BARCO Research and Development - Display Technology Group

Report : Quantization, noise, dither and noise shaping

PLB110298/JB4037

1. Introduction

The work was motivated by some common misunderstandings in the electronics industry about the effects of quantization. The existence of generalized and simplified rules, leads to oversimplification and incorrect interpretations. This is more than often the case when A-D and D-A converters are specified and tested[1], and leads to incorrect results for SNR (signal to noise ratio) and THD (total harmonic distortion).
Quantization and digital signal processing were originally used in seismologic survey. Dither signals were first used in PCM video systems in 1951. Theoretical research on quantization and dither mainly dates from the 80s and its use in digital audio signal processing. The relevant theory is scattered among industry journals and conference proceedings. There is a paper[2] that collects most of the significant theory on quantization and dither. This reports attempts to explain and illustrate the theorems with practical examples. It is hoped that a better understanding of the effects of quantization can correct some of the common misunderstandings. A part of this report originally appeared in Pro-Audio Magazine[3].

2. Quantization

Digitization of signals is usually implemented by sampling the signal repetitively at an adequate rate, and quantizing the resulting analog signal into finite steps. Truncating a digital word is theoretically the same as quantizing. Digital signal processing increases the wordlength, and reducing the wordlength to the input wordlength (requantization) always involves a truncation.
Most of the theory assumes the quantizing intervals are the same. In the examples the same assumption is made with a step size of 1 (LSB).
Midtread quantization can be expressed analytically in terms of the input signal as
$Q(x) = \mathrm{floor}(x + 0.5)$.

---

[1] B. Travis, “Demystifying ADCs”, EDN, March 27 1997 & B. Travis, “Remistifying ADCs”, EDN, October 9 1997
[2] S. Lipshitz and J. Vanderkooy, “ Quantization and Dither: A Theorethical Survey”, J. Audio Eng. Soc., vol. 40, no. 5, pp. 355-374, 1992 May
[3] P. Bracke, “Ruis en dither”, Pro-Audio Magazine, vol. 8, no. 11, 1996 November

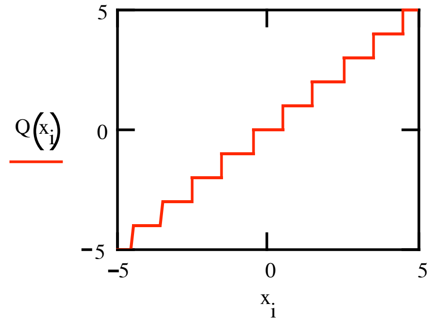


Fig. 1. Midtread quantizer transfer characteristic

Quantization of a DC signal with a level equal to an integer number of LSBs, e.g. 4, gives no quantization error. Quantization of a DC signal with a level equal to an integer number of LSBs ± 0.5, e.g. 3.5, gives a maximum quantization error of 0.5 LSB. The quantized signal is a DC-signal with a level equal to an integer number of LSBs, 4 in the example, without any noise. So the signal-to-noise ratio is **infinite**, but the level is wrong, the amplitude error is 0.5.
What about quantization of sine signals ?

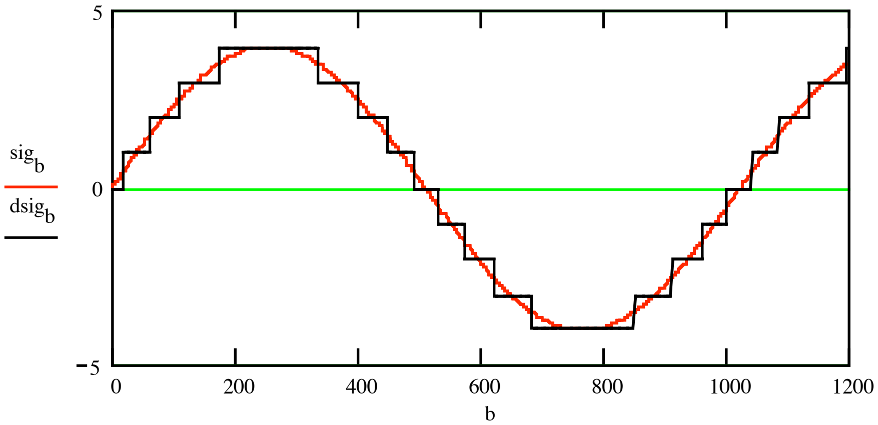


Fig. 2. Quantization of sine signal with 4 LSB amplitude

A quantized sine signal is the summation of square waves with 1 LSB amplitude and the same frequency as the sine signal, as one can see in fig. 2. Therefor, the spectrum consists only of harmonics of the sine signal.

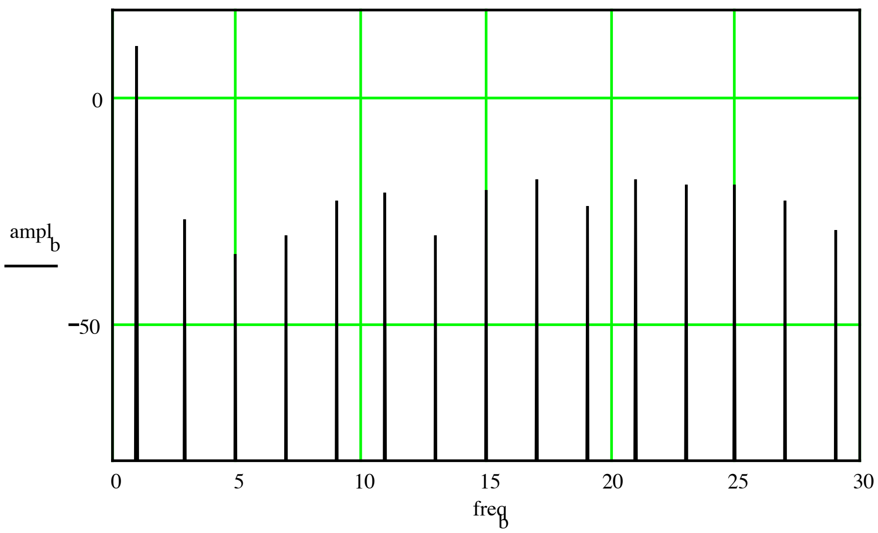


Fig. 3. Spectrum of quantized 4 LSB sine

The amplitude error of the first harmonic is 0.053 LSB and the THD is 9.5%. There is no noise, so the SNR of a quantized sine is **infinite**. If the amplitude of the sine is less than 0.5 LSB, the quantized output is zero and the amplitude error can be 0.5 LSB.

3. Noise

Noise is a stationary non-deterministic signal with average value zero. Only statistical properties can describe noise. The standard deviation σ is the RMS-value. The autocorrelation is non-deterministic and nearly 0 (typically between 0 and 0.1). The probability distribution can be anything. Natural occurring noise often has a normal distribution, the probability density is a Gaussian function. The frequency spectrum can be anything, like flat for white noise and 10 dB / decade descending for pink noise.
Quantization does not produce noise, so why is often 1.76 + 6.02 x (number of bits) quoted as the signal-to-noise ratio in decibels?
The quantization error is the difference between the original signal and the quantized signal.
The quantization error lies between —0.5 and 0.5 LSB and is correlated to the signal.

