Pulse-Code Modulation—An Overview*

STANLEY P. LIPSHITZ, AES Fellow, AND JOHN VANDERKOOPY, AES Fellow

Audio Research Group, University of Waterloo, Waterloo, Ont. N2L 3G1, Canada

Pulse-code-modulation (PCM) encoding of digital audio signals has had a long and successful history in the era of the Compact Disc (CD). This brief survey paper argues that it forms the logical way to extend either the bandwidth or the signal-to-noise ratio of a digital audio system, or both, to encompass even higher resolution. Underpinning its operation there are the iron-clad theorems that govern both the sampling-and-reconstruction and the dithered-quantizing processes that lie at its heart. It is adaptable enough to allow fully distortion-free noise shaping to be used if wordlength reduction is necessary, provided that the wordlength is not reduced so far as to cause quantizer overload when using proper dithering.

0 INTRODUCTION

Now that high-resolution digital audio is in regular use for mastering purposes in the professional audio field, and has also begun to make its appearance in the consumer marketplace, it seems appropriate to review briefly the technical merits and operating principles of the progenitor of all such formats, namely, linear pulse-code modulation (PCM). Digital audio depends for its operation on the validity of the two fundamental processes that are used to convert an analog signal to a digital form: sampling and quantization. These two operations are frequently combined in the analog-to-digital converter (ADC). The signal reconstruction operation takes place in the digital-to-analog converter (DAC), where the analog waveform is regenerated. Stripped down to its basics, the sampling rate determines the bandwidth of the signal conversion, and the number of steps, or least significant bits (LSBs), used in the quantization determines its signal-to-noise ratio (SNR). As such, it is in principle possible to envisage PCM systems of arbitrarily large bandwidth and/or SNR.

The validity of the sampling and reconstruction processes for band-limited analog signals is guaranteed by the sampling theorem. Frequently ascribed to Shannon (1948), Kotel’nikov (1933), or more usually Nyquist (1928), the sampling theorem was actually first published by Whittaker as early as 1915 [1]. The consequences of quantizing the signal samples, so that they can be represented by finite-precision digital words, are the introduction of an unavoidable quantization error. The discovery of how to dither the quantization operation in order to decorrelate the quantization error from the signal itself, so that it appears only as an innocuous noiselike signal, is much more recent [2]–[4]. When these uniformity quantized samples are represented by digital words, they form the digitized data stream of a linear PCM system. In the context of high-resolution digital audio we shall consider PCM to denote a uniformly sampled and uniformly quantized signal, using a number of binary digits (bits), which could range from as few as 1 to as many as 24 or more. As we shall see, if the number of bits used is insufficient to furnish the desired SNR, oversampling and/or noise shaping must be used. This permits trading sampling rate for number of bits by compromising the SNR over a less important portion of the signal band in exchange for an increased SNR over a more critical portion of the band. Thus one finds oversampled noise-shaped PCM systems in use in many places. An extreme case of this tradeoff is the use of low-bit or even 1-bit noise shapers (also known as sigma–delta modulators) for signal conversion. We shall also investigate the applicability of the theorems on dithered quantization to such low-bit systems.

This paper is intended to be tutorial in nature, and as such it will not try to present a rigorous development of the theory, or even to address all aspects of this theory. Rather, we shall use copious illustrations to emphasize some of the most important aspects of PCM systems, and try to clarify some issues which are frequent sources of misunderstanding and confusion. An outline of the paper is as follows. Section 1 addresses the sampling and reconstruction of signals. Section 2 discusses the proper quantization of such signals using dither. Section 3 illustrates how noise shaping can be applied to a PCM system. Finally, Section 4 summarizes our conclusions.

*Manuscript received 2003 October 9.
1 SAMPLING AND RECONSTRUCTION

The sampling theorem guarantees that any band-limited signal can be exactly reconstructed (in principle, at least—in practice nothing can be done exactly) from its time samples provided that the original signal contains no frequency components at or above one-half the sampling frequency \( f_s \). The frequency \( f_s/2 \) is called the Nyquist frequency. Mathematically this sampling and reconstruction process can be represented by the formula

\[
f(t) = \sum_{k=-\infty}^{\infty} f \left( \frac{k}{f_s} \right) \sin \left[ \pi f_s \left( t - \frac{k}{f_s} \right) \right] \frac{\sin \left[ \pi f_s \left( t - \frac{k}{f_s} \right) \right]}{\pi}.
\]

where \( f(t) \) is the original continuous-time signal whose samples, at the sampling rate \( f_s \), are \( f(kf_s) \), \( k = -\infty \ldots \infty \), and where \( \sin(x) \) denotes the reconstruction function \( \sin(\pi x)/\pi x \). There are two important things to note about this result. First the original analog signal must be band-limited to frequencies less than the Nyquist frequency \( f_s/2 \); otherwise the reconstructed signal will be falsified by the presence of aliases, caused by the “folding down” of higher frequency components into the baseband. To guarantee that this does not occur, either the input signal must be passed through a band-limiting anti-aliasing filter before entering the ADC, or else the sampling rate must be set high enough. (By oversampling the signal we mean sampling it at a rate greater than the rate that would otherwise suffice for its natural bandwidth. Thus the natural band limit of the input signal must still be below one-half of the new sampling frequency.) A suitable anti-aliasing filter is invariably included as part of the design of any digital audio system. Second, a perfect reconstruction of the original analog waveform from its samples requires that the samples be passed through an ideal “brick-wall” reconstruction filter. It is this filter that corresponds to the presence of the sinc function in the preceding formula since the impulse of an ideal brick-wall filter is a sinc (they are Fourier transforms of each other). We see that a perfect reconstruction is a sinc reconstruction—each sample must be multiplied by an appropriately positioned sinc waveform, and these scaled sincs must then be added together to reconstruct the original signal. We shall illustrate this process shortly.

