REDUCING QUANTIZATION ERROR BY MATCHING PSEUDOERROR STATISTICS

Stephen D. Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado 80305, USA, svoran@its.bldrdoc.gov

ABSTRACT

We investigate the use of an adaptive processor (a quantizer pseudoinverse) and the statistics of the associated pseudoerror signal to reduce quantization error in scalar quantizers when a small amount of prior knowledge about the signal x is available. This approach uses both the quantizer representation points and the thresholds at the receiver. No increase in the transmitted data rate is required. We discuss examples that use low-pass, high-pass, and band-pass signals along with an adaptive processor that consists of a set of filters and clippers. Matching a single pseudoerror statistic to a target value is sufficient to attain modest reductions in quantization error in situations with one degree of freedom. Adaptive processing based on a pair of pseudoerror statistics allows for quantization noise reduction in problems with two degrees of freedom.

Index Terms—quantizer, quantization noise reduction

1. INTRODUCTION

Quantization is fundamental to digital communications and digital signal processing. Audio, video, and other waveforms may be quantized at various resolutions (e.g., 8 to 24 bits/sample). Lower quantization resolutions may be used for extracted signal coding parameters inside of audio and video coders or in situations where channel capacity is severely limited (e.g., 2 to 8 bits/sample). A quantizer may be a uniform rounding quantizer (URQ) or it may be a Lloyd-Max quantizer (LMQ) [1] that is optimized to match the probability density function of a specific signal. Quantizers can be designed to operate on scalar or vector signals and a comprehensive overview of the topic is available in [2].

Figure 1 describes the basic case of memoryless scalar quantization using b bits/sample, or $N=2^b$ levels/sample. On the transmitting side Q_{TX} compares the signal with $N+1$ thresholds $\{t_i\}$ and sends one of N codes $\{c_i\}$. On the receiving side Q_{RX} uses the code c_i to look up one of N representation points $\{r_i\}$. When x falls into the j^{th} quantization cell, the quantizer operation is described by

$$t_j \le x < t_{j+1} \Rightarrow \hat{x} = r_j . \quad (1)$$

In a URQ the thresholds are uniformly spaced and the representation points are centered between the corresponding thresholds. In an LMQ the thresholds and representation points are jointly optimized to minimize the mean-squared quantization error $\varepsilon^2 = E\left(\hat{x} - x\right)^2$. The result is that each representation point is at the conditional mean of its quantization cell, and each threshold is midway between two representation points. Note, however, that when either type of quantizer is used, the thresholds are used only in Q_{TX} and the representation points are used only in Q_{RX}.

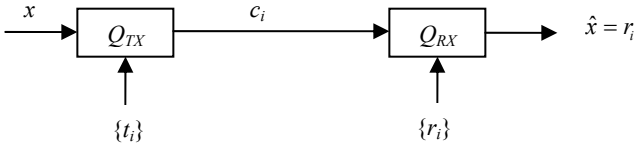

Fig 1. Conventional memoryless scalar quantization.

In this paper we investigate the use of a quantizer pseudoinverse and the statistics of the associated pseudoerror signal to reduce quantization error in scalar quantizers when a small amount of prior knowledge about the signal x is available. This approach can make use of both the representation points *and the thresholds* at the receiving side. The prior signal knowledge, the thresholds, and the representation points are all embedded at design time and thus do not add to the transmitted data rate during operations.

The only related prior work we are aware of is given in [3]. That work addresses two-dimensional quantization of line spectral frequencies (LSFs) for speech coding. The quantized LSF trajectory is smoothed to improve speech quality and that smoothing occurs under a soft constant that pulls the trajectory towards the Voronoi regions of the unsmoothed trajectory. Thus [3] uses the two-dimensional equivalent of the thresholds (Voronoi regions) at the receiving side. Note also that dither and noise shaping can reduce the perceptibility of quantization noise, but neither technique reduces the mean-squared quantization error.

