Dithering for Floating-Point Number Representation

Rezső Dunay[*], István Kollár[†], and Bernard Widrow[‡]

March 6, 1998

Abstract

Dithering is widely used for decreasing the bias in fixed-point quantization and rounding. Since floating-point digital signal processors (DSP's) and floating-point arithmetic are becoming widely used, it is timely to investigate the necessity and possibilities of dithering for floating-point numbers. The paper introduces a simple model of dithers for floating-point, and discusses its practical use.

Keywords: Dither, quantization, DSP, digital signal processor, floating-point, roundoff.

1 Introduction

Dithering is maybe the most popular method to decrease quantization distortion. It is well-known and also often in practice to add dither signals to data before fixed-point quantization. When the quantizer characteristic is precise, as it is in fixed-point arithmetic operations in computers and DSP's, usually a uniform dither in $(-q/2, q/2)$ or a triangular-shaped one in $(-q, q)$ is used.

With the spreading of floating-point DSP's and IEEE compatible computers, floating-point number representation is more and more widely used. Its quantization error is usually very small, but sometimes it is not negligible. Then, it is justified to add some dither. Let us assume for example that IEEE single precision

[*] Department of Measurement and Information Systems, Technical University of Budapest, Hungary, H-1521, phone: + 36 1 463-4116, email: dunay@mmt.bme.hu

[†] Department of Measurement and Information Systems, Technical University of Budapest, Hungary, H-1521, phone: + 36 1 463-1774, email: kollar@mmt.bme.hu

[‡] ISL, Department of Electrical Engineering, Stanford University, Stanford, CA 94305-4055, USA, phone: (650) 723-4949, fax: (650) 723-1783

is used. This means that we have $p = 24$ bits for the mantissa, and 8 bits for the exponent. The mantissa is signed, with suppressed leading bit, that is, usual numbers are normalized to have a leading bit equal to one, and since this is usually so, this bit is not stored. Therefore, the maximum relative quantization error is about $2^{-23}/2$ for the smallest mantissas, and about 2^{-23} for the largest ones. The latter number equals about $1.19 \cdot 10^{-7}$. This is still a small number, but when in the calculations the difference of numbers close to each other is calculated (cancellation), the relative error of the result can increase significantly.

Here two questions arise:

1. What are the properties of the proper dither for floating-point?
2. How can we apply the dither before arithmetic operations without immediately eliminating it during arithmetic addition of the number and the dither?

This paper explains the problem, and discusses the following answers.

- The dither for floating-point numbers is preferably a uniform or a triangular-shaped one FOR THE MANTISSA, with the same exponent as of the numbers. Some correction can be introduced to cope which the changing exponent.
- For proper dithering, we need more bits than the floating-point arithmetic usually provides. Therefore, we need one of the following special solutions:
  - special software solution to virtually increase the bit length
  - utilization of the extended precision of the accumulator if it is such a one
  - modification of the existing hardware.

The paper discusses how these principles are applied, and makes suggestions for future hardware design.

2 Basic Properties of Dithering

Dithering can be discussed in analogy to anti-aliasing in sampling. When a signal does not meet the conditions of the sampling theorem, it cannot be sampled in an error-free manner. In such cases we apply an *anti-alias* filter which restricts the bandwidth of the signal to the appropriate band. The filtered signal is then perfect for sampling theory.

In quantization, *bandwidth* is measured in the characteristic function (CF) domain. For proper quantization, the bandwidth of the characteristic function must

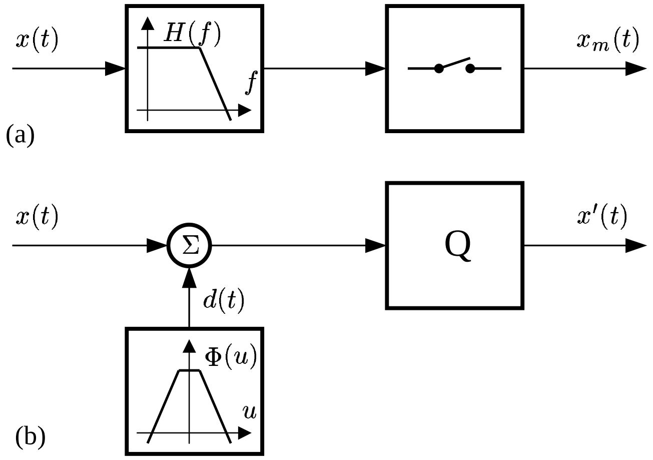

Figure 1: Sampling and quantization: (a) sampling; (b) quantization.

