Dithered Quantizers

Robert M. Gray, *Fellow, IEEE*, and Thomas G. Stockham, Jr., *Fellow, IEEE*

***Abstract*— Dithered quantization is a technique in which a signal called a dither is added to an input signal prior to quantization. This purposeful distortion of an input signal is common, because it can result in a more subjectively pleasing reproduction and because, under certain conditions, it can cause the quantization error to behave in a statistically nice fashion. In particular, suitably chosen random dither signals can cause the quantization error to be signal independent, uniformly distributed white noise. Unfortunately, however, these properties do not in general imply similar properties for the overall quantization noise in many systems, and this has caused some confusion in the understanding, application, and interpretation of the basic results. A theory of overall quantization noise for nonsubtractive dither was originally developed in unpublished work by Wright and by Stockham and subsequently expanded by Brinton, Lipshitz, Vanderkooy, and Wannamaker. These results are not as well known as the original results, however, and misunderstanding persists in the literature. New proofs of the aforementioned properties of quantizer dither, both subtractive and nonsubtractive are provided. The new proofs are based on elementary Fourier series and Rice's characteristic function method and do not require the traditional use of generalized functions (impulse trains of Dirac delta functions) and sampling theorem arguments. The goal is to provide a unified derivation and presentation of the two forms of dithered quantizer noise based on elementary Fourier techniques.**

***Index Terms*—Quantization, dither.**

I. Introduction

A UNIFORM quantizer maps an analog input into one of a collection of equally spaced output levels. The basic quantizer operation is depicted in the block diagram in Fig. 1 and Fig. 2 shows the typical uniform quantizer mapping for an even number of quantizer levels. The input is a discrete time stationary random process X_n; $n \in \mathcal{Z}$, where $\mathcal{Z}$ is the set of all integers. The quantizer output is the reproduction process $\hat{X}_n = q(X_n)$.

Deceptively simple in its description and construction, the uniform quantizer has proved to be surprisingly difficult to analyze, precisely because of its inherent nonlinearity. A common analysis technique is to linearize the quantizer by assuming that the quantizer error $e_n = q(X_n) - X_n$, the difference between the quantizer input and output, consists of a sequence of uniformly distributed random variables that are uncorrelated with each other (white) and with the input signal. In this case the quantizer is replaced by the simple system of Fig. 3, where e_n is an independent and identically distributed (i.i.d.) sequence of uniformly distributed random variables.

Fig. 1. Quantizer.

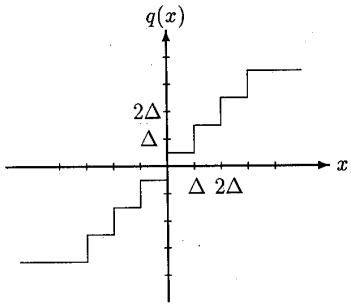

Fig. 2. Uniform quantizer.

This approximation was shown by Bennett [1] to be reasonable when the number of quanatizer levels N is large, the spacing Δ between the levels is small, and the input probability density function f_X is smooth. It has long been known that this approximation can be quite poor in certain cases, e.g., for sinusoidal inputs [2], [3]. Furthermore, many modern oversampled analog-to-digital convertors such as Sigma–Delta modulators clearly violate the Bennett conditions in that they have few levels, relatively large spacing Δ, and nonsmooth quantizer input density functions [4].

One common means of attempting to force the Bennett approximations to hold is to use a dither signal. In his pioneering paper, Roberts [5] argued that a random or pseudo-random noise added to an image before quantization and subtracted before reconstruction could "break up" the pattern of quantization noise and could result in a perceptually more pleasing reconstructed image. Simulations provided strong evidence that this simple randomization could indeed provide significant improvement in quality. In a dithered quantizer, instead of quantizing an input signal X_n directly, one quantizes a signal

$$U_n = X_n + W_n, \tag{1}$$

where W_n is a random process, independent of the signal X_n, called a *dither* process.

A simple dithered quantizer is shown in Fig. 4. The dither process is usually assumed to be i.i.d., and will be so assumed here. Unlike Roberts' system, we begin with a system that

Manuscript received August 30, 1991; revised July 17, 1992. This work was supported in part by the National Science Foundation under Grant MIP-8706539. This work was presented in part at the IEEE International Symposium on Information Theory, Budapest, Hungary, June 24–28, 1991.

R. M. Gray is with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, 133 Durand Building, Stanford, CA 94305-4055.

T. G. Stockham, Jr. is with the Department of Electrical Engineering, University of Utah, 1100 Vista View Drive, Salt Lake City, UT 84111.

IEEE Log Number 9205196.

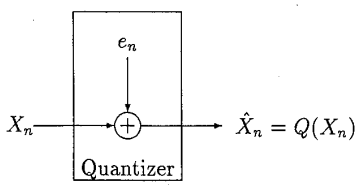

Fig. 3. Additive noise model of a quantizer.