2. QUANTIZER PSEUDOINVERSE AND PSEUDOERROR STATISTICS

The Lloyd-Max quantizer minimizes the mean-squared quantization error ε^2 for any given resolution or rate b, and in the most general case this error cannot be further reduced. However, in some slightly constrained cases ε^2 can be reduced through additional processing at the receiving side. This is possible when the receiving side has access to a small amount of prior knowledge about the signal x (e.g., sign of the spectral tilt of x). This additional processing can advantageously use the values of the quantizer thresholds $\{t_i\}$ to reduce ε^2.

Figure 2 summarizes our approach. Here $Q^{\dagger}$ is a processing element that can be viewed as an approximate inverse (a "pseudoinverse") of the quantizer Q, where Q is the composition of Q_{TX} and Q_{RX}. In the most general case, Q destroys information and no amount of further processing can retrieve that information. However in some slightly constrained cases (e.g., sign of spectral tilt known), some of that information can be retrieved.

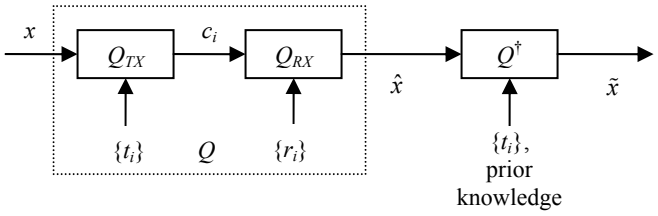

Fig 2. Conventional quantizer followed by a pseudoinverse.

We define the quantizer error signal to be $e = \hat{x} - x$ and the quantizer pseudoinverse pseudoerror signal as $p = \hat{x} - \tilde{x}$. Now consider the case of the theoretical continuum where $Q^\dagger$ evolves from the null processor, to a pseudoinverse for Q, and then on to an exact inverse for Q. This would have to cause the pseudoerror signal p to evolve from zero to an exact copy of the error signal e. Likewise at the end of this evolution, the total error $\tilde{x} - x$ would be reduced to zero. The key observation here is that if $Q^\dagger$ is to invert Q to the maximum extent possible, then the pseudoerror signal $p = \hat{x} - \tilde{x}$ must emulate the error signal $e = \hat{x} - x$ to the maximum extent possible. While the error signal e is unknown at the receiving side, the statistical properties of that error signal can be known and the pseudoerror signal can be tuned so that its statistics emulate those known error signal statistics to the maximum extent possible.

Here are four observations on quantization error. Quantizers are designed so that quantization error has zero mean. Quantization error is always bounded according to the size of the applicable quantization cell. For example, when x falls into the j^{th} quantization cell, the error e is bounded according to

$$r_j - t_{j+1} < e \le r_j - t_j . \tag{2}$$

In addition, if the signal pdf is relatively flat within a given quantization cell of width $t_{j+1} - t_j = 2\Delta$, the quantization error pdf associated with that cell will be nearly uniform on the interval $(-\Delta, +\Delta]$ and the associated variance and kurtosis are

$$\sigma_e^2 = E\left(\hat{x} - x\right)^2 = \frac{\Delta^2}{3}, \tag{3}$$

$$\beta_e = \frac{E\left(\hat{x} - x\right)^4}{\left(\sigma_e^2\right)^2} = \frac{9}{5} = 1.8 . \tag{4}$$

The goal of matching quantization error statistics leads to the adaptive processing structure shown in Figure 3. Here the statistics of the pseudoerror are calculated and compared with the *a priori* known statistics of the error signal. The difference between these two can then steer the adaptive processor to match these statistics to the maximum extent possible, and thus it may invert Q to the maximum extent possible. In general, these processes can advantageously use the values of the quantizer thresholds. Since these can be known and embedded at design time, they do not appear as inputs in Figures 3 or 4.

Note that matching pseudoerror statistics to quantization error statistics is necessary, but not sufficient for reducing the overall quantization error $E\left(\tilde{x} - x\right)^2$. For example, a processor could simply inject a suitable noise signal to achieve the desired match of statistics, but this cannot reduce the overall quantization error. Rather, if the processor is to reduce the overall quantization error, then some prior knowledge of the signal is required.

