Analogue to Digital Converter

Textbook pages 413-420

Dr D. Laurenson

13th November 2020

1 Introduction

Introduction

Material for this module is drawn from the course text, as well as "Principles of Digital Audio, 3rd edition" by Ken C. Pohlmann, published by McGraw-Hill, 1995, and "Principles of Oversampling A/D Conversion" by Max W. Hauser in the Journal of the Audio Engineering Society, Vol 39, No 1/2, Jan/Feb 1991

Many digital systems require some form of input from an analogue source such as a sensor. In order for the digital system to operate on such signals, not only does the signal need to be sampled, with all that is associated with that operation, but the signal must also be represented by a digital number. Representing an analogue quantity by a digital quantity is known as *quantisation*, as the analogue signal is represented by a restricted set of fixed values.

Quantisation introduces errors that are treated as noise, called quantisation noise.

The problem

If we wish to store, or process, an audio signal, such as a piece of music, or some speech, then we may either do this through an analogue medium, such as magnetic tape, or a vinyl disk, or by some digital representation stored in a computer, on a compact disk, or in a compressed audio format such as MP3. The former media, although they may offer an infinite range of possible amplitudes of the music or speech, and record it over continuous time, suffer from the effects of analogue noise, as well as storage problems such as magnetic leakage, and physical damage. The latter have the advantage of being more robust, and less susceptible to ageing effects, however they can only represent the audio signal by a set of samples in discrete time, and those samples are discretized, or quantised, into a finite set of binary patterns as shown below.[1]

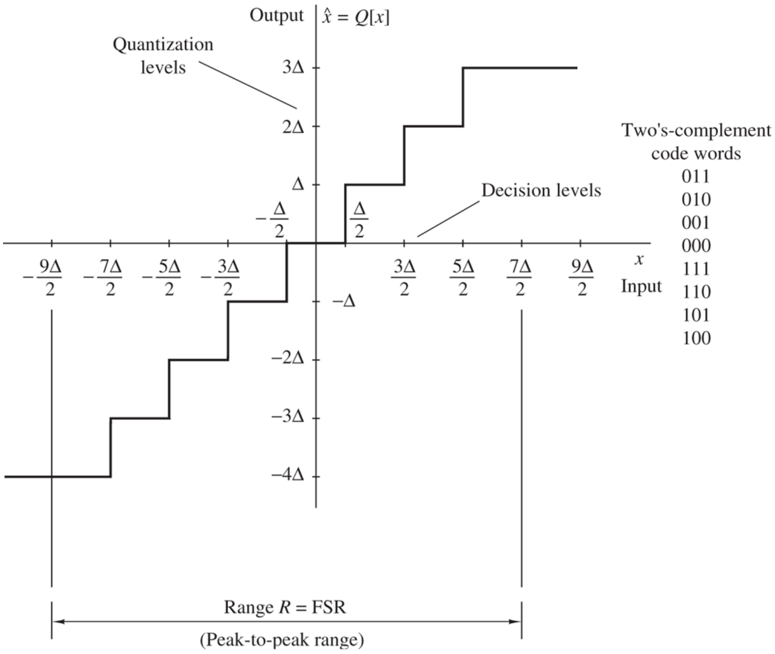


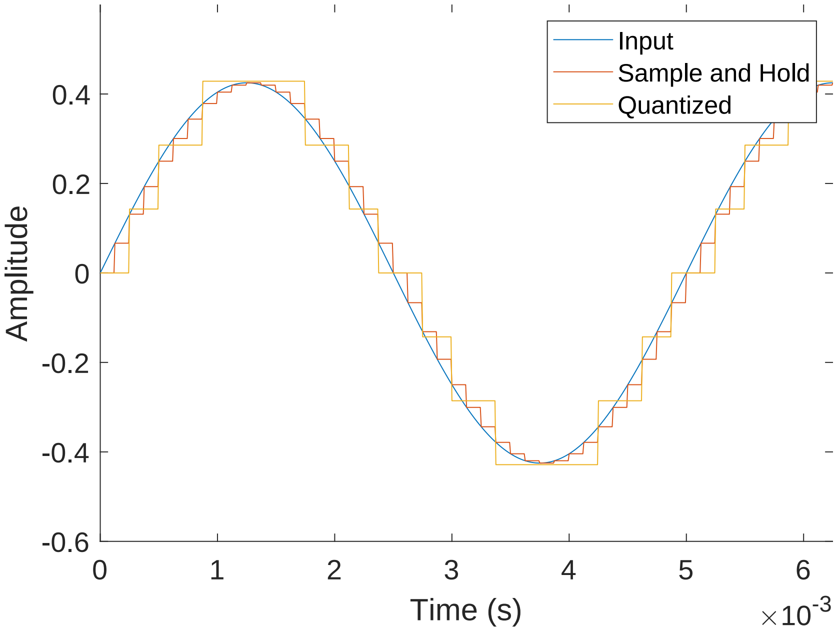
Figure 6.3.3 Example of a midtread quantizer.

The number of levels is determined by the number of binary digits (bits) used to represent the quantiser output. Because of this, the number of levels is a power of 2. The levels can either be placed equally across positive and negative voltages, giving rise to a midstep converter, or, as in the case above, one more level for negative values than positive values, a so called midtread converter. Midtread converters are common as 0 V corresponds to a quantisation level, rather than being halfway between two levels, even though the most negative quantisation level is largely unused.

If we are to represent the data by, for the sake of argument, 3 bits of binary data, the signal may be represented, for each sampling instant, by one of $2^3 = 8$ possible values. With so few quantisation levels, the effects of the difference between the actual analogue signal, and the quantised signal, are significant. Using a quantiser with more bits to represent the output will reduce the difference, but the effects still need to be considered.

