A describing function study of saturated quantization and its application to the stability analysis of Multi-Bit Sigma Delta modulators

Pieter Rombouts, Maarten De Bock, Jeroen De Maeyer and Ludo Weyten

This document is an author's draft version submitted for publication to IEEE Trans. Circ. Syst.-I.
The actual version was published as [1].

References

[1] P. Rombouts, M. De Bock, J. De Maeyer, and L. Weyten, "A Describing Function Study of Saturated Quantization and Its Application to the Stability Analysis of Multi-Bit Sigma Delta Modulators," *IEEE Trans. Circuits Syst.-I: Regular Papers*, vol. 60, no. 7, pp. 1740–1752, Jul. 2013.

A describing function study of saturated quantization and its application to the stability analysis of Multi-Bit Sigma Delta modulators

Pieter Rombouts, Maarten De Bock, Jeroen De Maeyer and Ludo Weyten

***Abstract*—Just as their single-bit counterparts, multi-bit sigma delta modulators exhibit nonlinear behavior due to the presence of the quantizer in the loop. In the multi-bit case this is caused by the fact that any quantizer has a limited output range and hence gives an implicit saturation effect. Due to this, any multi-bit modulator is prone to modulator overloading. Unfortunately, until now, designers had to rely on extensive time-domain simulations to predict the overloading level, because there is no adequate analytical theory to model this effect.**

In this work, we have developed such an analytical theory based on multiple input describing function analysis. This way, we obtained expressions for the signal gain, the noise gain and the variance of the quantization noise. Here, both the case of DC as well as sinusoidal signals was considered. These results were used for the stability analysis of multi-bit Sigma Delta modulators, which allows to predict the overloading level. Code implementing the proposed expressions is available for download at http://cas1.elis.ugent.be/cas/en/download/

I. Introduction

Sigma delta modulators are widely used for A/D-conversion and D/A-conversion. Such a modulator consists of a feedback loop with a linear filter, a quantizer and a feedback DAC. Originally most Sigma Delta modulators used a 2-level (single-bit) quantizer. As such the resulting system is heavily non-linear and difficult to analyze exactly. In particular the prediction of the stability is tricky and mathematically sophisticated [1]–[9]. It turns out that all single-bit sigma delta modulators become unstable when the magnitude of the input signal is to large. This phenomenon is called modulator overloading and the signal for which this occurs is called the overloading level. For the case of single-bit sigma delta modulation quite a lot of literature is devoted to stability analysis and attempts to predict the overloading level. A quite solid attempt in this field is the describing function theory of [4].

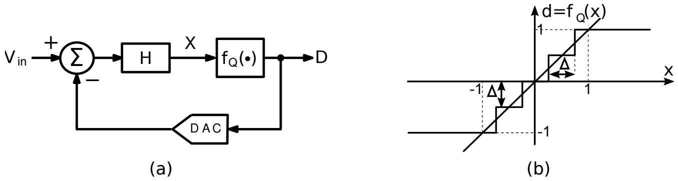

Fig. 1. (a) Block diagram of a MB ΣΔ modulator with (b) the input-output behavior of its nonlinear quantizers.

In the case of multi-bit modulators the analysis is considered to be much simpler, because it is commonly believed that multi-bit quantization can accurately be modeled by a linear gain and additive (white) quantization noise. This way, most literature on the stability of multi-bit sigma delta modulation focused on the robustness against parasitic effects [10]–[13]. However in practice, any quantizer has only a limited input (and output) range. As a result, any quantizer saturates. This way, every practical multi-bit modulator is non-linear as well and will also have a limited overloading level [14]. Moreover this overloading level will be smaller for more aggressive modulators. Surprisingly, very little literature about this effect in multi-bit modulators has been published. One notable exception is [15], [16, p. 104] that provides an iron-clad lower bound on the overloading level. Apparently independently, this bound was re-invented in [17], [18]. However the bound is not very tight and still lengthy time-domain simulations are needed to estimate the actual overloading level. In this work, we will apply the multiple-input describing function theory [19] to a saturating quantizer to obtain a more accurate analytical prediction of the quantizer behavior. Although the theory is still an approximation, we will see that it predicts the modulator behavior (including overloading level) considerably more accurate than [15]–[18], which were the only available analytical expression prior to this work. This way, the resulting expressions can be used within automated synthesis tools such as [20], [21].

The rest of this paper is structured as follows: in section II, we will review multi-bit sigma delta modulation. In section III, we will lift the multi-bit quantizer out of the sigma delta modulation loop and we will apply the describing function theory to the multi-bit quantizer alone. We will make a distinction between the case of DC and sinusoidal input signals. In section IV we will plug the multi-bit quantizer back into the sigma delta modulation loop and use the results of section III to analyze the multi-bit sigma delta modulator as a whole. In section V we will compare the proposed approach with prior art. Finally, we will present conclusions in section VI.

Pieter Rombouts, Maarten De Bock and Ludo Weyten (Pieter.Rombouts@ugent.be, Maarten.DeBock@UGent.be and Ludo.Weyten@UGent.be) are with the Department of Electronics and Information Systems (ELIS) of Ghent University, Belgium. Jeroen De Maeyer is with the Department of Electrical Energy, Systems and Automation of Ghent University, Belgium

This work is supported by the Fund for Scientific Research (F.W.O.-Vlaanderen)

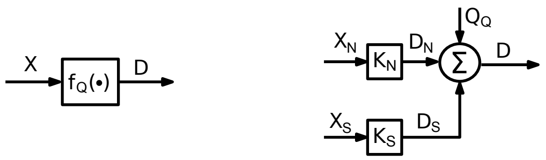

Fig. 2. Dual-input describing functions of a saturating quantizer.

II. Multi-bit Sigma Delta modulation

Fig. 1(a) shows a standard Multi-bit Sigma Delta modulator. It consists of a linear filter and a nonlinear quantization block. The input-output relationship[1] $d = f_Q(x)$ is show in Fig. 1(b). Mathematically this can be written as:

$$f_Q(x) = -1 + \sum_{i=1}^{n} \Delta \; u(x - \frac{\Delta}{2} - i\Delta), \qquad (1)$$

where Δ denotes the quantization step and n the number of discontinuities in the quantization function. With this definition of n, the number of quantization levels n_{lev} equals $n{+}1$. In a practical unit-element DAC implementation n would correspond to the number of unit-elements. The function $u(x)$ is the standard step function:

$$\begin{array}{rcl} u(x) & = & 0 \;\; \forall\, x < 0 \\ & & 1 \;\; else \end{array}$$

Since the quantizer of Fig. 1(b) is normalized to have a gain of 1 and a full scale input range of $[-1:1]$, this implies that:

$$\Delta = \frac{2}{n} \qquad (2)$$

Eq. (1) is cast in such a form that the odd symmetry of $f_Q(x)$ is hidden. To overcome this we can use the alternative notation:

$$\begin{aligned} f_Q(x) &= u(x) \sum_{j=0}^{\frac{n}{2}-1} \Delta u(x - \frac{\Delta}{2} j\Delta) \\ &\quad -u(-x) \sum_{j=0}^{\frac{n}{2}-1} \Delta u(-x - \frac{\Delta}{2} j\Delta). \qquad (3) \end{aligned}$$

This equation is only valid in the case where n is even (a so-called “mid-thread” quantizer). The equation for the case where n is odd (a so-called “mid-rise” quantizer) is similar. Without loss of generality we are going to restrict us here to the case where n is even.