First a comment about the anti-aliasing and reconstruction filters that the sampling theorem demands. Perfect brick-wall filters are a mathematical fiction. They can be very closely approximated, but never exactly achieved. Moreover, they have a “ringy” (namely, sinc-like) impulse response. There is thus much discussion of the tradeoffs between different filter alignments, and their corresponding magnitude and phase responses. One can trade off some of the desirable properties of the magnitude response against those of the phase response. We shall not address this question here, as it will be covered in detail in another paper in this issue of the Journal. These are linear filters, and so their effects on the signal are precisely predictable. For expositional purposes we shall assume the use of brick-wall filters in this paper. Suffice it to say that actual audio signals that do not have frequency components right up to (or beyond) the Nyquist frequency will not be modified by the brick-wall filter. The rational choice of anti-aliasing filter alignment is ultimately dependent on psychoacoustic criteria—if the band-limiting filter’s effect can be heard, then it will have audibly changed the signal being fed to the ADC, and hence the signal recovered from the DAC. In such cases oversampling could be used to allow gentler anti-aliasing and reconstruction filters to be used.

What bandwidth is necessary for audible perfection? Which filter impulse responses degrade the signal audibly? There is much pontification on this topic, but little evidence to support some of the claims of the ultrawide-bandwidth exponents. Whatever the ultimate outcome, once the required parameters are known, the sampling theorem is capable of capturing and reproducing the signal perfectly.

Fig. 1 illustrates the sampling theorem in action. Fig. 1(a) shows an original band-limited analog waveform (solid curve) and its samples (the scaled impulses shown at the sample points). To reconstruct the signal, one places a sinc function at each of the sample times, scaled by the amplitude of the corresponding sample, and adds all these (possibly infinitely many) waveforms, thus carrying out the summation prescribed by the sampling theorem. Fig. 1(b) shows the scaled sinc functions corresponding to the waveform portion shown in Fig. 1(a), and the heavy curve shows their sum, including the contributions from sinc functions (not shown) corresponding to samples lying outside the time interval shown. Since all sinc functions are zero at the individual sample times, except for the sinc function based at the sample in question, it is apparent that, with sinc reconstruction, the reconstructed waveform passes through all the original sample points. What is not so obvious, however, but as is clear from the figure, is that the reconstructed waveform is actually identical to the original analog waveform shown in Fig. 1(a), including everywhere between the sample points. This is in accord with the prediction of the sampling theorem. For a band-limited signal, sampled as prescribed by the sampling theorem, the samples contain all the information about the signal—absolutely nothing is lost.

We have illustrated the sampling theorem using impulse samples for the reconstruction process, as a direct application of the theorem. In practice the sample values are invariably held constant between samples (a zero-order hold operation), as this is easier to implement and results in a better achieved SNR. The difference is simply a small frequency response rolloff amounting to 3.9 dB at the Nyquist frequency, which is easily correctable (since this is only a frequency response change—a linear effect), and this correction is usually included in the design of the reconstruction filter.

One often misunderstood aspect of sampled-data systems is the question of their time resolution—can they resolve details that occur between samples, such as a time impulse or step? To show that the time resolution is in fact infinitely

fine for signals band-limited in conformity with the sampling theorem, and is completely independent of precisely where the samples happen to fall with respect to the time waveform, we shall now present some computed examples.

Fig. 2(a) shows a band-limited unit impulse occurring at time zero as well as its samples. In this illustration one sample falls precisely at time zero and captures the impulse directly; all the other samples are zero. During reconstruction only one nonzero sinc function occurs, and it is identical to the original waveform. The solid curve in Fig. 2(a) is thus the overlay of two identical curves—the original band-limited analog impulse and its reconstruction. What happens if the sample times are shifted relative to the time waveform? Fig. 2(b) shows a sampled and reconstructed version of the same waveform with the samples now shifted by one-half the sampling interval. Once again, the solid curve is the overlay of two waveforms—the original analog impulse and its reconstruction from its samples, including those from outside the time window shown. We get back the same waveform, even though now all the samples are nonzero. The time instant of the analog impulse (namely, zero) is exactly recovered, no matter where the samples fall with respect to the impulse.

To emphasize the point, we repeat the exercise with a band-limited step (from $-1$ to $+1$) in Fig. 3. Fig. 3(a) shows a sampled band-limited step, with one sample occurring at time zero, the zero-crossing time. The reconstructed waveform is once again overlaid on the analog original—they are identical. (Of course, the contributions of sincs from outside the time window have been included.