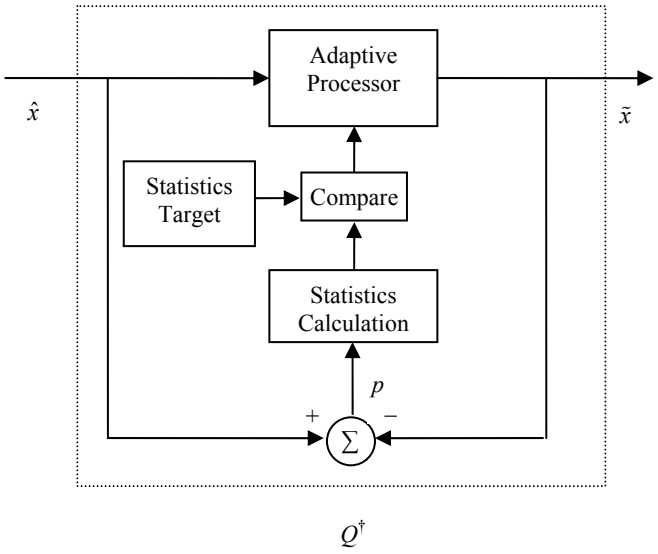

Fig 3. General structure of a quantizer pseudoinverse, $Q^\dagger$.

3. EXAMPLES

We have conducted several demonstrations of the quantizer pseudoinverse technique and have observed modest reductions in quantization noise under certain conditions. In each demonstration, the adaptive processor is in essence an adaptive filter followed by a clipper. For simplicity, our implementations actually use a set of fixed filters followed by clippers and an output selector switch, as depicted in Figure 4.

After filtering, the processor constrains each filtered sample to the proper quantization cell by clipping. This is easily done by comparing the filter output sample $\tilde{x}_i$ with the appropriate quantizer thresholds, based on the filter input sample $\hat{x}_i$. For example if $\hat{x}_i$ is located in the j^{th} quantization cell, then the clipping process is given by

$$\begin{aligned} \tilde{x}_i < t_j \quad &\Rightarrow \quad \text{replace } \tilde{x}_i \text{ with } t_j , \\ t_j \le \tilde{x}_i \le t_{j+1} \quad &\Rightarrow \quad \text{no change to } \tilde{x}_i , \\ t_{j+1} < \tilde{x}_i \quad &\Rightarrow \quad \text{replace } \tilde{x}_i \text{ with } t_{j+1} . \end{aligned} \tag{5}$$

By this process, $\tilde{x}_i$ is guaranteed to be in the same quantization cell as $\hat{x}_i$ and hence the same cell as x_i. This cannot increase quantization error, and in practice it often decreases quantization error. This clipping process makes use of the quantizer thresholds at the receive side, something that is not done in conventional quantization.

Note that if the filtering could somehow be perfectly matched to the original signal x, then the filter would move each input value $\hat{x}$ to an output value $\tilde{x}$ that exactly matches the original signal x. This

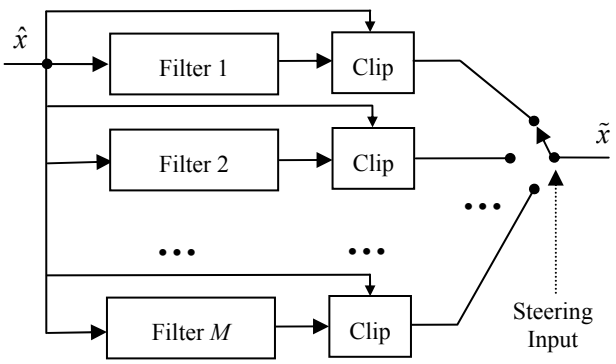

Fig 4. Adaptive processor used in examples.

would mean that every filter output value $\tilde{x}$ would be in the proper quantization cell, and no clipping would be necessary (i.e., the middle branch of (5) is always used). From this observation we arrive at the heuristic rule that desirable filters minimize the amount of clipping that is necessary.