III. Dual-input Describing function analysis of a saturating quantizer

It is clear that the quantization function $f_Q(x)$ of Fig. 1(b) is a hard nonlinear block. Such systems can be analyzed by a describing function analysis. In this analysis it is assumed that the input $x(t)$ of the non-linear function consists of several orthogonal components. In our analysis we will assume that there are 2 components: a signal component $x_S(t)$ and a Gaussian noise component $x_N(t)$. In principle the noise spectrum can be anything, as long as it is orthogonal to the signal component. Thus we have:

$$X = X_S + X_N \qquad (4)$$

Then we will associate a best-fit linear gain (K_N and K_S respectively) to each of these components (X_N and X_S respectively). Obviously in this linearization process we make a linearization error Q_Q. This linearization error is defined as:

$$q_Q = f_Q(x) - (K_S x_S + K_N X_N) \qquad (5)$$

Now the best-fit gains are determined in such a way that the variance σ_Q^2 of the linearization error is minimized. Or:

$$\begin{aligned} \sigma_Q^2 &= E\{(K_n x_N + K_S x_S - f_Q(x_N + x_S))^2\} \quad (6) \\ \frac{\delta(\sigma_Q^2)}{\delta K_N} &= 0 \Rightarrow K_N = \frac{E\{x_N f_Q(x_N + x_S)\}}{\sigma_N^2} \quad (7) \\ \frac{\delta(\sigma_Q^2)}{\delta K_S} &= 0 \Rightarrow K_S = \frac{E\{x_S f_Q(x_N + x_S)\}}{\sigma_S^2} \quad (8) \end{aligned}$$

Here, $E\{\cdot\}$ stands for the expectation operation and σ_N^2 and σ_S^2 correspond to the noise and signal variance respectively.

A. DC signal

A first important case to consider is the case of a DC-signal. Here $x_S = x_{DC}$. Then Eq. (7) can easily be evaluated using Eq. (1):

$$\begin{aligned} K_N &= \frac{1}{\sigma_N^2} \int_{-\infty}^{\infty} x_N f_Q(x_N + x_{DC}) p(x_N) dx_N \\ &= \frac{\Delta}{\sqrt{2\pi}\sigma_N} \sum_{i=1}^{n} e^{-(\frac{(-1-\frac{\Delta}{2}+i\Delta - x_{DC})^2}{2\sigma_N^2})} \qquad (9) \end{aligned}$$

Here $p(x_N)$ is the Gaussian probability density function (pdf) of the noise. Also Eq. (8) can be evaluated:

$$\begin{aligned} K_S &= \frac{1}{x_{DC}} \int_{-\infty}^{\infty} f_Q(x_N + x_{DC}) p(x_N) dx_N \\ &= \frac{\Delta}{2x_{DC}} \sum_{i=1}^{n} \mathrm{erf}\left(\frac{1 + \frac{\Delta}{2} + x_{DC} - i\Delta}{\sqrt{2}\sigma_N} \right) \quad (10) \end{aligned}$$

Here $\mathrm{erf}(\cdot)$ is the error function. Finally the variance σ_Q^2 of the linearization error can be evaluated as well:

$$\sigma_Q^2 = E(f_Q(x_N + x_{DC}))^2 - K_N^2 \sigma_N^2 - K_S^2 x_{DC}^2 \qquad (11)$$

This can be cast in closed form as:

$$\begin{aligned} \sigma_Q^2 &= \frac{\Delta^2}{2} \sum_{j=0}^{\frac{n}{2}-1} (2j+1) \left(2 - \mathrm{erf}\left(\frac{-x_{DC} + \frac{\Delta}{2} + j\Delta}{\sqrt{2}\sigma_N} \right) \right) + \\ &\quad \frac{\Delta^2}{2} \sum_{j=0}^{\frac{n}{2}-1} (2j+1) \left(\mathrm{erf}\left(\frac{-x_{DC} - \frac{\Delta}{2} - j\Delta}{\sqrt{2}\sigma_N} \right) \right) \\ &\quad -K_N^2 \sigma_N^2 - K_S^2 x_{DC}^2 \qquad (12) \end{aligned}$$

It is important to realize that the best fit gains K_N, K_S and the variance σ_Q^2 of the linearization error [Eqs. (9), (10) and (12)] all depend both on the magnitude of the quantizer input signal X_{DC} and on the variance σ_N^2 of the quantizer input noise. To illustrate this, some results for the case of

[1] Here the convention is used that lowercase letters are used for an explicit time-domain representation of a signal, while uppercase letters are used for a more abstract (potentially Z-domain or frequency domain) representation.

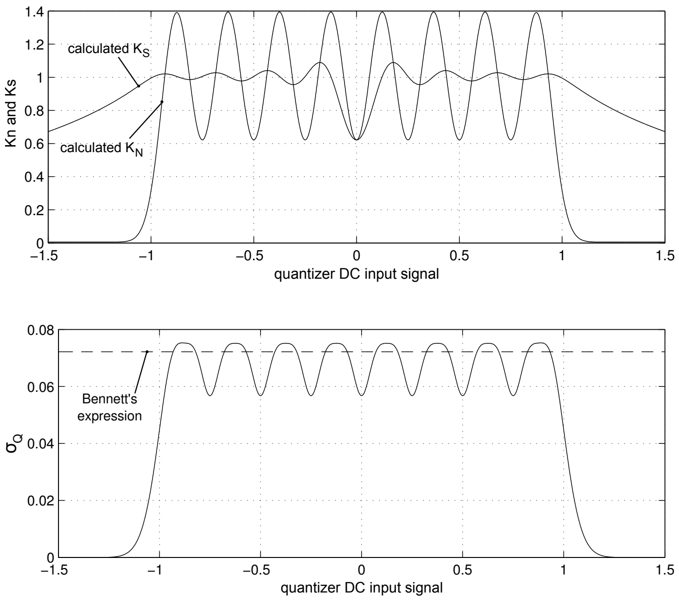

Fig. 3. (a) Best fit gains K_S and K_N and (b) the RMS value σ_Q of the linearization error vs. the magnitude of the DC signal for the case of a 9-level quantizer ($n = 8$) with a Gaussian noise RMS value $\sigma_N = \frac{\Delta}{\sqrt{12}}$.

$n = 8$ (a 9-level quantizer) are shown in Fig. 3. Here the input noise component x_N is arbitrarily assigned an RMS value σ_N equal to $\frac{\Delta}{\sqrt{12}}$. The results are plotted vs. the magnitude of the DC signal X_{DC}. The top figure shows the linearized gains K_N and K_S. It is clear that both gains are not equal. Moreover, as expected for a nonlinear block, both gains are signal dependent. In the quantizer's valid input range (the interval [-1:1]) both gains exhibit a ripple around the nominal gain of 1. Outside the valid input range the noise gain K_N rapidly drops. Fig. 3(b) shows the RMS value of the linearization error. According to Bennett's widely used approximation [22], this value should be equal to $\frac{\Delta}{\sqrt{12}}$, which is also indicated in the figure. It is clear that the actual value ripples around Bennett's approximation.

The results for the case where the RMS value of the Gaussian input noise is doubled compared to Fig. 3 (i.e. $\sigma_N = \frac{2\Delta}{\sqrt{12}}$) are shown in Fig. 4. It is clear that the traditional approximations (i.e. the gains are equal to 1 and the RMS value is equal to Bennett's expression) are a much better approximation, which is due to the increased dithering effect. However outside the valid input range, again the noise gain K_N rapidly drops.

B. Sinusoidal signal

The second important case is the case of a sinusoidal signal, where $x_s(t)$ consists of a sine wave. In this case Eq. (7) becomes:

$$K_N = \frac{1}{\sigma_N^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_N f_Q(x_N + x_S) p(x_N) q(x_S) dx_S dx_N, \tag{13}$$

