Audio Engineering Society

Convention Paper 5395

Presented at the 110th Convention
2001 May 12–15 Amsterdam, The Netherlands

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Why 1-Bit Sigma-Delta Conversion is Unsuitable for High-Quality Applications

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 imperfectible. We prove this fact. The reason, simply stated, is that, when properly dithered, they are in constant overload. Prevention of overload allows only partial dithering to be performed. The consequence is that distortion, limit cycles, instability, and noise modulation can never be totally avoided. We demonstrate these effects, and using coherent averaging techniques, are able to display the consequent profusion of nonlinear artefacts which are usually hidden in the noise floor. Recording, editing, storage, or conversion systems using single-stage, 1-bit sigma-delta modulators, are thus inimical to audio of the highest quality. In contrast, multi-bit sigma-delta converters, which output linear PCM code, are in principle infinitely perfectible. (Here, multi-bit refers to at least two bits in the converter.) They can be properly dithered so as to guarantee the absence of all distortion, limit cycles, and noise modulation. The audio industry is misguided if it adopts 1-bit sigma-delta conversion as the basis for any high-quality processing, archiving, or distribution format to replace multi-bit, linear PCM.

0. INTRODUCTION

This paper is an enlarged and extended version of [1], and its findings regarding 1-bit sigma-delta modulators are explored in greater detail in an associated paper [2].

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 or more 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 analogue-to-digital and digital-to-analogue conversions and intermediate digital signal processing steps were performed in the linear, multi-bit PCM format, using internal processing word-lengths greater than the desired final numerical precision. 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 artefacts, leaving a storage system with a constant, signal-independent, and hence benign noise floor (see [3] and [4]). 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 analogue realization can achieve this theoretical perfection, but in the digital domain the departure from perfection can indeed be zero due to the numerical precision of the arithmetical operations involved.

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 it is cheaper to manufacture, consumes 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 artefacts 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 more than two levels. Such converter architectures can 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 the adoption of the single-stage, 1-bit sigma-delta converter architecture as the encoding standard for a next-generation (and supposedly higher-quality) consumer 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, [5] and [6]). The original intention to have the digital audio data at every stage of the processing — from the original analogue-to-digital conversion, through all the editing and mastering operations — stored in the DSD 1-bit format has apparently now been abandoned. This was a wise decision. The conversion to the final 1-bit DSD format, however, still represents a required, and quite unnecessary, degradation of the quality of the audio signal. Every single 1-bit data conversion entails an inevitable loss of signal quality in a way which need not occur with multi-bit, linear PCM. The original rationale for storing a 1-bit DSD format signal on the Super Audio CD has now entirely vanished. The analogue-to digital and digital-to-analogue conversions, and all intermediate digital signal processing, will likely be done using multi-bit converters and storage formats. There really is no point in degrading the signal, by squeezing it onto a 1-bit Super Audio CD for transmission to the consumer, only to have it converted back to multi-bit PCM in the process of being played back. We shall now explain our reasoning in detail.

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

1. MULTI-BIT VERSUS 1-BIT CONVERTERS

In a normal multi-bit digital audio system, the intention is that the quantizer (*i.e.*, essentially 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 precisely two LSBs. For example, in a straight 16-bit system, the dither occupies only two out of the 65,536 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/or 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 [7] and [8]). As long as the quantizer inside the noise shaper does not saturate, and is properly dithered, one is guaranteed that this process is completely transparent, in that it is totally distortion free.

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

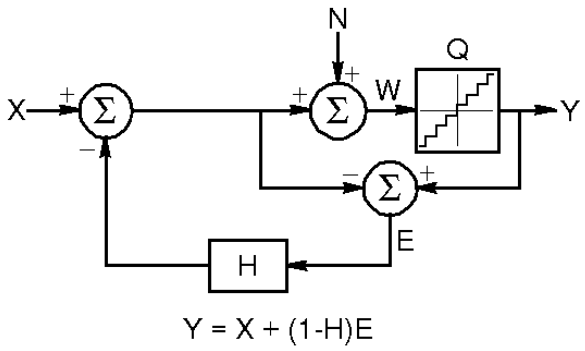

Figure 1. Simple dithered noise-shaping quantizer.

In this diagram, X is the input signal, N is the dither, W is the total input to the quantizer Q, and Y is the output signal. The quantization error E is extracted around the dithered quantizer (which can be multi-bit or single-bit), and subtracted from the input after passing through the noise-shaping filter H(z). H(z) can be either recursive or non-recursive. This is the error feedback loop. The signal in this loop is very small as long as the quantizer does not overload. The dither N controls the statistics of the error signal E such that, with TPDF dither, E has zero mean, constant variance, and a constant white power spectral density, independent of the input signal — indeed, E is then uncorrelated with X. This means that there is no distortion or noise modulation (see [3] and [4]). 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 [7], [8], and [9] 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 quantization error E(z) appears at the output shaped by the effective noise-transfer function {1 – H(z)}, to become the system's total error {1 – H(z)}·E(z). Proper TPDF dither N controls the statistical properties and power spectrum of the error signal E, and hence controls the power spectrum of the shaped output error {1 – H(z)}·E(z). For stable operation, and the least possible noise for a given noise-shaping curve, the equivalent noise-shaping filter {1 – H(z)} must be minimum phase. Of course, for computability, H(z) must incorporate at least a single sample delay.

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 artefact 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.

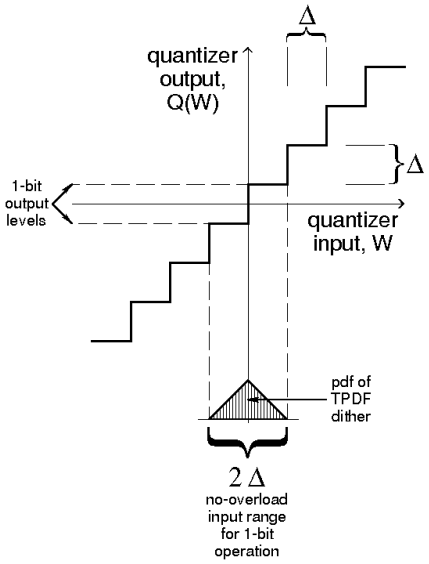

Figure 2. Mid-riser quantization characteristic adopted.

