OPEN ACCESS Freely available online
J. R. Stuart and P. G. Craven, "The Gentle Art of Dithering"
J. Audio Eng. Soc., vol. 67, no. 5, pp. 278–299, (2019 May.).
DOI: https://doi.org/10.17743/jaes.2019.0011

The Gentle Art of Dithering

J. ROBERT STUART,[1] *AES Life Fellow,* **AND PETER G. CRAVEN,**[2] *AES Life Member*
(jrs@mqa.co.uk) (peter@algol.co.uk)

[1]*MQA Ltd., Huntingdon, PE29 6YE, UK*
[2]*Algol Applications Ltd., BN44 3QG, UK*

On the topic of high-performance audio, there remains disagreement over the ways in which sound quality might benefit from higher sample-rates or bit-depths in a digital path. Here we consider the hypothesis that if a digital pathway includes any unintended or undithered quantizations, then several types of errors are imprinted, whose nature will change with increased sampling rate and wordsize. Although dither methods for ameliorating quantization error have been well understood in the literature for some time, these insights are not always applied in practice. We observe that it can be rare for a performance to be captured, produced, and played back with a chain "flawless" in this regard. The paper includes an overview of digital sampling and quantization with additive, subtractive, and noise-shaped dither. The paper also discusses more advanced topics such as cascaded quantizers, fixed and floating-point arithmetic, and time-domain aspects of quantization errors. The paper concludes with guidelines and recommendations, including for the design of listening tests.

0 SETTING THE SCENE

In [1] we suggested that "High Resolution" should be considered an attribute of a complete system in the analog domain (from microphone to loudspeaker)—rather than of the distributed signal or a specific technology.

If the system includes a digital path, higher sample rates enable wider bandwidth and it has been questioned whether a listening preference for wider-bandwidth systems could result from the reproduction of signal frequencies above 20 kHz, or alternatively, whether it might arise as a side-effect of filtering in the chain, such as may be encountered when constraining bandwidth to meet a Nyquist criterion.

In [3] and [4] we introduced a hierarchical method by which high resolution, defined as clear separation of temporal events, can be delivered efficiently. Prior to this there has been a tendency to describe resolution in the digital domain by the proxies of sample rate, bandwidth or data rate. We can't listen to a digital file without first converting it back to analog; this paper continues to consider the entire chain.

A third frequency-domain hypothesis suggests that wider-band signals may cause misbehavior in playback systems (shown to be improbable in [5]).

A fourth possibility, raised here, is that if a chain has defective quantizations in any part of a digital path, then the resulting errors (manifesting as distortion and/or modulation noise) also change with the inter-related variables of sampling rate and wordsize, with collateral consequences.

0.1. Outline of This Paper

This paper is both a tutorial and call to action, reminding about some nowadays-overlooked fundamentals.

Sec. 1 introduces the topic of modulation noise, a type of system error that can disturb or alter perception of background noise, or spatial or low-frequency elements.

Sec. 2 reviews quantization distortion and in Sec. 3 we recap the properties of additive, subtractive, and noise-shaped dither, including maintaining linearity to levels well below the LSB (least-significant bit). Sec. 3 also introduces dithering in the digital (DSP) domain.

Sec. 4 covers more advanced topics, including the analysis of quantization using histograms and synchronous averaging, cascaded quantizers, and a form of distortion we call "washboard," and closes with a topic on fixed and floating-point processors.

In Sec. 5 we look at temporal aspects of dithering: how quantization error settles with time; how "birdies" can be introduced by low-level, low-frequency sounds, and temporal cross-modulation dependent on a prior filter.

In Sec. 6 we examine signal chains in current music production and playback and consider their complexity and likelihood of avoiding quantization defects. The paper concludes with brief reference to audibility analysis and high-resolution topics in Secs. 7 and 8.

Sec. 11 Appendix, looks at discrete dither in more detail.

Due to the wide scope of the paper, many topics are introduced in References, which are grouped by topic in Sec. 13 and a Bibliography in Sec.14.

1 MODULATION NOISE

Sound reproduction systems suffer from commonly measured technical defects, such as frequency response irregularities, non-linear waveform distortion, and background noise. While non-linear distortion depends on signal level, we prefer system noise to be stationary, as a background, independent of the signal.

If an audible noise is stationary, we can perceive it as a separate "object" in the signal. But there are technical defects that can add so-called "modulation noise," an error responsive to the music waveform or envelope.

In analog systems, simple examples include level-dependent noise from the particles on magnetic tape [7] (particularly noticeable at low frequencies) and Barkhausen noise in transformers, microphones, tape heads or other ferrous-cored inductive components. Such modulation noise impacts transparency and tends to impair precise reproduction of low-frequencies or of spatial cues including of reverberation or instrument location.

When a digital path intervenes, other mechanisms can introduce noise-like errors that are dependent on the signal amplitude, rate of change, envelope or modulation index.[1]

One such error source is bit-weight mis-alignment in A/D or D/A converters (see e.g. Fig. 10), or quantization distortion caused by inadequate dither when processing single- or multi-bit or floating-point PCM [8, 21, 22].

From a sound quality perspective, quantization errors are somewhat distinctive, more intrusive on quiet signals, with no immediate parallel in the natural world.

A key problem is that once modulation noise has been added to a signal, in either the analog or digital domains, there is no straightforward way to remove it.

2 DIGITAL SAMPLING

When converting analog audio to a digital representation, the waveform is quantized in time and amplitude. As described in [4, 6, 15, and 14], provided that the conversion uses a suitable filter kernel and an appropriate dither, then time-base resolution can be effectively infinite. The proviso is important; if either the encoding or reconstruction kernels are inappropriate (or missing), or if dither is omitted, then modulation and imprecision will result.

2.1 Quantization Distortion

Quantization distortion is the error introduced when a continuous signal is quantized in amplitude, and its nature changes with sample rate, signal frequency, bit-depth, and signal level.

For an n-bit PCM channel coding the signal range –1 to +1, the smallest representable amplitude increment is $2^{(1-n)}$. We refer to this increment as a *quantum* Δ (also colloquially known as an "LSB").[2]

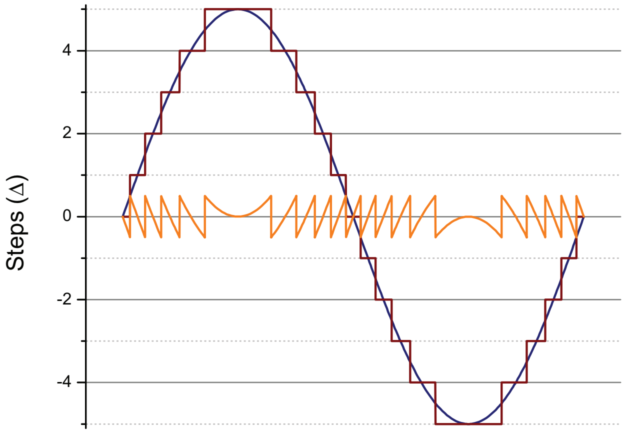

Fig. 1. Showing (navy) an input sinewave of peak amplitude 5 Δ; (wine) is the output of a rounding quantizer. The quantization error is shown in orange.

Fig. 1 shows a continuous sinewave with a peak level of 5 Δ, the quantized version, and the error. Two aspects are evident: the error waveform has sharp discontinuities (which may exceed the Nyquist bandwidth giving rise to aliases) and the error has a peak-peak level of Δ.

In a fixed-point system the peak-peak level of the quantization error remains constant, irrespective of the signal level[3], however, the error spectrum varies considerably with signal level and modulation index.

Although several sample rates (Fs) are in common use, for the purposes of illustration, this paper will concentrate on 48 and 96 kHz, while noting how effects scale to 192 kHz.

Similarly, because of widespread use we give examples of 16- and 24-bit channels, or generalized simulations based on a *quantum* Δ scale.

3 DITHER PRIMER

Quantizing to finite precision creates an error, but a suitable dither can endow some of the attributes of a linear channel by eliminating nonlinear distortion and modulation noise in the moving average (see Sec. 5).

Dither can be additive or, in its purest form, subtractive. In either case a dither noise is added immediately prior to or within the quantization operation. The optimum dither type depends on the context.

Amplitude quantization using dither has been well described in the literature and key references are in Sec. 13.4.

3.1 Additive Dither

To linearize a quantizer, the minimum dither power is a random noise signal uniformly distributed over an interval of width Δ—referred to as RPDF (rectangular probability distribution function) dither.

[1] Modulation index is used here to mean the ratio of the signal frequency to the sample rate.

[2] Quantization gives amplitude steps size Δ; the LSB is the first step on the ladder.

[3] Providing the input exceeds ½ Δ, below which there is no output.

Table 1. Enumerating higher order dithers produced by adding or subtracting "order" number of independent RPDF dithers, the noise power, and linearizing capability of each.

Order	PDF	Noise power[4]	Linearizing
0	none	1 σ^2	0th moment
1	Rectangular	2 σ^2	1st moment
2	Triangular	3 σ^2	2nd moment
3	Parabolic	4 σ^2	3rd moment
–		Etc.	
$\rightarrow\infty$	Gaussian		n/a

As we show later, although this dither linearizes the transfer function, it does not result in a quantization error that is statistically independent of the signal.

Combining multiple independent RPDF streams yields a series of dither types, summarized in Table 1, each with unique statistical properties and differing effects on both measurement instruments and human listeners. An important case is the second-order combination, TPDF (triangular probability distribution function) dither.

The literature strongly suggests that TPDF is the preferred choice for audio—higher order providing no discernible benefit while adding unnecessary noise. The quantization error when using TPDF dither is 4.77 dB higher than the undithered case, however the error is benign in the former and unpleasant in the latter.

There are a few often-overlooked constraints in the theory. For example, there must be zero correlation between the signal and the dither signal, consequently in a system where successive quantizations take place, each must use an independent dither signal.

Another complication arises in the digital domain where the dither itself will be quantized, or at least of finite precision, and so we must pay close attention to the word-sizes and number representation of signal and dither—a topic continued in Secs. 4.8 and 11.

Gaussian noise, which may be present in analog systems or signals at an rms level $> 0.5\ \Delta$, can appear to linearize a quantizer, but cannot guarantee zero modulation noise [15].

Figs. 2 and 3 examine the error spectra for a signal similar to that in Fig. 1 but sampled at 48 and 96 kHz. The quantization error changes character with increased Fs and the error power is distributed over a wider bandwidth. Raising Fs has the benefit of reducing the audio-band distortion while "whitening" the error because the signal/Nyquist frequency ratio is lower.[5]

Comparing the two figures we can also see that the RPDF and TPDF dither noise spectral levels are 3 dB lower at the higher rate. Fig. 3 also includes two examples of noise-shaped dithered quantizers designed for 96 kHz sampling; noise shaping is introduced later in Sec. 3.4.

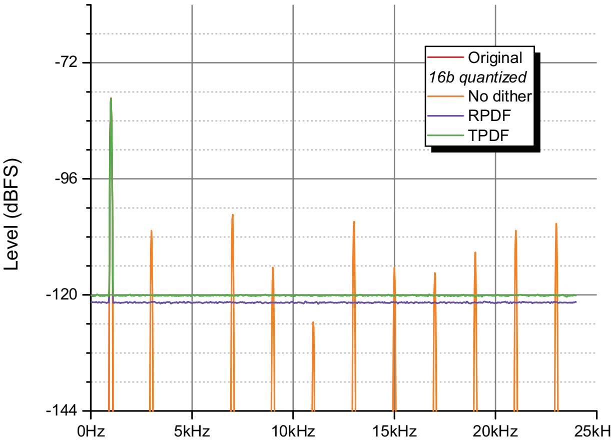

Fig. 2. Showing spectra for a 1 kHz tone @ –79 dBFS at a sample rate of 48 kHz and quantized to 16b with: no dither (orange), RPDF (violet) or TPDF (green).

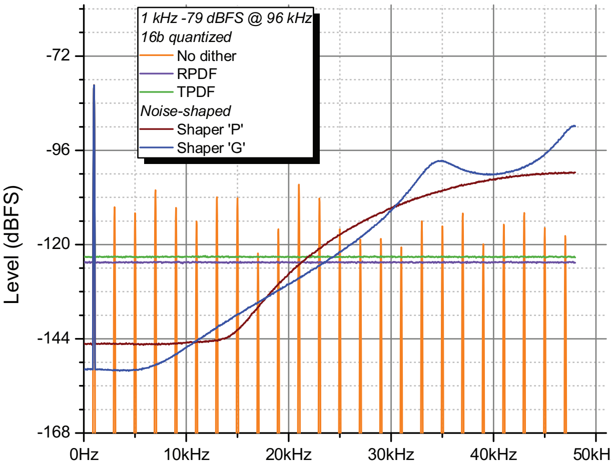

Fig. 3. The 1 kHz signal sampled at 96 kHz. In addition to quantizers with no (orange), RPDF (violet), and TPDF (green) dither, two low-complexity noise-shapers using TPDF dither are included. Shaper "P" increasing dynamic range by just over 3 bits at 4 kHz (wine); shaper "G" by Gerzon (blue) [6, 29] gives more advantage at 4 kHz but more out-of-band noise.

3.2 Resolution below the LSB

TPDF dither at the 16th bit has average rms power of –93.32 dB, meaning that the channel permits a signal/noise ratio of 93.32 dB. This fact occasionally confuses when we show lower levels in spectral plots.

If the noise has a flat spectrum, then the NSD (noise-spectral density) should be –137.12 $\text{dB}/\sqrt{\text{Hz}}$ over a 24-kHz bandwidth. In Figs. 2–5, the FFT bin-width is (48 kHz/2048) or (96 kHz/4096) = 23.4 Hz; after correcting for FFT bin-width and windowing, the NSD of the (–120 dB measured) TPDF dither plot is –137.1 $\text{dB}/\sqrt{\text{Hz}}$, showing an exact correspondence [27].

RPDF, TPDF, and higher orders of dither have the property of linearizing below the LSB level, which can be revealed by an averaging meter or FFT analysis. Fig. 4 shows a measurement of a 4 kHz –110 dB tone in a 16-bit channel quantized with TPDF dither; the signal is clearly visible.

[4] $\sigma^2 = \Delta^2/12$

[5] Increasing Fs might reduce the "impact" of the error, leading to a "preference" for the higher Fs.