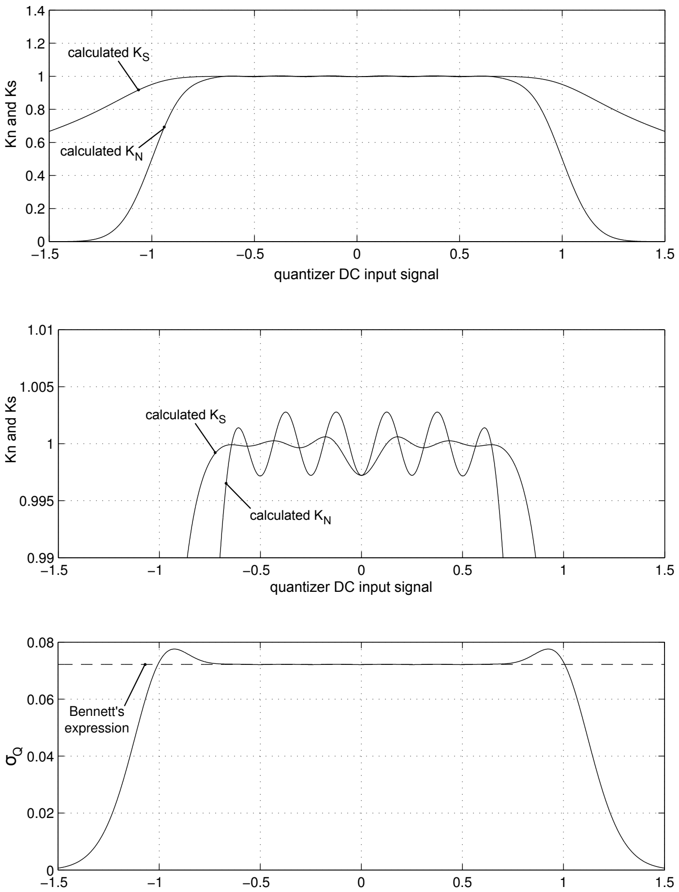

Fig. 4. (a) Best fit gains K_S and K_N with (b) a close in view and (c) the RMS value σ_Q of the linearization error vs. the magnitude of the DC signal for the case of a 9-level quantizer ($n = 8$) with a Gaussian noise RMS value $\sigma_N = \frac{2\Delta}{\sqrt{12}}$.

where $q(x_S)$ corresponds to the pdf of a sine wave with amplitude A:

$$\begin{aligned} q(x_S) &= \frac{1}{\pi\sqrt{A^2 - x_S^2}} \;\; \forall\, x_S \;\; with \;\; A^2 > x_S^2 \\ &= 0 \;\; else \end{aligned} \tag{14}$$

Filling this in into Eq. (13) and combining with Eq. (1):

$$K_N = \sum_{i=1}^{n}\int_{-A}^{A} \frac{\sigma \mathrm{e}^{-\frac{(x_S+1+\frac{\Delta}{2}-i\Delta)^2}{2\sigma^2}}}{\pi\sqrt{2\pi}\sqrt{a^2 - x_S{}^2}} dx_S \tag{15}$$

We were unable to find an analytical solution of the above integrals, but it is easy to obtain an efficient numerical evaluation e.g. by the transformation in the appendix. This way, K_N can be evaluated for any given n and σ_N. Hence, we can say that $K_{N,sin}(\sigma_N, n)$ is a known function:

$$K_N = K_{N,sin}(\sigma_N, n). \tag{16}$$

We can also evaluate Eq. (8) for the case of a sinusoidal signal:

$$K_S \;\; = \;\; \frac{2}{A^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_S f_Q(x_N + x_S) p(x_N) q(x_S) dx_S dx_N$$

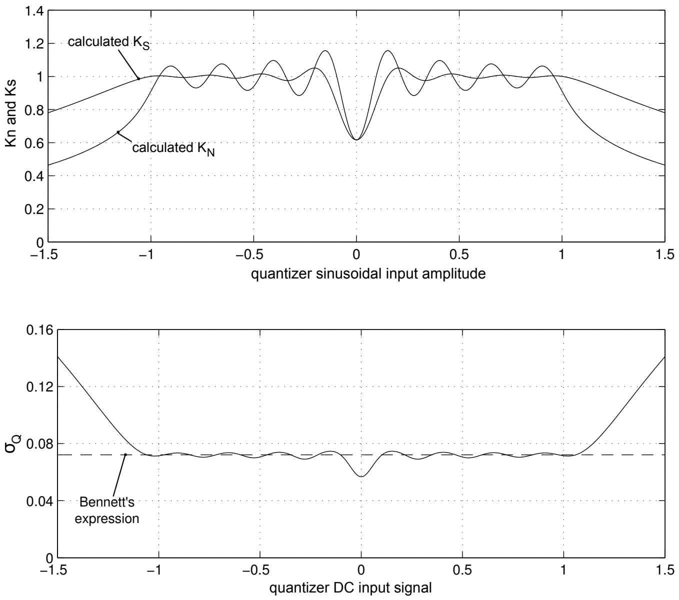

Fig. 5. (a) Best fit gains K_S and K_N and (b) the RMS value σ_Q of the linearization error vs. the amplitude of the sinusoidal signal for the case of a 9-level quantizer ($n = 8$) with a Gaussian noise RMS value $\sigma_N = \frac{\Delta}{\sqrt{12}}$.

$$= \sum_{i=1}^{n}\int_{-A}^{A}\frac{\left(\text{erf}\left(\frac{\left(x_S+1+\frac{\Delta}{2}-i\Delta\right)}{\sqrt{2}\sigma_N}\right)+1\right)x_S dx_S}{A^2\pi\sqrt{A^2-{x_S}^2}} \qquad (17)$$

Again we were unable to find an analytical solution of the above integrals, but also here it is easy to obtain an efficient numerical evaluation (see e.g. the appendix). This way, K_S can be evaluated for any given n and σ_N and we can say that $K_{S,sin}(\sigma_N, n)$ is a known function:

$$K_S = K_{S,sin}(\sigma_N, n). \qquad (18)$$

Finally we evaluate the variance σ_Q^2 of the linearization error for the sinusoidal signal case as follows:

$$\begin{aligned}\sigma_Q^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(f_Q(x_N+x_S))^2 p(x_n)q(x_s)dx_n dx_s \\ &\quad -K_N^2\sigma_N^2 - K_S^2\frac{A^2}{2} \qquad (19)\end{aligned}$$

This can be rewritten as:

$$\begin{aligned}&\sigma_Q^2 = \\ &\Delta^2\sum_{j=0}^{\frac{n}{2}-1}(2j+1)\int_{-A}^{A}\frac{\left(\text{erf}\left(\frac{(x_s-\Delta/2-j\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)}{2\pi\sqrt{A^2-{x_s}^2}}dx_s \\ &+\Delta^2\sum_{j=0}^{\frac{n}{2}-1}(2j+1)\int_{-A}^{A}\frac{\left(\text{erf}\left(\frac{(-x_s-\Delta/2-j\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)}{2\pi\sqrt{A^2-{x_s}^2}}dx_s \\ &-K_N^2\sigma_N^2 - K_S^2\frac{A^2}{2} \qquad (20)\end{aligned}$$

Again we had to evaluate the integrals numerically, but it turns out that this can be done efficiently.

Just as for the DC-case, the best fit gains K_N, K_S and the variance σ_Q^2 of the linearization error all depend both on the amplitude of the quantizer input signal A_X and on the variance σ_N^2 of the quantizer input noise. The results for the same situation as Fig. 3 (i.e. $n = 8$ and $\sigma_N = \frac{\Delta}{\sqrt{12}}$) are shown in Fig. 5. The top figure shows the linearized gains K_N and K_S. Here qualitatively the results are very similar to the case of a DC signal: i.e. both gains are not equal and signal dependent. In the quantizer's valid input range (the interval [-1:1]) both gains exhibit a ripple around the nominal gain of 1. Outside the valid input range the noise gain K_N drops, but not as rapidly as in the case of a DC input signal. Fig. 5(b) shows the RMS value of the linearization error. Also here, the actual value ripples around Bennett's approximation in the quantizer's valid input range. However, here the linearization error rapidly increases for overloading amplitudes, unlike the case of a DC-signal where the σ_Q goes to zero for large signals.