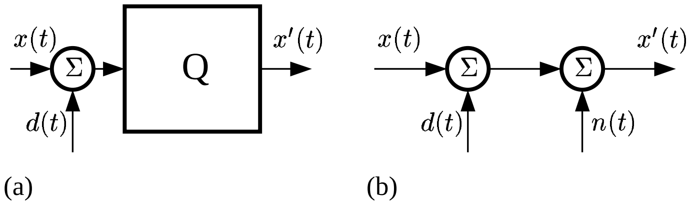

Figure 2: Noise model of dither: (a) adding a dither to the input signal; (b) representing a dithered quantizer via the noise model.

be limited. For this, we need to introduce an operation which restricts the width of the characteristic function before quantization. Since the CF's of independent input signals multiply, the best way is to add an *independent signal* to the input one in order to meet conditions of the quantizing theorems (see Fig. 1). Such an additional input signal is called *dither*.

In special cases, as e.g. for floating-point quantization, it may even depend on the input signal, but in most situations d is generated to be independent of the input signal x. Whenever a signal is properly band-limited, it can be perfectly represented by its samples. When a proper dither signal is applied, the quantization theorem is satisfied at the quantizer input, that is, the noise model can be applied (Fig. 2, see [19]), and the moments of the input signal $x + d$ can be perfectly expressed by moments of the output signal, via Sheppard's corrections. It is also possible that for some dither types, only a few selected Sheppard corrections can be applied, like the first one for uniform dither between $(-q/2, q/2)$.

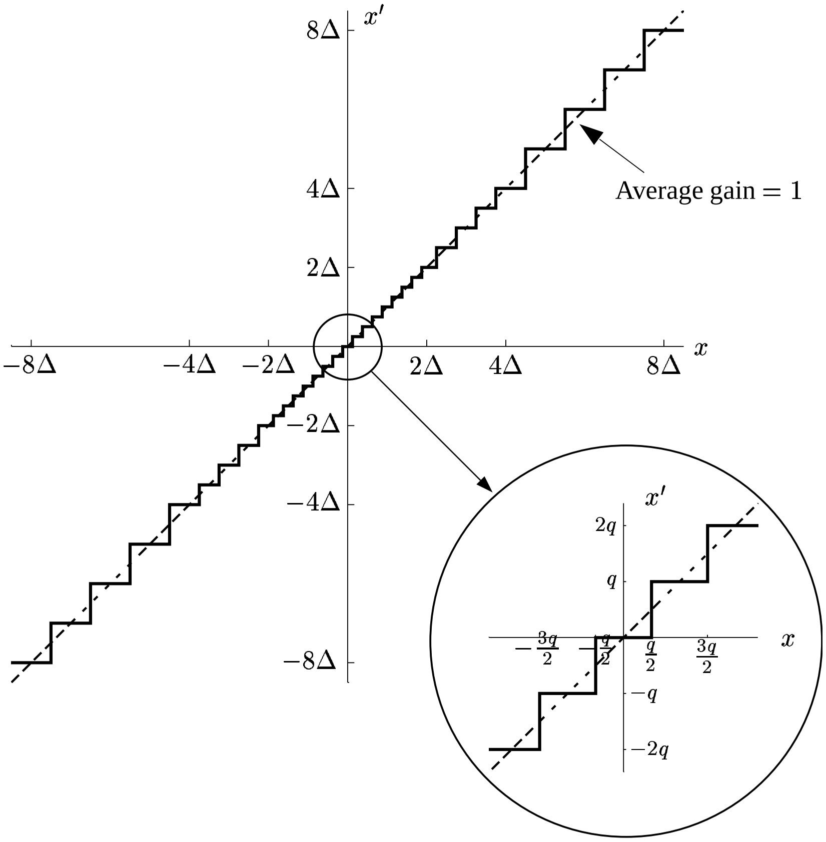

Figure 3: Input-output staircase function for a floating-point quantizer with a 3-bit mantissa.

3 Relations of Fixed-Point and Floating-Point Quantization

In order to understand how these two quantizers are related, let us briefly discuss a model which establishes their relationship.

An example for floating-point quantization is shown in Fig. 3.