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 indicated 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 –Δ ≤ W < Δ, 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, however, 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, which forms the heart of the DSD 1-bit encoder of the Super Audio CD. Simple circuit transformations allow one to convert the one configuration into the other. This is exhibited in Fig. 3, which shows in (a) the most general noise-shaper topology, and in (b) its equivalent sigma-delta form. In contradistinction to the error feedback of the noise-shaper topology (a), the equivalent sigma-delta topology (b) could be said to represent straight *negative feedback*. (To obtain the most-frequently used sigma-delta structure, where the filtering is done solely in the forward path, we set F ≡ 1, and the forward-path filter then becomes {G/(1 – G)}.) The advantage of looking at the circuit as a noise-shaper is that it is easier to understand than the sigma-delta circuit. Moreover, the error signal E is explicitly available in the former topology, but is implicit in the latter. We shall use the basic noise shaper circuit of Fig. 1 for the experiments to follow. Everything we have to say about the noise shaper thus applies equally to the corresponding sigma-delta converter under the transformation shown in Fig. 3. If the latter has enough bits so that it does not overload when properly dithered, it can thus in principle be perfect. That it *must* misbehave when it has only two levels is what we want to prove.

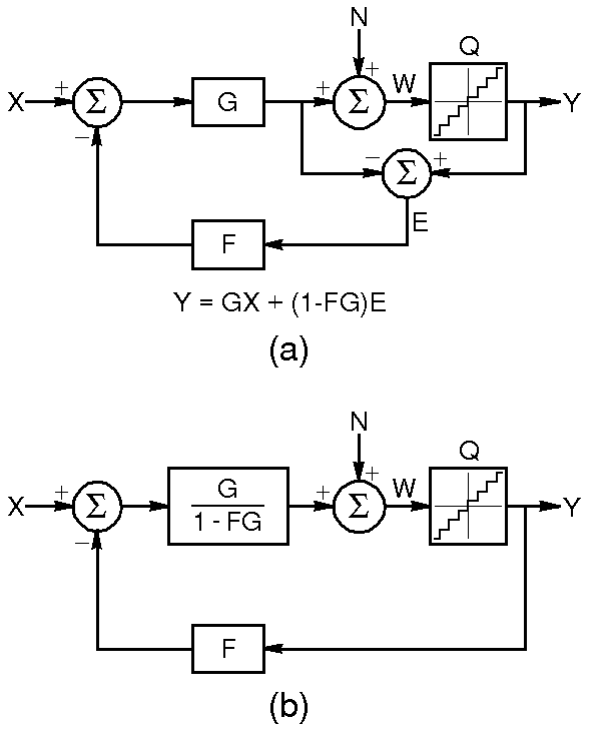

Figure 3. Showing the general equivalence of the noise-shaper (a) and sigma-delta (b) topologies.

We claim that a 1-bit sigma-delta converter *must* overload when properly dithered. This follows at once, since the TPDF dither N, which is needed to fully linearize the quantizer, itself swings the quantizer's input W over its full no-overload range of ±Δ. This is illustrated in Fig. 2. To obtain the total quantizer input W, we must add to N both the input signal X and the fed back error signal {–HE} (see Fig. 1). Now, the dither samples are statistically independent of the other signals in the loop, and so clearly the quantizer's total input W *has* to produce overload, and its consequences — distortion, noise modulation, and instability — if there is any input or feedback. One might think that one could maintain most of the benefits while preventing the overload caused by using full TPDF dither, by using only partial dithering and/or a reduced maximum input signal level. This is true. But, reduction in the maximum signal level is undesirable because of its direct impact on the signal-to-noise ratio;

and, as we shall see, the needed reduction in dither level is so great that its main remaining benefit is the prevention of limit cycles and noise modulation, and not the reduction of distortion. In the special case of the 1st-order sigma-delta modulator, one can prove that there exists a limited range of dither levels and of input signal levels which *guarantees* the absence of quantizer overload. Such results are more elusive in the higher-order case. These facts are demonstrated mathematically in the Appendix, to which we refer the interested reader. The important point is that full TPDF dither is *never* allowable in a 1-bit noise shaper or sigma-delta modulator, and hence full linearity is *never* achievable either in principle or, of course, in practice.

We wish to point out that we are not the first to have addressed the issue of dithering 1-bit sigma-delta converters. In this area, Norsworthy's research [10] is particularly comprehensive. Hawksford [11] and Stuart [12] have also previously voiced some of the same criticisms and comments as we do in this paper.

2. SIMULATIONS

Simulations will serve to make our points clear. For the sake of specificity we shall use a 64-times oversampled sigma-delta architecture, as envisaged by Philips and Sony in [5] and [6] for their DSD converter. Our design, called Lip7ZP, attempts to shape the noise floor in a psychoacoustically beneficial manner, by suppressing the noise spectral density in the 3-4 and 12 kHz regions where the human ear is most sensitive (see [7] and [8]). It is based on a 7th-order noise-shaping transfer function {1 – H(z)}, having seven zeros and seven poles in the complex z-plane. There is one real zero at dc, and three complex conjugate pairs of zeros just inside the unit circle, corresponding to frequencies of 4, 12, and 20 kHz. This results in a psychoacoustically-shaped noise floor in the audio band. The seven poles inside the unit circle have an approximate Butterworth alignment, and shelve the noise gain at frequencies above 70 kHz in order to maintain stability even in 1-bit mode. (The design is not optimized, but is simply intended to illustrate the type of psychoacoustic noise shaping that can be achieved, and to serve as an illustrative vehicle for the purposes of exposition.) The noise power spectral density rises at a rate of 140 dB per decade above 20 kHz. The theoretical spectral shape is shown in Fig. 4.

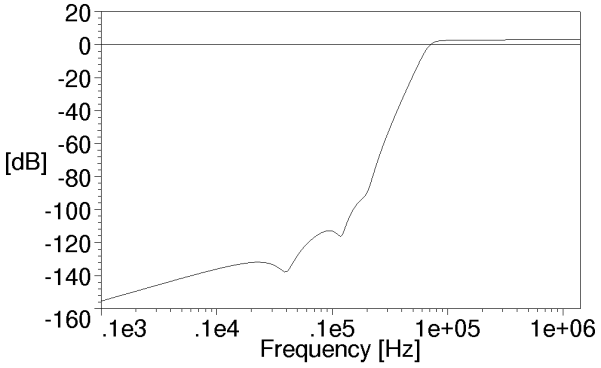

Figure 4. Theoretical power spectrum of the Lip7ZP noise shaper.