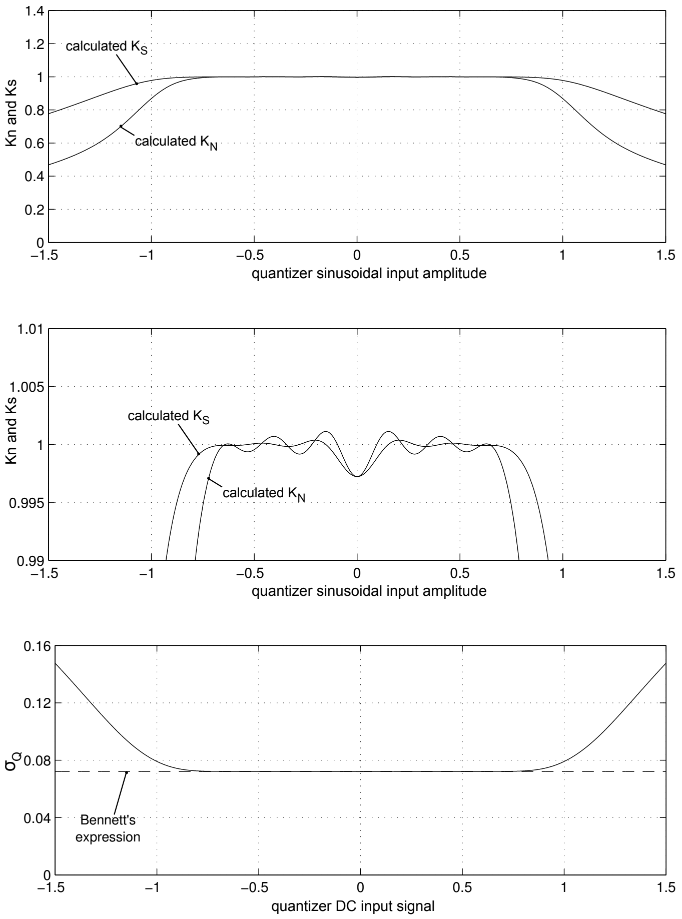

Fig. 6. (a) Best fit gains K_S and K_N with (b) a close in view and (c) the RMS value σ_Q of the linearization error vs. the magnitude of the DC signal for the case of a 9-level quantizer ($n = 8$) with a Gaussian noise RMS value $\sigma_N = \frac{2\Delta}{\sqrt{12}}$.

The results for the case where the RMS value of the Gaussian noise is doubled compared to Fig. 5 (i.e. $\sigma_N = \frac{2\Delta}{\sqrt{12}}$) are shown in Fig. 6. It is clear that the traditional approximations (i.e. the gains are equal to 1 and the RMS value is equal to Bennett's expression) are a much better approximation, which is due to the increased dithering effect. However outside the valid input range, again the noise gain K_N drops and the magnitude of the linearization error σ_Q rapidly increases.

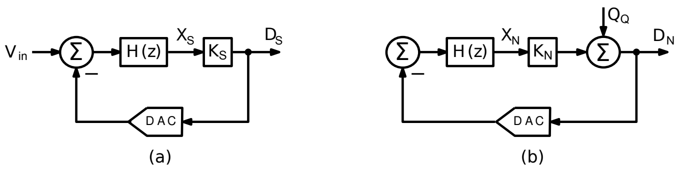

Fig. 7. Equivalent describing function model of a MB $\Sigma\Delta$ modulator: (a) system diagram corresponding to the signal component and (b) system diagram corresponding to the noise component.

IV. Application to a Multi-bit Sigma Delta Modulator

As illustrated in Figs. 3–6, the best-fit gains K_N and K_S of the quantizer are signal dependent. As a result, if such a quantizer is embedded inside a feedback loop, the linearized loop gain (and hence also the closed-loop poles) will be signal dependent. This way, there will be signal values for which the loop becomes unstable [14]. We will now quantify this effect.

In the theory elaborated in the previous section we did not yet make an approximation. This is due to the definition of the linearization error q_Q of Eq. (5) which conceals the nonlinearity. But it still is present since q_Q is a nonlinear function of the noise and the signal. To apply the above theory to a multi-bit Sigma Delta modulator we will now make two approximations. The first approximation is that q_Q is a noisy signal with a Gaussian pdf (the second approximation will come later). This approximation allows to draw the equivalent describing function model of Fig. 7. Here, the actual system is separated into two separated systems each dealing with one component (either the signal component or the noise component). In the system corresponding to the noise component, the noise is caused by the linearization error Q_Q, which we assumed to have a Gaussian pdf. As a result it will give rise to signals with a Gaussian pdf in the loop. In particular the signal X_N will have a Gaussian pdf as well and hence will exhibit the noise gain K_N.

To set up the equivalence with the conventional Multi-bit modulator of Fig. 1, we have:

$$D = D_S + D_N \tag{21}$$
$$X = X_S + X_N \tag{22}$$
$$Q_Q = D - K_N X_N - K_S X_S \tag{23}$$

Note that both systems are not completely decoupled because K_N depends on the magnitude of X_S and K_S depends on the variance of X_N. From the figure we readily obtain that:

$$D = \underbrace{\frac{H(z)K_S}{1+H(z)K_S}}_{STF(K_S)} V_{in} + \underbrace{\frac{1}{1+H(z)K_N}}_{NTF(K_N)} Q_Q \tag{24}$$
$$X = \frac{H(z)}{1+H(z)K_S} V_{in} - \frac{H(z)}{1+H(z)K_N} Q_Q \tag{25}$$
$$Q_Q = D - K_N X_N - K_S X_S \tag{26}$$

A. DC-input signal

Let us now further elaborate this for the case where V_{in} is a DC-input signal:

$$V_{in} = C \tag{27}$$

```
// initialization of the iterative procedure with the values
// of the case where no saturation occurs
$K_N \leftarrow 1$ ; $K_S \leftarrow 1$ ; $\sigma_Q \leftarrow \frac{2}{n\sqrt{(12)}}$;
// the actual iterative procedure
while (NOT accurate enough)
    $X_{DC} \leftarrow \frac{C}{K_S}$; // according to Eq. (29)
    $\sigma_N^2 \leftarrow$ evaluate Eq. (31);
    $K_{N,new} \leftarrow$ evaluate Eq. (9);
    $K_{S,new} \leftarrow$ evaluate Eq. (10);
    $K_N \leftarrow K_{N,new}$; $K_S \leftarrow K_{S,new}$;
    $\sigma_{Q,new} \leftarrow$ evaluate Eq. (12);
    $\sigma_Q \leftarrow \sigma_{Q,new}$;
```

Fig. 8. Iterative algorithm to find K_N, K_S, σ_Q, σ_N and X_{DC} for a DC input signal.

If we assume that the loop gain is infinite for DC (i.e. the loop filter contains at least one integration), then the DC-component in the output signal must be equal to $D_S = C$. In this way:

$$X_N = \frac{-H(z)}{1+H(z)K_N} Q_Q \tag{28}$$
$$X_S = \frac{C}{K_S} = X_{DC} \tag{29}$$
$$Q_Q = D - K_N X_N - K_S X_{DC} \tag{30}$$

Here the variance σ_Q^2 of Q_Q is defined by Eq. (12) and K_N and K_S are defined by Eqs. (9) and (10), which all depend on the variance σ_N^2 of the noise signal X_N. In principle Eq. (28) can be used to evaluate σ_N^2. To simplify this, we will make our second approximation: i.e. we will make the common assumption that the spectrum of the linearization error Q_Q is white. This way, we obtain the following relationship:

$$\sigma_N^2 = \frac{\sigma_Q^2}{f_s/2} \int_0^{\frac{f_s}{2}} \left| \frac{-H(e^{j2\pi\frac{f}{f_S}})}{1+H(e^{j2\pi\frac{f}{f_S}})K_N} \right| df, \tag{31}$$

where f_s stands for the sampling frequency.

Now we have 5 equations, i.e. Eqs. (9), (10), (12), (29) and (31) that fully define the 5 unknown quantities i.e.: K_N, K_S, σ_Q, σ_N and X_{DC} as a function of the known loop filter H and the modulator DC input signal C. We did not manage to find an analytical solution, but an iterative algorithm to obtain a numerical solution is quite straightforward and is illustrated in Fig. 8.

