Block floating point for radar data

Christensen, Erik Lintz

Published in:
I E E E Transactions on Aerospace and Electronic Systems

Link to article, DOI:
10.1109/7.745700

Publication date:
1999

Document Version
Publisher's PDF, also known as Version of record

Block Floating Point for Radar Data

E. LINTZ CHRISTENSEN
Technical University of Denmark

Integer, floating point, and block floating point (BFP) data formats are analyzed and compared in order to establish the mathematical tools for selection of an optimal format which fulfils the demands of high resolution radar (synthetic aperture radar (SAR)) data to large dynamic range and adequate S/N. The analysis takes quantization noise and saturation distortion into account and concludes that it is preferred to use small blocks and a (new) modified BFP format applying fractional exponents. Data from the EMISAR radar system are applied to illustrate the merits of the different schemes.

Manuscript received October 10, 1997; revised May 21, 1998.

IEEE Log No. T-AES/35/1/01508.

This work was supported by DCRS, one of the centers of the Danish National Research Foundation.

Authors' address: Dept. of Electromagnetic Systems, DCRS, Technical University of Denmark, Bldg. 348, DK-2800 Lyngby, Denmark, E-mail: (Lintz@emi.dtu.dk).

INTRODUCTION

Precision and dynamic range are usually not a problem in today's data processing equipment using large data words (integer, floating point, or double precision as appropriate). Similarly, data storage and archiving can rely on large mass storages if needed. However, there are several exceptions where trade-offs are necessary and one of these is the real-time storage (or data link transmission) of multichannel high resolution radar data.

For a linear radar system, e.g. a synthetic aperture radar (SAR), the data precision and dynamic range are usually limited either by 1) the analog to digital (A/D) converter, 2) the limited data rate of the data storage device, or 3) the downlink used to transfer the radar data from the (often airborne or spaceborne) measurement equipment to the facility for final processing. The received signal deviates from the ideal before being converted to digital due to additive thermal noise, interference, spectral distortion from filters, and possibly also distortion from receiver nonlinearity. However, for the purpose of this work, these deficiencies are considered as being part of the signal, while the term noise covers the power value of the deficiencies introduced by the conversion to the final digital data format.

The dynamic range available from A/D converters depends on the required sampling rate but for the data rates used for high resolution radar systems (100 MHz–1 GHz bandwidth) the state of the art is in the range 8–10 bits. This is often insufficient to cover the demands for both dynamic range and precision unless the range is extended by analog means such as time/range dependent analog attenuators or the like. Even then, the 8–10 bits are only just sufficient for linear radar systems using large time-bandwidth product signals, which reduce the peak responses of large point targets.

The A/D converter limits the signal-to-noise ratio (S/N). For a Gaussian distributed signal converted by an 8 bit A/D converter the optimal S/N is 40.5 dB which is only achieved if the rms level of the signal at the A/D converter is carefully and correctly adjusted. Unfortunately, the rms level is usually not stationary so the optimum adjustment is seldom attained during data acquisition.

When the signal has been converted to digital, e.g. by an 8 bit converter, it might be expected that 8 bit precision would be sufficient for the rest of the data storage but often further processing involves digital filtering before the data are transferred to the storage medium (or the downlink). Thus the dynamic range of the data is extended above the word length of the A/D converter.

This work analyzes various data formats aiming at identifying a way of extending the dynamic range with a negligible impact on the other quantities

influencing the total data rate. The data are assumed to be Gaussian distributed. This assumption is reasonable for most modern radar systems using large time-bandwidth product signals without pulse compression prior to A/D conversion. The distribution is assumed to be almost stationary in the sense that the rms value is constant for the number of samples considered (i.e., all the data for the integer case and the data within a single block for the case of block floating point). The analysis assumes the data representation to be continuous prior to the quantization. This is true for the A/D conversion but it is an approximation when already digitized data are truncated or rounded in order to be represented by fewer bits.

In order to establish a reference, equations for quantization noise and saturation distortion caused by limited size integer data representation (e.g. the A/D converters) are derived following the procedure presented by Gray and Zeoli [1] and these are extended to cover floating point representation.

The analysis is further extended to block floating point (BFP) (i.e., one exponent common to a block of samples). Quantization noise in BFP representation has been analyzed by K. Kalliojärvi [2] without assuming a Gaussian distribution of the signal. The present work 1) includes saturation distortion, 2) extends the concept further to include a modified BFP applying fractional scaling, which offers an improved S/N when few bits are allocated for the mantissa, and 3) presents a set of equations which are easily implemented for numeric computations.

This work is concluded by a discussion of reconstruction and removal of the bias introduced by quantization, and an evaluation of BFP with fractional scaling for improvement of the dynamic range of a SAR limited by the maximum data rate of the data storage device. Small data blocks are preferred since the rms value of the signal is changing. Furthermore, small blocks offer better performance than large blocks. Both a significant extension of the dynamic range and an improvement of the S/N can be achieved simultaneously, compared with integer data format, with little or no penalty on the data rate. An example is given to illustrate the result of using the different schemes on real SAR data.

INTEGER DATA REPRESENTATION

When a Gaussian signal is represented by a sequence of integer numbers of limited precision, the peaks of the signal must be limited and the signal below the upper limit must be quantized. This is what ideally takes place in an A/D converter.

Calculation of the combined noise power from quantizing and limiting is fairly straightforward [1]. A Gaussian distribution of the signal with an rms amplitude of σ is presumed throughout:

$$\left.\begin{aligned} p(x) &= \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \\ \int_{-z}^{+z} p(x)\,dx &= \mathrm{Erf}\left(\frac{z}{\sigma\sqrt{2}}\right) = 1 - \mathrm{Erfc}\left(\frac{z}{\sigma\sqrt{2}}\right) \end{aligned}\right\} \tag{1}$$

Assuming an M bit (sign included) linear quantizer (A/D converter) with mid scale at zero, the relation between the saturation levels $\pm XM$ and the quantizer step size Q is

$$XM = Q \cdot (2^{M-1} - 1). \tag{2}$$

The signal is distorted by saturation, when the absolute value of the signal exceeds the saturation level XM, and by quantizing when the absolute value of the signal is smaller. The total equivalent noise power caused by this distortion can be calculated as the sum of the quantization noise Nq and the saturation noise Ns.

