<table>
<thead>
<tr>
<th>Title</th>
<th>A Delta-Sigma ADC with Stochastic Quantization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Hirai, Yusaku; Yano, Shinya; Matsuoka, Toshimasa</td>
</tr>
<tr>
<td>Citation</td>
<td>IPSJ transactions on system LSI design methodology. 8 P.123-P.130</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2015-08</td>
</tr>
<tr>
<td>Text Version</td>
<td>publisher</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/11094/52087">http://hdl.handle.net/11094/52087</a></td>
</tr>
<tr>
<td>DOI</td>
<td></td>
</tr>
<tr>
<td>rights</td>
<td>Copyright © 2015 by the Information Processing Society of Japan</td>
</tr>
</tbody>
</table>
A Delta-Sigma ADC with Stochastic Quantization

YUSAKU HIRAI1,a) SHINYA YANO1 TOSHIMASA MATSUOKA1,b)

Received: December 5, 2014, Revised: March 13, 2015,
Accepted: April 29, 2015, Released: August 1, 2015

Abstract: An application of the stochastic A/D conversion to multi-bit delta-sigma modulators is considered, and a novel correction technique for D/A converter (DAC) error is proposed. The stochastic A/D conversion can reduce the area of the quantizer and allows large mismatches. The proposed calibration technique corrects DAC errors using a programmable quantizer. The programmable quantizer has a non-linear characteristic that cancels DAC errors. Using this technique, we can decrease the influence of DAC errors without using conventional dynamic element matching. This A/D converter has a non-linear quantization characteristic, so output digital code must be corrected using a programmable encoder. This code correction and setting of the quantization levels are carried out based on calibration data obtained using genetic algorithm.

Keywords: ADC, multi-bit delta-sigma modulator, DAC-error correction, mismatch, genetic algorithm

1. Introduction

Scaling of CMOS process technologies enable digital circuits to operate at high speed, low power consumption, and in small areas. On the other hand, increasing the device mismatch is inevitable. Therefore, decreasing the performance of analog circuits has become a more serious issue. In A/D converters (ADCs), device mismatch limits their resolution [1]. It has been difficult to manage both high resolution and scaling of devices in circuit design of ADCs.

In previous studies, decreasing device mismatches or their influence on circuits has been the primary focus. However, a signal detection technique called stochastic resonance has been proposed [2], [3], and this technique analyzes and uses device mismatches. Recently, a new prototype ADC that uses this technique, referred to as the stochastic flash analog-to-digital converter (SF-ADC), has been proposed [4]. The SF-ADC carries out A/D conversion using input referred offsets as reference levels. Therefore, large offsets can be allowed, and faint signals under the offset level can be detected.

In the present paper, we apply the SF-ADC to a multi-bit delta-sigma modulator (DSM) in order to achieve a high resolution ADC for fine CMOS processes. In multi-bit DSMs, the non-linearity of internal D/A converters (DACs) affects the resolution. Therefore, reducing the non-linearity of internal DACs is necessary for multi-bit delta-sigma ADCs. Recently, dynamic element matching (DEM) has been used as a non-linearity correction technique. The DEM averages element mismatches randomly (zero-order randomization) or in order (data weighted averaging, DWA). These schemes rely on oversampling and noise shaping and are thus ineffective for very low oversampling ratios (OSRs) [5]. In the present paper, the DAC non-linearity is corrected by a programmable quantizer with the SF-ADCs. The impact of DAC errors can be suppressed by adjusting the quantization characteristics to the non-linearity of the DAC without using the DEM. The proposed DAC-error correction technique is more effective than DEM for low OSR value and small input signal. However, the output digital codes correspond to non-linear analog values because the quantization characteristics become non-linear. In order to correct the non-linearity of the codes, the output codes are corrected by a programmable encoder. These corrections are carried out based on configuration data obtained using genetic algorithm.