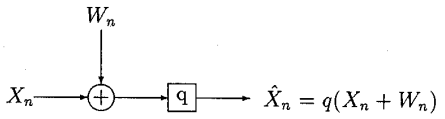

Fig. 4. Dithered quantizer.

does not subtract the dither signal following the quantizer (for reasons to be discussed later). This system is sometimes called a *nonsubtractive dithered quantizer* to distinguish it from the *subtractive dithered quantizer* considered by Roberts.

The idea of dithering is that such randomization can cause the quantization error

$$e_n = q(X_n + W_n) - (X_n + W_n)$$

to have the desired properties assumed in the Bennett approximation. As before, *quantization error* means the error between the input and output of the quantizer. We will refer to the overall difference between original input and final output,

$$\epsilon_n = q(X_n + W_n) - X_n = e_n + W_n .$$

as the *quantization noise.* The behavior of these two errors is the focus of this paper.

The principal theoretical property of dithering was developed by Schuchman [6]. He proved that if the following conditions are met:

1) the quantizer does not overload in the sense that the quantizer error never exceeds $\Delta/2$; i.e., the magnitude of the quantizer input cannot exceed $N\Delta/2$, and
2) the characteristic function of the marginal probability density function of the dither signal defined by

$$M_W(ju) = E\left(e^{juW}\right) \tag{2}$$

has the property that

$$M_W\left(\frac{j2\pi l}{\Delta}\right) = 0, \qquad l \neq 0 . \tag{3}$$

Recall that for $l = 0$, $M_W(0) = 1$, then the quantizer error e_n is uniformly distributed on $(-\Delta/2, \Delta/2]$ and is independent of the original input signal X_n. Schuchman's conditions are satisfied, for example, if the dither signal has a uniform probability density function on $(-\Delta/2, \Delta/2]$, in which case

$$M_W(ju) = \int_{-\Delta/2}^{\Delta/2} e^{juw}\,\frac{dw}{\Delta} = \frac{\sin(u\Delta/2)}{u\Delta/2} . \tag{4}$$

It follows from the work of Jayant and Rabiner [7] and Sripad and Snyder [8] (see also [4]) that this condition implies that the sequence of quantization errors $\{e_n\}$ is independent. The

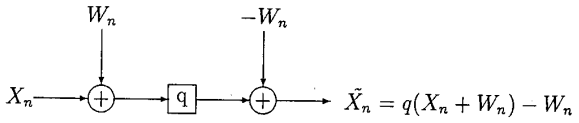

Fig. 5. Subtractive dither.

case of uniform dither remains by far the most widely studied in the literature.

In spite of these apparently nice statistical properties and the resulting accuracy of the simple additive signal-independent white noise model of quantization, dithering is controversial because it involves the addition of a corrupting noise to the input signal and forces a decrease in the dynamic range of the input if quantizer overload is to be avoided, that is, if the maximum quantizer error is to remain less than half the distance between the output levels.

It has been the authors' experience that it is a common misconception that if Schuchman's conditions hold for a dither signal in a simple scalar quantizer or PCM system, the overall quantizer noise ϵ_n will be independent of the input signal, uniform, and white. (See, for example, the statements made regarding nonsubtractive dither in [9, pp. 167, 170] and the discussion of misconceptions regarding dithering in Vanderkooy and Lipshitz [10].)

The difference between quantization error and quantization noise can be highlighted by distinguishing between subtractive dither and nonsubtractive dither. Consider the alternative system of Fig. 5, which is the form of dither introduced by Roberts. In this system, the receiver knows the dither signal and subtracts it from the reconstructed quantizer value to produce a final reproduction $\tilde{X}_n$ of the original input X_n. Although there is no guarantee that a dither signal added prior to the nonlinear quantization can be removed by subtracting it from the resulting digital signal, a startling thing happens if the overall quantizer noise in the subtractive dither system is considered:

$$\epsilon_n = \tilde{X}_n - X_n = q(X_n + W_n) - W_n - X_n = e_n . \tag{5}$$

That is, in this system the overall quantization noise is equal to the quantizer *error* in the original nonsubtractive dithered quantizer of Fig. 4, which is indeed uniform and i.i.d. if Schuchman's conditions are met.

The subtractive dither result is nice mathematically because it promises a well-behaved quantization noise as well as quantization error. It is impractical in many applications, however, for two reasons. First, the receiver will usually not have a perfect analog link to the transmitter (or else the original signal could be sent in analog form), and hence a pseudo-random deterministic sequence must be used at both transmitter and receiver as proposed by Roberts. In this case, however, there will be no mathematical guarantee that the quantization error and noise have the properties that hold for genuinely random i.i.d. dither. Second, subtractive dither of a signal that indeed resembles a sample function of a memoryless random process is complicated to implement, requiring storage of the dither signal, high-precision arithmetic, and perfect synchronization. As a result, it is of interest to study the

behavior of the quantization noise in a simple nonsubtractive dithered quantizer, that is, the simplest dithered quantizer of Fig. 4. One might correctly suspect that nonsubtractive dither does not possess the properties of the quantizer error, but it remains of interest to describe those properties it does possess and to make an intelligent choice of a dithering signal. It is not obvious at the outset that the standard choice of a uniformly distributed dither signal will be the correct choice in a nonsubtractive system.