The distortion caused by quantizing the complete signal results in a noise power equal to $Q^2/12$, [1], provided the signal can be assumed evenly distributed over the quantizer steps which is an accepted approximation for Gaussian signals with $\sigma \gg Q$. The mean noise contribution from quantizing the signal below the saturation limit is

$$Nq = \frac{Q^2}{12} \int_{-XM}^{+XM} p(x)\,dx = \frac{Q^2}{12} \mathrm{Erf}\left(\frac{XM}{\sigma\sqrt{2}}\right). \tag{3}$$

The mean noise contribution from the saturation is, [1]

$$\left.\begin{aligned} Ns &= 2\int_{XM}^{\infty} (x - XM)^2 \cdot p(x)\,dx \\ &= \sigma^2 \cdot \left(\left(\frac{XM^2}{\sigma^2} + 1 \right) \cdot \mathrm{Erfc}\left(\frac{XM}{\sigma\sqrt{2}}\right) \right. \\ &\quad \left. - \sqrt{\frac{2}{\pi}} \cdot \frac{XM}{\sigma} e^{-(XM^2/2\sigma^2)} \right) \end{aligned}\right\} \tag{4}$$

$$\frac{S}{N} = 10 \cdot \mathrm{Log}\left(\frac{\sigma^2}{Nq + Ns}\right) \quad (dB). \tag{5}$$

Fig. 1 shows the signal power (i.e., σ^2) to noise power (i.e., quantization noise plus saturation noise) ratio (S/N in dB) as a function of the rms signal amplitude, i.e., $\log_2[\sigma/Q]$, for M bit (including sign bit) integer representation with M = 6, 8, 10, and 12. For an 8 bit A/D converter the maximum S/N = 40.54 dB occurs for the rms signal equal $2^{5.02} = 32.4Q$, i.e., ca. 12 dB below the maximum value XM.

It is obvious from the figure that the signal level must be carefully adjusted to achieve the best S/N ratio in conflict with the fact that the signal is usually unknown and changing with time (or range), i.e., the

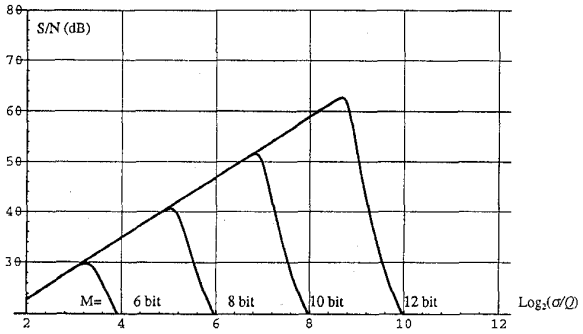


Fig. 1. S/N for M bit (including sign bit) quantizer with saturation versus $\text{Log}_2(\sigma/Q)$, $M = 6, 8, 10, 12$.

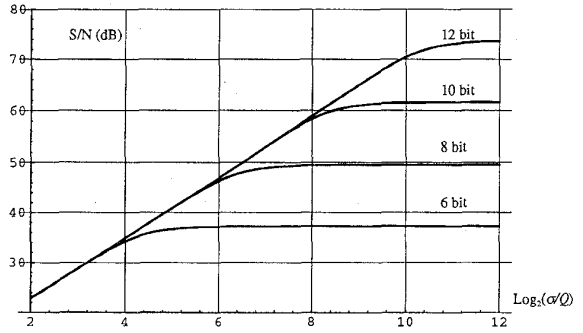


Fig. 2. S/N for M bit (including sign bit) mantissa floating point versus $\text{Log}_2(\sigma/Q)$, $M = 6, 8, 10, 12$ with unrestricted exponent.

Gaussian process modeling the signal is not stationary in real life.

FLOATING POINT DATA REPRESENTATION

For comparison and as a reference for the subsequent sections, Fig. 2 and (6) show the results using a hypothetical floating point quantizer using all the M bits for the mantissa and ignoring for now the number of bits used for the exponent. The floating point quantizer works as the integer when the signal is low. Instead of limiting large signal peaks the quantizer steps is increased by an appropriate factor 2^n and the quantization noise power is thus increased by the factor 4^n.

One observation from Fig. 2 is that the floating point representation with $M-2$ bit mantissa offers S/N that is comparable to the M bit integer representation unless the rms signal can be very precisely adjusted. The consequence of representing the data as floating point with an M bit mantissa and an MM bit exponent, with saturation when the signal exceeds the maximum range of the floating point format, is determined by (7), where the expression for Ns is equivalent to the one derived in (4) while the expression for Nq is similar to the one in (6) except

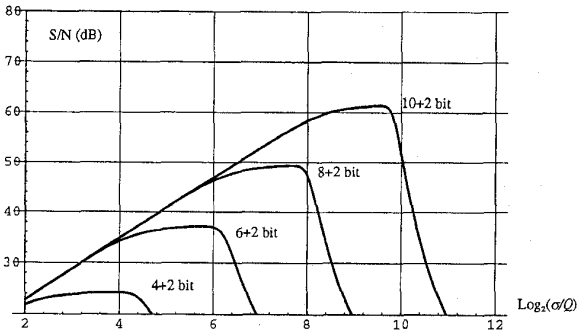


Fig. 3. S/N for M bit (including sign bit) floating point with saturation versus $\text{Log}_2(\sigma/Q)$, $M = 4, 6, 8, 10$ plus 2 bit exponent.

for the limited range of the summation

$$\left.\begin{aligned} Ns &= 0 \\ Nq &= \frac{Q^2}{12}(P(x < XM) + 4^1 P(XM \le x < 2XM) \\ &\quad + 4^2 P(2XM \le x < 4XM)\ldots) \\ &= \frac{Q^2}{12}\left(\text{Erf}\left(\frac{XM}{\sigma\sqrt{2}}\right) + \sum_{i=1}^{\infty} 4^i \right. \\ &\quad \left. \cdot\left(-\text{Erf}\left(\frac{2^{i-1}\cdot XM}{\sigma\sqrt{2}}\right) + \text{Erf}\left(\frac{2^i \cdot XM}{\sigma\sqrt{2}}\right)\right)\right) \end{aligned}\right\} \tag{6}$$

$$\left.\begin{aligned} ii &= 2^{MM} - 1 \\ Ns &= \sigma^2 \cdot \left(\left(\frac{(2^{ii}XM)^2}{\sigma^2} + 1\right)\cdot \text{Erfc}\left(\frac{2^{ii}XM}{\sigma\sqrt{2}}\right)\right. \\ &\quad \left. -\sqrt{\frac{2}{\pi}}\cdot\frac{2^{ii}XM}{\sigma} e^{-((2^{ii}XM)^2/2\sigma^2)}\right) \\ Nq &= \frac{Q^2}{12}\left(\text{Erf}\left(\frac{XM}{\sigma\sqrt{2}}\right) + \sum_{i=1}^{ii} 4^i \right. \\ &\quad \left. \cdot\left(-\text{Erf}\left(\frac{2^{i-1}\cdot XM}{\sigma\sqrt{2}}\right) + \text{Erf}\left(\frac{2^i \cdot XM}{\sigma\sqrt{2}}\right)\right)\right) \end{aligned}\right\}. \tag{7}$$