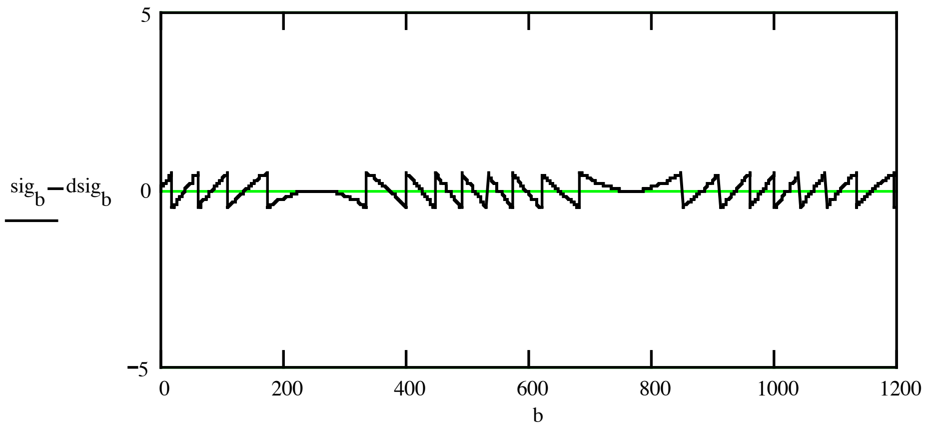


Fig. 4. Quantization error of 4 LSB sine

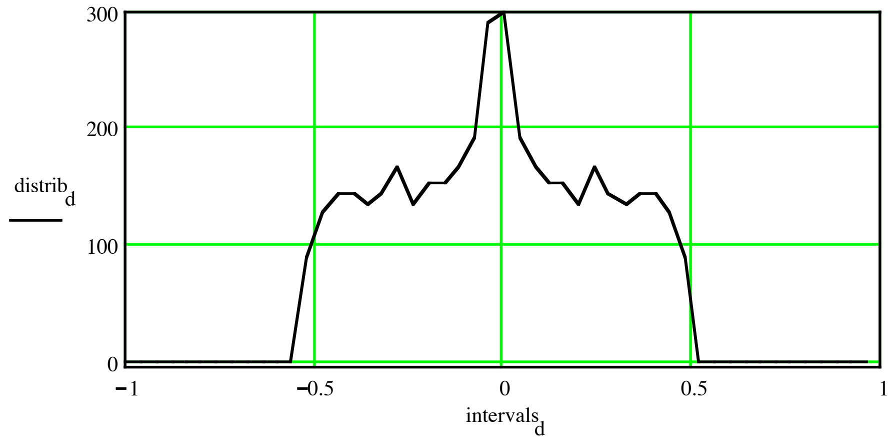


Fig. 5. Histogram of quantization error of 4 LSB sine

For complex signals the quantization error looks more like noise and the frequency distribution for the histogram is nearly uniform. This quantization error is sometimes called quantization noise. Noise with uniform distribution between —0.5 and 0.5 has a RMS level of $1/\sqrt{12}$. If the quantization has B bits (2^B intervals), the largest RMS sinusoidal level that can be quantized without clipping is $2^{B-1}/\sqrt{2}$. The ratio between these two levels in decibels is 1.76 + 6.02xB. This is the ratio between the quantization error and the largest sine, presuming the quantization error is noise with uniform distribution, which is certainly not the case for DC and sine signals.

The spectral components of the quantized signal that are not in the original spectrum, are also sometimes called quantization noise. This is different from the quantization error, due to the level difference after quantization. Using the above formula for the output “noise” after quantization is incorrect.

4. Quantization and sampling

Sampling the signal changes the quantization error. Sampling before quantizing or quantizing before sampling is mathematically the same. The spectrum of the sampled quantized signal is a periodic repetition of the spectrum of the quantized signal. Sampling gives aliasing : spectral components of the quantized signal above the Nyquist frequency (= half the sampling frequency) appear below the Nyquist frequency. If the sampling frequency is an integer multiple of the sine frequency, the aliased harmonics will coincide with other harmonic components.

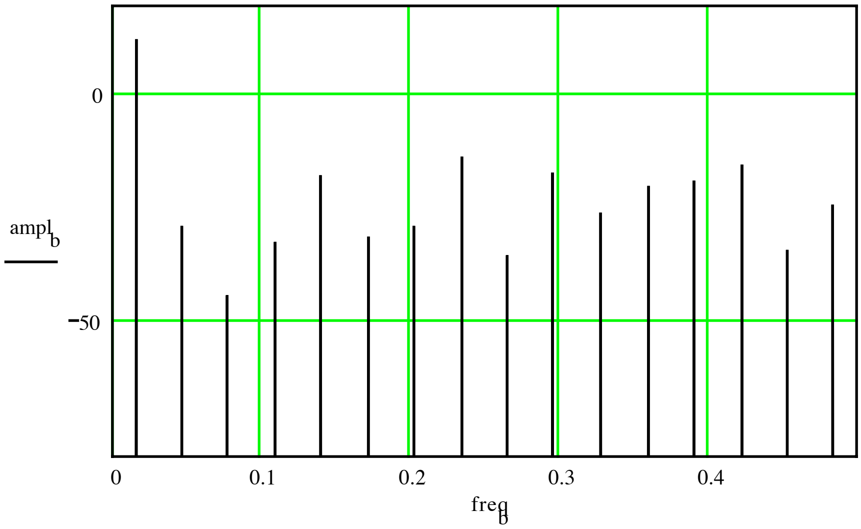


Fig. 6. Spectrum of digitized 4 LSB sine, frequency = 1/64 x sampling frequency

If the sampling frequency is not an integer multiple of the sine frequency, the aliased harmonics fall between the original harmonics.

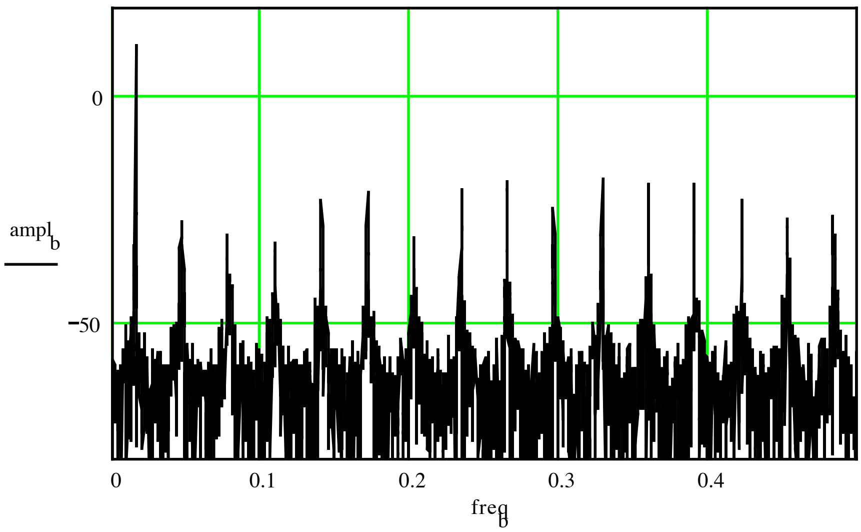


Fig. 7. Spectrum of digitized 4 LSB sine, frequency = 64/4095 x sampling frequency

Although there are spectral components between the lower harmonics, there still is no noise in this spectrum just aliased harmonics. Doubling the DFT length gives the same spectrum, while a noise floor would be 3 dB lower.

