A Theory of Nonsubtractive Dither

Robert A. Wannamaker, Stanley P. Lipshitz, *Member, IEEE*, John Vanderkooy, and J. Nelson Wright, *Senior Member, IEEE*

***Abstract*—A detailed mathematical investigation of multibit quantizing systems using nonsubtractive dither is presented. It is shown that by the use of dither having a suitably chosen probability density function, moments of the total error can be made independent of the system input signal but that statistical independence of the error and the input signals is not achievable. Similarly, it is demonstrated that values of the total error signal cannot generally be rendered statistically independent of one another but that their joint moments can be controlled and that, in particular, the error sequence can be rendered spectrally white. The properties of some practical dither signals are explored, and recommendations are made for dithering in audio, video, and measurement applications. The paper collects all of the important results on the subject of nonsubtractive dithering and introduces important new ones with the goal of alleviating persistent and widespread misunderstandings regarding the technique.**

***Index Terms*—Dither, quantization.**

I. Introduction

ANALOG-TO-DIGITAL conversion is customarily decomposed into two separate processes: *sampling* of the input analog waveform and *quantization* of the sample values in order to represent them with binary words of a prescribed length. The sampling operation incurs no loss of information as long as the input is appropriately bandlimited, but the approximating nature of the quantization operation *always* results in signal degradation. Another common operation with a similar problem is *requantization*, in which the wordlength of digital data is reduced after arithmetical processing in order to meet specifications for its storage or transmission.

Dither, straightforwardly put, is a random "noise" process added to a signal prior to its (re)quantization in order to control the statistical properties of the quantization error. This is not a new idea. Subtractively dithered (SD) quantizing systems, in which the dither is subsequently subtracted from the output signal after quantization, have been discussed and used for over 30 years in speech and video processing applications [1], [2], and a satisfactory theory of their operation exists in print [3], [4]. Nonsubtractively dithered (NSD) systems, in which the dither signal is not subtracted from the output, are a subject of more recent interest. The following provides a brief history of the theory of NSD quantization.

It must be acknowledged that all theoretical treatments of dithered quantization owe a substantial debt to the work of Widrow [5]–[8], who developed many of the essential mathematical tools while studying undithered quantization. Among Widrow's contributions was the "quantizing theorem," which is a counterpart in discrete-valued systems to the better-known "sampling theorem" in discrete-time systems. Important extensions to Widrow's treatment were made by Sripad and Snyder [9], Sherwood [4], and Gray [10].

Early investigations into nonsubtractive dither *per se* were conducted by Wright [11] in 1979, resulting in discovery of many of the important results that follow. This work remained unpublished until it was brought to the attention of the other authors of this article [12]. The results concerning moments of the error signal were rediscovered independently by Stockham [13] in 1980 and documented in an unpublished Master's thesis by Brinton [14], who was a student of Stockham's, in 1984. Stockham otherwise remained silent on the matter until the 1990's [15], making commercial use of triangular-pdf dither in digital recording/editing systems since the early 1980's.

The properties of nonsubtractive dither (and, in particular, triangular-pdf dither) were again discovered independently by Lipshitz and Vanderkooy in the mid-1980's. Vanderkooy and Lipshitz were the first researchers to publish their findings on nonsubtractive dither [16]–[20], and this prompted collation and extension of the theoretical aspects by Wannamaker [21]–[26].[1] Lipshitz, Wannamaker, and Vanderkooy have published a broad theoretical survey of multibit quantization treating both undithered and dithered systems [27]. They have also extended the treatment to include an analysis of dithered quantizing systems using noise-shaping error feedback [25], [28], [29] and multichannel quantizing systems [30]. The theory has been elegantly extended to cover dithered lattice quantization by Kirac and Vaidyanathan [31].

The results concerning error moments have been independently discovered by Gray [32], using arguments employing Fourier series along with characteristic functions. Gray and Stockham have published an important paper on the subject [15] that establishes the fundamental results regarding nonsubtractive dither using this approach, thus providing an alternative treatment of which readers interested in the topic should be aware.

Although a handful of individuals in the engineering community are aware of the correct results regarding nonsubtractive dither, a number of misconceptions concerning the technique are widespread. A persistent confusion regarding the quite

Manuscript received July 7, 1999; revised August 10, 1999. This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada. The associate editor coordinating the review of this paper and approving it for publication was Editor-in-Chief José M. F Moura. An earlier version of this manuscript was submitted on October 18, 1991 and formally accepted for publication by then associate editor Robert A. Gabel on March 1, 1993, subject to minor revisions. It was considered withdrawn, however, because the authors did not submit the revised manuscript until July 1999.

R. A. Wannamaker, S. P. Lipshitz, and J. Vanderkooy are with the Audio Research Group, University of Waterloo, Waterloo, Ont., N2L 3G1 Canada.

J. N. Wright is with the Parallax Group, Mountain View, CA 94043 USA.

Publisher Item Identifier S 1053-587X(00)01007-2.

[1]Reference [26] available online at http://audiolab.uwaterloo.ca/~rob/.

different properties of subtractive and nonsubtractive dithering (see, for instance, [33, p. 170]) is particularly serious. The aim of this paper is to provide a consistent and rigorous account of the theory of nonsubtractively dithered systems in order to promote a more universal understanding of this dithering technique. As such, it greatly extends and elaborates the treatment provided by our earlier presentation [23].

A. The Classical Model of Quantization

Quantization and requantization processes possess similar transfer characteristics, which are generally of either the *mid-tread* or *mid-riser* variety illustrated in Fig. 1. We will assume that the quantizers involved are *infinite*, which, for practical purposes, means that the system input signal is never clipped by saturation of the quantizer. (Some comments regarding the application of dither to 1-bit and sigma-delta converters will be reserved for the Conclusions.) In this case, the corresponding transfer functions relating the quantizer output to its input w can be expressed analytically in terms of the quantizer step size Δ

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} + \frac{1}{2} \right\rfloor$$

for a mid-tread quantizer, or

$$Q(w) = \Delta \left\lfloor \frac{w}{\Delta} \right\rfloor + \frac{\Delta}{2}$$

for a mid-riser quantizer, where the "floor" operator $\lfloor\ \rfloor$ returns the greatest integer less than or equal to its argument. The step size Δ is commonly referred to as a least significant bit (LSB), since a change in input signal level of one step width corresponds to a change in the LSB of binary coded output. Throughout the sequel, quantizers of the mid-tread variety will be assumed, but all derived results have obvious analogs for mid-riser quantizers, and all stated theorems are valid for both types.

Quantization or requantization introduces an error signal q into the digital data stream, which is simply the difference between the output of the quantizer and its input

$$q(w) \triangleq Q(w) - w$$

where we henceforth use $\triangleq$ to indicate equality by definition. This *quantization error* is shown as a function of w for a mid-tread quantizer in Fig. 2. It has a maximum magnitude of 0.5 LSB and is periodic in w with a period of 1 LSB.

Although q is clearly a deterministic function of the input, the classical model of quantization (CMQ) [34], [35] holds that the quantization error can be modeled as an additive random process that is independent of the system input and iid (i.e., that distinct samples of the error are statistically independent of one another and identically distributed). The CMQ further postulates that the error is *uniformly distributed*, meaning that its values exhibit a probability density function (pdf) of the form

$$p_q(q) = \Pi_\Delta(q)$$

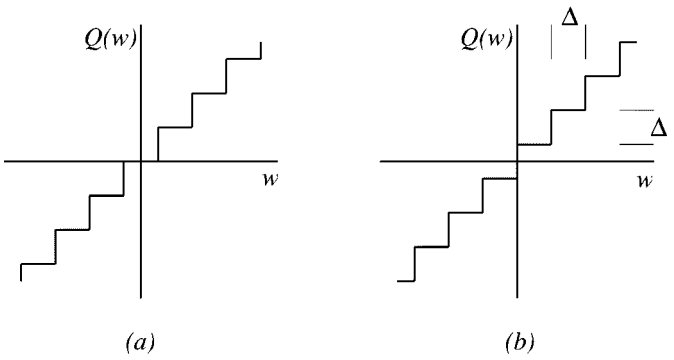

Fig. 1. Quantizer transfer characteristics. (a) Mid-tread and (b) mid-riser with Δ denoting the size of one LSB.

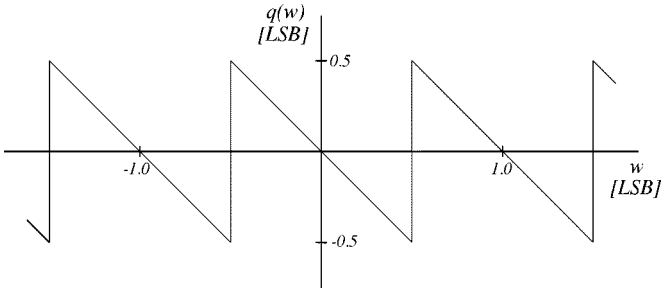

Fig. 2. Quantization error $q(w)$, as a function of quantizer input, w, for a mid-tread quantizer.

where the *rectangular window function* of width Γ, Π_Γ, is defined as

$$\Pi_\Gamma(q) \triangleq \begin{cases} \frac{1}{\Gamma}, & \frac{-\Gamma}{2} < q \leq \frac{\Gamma}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Such a pdf is referred to as a *uniform pdf* or *RPDF (rectangular pdf)*. If a quantization error signal is uniformly distributed, its moments are

$$E[q] = 0 \tag{1}$$

$$E[q^2] = \frac{\Delta^2}{12} \tag{2}$$

and

$$E[q^m] = \begin{cases} \frac{1}{m+1} \left(\frac{\Delta}{2}\right)^m, & \text{for } m \text{ even} \\ 0, & \text{for } m \text{ odd} \end{cases} \tag{3}$$

where (2) is the familiar expression for the variance of the quantization error in the classical model.

The CMQ is valid for input signals that exhibit smooth pdf's and are large relative to an LSB [26], [34], [35]. It fails catastrophically for small signals and many particularly simple (e.g., sinusoidal) signals, for which the quantization error retains the character of input-dependent distortion, rather than noise. The mid-tread quantization of a small signal of peak amplitude less than 0.5 LSB provides a simple example of this failure; the quantizer output is null, and the quantization error is just the input sign-inverted. Such an error is not uniformly distributed, iid, or independent of the input. In such cases, application of an appropriate dither can be used to temper the statistical properties of the error signal.

B. Dither: Subtractive versus Nonsubtractive

Schematics of subtractively dithered and nonsubtractively dithered quantizing systems are shown in Fig. 3. In each case, we denote the *system input* by x and the *system output* by y. We thus distinguish the system input from the *quantizer input*, which we continue to denote by w and which is given by $w = x + \nu$. ν represents the *dither* signal, which is a strict-sense stationary random process assumed to be statistically independent of x. Similarly, the *total error* of each quantizing system is defined as the difference between the system output and system input, and is denoted by $\varepsilon \triangleq y - x$ to distinguish it from the quantizer error $q \triangleq Q(w) - w$.

The total errors introduced by subtractively dithered and nonsubtractively dithered systems are not identical. In a subtractively dithered system, the dither is subtracted from the quantizer output to yield the system output. Hence, for such a system

$$\begin{aligned}\varepsilon &= Q(x+\nu) - (x+\nu) \\ &= q(x+\nu).\end{aligned}$$

On the other hand, for the nonsubtractively dithered system

$$\begin{aligned}\varepsilon &= Q(x+\nu) - x \\ &= q(x+\nu) + \nu.\end{aligned}$$

In neither case is the total error equal to $q(x)$, as in an undithered system (i.e., one for which $\nu \equiv 0$), although in an SD system, the total error does equal the quantization error associated with the *total* quantizer input w.

It has been shown by Schuchman [3] that the total error induced by an SD quantizing system can be rendered uniformly distributed for arbitrary input distributions if and only if the dither's *characteristic function* or *cf* (the Fourier transform of its pdf [36], [37]) obeys a certain condition. Defining the Fourier transform operator $\mathcal{F}[\,]$ by

$$\mathcal{F}[f](u) \triangleq F(u) \triangleq \int_{-\infty}^{\infty} f(x) e^{-j2\pi ux}\, dx$$

and denoting the dither pdf and cf as $p_\nu(\nu)$ and $P_\nu(u)$, respectively, Schuchman's condition is that

$$P_\nu\left(\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0 \tag{4}$$

where we take this opportunity to define the set $\mathbf{Z}_0^n$ as the set of all n-vectors with integer components with the exception of the zero vector $\mathbf{0} = (0, 0, \ldots, 0)$, i.e., $\mathbf{Z}_0^n = \mathbf{Z}^n \setminus \mathbf{0}$. (Note that this definition *does not* correspond to a Cartesian product of $\mathbf{Z}_0$'s.)

