Effect of Various Dither Forms on Quantization Errors of Ideal A/D Converters

MAHMOUD FAWZY WAGDY, SENIOR MEMBER, IEEE

Abstract—**The quantization error of a quantizer (ideal A/D converter (ADC)) is investigated. First, correlation between quantization error and quantizer input is considered. The input signal is taken as a sinusoid due to its importance in instrumentation systems. The cases of no dither, uniform dither, and discrete (digital) dither are considered. Effects of the dither probability density function (PDF) are discussed. The relationship between uniform dithered and discrete dithered quantizer inputs is derived.**

Second, spectra of the average quantization error, corresponding to an arbitrary input signal, are investigated. Different dither forms (Gaussian, uniform, and discrete) are compared and the effects of the dither PDF are discussed.

The paper provides a quantitative basis for comparing dither forms, and hence, selecting the one most appropriate for a particular application.

I. Inttroduction

DITHERING has been used in the past thirty years to reduce the effects of quantization noise on speech and visual signals. Roberts [1] showed that using dither signals with coarsely quantized TV pictures masks undesirable contours. Jayant and Rabiner [2] used the known idea of add-subtract dither in speech signals. Vanderkooy and Lipshitz [3] used dither to resolve audio signals smaller than the quantizing step. In their last paper, Vanderkooy and Lipshitz [4] continued their analysis on audio signals. They distinguished between various forms of dither and quantified their effects on differential and large signal nonlinearity, noise modulation, and total noise level.

Dithering has been effective recently in improving the resolution of certain systems. Holland [5] described a method for improving the resolution and linearity of conventional A/D converters by whitening quantization noise then reducing it by digital filtering. Carley [6] employed dithering techniques in a specific topology for an oversampling differential pulse code modulation (DPCM)-type A/D converter (ADC) to improve the linearity of ADC transfer function. White and Pickup [7] used dither to decrease the effects of quantization and differential nonlinearity on systematic errors of digital cross correlators.

From the above survey it is evident that more research should be done on dithering. Investigation of signal-noise correlation for different dither forms and PDF's is in order. Another potential research object is to investigate the effect of different dither forms and PDF's on the amplitude spectrum of quantization error. Quantified results in these two areas will help choose the most appropriate dither form and PDF for a particular application.

Manuscript received August 10, 1988; revised March 3, 1989.
The author is with the Department of Electrical Engineering, University of Lowell, Lowell, MA 01854.
IEEE Log Number 8928359.

II. Correlation Between Quantization Error and Quantizer Input

The correlation, R, between the quantizer input, y, and quantization error, e, is given by [2]

$$R = E(y \cdot e) \tag{1}$$

where E is the expectation. For a bipolar ADC, the correlation for the kth quantization slot would be given by

$$R(k) = \int_{(k-1/2)\Delta}^{(k+1/2)\Delta} (k-y)y \cdot f_y(y)\, dy \tag{2}$$

where Δ is the quantization step, and $f_y(y)$ is the PDF of y. By normalizing (2) with respect to y, which is equivalent to letting $\Delta = 1$, then

$$R(k) = \int_{k-1/2}^{k+1/2} (k-y)y \cdot f_y(y)\, dy. \tag{3}$$

Three cases are now considered.

A. With No Dither

The PDF of the undithered sinusoid, $y = A \sin \omega t$, is given by

$$f_y(y) = \frac{1}{\pi\sqrt{A^2 - y^2}} \tag{4}$$

where A is the amplitude of the sinusoid. By plugging (4) into (3), then:

$$R(k) = \int_{k-1/2}^{k+1/2} \frac{(k-y)y}{\pi\sqrt{A^2-y^2}}\, dy. \tag{5}$$

It can be shown that:

$$R(k) = -\frac{1}{\pi}[\phi(x_2) - \phi(x_1)]$$

$$\phi(x_i) = k\sqrt{A^2 - x_i^2} + \frac{A^2}{2}\sin^{-1} x_i - \frac{x_i}{2}\sqrt{A^2 - x_i^2}$$

$$x_1 = k - \frac{1}{2}, \quad x_2 = k + \frac{1}{2}. \tag{6}$$

0018-9456/89/0800-0850$01.00 © 1989 IEEE

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.

