Quantization and Dither: A Theoretical Survey*

STANLEY P. LIPSHITZ, *AES Fellow,* **ROBERT A. WANNAMAKER, AND JOHN VANDERKOOY,** *AES Fellow***

Audio Research Group, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada

A theoretical survey of multibit quantization is presented, beginning with the classical model of undithered quantization and proceeding to modern statistical models of undithered, subtractively dithered, and nonsubtractively dithered quantizing systems. Properties of the error and output signals for each type of system are examined in relation to the input, using both first- and second-order statistical approaches. The virtues of various dither signals are assessed for use in practical applications. It is hoped that this survey can help to clarify the differences between different types of dithered quantization schemes and to correct some of the common misunderstandings regarding the effects of dithered quantization.

0 INTRODUCTION

Dither and quantization are among the most frequently discussed topics in audio signal processing. Dithering techniques are fast becoming commonplace in applications where quantization or requantization is required in order to reduce the wordlength of audio data. The literature contains appropriate recommendations regarding dither signals which are suitable for audio applications [1], [2]. In spite of the widespread interest in dither and quantization, a comprehensive theory of their operation appears not to exist in print. The relevant theorems are scattered among sundry journals and conference proceedings, and many have not been published at all until very recently [3], [4]. This paper attempts to collect all of the significant theory and to supply references to its originators as best we can ascertain them. The treatment begins with fundamental concepts and proceeds to discuss models of undithered, subtractively dithered, and nonsubtractively dithered quantizing systems in a consistent manner which makes clear both their differences and similarities.

1 QUANTIZERS AND QUANTIZING SYSTEMS

Analog-to-digital conversion is customarily decomposed into two separate processes: time sampling of the input analog waveform and amplitude quantization of the signal values in order that the samples may be represented by binary words of a prescribed length (the order of these two processes being immaterial). The sampling operation incurs no loss of information as long as the input is band-limited [5] in accordance with the sampling theorem, but the approximating nature of the quantization operation generally results in signal degradation. An operation with a similar problem is *requantization*, in which the wordlength of digital data is reduced after processing in order to meet specifications for its storage or transmission. An optimal (re)quantizer is one that minimizes the deleterious effects of the aforementioned signal degradation by converting the audible signal-dependent artifacts into benign signal-independent ones as far as possible.

Quantization and requantization possess similar transfer characteristics, which are generally of either the *midtread* or the *midriser* variety illustrated in Fig. 1. Assuming that these represent *infinite* quantizers,[1] the corresponding transfer functions can be expressed analytically in terms of the input to the quantizer, w, and the quantizer step size Δ as

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

for a midtread quantizer, or

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

* Presented at the 91st Convention of the Audio Engineering Society, New York, 1991 October 4–8; revised 1991 October 21.

** The authors are members of the Guelph-Waterloo Program for Graduate Work in Physics.

[1] For practical purposes, this simply means that the signal is never clipped by saturation of the quantizer.

for a midriser 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 midtread variety will be assumed, but all derived results have obvious analogs for midriser quantizers, and all results stated as 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, $Q(w)$, and its input w

$$q(w) \triangleq Q(w) - w\ , \tag{3}$$

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

We shall refer to systems that restrict the accuracy of sample values using multibit quantization as *quantizing systems*, of which there exist three archetypes: undithered, subtractively dithered, and nonsubtractively dithered. Schematics of these systems are shown in Fig. 3.

Throughout the sequel we will refer to the *system input* as x, the *system output* as y, and the *total error* of the system as ε, where

$$\varepsilon \triangleq y - x \tag{4}$$

as distinguished from the quantization error q defined by Eq. (3). In an undithered quantizing system, the system input x is identical to the quantizer input w so that the total error equals the quantization error [$\varepsilon = q(x)$]. In the other two schemes, the quantizer input is comprised of the system input plus an additive random signal v, called *dither*, which is assumed to be stationary[2] and statistically independent of x. In such systems the quantizer input $w = x + v$ is not a deterministic function of x, and neither is the total error ε. In the subtractively dithered topology the dither signal is subtracted from the quantizer output, presumably after this output has been transmitted through some channel. This subtraction operation is omitted in a nonsubtractively dithered system.

The object of dithering is to control the statistical properties of the total error and its relationship to the system input. In undithered systems we know that the error is a deterministic function of the input. If the input is simple or comparable in magnitude to the quantization step size, the total error signal is strongly input-dependent and audible as gross distortion and noise modulation. We shall see that use of dither with proper statistical properties can render the total error signal audibly equivalent to a steady white noise.

The body of this paper proceeds to examine the three quantizing systems from a theoretical viewpoint. First, the classical model of undithered quantization is outlined with discussion of its limitations in Sec. 2. Secs. 3–5 survey generally valid statistical models for undithered, subtractively dithered, and nonsubtractively dithered quantizing systems, respectively. Each of these three sections is divided into subsections which address the statistics of the total error separately from those of the system output. In each case, a first-order statistical analysis of the error or output precedes a second-order analysis dealing with the relationship between error or output samples separated in time. For the dithered systems, properties of some practical dither signals are discussed.

2 CLASSICAL MODEL OF UNDITHERED QUANTIZATION

As mentioned, in an undithered quantizing system,

$$\varepsilon = q(x)\ . \tag{5}$$

Although it is a deterministic function of the input, the classical model of quantization treats this error as an additive random process which is independent of the input and *uniformly distributed*, meaning that the total error values have a probability density function (pdf) of the form

$$p_\varepsilon(\varepsilon) = \Pi_\Delta(\varepsilon) \tag{6}$$

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

$$\Pi_\Gamma(\varepsilon) \triangleq \begin{cases} \dfrac{1}{\Gamma}\ , & -\dfrac{\Gamma}{2} < \varepsilon \leq \dfrac{\Gamma}{2} \\ 0\ , & \text{otherwise.} \end{cases} \tag{7}$$

The pdf of Eq. (6) will be referred to as a uniform or rectangular pdf.

The mth *moment* of a random variable ε with pdf

[2] A stationary random process is one whose statistical properties are time invariant.

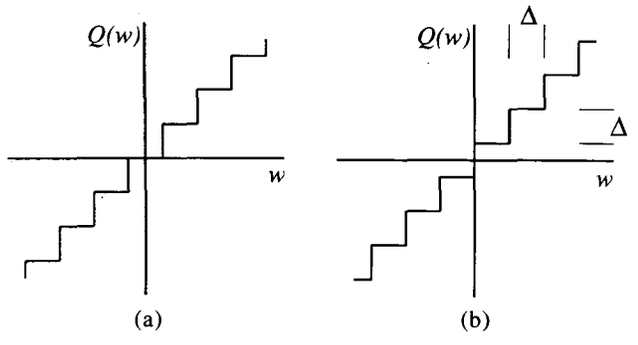


Fig. 1. Quantizer transfer characteristics. (a) Midtread. (b) Midriser. Size of 1 LSB is denoted by Δ.

$p_\varepsilon(\varepsilon)$ is defined as the expectation value of ε^m:

$$E[\varepsilon^m] \triangleq \int_{-\infty}^{\infty} \varepsilon^m p_\varepsilon(\varepsilon)\, d\varepsilon \tag{8}$$

where $E[\ ]$, the expectation value operator, is defined more generally by

$$E[f] \triangleq \int_{-\infty}^{\infty} f(\varepsilon) p_\varepsilon(\varepsilon)\, d\varepsilon\ . \tag{9}$$

The zeroth moment of any random process (that is, $E[\varepsilon^0]$) is identically equal to unity. The first moment is usually referred to as the *mean* of the process, whereas the term *variance* refers to the quantity $E[(\varepsilon - E[\varepsilon])^2] = E[\varepsilon^2] - E^2[\varepsilon]$. It is clear that if the mean of a random process is zero, its variance and second moment are equal.

For a random process uniformly distributed according to Eq. (6), the moments are

$$E[\varepsilon] = 0 \tag{10}$$

$$E[\varepsilon^2] = \frac{\Delta^2}{12} \tag{11}$$

$$E[\varepsilon^m] = \begin{cases} \dfrac{1}{m+1}\left(\dfrac{\Delta}{2}\right)^m, & m \text{ even} \\ 0, & m \text{ odd.} \end{cases} \tag{12}$$

Eq. (11) is the familiar expression for quantization error variance (or power) in the classical model.

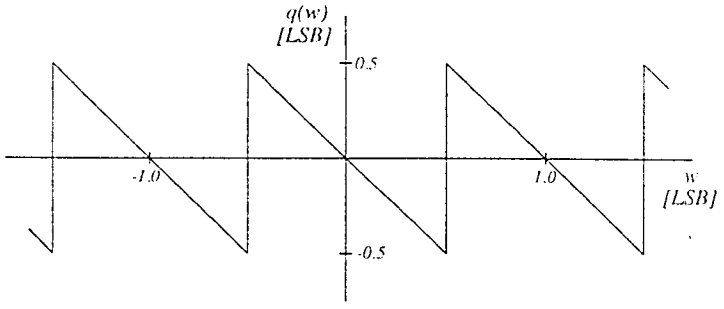


Fig. 2. Quantization error $q(w)$ as a function of quantizer input w for midtread quantizer.

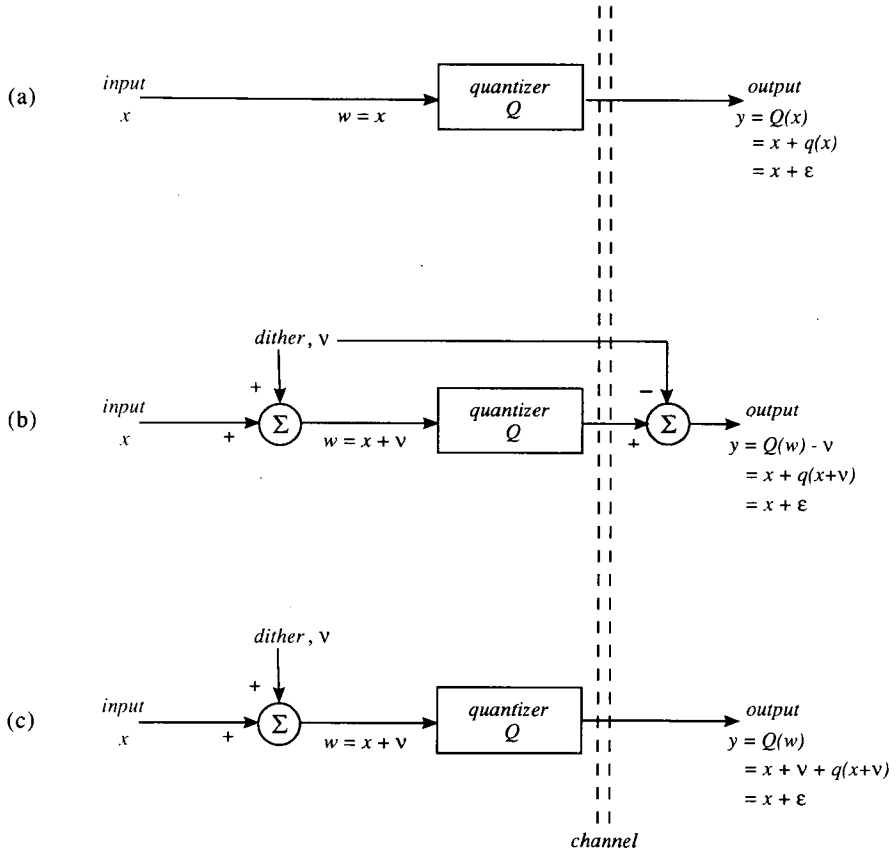


Fig. 3. Archetypal quantizing systems. (a) Undithered. (b) Subtractively dithered. (c) Nonsubtractively dithered. Shown are system input x, quantizer input w, and system output y.

This model of quantization error is valid for complex (quasi-random) input signals which are large relative to an LSB. It fails catastrophically for small or simple signals where, in undithered systems, the quantization error retains the character of input-dependent distortion or noise modulation.

The nonrandom nature of the error can be demonstrated by using a computer to simulate the undithered quantization of a very simple signal, say, a 1-kHz sine wave of 4.0 LSB peak-to-peak amplitude. Fig. 4 shows the system input and output from such a simulation as well as the resulting total error signal, and the power spectrum of the system output. Evidence of the input signal is clearly visible in the total error waveform. In the power spectrum, many sharp peaks fall at multiples of the input sine wave frequency, indicating not only a high degree of nonrandom structure (that is, harmonic distortion) in the error signal, but also a strong correlation between this signal and the system input. Inharmonic peaks are also present due to aliasing of distortion components above the Nyquist frequency (22.05 kHz in this simulation) into the baseband.