Furthermore, it can be shown [4], [9], [10], [27] that the total error in an SD quantizing system is statistically independent of the system input if and only if (4) holds. Thus, dither obeying Schuchman's condition renders the error statistically independent of the input and uniformly distributed. In particular, it exhibits a variance of $\Delta^2/12$. In these regards, then, it resembles the idealized quantization error of the CMQ. The simplest

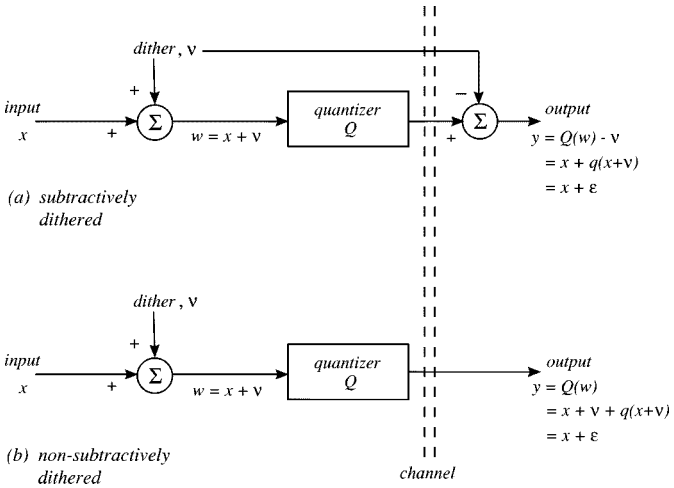

Fig. 3. Dithered quantizing systems. (a) Subtractively dithered (SD). (b) Nonsubtractively dithered (NSD).

random process satisfying Schuchman's condition is one exhibiting a uniform pdf $p_\nu(\nu) = \Pi_\Delta(\nu)$, whose associated characteristic function is a "sinc" function

$$P_\nu(u) = \operatorname{sinc}(u) \triangleq \frac{\sin(\pi \Delta u)}{\pi \Delta u}.$$

It can also be shown [4], [9], [10], [27] that subtractive dither will render distinct samples of the total error signal statistically independent of one another for arbitrary input distributions if and only if

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1, k_2) \in \mathbf{Z}_0^2 \tag{5}$$

where ν_1 and ν_2 represent dither values separated in time by $\tau \neq 0$, and where $P_{\nu_1,\nu_2}(u_1, u_2)$ represents their joint characteristic function [the two-dimensional (2-D) Fourier transform of their joint pdf $p_{\nu_1,\nu_2}(\nu_1, \nu_2)$]. This condition is satisfied by any dither that is iid so that $P_{\nu_1,\nu_2}(u_1, u_2) = P_\nu(u_1) P_\nu(u_2)$ and satisfies (4). For instance, this means that a subtractively dithered quantizing system employing iid dither of uniform distribution produces an iid (hence white) total error signal whose values are uniformly distributed and statistically independent of the input. The error thus behaves like a purely additive independent white noise process, as postulated by the CMQ. This beautiful result represents the ideal outcome for a quantization operation.

Unfortunately, subtractive dithering is difficult to use in many practical systems since the dither signal must be available at each end of the channel. This requires either the transmission of the dither values or the use of synchronized noise sources (pseudo-random number generators) separated, in general, by both time and distance. Furthermore, any digital processing of the dithered signal would necessitate processing of the dither prior to subtraction. For reasons such as these, the possibility of using dither without subsequently subtracting it is frequently of interest.

We will see that nonsubtractively dithered systems, as distinct from subtractively dithered ones, *cannot* render the total error statistically independent of the input. Neither can they make

temporally separated values of the total error statistically independent of one another. However, we will prove that they *can* render any desired statistical moments of the error signal independent of the input and regulate the joint moments of errors that are separated in time. The theory underlying these features of nonsubtractive dither is developed in Section II and is subsequently used to explore the properties of some practical dither signals in Section III. Important previously unpublished results regarding the optimality of triangular-pdf dither and the properties of a form of spectrally colored dither we call *highpass dither* are included. Section IV presents new results concerning the important special case of quantizing systems in which the available dither is discrete valued while Section V summarizes the most significant observations and conclusions.

II. Nonsubtractive Dither Theory

A. Total Error PDF's

The dependence of the total error on the system input can be analyzed in terms of its pdf as a function of a specified input value. This function is referred to as the *conditional pdf* (cpdf) of the total error and is denoted $p_{\varepsilon|x}(\varepsilon, x)$ throughout the following discussion.

In order to derive an expression for $p_{\varepsilon|x}(\varepsilon, x)$, we consider a nonsubtractively dithered quantizing system, as in Fig. 3(b), with a specified system input value x. The input to the quantizer is $w = x + \nu$, which is the sum of the system input and the statistically independent dither process. This sum has a cpdf $p_{w|x}(w, x) = p_\nu(w - x)$.

Fig. 4 shows that the total error depends not only on the system input value but on the value of the dither as well. In particular, if the input to the quantizer w is between $-\Delta/2$ and $+\Delta/2$, the output will be null (for a mid-tread characteristic) so that the error is $\varepsilon = -x$. Similarly, if the input to the quantizer is between $+\Delta/2$ and $+3\Delta/2$, the output will be $+\Delta$ so that $\varepsilon = -x + \Delta$. Hence, the pdf of the error for a fixed input is a series of delta functions separated by intervals of Δ, where each is weighted by the probability that w falls on the corresponding quantizer step

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k=-\infty}^{\infty} \delta(\varepsilon + x - k\Delta) \int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta} p_\nu(w - x)\, dw.$$

In the parlance of Widrow [7], the error cpdf is an *area sampled* version of the quantizer input cpdf.

Writing the integral in the last equation as a convolution (which is denoted by $\star$) of p_ν with a rectangular window function $\Delta\Pi_\Delta$ reduces it to

$$p_{\varepsilon|x}(\varepsilon, x) = [\Delta\Pi_\Delta \star p_\nu](\varepsilon) W_\Delta(\varepsilon + x) \tag{6}$$

where

$$W_\Gamma(\varepsilon) \triangleq \sum_{k=-\infty}^{\infty} \delta(\varepsilon - k\Gamma)$$

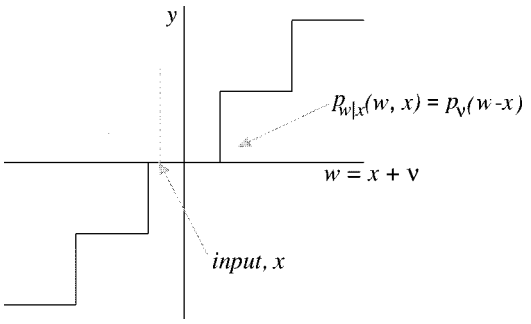

Fig. 4. Conditional pdf of the quantizer input showing its justification relative to the quantizer transfer characteristic.

is a train of Dirac delta functions separated by intervals of width Γ. [2] Thus, the pdf of ε is given by

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x) p_x(x)\, dx \\ &= [\Delta\Pi_\Delta \star p_\nu](\varepsilon)[W_\Delta \star p_x](-\varepsilon). \end{aligned} \tag{7}$$

As discussed in Section I-B in association with subtractively dithered systems, the quantization error $q(w)$ *will* be statistically independent of x and uniformly distributed if the dither statistics obey Schuchman's condition (4). Unfortunately, $q(w)$ is not the total error of a nonsubtractively dithered system. Indeed, we will now show the following:

Theorem 1: In an NSD quantizing system, it is not possible to render the total error either statistically independent of the system input or uniformly distributed for system inputs of arbitrary distribution.

Proof: In (6), it is clear that $p_{\varepsilon|x}(\varepsilon, x)$ *cannot* be rendered independent of x by any choice of dither pdf since the convolution of any dither pdf (which must be nonnegative everywhere) with a rectangular window function yields a function that is at least as wide as the rectangular window. Hence, at least one delta function always makes a contribution to the sum, and the position of that delta function is dependent on the system input. (A different proof that is based on the properties of characteristic functions may be found in [26].)

Taking the Fourier transform of (7), we find that the characteristic function of ε is given by

$$\begin{aligned} P_\varepsilon(u) &= [\mathrm{sinc}(u) P_\nu(u)] \star \left[W_{\frac{1}{\Delta}}(-u) P_x(-u) \right] \\ &= \sum_{k=-\infty}^{\infty} \mathrm{sinc}\left(u - \frac{k}{\Delta}\right) P_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right) \end{aligned} \tag{8}$$

where P_x is the arbitrary cf of the input signal, and P_ν is the cf of the dither. In order for ε to be uniformly distributed, this must reduce to $\mathrm{sinc}(u)$ for some choice of P_ν. Suppose that this is possible, in which case, we obtain

$$\mathrm{sinc}(u) = \sum_{k=-\infty}^{\infty} \mathrm{sinc}\left(u - \frac{k}{\Delta}\right) P_\nu\left(u - \frac{k}{\Delta}\right) P_x\left(-\frac{k}{\Delta}\right).$$

[2] A problem arises in the formalism if the dither is null [$p_\nu(\nu) = \delta(\nu)$] and the system input occurs at a quantizer step edge since the product of the generalized functions $W_\Delta(\nu - (2n+1)\Delta/2)$ and $\Pi_\Delta(\nu)$ is not conventionally defined. It is shown in [26] that an appropriate definition of this product for the purposes at hand is $1/2[\delta(\nu - n\Delta) + \delta(\nu - (n+1)\Delta)]$.

Now, let $u = \ell/\Delta$, where $\ell \in \mathbf{Z}_0$. Then, we have

$$\operatorname{sinc}\left(\frac{\ell}{\Delta}\right) = 0 = P_x\left(-\frac{\ell}{\Delta}\right)$$

which contradicts the assumption that P_x is arbitrary. Thus, the total error cannot be made uniformly distributed in a nonsubtractively dithered system for inputs of arbitrary distribution. □

The counterintuitive nature of this result is the source of much confusion regarding NSD systems. For instance, it is tempting to accept the following line of reasoning. Suppose that a dither satisfying Schuchman's condition (4) is used so that q is independent of x. Then, since ν is also independent of x, the total error $\varepsilon = q + \nu$ is the sum of two random processes , both of which are independent of x and, thus, should be independent of x as well. This conclusion is flatly false. In an NSD quantizing system, given the value of x, we know that the possible values of $q + \nu$ satisfy the equation $q + \nu = -x + k\Delta$, $k \in \mathbf{Z}$ so that the distribution of $q + \nu$ is highly dependent on x, in agreement with Theorem 1. To further elucidate the source of the problem, we consider the following. For arbitrary random variables q, ν, and x and a fourth $\varepsilon = q + \nu$ (none of these *necessarily* represent quantities in a quantizing system), it is clear that $p_{\varepsilon|q,\nu,x}(\varepsilon, q, \nu, x) = \delta(\varepsilon - q - \nu)$. Then, by direct integration, $p_{\varepsilon,x}(\varepsilon, x) = \int_{-\infty}^{\infty} p_{q,\nu,x}(\varepsilon - \nu, \nu, x)\, d\nu$, the Fourier transform of which yields the joint cf of ε and x as $P_{\varepsilon,x}(u_\varepsilon, u_x) = P_{q,\nu,x}(u_\varepsilon, u_\varepsilon, u_x)$. By definition, ε and x are statistically independent of each other if and only if $P_{\varepsilon,x}(u_\varepsilon, u_x)$ can be written as a product of two functions, where one involves u_ε alone, whereas the other involves u_x alone. We see that this is the case if and only if $P_{q,\nu,x}(u_\varepsilon, u_\varepsilon, u_x) = P_{q,\nu}(u_\varepsilon, u_\varepsilon)P_x(u_x)$. (We note that $P_{q,\nu}(u_\varepsilon, u_\varepsilon)$ is the cf of $\varepsilon = q + \nu$.) Unfortunately, we know only that $P_{q,x}(u_q, u_x) = P_q(u_q)P_x(u_x)$ and $P_{\nu,x}(u_\nu, u_x) = P_\nu(u_\nu)P_x(u_x)$, which is insufficient to establish the result. (The signals in an NSD quantizing system furnish a convenient counterexample.)

Since we have shown that statistical independence of the total error from the system input is not achievable, we now turn our attention to the possibility of controlling moments of the error. For many applications, controlling relevant error moments is just as good as having full statistical independence of the input and error processes.

B. A Condition for the Independence of Total Error Moments

The mth moment of the error signal is the expectation value of ε^m

$$E[\varepsilon^m] = \int_{-\infty}^{\infty} \varepsilon^m p_\varepsilon(\varepsilon)\, d\varepsilon.$$

It can be shown that these moments may also be expressed in terms of the cf of the given random variable as [37]

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0) \tag{9}$$

where $P_\varepsilon^{(m)}$ denotes the mth derivative of P_ε. (This expression is easily derived by differenting with respect to u the definition of $P_\varepsilon(u)$ as the Fourier transform of p_ε.)