The explicit expression for the Lip7ZP noise-shaping filter H(z) is given here so that interested readers can implement it and confirm the results which follow:

```
H(z) =
0.67147148261434210554
*(z^2 - 1.8188518739882731732*z + 0.82800290523865701387)
*(z^2 - 1.8703147300286237302*z + 0.88232721441979615814)
*(z^2 - 1.9638172195444934487*z + 0.98421718272643295544)/
((z^2 - 1.9124580000000000000*z + 0.93559300000000000000)
*(z^2 - 1.8047140000000000000*z + 0.82640000000000000000)
*(z^2 - 1.7373740000000000000*z + 0.75810400000000000000)
*(z - 0.85521900000000000000))
```

The reference sampling frequency is taken throughout to be the CD standard of 44.1 kHz, so that the DSD sampling rate is 64 × 44,100 Hz = 2.8224 MHz. For simplicity, we shall also set Δ = 1, so that the LSB is 1 V (see Fig. 2). All the time-domain figures which follow, display the time axis labelled in samples, at the DSD sampling rate of 2.8224 MHz, and use LSBs on the vertical axis. All the spectral figures show frequency on the horizontal axis (logarithmic or linear scale), up to the Nyquist frequency of 1.4112 MHz. On the vertical axis, 0 dB represents the power spectral density of the white quantization noise of a properly TPDF-dithered quantizer. This is the signal E in Fig. 1, and has a total noise power of $\Delta^2/4$ up to the Nyquist frequency, when the quantizer is dithered with a TPDF dither N. The spectral curves represent one of the following two things:

(a) The average of 64 successive power spectra, each obtained from a 16,384-point FFT of the output signal from the noise shaper, using a Hann[3] window. The power averaging results in a smoother curve (*i.e.*, in reduced variance), but does not affect the level of the noise.

(b) The FFT of the coherent (or synchronous) average of 65,536 (*i.e.*, 2^{16}) time records, each of 16,384 points. This averaging reduces uncorrelated signal components, while leaving correlated components unaffected. It lowers the uncorrelated part of the noise floor by 16 × 3 = 48 dB, while leaving its variance unchanged. This technique allows us to see correlated nonlinear components which would otherwise have been hidden in the noise floor of the 16,384-point FFT. It is equivalent noise-wise to averaging 65,536 complex spectra, or performing a single FFT on the impossibly long time record of $2^{16} \times 2^{14} = 2^{30}$ = 1,073,741,824 points, and then decimating the result! Again, a Hann[3] window has been used. (More information about the use of synchronous averaging will be found in the companion paper [2].)

When undithered, Lip7ZP exhibits all the defects of undithered noise shapers and sigma-delta converters. It needs only two output levels, even when driven to DSD full scale of ±Δ/4, provided that it is not fully dithered. It is also stable, within its overload range, even when operated in 1-bit mode. We shall illustrate that, when undithered, it can:

1) have audible limit cycles;
2) display gross modulation of the noise floor in the audio band, as a function of the input signal; and
3) produce harmonic and intermodulation distortion and other nonlinear effects.

All these defects are, of course, banished when Lip7ZP is fully TPDF-dithered, but then it requires at least eight levels to prevent overload and instability. These are not available in 1-bit mode, and so complete linearization is not possible if it is used in a 1-bit sigma-delta converter. Partial dithering does, however, allow us to suppress limit-cycle oscillations, control the noise modulation, and somewhat reduce the audio-band nonlinearities. Let us now show these effects.

In the absence of dither, and with zero initial conditions, Lip7ZP displays a low-level limit-cycle oscillation of period 328 samples (and a very audible fundamental frequency of 8.605 kHz). This "birdie" takes almost 16,000 samples before it is firmly established, but once established, it is very robust. This shows that, contrary to some opinions, limit cycles can occur, and can be pernicious, even in high-order undithered modulators. Fig. 5 shows precisely two periods of this complicated oscillation, as close inspection will reveal!

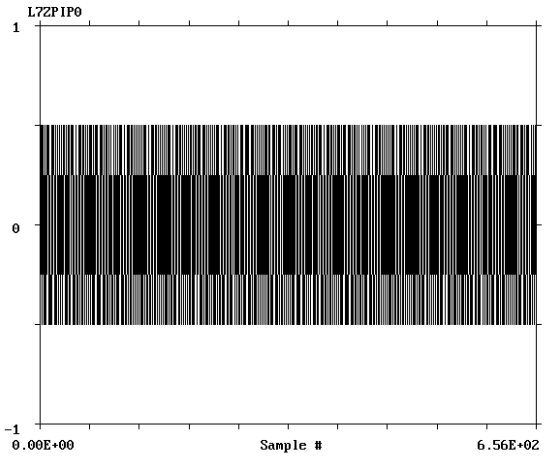

Figure 5. Output signal of the undithered Lip7ZP noise shaper, with no input, showing exactly two periods of its limit cycle.

An averaged power spectrum displays this periodicity clearly. Fig. 6 shows the result on a logarithmic frequency axis. Harmonics are seen to be present all the way up to the Nyquist frequency. The continuous curve running across the plot shows what this limit cycle changes into if we add to the quantizer 0.0013Δ peak-to-peak of rectangular probability density function (RPDF) dither — the minimum necessary to break up the limit cycle. (This is an amount of dither 60.7 dB less than full TPDF dither.) Note that this curve is *not* the proper shape (*cf.* Fig. 4), but this small amount of dither does serve (in this case) to produce a far more desirable type of background "noise".

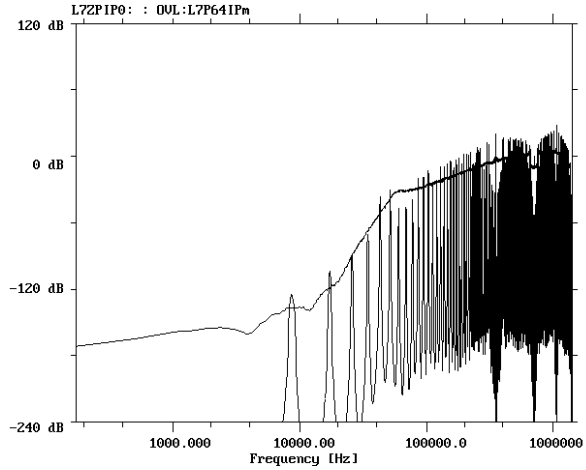

Figure 6. Averaged power spectrum of Fig. 5. (The continuous curve running across the plot is explained in the text.)