We also use the quantizer thresholds $\{t_i\}$ at the receive side to normalize the pseudoerror and hence the pseudoerror statistics. This normalization is based on the quantization cell associated with the quantized sample $\hat{x}_i$. For example if $\hat{x}_i$ is located in the j^{th} quantization cell then $\hat{x}_i = r_j$ and the corresponding normalized pseudoerror p_i is

$$p_i = \frac{\hat{x}_i - \tilde{x}_i}{t_{j+1} - r_j} = \frac{r_j - \tilde{x}_i}{t_{j+1} - r_j}, \text{ when } r_j < \tilde{x}_i,$$
$$p_i = \frac{\hat{x}_i - \tilde{x}_i}{r_j - t_j} = \frac{r_j - \tilde{x}_i}{r_j - t_j}, \text{ when } r_j \geq \tilde{x}_i. \quad (6)$$

That is, when the adaptive processor moves a sample to above the representation point for that cell, this pseudoerror is normalized by the upper "half width" of the cell. When the adaptive processor moves a sample to below the representation point for that cell, this pseudoerror is normalized by the lower "half width" of the cell. In general, these two "half widths" are not equal, in the special case of the URQ they are equal. In light of the clipping step (5) and the normalization (6), the normalized pseudoerror signal is constrained to the interval [-1,1] for every quantization cell and every possible quantizer.

3.1. Matching One Pseudoerror Statistic

For one class of examples, quantization noise can be reduced by adaptive filtering that seeks to match a single normalized pseudoerror statistic. In this class of examples, a single filtering parameter is to be optimized, and a single statistic suffices to guide this optimization. The mean of the normalized pseudoerror signal in our examples is inherently zero, and thus this statistic is not useful. Instead we use the sample variance of the normalized pseudoerror signal σ_p^2.

In our first experiment, we consider Gaussian signals x with constantly changing spectral tilt, and that spectral tilt is constrained to be negative (power generally decreases with frequency) or zero (spectrum generally flat). For some negative spectral tilts and quantization noise floors, the signal will fall below the quantization noise floor for all frequencies above some critical frequency. A lowpass filter with a cutoff frequency positioned at this critical frequency will eliminate more quantization noise (error) than signal and will thus reduce quantization error. An example of this situation is given in Figure 5.

We have passed a set of 13 signals with negative spectral tilts through a set of 20 symmetric FIR low-pass filters (order 32) to produce Figure 6. This figure shows the resulting total error $E\left(\tilde{x}-x\right)^2$ relative to the original quantization error $E\left(\hat{x}-x\right)^2$ in dB as a function of σ_p^2. This relative quantization error

$$\gamma = 10 \cdot \log_{10}\left(\frac{E\left(\tilde{x}-x\right)^2}{E\left(\hat{x}-x\right)^2}\right) \quad (7)$$

is a measure of how much the pseudoinverse has reduced the original quantization error.

At the left edge of Figure 6, no filtering is done, $\sigma_p^2 = 0$, and there is no change in the quantization error. At the right edge, filtering is such that significant portions of the signal x are lost, σ_p^2 is large, and the pseudoinverse has actually increased the error. Between these two extremes, we see that there is a range of σ_p^2 values (about 0 to 0.5) that is consistent with a reduction in the quantization error. Further, for this set of signals and filters, the value $\sigma_p^2 \approx 0.2$ is related to the maximal reduction of quantization error and this is largely independent of signal spectral tilt or filter cutoff frequency.

Thus in this first experiment we use a target value of $\sigma_p^2 \approx 0.2$ to continuously select filters that best reduce the quantization error as the spectral tilt of the signal continues to change. Figure 7 shows an example of the relative quantization error reduction attained vs. bits/sample for an LMQ using 1 to 7 bits/sample. Note that the greatest reduction in quantization error is about 2.7 dB at b=3 bits/sample. As b increases above 3 the quantization noise floor falls, the signal level drops below that noise floor less frequently, and there is less to be gained by this technique. Also, as b decreases below 3, the quantization noise becomes more correlated with the signal, the model of a flat quantization noise floor becomes less valid, and there is less to be gained by this technique.

It is important to realize that the *only prior knowledge* about the signal that is available to the processor is the single fact that the spectral tilt is negative or zero. In operation, the spectral tilt of the signal x varies and the pseudoinverse continuously selects the filter that will best eliminate signal components that fall below the quantization noise floor. The selection or adaptation process is *based solely* on the constraint that the sample variance of the normalized pseudoerror aligns with a single fixed target value, i.e. $\sigma_p^2 \approx 0.2$. Unlike in conventional quantization, the thresholds $\{t_i\}$ are used at the receive side both in the clipping process, and in the pseudoerror normalization that is prerequisite to the calculation of σ_p^2.