From (8), we obtain

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\Delta}\right) P_x\left(\frac{k}{\Delta}\right) \tag{10}$$

where

$$G_\nu(u) \triangleq \operatorname{sinc}(u) P_\nu(u). \tag{11}$$

Since the cf P_x of the system input is arbitrary, we obtain the following result [27].

Theorem 2: In an NSD quantizing system, $E[\varepsilon^m]$ is independent of the distribution of the system input x if and only if

$$G_\nu^{(m)}\left(\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0. \tag{12}$$

If the conditions of Theorem 2 are satisfied, from (10)

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m G_\nu^{(m)}(0),$$

which is precisely the mth moment of a notional random process with cf G_ν and pdf $\Delta\Pi_\Delta \star p_\nu$, although this is not, of course, the pdf of ε. We can derive the following expressions for the moments of the total error in terms of the moments of the dither signal by direct differentiation of $G_\nu(u)$:

$$E[\varepsilon] = E[\nu] \tag{13}$$

$$E[\varepsilon^2] = E[\nu^2] + \frac{\Delta^2}{12} \tag{14}$$

$$E[\varepsilon^m] = \sum_{\ell=0}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{2\ell} \left(\frac{\Delta}{2}\right)^{2\ell} \frac{E[\nu^{m-2\ell}]}{2\ell+1}. \tag{15}$$

We emphasize that each of these equations for $E[\varepsilon^m]$ is only valid when Theorem 2 is satisfied for that particular value of m and that the validity of one of these equations does not imply the validity of any others corresponding to different m values.

Equation (14) merits special comment. It indicates that if the total error variance in an NSD quantizing system is input independent, then it always exceeds that of an SD system (or a system described by the CMQ) by an amount equal to the variance of the dither. This characteristic increase in the error power is not problematic in most multibit applications, and the benefits of dithering typically far outweigh the slight noise penalty.

Two corollaries to Theorem 2 follow.

Corollary 1: In an NSD quantizing system, if the condition (12) is satisfied for any given m, then for any choice of n

$$E[\varepsilon^m x^n] = E[\varepsilon^m]E[x^n]$$

i.e., ε^m and x^n are uncorrelated.

Proof: We observe that if $p_x(x) = \delta(x - x_0)$, then

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \int_{-\infty}^{\infty} p_{\varepsilon,x}(\varepsilon, x)\, dx \\ &= \int_{-\infty}^{\infty} p_{\varepsilon|x}(\varepsilon, x)\delta(x - x_0)\, dx \\ &= p_{\varepsilon|x}(\varepsilon, x_0). \end{aligned} \tag{16}$$

By Theorem 2, $E[\varepsilon^m]$ is independent of the choice of p_x, and in particular, it is independent of the choice of x_0 when $p_x(x) = \delta(x - x_0)$, as above. Thus, $\int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)\, d\varepsilon = E[\varepsilon^m]$ for any x. In this case

$$\begin{aligned} E[\varepsilon^m x^n] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon^m x^n p_{\varepsilon,x}(\varepsilon, x)\, d\varepsilon\, dx \\ &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)\, d\varepsilon\right] x^n p_x(x)\, dx \\ &= \int_{-\infty}^{\infty} E[\varepsilon^m] x^n p_x(x)\, dx \\ &= E[\varepsilon^m]E[x^n]. \end{aligned}$$ □

In particular, if $E[\varepsilon]$ is independent of the distribution of x, then ε and x are uncorrelated in the usual mathematical sense that $E[\varepsilon x] = E[\varepsilon]E[x]$.

The second corollary is somewhat better known than Theorem 2 itself but demands satisfaction of a stronger condition [11], [32].

Corollary 2: In an NSD quantizing system, $E[\varepsilon^\ell]$ is independent of the distribution of the system input x for $\ell = 1, 2, \ldots, m$ if and only if

$$P_\nu^{(i)}\left(\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0 \quad \text{and} \quad i = 0, 1, 2, \ldots, m-1.$$

Proof: Proof of the "if" direction follows immediately from repeated differentiation of (11), where

$$G_\nu^{(\ell)}(u) = \sum_{i=0}^{\ell} \binom{\ell}{i} \operatorname{sinc}^{(\ell-i)}(u) P_\nu^{(i)}(u).$$

We see that the ℓth and all lower derivatives of G_ν will all go to zero at $u = k/\Delta$, $k \in \mathbf{Z}_0$ if the first $\ell - 1$ derivatives of P_ν do. The "only if" direction is easily proven using induction, but this requires more space than is justified here (see [26]). □

In most practical applications, we are interested in dither signals that satisfy the conditions of Corollary 2, and it turns out that the conditions of this corollary will be of interest when we examine the statistics of the quantizer output (Section II-D) and the special nature of digital dither signals (Section IV).

C. Second-Order Statistics of Total Error Values

We now begin an investigation into the joint statistics of temporally separated total error values, corresponding to input samples separated in time, in order to derive conclusions about the spectral characteristics of the total error sequence.

Consider two total error values ε_1 and ε_2, which are separated in time by $\tau \neq 0$. (In the special case where $\tau = 0$, the analysis reduces to that of Section II-B.) The corresponding system input values will be denoted as x_1 and x_2, respectively. Employing a derivation analogous to that of Section II-A, we find that

$$\begin{aligned} &p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2, x_1, x_2) \\ &\quad = \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \delta(\varepsilon_1 + x_1 - k_1\Delta)\delta(\varepsilon_2 + x_2 - k_2\Delta) \\ &\qquad \times \int_{-\frac{\Delta}{2}+k_1\Delta}^{\frac{\Delta}{2}+k_1\Delta}\int_{-\frac{\Delta}{2}+k_2\Delta}^{\frac{\Delta}{2}+k_2\Delta} p_{\nu_1,\nu_2}(w_1 - x_1, w_2 - x_2)\, dw_1 dw_2 \\ &\quad = [\Delta^2 \Pi_{\Delta\Delta} \star p_{\nu_1,\nu_2}](\varepsilon_1, \varepsilon_2) W_{\Delta\Delta}(\varepsilon_1 + x_1, \varepsilon_2 + x_2) \end{aligned}$$

where the convolution is 2-D, involving both ε_1 and ε_2, and where

$$\Pi_{\Gamma\Gamma}(\varepsilon_1, \varepsilon_2) \triangleq \Pi_\Gamma(\varepsilon_1)\Pi_\Gamma(\varepsilon_2)$$

and

$$W_{\Gamma\Gamma}(\varepsilon_1, \varepsilon_2) \triangleq W_\Gamma(\varepsilon_1) W_\Gamma(\varepsilon_2).$$

p_{ν_1,ν_2} represents the *joint pdf* of the dither values ν_1 and ν_2 associated with the inputs x_1 and x_2, respectively.

Hence, we have (17), shown at the bottom of the page. The joint characteristic function of ε_1 and ε_2 is found by taking the 2-D Fourier transform of (17) with respect to ε_1 and ε_2, resulting in an expression in the corresponding frequency variables u_1 and u_2

$$\begin{aligned} &P_{\varepsilon_1,\varepsilon_2}(u_1, u_2) \\ &\quad = \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} \operatorname{sinc}\left(u_1 - \frac{k_1}{\Delta}\right)\operatorname{sinc}\left(u_2 - \frac{k_2}{\Delta}\right) \\ &\qquad \times P_{\nu_1,\nu_2}\left(u_1 - \frac{k_1}{\Delta}, u_2 - \frac{k_2}{\Delta}\right) P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right). \end{aligned} \tag{18}$$

No choice of dither pdf will allow (18) to be expressed as a product of two characteristic functions, where one involves u_1 alone and the other u_2 alone for arbitrary choices of P_{x_1,x_2}. Thus, ε_1 and ε_2 cannot be rendered statistically independent for arbitrary joint input distributions. Let us therefore proceed to investigate the joint moments of ε_1 and ε_2 in the hope that we can exercise some control over them by an appropriate choice of the dither statistics.

The (m_1, m_2)*th joint moment* of the two signals of interest is given by

$$\begin{aligned} E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] &\triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varepsilon_1^{m_1}\varepsilon_2^{m_2} p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2)\, d\varepsilon_1\, d\varepsilon_2 \\ &= \left(\frac{j}{2\pi}\right)^{m_1+m_2} P_{\varepsilon_1,\varepsilon_2}^{(m_1,m_2)}(0, 0) \end{aligned} \tag{19}$$

$$\begin{aligned} p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2, x_1, x_2) p_{x_1,x_2}(x_1, x_2)\, dx_1\, dx_2 \\ &= [\Delta^2 \Pi_{\Delta\Delta} \star p_{\nu_1,\nu_2}](\varepsilon_1, \varepsilon_2)[W_{\Delta\Delta} \star p_{x_1,x_2}](-\varepsilon_1, -\varepsilon_2). \end{aligned} \tag{17}$$

where

$$P_{\varepsilon_1,\varepsilon_2}^{(m_1,m_2)}(u_1,u_2) \triangleq \frac{\partial^{(m_1+m_2)} P_{\varepsilon_1,\varepsilon_2}}{\partial u_1^{m_1} \partial u_2^{m_2}}(u_1,u_2).$$

Substituting (18) into (19), we find that

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(-\frac{k_1}{\Delta},-\frac{k_2}{\Delta}\right) \tag{20}$$

where

$$G_{\nu_1,\nu_2}(u_1,u_2) \triangleq \operatorname{sinc}(u_1)\operatorname{sinc}(u_2) P_{\nu_1,\nu_2}(u_1,u_2).$$

At this point, we may deduce a theorem that represents a second-order analog of Theorem 2.

Theorem 3: In an NSD quantizing system, the (m_1,m_2)th joint moment $E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}]$ of two total error values ε_1 and ε_2 separated in time by $\tau \neq 0$ is independent of the system input for arbitrary input distributions if and only if

$$G_{\nu_1,\nu_2}^{(m_1,m_2)}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1,k_2) \in \mathbf{Z}_0^2. \tag{21}$$

The proof is completely analogous to that of Theorem 2. When (21) is satisfied, we have

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} G_{\nu_1,\nu_2}^{(m_1,m_2)}(0,0) \tag{22}$$

so that by explicitly performing the differentiation, we can write an expression that is analogous to (15), relating the joint moments of the total error to those of the dither

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \sum_{\ell_1=0}^{\lfloor \frac{m_1}{2} \rfloor}\sum_{\ell_2=0}^{\lfloor \frac{m_2}{2} \rfloor} \binom{m_1}{2\ell_1}\binom{m_2}{2\ell_2}\left(\frac{\Delta}{2}\right)^{2(\ell_1+\ell_2)} \times \frac{E[\nu_1^{m_1-2\ell_1}\nu_2^{m_2-2\ell_2}]}{(2\ell_1+1)(2\ell_2+1)}. \tag{23}$$

We attach the caveat that satisfaction of (23) for some particular m_1 and m_2 does not imply its satisfaction for any other values thereof.

If the dither process is iid so that ν_1 and ν_2 are statistically independent, we have $P_{\nu_1,\nu_2}(u_1,u_2) = P_\nu(u_1)P_\nu(u_2)$. Then, if the conditions of Corollary 2 are satisfied for $m = \max(m_1,m_2)$, we have

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = \left(\frac{j}{2\pi}\right)^{m_1+m_2} G_\nu^{(m_1)}(0) G_\nu^{(m_2)}(0) \tag{24}$$

$$= E[\varepsilon_1^{m_1}]E[\varepsilon_2^{m_2}] \tag{25}$$

so that $\varepsilon_1^{m_1}$ and $\varepsilon_2^{m_2}$ are uncorrelated. In this case, of course, $E[\varepsilon_1^m] = E[\varepsilon_2^m] = E[\varepsilon^m]$. Hence, we have the following corollary.

Corollary 3: Any iid nonsubtractive dither signal that satisfies the conditions of Corollary 2 for $m = \max(m_1,m_2)$ will ensure that for two error values ε_1 and ε_2 separated in time by $\tau \neq 0$, we have

$$E[\varepsilon_1^{m_1}\varepsilon_2^{m_2}] = E[\varepsilon^{m_1}]E[\varepsilon^{m_2}].$$

In this case, $E[\varepsilon^{m_2}]$ and $E[\varepsilon^{m_2}]$ will be given by (15). In particular, for an iid dither with zero mean, we note that $E[\varepsilon_1\varepsilon_2] = 0$.

