Proceedings of The Institute of Acoustics

QUANTISATION EFFECTS IN HIGH QUALITY DIGITAL FILTERING

J.A.S. Angus

Dept. of Electronics, University of York, Heslington, York

INTRODUCTION

The aim of high quality digital filtering is to add the minimum degradation to the signal during processing. In practice such processing can degrade the signal because practical digital filtering must be done with finite precision arithmetic. This paper describes optimum structures for digital filtering which cause minimum degradation to the signal. It first considers the type of degradations which occur and then goes on to consider the effects and implications of these degradations on a signal processing system.

DEGRADATIONS DUE TO FINITE PRECISION ARITHMETIC

The theory used to describe and calculate digital filters assumes that the numbers are represented with infinite precision. If this is the case then the theory shows that the many different structures which can realise a given filtering function are equivalent. However, in practice digital filters must be realised with finite wordlength arithmetic and this imposes a limit on the precision of both the necessary multipliers and the calculations required to realise a given filtering function. There are two major effects of this finite precision.

1) Inaccuracy in the transfer function: Digital filters achieve their filtering function by adding and subtracting weighted, through multiplication by a constant, and delayed versions of the signal to be filtered. Because these operations have to be carried out with finite precision the actual transfer function achieved will be different to the desired one. In this respect the different filter structures are not equivalent with some being superior to others for a given application.

2) Increased noise due to roundoff: The addition of two N-bit numbers produces a result with a potential range of N+1 bits. In general the addition of K N-bit numbers results in an output with a potential range of $N+\log_2 K$ bits. The multiplication of two N bit numbers results in a product which requires 2N bits to retain the resulting precision. Therefore, if full precision is to be retained, a digital filtering system must have an expanding word length. This is generally inconvenient so the wordlength is reduced to its normal size in the structure when necessary. This may be achieved either by ignoring the unwanted least significant bits (truncation) or by adding one to the desired result if the unwanted bits are greater than half an LSB of the wanted bits (rounding). Both of these procedures add noise to the

signal of the same equivalent power [1]. The only difference is that truncation is easier to implement but has a dc bias when compared with rounding. Again the choice of filtering structure can have a significant effect on this noise.

Clearly both of the above effects are undesirable and for high quality audio signal processing we want to minimise the perceived effect of these degradations.

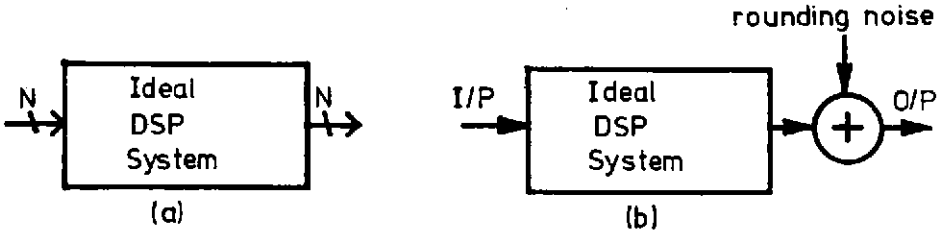

Figure 1

THE EFFECT OF COMPUTATION NOISE ON THE SIGNAL PROCESSING SYSTEM

Figure 1(a) shows an ideal digital filtering system with an input and output wordlength of N bits and infinite precision arithmetic inside (impractical even for VLSI). Figure 1(b) shows how we can model the reduction from infinite precision to the N output bits required as an injection of noise into output of the system. The level of noise injected depends on the output wordlength and is less for larger wordlengths. If we assume that the input to the filtering system has come from an A/D convertor (which would have added the same amount of noise to the signal as quantisation of the output of the system) [1] then we can see that the ideal signal processing system will degrade the S/N ratio of the signal by 3db. The implication of this is that if one anticipates passing recorded signals through digital signal processing systems several times then one must store the intermediate processed results at a much higher precision than the required final output if one is to avoid a build-up of noise through the processing. Also if the input precision is the same as the output precision, then the ideal system will still result in a S/N ratio loss of 3db. This implies that ideally, for professional work, the input wordlength (recording wordlength) should be greater than the desired final one.

IDEAL FILTERING STRUCTURES

We can realise an ideal signal processing system with practical finite wordlength hardware by the correct choice of digital filter structure. Previous work in this field was done in the sixties when most digital hardware was comparatively expensive and so the

structures proposed usually considered the adders to have the same precision as the multipliers. However current technology is less expensive and this assumption is no longer valid. For the purposes of this paper, one makes the following assumptions.

1) Storage required for delaying the signal is relatively inexpensive.

2) Adder/subtractor logic is more expensive than memory but is still comparatively cheap.

3) Multiplication is approximately N times more expensive than addition where N is the wordlength of the multiplier.

4) Data movement is comparatively expensive.