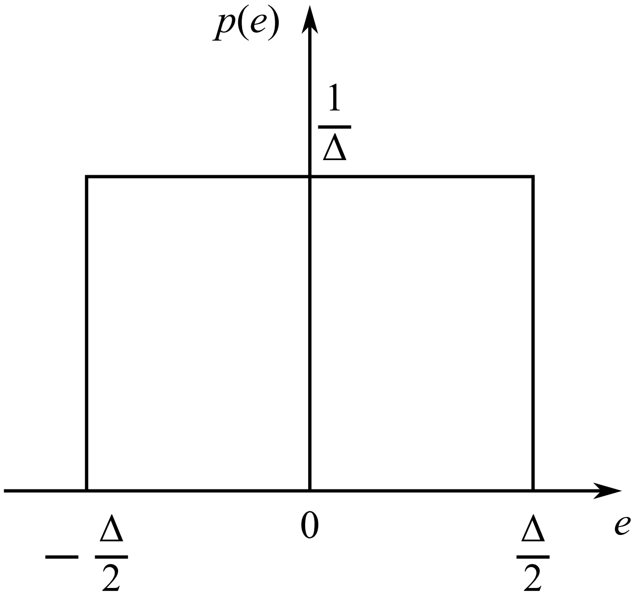
Consider a sine wave of 200 Hz of amplitude 0.85 V peak to peak, sampled at 8 kHz, with a 3 bit A/D converter with a quantiser range from +0.5 V to -0.625 V.

[1] Selected figures taken from “Digital Signal Processing, New International Edition/4th”, Proakis & Manolakis, ©Pearson Education Limited, 2014. ISBN: 978-1-29202-573-5



The figure shows the original signal, the output of the analogue sample and hold device, and then the digitised output. The sample and hold operates at the sampling rate of the converter, and the quantiser selects the nearest quantisation level to each of these samples. For this converter, the quantisation levels are separated by 0.141 V.

The error is often assumed to be uniformly distributed:



thus it can be described by a uniform probability density function (pdf).

The resolution of the converter, Δ, is given by

$$\Delta = \frac{R}{2^b} \tag{6.3.5}$$

where R is the range of the converter, and b the number of bits. (Note, the textbook incorrectly includes a +1 term in the power of the denominator).

Signal to Quantisation Noise Power (SQNR)

The SQNR is given by

$$\text{SQNR} = 10\log_{10}\frac{P_x}{P_n} \tag{6.3.6}$$

where

$$P_n = \int_{-\infty}^{\infty} e^2 p(e)de = \frac{\Delta^2}{12} \tag{6.3.7}$$

Substituting in for Δ,

$$P_n = \frac{R^2}{12(2^{2b})} = \frac{R^2}{12(4^b)}$$

For our example A/D converter, $R = 1.125$ and $b = 3$, thus $P_n = 1.65 \times 10^{-3} = -27.8$ dB.

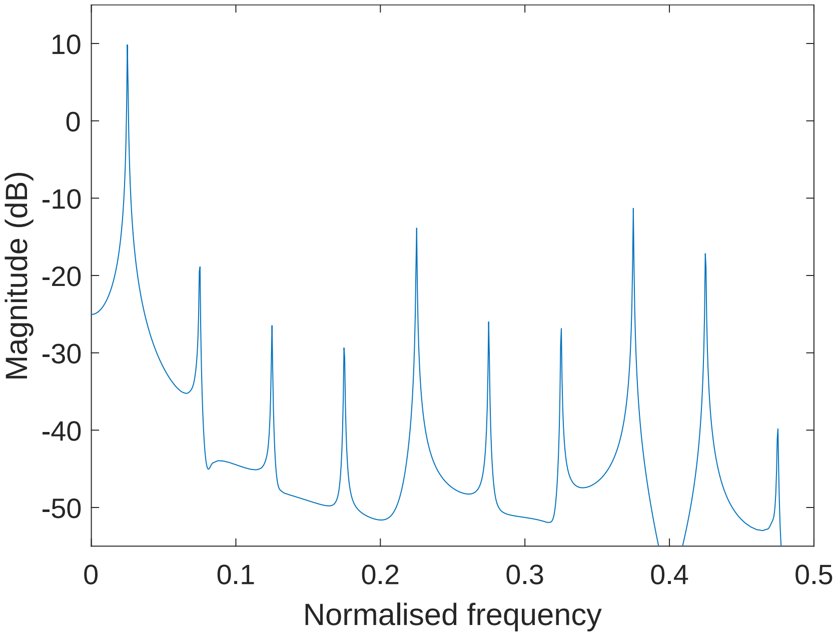
Then,

$$\begin{aligned}\text{SQNR} &= 10\log_{10}\frac{P_x 12(4^b)}{R^2}\\ &= 6.02b + 10.79 - 20\log_{10}\frac{R}{\sqrt{P_x}}\end{aligned}$$

Note that this corrects (6.3.8) caused by the error in the book in (6.3.5).

In our example, $R = 1.125$, $\sqrt{P_x} = 0.425/\sqrt{2}$ (rms of a sine wave of peak value 0.425), and $b = 3$. Thus SQNR = 17.4 dB.