When the sine frequency is irrational versus the sampling frequency, you can not directly evaluate the spectrum with a DFT. Since the data length can not be infinite, it must be limited in time. This is called windowing of the time-domain or apodization of the frequency domain. Applying a DFT means calculating the spectrum of a signal that is the periodic repetition of the windowed one, which is not correct in this case. The effects in the spectrum are called leakage or smearing. Weighting functions are used to reduce some consequences of leakage : broadening of the spectral peaks and wrong amplitudes of the spectral peaks.

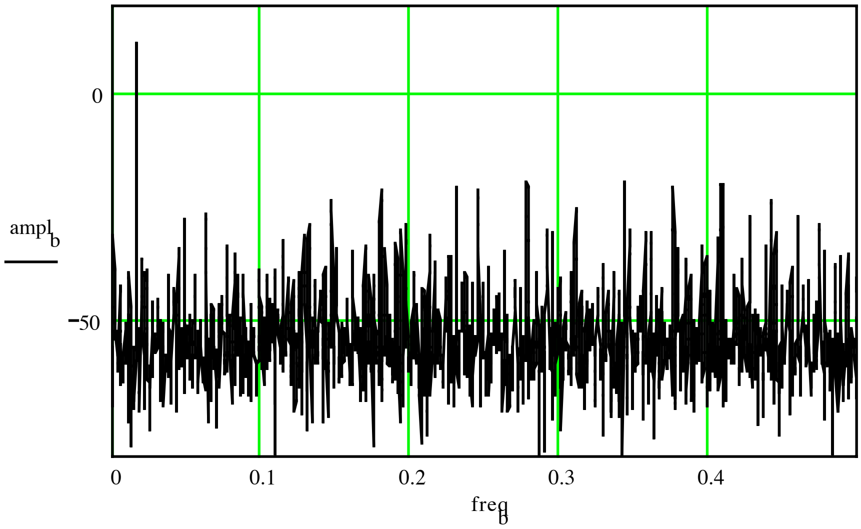


fig. 8. Spectrum of digitized 4 LSB sine, frequency = Π/192 x sampling frequency

It is difficult to recognize the lower harmonics midst the aliases and leakage. Once again there is no noise present, just smeared out harmonics, but this time it looks like a distortion-free noisy spectrum.

Averaging a few of these spectra, flattens out the spectrum except for the first harmonic. Since the sine frequency is irrational versus the sampling frequency, the averaging is not synchronous with the sine frequency and the aliases and leakage are smeared out even more.

Conclusion : Using a pure sine signal to evaluate the performance of quantizers like A-D converters just gives distortion components. Using a DFT to evaluate the spectrum can make the distortion look more like noise depending on the ratio between the sine signal and the sampling frequency. Presuming that there are only baseband harmonics and thus deducing THD and SNR from the DFT spectrum is wrong and gives arbitrary results, because the energy of the quantization error will be distributed between the THD and SNR depending on the frequency ratio. In the examples the RMS of the quantization error was always the same 0.27 LSB. So beware when you see THD and SNR specifications.

This misconception is probably due to Philips who tried to convince the consumer public at the launch of the compact disc that digitization of audio gave no distortion, just a very low noise floor. To prove their point they showed a spectrum of a truncated sine wave with frequency = sampling frequency / 44.55555… . This irrational frequency and the averaged spectrum gave the wanted figure, as concluded above.

5. Non-ideal digitization

Whereas requantization in the digital domain is fully described above, A-D and D-A conversion has some deviations.

In an ideal converter, the quantizing intervals are exactly 1 LSB apart. In a real convereter there is “nonlinearity”[4], usually expressed as DNL (differential nonlinearity) and INL (integral nonlinearity). The quantization steps deviate from the ideal levels and the quantizing intervals are not the same. This makes the quantization error and amplitude error larger : 0.33 LSB RMS in stead of 0.27 LSB RMS for the quantization error in the example of figure 9. With DC and sine signals there still is no noise but more distortion.

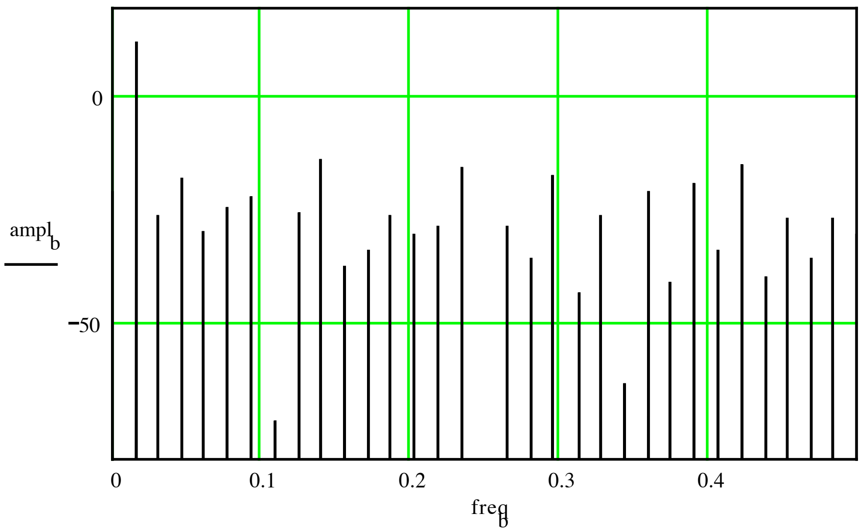


Fig. 9. Same as fig. 6 with one 0.5 LSB deviation in the quantization

In every analog component noise is generated, so the quantization levels are noisy due to voltage reference noise, resistor noise, comparator noise etc. For A-D conversion this is the same as adding noise before quantizing, which is dithering and will be discussed in the next chapter. For D-A conversion this is the same as adding noise to the output signal.

There is also noise on the clock, which is called jitter. Describing the effects of jitter is too comprehensive for this report. The effect on the spectrum of a digitized sine is smearing of the energy of the harmonics into adjacent frequencies.

6. Dither

The effect of quantization on an image with a slowly varying gray scale is contouring. The use of noise in masking contour effects was studied by Roberts[5]. Noise was added before quantization and the same noise was subtracted after reconversion, and the term dither became widespread. Blesser[6] examined digitization of audio and used the term dither for noise added before quantization, without the addition-subtraction scheme.

Adding a noise signal with a RMS amplitude of half the step size, linearizes the quantization and increases the resolution well below the quantizer step size[7]. The first noise sources used were

[4] A deviation from a uniform quantizer is meant. Since the tranfer function is a staircase, a linear quantizer is a contradiction.

