Stanley P. Lipshitz and John Vanderkooy
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada

Presented at the 109th Convention 2000 September 22-25 Los Angeles, California, USA

AES

AN AUDIO ENGINEERING SOCIETY PREPRINT

Why Professional 1-Bit Sigma-Delta Conversion is a Bad Idea

by

Stanley P. Lipshitz and John Vanderkooy
Audio Research Group
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada

Abstract: Single-stage, 1-bit sigma-delta converters are in principle unperfectible. We prove this fact. The reason, simply stated, is that, when properly dithered, they are in constant overload. The consequence is that distortion, limit cycles, instability, and noise modulation can never be totally avoided. Recording, editing, or storage systems based upon single-stage, 1-bit sigma-delta conversion, and in particular professional systems using this type of conversion, are thus a bad idea. In contrast, multi-bit sigma-delta converters which output linear PCM code (and here, multi-bit refers to five or so bits in the converter) are in principle infinitely perfectible. They can be properly dithered so as to guarantee the absence of all distortion, limit cycles, and noise modulation. The audio industry is making a tragic mistake if it adopts 1-bit sigma-delta conversion as an archival format to replace multi-bit, linear PCM.

0. Introduction

In the past twenty or so years we have seen the multi-bit converter technology used in professional and consumer equipment progress from 14 through 16 and 18 to 20 bits of resolution. Indeed, the 16-bit linear PCM format became enshrined in the CD standard, and was the basis of most digital audio storage devices for many years. All analog-to-digital and digital-to-analog conversions and intermediate signal processing steps were performed in the linear, multi-bit PCM format. One primary benefit of this format is the fact that such systems can be rendered completely linear, with infinite resolution below the least significant bit (LSB), by the adoption of proper dithering at each quantizing, or (in the case of editing and signal processing) at each requantizing, stage. Such dithering, with the optimal triangular probability density function (TPDF) dither, in principle completely eliminates all distortion, noise modulation, and other signal-dependent artifacts, leaving a storage system with a constant, signal-independent, and hence benign noise floor [1]. This is now well understood, and such practices have been the norm in the industry for over a decade. In practice, of course, no actual realization can achieve this theoretical perfection, but the departure from perfection can theoretically be zero with a multi-bit converter.

In recent years, we have seen the consumer audio industry perform a remarkable feat of salesmanship by proclaiming that 1-bit converters are better than multi-bit converters, and succeeding in marketing 1-bit products as preferable for the highest-quality performance. The original primary motivation for pursuing the 1-bit converter architecture was not superior performance, but rather the fact that they are cheaper to manufacture, consume less power, and can operate well at the voltages used in battery-powered portable equipment. This has now become secondary, as 1-bit converters are currently used in consumer audio equipment at all price and quality levels. The manufacturers of high-quality converters struggled mightily to produce 1-bit devices that met the performance goals of the industry. But, they could never eliminate all the undesirable artifacts of such converters, and after more than a decade of trying, they came to the realization that they could produce better performance by using multi-bit converter architectures in their products. The one inherent advantage of the 1-bit architecture, namely its avoidance of the level-matching difficulties found in multi-bit converters, turned out not to be as significant a benefit as one might have thought. If one examines the current data-sheets of all the major high-quality converter manufacturers, one finds that they have almost universally given up on the 1-bit sigma-delta topology in favor of oversampling converters using five or so bits. Such converter architectures avoid the intractabilities of both the 1-bit and the 20+ -bit designs. They can be properly dithered, and can thus be guaranteed to be free of low-level,

limit-cycle oscillations ("birdies"). Moreover, they do not suffer from the high-level instability problems of the higher-order, 1-bit sigma-delta converters.

In light of the above, it is with alarm that we note a recent effort to have the single-loop, 1-bit sigma-delta converter architecture adopted by both the professional and consumer audio industries as the encoding standard for a next-generation digital audio format. We refer, of course, to the Direct Stream Digital (DSD)[1] encoding which forms the basis of the Super Audio CD[2] format introduced recently by Philips and Sony (see, for example, [2] and [3]). Their intention is to have the digital audio data at every stage of the processing, from the original analog-to-digital conversion, through all the editing and mastering operations, stored in the DSD 1-bit format. We contend that this will be detrimental to audio quality, because every one of these 1-bit data conversions or reconversions entails an inevitable loss of signal quality in a way which need not occur with multi-bit, linear PCM. We shall now explain our reasoning in detail.