3 WIDROW'S MODEL OF UNDITHERED QUANTIZATION

A generalized statistical model of undithered quantization, valid for inputs with arbitrary statistical properties, was first developed by Widrow [6]–[8] in the 1950s. Widrow realized that quantizing a signal transforms its pdf from a continuous function to a train of impulse functions in a fashion reminiscent of time sampling, so that recovery of the system input statistics from those of the system output must require conditions analogous to those of the sampling theorem [5].

We pause to mention that the theoretical results presented in this section have limited practical significance, since detailed statistical knowledge of the total error produced by an undithered quantizer is not very helpful in alleviating the undesirable audible effects of this error. The development which follows is included both for completeness and in order to provide the mathematical machinery necessary for the treatment of dithered systems. In particular, the results of this section can be applied to subtractively dithered systems almost directly with the aid of a very simple transformation.

3.1 Statistics of the Total Error

3.1.1 First-Order Statistics

We begin by deriving Widrow's expression for the pdf of the total error, p_ε, in terms of the input pdf p_x. In an undithered system we have seen that the total error of the system is a deterministic function of the input. This is reflected in the form of the conditional probability density function (cpdf) of ε given x, denoted[3] by $p_{\varepsilon|x}(\varepsilon, x)$. This bivariate function represents the

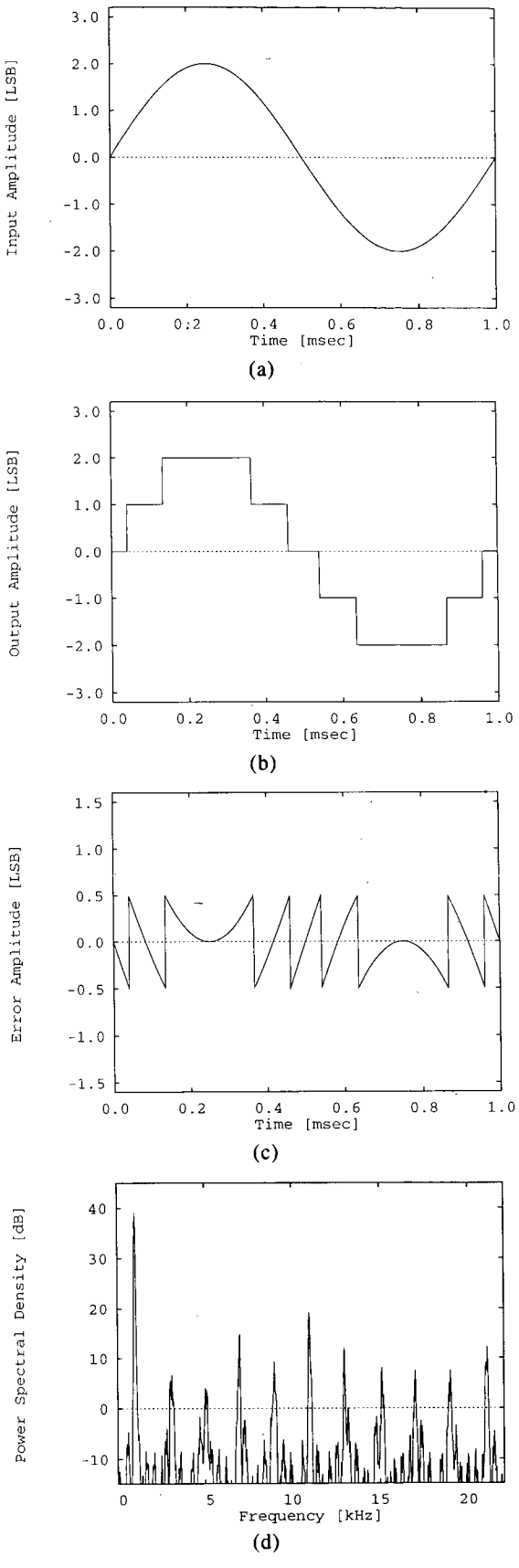


Fig. 4. Results from computer-simulated quantization of 1-kHz sine wave of 4.0-LSB peak-to-peak amplitude without dither. (a) System input signal. (b) System output signal. (c) Resulting total error signal. (d) Power spectrum of system output signal (as estimated from sixty 50% overlapping Hann-windowed 512-point time records with assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$, T being the sampling period).

[3] A more conventional notation is $p_{\varepsilon|x}(\varepsilon|x)$, but in many cases this can be confusing since, in fact, $p_{\varepsilon|x}$ is simply a function of the two variables ε and x.

pdf of the error for a specified input value. From Eqs. (1) and (3) the cpdf in question is expressible in terms of the Dirac delta function δ as

$$p_{\varepsilon|x}(\varepsilon, x) = \delta(\varepsilon - q(x))$$

$$= \delta\left(\varepsilon - \Delta\left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor + x\right) \quad (13)$$

$$= \Delta \Pi_\Delta(\varepsilon) \cdot W_\Delta(\varepsilon + x) ,$$

where

$$W_\Gamma(\varepsilon) \triangleq \sum_{k=-\infty}^{\infty} \delta(\varepsilon - k\Gamma) \quad (14)$$

is a train of Dirac delta functions separated by intervals of width Γ. The rectangular window function in Eq. (13) serves to select from the impulse train W_Δ that delta function which is closest to the origin (since $|\varepsilon| \leq \Delta/2$, of course).

We could easily use Eq. (13) to derive the marginal pdf of the total error, $p_\varepsilon(\varepsilon)$, by integrating $p_{\varepsilon|x}(\varepsilon, x)p_x(x)$ with respect to x, but it is more instructive to reason as follows. If $-\Delta/2 \leq x < \Delta/2$, then, by inspection of Fig. 5, we see that the distribution of errors arising from such inputs is equal to $p_x(-x)$ between $-\Delta/2$ and $\Delta/2$ and zero elsewhere. Similarly, the distribution of errors arising from inputs x such that $-3\Delta/2 \leq x < -\Delta/2$ is given by that portion of $p_x(-x)$ which resides between $\Delta/2$ and $3\Delta/2$, but recentered around $x = 0$. We conclude that

$$p_\varepsilon(\varepsilon) = \begin{cases} \sum_{k=-\infty}^{\infty} p_x(-\varepsilon + k\Delta), & -\frac{\Delta}{2} \leq \varepsilon < \frac{\Delta}{2} \\ 0, & \text{otherwise} \end{cases}$$

$$= \Delta \Pi_\Delta(\varepsilon) \sum_{k=-\infty}^{\infty} p_x(-\varepsilon + k\Delta)$$

$$= \Delta \Pi_\Delta(\varepsilon) \cdot [W_\Delta * p_x](-\varepsilon) , \quad (15)$$

where * denotes the operation of convolution.

The characteristic function (cf) of a random variable is the Fourier transform of its pdf, and is often easier to interpret than the pdf itself. We define the Fourier transform operator $\mathcal{F}[\ ]$ by[4]

$$\mathcal{F}[f](u) \triangleq F(u) \triangleq \int_{-\infty}^{\infty} f(\varepsilon)\, e^{-j2\pi u\varepsilon}\, d\varepsilon . \quad (16)$$

Then from Eq. (15) the cf P_ε of ε is given by

$$P_\varepsilon(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u} * [W_{\frac{1}{\Delta}}(-u)P_x(-u)]$$

$$= \sum_{k=-\infty}^{\infty} P_x\left(-\frac{k}{\Delta}\right) \frac{\sin[\pi\Delta(u - k/\Delta)]}{\pi\Delta(u - k/\Delta)} , \quad (17)$$

where P_x represents the cf of the system input. If the error is to be uniformly distributed, Eq. (17) must reduce to a single sinc function $\sin(\pi\Delta u)/(\pi\Delta u)$ centered at the origin, in which case it will be independent of P_x. We observe that this can occur only under the conditions of the following theorem:

Theorem 1 The total error induced by an undithered quantizing system is uniformly distributed if and only if the cf of the system input, P_x, satisfies the condition that

$$P_x(u)\Big|_{u=k/\Delta} = 0$$

$$\text{for } k = \pm 1, \pm 2, \pm 3, \ldots . \quad (18)$$

This condition is not actually due to Widrow,[5] but to Sripad and Snyder [9]. Note that if the requirements of Theorem 1 are satisfied, then the error is precisely of the sort which is postulated by the classical model and the moments of the total error are given by Eq. (12).

3.1.2 Second-Order Statistics

The statistical relationships between total error values separated in time are of particular interest since these determine the power spectral characteristics of the total error signal. Consider two system input values x_1 and x_2 occurring at times t_1 and t_2, respectively, so that they are separated in time by $\tau = t_2 - t_1$, where $\tau \neq 0$.[6] Their statistical relationship is described by their joint pdf $p_{x_1,x_2}(x_1, x_2)$, representing the probability

[4] This definition is retained throughout the sequel. $j = \sqrt{-1}$.

[5] Widrow [8] cites a different condition, which is sufficient but not necessary; namely, $P_x(u) = 0$ for $|u| \geq 1/\Delta$. Widrow calls this requirement "half-satisfaction" of the conditions of the quantizing theorem (see Theorem 3).

[6] In the special case where $\tau = 0$, the analysis reduces to that of Section 3.1.1.

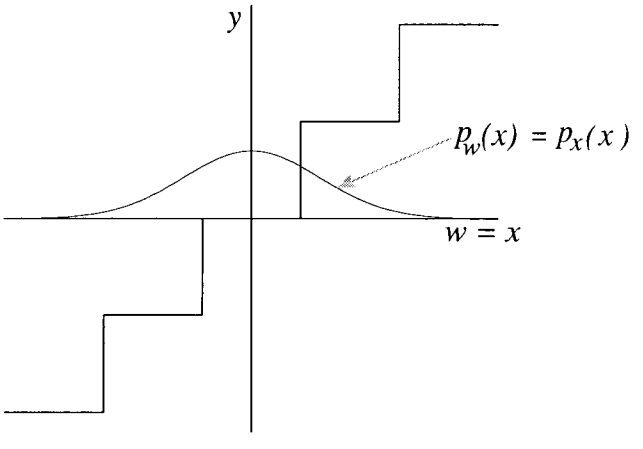


Fig. 5. Pdf of quantizer input in undithered quantizing system, showing its justification relative to the quantizer characteristic.

that the inputs at times t_1 and t_2 have the indicated values. Since the error is a deterministic function of the input in an undithered system, we can immediately write down the cpdf for a pair of error values, call them ε_1 and ε_2, given the inputs x_1 and x_2 [see Eq. (13)]:

$$p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2; x_1, x_2)$$

$$= \delta(\varepsilon_1 - q(x_1))\delta(\varepsilon_2 - q(x_2)) \tag{19}$$

$$= \Delta^2 \Pi_{\Delta\Delta}(\varepsilon_1, \varepsilon_2) \cdot W_{\Delta\Delta}(\varepsilon_1 + x_1, \varepsilon_2 + x_2) \tag{20}$$

where

$$\Pi_{\Gamma\Gamma}(\varepsilon_1, \varepsilon_2) \triangleq \Pi_\Gamma(\varepsilon_1)\Pi_\Gamma(\varepsilon_2) \tag{21}$$

$$W_{\Gamma\Gamma}(\varepsilon_1, \varepsilon_2) \triangleq W_\Gamma(\varepsilon_1)W_\Gamma(\varepsilon_2) \ . \tag{22}$$

Now we can straightforwardly compute the joint pdf of the two error values [see Eq. (15)]:

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2; x_1, x_2)$$

$$\times \ p_{x_1,x_2}(x_1, x_2)\, dx_1\, dx_2$$

$$= \Delta^2 \Pi_{\Delta\Delta}(\varepsilon_1, \varepsilon_2) \cdot [W_{\Delta\Delta} * p_{x_1,x_2}](-\varepsilon_1, -\varepsilon_2) \tag{23}$$

where the convolution is two-dimensional, involving both ε_1 and ε_2.

