ADC Post-Correction Using Limited Resolution Correction Values

Henrik Lundin, Mikael Skoglund and Peter Händel

Signal Processing, KTH Electrical Engineering
Royal Institute of Technology, SE-100 44 Stockholm, Sweden
Telephone: +46 8 790 8462, Fax: +46 8 790 7260
E-mail: henrik.lundin@s3.kth.se, Web: http://www.s3.kth.se/~hlu

Abstract–Analog-to-digital converter additive post-correction using look-up-tables is considered. In a practical post-correction system, the correction values must be stored using limited (fixed-point) precision. In this paper the effects of limited precision in the correction values is investigated. A theory for approximating the SINAD after correction using fixed-point values is presented. The theory shows good agreement when compared with simulation results.

I. Background

Analog-to-digital converter (ADC) post-correction has been proposed in many different forms, many of these applying look-up tables (LUTs) [1]. In a practical post-correction application it is very likely that the correction values will be stored with fixed-point precision. However, most of the evaluations and experiments reported in the literature have been conducted with infinite (or floating-point) precision in the representation of the correction values stored in the LUT. One of few exceptions is [2], where experimental results indicated that the precision of the correction values strongly affect the outcome of the correction.

In this paper we present a theory for the relationship between the precision of the correction values and the resulting ADC performance after correction.

II. ADC and Correction System Model

The ADC to be corrected is assumed to consist of an ideal sample-and-hold circuit followed by a non-ideal quantizer. Thus, we can omit the sample-and-hold and perform a discrete-time analysis. We assume a static model for the quantizer. The mapping of a value s into a quantized value $x = \mathcal{Q}(s)$ is determined by a partition of the real numbers into disjoint sets $\{\mathcal{S}_j\}_{j=0}^{2^b-1}$; if $s \in \mathcal{S}_i$, then the quantized value $x = x_i$ is produced.

We assume that the input value $s(n)$ is drawn from a stochastic variable s with probability density function (PDF) $f_s(s)$. The temporal properties for s are immaterial since the quantizer is, for now, assumed to be non-dynamic, i.e., the output of the quantizer at time n depends only on the input at the same time. The MSE for the quantizer without correction is

$$\mathrm{MSE}_Q = \mathrm{E}[(s-x)^2] = \int (s-\mathcal{Q}(s))^2\, f_s(s)\, ds = \sum_i \int_{s\in\mathcal{S}_i} (s-x_i)^2\, f_s(s)\, ds. \tag{1}$$

Assume further that an additive correction is employed. The corrected value y is produced by adding a correction term $e(x)$ to the output x so that $y = x + e(x)$ as in Fig. 1. Every possible output value $\{x_j\}_{j=0}^{2^b-1}$ is associated with a correction term $\{e_j\}_{j=0}^{2^b-1}$.

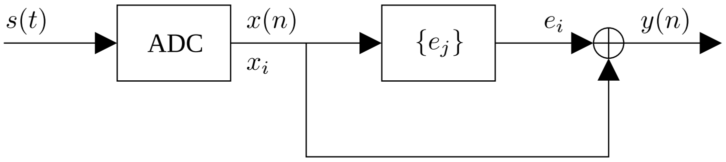

Fig. 1. Additive correction system.

A. Optimal Correction Values

Optimal correction values for minimizing the mean-square error $\mathrm{E}[(s-y)^2]$ have been derived in [3] ($\mathrm{E}[\cdot]$ is the expected value operator taken with respect to $s(n)$, noting that y is a function of s). In [3] a quantizer operating on a value s is considered; the quantization region model above is used. Just as above, the input is regarded to be drawn from a stochastic variable with a PDF $f_s(s)$. If the quantization regions $\{\mathcal{S}_j\}$ are assumed fixed, then it is proved that the optimal reconstruction values[1] $\{x_j\}$, in the mean-squared sense, are given by

$$y_{j,\,\mathrm{opt}} = \arg\min_{y} \mathrm{E}[(y-s)^2 | s \in \mathcal{S}_j] = \frac{\int_{s\in\mathcal{S}_j} s\, f_s(s)\, ds}{\int_{s\in\mathcal{S}_j} f_s(s)\, ds}, \tag{2}$$