1. Multi-Bit versus 1-Bit Converters

In a normal multi-bit digital audio system, the intention is that the quantizer (i.e., the number system) is never deliberately driven into saturation. Because one has enough levels available, avoiding saturation is not a significant problem in practice. Moreover, there is no problem in devoting a few LSBs of headroom to ensuring that quantization errors are properly dithered. In straight linear PCM encoding, the proper (i.e., TPDF) dither spans just two LSBs. For example, in a 16-bit system, the dither would occupy only two out of the 65536 levels available. This causes a negligible reduction in system headroom in return for all the acknowledged benefits of properly-dithered signal manipulation. If one wishes to reduce the data word-length used, one can recover the lost signal-to-noise ratio by a combination of oversampling and noise shaping. Alternatively, one can increase the system's signal-to-noise ratio by the use of oversampling and noise shaping, while leaving the word-length unchanged. Noise shaping allows one to increase the signal-to-noise ratio in the audio band at the expense of decreasing it at frequencies above the audio band. One can even use *in-band* noise shaping without oversampling to significantly increase the perceived signal-to-noise ratio (see [4] and [5]). As long as the quantizer in the noise shaper does not saturate, and is properly dithered, one is guaranteed that this process is in principle completely transparent.

Noise shaping entails negative error feedback around the quantizer. In a noise shaper, a filter H is used to spectrally shape the quantization errors. Fig. 1 shows the architecture of a dithered noise-shaping quantizer.

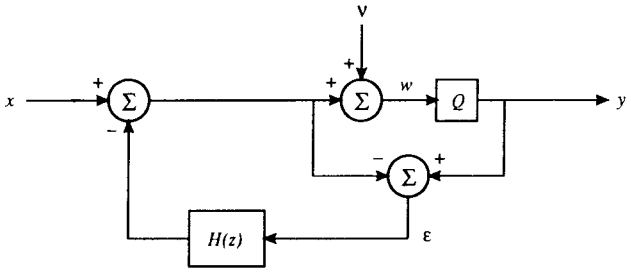

Figure 1. A dithered noise-shaping quantizer.

In this diagram, x is the input signal, v is the dither, and w is the total input to the quantizer Q. The quantization error ε is extracted around the dithered quantizer, and subtracted from the input after passing

[1, 2] Trademarks of Philips Electronics NV and Sony Electronics Inc.

through the noise-shaping filter H. This is the error feedback loop. The signal in this loop is very small as long as the quantizer does not overload. The dither v controls the statistics of the error signal ε such that, with TPDF dither, ε has zero mean, constant variance, and constant power spectral density, independent of the input signal. This means that there is no distortion or noise modulation [1]. In addition, the negative feedback loop is stable as long as there is no overload, and this is easily achieved with a multi-bit quantizer Q. The theory of such dithered noise shapers can be found in [3], [4], and [5] for example. In a sampled-data realization, the z-transforms of the input, $X(z)$, output, $Y(z)$, and error, $E(z)$, are related by

$$Y(z) = X(z) + \{1 - H(z)\}\cdot E(z).$$

The signal thus passes through the system unchanged, and the total error $E(z)$ appears at the output shaped by the effective noise-transfer function $\{1 - H(z)\}$. Proper dither v controls the statistical properties and power spectrum of the signal ε, and hence the power spectrum of the shaped output error $\{1 - H(z)\}\cdot E(z)$.

It should be noted that, in the absence of proper dither in Fig. 1, the circuit exhibits not only the expected signal-dependent quantization distortions and noise modulations, but also low-level limit-cycle oscillations because of the nonlinearity Q within the feedback loop. These "birdies" are input dc-offset dependent, and are frequency modulated by the audio signal. They can be quite pernicious and audible, and are an artifact of undithered noise shapers in general, but are *completely* eliminated by proper dithering. We shall consider quantizers $Q(w)$ of the mid-riser type shown in Fig. 2, since this characteristic is most appropriate for the 1-bit case, which we shall shortly be considering.

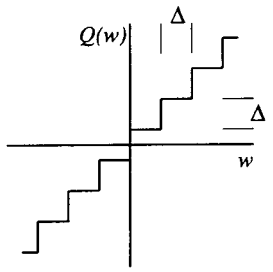

Figure 2. The mid-riser quantization characteristic adopted.

In Fig. 2, the size of the LSB is represented by Δ, so that the quantized output levels are ±Δ/2, ±3Δ/2, ±5Δ/2, etc. In an N-bit quantizer, there are 2^N levels (i.e., LSBs). In a 1-bit quantizer, however, there are *only* the two central levels present, namely ±Δ/2.