The quantum size increases gradually with increasing signal amplitudes. Therefore, uniform quantization theory cannot be directly applied. A possibility to establish a relationship with uniform quantization is to use the so-called compandor concept as in [19]: floating-point quantization is transformed to a fixed-point one by means of a compressor before, and an expandor after a fixed-point quantizer (Fig. 4b). Both nonlinear elements are stepwise linear (see Fig. 5), approximating a logarithmic and an exponential characteristics, respectively.

By this, we have transformed the quantization operation itself to a uniform

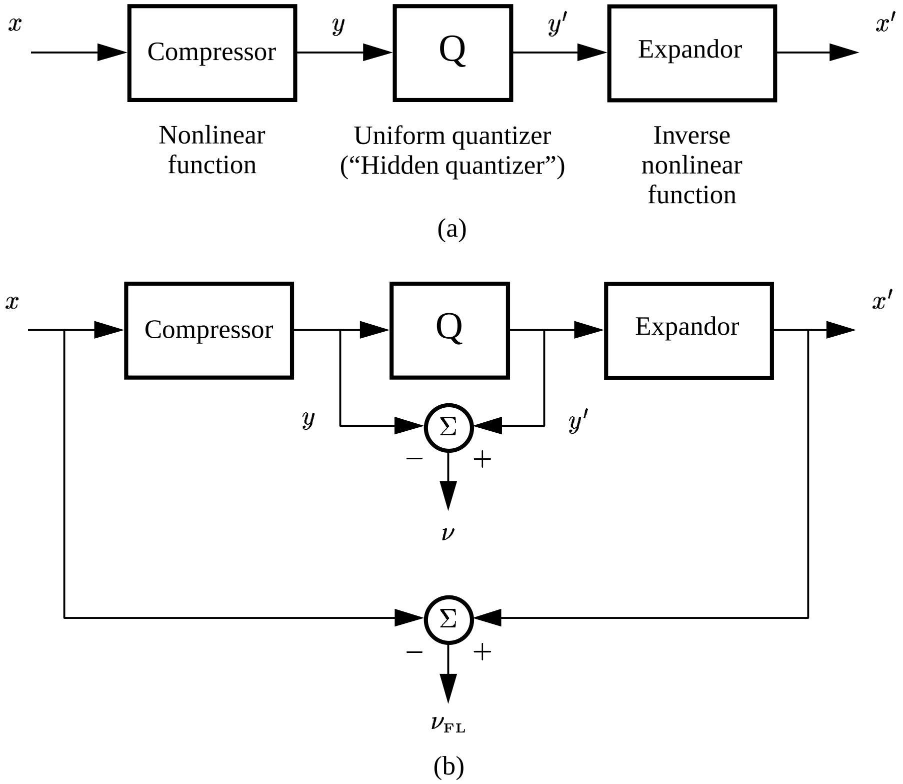

Figure 4: A model of a floating-point quantizer: (a) block diagram; (b) definition of quantization noises.

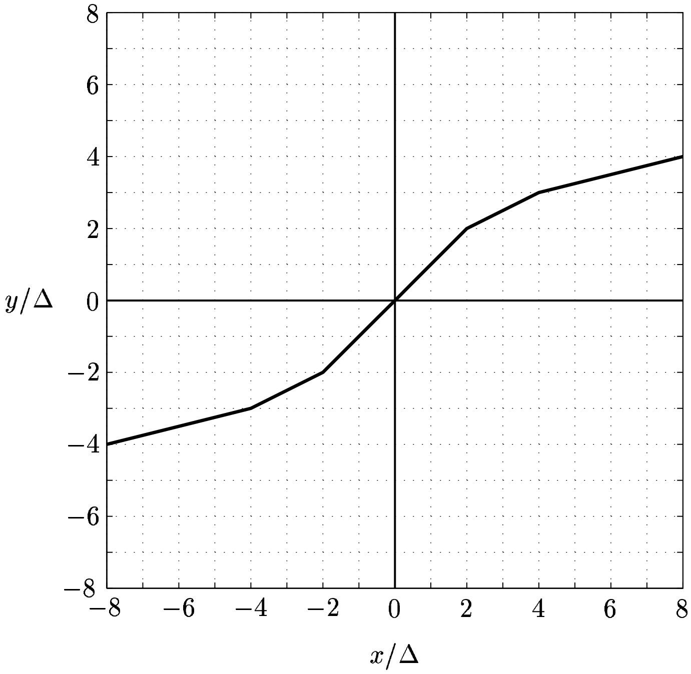

Figure 5: The compressor's input-output characteristic

one: therefore, we can try to apply a certain dither to the "hidden quantizer", in order to diminish or eliminate quantization bias. Since the compressor essentially removes the exponent from the input, we can say that in this concept, we apply a dither signal to the *mantissa or significand of the input*.