i.e., the optimal reconstruction value for each region is the "center of mass" of the region. Hence, the optimal *correction* values are

$$e_j = y_{j,\,\mathrm{opt}} - x_j. \tag{3}$$

When representing the correction values with infinite precision, i.e., using the values (3), the resulting MSE after correction is

$$\begin{aligned} \mathrm{MSE}_o &= \mathrm{E}[(s-y)^2] = \mathrm{E}\left[(s-x-e(x))^2\right] = \mathrm{E}\left[(s-x)^2 - 2(s-x)\,e(x) + e(x)\right] \\ &= \mathrm{MSE}_Q + \mathrm{E}\left[e(\mathcal{Q}(s))^2\right] - 2\,\mathrm{E}\left[(s-\mathcal{Q}(s))\,e(\mathcal{Q}(s))\right], \end{aligned} \tag{4}$$

where the last equality comes from applying (1). In order to simplify the expression, we use (3) in the last term of the expression above. Thus,

$$\begin{aligned} \mathrm{E}[(s-\mathcal{Q}(s))\,e(\mathcal{Q}(s))] &= \int (s-\mathcal{Q}(s))\,e(\mathcal{Q}(s))\,f_s(s)\,ds = \sum_i \int_{s\in\mathcal{S}_i} (s-x_i)\,e_i\,f_s(s)\,ds \\ &= \sum_i \left(e_i \int_{s\in\mathcal{S}_i} s\,f_s(s)\,ds - x_i\,e_i \int_{s\in\mathcal{S}_i} f_s(s)\,ds \right) \\ &= \Bigg/ \text{from (3): } \int_{s\in\mathcal{S}_i} s\,f_s(s)\,ds = (e_i + x_i) \int_{s\in\mathcal{S}_i} f_s(s)\,ds \Bigg/ \\ &= \sum_i \left(e_i\,(e_i + x_i) \int_{s\in\mathcal{S}_i} f_s(s)\,ds - x_i\,e_i \int_{s\in\mathcal{S}_i} f_s(s)\,ds \right) \\ &= \sum_i e_i^2 \int_{s\in\mathcal{S}_i} f_s(s) ds = \mathrm{E}\left[e(\mathcal{Q}(s))^2\right], \end{aligned} \tag{5}$$

which is in fact nothing but the variance of the correction value e. Reapplying this in (4) yields

$$\mathrm{MSE}_o = \mathrm{MSE}_Q - \mathrm{E}\left[e(\mathcal{Q}(s))^2\right]. \tag{6}$$

We have now seen how an ideal, or infinite precision, correction affects the MSE of the quantizer. In the next section we will study the effects of fixed-point resolution for the correction values.

III. Fixed-Point Resolution for Correction Values

It is not always feasible, let alone practical, to implement an ADC post-correction system, such as the one in Fig. 1, using floating-point representation for the stored correction values $\{e_j\}_{j=0}^{2^b-1}$. It is natural to settle for a specific precision with which the correction terms are stored, e.g., a certain number of bits. Obviously, the performance of the corrected ADC will depend on which precision that is used.

The precision of digitally stored values is often stated as a number of bits. Assume that the table is stored using τ bits and that the ADC to be corrected converts the signal into b-bit values. If we know that the ADC only has error in the lower bits, then we can "shift" the bits of the correction table and obtain a correction with higher effective precision. For example, if the ADC has 10 bits, but only the 2 LSBs need correction, then the remaining bits of the correction values (minus the sign bit) can be used to get a better precision.

The problem gets easier to analyze if the resolution Δ, being the smallest possible difference between two different correction values, is used instead of the actual number of bits τ. The relationship between the two is straightforward: $\Delta = 2^{-\delta b}$ LSBs, where δb is the number of extra bits of precision the table adds to the ADC. See Fig. 2 for an illustration. It is assumed that the correction values never exceed the largest number that can be represented by the τ correction bits.