However, the error is correlated with the signal being quantised. The correlated errors create distortion in the output signal with harmonics of the signal component being present. Consider the spectrum of the quantised signal (produced by a minimum variance spectral estimate):



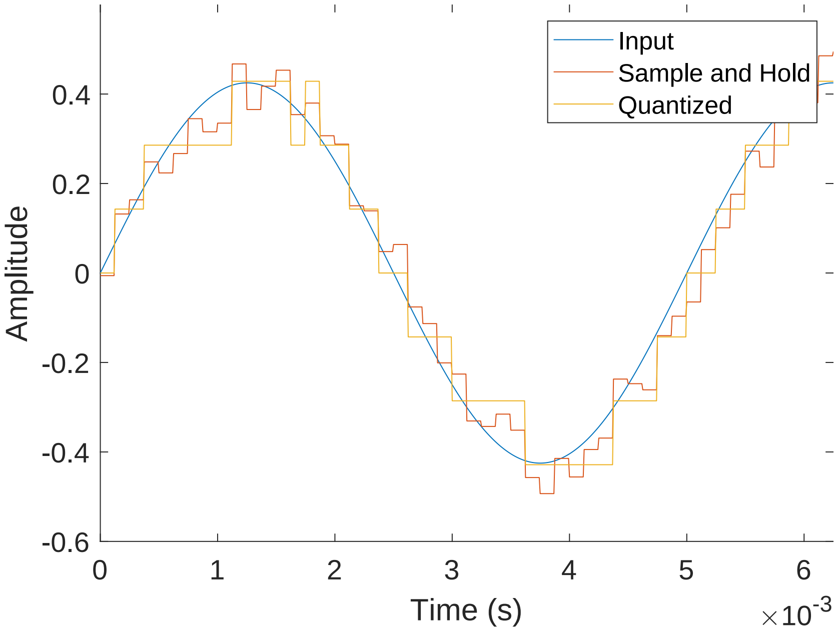
In the spectrum, clear peaks indicating the harmonics of the 200 Hz signal are evident at odd harmonics of 200 Hz, 600 Hz, 1 kHz, etc. These peaks are caused by the quantisation error being correlated to the signal that it is quantising.

Obviously, the quantisation error may be reduced by representing signals in more bits, and then the quantisation noise would become lower in comparison to the signal level, however practical considerations, such as expense and complexity of constructing a converter with a large number of bits, put practical limits on this approach. In order to overcome the problem presented by the finite number of levels for signal representation creating correlated noise, a technique of *adding* noise, called dithering, will be used to improve the quality of the output.

Dithering

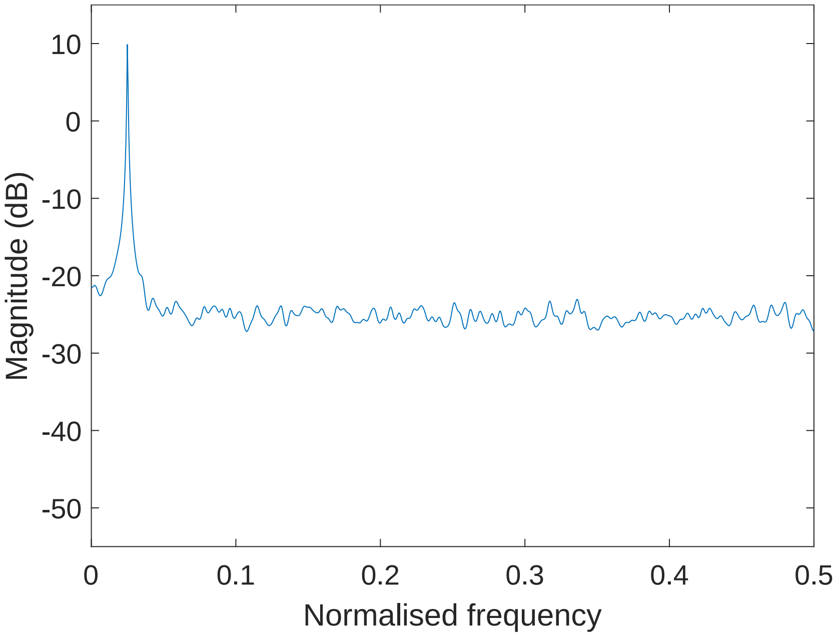
This technique may, at first, seem counterintuitive to all Engineering principles—adding noise to reduce the detrimental effects of an error present in the system, however it is a very effective approach. The input to the sample and hold, $x'(t)$, is given by $x'(t) = x(t) + d(t)$, where $x(t)$ is the analogue input, and $d(t)$ is a noise term that is white, and is uniformly distributed over the range $-\Delta/2 < d(t) < \Delta/2$.

The technique relies upon the principle that quantisation error is correlated to the signal, which is where the deleterious effects arise from. By adding noise to the signal, the quantisation error becomes correlated to the noisy signal, thus is no longer correlated to just the signal.



This results in the output of the quantiser not always selecting the nearest quantisation level to the input signal, thus the error is larger. However, the spectral characteristics are much improved.

Dithered spectrum



Noise and error power

For the example shown, the dither that was added was uniformly distributed noise in the range $(-\Delta/2, \Delta/2)$. We can determine the power of the combined quantization, Q, and dither, D, noise as follows:

$$\text{Power} = E\left[(D+Q)^2\right] = E\left[D^2 + 2DQ + Q^2\right]$$

As dither and quantization noise are independent, and zero mean:

$$\text{Power} = E\left[D^2\right] + 2E\left[D\right]E\left[Q\right] + E\left[Q^2\right] = E\left[D^2\right] + E\left[Q^2\right]$$

As both dither and quantization noise sources have equal power:

$$\text{Power} = 2E\left[D^2\right] = 2E\left[Q^2\right]$$

So, as the noise power is increased by a factor of 2, the SQNR is reduced by 3.01 dB, which is equivalent to half a bit of accuracy. For our example, this results in a noise power of -24.8 dB as shown in the figure above. Other types of noise sources may be used for dithering purposes, in particular noise with a triangular pdf is particularly suited to audio applications, and noise with a Gaussian pdf is particularly suited to generation using analogue electronics. (Gaussian noise may readily be produced using a diode as a noise source). Dither may also be applied in the digital domain for reducing the number of bits after processing, and is also used prior to digital to analogue conversion. In all cases, the additional noise reduces the SQNR of the converter output, however having the quantisation noise decorrelated is worth this cost.