Taking the two-dimensional Fourier transform of Eq. (23) with respect to both ε_1 and ε_2, we find that the joint cf of ε_1 and ε_2 is given by [see Eq. (17)]

$$P_{\varepsilon_1,\varepsilon_2}(u_1, u_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right)$$

$$\times \frac{\sin[\pi\Delta(u_1 - k_1/\Delta)]}{\pi\Delta(u_1 - k_1/\Delta)} \frac{\sin[\pi\Delta(u_2 - k_2/\Delta)]}{\pi\Delta(u_2 - k_2/\Delta)} \ , \tag{24}$$

where P_{x_1,x_2} denotes the joint cf of x_1 and x_2. We can now write an obvious two-dimensional analog of Theorem 1:

Theorem 2 In an undithered quantizing system, the joint pdf $p_{\varepsilon_1,\varepsilon_2}$ of total error values ε_1 and ε_2, separated in time by $\tau \neq 0$, is given by

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \Pi_\Delta(\varepsilon_1)\Pi_\Delta(\varepsilon_2) \tag{25}$$

if and only if the joint cf P_{x_1,x_2} of the corresponding system inputs x_1 and x_2 satisfies the condition that

$$P_{x_1,x_2}(u_1, u_2)\Big|_{\substack{u_1=k_1/\Delta \\ u_2=k_2/\Delta}} = 0$$

for all integers k_1, k_2 with $(k_1, k_2) \neq (0, 0)$. (26)

Eq. (25) shows that, subject to the specified conditions, the joint pdf of ε_1 and ε_2 is a product of two rectangular window functions, one of which is a function of ε_1 alone and the other of ε_2 alone. Hence the two error values are statistically independent and each is uniformly distributed, so that for $\tau \neq 0$ we can write

$$E[\varepsilon_1^m \varepsilon_2^n] = E[\varepsilon_1^m]E[\varepsilon_2^n] \ . \tag{27}$$

In a digital system the total error is a discrete-time signal, so that $\tau = kT$, where T represents the sampling period and k is an integer. The *autocorrelation function* of such a signal is defined to be the function $E[\varepsilon_1\varepsilon_2](k)$ of the lag parameter k. According to the Wiener–Khinchin theorem [10], the power spectral density (PSD) of a discrete-time signal is equal to the discrete-time Fourier transform (DTFT) of its autocorrelation function, where we define the DTFT as

$$\mathcal{F}_{\mathrm{DT}}[h](f) \triangleq 2T \sum_{k=-\infty}^{\infty} h(k)\mathrm{e}^{-\mathrm{j}2\pi fkT} \ , \tag{28}$$

and where the 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.)

For an undithered system satisfying the conditions of Theorems 1 and 2, the autocorrelation function of the error is given by

$$E[\varepsilon_1\varepsilon_2](k) = \begin{cases} E[\varepsilon^2], & k = 0 \\ E[\varepsilon_1]E[\varepsilon_2], & \text{otherwise} \end{cases} \tag{29}$$

$$= \begin{cases} \dfrac{\Delta^2}{12}, & k = 0 \\ 0, & \text{otherwise} \ . \end{cases} \tag{30}$$

Thus its PSD is given by

$$\mathrm{PSD}_\varepsilon(f) = \frac{\Delta^2 T}{6} \ , \tag{31}$$

which is obviously constant with respect to frequency so that the error signal is spectrally white and exhibits a total power of $\Delta^2/12$ up to the Nyquist frequency.

3.2 Statistics of the System Output

We now proceed to investigate the relationship between the input and the output of an undithered quantizing system.

3.2.1 First-Order Statistics

The output can only assume values which are integer multiples of the quantization step size Δ. Referring to Fig. 5, we see that the probability of an output having value $y = k\Delta$, for some specified integer k, is equal to the probability that the input lies between $-\Delta/2 + k\Delta$ and $\Delta/2 + k\Delta$. Hence,

$$p_y(y) = \sum_{k=-\infty}^{\infty} \delta(y - k\Delta) \int_{-\Delta/2+k\Delta}^{\Delta/2+k\Delta} p_x(x)\, dx \ . \tag{32}$$

Borrowing Widrow's terminology, we say that the quantization operation performs *area sampling* of the input distribution. Writing the integral in Eq. (32) as a convolution of p_x with a rectangular window function, it reduces to

$$p_y(y) = [\Delta \Pi_\Delta * p_x](y) \cdot W_\Delta(y) \ . \tag{33}$$

Clearly, $p_y(y)$ cannot be equal to $p_x(y)$ unless the input pdf is itself a train of impulses separated by intervals of Δ (that is, unless x is already quantized at the LSB level). We will see, however, that the statistical properties of the input can be recovered from the output subject to certain less restrictive conditions.

From Eq. (33), the cf of y, P_y, is given by

$$\begin{aligned} P_y(u) &= G_x(u) * W_{\frac{1}{\Delta}}(u) \\ &= \sum_{k=-\infty}^{\infty} G_x\left(u - \frac{k}{\Delta}\right) , \end{aligned} \tag{34}$$

where

$$G_x(u) \triangleq \frac{\sin(\pi \Delta u)}{\pi \Delta u} \cdot P_x(u) \tag{35}$$

and where P_x is the cf of the input. Hence $P_y(u)$ consists of copies of the function $G_x(u)$ separated by intervals of $1/\Delta$. Note, however, that if P_x is band-limited such that $P_x(u) = 0$ for $|u| \geq 1/(2\Delta)$, then the repeated versions of $G_x(u)$ do not overlap, allowing recovery of the input cf (and hence the input pdf) from that of the output. Indeed, this is [6]–[8]:

Theorem 3 (Widrow's Quantizing Theorem) The pdf of the input to an undithered infinite linear quantizing system is recoverable from the pdf of its output if the cf of the input, P_x, is band-limited such that $P_x(u) = 0$ for $|u| \geq 1/(2\Delta)$.

Obviously, this theorem closely resembles the sampling theorem, which allows recovery of an appropriately band-limited analog signal from discrete-time samples thereof. The notable difference, of course, is that the quantizing theorem pertains not to time sampling, but to amplitude quantizing of a signal (that is, to area sampling of the pdf of a signal).

In practice, recovering the pdf of the input is often unnecessary and it is sufficient to recover the moments of the input signal from the output. The mth moment of the output signal can be expressed in terms of either the pdf or the cf of y by[7]

$$E[y^m] = \int_{-\infty}^{\infty} y^m p_y(y)\, dy \tag{36}$$

$$= \left(\frac{j}{2\pi}\right)^m \frac{d^m P_y}{du^m}(u)\bigg|_{u=0} \ . \tag{37}$$

It is obvious that if the quantizing theorem is satisfied then the shifted versions of $G_x(u)$ do not overlap, so that the mth derivative of $P_y(u)$ at the origin is determined only by the "baseband" ($k = 0$) term in Eq. (34). This is also true, however, subject to the weaker condition that the quantizing theorem is only half-satisfied (see footnote 5) or the still weaker condition that

$$\frac{d^m G_x}{du^m}(u)\bigg|_{u=k/\Delta} = 0 \qquad \text{for } k = \pm 1, \pm 2, \pm 3, \ldots . \tag{38}$$

If the input statistics obey this condition, then

$$E[y^m] = \left(\frac{j}{2\pi}\right)^m \frac{d^m G_x}{du^m}(u)\bigg|_{u=0} \tag{39}$$

so that by repeated differentiation of Eq. (35) we can express the moments of y in terms of the moments of x. In particular, for the first few moments we can write the following useful relationships, which give rise to Sheppard's corrections for grouping:

$$E[y] = E[x] \tag{40}$$

$$E[y^2] = E[x^2] + \frac{\Delta^2}{12} \tag{41}$$

$$E[y^m] = \sum_{l=0}^{[m/2]} \binom{m}{2l} \left(\frac{\Delta}{2}\right)^{2l} \frac{E[x^{m-2l}]}{2l + 1} \ . \tag{42}$$

We emphasize that, in an undithered quantizing system, each of these equations for $E[y^m]$ is only valid when Eq. (38) 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.

Furthermore, these equations show that if Eq. (38) is satisfied for some m, then the mth moment of $y = x + \varepsilon$ is the same as that of x plus a statistically independent additive random process with uniform pdf.

[7] Eq. (37) can be straightforwardly derived by writing $\mathcal{F}[p_y](u) = E[e^{-j2\pi uy}]$, expanding the complex exponential in a Taylor series about $y = 0$ and using the fact that the expectation value operator is linear.

It is important to remember, however, that x and ε are not, in fact, statistically independent and that, for an undithered quantizing system, they are deterministically related as shown in Fig. 2.

We note in passing that by repeated differentiation of Eq. (35) for $G_x(u)$ we can derive from Eq. (38) a stronger, but perhaps more practical, condition in terms of the input cf, which ensures that $E[y^m]$ obeys Eqs. (39) and (42) for $m = 1, 2, \ldots, M$:

$$\left.\frac{d^i P_x}{du^i}(u)\right|_{u=k/\Delta} = 0 \qquad \text{for } k = \pm 1, \pm 2, \pm 3, \ldots; \quad i = 0, 1, 2, \ldots, M-1. \tag{43}$$

3.2.2 Second-Order Statistics

Proceeding in a fashion similar to that of Sec. 3.2.1, we find that the joint pdf of two system output values y_1 and y_2, separated in time by $\tau \neq 0$, is given [see Eq. (33)] by

$$p_{y_1,y_2}(y_1, y_2) = [\Delta^2 \Pi_{\Delta\Delta} * p_{x_1,x_2}](y_1, y_2) \cdot W_{\Delta\Delta}(y_1, y_2) \tag{44}$$

with corresponding joint cf [see Eq. (34)]

$$P_{y_1,y_2}(u_1, u_2) = G_{x_1,x_2}(u_1, u_2) * W_{\frac{1}{\Delta}\frac{1}{\Delta}}(u_1, u_2) , \tag{45}$$

where

$$G_{x_1,x_2}(u_1, u_2) \triangleq \frac{\sin(\pi\Delta u_1)}{\pi\Delta u_1}\frac{\sin(\pi\Delta u_2)}{\pi\Delta u_2} \cdot P_{x_1,x_2}(u_1, u_2). \tag{46}$$

We can now write a two-dimensional analog of the quantizing theorem, namely that the joint pdf of the input is recoverable from that of the output if $P_{x_1,x_2}(u_1, u_2) = 0$ for $|u_1| \geq 1/(2\Delta)$ or $|u_2| \geq 1/(2\Delta)$, or both. Of potentially greater interest, however, is the two-dimensional analog of Eq. (39), which allows us to recover the joint moments of the system input from those of the output. That is, if

$$\left.\frac{\partial^{m+n} G_{x_1,x_2}}{\partial u_1^m \partial u_2^n}(u_1,u_2)\right|_{\substack{u_1=k_1/\Delta \\ u_2=k_2/\Delta}} = 0$$

$$\text{for all integers } k_1, k_2 \text{ with } (k_1, k_2) \neq (0, 0) \tag{47}$$

then

$$E[y_1^m y_2^n] = \left(\frac{j}{2\pi}\right)^{m+n} \left.\frac{\partial^{m+n} G_{x_1,x_2}}{\partial u_1^m \partial u_2^n}(u_1, u_2)\right|_{\substack{u_1=0 \\ u_2=0}} . \tag{48}$$

In this case we can write an expression analogous to Eq. (42), relating the joint moments of the output to those of the input:

$$E[y_1^m y_2^n] = \sum_{l_1=0}^{\lfloor m/2 \rfloor} \sum_{l_2=0}^{\lfloor n/2 \rfloor} \binom{m}{2l_1}\binom{n}{2l_2}\left(\frac{\Delta}{2}\right)^{2(l_1+l_2)} \times \frac{E[x_1^{m-2l_1} x_2^{n-2l_2}]}{(2l_1+1)(2l_2+1)} . \tag{49}$$

In particular, we can write that

$$E[y_1 y_2](k) = \begin{cases} E[x^2] + \dfrac{\Delta^2}{12}, & \text{for } k = 0 \\ E[x_1 x_2], & \text{otherwise} \end{cases} \tag{50}$$

so that the power spectral density of the output is identical to that of the input apart from an additive white-noise component arising from the quantization operation; that is

$$\text{PSD}_y(f) = \text{PSD}_x(f) + \frac{\Delta^2 T}{6} . \tag{51}$$

3.3 Summary of Undithered Quantization

It is clear that the results of this section are primarily of theoretical, rather than practical, interest. All of the theorems given impose conditions on the statistics of the system input, and such restrictions are usually undesirable in practice. Some not uncommon system inputs satisfy the conditions of Theorem 1 (for example, a uniformly distributed random input), but the conditions of the quantizing theorem (Theorem 3) cannot be met by any system input whose pdf is only nonzero on a finite interval (although some signals, such as large ones with Gaussian distributions, can come very close to satisfying the conditions [6]–[8]).