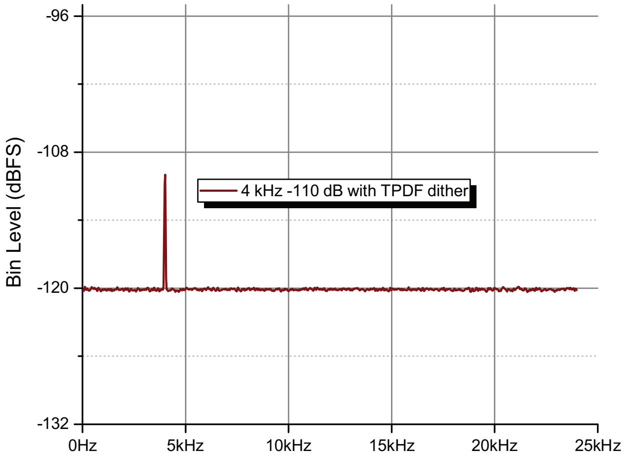

Fig. 4. Showing a 4 kHz tone @ –110 dBFS, reproduced in a 16-bit channel using TPDF dither. Without dither, there is no output for an input signal below –96 dB.

In fact, using synchronous averaging, we can demonstrate that this signal remains detectable well below the –120 dB noise spectrum. This result is also true for a quantizer using RPDF, although in that case there will be modulation noise. With no dither, a rounding ("mid-tread") quantizer gives no output for a signal below –96 dB [16].

This property of a TPDF-dithered quantizer is important in the gateways between analog and digital and throughout a delivery chain.

The fact that we can resolve a measurement of a tone at these levels does not necessarily mean that it will be audible—human hearing is quite non-linear with detection thresholds depending on level, frequency, and adjacent sounds. Methods for estimating the audibility of these low-level signals and errors are given in [27] and [6].

In Figs. 2–5, rather than plotting NSD, we chose an FFT with 23 Hz noise bandwidth because this corresponds to the narrowest (equivalent rectangular bandwidth) ERB for human listeners [27].

3.3 Subtractive Dither

Subtractive dither was first described by Roberts [20].

In some systems we can add dither prior to quantization to optimize a transmission channel, but then subtract the same dither noise when the audio signal is reconstructed at high resolution. Providing a suitable dither is chosen, subtraction can result in an error which is both the minimum possible power ($\Delta^2/12 = \sigma^2$) and uncorrelated to the signal.

The minimum dither that achieves this result is 1 Δ RPDF, but if we anticipate some listeners without a decoder (i.e., no subsequent dither subtraction), then using TPDF dither ensures good sound, but for them the quantization-related noise will be 4.77 dB higher. In distribution there are significant benefits to subtractive dither, but a synchronous end-to-end coding system is needed [12, 17].

Subtractive dither can also be used to improve analog interfaces, for example around an A/D or combinations of coupled D/A converters such that dither applied to the input can be subtracted in the summed analog output.

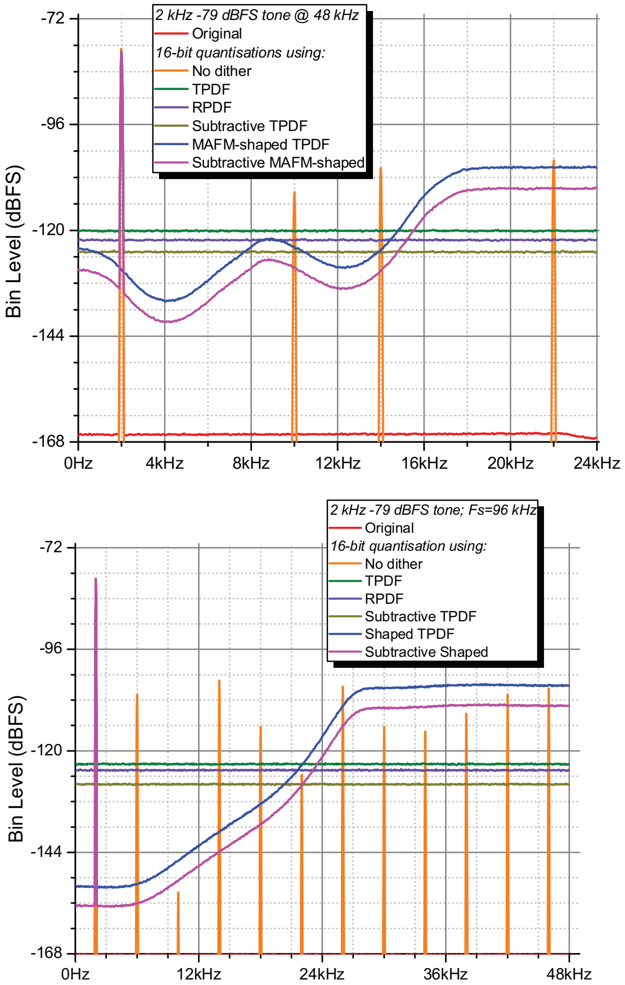

Fig. 5. Showing noise-shaped examples at 48 kHz (upper) and 96 kHz (lower) in a 16-bit channel (23.4 Hz analysis bin). In each the original signal (red) is quantized to 16 bits: without dither (orange); RPDF (violet); TPDF (green); subtractive TPDF (dark yellow). In both plots a noise-shaper is used with additive (blue) and subtractive (magenta) dither.

3.4 Noise and Error Shaping

While quantization errors can't be eliminated, it is possible to redistribute their spectrum. This technique, known as noise-shaping, is extensively covered in the literature; see Sec. 13.5.

By feeding back the error around a quantizer, via a suitable filter, its error can be shaped to be lower where the ear is most sensitive, e.g., around 4 kHz.[6] Some examples of noise-shaping are included in Fig. 5.

Noise-shaping can be used with or without additive or subtractive dither. Although in some cases designers try to omit the dither (error-shaping), it is necessary to dither the noise-shaped quantizer to guarantee transparency.

[6] Noise shapers must conform to the performance criterion first described by Gerzon and Craven in [25]. Lower in one part of the band must be balanced by higher elsewhere.

3.5 Digital Dither

In Sec. 2 we introduced quantization in the context of conversion from analog to digital. But what if the signal we start with is already digital and properly dithered?

One benefit in the digital domain is that the bit-weightings are perfect; each maintains a precise 2:1 relationship with its neighbor. Still, whenever a signal is processed in the digital domain, by even the simplest gain-change operation, the output signal ends up with wider wordwidth than the input.[7]

At each step the dither must be adequate. But is TPDF always the best? And at which level do we apply it? What if the signal is processed without a large accumulator? What about systems where the signal is moved from fixed to floating-point and back? These thorny topics are tackled next and also in Sec. 11.

4 ADVANCED DITHER TOPICS

If it is important to use a properly dithered quantizer at each step, can incorrect application of dither be revealed by subsequent tests?

If we have a recording but no access to the equipment that produced it, certain faults can be revealed by histogram analysis, though this is fragile as noted below. If we do have access to original equipment and can feed it with test tones, synchronous averaging can reveal faults more robustly.

4.1 Histogram Analysis

Histogram analysis records the number of occurrences of each digital code in a recorded signal. A 16-bit signal has 65536 potential values, so the histogram occupies 65536 bins. On typical music signals, the low values have a higher probability, and the histogram generally has a smooth bell-shaped appearance.[8] Fig. 6 shows analysis of a commercial CD.

Zooming in we may see “structure” in the histogram. Sometimes the pattern is complicated and difficult to interpret, but there are also classic faults that can be revealed. As a trivial example, if a supposedly 16-bit signal exercises only 15 bits, the odd-numbered bins will register as zero, cf., Fig. 7.

Quantization of a high-resolution signal using truncation-towards-zero (not recommended, see Sec. 11.4) will result in over-population of the zero value by a factor of two. Unfortunately, in this case only some aspects are improved by dither. Values near zero will still have slightly increased occupancy and the sound will be compromised by a kink in the transfer characteristic, somewhat similar to crossover distortion in Class-B amplifiers.

Another common example is a 16-bit recording that is deemed to require gain adjustment before release. The relevant multiplication will generally yield a result having more than 16 bits, so a further quantization may be required.

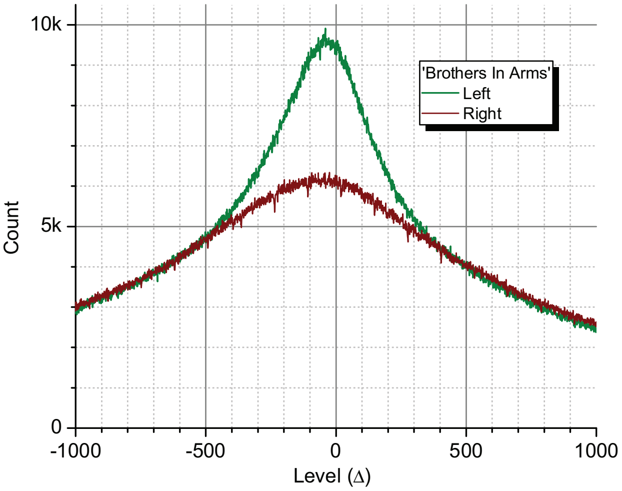

Fig. 6. Histogram of typical well-behaved CD recording.

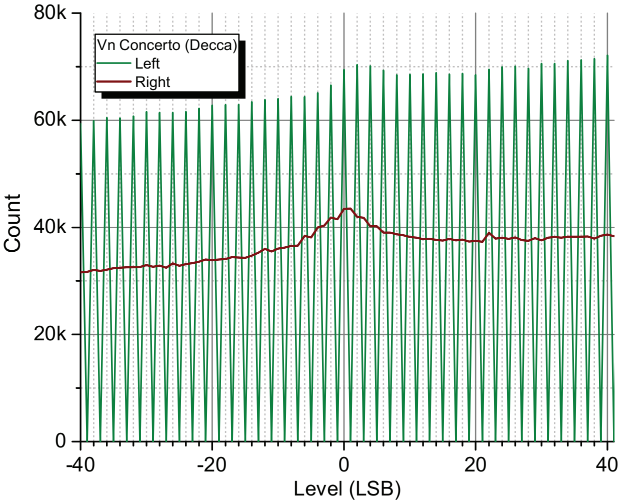

Fig. 7. An example of a CD release where we deduce the left channel was amplified without dither; the LSB is not exercised, resulting in double occupancy rate of the 15th bit.

Assuming a quantized input, a gain increase followed by an undithered quantization will cause missing codes in the histogram, as noted in Fig. 7. Conversely an undithered gain decrease results in spikes in the histogram, as we can see in Fig. 8 (a not uncommon real-world example) and Fig. 9 (a simulation using full scale RPDF white noise: 100,000 samples covering the region of the plot so an average density of ∼1020 samples per quantum).

In Fig. 9 the spike at zero is larger because, contrary to recommendations, the quantizer performs a truncate-towards-zero operation. As illustrated, smaller gain changes result in more widely spaced spikes or holes in the histogram (for negative or positive gain respectively), so the sign and magnitude of the gain change can be deduced from the histogram. However, it is not unknown for several features of this type to appear in a histogram when a recorded stream has been subjected to several processing stages: in some cases, it may be difficult to disentangle the various

[7] For example, in a DSP engine, perfect multiplication of two 24-bit numbers requires a 48-bit accumulator (actually 47 bits would be sufficient). To bring this back to a 24-bit number we need to quantize and, when we do, the correct procedure would be to use independent dither, each and every time.

[8] Assuming silence has first been excluded from the analysis.

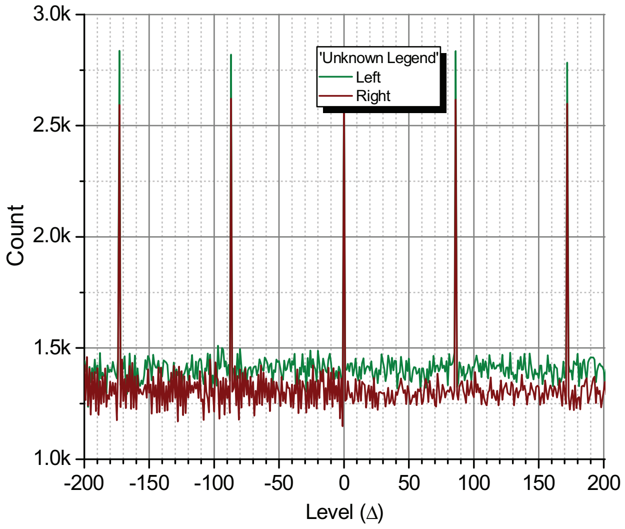

Fig. 8. A recording where the level seems to have been reduced without dither (spike spacing suggests by 0.1 dB).

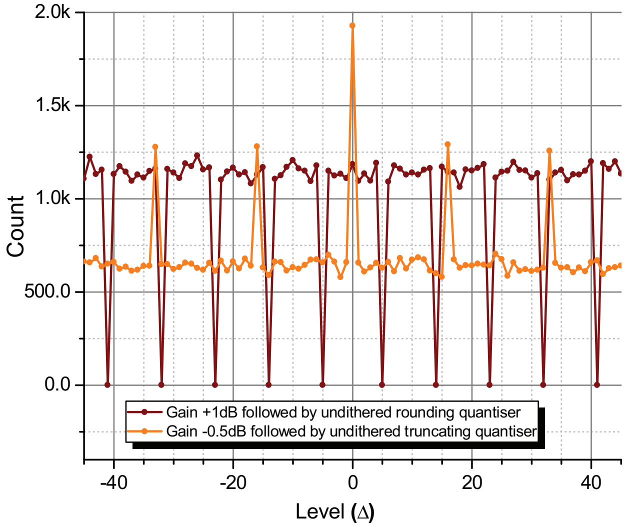

Fig. 9. Histograms for two cases: (Wine) +1 dB gain with undithered rounding quantization; (Orange) –0.5 dB gain with an undithered truncating quantizer.

manipulations. Subsequent signal processing such as equalization may result in a smoothed-out histogram that does not reveal prior undithered quantizations, which is why we describe the histogram analysis as "fragile." In practice histogram analysis can reveal quantization problems, but it cannot be used to prove that there are no problems.