B. With Uniform Dither

A dither signal whose PDF is given by

$$f_d(d) = \begin{cases} \frac{1}{2c}, & -c \le d \le c \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

is added to the input signal whose PDF is given by (4). $f_y(y)$ is derived by convolving (4) and (7). Two cases are considered.

1) $c < A$: When the dither amplitude is smaller than the sinusoid amplitude, the correlation for the kth quantization slot is derived. Using mathematical tables [8] for integration it can be shown that

$$R(k) = \frac{1}{2\pi}[\phi(x_2) - \phi(x_1) - \phi(x_4) + \phi(x_3)],$$
$$|k| \le A - c - \frac{1}{2} \quad (8.a)$$

where:

$$\phi(x_i) = -\left(\frac{A^3}{3c}\right)x_i^3 \sin^{-1} x_i + A^2\left(\frac{k}{2c} + 1\right)x_i^2 \sin^{-1} x_i - A(k+c)x_i \sin^{-1} x_i - A^2\left(\frac{k}{4c} + \frac{1}{2}\right)\sin^{-1} x_i - \left(\frac{A^3}{9c}\right)x_i^2\sqrt{1 - x_i^2} + A^2\left(\frac{k}{4c} + \frac{1}{2}\right)x_i\sqrt{1 - x_i^2} - A\left(k + c + \frac{2A^2}{9c}\right)\sqrt{1 - x_i^2}$$

$$x_{1,2} = \frac{k \mp \frac{1}{2} + c}{A}, \quad x_{3,4} = \frac{k \mp \frac{1}{2} - c}{A}. \quad (8.b)$$

Also:

$$R(k) = \frac{1}{2\pi}[\phi(x_2) - \phi(x_1)],$$
$$A - c + \frac{1}{2} \le |k| \le A - c - \frac{1}{2} \quad (9.a)$$

where:

$$\phi(x_i) = -\left(\frac{A^3}{3c}\right)x_i^3 \cos^{-1} x_i + A^2\left(\frac{k}{2c} - 1\right)x_i^2 \cos^{-1} x_i + A(k-c)x_i \cos^{-1} x_i - A^2\left(\frac{k}{4c} - \frac{1}{2}\right)\cos^{-1} x_i - \left(\frac{A^3}{9c}\right)x_i^2\sqrt{1 - x_i^2} - A^2\left(\frac{k}{4c} - \frac{1}{2}\right)x_i\sqrt{1 - x_i^2} - A\left(k - c + \frac{2A^2}{9c}\right)\sqrt{1 - x_i^2}$$

$$x_{1,2} = \frac{k \mp \frac{1}{2} - c}{A} \quad (9.b)$$

The variation of $R(k)$ with k for different values of c is shown in Fig. 1. In this example, A is taken as 40.5Δ, i.e., a normalized value of 40.5. Due to the symmetry of the quantizer characteristics (bipolar ADC), it is sufficient to show $R(k)$ for only positive k values. It is obvious that as c increases, the quantization error is less correlated to the quantizer input, an advantage of larger amplitude dither.

2) $c > A$: When the dither amplitude is greater than the sinusoid amplitude, derivations utilizing mathematical tables [8] lead to

$$R(k) = 0, \quad |k| \le c - A - \frac{1}{2} \quad (10)$$

$$R(k) = \frac{1}{2\pi}[\phi(x_2) - \phi(x_1)],$$
$$c - A + \frac{1}{2} \le |k| \le c + A - \frac{1}{2} \quad (11)$$

where $\phi(x_i)$, x_1, and x_2, are given by (9.b).

The variation of $R(k)$ with positive k for the example cases ($A = 8.5\Delta$ and $c = 16\Delta, 32\Delta$) is shown in Fig. 2. It is obvious that as c increases $|R(k)|$ decreases, an advantage, again, of larger amplitude dither.