Low bit analogue to digital conversion

One of the biggest problems in creating an accurate multiple bit analogue to digital converter is that of ensuring a uniform step size over the entire range of possible output values. If this uniformity is not maintained, then the analogue to digital converter will introduce distortion into the signal. A second problem is that digital to analogue converters with a large number of output bits are difficult to construct if a simple threshold comparison system is used.

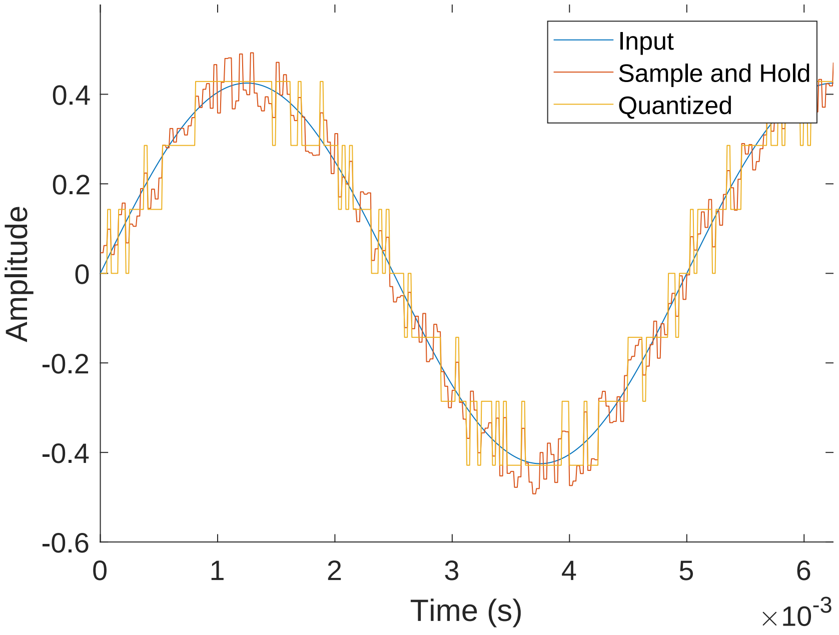
An alternative strategy to overcome some of these problems is to investigate the advantages of oversampling with low resolution analogue to digital converters, and then applying decimation to achieve the desired output resolution. As we shall see, this technique simplifies many of the components in the analogue to digital conversion process.

Oversampling

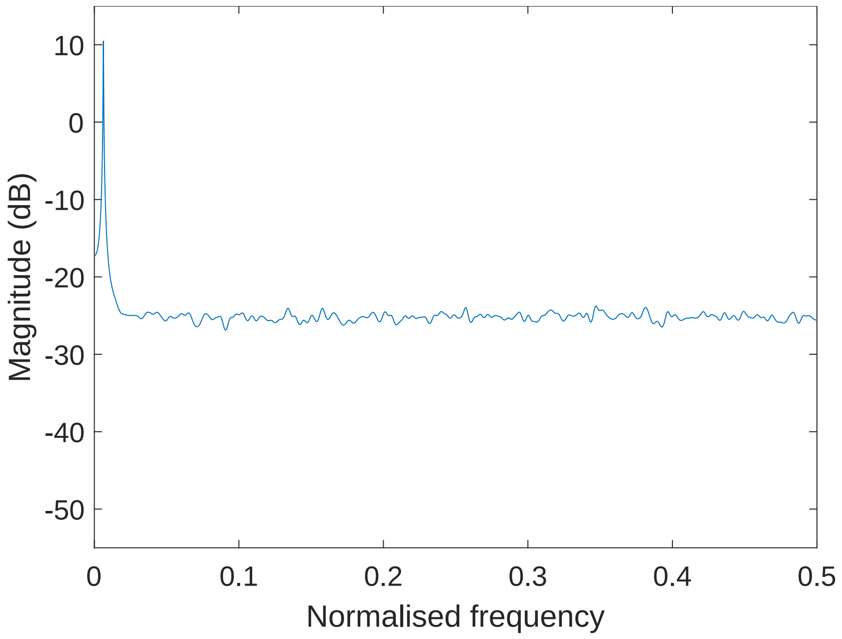
Oversampling is sampling the input at a higher frequency than we desire at the output. For example we may sample at a rate of OF_s, where O is a positive integer, and F_s is the desired sampling rate at the output. Note that this is not the same as upsampling, where the data is already in the sampled domain. Oversampling is an operation carried out on the continuous time signal.

The desired output is obtained after decimation by O. The first obvious advantage is that the antialias filter prior to the sampling does not need to be a high order approximation to a brick wall filter, but instead may be as simple as a second order filter which is approximately flat in the band of interest, $[0, F_s/2]$, and removes all components of frequencies higher than $OF_s/2$. It is once the signal is in the digital domain that a more accurate approximation of a brick wall filters is applied as part of the decimation process.