---

![Fig. 1. Sampling and reconstruction process. (a) Original band-limited analog signal (solid curve) and its delta-function samples (arrows). (b) Scaled sinc functions corresponding to samples, and their sum (heavy curve). Heavy curves in (a) and (b) are identical.](image-url)
in this reconstruction, in accordance with the sampling theorem.) In Fig. 3(b) the sample times are shifted by one-half a sample period, so that now no sample occurs at the zero crossing. The solid curve is again the overlay of both the original step and its reconstruction. The reconstructed waveform is identical to the original, and consequently the zero-crossing time is exactly correct. Perhaps things would not work out correctly if the samples occurred asymmetrically with respect to the waveform? Fig. 3(c) shows that this is not so. The time resolution of band-limited waveforms is exactly preserved by the sampling and reconstruction process. It is not necessary to use a high oversampling ratio to ensure the precise localization of signal events in time.

Fig. 2. (a) Sampled band-limited unit impulse at time zero. Sample values are indicated by circles. Since only one sample is nonzero, the single sinc function shown reconstructs the original impulse. (b) Same band-limited impulse as in (a), but samples occur shifted by one-half a sample period relative to those in (a). All samples are now nonzero, but the solid curve (representing the sum of all the reconstruction sinc functions) is exactly the same as before, with its peak at time zero.
This being the case, the only question remaining is: how rapidly should we sample? The answer is: not more slowly than required to achieve the desired bandwidth and/or transient response. If, for whatever reason, a faster rise time is desired, the sampling rate (and hence the baseband width) must be increased. Then within the limits of our ability to actually build ADCs and DACs of sufficient quality running at sampling rate $f_s$, we can approximate the theoretical ideal to an extremely high precision. Other papers in this issue will address the actual performance of high-resolution ADCs and DACs.

2 QUANTIZATION AND DITHERING

The order in which the sampling and quantizing operations are carried out is, in principle, irrelevant, but in prac-

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{fig3}
\caption{Reconstructed signal is independent of sample location relative to time waveform (here a band-limited step). (a) One sample occurs at zero-crossing time. (b) One-half sample shift places samples symmetrically about time zero. (c) Sample locations are asymmetrically distributed relative to time zero. In all three cases, reconstructed waveforms (solid curves) are identical, and zero-crossing time is the same.}
\end{figure}
The sampling operation is fundamentally lossless for band-limited signals, the quantizing process must degrade the signal. When the analog sample voltage is converted to a finite-precision representation, a quantization error is inevitable, no matter how many bits are used in the representation. Even high-resolution audio of 24 bit suffers from quantization errors, albeit at an extremely low level. An ideal noiseless analog system (theoretically unrealizable) could be considered to be the infinite-wordlength limit of a digital system. Quantization errors are always present near the noise floor of any digital system. The best we can hope to do is to control their nature. The number of bits of precision used in the quantization determines the level of this noise floor—the greater the number of bits, the lower the noise floor relative to full scale, and hence the greater the SNR. Quantization errors can be either innocuous or pernicious, depending on the level and properties of the signal being quantized. It is a signal-dependent error. For loud, complex signals it may sound like a constant low-level background white noise accompanying the signal. For low-level, simple signals it can manifest itself as significant harmonic and intermodulation distortion, accompanied by severe modulation of the background noise. This is clearly undesirable. The ideal would be to have the quantization error appear as a low-level white noise whose level is signal independent. This can be achieved by the addition of a suitable dither noise signal during the analog-to-digital conversion process (and indeed during any subsequent purely digital signal manipulations). The theory behind the operation of such nonsubtractive dither is fully developed (see, for example, [2]–[4]). We shall now discuss and illustrate both the undithered and the properly dithered quantization operations.

Fig. 4 shows how a dither signal $v$ should be added to the audio signal $x$, to be quantized before being sent to the quantizer $Q$. It is the sum signal $w$ that gets quantized, producing the quantized output signal $y$. By choosing the properties of the dither signal appropriately, one can control the nature of the resulting quantization errors. The figure also shows the two most commonly used uniform quantization “staircase” functions—the midtread (or

Fig. 4. In a dithered quantizing system, dither noise signal $v$ is added to input signal $x$, and their sum $w$ is fed to quantizer input $Q$. Quantizer output $Q(w)$ is output signal $y$. $Q$ can have either a midtread or a midriser characteristic. $\Delta$—quantizer step size, or least significant bit (LSB).
rounding) quantizer and the midriser quantizer. The former is usually used in multibit systems (although dc drift may make the distinction meaningless), whereas the latter must be used in low-bit systems, and especially in 1-bit systems (which only have two levels). In such systems the available levels must be arranged symmetrically above and below zero to accommodate optimally the bipolar nature of the audio signal, and so maximize the dynamic range.

