The Properties of Floating Point Single Quantization including Under- and Overflow

F. Hartwig and A. Lacroix
University of Frankfurt, Institute of Applied Physics
D - 60325 Frankfurt am Main, Robert - Mayer - Straße 2-4

ABSTRACT

The quantization of floating point numbers is well investigated for situations where no under- or overflow occurs [1-3]. In this paper results are presented including these cases for the quantization of uniform, gaussian and sinusoidal distributed numbers. For underflow two different cases are considered: (1) no unnormalized mantissas occur and numbers whith magnitudes lower than a certain limit are set to zero, (2) unnormalized mantissas are used in the underflow region which leads to a behaviour similar to that of fixed point quantization. It can be seen that different slopes of the SNR vs. S curves in the underflow regions characterize the utilization of normalized or unnormalized mantissas. In the overflow-region it is assumed that saturation is utilized, which means that numbers with magnitude greater than a certain limit are set to fixed overflow values.

1. Quantization Analysis

The theoretical analysis of the SNR-curves has to take into account three contributions. These contributions are due to:
Underflow region: The quantized values are set to zero or represented by denormalized mantissas.
Region of mantissa quantization: Utilizes the concept of the relative error in a simplified model or power of absolute error in an improved model.
Region of overflow: All quantized numbers are set to a certain maximum value. The power of the absolute error in this domain is the significant portion here.

1.1 Underflow region

1.1.1 *Normalized Mantissas*

All numbers with a magnitude smaller than a certain limit ξ_{min} are set to zero. Therefore the underflow portion for the total noise power is:

$$\sigma^2_{e_{abs}} = 2\int_0^{\xi_{min}} x^2 p_\xi(x)dx$$

where $p_\xi(x)$ dennotes the probability density function of the quantized signal.

1.1.2 *Nonnormalized Mantissas*

If denormalized numbers are used we have a situation similar to fixed point quantization, that means the error power in that region is independent of the signal. The smallest number in this underflow situation is $\xi_{min}2^{-t+1}$

Values smaller than this limit are set to zero. So the underflow region is split into two different regions an the total noise power contribution is:

$$\sigma^2_{e_{abs}} = 2\left(\int_0^{\xi_{min}2^{-t+1}} x^2 p_\xi(x)dx + \frac{Q^2}{12}\int_{\xi_{min}2^{-t+1}}^{\xi_{min}} p_\xi(x)dx\right)$$

The corresponding quantization step size in the region of denormalized mantissas is given by $Q = x_{min}2^{-t+1}$.

1.2 Contribution of mantissa quantization

The simplified model assumes constant power of relative error between ξ_{min} and ξ_{max} , which leads to a contribution of

$$\sigma^2_{e_{abs}} = 2\,\frac{2^{-2t}}{8\ln 2}\int_{\xi_{min}}^{\xi_{max}} x^2 p_\xi(x)dx \quad .$$

Here the assumtion of reciprocal distributed mantissas is included, which is a good approximation for gaussian ditributed signals [3].

1.3 Contribution of Overflow

In the region of overflow with saturation the absolute

quantization error is given by

$$\varepsilon_{abs}^2 = (\xi - \xi_{max})^2$$

Therefore the total power of the overflow quantization error is given by:

$$\sigma_{e_{abs}}^2 = 2 \int_{\xi_{max}}^{\infty} (x - \xi_{max})^2 p_\xi(x) dx$$

2. Improved Model

As can be seen by simulation results there is a signifcant ripple in the region of mantissa quantization. This effect is due to a nonreciprocal mantissa distribution in the highest dyade. The deviation of this reciprocal mantissa distribution is small for gaussian distributed numbers, visible for uniform distributed numbers and stronger for sinusoidal signals.
In one dyade there is a constant power of the absolute error, i.e. $Q^2/12$. Here the local quantization step size according to the current exponenent has to be used. Every dyade leads therefore to a different contribution to the absolute noise power. All different contributions have to be summed up.
The contributions of mantissa quantization and overflow lead to the same expressions as in the simplified model.