4.2 Synchronous Averaging

The synchronous averaging [SA] method applies a periodic test signal to a device's input and accumulates the output over successive periods. Noise from dither or other sources can be reduced to insignificance by the averaging so that small deviations from linearity can clearly be seen.[9]

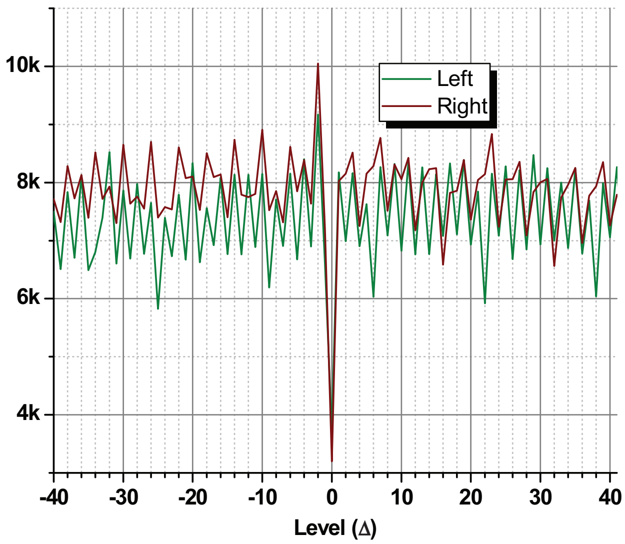

Fig. 10. An example histogram from a recording that reveals two more fundamental problems (typical in early A/D converters) where the MSB and other bits are mismatched, and droop occurs in a preceding analog sample-and-hold.

One might at first think to use a repeating linear ramp (a sawtooth wave) as the test signal exercising each input quantization level in turn[10], but this is not useful if there may be equalization or filtering that generate pre- or post-responses when excited by the sharp corners of a sawtooth, such as to complicate the interpretation of the linear parts.

It seems better to start with a sinewave so that irregularities in the output can unambiguously be ascribed to nonlinearity. As a sinewave is almost linear near a zero-crossing, a sinewave whose amplitude is somewhat larger than the range of signal amplitudes it is desired to explore can serve in place of a ramp. Interpretation is easiest if the frequency is low, so that the individual input quantization levels are exercised in turn, as would be the case with the ramp.

Fig. 11 provides an example of the SA method's ability to reveal extremely small deviations from linearity. It relates to a professional digital audio workstation exercised at 192 kHz with a –90 dBFS 24 Hz 24-bit sinewave as a test signal. The device was asked to provide a –3 dB gain and the plot shows a synchronous average over 10 minutes (14,400 periods) from the workstation's 24-bit output.

One can see a general waviness, repeating about 5 times per input Δ, indicating "something curious" happening internally at about the 26th or 27th bit We can also see that the general trend of the right half of the plot (corresponding to positive output signals) is offset by about –0.05 Δ from the left half, suggesting a truncation-towards-zero at the 28th bit.

[9] SA provides a direct picture of an amplitude nonlinearity. By contrast, spectral analysis with a long window (hence fine frequency resolution) can isolate minute tonal signals from surrounding noise and presents the consequence of non-linearity in terms of unwanted tones, as in Fig. 21.

[10] E.g., a ramp incrementing by 1 Δ per sample period.

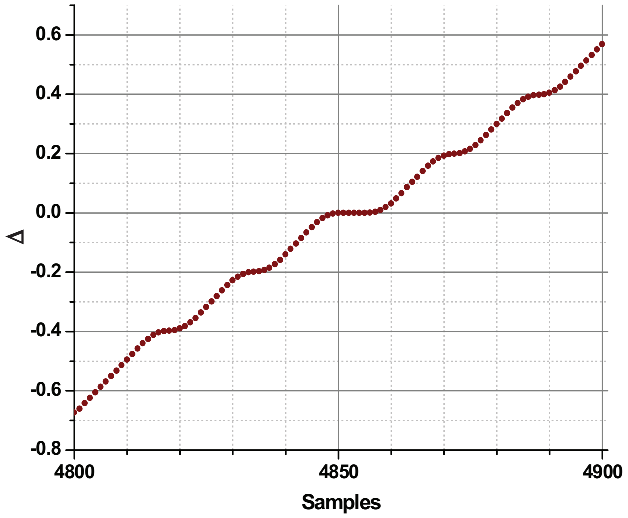

Fig. 11. An example of the SA method's ability to reveal extremely small deviations from linearity. A professional digital audio workstation was exercised at 192 kHz with a 24 Hz –90 dBFS 24-bit sinewave as a test signal. The device was asked to provide a –3 dB gain and the plot shows a synchronous average over 10 minutes from the workstation's 24-bit output.

4.3 Histogram versus Synchronous Average

In general, the histogram can be recommended as a quick way of revealing problems in the output quantization, with the caveat that faults earlier in the processing can easily be obscured by subsequent linear filtering or by the addition of dither at a later point.

A histogram is plotting bin occupancy as a function of *output* level; conversely synchronous averaging uses the sample number as a proxy for *input* level, making use here of the almost-linear slope of the input sinewave near a zero-crossing. A static nonlinearity anywhere along the chain will be visible on the synchronous average, possibly smeared by linear filtering at a later point, but still visible unless the filtering is drastic. Adding dither at a later point in the chain may add noise to an SA plot but will not reduce the amplitude of the nonlinearity's signature. The noise can be reduced by averaging over longer period, or if that is not practical by using a test signal of higher frequency, simultaneously reducing the amplitude to ensure that a contiguous set of input codes is exercised.

4.4 Mean-Square-Error and TPDF Dither

Plain synchronous averaging suppresses the noise and reveals the mean. To investigate noise modulation, we can accumulate the square of deviation from the mean as well as the mean itself. Fig. 12 shows plots accumulated over 1 minute, relating to a 48 kHz 16-bit rounding quantization of a 240 Hz sinewave of amplitude –79 dBFS. The blue trace in the upper plot is of the mean value when the quantizer used RPDF dither, seemingly perfect; the lower plot shows the corresponding mean-square error (MSE). Comparing plots, we see that the MSE goes to zero when the input sample value is an exact integer number of Δ steps, resulting in modulation noise.

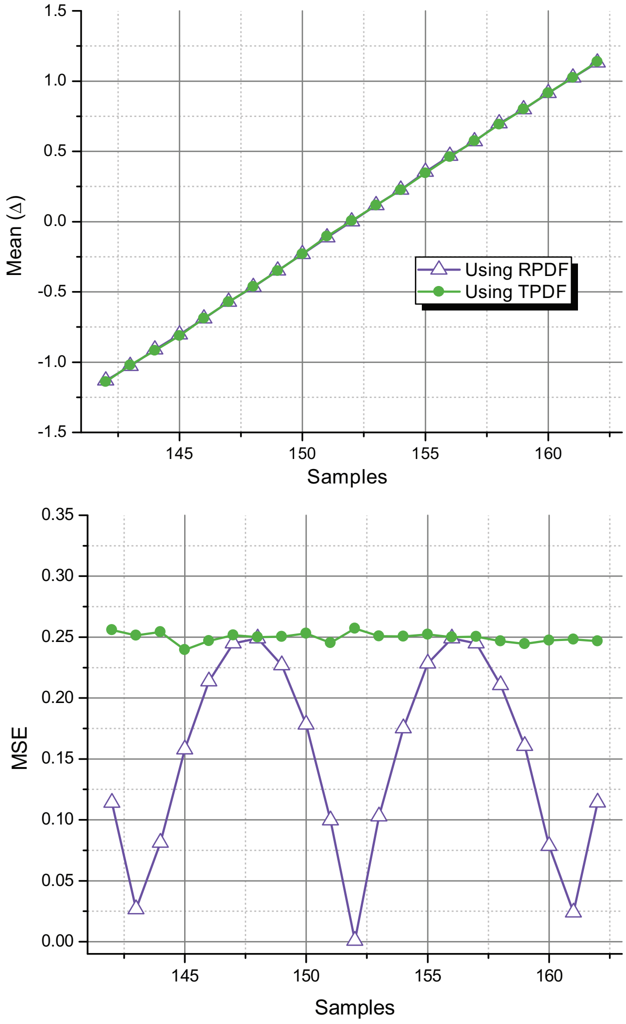

Fig. 12. SA (1 min.) for a 48 kHz 16-bit rounding quantization of a –79 dBFS 240 Hz tone. Upper: the apparent perfect average level using RPDF (violet) and (overlaid) TPDF (green) dither. Lower: MSE for the same conditions.

On the same plots, in green, we can see the impact of using TPDF dither. Linearity is unchanged (perfect) but now the noise is constant to within the statistical fluctuation in the measurement.[11]

4.5. Cascaded Quantizers: Washboard and Moiré

Figs. 13 and 14 relate to quantization of a downward linear ramp. The orange traces relate to a quantization step size 1 Δ; the navy curves relate to a gain change of –0.915 dB followed by a similar quantization. Magenta relates to a cascade of both quantizers (including the gain change).

In each plot the three upper traces showing the system output are hardly distinguishable. The lower traces show,

[11] This phenomenon was extensively investigated by Lipshitz et al. [14] who identified that the simplest practical way to achieve zero modulation noise was to use TPDF dither with a pk-pk amplitude of two quantization steps (2 Δ).

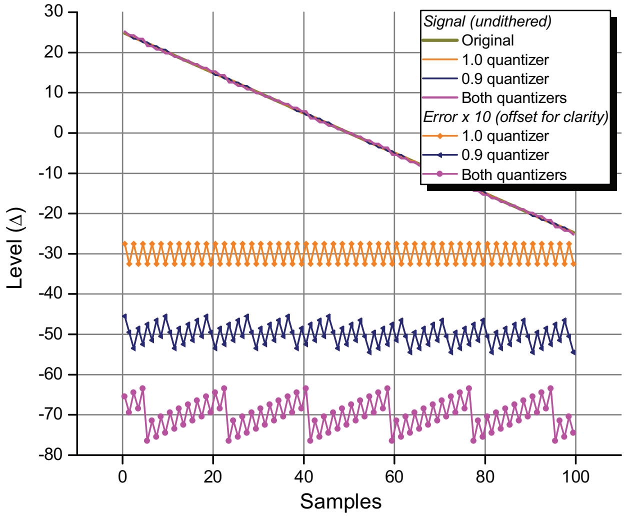

Fig. 13. A linear ramp (olive); (orange) quantized with stepsize Δ without dither; (navy) after gain and requantization having effective quantization step size 0.9 Δ,[12] and finally (magenta) with both quantizations cascaded. In each case the quantization error is offset and multiplied by 10 for clarity.

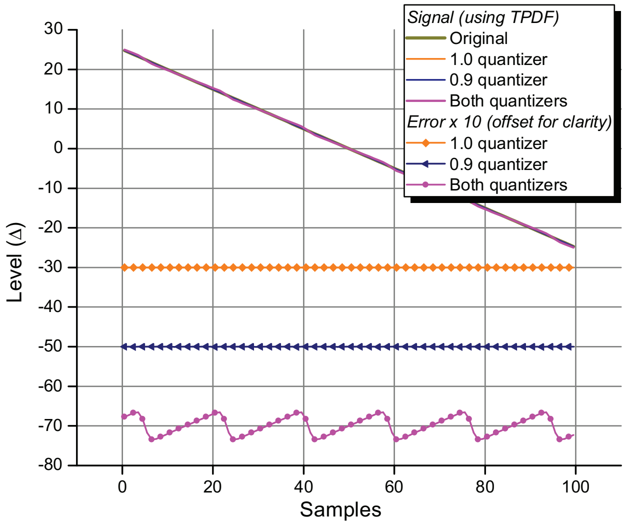

Fig. 14. Here the plots are repeated to show the mean (expected) value when TPDF dither of pk-pk amplitude 2 Δ is added before the quantizations.

in each case, the quantization error (multiplied by 10 and offset for clarity).

Fig. 13 is undithered while Fig. 14 is a simulated synchronous average showing the mean (expected) values when TPDF dither of pk-pk amplitude 2 Δ is added before the first quantization.

The TPDF dither has the correct amplitude to perfectly linearize the plain quantizer and to linearize almost completely the combination of gain change and quantizer (the maximum deviation of about 0.002 Δ being too small to see on the plot). However, for the two quantizers in cascade (with gain change but no dither prior to the second quantizer) there is a wavy component in the transfer characteristic; see Fig. 14 (magenta trace).[13]

The wavy component represents distortion which we refer to as "washboard distortion," on account of the visual similarity of the transfer characteristic to the ribbed surface of the "washboard," a household item formerly pressed into service as a musical instrument [37].

The ribs on a washboard represent a beat pattern between the two effective quantization stepsizes, analogous to the Moiré patterning seen when a pixelated image is resampled with a slightly different pixel spacing [38].

If the pixelated image is firstly blurred just enough to hide the pixels before resampling, the Moiré pattern will disappear also, which is vastly better than trying to get rid of the Moiré after the event. So, with quantization each quantization should be given its own dither. If an equipment applies a gain change to a digital input signal and requantizes without dither, the washboard will result. Such a situation can be ameliorated by applying dither before presenting the signal, but this would typically require a much larger amount of dither, especially if there are several quantizations and perhaps EQ stages creating a complex pattern. It seems plausible that a shaped dither having a high spectral density only in the ultrasonic region, would be the best way to "lubricate" such a situation without compromising the signal-to-noise ratio in the audio band.

A gain change that is only "small" does not result in a smaller washboard effect but rather in an increased spacing between the ribs of the washboard. For example, a gain reduction of 0.1 dB is sometimes used to adjust the peak excursion of a recording. If undithered, this adjustment results in a spacing of 87.4 quanta between the ribs and the effect is also visible in a histogram if not obscured by subsequent processing, cf., Fig. 8.

The washboard effect does not necessarily increase the total quantization noise power beyond what one would ordinarily expect; it is the slower dependency on signal amplitude that is the problem, greatly increasing the likelihood of correlation with the original signal when at a low level, for example in a reverberation tail.

4.6. Can Analog Noise Provide Necessary Dither?

Analog noise generally has a Gaussian probability distribution and if (as is typical) its rms amplitude is greater than one or two steps of a single quantizer then the quantization of a signal containing the noise will be substantially linearized, i.e., the mean value will be correct and the MSE will be substantially constant [15].

From this truth has grown the myth that if an equipment is generally fed from an analog source there will be "plenty"

[12] A gain 10/9 (≈ + 0.915 dB) is applied, then the signal is requantized with stepsize Δ, then, for plotting purposes, a gain of 9/10 (≈ – 0.915 dB) is applied to restore the original signal level. This sequence is equivalent to quantization with a stepsize of 0.9 Δ.

[13] The qualitative similarity to Fig. 11 will not have escaped the reader's attention.

of dither to cover internal quantizations that therefore do not need to be dithered individually.

This reasoning is of course false: it does not consider the likely accumulation of washboard distortion.

4.7. Dither versus Bit-Depth

As noted earlier, dither should ideally be applied at each significant quantization but unfortunately washboard effects will not be eliminated if the same dither is applied to each quantization in a chain. For the system to work as intended, each dither should be statistically independent of the others. Generating statistically independent multibit dither streams is, however, computationally intensive and, in some cases more expensive than the signal-processing itself. An alternative is to increase the bit-depth sufficiently that quantization distortion is below any plausible threshold of audibility. To determine the required bit-depth is extremely difficult—at very low levels the distortion, in isolation, is not directly audible as such. However, over many years of working on this topic, we have anecdotes from more than one source that a new algorithm or piece of equipment has been judged sonically "not quite right" and that, on investigation, an undithered quantization has been found around the 22nd, 23rd or even 24th bit; and that smiles returned to the listeners' faces when this was corrected.