This problem is of practical as well as theoretical importance. The second author was working in 1980 as a technical consultant on one of the first digital recordings of rock and roll, the album *Tusk* by Fleetwood Mac. At the end of one track, there is a long fadeout of the University of Southern California, Trojan Marching Band. The quantizer was dithered using a uniform dither, which, to the author's mind, should have forced the quantization noise to be independent of the signal. This was clearly not the case, however, as a good ear could hear the quantization noise fading in and out as the signal decrease, instead of remaining constant and small. The energy in the quantization noise was clearly signal-dependent. It was found, however, that by using a triangle density for the dither noise instead of the uniform density (the triangle density corresponds to a random variable formed by summing two independent uniform variates together), the perceived noise level did stay constant, regardless of the signal. The density also satisfies Schuchman's conditions. Similar phenomena were found when dithering quantized images. A simple uniform dither led to quantization noise whose strength was visibly dependent on the input signal magnitude. Triangular dither smoothed out these variations and yielded roughly equal quantization noise energy throughout the image.

These initial observations led to a study of the problem by T. J. Stockham, Jr. and a student, L. K. Brinton, which resulted in a Master's thesis [11]. This work provided a development for the behavior of the conditional moments of the quantizer noise when the dither signal had a uniform or triangular distribution, and inferred the general result for the case of dithering by the sum of a finite number of independent uniformly distributed random variables. The proof was effectively a variation on the traditional engineering-oriented proof of the sampling theorem and made heavy use of generalized functions, impulse trains or Dirac delta sequences in particular. A more detailed development following similar lines was subsequently published by Vanderkooy, Lipshitz, and Wannamaker in a series of papers [12]–[15] culminating in a joint work with Wright [16], [17], who had developed the basic results in unpublished work in 1979.

The origins of this paper lay in the first author's attempts to understand the basic results without recourse to generalized functions. Just as the sampling theorem can be proved by simply expanding the spectrum of a band-limited function in a Fourier series and evaluating the coefficients, rather than by using the traditional engineering arguments of multiplication by an impulse train and low pass filtering, it is reasonable to suspect that the similar looking results for dithering quantizers can be derived from elementary Fourier series arguments as well as from generalized functions and Fourier transforms. The results of this effort are an alternative development of the theory of dithered quantizers that is based on ordinary Fourier series and (like the traditional approach pioneered by Widrow for quantization) Rice's characteristic function method. The quantizer results are here proved via intermediate results describing the characteristic functions of fractional and integer parts of random variables. These results are of some interest in their own right and provide a contrasting insight into the dithering results.

The impulse train/sampling theorem arguments have their intuitive value for those familiar with the techniques. The Fourier series approach, however, is arguably simpler mathematics which provide an alternative intuition for those more familiar with elementary Fourier series than with generalized functions.

We begin in the next section with a description of the principal results and a discussion of some of their implications. The remainder of the paper is devoted to their proof. The following notation will be used throughout. Every real number r can be uniquely written in the form

$$r = \lfloor r \rfloor + \langle r \rangle, \tag{6}$$

where $\lfloor r \rfloor$ denotes the greatest integer less than or equal to r and $0 \le \langle r \rangle < 1$ is the fractional part of r (or r mod 1).

The derivatives of the characteristic function will be denoted

$$M_Z^{(k)}(j\beta) = \frac{d^k}{d\alpha^k} M_Z(j\alpha)|_{\alpha=\beta}.$$

Recall the moment generating property of characteristic functions:

$$E(Z^K) = j^{-k} M_Z^{(k)}(0). \tag{7}$$

Note that if the random variable Z has a probability density function (pdf) f_Z, then the characteristic function is related to the Fourier transform $\hat{f}_Z$ of f_Z by

$$\hat{F}_Z(u) = M_Z(-j2\pi u), \tag{8}$$

where

$$\hat{f}_Z(u) = \int_{-\infty}^{\infty} f_Z(z) e^{-j2\pi zu}\, dz.$$

II. Dithered Quantizers

Consider the dithered quantizer of Fig. 4, and define the quantizer error $e_n = q(X_n + W_n) - (X_n + W_n)$. We require that the input and dither signals are chosen so that the quantizer does not overload. In particular, if there are M quantizer levels spaced Δ apart, we assume that with probability 1

$$|X_n + W_n| \le \frac{M\Delta}{2}. \tag{9}$$

In this case, the normalized quantizer error can be expressed as (see, e.g., [4]):

$$\frac{e_n}{\Delta} = \frac{1}{2} - \left\langle \frac{X_n}{\Delta} + \frac{W_n}{\Delta} \right\rangle, \tag{10}$$

where $\langle\cdot\rangle$ is the fractional part operator defined in (6), and hence, the quantizer noise (or nonsubtractive error) is

$$\frac{\epsilon_n}{\Delta} = \frac{W_n}{\Delta} + \frac{1}{2} - \left\langle \frac{X_n}{\Delta} + \frac{W_n}{\Delta} \right\rangle. \tag{11}$$

Application of Corollary 1 of Section IV with X_n/Δ replacing X_n and $Z_n = W_n/\Delta$ yields the following result.