Fig. 3 shows the results if 2 bits are converted from mantissa to exponent with signal saturation when the signal would require a larger exponent than 3. The loss in S/N (compared with using all bits for integer representation with optimal signal magnitude) obviously becomes smaller the larger the total number of bits since the quantization noise then has decreasing importance and for e.g., 18 bits in total, the maximum S/N is better when 2 of the bits are used as an exponent.

BLOCK FLOATING POINT

When the dynamic range of the available data word size is not sufficient, this can be improved by

using floating point, but the increased number of bits (or the reduced precision) due to the exponent may not be acceptable. BFP combines a number of signal samples with one common exponent. This saves data overhead (compared with floating point) at the cost of an increased quantization noise since one large sample causes all the samples in a block to be quantized coarser.

When all data in a block are small enough to be represented by the mantissa alone (i.e., the exponent 0), the quantization noise has the lowest value. When the largest value in a block requires an exponent of n, the quantization noise of all the data in the block is increased to 4^n times the lowest value.

The mean noise is then the lowest noise times the probability that all data in the block are smaller than the maximum number in the mantissa plus the lowest noise times 4^1 times the probability that the largest value falls in the interval between the maximum mantissa and 2 times the maximum mantissa, etc. Using the same definitions as before, (7) can be extended to cover this case by introducing the block size bz (the number of samples per block) and the probability that all samples within a block are below a limit, i.e., $2^i XM$:

$$P((\text{all } |x| \text{ in group of } bz) < 2^i \cdot XM) = \left(\text{Erf}\left(\frac{2^i \cdot XM}{\sigma\sqrt{2}}\right)\right)^{bz}. \tag{8}$$

In any practical system there will be an upper limit for the exponent and thus for the maximum signal which can be represented. Consequently, saturation will take place at some signal level. When some samples are saturated, the rest of the samples in the same block are quantized with the same stepsize as used when the largest sample is just below the limit, i.e., in addition to the distortion from saturation we get noise from quantizing the samples which are members of blocks with some samples above the limit but are themselves below the limit. The number of such samples are, assuming the limit to be $2^{ii}XM$:

$$bz \cdot \left[\left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}}\right)\right) - \left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}}\right)\right)^{bz}\right] \tag{9}$$

where the first part of (9) is the average number of samples with magnitude below the limit and the second part is the average number of samples being included in blocks where all samples are below the limit.

Applying the principles of (7) together with (8) and (9) we get the noise contributions:

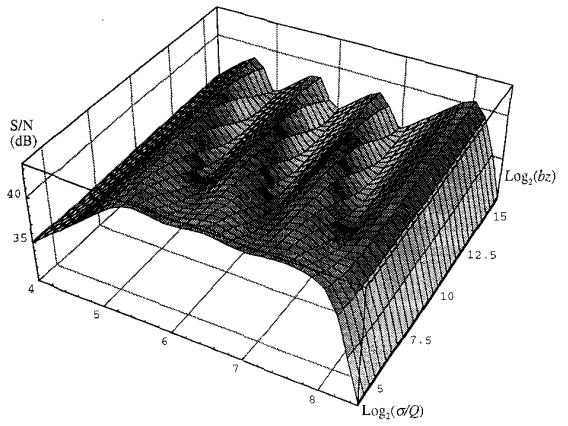


Fig. 4. S/N for BFP (8 bit mantissa, 2 bit exponent and saturation at maximum) versus $\text{Log}_2(\sigma/Q)$ and Log_2 (block size).

$$\left.\begin{aligned} ii &= 2^{MM} - 1 \\ Ns &= \sigma^2 \cdot \left(\left(\frac{(2^{ii}XM)^2}{\sigma^2} + 1\right) \cdot \text{Erfc}\left(\frac{2^{ii}XM}{\sigma\sqrt{2}}\right) \right. \\ &\qquad \left. -\sqrt{\frac{2}{\pi}} \cdot \frac{2^{ii}XM}{\sigma} e^{-((2^{ii}XM)^2/2\sigma^2)}\right) \\ Nq &= \frac{Q^2}{12} \cdot \left[\left(\text{Erf}\left(\frac{XM}{\sigma\sqrt{2}}\right)\right)^{bz} \right. \\ &\qquad + \sum_{i=1}^{ii} 4^i \cdot \left(-\left(\text{Erf}\left(\frac{2^{i-1} \cdot XM}{\sigma\sqrt{2}}\right)\right)^{bz} \right. \\ &\qquad \left. + \left(\text{Erf}\left(\frac{2^i \cdot XM}{\sigma\sqrt{2}}\right)\right)^{bz}\right) \\ &\qquad + 4^{ii} \cdot \left(\left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}}\right)\right) \right. \\ &\qquad \left. \left. - \left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}}\right)\right)^{bz}\right)\right] \end{aligned}\right\}. \tag{10}$$

The assumptions that the quantization noise is $Q^2/12$ (assuming an equal distribution over the quantizer interval Q) does not hold in general for the signal peaks much larger than the signal rms value (some of the signal samples may then be smaller than the quantizer step). Such large peaks will occur with low probability and the influence on the total noise will be small except for cases with very small number of bits in the mantissa.

Fig. 4 shows an example with 8 bit mantissa and 2 bit exponent and saturation at the top of the range. It is noted that for small block sizes (the smallest block size bz on the figure is $2^4 = 16$) and for large rms signal magnitudes the S/N is almost constant (around 43 dB) up to the point where saturation occur often and the S/N decreases.