[5] L. Roberts, “Picture Coding Using Pseudo-Random Noise”, IRE Trans. Inform. Theory, vol. IT-8, pp. 145-154, 1962

[6] B. Blesser, “Digitization of audio: A Comprehensive Examination of Theory, Implementation and Current Practice", J. Audio Eng. Soc., vol. 26, 1978 Oct.

[7] J. Vanderkooy and S. Lipschitz, “Resolution Below the Least Significant Bit in Digital Sysytems with Dither”, J. Audio Eng. Soc., vol. 32, no. 3, 1984 March

analog white noise which has a normal distribution: a probability density corresponding to a Gaussian function. For quantization in the digital domain, generation of noise with a Gaussian probability density function (GPDF) is very complex compared to generating a random number. A random or pseudo-random number has a rectangular probability density function (RPDF). For perfect linearization the peak-peak amplitude needs to be an integer number of LSBs[8]. Quantization of a DC signal with a level equal to 3.5 LSB with 1 LSBpp RPDF dither (0.289 LSB RMS), gives a signal that is by turns 3 or 4 LSB. The average value is 3.5 LSB, so the original amplitude error of 0.5 is gone which illustrates the linearizing effect of dither. The dither + quantizer transfer characteristic is no longer the staircase of figure 1 but becomes a straight line thus realizing infinite resolution.

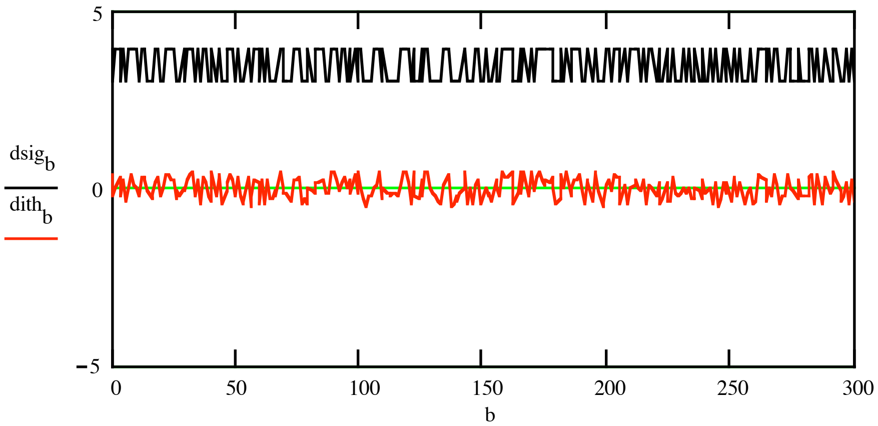
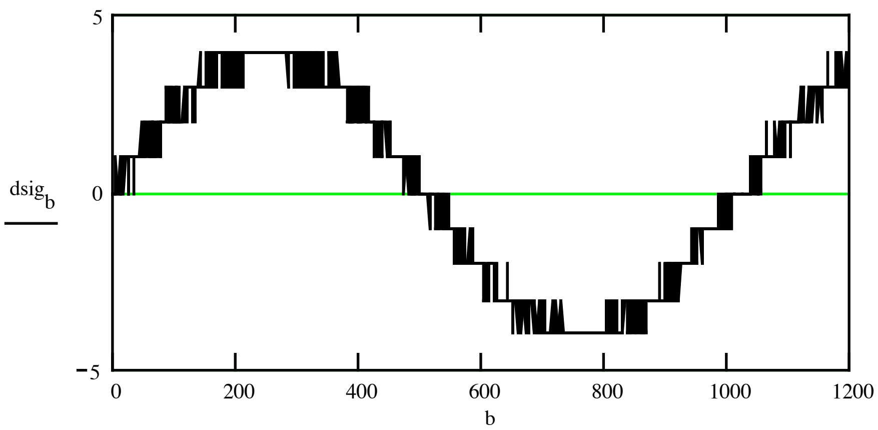


Fig. 10. Quantization of 3.5 LSB DC with 1 LSBpp RPDF dither

This time there is noise in the quantized signal with a discrete PDF with two levels: + and — 0.5 LSB. The RMS amplitude of the noise is 0.5 LSB. Quantization of a DC signal with a level equal to an integer number of LSBs gives the original signal without noise. Depending on the input level, the output noise varies between 0 and 0.5 LSB. This noise modulation can be an advantage in a digital PLL or unwanted in audio. Since the linearization is not perfect for other amplitudes of RPDF dither, it can't fully linearize non-uniform quantizers like real A-D and D-A converters.

What about quantization and dithering of sine signals ?



[8] P. Carbone and D. Petri, "Effect of Additive Dither on the Resolution of Ideal Quantizers", IEEE Trans. Instrum. Meas., vol. 43, no. 3, 1994 June

Fig. 11. Quantization of 4 LSB sine with 1LSBpp RPDF dither

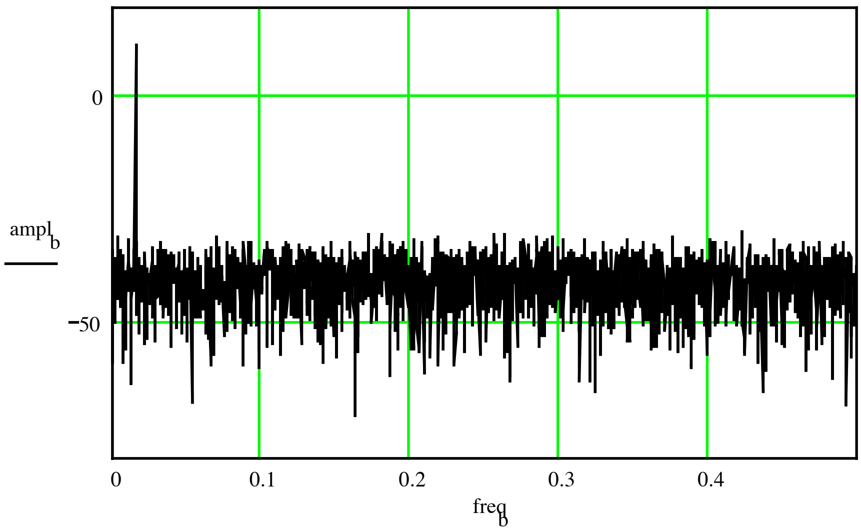


Fig. 12. Spectrum of digitized 4 LSB sine with 1 LSBpp RPDF dither

Comparing the spectrum of fig. 12 with the spectrum of fig. 6, it is obvious that the distortion is gone and is replaced with noise. A plot of the quantization error shows a noise signal, whereas in fig. 4 this was not the case. The amplitude of the first harmonic is correct, so the output noise is equal to the quantization error and in this case can be correctly called quantization noise. The dither signal has a 0.289 LSB RMS level, the output noise has a 0.386 LSB RMS level, but depends on the sine level. Compared to the largest sine that can be quantized without clipping, the signal-to-noise ratio in decibels is $-0.76 + 6.02xB$, this is 2.52 dB less than the often quoted value. If you see a larger SNR quoted, it must be wrong : either it is not noise but distortion or the signal and noise were measured separately, which excludes the quantization error.