In a digital system, the total error is a discrete-time signal; therefore, $\tau = kT$, where T represents the sampling period, and $k \in \mathbf{Z}$. The *autocorrelation function* of such a signal is defined to be $E[\varepsilon_1\varepsilon_2](k)$. The *power spectral density* (PSD) of a discrete-time random process is equal, by definition, to the discrete-time Fourier transform (DTFT) of its autocorrelation function, where we define the DTFT as

$$\mathcal{F}_{\mathrm{DT}}[h](f) \triangleq 2T \sum_{k=-\infty}^{\infty} h(k) e^{-j2\pi fkT} \tag{26}$$

where the continuous frequency variable f is in hertz if T is in seconds. This definition is normalized such that the integral of the PSD from zero to the Nyquist frequency $1/2T$ yields the variance of the signal.

Using (14), we find that for an NSD quantizing system using iid dither satisfying the conditions of Corollary 2 for $m = 2$, the autocorrelation function of the error is

$$E[\varepsilon_1\varepsilon_2](k) = \begin{cases} E[\nu^2] + \frac{\Delta^2}{12}, & k = 0 \\ E^2[\nu], & \text{otherwise.} \end{cases}$$

Comparing this with the autocorrelation function of the dither sequence

$$E[\nu_1\nu_2](k) = \begin{cases} E[\nu^2], & k = 0 \\ E^2[\nu], & \text{otherwise} \end{cases}$$

we conclude that

$$\mathrm{PSD}_\varepsilon(f) = \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6}$$

so that the total error signal must be spectrally white since the dither is spectrally white (apart from a dc component if the dither is not zero mean).

Using the product rule to explicitly perform the differentiations in (21), it is straightforward to derive conditions that ensure its satisfaction for the case where $m_1 = m_2 = 1$ but do not require statistical independence of distinct dither values [25], [26]. This will allow the use of certain dither signals that are not spectrally white.

Theorem 4: In an NSD system where all dither values are statistically independent of all system input values

$$E[\varepsilon_1\varepsilon_2] = E[\nu_1\nu_2] \tag{27}$$

for arbitrary input distributions if and only if the following three conditions are satisfied:

$$P_{\nu_1,\nu_2}\left(\frac{k_1}{\Delta},\frac{k_2}{\Delta}\right) = 0 \quad \forall (k_1,k_2) \in \mathbf{Z}_0^2 \tag{28}$$

$$P_{\nu_1,\nu_2}^{(0,1)}\left(\frac{k_1}{\Delta},0\right) = 0 \quad \forall k_1 \in \mathbf{Z}_0 \tag{29}$$

and

$$P^{(1,0)}_{\nu_1,\nu_2}\left(0,\frac{k_2}{\Delta}\right)=0 \quad \forall k_2\in\mathbf{Z}_0. \tag{30}$$

This can be thought of as a second-order counterpart to Corollary 2 for the simple case $m_1=m_2=1$. When the conditions of the theorem are satisfied, (27) follows immediately from direct differentiation of (22). It is possible to gain further insight into the meaning of the conditions involved by noting that

$$\begin{aligned}E[\varepsilon_1\varepsilon_2]=&E[(q_1+\nu_1)(q_2+\nu_2)]\\ =&E[\nu_1\nu_2]+E[q_1\nu_2]+E[q_2\nu_1]+E[q_1q_2].\end{aligned}$$

The last term is equal to zero as long as (28) is satisfied (see [27, (5), Th. 2]), whereas it can be shown that the second and third terms vanish subject to the satisfaction of (29) and (30), thus yielding (27). Necessity of the conditions follows from the arbitrariness of the input distribution.

Suppose that the conditions of both Theorem 4 and Corollary 2 with $m=2$ are satisfied. Then, the autocorrelation function of the error is given by

$$E[\varepsilon_1\varepsilon_2](k)=\begin{cases}E[\nu^2]+\frac{\Delta^2}{12}, & k=0\\ E[\nu_1\nu_2], & \text{otherwise.}\end{cases} \tag{31}$$

This indicates that the power spectrum of the error will be identical to the power spectrum of the dither, apart from a contribution due to the $k=0$ case, manifested as an additive constant present at all frequencies (i.e., a white spectral component introduced by the properly dithered quantization operation). Hence, as before

$$\text{PSD}_\varepsilon(f)=\text{PSD}_\nu(f)+\frac{\Delta^2 T}{6} \tag{32}$$

except that now, the dither PSD is not necessarily white. This will be illustrated by the discussion of highpass dither in Section III-E.

D. Statistics of the System Output

It is of interest to express the statistical attributes of the output y in terms of those of the input x since it is frequently required that one be deduced from the other. We apply the same brand of reasoning as used to determine the cpdf of the total error in Section II-A. The values of the quantizer output are restricted to values of $k\Delta$, $k\in\mathbf{Z}$. Therefore, $p_y(y)$ will consist of delta functions at these locations weighted by the probability that the quantizer input $w=x+\nu$ falls in the range of the corresponding quantizer step $(2k-1)\Delta/2<w<(2k+1)\Delta/2$. This probability is just the integral of $p_w(w)$ over this range, where, since x and ν are statistically independent, p_w is given by [39] $p_w(w)=[p_\nu\star p_x](w)$. Thus, we have

$$\begin{aligned}p_y(y)&=\sum_{k=-\infty}^{\infty}\delta(y-k\Delta)\int_{-\frac{\Delta}{2}+k\Delta}^{\frac{\Delta}{2}+k\Delta}[p_\nu\star p_x](w)\,dw\\ &=[\Delta\Pi_\Delta\star p_\nu\star p_x](y)W_\Delta(y).\end{aligned}$$

Taking the Fourier transform of this expression yields

$$\begin{aligned}P_y(u)&=[G_\nu(u)P_x(u)]\star W_{\frac{1}{\Delta}}(u)\\ &=\sum_{k=-\infty}^{\infty}G_\nu\left(u-\frac{k}{\Delta}\right)P_x\left(u-\frac{k}{\Delta}\right)\end{aligned} \tag{33}$$

and therefore

$$\begin{aligned}E[y^m]&=\left(\frac{j}{2\pi}\right)^m P_y^{(m)}(0)\\ &=\sum_{k=-\infty}^{\infty}\sum_{r=0}^{m}\binom{m}{r}\left[\left(\frac{j}{2\pi}\right)^r G_\nu^{(r)}\left(\frac{k}{\Delta}\right)\right]\\ &\quad\times\left[\left(\frac{j}{2\pi}\right)^{m-r}P_x^{(m-r)}\left(\frac{k}{\Delta}\right)\right].\end{aligned} \tag{34}$$

Now, if the first m derivatives of $G_\nu(u)$ are zero at all nonzero multiples of $1/\Delta$, then (34) reduces to

$$E[y^m]=\sum_{r=0}^{m}\binom{m}{r}E[\varepsilon^r]E[x^{m-r}] \tag{35}$$

where the expectation values of the total error are given in terms of the expectation values of the dither by (15). By direct differentiation of $G_\nu(u)$, the above condition is easily shown to be equivalent to the condition of Corollary 2. For the special cases $m=1$ and $m=2$, we note that

$$E[y]=E[x]+E[\varepsilon]=E[x]+E[\nu]$$

and

$$E[y^2]=E[x^2]+2E[x]E[\nu]+E[\nu^2]+\frac{\Delta^2}{12}$$

where (13) and (14) have been substituted for the error moments.[3]

Proceeding similarly for the joint moments of output values y_1 and y_2 separated in time by $\tau\neq 0$, we find (36), shown at the bottom of the next page. If the indicated partial derivatives of G_{ν_1,ν_2} are zero for all $(k_1,k_2)\in\mathbf{Z}_0^2, r_i=1,2,\ldots,m_i$, $i\in\{1,2\}$, then (36) reduces to

$$\begin{aligned}E[y_1^{m_1}y_2^{m_2}]=&\sum_{r_1=0}^{m_1}\sum_{r_2=0}^{m_2}\binom{m_1}{r_1}\binom{m_2}{r_2}\\ &\times E[\varepsilon_1^{r_1}\varepsilon_2^{r_2}]E[x_1^{m_1-r_1}x_2^{m_2-r_2}]\end{aligned} \tag{37}$$

[3]Note that the so-called *dither averaged transfer characteristic* $E[y\mid x]$ is given by

$$E[y\mid x]=E[Q(x+\nu)\mid x]=\int_{-\infty}^{\infty}Q(x+\nu)p_\nu(\nu)d\nu=Q(x)*p_\nu(-x)$$

which is the convolution of the quantizer staircase with the dither pdf. For the $m=1$ case, this defines the line $y=x$ (see the illustrations in [17] and [18]).

where the joint moments of the total error are given in terms of those of the dither by (23).

Beginning from (36) with $m_1 = m_2 = 1$, it is straightforward to show that if the conditions of Theorem 4 are satisfied, i.e., (28)–(30), then $E[y_1 y_2] = E[x_1 x_2] + E[\nu_1 \nu_2]$ so that, with the aid of (35) and (15), we find that the output has an autocorrelation function

$$E[y_1 y_2](k) = \begin{cases} E[x^2] + 2E[x]E[\nu] + E[\nu^2] + \frac{\Delta^2}{12}, & k = 0 \\ E[x_1 x_2] + E[\nu_1 \nu_2], & \text{otherwise.} \end{cases} \tag{38}$$

Then, the spectrum of the output is the sum of the input and dither spectra apart from a white noise component, which is contributed by the $k = 0$ case of (38). The latter component is comparable with the white "quantization noise" posited in the CMQ. In particular, for a system using a zero-mean dither

$$\text{PSD}_y(f) = \text{PSD}_x(f) + \text{PSD}_\nu(f) + \frac{\Delta^2 T}{6}.$$

III. Error Moments in Some Representative Systems

We proceed to apply the above results to realizable quantizing systems using those dither signals that we consider to be of greatest interest in practical applications. Some of the many other possible dither signals are investigated in [16], [18], and [38].

A. Null Dither

We begin by considering an undithered system. The pdf of a "null dither" is $p_\nu(\nu) = \delta(\nu)$, the Fourier transform of which is equal to unity everywhere. Hence, by (11)

$$G_\nu(u) = \text{sinc}(u).$$

No derivatives of this function vanish at nonzero multiples of $1/\Delta$; therefore, no moments of the total error (excepting the zeroth) will be independent of the input distribution. Of course, it is not expected that they would be. We know that in the absence of dither, the error is a *deterministic function* of the input. The error $q(x)$ as a function of the input is shown in Fig. 2.

B. Rectangular-PDF Dither

Now, consider a system using dither with a simple rectangular (i.e., uniform) pdf of 1 LSB peak-to-peak amplitude $p_\nu(\nu) = \Pi_\Delta(\nu)$, with a corresponding cf $P_\nu(u) = \text{sinc}(u)$. Hence, from (11)

$$G_\nu(u) = \text{sinc}^2(u).$$

The first two derivatives of this function are plotted in Fig. 5. The first derivative clearly satisfies the condition of going to zero at the regularly spaced points stipulated by (12), whereas the second derivative does not (nor do higher derivatives). This indicates that the first moment of the error signal is independent of the input, but its variance remains dependent. These conclusions are borne out by the accompanying plots in Fig. 5 of the *conditional moments*

$$E[\varepsilon^m \mid x] \triangleq \int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)\, d\varepsilon$$

as computed using (6). The first moment, or mean error, is zero for all inputs, indicating that the quantizer has been *linearized* by the use of this dither. The error variance, on the other hand, is clearly signal dependent; therefore, the noise power in the signal varies with the input. This is sometimes referred to as *noise modulation* and is undesirable in audio or video signals.

If the dither is iid, then (as shown in Section II-C) temporally separated error values will be uncorrelated. Thus, short-time error spectra will appear flat, but their level will be input dependent.