The symbol $\Delta$ denotes the step size of the quantizer, the least significant bit (LSB). Let $b$ denote the number of binary bits used by the quantizer. It thus has $2^b$ quantum levels (LSBs). Each increase of 1 in $b$ results in a halving of the quantization error, and so a reduction in the error of 6.02 dB. The SNR (in decibels) will thus depend on the product $6.02b$. For an undithered midtread quantizer with $b$ bits and LSB $\Delta$, the “classical” quantization error power is usually assumed to be $\Delta^2/12$. The SNR for full-scale sinewaves is given by the formula

$$\text{SNR} = (6.02b + 1.76) \text{ dB}.$$ 

The additional term of 1.76 dB accounts for the fact that the SNR is referred to a full-scale sinewave of peak amplitude $2^{b-1}$ LSBs ($2^b$ LSBs peak to peak). This formula is sufficiently accurate provided $b$ is at least 8 but is overly optimistic for smaller values of $b$ due to the fact that, for a 2’s complement midtread quantizer, there is one fewer positive than negative output codes available. (In such cases one would use a midriser quantizer characteristic instead.) On this basis, a 20-bit system, for example, should have an SNR of 122.2 dB. The problem is that the quantization error (noise plus distortion) is not constant for an undithered system, and so the term SNR is not really meaningful. We shall return to this issue shortly.

Fig. 5 illustrates the undithered midtread quantization of a small sinewave in a 20-bit system. Fig. 5(a) is the original analog input signal. The sine-wave frequency is 1378.125 Hz, a number chosen so as to make the waveform commensurate with the fast Fourier transform (FFT) record length shortly to be used to view the signal in the frequency domain. The sine-wave amplitude is 2 LSBs, so that this is a very low-level signal (the units on the vertical axis are LSBs). Its quantized version is shown in Fig. 5(b); only five quantum levels are in use. This is clearly a rather distorted “sinewave,” the quantization error $[(y - x)\Delta$ in Fig. 4] being the waveform of Fig. 5(c). The quantization error (bounded by $\pm\Delta/2$) is obviously not noise-like—it is a distortion, highly correlated with the input sinewave. Using a sampling rate of 176.4 kHz (chosen because we are trying to illustrate the aberrant behavior of even high-resolution systems—here we have in mind a 176.4-kHz 20-bit system), we have computed the power spectrum of the quantized waveform of Fig. 5(b) to obtain Fig. 5(d). This power spectrum was obtained from a windowless 16 384-point FFT. The biggest spectral line on the left is the input fundamental. All the remaining lines represent the quantization “noise.” The vertical scale is in decibels relative to the $-122.2$-dB theoretical noise floor of a 20-bit system; that is, 0 dB on this scale represents the classical noise floor of $\Delta^2/12$ total power. Not only is the noise actually a distortion, but many lines have levels far above 0 dB. Undithered quantization is not a good idea.

The proper dither to use is random (or pseudorandom)

---

**Fig. 5.** Undithered quantization. (a) Original analog sine waveform, 2 LSBs in amplitude. (b) midtread quantized version of (a). Only five quantum levels are needed to represent this low-level signal. (c) Quantization error, difference between (b) and (a). This error waveform is highly correlated with original signal (a). (d) Power spectrum of quantized waveform (b).
white noise having a probability density function (PDF) that is triangular and of a peak-to-peak width equal to 2 LSBs. We call this white TPDF dither. Such dither is easily generated. The mathematical theory shows that this is the lowest power dither that will guarantee that there is both zero mean quantization error (that is, there is absolutely no signal-correlated distortion) and constant error variance and power spectral density (that is, the total quantization error power is constant and its spectrum is white). There is consequently absolutely no signal-dependent noise modulation. From the point of view of audibility, the digital system behaves exactly like an ideal analog system, having infinite resolution below the LSB, no distortion, and no noise modulation. The power of the added TPDF dither is equal to $\Delta^2/6$, and so the resultant total quantization noise power is constant at $\Delta^2/4(=\Delta^2/6 + \Delta^2/12)$, three times that of the classical undithered quantizer. Fig. 6 shows the TPDF dither covering a width of 2 LSBs along the hori-

Fig. 5. Continued
horizontal (quantizer input) axis of a notional midriser quantizer. (Of course, TPDF dither is equally applicable to either midriser or midtread quantizers, and is equally effective for both.)

This result immediately raises the problem that 1-bit systems cannot be properly dithered, as Fig. 6 shows that the dither takes up the full no-overload input range of a 1-bit (two-level) quantizer. The consequences of this will be addressed briefly in Section 3. For the moment all we need to note is that TPDF dither occupies 2 LSBs out of the 2^b LSBs possessed by a b-bit system. For multibit systems (say, b ≥ 8) this represents a negligible reduction in dynamic range. The SNR is, however, reduced by 4.77 dB by the effective tripling of the quantization noise power from Δ^2/12 to Δ^2/4. This is an acceptable loss in a multibit system when one considers the benefits; in low-bit quantizers it might be considered unacceptable. The penalty in such systems is that they cannot be totally free of distortion and noise modulation. The SNR of a TPDF-dithered b-bit digital system is given by

\[ \text{SNR} = (6.02b - 3.01) \text{ dB}. \]

For a 20-bit quantizer the SNR is thus 117.4 dB.