Another dither signal is noise with a triangular PDF. For perfect linearization the peak-peak amplitude needs to be an integer multiple of 2 LSB. TPDF dither is less sensitive to the correct amplitude than RPDF dither. With 2 LSBpp there is no noise modulation. The amplitude of the quantization noise is 0.5 LSB RMS, 2.25 dB more than with RPDF dither. TPDF dither is easy to generate in the digital domain by adding two uncorrelated RPDF dither signals.

Analog noise has a Gaussian PDF. There is sufficient linearization with a 0.33 LSB amplitude of the dither signal. The quantization noise has a 0.43 LSB amplitude with very little noise modulation. There is no specific value were the linearization is perfect but the deviation from linearity goes monotonous to zero with increasing amplitude[9]. This is better for linearizing non-uniform quantizers such as A-D and D-A converters. The dither amplitude has to be increased depending on the nonlinearity in the converter.

This is illustrated with the quantization of a 4 LSB sine signal by a quantizer with a 0.5 LSB error located at 2 LSB.

[9] M. Wagdy and M. Goff, “Linearizing Average Transfer Characteristics of Ideal ADC’s via Analog and Digital Dither”, IEEE Trans. Instrum. Meas., vol. 43, no. 2, 1994 April

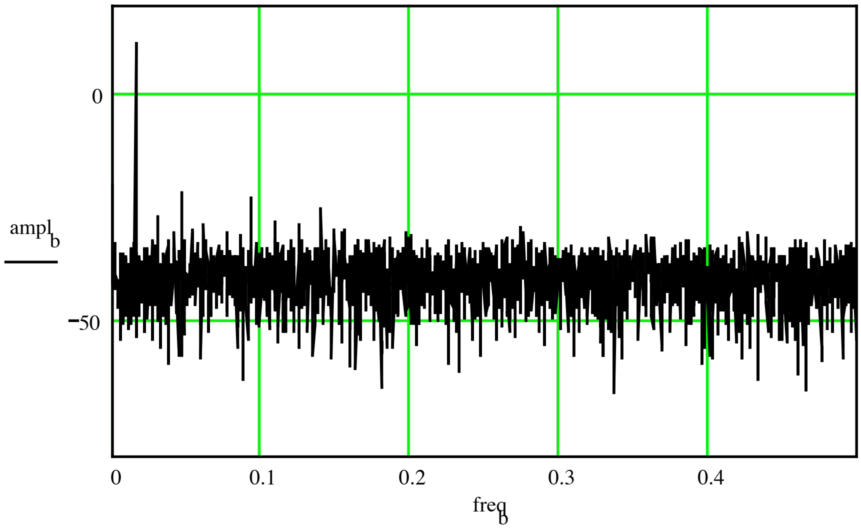


Fig. 13 Non-uniform quantizer with RPDF dither

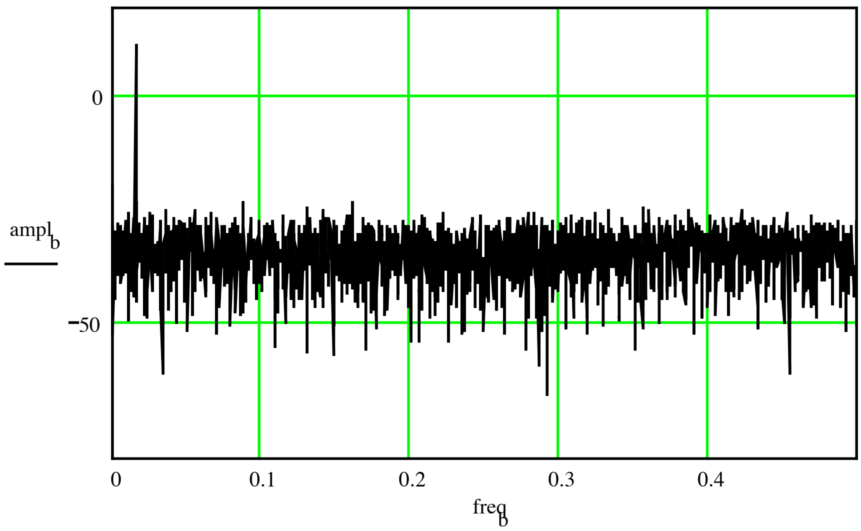


Fig. 14 Non-uniform quantizer with 0.7 LSB GPDF dither

With the rectangular dither there still is distortion with a non-uniform quantizer. The distortion disappears with Gaussian dither of sufficient amplitude, but there is more noise.

The dither signal used in the examples was always white noise, which means having a flat spectrum. Since all remarks only refer to the amplitude of the dither signal and the PDF, it is clear that the dither may be spectrally shaped, even narrowband. A blue noise dither has a 10 dB per decade rising spectrum, as shown in figure 15 on a logarithmic frequency axis. This dither signal is a differentiated 1 LSBpp RPDF white noise, which gives a 2 LSBpp TPDF blue noise.