C. With Discrete (Digital) Dither

Discrete (digital) dither is typically applied via a D/A converter [4]. It is sometimes desirable to add discrete dither to the signal at the ADC input. The PDF of discrete dither is in the form of two or more equispaced impulses with equal weights. Two cases are considered.

1) A PDF With Two Impulses: With two impulses, let the PDF of dither be given by

$$f_d(d) = \begin{cases} \frac{1}{2}, & d = c, -c \text{ with } c < A \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

After derivations, using mathematical tables [8] for integration, it can be shown that

$$R(k) = \frac{1}{2\pi}\left[\phi\left(k + \frac{1}{2}\right) - \phi\left(k - \frac{1}{2}\right)\right]. \quad (13)$$

For $|k| \le A - c - 1/2$ we have:

$$\phi(y) = \psi(c) + \psi(-c),$$
$$\psi(c) = \left(\frac{y}{2} + \frac{3c}{2} - k\right)\sqrt{A^2 - (y-c)^2} + \left(ck - \frac{A^2}{2} - c^2\right)\sin^{-1}\left(\frac{y-c}{A}\right) \quad (14.a)$$

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.

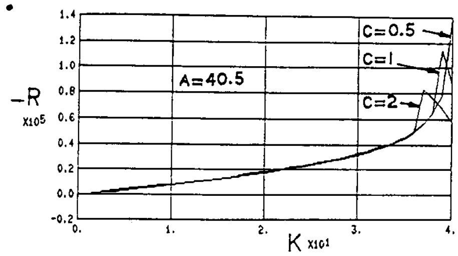


Fig. 1. $R(k)$ versus k for different amplitudes of uniform dither ($c < A$).

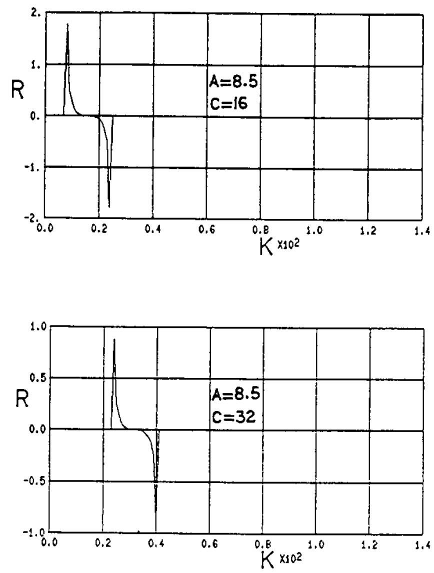


Fig. 2. $R(k)$ versus k for two different amplitudes of uniform dither ($c > A$).

and for $A - c + 1/2 \le |k| \le A + c - 1/2$, we have

$$\phi(y) = \psi(c) \tag{14.b}$$

where $\psi(c)$ is given by (14.a).

The change of $R(k)$ with positive k is shown in Fig. 3 for the cases: 1) $c = 1/2$, $A = 16$; 2) $c = 1$, $A = 15.5$; and 3) $c = 2$, $A = 15.5$.

2) A PDF With Three Impulses: In this case the PDF of dither is given by

$$f_d(d) = \begin{cases} \frac{1}{3}, & d = +c, 0, -c \text{ with } c < A \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

When c is an integer, it can be shown that

$$R(k) = \frac{1}{3\pi}\left[\phi\left(k + \frac{1}{2}\right) - \phi\left(k - \frac{1}{2}\right)\right] \tag{16}$$