In another, largely equivalent experiment, we attained similar results for the case of Gaussian signals with spectral tilt that is zero or positive. This experiment parallels the first, but a set of high-pass filters was used. Here again filter output selection was based solely on matching the sample variance of the normalized pseudoerror to a fixed target.

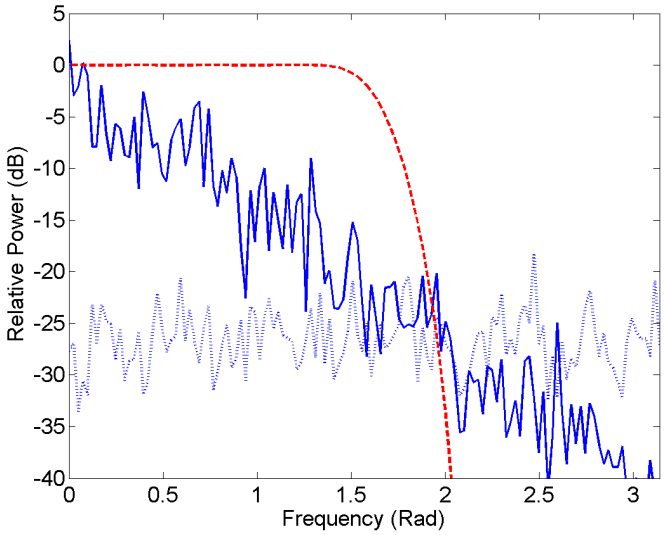

Fig 5. Example of signal with negative spectral tilt (solid), quantization noise (dotted), and response of a lowpass filter that can reduce quantization noise (dashed).

In additional work we considered the case of a bandpass signal with a fixed center frequency but an unknown bandwidth. Again based only on the matching of σ_p^2 to a target value, we were able to reliably select the bandpass filter that best reduced quantization error (because it best matched the signal). We also considered the case of a bandpass signal with a fixed bandwidth but an unknown center frequency and similar results were obtained.

Finally we have performed an additional cursory demonstration of this technique using a single image. Treating each row of pixels as a signal, some rows exhibit varying degrees of low pass nature and quantization noise can be reduced by low-pass filtering. But other rows may contain fine details that would be damaged by low-pass filtering, so a single fixed filter is not practical. For an image coarsely quantized to 3 bits/pixel, low-pass filtering to satisfy $\sigma_p^2 \approx 0.3$ resulted in an average reduction in quantization noise of about 0.7 dB. When 4 or more bits/pixel were used, no reduction in quantization noise was possible.

Throughout these examples, we have found that a single target value of normalized pseudoerror variance σ_p^2 in the range 0.15 to 0.50 serves to guide us to the proper filter. If an exact inverse were possible, it would have to produce a normalized pseudoerror that is uniformly distributed on [-1,1] (this range is due to the normalization in (6)) with variance $\sigma_p^2 = \frac{1}{3} = 0.33$ and kurtosis $\beta_p = 1.8$. But the pseudoinverses we have constructed use linear filters followed by a clipper. These linear filters produce approximate Gaussian outputs (since the output is a linear combination of quantized Gaussian input samples). The pseudoerror is the difference between this approximate Gaussian signal and a partially correlated, quantized Gaussian signal. As a consequence, the pseudoerror and normalized pseudoerror are approximately Gaussian. Thus one way to describe the task of the pseudoinverse is that it must produce a pseudoerror signal that (when normalized) approximates a uniform distribution on [-1,1] and it must do this by means of an approximate Gaussian distribution that is clipped at ±1.