Fig. 7 shows the same sinewave as Fig. 5, this time properly midtread quantized with TPDF dither. The TPDF-quantized version of the sinewave of Fig. 7(a) is shown in Fig. 7(b). Note how the dither causes the quantized signal to toggle in a stochastic way between adjacent quantizer levels—two more levels than in Fig. 5(b), as predicted. The TPDF-dithered quantization error [(y - x) in Fig. 4] is shown in Fig. 7(c). It is bounded by ±3Δ/2. Although clearly not statistically independent of the input sinewave, it is in fact completely uncorrelated with both the input signal and itself. It is the white noise component in the quantized output spectrum of Fig. 7(d). This graph represents the average of 64 power spectra, each obtained from a 16 384-point windowless FFT of the simulator data. The correlated distortion lines of Fig. 5(d) have been

![Diagram of quantizer input and output with TPDF dither](image)

Fig. 6. Triangular PDF (TPDF) dither of 2-LSB peak-to-peak amplitude registered at origin of midriser quantizer characteristic. Although TPDF dither occupies only 2 LSBs, this is the full no-overload input range of a 1-bit quantizer.
turned into an innocuous white-noise floor by the TPDF-dithered quantization. Notice that the noise floor appears at a level of +4.77 dB, as predicted by the theory. The fundamental component now also appears at the correct height of +52.9 dB relative to 0 dB. (Its height is 0.3 dB too high in Fig. 5(d) due to the undithered quantization.) A comparison of Figs. 5(d) and 7(d) also shows that many of the distortion lines in Fig. 5(d) rise well above the white-noise floor of Fig. 7(d). The latter is obviously far preferable. It should thus be clear, and it is important to realize, that the distortion has actually been converted into a benign noise. It is not a question of noise masking or

Fig. 7. TPDF-dithered quantization. (a) Original analog sine waveform, 2 LSBs in amplitude. (b) Midtread quantized version of (a) using proper TPDF dither. Seven quantum levels (two more than before) are now needed to represent signal. (c) Resulting total quantization error, difference between (b) and (a). Although clearly not statistically independent of original signal (a), this error waveform is now completely uncorrelated with it. (d) Average of 64 power spectra of quantized waveform (b). Distortion of Fig. 5(d) has been converted into an uncorrelated noise floor. Only the fundamental spectral line is left.
“covering up” the distortion. It should also be noted that the standard AES17 [5] mandates the use of TPDF dither for the proper assessment of digital audio systems.

One adds a dither noise signal to the quantizer input in order to ensure that no matter what its input signal may be, the quantization is free from distortion and noise modulation. Theory shows that independent TPDF white noise is the ideal dither. If the input signal already contains a suitable independent noise component, like the thermal noise generated by analog circuitry, it can act as the dither for an ADC. If this noise is Gaussian, about ½ LSB rms is the correct level to use for adequate linearization of the quantizer (although the linearization is not perfect, and the total quantization noise is about 1.25 dB higher than that produced by the ideal TPDF dither). It goes without saying that full TPDF (digital) dither should also be used in any editing or postprocessing operation on the digital data, which results in a wordlength increase, necessitating a subsequent wordlength reduction, for this is a (digital) requantization of the signal. Failure to do so will introduce
quantization artifacts into the audio signal. It should also be clear that properly dithered quantization preserves the perfect time resolution of proper sampling, since the effect of TPDF-dithered quantization is simply to add an uncorrelated noise to the signal.

Finally it is sometimes misleadingly and improperly suggested that having an adjustable dither PDF and/or width (that is, power level) allows one to “tailor” the dither to the signal. It should be clearly understood that only TPDF dither of the correct width (namely, \(2\Delta\)) has these properties. If the width is reduced, one no longer has zero distortion and zero noise modulation. Furthermore, uniform or rectangular PDF (RPDF) dither of width 0 eliminates all distortion, does not prevent noise modulation, and is not recommendable for high-quality digital audio (and RPDF dither whose width is not an integer multiple of \(\Delta\) eliminates neither the distortion nor the noise modulation).

3 OVERSAMPLING AND NOISE SHAPING

Sections 1 and 2 have outlined the correct way to sample and quantize an audio signal so that the digitized signal has the same characteristics as an analog signal. The bandwidth of the digital recording system can in principle be as wide, and its SNR can be as great, as desired, by choosing \(f_s\) and \(b\) appropriately. Its frequency response can be absolutely flat, and its noise floor can be a signal-independent white noise. It will have no distortion and no noise modulation. It can resolve signals arbitrarily far below the noise floor. It thus behaves like an ideal analog recording system, albeit one with much greater linearity and SNR. So what more could one want? Well, the system just described has a data rate of \(bf_s\) bit/s per channel. It would be nice if the data rate could be reduced. If we drop the demand that the SNR be constant across the whole Nyquist band, it is possible to keep the noise floor low, below, say, 20 kHz, but allow it to rise at high frequencies (where it is psychoacoustically less audible), and so trade wordlength for reduced high-frequency dynamic range.

If we also oversample the audio signal, we can create more space at high frequencies into which the noise can be placed, and potentially further reduce the wordlength needed to achieve the desired SNR below 20 kHz. There is an encoding technique available which can allow us to do this. It is variously called noise shaping or sigma–delta modulation, and it can let us tailor the noise power spectral density curve as a function of frequency, and to an extent trade sampling rate for wordlength.