So, given that a little dither is a good thing, let us try using the proper amount of dither, namely, full TPDF dither of ±1 LSB width. We already know that the system will need more than two quantizer levels, and indeed it does. Fig. 7 shows that it occasionally needs up to ±4 levels; that is, a 3-bit quantizer is needed to avoid overload. In fact, if one restricts this fully-dithered quantizer to just two levels, the system is unstable. 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.

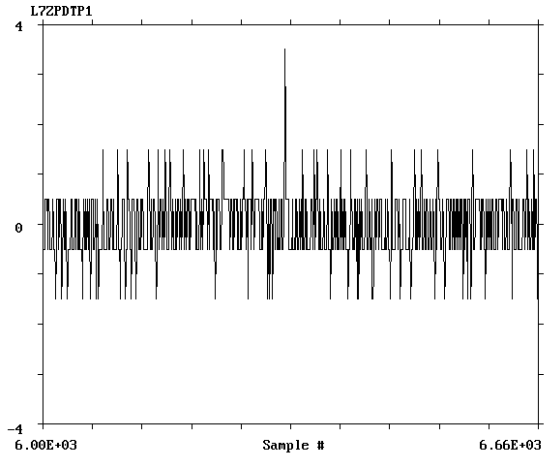

Figure 7. TPDF-dithered Lip7ZP noise shaper output signal with no input, showing the need for a 3-bit quantizer (8 levels).

The noise power spectral density corresponding to Fig. 7, shown in Fig. 8, follows precisely the theoretical shaper curve (compare with Fig. 4). All limit cycles have been completely banished, and the output spectrum is noise-like rather than tonal. Fig. 8 actually shows *two* superimposed curves for the TPDF-dithered Lip7ZP noise shaper. One is the zero-input power spectrum, and the other is the power spectrum when a full-scale sine-wave signal of 22.05 kHz (*i.e.*, on FFT bin 128) is applied to the input X. (The full-scale signal is set at an amplitude of Δ/4, in accordance with the DSD specification, and reads +26.5 dB on our TPDF-normalized vertical axis.) Note the complete absence of any noise modulation — the two curves overlay each other perfectly. The only difference between these curves is the change due to the presence of the sinusoidal input.

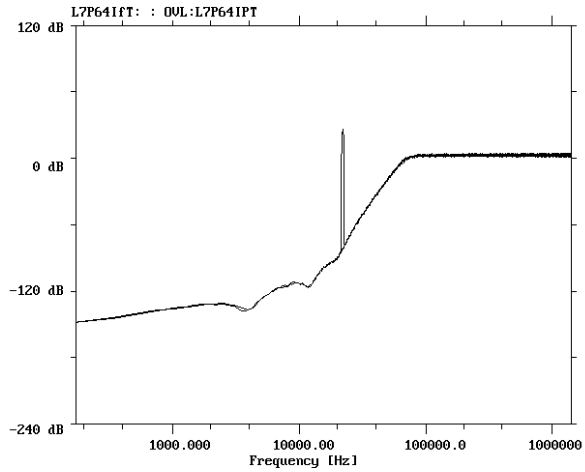

Figure 8. Overlaid responses of a TPDF-dithered Lip7ZP noise shaper, both without and with a DSD full-scale 22.05 kHz sine-wave input. The noise floor is completely unchanged when the input signal varies from zero to full scale.

This superb 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, in that its only contribution is the addition of a benign constant background noise spectrum. This is the best that *any* system, digital or analogue, can do. (The addition of noise is inevitable in either domain; the addition of nonlinear artefacts isn't.)

Synchronous averaging will help to demonstrate the complete lack of nonlinear distortions in this converter when TPDF dither is used. In Fig. 9 we show two output spectra under the simultaneous input of *two* sine-wave signals, each of half full-scale amplitude (*i.e.*, $\Delta/8$ each), on FFT bins 32 (5512.5 Hz) and 48 (8268.75 Hz). Placing these signals precisely on FFT frequency bins ensures that they are indeed fully synchronous with the time records being transformed. Only eight quantizer levels are needed. The upper trace is the result of incoherently averaging 64 power spectra, while the lower trace shows the FFT of 65,536 ($= 2^{16}$) coherently averaged data sets. The noise floor in the lower trace has the same variance as that of a single transformed data set, but its level has been reduced by 48 dB ($= 16 \times 3$ dB). Each doubling of the number of coherently-averaged data sets lowers the noise floor by 3 dB, provided that it is incoherent, as is the case here in consequence of the TPDF dithering. The synchronous input signals, on the other hand, are left unchanged by the averaging. Note the complete absence of any signal-related artefacts poking out from the lowered noise floor. This is the beauty of proper dither! There is *no* distortion.

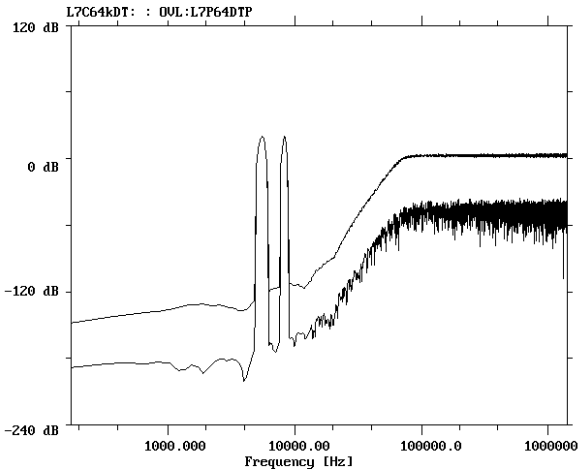

Figure 9. Here Lip7ZP is TPDF-dithered and subjected to the simultaneous application of two half-full-scale sine-wave signals. The upper trace shows the result of performing 64 power spectral averages on the quantizer output. The lower trace shows how 65,536 synchronous averages have uniformly lowered the uncorrelated noise floor by the expected 48 dB.