If quantizing an analog or other high-precision source, quantization distortion will reduce in amplitude commensurately with the bit depth, for example, 24 dB lower at 20 bits than if the quantizer were at the 16th bit. Further, errors from a single undithered quantizer are likely to be less correlated with the original audio if quantizing to a higher bit depth, since the quantization grid will then be finer so each sample is more likely to arrive at a "random" position relative to the grid.[14]

With only 16 bits a reverberation tail from a low-noise original recording can easily remain in a similar position relative to the grid for several samples in succession, so a similar quantization error will apply to each sample, resulting in a "gritty" sound.

With more than one quantizer in the chain, increasing bit depth will reduce the amplitude of quantization artefacts, but the decorrelation advantage cannot now be guaranteed because of the likely "washboard" effects. To avoid such effects, one may either apply an independent dither prior to each quantization or increase the bit-depth to the point where quantization errors become insignificant, but that strategy is inefficient for distribution.

4.8. Fixed and Floating-Point Arithmetic

The theory of quantization is straightforward in principle, but correct implementation typically requires precise knowledge of the platform. For example, RPDF where the dither is a random number between 0 and 1 requires a quantizer which truncates towards $-\infty$ (to cancel the dither's DC offset of 0.5 Δ).

By contrast, with TPDF generated by subtracting two RPDF random numbers, then the quantizer preferably uses a rounding operation to avoid a DC offset.

DSP operations are often carried out in fixed-point processors. A well-known 24-bit platform can multiply two 24-bit numbers, the result ending up in a 56-bit accumulator. After operations are accumulated it will be necessary to quantize to a 24-bit number and it is important that the dither number has enough bits, as shown later in Sec. 11.6, and is scaled correctly in the 56-bit accumulator.

Other fixed-point platforms exist including asymmetric engines in integrated circuit filtering or gain stages (where signal and coefficient word-sizes differ). We have seen processors operating with 16, 18, 24, and 32-bit word widths and intermediate numbers in silicon implementation.

Much of the theory of quantization has been based on integer models. In the last two decades it has become more common to encounter signal paths and processors that use floating-point arithmetic. A significant advantage of a floating-point (FP) representation is the large dynamic range, so designers do not have to worry about overshoots in internal processing. Another is that quantization artefacts tend to reduce with signal level so dithering of internal processing is arguably less necessary.

As examples, 32-bit FP is seen in studio workstations, audio desktop software, plug-ins, and intermediate files. 32-bit FP is also used in computer or "phone" operating systems.

As specified by the IEEE 754 standard, a 32-bit float can represent all the values that a 24-bit or even a 25-bit fixed-point integer format can represent, and more too. It is therefore tempting to assume that the floating format is "clearly better." However:

- Adding dither within the format is problematic because without special hardware some steps may not be accessible (well described in [12] and [13]).[15]
- Conversion from a 32-bit float back to a 24-bit fixed[16] requires care and understanding and perfection may be impossible in some architectures.[17]

[14] This is the converse of the observation made in the last paragraph of Sec. 4.5.

[15] A key problem is having access to internal signals during arithmetic operations. In [13] the authors suggest that custom hardware offers the only "perfect" solution.

[16] This operation is inevitably needed to create a distribution file or to pass over an interface, including to a converter.

[17] To convert from 32-bit float to 24-bit fixed, dither (preferably TPDF) needs to be added appropriate to the 24-bit stepsize because, for small signals, the floating resolution is much higher. The sum of signal plus dither should ideally not be stored back to a 32-bit float before rounding to integer, otherwise at high signal levels the dither is compromised by the coarser quantization, leading to the complications of "Discrete Dither" and potentially giving rise to a "23-bit-ism" (c.f., Sec. 11.4). If programming in a high-level language it may be hard to know the details of intermediate storage. One may ideally convert to 64 bits first to avoid these problems. Failing that, an undithered conversion to 32-bit fixed, followed by a dithered rounding to 24 bits is a possible option.

- Floating-point operation does not solve all quantization problems; e.g., when recursive processing is implemented, such as the IIR filters illustrated in [35] or in plug-ins for adding equalization or reverberation.[18]

Some platforms now use 64-bit FP and there is very much less risk. In this case the mantissa is 53 bits, so computational noise and undithered quantizations can be at a very low level in the FP domain. Even so, dither is needed when transitioning back to a 24-bit fixed-point representation.

It is possible that some (not all) of the non-linear behavior analyzed in Fig. 11 manifests inadequate conversions from 32-bit FP to 24-bit fixed.

5 TEMPORAL ASPECTS OF DITHER

Although presented in the theory as an "amplitude" topic, quantization and dither introduce or manage an error signal that evolves causally over time. Put another way, quantization may be a bridge between amplitude and temporal errors, as we shall illustrate in this section.

Dither is implicitly an averaging process and applies neatly to continuous signals; it seems to be effective at removing errors we hear and measure providing the 1st and 2nd moments of error remain independent of the signal. These conditions are fully satisfied with subtractive dither and sufficiently satisfied by the use of TPDF additive dither in a single quantization [16].

The theory of dither is founded on a statistical approach, the unstated assumption proposing commonality between measuring or analyzing instruments that implicitly average stationary signals (e.g., distortion meter or FFT analyzer) and the human listener. There is a good basis for assuming some averaging or integration in the auditory process, particularly on tones; our hearing process involves auditory pathway processing on various timescales between 10 us and 250 ms (and longer in the cortex). (See bibliography in Sec. 14, particularly 14.4.) On steady tonal signals TPDF is a great success.

However, most useful sounds are not stationary and for survival and for everyday function human perception groups sound elements together into "objects" with locations. The timescale is rapid, and percepts are continually confirmed or rejected over time.[19] If a signal contains a modulation noise, even at a very low level, we might "attach" that noise to an "object."

The error in a dithered quantization is causal; its moving average follows the stimulus and may or may not converge.

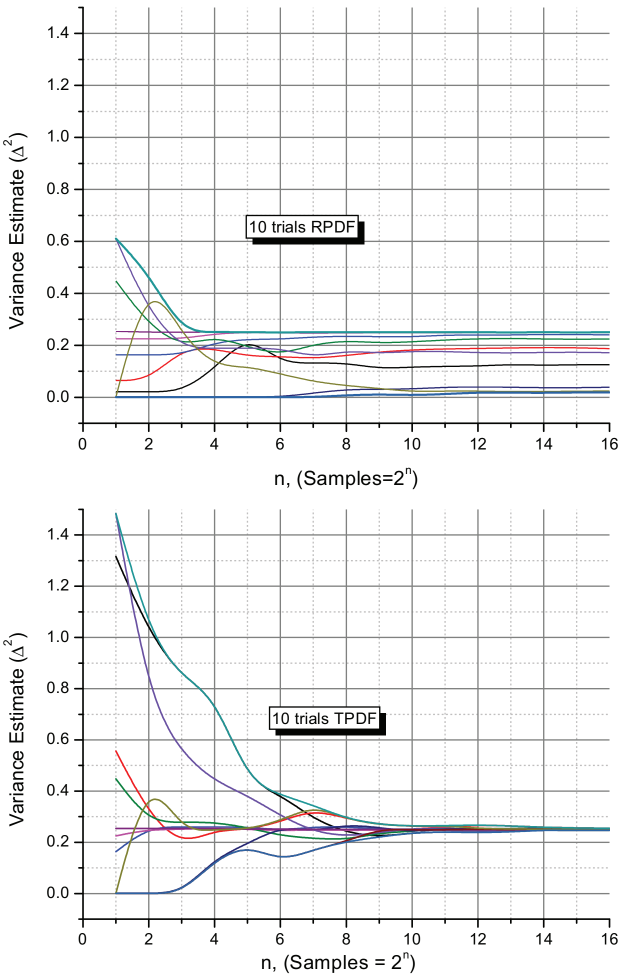

Fig. 15. Time-course of MSE variance for 10 trials each using (upper) RPDF and (lower) TPDF dithered quantizers.

5.1. Pursuit and Convergence

What can we discover about the time course of quantization distortion? Does dither change the picture? The literature says very little on this topic. In [19] examples were given for dithered quantizers and a variety of stimuli.

Strictly speaking, the sample rate should be an independent matter when asking, "how many sampling intervals are required for the output to settle following a small perturbation."

In Fig. 15 we repeat the simulation of [19] for RPDF and TPDF dither, observing the MSE (mean-square error) over 64 K subsequent samples.

Using TPDF dither, despite some higher initial values, the MSE error fully converges to the expected value of ¼ Δ^2 after around 1000 samples (23 ms @ 44.1 kHz). With RPDF dither, for some input levels the output never converges, and we see an occupied region between 0 and ¼ Δ^2 corresponding to modulation noise. This is another view of the mechanism in Fig. 12 (lower).

We are interested to see how the MSE behaves with higher sample rates and with other dithers.

Fig. 16 plots the maximum and minimum MSE recorded for each dither type over a very large number of trials,

[18] Further, not all processors conform to the IEEE standard. We have been disappointed to observe that a direct conversion of a 24-bit value to a 32-bit float and back (without dither) does not always return the original value. These conversions would be exact under IEEE rules.

[19] For example, in a similar manner to synchronous averaging.

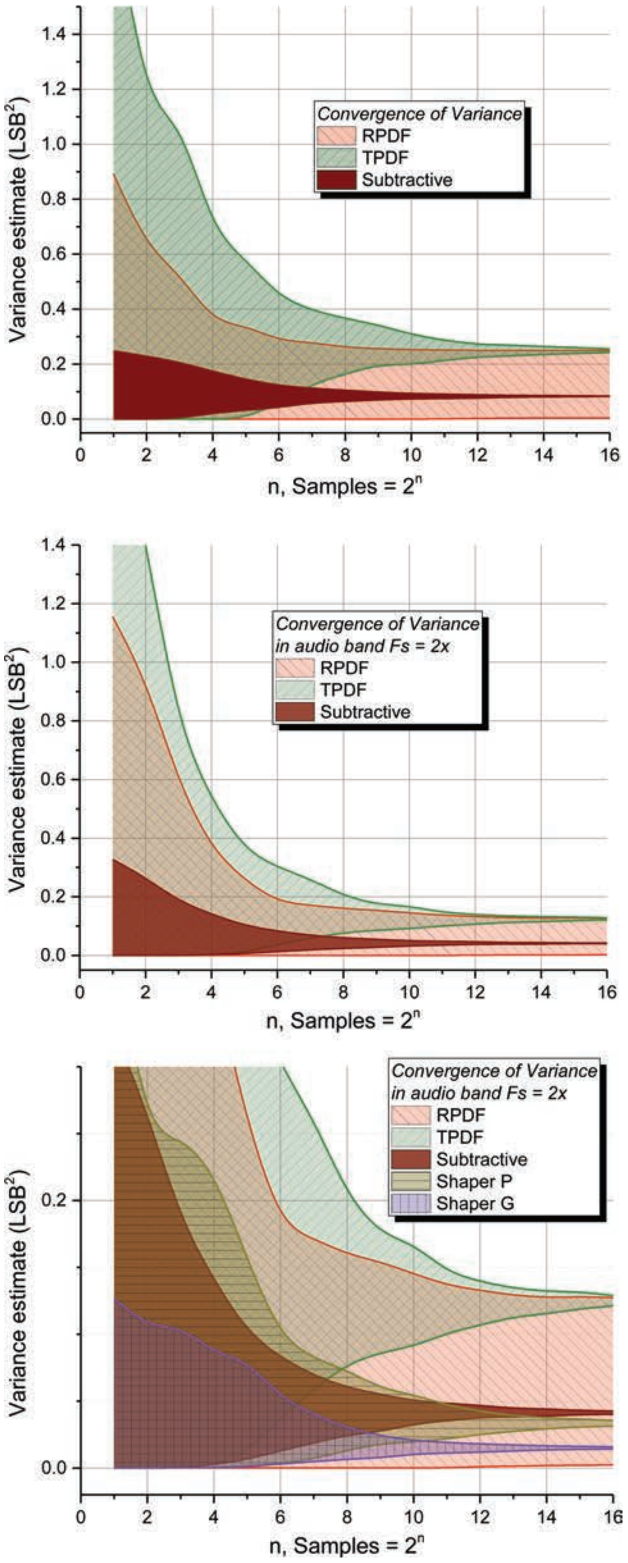

Fig. 16. Showing the existence region of modulation noise as the area between the minimum and maximum MSE seen in a large number of trials of the time-course of dithered quantizations—as illustrated individually in Fig. 15. The top and middle graphs show RPDF, TPDF, and subtractive dither in 48- and 96-kHz channels. The bottom graph, of the 96-kHz channel, includes two noise-shapers seen in Fig. 5.

so rather than seeing how each value evolves, we can concentrate on the region of uncertainty. For each plotted max./min. pair, the enclosed area indicates the existence region for modulation noise.

Fig. 16 (top) plots the convergence of variance for RPDF, TPDF, and subtractive dithers in a 48-kHz channel. As before, RPDF never converges, whereas TPDF and subtractive dithers converge to the expected values of ¼ Δ^2 and $1/12\ \Delta^2$ respectively. Note that the MSE for subtractive dither converges in a quarter of the time compared to TPDF.

The lower two graphs of Fig. 16 plot the minimum and maximum variance for several dither types, viewing the output of a short minimum-phase filter of bandwidth Fs/4, allowing us to consider the MSE in the audio band when the sample rate is 2x (i.e., 96 kHz).

In the middle graph we see TPDF and subtractive dithers converging to $1/8\ \Delta^2$ and $1/24\ \Delta^2$ respectively. Convergence is comparable to the 1x (48 kHz) case, but since the sample rate is 96 kHz, settling is in half the time.

The bottom graph repeats the middle, with the vertical axis zoomed into the region 0 to 0.3 and adds the two noise-shapers designed for use at 96 kHz (plotted in Fig. 3). Two phenomena can be observed: first, the noise-shapers converge to different levels, reflecting their mid-audio-band dynamic-range advantage and, second, their rate of convergence is different—shaper "G" converging 4-times faster than "P"—and more rapidly than subtractive dither.

Although that is a satisfying result, it does remind us that the errors take significant periods of time to settle, from 43 ms for TPDF at 48 kHz to 1 ms for a noise-shaped TPDF quantizer at 96 kHz. Without commenting on the significance, it is interesting to contemplate that simply increasing sample rate will proportionally reduce the duration of the tail.