There now becomes apparent, however, the possibility of dithering the system input with a suitably chosen dither signal v so as to ensure that the quantizer input $w = x + v$ [Fig. 3(b) and (c)] satisfies some of the aforementioned conditions. In particular, if the dither is statistically independent of the system input, then the pdf p_w of w is the convolution $p_w = p_x * p_v$, and hence its cf is the product $P_w = P_x \cdot P_v$. In this case the dither statistics can be freely chosen so as to cause P_w to vanish at the required places, and this accomplishment cannot be undone by any system input x that is statistically independent of v.

These ideas will now all be quantified using the mathematical techniques developed in this section.

4 SUBTRACTIVE DITHER

The first use of subtractive dither must be credited to Roberts [11] in the early 1960s, who applied it to picture coding. Roberts added uniformly distributed random noise of 1-LSB peak-to-peak amplitude (statistically independent of the system input) to a video

signal prior to its quantization, and subsequently subtracted the same dither signal from the quantizer's output. He found that the total error of the quantizing system was uniformly distributed and statistically independent of the input signal.

Prompted by Roberts' discovery, a theoretical investigation of subtractive dithering was undertaken by Schuchman [12], a student of Widrow's. Schuchman derived a condition on the dither pdf [see Eq. (56)] which guarantees that the total error and the input to the quantizer are statistically independent of each other. His work was later recast and generalized to include a second-order statistical analysis by Sherwood [13].

4.1 Statistics of the Total Error

4.1.1 First-Order Statistics

The quantizer input is $w = x + v$ so that the output of the system is [see Fig. 3(b)]

$$y = Q(x + v) - v \tag{52}$$

and hence the total error is given by

$$\begin{aligned} \varepsilon &= y - x \\ &= Q(x + v) - (x + v) \\ &= q(x + v) \end{aligned} \tag{53}$$

which is simply the quantization error of the total quantizer input w. Therefore we can treat the system as one of the sort described by the analysis of Sec. 3, but with w rather than x as its input. From Eq. (15) and $p_w(\varepsilon) = p_x(\varepsilon) * p_v(\varepsilon)$ (since x and v are statistically independent) we immediately obtain

$$\begin{aligned} p_\varepsilon(\varepsilon) &= \Delta\Pi_\Delta(\varepsilon) \cdot [W_\Delta * p_w](-\varepsilon) \\ &= \Delta\Pi_\Delta(\varepsilon) \cdot [W_\Delta * p_x * p_v](-\varepsilon) \end{aligned} \tag{54}$$

and

$$P_\varepsilon(u) = \frac{\sin(\pi\Delta u)}{\pi\Delta u} * [W_{\frac{1}{\Delta}}(-u) \cdot P_x(-u) \cdot P_v(-u)] . \tag{55}$$

Note that the statistical properties of the dither can be chosen to control the properties of the error. In particular, it can be shown that [12]:

Theorem 4 (Schuchman's Condition) In a subtractively dithered quantizing system the total error will be uniformly distributed and statistically independent of the input for arbitrary input distributions if and only if the cf of the dither, P_v, satisfies the condition that

$$P_v(u)\Big|_{u=k/\Delta} = 0 \qquad \text{for } k = \pm 1, \pm 2, \pm 3, \ldots . \tag{56}$$

This is clear from Eq. (55), which reduces to a single sinc function centered at the origin under the conditions of the theorem. In this case, the total error is of the form postulated by the classical model, and its moments are given by Eq. (12) for a system input of arbitrary distribution.

4.1.2 Second-Order Statistics

Proceeding as for the first-order statistics, we use Eq. (24) to deduce that for two total error values ε_1 and ε_2, separated in time by $\tau \neq 0$:[8]

$$\begin{aligned} P_{\varepsilon_1,\varepsilon_2}(u_1, u_2) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} & P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right) P_{v_1,v_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right) \\ & \times \frac{\sin[\pi\Delta(u_1 - k_1/\Delta)]}{\pi\Delta(u_1 - k_1/\Delta)} \frac{\sin[\pi\Delta(u_2 - k_2/\Delta)]}{\pi\Delta(u_2 - k_2/\Delta)} \end{aligned} \tag{57}$$

where P_{v_1,v_2} represents the joint pdf of dither values v_1 and v_2, applied to input values x_1 and x_2, respectively. We immediately draw the following conclusion:

Theorem 5 In a subtractively dithered quantizing system the joint pdf $p_{\varepsilon_1,\varepsilon_2}$ of two total error values ε_1 and ε_2, separated in time by $\tau \neq 0$, is given by

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \Pi_\Delta(\varepsilon_1)\Pi_\Delta(\varepsilon_2) \tag{58}$$

for arbitrary input distributions if and only if the joint cf P_{v_1,v_2} of the corresponding dither values v_1 and v_2 satisfies the condition that

$$P_{v_1,v_1}(u_1, u_2)\Big|_{\substack{u_1=k_1/\Delta \\ u_2=k_2/\Delta}} = 0$$

for all integers k_1, k_2 with $(k_1, k_2) \neq (0, 0)$ (59)

We observe that if the conditions of this theorem are satisfied, then ε_1 and ε_2 are both uniformly distributed and statistically independent of each other.

It should be noted that if v_1 and v_2 are statistically independent of each other, and the cf of each satisfies Eq. (56), then Eq. (59) will be satisfied. This is the situation of interest in most practical applications using subtractive dither.

Subject to satisfaction of Eq. (59), the joint moments

[8] In the special case where $\tau = 0$, the analysis reduces to that of Sec. 4.1.1.

of ε_1 and ε_2 are given by Eq. (27), reproduced here for reference:

$$E[\varepsilon_1^m \varepsilon_2^n] = E[\varepsilon_1^m]E[\varepsilon_2^n]$$

so that ε_1^m and ε_2^n are uncorrelated. In particular, for $m = n = 1$ and a stationary dither signal,

$$\begin{aligned} E[\varepsilon_1 \varepsilon_2] &= E[\varepsilon_1]E[\varepsilon_2] \\ &= E^2[\varepsilon] \\ &= 0 \end{aligned} \tag{60}$$

independent of τ, for $\tau \neq 0$, where we have assumed that the dither satisfies the conditions of Theorem 4 so that the moments of the total error will be those predicted by Eq. (12) of the classical model. Eq. (60) indicates, of course, that the total error signal will be spectrally white even if the dither signal is not.

4.2 Statistics of the System Output

4.2.1 First-Order Statistics

For pragmatic reasons, we will not derive an expression for the output pdf, given an arbitrary dither distribution. Instead, we will assume that the dither signal satisfies the conditions of Theorem 4. Then, since the total error is statistically independent of the input, and since the output is given by $y = x + \varepsilon$, we can immediately write that

$$\begin{aligned} p_y(y) &= [p_\varepsilon * p_x](y) \\ &= [\Delta \Pi_\Delta * p_x](y) \end{aligned} \tag{61}$$

so that

$$\begin{aligned} P_y(u) &= \frac{\sin(\pi \Delta u)}{\pi \Delta u} \cdot P_x(u) \\ &= G_x(u) \ . \end{aligned} \tag{62}$$

Obviously, the output is precisely the sum of the input plus a statistically independent, uniformly distributed random process. Hence, the moments of the output in terms of the moments of the input will be given by Eq. (42), which, in this case, is valid for all m.

4.2.2 Second-Order Statistics

If P_{v_1,v_2} satisfies the conditions of Theorem 5, then

$$p_{y_1,y_2}(y_1, y_2) = [\Delta^2 \Pi_{\Delta\Delta} * p_{x_1,x_2}](y_1, y_2) \tag{63}$$

so that

$$\begin{aligned} P_{y_1,y_2}(u_1, u_2) &= \frac{\sin(\pi \Delta u_1)}{\pi \Delta u_1} \frac{\sin(\pi \Delta u_2)}{\pi \Delta u_2} \cdot P_{x_1,x_2}(u_1, u_2) \\ &= G_{x_1,x_2}(u_1, u_2) \ . \end{aligned} \tag{64}$$

Hence, the joint moments of the output in terms of the moments of the input will be given by Eq. (49), and Eqs. (50) and (51) will hold. The quantization process has thus merely added to the input signal a white-noise process of total power $\Delta^2/12$ (up to the Nyquist frequency).

4.3 Properties of Practical Dither Signals

It is naturally of interest to inquire as to which common random signals satisfy the criterion of Theorem 4. Perhaps the simplest imaginable candidate is dither with the uniform pdf

$$p_v(v) = \Pi_\Delta(v) \tag{65}$$

whose corresponding cf is the sinc function

$$P_v(u) = \frac{\sin(\pi \Delta u)}{\pi \Delta u} \ . \tag{66}$$

This cf obviously satisfies the desired conditions. We conclude that dither of uniform pdf will render the total error statistically independent of the input and uniformly distributed in a subtractively dithered system. We assume that values in the dither sequence are statistically independent of one another so that the criterion of Theorem 5 is also satisfied and distinct values in the total error sequence are statistically independent of one another (thus ensuring that this sequence meets the weaker requirement of being spectrally white).

Of course, there are other cf's which meet the requirement of vanishing at all nonzero multiples of $1/\Delta$. For instance, a dither produced by summing n independent uniformly distributed random processes, each of 1-LSB peak-to-peak amplitude, will yield a dither which satisfies the criterion. The summation operation convolves the pdf's of the random processes together, thus multiplying their cf's, so that such a dither will exhibit a cf of the form

$$P_v(u) = \left[\frac{\sin(\pi \Delta u)}{\pi \Delta u}\right]^n \ . \tag{67}$$

However, in a subtractively dithered system such dithers have no inherent advantage over the uniform-pdf dither of Eq. (65).

4.4 Summary of Subtractive Dither

The most practically important theoretical results concerning subtractively dithered quantizing systems are that:

1) The total error can be rendered uniformly distributed and statistically independent of the system input by choosing a dither which satisfies the conditions of Theorem 4.

2) Values of the total error separated in time can be rendered statistically independent of one another (so that the total error signal is spectrally white) by using a dither whose values, in addition to satisfying Theorem

4, are statistically independent of one another.

A familiar dither which satisfies all the required conditions is one with a rectangular pdf of 1-LSB peak-to-peak amplitude, and whose values are statistically independent of one another. Fig. 6 shows the results of a computer-simulated quantization operation performed upon a 1-kHz sine wave of 4.0-LSB peak-to-peak amplitude and using this type of subtractive dither. Shown are the system input and output, the total error, and the power spectrum of the system output. Note that the system output resembles a sine wave plus an independent additive noise, and that no trace of the input signal is visible in the noiselike total error waveform. Furthermore, the power spectrum of the system output exhibits no distortion components whatsoever and indicates that the total error is spectrally white. (The 0-dB noise floor in Fig. 6 represents a power spectral density of $\Delta^2 T/6$, which has an integrated noise power of $\Delta^2/12$ up to the Nyquist frequency.) These results should be compared with those in Fig. 4, which illustrate the signal-dependent distortions produced by an undithered quantizing system with the same system input signal.

Subtractively dithered quantizing systems are clearly ideal in the sense that they render the total error an input-independent additive noise process. The requirement of dither subtraction at the system output, however, imposes restrictions which make it difficult to implement in practical audio applications. For example, the dither signal must be available at the output, and so either the dither must be transmitted along with the signal or synchronized dither generators must be present at either end of the channel. Even more seriously, any signal editing or modification occurring between the original quantization and the subtraction of the dither necessitates a like operation on the dither sequence. It is for such reasons that subtractive dither is generally not a feasible option, and nonsubtractive dithering schemes are of interest. Although many of the same benefits can be achieved (see Sec. 5), the total error variance is inevitably greater, and the beautiful result regarding full statistical independence of the total error is unattainable.

5 NONSUBTRACTIVE DITHER

We now examine the possibility of using dither without subsequently subtracting it. Early investigations into nonsubtractive dither were conducted by Wright [14] in 1979, resulting in the discovery of many of the important results that follow. This work has remained unpublished, and has only very recently become known to the authors [15], [16].