This procedure was applied to a typical 3rd-order modulator designed according to [23], [24] with $h_\infty = 4$, for the case of $n = 8$ (a 9-level quantizer). The results are shown in Fig. 9, where K_N, K_S, σ_Q and σ_N are shown vs. the overall modulator input level C. In addition to the calculated values, also the experimental results are shown. Here, for each value of the modulator input signal C an extensive time domain simulation was performed to determine the corresponding quantizer's input sequence x as well as the output sequence d. Then the experimental value of X_{DC} can be estimated as the average of the x-sequence and K_S as the ratio of X_{DC} and the average of the d-sequence D_S. Then we obtain the noise sequence x_N by removing the DC-component from the x-sequence. By calculating the variance of the x_N-sequence, we

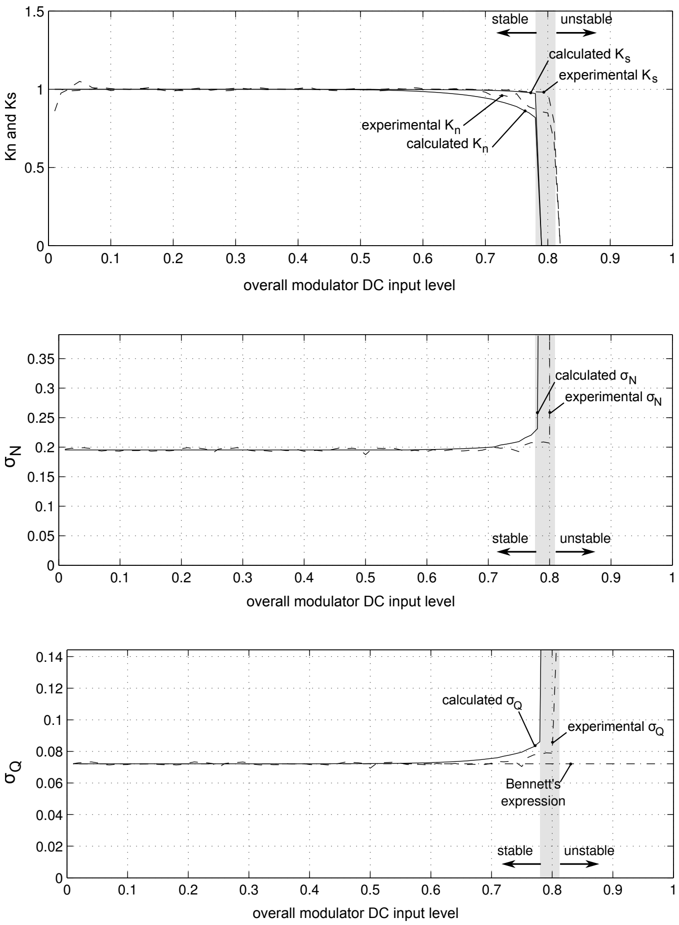

Fig. 9. Describing function results for a maximally flat 3rd-order modulator with $h_\infty = 4$ for the case of a 9-level quantizer ($n = 8$) vs. the modulator DC input level: (a) Best fit gains K_S and K_N with (b) the RMS noise level σ_N in front of the quantizer and (c) the RMS value σ_Q of the linearization error.

obtain the experimental value of σ_N. Finally, the experimental value of the noise gain K_N is determined by minimizing the variance of the sequence $d - K_N x_N$ with respect to K_N. We simply implemented this by sweeping K_N and selecting the value which corresponds to the minimum. The variance of the corresponding $(d - K_N x_N)$-sequence provides the experimental value of σ_Q. For the experimental plot C was swept with a sweeping step of 0.01. For each value of C, 32 simulation runs of 2^{19} clock cycles were averaged. Each simulation run was launched with a randomly dithered initial state of the modulator state variables. The total time to execute the matlab script to generate the experimental plot was 1hour 16minutes on 1 core of an Intel Core2 Q9400 CPU (2.66GHz), whereas the theoretical plot took less than 1 second.

From the plot we observe the well known fact that the modulator is stable and has well defined quantizer gains and noise levels for sufficiently small input signals. Moreover the commonly accepted white noise expression turns out to be very accurate for small input signals. For larger input signals the modulator becomes unstable. Moreover the transition (both according to the theory as well as according to the experiment) is very steep. It is also clear that the theory matches the experiment reasonably well, although the prediction of the overloading level is somewhat pessimistic in this case.

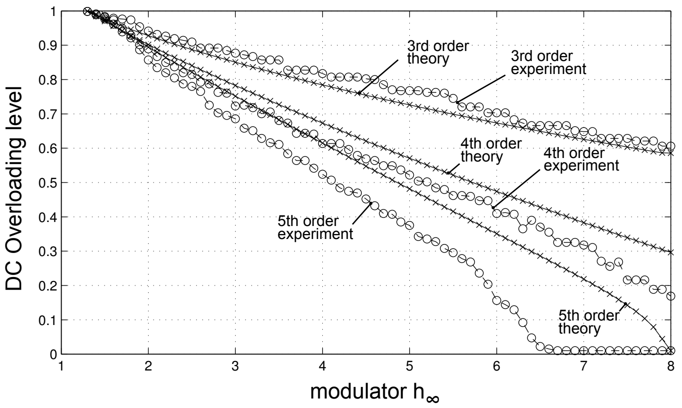

Fig. 10. Describing function prediction of the DC overloading level (marked with "×") as well as the simulated value (marked with "o") vs. h_∞ results for maximally flat modulators of order 3,4 and 5 for the case of a 9-level quantizer ($n = 8$).

Fig. 10 shows the describing function prediction of the DC overloading level for maximally flat modulators designed according to [23], [24] vs. h_∞. The cases of 3rd-order, 4th-order and 5th-order modulators are considered. Also the corresponding experimental plots based on time domain simulation are shown. Here, a modulator was considered to be stable for a given input level if the state variables remain bounded in each of the 32 simulation runs of 2^{19} clock cycles. For this case the CPU time on the same machine was 1hr42min. From the figure it is clear that, the theory matches the experiment very well for the case of the 3rd-order modulators. For the 4th- and 5th-order modulators the theory matches the experiment qualitatively, but the quantitative matching is somewhat less good. Additionally, it is observed that the correspondence between the theory and the experiment becomes less good for more aggressive modulators (with higher h_∞).

B. Sinusoidal input signal

Let us now consider the case where V_{in} is a sinusoidal signal with amplitude A_{in}. Obviously, the signal component in the output signal is also sinusoidal and given by $D_s = STF(K_s)V_{in}$. For the elaboration of the theory we will assume that the output signal level is known. In many cases, this is readily fulfilled because, we are usually interested in signals in the signal band, where the loop gain is so large that in practice the signal transfer function is very close to unity[2]. This way, we will assume that $D_S = V_{in}$ and hence:

$$X_N = \frac{-H(z)}{1 + H(z)K_N} Q_Q \qquad (32)$$

$$X_S = \frac{D_s}{K_S} = \frac{V_{in}}{K_S} \qquad (33)$$

$$Q_Q = D - K_N X_N - K_S X_S \qquad (34)$$

Here the variance σ_Q^2 of Q_Q is defined by Eq. (20) and K_N and K_S are defined by Eqs. (15) and (17), which all depend

[2] If the signal transfer function is not close to unity, the corresponding input level has to be corrected by a factor $1/STF(K_s)$, where K_s is the corresponding describing function gain.

```
// initialization of the iterative procedure with the values
// of the case where no saturation occurs
$K_N \leftarrow 1$ ; $K_S \leftarrow 1$ ; $\sigma_Q \leftarrow \frac{2}{n\sqrt{(12)}}$;
// the actual iterative procedure
while (NOT accurate enough)
    $A_X$          $\leftarrow \frac{A_{in}}{K_S}$; // according to Eq. (33)
    $\sigma_N^2$          $\leftarrow$ evaluate Eq. (31);
    $K_{N,new}$ $\leftarrow$ evaluate Eq. (15);
    $K_{S,new}$ $\leftarrow$ evaluate Eq. (17);
    $K_N \leftarrow K_{N,new}$; $K_S \leftarrow K_{S,new}$;
    $\sigma_{Q,new}$ $\leftarrow$ evaluate Eq. (20);
    $\sigma_Q$          $\leftarrow \sigma_{Q,new}$;
```

Fig. 11. Iterative algorithm to find K_N, K_S, σ_Q, σ_N and A_X for a sinusoidal input signal.