It is also interesting that the convergence time is inversely proportional to the bandwidth of the noise: hence the reduction in convergence time with higher sample rates.

However, in the case of shaped noise, much of the energy might have been concentrated in a narrow band (perhaps close to the Nyquist frequency), leading to a longer convergence time. It is then helpful to introduce the concept of an "Equivalent Rectangular Bandwidth" (ERB)[20] which, for a conventional rms detector, would be given by:

$$ERB = \frac{(\int p(f)\,df)^2}{\int p(f)^2 df}$$

where *p(f)* is spectral density at frequency *f* and the integrals are taken over the Nyquist range. Reduced convergence time is obtained by maximizing the ERB which in turn argues for substantially flat top to a noise shaping curve (cf., Fig. 5). However, it is yet to be determined whether the ear responds in the same way as a plain rms detector.

The above results show us clearly that a PCM system will have smaller errors in the audio band if either subtractive dither is used or when quantizers with added dither employ noise-shaping, whose errors will propagate for shorter

[20] This is distinct from the ERB discussed in Sec. 3.2.

Table 2. Listing orders of dither and comparing convergence of MSE limits to 0.05 Δ^2 for channels running at 48 kHz and 96 kHz, in samples and also time for 96 kHz.

		Converge to 0.05 Δ^2		
			96 kHz	
Dither		48 kHz Samples	Samples	time
None		never	never	
Additive	RPDF	never	never	
	TPDF	4096	2048	21.3 ms
Subtractive	RPDF	128	128	1.3 ms
	TPDF	128	128	1.3 ms
Shaped	2x P	na	360	3.8 ms
	2x G	na	64	0.7 ms

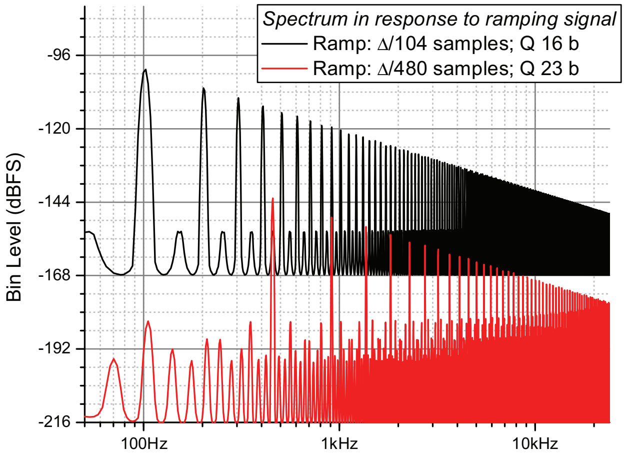

Fig. 17. Showing the spectrum of quantization error seen when a ramp moves: (black) $\Delta/104$ samples with a 16-bit quantizer; (red) $\Delta/480$ samples with a 23-bit quantizer. Fs = 48 kHz.

times, giving less temporal "smear." Table 2 shows some comparisons.

5.2. Signal Rate-of-Change

Here we briefly examine quantization errors that are responsive to the rate-of-change (slew-rate) of the signal and in Fig. 17 illustrate with two examples how quantizing signals that are slowly changing with respect to the sampling rate can introduce "birdies," manifesting as errors of quiet or low-frequency signals but in an inharmonic fashion, to more audible frequencies in the midrange.

In this example the signals are moving by 1 quantum every 104 samples (black) and every 480 samples (red)—note the quantum step is different in each case (16b and 23b, respectively).

5.3. Temporal Cross Modulation

In a typical digital system, the capture process assumes a band-limiting (anti-aliasing) filter followed by instantaneous sampling and quantization. The filter is often linear-phase, but it may be minimum-phase; the filter kernel redistributes signals so that each input sample contributes to a precession or succession of output samples.

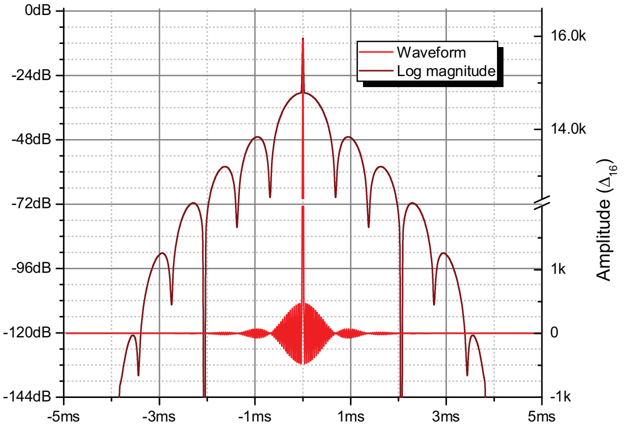

Fig. 18. Resulting signal when a unit impulse is converted in a workstation from 96 kHz to 48 kHz sample rate using a "high-quality" option. The output is shown as: (red) waveform; (wine) dB magnitude of the waveform. Note that the response extends more than 3 ms prior and post the signal event.

A simplistic example considers an impulse. In Fig. 18 we can see the output waveform when such a unit impulse is converted in a workstation from 96 kHz to 48 kHz 32b float; the familiar *sinc* response is seen.

But what happens when the bit depth of this signal is reduced? If this signal is fed to a TPDF-dithered quantizer then the waveform is preserved, including the (approx. 22 kHz) pre- and post-rings. But what if the signal encounters an undithered quantizer?

Fig. 19 shows (upper) the error waveform (difference between the input and output) of an undithered 16-bit quantizer and (lower), its spectrum. Two notable features are: (i) the error adds a burst of noise preceding the signal, (ii) whereas the filter ringing is at a very high frequency, the quantizer error distributes a broadband noise.

Here is a mechanism that co-implicates the filter and quantizer. Even if the converter turns out to be "correctly dithered," under some circumstances, one or more undithered downstream quantizations (e.g., in a volume control or a filter in a converter), may reveal an A/D or workstation or playback EQ kernel through noise modulation.

6 CONVERTERS, WORKSTATIONS, AND SYSTEMS

So far, we have examined topics of quantization and dither from both ideal and pragmatic viewpoints, including illustrating how modulation noise will appear and accumulate if not all quantizations in a chain are correctly dithered independently.

Fig. 20 highlights relevant signal processing stages in a typical delivery path; the complexity reminds us how difficult it can be to be sure everything is right in this regard.

In the diagram, the signal cascades through four groups.[21]

[21] Within the groups, black boxes normally contain integer or fixed-point PCM processing; purple boxes may or may not contain

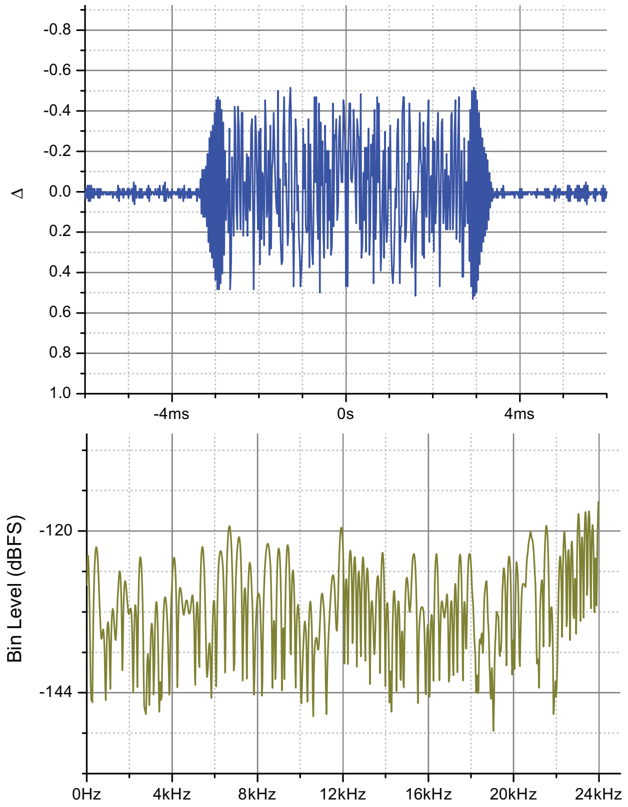

Fig. 19. (Upper) shows the error signal when the signal of Fig. 18 encounters an undithered 16-bit quantizer and (lower), its spectrum.

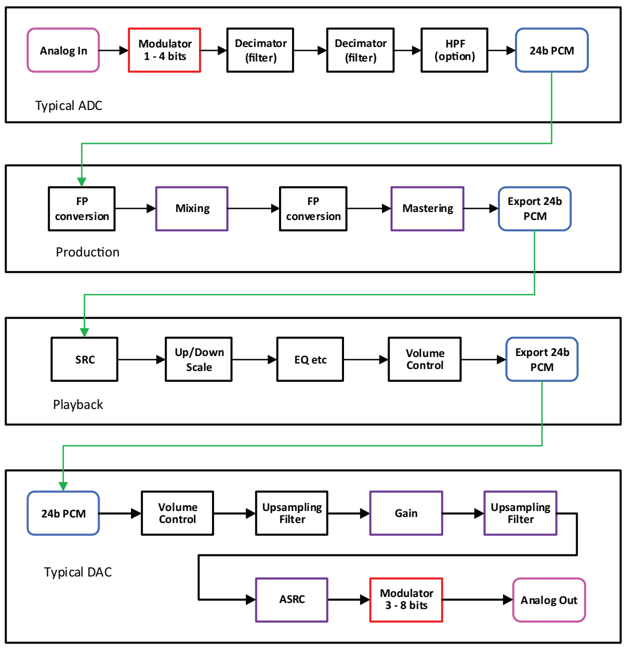

Fig. 20. Schematic diagram to represent signal processing stages in a delivery path. The top and bottom regions illustrate processes in typical integrated delta-sigma A/D and D/A converters. From the A/D the signal travels through several stages in studio production. The next region includes steps commonly seen during consumer playback.

The top and bottom groups illustrate typical delta-sigma A/D and D/A converters (expanded from Fig. 1 in [4]). These days professional A/D converters normally use dither for internal processing. Unfortunately, the same cannot be said of all IC D/A converters; sometimes the use of dither is a configuration option, in others it is simply omitted (even from the volume control) “because dither would worsen the noise specification” or “for cost reasons.”

Following A/D conversion, the signal travels through several processes in studio production—which can be quite an unruly path through mixing and mastering workstations with plug-ins for equalization, gain, reverberation, compression, restoration, click-removal, etc.

Normally a workstation will operate in 32- or 64-bit FP, but to be certain of performance one needs to test each setup. Apart from the problems converting from 32b FP to 24b (highlighted in Sec. 4.8), we commonly see implementation errors within plug-ins (such as under-sampling or “16-in-24b” quantization).

The resulting export (file) goes to distribution and *may* arrive without transcoding. If it is to make an optical disk, we need to be sure the pressing plant doesn’t compress the audio or drop the level (to “be on the safe side”)—as analyzed in Sec. 4.1.[22]

The penultimate group includes processes commonly seen during playback. Cost, convenience, expediency, inflexibility of operating systems, bandwidth restrictions in domestic wireless distribution, etc., all conspire to increase the number of signal processing steps.

In summary, between the high-speed modulators (in red) we have highlighted 19 stages where the signal can be manipulated. If just one has an undithered quantization, then modulation noise and temporal blur can appear. What are the chances of purity?

6.1 Workstation Example

A test was prompted by independent observations that a 24-bit 192 kHz recording lost “articulation and presence” when it was returned to a workstation to have some songs altered slightly in level.

On the face of it, this was puzzling, since the signal processing should only have been a benign gain change. Two test signals were put through the same processing path: (i) a 24 Hz tone (see Fig. 21 upper) that was also analyzed by SA and shown earlier in Fig. 11, and (ii) a tone at 2.4 kHz (see Fig. 21 lower).

The output shows clear evidence of inadequate dithering.

The experiments were repeated while forcing the workstation to “export” while applying TPDF dither at the 23rd and 24th bits. In both cases the obvious harmonics vanished. Although it’s hard to see on the plot, the 24-bit dithered case

FP. The high-speed modulators (shown red) must contain at least minimal dither to remain stable [36]. Signals in green are typically 16- or 24-bit fixed-point files, disks or signals on interfaces.

[22] Although in Sec. 4.1 we described these errors as common on old recordings, our direct experience is that even now, occasionally not all plants make bit-accurate transfers to CD.

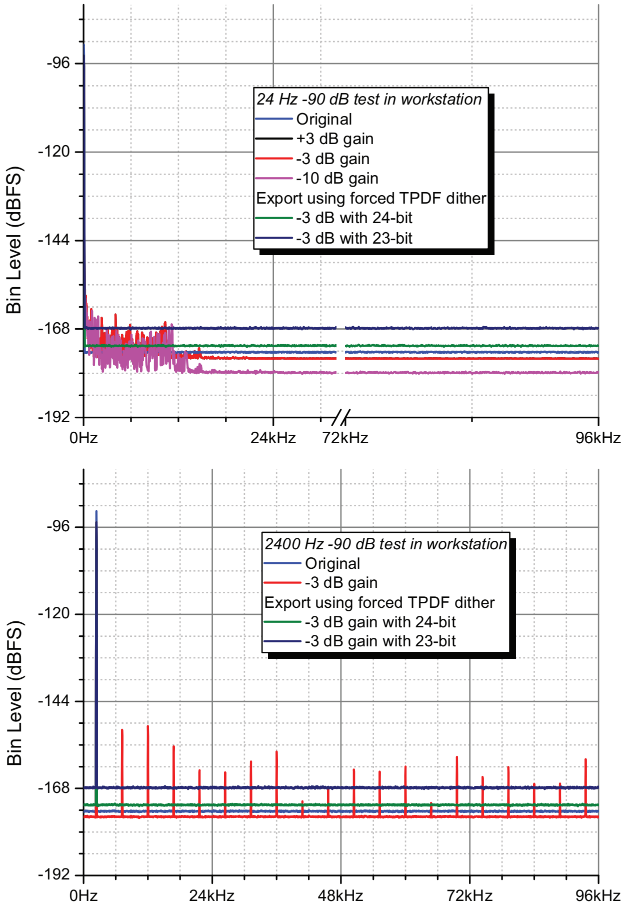

Fig. 21. Showing FFT analysis of the test signals used to probe a quantization error in a workstation.

still has “grainy structure” in the noisefloor. Only in the 23-bit case were listeners happy that the sound had not been degraded by the simple mastering process.[23]

This was a useful episode because the problem was discovered by relaxed listening after mastering and the content was not especially low noise, the background spectrum being equivalent to 14 bits (see Fig. 8 in [4]).