These assumptions are based on the approximate silicon area taken up by these functions. The net result is that an ideal digital filtering structure from the cost point of view should minimise the multiplications even at the expense of more additions or delays. It also means that a double precision addition to accumulate the full-width product is perfectly feasible, and is in fact provided by a number of signal processing products currently available (e.g multiplier-accumulators).

Higher order digital filtering functions may be realised in several ways.

1) Direct implementation in one filter.

2) Cascade implementation of (usually) 2nd order sections.

3) Parallel connection of 2nd order sections.

4) Coupled structures (e.g. lattice and wave filters).

The first three of these structures are shown in figs. 2-4 with the points at which rounding or truncation must take place being indicated. Coupled structures are not shown because, although they have a low sensitivity of transfer function variation due to coefficient precision, they require more truncation or rounding nodes and thus exhibit a poorer noise performance.

If we examine the three higher order structures we can see that the cascade connection requires more rounding nodes than the other two and so will exhibit an increased noise. The parallel and direct forms at first appear to have the same number of rounding nodes. However, this is not the case as there will be some form of truncation or roundoff in the 2nd order section used in the parallel connection and so the parallel connection will exhibit a lower value of quantisation noise over a cascade connection but will have a higher level of roundoff noise when compared with a direct connection.

Thus it would seem that in order to minimise the amount of quantization noise in the output of a filter, one should use a direct form of structure. However the direct form is known to exhibit an unacceptable sensitivity of the transfer-function to