[1] In [3] the reconstruction values are denoted 'quanta'.

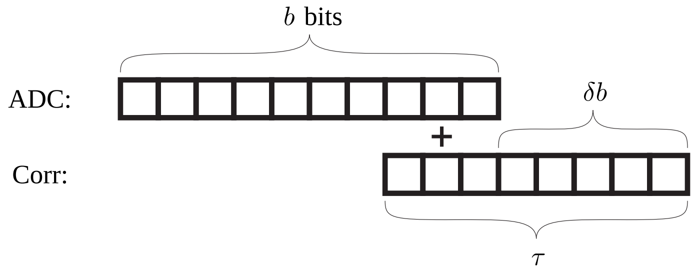

Fig. 2. Addition of the ADC output with a correction value. The bits of the table value are shifted in order to enhance the precision of the corrected ADC.

Let $\tilde{e}_i$ be the (uniformly) quantized version of the table entry e_i, i.e.,

$$\tilde{e}_i = \mathcal{Q}_\Delta(e_i). \tag{7}$$

The notation $\mathcal{Q}_\Delta$ is used to distinguish this quantization from the one performed in the converter. We assume that one of the quantized cells is centered at zero. That is, if a certain correction term is within the interval $[-\Delta/2,\, \Delta/2]$ it will be quantized to zero, and, since the quantization of correction terms is uniform, all other possible quantized values are located at an integer multiple of Δ. Hence, we can say that

$$\tilde{e}_i \in \{k\Delta :\; k = \ldots,\, -2,\, -1,\, 0,\, 1,\, 2,\, \ldots\}. \tag{8}$$

Also let $\delta_i = \tilde{e}_i - e_i$ be the error introduced by the quantization of the correction term. The notation $\delta(x)$ denotes the correction term quantization error associated with a specific x, i.e., $\delta(x) = \delta_i$ if $x = x_i$. Thus, $\delta(x)$ ultimately depends on the stochastic variable s.

A. MSE

The MSE obtained using the quantized correction terms becomes

$$\begin{aligned}\mathrm{MSE}_\Delta &= \mathrm{E}\left[(s - x - \tilde{e}(x))^2\right] = \mathrm{E}\left[(s - x - e(x) - \delta(x))^2\right]\\ &= \mathrm{E}\left[(s - x - e(x))^2\right] + \mathrm{E}[\delta(x)^2] - 2\,\mathrm{E}[(s - x - e(x))\,\delta(x)]\\ &= \mathrm{MSE}_o + \mathrm{E}[\delta(x)^2] - 2\,\mathrm{E}[(s - x - e(x))\delta(x)].\end{aligned} \tag{9}$$

Analyze the last term to find that

$$\begin{aligned}\mathrm{E}[(s - x - e(x))\,\delta(x)] &= \sum_i \int_{s\in\mathcal{S}_i} (s - x_i - e_i)\,\delta_i\, f_s(s)\, ds\\ &= \sum_i \delta_i \left(\int_{s\in\mathcal{S}_i} (s - x_i)\, f_s(s)\, ds - \int_{s\in\mathcal{S}_i} e_i\, f_s(s)\, ds \right) = 0,\end{aligned} \tag{10}$$

where, again, the relationship in (3) has been used in the last term. This results in that the MSE when using quantized correction terms is

$$\mathrm{MSE}_\Delta = \mathrm{MSE}_o + \mathrm{E}[\delta(x)^2]. \tag{11}$$

The error variance $\mathrm{E}[\delta(x)^2]$ can further be written as

$$\begin{aligned}\mathrm{E}[\delta(x)^2] &= \int \delta(x)^2\, f_s(s)\, ds = \sum_i \int_{s\in\mathcal{S}_i} \delta_i^2\, f_s(s)\, ds = \sum_i \int_{s\in\mathcal{S}_i} f_s(s)\, ds \frac{\int_{s\in\mathcal{S}_i} \delta_i^2\, f_s(s)\, ds}{\int_{s\in\mathcal{S}_i} f_s(s)\, ds}\\ &= \sum_i \int_{s\in\mathcal{S}_i} f_s(s)\, ds\, \mathrm{E}[\delta^2 | s \in \mathcal{S}_i].\end{aligned} \tag{12}$$