At this point one should note a couple of very important facts:

1) If the *total* input w to the quantizer always lies in the range $-\Delta \le w < \Delta$, no additional output levels beyond ±Δ/2 will be called upon, and a 1-bit quantizer will behave just like a multi-bit one. Under these conditions, the full theory of dithered multi-bit quantizers can be applied to deduce the system's behaviour. If w lies outside this range, the 1-bit quantizer overloads (i.e., saturates), and the multi-bit theory breaks down.
2) The noise-shaper circuit of Fig. 1 is functionally *completely* equivalent to the single-stage sigma-delta converter. Simple circuit transformations allow one to convert the one configuration into the other. The advantage of looking at it as a noise-shaper is that it is easier to understand than the sigma-delta circuit. Thus everything we have to say about the noise shaper applies equally to the sigma-delta converter. If the latter has many bits, it can thus in principle be perfect. That it *must* misbehave when it has only two levels is what we want to prove.

We claim that a 1-bit sigma-delta converter *must* overload when properly dithered. This follows since the TPDF dither v, needed to fully linearize the quantizer, on its own swings the quantizer's input *w* over its full no-overload range of $\pm\Delta$. Add to this the input signal and the error feedback, and the quantizer's total input *w has* to produce clipping, and its consequences — distortion, noise modulation, and instability. One might think that one could get away with partial dithering and/or a reduced input signal level, and so avoid overload; but this is not so either, except in the special case of the 1st-order sigma-delta modulator, where there does exist a limited range of possibilities. These facts are demonstrated mathematically in the Appendix, to which we refer the interested reader.

A few simulations will serve to make our points clear. For the sake of specificity we shall use a 5th-order, 64-times oversampled sigma-delta architecture, as envisaged by Philips and Sony in [2] and [3] for their DSD converter. We shall take the simplest such design, with all the noise-shaper zeros at dc. The effective noise-shaping filter $\{1 - H(z)\}$, which operates linearly on the error signal ε, is then given by $\{1 - z^{-1}\}^5$. There is a complete noise null at dc, and the noise power spectral density rises at the rate of 100 dB per decade initially. The reference sampling frequency is taken throughout to be the CD standard of 44.1 kHz, so that the DSD sampling rate is $64\times44100 = 2.8224$ MHz. For simplicity, we shall also set $\Delta = 1$, so that the LSB is 1 V. In the absence of dither, the system displays a low-level limit-cycle oscillation, whose instantaneous frequency is dependent upon the input signal. To show that these limit cycles occur, and can be pernicious even in high-order undithered modulators, Fig. 3 shows the quantizer output signal *y* when the input is just a small dc offset of amount $\Delta/128 = 0.0078125$ V.

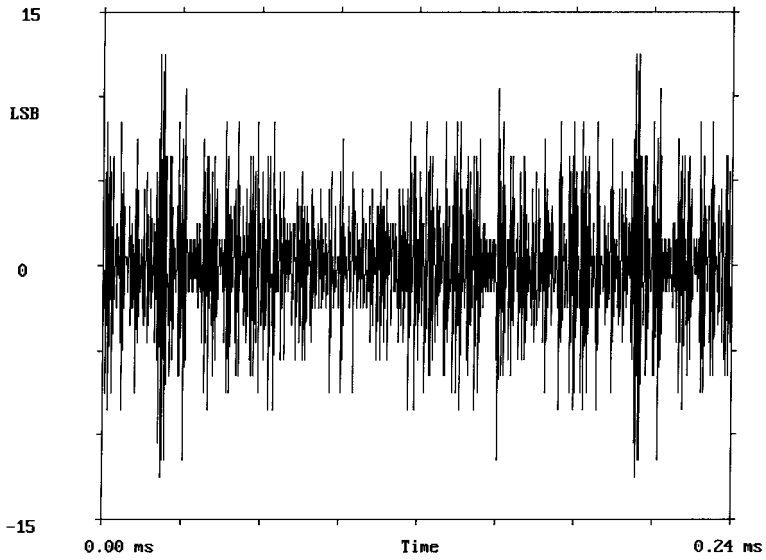

Figure 3. The undithered quantizer output signal with a dc input of +1/128 of an LSB.

We see at once that this *undithered* multi-bit sigma-delta converter needs at least ±13 LSBs if overload is to be avoided! We also see that the circuit is executing a very complicated dance, which repeats every 512 samples. The time axis shows just more than one repetition, making the periodicity evident. This "birdie" has a fundamental frequency of 5512.5 Hz, a very audible pitch. Changing the dc offset will change the frequency of this oscillation; increasing the offset raises the pitch, and reducing the offset lowers the pitch. A 1024-point unwindowed FFT displays this periodicity clearly: in Fig. 4, only every *alternate* frequency bin is present, representing all the harmonics of the waveform of Fig. 3. Note how the spectrum attempts to follow the desired 5th-order shape. This is more clearly evident in Fig. 5, which is the same data as shown in Fig. 4, but presented on a logarithmic frequency axis.