QUANTISATION EFFECTS IN HIGH QUALITY DIGITAL FILTERING

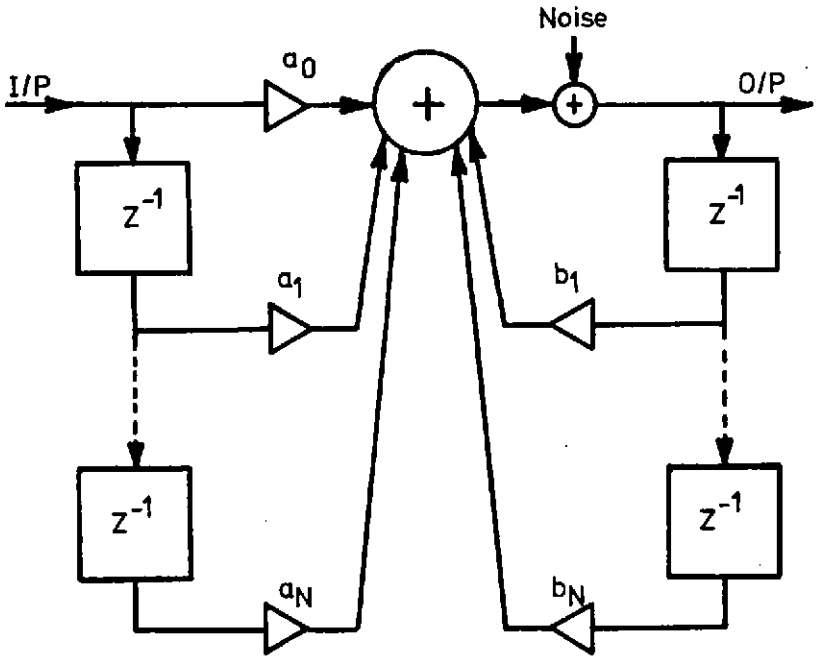

Figure 2 Direct Form

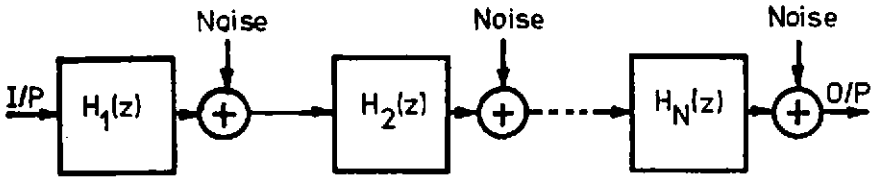

Figure 3 Cascade Form

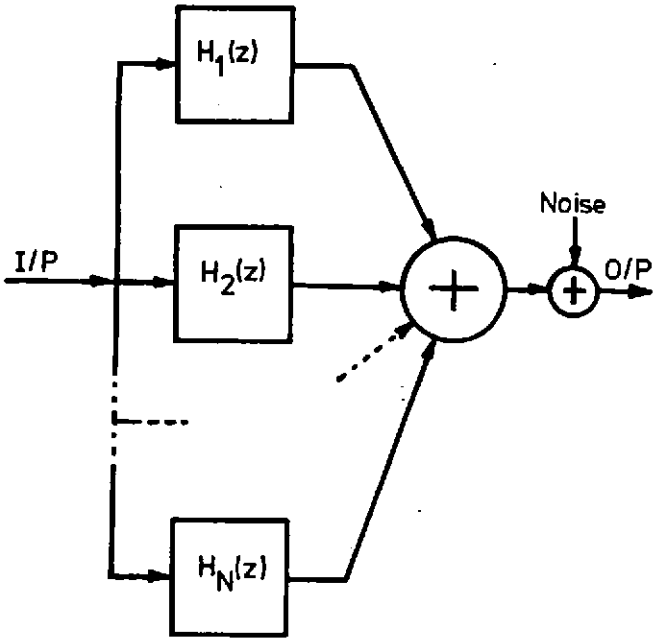

Figure 4 Parallel Form

the coefficient precision for higher order filters [2]. Therefore it would seem that the parallel structure is the best compromise for higher order filters. The parallel form does have a disadvantage over the cascade form in that the sensitivity of stop-band performance to the coefficient precision is higher than that of the cascade structure [3]. This may cause problems for filters which are designed to remove hum or low frequency noise from the signal.

IDEAL 2ND ORDER SECTIONS

The ideal way of implementing the second order sections is by using the direct form of structure shown in fig. 2. This structure has the advantage that the multiplier outputs can be accumulated in double precision. Thus the only noise source is the rounding or truncation of the final results to the output word length. Also, because all the additions occur at one node, scaling to avoid overflow is easier. In fact the accumulation register can be made large enough to handle the maximum result of any given calculation. Logic on the rounding or truncation stage can be then be used to implement a saturation characteristic. This allows for a maximum dynamic range from the filter.

QUANTISATION EFFECTS IN HIGH QUALITY DIGITAL FILTERING

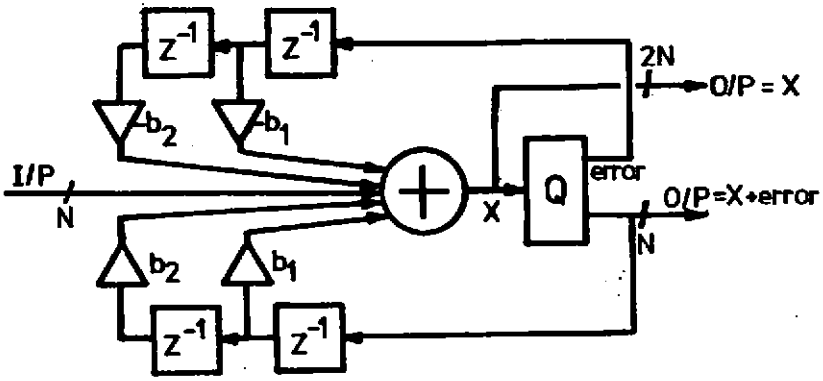

Figure 7 Noise Shaping

In the midrange (> 500Hz) 16 bit coefficients provide adequate accuracy. However, for lower frequencies, especially if a high Q is required, this may not be the case. In these filters the coefficients are near to 2 or 1, so by implementing them as $(2-b_1)$ and $(1-b_2)$, where b_1 and b_2 are two small positive numbers, and by using some judicious scaling of b_1 and b_2 it is possible to implement the direct form with sufficient accuracy even for low frequency filters [4]. The same approach can also be applied to the zeros of the filter. The penalty is an increase in the required number of additions (up to 4 extra) and some shifting operations (fig. 8).

THE EFFECT OF DITHER

All the above discussion has assumed that the error due to quantization can be modelled as white noise which is uncorrelated with the signal. This is not the case when low level signals are considered [5] because in that situation the quantisation noise is correlated with the signal and is therefore subjectively more noticeable. The addition of dither, of about one LSB in magnitude, is known to be beneficial as it can force the quantisation noise to be uncorrelated with the signal [6]. The penalty of using dither is that the total noise power in the signal is increased.

As we have already seen one of the advantages of the parallel connection is that truncation or roundoff need only be carried out at the output of the final adder. Thus it is very easy to add dither at this stage in order to ensure that the roundoff noise is white and uncorrelated with the signal. If one adds dither, with a uniform probability density function and a range of one LSB, the net loss in S/N in the system due to the rounding and dither is 4.8db, assuming input and output wordlengths are equal. It is

