NOISE ANALYSIS OF MODULATED QUANTIZER BASED ON OVERSAMPLED SIGNALS

Charlotte Yuk-Fan Ho

Dept. of Electronic Engineering and School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London, E1 4NS, U.K.

Bingo Wing-Kuen Ling

Dept. of Electronic Engineering, Division of Engineering, King's College London, Strand, London, WC2R 2LS, U.K.

Joshua D. Reiss

Dept. of Electronic Engineering, Queen Mary, University of London, Mile End Road, London, E1 4NS, U.K.

ABSTRACT

In this paper, a noise analysis of a modulated quantizer is performed. If input signals are oversampled, then the quantization error could be reduced by modulating both the input and the output of the quantizer. The working principle is based on the fact that convolutions of bandpass signals would spread wider in the frequency spectrum than that of lowpass signals. Hence, by filtering the high frequency components, the signal-to-noise ratio (SNR) could be increased. Numerical simulation results show that the modulated quantization scheme could achieve an average of 13.0960dB to 21.4700dB improvements on SNR over the conventional scheme, depends on the types of bandlimited input signals.

1. INTRODUCTION

Quantization is widely employed in many signal processing applications, such as in data compression [1] and analog-to-digital conversion [2], etc. However, as quantization is not a reversible process because it is a many-to-one mapping, the system cannot be recovered once it is generated [3]. Hence, it is very important to minimize the quantization error.

The most common method to minimize the quantization error is based on the statistics of input signals [4]. Finer resolutions are assigned to the ranges of input signals which occur most frequently, and vice versa. However, this kind of quantization scheme requires a prior knowledge of statistics of inputs signals. In many situations, the statistics of input signals are unknown and this method cannot be applied directly.

Another common method to minimize the quantization error is via a sigma delta modulation technique [5]-[7]. If input signals are oversampled, then the signal band is very narrow. Hence, the overlap between the noise spectrum and the signal band is small. As a result, a very high SNR can be achieved. In this paper, we further utilize the oversampling technique to reduce the quantization error. The input and the output of a quantizer are modulated via a bank of modulators. Based on the obtained numerical simulation results, an average of 13.0960dB to 21.4700dB improvements on SNR over the conventional scheme can be achieved.

The outline of this paper is as follow: In Section 2, an approximated model for the quantizer is introduced. Based on the model, detail error analysis is performed. It is shown that the quantization error could be reduced by applying a modulation technique on the input and output of the quantizer. In Section 3, we further extend the results in Section 2 from a single modulator to a bank of modulators. Finally, a conclusion and future work is summarized and discussed in Section 4.

2. REDUCTION OF NOISE VIA MODULATED QUANTIZER

The block diagrams of systems using a conventional quantizer and a modulated quantizer are shown in, respectively, Figure 1a and Figure 1b. Denote the input to these two quantizers, the quantizers, the frequency response of these two linear time-invariant filters, the output of the conventional quantizer, that of the modulated quantizer, the reconstructed signal using the conventional quantizer, and that of using the modulated quantizer as, respectively, $u(k)$, $Q(\cdot)$, $H(\omega)$, $s_1(k)$, $s_2(k)$, $y_1(k)$ and $y_2(k)$. We assume that $u(k)$ is oversampled. That means, most of the energy of $u(k)$ is within the frequency band $\left(-\frac{\pi}{R},\frac{\pi}{R}\right)$, where R is the oversampling ratio. Consider an N bit quantizer with the quantization range $[-L,L]$. Then

$$Q(y)\equiv\begin{cases}\Delta sign(y)\left(ceil\left(\frac{|y|}{\Delta}\right)-\frac{1}{2}\right) & |y|\le L\\ sign(y)\left(2^{N-1}-\frac{1}{2}\right) & |y|>L\end{cases}, \qquad (1)$$

where $sign(y) \equiv \begin{cases} \frac{y}{|y|} & y \neq 0 \\ 0 & y = 0 \end{cases}$, $ceil(y)$ denotes the rounding operator towards the plus infinity, $|\cdot|$ denotes the absolute operator, and $\Delta \equiv \frac{L}{2^{N-1}}$ is the step size of the quantizer. To approximate $Q(y)$ as a polynomial of y, denoting $\mathbf{y} \equiv \begin{bmatrix} y & y^3 & \cdots & y^{2M-1} \end{bmatrix}^T$ and $\mathbf{p} \equiv \begin{bmatrix} p_1 & \cdots & p_M \end{bmatrix}^T$, where the superscript ^{T} denotes the transpose operator, p_m for $m = 1,2,\cdots,M$ and $2M-1$ are, respectively, the coefficients and the order of the polynomial of y, then $\mathbf{p}$ can be found via solving the optimization problem with the objective being minimizing the total absolute square difference between the actual quantizer and the approximated quantizer, that is:

$$\min_{\mathbf{p}} \int_{-L}^{L} \left| \mathbf{y}^T \mathbf{p} - Q(y) \right|^2 dy \cdot \tag{2}$$

