Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science

6.341: Discrete-Time Signal Processing

OpenCourseWare 2006

Lecture 6
Quantization and Oversampled Noise Shaping

Reading: Sections 4.8 - 4.9 in Oppenheim, Schafer & Buck (OSB).

While the title of the course is *Discrete-Time Signal Processing*, practical implementations for many of the discussed systems rely on discrete-value data representations as well. This business of representing discrte-time signals using finite precision will be the focus of the lecture.

The topic plays a key role in the problems of analog-to-digital and digital-to-analog conversion. So far we've explicitly studied C/D and D/C conversion, and within this context we have looked at how CT signals are processed by a DT system. In practice, however, a common scheme for operating in discrete-time on continuous-time signals consists of a sample-and-hold stage, followed by digitization of the resulting analog signal. This overall scheme is depicted in OSB Figure 4.45, and the effect of the sample-and-hold block is illustrated in OSB Figure 4.46(b).

Decomposing the system in OSB Figure 4.45 into a more easily analyzable conceptual representation, we arrive at OSB Figure 4.47, which contains the familiar C/D block. Typical behavior of the quantizer and coder blocks from the figure are illustrated in OSB Figure 4.48 for 3-bit uniform quantization. With a uniform quantization scheme, representing $x[n]$ using $B+1$ bits means that the signal is quantized to 2^{B+1} levels, so if $x[n]$ has maximum value X_M such that $-X_M \leq x[n] < +X_M$, these 2^{B+1} levels must cover the range $\pm X_M$. This implies that the spacing Δ between adjacent quantization levels is therefore

$$\Delta = X_M 2^{-B}.$$

The question then arises of how to analyze the error introduced by the process of quantization. Since a quantizer is generally a highly nonlinear system, we instead choose to use an additive noise model in its place, as depicted in OSB Figure 4.50. While the appropriateness of using such a model is perhaps best evaluated on a case-by-case basis, we'll see that this model does allow us to conveniently analyze the effects of signal quantization for a number of systems.

The output of a quantizer $\hat{x}[n]$ is represented in the additive noise model as

$$\hat{x}[n] = x[n] + e[n],$$

where $e[n]$ is the additive noise source. We'll now discuss the statistics of $e[n]$. As a starting point, we know that $e[n]$ cannot take on values greater than $\Delta/2$ or less than $-\Delta/2$. It may furthermore seem reasonable that the distribution of $e[n]$ is uniform for many signals encountered in practice. The full set of assumptions made by the additive noise model are discussed in Section 4.8.3 of OSB. Briefly, they are:

- $e[n]$ is a sample sequence of a stationary random process.
- $e[n]$ is uncorreletaed with $x[n]$.
- $e[n]$ is a white-noise process.
- The probability distribution of $e[n]$ is uniform over the range $-\Delta/2$ to $\Delta/2$, as illustrated in OSB Figure 4.52.

The expected value of $e[n]$ is therefore

$$\mathcal{E}\{e[n]\} = 0,$$

and its variance is

$$\sigma_e^2 = \frac{\Delta^2}{12}.$$

Because it is a white-noise process, its autocorrelation is

$$\phi_{ee}[m] = \frac{\Delta^2}{12}\delta[m],$$

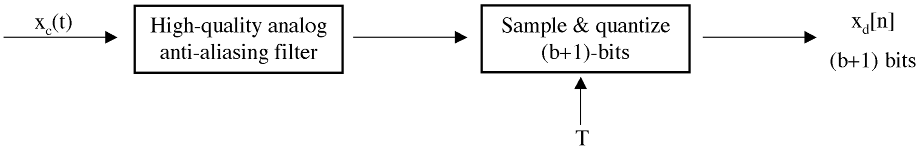
and its power spectral density is

$$\Phi_{ee}(e^{j\omega}) = \frac{\Delta^2}{12}.$$

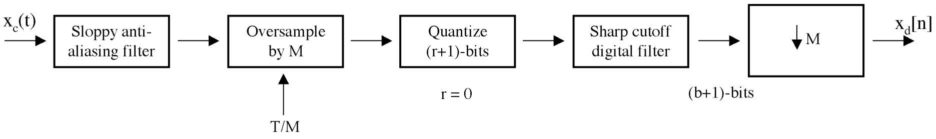
The total noise power is therefore

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\Delta^2}{12} d\omega = \phi_{ee}[0].$$

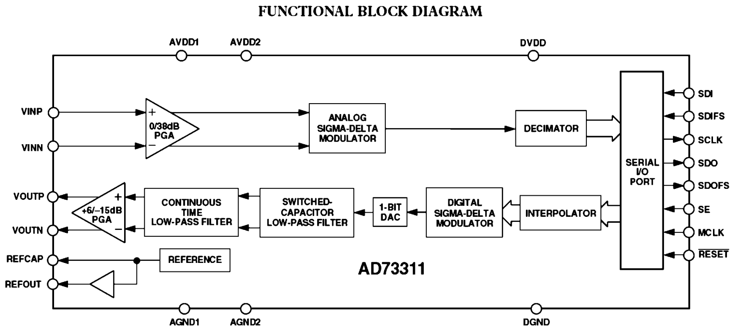
Now that we've introduced and analyzed this additive noise model, take a few seconds to think about how systems such as this one