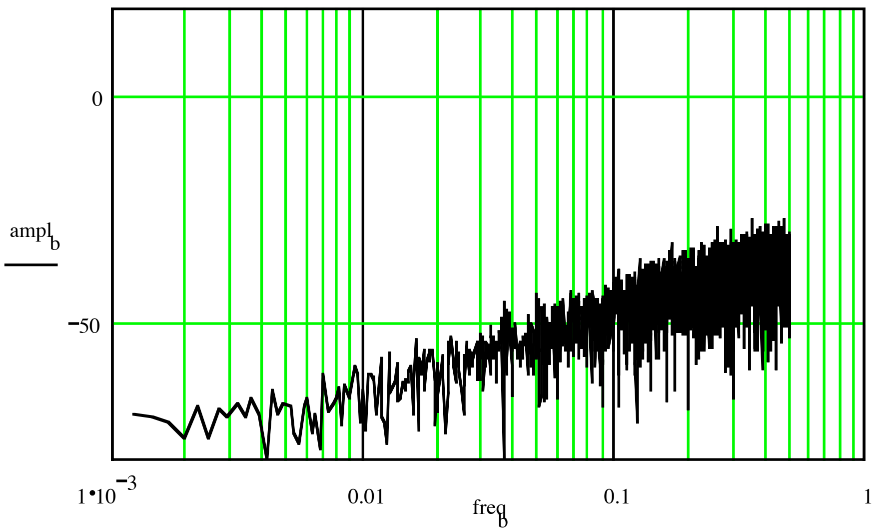


Fig. 15. Spectrum of 2 LSBpp TPDF blue dither

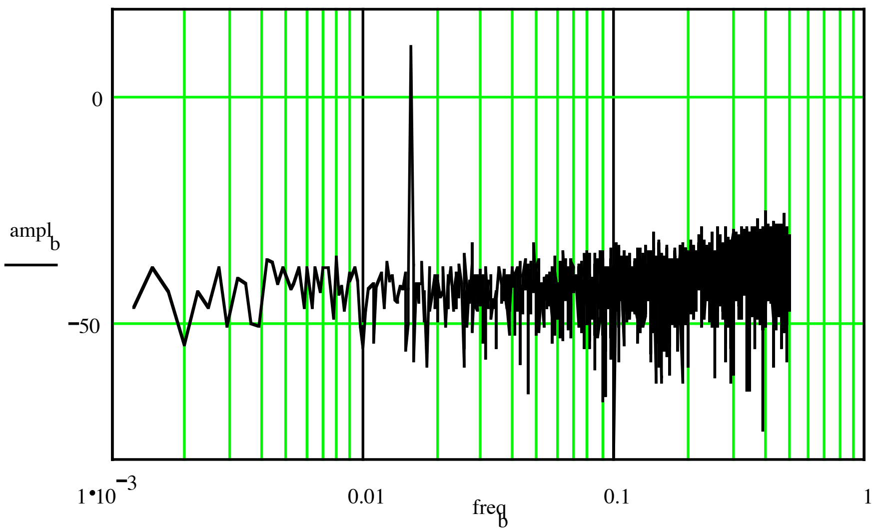


Fig. 16. Spectrum of quantized 4 LSB sine with 2 LSBpp TPDF blue dither

For high frequencies where the noise energy of the dither signal is largest, the spectrum of the quantized signal has the same noise energy as the dither source. For low frequencies, the spectrum of the quantized signal has a flat noise floor that is 4 dB less than with white TPDF dither. The amplitude of the quantization noise is 0.5 LSB RMS, the same as with white TPDF dither.

7. Subtractive dither

For large "nonlinearities" of the converter, the dither amplitude has to be increased and the output noise increases. With subtractive dither, the dither signal is subtracted from the quantizer output. This is only useful when doing A-D or D-A conversions and not when requantizing in the digital domain. The same dither signal has to be available in the analog and digital domain. A digitally generated dither can be converted to the analog domain, or an analog noise signal can be digitized. Whatever the amplitude of the dither, the noise amplitude will remain the same. So large dither amplitudes can be used to increase the linearity of the conversion process. There is a tradeoff because there is an input overhead for the dither amplitude, so the maximum input-signal amplitude is less. HP uses a peak to peak dither level equal to a half scale of the converter range in the HP 89410A vector signal analyzer[10], the maximum input-signal amplitude is then also a half scale. 6 dB SNR is lost but 4 bits (24 dB) linearity are gained.

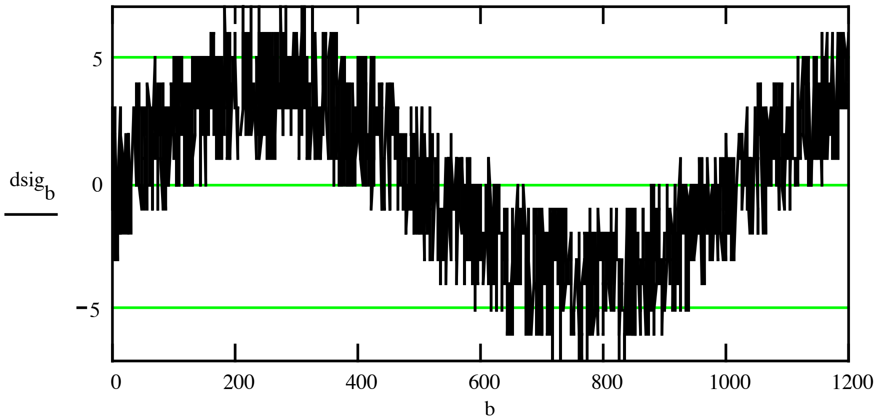


Fig. 17. Quantization of 4 LSB sine with 6 LSBpp RPDF dither

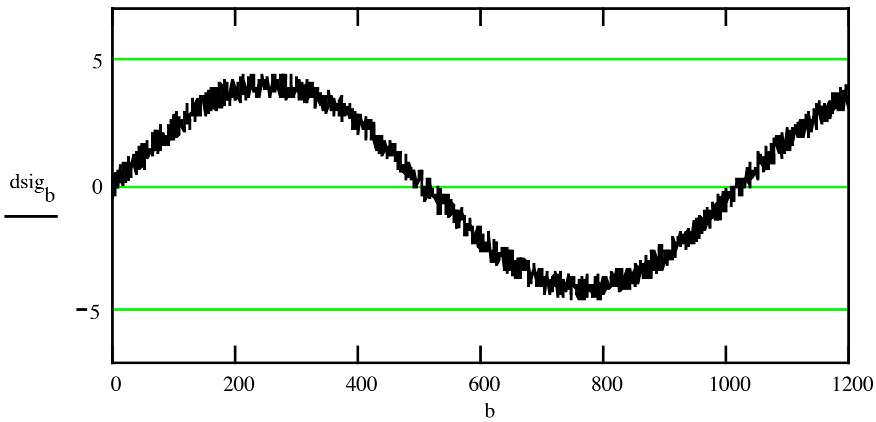


Fig. 18. Quantization of 4 LSB sine with 6 LSBpp RPDF subtractive dither

The RMS noise level in figure 17 is 1.74 LSB and 0.289 LSB in figure 18. Whatever the dither signal (but with more than 0.5 LSB amplitude), the SNR in decibels is 1.76 + 6.02xB. So the often quoted formula for SNR is correct for subtractive dither. Great, isn't it? Well, the B in the formula stands for the number of bits in the quantizer, not the number of bits in the output signal which is much larger. Limiting the extra number of bits in the output to 2, gives acceptable results. This is requantization, but since there is noise present, it is dithered quantization. The noise level after the requantizer is 0.297 LSB which is 0.24 dB more. To achieve the low noise, the digital and analog level of the dither signal must be exactly the same. This demands a trimming or an automatic loop.

Conclusion : Dither is very useful for linearizing quantizers, even non-uniform quantizers. The

[10] M. Bartz e.a., "Baseband Vector Signal Analyzer Hardware Design", HP Journal, vol. 44, no. 6, 1993 Dec.

importance of dithering can not be emphasized enough and it is used too little. The linearity of a dithered quantizer is not related to the number of bits, but the SNR is. In measurements, the noise floor is lowered with averaging and the resolution becomes independent of the number of bits in the quantizer.

8. Noise shaping

With noise shaping, the spectrum of the quantization error is shaped. Noise shaping and quantizing is also known as sigma-delta modulation.