While this concept seems to be very logical, there is one flaw. There is no deviation from the uniform case as long as we operate on linear sections of both the compressor and the expandor. However, nonlinear effects appear when a corner-point is involved. Moreover, the corner-points of the expandor also change the statistical properties of the quantized signal. Therefore, when a predefined-form dither is applied at the input, this concept is only approximate. We can say that it is a good approximation if the number of mantissa bits is not very low (let us say, it is larger than 6-8), because it is quite unlikely that the signal is close to a corner-point. A precise theory has to be still developed.

If we want to treat the corner-points properly, a special solution can be introduced. Let us discuss first the case of a uniform dither. In the hidden quantizer, we would like to have a uniform dither of the same width everywhere. We can directly influence the shape of the dither distribution only *before* the input compressor. If we calculate the shape of the dither which is transformed by the compressor into a uniform one, we obtain a rule illustrated in Fig. 6.

When the "hidden" dither PDF around the signal value does not include a

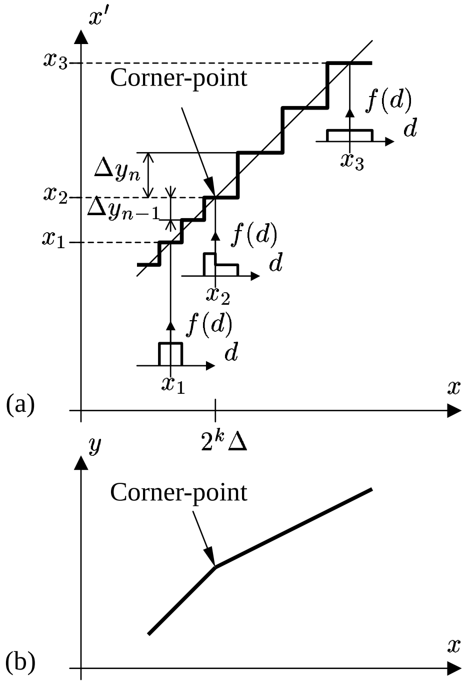

Figure 6: Generation of uniform hidden dither in a floating-point quantizer. (a) portion of the floating-point characteristic and the input dither distributions; (b) corresponding portion of the compressor.

corner-point of the compandor, the dither at the floating-point input is simply a uniform one. When however a corner-point is included, the distribution at the right side of the corner-point is stretched horizontally by a factor of 2. The compressor on the other hand will produce a uniform hidden dither by squeezing this portion by a factor of 0.5. Algorithmically, when the input signal is close to a corner-point so that the support of the PDF of the generated dither includes it (that is, when the signal is below a corner-point, and it is closer to it than $0.5\Delta y_{n-1}$, and when it is larger, then it is closer to it than $0.5\Delta y_n$, where Δy_{n-1} and Δy_n are the two neighboring quantum sizes, i.e. $\Delta y_{n-1} = 0.5\Delta y_n$) the rule is as follows.

- When the input value is just below a corner-point, a dither between

  $$(-\Delta y_{n-1}/2, \Delta y_{n-1}/2)$$

  is added, and if the result is now above the corner-point, the difference between the result and the corner-point is doubled.

- When the input value is just above a corner-point, a dither between

  $$(-\Delta y_n/2, \Delta y_n/2)$$

  is added, and if the result is now below the corner-point, the difference between the result and the corner-point is halved.

It is not certain from the above arguments however whether this dither will indeed eliminate the quantization distortion, since the effect of the corner-points is not clear. Let us discuss this question now.

First let us make a note. The first moment is unbiased in the hidden quantizer, because the output of the quantizer contains information about the position of the input signal everywhere: when it is moved in either direction, an increasing fraction of the PDF of the dither crosses the next quantization level in this direction, and the average of the quantized (binary) distribution increases accordingly. By the above described stretching, we also achieve that quantization level crossings occur at every position (except when the dither PDF around the input signal exactly falls between two neighboring quantization levels). This means that there is a chance that the quantized signal contains indeed proper information about the value of the input signal.