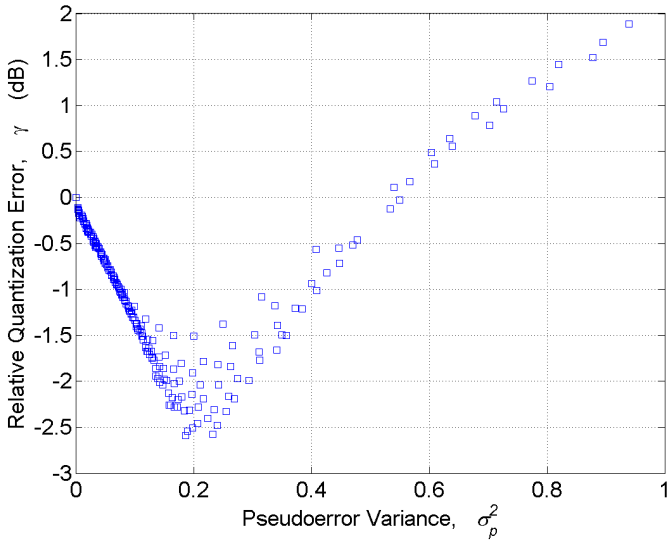

Fig 6. Relative quantization error as a function of normalized pseudoerror variance.

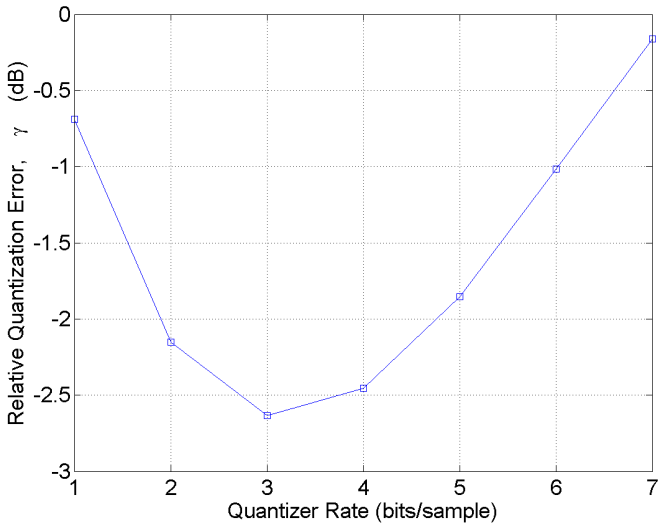

Fig 7. Relative quantization error for Gaussian signals with negative spectral tilt for LMQ using 1 to 7 bits/sample.

On the other hand, the discussion immediately following (5) shows that desirable filters minimize the amount of clipping that is necessary. This minimal clipping will lead to a pseudoerror signal that is far closer to Gaussian than to uniform. Thus the pseudoinverse is caught between these two competing goals: approximate a uniform pseudoerror distribution via a clipped approximate Gaussian distribution, yet minimize the number of samples that must be clipped.

The compromise solutions fall into a range, depending on the exact problem. As noted above, the target variance for the normalized pseudoerror tends to fall in the range 0.15 to 0.50, which includes the variance of uniformly distributed error (0.33). Approximately 1 to 5% of the samples tend to be clipped in practice. The normalized pseudoerror kurtosis tends to range from 2.0 to 3.5. This nearly covers the range from the uniform kurtosis (1.8) to the Gaussian kurtosis (3.0) and extends beyond. In all of our examples, we have seen that one can use a clipping target (in the range of 1 to

5%) in place of a variance target and obtain nearly identical results. Another option that yields similar results is to set a target for the 90th or 95th percentile of the normalized pseudoerror values.

Reductions in quantization noise always depend on how signal power and quantization noise power are distributed across frequency. If there are large portions of the spectrum where the quantization noise power is well above the signal power, then greater reductions are possible. If this is not the case, then smaller reductions, if any, will be possible.

In all of this work we have used a target value of σ_p^2 to select the best filter from a bank of fixed filters. We argue that σ_p^2 could also be used to adapt a single filter to an optimal state in many cases. Figure 8 shows examples of the empirical relationships between σ_p^2 and cutoff frequency for low-pass filters operating on 3 different signals with negative spectral tilt. The relationships are monotonically decreasing. Independent of the signal, when σ_p^2 is below target the filter cutoff frequency (f_c) should be decreased, and when σ_p^2 is above target the filter cutoff frequency should be increased. A local estimate of the slope $\partial\sigma_p^2/\partial f_c$ could be used to quickly converge on the cutoff frequency that gives the desired value of σ_p^2. Analogous situations exist for the cases of high-pass filtering and bandpass filtering with a known center frequency and an unknown bandwidth.