QUANTISATION EFFECTS IN HIGH QUALITY DIGITAL FILTERING

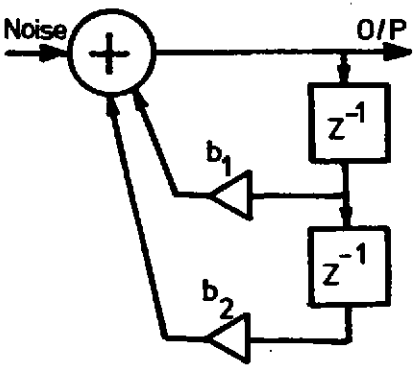

Figure 5

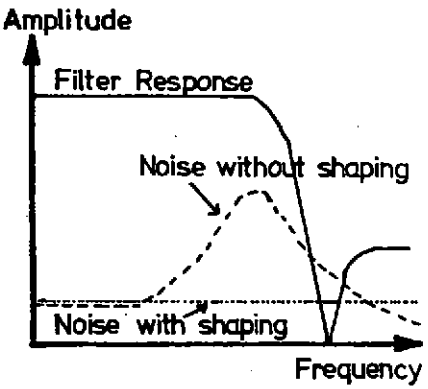

Figure 6

The structure does have some problems however. The first is that the round-off noise is not flat with frequency. If one models the noise introduced by the rounding as additive white noise. Then one can see that the noise is effectively presented to the input of the pole section of the filter (fig. 5). Thus the quantisation noise is shaped by the frequency response of pole portion of the filter. Therefore the effect of a high Q filter is to cause a large peak in the noise output by the filter. This is especialy serious where poles and zeros are close to each other (fig. 6) as an apparently flat response can exhibit severe noise peaking at the band edge due to the presence of a high Q pole-pair. One way of reducing this effect is to use noise shaping (fig. 7). This works by feeding back the quantization error to the accumulator via another set of coefficients. The net result is to cause the noise to be shaped by a filter whose transfer function is given by the combined effect of the two filters in the feedback path. If the coefficients of the noise filter equals those of the pole part of the filter then the transfer function seen by the noise is flat and so the noise output of the filter is flat. This requires two more delays and multiplies but results in a significant improvement in the noise output of a filter. Even more interestingly if one uses the double precision output of the accumulator of the second order section, when noise shaping is applied, one finds that the noise introduced by the truncation of the feedback path has been cancelled out (fig. 7). This means that one can implement a parallel structure, using double precision addition, which has the same noise performance as the ideal filter!

The other problem with the direct realization is the sensitivity of the transfer function to the coefficient's precision. This is particularly acute when low frequency high Q filters are required.

QUANTISATION EFFECTS IN HIGH QUALITY DIGITAL FILTERING

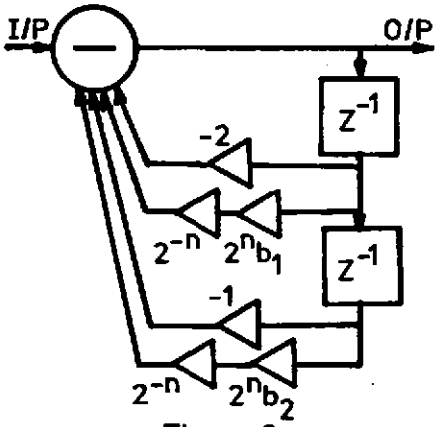

Figure 8

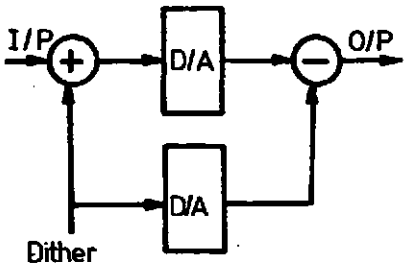

Figure 9

possible to convert the dither into analogue form and subtract it from the signal (fig. 9) but this would only result in a 1.8db improvement in S/N.

CONCLUSIONS

For high quality digital filtering one must minimize the sources of computation noise. In this respect a parallel connection of 2nd order sections is better than a cascade connection. By using a parallel connection of 2nd order sections with double precision addition, one can realize an audio signal processing system which adds the minimum computation noise to the system while being efficient to implement on available digital signal processing hardware.

REFERENCES

[1] L.R. Rabiner and B. Gold, "Theory and application of digital signal processing", Chapter 5, 295-309, pub. 1975 by Prentice Hall.

[2] A.V. Oppenheim and C.J. Weinstein, "Effects of finite register length in digital filtering and the fast fourier transform", Proc. IEEE, vol. 60, 957-976, August 1972.

[3] E. Avenhaus, "On the design of digital filters with coefficients of limited word length", IEEE Trans. Audio Electroacoust., vol. AU-20, no.3, 206-212, August 1972.

[4] R.C. Agarwal and C.S. Burrus, "New recursive digital filter structures having very low sensitivity and roundoff noise", IEEE Trans. Ccts. and Syst., vol. CAS-22, no.12, 921-927, December 1975.

[5] W.R. Bennett, "Spectra of quantized signals", BSTJ., vol. 27, 446-472, 1948.

[6] J. Vanderkooy and S.P. Lipshitz, "Resolution Below the least significant bit in digital systems with dither", J. Audio Eng. Soc., vol. 32, no.3, 106-112, March 1984.