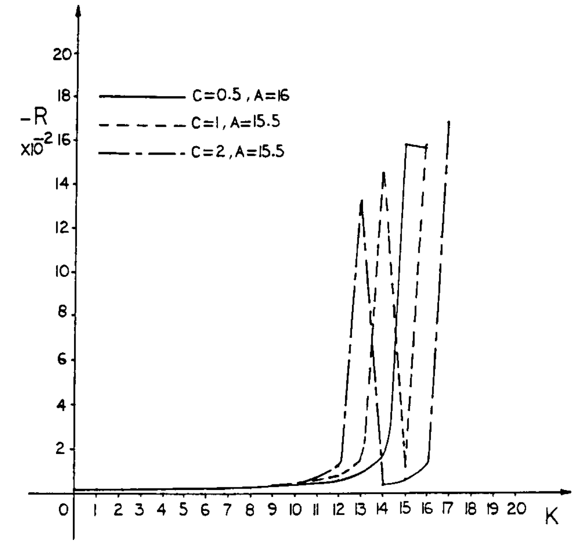


Fig. 3. $R(k)$ versus k for different amplitudes of discrete dither (PDF with 2 impulses).

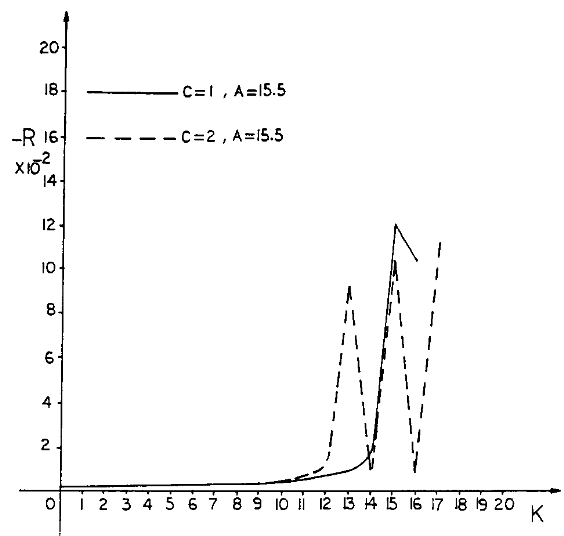


Fig. 4. $R(k)$ versus k for two different amplitudes of discrete dither (PDF with 3 impulses).

where $\phi(y)$ is given by

$$\phi(y) = \begin{cases} \psi(-c) + \psi(0) + \psi(c), \\ \qquad |k| \le A - c - \frac{1}{2} \\ \psi(0) + \psi(c), \\ \qquad A - c + \frac{1}{2} \le |k| \le A - \frac{1}{2} \\ \psi(c), \\ \qquad A + \frac{1}{2} \le |k| \le A + c - \frac{1}{2} \end{cases} \tag{17}$$

and $\psi(c)$ is given by (14.a).

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.

The change of $R(k)$ with positive k is shown in Fig. 4 for the example case: $A = 15.5$ with $c = 1$ and 2. It is interesting to notice that correlation corresponding to three impulses is less than that corresponding to two impulses. This implies that increasing dither impulses, while the same dither amplitude c, results in less correlation. This issue will be discussed more fundamentally in Section III.

III. Relationship Between Uniform Dithered and Discrete Dithered Quantizer Input

Let the PDF of discrete dither be as shown in Fig. 5(a). Let this PDF be convolved with the PDF of a sinusoid, thus

$$f_y(y) = f_x(x) * f_d(d). \tag{18}$$

Let us choose, for example, the convolution range: $y + c > -A$ and $y - c < -A$, i.e., $-(A + c) < y < -(A - c)$. A graphical representation of the convolution process is shown in Fig. 5(b), thus

$$f_y(y) = \int_{-A}^{y+n\Delta c} \frac{1}{2n+1} \left\{ \sum_{i \le n} \delta[x - (y + i\Delta c)] \right\} \cdot \frac{dx}{\pi\sqrt{A^2 - x^2}} \tag{19}$$

where $\delta[\cdot]$ is the Delta–Dirac function. From (19) it follows that:

$$f_y(y) = \frac{1}{(2n+1)\pi} \sum_{y+i\Delta c = -A}^{y+c} \frac{1}{\sqrt{A^2 - [y + i\Delta c]^2}}. \tag{20}$$

Let $z = y + i\Delta c$, then (20) can be rewritten as:

$$f_y(y) = \frac{1}{(2n+1)\pi \cdot \Delta c} \sum_{z=-A}^{y+c} \frac{\Delta c}{\sqrt{A^2 - z^2}}. \tag{21}$$