3. Conclusion

The comparison of theory and measurement show an excellent aggreement. In Fig. 1 two examples for uniform and sinusoidal signals utilizing denormalized mantissas in the underflow region are shown in theory and measurement.
Hardware designers for low cost floating point hardware may use this results. The results are also of great interest in coding and storing of high quality signals like broad band music signals.

References

[1] : F. Hartwig, A. Lacroix, "Analysis of Floating Point Quantization Errors Using Stochastic Models", Proc. European Signal Processing Conf., pp. 247-250, Brussels 1992.

[2] : T. Kaneko, B. Liu, "On Local Roundoff Errors in Floating Point Arithmetics", Jounal Assoc. Comp. 20, pp. 391-398 (1973).

[3] : A. Fettweis, " On Properties of Floating Point Roundoff Noise", IEEE Trans. on Acoustics, Speech an Signal Processing ASSP-22, pp. 149-151, 1974.

APPENDIX

SNR for gaussian distributed numbers and only normalized mantissas

$$SNR = \frac{\sigma_\xi^2}{\sigma_{e_{abs}}^2} = \frac{\frac{1}{2}}{T_1 + T_2 + T_3}$$

T_1 represents the underflow part and is given by:

$$T_1 = \mathrm{errf}(\xi_{min}/\sigma_\xi) - \xi_{min} p_\xi(\xi_{min})$$

$p_\xi(x)$ dennotes the gaussian density function.The numerical computation of this quantity is critical for big signal power and has in this case to be replaced by:

$$T_1 = \frac{1}{\sqrt{(2\pi)}\sigma_\xi^3}\left(\frac{\xi_{min}^3}{3} - \frac{\xi_{min}^5}{10\,\sigma_\xi^2} + \frac{\xi_{min}^7}{56\,\sigma_\xi^2}\right) \quad ,$$

which is derived from a series expansion of $p_\xi(x)$.

The two other parts are given by:

$$T_2 = \frac{2^{-2t}}{8\ln 2}[-\xi_{max} p_\xi(\xi_{max}) + \mathrm{errf}(\xi_{max}/\sigma_\xi) + \xi_{min} p_\xi(\xi_{min}) - \mathrm{errf}(\xi_{min}/\sigma_\xi)]$$

$$T_3 = \left(1 + \frac{\xi_{max}^2}{\sigma_\xi^2}\right)\left(\frac{1}{2} - \mathrm{errf}(\xi_{max}/\sigma_\xi)\right) - \xi_{max} p_\xi(\xi_{max})$$

SNR for uniform distributed signal (improved model, normalized mantissas)

The situation for signal which is uniformly distributed between $-\xi_d$ and ξ_d splits up into three different cases:

$\xi_d < \xi_{min}$: SNR = 1

$\xi_{min} < \xi_d < \xi_{max}$: The absolute error power is given by:

$$\frac{1}{3\xi_d}\left(\xi_{min}^3 + \frac{2^{-2t}}{56}(2^{3\beta(\xi_d)} - 2^{3\beta(\xi_{min})}) + \frac{\xi_d - 2^{\beta(\xi_d)-1}}{4} 2^{2(\beta(\xi_d)-t)}\right)$$

$\xi_d > \xi_{max}$: The absolute error power is given by

$$\frac{1}{3\xi_d}\left(\xi_{min}^3+\frac{2^{-2t}}{56}(2^{3\beta(\xi_{max})}-2^{3\beta(\xi_{min})})+\xi_d^3-\xi_{max}^3-3\xi_{max}\xi_d^2+3\xi_{max}^2\xi_d\right)$$

SNR for uniform distributed signal (improved model, denormalized mantissas)

The situation for signal which is uniformly distributed between $-\xi_d$ and ξ_d splits up into three different cases:

$\xi_d < \xi_{min}2^{-t+1}$: SNR = 1

$\xi_{min}\,2^{-t+1} < \xi_d < \xi_{max}$: The absolute error power is given by

$$\frac{1}{3\xi_d}\left(\xi_{min}^3 2^{-3(t-1)}+\xi_{min}^2 2^{-2t}(\xi_d-\xi_{min}2^{-t+1})\right)$$

$\xi_{min} < \xi_d < \xi_{max}$:

$$\frac{1}{3\xi_d}[\xi_{min}^3 2^{-3(t-1)}+\xi_{min}^2 2^{-2t}(\xi_{min}-\xi_{min}2^{-t+1}) + .\frac{2^{-2t}}{56}(2^{3\beta(\xi_d)}-2^{3\beta(\xi_{min})})+\frac{\xi_d-2^{\beta(\xi_d)-1}}{4}2^{2(\beta(\xi_d)-t)}]$$

$\xi_d > \xi_{max}$:

$$\frac{1}{3\xi_d}[\xi_{min}^3 2^{-3(t-1)}+\xi_{min}^2 2^{-2t}(\xi_{min}-\xi_{min}2^{-t+1}) + .\frac{2^{-2t}}{56}(2^{3\beta(\xi_{max})}-2^{3\beta(\xi_{min})})+\xi_d^3-\xi_{max}^3-3\xi_{max}\xi_d^2+3\xi_{max}^2\xi_d]$$

Sinusoidal Signal (denormalized mantissas, improved model)

a = amplitude of sinusoid
t = mantissa wordlength
$\beta(x)$ = exponent of x

$a < \xi_{min}\,2^{-t+1}$: SNR = 1;

$2^{-t+1}\xi_{min} < a < \xi_{max}$:

The absolute error power is given by:

$$\frac{2}{\pi}\left[\frac{a^2}{2}\arcsin\left(\frac{\xi_{min}2^{-t+1}}{a}\right)-\xi_{min}2^{-t}\sqrt{a^2-\xi_{min}^2 2^{-2t+2}}+\frac{\xi_{min}^2 2^{-2t}}{3}\left(\frac{\pi}{2}-\arcsin(\frac{\xi_{min}2^{-t+1}}{a})\right)\right]$$

$\xi_{min} < a < \xi_{max}$:

The absolute error power is given by:

$$\frac{2}{\pi}\left[\frac{a^2}{2}\arcsin\left(\frac{\xi_{min}2^{-t+1}}{a}\right)-\xi_{min}2^{-t}\sqrt{a^2-\xi_{min}^2 2^{-2t+2}}+\frac{\xi_{min}^2 2^{-2t}}{3}\left(\arcsin\frac{\xi_{min}}{a}-\arcsin\frac{\xi_{min}2^{-t+1}}{a}\right)+\sum_{i=\beta(\xi_{min})}^{\beta(a)}\frac{2^{2(i-t)}}{12}\left(\arcsin\frac{2^i}{a}-\arcsin\frac{2^{i+1}}{a}\right)+\frac{2^{2(\beta(a)-t)}}{12}\left(\frac{\pi}{2}-\arcsin\frac{2^{\beta(a)-1}}{a}\right)\right]$$

$a > \xi_{max}$:

The absolute error power is given by:

$$\frac{2}{\pi}\left[\frac{a^2}{2}\arcsin\left(\frac{\xi_{min}2^{-t+1}}{a}\right)-\xi_{min}2^{-t}\sqrt{a^2-\xi_{min}^2 2^{-2t+2}}+\frac{\xi_{min}^2 2^{-2t}}{3}\left(\arcsin\frac{\xi_{min}}{a}-\arcsin\frac{\xi_{min}2^{-t+1}}{a}\right)+\sum_{i=\beta(\xi_{min})}^{\beta(\xi_{max})}\frac{2^{2(i-t)}}{12}\left(\arcsin\frac{2^i}{a}-\arcsin\frac{2^{i+1}}{a}\right)+\left(\frac{a^2}{2}+\xi_{max}^2\right)\left(\frac{\pi}{2}-\arcsin\frac{\xi_{max}}{s}\right)-\frac{3}{2}\xi_{max}\sqrt{a^2-\xi_{max}^2}\right]$$