Theorem 1: Suppose that a dither signal W_n is independent of the input process X_n and is i.i.d., and that the quantizer does not overload. Then Schuchman's condition (3) is necessary and sufficient for the following properties.

- X_k is independent of the quantizer error $e_n = q(X_n + W_n) - (X_n + W_n)$ for all n and k.
- The quantizer error e_n is an i.i.d. sequence of uniform random variables on $(-\Delta/2, \Delta/2]$.

In Theorem 1 and in later results, necessity of the condition means that it is necessary for the properties to hold for *all* stationary distributions on the input process $\{X_n\}$. The properties might, of course, hold for a specific stationary distribution even if the necessary conditions of the theorem are violated.

Similarly, Corollary 2 of Section IV then provides a characterization of the quantizer noise (nonsubtractive error).

Theorem 2: Suppose that a dither signal W_n is independent of the input process X_n and is i.i.d., and that the quantizer does not overload. Then, the condition

$$\frac{d^k}{du^k}\left[M_W(ju)M_V(ju)\right]\Big|_{u=2\pi l/\Delta} = 0, \qquad \text{all } l \neq 0, \tag{12}$$

where V is a random variable uniformly distributed on $(-\Delta/2, \Delta/2]$ and independent of the W_n, is necessary and sufficient for the conditional kth moment of the quantization noise $\epsilon_n = q(X_n + W_n) - X_n$ not to depend on X_n:

$$\begin{aligned} E[\epsilon^k|X] &= E[\epsilon^k] \\ &= \Delta^k E\left[\left(\left\lfloor \frac{W_n}{\Delta} \right\rfloor + \frac{1}{2}\right)^k\right] \\ &= \frac{1}{j^k}\frac{d^k}{du^k}[M_W(ju)M_V(ju)]|_{u=0} \\ &= E\left[(W+V)^k\right]. \end{aligned} \tag{13}$$

In other words, the kth conditional moment of ϵ is the kth unconditional moment of the dither signal plus an independent random variable uniformly distributed on one quantizer bin width.

Equation (12) holding with $k = 1$ is necessary and sufficient for the following properties.

- ϵ_n and X_n are uncorrelated processes; that is, for all integers n and m

$$E(\epsilon_n X_m) = E(\epsilon_n)E(X_m).$$

- The sequence ϵ_n is uncorrelated (white); that is, the autocorrelation has the form

$$R_\epsilon(n,m) = E(\epsilon_n\epsilon_m) = \begin{cases} E(\epsilon^2), & \text{if } n = m, \\ E(\epsilon)^2, & \text{if } n \neq m. \end{cases}$$

Lastly, a sufficient condition for (12) to hold for a given $k \geq 1$ is that

$$M_W^{(m)}\left(j\frac{2\pi l}{\Delta}\right) = 0, \qquad \text{all } l \neq 0, m = 0, 1, \cdots, k-1. \tag{14}$$

Observe in particular that if (12) holds for $k = 2$, then the quantizer noise power (the mean-squared error) will be

$$\begin{aligned} E[\epsilon^2|X] = E[\epsilon^2] &= E[W^2] \\ &+ 2E[W]E[V] + E[V^2] = E[W^2] + \frac{\Delta^2}{12}, \end{aligned} \tag{15}$$

since V is uniform on $(-\Delta/2, \Delta/2]$. This means that the power in the dither signal is directly added to that of the quantizer error in order to form the overall mean-squared error. Thus, one can have the conditional second order moment of the noise be independent of the input only at the cost of increase quantizer noise power.

As an example, the conditions of the theorem are satisfied if W is the sum of k independent random variables identically distributed as V. If $k = 2$, this says that the variance of the quantizer noise does not depend on the input if the dither signal is chosen as the sum of two uniform random variables (the triangle distribution), but its value is

$$E[\epsilon^2] = 3E[V^2] = \frac{\Delta^2}{4}.$$

If a uniform dither is used, however, the resulting noise power is only $2E[V^2] = \Delta^2/6$, but the noise energy is no longer independent of the signal. If k uniform variates are added to form the dither, then, as stated by Brinton [11], the noise power is

$$E[\epsilon^2] = (k+1)\frac{\Delta^2}{12}, \qquad \text{for } k \geq 2.$$

III. Characteristic Functions and Fractional Parts

We begin the proofs of the basic results with lemmas describing the behavior of fractional parts of random variables in terms of their characteristic functions.