A simple noise shaper is shown in Fig. 8. The dithered quantizer \(Q\) (with dither noise \(N\)) has its quantization error \(E\) extracted, fed back through the feedback filter \(H\), and subtracted from the input signal \(X\). This is error feedback. It is easy to show that the equation relating the output \(Y\) to \(X\) and \(E\) in the frequency domain is \(Y = X + (1 - H)E\). The signal \(X\) passes through the noise shaper unchanged, but the error \(E\) is shaped by the effective noise-shaping filter \((1 - H)\) and appears at the output as an additive shaped noise \((1 - H)E\). By dithering the noise shaper with a TPDF dither \(N\), we ensure that the error \(E\) is indeed a constant-power white-noise signal, as explained in Section 2, and hence that \((1 - H)E\) is truly a shaped signal-independent noise. Without proper dithering, this cannot be guaranteed, and so \(E\), and \((1 - H)E\), can (and will) contain signal-dependent distortion and noise modulation. Moreover, because of the error feedback and the presence of the nonlinear quantizer, the output can exhibit low-level limit-cycle oscillations if the system is inadequately dithered. The general single-stage sigma–delta modulator topology, shown in Fig. 9(a), is completely equivalent to the general noise-shaper topology shown in Fig. 9(b). So the preceding comments apply equally to the sigma–delta modulator.

However, it is at this point that an important fact must be realized. The optimum design of a noise shaper is one for which the noise transfer function (NTF), \(1 - H\), is minimum phase [6]. (A non-minimum-phase NTF can be shown to produce more noise, for a given shape, than a minimum-phase NTF.) Moreover, such an optimum noise shaper (or

![Figure 8](image1.png)

**Fig. 8.** Simple noise shaper. Error signal \(E\) is extracted around (possibly dithered) quantizer \(Q\) and fed back to input through noise shaping filter \(H\). Output \(Y\) differs from input \(X\) by noise-shaped error \((1 - H)E\).

![Figure 9](image2.png)

**Fig. 9.** Equivalence of noise shapers and sigma–delta modulators. (a) Structure of general single-stage sigma–delta modulator (possibly multibit). (b) Equivalent noise-shaper topology.
sigma–delta modulator) satisfies the Gerzon–Craven noise-shaping theorem, which states that the areas of the shaped NTF above and below the unshaped noise level must be equal on a decibel vertical axis and linear frequency axis. The implications for noise shaping are profound. If we wish to pull the noise floor down over part of the Nyquist band, it will inevitably be pushed up over another part of the band. Moreover, since this “equal areas” requirement holds on a logarithmic vertical axis, it follows that the total noise power will be increased by the shaping. There is thus always a total noise power gain, and the quantizer design must allow for the extra quantizer levels required to accommodate this noise gain. TPDF dithered on its own takes up only 2 LSBs of the quantizer’s dynamic range. However, when error feedback is used around the quantizer, many more LSBs must be devoted to handle the noise if quantizer overload and instability are to be prevented. We shall see shortly how such considerations affect noise-shaper design. As was mentioned earlier in connection with Fig. 6, it is evident that a 1-bit quantizer used in a noise shaper or sigma–delta modulator cannot be fully TPDF dithered. Even without the error feedback, the full dynamic range of the quantizer will be devoted to handling the dither. With the use of error feedback, and its concomitant noise gain, far less than TPDF dither is the maximum that can actually be applied. Indeed, one cannot even use RPDF dither without severe quantizer overload. Only a small amount of dither is usable, and so only a partial (but nevertheless most worthwhile) linearization of the quantizer is possible. The limitations of 1-bit quantization are addressed in more detail in [7].

A couple of concrete examples of noise-shaper design might be helpful here. We shall compare two extreme examples of high-resolution wide-bandwidth design. The first (called Example (c) in [7]) is that of a system sampling at four times the CD rate, namely, \( f_s = 176.4 \) kHz, but using only 8-bit words. It has a Nyquist bandwidth of 88.2 kHz, and it is desired to achieve a SNR of around 120 dB over the 0–20-kHz band. Not coincidentally this gives this multibit PCM system approximately the same dynamic range and bandwidth as the Super Audio CD\(^1\) system, which uses the 1-bit Direct Stream Digital (DSD)\(^1\) sigma–delta encoding format. However, it has only one-half the data rate of DSD. Fig. 10(a) shows the prototype noise-shaping filter specification for this design, in which 0 dB denotes the level of the unshaped noise floor.

The reasoning for the steps in its design are as follows. The four-times oversampling spreads the quantization noise power over four times the CD’s bandwidth, and so reduces its noise power spectral density (PSD) by 6.02 dB. Unshaped, but fully TPDF dithered, this 8-bit system would thus have a noise PSD lying at \(-51.2\) dB relative to full scale, denoted by dBFS (=45.2 dB for a TPDF-dithered 8-bit system + 6.02 dB for the oversampling). Allowing a generous 3-dB headroom reduction because of the noise-shaped dither, the noise PSD lies at \(-48.2\) dBFS. Now the Gerzon–Craven noise-shaping theorem tells us that since we want to pull the shaped noise floor down to \(-120\) dBFS over about one-quarter of the Nyquist band (say, to 22.05 kHz), the idealized shaping required is as shown in Fig. 10(a). The noise PSD below 20 kHz needs to be pulled down by 72 dB (=120 – 48 dB). The equal-areas theorem mandates that the PSD over the remaining three-quarters of the Nyquist band be consequently elevated by 24 dB (=72/3 dB), so that the total amount of shaping is 96 dB (=72 + 24 dB). The total noise power can now be computed to be \(-19\) dBFS. This can be seen as follows. The Nyquist-band TPDF-dithered quantization noise power is \(2\Delta^2/4\). The idealized noise shaping of Fig. 10(a) has a noise power gain of \((10^{-7.2} + 3 \times 10^{2.4})/4 = 188.39\) (=22.75 dB). The total system noise power is thus 188.39\(\Delta^2/4 = 47.10\Delta^2\). The 3-dB headroom allowance leaves \(\pm 90\) of the 256 levels of the 8-bit system available for the input signal. Hence, the full-scale sinewave power is \(90\Delta^2/2 = 4050\Delta^2\). So the full-scale SNR is 4050/47.10 = 85.99 = 19.34 dB.