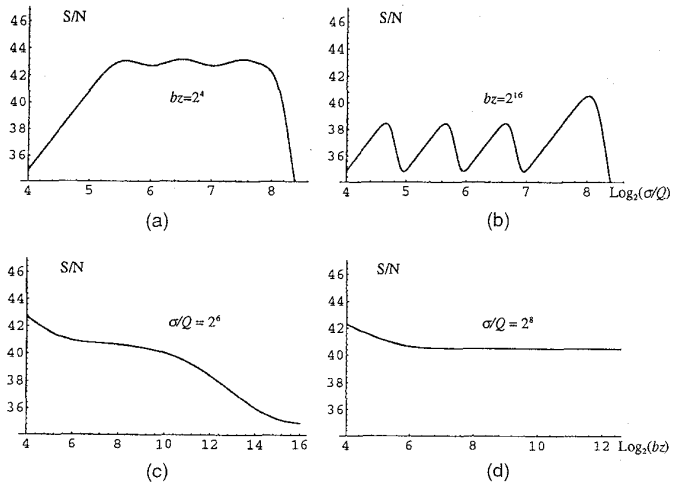


Fig. 5. S/N for BFP (8 bit mantissa, 2 bit exponent and saturation at maximum) (a) and (b) versus $\text{Log}_2(\sigma/Q)$ for block size = 2^4 and 2^{16}. (c) and (d) versus Log_2 (block size) and $\sigma/Q = 2^6$ and 2^8.

However, it is also noted that intentional limitation of the signal peaks can be advantageous. For large block sizes a large number of samples may suffer a coarse quantization when one sample increases slightly, i.e., the noise added by coarser quantization may be larger than the alternatively added noise by limiting the signal peaks. For the example given in Fig. 4 (8 bit mantissa) the S/N for large blocks has a maximum of about 40.5 dB (i.e., the same as the maximum for 8 bit integer representation), obtained at the rms signal $\sigma = 2^8 Q$, i.e., 12 dB below the limit.

The signal must be scaled accurately to take advantage of the peak in the S/N and thus the rms value of the signal must be the same for all the samples in a block. Furthermore, it is noted that if the signal is not properly scaled, the S/N is worse than it is for small block sizes. This is further illustrated in Fig. 5 which highlights selected subsets of Fig. 4. The conclusion which may be drawn from this is that one should choose the smallest block size consistent with the acceptable overhead for the exponent.

BLOCK FLOATING POINT WITH FRACTIONAL EXPONENT

The BFP format described in the previous section simply uses the exponent which is necessary to bring the largest value in a block within the range of the mantissa and this means that the quantization noise changes by a factor 4 for each increment of the exponent. This is a consequence of the (implementation driven) choice of using a base 2 number system. Using a smaller base (> 1) number system would reduce this effect but would also require more digits for the representation.

For the case of BFP, where the exponent is shared by several samples, a smaller base number can be applied for the exponent without too high costs with regard to the total number of bits. One way of implementing this is to normalize all samples in a block to the largest sample in the block and then use the value of the largest sample instead of the exponent. The theoretical quantization noise and saturation noise (still assuming an upper limit for the largest sample) is calculated in (11) and the resultant S/N for different block sizes is displayed as the upper limit in Fig. 6

$$\left.\begin{aligned} ii &= 2^{MM} - 1 \\ Ns &= \sigma^2 \cdot \left(\left(\frac{(2^{ii} XM)^2}{\sigma^2} + 1 \right) \cdot \text{Erfc}\left(\frac{2^{ii} XM}{\sigma\sqrt{2}} \right) - \sqrt{\frac{2}{\pi}} \cdot \frac{2^{ii} XM}{\sigma} e^{-((2^{ii} XM)^2/2\sigma^2)} \right) \\ Nq &= \frac{Q^2}{12} \cdot \left[\sum_{X=\delta}^{X=2^{ii}\cdot XM} \left(-\left(\text{Erf}\left(\frac{X-\delta}{\sigma\sqrt{2}} \right) \right)^{bz} + \left(\text{Erf}\left(\frac{X}{\sigma\sqrt{2}} \right) \right)^{bz} \right) \cdot \left(\frac{X}{XM} \right)^2 + 4^{ii} \left(-\left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} + \left(\text{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}} \right) \right) \right) \right] \\ \delta &= \text{small increment} \end{aligned}\right\} \quad (11)$$

The approach described above (11) is not practical for a real-time system for several reasons including the assumption of continuous data representation (i.e.,

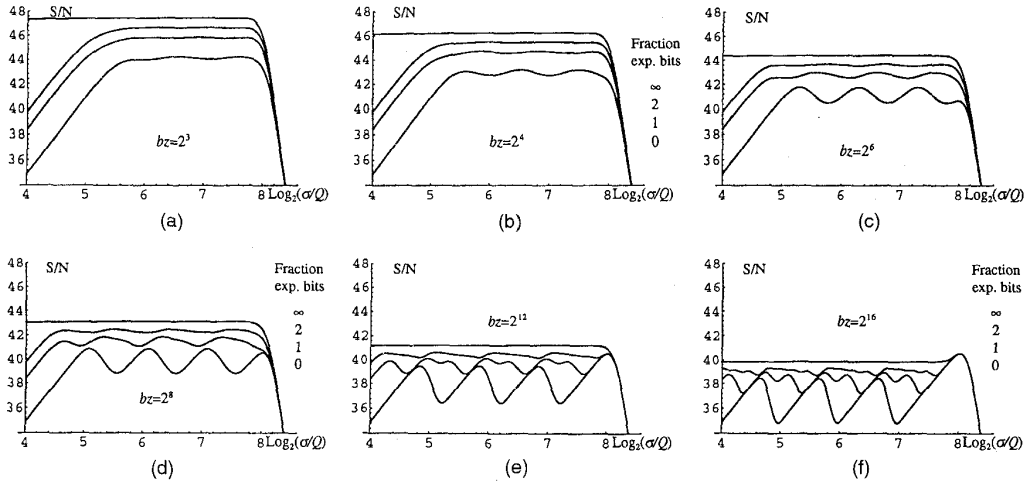


Fig. 6. S/N for BFP with saturation ($M = 8$ $MM = 2$) without and with fractional scaling (1, 2, ∞ bits) versus $\mathrm{Log}_2(\sigma/Q)$. (a) Block size = 2^3. (b) Block size = 2^4. (c) Block size = 2^6. (d) Block size = 2^8. (e) Block size = 2^{12}. (f) Block size = 2^{16}.

many significant bits below the quantizing level) and the use of many bits to represent the largest sample in each block. An approximation, which takes these problems and the implementation into account, could be to scale all the data values within a block with a factor, which brings the largest data value close to the top of the range of the mantissa. In this way a reduction in quantization noise, compared with the normal BFP representation, can be obtained. This requires the exponent to be extended by e.g., 1 or 2 (fractions) bits. A further reduction can be obtained by using more bits on the fractional exponent but the payoff per bit will be smaller, as can be verified from Fig. 6.