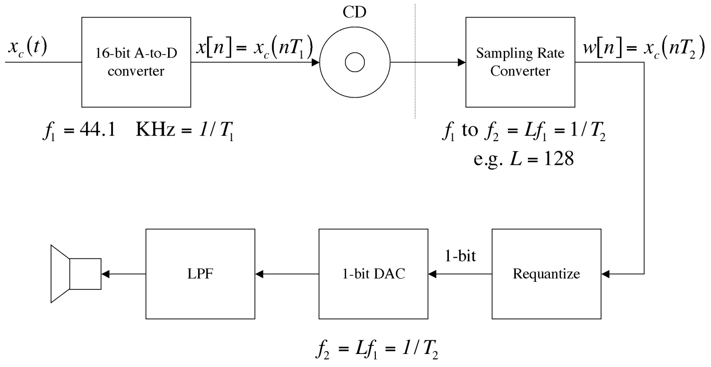

$x_c(t)$
High-quality analog anti-aliasing filter
Sample & quantize (b+1)-bits
T
$x_d[n]$
(b+1) bits

might be replaced by something like the following system, which typically uses less-expensive analog hardware:



As a further teaser, consider that a related configuration appears in a number of applications in practice, such as the AD73311 CODEC and many compact disk players:





A common theme in each of these systems is that sampling rate conversion is used to help mitigate quantization effects. But how exactly does the process of rate conversion affect quantization noise? We'll use the system depicted in OSB Figure 4.56 and its corresponding linear noise model (OSB Figure 4.57) to address this question. The total noise power at $\hat{x}[n]$ in OSB Figure 4.57 is, according to the linear noise model, $\phi_{ee}[0] = \frac{\Delta^2}{12}$. When this signal is lowpass filtered by the ideal LPF with $\omega_c = \pi/M$, the total noise power is divided by M, resulting in a total noise power of $\frac{\Delta^2}{12M}$ after lowpass filtering. As long as no aliasing occurs, the total noise power at the output of a compressor-by-M is the same as that of its input signal, so we know that the total noise power at the output of the system is

$$\text{Total noise power at output of system in OSB Figure 4.57} = \frac{\Delta^2}{12M}.$$

Furthermore, the resulting noise at the output of the system is still a white-noise process, and its PSD is $\Phi_{ee}(e^{j\omega}) = \frac{\Delta^2}{12M}$. This is demonstrated graphically in OSB Figures 4.59 and 4.60. Note that when M is doubled, the total noise power is halved. Since $\Delta = X_M 2^{-B}$,

$$\frac{\Delta^2}{12} = \frac{X_M^2 4^{-B}}{12}.$$

Therefore, doubling M corresponds to halving the total noise power, which implies

$$\frac{1}{2}\frac{\Delta^2}{12} = \frac{X_M^2 2^{-B}}{12} = \frac{X_M^2 4^{-(B+\frac{1}{2})}}{12}.$$

A doubling of M, then, has the same effect in terms of total noise power as adding an extra half-bit of precision to the quantizer.

We'll now discuss noise shaping, a method for further controlling quantization noise which has the ability to reduce in-band noise to a still greater degree. Let's first consider the system in OSB Figure 4.68 and its associated additive noise model (OSB Figure 4.69).

The system in OSB Figure 4.68 implements what might be considered a "first logical stab" at designing a linear system to reduce the effects of quantization noise. Looking at OSB Figure 4.69 for further insight, the system works by first obtaining the error signal $e[n]$, delaying it, and then using it to pre-compensate the input signal so that as long as $e[n]$ is changing very slowly, it is mostly cancelled at $y[n]$ by this pre-compensation. (Removing the delay block would cause the error at $y[n]$ to be exactly 0, but it would also prevent the system from being physically realizable.)

Exactly how slowly does $e[n]$ need to be changing so that it is singificantly reduced? The question can be answered by determining how the system responds at $y[n]$ to an input at $e[n]$. Using superposition, $\hat{y}[n] = 0$, and so $y[n] = e[n] - e[n-1]$. The deterministic transfer function from $e[n]$ to $y[n]$ is therefore

$$\frac{Y(e^{j\omega})}{E(e^{j\omega})} = 1 - e^{-j\omega},$$

with magnitude-squared response

$$\left|\frac{Y(e^{j\omega})}{E(e^{j\omega})}\right|^2 = \left(1 - e^{-j\omega}\right)\left(1 - e^{j\omega}\right) = 2 - 2\cos\omega = 4\sin^2(\omega/2).$$

The PSD of the noise at the system output then becomes

$$\Phi_{\hat{e}\hat{e}}(e^{j\omega}) = \frac{\Delta^2}{12} 4 \sin^2(\omega/2),$$

as depicted in OSB Figure 4.64. Lowpass filtering the resulting signal by an ideal LPF with $\omega_c = \pi/M$ and compressing by M eliminates out-of-band noise and therefore reduces the overall noise power. This result is shown in OSB Figure 4.65, and a number of implementations of this type of cascaded system are discussed in OSB.

Note that in practice, higher-order noise shaping systems are often designed to provide more detailed control of the resultant noise power spectrum. There are a number of practical issues associated with implementing such systems, and challenges in many of these cases relate to the additive noise model as an analytically convenient but nonetheless approximate representation.

In parting, consider OSB Table 4.2, which compares the order of a noise shaper p and oversampling factor M to the equivalent reduction in quantizer bits.