Lemma 1: Let Z be a random variable with characteristic function $M_z(ju)$. Then, $1/2 - \langle a + Z\rangle$ is uniformly distributed on $(-\Delta/2, \Delta/2]$ for all real a, if and only if

$$M_Z(j2\pi l) = 0, \qquad \text{all } l \neq 0. \tag{16}$$

Comments:

- We focus on $1/2 - \langle\cdot\rangle$ instead of the fractional part $\langle\cdot\rangle$ alone, in order to have a variable in $(-\Delta/2, \Delta/2]$ instead of in $[0, 1)$.
- If U is a uniformly distributed random variable on $(-\Delta/2, \Delta/2]$, then from (4)

$$M_U(ju) = \frac{\sin u/2}{u/2}. \tag{17}$$

Thus, for example, if $Z = U$ for U uniformly distributed, the condition of (16) is met.

- If U_i, $i = 1, 2, \cdots, N$ are independent identically distributed uniform random variables with the same distribution as U, then,

$$S_N = \sum_{i=1}^{N} U_i \tag{18}$$

has a characteristic function $M_U(ju)^N$. The random variable S_N thus clearly has a characteristic function satisfying (16).

Proof of Lemma 1: First consider the "only if" or necessity part by assuming that for any fixed a, $1/2 - \langle a + Z \rangle$ has a uniform probability density function on $(-1/2, 1/2]$ and hence

$$M_{1/2-\langle a+Z\rangle}(j2\pi l) = 0, \qquad \text{all } l \neq 0\,.$$

Then,

$$\begin{aligned} M_Z(j2\pi l) &= E\left(e^{j2\pi lZ}\right) = e^{-j2\pi la}E\left(e^{j2\pi l(a+Z)}\right) \\ &= e^{-j2\pi la}E\left(e^{j2\pi l(\lfloor a+Z\rfloor+\langle a+Z\rangle)}\right) \\ &= e^{-j2\pi la}E\left(e^{j2\pi l\langle a+Z\rangle}\right) \\ &= e^{-j2\pi l(a-1/2)}E\left(e^{j2\pi(-l)(1/2-\langle a+Z\rangle)}\right) \\ &= e^{-j2\pi(a-1/2)}M_{1/2-\langle a+Z\rangle}(-j2\pi l) = 0, \\ &\quad \text{all } l \neq 0\,, \end{aligned}$$

proving necessity of (16).

To consider the "if" or sufficiency part, assume that (16) holds. Since

$$1/2 - \langle a + Z\rangle \in (-1/2, 1/2]\,,$$

its probability density function $f_{1/2-\langle a+Z\rangle}$ can be taken to be nonzero only in that region, and hence, we can expand it in a Fourier series as

$$f_{1/2-\langle a+Z\rangle}(\alpha) = \sum_{l=-\infty}^{\infty} c_l e^{-j2\pi l\alpha}, \qquad \alpha \in (-1/2, 1/2]\,, \tag{19}$$

where the Fourier coefficients are given by

$$\begin{aligned} c_l &= \int_{-1/2}^{1/2} f_{1/2-\langle a+Z\rangle}(\alpha)e^{j2\pi l\alpha}\,d\alpha \\ &= E\left(e^{j2\pi l(1/2-\langle a+Z\rangle)}\right) \\ &= E\left(e^{j2\pi l(1/2-(a+Z)+\lfloor a+Z\rfloor)}\right) \\ &= e^{j2\pi l(1/2-a)}E\left(e^{-j2\pi lZ}\right) \\ &= e^{j2\pi l(1/2-a)}M_Z(-j2\pi l)\,. \end{aligned}$$

Hence,

$$f_{1/2} - \langle a + Z\rangle(\alpha) = \sum_{l=-\infty}^{\infty} e^{j2\pi l(1/2-a-\alpha)}M_Z(-j2\pi l)\,, \quad \alpha \in (-1/2, 1/2]\,. \tag{20}$$

This Fourier series points out the connection with the sampling theorem: probability density function with domain $(-1/2, 1/2]$ can be expressed in terms of the samples of its characteristic function, just as a band-limited spectrum can be expressed in terms of the samples of the corresponding time signal. The point is that one need not introduce Dirac deltas and impulse trains to write this; it is just a simple Fourier series.

Since

$$c_0 = M_Z(0) = \int_{-1/2}^{1/2} f_{1/2-\langle a+Z\rangle}(\alpha)\,d\alpha = 1\,,$$

(16) implies that

$$f_{1/2-\langle a+Z\rangle}(\alpha) = 1, \qquad \alpha \in (-1/2, 1/2]\,. \qquad \square$$

The first lemma relates the zeros of the characteristic function of a random variable to the uniformity of the distribution of the random variable mod 1. It also implies that the *only* pdf on $(-1/2, 1/2]$ satisfying (16) is the uniform pdf. (Nonuniform pdfs with larger support can, of course, satisfy (16).)

Lemma 2: Let Z be a random variable with characteristic function $M_Z(ju)$. Then for a positive integer k,

$$E\left[\left(Z + \frac{1}{2} - \langle a + Z\rangle\right)^k\right] = E\left[\left(\lfloor Z\rfloor + \frac{1}{2}\right)^k\right] \tag{21}$$

for all real a, if and only if

$$\frac{1}{j^k}\frac{d^k}{du^k}[M_Z(ju)M_U(ju)]|_{u=2\pi l} = C\delta_l, \qquad \text{all integer } l\,, \tag{22}$$

where δ_l is the Kronecker delta and

$$\begin{aligned} C &= E\left[\left(\lfloor Z\rfloor + \frac{1}{2}\right)^k\right] \\ &= \frac{1}{j^k}\frac{d^k}{du^k}[M_Z(ju)M_U(ju)]|_{u=0} = E\left[(Z+U)^k\right], \end{aligned} \tag{23}$$

where U is a uniform $(-1/2, 1/2]$ random variable independent of Z.