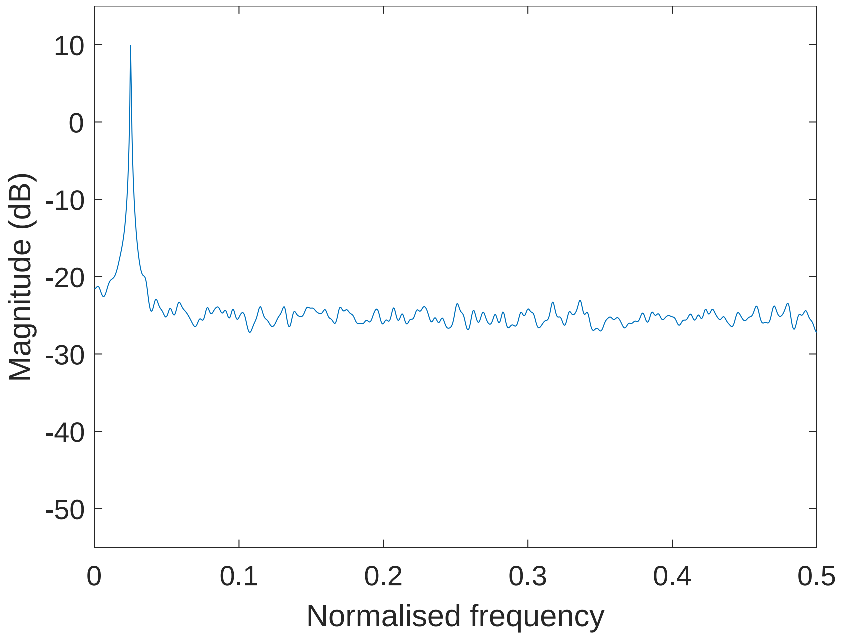
First, the signal is oversampled (O=4)



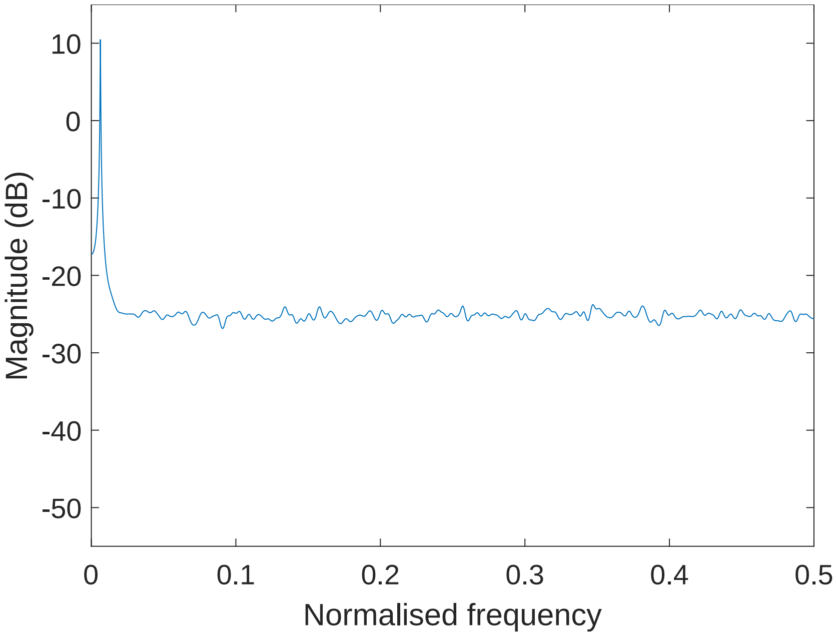
The spectrum of this signal is:



The non-oversampled spectrum is:

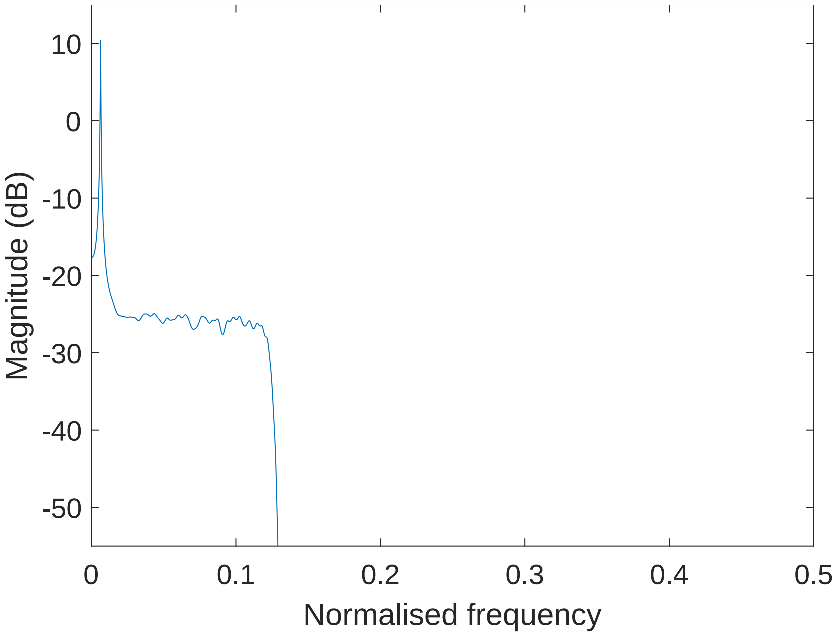


The spectrum of this signal is:



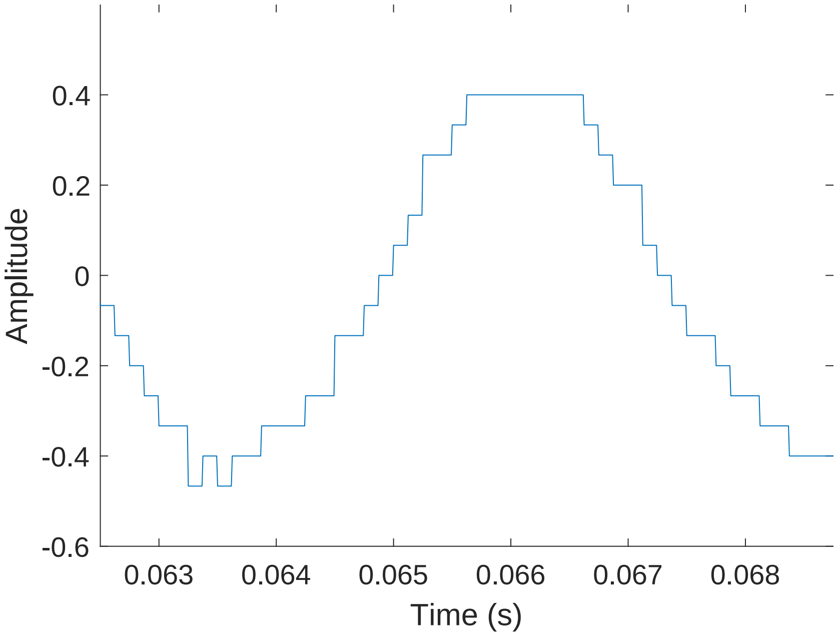
Oversampling does not alter the SQNR of the signal. The spectrum of the signal is constrained to lie in the range of $[-F_s/2, F_s/2]$, however the spectrum of the noise, as it is white, is spread over the range of $[-OF_s/2, OF_s/2]$.

The filter within the decimator alters this spectrum to become

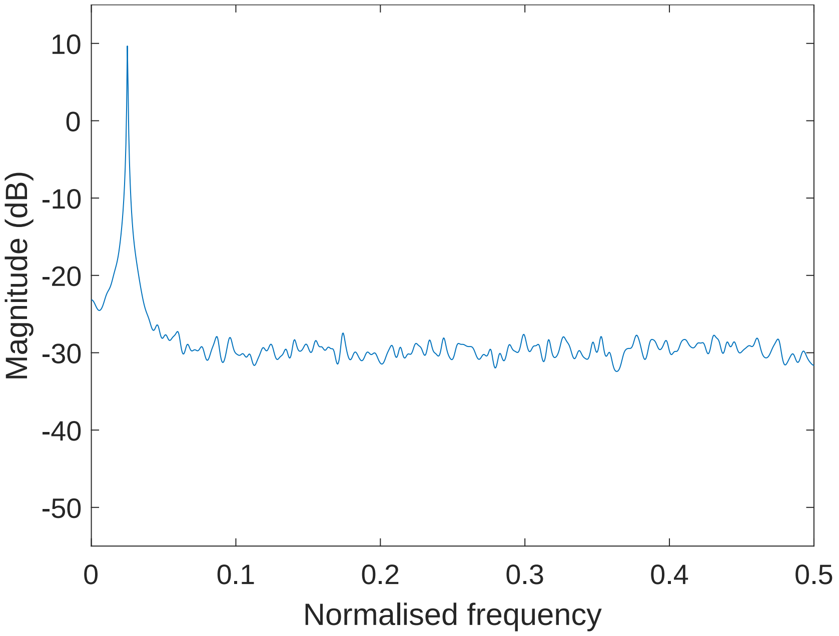


Now that quantisation and noise has been removed from the bands $[-OF_s/2, -F_s/2]$ and $[F_s/2, OF_s/2]$, the quantisation and noise power has reduced. The signal is unaffected by this as it lies in the band $[-F_s/2, F_s/2]$, thus the SQNR is increased by decimation.

The final downsampled output of the decimator is



The higher SQNR in the spectrum of the output can be observed from the lower noise level:



An intuitive time domain consideration reveals that if the decimator is considered as an averaging process, then the generation of a more accurate estimation of the signal amplitude can be obtained from limited resolution samples. Thus, the quantisation noise has been reduced at the output of the decimator.

The quantisation and noise power has been reduced by a factor of O, thus the SQNR of the output is given by:

$$\text{SQNR} = 6.02b + 10.79 - 20\log_{10}\frac{R}{\sqrt{P_x}} + 10\log_{10} O$$

Note that the output of the decimator will need $b + \left\lceil \frac{\log_{10} O}{0.602} \right\rceil$ bits to properly represent the higher resolution output. Without increasing the number of bits, the gains achieved by oversampling would be lost as the quantisation noise level is limited by Δ. To increase the number of bits, within the decimator, more bits are used to represent the filter multiplication outputs, $h(m)x(n-m)$ than used for the input $x(n)$.

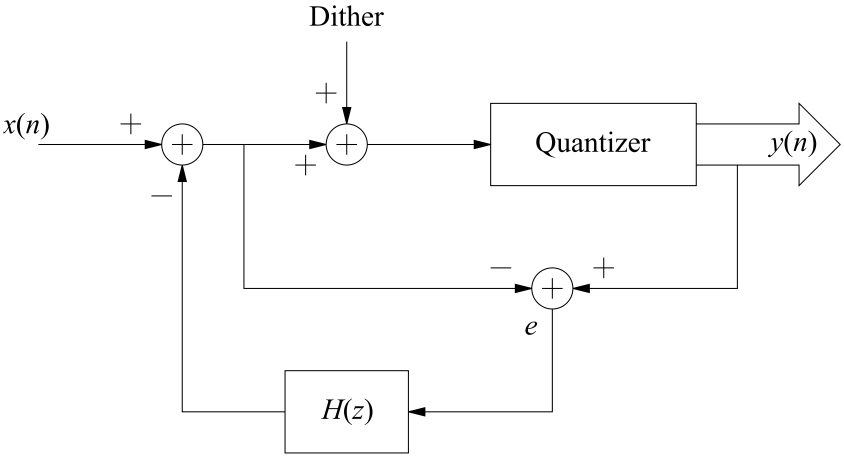
Example: If $O = 4$, $\log_{10} O/0.602 = 1$. One additional bit is required.