The consequence of using fractional scaling is calculated following the same principles as for BFP in the previous section with the modification that all numbers in a block are multiplied by a scale factor, depending on what is required to scale the largest value to the upper part of the mantissa. The scale factor may be $1/w$ for values below $w < 1$ for the simplest case (1 bit) or $\{1/w1, 1/w2, 1/w3\}$ for values below $\{w1 < w2 < w3 < 1\}$ for the slightly more complex case (2 bits). This multiplication before conversion to BFP reduces the quantization noise inversely proportional to the square of the scaling factor. The calculations are split up in intervals reflecting the probability for using the different scaling factors.

The particular scaling factors used in the examples here are selected to make the hardware implementation simple and they are not optimal in any other sense. The coefficients used are $w = 2/3$ for the 1 bit case $\{w1, w2, w3\} = \{4/7, 4/6, 4/5\}$ for the 2 bit case. Thus the scaling with $1/w$ becomes a simple integer multiplication and a division by 2 or 4, which can be performed by adjusting the exponent. The calculation of noise, and thus S/N ratio, follows the same procedures as used for (10). Equation (12) gives the analytic expression for the case with one scale coefficient

$$
\begin{aligned}
&ii = 2^{MM} - 1 \\
&Ns = \sigma^2 \cdot \left(\left(\frac{(2^{ii} XM)^2}{\sigma^2} + 1 \right) \cdot \mathrm{Erfc}\left(\frac{2^{ii} XM}{\sigma\sqrt{2}} \right) \right. \\
&\qquad \left. - \sqrt{\frac{2}{\pi}} \cdot \frac{2^{ii} XM}{\sigma} e^{-((2^{ii} XM)^2/2\sigma^2)} \right) \\
&Nq = \frac{Q^2}{12} \cdot \left[w^2 \cdot \left(\mathrm{Erf}\left(\frac{w \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right. \\
&\qquad + \left(- \left(\mathrm{Erf}\left(\frac{w \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right. \\
&\qquad\quad \left. + \left(\mathrm{Erf}\left(\frac{XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right) \\
&\qquad + \sum_{i=1}^{ii} 4^i \cdot \left[w^2 \cdot \left(- \left(\mathrm{Erf}\left(\frac{2^{i-1} \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right. \right. \\
&\qquad\quad \left. \left. + \left(\mathrm{Erf}\left(\frac{2^i \cdot w \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right) \right. \\
&\qquad + 1 \cdot \left(- \left(\mathrm{Erf}\left(\frac{2^i \cdot w \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right. \\
&\qquad\quad \left. \left. + \left(\mathrm{Erf}\left(\frac{2^i \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right) \right] \\
&\qquad + 4^{ii} \cdot \left(- \left(\mathrm{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}} \right) \right)^{bz} \right. \\
&\qquad\quad \left. \left. + \left(\mathrm{Erf}\left(\frac{2^{ii} \cdot XM}{\sigma\sqrt{2}} \right) \right) \right) \right]
\end{aligned}
\tag{12}
$$

The results of using fractional scaling are displayed in Fig. 6 for block sizes 8, 16, 64, 256, 4096, and 65536, respectively, together with BFP without fractional scaling. In all cases the lowest S/N is obtained without fractional scaling while the highest S/N is obtained using 2 bit fractional scaling (except for the even higher theoretical limit with ∞ bit fraction). The upper limit curves are based on (11).

QUANTIZATION AND RECONSTRUCTION

The results displayed in the previous sections were all based on the assumption that a continuous signal was quantized, i.e., the signal was either digital, with a much better precision than the one actually utilized, or analog. This assumption is true for an A/D converter but it is less likely to be true for digitized signals to be converted to BFP format.

There is a fundamental difference, which is often overlooked, between quantizing an analog signal and quantizing a digital signal, e.g. in binary representation. The A/D converter output is (assumed to be) representing the signal range from $-Q/2$ to $+Q/2$ symmetric around each output value. When binary data are quantized by discarding the least significant bits a bias is introduced. Fewer bits cannot represent the center value. The best estimate is obtained by rounding but even then the output values will be $Q/2$ too large (Q being the value of the least significant bit before quantization) in 2s complement representation and for the positive part in sign/magnitude representation, and $Q/2$ too small for the negative part in sign/magnitude representation (assuming rounding by addition of 0.5 and truncation). This bias is significant, when a signal already represented by few bits is further quantized, and it is important to remove it before further processing.

When the binary data have been quantized, the information on the number of bits discarded is given in the exponent of floating point and BFP representations so the central (bias free) estimate of the data values may later be reconstructed (assuming rounding) by appending a number of zeros and subtracting (or adding for the negative part if sign/magnitude representation) $Q/2$ where Q is the value of the least significant of the discarded bits. When the exponent is zero there is no bias to remove.

When fractional scaling, as defined in the previous section, has been applied, the scaling changes the value of the least significant bit. For the case of 1 bit fraction, using the weight factors of 1 or 2/3, the data were multiplied by 1 or 3/2 before rounding, i.e., the value to be subtracted or added is $Q/2$ or $Q/4$ depending on the fractional exponent. For the case of 2 bit fraction the data were multiplied by 1, 7/4, 3/2, or 5/4, i.e., the value to be subtracted or added is $Q/2$, $Q/4$, or $Q/8$ depending on the fractional exponent.

A few additional facts need consideration for the case of few bits in the mantissa. 1) It is important to utilize all possible bit combinations. This means that sign/magnitude representation is not a good choice since one of the combinations are not used (e,g, for 5 bits the range covered is ± 15; including zero this offers 31 values instead of 32). 2) It is also important to use all bit combinations equally. This implies that truncation (and the addition of 0,5 at reconstruction rather than at encoding) is preferred to rounding because the latter does only use the most negative combination for half an interval and gives overflow for the most positive values. 3) Further to be considered is the fact that utilization of fractional scaling results in output values which are in general not equally distributed over the possible values, i.e., the best reconstruction value may not always be as simple as just adding 0.5 and correcting for the small offset $-Q/4$.