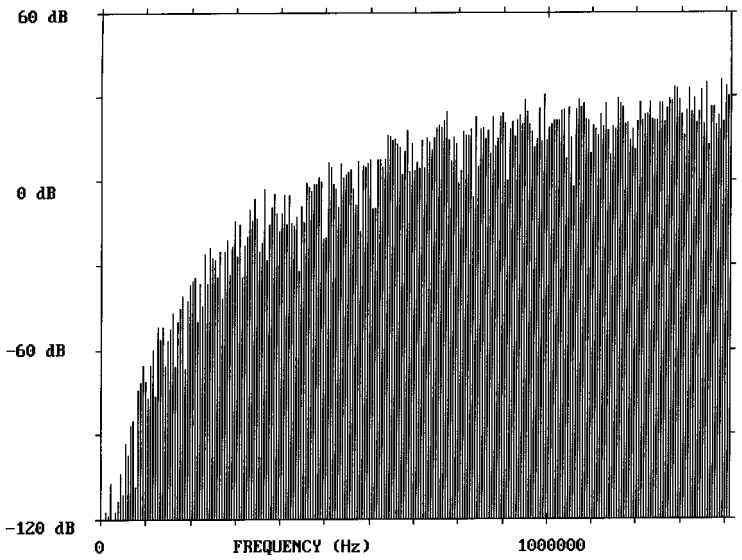

Figure 4. Linear-frequency display of the spectrum of Fig. 3.

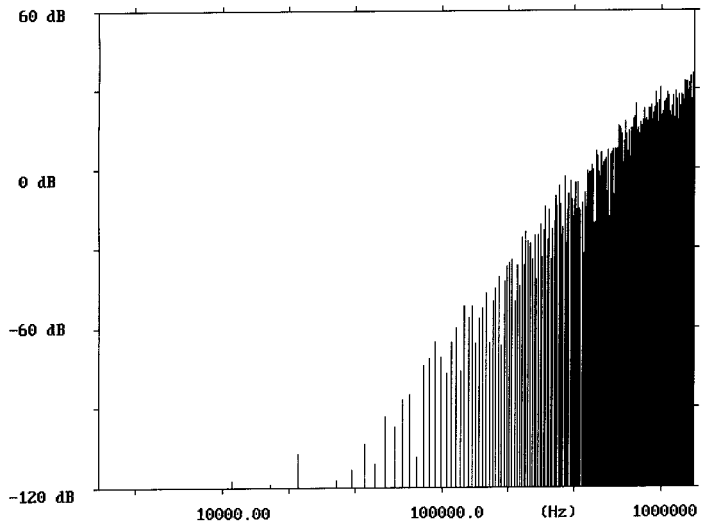

Figure 5. Log-frequency display of the spectrum of Fig. 3.

To make this quantizer behave properly, we must dither it with TPDF dither of ±1 LSB width. Figs 6 – 8 correspond respectively to Figs 3 – 5, and show the time- and frequency-domain behavior of this same circuit, with the same dc input offset, but now properly dithered. The time behaviour is now completely aperiodic, and this is reflected in the fact that the noise spectrum is now a continuous, rather than a line, spectrum. The noise power spectral density (computed using a Hann data window) follows precisely the 5^{th}-order shaper curve. All limit cycles have been completely banished, and the output spectrum is noise-like rather than tonal. Note, however, that the quantizer output now ranges over ±30 LSBs! Given that it has at least this many levels available to it, plus any additional levels necessitated by the input signal, the

quantizer will behave completely linearly, and free of any distortion or noise modulation. (The number of levels required of the quantizer before saturation generally increases with the order of the noise shaper.) This performance will hold for *any* input signal, bandlimited to less than the Nyquist frequency, which does not drive the converter into saturation. This means that such a multi-bit noise shaper/sigma-delta modulator is in principle perfect. The curve of Fig. 7 is seen to obey the Gerzon/Craven normalization condition [6] of subtending equal areas above and below the 0 dB reference line, which has been placed on all these graphs at the level corresponding to the power spectral density of a TPDF-dithered quantizer.