But $\Delta z = [y + i \cdot \Delta c] - [y + (i-1)\Delta c] = \Delta c$. Also, as $n \to \infty$, $\Delta c \to 0$, consequently $(2n+1)\Delta c \to 2c$. Thus the summation of (21) boils down to an integration as follows:

$$\begin{aligned} f_y(y) &= \frac{1}{2\pi c} \int_{-A}^{y+c} \frac{dz}{\sqrt{A^2 - z^2}} \\ &= \frac{1}{2\pi c}\left[\sin^{-1}\left(\frac{y+c}{A}\right) - \sin^{-1}\left(\frac{-A}{A}\right)\right] \\ &= \frac{1}{2\pi c}\cos^{-1}\left(\frac{-y-c}{A}\right). \end{aligned} \tag{22}$$

This result is exactly the same as that when uniform dither, given by (7), is employed. It indicates that uniform dither is equivalent to discrete dither, when the latter has a PDF with an equi-probable infinite number of impulses. This is in agreement with some of the results of Section II, namely the fact that 3-impulse dithering gives less correlation than 2-impulse dithering. Ultimately, uniform dithering gives less correlation than discrete dithering.

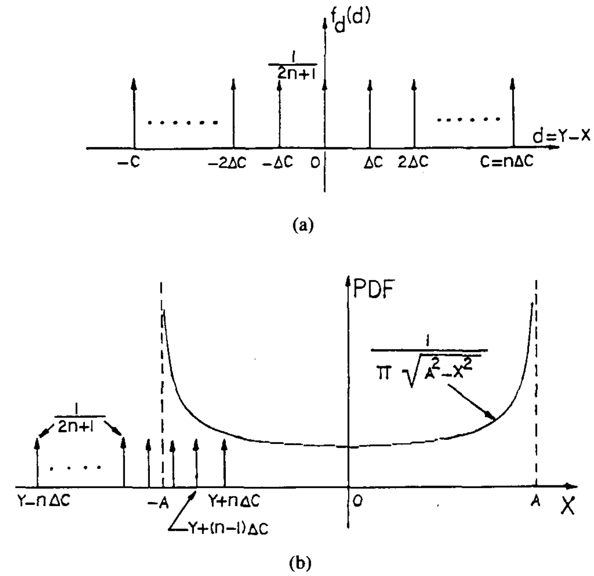


Fig. 5. Discrete dither. (a) PDF with many impulses. (b) Convolving dither with the input sinusoid.

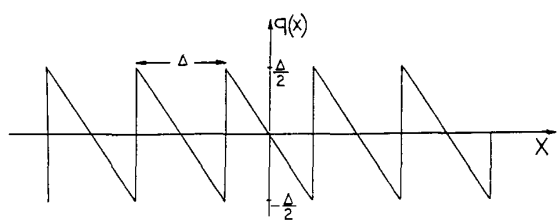


Fig. 6. The quantization error function of an ideal A/D converter.

IV. Amplitude Spectra of Quantization Errors for Different Dither Forms

The quantization error function for an ideal ADC is shown in Fig. 6, with x being the quantizer input voltage. The quantization error function is readily expanded as a Fourier series in x [7]:

$$q(x) = \sum_{n=1}^{\infty} \frac{\Delta}{n\pi} \sin\left(\frac{2\pi n x}{\Delta}\right) \cdot (-1)^n. \tag{23}$$

When a dither voltage d is added to the signal voltage x at the quantizer input, then the average observed error at the ADC output is

$$\bar{q}(x) = \int_{-\infty}^{\infty} q(x + d) \cdot p(d)\, dd \tag{24}$$

where $p(d)$ is the PDF of dither. Equation (24) can be recognized as a convolution integral and transformed to the frequency domain as

$$\overline{Q}(f) = Q(f) \cdot P(f) \tag{25}$$

where $\overline{Q}(f)$, $Q(f)$, and $P(f)$, are the Fourier transforms of $\bar{q}(x)$, $q(x)$, and $p(d)$, respectively.