APPLICATION EXAMPLE

EMISAR is a dual frequency fully polarimetric SAR [3] acquiring 8 complex numbers for each resolution cell (2 frequencies, each with 4 polarizations). The present equipment uses 8 bit A/D converters and transfers the data to tape as 2×8 bit integer per complex number. Considering the normal sampling density (1.5×1.5 m), swath width (8192 samples), and aircraft velocity (240 m/s) this amounts to around 220 Mbit/s including ancillary data. The tape recorder is an Ampex DCRSi 240 which can record with a sustained data rate of up to 240 Mbit/s, i.e., there is room for an overhead of up to 9% but for various practical reasons it is preferred to keep the overhead below 6% (i.e., less than 1/16).

The 8 bit A/D converters limit the range of the signal but the following range and azimuth filters potentially extend the dynamic range (both reducing the noise and increasing the maximum). Although these data cannot be considered exactly Gaussian distributed, it is obvious that simply limiting the data to 8 bit integer after filtering (see Fig. 1) will reduce the data quality. Furthermore, the rms value is usually changing over the radar swath and a simple way of handling this fact is desirable.

An update of the A/D converters to 10 bit converters is planned to improve the adaptation to changing signal levels. When online filtering is taken into account the useful information may require 12 or even more bits but the capacity of the tape recorder does not permit the word length to be increased correspondingly for the tape storage format. Even 9 bits per word is too much unless the swath width is sacrificed.

A 3 bit exponent permits the dynamic range to be extended by a factor of 2^7 ($2^3 - 1 = 7$) which, for an 8 bit mantissa, gives the same dynamic range as an

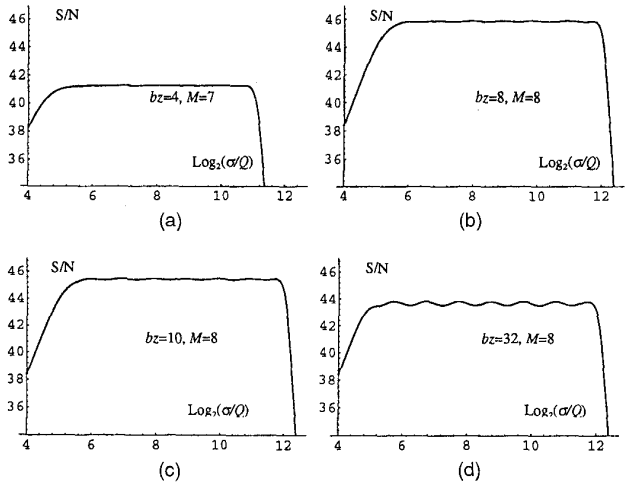


Fig. 7. S/N for BFP (3 bit exponent and saturation), with 1 bit fractional scaling versus $\mathrm{Log}_2(\sigma/Q)$. (a) $M = 7$ bits, $bz = 4$. (b) $M = 8$ bits, $bz = 8$. (c) $M = 8$ bits, $bz = 10$. (d) $M = 8$ bits, $bz = 32$.

integer with 15 bit. This is sufficient to cope with a 12 bit A/D converters and the additional dynamic range achieved by the filters. Taking advantage of the fractional scaling, 4 bits per exponent is desired.

One possible solution is illustrated in Fig. 7(a): using a block size $bz = 4$ (i.e., 2 complex samples), gives a $S/N = 41.2$ dB for 7 bit mantissa and fractional scaling. Using a 3 bit exponent (plus the 1 bit fractional exponent), permits each block to be contained in 32 bits. This format offers a better S/N than the best possible for 8 bit integer (see Fig. 1), with a much larger dynamic range and with exactly the same data rate.

As illustrated by Fig. 7(c), a considerable improvement of the S/N is possible. Using a block size = 10 (5 complex samples) gives $S/N = 45.5$ dB for 8 bit mantissa and fractional scaling. Using 3 bit exponent (plus the 1 bit fractional exponent) results in a 5% overhead.

It might be considered desirable for implementation reasons to operate with 2^n as block size and using 8 bits for the exponent. However, this would require a block size of 32 (Fig. 7(d)) to keep the overhead below 6% and the S/N would drop to 43.6 dB.

Instead, it could be decided to reduce the total number of data samples slightly and combine two 4 bit exponents in one 8 bit byte. Then blocks with 8 samples (Fig. 7(b)) and their exponents could meet the data rate requirement with a S/N of 45.9 dB.

All four solutions are fully compatible with the requirements to EMISAR [4]. The data rates are all within the stipulated limit set by the tape recorder (max 6% above that needed for 8 bit integer representation. The solution with 7 bit mantissa is preferred for several reasons as follows.

1) The solution offers the same S/N as the present 8 bit integer representation for small signal levels and a better S/N (41 dB) than the 8 bit integer for all signal levels above $2^5 = 32$ (see Fig. 1 versus Fig. 7(a)).

2) A 41 dB rms signal to quantization and saturation noise is adequate for the data quality required by remote sensing applications provided the analog signal at the A/D converter input fulfills the requirements.

3) The limiting value of 2^{14} achieved with 7 bit mantissa and 3 bit exponent is sufficient even for a 10 bit A/D converter and the expansion achieved by online preprocessing.

4) The reduction in quantization noise achieved by implementing the 1 bit fractional exponent is around 1.8 dB for the small blocks considered for EMISAR and since the costs are low the feature is included.

5) An implementation applying very small blocks has the advantage that it adapts rapidly to signal rms variations (and thus is also more tolerant to the actual distribution function) and the damage from pulse interference in the input data and possible bit errors in the exponent are minimized.

6) The data rate is the same as for 8 bit integer representation and 2 complex samples including exponent can be packed in a 32 bit data word which simplifies the unpacking. The other suggested solutions all results in higher data rates and more complicated data packing/unpacking while their higher S/N and dynamic range are not required.

PERFORMANCE ILLUSTRATION

Experimental verification of the performance of the solution discussed in the previous section can be performed although real SAR data with sufficient dynamic range are not readily available since the present EMISAR system is limited to the recording of 8 bit integer representation and an upgrading,

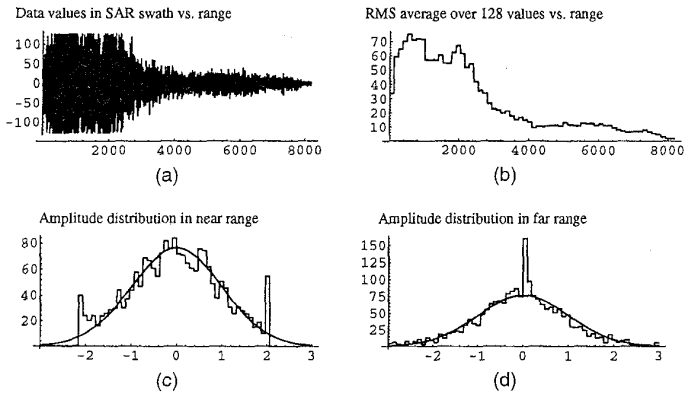


Fig. 8. (a) Sample of SAR swath with 4096 complex samples (8192 data values) showing signal versus range. (b) RMS value averaged over 128 values versus range. (c) and (d). Sample histograms of 2048 values of SAR swath. (c) RMS = 62 in near range. (d) RMS = 11 in far range. Abscissa axes are normalized to rms values (±3 rms). Smooth curves show normal distribution as reference.