The solution of this optimization problem is $\mathbf{p} = -\mathbf{A}^{-1}\mathbf{b}$, where $\mathbf{A} \equiv 2\int_{-L}^{L} \mathbf{y}\mathbf{y}^T dy$ and $\mathbf{b} \equiv -2\int_{-L}^{L} Q(y)\mathbf{y} dy$. Figure 2 shows examples of input-output relationships of actual quantizers with $L = 1$ and the approximated quantizers $\mathbf{y}^T\mathbf{p}$ with $M = 10$ for 1-bit, 2-bit and 8-bit cases. Figure 3 show the corresponding differences, that is $Q(y) - \mathbf{y}^T\mathbf{p}$. It can be seen from Figure 3 that the differences between the actual quantizers and the approximated quantizers get smaller and smaller as N increases. Hence, the approximation is valid.

Now, let's analyze the quantization noise using the above approximated model. That is, replacing the actual quantizer $Q(y)$ by the approximated quantizer $\mathbf{y}^T\mathbf{p}$. Denote the Fourier transform of $u(k)$, $s_1(k)$, $s_2(k)$, $y_1(k)$ and $y_2(k)$ as, respectively, $U(\omega)$, $S_1(\omega)$, $S_2(\omega)$, $Y_1(\omega)$ and $Y_2(\omega)$. Denote $U_{2m-1}(\omega) \equiv U(\omega) * \cdots * U(\omega)$, where $*$ denotes the convolution operator and there are $2m-1$ terms in $U_{2m-1}(\omega)$. For the system with the conventional quantizer shown in Figure 1a,

$$Y_1(\omega) = H(\omega)S_1(\omega) \approx H(\omega)\sum_{m=1}^{M} p_m U_{2m-1}(\omega) \cdot \tag{3}$$

Since we assume that $u(k)$ is oversampled, $U(\omega)$ is approximately bandlimited within $\left(-\frac{\pi}{R}, \frac{\pi}{R}\right)$. As a result, $U_{2m-1}(\omega)$ is approximately bandlimited within $\left(-\frac{(2m-1)\pi}{R}, \frac{(2m-1)\pi}{R}\right)$. However, since all these M terms have zero center frequency, all the higher order terms are overlapped to the signal band $\left(-\frac{\pi}{R}, \frac{\pi}{R}\right)$. If we regard all higher order terms ($m \geq 2$) as the quantization noise, then the quantization noise would corrupt the signal seriously. Since

$$SNR \approx 10\log_{10} \frac{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| H(\omega) p_1 U(\omega) \right|^2 d\omega}{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| H(\omega) \sum_{m=2}^{M} p_m U_{2m-1}(\omega) \right|^2 d\omega}, \tag{4a}$$

if we further assume that $H(\omega)$ is an ideal lowpass filter with $H(\omega) = \begin{cases} 1 & |\omega| < \frac{\pi}{R} \\ 0 & otherwise \end{cases}$, then

$$SNR \approx 10\log_{10} \frac{p_1^2 \int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| U(\omega) \right|^2 d\omega}{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| \sum_{m=2}^{M} p_m U_{2m-1}(\omega) \right|^2 d\omega}, \tag{4b}$$

which would be quite low for the conventional quantizer.

Now consider the system with modulators as shown in Figure 1b. Denote the input to the quantizer as $\tilde{u}(k)$ and $\tilde{U}_{2m-1}(\omega) \equiv \tilde{U}(\omega) * \cdots * \tilde{U}(\omega)$, in which there are $2m-1$ terms in $\tilde{U}_{2m-1}(\omega)$. Then

$$\tilde{U}_{2m-1}(\omega) = \frac{U_{2m-1}(\omega)}{2^{2m-1}} * \sum_{r=0}^{2m-1} \frac{(2m-1)!}{r!(2m-1-r)!} \delta(\omega + (2m-2r-1)\omega_0), \tag{5a}$$