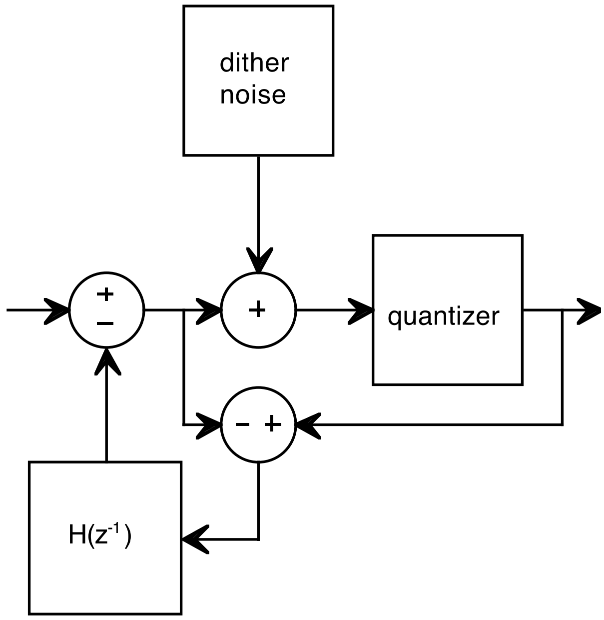


Fig. 19. Noise shaped quantizer

The simplest application is without dither and the filter being just one delay : $H(z^{-1}) = z^{-1}$. This is called a first order noise shaper. Quantization of a 3.1 LSB DC signal gives at the output a signal that has 9 samples a value of 3 LSB, then once 4 LSB, this sequence always repeating. The DC value of the output signal is 3.1 LSB, so noise shaping linearizes the quantization. The repeating sequence gives peaks at multiples of 1/10 of the sampling frequency in the output spectrum, these are called idle tones. The quantization error contains no noise, just idle tones and the amplitude is 0.3 LSB RMS. Quantization of a DC signal with a level equal to an integer number of LSBs, e.g. 3, gives no quantization error. Quantization of a DC signal with a level equal to an integer number of LSBs ± 0.5, e.g. 3.5, gives a maximum quantization error of 0.5 LSB and the maximum idle tone frequency equal to the Nyquist frequency.

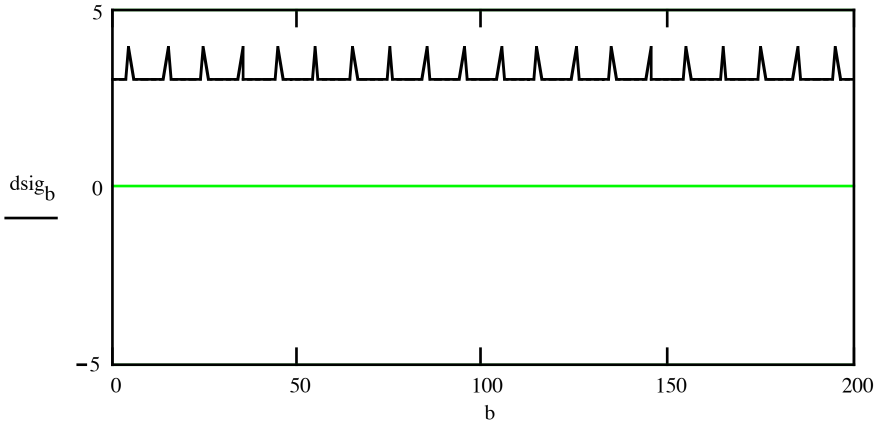


Fig. 20. Quantization of 3.1 LSB DC signal with 1st order noise shaper

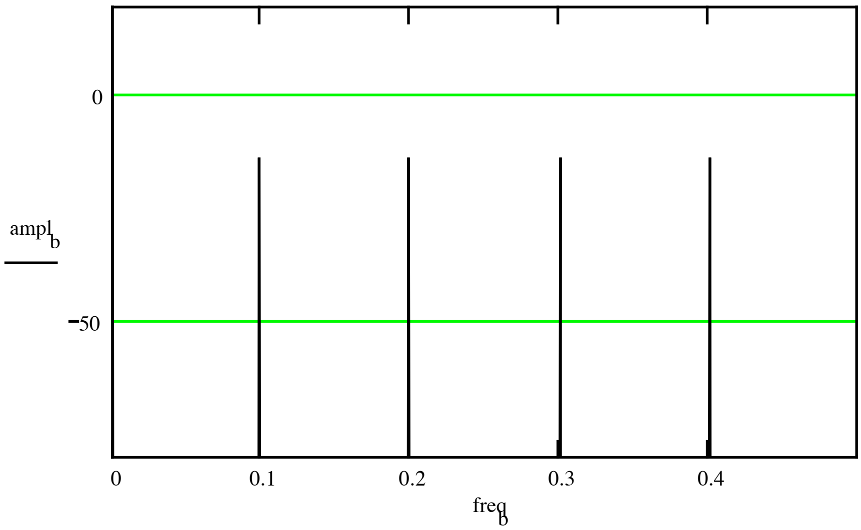


Fig. 21. Spectrum of quantized 3.1 LSB DC with 1st order noise shaper

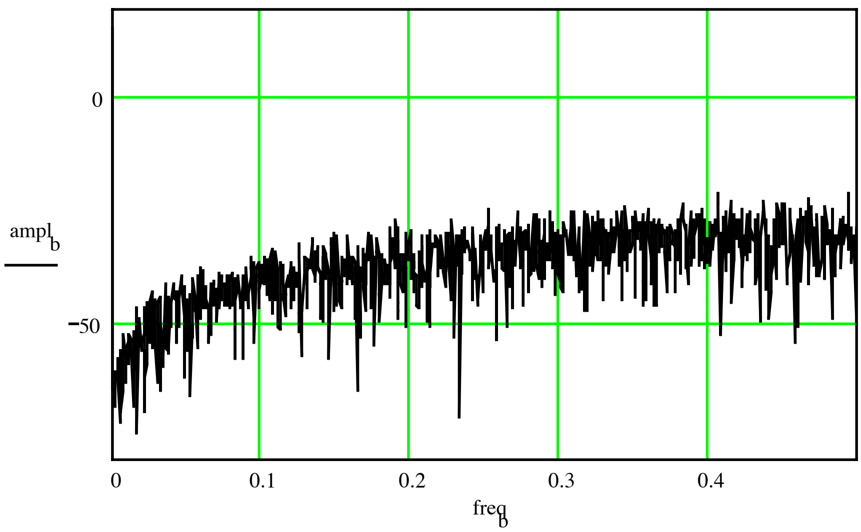


Fig. 22. Spectrum of 3.1 LSB DC with 1st order noise shaper and 1 LSBpp RPDF dither

The unwanted idle tones can be eliminated by dithering the quantizer.

The quantization error is spectrally shaped and has the same spectrum as figure 15 on a logarithmic scale. The dither and quantization noise are filtered by the frequency response of the noise shaping filter : $1 - H(z^{-1})$.

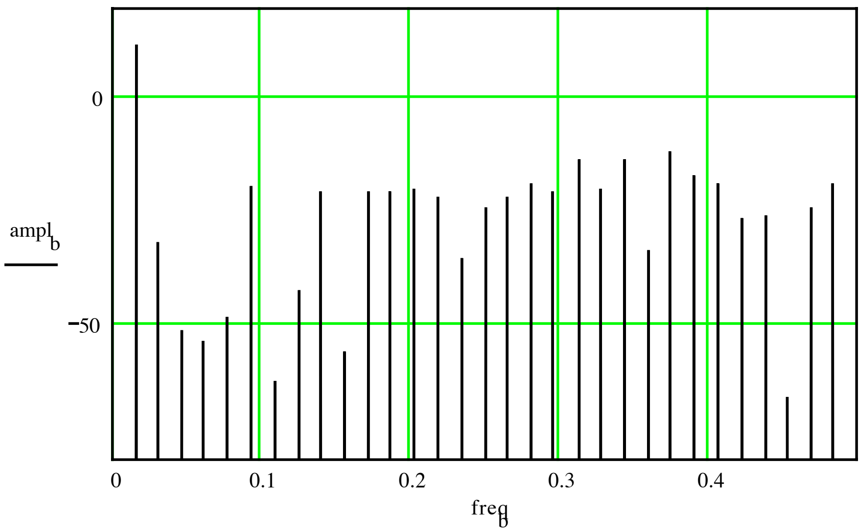


Fig. 23. Spectrum of 4 LSB sine signal with 1st order noise shaper

Noise shaping does not eliminate the harmonics when quantizing sine signals. Dithering is necessary. An important difference is that with noise shaping, the amplitude of the quantization noise is **higher**. For the example of the 4 LSB sine signal with 1 LSBpp RPDF dither, the quantization noise is 0.386 LSB RMS without noise shaping and 0.654 LSB RMS with noise shaping, The signal-to-noise ratio is in this example 4.58 dB worse with noise shaping.