C. Triangular-PDF Dither

The most straightforward means of generating dither signals with more complicated pdf's is to simply sum two or more statistically independent RPDF random processes. For instance, the sum of two such processes ν_1 and ν_2, each of 1 LSB peak-to-peak amplitude, yields a dither with a triangular pdf (TPDF) of two LSB peak-to-peak amplitude since the summation of statistically independent random processes convolves their pdf's (see Fig. 6)

$$\begin{aligned} p_\nu(\nu) &= [p_{\nu_1} \star p_{\nu_2}](\nu) \\ &= [\Pi_\Delta \star \Pi_\Delta](\nu). \end{aligned} \tag{39}$$

Convolution of pdf's corresponds to multiplication of the respective cf's [39] so that in a system employing this kind of dither, P_ν is a squared sinc function, and G_ν is given by

$$G_\nu(u) = \text{sinc}^3(u).$$

$$E[y_1^{m_1} y_2^{m_2}] = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{r_1=0}^{m_1} \sum_{r_2=0}^{m_2} \binom{m_1}{r_1} \binom{m_2}{r_2} \left[\left(\frac{j}{2\pi} \right)^{r_1+r_2} G_{\nu_1,\nu_2}^{(r_1,r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right] \times \left[\left(\frac{j}{2\pi} \right)^{(m_1-r_1)+(m_2-r_2)} P_{x_1,x_2}^{(m_1-r_1,m_2-r_2)} \left(\frac{k_1}{\Delta}, \frac{k_2}{\Delta} \right) \right]. \tag{36}$$

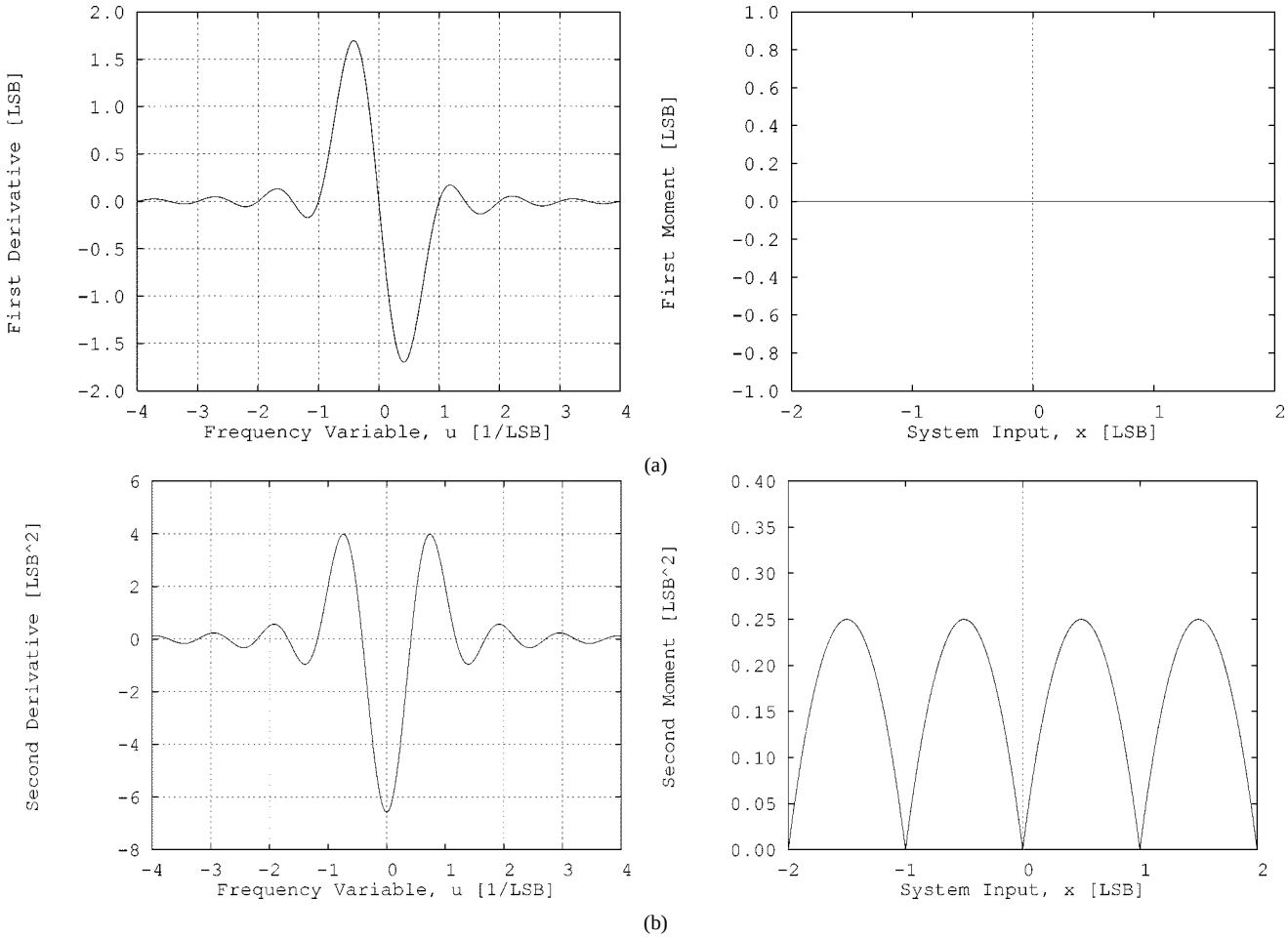

Fig. 5. Derivatives of $G_\nu(u)$ (left) and conditional moments of the error (right) for a quantizer using RPDF dither of 1 LSB peak-to-peak amplitude. (a) $G_\nu^{(1)}(u)$ and $E[\varepsilon \mid x]$ (both in units of Δ). (b) $G_\nu^{(2)}(u)$ and $E[\varepsilon^2 \mid x]$ (both in units of Δ^2). The frequency variable, u is plotted in units of $1/\Delta$ and the system input, x in units of Δ.

The first *and* second derivatives of this function go to zero at the required places; therefore, this dither renders both the first *and* second moments of the total error independent of the system input.[4] The second derivative of G_ν is shown in Fig. 7, along with the second conditional moment of the total error, which is a constant $\Delta^2/4$ for all inputs, in agreement with (14). Higher derivatives of G_ν do not meet the required conditions; therefore, higher moments of the error remain dependent on the input.

We now show that a triangular pdf of two LSB peak-to-peak amplitude is the only choice of zero-mean dither pdf, which renders the first two moments of the total error independent of the input while minimizing the second. We will begin by noting, from Corollary 2 and the stipulation of zero mean, that

$$P_\nu\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbf{Z}_0$$

and

$$P_\nu^{(1)}\left(\frac{k}{\Delta}\right) = 0, \quad \forall k \in \mathbf{Z}.$$

[4]A different proof of this result, using a direct method, was given in [17].

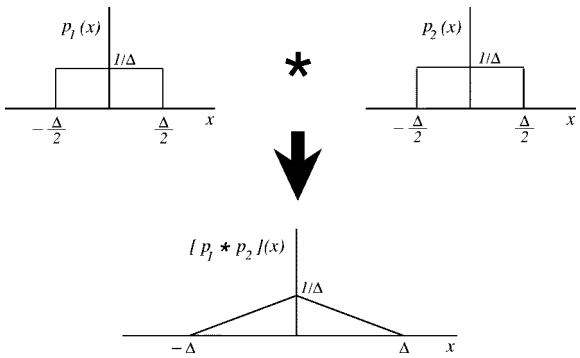

Fig. 6. Triangular pdf, formed by the convolution of two rectangular pdf's.

In addition, $P_\nu(u)$ must be equal to unity at $u = 0$ if it is to be a valid characteristic function since

$$P_\nu(0) = \int_{-\infty}^{\infty} e^{-j2\pi(0)\nu} p_\nu(\nu)\, d\nu = 1.$$

We conclude that the dither cf and its first derivative are completely specified at *all* integer multiples of $1/\Delta$. According to the generalized sampling theorem [39], this is sufficient to

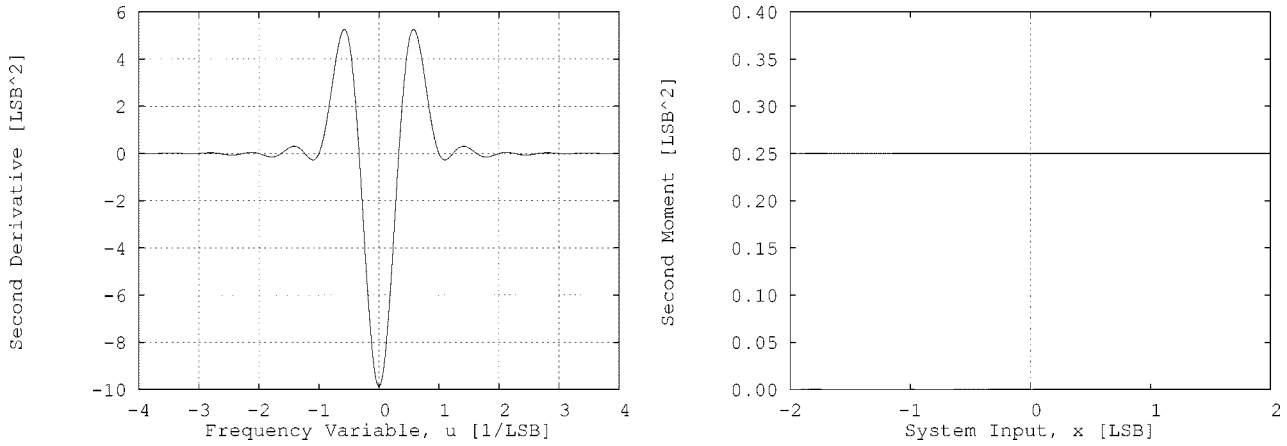

Fig. 7. $G_{\nu}^{(2)}(u)$ (left) and $E[\varepsilon^2 \mid x]$ (right) (both in units of Δ^2) for a quantizer using triangular-pdf dither of 2 LSB peak-to-peak amplitude. The frequency variable u is plotted in units of $1/\Delta$, and the system input, x in units of Δ.

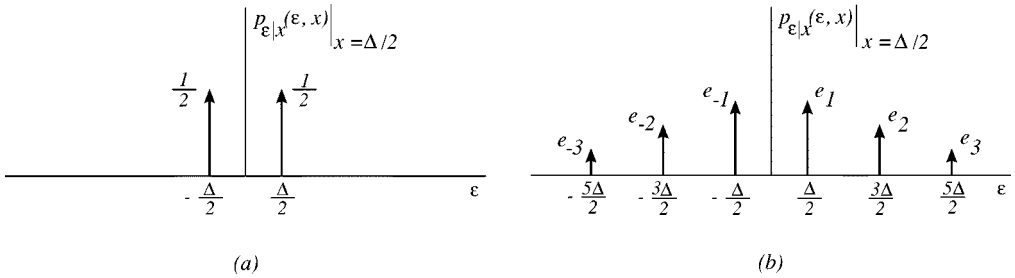

Fig. 8. $p_{\varepsilon|x}(\varepsilon, x)$ evaluated at $x = \Delta/2$ for systems using (a) a triangular-pdf dither of 2 LSB peak-to-peak amplitude and (b) a wider dither pdf (the delta functions possess the indicated weightings).

uniquely specify $P_{\nu}(u)$ for all u if $p_{\nu}(\nu)$ is Δ-bandlimited (i.e., if $p_{\nu}(\nu) = 0$ for $|\nu| \geq \Delta$). Since the pdf of (39) *is* Δ-bandlimited, and its corresponding cf satisfies all the given conditions, it must be the unique pdf in question.

It remains to be shown that any dither pdf that is not thus bandlimited will produce a greater error variance. Since this variance is assumed to be constant with respect to the input, it is sufficient to show that this holds for a single input value. We will do so for $x = \Delta/2$.

$p_{\varepsilon|x}(\varepsilon, x)$ for $x = \Delta/2$ is obtained from (7) using $p_x(x) = \delta(x - \Delta/2)$ [see (16)]. As is shown in Fig. 8(a), it consists of two equally weighted delta functions at $\varepsilon = \pm\Delta/2$ when triangular-pdf dither of two LSB peak-to-peak amplitude is employed. Use of a wider dither pdf will result in the appearance of more delta functions in the error's cpdf, as shown in Fig. 8(b), where we denote the weighting of the delta function at $\varepsilon = \pm(2i-1)\Delta/2$, $i \geq 1$, by $e_{\pm i}$, so that we have (40), shown at the bottom of the page. We proceed by expressing the fundamental condition that the integral of this pdf must equal unity

$$(e_1 + e_{-1}) + \sum_{i=2}^{\infty}(e_i + e_{-i}) = 1. \tag{41}$$

Now, by direct integration of (40), we have

$$E[\varepsilon^2 \mid x=\Delta/2] = \sum_{i=1}^{\infty}\left[(2i-1)\frac{\Delta}{2}\right]^2 (e_i+e_{-i}) = \frac{\Delta^2}{4}\left[(e_1+e_{-1}) + \sum_{i=2}^{\infty}(2i-1)^2(e_i+e_{-i})\right]. \tag{42}$$

Substituting (41) yields

$$E[\varepsilon^2 \mid x = \Delta/2] = \frac{\Delta^2}{4}\left[1 + 4\sum_{i=2}^{\infty} i(i-1)(e_i + e_{-i})\right]$$

which is always greater than $\Delta^2/4$ since the e_i's must be positive. We have thus shown the following theorem.

Theorem 5: The choice of zero-mean dither pdf that renders the first and second moments of the total error independent of the input, such that the first moment is zero and the second is minimized, is unique and is a triangular pdf of two LSB peak-to-peak amplitude.

$$p_{\varepsilon|x}\left(\varepsilon, \frac{\Delta}{2}\right) = \sum_{i=1}^{\infty}\left[e_i\delta\left(\varepsilon - (2i-1)\frac{\Delta}{2}\right) + e_{-i}\delta\left(\varepsilon + (2i-1)\frac{\Delta}{2}\right)\right]. \tag{40}$$

D. The Sum of n Independent Rectangular-PDF Random Processes

Theorem 6: A nonsubtractive dither signal generated by the summation of n statistically independent RPDF random processes renders $E[\varepsilon^\ell]$ independent of the system input distribution for $\ell = 0, 1, \ldots, n$ and results in a total error variance for $n \geq 2$ of $(n+1)\Delta^2/12$.