7 AUDIBILITY ANALYSIS

Psychoacoustic modeling will indicate a high probability that the quantization distortion and noise shown in Figs. 2, 3, and 5 will be detectable at modest playback gains—the noise exceeding the normal threshold. There are appropriate examples in [27] and [33].

Fig. 22 plots the data of Fig. 3 adjusted for a playback gain of 108 dBSPL, alongside some fundamental limits and channel capacities. We can see that at this gain, the quantization products are predicted to be audible based on the simple monaural psychoacoustic model. Higher gain increases the likelihood and vice-versa.

But we encounter several situations where mastering engineers or equipment designers are reacting to distortions at lower levels—yet we are not suggesting that the error signals in Fig. 21 would be detectable in isolation—even though it can have audible consequences on content. This is a situation that requires further understanding of auditory grouping mechanisms and further confirming experiments.

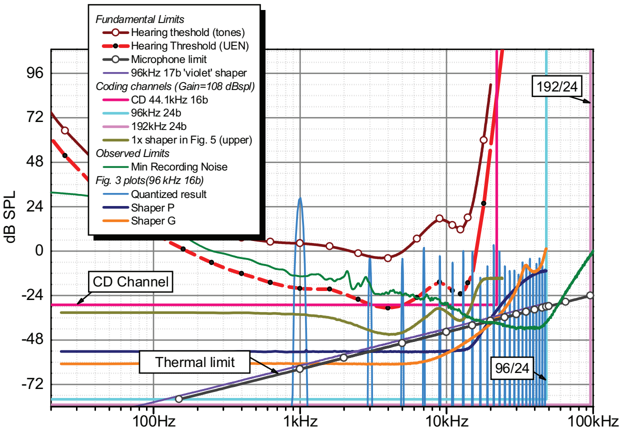

Fig. 22. Audibility analysis perspective. Thresholds: (wine + circles) is the minimum audible field threshold for single tones; (red + dots) uniformly-exciting noise at threshold [27] (depending on its bandwidth, noise between these curves may be detectable). The thermal limit for a microphone [34] is plotted (dark grey with circles)[24] and (violet) a TPDF shaper at 96 kHz/17b. Coding spaces are shown for 192 kHz/24b (magenta), 96 kHz/24b (cyan), and CD (pink). From Fig. 3, the quantized signal spectrum (dark cyan), shapers P (navy), G (orange), and (olive) the shaper from Fig. 5 (upper). Finally (green) is the background spectrum of one of the lowest-noise recordings in the survey referred to in Fig. 8 of [4].

8 HIGH-RESOLUTION

In Sec. 0 we speculated that if a chain has defective quantizations in any part of a digital path, then the resulting errors (manifesting as distortion and/or modulation noise) change with the inter-related variables of sampling rate, modulation index, and wordsize. And therefore, might confound experiments to determine, in isolation, the significance of higher sample rates or increased word-widths.

For example, when we increase the sample rate from 48 to 96 kHz or to 192 kHz are we responding to finer temporal resolution (including lower noise modulation with faster convergence—Sec. 5.1)? Is it that quantization noise lowers 3 dB with each doubling, or that some quantization errors move outside the conventional “audio band” (see Sec. 3)? Are we responding to shorter filter chains and hence lower computational noise within A/D and D/A converters (Secs. 5.3 and 6)?

For many questions on high resolution, factors tend to be inter-related and enquiry needs careful framing and interpretation of results.

[23] This story is related because we suspect the least bias happens in an unwitting experiment—no one expected a problem.

[24] Note that the thermal limit can be emulated by a 17.5-bit noise-shaper in a 96 kHz channel (or 17 bits at 192 kHz). Fig. 22 shows a 96-kHz/17-bit example (appropriately in violet)!

8.1 Designing Listening Tests

The topic of high-resolution listening tests has proven to be very challenging, for example, to compare sample rates. It is difficult to isolate a single parameter (e.g., Fs) and avoid inadvertently incorporating other variables associated with the reconfiguration of signal paths at different rates or changes in quantization.

Sec. 13.7 includes some references including the meta-study by Reiss [41] and one in which great pains were taken to isolate parameters [42].

Moving forward, it seems logical that to design a listening experiment with the minimum risk of accidental variables, we would do better with a very short signal path from microphone to loudspeaker, passing through only the variable under test.

Fig. 20 reminds us how difficult it can be to "clean the pipe" in everyday music distribution—the problem is much more acute for a test.

Without clear proof from measurement, we can't arbitrarily pass a signal through any "black box" such as a workstation, or between fixed and floating-point representation, nor can we select a converter without a clear understanding of its internal signal path.

So, the test setup should be carefully measured in the test configuration. We can use long-duration signals and SA to check for non-linearities and autocorrelation methods to document the system impulse response. Until we have a more complete understanding of the human hearing process, we should take care to avoid all known errors without expedient or pre-conceived exclusions.

For a simple test of sample rate discrimination, maybe we should select converters that are discrete, or implemented in FPGA, just so that each processing step can be forensically examined before investing in the biggest cost of a listening test, which is the listeners' time?

9 TUTORIAL SUMMARY

In Sec. 1 we pointed out that modulation noise can be insidious. It may not always be directly audible as such, but experience suggests it can modify our perception by subtly changing the way perceived objects separate from each other or from the background. The paper highlights several points in a digital chain where this defect can arise.

Sec. 2 laid the foundations for discussion of quantization artifacts and in Sec. 3 we showed how by using dither, such artifacts can be rendered substantially independent of the input signal, leading to zero measured distortion and also the ability to resolve signals "below the LSB." We indicated that subtractive dither can achieve this advantage without incurring a noise penalty and how spectral shaping can further reduce the audibility of quantization noise.

Sec. 4 covered some advanced dither topics. We showed how histogram analysis can sometimes reveal problems in an existing recording, how "synchronous averaging" can be used to characterize a complete equipment and we highlighted the difference between these two investigative tools.

Sec. 4.4 demonstrated the ability of TPDF dither to avoid modulation noise, while in Sec. 4.5 we advocated the use of independent dithers for each significant quantization within an equipment and illustrated the "washboard" distortion that can result if this is not done.

Sec. 4.6 challenged the widespread assumption that analog noise can generally be relied upon to dither a complete system.

Sec. 4.7 considered the tactic of increasing bit-depth as a possibly more cost-effective alternative to adding dither within an equipment (but not for a distribution file). In Sec. 4.8 we pointed out that, while 64-bit floating-point arithmetic can be helpful, the 32-bit equivalent does not necessarily solve all quantization problems.

In Sec. 5 we examined quantization from the time-domain perspective, reminding us that errors in a quantized system evolve over time and that analysis often assumes an averaging process. We showed an existence region for modulation noise and how it varies with sample rate, noise-shaping, additive or subtractive dither.

In Sec. 5.2 we illustrated how signals that move very slowly with respect to an undithered quantization grid can introduce tonal errors at much higher frequencies—adding so-called "birdies."

Sec. 5.3 illustrated how an undithered quantizer encountering a filtered signal can add a noise error proportional to the length of the filter kernel, at least in principle impacting both amplitude and time domains.

Sec. 6 illustrated the complexity of a typical recording/reproducing chain and its many opportunities for incorrectly-dithered quantization. As an example, we reported how an audible problem with a highly-regarded professional workstation (when used with its default settings) was resolved by adding TPDF dither at the 23rd bit.

Sec. 7 examined the theoretical audible consequences of a measured spectrum from a 16-bit undithered quantization, concluding that further work is needed to explain why some listeners apparently react to distortions that should not be audible according to simplistic criteria.

Sec. 8 returned to the topic of high resolution and argued for listening tests to be designed with greater care to isolate some variables than has been typical in the past, especially considering that the effect of a quantization will be different at different sample rates.

10 CONCLUDING REMARKS

10.1 A Gentle Art?

It wasn't whimsy that titled this paper "The Gentle Art of Dithering." Practitioners of high-resolution techniques know that valued sonic attributes of "high resolution" include transparency, sound separation, the reproduction of space and acoustic in the performance, etc. Many of the sounds that are "lost" with lower resolution don't prevent us following the song; they are low level yet contribute to the interested listener's enjoyment or engagement.

It is well established from experience that many times we can identify an improvement in signal processing, not by listening to loud or impulsive sounds but to realize that there is a difference in the subliminal "air" or "sound of the hall," even before the musicians start playing.

The relative importance of preserving micro-sounds or small details should put emphasis on eliminating any mechanism that introduces low-level modulation noise, for which quantization errors are obvious candidates. And yet, as we work and examine the entire signal chain, we continue to see errors, oversights, or misplaced pragmatism that gradually, by small cuts, put the resolution at risk. It is somewhat akin to carefully assembling a precision bearing or gear-train and then forgetting to lubricate it—eventually it will squeak.

Dither is a "gentle art" because it doesn't deal with large spectacular things, but it preserves essential naturalness.

10.2 CONCLUSION

Despite having been understood for over three decades, in practice we too often see a careless and expedient approach to dithering multibit PCM. Unless dither is used correctly at each stage in a digital recording and playback, transparency will suffer.

It is often asked: "Why should I add noise to my recording" or, "How can adding noise make things clearer?" This paper gives a tour through these questions as well as aspects of time-domain and highlighting the need for care if a signal is moved within or between integer and floating-point representations.

By illustrating "washboard" distortion we can be reminded that the strategy of "hoping the errors will all be covered by the microphone noise" can come unstuck downstream.

The paper reminds us that bigger numbers are not always better—a 16-bit delivery where dither is used imaginatively can easily carry modern recordings without impacting noisefloor (see Fig. 8 in [4]) and may even be preferred to one using 24-bits but where implementation had assumed that "the errors were too low to hear." Fig. 22 and footnote 24 remind us that a 96 kHz 18-bit channel can encode down to the thermal noise of our planet's atmosphere.

We hope a few common misconceptions or regrettable decisions have been pointed out.[25]

Dither should not be looked on as an added noise but an essential lubricant. If noise-floor is a problem, then noise-shaping—or even better, subtractive dither—provide the perfect solutions.

We hope that, in some small way, this paper can remind us all to be vigilant.

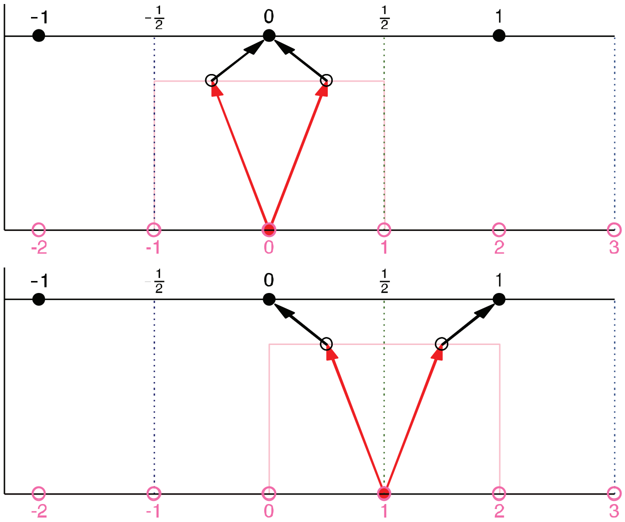

Fig. 23. RPDF requantization of a 17-bit signal to 16 bits using a discrete binary dither. Upper: an original 17-bit sample lies on an even 17-bit grid point. Lower: sample lies on an odd grid point. Red arrows show the changes in sample values from dithering; black arrows show the further changes from quantization.

11 APPENDIX: DISCRETE DITHER

Up to now we have assumed that RPDF and TPDF dithers are available as continuous distributions. Now we ask whether there is compromise in using digital dither samples having only a finite number of dither bits.

It may seem "obvious" that if we have a signal already quantized to 24 bits and we requantize to 16, there is no point in generating dither that extends beyond the 24th bit. That is, in fact, correct for RPDF dither, but we shall show that TPDF performance is compromised very slightly; more severely as the number of bits shaved off is reduced. Accordingly, we shall start by considering the limiting case where the input wordlength is to be reduced by just one bit.

11.1 Continuous and Discrete RPDF Quantization

For definiteness we consider requantization of a 17-bit signal to 16 bits. In Fig. 23, horizontal axes represent signal amplitude. The lower axis is labeled (pink) in quanta of the original 17-bit signal and the upper axis is labeled (black) in quanta of the 16-bit quantized signal. The pink rectangles show the probability distributions that would result if continuous RPDF dither were added to the original sample shown as a red blob. In contrast, a 1-bit discrete dither will oscillate randomly between two values (black open circles) separated by half a 16-bit quantum.

After quantization by a rounding quantizer, an even 17-bit value (upper diagram) is mapped to the same value as itself while an odd 17-bit input sample is mapped to a random choice between the two nearest 16-bit grid points.

This scheme preserves expected value (the *first moment* of error is zero) while minimizing the mean-square error (MSE, aka second moment of error). However, the resulting MSE is not constant, being zero for even 17-bit sample

[25] Including: "I don't need to add dither because there is so much noise coming in from the microphone"; "I don't need to add dither because my incoming and outgoing signals are both 24-bit."; "Dithering-down is only used for reducing wordsize"; "We can't add dither to our D/A (SRC, A/D) chip because if we do it will reduce the dynamic-range specification and not be selected."

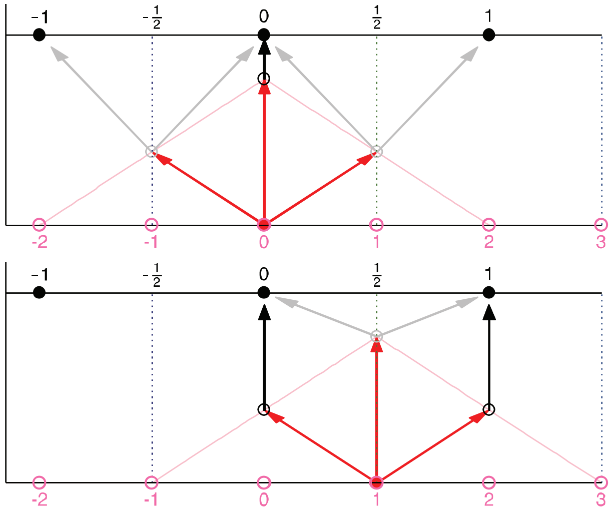

Fig. 24. Akin to Fig. 23 but using a ternary dither with probabilities given by a discretized TPDF. The grey arrows show ambiguous quantization when the dithered sample is on a decision threshold (dotted vertical lines).

values but $\frac{1}{4}\,\Delta^2$ for odd values where Δ refers to the step size of the 16-bit grid. This is similar to the situation with continuous dither, cf., Fig. 12 (lower).