Comments:

- The lemma gives necessary and sufficient conditions under which the kth moment of $Z + 1/2 - \langle a + Z\rangle$ does not depend on a. When the conditions are met, the kth moment of $Z + 1/2 - \langle a + Z\rangle$ equals the kth moment of $Z + U$ with U uniform and independent of Z. Although the previous lemma promises that $1/2 - \langle a + Z\rangle$ will indeed be uniform if the condition holds for $k = 0$, it is clearly not independent of Z since it is a deterministic function of Z. The lemma shows that a weaker form of independence does, however, hold.
- The moment generating property of (7) and the fact that $M_Z(ju)M_U(ju)$ is the characteristic function of $Z + U$ yield the final equality of the lemma.
- Since

$$\frac{d^k}{du^k}[M_Z(ju)M_U(ju)]|_{u=2\pi l} = \sum_{m=0}^{k-1}\binom{k}{m}M_Z^{(m)}(j2\pi l)M_U^{(k-m)}(j2\pi l)\,.$$

a sufficient condition for (12) to hold for a given $k \geq 1$ is that

$$M_Z^{(m)}(j2\pi l) = 0, \qquad \text{all } l \neq 0, m = 0, 1, \cdots, k-1. \tag{24}$$

- A random variable Z meeting the conditions of (24) is the sum of k uniform random variables S_k defined in (18).

Proof: Define

$$e = \frac{1}{2} - \langle a + Z \rangle$$

and

$$\epsilon = Z + e$$

so that the problem is to determine the kth moment of ϵ. We have that

$$\begin{aligned} \epsilon &= Z + \frac{1}{2} - (a + Z) + \lfloor a + Z \rfloor \\ &= \frac{1}{2} - a + \lfloor a + Z \rfloor, \end{aligned}$$

so that ϵ is discrete and described by a probability mass function

$$p_\epsilon(\zeta) = \Pr(\epsilon = \zeta).$$

This probability will be nonzero only for $\zeta \in A_a$ where

$$A_a = \left\{ \frac{1}{2} - a + k; k \pm \mathcal{Z} \right\}.$$

If $\zeta = 1/2 - a + k$, then

$$\begin{aligned} p_\epsilon(\zeta) &= \Pr(\lfloor Z + a \rfloor = k) \\ &= \Pr(k \leq Z + a < k + 1) \\ &= \int_{k-a}^{k+1-a} f_Z(z)\, dz \\ &= \int_{\zeta-1/2}^{\zeta+1/2} f_Z(z)\, dz \\ &= \int_{-1/2}^{1/2} f_Z(\zeta - z)\, dz, \qquad \zeta \in A_a, \end{aligned} \tag{25}$$

which has the form of a convolution of a probability density function f_Z with a uniform probability density function over $(-1/2, 1/2]$. This form is at the heart of Wright's and Stockham's original arguments, although their derivations differ. Although the pmf is defined only for $\zeta \in A_a$, the right-hand side of (25) is clearly well defined for all read ζ. For later use, we define

$$h(\zeta) = \int_{-1/2}^{1/2} f_Z(\zeta - z)\, dz, \tag{26}$$

for all real ζ and observe that $p_\epsilon(\zeta) = h(\zeta)$ for $\zeta \in A_a$.

The task now is to evaluate the moments. To accomplish this, consider the characteristic function

$$M_\epsilon(ju) = \sum_{k=-\infty}^{\infty} e^{ju\left(\frac{1}{2}-a+k\right)} p_\epsilon\left(\frac{1}{2} - a + k\right).$$

Sums of this form can be evaluated using the Poisson summation formula (cf. [18, p. 316]) that states if $\hat{g}$ is the (continuous parameter) Fourier transform of g, then

$$\sum_{l=-\infty}^{\infty} g(\alpha + l) = \sum_{l=-\infty}^{\infty} \hat{g}(l) e^{i2\pi l\alpha}. \tag{27}$$

Setting $g(y) = e^{juy} h(y)$, then $\hat{g}(f) = \hat{h}(f - u/2\pi)$ and the Poisson summation formula yields

$$M_\epsilon(ju) = \sum_{l=-\infty}^{\infty} \hat{h}\left(l - \frac{u}{2\pi}\right) e^{i2\pi\left(\frac{1}{2}-a\right)l}.$$