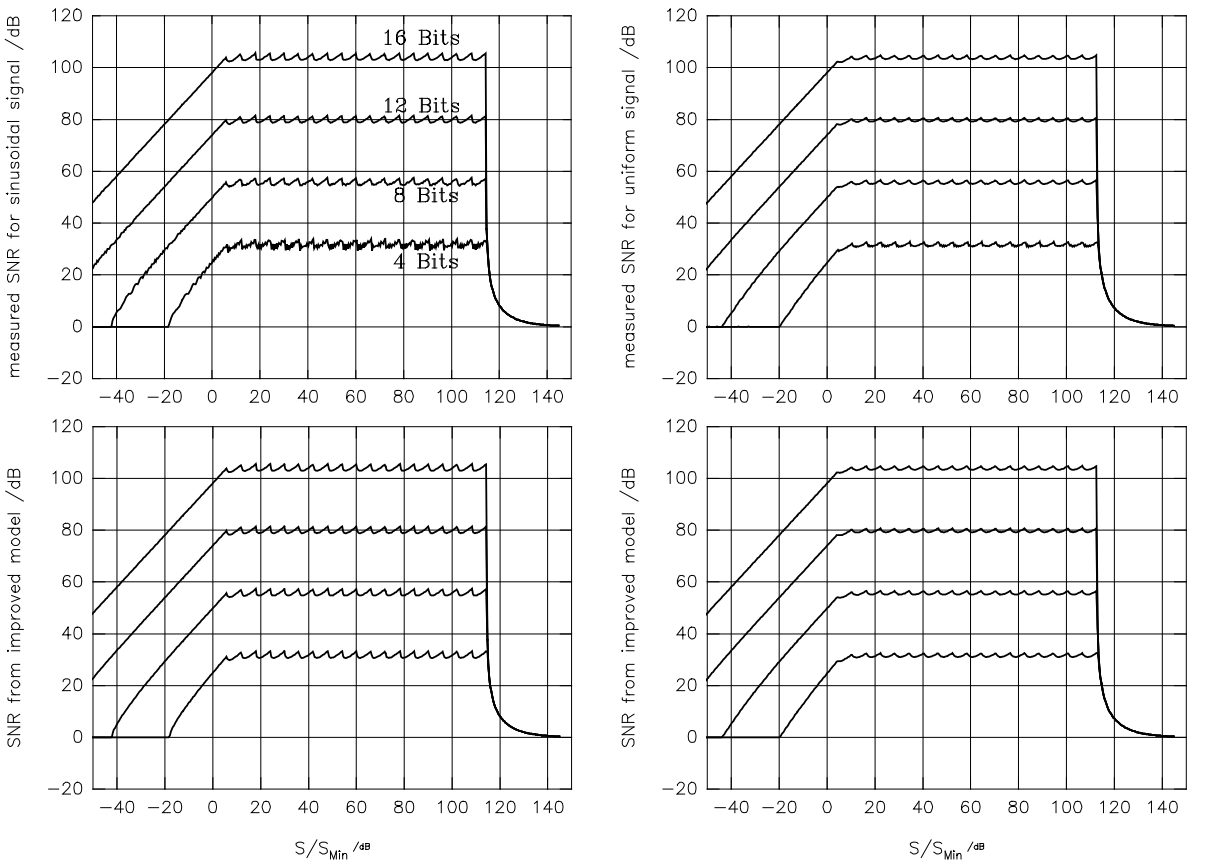

Fig. 1: SNR vs. S/S_{min} for sinusoidal and uniform signal utilizing denormalized mantissas in underflow