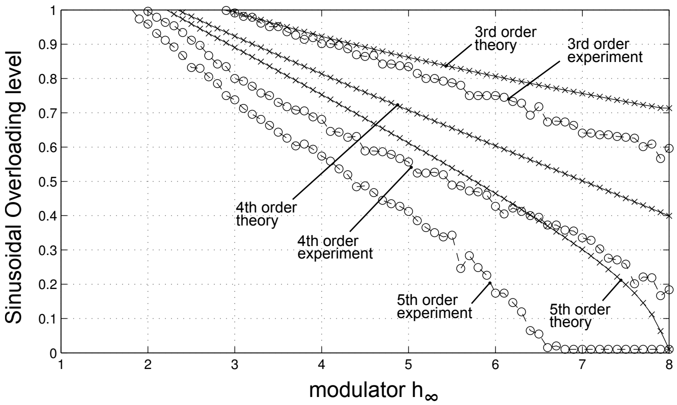

Fig. 12. Describing function prediction of the sinusoidal overloading level (marked with "×") as well as the simulated value (marked with "o") vs. h_∞ results for maximally flat modulators of order 3,4 and 5.

on the variance σ_N^2 of the noise signal X_N. Now the rest of the analysis is very similar to the case of a DC input signal. To simplify the evaluation of σ_N^2, we will again make the approximation that the spectrum of the linearization error Q_Q is white. This way, we obtain again Eq. (31).

Now we have 5 equations, i.e. Eqs. (15), (17), (20), (33) and (31) that fully define the 5 unknown quantities i.e.: K_N, K_S, σ_Q, σ_N and X_S as a function of the known loop filter H and the modulator input signal amplitude A_{in}. To find the solution, we had to resort again to an iterative algorithm (illustrated in Fig. 11) to obtain a numerical solution.

To verify the theory, the same experiments as for the DC-case were performed. Some of the corresponding results are shown in Fig. 12, where the describing function prediction of the overloading level as well as the corresponding experimental plots based on time domain simulation are shown. Unfortunately, the correspondence of the theory and the simulation experiment is only very moderate. Moreover, it was found that the experimental results depend on the input frequency, especially if the input frequency is low compared to the clock frequency. This phenomenon is not predicted by the theory. The plot in the figure is for a relatively low input frequency around $f_S/84$ (i.e. a typical sigma delta modulator input frequency). When the plot is done for even lower input frequencies (say $f_S/1000$), the experimental overloading levels completely collapse to the same values as for the DC case. This effect as well as the relatively poor matching between our theory and the experiment, can be explained by the fact that we have made one additional approximation in the case of sinusoidal signals compared to the DC case: i.e. we have implicitly assumed that a sinusoid can be modeled by a random signal with the same pdf as the sinusoidal signal [Eq. (14)]. This way, correlation between successive samples is neglected and the describing function prediction becomes independent of the input frequency. However in an actual sinusoidal signal, successive signal values are of course heavily correlated: e.g. near the top of a relatively low-frequency sinusoid (as in a Sigma Delta modulator), there are many successive occurrences of high input values. As a result the describing function result for a sinusoidal input signal is optimistic, and the actual overloading level is much closer to the overloading level for a DC input signal. Therefore, for a practical prediction of the overloading level we advise to use the describing function result for the DC-case.

V. Comparison with prior art

Before this work, the only analytical prediction of the overloading level of a multi-bit sigma delta modulator was based on Kenney and Carley's work [15] which is quantified in the lower bound on the overloading level of [16, p. 104]. When using our notation convention this lower bound can be written as:

$$OL \geq 1 - \frac{\Delta}{2}(||NTF||_1 - 2) \tag{35}$$

Note that this bound does not make a distinction between the type of signal (sinusoid vs. DC or whatever). Also, this expression is not an estimation but instead a lower bound.

This bound is shown in Fig. 13 for the case of $n = 8$, together with the results of Figs. 10 and 12, which are now re-arranged per modulator order. For the reasons explained above, the describing function results for the sinusoidal case are omitted and only the describing function predictions for the DC-case are shown.

When comparing Kenney and Carley's bound [Eq. (35)] with the describing function result, it is clear that the describing function prediction is considerably more accurate. Moreover in practice only modulators that *work*, i.e. with a sufficiently large overloading level (say above -6dB) are useful. Therefore it makes sense to limit the comparison to these modulators. In this case, the describing function prediction matches the experiment within 0.7 dB for 3rd order, within 0.9 dB for 4th order and within 1.5 dB for 5th order modulators. In comparison, in the same situation, Kenney and Carley's bound [Eq. (35)] has an error of 7.3 dB for 3rd order, 6.5 dB for 4th order and 6.1 dB for 5th order modulators.

Fig. 14 shows the comparison of the describing function result and Eq. (35) for the case of the lowest quantizer resolution where this theory still makes sense, i.e. 3 quantizer levels ($n = 2$). For the sinusoidal experiment, again a frequency of about $f_S/84$ was used. Surprisingly, for some of the 4th and 5th order modulators, the experimental overloading level for sinusoidal input signals is lower than that for a DC input

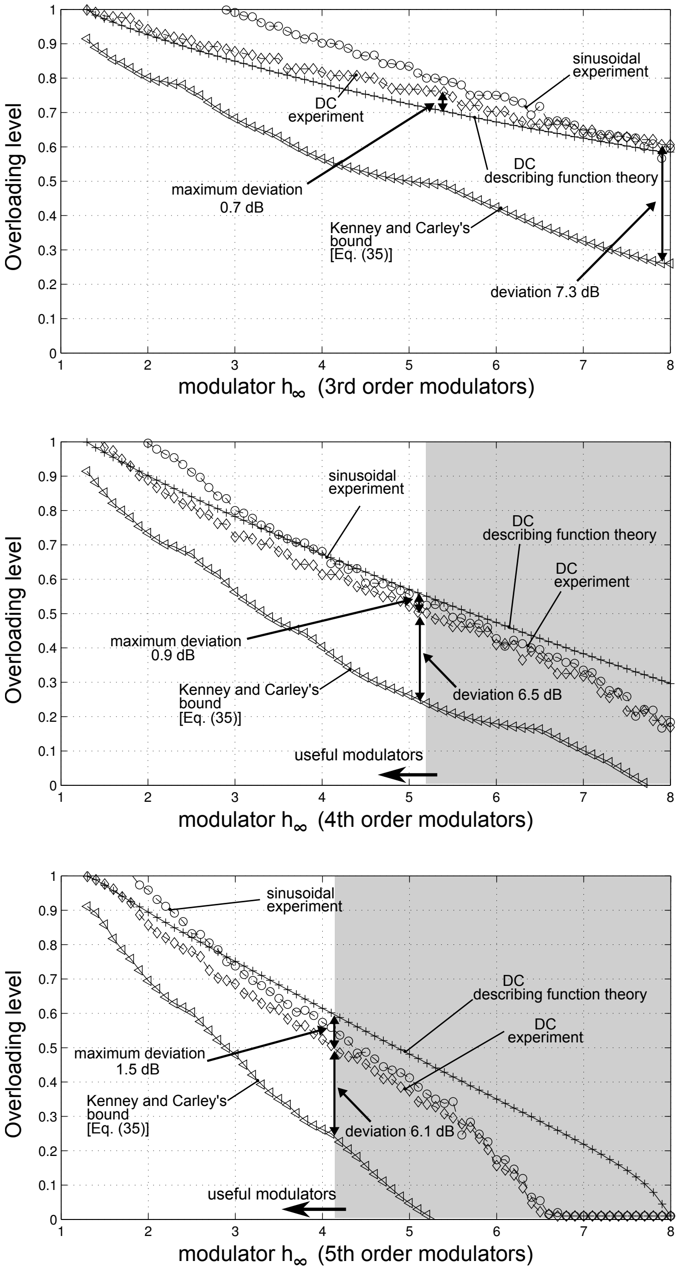

Fig. 13. Modulator overloading level vs. h_∞ results for maximally flat modulators with a 9-level quantizer ($n = 8$): describing function prediction (marked with "$\times$") and experimental value (marked with "$\circ$") for the sinusoidal case, describing function prediction (marked with "$+$")and experimental value (marked with "$\diamond$") for the DC case and Kenney and Carley's bound [15], [16] (marked with "$\triangleleft$").

signal. This is counter-intuitive and not predicted by the theory (note that unity STF's were used in the experiment). Again, the experimental curve for sinusoidal input signals was found to be dependent on the input frequency and converges toward that for a DC input signal, when the input frequency is decreased.