Although previous simulations [3] showed that dithering reduces harmonic distortion, there was no mathemat-

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.

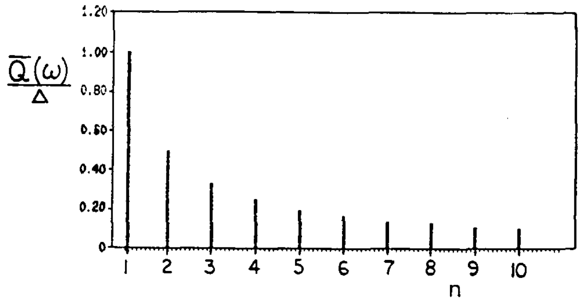


Fig. 7. Spectrum of the average quantization error without dither.

ical basis. To fill this gap, we studied the spectra of the average quantization errors for different dither forms. Let us first consider (23). It can be shown that the positive frequency side of the dither-free amplitude spectrum is given by

$$\left|Q(\omega)\right| = \sum_{n=1}^{\infty} \frac{\Delta}{n} \delta(\omega - n\omega_0) \tag{26}$$

where $\omega_0 = 2\pi/\Delta$. Fig. 7 shows $|Q(\omega)|/\Delta$ versus n, the harmonic index, without dither. Three dither forms are now considered.

A. Gaussian Dither

The PDF of Gaussian dither is given by:

$$p(d) = \frac{1}{\sigma\sqrt{2\pi}} e^{-d^2/2\sigma^2} \tag{27}$$

where σ is the dither rms value. The Fourier transform of (27) is given by

$$P(f) = e^{-2\pi^2 f^2 \sigma^2}. \tag{28}$$

By substituting (26) and (28) in (25), the amplitude spectrum of the average observed quantization error is given by

$$\overline{Q}(\omega) = \sum_{n=1}^{\infty} \frac{\Delta}{n} \exp\left[-\frac{2\pi^2\sigma^2 n^2}{\Delta^2}\right] \cdot \delta(\omega - n\omega_0). \tag{29}$$

Fig. 8 shows $\overline{Q}(\omega)/\Delta$ versus n for the cases: 1) $\sigma = \Delta/2$, and b) $\sigma = \Delta$. It is evident that the first component exists, whereas the contribution from the higher spectral components is negligible.

B. Uniform Dither

The PDF of uniform dither is given by (7), and the Fourier transform is given by

$$P(\omega) = \frac{\sin(\omega c)}{\omega c}. \tag{30}$$

From (25), (26), and (30), it can be shown that

$$\left|\overline{Q}(\omega)\right| = \sum_{n=1}^{\infty} \frac{\Delta}{n} \left|\frac{\sin(2\pi nc/\Delta)}{2\pi nc/\Delta}\right| \cdot \delta(\omega - n\omega_0). \tag{31}$$

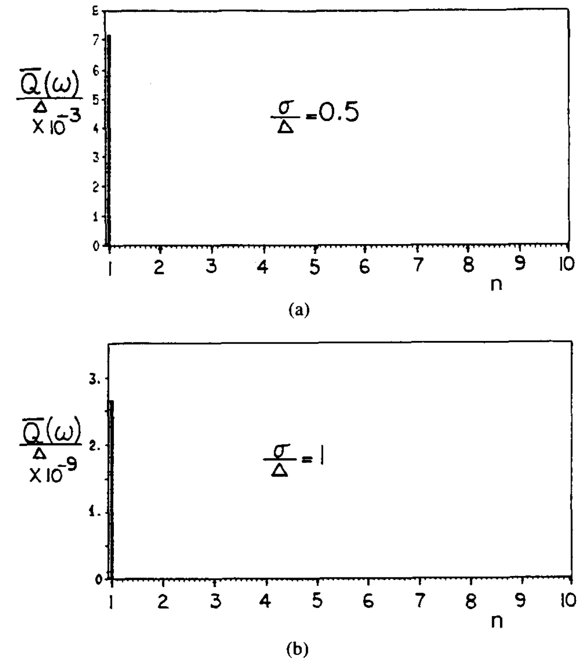


Fig. 8. Spectrum of the average quantization error for Gaussian dither. (a) $\sigma/\Delta = 0.5$. (b) $\sigma/\Delta = 1$.