Let us consider now the signal value which causes the floating-point input dither to be at the position shown in the center of Fig. 6a. The output is constant, and exactly equals the input x_2. The bias is therefore zero. Let the value of the PDF at the left side be $f_l = 1/\Delta y_{n-1}$, and let us shift x by $+\mathrm{d}x$. We are now *above* the corner-point with $x + \mathrm{d}x$. Then, at the right side a new output amplitude level appears with probability $P_r = 1/\Delta y_n \cdot \mathrm{d}x$. Its position is at distance Δy_n above

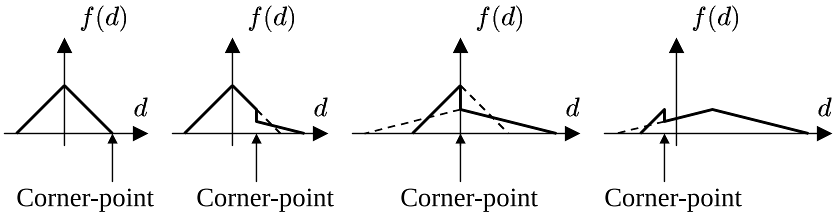

Figure 7: Triangular hidden dither transformed back to the floating-point input.

the previous level. Therefore, the change in the mean value is $P_r \cdot \Delta y_n = \mathrm{d}x$. The mean value changes by exactly the same amount as at the input. Since this the same argument can be applied for each input signal value, we have proved that the first moment (the expected value) is unbiased for all input amplitudes.

Further research has to be carried out concerning more complicated dither forms. We think that e.g. a properly stretched triangular dither (Fig. 7) can similarly diminish the already small correlation between the output signal and the quantization error.

We can also realize that subtraction of the dither from the output destroys finite bit length behavior (see the next section). Therefore, for floating-point, basically only non-subtractive dither can be applied (see e.g. [11]).

We finally mention that concerning the interrelation between the signal and the quantization noise, for large-scale signal variations the expandor can be approximated by an exponential function, therefore the additive relationship in the hidden quantizer becomes multiplicative at the output ([19]). Meanwhile, for relatively small signal variations, when we stay on the same segment of the quasi-linear compandor, the relationship remains linear. Therefore, no general rule can be established for relationship of x and ν_{FL}.

4 Suggested Solutions

Knowing what kind of dither has to be added, we have to implement this scheme. Here another difficulty arises. We should inject the dither after the arithmetic operation, but just before quantization. In other words, we should deal with the long-mantissa results, before re-quantizing it to the memory bit length. The practical problem is that this intermediate number is usually not available. Most arithmetic processors generate the results without providing access to the intermediate result before quantization. In such a case, we cannot properly add a dither. An idea would be to add dithers to both inputs before the arithmetic operation, but this is usually not usable, either: the necessary dither is at LSB level, and after any addition, the result is immediately quantized: we immediately lose the dither bits.

In order to overcome the above difficulty, we suggest the following solutions.

- If we cannot have direct access to the high-precision results, we may still be able to indirectly generate them. This means that we can virtually increase precision. For example, by multiplication, we make use of the fact that if a mantissa is represented by p bits, and we denote the upper half by HW, the lower half by LW, we can exactly calculate HW1*HW2, HW1*LW2 and LW1*HW2 with the available arithmetic processor, and by proper addition of these, we can represent the significand of the result on $2p$ bits. Then, we can properly add the dither to the lower bits, and quantize after that. This is a little tedious procedure, especially for more complicated operations, and sometimes, when the coprocessor can only yield an already quantized result, it is not even possible. Still, in most cases this seems to be a doable solution. The speed is acceptable, by utilizing the coprocessor arithmetic.

- In certain cases, like in the PC coprocessor, the accumulator bit length is higher than that of the memory. In a PC, it uses 64 mantissa bits instead of 53. When the result is generated in the accumulator, it has bits below the LSB level of the memory. This means that while the result is still in the accumulator, we can add the dither properly, and then move the data to the memory (during which re-quantization happens).

- If we realize the above difficulties, it is straightforward to speculate that the best way would be to add the dither *in hardware* at the right place. This implies redesign of the hardware which is not possible at this moment, but it can be incorporated in future designs. Therefore, an improved coprocessor hardware can be suggested which allows addition of a dither before quantization.

  While this is a very reasonable statement, we have to realize that even a pseudorandom dither makes the result slightly (pseudo)random. This means that without the dither being synchronized from outside – and we may not want to do synchronization because thus we may loose the advantage of low bias – this makes the result non-repeatable. This causes additional difficulties in the evaluation of algorithms and so on. Therefore, dithering must be selectable in a well designed hardware, and the possibility of synchronization of the pseudo random generator is also desirable.