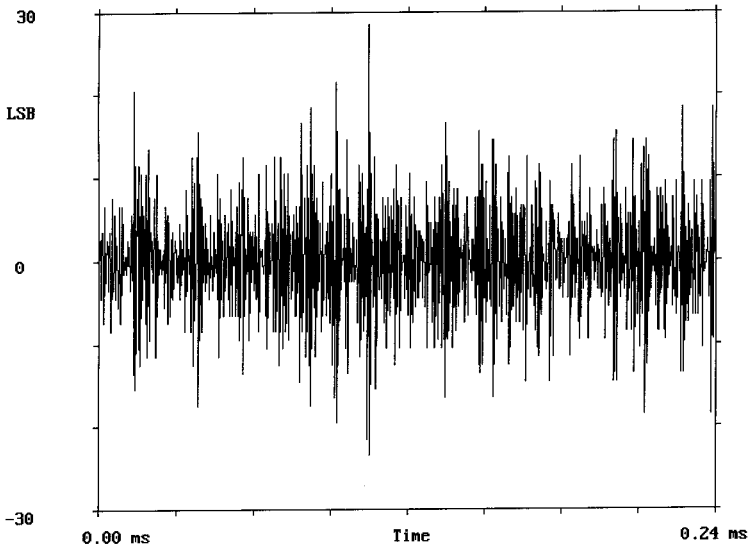

Figure 6. The TPDF-dithered quantizer output signal with a dc input of 1/128 of an LSB.

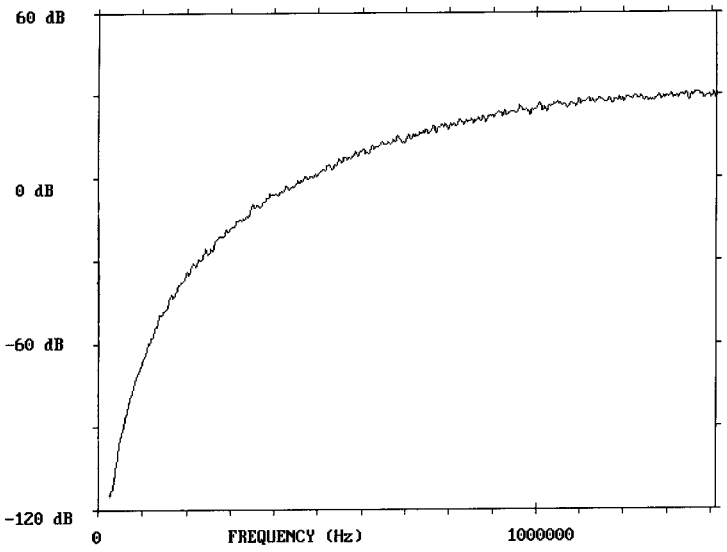

Figure 7. Linear-frequency display of the power spectral density of Fig. 6.

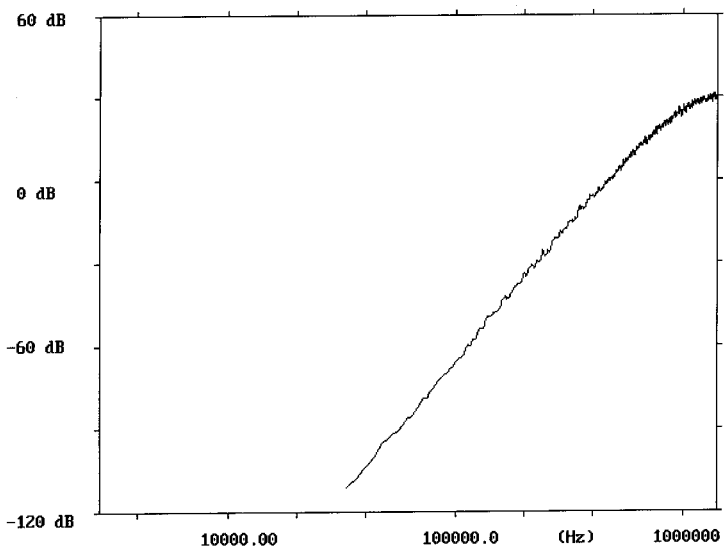

Figure 8. Log-frequency display of the power spectral density of Fig. 6.

Summarizing what we have learned from these simulations, a high-order sigma-delta converter requires *many more* than two levels in order to operate free of nonlinear artifacts, which it can do when properly dithered. If it is restricted to just the central two levels it *will* be constantly overloaded. Under these conditions, it is *impossible* to linearize it completely. This clipping nonlinearity is so severe that it is unfixable no matter how much negative feedback (i.e., noise shaping) is applied around the quantizer. More feedback will enable one to reduce the errors in *some* parts of the band under *some* signal conditions, but not throughout the audio band and not with *all* legitimate inputs. Negative feedback can work wonders but not miracles: it cannot reduce all errors to zero! In this case there is a better alternative available — the multi-bit converter. The 1-bit sigma-delta system is in principle unperfectible, while there is no theoretical limit to how far the multi-bit sigma-delta converter can be improved. One is tempted to paraphrase Albert Einstein here: "A system should be as simple as possible, but no simpler." The single-stage, 1-bit sigma-delta converter is just too simple! It *is* possible to make it surprisingly good for a system with such a gross nonlinearity, but this very nonlinearity severely limits its ultimate performance capability. Multi-bit converters do not have this limitation.