Under the assumption that the quantization error δ_i is uniformly distributed in $[-\Delta/2,\, \Delta/2]$, then each $\mathrm{E}[\delta^2 | s \in \mathcal{S}_i] = \Delta^2/12$ for all i, and (12) becomes

$$E[\delta(x)^2] = \frac{\Delta^2}{12} \sum_i \int_{s\in\mathcal{S}_i} f_s(s)\, ds = \frac{\Delta^2}{12}. \tag{13}$$

The MSE in (11) then boils down to

$$\mathrm{MSE}_\Delta = \mathrm{MSE}_o + \frac{\Delta^2}{12}. \tag{14}$$

It is reasonable to believe that the assumption is valid for small Δ, i.e., when the quantization is assumed to be "high-rate". (See [4] for a thorough discussion and precise conditions for the uniformity of the quantization noise.) However, as Δ grows large the assumption will become invalid, motivating the asymptotic analysis.

B. Asymptotic MSE

Recall that one of the quantization cells is centered at zero and that all table values e_i that fall within $[-\Delta/2, \Delta/2]$ will be quantized to $\tilde{e}_i = 0$. When we enlarge the quantization step, i.e. when $\Delta \to \infty$, all $\tilde{e}_i$ will inevitably be zero, since all table values will fall into the center region at zero. Consequently, the resulting MSE becomes

$$\mathrm{MSE}_\Delta = \mathrm{E}[(s - x - 0)^2] = \mathrm{E}[(s - x)^2] = \mathrm{MSE}_Q \tag{15}$$

when the resolution tends to zero.

C. Approximation and Lower Bound on MSE$_o$

The performance description provided in (14) above is dependent on MSE$_o$ – a quantity that is dependent on the actual transfer characteristics of the ADC under test, the accuracy of the calibration and correction schemes, and on the signal considered. A coarse approximation can be obtained through the following discussion.

We make these assumptions:

- The quantization step size T of the ADC is assumed to be small compared with the variability of the source PDF – i.e., the requirements for high-rate quantization are fulfilled;
- The actual quantization regions' deviation from an ideal uniform quantizer is small;
- The correction values are perfect in accordance with (3).

Then, we can say that the corrected ADC acts like a perfect uniform quantizer having the classical MSE

$$\mathrm{MSE}_{\text{uniform}} = \frac{T^2}{12}. \tag{16}$$

This can be used as an approximation of MSE$_o$ in (14).

Now, a counter-argument to the discussion above is that the ADC considered does in fact deviate from an ideal uniform quantizer – otherwise there would be no need for a post-correction – although we assumed a perfect uniform quantizer in the derivation. We can therefore see (16) as a practical lower bound on the MSE of a perfect correction. However, if we want to be rigourous we must consider the (unlikely) possibility that the quantization regions actually deviate from the uniform quantizer to a configuration which is *more beneficial for the considered test signal*. We will therefore resort to results from information theory to derive a true lower bound for the MSE of a perfectly corrected ADC.

From information theory (see e.g. [5]) we learn that the *rate-distortion function* tells us how small the resulting distortion can be when describing the outcomes of a certain random variable with a specific rate (resolution). The inverse *distortion-rate function* provides the reverse relation[2]. In numerous situations the rate-distortion function is inherently difficult to calculate, therefore, the *Shannon lower bound* on the rate-distortion function is frequently used. The lower bound has the advantage that it is often easier to compute.

If we are quantizing a random variable s and using a squared-error criterion (MSE), the Shannon lower bound is defined as

$$R_{\mathrm{SLB}}(D) = h(s) - \frac{1}{2}\log_2(2\pi e D), \tag{17}$$

where $h(s)$ is the differential entropy of s, D is the squared-error distortion and R_{SLB} is the rate in bits. The result says that it is impossible to represent a random variable s with less than R_{SLB} bits if the MSE should be no more than D.