This must be the case since the use of n such dithers gives

$$G_\nu(u) = \text{sinc}^{n+1}(u)$$

the first n derivatives of which will consist entirely of terms containing nonzero powers of $\text{sinc}(u)$. Since this function goes to zero at the required places, the first n moments of the error will always be independent of the input. Higher derivatives will not share this property [26]. Dithers of this form are sometimes referred to as nRPDF so that, for instance, TPDF dither may also be referred to as 2RPDF.

It is important to note that using uniformly distributed processes of peak-to-peak amplitude not equal to one LSB (or, rather, not equal to an integral number of LSB's) will not render error moments independent of the input since the zeros of the associated sinc functions will not fall at integral multiples of $1/\Delta$ (see illustrations in [17]).

Finally, it is easily shown from the generalized sampling theorem that the $(n\Delta/2)$-bandlimited dither pdf that renders the first n moments of the total error independent of the input is unique and must therefore be the pdf of Theorem 6.

E. Highpass Dither

A very simple discrete-time noise generator capable of producing dither with a highpass spectrum [19], [20] is shown in Fig. 9. The system contains a pseudo-random number generator that is marked PRN, producing iid, uniformly distributed random numbers and a one-sample delay element marked z^{-1}. The output is the difference between the pseudo-random number most recently generated by the PRN η_n and the previous one η_{n-1}, i.e., $\nu_n = \eta_n - \eta_{n-1}$. The (first-order) pdf of the resulting dither sequence is triangular (TPDF) since it results from the summation of two statistically independent RPDF sequences, albeit one of these is simply a delayed version of the other. This means that all the beneficial effects of the TPDF dither discussed in Section III-C will also be associated with dither thus generated. Such highpass TPDF dither may be preferable in some audio applications since it is less audible than spectrally white TPDF dither due to the ear's reduced sensitivity at high frequencies (although these dithers have equal variances of $\Delta^2/6$). Similar comments apply regarding reduced error visibility in imaging applications. Furthermore, the use of highpass TPDF dither is more computationally efficient since it requires the calculation of only *one* new RPDF random number per sample as compared with two when iid TPDF dither is used.

In order to investigate the spectral characteristics of the total error associated with this sort of dither, we must derive an expression for $p_{\nu_1,\nu_2}(\nu_1, \nu_2)$ as defined in Section II-C. Suppose

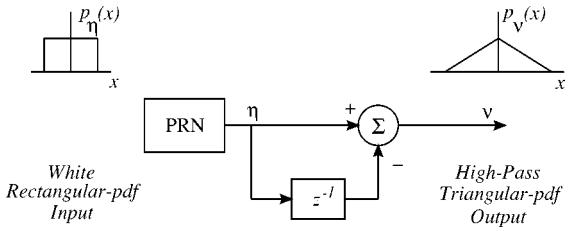

Fig. 9. Highpass dither generator.

that the sampling period of the system is T. For time lags $|\tau| > T$, the dither values are statistically independent so that

$$\begin{aligned} p_{\nu_1,\nu_2}(\nu_1,\nu_2) &= p_{\nu_1}(\nu_1) \\ p_{\nu_2}(\nu_2) &= [\Pi_\Delta \star \Pi_\Delta](\nu_1)[\Pi_\Delta \star \Pi_\Delta](\nu_2) \end{aligned}$$

and

$$P_{\nu_1,\nu_2}(u_1,u_2) = \text{sinc}^2(u_1)\,\text{sinc}^2(u_2).$$

The nontrivial cases are those for $\tau = \pm T$. Consider two successive dither values (i.e., $\tau = T$) $\nu_1 = \eta_1 - \eta_0$ and $\nu_2 = \eta_2 - \eta_1$. Then, we have the expression at the bottom of the next page. Taking the Fourier transform of this expression with respect to all variables present yields

$$\begin{aligned} &P_{\nu_1,\nu_2,\eta_0,\eta_1,\eta_2}(u_1,u_2,w_0,w_1,w_2) \\ &\quad = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi u_1(\eta_1-\eta_0)} e^{-j2\pi u_2(\eta_2-\eta_1)} \\ &\qquad \times p_{\eta_0}(\eta_0)p_{\eta_1}(\eta_1)p_{\eta_2}(\eta_2) \\ &\qquad \times e^{-j2\pi(w_0\eta_0+w_1\eta_1+w_2\eta_2)}\,d\eta_0\,d\eta_1\,d\eta_2 \\ &\quad = P_{\eta_0}(w_0-u_1)P_{\eta_1}(w_1+u_1-u_2)P_{\eta_2}(w_2+u_2). \end{aligned}$$

Desired marginal cf's can be obtained from a given joint cf by simply setting unwanted variables to zero since

$$P_{x,y}(u,0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-j2\pi(xu+y\times(0))} p_{x,y}(x,y)\,dx\,dy = P_x(u).$$

Thus, we have $P_{\nu_1,\nu_2}(u_1,u_2) = P_{\eta_0}(-u_1)P_{\eta_1}(u_1 - u_2)P_{\eta_2}(u_2)$.

Proceeding similarly for the case of $\tau = -T$, we find that $P_{\nu_1,\nu_2}(u_1,u_2) = P_{\eta_0}(-u_2)P_{\eta_1}(u_2-u_1)P_{\eta_2}(u_1)$. For our purposes, $P_{\eta_0}(u) = P_{\eta_1}(u) = P_{\eta_2}(u) = \text{sinc}(u)$ so that for both cases ($\tau = \pm T$), we have

$$P_{\nu_1,\nu_2}(u_1,u_2) = \text{sinc}(u_2-u_1)\,\text{sinc}(u_1)\,\text{sinc}(u_2).$$

Finally, using $\tau = kT$, $k \in \mathbf{Z}$, we can write that

$$\begin{aligned} &P_{\nu_1,\nu_2}(u_1,u_2;k) \\ &\quad = \begin{cases} \text{sinc}(u_2-u_1)\,\text{sinc}(u_1)\,\text{sinc}(u_2), & k = \pm 1 \\ \text{sinc}^2(u_1)\,\text{sinc}^2(u_2), & |k| > 1. \end{cases} \end{aligned} \tag{43}$$

It is straightforward to check that this joint cf satisfies all three conditions of Theorem 4.

Using (43) and the knowledge that TPDF dither has a variance of $\Delta^2/6$, we find using (19) that the autocorrelation function of the dither under consideration is

$$E[\nu_1\nu_2](k) = \frac{\Delta^2}{6} \times \begin{cases} 1, & k = 0 \\ -\frac{1}{2}, & k = \pm 1 \\ 0 & \text{otherwise.} \end{cases}$$

This corresponds to a simple highpass power spectral density

$$\text{PSD}_\nu(f) = \frac{\Delta^2 T}{3}[1 - \cos(2\pi f T)]$$

where the frequency variable f is in units of hertz if T is in seconds. (Of course, this is just as expected from inspection of Fig. 9, which is nothing but a linear discrete-time filter with transfer function $1 - z^{-1}$.) Then, according to (31)

$$E[\varepsilon_1\varepsilon_2](k) = \frac{\Delta^2}{4} \times \begin{cases} 1, & k = 0 \\ -\frac{1}{3}, & k = \pm 1 \\ 0, & \text{otherwise} \end{cases}$$

which corresponds to a power spectral density of

$$\text{PSD}_\varepsilon(f) = \frac{\Delta^2 T}{6}[3 - 2\cos(2\pi f T)].$$

This is simply the highpass spectrum of the dither plus a white "quantization noise" component of $\Delta^2 T/6$ (which has a total power of $\Delta^2/12$ up to the Nyquist frequency $1/2T$) in agreement with (32).

Of course, it is possible to imagine many other spectrally shaped dither signals. The properties of such signals have now been investigated in detail. In particular, there is the following theorem, which is proven and extensively illustrated in [26] and [28]:

Theorem 7: In an NSD quantizing system using dither of the form

$$\nu_n = \sum_{i=-\infty}^{\infty} c_i \eta_{n-i}$$

where η is an iid nRPDF random process, the total error will be wide-sense stationary and independent of the system input with a PSD given by

$$\text{PSD}_\varepsilon(f) = \text{PSD}_\nu(f) + \frac{\Delta^2 T}{6}$$

under the following conditions.

1) For each $\ell \in \mathbf{Z}_0$, there exists an i such that of c_i and $c_{i+\ell}$, one is zero, and the other is a nonzero integer, and
2) either η is nRPDF with $n \geq 1$ and there exist at least two distinct values of i such that c_i is a nonzero integer or η is nRPDF with $n \geq 2$, and there exists at least one value of i such that c_i is a nonzero integer.

In particular, simple highpass TPDF dither satisfies the above conditions.

IV. Digital Dither

Some comment is required concerning the special nature of *re*quantization operations, in which the binary wordlength of data is reduced prior to its storage or transmission. This operation takes place entirely within the digital domain so that both the input and dither signals are discrete valued due to the finite wordlengths available in practical digital systems. The continuous pdf's discussed thus far are unattainable in a purely digital scheme so that the properties of true digital dither signals require further investigation.

The following discussion represents a theoretical complement to empirical results presented in [17]. It is not intended to be exhaustive but merely to demonstrate that there is no great difficulty in extending the results obtained for analog systems to digital ones and to illustrate how this may be done. In particular, the discussion will be restricted to a treatment of first-order statistics with the extension to second-order being straightforward.

Consider a quantizing system that applies digital dither to digital data before removing its L least significant bits. We will use δ to denote the magnitude of an LSB of the higher precision signal to be requantized and $\Delta = 2^L\delta$ for an LSB of the requantized output.

Let us consider the following digital dither pdf:

$$p_\nu(\nu) = \delta\tilde{p}_\nu(\nu)W_\delta(\nu) \tag{44}$$

where $\tilde{p}_\nu(\nu)$ represents an absolutely integrable function that serves as a "weighting" for the impulse train. $\tilde{p}_\nu$ is assumed to be normalized such that

$$\int_{-\infty}^{\infty} p_\nu(\nu)\,d\nu = \delta \sum_{\ell=-\infty}^{\infty} \tilde{p}_\nu(\ell\delta) = 1.$$

For instance, $\tilde{p}_\nu$ might be the pdf of a dither of order n, such as an nRPDF dither, in which case, it is straightforward to show

$$\begin{aligned} &p_{\nu_1,\nu_2,\eta_0,\eta_1,\eta_2}(\nu_1,\nu_2,\eta_0,\eta_1,\eta_2) \\ &\quad = p_{\nu_1|(\nu_2,\eta_0,\eta_1,\eta_2)}(\nu_1,\nu_2,\eta_0,\eta_1,\eta_2)p_{\nu_2|\eta_0,\eta_1,\eta_2}(\nu_2,\eta_0,\eta_1,\eta_2)p_{\eta_0,\eta_1,\eta_2}(\eta_0,\eta_1,\eta_2) \\ &\quad = \delta(\nu_1 - \eta_1 + \eta_0)\delta(\nu_2 - \eta_2 + \eta_1)p_{\eta_0}(\eta_0)p_{\eta_1}(\eta_1)p_{\eta_2}(\eta_2). \end{aligned}$$

using Poisson's summation formula [40] that $\tilde{p}_\nu$ has the above normalization. In general, however, $\tilde{p}_\nu$ is merely a mathematical device used to specify arbitrary weights for the delta functions in (44) and need not even correspond to a pdf since it need not subtend unit area.