Let us now try the same experiment using absolutely no dither, which is the way sigma-delta converters are often used. Fig. 10 shows what we find. The two traces correspond to those in Fig. 9. We see that the coherent averaging has revealed a forest of harmonic and intermodulation products in the audio band lying just below the incoherent (upper) noise floor curve. Only three of them actually poke out above this upper curve, and so without the benefit of synchronous averaging we would have been led to believe that there was very little distortion being produced. Now we see many sum and difference products, including a strong first-order difference tone on bin 16 (2756.25 Hz), as expected. Notice also that the high-frequency portion of the lower curve (actually the whole upper half of the Nyquist band) has *not* been lowered by the coherent averaging. This indicates that these components are fully correlated with the input signal, and indeed represent the bulk of the total output error power. Fig. 11 shows the same data as Fig. 10 on a linear frequency axis, to display the Nyquist region artefacts more clearly. There is definitely something interesting going on here, which wouldn't be nearly as apparent had we not performed the synchronous averaging. We shall show in [2] that these spectral lines are the FM sidebands of an idle tone whose frequency is being modulated by the input signal. Finally, comparing the upper traces of Figs 9 and 10, we see that both the level and shape of the noise spectrum in Fig. 10 are wrong (also *cf.* Fig. 4).

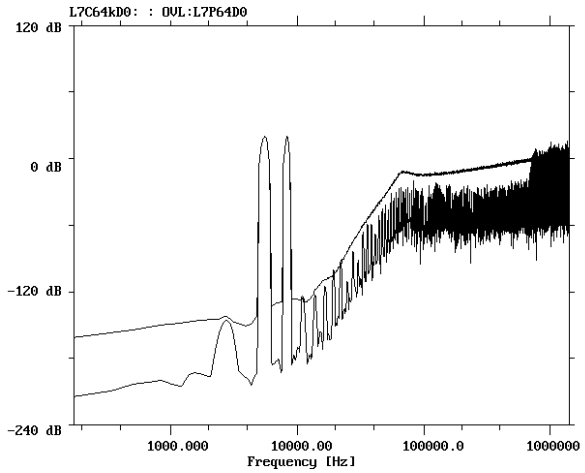

Figure 10. Here Lip7ZP is undithered and subjected to the same input signals as in Fig. 9. The upper trace shows the result of performing 64 power spectral averages on the quantizer output. The lower trace shows how 65,536 synchronous averages have revealed a plethora of nonlinear artefacts below the noise floor.

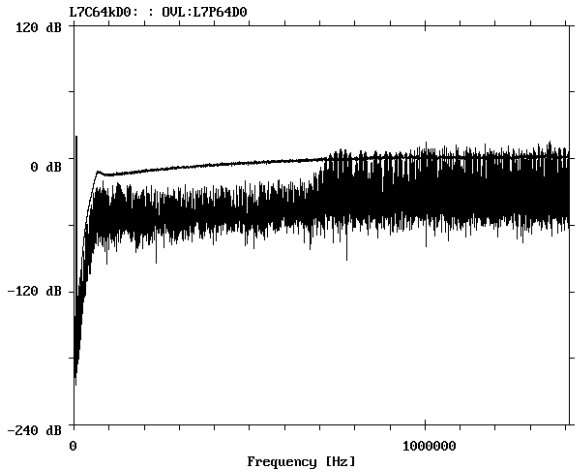

Figure 11. The same as Fig. 10, but displayed using a linear frequency axis.

Figs 9 and 10 have revealed respectively the best and worst behaviour of Lip7ZP. They are in fact quite typical of the generic behaviour of 1-bit noise shapers, as is shown in [2]. In Fig. 9 the quantizer is operating multi-bit due to the TPDF dither. In Fig. 10 it is operating as a 1-bit device, exercising only the central two levels. Let us add as much dither as it can tolerate without 1-bit overload. We find that we can inject only 0.17 LSBs peak-to-peak of RPDF dither before additional levels become necessary. This is only about

$1/69^{th}$ of the power of full TPDF dither (*i.e.*, −18.4 dB relative to the power of TPDF dither). How much benefit does this small dither provide? Figure 12 shows the result. Compared to Fig. 10 we see a general reduction in the higher-order audio-band products, but only modest decreases in most of the lower-order ones, which are close in frequency to the input signals. (Interestingly, one will see on close inspection that some distortion components have actually *increased* as a result of this partial dither — look at the sixth spectral line to the right of the two input signals in Fig. 10, and compare it to Fig. 12!) None of the audio-band artefacts now pokes above the upper curve, and so they are invisible to straight power spectral averaging of 16,384-point resolution. One might have been led to wrongly conclude that this maximally-dithered 1-bit sigma-delta converter was free of audio-band distortion. It is, however, still decidedly nonlinear.

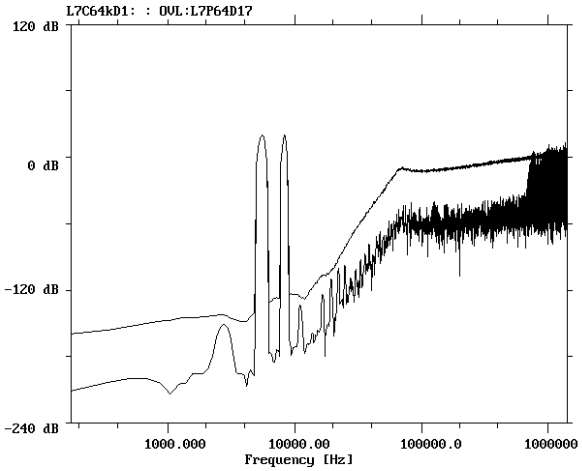

Figure 12. Here Lip7ZP is given the maximum dither possible subject to no 1-bit overload (namely 0.17Δ peak-to-peak of RPDF dither), and subjected to the same input signals as in Fig. 9. The upper trace shows the result of performing 64 power spectral averages on the quantizer output. The lower trace (65,536 synchronous averages) shows a modest reduction in distortion products compared to the undithered case (Fig. 10).

So, partial dithering to the maximal extent possible within overload margins has helped, but only in a limited way. It is, nevertheless, extremely desirable to do this for two other reasons: it reduces signal-dependent noise modulation, and it prevents limit-cycle oscillations from occurring. We have already demonstrated the latter benefit in Fig. 6. To demonstrate the former benefit, we shall perform two further experiments. First, we compare the continuous curve from Fig. 6, corresponding to the zero-input case with the addition of the smallest smidgen of dither (0.013Δ peak-to-peak of RPDF) to break up the zero-input limit cycle, to the case of an undithered full-scale sine-wave of 22.05 kHz (*i.e.*, on FFT bin 128). These power-averaged curves are overlaid in Fig. 13, and demonstrate the extent of the modulation of the noise floor due to the presence of signal — a substantial 10 dB over the audio band — and major changes in shape above 50 kHz. Notice also the presence of the third- and fifth-harmonic distortion products of the input signal poking out from the undithered (upper) curve, and the many spectral lines near the Nyquist frequency. The third harmonic is only 32 dB below the fundamental.

