Remarks on SNR, Floating point FX correlator

B.G. Clark

July, 1986

Suppose the input to an FFT butterfly is continuous and is so sized that the output has RMS of unity. Looking in the table of the error function, we find that with five-bit nonphantom floating point, the value of the least significant bit has the values given for the percentages of time given:

Value range	LSB	% of time	Trunc. var	Trunc. Noise
.125 to .25	.016	9.8 %	.00002	.00000
.25 to .50	.031	18.6	.00008	.00002
.5 to 1.0	.062	29.6	.00033	.00010
1.0 to 2.0	.125	27.2	.00130	.00035
2.0 to 4.0	.25	4.55	.00521	.00024
4.0 to 8.0	.50	0.0063	.0208	.00000

With an LSB of 0.016, the maximum quantization error (with rounding) is 0.008, and the contribution to the variance is one-third of the square ofthat. This is tabulated in the fourth column above. The fifth column above is the product of the third and fourth. The sum of the fifth column is 0.00071. This is approximately the loss in signal-to-noise ratio in each requantization.

The factor of two that sometimes appears in such calculations is not present because there is a cancelling factor of two from sine and cosine components. The assumption has been implicitly made that the choice of exponent has been made on the magnitude of the complex number, rather than on the largest of (real,complex), which will cause the loss of SNR to be slightly overestimated. The assumption has been made that all values of the signal within the levels set by the LSB are equally probable; this is not the case for the very large deviations (greater than two standard deviations). This will cause the loss of SNR to be slightly underestimated. The calculation has been done for rounding, rather than for truncation. Rounding can be difficult for hardware implementation because the process of rounding can cause the exponent to change by one. The SNR loss for truncation is not a factor of four greater, as one might naively assume following the above analysis, because it is equivalent to a gain different from unity as well as the addition of noise. Without doing the calculation, I suspect that the loss of SNR is about the same as for rounding, but that the gain effects may be serious, in that they may introduce amplitude closure errors.

Repeating the calculation for other values of the output variance give essentially the same loss of SNR, so long as the values are within the range that underflow or overflow of the floating point numbers are not a significant effect.

For multiple stages of this process, the loss is merely summed; for a 12 stage radix two machine, the loss would be 0.9%. For a 6 stage radix four machine, it would be 0.4%, provided that sufficient precision is maintained internally within the stage (that is, that the floating point adder preserves internally at least a couple of bits beyond the five seen externally).

Although these losses seem rather low, dropping to four bits does

seem a little risky. The SNR losses (increased by a factor of four over those quoted above) are still acceptable, but one worries about where the loss of power is going. My suspicion is that an appreciable part of it is going into a loss of spectral dynamic range, and for this reason I oppose going to a four bit system.