Taking the Fourier transform of (44), we find that

$$P_\nu(u) = [\tilde{P}_\nu \star W_{\frac{1}{\delta}}](u) = \sum_{\ell=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{\ell}{\delta}\right) \tag{45}$$

where $\tilde{P}_\nu(u)$ is the Fourier transform of $\tilde{P}_\nu(\nu)$. Note that even if $\tilde{p}_\nu$ satisfies the conditions of Corollary 2 (for some m), P_ν will not, due to the modulation of $\tilde{P}_\nu(u)$ by the impulse train $W_{\frac{1}{\delta}}(u)$. Fortunately, we do not require that these conditions be satisfied in a digital system since the requirement that $E[\varepsilon^m \mid x]$ be constant for *all* values of the system input is not of interest. Instead, we require only that the moments be constant for a subset of all conceivable x values, namely, $\{x \mid x = n\delta, n \in \mathbf{Z}\}$, which includes all values that are representable in the digital system. Thus, we assume that the pdf of the system input can be expressed in the form

$$p_x(x) = \delta \tilde{p}_x(x) W_\delta(x) \tag{46}$$

where $\tilde{p}_x$ is an absolutely integrable function normalized such that the integral of (46) is unity. Then

$$P_x(u) = [\tilde{P}_x \star W_{\frac{1}{\delta}}](u) \tag{47}$$

$$= \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(u - \frac{\ell}{\delta}\right). \tag{48}$$

Now, from (11), we have

$$G_\nu(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u} P_\nu(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u} \sum_{k=-\infty}^{\infty} \tilde{P}_\nu\left(u - \frac{k}{\delta}\right). \tag{49}$$

Then, from (8)

$$P_\varepsilon(u) = \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} G_\nu\left(u - \frac{k}{\Delta}\right) \tilde{P}_x\left(-\frac{k + 2^L\ell}{\Delta}\right)$$

so that

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m P_\varepsilon^{(m)}(0) = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} \sum_{\ell=-\infty}^{\infty} G_\nu^{(m)}\left(-\frac{k}{\Delta}\right) \tilde{P}_x\left(-\frac{k+2^L\ell}{\Delta}\right). \tag{50}$$

The only way that this quantity can be independent of $\tilde{P}_x$ is if we require that

$$G_\nu^{(m)}\left(\frac{k}{\Delta}\right) = 0 \tag{51}$$

for all $k \in \mathbf{Z}$, except possibly for those values of k such that $(k/2^L) \in \mathbf{Z}$.

That is, the indicated derivative must vanish for all integral values of k, except those that are integral multiples of 2^L, the value of this derivative being immaterial in the latter cases. In order to see that this is so, note that if a dither is chosen such that (51) holds, then many terms vanish from (50), leaving

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\delta}\right) \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(\frac{\ell}{\delta}\right).$$

Now, from (48), we know that

$$P_x(0) = \sum_{\ell=-\infty}^{\infty} \tilde{P}_x\left(\frac{\ell}{\delta}\right) = 1.$$

This leaves

$$E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \sum_{k=-\infty}^{\infty} G_\nu^{(m)}\left(\frac{k}{\delta}\right) \tag{52}$$

which does not depend on the input distribution. The necessity of (51) follows from the arbitrariness of $\tilde{P}_x$ (apart from its normalization). Furthermore, by inspection, (52) is precisely the mth moment of a notional random variable with pdf

$$\left[\frac{\Delta}{2^L}\Pi_\Delta \star p_\nu\right](\varepsilon) W_\delta(\varepsilon)$$

although this is not, of course, the pdf of ε. Some algebraic manipulation of this expression, exploiting the discrete-valued character of ν, reveals that it is equivalent to

$$\left[\frac{\Delta}{2^L}\Pi_\Delta \cdot W_\delta\right](\varepsilon) \star p_\nu(\varepsilon). \tag{53}$$

This may be regarded as the pdf of a notional random variable that is the sum of the dither and an independent discrete-valued "quantization noise."

In addition, note that in the limit as $\delta \to 0$ (i.e., as $L \to \infty$), (51) becomes (12), which is the condition of Theorem 2 for analog systems.

Returning to (49) and differentiating, we have

$$\frac{d^m G_\nu}{du^m}(u) = \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r} \times \frac{d^r}{du^r}\left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right] \tilde{P}_\nu^{(m-r)}\left(u - \frac{k}{\delta}\right). \tag{54}$$

If $\tilde{P}_\nu$ meets the conditions of Corollary 2, then all terms in (54) involving the derivatives of $\tilde{P}_\nu$ go to zero at the places required by (51), except for the single $(r = 0)$ term involving the mth derivative. Fortunately, this term involves the zeroth derivative of the leading sinc function, which goes to zero at all the required places. This yields the following theorem.

Theorem 8: For a digital NSD system in which requantization is used to remove the L least significant bits of binary data, $E[\varepsilon^\ell]$ is independent of the input distribution for $\ell =$

$1, 2, \ldots, m$ if a nonsubtractive digital dither (with the same precision as the input data) is applied for which

$$\tilde{P}_\nu^{(i)}\left(\frac{k}{\Delta}\right) = 0 \quad \forall k \in \mathbf{Z}_0 \quad \text{and} \quad i = 0, 1, 2, \ldots, m-1.$$

This theorem is the digital counterpart of Corollary 2. It is interesting to note that no such counterpart exists for Theorem 2 in terms of $\tilde{P}_\nu$.

We observe that using a dither of higher precision than the input signal is of no benefit. For instance, a dither cf that satisfies the conditions of (51) with $m = 1$ for $L = 8$ will also satisfy them for $L = 4$, but for a quantizing system in which the precision is reduced by only four bits, there is no advantage associated with this cf over one that only satisfies the conditions for $L = 4$.

Frequently, dithers in digital systems will be given a 2's-complement [33] representation and, thus, will exhibit a mean that differs slightly from zero. This will be reflected in the appearance of a small nonzero mean error that, of course, will be input independent if an appropriate dither pdf has been chosen.

To express the moments of the system output, we impose the conditions of Theorem 8 on (34), obtaining

$$\begin{aligned} E[y^m] &= \sum_{r=0}^{m} \binom{m}{r} \sum_{k=-\infty}^{\infty} \left[\left(\frac{j}{2\pi}\right)^r G_\nu^{(r)}\left(\frac{k}{\delta}\right)\right] \\ &\quad \times \left[\left(\frac{j}{2\pi}\right)^{m-r} P_x^{(m-r)}\left(\frac{k}{\delta}\right)\right] \\ &= \sum_{r=0}^{m} \binom{m}{r} E[\varepsilon^r] E[x^{m-r}] \end{aligned}$$

where we have observed from (48) that $P_x(u)$ is periodic with period $1/\delta$ so that for any $k \in \mathbf{Z}$

$$\left(\frac{j}{2\pi}\right)^{m-r} P_x^{(m-r)}\left(\frac{k}{\delta}\right) = \left(\frac{j}{2\pi}\right)^{m-r} P_x^{(m-r)}(0) = E[x^{m-r}].$$

$E[\varepsilon^r]$ is given by (52).

The treatment presented above is most appropriate to dithers generated entirely in the digital domain using, for instance, pseudo-random number generation algorithms. In particular, we have shown that whenever the weighting function $\tilde{p}_\nu$ corresponds to the pdf of an analog nRPDF dither, the associated digital dither with pdf given by (44) shares the beneficial properties of its analog counterpart.

In the case where a digital dither signal is generated by fine quantization of an analog dither signal, the details of the derivation change only slightly. The forms of the theorems, however, remain the same, with $\tilde{P}_\nu$ representing the cf of the analog signal. This can be seen directly using (33) (with null dither) because the pdf of the digital dither generated by quantization will be

$$p_\nu(\nu) = [\delta \Pi_\delta \star \tilde{P} p_\nu](\nu) W_\delta(\nu)$$

with

$$P_\nu(u) = \left[\frac{\sin(\pi\delta u)}{\pi\delta u} \tilde{P}_\nu(u)\right] \star W_{\frac{1}{\delta}}(u).$$

This expression should be compared with (45). Note that if $\tilde{P}_\nu$ satisfies the conditions of the theorems, then so will the quantity

$$\frac{\sin(\pi\delta u)}{\pi\delta u} \tilde{P}_\nu(u).$$

Thus far, the behavior of the quantizer at step edges (i.e., when $w = (2k-1)\Delta/2$, $k \in \mathbf{Z}$) has not been explicitly considered. This is not a problem if the signals in question are continuous valued. In this case, the addition of dither will ensure that the quantizer input resides at a quantizer-step edge with zero probability. On the other hand, if digital signals are in use, the probability that the quantizer input resides at a step edge is always greater than zero. In this instance, it makes a considerable difference to the quantizer output (and total error) whether the quantizer rounds up, down, or stochastically (up or down with equal probability) at these edges. Technically, it can be shown [26] that the above formalism yields correct predictions if a stochastic quantizer is used.

The extension of the results to deterministic (i.e., nonstochastic) quantizers employs a simple trick. Consider, for instance, the consequences of choosing a quantizer that always rounds up at step edges (a similar argument applies to quantizers that round down). We note that if a (dc) *virtual offset* μ such that $0 < \mu < \delta$ is introduced into the dither signal, the quantizer output is unaffected, except that quantizer inputs residing at step edges are consistently rounded up. We can thus analyze digitally dithered systems with deterministic requantizers using such a notional dc offset, which is a purely mathematical device without physical counterpart. It can be shown [26] that Theorem 8 holds precisely as before. Equation (52) holds if the virtually offset dither pdf $p_\nu(\nu) = \delta \tilde{p}_\nu(\nu - \mu) W_\delta(\nu - \mu)$ is used in the calculations. In this case, (53) becomes

$$\left[\frac{\Delta}{2^L} \Pi_\Delta(\varepsilon - \mu) W_\delta(\varepsilon)\right] \star p_\nu(\varepsilon + \mu).$$

V. Conclusions

The following conclusions bear repeating.

1) Nonsubtractive dithering, unlike subtractive dithering, cannot render the total error statistically independent of the system input, but it *can* render any desired moments of the total error independent of the input distribution provided that certain conditions on the cf of the dither are met (see Theorems 1 and 2). In particular, a nRPDF dither will render the first n moments of the total error input independent.
2) Nonsubtractive dithering, unlike subtractive dithering, cannot render total error values separated in time statistically independent of one another. It can, however,

regulate the joint moments of such errors, and in particular, it can render the power spectrum of the total error signal equal to the power spectrum of the dither signal plus a white "quantization noise" component [see Theorem 4 and (32)].

3) Nonsubtractive dithering can render any desired moments of the system input recoverable from those of the system output, provided that the statistical attributes of the dither are properly chosen (see Section II-D). This includes joint moments of system inputs separated in time; therefore, the spectrum of the input can be recovered from the spectrum of the output.
4) Proper nonsubtractive dithering always results in a total error variance greater than $\Delta^2/12$ [see (14) and Theorem 6].

It is also worth noting that since the dither is simply an additive signal that is independent of the system input, we are free to add it at *any* time prior to (re)quantization. In particular, once a signal is properly dithered, other signals (which are statistically independent of the dither) may be added to it, and the resulting total signal will still be properly dithered for (re)quantization purposes.

For audio signal processing purposes, there seems to be little point in rendering any error moments other than the first and second independent of the input. Variations in higher moments are believed to be inaudible, and this has been corroborated by a large number of psycho-acoustic tests conducted by the authors and others [14], [22]. These tests involved listening to a large variety of signals (sinusoids, sinusoidal chirps, slow ramps, various periodically switched inputs, piano and orchestral music, etc.), which had been very coarsely requantized (from 16 to 8 bits) in order to render the requantization error essentially independent of low-level nonlinearities in the digital-to-analog conversion system used for listening purposes. In addition, the corresponding error signals (output minus input) were used in listening tests in order to check for any vestiges of audible dependence on the input. Using undithered quantizers resulted in clearly audible distortion and noise modulation in the output and error signals. Rectangular-pdf dither of one LSB peak-to-peak amplitude eliminated all distortion, but the residual noise level was found to vary audibly in an input-dependent fashion. When triangular-pdf dither of two LSB peak-to-peak amplitude (either white or highpass) was employed, no instance was found in which the error was audibly distinguishable from a steady random noise entirely unrelated to the input. Admittedly, these tests were informal, and there remains a need for formal psychoacoustic tests of this sort involving many participants under controlled conditions.

We recommend the use of spectrally white triangular-pdf (TPDF) dither of two LSB peak-to-peak amplitude for most audio applications requiring nonsubtractively dithered multibit quantization or requantization operations since this type of dither renders the first and second moments of the total error signal constant with respect to the system input while incurring the minimum increase in error variance. This kind of dither is easy to generate for digital requantization by simply summing two independent rectangular-pdf (RPDF) pseudo-random processes, each of one LSB peak-to-peak amplitude, which may easily be generated using a linear congruential algorithm [22], [41]. The resulting digital dither can be used to feed a digital-to-analog converter for analog dithering applications. It should be noted, however, that many analog signals and digital conversion systems exhibit a Gaussian noise component that is of large enough amplitude to act as a satisfactory dither without the requirement of an explicit dithering operation [7], [18].

Highpass TPDF dither is of interest for audio processing since it yields a total error that is *audibly* quieter than that associated with iid TPDF dither. It also can be generated with greater computational efficiency since only one new pseudo-random number needs to be calculated per sampling period instead of two. Other spectrally shaped dithers can also be used [25], [26]. The use of spectrally shaped dither will often be superseded, however, by the powerful technique of *noise shaping* in applications where the total audibility of the error signal needs to be reduced [42]–[44]. This technique employs error feedback in order to spectrally shape the total error of a quantizing system, including the white component arising from a properly dithered quantization. The necessity of and criteria for proper dithering of such systems have now been explored in some detail [25], [28].