The results concerning conditional moments of the error signal were rediscovered independently by Stockham [17] in 1980, as documented in an unpublished Master's thesis by Brinton [18], a student of Stockham's, in 1984. Stockham remained silent on the matter until recently [4] for commercial reasons (triangular-pdf dither was used in the Soundstream digital

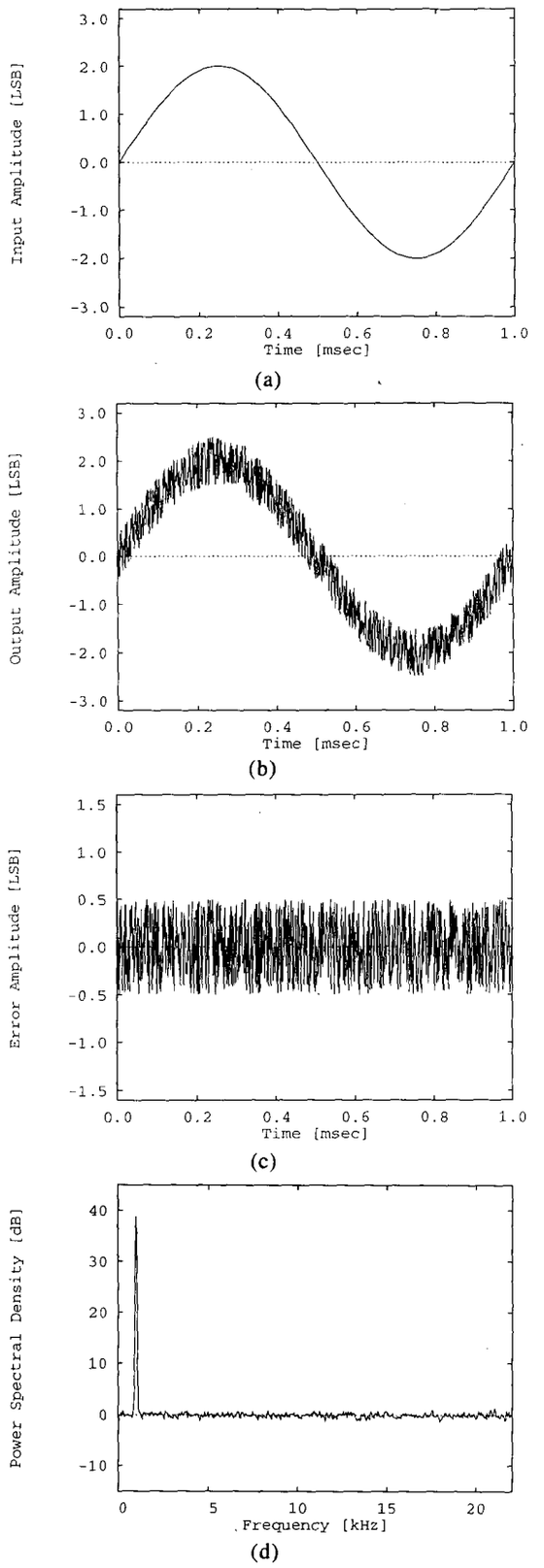


Fig. 6. Results from computer-simulated quantization of 1-kHz sine wave of 4.0-LSB peak-to-peak amplitude using rectangular-pdf subtractive dither of 1-LSB peak-to-peak amplitude. (a) System input signal. (b) System output signal. (c) Resulting total error signal. (d) Power spectrum of system output signal (as estimated from sixty 50% overlapping Hann-windowed 512-point time records with assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$, T being the sampling period).

recording/editing system in the early 1980s).

The properties of nonsubtractive dither were again investigated independently by two of the authors, Lipshitz and Vanderkooy, in the mid-1980s. Vanderkooy and Lipshitz appear to have been the first researchers to publish their findings on nonsubtractive dither [1], [2], [19]–[21] (and in particular on triangular-pdf dither), and this prompted collation and extension of the theoretical aspects by another of the authors, Wannamaker [16], [22], [23]. The ultimate outcome has been the publication of a detailed analysis of nonsubtractively dithered systems coauthored with Wright [3]. The treatment of the subject presented hereafter represents an abridgement of [3], omitting many of the mathematical details and some of the more esoteric results.

Recently the results concerning conditional moments have again been independently discovered by Gray [24], using a somewhat different formal approach from the one adopted herein. Gray and Stockham are currently preparing a paper on the subject [4].

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. Particularly serious is a persistent confusion of subtractive and nonsubtractive dithering, which have quite different properties (see, for instance, [25, p. 170]). We will see that nonsubtractively dithered systems *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. They *can*, however, render certain statistical moments of the total error independent of the system input, and regulate the joint moments of total error values which are separated in time. For many applications this is as good as full statistical independence.

5.1 Statistics of the Total Error

5.1.1 First-Order Statistics

The quantizer output in a nonsubtractively dithered quantizing system is given by [see Fig. 3(c)]

$$y = Q(x + v) \tag{68}$$

so that the total error is given by

$$\varepsilon = y - x$$

$$= Q(x + v) - x \tag{69}$$

$$= q(x + v) + v \; . \tag{70}$$

Obviously the total error is not simply the quantization error alone, but also involves the dither. This fact somewhat complicates the statistical treatment of nonsubtractively dithered systems. We begin by investigating the conditional pdf of the total error given the system input.

In order to derive an expression for $p_{\varepsilon|x}(\varepsilon, x)$, we consider a nonsubtractively dithered quantizing system with a specified input value x. The input to the quantizer is $w = x + v$, which has the cpdf

$$p_{w|x}(w, x) = p_{(x+v)|x}(w, x) \; . \tag{71}$$

Since x and v are statistically independent, we can write (see Fig. 7)

$$p_{w|x}(w, x) = p_{x|x}(w, x) * p_{v|x}(w, x)$$

$$= \delta(w - x) * p_v(w)$$

$$= p_v(w - x) \; . \tag{72}$$

We observe that the total error depends on the values of both the input and the dither. In particular, if the input $w = x + v$ to the quantizer is between $-\Delta/2$ and $+\Delta/2$, the output will be nil (for a midtread 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, each weighted by the probability that w falls on the corresponding quantizer step:

$$p_{\varepsilon|x}(\varepsilon, x) = \sum_{k=-\infty}^{\infty} \delta(\varepsilon + x - k\Delta)$$

$$\times \int_{-\Delta/2+k\Delta}^{\Delta/2+k\Delta} p_v(w - x) \, dw \; . \tag{73}$$

Writing the integral in Eq. (73) as a convolution of p_v with a rectangular window function, it reduces to[9]

$$p_{\varepsilon|x}(\varepsilon, x) = [\Delta \Pi_\Delta * p_v](\varepsilon) \cdot W_\Delta(\varepsilon + x) \; . \tag{74}$$

Eq. (74) makes it clear that $p_{\varepsilon|x}(\varepsilon, x)$ cannot be rendered independent of x by any choice of dither pdf,

[9] Note that Eq. (74) reduces to Eq. (13) if dither is not used [that is, if $p_v(v) = \delta(v)$].

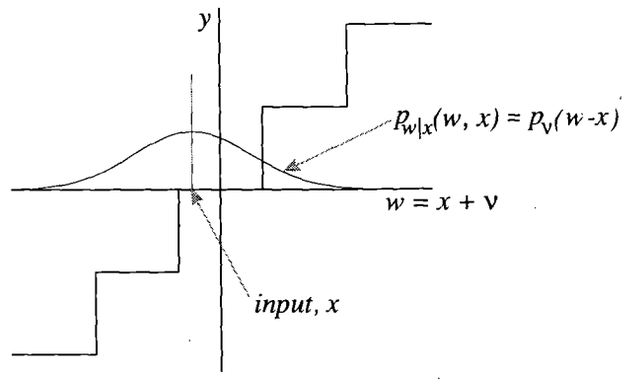


Fig. 7. Conditional pdf of quantizer input in nonsubtractively dithered quantizing system, showing its justification relative to the quantizer transfer characteristic.

since the convolution of any dither pdf (which must be nonegative everywhere) with a rectangular window function yields a function 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 the delta function is dependent on the input.[10] Furthermore it can be shown [3] that the marginal pdf of the error $p_\varepsilon(\varepsilon)$ is not uniform for arbitrary inputs.

Since we have seen that statistical independence of the total error from the input is not achievable, we turn our attention to the possibility of controlling moments of the error.

The *m*th *conditional moment* of the error signal given x is defined in the obvious fashion:

$$E[\varepsilon^m|x] \triangleq \int_{-\infty}^{\infty} \varepsilon^m p_{\varepsilon|x}(\varepsilon, x)\, d\varepsilon \tag{75}$$

$$= \left(\frac{j}{2\pi}\right)^m \frac{d^m P_{\varepsilon|x}}{du_\varepsilon^m}(u_\varepsilon, x)\bigg|_{u_\varepsilon=0} . \tag{76}$$

Taking the two-dimensional Fourier transform[11] (with respect to both ε and x) of Eq. (74) and substituting into Eq. (76) yields the Fourier transform with respect to x of $E[\varepsilon^m|x]$:

$$\mathcal{F}\left[E[\varepsilon^m|x]\right](u_x) = \left(\frac{j}{2\pi}\right)^m \frac{d^m}{du_x^m}\left[\frac{\sin(\pi\Delta u_x)}{\pi\Delta u_x} P_v(-u_x)\right] W_{\frac{1}{\Delta}}(u_x) . \tag{77}$$

If $E[\varepsilon^m|x]$ is to be independent of x, we require that its Fourier transform reduce to a constant times a single delta function at the origin. Imposing this restriction on Eq. (77) yields the following theorem:

Theorem 6 In a nonsubtractively dithered quantizing system, $E[\varepsilon^m|x]$ is functionally independent of x if and only if

$$\frac{d^m G_v}{du^m}(u)\bigg|_{u=k/\Delta} = 0 \qquad \text{for } k = \pm 1, \pm 2, \pm 3, \ldots \tag{78}$$

where

$$G_v(u) \triangleq \frac{\sin(\pi\Delta u)}{\pi\Delta u} \cdot P_v(u) . \tag{79}$$

If the conditions of Theorem 6 are satisfied, then from Eq. (77),

$$E[\varepsilon^m|x] = E[\varepsilon^m] = \left(\frac{j}{2\pi}\right)^m \frac{d^m G_v}{du^m}(u)\bigg|_{u=0} \tag{80}$$

so that we can derive useful expressions for the moments of the total error in terms of the moments of the dither signal by direct differentiation of Eq. (79):

$$E[\varepsilon] = E[v] \tag{81}$$

$$E[\varepsilon^2] = E[v^2] + \frac{\Delta^2}{12} \tag{82}$$

$$E[\varepsilon^m] = \sum_{l=0}^{\lfloor m/2 \rfloor} \binom{m}{2l}\left(\frac{\Delta}{2}\right)^{2l} \frac{E[v^{m-2l}]}{2l+1} \tag{83}$$

which, again, are related to Sheppard's corrections for grouping. We emphasize that each of these equations for $E[\varepsilon^m]$ is valid only when Theorem 6 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.

Eq. (82) clearly shows that with proper nonsubtractive dither satisfying the conditions of Theorem 6, the total error variance is greater than that of classical or subtractively dithered quantization (namely, $\Delta^2/12$) by the dither variance.

The following weaker theorem is really a corollary to Theorem 6, but is somewhat better known [14], [24]. It follows immediately from repeated differentiation of Eq. (79), resulting in binomial expansions involving the derivatives of $P_v(u)$ and the sinc function.

Theorem 7 $E[\varepsilon^l|x]$ is functionally independent of x for $l = 1, 2, \ldots, M$ if and only if

$$\frac{d^i P_v}{du^i}(u)\bigg|_{u=k/\Delta} = 0 \quad \text{for } k = \pm 1, \pm 2, \pm 3, \ldots; \quad i = 0, 1, 2, \ldots, M-1 . \tag{84}$$

5.1.2 Second-Order Statistics

Consider two total error values ε_1 and ε_2 which are separated in time by $\tau \neq 0$,[12] and the two corresponding input signal values x_1 and x_2. Using a derivation analogous to that of Sec. 5.1.1, we find that

$$p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2; x_1, x_2) = [\Delta^2 \Pi_{\Delta\Delta} * p_{v_1,v_2}](\varepsilon_1, \varepsilon_2) \cdot W_{\Delta\Delta}(\varepsilon_1 + x_1, \varepsilon_2 + x_2) , \tag{85}$$

[10] As discussed in Sec. 4 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, Eq. (56). Unfortunately $q(w)$ is not the total error of a nonsubtractively dithered system.

[11] This approach is due to Brinton [18], who also used it in deriving an expression for the conditional moments. Unfortunately his method entailed two invalid assumptions: 1) that the input was uniformly distributed, and 2) that the rectangular window function in Eq. (74) represented the pdf of a quantization error which was statistically independent of the dither.