2. Further Comments on DSD and 1-Bit Sigma-Delta Conversion

Referring now more specifically to the DSD encoding format, let us recall ([2], [3]) that this mandates the use of a single-stage, 1-bit sigma-delta converter running at 2.8224 megasamples/s per channel. This is four times the data rate of a single CD audio channel, and is very wasteful from an information-theoretic point of view [7], when compared with the information capacity of the human hearing system. Be this as it may, it is nevertheless instructive to see what linear, multi-bit PCM is capable of at the same, or lower, data rates. There are many possible comparisons that could be made. Using the Gerzon/Craven "noise-shaping theorem" [6], it is easy to construct possible scenarios. We shall consider just four. Bear in mind that a 1-bit quantizer, switching between the two output levels of $\pm\Delta/2$, has a *constant total* output power of $\Delta^2/4$. Since the power is constant, the signal component of the output must come at the expense of the remainder. This argument shows that there inevitably *must* be correlated error modulation accompanying its operation. The best that could be hoped for, would be to keep all such modulation effects above the audio band. This cannot, however, be guaranteed. DSD defines a full-scale sine-wave signal to be 9 dB below this total output power (i.e., the total output power is +9 dBFS). This corresponds to a peak amplitude of $\Delta/4$ for a full-scale sine wave. Calling this level 0 dBFS for DSD, and given that the system is

to produce a noise floor more than 120 dB below full scale up to 20 kHz, rising rapidly above this frequency, one can compute that the noise power spectral density must be shaped by more than 115 dB. This *enormous* amount of noise shaping is the penalty for using a 1-bit converter. The shaping is what produces the in-band signal-to-noise ratio. Any multi-bit converter needs *much* less noise shaping to produce an equivalent result.

(a) Let us consider 16-bit, four-times oversampled PCM with noise shaping. One of the claims for the superiority of DSD is its 100-kHz bandwidth. This must be tempered by knowledge that the steeply-rising (5^{th}-order) noise curve necessitates either an even steeper lowpass filter in the digital-to-analog converter so as to control the potentially destructive high-frequency output noise, or else a premature roll-off of the band below 100 kHz. The latter seems to be the current approach being adopted by the DSD originators, as their products roll off above 50 kHz. The following scenarios are easily possible with properly-dithered, 16-bit PCM at a sampling rate of $4 \times 44100 = 176400$ Hz:
   - A noise floor 123 dB below full scale all the way up to 40 kHz, using 48 dB of noise shaping, and a total noise power of –72 dBFS.
   - A noise floor 123 dB below full scale up to 20 kHz, using only 32 dB of noise shaping, and a total noise power of only –86 dBFS.

   Both these scenarios would have a frequency response flat to 80 kHz. Either is infinitely preferable to the DSD performance at the same data rate.

(b) Next, consider 16-bit, two-times oversampled PCM with noise shaping. This is a data rate *half* that of DSD, with a sampling rate of $2 \times 44100 = 88200$ Hz. It can achieve a noise floor 120 dB below full scale up to 20 kHz, using 48 dB of noise shaping, and a total noise power of –72 dBFS. Its frequency response would be flat to 40 kHz.

(c) Finally, consider 8-bit, four-times oversampled PCM with noise shaping. This is a data rate ½ that of DSD and double that of CD, with a sampling rate of $4 \times 44100 = 176400$ Hz. It can achieve a noise floor 120 dB below full scale up to 20 kHz, using 96 dB of noise shaping, and a total noise power of –19 dBFS. Its frequency response would be flat to 80 kHz. This example is perhaps the most instructive of the lot. For a data rate ½ that of DSD, it achieves a comparable signal bandwidth, with a similar noise power density up to 20 kHz, but much lower power above this frequency, and 28 dB lower total noise power. It is fully TPDF-dithered, and so is completely artifact free. At ½ the data rate it outperforms DSD on *every* count! DSD is a profligate wastrel of capacity.