$$S_2(\omega) \approx \sum_{m=1}^{M} \left(\frac{p_m U_{2m-1}(\omega)}{2^{2m-1}} * \sum_{r=0}^{2m-1} \frac{(2m-1)!}{r!(2m-1-r)!} \delta(\omega + (2m-2r-1)\omega_0) \right), \tag{5b}$$

and

$$Y_2(\omega) \approx H(\omega)\sum_{m=1}^{M} \left(\frac{p_m U_{2m-1}(\omega)}{2^{2m}} * \sum_{r=0}^{2m} \frac{(2m)!}{r!(2m-r)!} \delta(\omega + 2(m-r)\omega_0) \right), \tag{5c}$$

where ! denotes the factorial operator. Hence,

$$SNR \approx 10\log_{10} \frac{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| H(\omega) \frac{p_1 U(\omega)}{2^2} * \sum_{r=0}^{2} \frac{(2)!}{r!(2-r)!} \delta(\omega + 2(1-r)\omega_0) \right|^2 d\omega}{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}} \left| H(\omega) \sum_{m=2}^{M} \left(\frac{p_m U_{2m-1}(\omega)}{2^{2m}} * \sum_{r=0}^{2m} \frac{(2m)!}{r!(2m-r)!} \delta(\omega + 2(m-r)\omega_0) \right) \right|^2 d\omega}. \tag{6a}$$

If ω_0 are selected in such a way that $\omega_0 \geq \frac{(2M-1)\pi}{R}$, then the mirror signals $U_{2m-1}(\omega + 2(m-r)\omega_0)$ for $r = 0,1,\cdots 2m$ and for $m = 1,2,\cdots M$ do not overlap each others in the frequency spectrum. Hence, (6a) can be further simplified as:

$$SNR \approx 10\log_{10}\frac{\frac{p_1^2}{4}\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}}|U(\omega)|^2 d\omega}{\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}}\left|\sum_{m=2}^{M}\frac{p_m U_{2m-1}(\omega)(2m)!}{2^{2m}(m!)^2}\right|^2 d\omega}. \quad (6b)$$

Since $2^{2m}=\sum_{r=0}^{2m}\frac{(2m)!}{r!(2m-r)!}$, if $\sum_{\substack{r=0\\ r\neq m}}^{2m}\frac{(2m)!}{r!(2m-r)!}>\frac{3(2m)!}{(m!)^2}$, then $\frac{4(2m)!}{2^{2m}(m!)^2}<1$ and

$$\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}}\left|\sum_{m=2}^{M}p_m U_{2m-1}(\omega)\right|^2 d\omega > 4\int_{-\frac{\pi}{R}}^{\frac{\pi}{R}}\left|\sum_{m=2}^{M}\frac{p_m U_{2m-1}(\omega)(2m)!}{2^{2m}(m!)^2}\right|^2 d\omega. \quad (7)$$

By comparing (4b) to (6b), the modulated system will provide improvement on SNR compared to the conventional system.

To verify the approximations and the above analysis, we have performed some simulation results. Denote $u'(k)$ by a random signal with zero mean uniform distribution between -1 and 1. The bandlimited input is generated via filtering $u'(k)$ through $H(\omega)$ and normalizing the maximum absolute value to 1, that is $U(\omega)=\frac{U'(\omega)H(\omega)}{K}$, where $U'(\omega)$ is the Fourier transform of $u'(k)$ and K is selected such that $\max_{\forall k\geq 0}|u(k)|=1$. In the following simulation results, we choose an elliptic filter with the following transfer function:

$$H(z)=\frac{10^{-4}\left(0.0761-0.2027z^{-1}+0.1283z^{-2}+0.1283z^{-3}-0.2027z^{-4}+0.0761z^{-5}\right)}{1-4.8675z^{-1}+9.4816z^{-2}-9.2392z^{-3}+4.5036z^{-4}-0.8785z^{-5}}$$