[12] In the special case where $\tau = 0$, the analysis reduces to that of Sec. 5.1.1.

where the convolution is two-dimensional, involving both ε_1 and ε_2. Here p_{ν_1,ν_2} is the joint pdf of two dither values ν_1 and ν_2 corresponding to the inputs x_1 and x_2, respectively. Hence,

$$p_{\varepsilon_1,\varepsilon_2}(\varepsilon_1, \varepsilon_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{(\varepsilon_1,\varepsilon_2)|(x_1,x_2)}(\varepsilon_1, \varepsilon_2; x_1, x_2)p_{x_1,x_2}(x_1, x_2)\,dx_1\,dx_2$$

$$= [\Delta^2\Pi_{\Delta\Delta} * p_{\nu_1,\nu_2}](\varepsilon_1, \varepsilon_2) \cdot [W_{\Delta\Delta} * p_{x_1,x_2}](-\varepsilon_1, -\varepsilon_2) \tag{86}$$

so that

$$P_{\varepsilon_1,\varepsilon_2}(u_1, u_2) = \sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right)\frac{\sin[\pi\Delta(u_1 - k_1/\Delta)]}{\pi\Delta(u_1 - k_1/\Delta)}$$

$$\times \frac{\sin[\pi\Delta(u_2 - k_2/\Delta)]}{\pi\Delta(u_2 - k_2/\Delta)} P_{\nu_1,\nu_2}\left(u_1 - \frac{k_1}{\Delta}, u_2 - \frac{k_2}{\Delta}\right). \tag{87}$$

Clearly, no choice of joint dither cf P_{ν_1,ν_2} will allow Eq. (87) to be expressed as a product of two characteristic functions, one involving u_1 alone and the other u_2 alone, for arbitrary choices of the joint input cf P_{x_1,x_2}. We therefore conclude that 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 correct choice of the dither statistics.

From Eq. (87) we proceed to calculate the joint moments of ε_1 and ε_2, finding that

$$E[\varepsilon_1^m\varepsilon_2^n] = \left(\frac{j}{2\pi}\right)^{m+n}\sum_{k_1=-\infty}^{\infty}\sum_{k_2=-\infty}^{\infty} P_{x_1,x_2}\left(-\frac{k_1}{\Delta}, -\frac{k_2}{\Delta}\right)$$

$$\times \frac{\partial^{m+n}G_{\nu_1,\nu_2}}{\partial u_1^m\partial u_2^n}(u_1, u_2)\bigg|_{\substack{u_1=-k_1/\Delta\\ u_2=-k_2/\Delta}} \tag{88}$$

where

$$G_{\nu_1,\nu_2}(u_1, u_2) \triangleq \frac{\sin(\pi\Delta u_1)}{\pi\Delta u_1}\frac{\sin(\pi\Delta u_2)}{\pi\Delta u_2} \cdot P_{\nu_1,\nu_2}(u_1, u_2). \tag{89}$$

Now if

$$\frac{\partial^{m+n}G_{\nu_1,\nu_2}}{\partial u_1^m\partial u_2^n}(u_1, u_2)\bigg|_{\substack{u_1=k_1/\Delta\\ u_2=k_2/\Delta}} = 0$$

for all integers k_1, k_2 with $(k_1, k_2) \neq (0, 0)$ (90)

then

$$E[\varepsilon_1^m\varepsilon_2^n] = \left(\frac{j}{2\pi}\right)^{m+n}\frac{\partial^{m+n}G_{\nu_1,\nu_2}}{\partial u_1^m\partial u_2^n}(u_1, u_2)\bigg|_{\substack{u_1=0\\ u_2=0}} \tag{91}$$

which no longer depends on the joint pdf of the input. In this case we can write an expression analogous to Eq. (83), relating the joint moments of the total error to those of the dither:

$$E[\varepsilon_1^m\varepsilon_2^n] = \sum_{l_1=0}^{\lfloor m/2\rfloor}\sum_{l_2=0}^{\lfloor n/2\rfloor}\binom{m}{2l_1}\binom{n}{2l_2}\left(\frac{\Delta}{2}\right)^{2(l_1+l_2)}$$

$$\times \frac{E[\nu_1^{m-2l_1}\nu_2^{n-2l_2}]}{(2l_1 + 1)(2l_2 + 1)}. \tag{92}$$

Note that if ν_1 and ν_2 are statistically independent, and each satisfies Eq. (78), then ε_1^m and ε_2^n are uncorrelated (that is, $E[\varepsilon_1^m\varepsilon_2^n] = E[\varepsilon_1^m]E[\varepsilon_2^n]$) and Eq. (90) is automatically satisfied. Furthermore, if the dither represents a stationary random process, $E[\nu_1^m\nu_2^n] = E[\nu^m]E[\nu^n]$. In particular, for $m = n = 1$, using Eq. (92) we see that under these conditions

$$E[\varepsilon_1\varepsilon_2](k) = \begin{cases} E[\nu^2] + \dfrac{\Delta^2}{12}, & k = 0 \\ E^2[\nu], & \text{otherwise} \end{cases} \tag{93}$$

so that

$$\mathrm{PSD}_\varepsilon(f) = \mathrm{PSD}_\nu(f) + \frac{\Delta^2 T}{6}. \tag{94}$$

Thus the power spectrum of the total error is white since the dither spectrum is white. This will be the case in most practical situations where proper nonsubtractive dither is used. The possibility of using nonwhite dithers is explored in detail in [3], which shows that, under certain conditions, some such dithers can satisfy the requirements of Theorem 6 while yielding a total error spectrum which is the sum of the dither spectrum and a white-noise component.

5.2 Statistics of the System Output

5.2.1 First-Order Statistics

We now turn our attention to the output of the system. Applying the same reasoning used to determine the cpdf of the total error in Sec. 5.1.1, we find that the

cpdf of the output is given by

$$p_{y|x}(y, x) = [\Delta\Pi_\Delta * p_v](y - x) \cdot W_\Delta(y) . \quad (95)$$

Hence,

$$p_y(y) = \int_{-\infty}^{\infty} p_{y|x}(y, x)p_x(x)\,dx$$

$$= [\Delta\Pi_\Delta * p_v * p_x](y) \cdot W_\Delta(y) . \quad (96)$$

Moving to the Fourier transform domain,

$$P_y(u) = [G_v(u)P_x(u)] * W_{\frac{1}{\Delta}}(u)$$

$$= \sum_{k=-\infty}^{\infty} G_v\left(u - \frac{k}{\Delta}\right)P_x\left(u - \frac{k}{\Delta}\right) \quad (97)$$

and hence

$$E[y^m] = \sum_{k=-\infty}^{\infty} \sum_{r=0}^{m} \binom{m}{r}\left[\left(\frac{j}{2\pi}\right)^r \frac{d^r G_v}{du^r}(u)\right]$$

$$\cdot \left[\left(\frac{j}{2\pi}\right)^{m-r} \frac{d^{m-r}P_x}{du^{m-r}}(u)\right]\Bigg|_{u=k/\Delta} . \quad (98)$$

Now if the first m derivatives of $G_v(u)$ vanish at all nonzero multiples of $1/\Delta$, then Eq. (98) reduces to

$$E[y^m] = \sum_{r=0}^{m} \binom{m}{r} E[\varepsilon^r]E[x^{m-r}] \quad (99)$$

where the expectation values of the total error are given in terms of the expectation values of the dither by Eq. (83). By direct differentiation of $G_v(u)$, this condition is easily shown to be equivalent to the condition of Theorem 7 with $M = m$.

5.2.2 Second-Order Statistics

Proceeding in the usual fashion, we find that the joint moments of output values y_1 and y_2, separated in time by $\tau \neq 0$, are given by

$$E[y_1^m y_2^n] = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \sum_{r=0}^{m} \sum_{s=0}^{n} \binom{m}{r}\binom{n}{s}\left[\left(\frac{j}{2\pi}\right)^{r+s} \frac{\partial^{r+s}G_{v_1,v_2}}{\partial u_1^r \partial u_2^s}(u_1, u_2)\right]$$

$$\times \left[\left(\frac{j}{2\pi}\right)^{m+n-r-s} \frac{\partial^{m+n-r-s}P_{x_1,x_2}}{\partial u_1^{m-r}\partial u_2^{n-s}}(u_1, u_2)\right]\Bigg|_{\substack{u_1=k_1/\Delta \\ u_2=k_2/\Delta}} . \quad (100)$$

If the indicated partial derivatives of G_{v_1,v_2} are zero for all integers k_1, k_2 with $(k_1, k_2) \neq (0, 0)$ and for $r = 1, 2, \ldots, m$ and $s = 1, 2, \ldots, n$, then Eq. (100) reduces to

$$E[y_1^m y_2^n] = \sum_{r=0}^{m} \sum_{s=0}^{n} \binom{m}{r}\binom{n}{s} E[\varepsilon_1^r \varepsilon_2^s]E[x_1^{m-r}x_2^{n-s}] , \quad (101)$$

where the joint moments of the total error are given in terms of those of the dither by Eq. (92). In particular, note that if these conditions are satisfied for $m = n = 1$, then

$$E[y_1y_2] = E[x_1x_2] + E[\varepsilon_1\varepsilon_2] \quad (102)$$

so that, substituting Eqs. (83), (92), and (99), we have

$$E[y_1y_2](k) =$$

$$\begin{cases} E[x^2] + 2E[x]E[v] + E[v^2] + \dfrac{\Delta^2}{12}, & k = 0 \\ E[x_1x_2] + E[v_1v_2], & \text{otherwise.} \end{cases} \quad (103)$$

Hence, under these conditions, if the dither signal has zero mean,

$$\text{PSD}_y(f) = \text{PSD}_x(f) + \text{PSD}_v(f) + \frac{\Delta^2 T}{6} \quad (104)$$

so that the spectrum of the output is the sum of the input and dither spectra, apart from a white-noise component contributed by the $k = 0$ term in Eq. (103).

5.3 Properties of Practical Dither Signals

Theorem 8 A nonsubtractive dither signal generated by the summation of n statistically independent zero-mean rectangular-pdf random processes, each of 1-LSB peak-to-peak amplitude, renders the first n moments of the total error independent of the system input, and results, for $n \geq 2$, in a total error power of $(n + 1)\Delta^2/12$.

This must be the case since the addition of n such random processes convolves their pdf's, hence multiplying their cf's and yielding

$$G_v(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^{n+1} , \quad (105)$$

the first n derivatives of which will consist entirely of terms containing nonzero powers of $\sin(\pi\Delta u)/(\pi\Delta u)$. 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.

Furthermore, it is important to note that using rectangular-pdf noises of peak-to-peak amplitude not equal to 1 LSB (or, rather, not equal to an integral number of LSB) 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 [1]).

We proceed to examine two important examples of nonsubtractive dither pdf's. First, consider a system using dither with a simple rectangular (that is, uniform) pdf of 1-LSB peak-to-peak amplitude

$$p_\nu(v) = \Pi_\Delta(v) \tag{106}$$

for which

$$G_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^2 . \tag{107}$$

The first three derivatives of this function, and the corresponding moments as a function of the input, are plotted in Fig. 8. The first derivative clearly satisfies the condition of going to zero at the regularly spaced intervals stipulated by Eq. (78), while the second derivative and higher derivatives do not. This indicates that the first moment of the error signal is independent of the input, but that its variance and higher moments remain dependent. These conclusions are borne out by