These calculations assume an ideal rectangular-shaped noise curve, as shown in Fig. 10(a). This is, of course, not achievable with finite-order filters, and so these numbers must be taken merely as a useful approximate starting point for the actual design work. We have implemented this design using a twelfth-order recursive filter for \(H(z)\), and have also taken the opportunity to provide some crude

\(^{1}\)Super Audio CD and Direct Stream Digital (DSD) are trademarks of both Philips Electronics NV and Sony Electronics Inc.
(nonoptimized) psychoacoustic noise shaping,\(^2\) with PSD dips around 3 and 12 kHz. The actual result is shown in Fig. 10(b). Notice how the curve has the equal-areas property above and below the 0-dB unshaped noise PSD floor. The simulated performance of this design shows an SNR of 120.4 dB up to 20 kHz. The TPDF dither uses up about 70 of the 256 levels available in this 8-bit system, so that our 3-dB allowance for it was indeed conservative. (Incidentally, this is a good example of the loss of dynamic range due to the noise gain of the shaper. The error feedback has increased the 2 LSBs, which straight TPDF dithering would require, to 70 LSBs.) The system's frequency response would be flat to 80 kHz, although the noise floor would rise rapidly above 20 kHz. (This very rapid rise is probably undesirable, but this is just intended as an example of the sort of thing that can be achieved by noise shaping.) It is fully TPDF dithered, and so is completely artifact free—no distortion, noise modulation, or limit-cycle oscillations will occur.

The second example (called Lip7ZP and described in detail in [7]) is designed to use a 1-bit sigma–delta modulator whose order of 120 dBFS. These are the parameters of the Super Audio CD system. It uses a seventh-order noise-shaping filter and attempts to shape the noise floor in a psychoacoustically beneficial manner by suppressing the noise PSD in the 4- and 12-kHz regions, where the human ear is most sensitive. (This design has also not been optimized.) This is shown in Fig. 11. Note that this plot is on a logarithmic, not linear, frequency axis, so that the audio portion of the Nyquist band can be discerned. Consequently neither the equal-areas property of the noise-shaping curve nor the enormous amounts of high-frequency noise are apparent. As before, the 0-dB line represents the level of the unshaped noise floor. At high frequencies the noise rises to a level of +2.92 dB. This 1-bit system has an output swing of \(\pm \Delta/2\), but the full-scale level is set at \(\pm \Delta/4\) to prevent quantizer overload, in accord with the DSD specification. The integrated total wide-band noise power is 28 dB higher than that of the previous 8-bit example, as the 1-bit noise penalty far outweighs the benefit of the further 16 times increased oversampling ratio. As we have indicated before, a 1-bit system can be only partially dithered. This design can accept only 0.17\(\Delta\) peak to peak of RPDF dither without overloading the 1-bit quantizer, and so cannot be fully linearized. This is only about 1/69 the power of full TPDF dither. (To be able to accept full TPDF dither, simulations of the quantizer show that it would need eight levels, that is 3 bits.)

The noise floors of the two noise-shaper designs are compared in Fig. 12 on an equal full-scale signal-output basis, with the noise floor of a straight (unshaped), 4 times oversampled 20-bit PCM system. The frequency axis is linear up to 88.2 kHz. In this figure, 0 dB corresponds to the full-scale sine-wave signal output level, and the PSD curves have been normalized so that they display the noise power per hertz, shown up to 88.2 kHz. The horizontal line at 166.86 dBFS/Hz represents the noise floor of the unshaped 176.4-kHz 20-bit PCM system. Being unshaped, the dynamic range is constant over the whole band up to the Nyquist frequency of 88.2 kHz. Over the range from dc to 22.05 kHz, the total noise power is \(-123.4\) dB \((-166.86 + 10\log[22050])\) relative to full-scale signal level. The heavy curve is the 176.4-kHz 8-bit noise-shaped system. The light curve is the 2.8224-MHz 1-bit noise-shaped system. All three systems have comparable total noise powers up to 20 kHz, as intended. The 1-bit system’s elevated PSD above 70 kHz (all the way out to 1.4112 MHz) is responsible for its much higher total noise power. The correct registration of the curves in Fig. 12 requires taking into account the full-scale signal level of each system, in comparison with its shaped noise floor, on a power-per-hertz basis. Integrating with respect to frequency yields each system’s SNR over the corresponding band.

Noise shaping can thus be used in a variety of different

\(^2\)The design of psychoacoustically optimal noise shapers is beyond the scope of this survey paper. The interested reader is referred, for example, to [8]–[10] and the references contained therein.