as the ideal lowpass filter because this filter can be obtained easily from the Matlab toolbox. Also, the saturation level of the quantizer is selected as 1, that is $L=1$. This is because of the normalization reason. Moreover, we select the oversampling ratio as $R=64$ because this is the most common value employed in industry. Figure 4 shows the improvements of SNR of the modulated quantizer over the conventional quantizer, where $SNR\equiv 10\log_{10}\frac{\sum_{\forall k\geq 0}|u(k)|^2}{\sum_{\forall k\geq 0}|u(k)-y_i(k)|^2}$ for $i=1,2$. It is worth noting that the equation for calculating SNR here is different from that in the previous Section because the one in the previous Section is based on the approximated model, while the one in this Section is from the definition. According to the simulation results, it can be seen from Figure 4a that there is an average of 5.3136dB improvement when $\omega_0=\frac{\pi}{R}$ and 5.6084dB improvement when $\omega_0\geq\frac{3\pi}{R}$, but there is no significant change on the improvement when the modulating frequency is higher than $\frac{3\pi}{R}$. Compared to the theory we have developed, that is, if $\omega_0\geq\frac{(2M-1)\pi}{R}$ and $\sum_{\substack{r=0\\ r\neq m}}^{2m}\frac{(2m)!}{r!(2m-r)!}>\frac{3(2m)!}{(m!)^2}$, then the SNR could be improved, it is interesting to see from Figure 4a that when $\omega_0\geq\frac{3\pi}{R}$ it is already enough to satisfy the condition. Besides, there is an average of 6.9007dB improvement when $N=1$ and the average improvement drops monotonically and converges to 5.3070dB when $N=16$. This is because as N increases, the effects of nonlinearity decrease. As a result, the improvement based on the modulation technique will be less significant. Figure 4b shows the corresponding results for a sinusoidal input $u(k)=\sin\left(\frac{2\pi k}{3R}\right)$ for $k\geq 0$. We choose this sinusoidal input because this operating frequency is the most common test frequency employed for the analog-digital conversion and the magnitude of the sinusoidal input is chosen to be 1 because of the normalization reason. It can be seen from Figure 4b that there is an average of 5.3821dB improvement when $\omega_0=\frac{\pi}{R}$ and 5.7670dB improvement when $\omega_0\geq\frac{3\pi}{R}$, but there is no significant change on the improvement when the modulating frequency is higher than $\frac{3\pi}{R}$. This phenomenon occurs similarly for the bandlimited random input case. However, we observe that there is an average of 4.8981dB improvement when $N=1$ and the average improvement increases and converges to 5.7619dB when $N=16$ for the sinusoidal input.

3. EXTENSION FROM A SINGLE MODULATOR TO A BANK OF MODULATORS

The technique discussed in Section 2 can actually be further extended to the case if a bank of modulators is employed. Denote N_q as the number of modulators employed in the system as shown in Figure 5. Figure 6 show simulation results of various quantizers with same values of L, R, and the filter as in the previous Section. It can be seen from Figure 6a that there is an average of 6.2730dB improvement when $N_q=1$ and the average improvement increases and converges to 11.0943dB when $N_q=40$. Besides, there is an average of 19.6950dB improvement when $N=1$ and the average improvement decreases monotonically and converges to 9.4297dB when $N=16$ for a bandlimited random input. Figure 6b shows the corresponding results for a sinusoidal input. It can be seen from Figure 6b that there is an average of 5.7679dB improvement when $N_q=1$ and the

average improvement increases and converges to 11.0636dB when $N_q = 40$. Besides, there is an average of 12.1371dB improvement when $N = 1$ and the average improvement decreases monotonically and converges to 10.3246dB when $N = 16$. According to the simulation results, it is found that the highest improvement occurs at $N = 1$ and $N_q = 30$ for both a bandlimited random input and a sinusoidal input. The corresponding improvements are 21.4700dB and 13.0960dB, respectively.

4. CONCLUSION

In this paper, we propose to employ a bank of modulators for reducing the quantization error. Since bandpass signals spread wider in the frequency spectrum than that of the lowpass signals, quantization error could be reduced by filtering the high frequency components. Numerical simulation results show that an average of 13.0960dB to 21.4700dB improvements on SNR over the conventional scheme could be achieved. It is worth noting that this technique is different from the dithering approach because a signal is *added* to the quantizer output for the dithering approach, while we propose to *multiply* a signal at the input and the output of the quantizer.

5. ACKNOWLEDGEMENTS

The work obtained in this paper was supported by a research grant from Queen Mary, University of London.

6. REFERENCES

[1] E.H. Yang, and Z. Zhang, "An On-line Universal Lossy Data Compression Algorithm via Continuous Codebook Refinement—Part III: Redundancy Analysis," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1782-1801, 1998.

[2] D.E. Quevedo, and G.C. Goodwin, "Multistep Optimal Analog-to-Digital Conversion," *IEEE Transactions Circuits and Systems—I: Regular Papers*, vol. 52, no. 3, pp. 503-515, 2005.

[3] N.T. Thao, "Vector Quantization Analysis of ΣΔ Modulation," *IEEE Transactions on Signal Processing*, vol. 44, no. 4, pp. 808-817, 1996.

