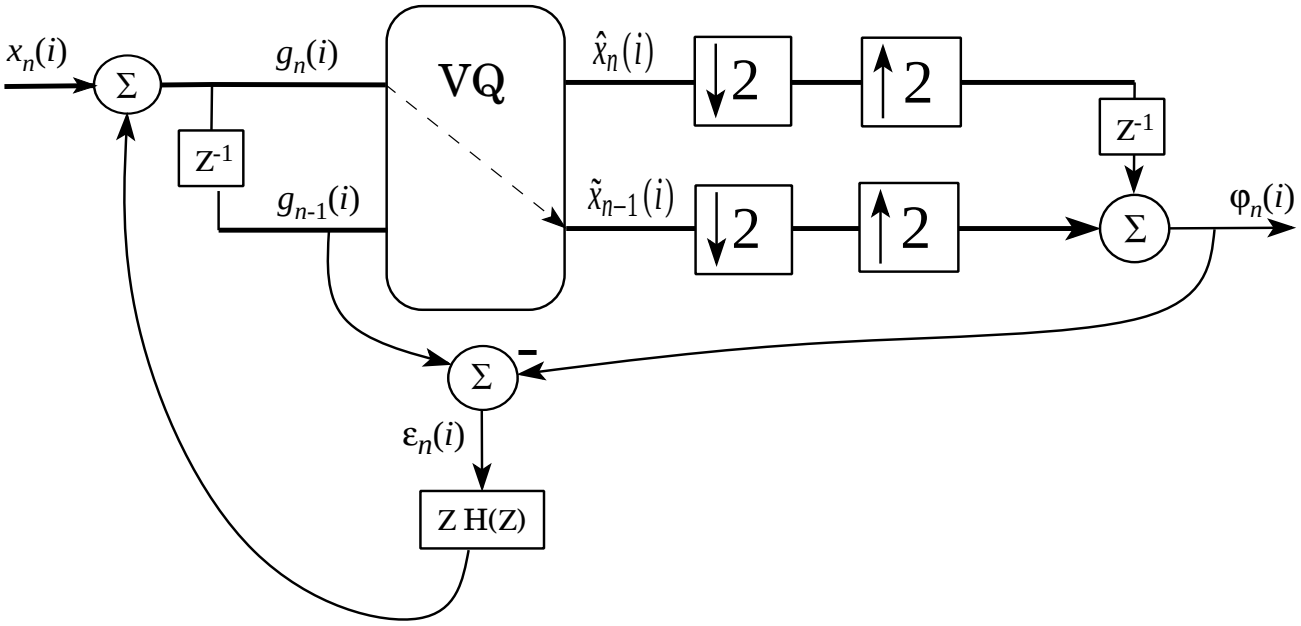


Figure 5. Flawed 2D vector quantization scheme for 1D audio data.

The multiplication of $H(z)$ by z is a compensation required by the formulation of $\varepsilon_n(i)$ using delayed signals. Normally, $\varepsilon_n(i)$ would be formulated using the instantaneous input and output of the quantizer, as in Figure 1. We believe that our formulation of $\varepsilon_n(i)$ is the only logical choice. Choosing $\tilde{x}_{n-1}(i)$ instead of $\varphi_n(i)$ as the tap point, for example, would subject the error spectrum to aliasing caused by the downsampling which is indigenous to vector quantization of 1D signals.

The reason that the circuit in Figure 5 fails is because it is not realizable due to the implicit cross-coupling indicated by the dotted line. Note that $z\,H(z)$ has a straight-through path [Eq. (5a)]. The consequence of that cross-path is to create a delay-free loop in the error feedback scheme. Hence, the circuit in Figure 5 was never implemented, and our intended application to audio was abandoned.

5. Application of Error Feedback to Images

Not having found success in devising a scheme to shape the spectral error of audio, we considered application of the idea to the vector quantization of images. Figure 6 shows the scheme we selected for the 2D vector quantization (VQ) of images. The circuit in Figure 6 codes two rows of the image, number i and i+1, in parallel while applying error feedback.