The case of bandpass filtering with a known bandwidth and an unknown center frequency is more complex because ambiguities can arise between the cases where the filter passband is positioned too high and too low. Here additional steps exploring both the higher and lower frequencies may be required to arrive at the proper center frequency.

3.2. Matching Two Pseudoerror Statistics

Next we consider a more general problem that involves two degrees of freedom. As an example, we generated a variety of bandpass signals with center frequencies and bandwidths that were unknown at the receiving side. Because there are now two degrees of freedom (center frequency and bandwidth) we lose the opportunity to optimize a filter using a single simple monotonic relationship like that shown in Figure 8. However, intuition suggests that the use of a second pseudoerror statistic might help to address this problem, and this turns out to be true in this example. Like the normalized pseudoerror mean, the normalized pseudoerror skewness is not useful. But the sample kurtosis serves as a useful second normalized pseudoerror statistic.

We sent the received quantized signals through a set of bandpass filters of various center frequencies and bandwidths. For each combination of signal and filter, we have calculated the normalized pseudoerror variance and kurtosis, as well as the change in quantization error associated with this combination of signal and filter. Figure 9 shows the resulting values of variance and kurtosis. In that figure, when a filter decreases quantization error, that data point is represented with a triangle. When a filter increases quantization error, a circle is used. From this figure it is clear that in this example, a single target or threshold on variance will not lead us to consistent decreases in quantization error. The same is true for kurtosis. But together, the pair of normalized pseudoerror statistics can be used to select filters that will reduce quantization error in

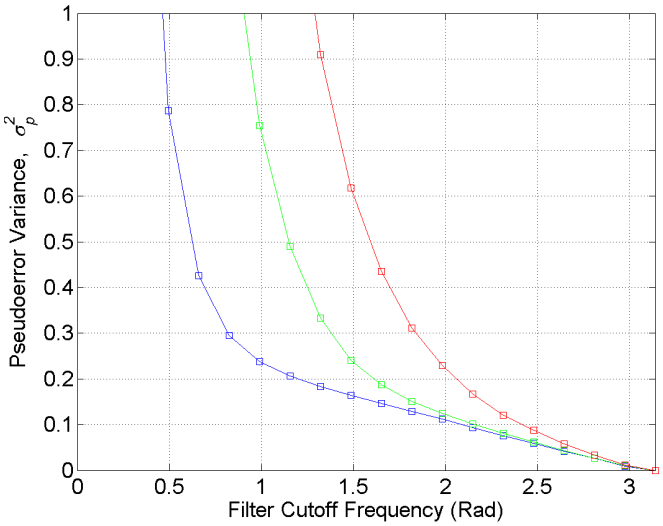

Fig 8. Monotonic relationships between filter cutoff frequency and normalized pseudoerror variance for three signals with negative spectral tilt.

the vast majority of the cases. For example, one might select filters where the normalized pseudoerror variance and kurtosis conform with the linear discriminant function

$$\beta_p > 1.8 + 1.7 \cdot \sigma_p^2 . \qquad (8)$$

This function is shown as a broken line in the figure. This is an example result and other selection rules will likely be more appropriate in other situations.

Adapting a single filter to obtain pseudoerrors with these properties remains as a development project. In this present example, variance decreases and kurtosis increases as the filter center frequency moves towards the signal center frequency. Further, this rate of change is greater when the filter width is matched to the signal width and smaller when these widths are not well matched. These observations may facilitate the development of a single adaptive filter that produces the desired pseudoerror.

4. SUMMARY AND INTERPRETATIONS

We have proposed a technique for reducing the quantization error associated with scalar quantization. The technique is implemented at the receive side only, and it uses the quantization thresholds and a small amount of prior knowledge of the signal. No increase in the transmitted data rate is required. The key ideas are a quantizer pseudoinverse and the associated pseudoerror signal. To remove the quantization error it is necessary, but not sufficient, to subtract a pseudoerror signal bearing statistics that match (to the maximum extent possible) the statistics of the original error signal.