From the figure, it is clear that Eq. (35), does not at all manage to give an adequate prediction of the overloading level in this case and estimates a zero overloading level for modulators that, in simulation, have an overloading level well above -6dB. If we restrict ourself again to modulators with an overloading level of at least -6 dB, our describing function prediction is within 0.5 dB for 3rd order, within 1.7 dB for 4th order and within 2.2 dB for 5th order modulators.

Fig. 15 shows the case of a higher quantizer resolution (17 quantizer levels, $n = 16$). Now, all considered modulators have an overloading level that is larger than -6dB. In this case, Kenney and Carley's bound is considerably closer to the experimental curve than in the lower-resolution cases (within 2.2 dB for 3rd and 4th order, within 3.3 dB for the 5th order case). Also the accuracy of the describing function estimation for a DC signal is improved compared to lower resolutions and is now within 0.2 dB for the 3rd and 4th order case and within 1.0 dB for the 5th order case.

VI. Conclusion

Every multi-bit quantizer has a finite input voltage range and hence is actually a saturating quantizer. When such a multi-bit quantizer is used inside a sigma delta loop, this saturation is the cause of its finite overloading level. To analyze this phenomenon quantitatively, we have applied dual-input describing function theory to such a multi-bit quantizer (with saturation). We have obtained analytical expressions for the case of DC signals and nearly analytical (easy to evaluate) expressions for the case of sinusoidal signals. Next, we have shown how these describing function results can be used to analyze multi-bit Sigma Delta modulators and to predict their overloading level. The results of the theory were compared to the results of time domain simulations. For the case of DC-signals, it was found that the matching between theory and experiment was quite good. For the case of sinusoidal signals the quantitative matching was less good. This is attributed to the fact that the theory does not take correlation between successive signal values into account. This way, for practical overloading level predictions it seems advisable to use the describing function results for the DC case. Here, the proposed approach is considerably more accurate than the prior art [15], [16].

Appendix

The numerical evaluation of Eqs. (15), (17) and (20) (through an approximation as a finite sum) is awkward due to the singularity at the edges of the integrandum caused by the factor $\frac{1}{\sqrt{A^2-x_S{}^2}}$. To get rid of this we use the following substitution:

$$x_S = A\sin(\theta)$$

This will map the integrandum $x_S \in [-A; A]$ on $\theta \in [-\frac{\pi}{2}; \frac{\pi}{2}]$. The singular factor will be removed because:

$$\frac{dx_S}{\sqrt{A^2-x_S{}^2}} = Ad\theta \quad \forall x_S \in [-A; A]$$

This way, the integrals in Eq. (15) are transformed into:

$$\int_{-A}^{A} \frac{\sigma \mathrm{e}^{-\frac{(x_S+1+\frac{\Delta}{2}-i\Delta)^2}{2\sigma^2}}}{\pi\sqrt{2\pi}\sqrt{A^2-x_S{}^2}}dx_S = \int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} \frac{\sigma \mathrm{e}^{-\frac{(A\sin(\theta)+1+\frac{\Delta}{2}-i\Delta)^2}{2\sigma^2}}}{\pi\sqrt{2\pi}}d\theta$$

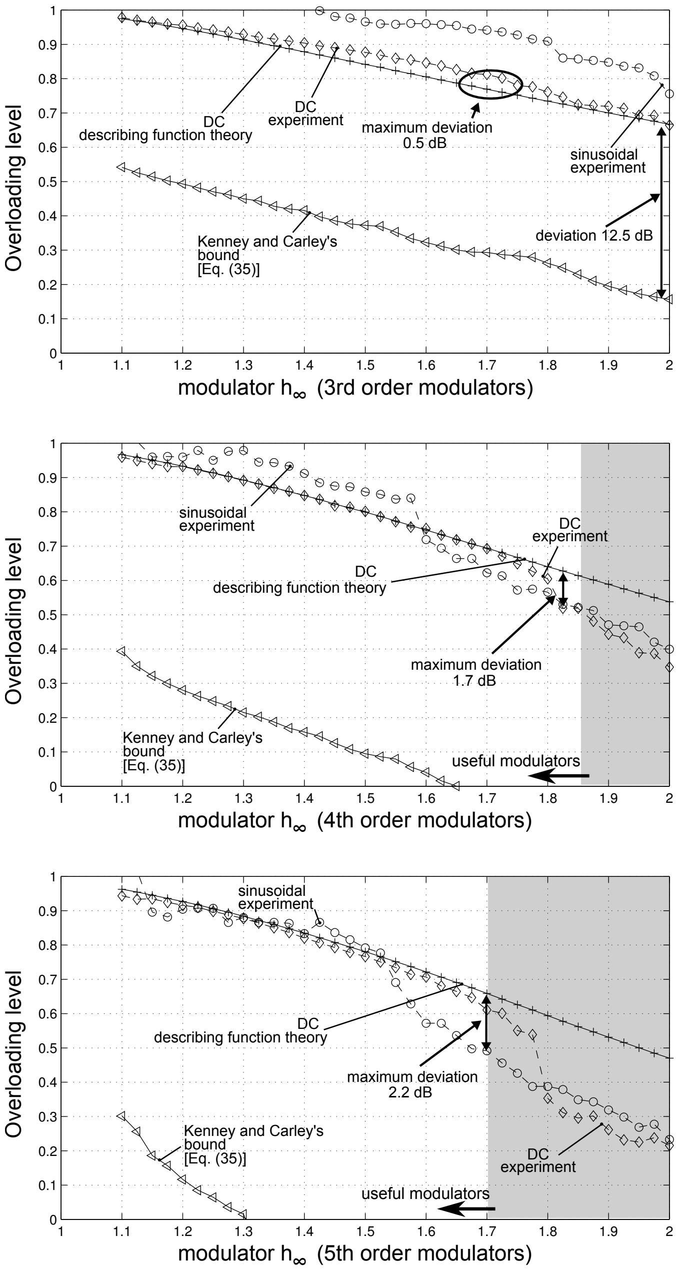

Fig. 14. Modulator overloading level vs. h_∞ results for maximally flat modulators with a 3-level quantizer ($n = 2$): describing function prediction (marked with "$\times$") and experimental value (marked with "$\circ$") for the sinusoidal case, describing function prediction (marked with "$+$")and experimental value (marked with "$\diamond$") for the DC case and Kenney and Carley's bound [15], [16] (marked with "$\triangleleft$").

and those in Eq. (17) into:

$$\int_{-A}^{A} \frac{\left(\mathrm{erf}\left(\frac{(x_S+1+\frac{\Delta}{2}-i\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)x_S dx_S}{A^2\pi\sqrt{A^2-{x_S}^2}} =$$

$$\int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} \frac{\left(\mathrm{erf}\left(\frac{(A\sin(\theta)+1+\frac{\Delta}{2}-i\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)\sin(\theta)d\theta}{A\pi}$$