Now, since (17) is a lower bound on the rate-distortion function and is strictly decreasing, the inverse of (17) – D as a function of R – is a lower bound on the distortion-rate function, which is of greater interest to us. We get

$$D = \frac{1}{2\pi e} 2^{2h(s) - 2R}. \tag{18}$$

The differential entropy $h(s)$ is a function of the distribution of s. Therefore we cannot say anything more about the lower bound before we choose a PDF for s. In this case, we let s be a sample function

[2] The distortion-rate function is the inverse of the rate-distortion function whenever the latter is strictly decreasing.

of a sinusoid with amplitude A because it is the predominant test signal in ADC testing. The PDF of s is then given by

$$f_s(s) = \frac{1}{\pi A \sqrt{1 - \left(\frac{s}{A}\right)^2}}, \quad |s| < A, \tag{19}$$

and the differential entropy (in bits) can be shown to be

$$h(s) = \log_2 \left(\frac{\pi A}{2}\right). \tag{20}$$

Inserting this result into (18) we get the lower bound

$$D = \frac{\pi A^2}{16} 2^{-2R}. \tag{21}$$

Using this last result we can obtain a lower bound on the distortion when quantizing sinusoids. For instance, when quantizing a full-scale sinusoid with a b-bit quantizer, the amplitude is $A = 2^{b-1}$ (LSBs) and the rate is $R = b$ (bits). The squared-error distortion can in this case never be lower than

$$D_{\text{FS}} = \frac{\pi}{64}. \tag{22}$$

This result can be compared with the MSE of a uniform quantizer in (16), which upon inserting $T = 1$ LSB equates to $\frac{1}{12} > \frac{\pi}{32e}$. Note, however, that D_{FS} is a *lower bound* on the distortion of a quantizer *tailored* for a sinusoid input, while $\text{MSE}_{\text{uniform}}$ applies when we are restricted to uniform quantizers.

D. SINAD

When characterizing ADCs the signal-to-noise and distortion ratio (SINAD) is more frequently used that the MSE. It is therefore interesting to state the results obtained above in terms of SINAD instead of MSE. The SINAD is defined as [6]

$$\text{SINAD} = 20 \log_{10} \frac{A}{\sqrt{2} P_{\text{noise}}} \quad [\text{dB}], \tag{23}$$

where A is the amplitude of the sine-wave test signal and P_{noise} is the rms noise amplitude. For this purpose we can use the MSE expressions above so that

$$P_{\text{noise}} = \sqrt{\text{MSE}}. \tag{24}$$

To obtain the SINAD as a function of correction table resolution, we apply the result in (14) and use either (16) or (21) as MSE_o. We obtain the expressions

$$\text{SINAD}_{\text{uniform}} = b\, 20 \log_{10} 2 + 10 \log_{10} \frac{3}{2} - 10 \log_{10}(1 + \Delta^2) \approx 6.02b + 1.76 - 10 \log_{10}(1 + \Delta^2) \tag{25}$$

and

$$\text{SINAD}_{\text{SLB}} = b\, 20 \log_{10} 2 - 10 \log_{10} \left(\frac{\pi}{8} + \frac{2\Delta^2}{3}\right) \approx 6.02b + 1.76 - 10 \log_{10}(0.59 + \Delta^2), \tag{26}$$

respectively, for the two different alternatives.

IV. Simulations

Simulations with a mathematical ADC model has been conducted in order to verify the results above. A 10-bit ADC is simulated using an ADC model with a 4th order polynomial describing the distorted INL followed by an ideal quantizer. A sinusoid with 16384 samples is generated and fed through the model. The output from the model is used to calibrate the LUT. Subsequently, 16 sinusoids having 4096 samples each are generated and the ADC model is applied to them. The output samples of the model are corrected using the LUT, and the SINAD is calculated for the corrected signals (mean SINAD over the 16 signals). The values of the LUT are subsequently quantized to a lower precision and the correction and evaluation procedure is repeated for several different Δ.