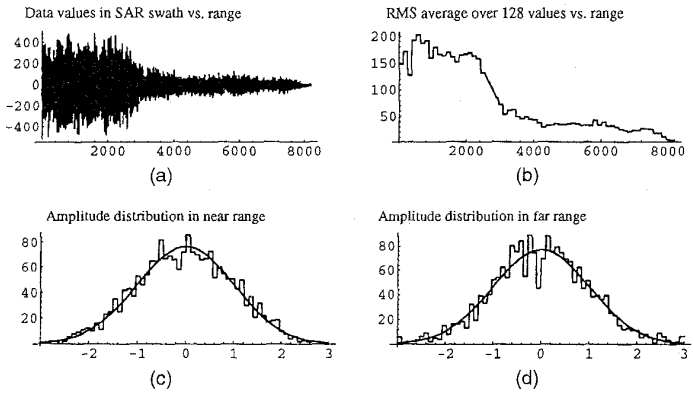


Fig. 9. (a) and (b). Sum of 4 SAR swaths with 4096 complex samples (8192 data values) showing signal versus range and rms value averaged over 128 values versus range. (c) and (d). Sample histograms of 2048 values of sum of 4 SAR swaths. (c) RMS = 170 in near range. (d) RMS = 32 in far range. Abscissa axes are normalized to rms values (±3 rms). Smooth curves show normal distribution as reference.

presently under construction, will deliver data in the BFP representation without the original data.

Test data were acquired by EMISAR at L-band using a mode, where the data from the A/D converters are recorded directly on tape without any on-line processing but applying the high sampling density (37.5 cm in azimuth) normally used together with on-line azimuth filtering and decimation. In order to get a wide swath and still keep the data rate within limits, the swath was reduced to 4096 samples at a sampling density of 6 m in range. A swath thus consists of 4096 complex samples or 8192 real values for each polarization. Fig. 8 offers a description of a single polarization swath by displaying a) an example of the uncompressed SAR data, b) the rms value versus range, and histograms of c) the near range (sample 500–2548), and d) the far range (sample 5000–7048). The average rms value is 36 being fairly close to the optimum of 32 for an 8 bit system.

The data acquired from a scene with significant changes in the signal rms versus range will either be saturated at the large rms parts or will only be utilizing a few of the bits in the low rms parts. In either case the amplitude distribution will deviate significantly from that of the original analog signal. It is obvious from the histograms in Fig. 8(c), (d) that the data have been limited in near range and have an exceptionally large number of very small samples in far range.

The physical antenna pattern of a strip mapping SAR illuminates a wide area so subsequent range lines cover reflections from virtually the same objects. Consequently, the rms value of the data versus range changes little from one range swath to the next and the summation of a number of swaths will increase the dynamic range of the data values without changing the relative variation of rms versus range.

Fig. 9 displays the same 4 elements as Fig. 8, however, for the sum of 4 range swaths acquired with 37.5 cm separation in azimuth. The sum of 4 swaths is quite close to the output of the azimuth filter in the SAR when that is set for filtering and

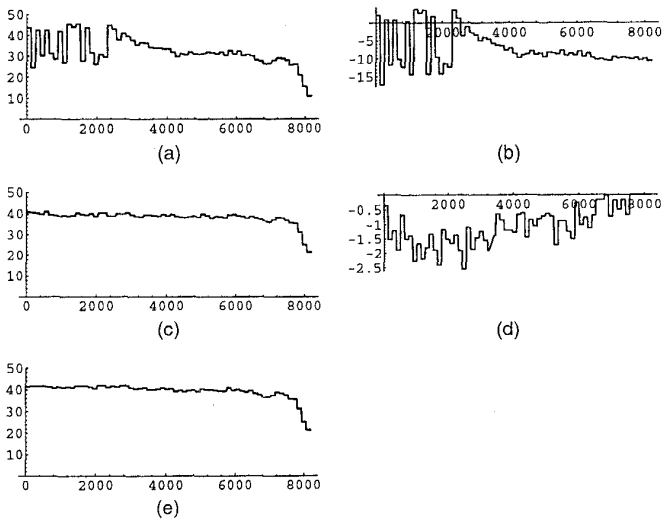


Fig. 10. S/N versus range (and thus signal level) for signal of Fig. 9 and 3 encoding schemes. (a) 8 bit integer. (c) 7 bit BFP with $bz = 4$. (e) 7 bit BFP and fractional scaling with $bz = 4$. (b) and (d). Display performance of integer and BFP relative to that of BFP with fractional scaling.

decimation by 4. It can be seen from the ratio between the rms values (ca. 3 : 1) that the 4 swaths are highly correlated even if the usual motion compensation was omitted in this case

The data displayed in Fig. 9 have been encoded directly by the algorithms for BFP with and without fractional scaling using 7 bit mantissa and a block size of 4, i.e. the preferred solution from the previous section which was displayed in Fig. 7(a). For comparison, the data have also been encoded to 8 bit integer representation after scaling to the optimum average (i.e., $2^5 = 32$) rms over the swath.

After encoding the best estimate of the data were reconstructed for all 3 sets and the deviations from the originals were calculated. The power values of these deviations were averaged over 128 range samples and increased by the $Q^2/12$ inherent quantization noise. The signal power versus this total distortion power in dB (S/N) is displayed in Fig. 10 for the 3 schemes as a function of range (and thus as a function of the local signal level). Fig. 10 also displays the differences in dB between the BFP with fractional scaling and the two other schemes.

Fig. 10(a), (b) verifies that the integer formats is not suited for data with a large variation of the rms value. The 8 bit integer format does offer better S/N (i.e., smaller reconstruction errors) than the 7 bit BFP for data samples where all the 8 bits are utilized but the penalty of the necessary scaling and limiting is high both for large data values causing saturation and for small data values for which the quantization noise is increased.