Clearly this technique could be used to obtain high resolution digitised signals from a low resolution quantiser operating at higher than the desired sampling rate. Using the above system, 16 bit resolution conversion may be obtained from a 12 bit quantiser using 512 times oversampling. This calculation assumes that 1 bit of dither noise is added to the system reducing the overall accuracy by half of one bit.

The principle can be extended to the point where a simple 1 bit quantiser is used, which results in a high quality uniform converter. The implication is that to obtain 16 bit resolution from a 1 bit converter, then 2.15×10^9 times oversampling must be employed, which is quite impractical. However, this is not the end of the story, as better gains may be had from oversampling if we alter the noise spectrum so that more of its power per unit bandwidth lies in the frequencies between $F_s/2$ and $F_s(D-1/2)$ than in the bands of interest. This may be achieved by a technique called noise shaping.

Noise shaping

By placing the dither noise and quantiser within a feedback loop, it is possible to arrange for the quantisation noise to be subject to a different transfer function to the signal input.

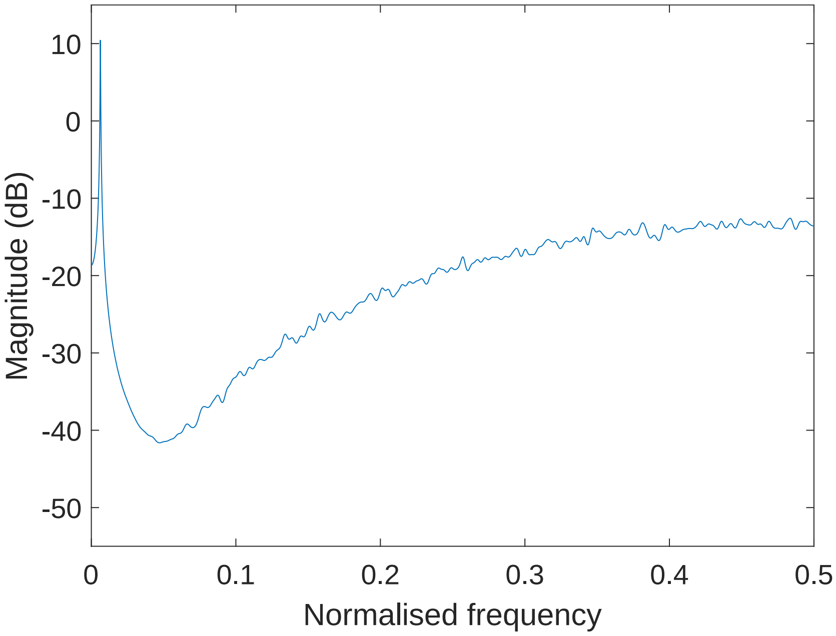


$$Y(z) = X(z) + Q(z)(1 - H(z))$$

where $Q(z)$ represents dither plus quantisation noise.

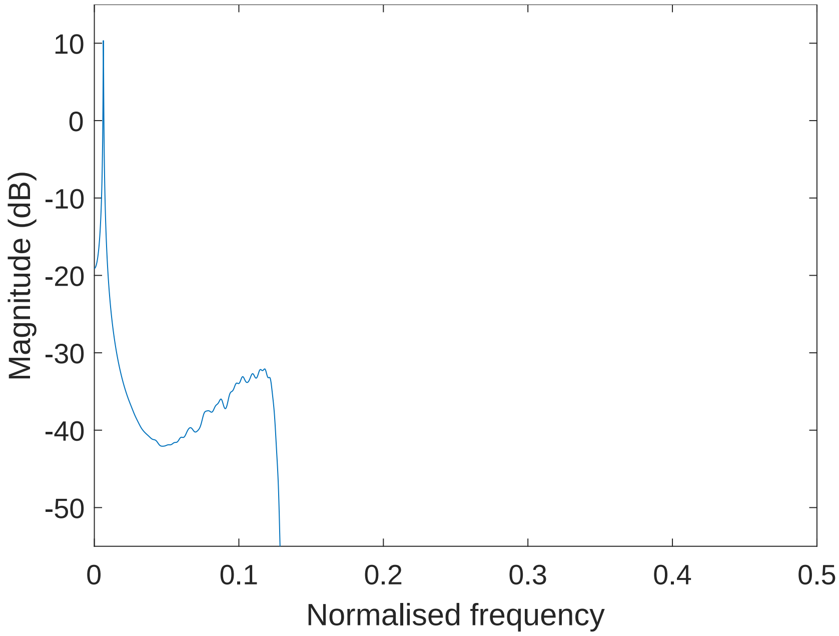
$H(z)$ can be chosen such that $1 - H(z)$ is a high-pass filter. This will reduce the noise in the region of the desired input signal, at the expense of increasing the noise around the oversampled sampling frequency. Selecting $H(z) = z^{-1}$ results in a first order response, and $H(z) = 2z^{-1} - z^{-2}$ a second order response.