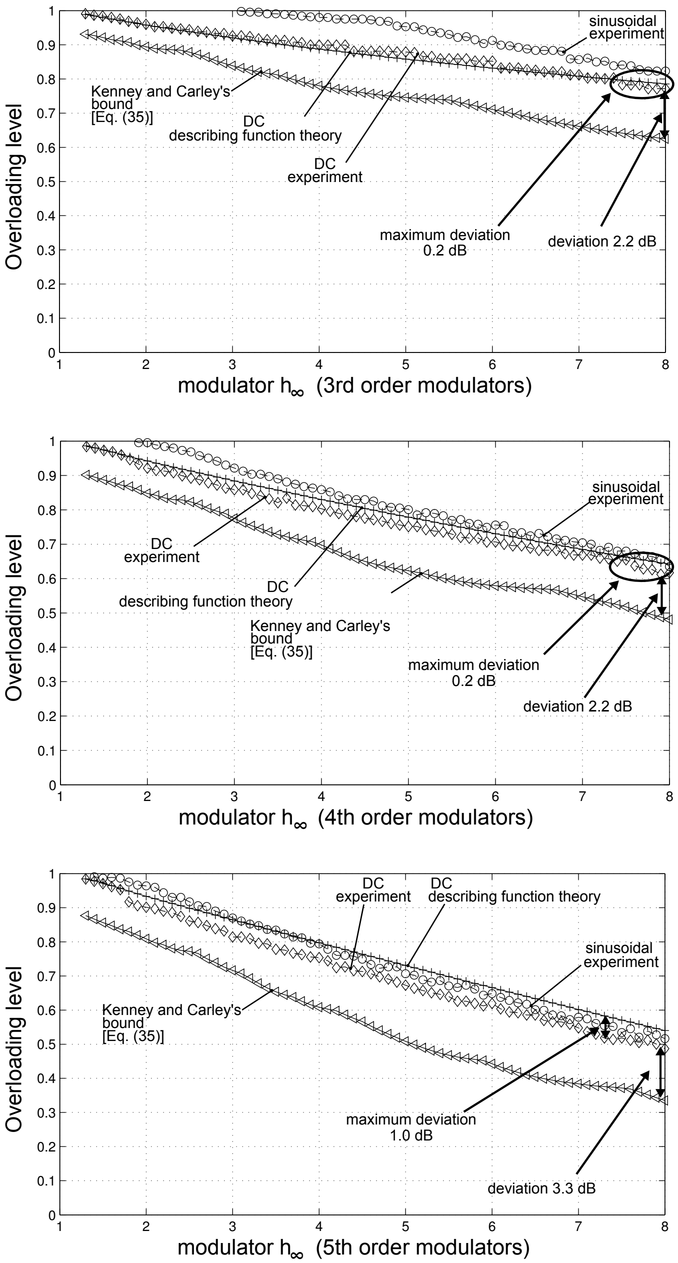

Fig. 15. Modulator overloading level vs. h_∞ results for maximally flat modulators with a 17-level quantizer ($n = 4$): describing function prediction (marked with "$\times$") and experimental value (marked with "$\circ$") for the sinusoidal case, describing function prediction (marked with "$+$")and experimental value (marked with "$\diamond$") for the DC case and Kenney and Carley's bound [15], [16] (marked with "$\triangleleft$").

and finally those in Eq. (20) into:

$$\int_{-A}^{A}\frac{\left(\mathrm{erf}\left(\frac{(x_s-\Delta/2-j\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)}{2\pi\sqrt{A^2-{x_s}^2}}dx_s =$$

$$\int_{\frac{-\pi}{2}}^{\frac{\pi}{2}}\frac{\left(\mathrm{erf}\left(\frac{(A\sin(\theta)-\Delta/2-j\Delta)}{\sqrt{2}\sigma_N}\right)+1\right)}{2\pi}d\theta$$

All these integrals can efficiently be evaluated by discretizing the integrandum in a few 100 data points.

References

[1] S. Hein and A. Zakhor, "On the stability of sigma-delta modulators," *IEEE Trans. Sign. Proc.*, vol. 41, no. 7, pp. 2322–2348, JUL 1993.

[2] E. Stikvoort, "Some remarks on the stability and performance of the noise shaper or sigma-delta-modulator," *IEEE Trans. Communications*, vol. 36, no. 10, pp. 1157–1162, OCT 1988.

[3] R. Gray, W. Chou, and P. Wong, "Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs," *IEEE Trans. Communications*, vol. 37, no. 9, pp. 956–968, Sep 1989.

[4] S. Ardalan and J. Paulos, "An analysis of nonlinear behavior in delta-sigma modulators," *IEEE Trans. Circuits Syst.*, vol. 34, no. 6, pp. 593–603, Jun 1987.

[5] J. Lota, M. Al-Janabi, and I. Kale, "Nonlinear-stability analysis of higher order Delta-Sigma modulators for DC and sinusoidal inputs," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 3, pp. 530–542, Mar 2008.

[6] R. Farrell and O. Feely, "Bounding the integrator outputs of second-order sigma-delta modulators," *IEEE Trans. Circuits Syst.-II*, vol. 45, no. 6, pp. 691–702, JUN 1998.

[7] R. Schreier, M. Goodson, and B. Zhang, "An algorithm for computing convex positively invariant sets for delta-sigma modulators," *IEEE Trans. Circuits Syst.-I*, vol. 44, no. 1, pp. 38–44, Jan 1997.

[8] J.-M. Liu, S.-H. Chien, and T.-H. Kuo, "Optimal design for delta sigma modulators with root loci inside unit circle," *IEEE Trans. Circuits Syst.-II*, vol. 59, no. 2, pp. 83 –87, feb. 2012.

[9] G. Bourdopoulos, A. Pnevmatikakis, and T. Deliyannis, "Numerical method for determining the quantization error PDF of single-bit Sigma Delta modulators," *IEEE Trans. Circuits Syst.-I*, vol. 51, no. 4, pp. 718–731, Apr 2004.

[10] M. Keller, A. Buhmann, J. Sauerbrey, M. Ortmanns, and Y. Manoli, "A comparative study on excess-loop-delay compensation techniques for continuous-time sigma-delta modulators," *IEEE Trans. Circuits Syst.-I*, vol. 55, no. 11, pp. 3480–3487, 2008.

[11] B. De Vuyst, P. Rombouts, J. De Maeyer, and G. Gielen, "The Nyquist Criterion: A Useful Tool for the Robust Design of Continuous-Time $\Sigma\Delta$ Modulators," *IEEE Trans. Circuits Syst.-II*, vol. 57, no. 6, pp. 416 –420, june 2010.

[12] M. Ranjbar and O. Oliaei, "A Multibit Dual-Feedback CT Delta Sigma Modulator With Lowpass Signal Transfer Function," *IEEE Trans. Circuits Syst.-I*, vol. 58, no. 9, pp. 2083–2095, SEP 2011.

[13] V. Singh, N. Krishnapura, and S. Pavan, "Compensating for Quantizer Delay in Excess of One Clock Cycle in Continuous-Time Delta Sigma Modulators ," *IEEE Trans. Circuits Syst.-II*, vol. 57, no. 9, pp. 676–680, SEP 2010.

[14] R. Baird and T. Fiez, "Stability analysis of high-order delta-sigma modulation for ADCs," *IEEE Trans. Circuits Syst.-II*, vol. 41, no. 1, pp. 59–62, JAN 1994.

[15] J. Kenney and L. Carley, "Design of multibit noise-shaping data converters," *Analog Integr. Circuits Process.*, vol. 3, no. 3, pp. 259–272, May 1993.

[16] R. Schreier and G. C. Temes, *Understanding Delta-Sigma Data Converters*. IEEE, 2005.

[17] I. Lokken, A. Vinje, T. Saether, and B. Hernes, "Quantizer nonoverload criteria in sigma-delta modulators," *IEEE Trans. Circuits Syst.-II*, vol. 53, no. 12, pp. 1383–1387, DEC 2006.

[18] P. Kiss, J. Arias, D. Li, and V. Boccuzzi, "Stable high-order delta-sigma digital-to-analog converters," *IEEE Trans. Circuits Syst.-I*, vol. 51, no. 1, pp. 200–205, JAN 2004.

[19] A. Gelb and W. E. Vander Velde, *Multiple-Input Describing Functions and Nonlinear System Design*. McGraw-Hill, 1968.

[20] M. Keller, A. Buhmann, M. Ortmanns, and Y. Manoli, "Systematic approach to the synthesis of continuous-time cascaded sigma-delta modulators," *Analog Integr. Circuits Process.*, vol. 60, no. 1-2, pp. 155–164, AUG 2009.

[21] B. De Vuyst, P. Rombouts, and G. Gielen, "A Rigorous Approach to the Robust Design of Continuous-Time Sigma Delta Modulators," *IEEE Trans. Circuits Syst.-I*, vol. 58, no. 12, pp. 2829–2837, Dec 2011.

[22] W. Bennett, "Spectra of quantized Signals," *Bell Syst. Tech. J*, vol. 27, no. , pp. 446–473, Jul. 1948.

[23] R. Schreier, "An empirical study of high-order single-bit delta-sigma modulators," *IEEE Trans. Circuits Syst.-II*, vol. 40, no. 8, pp. 461 –466, aug 1993.

[24] ——, "The $\Sigma\Delta$ Toolbox."