With regard to video signals, at least the first two moments of the total error should be rendered independent of the input by the use of an appropriate dither. Some evidence exists that input-dependent variations in the third moment of the total error can be perceptually significant in some video signals [14], but the effects of such variations are probably not noticeable in most cases.

For signal measurement and statistical signal analysis applications in which signal moments are being measured, appropriate dither (possibly of a quite high order) must be used to render the input signal statistics correctly determinable from the statistics of the quantized output, in accordance with (35) and (37).

Some of the results obtained above for multi-bit systems may be applied to 1-bit quantizers with certain caveats. For instance, a 1-bit quantizing system using a full-scale RPDF dither signal will exhibit a total error signal that is zero-mean and spectrally white whenever its peak input value always remains less than the peak dither amplitude. In this instance, the 1-bit quantizer may be regarded as a RPDF-dithered multi-bit mid-riser quantizer whose peak input amplitude is restricted to less than $\Delta/2$ so that the analysis presented above applies without modification. Granted, the amount of dither required is large, but this may be acceptable when no distortion is tolerable and the oversampling ratio is high. A sigma-delta converter topology [45] may be employed to move the noise power out-of-band as long as the dither is iid (in order to avoid statistical dependences between the dither and input, which would otherwise be introduced by the error feedback [25], [28]). We observe that 2RPDF dither cannot be used in such 1-bit systems to eliminate noise modulation since no nonzero input could then be accommodated without quantizer saturation. The potential benefits of using lower dither amplitudes in sigma-delta conversion systems have been explored in [45], and a detailed statistical analysis of one dithered implementation is presented in [46].

We maintain that the use of appropriate dither prior to (re)quantization is as fundamental as the use of an appropriate anti-aliasing filter prior to sampling; both serve to eliminate classes of signal-dependent errors. In each case, this is accomplished by protecting the system from inputs which, if left unmodified, may introduce such errors.

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. Tech. J.*, vol. 51, pp. 1293–1304, July–Aug. 1972.

[3] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun. Technol.*, vol. COMM-12, pp. 162–165, Dec. 1964.

[4] D. T. Sherwood, "Some theorems on quantization and an example using dither," presented at the 19th Asilomar Conf. Circuits, Syst., Comput., Pacific Grove, CA, Nov. 1985.

[5] B. Widrow, "A study of rough amplitude quantization by means of Nyquist sampling theory," Sc.D. thesis, Dept. Elec. Eng., Mass. Inst. Technol., Cambridge, June 1956.

[6] ——, "A study of rough amplitude quantization by means of Nyquist sampling theory," *IRE Trans. Circuit Theory*, vol. CT-3, pp. 266–276, Dec. 1956.

[7] ——, "Statistical analysis of amplitude-quantized sampled-data systems," *Trans. Amer. Inst. Elec. Eng.*, pt. II, vol. 79, pp. 555–568, Jan. 1961.

[8] B. Widrow, I. Kollár, and M.-C. Liu, "Statistical theory of quantization," *IEEE Trans. Instrum. Meas.*, vol. 45, pp. 389–396, June 1996.

[9] 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.

[10] R. M. Gray, "Quantization noise spectra," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1220–1244, Nov. 1990.

[11] J. N. Wright, unpublished manuscripts, June–Aug. 1979.

[12] ——, private communication, Apr. 1991.

[13] T. G. Stockham, private communication, 1988.

[14] L. K. Brinton, "Nonsubtractive dither," M.Sc. thesis, Dept. Elect. Eng., Univ. Utah, Salt Lake City, UT, Aug. 1984.

[15] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Inform. Theory*, vol. 39, pp. 805–811, May 1993.

[16] 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, Mar. 1984. Correction ibid, p. 889, Nov. 1984.

[17] S. P. Lipshitz and J. Vanderkooy, "Digital dither," in *Proc. 81st Conv. Audio Eng. Soc.*, Los Angeles, CA, Nov. 1986. Preprint 2412; *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 1030, Dec. 1986.

[18] J. Vanderkooy and S. P. Lipshitz, "Dither in digital audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975, Dec. 1987.

[19] ——, "Digital dither: Signal processing with resolution far below the least significant bit," in *Proc. Audio Eng. Soc. 7th Int. Conf.: Audio in Digital Times*, Toronto, Ont., Canada, May 1989, pp. 87–96.

[20] S. P. Lipshitz and J. Vanderkooy, "High-pass dither," in *Proc. 4th Reg. Conv. Audio Eng. Soc.*, Tokyo, Japan, June 1989, pp. 72–75. *Collected Preprints* (Audio Eng. Soc., Japan Section, Tokyo, 1989).

[21] R. A. Wannamaker, S. P. Lipshitz, and J. Vanderkooy, "Dithering to eliminate quantization distortion," in *Proc. Annu. Meeting Can. Acoust. Assoc.*, Halifax, N.S., Canada, Oct. 1989, pp. 78–86.

[22] R. A. Wannamaker, "Dither and noise shaping in audio applications," M.Sc. Thesis, Dept. Phys., Univ. Waterloo, Waterloo, Ont., Canada, Dec. 1990.

[23] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, and J. N. Wright, "Non-subtractive dither," presented at the 1991 IEEE Workshop on Appl. Signal Process. Audio Acoust., New Paltz, NY, Oct. 1991. Paper no. 6.2.

[24] R. A. Wannamaker and S. P. Lipshitz, "Time domain behavior of dithered quantizers," presented at the 93rd Conv. Audio Eng. Soc., San Francisco, CA, Oct. 1992. preprint 3418.

[25] R. A. Wannamaker, "Subtractive and nonsubtractive dithering: A comparative analysis," presented at the 97th Conv. Audio Eng. Soc., San Francisco, CA, Nov. 1994. preprint 3920.

[26] R. A. Wannamaker, "The theory of dithered quantization," Ph.D. dissertation, Dept. Appl. Math., Univ. Waterloo, Waterloo, Ont., Canada, June 1997.

[27] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, May 1992.

[28] ——, "Dithered noise shapers and recursive digital filters," presented at the 94th Conv. Audio Eng. Soc., Berlin, Germany, Mar. 1993. preprint 3515.

[29] R. A. Wannamaker and S. P. Lipshitz, "Dithered quantizers with and without feedback," presented at the 1993 IEEE Workshop Appl. Signal Process. Audio Acoust., New Paltz, NY, Oct. 1993.

[30] R. A. Wannamaker, "Efficient generation of multichannel dither signals," presented at the 103rd Conv. Audio Eng. Soc., New York, NY, Sept. 1997. preprint 4533.

[31] A. Kirac and P. P. Vaidyanathan, "Results on lattice vector quantization with dithering," *IEEE Trans. Circuits Syst. II*, vol. 43, pp. 811–826, Dec. 1996.

[32] R. M. Gray, private communication, Apr. 1991.

[33] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[34] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 33, pp. 446–472, July 1948.

[35] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2325–2383, Oct. 1998.

[36] E. Lukacs, *Characteristic Functions*. London, U.K.: Griffin, 1960.

[37] T. Kawata, *Fourier Analysis in Probability Theory*. New York, NY: Academic, 1972.

[38] P. Carbone and D. Petri, "Effect of additive dither on the resolution of ideal quantizers," *IEEE Trans. Instrum. Meas.*, vol. 43, pp. 389–396, June 1994.

[39] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd ed. New York: McGraw-Hill, 1984.

[40] A. Papoulis, *The Fourier Integral and Its Applications*. New York: McGraw-Hill, 1962.

[41] D. Knuth, *The Art of Computer Programming*, 2nd ed. Reading, MA: Addison-Wesley, 1981, vol. 2.

[42] H. A. Spang and P. M. Schultheiss, "Reduction of quantizing noise by use of feedback," *IRE Trans. Commun. Syst.*, vol. COMM-10, pp. 373–380, Dec. 1962.

[43] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally audible noise shaping," *J. Audio Eng. Soc.*, vol. 39, pp. 836–852, Nov. 1991.

[44] R. A. Wannamaker, "Psychoacoustically optimal noise shaping," *J. Audio Eng. Soc.*, vol. 40, pp. 611–620, July–Aug. 1992.

[45] S. R. Norsworthy, R. Schreier, and G. C. Temes, Eds., *Delta-Sigma Data Converters: Theory, Design and Simulation*. Piscataway, NJ: IEEE, 1997.

[46] W. Chou and R. M. Gray, "Dithering and its effects on sigma-delta and multi-stage sigma-delta modulation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 500–513, May 1991.

Robert A. Wannamaker received the B.Sc.E. degree in engineering physics from Queen's University, Kingston, Ont., Canada, in 1988 and the M.Sc. degree in physics in 1991 from the University of Waterloo, Waterloo, Ont., Canada. He received the Ph.D. degree in applied mathematics in 1997 from the University of Waterloo for research conducted into the mathematical modeling of dithered quantizing systems. He is also a composer of experimental acoustic and electroacoustic music and holds the B.F.A. degree in music from York University Toronto, Ont., Canada.

He continues to conduct research as a member of the Audio Research Group, University of Waterloo. His interests include quantization, dither, noise-shaping converters, stochastic resonance, fractals, wavelet analysis, data compression, and psychoacoustics. His publications have appeared in *Physical Review*, the *Journal of the Audio Engineering Society*, and elsewhere.

Stanley P. Lipshitz (M'83) received the B.Sc.Hons. degree in applied mathematics from the University of Natal, Durban, South Africa, in 1964, the M.Sc. degree in applied mathematics from the University of South Africa, Pretoria, in 1966, and the Ph.D. degree in mathematics from the University of the Witwatersrand, Johannesburg, South Africa.

He is a Professor with the Departments of Applied Mathematics and Physics, University of Waterloo, Waterloo, Ont., Canada, which he joined in 1970. He is one of the founding members of the Audio Research Group at the University of Waterloo, which conducts research in many areas of audio and electroacoustics. His current research interests include the mathematical theory of dithered quantizers and noise shapers (and their relation to stochastic resonance and chaos), physical acoustics, and active noise control. He has presented numerous technical papers, on a wide range of topics, at conferences both in North America and overseas.

Dr. Lipshitz is a Fellow of the Audio Engineering Society, a recipient of its Silver Medal for his research contributions to digital audio, and a recipient of its Publication Award (jointly with J. Vanderkooy and R. Wannamaker) for a survey paper on quantization and dither in the *Journal of the Audio Engineering Society*. He has served as a Governor of that society and was its President from 1988 to 1989. Other society memberships include, the Acoustical Society of America and the Canadian Acoustical Association.

John Vanderkooy received the B.Eng. degree in engineering physics in 1963 from McMaster University, Hamilton, Ont., Canada. He continued studies there in low temperature physics of metals, receiving the Ph.D. degree in 1967.

He spent two years as a postdoctoral fellow at the University of Cambridge, Cambridge, U.K., and joined the faculty at the University of Waterloo, Waterloo, Ont., Canada, in 1969. For some years, he followed his doctoral interests in magnetic properties of electrons in metals, but his research interests have slowly shifted since the late 1970's to audio and electroacoustics. He is presently a Professor of physics at the University of Waterloo. He has contributed a wide variety of technical papers in such areas as loudspeaker crossover design, electroacoustic measurement technique, dithered quantizers, and acoustics. He is also a founding member of the Audio Research Group at the University of Waterloo. His current research interests are dithered quantization, transducers, acoustic diffraction from edges, and loudspeaker ports.

Dr. Vanderkooy is a Fellow of the Audio Engineering Society, a recipient of its Silver Medal, and of several publication awards, one of which concerns the measurement of transfer functions with maximum-length sequences.

J. Nelson Wright (SM'97) received the B.S.E.E. degree in 1976 and the M.S.E.E. degree in 1978 from the Massachusetts Institute of Technology (MIT), Cambridge.

He was subsequently employed by the MIT Lincoln Laboratory, Lexington, as a Member of Technical Staff, where he performed research in geosynchronous communication satellite technologies. In 1981, he joined Acuson, where he was project manager for the scanner subsystem of the company's first product (the Acuson 128) and a co-inventor of its patented summing delay line architecture. He was later responsible for Acuson's Sequoia project, which was chartered to develop a new generation imaging technology for diagnostic ultrasound. In this capacity, he was responsible for conceptualizing a new imaging architecture, leading the research team to analyze and simulate it, and managing the development of the first two Sequoia products. During his tenure at Acuson, he served in various positions including over five years as Vice President of Advanced Development. The Acuson 128 was introduced in 1983. The basic architecture is still in production and has generated in excess of $1 billion in revenue. The Sequoia products were introduced in 1996 and have received unprecedented acceptance and recognition, including acquisition by the Smithsonian as part of their permanent collection of important medical technology. He founded the Parallax Group and now acts as a consultant to industry and the venture community. He has 19 patents in the field of ultrasonic imaging.

Mr. Wright is a member of SPIE, ASA, and Sigma Xi.