It is obvious that there are c values that nullify the spectrum of the averaged quantization error, e.g., $c = \Delta/2$, Δ, 2Δ, $\cdots$. These c values correspond to integral values of nc/Δ or $2nc/\Delta$. Other c values do not nullify the spectrum. However, this result is sensitive to the value of c so that it seems dangerous to conclude that uniform dithering has better properties than Gaussian dithering. To compare the two forms, the noise power has to be the same, i.e.,

$$\frac{c^2}{3} = \sigma^2. \tag{32}$$

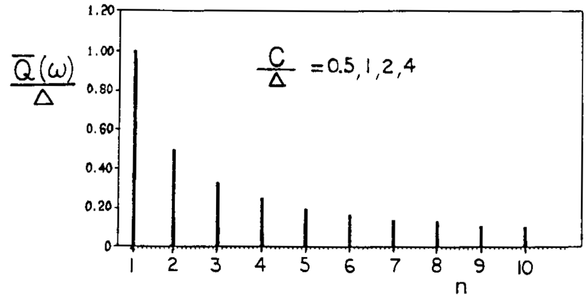
As a numerical example the fundamental components ($n = 1$) for both dither forms are compared. For $\sigma = 0.2887$ (corresponding to $c = 0.5$), (29) gives $\overline{Q}(\omega_0) = 0.193$, which is also obtained from (31) when $c = 0.642$, i.e., a deviation of 0.142Δ from $\Delta/2$. On the other hand for $\sigma = 0.5774$ (corresponding to $c = 1$), (29) gives 0.001 39, which is also obtained from (31) when $c = 1.001$ 39, i.e., a deviation of $0.001\,39\Delta$ from Δ. This means that as c increases, uniform dithering produces a spectrum which is more sensitive to slight uncertainties in c. Therefore, although uniform dithering can theoretically nullify the spectrum, Gaussian dithering is more appropriate for most practical cases due to fewer sensitivity problems and its steep attenuation of higher spectral components.

C. Discrete Dither

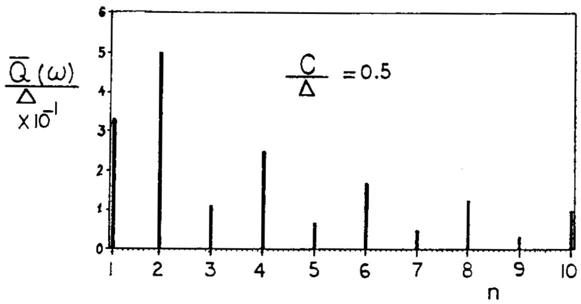
When the PDF of dither is composed of two impulses as given by (12), the Fourier transform is given by

$$P(\omega) = \frac{1}{2}\left[e^{-j\omega c} + e^{-j\omega(-c)}\right] = \cos \omega c. \tag{33}$$

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.



(a)



(b)

Fig. 9. Spectrum of the average quantization error for discrete dither. (a) PDF with two impulses. (b) PDF with three impulses.

From (25), (26), and (33) it follows that:

$$|\overline{Q}(\omega)| = \sum_{n=1}^{\infty} \frac{\Delta}{n} \left| \cos\left(\frac{2\pi nc}{\Delta}\right) \right| \cdot \delta(\omega - n\omega_0). \quad (34)$$

Fig. 9(a) shows $|\overline{Q}(\omega)|/\Delta$ versus n for $c/\Delta = 1/2$, 1, 2, 4. It is obvious that for these c values, $|\overline{Q}(\omega)|/\Delta$ changes as $1/n$. To nullify the amplitude spectrum, we should have

$$\frac{c}{\Delta} = \frac{m}{4n}, m = 1, 3, 5, \cdots \quad (35)$$

which cannot be fulfilled. Thus the spectrum would be null for some values of n, and finite for other values.