Three final comments should be made. MASH-type multi-stage converters, using say 5-bit quantization at the first-stage, are *not* subject to the same criticism as the single-stage, 1-bit DSD. Also, the repeated 1-bit sigma-delta reconversions entailed by the desire to store the data in DSD format after *each* intermediate processing stage, will result in the accumulation of significantly greater noise and nonlinear artifacts than would occur with any of the dithered multi-bit systems under corresponding conditions. This is not a trivial matter, because *each* signal processing operation (even a trivial one) results in the 1-bit DSD data stream turning into a multi-bit data stream! We wish to point out that other commentators (e.g., Hawksford [8] and Stuart [9]) have also voiced some of the same criticisms and comments as we have in this paper.

3. Acknowledgment

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada.

4. References

[1] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and Dither: A Theoretical Survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355-375 (1992 May).

[2] A. Nishio, G. Ichimura, Y. Inazawa, N. Horikawa, and T. Suzuki, "Direct Stream Digital Audio System," presented at the 100th Convention of the Audio Engineering Society, Copenhagen, 1996 May 11-14, preprint 4163.
[3] "Super Audio Compact Disc: A Technical Proposal," Philips/Sony white paper, 12 pp. (1997).
[4] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally Audible Noise Shaping," *J. Audio Eng. Soc.*, vol. 39, pp. 836-852 (1991 Nov.).
[5] R. A. Wannamaker, "Psychoacoustically Optimal Noise Shaping," *J. Audio Eng. Soc.*, vol. 40, pp. 611-620 (1992 July/Aug.).
[6] M. A. Gerzon and P.G. Craven, "Optimal Noise Shaping and Dither of Digital Signals," presented at the 87th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 37, p. 1072 (1989 Dec.), preprint 2822.
[7] J. R. Stuart, "Coding High Quality Digital Audio," presented to the Japan Audio Society, (1998 June); available for download at http://www.meridian-audio.com/ara/.
[8] M. O. J. Hawksford, "Bitstream versus PCM Debate for High-Density Compact Disc," private publication (1995 April) available for download at http://www.meridian-audio.com/ara/.
[9] R. M. Gray, "Oversampled Sigma-Delta Modulation," *IEEE Trans. Commun.*, vol. COM-35, pp. 481-489 (1987 May).

Appendix: Proof of Sigma-Delta Overload Inevitability

We present the full mathematical proof of the inevitability of quantizer overload in the case of the 1st-order, 1-bit noise-shaper. Our argument is an extension of that given by Gray [9]. In Fig. 1, this corresponds to setting $H(z) = z^{-1}$, a single-sample delay, and allowing $Q(w)$ to saturate at the $\pm\Delta/2$ levels. As we have mentioned, this circuit corresponds precisely to the single-stage, 1st-order, 1-bit sigma-delta converter. We use subscripts to denote the signals in the circuit at different sample-time instants 0, 1, 2,…, n,… , and define the quantizer output at the decision level by $Q(0) = +\Delta/2$. From Fig. 1, we deduce that these signals are related by the equations

$$w_n = x_n - \varepsilon_{n-1} + v_n \text{ for } n = 1, 2,\ldots, \tag{A.1}$$

and

$$y_n = w_n + \varepsilon_n - v_n \text{ for } n = 1, 2,\ldots . \tag{A.2}$$

Without loss of generality, we may assume that the initial state is

$$\varepsilon_0 = 0. \tag{A.3}$$

[If the initial state causes quantizer overload, then the circuit may take a finite number of steps before it comes within the no-overload region of operation, after which the analysis below will apply.] As already discussed, the 1-bit quantizer Q will operate without overload if, for each $k = 1, 2,\ldots$, we have

$$-\Delta \le w_k < \Delta, \tag{A.4}$$

or equivalently

$$-\Delta/2 < y_k - w_k \le \Delta/2;$$

i.e., by (A.2), if

$$-\Delta/2 + v_k < \varepsilon_k \le \Delta/2 + v_k \text{ for } k = 1, 2,\ldots . \tag{A.5}$$

Case (1): No dither: $\nu_n \equiv 0$

By (A.1) and (A.3), $w_1 = x_1 - \varepsilon_0 = x_1$, and so, by (A.4), no overload occurs at step number 1, provided that

$$-\Delta \le x_1 < \Delta,$$

and so certainly also under the more restrictive condition

$$-\Delta/2 \le x_1 \le \Delta/2. \tag{A.6}$$

We now use mathematical induction. Suppose that no overload has occurred at steps $k = 1, 2, \ldots, n$ with the input restricted by

$$-\Delta/2 \le x_k \le \Delta/2, \tag{A.7}$$

a condition which includes (A.6). Then, by the hypothesis and (A.5), we have

$$-\Delta/2 < \varepsilon_k \le \Delta/2 \text{ for } k = 1, 2, \ldots, n,$$

and so, by (A.1),

$$x_n - \Delta/2 \le w_{n+1} < x_{n+1} + \Delta/2;$$

i.e., also

$$-\Delta \le w_{n+1} < \Delta$$

under condition (A.7). Thus, by induction, no overload occurs for all k under condition (A.7). In this case, the 1st-order, 1-bit sigma-delta converter operates without overload. But, being undithered, it isn't linear; it exhibits distortion, noise modulation, and low-level limit-cycle oscillations just like any undithered multi-bit noise shaper.