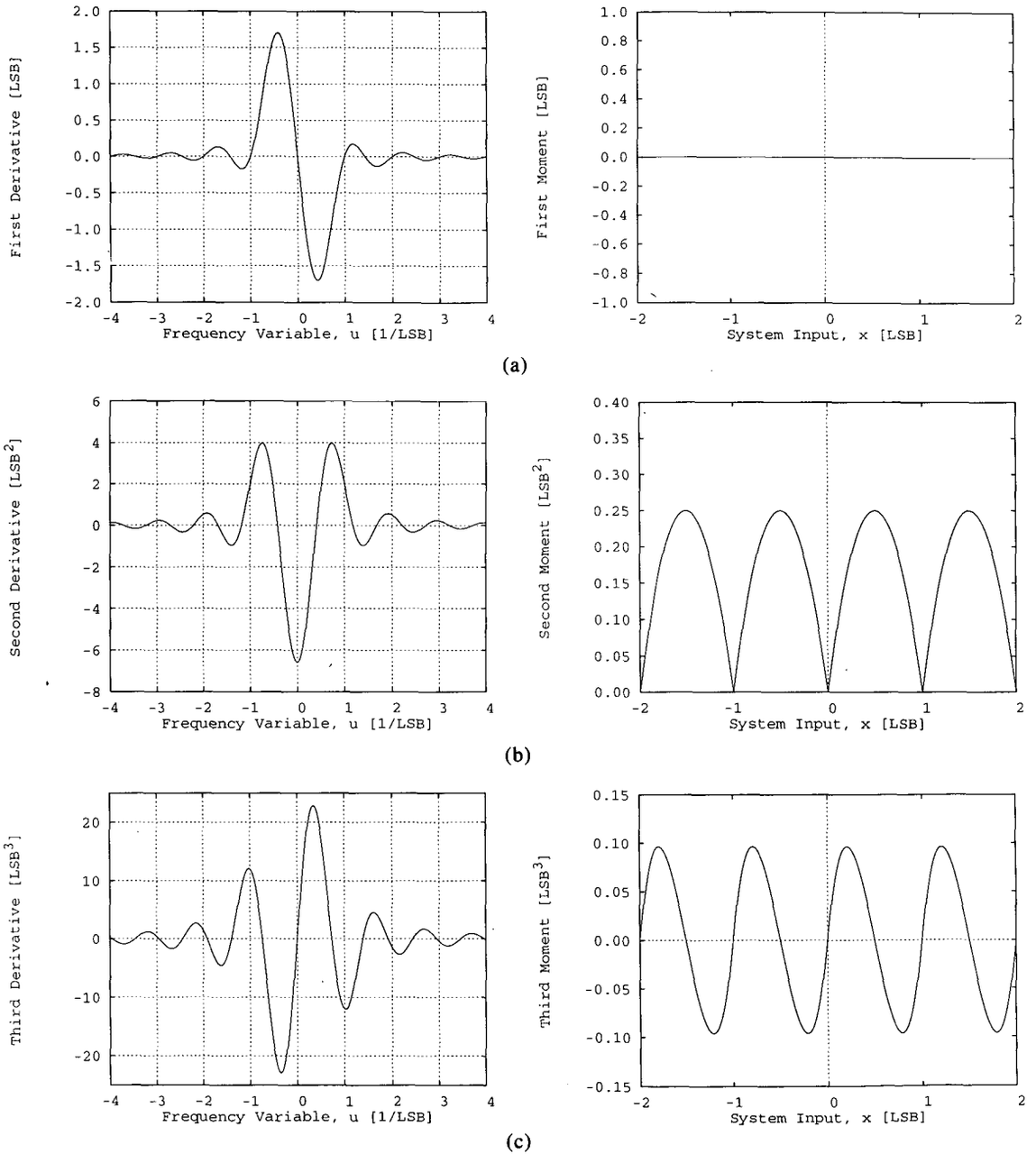


Fig. 8. Derivatives of $G_\nu(u)$ (left) and conditional moments of error (right) for quantizer using rectangular-pdf dither of 1-LSB peak-to-peak amplitude. (a) dG_ν/du and $E[\varepsilon|x]$ (in units of Δ). (b) d^2G_ν/du^2 and $E[\varepsilon^2|x]$ (in units of Δ^2). (c) d^3G_ν/du^3 and $E[\varepsilon^3|x]$ (in units of Δ^3). Frequency variable u is plotted in units of $1/\Delta$ and input x in units of Δ.

the plots of the moments themselves. 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, so that the noise power in the signal varies with the system input. This is sometimes referred to as noise modulation and is audibly undesirable.

Now consider a triangular-pdf dither of 2-LSB peak-to-peak amplitude resulting from the sum of two independent rectangular-pdf noises ν_1 and ν_2, each of 1-LSB peak-to-peak amplitude:

$$p_\nu(\nu) = [p_{\nu_1} * p_{\nu_2}](\nu) = [\Pi_\Delta * \Pi_\Delta](\nu) \ . \quad (108)$$

In a system using this kind of dither, $p_{\varepsilon|x}(\varepsilon, x)$ involves the convolution of three rectangular window functions, so that $G_\nu(u)$ is given by

$$G_\nu(u) = \left[\frac{\sin(\pi\Delta u)}{\pi\Delta u}\right]^3 \ . \quad (109)$$

The first three derivatives of this function, and the corresponding moments as a function of the input, are plotted in Fig. 9. The first *and* second derivatives of this function go to zero at the required places, so this dither renders both the first and second moments of the total error independent of x. The second moment of the total error is a constant $\Delta^2/4$ for all inputs, in agree-

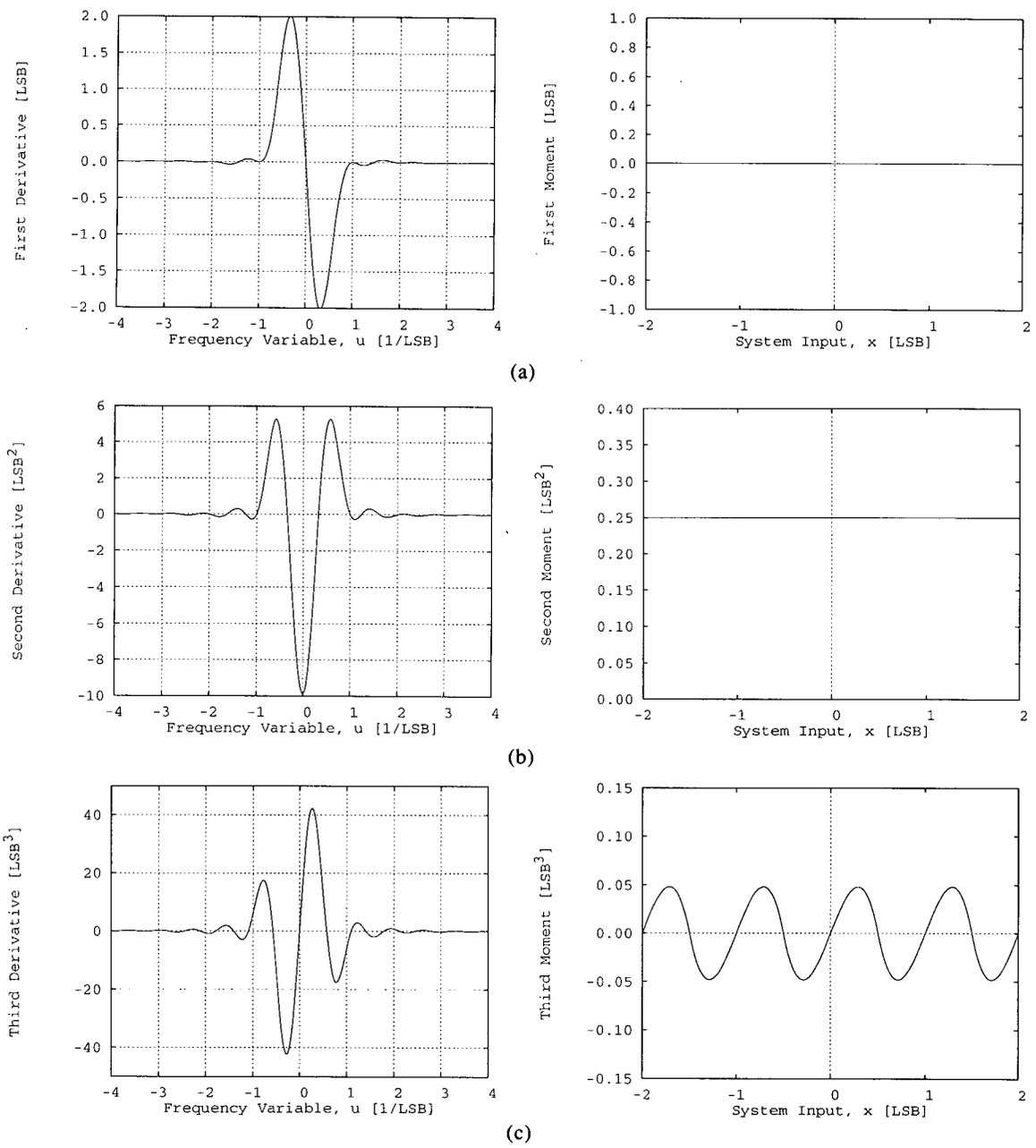


Fig. 9. Derivatives of $G_\nu(u)$ (left) and conditional moments of error (right) for quantizer using triangular-pdf dither of 2-LSB peak-to-peak amplitude. (a) dG_ν/du and $E[\varepsilon|x]$ (in units of Δ). (b) d^2G_ν/du^2 and $E[\varepsilon^2|x]$ (in units of Δ^2). (c) d^3G_ν/du^3 and $E[\varepsilon^3|x]$ (in units of Δ^3). Frequency variable u is plotted in units of $1/\Delta$ and input x in units of Δ.

ment with Eq. (82). Higher derivatives of $G_v(u)$ do not meet the required conditions, so that higher moments of the error remain dependent on the input.

It can be shown [3] that triangular-pdf dither of 2-LSB peak-to-peak amplitude is unique and optimal in the sense that it renders the first and second moments of the total error input independent, while minimizing the second moment. That is, when used in a nonsubtractively dithered quantizing system, this dither incurs the least possible increase in the rms noise level of any dither which eliminates input-dependent distortion and noise modulation.

5.4 Summary of Nonsubtractive Dither

The results of greatest practical importance concerning nonsubtractively dithered quantizing systems are as follows:

1) Nonsubtractive dithering, unlike subtractive dithering, cannot render the total error independent of the system input. It *can* render any desired conditional moments of the total error independent provided that certain conditions on the cf of the dither are met (see Theorem 6).

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. In particular, it can render the power spectrum of the total error signal white [see discussion leading to Eq. (94)].

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 Sec. 5.2). This includes joint moments of system inputs separated in time, so that 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 Eq. (82)].

5) Triangular-pdf dither of 2-LSB peak-to-peak amplitude incurs the least increase in the rms total error level of any nonsubtractive dither which eliminates input-dependent distortion and noise modulation.

Fig. 10 shows the results of a computer-simulated quantization operation performed upon a 1-kHz sine wave of 4.0-LSB peak-to-peak amplitude and using the aforementioned triangular-pdf dither. Shown are the system input and output, the total error, and the power spectrum of the system output. Note that vestiges of the input signal are clearly visible in the total error waveform, indicating that the two signals are *not* statistically independent. Also, the system output does not resemble a sine wave plus an independent additive noise. Surprising as it may seem, listening experiments (see Sec. 6) show that the total error signal of Fig. 10(c) sounds like a constant white noise, independent of the nature of the input signal (with which it is indeed uncorrelated), and that Fig. 10(b) sounds identical to a noisy sine wave. Furthermore, the power spectrum of the system output, in Fig. 10(d), exhibits no distortion components and indicates that the total error *is* spectrally

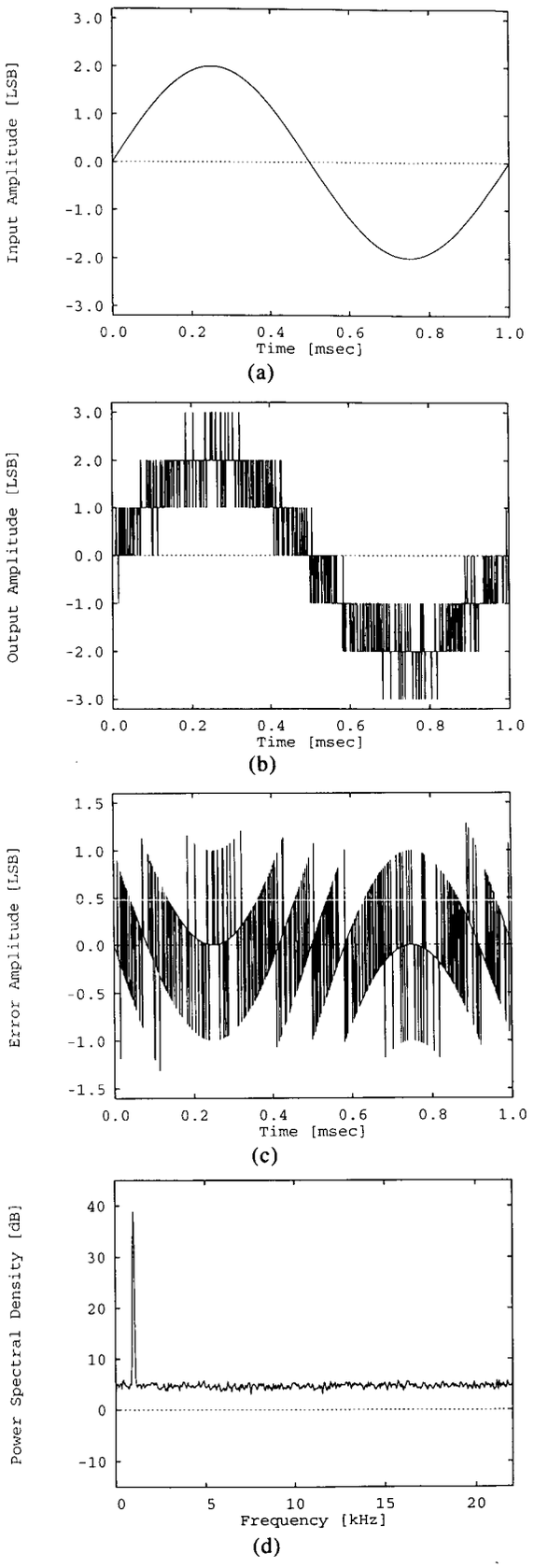


Fig. 10. Results from computer-simulated quantization of 1-kHz sine wave of 4.0-LSB peak-to-peak amplitude using triangular-pdf nonsubtractive dither of 2-LSB peak-to-peak amplitude. (a) System input signal. (b) System output signal. (c) Resulting total error signal. (d) Power spectrum of system output signal (as estimated from sixty 50% overlapping Hann-windowed 512-point time records with assumed sampling frequency of 44.1 kHz; 0 dB represents a power spectral density of $\Delta^2 T/6$, T being the sampling period).

white. These results should be compared with those in Figs. 4 and 6, which illustrate the results of quantizing a sine wave using undithered and subtractively dithered systems, respectively. In particular, it should be noted that the noise floor in Fig. 10(d) is up by 4.8 dB relative to that of Fig. 6(d) due to the tripling of the noise spectral density, in accordance with Eq. (104).