[4] M.G. Strintzis, and D. Tzovaras, "Optimal Pyramidal Decomposition for Progressive Multidimensional Signal Coding Using Optimal Quantizers," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 1054-1068, 1998.

[5] C.Y.F. Ho, B.W.K. Ling, J.D. Reiss, and X. Yu, "Nonlinear Behaviors of Bandpass Sigma Delta Modulators with Stable System Matrices," to appear in *IEEE Transactions on Circuits and Systems—II: Express Briefs*.

[6] C.Y.F. Ho, B.W.K. Ling, and J.D. Reiss, "Fuzzy Impulsive Control of High Order Interpolative Lowpass Sigma Delta Modulators," to appear in *IEEE Transactions on Circuits and Systems—I: Regular Papers*.

[7] C.Y.F. Ho, B.W.K. Ling, J.D. Reiss, Y.Q. Liu, and K.L. Teo, "Design of Interpolative Sigma Delta Modulators via Semi-infinite Programming," to appear in *IEEE Transactions on Signal Processing*.

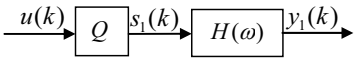

Figure 1a. A system using conventional quantizer.

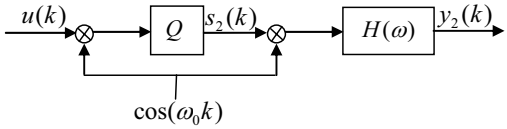

Figure 1b. A system using modulated quantizer.

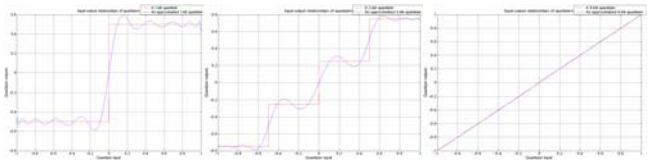

Figure 2. Input-output relationships of the original quantizers and the approximated quantizers. (a) 1-bit case. (b) 2-bit case. (c) 8-bit case.

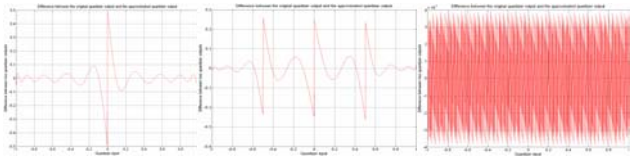

Figure 3. Differences between the original quantizers and the approximated quantizers. (a) 1-bit case. (b) 2-bit case. (c) 8-bit case.

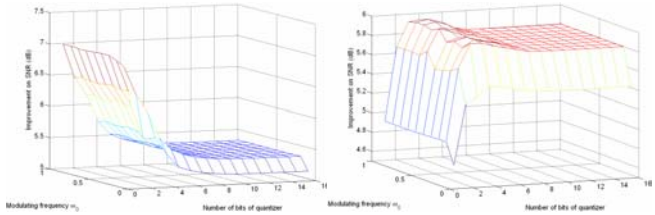

Figure 4. Effect of different number of bits of quantizers and modulating frequencies on the improvements of SNR. (a) a bandlimited random input. (b) a sinusoidal input.

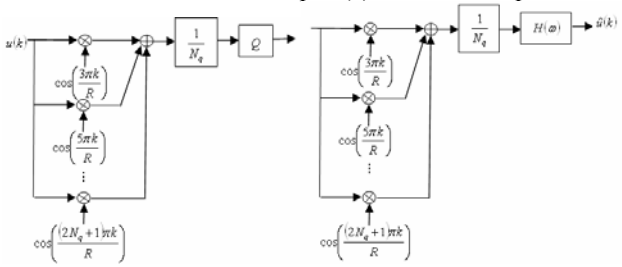

Figure 5. A system for noise reduction using a bank of modulators.

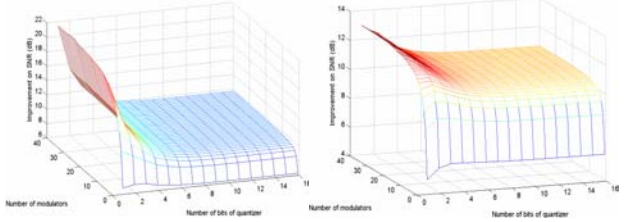

Figure 6. Effect of different number of bits of quantizers and number of modulators on the improvements of SNR. (a) a bandlimited random input.(b) a sinusoidal input.