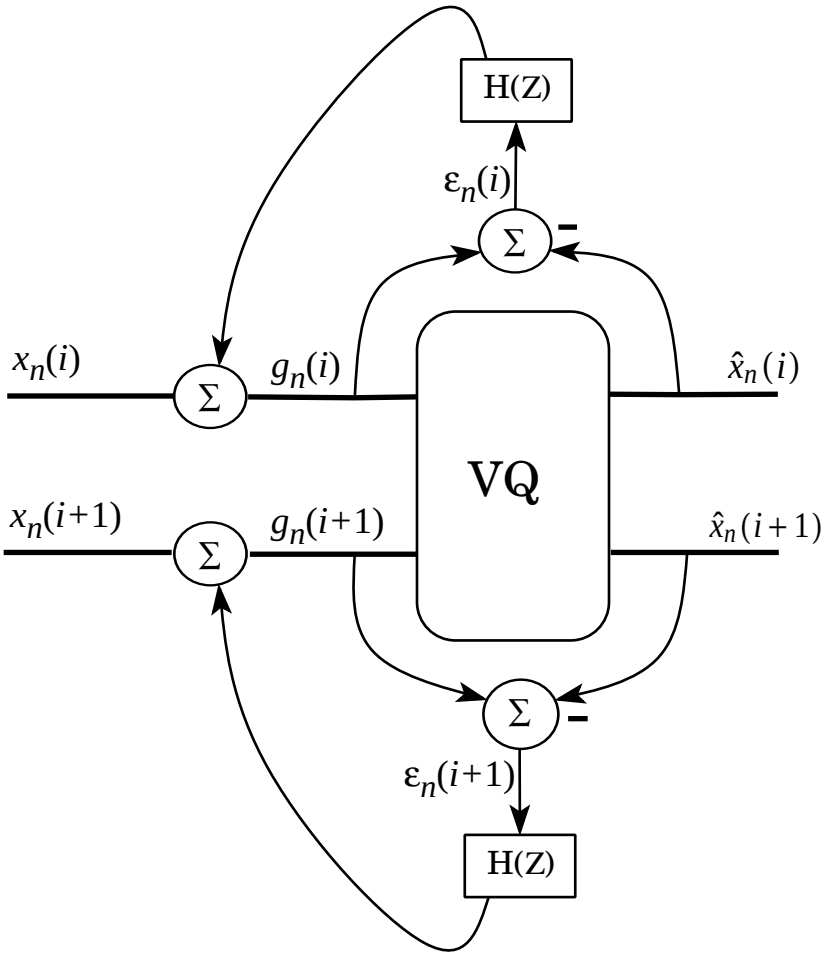


Figure 6. Row-pair 2D vector quantization scheme.

A vector quantizer codebook of size 16 (NUMVECS) two-dimensional vectors is created using the LBG [2] algorithm in a program written in C by the present author. (See the attachments.) The training set is taken to be the entire image. A new codebook is determined for each image. Truncation error feedback is applied only during the encoding process; that is, only after training independently of any considerations regarding error feedback.

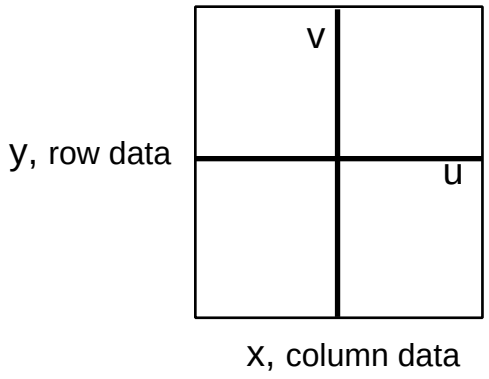


Figure 7. Nomenclature of signal and Fourier transformed data.

We show in Figure 7 the nomenclature for all our graphical presentation of data. The spatial-domain signal data organization is the same as that of the image. The rows and columns of input signal data are respectively labelled y and x, while their Fourier transformed counterparts are v and u. Hence the variable $x_n(i)$ denotes the one-dimensional signal found by reading the spatial sequence in the i^{th} row. The Fourier transform of all our data is organized so that DC in both spatial directions resides at the center of the graph.