Some comment is required concerning the special nature of requantization. In a purely digital system, random processes exhibiting the continuous pdf's described in this section are not, strictly speaking, available since not all real numbers are representable using a finite number of binary digits. In fact, digital dither pdf's of necessity resemble discretized or "sampled" versions of the continuous pdf's (rectangular, triangular, etc.) described. It is not immediately obvious that such dithers will retain the desirable properties of their analog counterparts with respect to rendering total error moments independent of the system input. It is rigorously proven in [3] that such dithers *do indeed* retain these properties, and empirical evidence (and a less general proof) corroborating this conclusion is presented in [1].

6 CONCLUSIONS

For audio signal processing purposes there seems to be little point in rendering any moments of the total error 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 psychoacoustic tests conducted by the authors and others [18], [23]. These tests involved listening to a large variety of signals (such as sinusoids, sinusoidal chirps, slow ramps, various periodically switched inputs, piano and orchestral music) which had been requantized very coarsely (to 8 bits from 16) in order to render the requantization error essentially independent of low-level nonlinearities in the digital-to-analog conversion system through which the listening took place. In addition, the corresponding total error signals (output minus input) were used in listening tests in order to check for any audible dependences on the input. Using undithered quantizers resulted in clearly audible distortion and noise modulation in the output and error signals. A subtractively dithered quantizing system using rectangular-pdf dither of 1-LSB peak-to-peak amplitude was found to eliminate all audible input dependences in the error signal, which was discovered to be audibly equivalent to a steady white noise. A nonsubtractively dithered quantizing system using the same dither eliminated all distortion, but the residual noise level was found to vary audibly in an input-dependent fashion. When triangular-pdf dither of 2-LSB peak-to-peak amplitude was used, no instance was found in which the error was audibly distinguishable from a steady white noise entirely unrelated to the input, although the level of this noise was somewhat higher than that observed in the subtractively dithered system. Admittedly, these tests were informal, and there remains a need for formal psycoacoustic tests of this sort involving many participants under carefully controlled conditions.

We recommend the use of nonsubtractive, triangular-pdf dither of 2-LSB peak-to-peak amplitude for most audio applications requiring multibit quantization or requantization operations, since this type of dither renders the first and second moments of the error signal constant with respect to the input, while incurring the minimum increase in error variance. This kind of dither is easy to generate for digital requantization purposes by simply summing two independent rectangular-pdf pseudo-random processes, which are easily generated by linear congruential algorithms [23], [26]. The resulting digital dither can be used to feed a digital-to-analog converter for analog dithering applications.[13]

This paper has attempted to underscore the close mathematical relationship between time sampling and amplitude quantization. In closing, we emphasize that appropriate dithering prior to (re)quantization is as fundamental as appropriate antialias filtering prior to sampling—both serve to eliminate classes of signal-dependent errors.

7 ACKNOWLEDGMENTS

Stanley P. Lipshitz and John Vanderkooy have been supported by operating grants from the Natural Sciences and Engineering Research Council of Canada.

8 REFERENCES

[1] S. P. Lipshitz and J. Vanderkooy, "Digital Dither," presented at the 81st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 34, p. 1030 (1986 Dec.), preprint 2412.

[2] J. Vanderkooy and S. P. Lipshitz, "Dither in Digital Audio," *J. Audio Eng. Soc.*, vol. 35, pp. 966–975 (1987 Dec.).

[3] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Non-Subtractive Dither," *IEEE Trans. Signal Process.*, submitted for publication.

[4] R. M. Gray and T. G. Stockham, "Dithered Quantizers," *IEEE Trans. Inform. Theory*, submitted for publication.

[5] E. T. Whittaker, "On the Functions which Are Represented by the Expansions of the Interpolation-Theory," *Proc. R. Soc. Edinburgh*, vol. 35, pp. 181–194 (1915).

[6] B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," Ph.D. thesis, Dept. of Elect. Eng., M.I.T., Cambridge, MA (1956 June).

[7] B. Widrow, "A Study of Rough Amplitude

[13] It should be noted, however, that many analog signals and digital-to-analog conversion systems exhibit a Gaussian noise component which is of large enough amplitude to act as a satisfactory inherent dither without the requirement of an explicit dithering operation [8], [2], [23].

Quantization by Means of Nyquist Sampling Theory," *IRE Trans. Circuit Theory*, vol. PGCT-3, pp. 266–276 (1956 Dec.).

[8] B. Widrow, "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems," *Trans. AIEE*, pt. II, vol. 79, pp. 555–568 (1961 Jan.).

[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 Process.*, vol. ASSP-25, pp. 442–448 (1977 Oct.).

[10] J. S. Lim and A. V. Oppenheim, *Advanced Topics in Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1988).

[11] L. G. Roberts, "Picture Coding Using Pseudo-Random Noise," *IRE Trans. Inform. Theory*, vol. IT-8, pp. 145–154 (1962 Feb.).

[12] L. Schuchman, "Dither Signals and Their Effect on Quantization Noise," *IEEE Trans. Commun. Technol.*, vol. COM-12, pp. 162–165 (1964 Dec.).

[13] D. T. Sherwood, "Some Theorems on Quantization and an Example Using Dither," in *Conf. Rec., 19th Asilomar Conf. on Circuits, Systems, and Computers* (Pacific Grove, CA, 1985 Nov.).

[14] J. N. Wright, unpublished manuscripts (1979 June–Aug.).

[15] J. N. Wright, private communication (1991 Apr.).

[16] S. P. Lipshitz, R. A. Wannamaker, J. Vanderkooy, and J. N. Wright, "Non-Subtractive Dither," presented at the 1991 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY (1991 Oct. 20–23), paper 6.2.

[17] T. G. Stockham, private communication (1988).

[18] L. K. Brinton, "Nonsubtractive Dither," M.Sc. thesis, Dept. of Elec. Eng., Univ. of Utah, Salt Lake City, UT (1984 Aug.).

[19] 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 (1984 Mar.); correction, *ibid.* (*Letters to the Editor*), p. 889 (1984 Nov.).

[20] J. Vanderkooy and S. P. Lipshitz, "Digital Dither: Signal Processing with Resolution Far Below the Least Significant Bit," in *Proc. AES 7th International Conf.: Audio in Digital Times* (Toronto, Ont., Canada, 1989 May), pp. 87–96.

[21] S. P. Lipshitz and J. Vanderkooy, "High-Pass Dither," presented at the 4th Regional Convention of the Audio Engineering Society, Tokyo (1989 June); in *Collected Preprints* (AES Japan Section, Tokyo, 1989), pp. 72–75.

[22] R. A. Wannamaker, S. P. Lipshitz, and J. Vanderkooy, "Dithering to Eliminate Quantization Distortion," in *Proc. Annual Meeting Can. Acoustical Assoc.* (Halifax, N.S., Canada, 1989 Oct.), pp. 78–86.

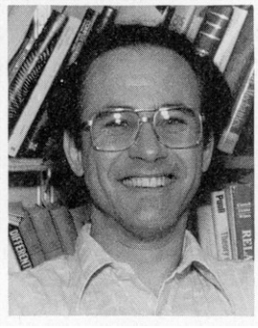
[23] R. A. Wannamaker, "Dither and Noise Shaping in Audio Applications," M.Sc. thesis, Dept. of Physics, Univ. of Waterloo, Waterloo, Ont., Canada (1990 Dec.).

[24] R. M. Gray, private communication (1991 Apr.).

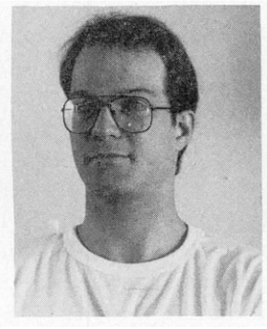
[25] N. S. Jayant and P. Noll, *Digital Coding of Waveforms* (Prentice Hall, Englewood Cliffs, NJ, 1984).

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

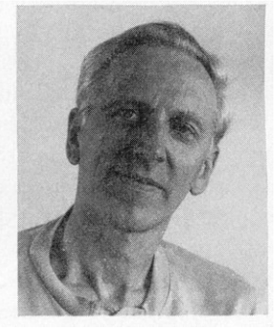
THE AUTHORS



S. P. Lipshitz



R. Wannamaker



J. Vanderkooy

Stanley P. Lipshitz is a professor in both the departments of Applied Mathematics and Physics at the University of Waterloo in Waterloo, Ontario, Canada. There, in addition to the normal teaching associated with a university position, he conducts his own and supervises graduate student research in audio and electroacoustics as a member of the university's Audio Research Group. Prior to joining the faculty of the University of Waterloo in 1970, he studied in South Africa, the country of his birth, where he received his Bachelors (1964), Masters (1966) and Ph.D. (1970) degrees in applied mathematics and physics. He is now a Canadian citizen.

Dr. Lipshitz is a fellow of the AES. He served as a governor of the Society for the periods 1984–1986 and 1987–1992, and as its president for the year 1988–1989. Other society memberships include the IEEE, the Acoustical Society of America, and the Canadian Acoustical Association.

He has presented numerous technical papers at AES

conventions, both in North America and overseas, on a wide range of topics including amplifier design, psychoacoustic, loudspeaker crossover design, electroacoustic transducer measurement, acoustics, and digital signal processing for audio. He has participated in many educational workshops and seminars on audio topics including loudspeaker measurement, stereo microphone techniques, and the fundamentals of digital audio.

Dr. Lipshitz's publications have appeared frequently in the AES *Journal* and elsewhere. His current research interests include transducer design and measurement, digital signal processing for audio, and the characterization and design of surround-sound systems. He has consulted for a number of companies on audio-related questions.

•

Robert Wannamaker was born in 1967 in Ontario, Canada. In 1988 he received the degree of B.Sc.E. in engineering physics from Queen's University in Kingston, Ontario. He was granted the degree of M.Sc. in physics by the University of Waterloo in 1990 for research on dither and noise-shaping conducted under the aegis of the Audio Research Group. Since graduating, he has continued to work with the ARG as a research assistant. He plays a pretty mean piano and some guitar as well.

•

John Vanderkooy was born in The Netherlands in 1941, but received all of his education in Canada, with a B.Eng. degree in engineering physics in 1963 and a Ph.D. in physics in 1967, both from McMaster University in Hamilton, Ontario. For some years he followed his doctoral interests in low-temperature physics of metals at the University of Waterloo, where he is currently a professor of physics. However, since the late 1970s, his research interests have been mainly in audio and electroacoustics.

A fellow of the AES and a member of the IEEE, Dr. Vanderkooy has contributed a variety of papers at conventions and to the *Journal*. Together with his collegue Stanley Lipshitz and a number of graduate and undergraduate students, they form the Audio Research Group at the University of Waterloo.

Dr. Vanderkooy's current interests are digital audio signal processing, measurement of transfer functions with maximum-length sequences, transducers, diffraction of loudspeaker cabinet edges, and most recently sub-surface analysis techniques using maxiumum-length sequences.