Thus the kth-order moment is given by

$$\begin{aligned} E(\epsilon^k) &= \frac{1}{j^k} \frac{d^k}{du^k} M_\epsilon(ju)|_{u=0} \\ &= \sum_{l=-\infty}^{\infty} \hat{h}^{(k)}(l) e^{i2\pi\left(\frac{1}{2}-a\right)l}, \end{aligned}$$

where $\hat{h}^{(k)}(y)$ denotes the kth derivative of $\hat{h}$ evaluated at y. This is simply the discrete parameter Fourier transform of the sequence $\hat{h}^{(k)}(l)$ evaluated at frequency $1/2 - a$ (mod 1). This will be constant (not a function of frequency), if and only if

$$\hat{h}^{(k)}(l) = C\delta_l, \qquad l \in \mathcal{Z}.$$

Since $h(\zeta)$ is defined as the convolution of a uniform pdf and f_Z, its Fourier transform is just the product of the Fourier transforms of the uniform pdf and f_Z. Applying (8) results in condition (22). The remainder of the lemma follows. □

IV. Proofs of Theorems

The lemmas of the previous section are now used to derive the results that imply the theorems of Section II. The corollaries just replace the constant a by a random variable and use conditional expectation to apply the lemmas. As in the previous section, the dither random variable is implicitly scaled to $(-1/2, 1/2]$, that is, $\Delta = 1$.

Corollary 1: Suppose $\{X_n\}$ and $\{Z_n\}$ are two random processes such that

a) X_n and Z_k are independent for all n and k,
b) $\{Z_n\}$ is an i.i.d. sequence.

Then the condition that the characteristic function $M_Z(ju)$ satisfy (16) is necessary and sufficient for the following to hold.

a) $1/2 - \langle X_n + Z_n \rangle$ is uniformly distributed on $(-1/2, 1/2]$ for all n.
b) $1/2 - \langle X_n + Z_n \rangle$ and X_k are independent for all n and k.
c) the sequence $1/2 - \langle X_n + Z_n \rangle$ is an independent and identically distributed sequence.

Proof: We begin with the second property. Define

$$e_n = \frac{1}{2} - \langle X_n + Z_n \rangle .$$

The second property will be proved if we can show that the joint characteristic function factors, that is,

$$\begin{aligned} M_{e_n,X_k}(ju,jv) &= E\left(e^{jue_n+juX_k}\right) \\ &= M_{e_n}(ju)M_{X_k}(jv) . \end{aligned}$$

To do this, consider the nested expectation

$$M_{e_n,X_k}(ju,jv) = E\left[e^{juX_k}E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)}|X_n,X_k\right]\right] . \tag{28}$$

Since Z_n is independent of X_n and X_k, the conditional expectation evaluated at $X_n = x_n$ and $X_k = x_k$ is given by

$$E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)}|X_n = x_n, X_k = x_k\right] = E\left[e^{ju(1/2-\langle x_n+Z_n\rangle)}\right] .$$

From Lemma 1, the given conditions imply that for a fixed x_n, $1/2 - \langle x_n + Z_n\rangle$ is uniformly distributed on $(-1/2, 1/2]$ regardless of the value of x_n. Hence, from (17),

$$E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)}|X_n,X_k\right] = M_U(ju) , \tag{29}$$

and using (28)

$$\begin{aligned} M_{e_n,X_k}(ju,jv) &= E\left[e^{jvX_k}\right]M_U(ju) \\ &= M_X(jv)M_U(ju) . \end{aligned} \tag{30}$$

Equation (29) also implies that

$$M_{e_n}(ju) = E\left[E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)}|X_n,X_k\right]\right] = M_U(ju) . \tag{31}$$

Equations (30) and (31) together imply the first two properties.

The third property follows by a similar argument. We have that

$$\begin{aligned} M_{e_n,e_l}(ju,jv) &= E\left[e^{jue_n+jve_l}\right] \\ &= E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)+jv(1/2-\langle X_l+Z_l\rangle)}\right] \\ &= E\left[E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)+jv(1/2-\langle X_l+Z_l\rangle)} \mid X_n, X_l\right]\right] . \end{aligned}$$

But given $X_n = x_n$ and $X_l = x_l$, the random variables $(1/2 - \langle x_n + Z_n\rangle)$ and $(1/2 - \langle x_l + Z_l\rangle)$ are (conditionally) independent, since Z_n and Z_l are independent and are both uniformly distributed from Lemma 1. Thus the conditional expectation is

$$E\left[e^{ju(1/2-\langle X_n+Z_n\rangle)+jv(1/2-\langle X_l+Z_l\rangle)}|X_n,X_l\right] = M_U(ju)M_U(jv) = M_{e_n}(ju)M_{e_l}(jv) ,$$

completing the proof of sufficiency. Necessity follows from Lemma 1 since a stationary process can assign $\Pr(X_n = a) = 1$ for any real a and, hence, if the condition does not hold, a stationary process can be constructed which violates the first property. □