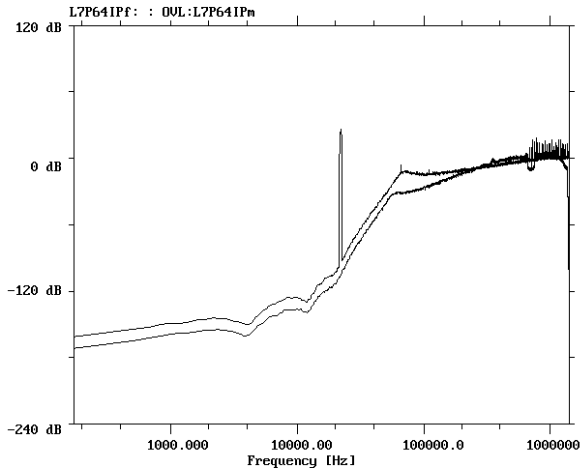

Figure 13. Showing the extent of the noise modulation in Lip7ZP when undithered. See the text for details. Both curves are the result of 64 power spectral averages.

Second, we apply the maximum amount of RPDF dither which is compatible with the 1-bit no-overload criterion, namely 0.17 LSBs peak-to-peak of RPDF. Fig. 14 shows two power spectral average curves overlaid, one corresponding to no input, and the other corresponding to a full-scale sine-wave input on bin 128 (*i.e.*, 22.05 kHz). Compared to Fig. 13 we see that the noise modulation has been drastically reduced (to around 2 dB), but the third-harmonic distortion spike, and the Nyquist products are only slightly suppressed (the third harmonic has fallen by only 2 dB). Nevertheless, these results serve to confirm the benefits of using the maximum amount of dither that can be accommodated.

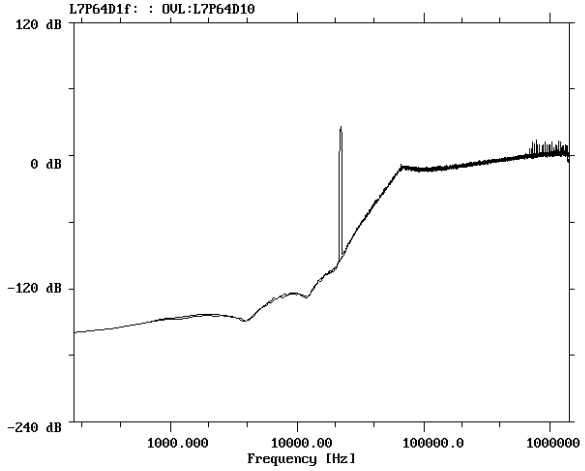

Figure 14. The same as Fig. 13, but using the maximum amount of dither possible without causing overload (0.17Δ peak-to-peak RPDF). The noise modulation has been reduced to around 2 dB.

Summarizing what we have learned from these simulations, a high-order sigma-delta converter requires *more* than two levels in order to operate free of nonlinear artefacts, which it can do when properly dithered. If it is then restricted to just the central two levels it *will* be constantly overloaded, and probably unstable. Under these conditions, it can be only partially dithered, and so *cannot* be completely linearized. The remaining nonlinearity is unfixable no matter how much negative feedback (*i.e.*, noise shaping) is applied

around the quantizer. (See [2] for more discussion of these matters.) More feedback will enable one to reduce the errors in *some* parts of the band under *some* signal conditions, but not throughout the whole 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 imperfectible, 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.

3. FURTHER COMMENTS ON DSD AND 1-BIT SIGMA-DELTA CONVERSION

Referring now more specifically to the DSD encoding format, let us recall ([5], [6]) 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 [10], 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 rate. There are many possible comparisons that could be made. Using the Gerzon/Craven "noise-shaping theorem" [9], 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 output 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 noise and error modulation accompanying its operation. The best that could be hoped for would be that all such modulation effects occur only 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 at least 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, because it starts with a better signal-to-noise ratio to begin with.

(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}- to 7^{th}-order) noise curve necessitates either an even steeper lowpass filter in the digital-to-analogue 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 $4 \times 44{,}100 = 176{,}400$ Hz, which is the *same* data rate as DSD:
   - 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 *one-half* that of DSD, with a sampling rate of $2 \times 44{,}100 = 88{,}200$ 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 also a data rate *one-half* that of DSD and double that of CD, with a sampling rate of $4 \times 44{,}100 = 176{,}400$ 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 one-half 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 artefact free. At one-half the data rate it outperforms DSD on *every* count! DSD is a profligate wastrel of capacity.

It is instructive to see how the above numbers are calculated. For the purposes of illustration, let us consider Example (c) above. The four-times oversampling spreads the quantization noise power over four times the CD's bandwidth, and so reduces its noise power spectral density (PSD) by 6 dB. Unshaped, but fully TPDF dithered, this 8-bit system would thus have a noise PSD lying at –51.1 dBFS (= 49.9 dB for an 8-bit system + 6 dB for the oversampling – 4.8 dB for the dither). Allowing a generous 3-dB headroom reduction because of the dither, the noise PSD lies at –48.1 dBFS. Now, the Gerzon/Craven "noise-shaping theorem" [9] tells us that the areas of any optimal noise shaper curve above and below the unshaped PSD must be *equal* when plotted on a linear-frequency/logarithmic-amplitude basis. Since we want to pull the shaped noise floor down to –120 dBFS over about one-quarter of the Nyquist band (say to 22.05 kHz), the idealized shaping required is as shown in Fig. 15.

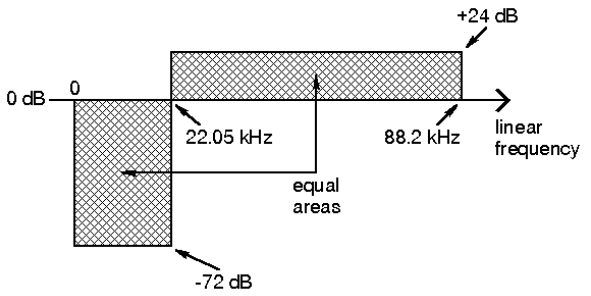

Figure 15. Illustrating the Gerzon/Craven "noise-shaping theorem" for the case of Example (c).