The spectrum of the quantiser output when $H(z) = 2z^{-1} - z^{-2}$ is



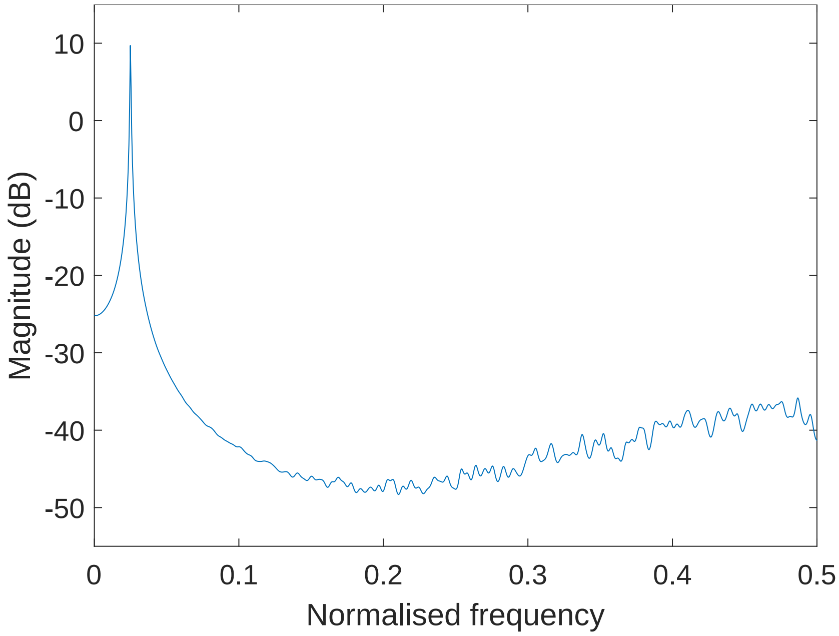
It is clear, when compared with the case without noise shaping, that the noise in the bandwidth of the signal is significantly reduced. In this example, $O = 4$.

The filter within the decimator removes more noise:



Clearly, noise above $F_s/2$ has been filtered out, leaving frequency content only in the range $[-F_s/2, F_s/2]$. As the signal in the region $[F_s/2, OF_s/2]$ was noise, the noise power has been reduced. This results in a higher SQNR.

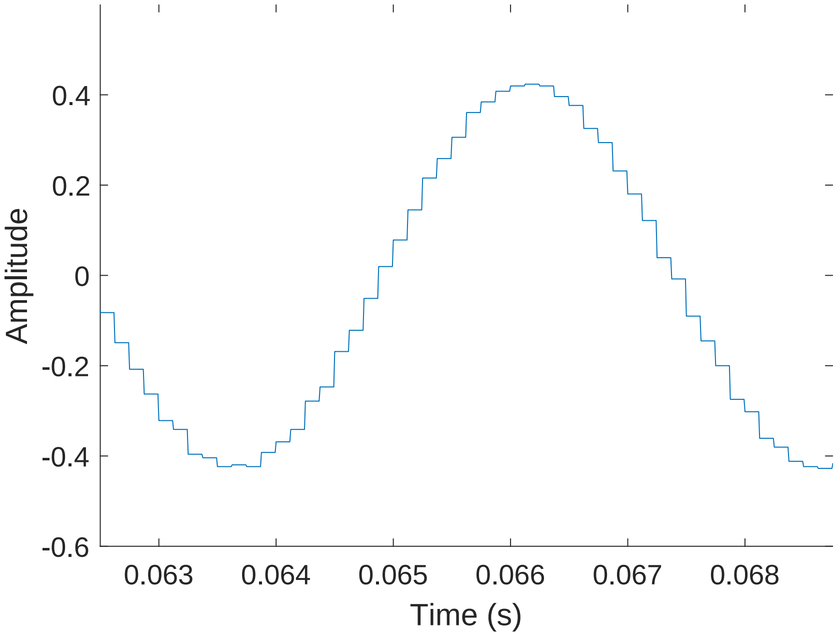
The spectrum of the decimated signal is



As the noise above $F_s/2$ is removed, when the signal is decimated, only residual frequency content above $F_s/2$ is aliased down. Note that the profile of the noise shaping means that the noise content near to half the sampling frequency is higher than that near to d.c.

In practice, the efficient decimation structure shown in the Multirate DSP module is used, which means that the decimation and filtering are combined into a single process, and not separated out as shown in the two figures above.

The final output is



In order to successfully represent this as sampled values, without losing the benefits of the noise shaping, additional bits are required. These bits must be included in the decimation filter when multiplying by the coefficients, and summing the final output.

SQNR

The SQNR is given by:

$$\begin{aligned} \text{SQNR} &= 6.02b + 10.79 - 20\log_{10}\frac{R}{\sqrt{P_x}} \\ &\quad +10(2L+1)\log_{10} O \end{aligned}$$

where L is the filter order. In this case the output of the decimator will need $b + \left\lceil \frac{(2L+1)\log_{10} O}{0.602} \right\rceil$ bits to properly represent the higher resolution output.

If $O = 4^l$, then the output resolution is approximately $b + (2L+1)l$.

Examples:

- 16-bit resolution can be obtained by oversampling a 12-bit quantiser, with a 1-bit dither, with first order noise shaping, by a factor of $O = \left\lceil 4^{(16-(12-0.5))/3} \right\rceil = 8$.
- 16-bit resolution can be obtained by oversampling a 1-bit quantiser, with a 1-bit dither, and second order noise shaping, by a factor of $O = \left\lceil 4^{(16-(1-0.5))/5} \right\rceil = 74$.

Because of the approximation and non-ideal filters, it is common to round up to the next power of 2, i.e. 16 times oversampling and 128 times oversampling.