1.6. QUANTIZATION

In mathematics and digital signal processing, is the process of mapping input values from a large set to output values in a smaller set, often with a finite number of elements. Rounding and truncation are typical examples of quantization processes. Quantization is involved to some degree in nearly all digital signals processing, as the process of representing a signal in digital form ordinarily involves rounding. Quantization also forms the core of essentially all lossy compression algorithms.

The difference between an input value and its quantized value is referred to as quantization error. A device or algorithmic function that performs quantization is called a quantizer. An analog-to-digital converter is an example of a quantizer.

1.6.1. ANALOG-TO-DIGITAL CONVERTER

An analog-to-digital converter (ADC) can be modeled as two processes: sampling and quantization. Sampling converts a time-varying voltage signal into a discrete-time signal, a sequence of real numbers. Quantization replaces each real number with an approximation from a finite set of discrete values. Most commonly, these discrete values are represented as fixed-point words. Though any number of quantization levels is possible, common word-lengths are 8-bit (256 levels), 16-bit (65,536 levels) and 24-bit (16.8 million levels). Quantizing a sequence of numbers produces a sequence of quantization errors which is sometimes modeled as an additive random signal called quantization noise because of its stochastic behavior. The more levels a quantizer uses, the lower is its quantization noise power.

1.6.2. RATE–DISTORTION OPTIMIZATION

Rate–distortion optimized quantization is encountered in source coding for lossy data compression algorithms, where the purpose is to manage distortion within the limits of the bit rate supported by a communication channel or storage medium. The analysis of quantization in this context involves studying the amount of data that is used to represent the output of the quantizer, and studying the loss of precision that is introduced by the quantization process.

1.6.3. MID-RISER AND MID-TREAD UNIFORM QUANTIZERS

Most uniform quantizes for signed input data can be classified as being of one of two types: mid-riser and mid-tread. The terminology is based on what happens in the

region around the value 0, and uses the analogy of viewing the input-output function of the quantizer as a stairway. Mid-tread quantizers have a zero-valued reconstruction level (corresponding to a tread of a stairway), while mid-riser quantizers have a zero-valued classification threshold (corresponding to a riser of a stairway).

Mid-riser quantization involves truncation. The input-output formula for a mid-riser uniform quantizer is given by:

$Q(x)=\Delta.([\frac{x}{\Delta}]+\frac{1}{2})$

Where the classification rule is given by

$K=\frac{x}{\Delta}$

and the reconstruction rule is

$yk=\Delta.(k+\frac{1}{2})$

Note that mid-riser uniform quantizers do not have a zero output value – their minimum output magnitude is half the step size. In contrast, mid-tread quantizers do have a zero output level. For some applications, having a zero output signal representation may be a necessity.

In general, a mid-riser or mid-tread quantizer may not actually be a uniform quantizer – i.e., the size of the quantizer's classification intervals may not all be the same, or the spacing between its possible output values may not all be the same. The distinguishing characteristic of a mid-riser quantizer is that it has a classification threshold value that is exactly zero, and the distinguishing characteristic of a mid-tread quantizer is that is it has a reconstruction value that is exactly zero.

1.6.4. DEAD-ZONE QUANTIZERS

A dead-zone quantizer is a type of mid-tread quantizer with symmetric behavior around 0. The region around the zero output value of such a quantizer is referred to as the dead zone or dead band. The dead zone can sometimes serve the same purpose as a noise gate or squelch function. Especially for compression applications, the dead-zone may be given a different width than that for the other steps.

$K=\text{sgn}(x).\max(0,[\frac{|x|-\omega/2}{\Delta}+1]$

The general reconstruction rule for such a dead-zone quantizer is given by

$Y_k = \text{sgn}(k).(\frac{\omega}{2} + \Delta.(|k|-1+rk))$

In an ideal analog-to-digital converter, where the quantization error is uniformly distributed between −1/2 LSB and +1/2 LSB, and the signal has a uniform distribution covering all quantization levels, the Signal-to-quantization-noise ratio (SQNR) can be calculated from

$\text{SQNR}=20 \log_{10}(2^Q) \approx 6.02.Q \text{ dB}$

Where Q is the number of quantization bits.

Quantization noise model

$\text{SQNR} \approx 1.761 + 6.02.Q \text{ dB}$

For example, a 16-bit ADC has a maximum signal-to-quantization-noise ratio of 6.02 × 16 = 96.3 dB.

The most common test signals that fulfill this are full amplitude triangle waves and waves. For example, a 16-bit ADC has a maximum signal-to-quantization-noise ratio of 6.02 × 16 = 96.3 dB.

Here, the quantization noise is once again assumed to be uniformly distributed. In this case a 16-bit ADC has a maximum signal-to-noise ratio of 98.09 dB. The 1.761 difference in signal-to-noise only occurs due to the signal being a full-scale sine wave instead of a triangle or saw tooth.

For complex signals in high-resolution ADCs this is an accurate model. For low-resolution ADCs, low-level signals in high-resolution ADCs, and for simple waveforms the quantization noise is not uniformly distributed, making this model inaccurate. In these cases the quantization noise distribution is strongly affected by the exact amplitude of the signal.

The calculations are relative to full-scale input. For smaller signals, the relative quantization distortion can be very large. To circumvent this issue, analog companding can be used, but this can introduce distortion. Quantization noise for a 2-bit ADC operating at infinite sample rate. The difference between the blue and red signals in the upper graph is the quantization error, which is "added" to the quantized signal and is the source of noise.