---

**Fig. 11.** Second noise-shaper design example (Lip7ZP from [7]). 0 dB is level of unshaped noise floor. Seventh-order design intended to run at DSD sampling rate of 2.8224 MHz and use a 1-bit quantizer. Some psychoacoustic scalloping in audio band has been incorporated.

**Fig. 12.** Comparison (on a linear frequency scale up to 88.2 kHz) of noise PSDs of examples of Figs. 10 and 11, registered so that a full-scale signal corresponds to 0 dB. — noise shaper of Fig. 10; — noise shaper of Fig. 11; horizontal line — noise floor PSD of unshaped but TPDF-dithered 176.4-kHz 20-bit uniform PCM system. These noise-shaped systems trade greatly increased high-frequency noise for reduced wordlength. Total noise power (relative to full-scale output) of 1-bit Lip7ZP example (—) is 28 dB higher than that of 8-bit system (—–) and 126 dB higher than that of flat 20-bit system (whose data rate is only 25% greater than that of 1-bit system).
engineering compromises, trading off against one another—sampling rate, wordlength, in-band PSD, out-of-band PSD, total noise power, and the ability to implement full TPDF-dithered linearization. It should also be understood that low-bit systems, and especially 1-bit systems, impose particularly severe hardships when signal editing or postprocessing is undertaken. The entailed digital arithmetic inevitably results in a wordlength increase, but since the wordlength to which the signal must be requantized is not adequate to furnish the desired system SNR without noise shaping, each such operation necessitates further noise shaping and a consequent noise power increase. In low-bit systems, low-pass filtering may also be required at each processing stage in order to maintain modulator stability and loudspeaker longevity in the presence of the elevated high-frequency noise level, and this can result in a restricted audio passband. If each operation cannot be fully TPDF dithered, there will also be an accumulation of nonlinear artifacts. (In contrast, digital signal processing is very easy in PCM systems using a wordlength longer than that necessary to deliver the desired final SNR.) In many ways the most sensible approach is to use a wordlength such that no noise shaping is needed at any stage of the process in order to furnish the desired final baseband SNR. Failing this, the only final reduction to consumer format should be noise shaped.

4 SUMMARY

The beauty of the uniform PCM system of signal encoding is its extensibility to accommodate almost any desired signal bandwidth and SNR while allowing full linearizing TPDF dither to be used in all but the smallest wordlength cases. For its operation it rests upon an unassailable mathematical foundation governing signal sampling and quantizing. It is flexible enough to allow for fully dithered noise shaping when the need to reduce the wordlength forces one to tailor the shape of the noise PSD. If the wordlength reduction is extreme, only partial dithering, and hence only partial linearization of the quantization, may be possible.

5 ACKNOWLEDGMENT

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

6 REFERENCES


THE AUTHORS

S. P. Lipshitz

J. Vanderkooy
Stanley P. Lipshitz received a B.Sc. Hons. degree from the University of Natal, Durban, South Africa, in 1964 and an M.Sc. degree from the University of South Africa, Pretoria, in 1966, both in applied mathematics. He is a professor in the departments of Applied Mathematics and of Physics at the University of Waterloo in Waterloo, Ont., Canada, which he joined in 1970 after completing a Ph.D. degree in mathematics at the University of the Witwatersrand, Johannesburg. He is one of the founding members of the Audio Research Group at the University of Waterloo, which conducts research in many areas of audio and electroacoustics. His current research interests include the mathematical theory of dithered quantizers, noise shapers, and sigma–delta modulators (and their connection with stochastic resonance and chaos), physical acoustics, and active noise control.

Dr. Lipshitz is a fellow of the Audio Engineering Society, a recipient of its Silver Medal for his research contributions to digital audio, and of its Publication Award (jointly with John Vanderkooy and Robert Wannamaker) for a survey paper on quantization and dither in the Journal of the Audio Engineering Society. He has served as a governor of that Society, and was its president in 1988–1989. Other society memberships include the IEEE, the Acoustical Society of America, and the Canadian Acoustical Association. He has presented numerous technical papers, on a wide range of topics, at conferences both in North America and overseas. As vice president of the Kitchener–Waterloo Chamber Music Society, he records and broadcasts about fifty local classical music concerts annually, and has been doing so digitally for more than 20 years.

John Vanderkooy was born on January 1, 1943, in Maasland, The Netherlands. He emigrated to Canada with his family at an early age. He received all of his education in Canada, with a B. Eng. degree in engineering physics in 1963 and a Ph.D. degree in physics in 1967, both from McMaster University in Hamilton, Ontario. After a two-year postdoctoral appointment at the University of Cambridge, Cambridge, UK, he went to the University of Waterloo. For some years he followed his doctoral interests in high magnetic-field low-temperature physics of metals. His research interests since the late 1970s, however, have been mainly in audio and electroacoustics. He is a professor of physics at the University of Waterloo.

Dr. Vanderkooy is a fellow of the Audio Engineering Society, a recipient of its Silver Medal, and has contributed a variety of papers at conventions and to the Journal. Together with his colleague Stanley Lipshitz and a number of graduate and undergraduate students they form the Audio Research Group at the University of Waterloo. The group has produced a series of papers on dither in digital audio over the years. His current interests are digital audio, transducers, diffraction, and the measurement of acoustic impedance.