11.2 TPDF Dither

TPDF dither attempts to balance the MSEs from even and odd 17-bit samples and so to eliminate modulation noise.

The pink triangles in Fig. 24 show the PDFs that would result from adding continuous TPDF dither to the signal. The open black circles are the dithered values resulting from adding discrete ternary dither taking the 17-bit values (–1, 0, 1) with probabilities ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$). Thus, the central dithered value has twice the probability of each of the two outliers, irrespective of whether the 17-bit input is even or odd.

The new phenomenon is that some of the dithered values now lie on decision thresholds, shown as vertical dotted lines.

We show firstly an ideal case where threshold values are quantized up or down with equal probability, as shown by the grey arrows in Fig. 24. The probabilities of quantizing the 17-bit value "0" to the 16-bit values (–1, 0, 1) are ($\frac{1}{8}$, $\frac{3}{4}$, $\frac{1}{8}$) respectively[26], from which we deduce an absolute error of Δ with probability $\frac{1}{4}$ and zero with probability $\frac{3}{4}$, leading to an MSE of $\frac{1}{4}\,\Delta^2$.

For odd 17-bit input values the quantized result is a random choice between the two nearest 16-bit values, so the absolute error is always $\frac{1}{2}\,\Delta$ so the MSE is again $\frac{1}{4}\,\Delta^2$. The MSE is thus independent of signal and there is no modulation noise.

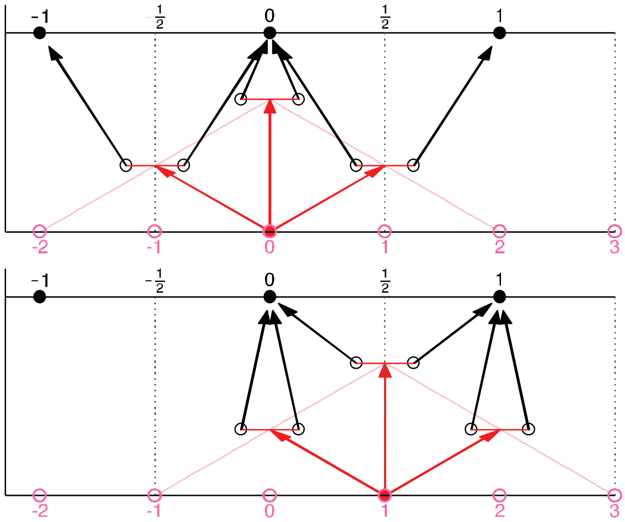

Fig. 25. Akin to Fig. 24 but avoiding decision thresholds.

11.3 Avoiding Decision Thresholds

With continuous dither, it is not important how the subsequent rounding (or truncation) operation will behave at a quantization decision threshold because the probability of a dithered sample landing precisely on a threshold is vanishingly small. This is not the case with discrete dither, and we need either to avoid thresholds or to consider the threshold behavior carefully. In particular, the "ideal" random up or down quantization of threshold values will not be obtained by default.

A random up/down choice can be forced by adding a smaller binary perturbation as shown in Fig. 25.[27]

11.4 Rounding Modes

Common rounding modes are:

• Truncate towards zero	(not recommended)
• Truncate towards $\pm\infty$	(*floor(), ceil()*)
• Round ties[28] consistently	(up or down)
• Round ties to even	(IEEE 754)

"Truncation towards zero" must be avoided but unfortunately it can happen unintentionally because some programming languages silently do it when a floating-point value is assigned to an integer variable.[29] If decision thresholds have been avoided (or will occur with negligible probability) then the remaining modes are all fine regarding distortion and modulation noise.

[26] The same probabilities are obtained using continuous dither. They correspond to the areas of the pink triangle bounded by the decision thresholds in Fig. 24 (upper).

[27] This combination of TPDF and binary dither can be implemented as the sum of a binary dither at the 17th bit and a two-bit RPDF dither occupying the 17th and 18th bit positions. Six perturbed values (black open circles) are thus generated with probabilities proportional to (1, 1, 2, 2, 1, 1): $(1, 0, 1) \otimes (1, 1, 1, 1) = (1, 1, 2, 2, 1, 1)$.

[28] "Ties" refers to values on a decision threshold.

[29] Truncation towards zero creates a kink in the transfer function akin to classic crossover distortion. It affects all sample values, not just those on a decision threshold.

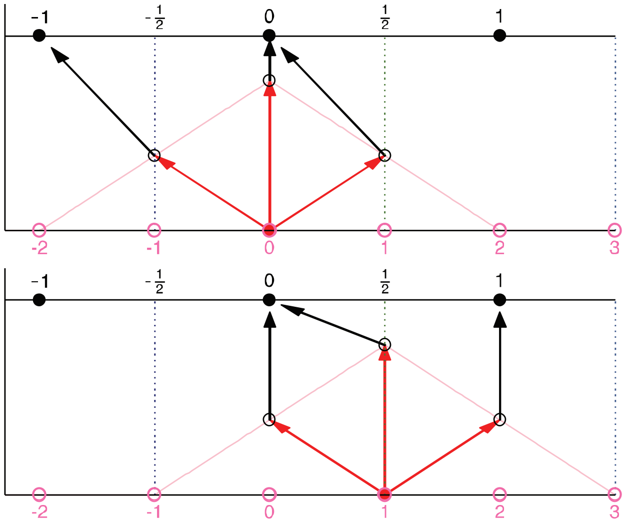

Fig. 26. Using floor in place of round in Fig. 24.

The IEEE 754 standard for floating point arithmetic specifies “Round to nearest, ties to even” as the default behavior. This is good for general numerical use but not for discrete dither that does not avoid decision thresholds: any ties will result in a quantized signal in which the even 16-bit values will occur more often than odd values–a “15–bit-ism!”

11.5 A Lower-Noise Option for (n+1) → n Bits

Let us take Fig. 24 and replace the special random–choice quantizer by a plain *floor* (truncate towards $-\infty$) operation. Please see Fig. 26.

We see that the 17-bit input value “0” has been quantized to 16-bit “–1” with probability $\frac{1}{4}$ and to “0” with probability $\frac{3}{4}$, giving a mean value $-\frac{1}{4}$. The 17-bit input value “1” has been quantized to 16-bit “0” with probability $\frac{3}{4}$ and to “1” with probability $\frac{1}{4}$, giving a mean value $+\frac{1}{4}$.

Thus, the mean output value is offset by $-\frac{1}{4}\ \Delta$ (on the 16-bit grid) in both cases, i.e., there is a tiny DC shift but no distortion. After allowing for the DC shift, the MSE is:

$$\tfrac{1}{4} \times (9/16) + \tfrac{3}{4} \times (1/16) = 3/16\Delta^2$$

for both the even and odd 17-bit input values.

There is therefore no modulation noise and the MSE is 1.25 dB lower than the $\frac{1}{4}\ \Delta^2$ obtained with conventional TPDF dither.[30]

So, is continuous TPDF dither no longer considered optimal? Only for the special cases of needing to reduce the wordlength by just one or two bits. If we were starting from 24 bits and used the same principle repeatedly to remove bits one at a time, the MSEs would accumulate and the total would be the same as when conventional TPDF is used.

Nevertheless, this recipe to reduce the wordlength one bit at a time with minimal added noise may be interesting in applications such as scalable lossless transmission.

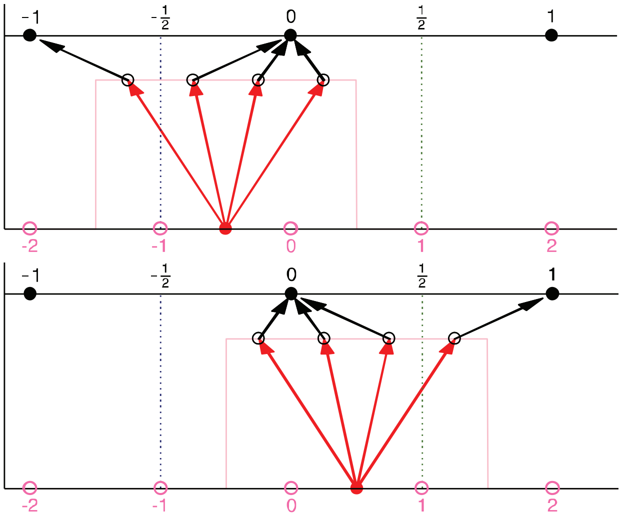

Fig. 27. Use of a two-bit rectangular dither to achieve the same result as Fig. 26.

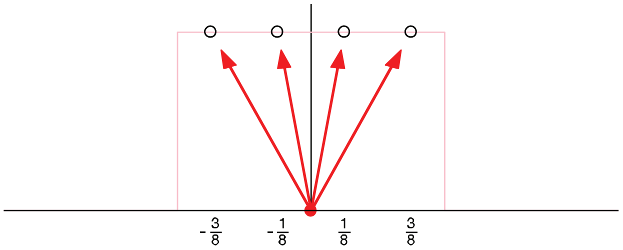

Fig. 28. Two-bit symmetrical rectangular dither to requantize 18 bits down to 16.

It may be of interest that the same result can be obtained by first adding an offset of $-\frac{1}{2}\ \Delta$ to the 17-bit input signal and then quantizing with a two-bit RPDF dither occupying the 17th and 18th bit positions. The diagram, Fig. 27., looks different but the final result is identical.

This implementation avoids decision thresholds so any of the rounding quantizers listed above, with the exception of truncation towards zero, will be fine.

11.6. Multibit Discrete Dither

A more usual requirement is to reduce a wordwidth by several bits in one operation—often starting from an unknown bit depth. If we can afford to generate a TPDF dither by adding or subtracting two 8-bit RPDF dithers, that should be good enough for most practical purposes. However, we will illustrate precautions that will replicate precisely the desirable properties of continuous RPDF and TPDF dithers by considering how to reduce a signal’s wordwidth from 18 bits down to 16 bits with a minimal number of dither bits.

For RPDF we can dispose the dither values symmetrically about zero as in Fig. 28 and use a rounding quantizer. All samples of the 18-bit input signal are integer multiples of $\frac{1}{4}\ \Delta$ (referred to the step size of the 16-bit grid) so if dither composed only of odd multiples of 1/8 Δ is added, the dithered sample will never be at a decision threshold of the quantizer and the quantizer’s behavior on a “tie” need not concern us.

[30] An interesting question is whether to attach significance to the lop-sided nature of the error—a skew distribution with nonzero third moment. Conventional wisdom is that the ear is insensitive to moments beyond the second.

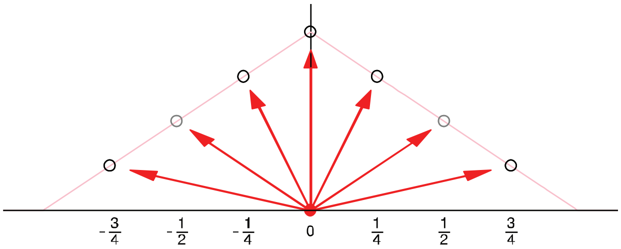

Fig. 29. Triangular distribution from adding or subtracting two instances of Fig. 28.

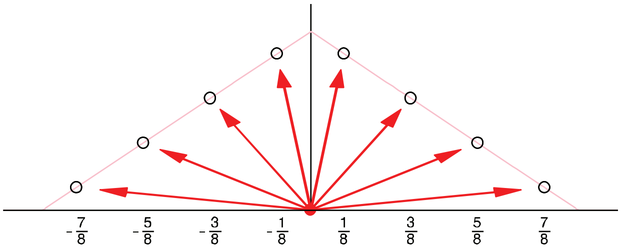

Fig. 30. Improved triangular distribution avoiding decision thresholds.

For TPDF, we might think to create two independent 2-bit RPDF dithers (each like Fig. 28) and add or subtract them to create the discrete triangular dither of Fig. 29.

This triangular dither does not avoid decision thresholds whichever type (*round, floor* or *ceiling)* of quantizer is used. A small DC offset could enable it to do so but we prefer to add a random binary offset taking the values ±1/8 relative to the 16-bit quantization grid, giving the discrete PDF indicated in Fig. 30.

This approach generalizes to cases where more bits need to be shaved.[31] Using a rounding quantizer, the mean output (ensemble average over dither instances) faithfully reproduces the input signal and there is zero modulation noise. The second moment of the error is constant at $\frac{1}{4}\Delta^2$, exactly[32] as for continuous TPDF dither.

12 ACKNOWLEDGMENTS

Please see the acknowledgments in [4]. This paper covers a sustained enquiry and the authors are particularly grateful to our co-workers in MQA and Algol; especially in this case: Spencer Chrislu, Trefor Roberts, Alan Wood, Michael Capp, and Malcolm Law.

We are very grateful to Michael Capp who produced the simulation data for Sec. 5.1.

We are particularly grateful to Morten Lindberg, Bob Ludwig, Mick Sawaguchi, Peter McGrath, and John Atkinson who participated in testing their recordings on workstations.

Thanks also for many stimulating discussions on the dither topic over the years with Stanley Lipshitz, John Vanderkooy, Vicki Melchior, Rhonda Wilson, Wieslaw Woszczyk, and, of course, our co-conspirator, the late Michael Gerzon.

13. REFERENCES

References are grouped into topic groups in approximate order of introduction in the paper. Sec. 14 offers a bibliography.

13.1. High Resolution

[1] J. R. Stuart, "Soundboard: High-Resolution Audio," *J. Audio Eng. Soc.*, vol. 63, pp. 831–832 (2015 Oct.). Open Access http://www.aes.org/e-lib/browse.cfm?elib=18046

[2] Recording Academy Producers & Engineers Wing, "Recommendations for Hi-Resolution Music Production," https://www.grammy.com/sites/com/files/recommendations_for_hires_music_production_09_28_18.pdf (2018 Sep.).