The noise PSD below 20 kHz needs to be pulled down by 72 dB (= 120 – 48 dB). The equal-areas theorem mandates that the PSD over the remaining three-quarters of the Nyquist band will consequently be elevated by 24 dB (= 72/3 dB), so that the total amount of shaping is 96 dB (= 72 + 24 dB). The *total* noise power can now be computed to be –19 dBFS. Now, these calculations assume a ideal rectangular-shaped noise curve, as shown in Fig. 15. This is, of course, not achievable with finite-order filters, and so these numbers must be taken merely as a useful approximate starting point for the actual design work. We have implemented the design of Example (c)

using a 12th-order recursive filter for H(z), and have also taken the opportunity to provide some crude (non-optimized) psychoacoustic shaping, with dips around 3 and 12 kHz. The actual result is shown in Fig. 16. Notice how the curve obeys the equal-areas property. The simulated performance of this design shows a signal-to-noise ratio of 120.4 dB up to 20 kHz. The TPDF dither uses up about 70 of the 256 levels available in this 8-bit system, so that our allowance for it was indeed conservative.

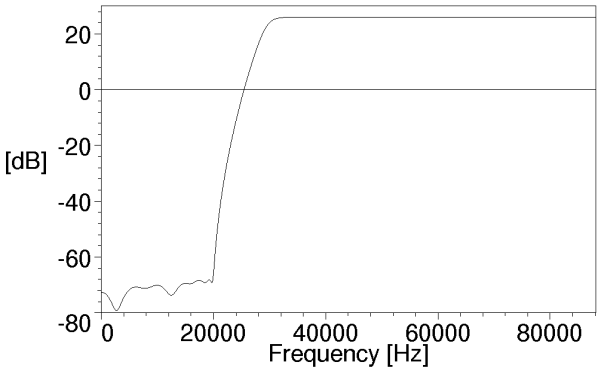

Figure 16. The realization of Example (c), incorporating a crude attempt at psychoacoustic shaping. This figure should be compared with the idealized diagram shown in Fig. 15, which formed the basis for the design.

4. CONCLUDING REMARKS

Some final comments and speculations can be made:

- MASH-type multi-stage converters, using multi-bit quantization at the first-stage, are *not* subject to the same criticism as the single-stage 1-bit sigma-delta converter, provided that their quantizers do not overload.
- The repeated 1-bit sigma-delta reconversions entailed by a misguided desire to store the data in DSD format after *each* intermediate processing stage, would result in the accumulation of significantly greater noise and nonlinear artefacts 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, such as a gain change) results in the 1-bit DSD data stream turning into a multi-bit data stream!
- Because of the insoluble theoretical problems discussed in this paper, we are unaware of *any* way to generate a Super Audio CD test disc which is both distortion-free and has a constant, signal-independent noise floor! In the multi-bit domain, this is easily done using standard dithering methods. Indeed, the measurement standards for PCM-based audio, developed by the Audio Engineering Society, *mandate* the use of TPDF dither. This is an impossibility for 1-bit digital audio.
- The amount of negative feedback used in a 1-bit sigma-delta modulator striving to straighten its quantizer transfer characteristic, and simultaneously achieve a signal-to-noise ratio of 120 dB, *far* exceeds anything ever used before in high-quality audio design. Ironically, while a part of the industry mistakenly espouses low feedback for top quality, what we have here is the exact opposite touted as being even better!
- Since it is the high amount of negative feedback at low frequencies that reduces the 1-bit distortion products to low levels in the audio band, it is not unexpected that we find the distortion products rising at high frequencies, where the corrective negative feedback has actually turned into *positive* feedback!
- The high levels of ultrasonic noise and spuriae produced by an inadequately-filtered 1-bit sigma-delta converter pose a problem for audio amplifiers and loudspeakers, which can generate nonlinear distortion products *in the baseband* when subjected to this type of indignity. One wonders how many of the perceived "differences" noted in Super Audio CD listening comparisons might be due to such nonlinear effects.
- Just as it might be true that one can perceive ultrasonic signals that are correlated with the baseband signal, so too might the low-level, but correlated, distortion products that we have shown to exist *within the baseband* be perceptible, even though they would normally be thought to be below audibility. Further research is needed here.
- Since the Gerzon/Craven "noise-shaping theorem" implies equal areas above and below the unshaped noise floor on a *logarithmic* vertical axis, it follows that there will *always* be a net increase in the total noise power as a result of noise shaping, if the theoretical curve is adhered to. Since the total output power of a 1-bit shaper is constant, it follows that, even in the absence of signal, the noise PSD of a 1-bit noise shaper *cannot* follow the theoretically-prescribed curve. Its total output noise power *must* be less than the curve predicts. Then, if an input signal is added, the noise power at the output *must* drop even further, so causing further noise PSD changes.
- As a general principle, it is undesirable to repeatedly noise-shape a signal as it progresses through various processing stages, since the total high-frequency noise power keeps accumulating. This means that a signal should ideally always be kept at a wordlength at least as long as that required to preserve the signal's baseband noise floor, irrespective of the incoming wordlength. Thus a signal with 120 dB baseband signal-to-noise ratio, should never be processed and stored with less than a 24-bit wordlength all the way through to the final digital-to-analogue conversion, whether it is a shaped 1-bit DSD signal, a shaped 8-bit signal as envisaged in Example (c) of Section 3, or an unshaped 20- or 24-bit signal. This principle should apply to both consumer and professional digital audio. Noise shaping should only be used when storage or transmission limitations *require* data rates to be reduced. (Parenthetically, it seems perverse of DSD to unnecessarily go in the opposite direction by actually *increasing* the data rate!)
- That the 1-bit converter is quite aberrant can also be gleaned from the following arguments. The classical model for the quantization error E of an undithered multi-bit quantizer postulates that it has a power of $\Delta^2/12$. The error power of a 1-bit quantizer is $(\Delta/2)^2 = \Delta^2/4$ under no-signal conditions — three times too large. This error power is, moreover, independent of whether or not the 1-bit quantizer is dithered. [Interestingly, a TPDF-dithered multi-bit quantizer's error power is also $\Delta^2/4$ (= $3\Delta^2/12$).] We have already noted (in Section 3) that the 1-bit quantizer's "noise" power must drop in the presence of any input signal, since its total output power is absolutely constant, thus unavoidably causing noise modulation. We now see that, when outputting any signal, its error power *must be less* than the $\Delta^2/4$ error power of a TPDF-dithered multi-bit quantizer. In addition, its error power spectrum is only approximately white, while that of the TPDF-dithered multi-bit quantizer is *truly* white, as guaranteed by dither theory ([3], [4]). Hence, when modulated, its noise spectrum would be expected, on average, to lie *below*, and parallel to, that of the corresponding TPDF-dithered multi-bit quantizer. (Were the spectrum of E truly white for the 1-bit quantizer, the two curves *would* be exactly parallel, for the output error is just E as shaped by the linear filter $\{1 - H\}$.) This is confirmed by what we found in Section 2, and can be clearly exhibited by overlaying the upper curves of Figs 9 and 10. This is done in Fig. 17, where in each case the input to the Lip7ZP shaper is two half-full-scale (*i.e.*, $\Delta/8$ amplitude each) sine-waves on FFT bins 32 and 48. Note too the considerable discrepancy between the shapes of the upper TPDF-dithered noise curve (which is the *intended* noise curve)