Let us now consider what happens if we dither the modulator.

Case (2): $-\mu \le \nu_n < \mu$ for all n

Here μ represents the peak dither amplitude. By (A.1) and (A.3)

$$w_1 = x_1 - \varepsilon_0 + \nu_1 = x_1 + \nu_1,$$

and so

$$x_1 - \mu \le w_1 < x_1 + \mu;$$

i.e., by (A.4), no overload occurs at step number 1 provided that

$$-\Delta + \mu \le x_1 \le \Delta - \mu. \tag{A.8}$$

Let us now suppose that no overload occurs for $k = 1, 2, \ldots, n$ under some (to be determined) condition on the input x_k which also satisfies (A.8). Then, by (A.5), we have

$$-\Delta/2 + \nu_k < \varepsilon_k \le \Delta/2 + \nu_k \text{ for } k = 1, 2, \ldots, n,$$

and so, by (A.1),

$$x_{n+1} - \Delta/2 - v_n + v_{n+1} \le w_{n+1} < x_{n+1} + \Delta/2 - v_n + v_{n+1}.$$

But since

$$-\mu \le v_k < \mu \text{ for all } k,$$

we have

$$-2\mu \le v_{n+1} - v_n < 2\mu,$$

and so

$$x_{n+1} - \Delta/2 - 2\mu \le w_{n+1} < x_{n+1} + \Delta/2 + 2\mu.$$

Thus, the no-overload condition

$$-\Delta \le w_{n+1} < \Delta$$

holds provided

$$-\Delta/2 + 2\mu \le x_k \le \Delta/2 - 2\mu \text{ for all } k. \qquad \text{(A.9)}$$

Note that this condition also guarantees the validity of (A.8). Since (A.9) also requires that $\Delta - 4\mu \ge 0$, we must have $\mu \le \Delta/4$ for compatibility. This compatibility condition limits the dither which can be applied before causing the quantizer to overload. Now, mathematical induction leads us to conclude that for:

(a) $\mu = 0$: By (A.9) we recover again the condition (A.7) of Case (1).
(b) $0 < \mu < \Delta/4$: No overload occurs provided that the input is restricted by (A.9). The input range is now less than (A.7), and since $\mu < \Delta/4$, the dither is only partial (full TPDF dither would require that $\mu = \Delta$), and thus distortion, noise modulation, and limit-cycle oscillations can still occur.
(c) $\mu = \Delta/4$: For no overload to occur, we can allow *no input at all*; i.e., we must have $x_k \equiv 0$ for all k.
(d) $\mu > \Delta/4$: The compatibility condition is now violated, and overload is guaranteed to occur (even with no input) at some step n.

Summarizing: A dithered 1st-order, 1-bit sigma-delta quantizer (or the equivalent 1st-order noise shaper) can operate without overload only if $\mu < \Delta/4$, and then only if its input x_k is restricted by (A.9). Since it is then under-dithered, distortion, noise modulation, and limit-cycle oscillations are not eliminated. [Interestingly, choosing $\mu = \Delta/6$ gives the maximum possible no-overload input range of $-\Delta/6 \le x_k \le \Delta/6$.] When properly dithered [Case (2d) above with $\mu = \Delta$ for TPDF dither], it is *impossible* to prevent it from overloading.

Having now proven that even the simplest (i.e., 1st-order) 1-bit sigma-delta converter cannot be properly dithered, and hence completely linearized, we ask what the situation is for higher-order, single-stage, 1-bit sigma-delta modulators. These, as we have said, are equivalent to the general noise shaper shown in Fig. 1, with more complicated filters H(z). In light of the above analysis, it should come as no surprise that the higher-order circuits, with their higher noise gains, are even *more* likely to overload than the simple example discussed above. We will not attempt to present a full analysis here. The 5th-order example used to illustrate Section 1 represents the minimum needed by a DSD modulator in order to achieve the desired signal-to-noise ratio of more than 120 dB up to 20 kHz using 64 times oversampling. We saw that such a modulator is operating well into overload even when undithered. This represents the generic situation.