13.2. The audio channel

[3] J. R. Stuart and P. G. Craven, "A Hierarchical Approach to Archiving and Distribution," presented at the 137th Convention of the Audio Engineering Society (2014 Oct.), convention paper 9178. Open Access: http://www.aes.org/e-lib/browse.cfm?elib=17501

[4] J. R. Stuart and P. G. Craven, "A Hierarchical Approach for Audio Capture, Archive and Distribution," *J. Audio Eng. Soc.*, Vol. 67, (2019 May) (this issue). Open access: https://doi.org/10.17743/jaes.2018.0062

[5] J.R. Stuart, R. Hollinshead and M. Capp, "Is High-frequency Intermodulation Distortion a Significant Factor in High-resolution Audio?" *J. Audio Eng. Soc.*, Vol. 67, (2019 May) (this issue) Open access: https://doi.org/10.17743/jaes.2018.0060

[6] J. R. Stuart "Coding for High-Resolution Audio Systems," *J. Audio Eng. Soc.*, vol. 52, pp. 117–144 (2004 Mar.). http://www.aes.org/e-lib/browse.cfm?elib=12986

[7] R. Z. Langevin, "Modulation Noise in Tape Recordings," presented at the 17th Convention of the Audio Engineering Society (1965 Oct.), convention paper 391. http://www.aes.org/e-lib/browse.cfm?elib=1049

[8] L. Fielder, "The Audibility of Modulation Noise in Floating-Point Conversion Systems," *J. Audio Eng. Soc.*, vol. 33, pp. 770–781 (1985 Oct.). http://www.aes.org/e-lib/browse.cfm?elib=4427

13.3. Sampling theory (ancient and modern)

[9] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21 (1949 Jan.).

[31] We generate discrete TPDF dither by subtracting two independent RPDF dithers having the same (or greater) bit-depth as the input signal, then add a further independent binary dither oscillating symmetrically about zero between two values separated by the stepsize of the two RPDF dithers.

[32] More precisely, the MSE is exactly constant with respect to time and its level approaches $\frac{1}{4}\Delta^2$ rapidly as the number of bits shaved off increases.

[10] H. Nyquist, “Certain Topics in Telegraph Transmission Theory,” *Trans. Amer. IEE*, vol. 47, pp. 617–644 (1928).

[11] M. Unser “Sampling—50 Years after Shannon,” *Proc. IEEE,* vol. 88, no. 4, pp. 569–587 (2000 Apr.). DOI:http://dx.doi.org/10.1109/5.843002

13.4. Dither and quantization

[12] B. Widrow and I. Kollár, Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications (CUP, Cambridge, UK, 2008).

[13] R. Dunay, I. Kollár, and B. Widrow, “Dithering for Floating-Point Number Representation,” 1st International On-Line Workshop on Dithering in Measurement (1998).

[14] S. P. Lipshitz and J. Vanderkooy, “Pulse-Code Modulation—An Overview,” *J. Audio Eng. Soc.*, vol. 52, pp. 200–214 (2004 Mar.). http://www.aes.org/e-lib/browse.cfm?elib=12991

[15] J. Vanderkooy and S. P. Lipshitz, “Dither in Digital Audio,” *J. Audio Eng. Soc.*, vol. 35, pp. 966–975 (1987 Dec.). http://www.aes.org/e-lib/browse.cfm?elib=5173

[16] J. Vanderkooy and S. P. Lipshitz, “Digital Dither: Signal Processing with Resolution Far Below the Least Significant Bit,” presented at the AES 7th International Conference: Audio in Digital Times (1989 May), conference paper 7-014. http://www.aes.org/e-lib/browse.cfm?elib=5482

[17] P. G. Craven and M. A. Gerzon, “Compatible Improvement of 16-Bit Systems Using Subtractive Dither,” presented at the 93rd Convention of the Audio Engineering Society (1992 Oct.), convention paper 3356. http://www.aes.org/e-lib/browse.cfm?elib=6777

[18] R. Lagadec, “New Frontiers in Digital Audio,” presented at the 89th Convention of the Audio Engineering Society (1990 Sep.), convention paper 3002. http://www.aes.org/e-lib/browse.cfm?elib=5691

[19] R. A. Wannamaker and S. P. Lipshitz, “Time Domain Behavior of Dithered Quantizers,” presented at the 93rd Convention of the Audio Engineering Society (1992 Oct.), convention paper 3418.

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

[21] R. Cabot, “Noise Modulation in Digital Audio Devices,” presented at the 90th Convention of the Audio Engineering Society (1991 Feb.), convention paper 3021. http://www.aes.org/e-lib/browse.cfm?elib=5672

[22] R. Cabot, “Performance Limitations of Digital Filter Architectures,” presented at the 89th Convention of the Audio Engineering Society (1990 Sep.), convention paper 2964. http://www.aes.org/e-lib/browse.cfm?elib=5729

[23] S. P. Lipshitz and J. Vanderkoy, “Dither Myths and Facts,” presented at the 117th Convention of the Audio Engineering Society (2004 Oct.), convention paper 6279. http://www.aes.org/e-lib/browse.cfm?elib=12936

13.5. Noise shaping

[24] M. A. Gerzon, P. G. Craven, J. R. Stuart, and R. J. Wilson, “Psychoacoustic Noise Shaped Improvements in CD and Other Linear Digital Media,” presented at the 94th Convention of the Audio Engineering Society (1993 Mar.), convention paper 3501. http://www.aes.org/e-lib/browse.cfm?elib=6647

[25] 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 (1989 Oct.), convention paper 2822. http://www.aes.org/e-lib/browse.cfm?elib=5872

[26] M. A. Gerzon and P. G. Craven, “A High-Rate Buried Data Channel for Audio CD,” presented at the 94th Convention of the Audio Engineering Society (1993 Mar.), convention paper 3551.

[27] J. R. Stuart, “Noise: Methods for Estimating Detectability and Threshold,” *J. Audio Eng. Soc.*, vol. 42, pp. 124–140 (1994 Mar.). http://www.aes.org/e-lib/browse.cfm?elib=6959

[28] J. R. Stuart, and R. J. Wilson, “Dynamic Range Enhancement Using Noise-Shaped Dither at 44.1, 48, and 96 kHz,” presented at the 100th Convention of the Audio Engineering Society (1996 May), convention paper 4236. http://www.aes.org/e-lib/browse.cfm?elib=7538

[29] Acoustic Renaissance for Audio, “DVD: Pre-Emphasis for Use at 96 kHz or 88.2 kHz” (1996 Nov.) (described in [3]).

[30] J. R. Stuart and R. J. Wilson, “A Search for Efficient Dither for DSP Applications,” presented at the 92nd Convention of the Audio Engineering Society (1992 Mar.), convention paper 3334. http://www.aes.org/e-lib/browse.cfm?elib=6799

[31] J. R. Stuart and R. J. Wilson, “Dynamic Range Enhancement Using Noise-Shaped Dither Applied to Signals with and without Pre-emphasis,” presented at the 96th Convention of the Audio Engineering Society (1994 Feb.), convention paper 3871. http://www.aes.org/e-lib/browse.cfm?elib=6361

[32] M. Akune, R. M. Heddle, and K. Akagiri, “Super Bit Mapping: Psychoacoustically Optimized Digital Recording,” presented at the 93rd Convention of the Audio Engineering Society (1992 Oct.), convention paper 3371. http://www.aes.org/e-lib/browse.cfm?elib=6762

[33] J. R. Stuart, “Auditory Modelling Related to the Bit Budget,” presented at the AES UK 9th Conference: Managing the Bit Budget’ (1994 May), conference paper MBB-18. http://www.aes.org/e-lib/browse.cfm?elib=6110

[34] P. B. Fellgett, “Thermal Noise Limits of Microphones,” *J. IERE*, vol. 57, no. 4, pp. 161–166 (1987).

13.6. Signal Processing

[35] G. J. Barton and P. G. Craven, “Digital Analysis of Quantization Errors in Digital Signal Processing,” presented at the 77th Convention of the Audio Engineering Society (Mar. 1985 Mar.), convention paper 2180. http://www.aes.org/e-lib/browse.cfm?elib=11575

[36] S. R. Norsworthy, R. Schrier, and G. C. Temes (eds.), Delta-Sigma Data Converters: Theory, Design and Simulation (IEEE Press, 1997).

[37] https://en.wikipedia.org/wiki/Washboard_(musical_instrument)

[38] https://en.wikipedia.org/wiki/Moir%C3%A9_pattern

[39] https://en.wikipedia.org/wiki/IEEE_754#Rounding_rules

[40] https://en.wikipedia.org/wiki/Quantization_(signal_processing)#Mid-riser_and_mid-tread_uniform_quantizers

13.7. Listening tests and evaluation

[41] J. D. Reiss, "A Meta-Analysis of High Resolution Audio Perceptual Evaluation," *J. Audio Eng. Soc.,* vol. 64, pp. 364–379 (2016 Jun.). https://doi.org/10.17743/jaes.2016.0015

[42] H. M. Jackson, M. D. Capp, and J. R. Stuart, "The Audibility of Typical Digital Audio Filters in a High-Fidelity Playback System," presented at the 137th Convention of the Audio Engineering Society (2014 Oct.), convention paper 9174. http://www.aes.org/e-lib/browse.cfm?elib=17497

[43] A. Pras and C. Guastavino, "Sampling Rate Discrimination: 44.1 kHz vs. 88.2 kHz," presented at the 128th Convention of the Audio Engineering Society (2010 May), convention paper 8101. http://www.aes.org/e-lib/browse.cfm?elib = 15398

[44] T. Nishiguchi and K. Hamasaki, "Differences of Hearing Impressions among Several High Sampling Digital Recording Formats," presented at the 118th Convention of the Audio Engineering Society (2005 May), convention paper 6469. http://www.aes.org/e-lib/browse.cfm?elib=13185

[45] W. Woszczyk, "Physical and Perceptual Considerations for High-Resolution Audio," presented at the 115th Convention of the Audio Engineering Society (2003 Oct.), convention paper 5931. http://www.aes.org/e-lib/browse.cfm?elib = 12372

[46] S. Yoshikawa, S. Noge, M. Ohsu, S. Toyama, H. Yanagawa, and T. Yamamoto, "Sound Quality Evaluation of 96-kHz Sampling Digital Audio," presented at the 99th Convention of the Audio Engineering Society (1995 Oct.), convention paper 4112. http://www.aes.org/e-lib/browse.cfm?elib=7654

14. BIBLIOGRAPHY

14.1. Human auditory perception

[47] C. J. Plack (ed.), The Oxford Handbook of Auditory Science: Hearing, vol. **3** (OUP, 2010).

[48] A. S. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound (The MIT Press, 1990).

[49] E. Zwicker, and H. Fastl, Psychoacoustics, Facts and Models (Springer, Berlin, 1990), vol. 22.

14.2. Natural sounds and ethology

[50] M. S. Lewicki "Efficient Coding of Natural Sounds," Nature Neurosci., vol. 5, pp. 356–363 (2002). DOI:http://dx.doi.org/10.1038/nn831

[51] E. C. Bluvas and T. Q. Gentner, "Attention to Natural Auditory Signals," *Hearing Research*, vol. 305, pp. 10–18 (2013). https://doi.org/10.1016/j.heares.2013.08.007

[52] E. C. Smith and M. S. Lewicki, "Efficient Auditory Coding," *Nature*, vol. 439, pp. 978–982 (2006 Feb.). https://doi.org/10.1038/nature04485

14.3. Auditory neuroscience

[53] W. A. Yost et al., "Auditory Perception of Sound Sources," in *Springer Handbook of Auditory Research*, vol. **29** (Springer Science+Business Media, 2008).

[54] J. Schnupp et al., Auditory Neuroscience: Making Sense of Sound (MIT Press, 2011).

[55] A. Rees and A. R. Palmer (eds.), The Oxford Handbook of Auditory Science: The Auditory Brain, vol. **2** (OUP, 2010).

[56] D. Oertel et al., "Integrative Functions in the Mammalian Auditory Pathway," in *Springer Handbook of Auditory Research*, vol.**15** (Springer Verlag, 2002).

[57] F. Rieke et al., Spikes: Exploring the Neural Code (MIT Press, 1997).

14.4. Audio temporal discrimination

[58] T. Deneux et al., "Temporal Asymmetries in Auditory Coding and Perception Reflect Multi-Layered Nonlinearities," Nature Communications, (2016 Sep.). DOI: https://doi.org/10.1038/ncomms12682

[59] D. A. Abrams, "Population Responses in Primary Auditory Cortex Simultaneously Represent the Temporal Envelope and Periodicity Features in Natural Speech," Hearing Research, vol. 348, pp. 31–43 (2017). https://doi.org/10.1016/j.heares.2017.02.010

[60] K. Krumbholz, R. D. Patterson, et al., "Microsecond Temporal Resolution in Monaural Hearing without Spectral Cues," *J. Acoust. Soc. Amer.*, vol. 113, no. 5, 2790–2800 (2003). https://doi.org/10.1121/1.1547438

[61] P. Heil and H. Neubauer, "A Unifying Basis for Auditory Thresholds Based on Temporal Summation," *PNAS,* vol. 100, no. 10, pp. 6151–6156 (2003). https://doi.org/10.1073/pnas.1030017100

THE AUTHORS

J. Robert Stuart

Peter G. Craven

J. Robert (Bob) Stuart was born in 1948. He studied electronic engineering and acoustics at the University of Birmingham and took an M.Sc. in operations research at Imperial College, London. While at Birmingham he studied psychoacoustics under Professor Jack Allison, which began a lifelong fascination with the subject.

In 1977 he co-founded Meridian Audio and served as CTO until early 2015. In 2014 he founded MQA Ltd. where he is currently full time as Chairman and CTO.

At the request of Hiro Negishi and Raymond Cooke, Bob chaired the advocacy group Acoustic Renaissance for Audio between 1994 and 2002.

In the 1990s he worked with Michael Gerzon and Peter Craven on lossless compression and was instrumental in its adoption for optical discs.

Bob has contributed to DVD-Audio and BluRay standards and has served on the technical committees of the National Sound Archive, JAS, and the ADA (Japan).

Bob's professional interests are the furthering of analog and digital audio and developing understanding of human auditory perception mechanisms relevant to live and recorded music. His specialties include the auditory sciences and the design of analog and digital electronics, loudspeakers, audio coding, and signal processing.

Bob joined AES in 1971, has been a fellow since 1992, and is a member of ASA, IEEE, and the Hearing Group at Cambridge.

Bob has a deep interest in music and spends a good deal of time listening to live and recorded material.

•

Peter Craven was born in 1948. He studied maths and then astrophysics at Oxford University while collaborating with Michael Gerzon and others at the Oxford University Tape Recording Society on making live recordings and on audio quality topics generally, including the ideas (1972) leading to the Ambisonic Soundfield Microphone.

After some years in research and academia, Peter became independent in 1981, specializing in digital signal processing for audio. Further collaborations with Gerzon resulted in a seminal paper on noise-shaping in 1989. Work on room equalization with B&W loudspeakers was followed by collaboration with Bob Stuart on noise shaping and buried data, leading eventually to the MLP lossless compression system and, more recently, to the MQA music delivery system.

A collector of pre-war coarse-groove 78 r.p.m. records, Peter Craven seeks a synergy between modern high definition multichannel technology and the simpler recording techniques of yesteryear.

Bob and Peter have worked together on audio topics since they first met in 1975.