Corollary 2: Suppose that X_n and Z_n are independent random processes, that Z_n is i.i.d., and that

$$\epsilon_n = Z_n + 1/2 - \langle X_n + Z_n \rangle .$$

Then, the condition

$$\frac{d^k}{du^k}[M_U(ju)M_Z(ju)]|_{u=2\pi l} = 0, \qquad \text{all } l \neq 0 , \tag{32}$$

where U is a uniformly distributed random variable on $(-1/2, 1/2]$ that is independent of Z, is necessary and sufficient for the following property:

$$E\left[\epsilon_n^k|X_n\right] = E\left[(Z_n + U)^k\right] . \tag{33}$$

If (32) holds for $k = 1$, then also

- $$E(\epsilon_n X_l) = E(\epsilon_n)E(X_l) ; \tag{34}$$

that is, the input and the nonsubtractive error are uncorrelated, and

- $$E(\epsilon_n\epsilon_l) = E(\epsilon_n)E(\epsilon_l); \qquad n \neq l . \tag{35}$$

Proof: The first part follows directly from Lemma 2, since Z_n and X_n are independent. Hence,

$$p_{Z_n+1/2-\langle X_n+Z_n\rangle|X_n}(\zeta|x) = p_{z_n+1/2-\langle x+Z_n\rangle}(\zeta) ,$$

and the unconditional expectations for fixed a in the lemma become conditional expectations given X_n. The second part follows by manipulating the conditional expectations in a manner similar to that used in the proof of Corollary 1 to manipulate the joint and conditional characteristic functions. We have using the first part of the corollary for $k = 1$ that

$$\begin{aligned} E(\epsilon_n X_l) &= E[E(\epsilon_n X_l|X_n X_l)] \\ &= E[X_l E(\epsilon_n|X_n)] = E[X_l E(\epsilon_n)] \\ &= E(X_l)E(\epsilon_n) . \end{aligned}$$

Similarly, since as in the proof of Corollary 1 ϵ_n and ϵ_l are conditionally independent given X_n and X_l if $n \neq l$:

$$\begin{aligned} E(\epsilon_n\epsilon_l) &= E[E(\epsilon_n\epsilon_l|X_n X_l)] \\ &= E[E(\epsilon_l|X_n X_l)E(\epsilon_n|X_n X_l)] \\ &= E[E(\epsilon_l|X_l)E(\epsilon_n|X_n)] \\ &= E(\epsilon_n)E(\epsilon_l) . \end{aligned}$$

□

The first corollary implies that, under the assumed conditions, $e_n = 1/2 - \langle X_n + Z_n\rangle$ is a uniformly distributed random variable. The second corollary shows that when the quantizer error e_n is added to Z_n and the kth moment is formed, the result is the same as if a uniform random variable independent of Z_n has been added and the kth moment formed. If the condition holds for $k = 1$, then the nonsubtractive quantizer error is uncorrelated with the input signal and is white. This is in contrast to the subtractive quantizer error which is actually independent of the input signal and i.i.d.

ACKNOWLEDGMENT

R. M. Gray acknowledges helpful discussions with Prof. S. Lipshitz which led to several simplifications and corrections in statements and proofs.

REFERENCES

[1] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.
[2] A. G. Clavier, P. F. Panter, and D. D. Grieg, "Distortion in a pulse count modulation system," *AIEE Trans.*, vol. 66, pp. 989–1005, 1947.
[3] ——, "PCM distortion analysis," *Elec. Eng.*, pp. 1110–1122, Nov. 1947.
[4] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.
[5] L. G. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145–154, Feb. 1962.
[6] L. Schuchman, "Dither signals and their effects on quantization noise," *IEEE Trans. Commun. Technol.*, vol. COM-12, pp. 162–165, Dec. 1964.
[7] N. S. Jayant and L. R. Rabiner, "The application of dither to the quantization of speech signals," *Bell Syst. Tech. J.*, vol. 51, pp. 1293–1304, July–Aug. 1972.
[8] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust. Speech Signal Processing*, vol. ASSP-25, pp. 442–448, Oct. 1977.
[9] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
[10] J. Vanderkooy and S. P. Lipshitz, "Resolution below the least significant bit in digital systems with dither," *J. Audio Eng. Soc.*, vol. 32, pp. 106–113, Nov. 1984. (Correction, p. 889)
[11] L. K. Brinton, "Nonsubtractive dither," Master's thesis, Univ. of Utah, Salt Lake City, UT, Aug. 1984.
[12] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975, Dec. 1987.
[13] R. A. Wannamaker, S. P. Lipshitz, and J. Vanderkooy, "Dithering to eliminate quantization distortion," in *Proc. Ann. Meeting Canadian Acoust. Assoc.*, Halifax, Canada, Oct. 1989, pp. 78–86.
[14] R. A. Wannamaker, "Dither and noise shaping in audio applications," Master's thesis, Univ. of Waterloo, Waterloo, Canada, 1991.
[15] S. P. Lipshitz, R. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical study," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, May 1992.
[16] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, and J. N. Wright, "Nonsubtractive dither," submitted to the *IEEE Trans. Signal Processing*, 1991.
[17] ——, "Nonsubtractive dither," presented at *1991 Workshop Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 1991.
[18] J. S. Walker, *Fourier Analysis.* New York: Oxford Univ. Press, 1988.