and the lower noise curve, which is that actually delivered by the undithered 1-bit sigma-delta modulator.

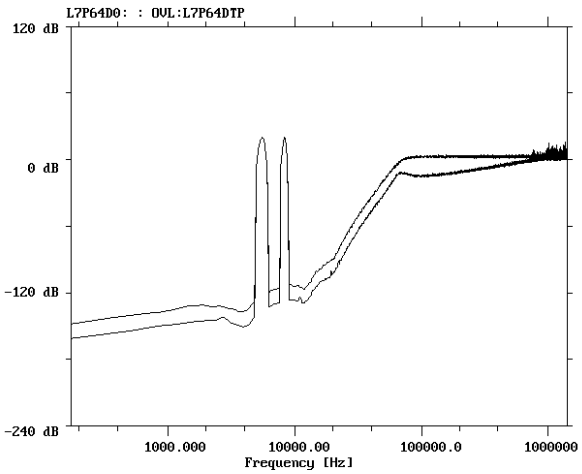

Figure 17. Overlay of the upper curves of Figs 9 and 10, showing that the 1-bit sigma-delta modulator's noise spectrum lies, on average, below that of a TPDF-dithered multi-bit noise shaper. Each trace is the result of performing 64 power spectral averages on the corresponding quantizer output.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] S. P. Lipshitz and J. Vanderkooy, "Why Professional 1-Bit Sigma-Delta Conversion is a Bad Idea", presented at the 109th Convention of the Audio Engineering Society, Los Angeles, CA, 2000 Sept. 22-25, preprint 5188; a Supplementary Information booklet (9 pp.), handed out at the presentation, is obtainable from SPL by e-mail at spl@audiolab.uwaterloo.ca. (This reference should be considered superseded by the present paper.)

[2] J. Vanderkooy and S. P. Lipshitz, "Towards a Better Understanding of 1-Bit Sigma-Delta Modulators", presented at the 110th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2001 May 12-15. (Companion to the present paper.)

[3] 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).

[4] R. A. Wannamaker, S. P. Lipshitz, J. Vanderkooy, and J. N. Wright, "A Theory of Nonsubtractive Dither", *IEEE Transactions On Signal Processing*, vol. 48, pp. 499-516 (2000 Feb.); erratum *ibid*., p. 3258 (2000 Nov.).

[5] 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.

[6] "Super Audio Compact Disc: A Technical Proposal", Philips/Sony white paper, 12 pp. (1997).

[7] S. P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker, "Minimally Audible Noise Shaping", *J. Audio Eng. Soc.*, vol. 39, pp. 836-852 (1991 Nov.).

[8] R. A. Wannamaker, "Psychoacoustically Optimal Noise Shaping", *J. Audio Eng. Soc.*, vol. 40, pp. 611-620 (1992 July/Aug.).

[9] 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.

[10] 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/.

[11] S. R. Norsworthy, "Quantization Errors and Dithering in ΔΣ Modulators", being Chapter 3 of *Delta-Sigma Data Converters: Theory, Design, and Simulation*, S. R. Norsworthy, R. Schreier, and G. C. Temes, eds (IEEE Press, New York, NY, 1997).

[12] 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/.

[13] 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 in the 1st-Order Case

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 [13]. In Fig. 1, this corresponds to setting $H(z) = z^{-1}$, a single-sample delay, and allowing Q 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. In this Appendix, we use lower case symbols to represent the time-domain signal quantities, and subscripts to denote the different sample-time instants 0, 1, 2, …, n, … . So, x_n denotes the input signal X at time instant n; y_n denotes the output signal Y; ν_n denotes the dither signal N; w_n denotes the quantizer input W; and ε_n denotes the quantization error signal E. We 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

$$w_n = x_n - \varepsilon_{n-1} + \nu_n \text{ for } n = 1, 2, \ldots, \qquad (A.1)$$

and

$$y_n = w_n + \varepsilon_n - \nu_n \text{ for } n = 1, 2, \ldots . \qquad (A.2)$$

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

$$\varepsilon_0 = 0. \qquad (A.3)$$

[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

$$-\Delta \le w_k < \Delta, \qquad (A.4)$$

or equivalently

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

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

$$-\Delta/2 + \nu_k < \varepsilon_k \le \Delta/2 + \nu_k \text{ for } k = 1, 2, \ldots . \qquad (A.5)$$

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

$$-\Delta/2 \le x_1 \le \Delta/2. \qquad \text{(A.6)}$$

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

$$-\Delta/2 \le x_k \le \Delta/2, \qquad \text{(A.7)}$$

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

$$-\Delta + \mu \le x_1 \le \Delta - \mu. \qquad \text{(A.8)}$$

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),

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

But since

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

we have

$$-2\mu \le \nu_{n+1} - \nu_n < 2\mu,$$

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, but 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. They can only accept a small amount of dither without overload — *very* much less than full TPDF dither. The 7th-order example Lip7ZP, used to illustrate Section 2, represents the type of shaping needed by a DSD modulator in order to achieve a signal-to-noise ratio on the order of 120 dB up to 20 kHz using 64-times oversampling. We saw that this modulator could accept only 0.17Δ peak-to-peak of RPDF dither without overload. This represents a dither power 18.4 dB less than full TPDF dither! This is the generic situation: the allowable dither cannot fully linearize a 1-bit quantizer.