5 Conclusions

The necessity, possibilities and difficulties of applying dither to floating-point numbers is discussed. Realization of this dither is an option for future hardware/software developments.

References

[1] L. K. Brinton, “Nonsubtractive dither,” M.S. thesis, University of Utah, Salt Lake City, Utah, 1984.

[2] P. Carbone et al., “Effect of additive dither on the resolution of ideal quantizers,” *IEEE Trans. on Instrumentation and Measurement*, vol. 43, no. 3, pp. 389–96, Jun 1994.

[3] I. De Lotto and G. E. Paglia, “Dithering improves A/D converter nonlinearity,” *IEEE Trans. on Instrumentation and Measurement*, vol. 35, no. 2, pp. 170–77, June 1986.

[4] G. G. Furman, “Improvement of quantized feedback systems by means of external dither,” M.S. thesis, MIT, 1957.

[5] R. M. Gray and T. G. Stockham, Jr, “Dithered quantizers,” *IEEE Trans. on Information Theory*, vol. 39, no. 3, pp. 805–12, May 1993.

[6] N. S. Jayant and L. R. Rabiner, “The application of dither to the quantization of speech signals,” *The Bell System Technical Journal*, vol. 51, no. 6, pp. 1293–304, July-August 1972.

[7] I. Kollár, “Quantization noise,” Doct. Sci. Thesis, Hungarian Academy of Sciences, Budapest, 1997.

[8] I. Kollár, “Statistical theory of quantization: Results and limits,” *Periodica Polytechnica Ser. Electrical Engineering*, vol. 28, no. 2-3, pp. 173–190, n/a 1984.

[9] J. O. Limb, “Design of dither waveforms for quantized visual signals,” *The Bell System Technical Journal*, vol. 48, no. 7, pp. 2555–82, Sept 1969.

[10] S. P. Lipshitz and J. Vanderkooy, “Digital dither,” in *81st Convention of the Audio Engineering Society, Los Angeles, CA, Nov. 12–16, 1986. Preprint No. 2412*, 1986.

[11] S. P. Lipshitz and R. A. Wannamaker, “Quantization and dither: a theoretical survey,” *Journal of the Audio Engineering Society*, vol. 40, no. 5, pp. 355–75, May 1992.

[12] S. P. Lipshitz and R. A. Wannamaker, "Dithered noise shapers and recursive digital filters," in *94th Convention of the Audio Engineering Society, Berlin, Germany, 16–19 March, 1993, Preprint # 3515 (D2-2)*, 1993.

[13] G. L. Roberts, "Picture coding using pseudo-random noise," *IRE Trans. on Information Theory*, vol. IT-8, no. 2, pp. 145–54, Feb 1962.

[14] R. A. Schaporst, "Dither coding in facsimile systems," in *Electronic Imaging '88: International Electronic Imaging Exposition and Conference, Boston, MA, Oct. 3-6, 1988*, 1988, vol. 2, pp. 723–28.

[15] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. on Communication Theory*, vol. 12, pp. 162–65, Dec 1964.

[16] D. T. Sherwood, "Some theorems on quantization and an example using dither," in *Conference Record of the 19th Asilomar Conference on Circuits, Systems & Computers, Pacific Grove, CA, Nov. 6–8, 1986, 86CH2331-7*, 1986, pp. 207–12.

[17] A. K. Sinha and P. R. Chadha, "Investigation on the effect of various dithers in a noisy communication channel," in *Proc. TENCON'89, Fourth IEEE Region 10 International Conference on Information Technologies for the 90's. Bombay, India, 22-24 Nov. 1989. A89CH2766-4*, 1989, pp. 924–27.

[18] R. A. Wannamaker, "Dither and noise shaping in audio applications," M.S. thesis, University of Waterloo, Waterloo, Canada, 1991.

[19] B. Widrow, I. Kollár, and M.-C. Liu, "Statistical theory of quantization," *IEEE Trans. on Instrumentation and Measurement*, vol. 45, no. 6, pp. 353–61, 1995.

[20] B. Widrow and I. Kollár, *Quantization Noise*, Prentice-Hall, Englewood Cliffs, NJ, in preparation.

[21] P. W. Wong, "Quantization noise, fixed-point multiplicative roundoff noise, and dithering," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 2, pp. 286–300, Feb 1990.

[22] J. N. Wright, "Quantization and dither," Manuscript, 1979.