From the schematic of figure 19, it is easy to understand why the quantization noise is higher with noise shaping. At the input of the quantizer there is the dither noise summed with the differentiated quantization error, and the sum of these two uncorrelated noise signals is always higher in amplitude than the amplitude of one noise signal.

Since the amplitude of the noise at the input of the quantizer is higher than the amplitude of the dither noise, perhaps a lower amplitude of the dither is sufficient? This is indeed the case, but I did not find an exact analysis, the noise shaping loop makes an analysis difficult and the simulations become very time-consuming. For RPDF dither, there is no more an exact amplitude for perfect linearization and a 0.5 LSB peak to peak amplitude seems to be a good value. For the example of the 4 LSB sine signal, the quantization noise is 0.53 LSB RMS with 0.5 LSBpp RPDF dither.

With a Gaussian PDF there is sufficient linearization with a 0.2 LSB amplitude of the dither signal. The quantization noise has a 0.55 LSB amplitude, but there is noise modulation.

There is more quantization noise with noise shaping[11]. But the fact that the noise is spectrally shaped can be an advantage. For digital audio, the psycho-acoustic weighted noise energy is slightly less with first order noise shaping. Higher order filters can minimize the weighted noise energy. If the signal bandwidth is much less than the Nyquist frequency, the signal is said to be **oversampled**. If a filter is applied at the output that rejects the noise energy outside the signal bandwidth, the noise amplitude is lowered. With 4 times oversampling and filtering, the noise bandwidth is one fourth of the Nyquist frequency and the noise amplitude is 6 dB less for a white noise spectrum. So 4 times oversampling and filtering gives one bit extra resolution. With first order noise shaping, the noise amplitude is 18 dB less for 4 times oversampling and there are three bits extra resolution. With second order noise shaping, the noise amplitude is 30 dB less after filtering. There are many significant bits present at the output of the filter and the order, filter characteristic and oversampling ratio can be chosen so that only a 1 bit quantizer is needed. Converters working on this principle are often called 1 bit converters, although sometimes the dithering is "forgotten". In 1 bit converters most processing can be done in the digital domain which is an advantage for integration. The effect of clock jitter is worse with oversampling.

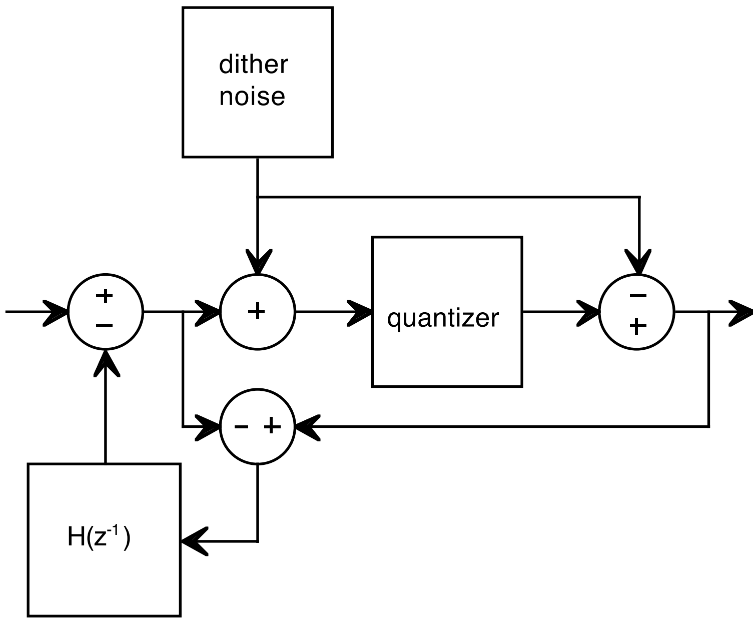


Fig. 24. Noise shaped quantizer with subtractive dither

Subtractive dither is also possible with noise shaping. The subtractive dither is applied directly to the quantizer, within the error feedback loop.

[11] No reference was found for this. There are no results published about the dither and noise amplitude with noise shaping

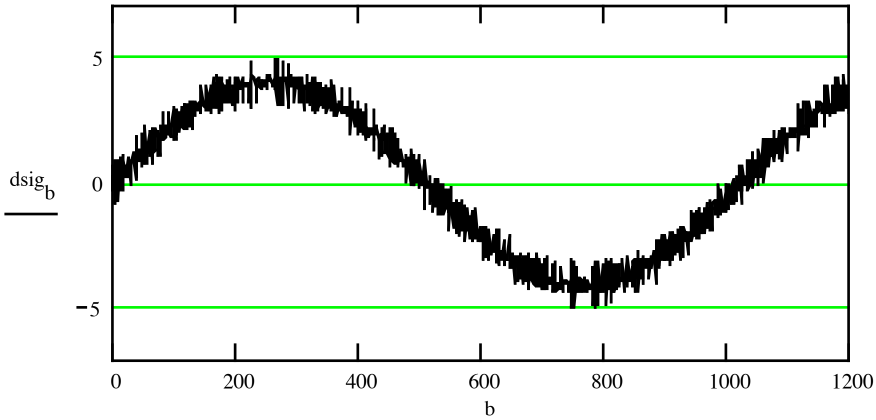


Fig. 25. Noise shaped quantization of 4 LSB sine with 6 LSBpp RPDF subtractive dither

The RMS noise level in figure 25 is 0.5 LSB and is 4.76 dB higher than without noise shaping. Subtractive dither can also be used to bury data in a low level noise, as is done in the HDCD process.

9. Conclusion

Beware when SNR and THD are specified for converters.
Quantization is a non-linear process and gives amplitude errors and distortion, not noise.
Sampling can make the distortion look like noise.
Dithering with correct amplitude removes the amplitude errors and distortion in exchange for added noise. A higher amplitude dither removes distortion in non-uniform quantizers in exchange for more noise, but subtractive dither eliminates this extra noise.
Noise shaping eliminates the amplitude errors but not the distortion in exchange for idle tones.
Adding dither to noise shaping eliminates the distortion and idle tones. The noise shaping allows one to choose the optimum noise spectrum but the non-weighted noise level is higher.
For noise shaping (ΣΔ) there is a lot of research going on and a more exact analysis and other applications will evolve.

Peter BRACKE ((3445)