The important factor in S/N degradation is the ratio between the larger rms and the smaller rms within the same scene (i.e., the area where the same average rms value is assumed in adjusting the signal levels). For the example given in Fig. 9 and 10 the overall average rms value is 101, the near range rms is 170 (1.68 times the average), and the far range rms is 32 (0.32 times the average).

When the average rms value is scaled to the optimum value of 2^5, the near range rms is scaled to $2^{5.75}$ and the far range rms to $2^{3.34}$. The expected S/N values are 24.3 dB and 31.0 dB, respectively (Fig. 1). The results displayed in Fig. 10(a), (b) are reasonably close to this. When the number of significant bits in the input signal is increased, i.e., by summing a larger number of range swaths with the same rms versus range, the S/N in both the high rms and the low rms areas comes even closer to the theoretical values (Fig. 11(a), (b)).

BFP using fractional scaling (Fig. 10(e)) is also seen to be better than conventional BFP (Fig. 10(c), (d)) although the gain in S/N is limited to be between 1 and 2 dB for the actual test data. The theoretical value is 1.8 dB for 7 bit mantissa and a block size of 4. There is no advantage in scaling when the data values are small enough to allow all significant bits to be included in the mantissa and this occurs frequently when the rms value is small, i.e., in the far range of the example) but also occasionally when the rms value is large.

The average difference between the S/N for the 2 BFP formats is very close to the theoretical value when the $Q^2/12$ inherent quantization noise is not included in the S/N calculation, i.e., when the unlimited quantized signal is considered as the reference rather than the original analog signal. The results also come closer to the theoretical values when the number of significant bits in the input signal is increased (Fig. 11(c), (d), (e)). In this case it is especially reflected in the far range values.

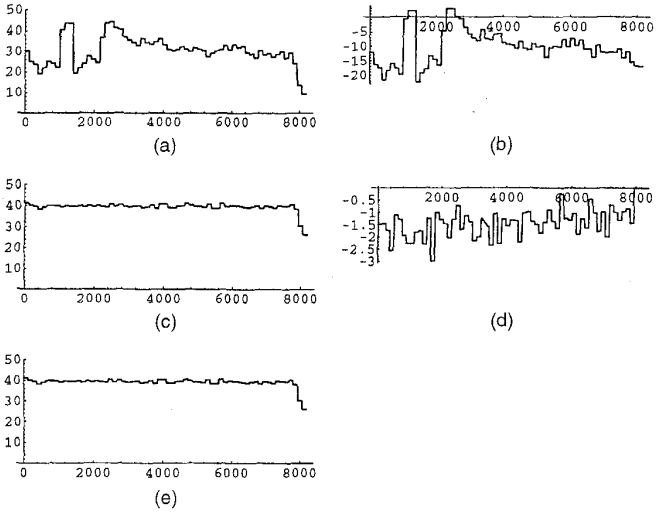


Fig. 11. S/N versus range (and thus signal level) for sum of 16 SAR swaths with 4096 complex samples (8192 data values) and 3 encoding schemes. (a) 8 bit integer. (c) 7 bit BFP with $bz = 4$. (e) 7 bit BFP and fractional scaling with $bz = 4$. (b) and (d) Display performance of integer and BFP relative to that of BFP with fractional scaling.

CONCLUSION

This paper has presented a unified theoretical analysis of the distortion caused by quantization and saturation for integer, floating point, and BFP data formats used to represent, with a limited number of bits, a continuous Gaussian distributed signal. A modified BFP format with fractional exponent has been introduced in order to improve the performance of the BFP format.

The merits of the various formats have been demonstrated on SAR data and the advantages of the BFP formats in handling data with large variations in the rms value (i.e., non-Gaussian distributed) have been verified. It is concluded that significantly improved data quality can be achieved for a SAR system without any increase in the data rate by using BFP instead of integer.

ACKNOWLEDGMENT

The author is grateful to Professor S. Nørvang Madsen of the Danish Center for Remote Sensing (DCRS), DTU, for fruitful discussions and many helpful suggestions. S. Savstrup Kristensen is acknowledged for his valuable contributions regarding implementation of algorithms.

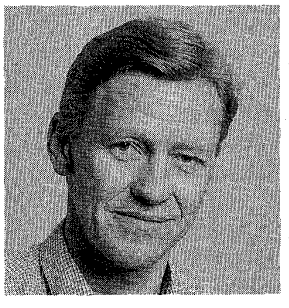
REFERENCES

[1] Gray, G. A., and Zeoli, G. W. (1971)
Quantization and saturation noise due to analog to digital conversion.
IEEE Transactions on Aerospace and Electronic Systems (Jan. 1971), 222–223.

[2] Kalliojärvi, K. (1993)
Analysis of Block-Floating-Point Quantization Error.
Amsterdam: Elsevier, 1993, 791–796; also in *Proceedings of the 1st European Conference on Circuit and Design*, Davos, Switzerland, Aug. 30–Sept. 3, 1993.

[3] Christensen, E. L., Dall, J., Skou, N., Woelders, K., Granholm, J., and Madsen, S. N. (1996)
EMISAR: C- and L-band polarimetric and interferometric SAR.
In *Proceedings of the International Geoscience and Remote Sensing Symposium*, IGARSS'96, Lincoln, NE, May 27–31, 1996, 1629–1632.

[4] Christensen, E. L., Skou, N., Dall, J., Woelders, K. W., Netterstrøm, A., Jørgensen, J. H., Granholm, J., and Madsen, S. N.
EMISAR: An absolutely calibrated polarimetric L- and C-band SAR.
IEEE Transactions on Geoscience and Remote Sensing (Nov. 1998), 1852–1865.



Erik Lintz Christensen received the M.Sc.E.E. in 1966.

In 1968 he joined the Electromagnetics Institute, now Dept. of Electromagnetic Systems (EMI), Technical University of Denmark, Lyngby, where he is now an Associate Professor. His work has covered many aspects of radar, radio communications, measurement systems, and high frequency and microwave electronics. This includes the design of 60 MHz and 300 MHz radar systems for recording the thickness of the ice sheets of Greenland and Antarctica and various microwave measurement systems. He was project manager of the Danish Airborne SAR program implementing a dual frequency fully polarimetric SAR system completed in early 1996. Since February 1994 he has been comanager of the Danish Center for Remote Sensing.