We have demonstrated that matching one or more statistics of the normalized pseudoerror (variance, fraction of samples clipped, value of 90th or 95th percentile, kurtosis) to a fixed target can indeed guide an adaptive processor to reduce quantization error in a range of cases including low-pass, high-pass and band-pass Gaussian signals, as well as a coarsely quantized image. We have noted that while the quantization error is typically a uniformly distributed signal, the pseudoerror is much closer to a clipped Gaussian signal. The amount that quantization error can be reduced depends on the relationship

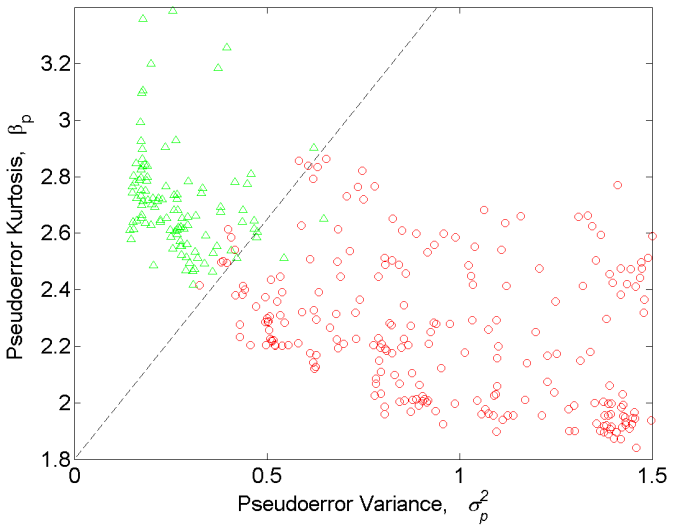

Fig 9. Together, variance and kurtosis can separate filters that reduce quantization error (triangles) from filters that increase quantization error (circles) across a wide range of signals.

between the signal spectrum and the quantization noise floor. Our examples have used a set of fixed filters and clippers followed by a selector switch, but we have also addressed the key issues associated with the extension to processing structures that are more innately adaptive.

A time-domain or per-sample interpretation centers on refinement of quantizer output sample values. The conventional quantizer produces a single value r_j for all input samples with values in the j^{th} quantization cell. The pseudoinverse uses some prior knowledge of the input signal to attempt to refine the value r_j to a value that is closer to the true input value. We can judge the merit of these attempted refinements (and hence judge the appropriateness and effectiveness of the pseudoinverse) without access to the input signal by considering some aggregate properties (statistics) of the refinements (the pseudoerror signal) themselves. For example, are the refinements sufficiently spread across the widths of the quantization cells (is σ_p^2 big enough)? Do the refinements result in refined sample values that are mostly within the proper quantization cell (is σ_p^2 small enough, is the percentage of samples clipped small enough)?

A frequency-domain or spectral interpretation focuses on matching filters to the original input signal, but without access to that input signal. When the match is a good one, minimal signal is removed but maximal quantization noise is removed. We can judge the match (and hence judge the appropriateness and effectiveness of the pseudoinverse) without any access to the input signal by considering properties of the signal removed by the filtering (the pseudoerror signal). Is that signal variance approximately consistent with the variance of the quantization noise that we wish to remove (is σ_p^2 close to the target value)? To further enhance this process, one might invoke one or more band-limited spectral flatness measure(s) covering the filter stopband(s). If the signal removed (pseudoerror signal) is flatter, that will often indicate that more quantization noise and less signal is being removed (an indication of a better match). If the signal removed is less flat, that will often indicate that more signal and less quantization noise is being removed (an indication of a worse match).

5. REFERENCES

[1] J. Max, "Quantizing for Minimum Distortion," *IEEE Trans. Information Theory,* vol. 6, no. 1, pp. 7-12, Mar. 1960.

[2] R. Gray & D. Neuhoff, "Quantization," *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2325-2383, Oct. 1998.

[3] H. Knagenhjelm & B. Kleijn, "Spectral Dynamics is More Important than Spectral Distortion," in *Proc. IEEE ICASSP '95,* Detroit, 1995.