The SINAD for the corrected output is presented in Fig. 3. We see that the results align well with the predicted value (25) (labelled 'Prediction') up to a table resolution of approximately 2 bits less resolution than the ADC (i.e., $\Delta = 2^2 = 4$ LSBs or *converter* least significant bits). For lower resolution, however, the SINAD becomes that of the uncorrected converter marked with '+' in the plot.

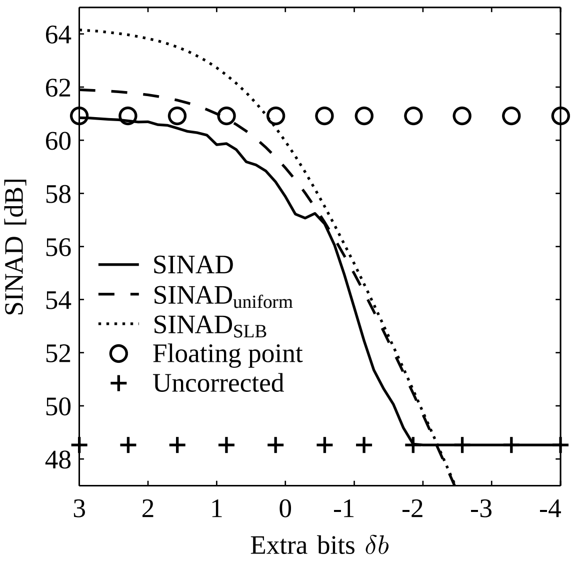

Fig. 3. Simulation results using an ADC model with a polynomial INL function. The plot shows the resulting SINAD after correction with different levels of quantization on the correction values. The x-axis grades how many extra bits in excess of the resolution of the quantizer that the correction values are represented with. Negative values correspond to where the precision of the correction values are worse than 1 LSB of the ADC.

V. Conclusions and Discussion

In this paper we have presented methods to predict the effects of limited resolution in additive ADC (or quantizer) post-correction systems. The first results explained how the MSE of the corrected quantizer increases with decreased resolution. This first result was dependent on the outcome of infinite resolution correction, a quantity that is not always known. Therefore, as a second result, two theoretical values for this starting point was suggested. The first was based on ideal uniform quantization, while the second was based on optimal quantization. Finally, simulation results using a behavioral ADC model showed good agreement with the presented theory. The theories presented here can be used as a design tool to decide how many bits should be used for storing the correction values (τ), and how much the correction value should be shifted relative to the ADC bits (δb).

It is evident from the simulation results that the choice of starting point, or MSE$_o$, adds to the uncertainty of the overall results. We see that the uniform quantization approximation SINAD$_{\text{uniform}}$, where MSE$_o$ is selected as in (16), is closer to the simulation results than SINAD$_{\text{SLB}}$. This is in fact reasonable, since the actual quantizers transfer function is likely closer to uniform than to the optimal non-uniform quantizer needed to get close to the Shannon lower bound. (It may even be impossible to construct a quantizer that achieves the Shannon lower bound.)

One reason for discrepancy between the actual outcome and the theories is that the theories presented here does not take estimation errors into account. That is, the non-quantized table values may also have errors, which stem from the calibration process.

References

[1] Eulalia Balestrieri, Pasquale Daponte, and Sergio Rapuano. A state of the art on ADC error compensation methods. In *Proceedings IEEE Instrumentation and Measurement Technology Conference*, volume 1, pages 711–716, Como, Italy, May 2004.

[2] Fred H. Irons and Alan I. Chaiken. Analog-to-digital converter compensation precision effects. In *29th Midwest Symposium on Symposium on Circuits and Systems*, pages 849–852, Lincoln, Nebraska, August 1986.

[3] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28(2):129–137, March 1982.

[4] Bernard Widrow, István Kollár, and Ming-Chang Liu. Statistical theory of quantization. *IEEE Transactions on Instrumentation and Measurement*, 45(2):353–361, April 1996.

[5] Robert M. Gray. *Source Coding Theory*. Kluwer Academic Publishers, Boston, MA, 1990.

[6] IEEE. *IEEE Standard for Terminology and Test Methods for Analog-to-Digital Converters*. IEEE Std. 1241. 2000.