When the PDF of dither is composed of three impulses as given by (15), the Fourier transform is given by

$$P(\omega) = \frac{1}{3}\delta(\omega) + \frac{2}{3}\cos \omega c. \quad (36)$$

Again, it can be shown that

$$|\overline{Q}(\omega)| = \sum_{n=1}^{\infty} \frac{\Delta}{n} \left| \frac{1}{3} + \frac{2}{3}\cos\left(2\pi n \frac{c}{\Delta}\right) \right| \cdot \delta(\omega - n\omega_0). \quad (37)$$

When $c/\Delta = 1, 2, 4, \cdots$, the amplitude spectrum is also shown in Fig. 9(a), whereas for $c/\Delta = 1/2$, the spectrum does not monotonically decay which is shown in Fig. 9(b). This last observation was also verified by using a 5-impulse discrete dither. It can be shown that

$$|\overline{Q}(\omega)| = \sum_{n=1}^{\infty} \frac{\Delta}{n} \left| \frac{1}{5} + \frac{2}{5}\cos\left(2\pi n \frac{c}{\Delta}\right) + \frac{2}{5}\cos\left(\pi n \frac{c}{\Delta}\right) \right| \cdot \delta(\omega - n\omega_0). \quad (38)$$

When $c/\Delta = 2, 4, \cdots$, the amplitude spectrum is shown in Fig. 8(a), whereas for $c/\Delta = 1/2$ and 1, it does not decay monotonically.

V. Conclusions

The effect of uniform and discrete dither on the correlation between quantization error and the dithered sinusoidal input signal has been investigated theoretically. Increasing dither amplitude results in less correlation for uniform dither. When the amplitude of discrete dither is kept constant, less correlation is obtained as the number of PDF impulses is increased. Uniform dither is thus better than discrete dither in decorrelating quantization error and input signal, since in the limiting case when the number of impulses is infinite the effect of discrete dither tends toward that of a uniform dither.

The effect of Gaussian, uniform, and discrete dither on the amplitude spectra of the average observed quantization error for an arbitrary quantizer input signal has been investigated theoretically. Gaussian dither results in a very small first harmonic which even decreases with increasing dither variance, whereas higher order harmonics are negligible. The effect of uniform dither can be theoretically made better than that of Gaussian dither by the proper selection of dither amplitude to nullify the noise spectrum. However, it is recommended that Gaussian dither is a better solution for practical cases, corresponding to a uniform dither amplitude $c = M\Delta/2$, where $M \geq 2$. Discrete dither is not as good as uniform dither since it cannot theoretically nullify all harmonic components.

References

[1] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145-154, Feb. 1962.
[2] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," *Bell Syst. Techn. J.*, vol. 51, no. 6, pp. 1293-1304, July-Aug. 1972.
[3] J. Vanderkooy and S. P. Lipshitz, "Resolution below the least significant bit in digital systems with dither," *J. Audio Eng. Soc.*, vol. 32, no. 3, pp. 106-113, Mar. 1984.
[4] ——, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, no. 12, pp. 966-975, Dec. 1987.
[5] A. Holland, "High resolution, high linearity integrating A/D converter," in *Proc. IEEE Int. Test Conf.*, pp. 96-104, 1984.
[6] L. R. Carley, "An oversampling analog-to-digital converter topology for high-resolution signal acquisition systems," *IEEE Trans. Circuits Syst.*, vol. CAS-34, Jan. 1987.
[7] D. R. White and C. P. Pickup, "Systematic errors in digital cross correlators due to quantization and differential nonlinearity," *IEEE Trans. Instrum. Meas.*, vol. IM-36, pp. 47-53, Mar. 1987.
[8] R. S. Burington, *Handbook of Mathematical Tables and Formulas*. 3rd ed. Sandusky, OH: Handbook, 1949.

Authorized licensed use limited to: IEEE Xplore. Downloaded on January 10, 2009 at 22:10 from IEEE Xplore. Restrictions apply.