We quantize all the pixel pairs (as in Figure 8) in adjacent rows, number i and i+1, then we proceed to the next pair of rows. Each time a row pair is begun, all previous memory is cleared. The memory in row i is of the previous column n-1 mod 256. Thus the columns become our spatial analogue to time.

row $\boldsymbol{i}$	•	•	•	…	•
row $\boldsymbol{i}$+1	•	•	•	…	•
	0	1	$\boldsymbol{n}^{th}$ *column*→		255

Figure 8. Pixel pairs.

6. Attachments

The attachments show the results of applying the VQ circuit in Figure 6 to images. There we show:

◦ the original image and its 2D Fourier transform,

◦ a three–dimensional graph of original image luminosity,

◦ the vector quantized image with and without error feedback,

◦ the error image $X - \hat{X}$ and its 2D lfftl,

◦ and the actual magnitude response of the error feedback filter factor $(1 \pm H)$.

The graph of image luminosity was generated so as to aid understanding of the corresponding 2D lfftl.

The lfftls of the error images without error feedback show some spikiness; energy is concentrated. But there are no pronounced trends in the error image Fourier magnitude.

The actual magnitude response of the error feedback filter factor was determined by dividing the DFT of the error image $X - \hat{X}$ by E (Eq.(3)). For that to work, it became necessary to make the convolution $\varepsilon * h$ in Equation (2) circular in the C program.

Table 1. **SQNR** of VQ vs. VQ with first order truncation error feedback.

Image	VQ	VQ with lowpass EFB	VQ with highpass EFB	entropy
cman	17.56 dB	15.00 dB	13.00 dB	1.37 bit
mri	16.90 dB	14.51 dB	11.87 dB	1.47 bit
einstein	13.96 dB	11.50 dB	9.96 dB	1.64 bit
reagan	16.48 dB	14.10 dB	12.26 dB	1.87 bit
smandril	13.21 dB	10.78 dB	9.74 dB	1.88 bit

In Table 1 we compare the performance, via the MSE in dB which we call the **SQNR** (signal to quantization noise ratio), of ordinary vector quantization to the performance of the same vector quantizer when truncation error feedback (EFB) is incorporated into the encoding process as in Figure 6. Our data reflects both the high (Eq.(3)) and lowpass (Eq.(3b)) sense of the error filter factor $(1 \pm H)$.

The lowpass sense seems to be favored but none of the EFB data surpasses ordinary VQ. The lowpass favor is a surprise because the spikiness observed in the error images having no error feedback would seem to favor the highpass sense.

Since the codebook vectors are unaffected by the use of error feedback, the entropy is the same in all VQ cases.

7. Conclusions

Observation of the 2D |fft| of the error images in the attachments show no preponderance of energy in localized regions. Hence it is difficult to incorporate a simple digital filter such as those we propose here into the error feedback scheme in order to minimize the overall error. The FIR error filters $H(z)$ are constrained to be monic thus limiting our latitude in the design procedure. Further we have found that at very low bit rate, the error filter order cannot be greater than 1, for then the system becomes unstable. In light of these findings, we conclude that this technique is *not* very useful when applied in this manner. Further research is required; perhaps incorporating the error feedback into the LBG codebook design procedure warrants investigation.

References

[1] R. J. van de Plassche, E. C. Dijkmans, 'A monolithic 16-bit D/A Conversion System for Digital Audio', B. A. Blesser, B. Locanthi, T.G. Stockham, Jr. (eds.), *Digital Audio*, The Proceedings of the Audio Engineering Society Premiere Conference, New York, pp.54-60, 1982 June 3-6

[2] Allen Gersho, Robert M. Gray, *Vector Quantization and Signal Compression*, 1991, Kluwer Academic Publishers

[3] Robert M. Gray, *Source Coding Theory*, 1990, Kluwer Academic Publishers

[4] Wu Chou, Ping Wah Wong, Robert M. Gray, 'Multistage Sigma-Delta Modulation', IEEE *Transactions on Information Theory*, vol.35, no.4, pp.784-796, 1989 July.

[5] R. C. Ledzius, J. Irwin, 'The Basis and Architecture for Reduction of Tones in a Sigma-Delta DAC', IEEE *Transactions on Circuits and Systems II*, vol.40, no.7, pp.429-439, 1993 July.

[6] Ronald N. Bracewell, *Two-Dimensional Imaging*, 1995, Prentice-Hall