Section 2 presents an explanation of the SF-ADC. Sections 3 and 4 introduce the proposed DAC-error correction technique and the code correction method, respectively. In Section 5, the optimization method using genetic algorithm is presented in order to optimize the configuration parameters of the proposed DSM. Section 6 presents the results of a system-level simulation for a eight-level, fourth-order, feed-forward delta-sigma ADC obtained using the proposed technique. In Section 7, we summarize the results of the present study.

2. Stochastic Flash ADC

The conventional flash ADC uses \(2^n - 1\) reference levels \(m\) is the number of bits). The reference levels (voltages) are generated using resistor ladders, for example. In real circuits, the threshold voltages of MOSFETs disperse due to device mismatches. The threshold voltage mismatch \(\Delta V_{th}\) is inversely proportional to the square root of the gate area. Thus, \(\Delta V_{th}\) can be written as Ref. [6]:

\[
\Delta V_{th} = \frac{AV_{TH}}{\sqrt{LW}} \tag{1}
\]

where \(L\) is the gate length, \(W\) is the gate width, and \(AV_{TH}\) is a

---

1. Graduate School of Engineering, Osaka University, Suita, Osaka 565–0871, Japan
a) hirai@si.eei.eng.osaka-u.ac.jp
b) matsuoka@eei.eng.osaka-u.ac.jp
The transconductance parameter \( \beta \), which is related to carrier mobility, gate oxide thickness, and gate size, also has a similar dependence, as follows:

\[
\frac{\Delta \beta}{\beta} \approx \frac{A_K}{\sqrt{VW}}
\]

(2)

where \( A_K \) is a proportionality factor.

**Figure 1** shows a differential pair with device mismatches. The drain current in a saturation region without channel length modulation \( (I_{DL2}) \) is given as

\[
I_{DL2} = \frac{1}{2} \beta_{1,2} (V_{GS1,2} - V_{th1,2})^2
\]

(3)

where \( V_{th1,2} \) and \( \beta_{1,2} \) are the threshold voltages and transconductance parameters of M1 and M2, respectively. The mismatch of the threshold voltage is incorporated as \( V_{th1} = V_{th0} + \Delta V_{th} \). When the output voltage \( V_{out} \) is 0, \( V_{in} \) becomes the input-referred offset \( V_{OS,in} \). Since the input-referred offset \( V_{OS,in} \) is equal to \( V_{OS1} - V_{OS2} \), we can obtain [7]

\[
V_{OS,in} = \left( \frac{I_{SS} + \Delta I_D}{\beta_1} + V_{th0} - \left( \frac{I_{SS} - \Delta I_D}{\beta_2} - V_{th2} \right) \right)
\]

(4)

Here, \( \Delta \beta = \beta_2 - \beta_1 \), \( \Delta R_D = R_{D2} - R_{D1} \), and \( \Delta I_D = I_{D2} - I_{D1} \approx (I_{SS}/2)(\Delta R_D/R_D) \). For simplification, assuming that the mismatch of \( R_D \) influences \( V_{OS,in} \) only slightly, based on Eqs. (1) and (2), \( V_{OS,in} \) is inversely proportional to the square root of the gate area.

In conventional flash ADC architectures, the comparator offset limits the ADC resolution. Therefore, devices must be large in order to achieve the desired resolution. The offset can be reduced by using offset calibration techniques with small area. However, the offset calibration technique becomes more difficult as the CMOS process is refined. On the other hand, the SF-ADC uses the comparator offset. **Figure 2** shows the architecture of the SF-ADC. In the SF-ADC, a number of comparators are connected in parallel, and input referred offset voltages are used as reference voltages. Each offset \( \Delta_{off,i} (i = 1, 2, \cdots N_{comp}) \) is random and difficult to estimate, where \( N_{comp} \) is the total number of comparators. However, the offsets are usually modeled as a Gaussian distribution, as shown in **Fig. 3**. Therefore, the standard deviation of the offsets \( \sigma_{off} \) can be estimated to some extent in the circuit design [8].

Next, considering a set of \( N_{comp} \) comparators, the probability \( P \) of obtaining outputs equal to one depends on the cumulative distribution function of the Gaussian distribution. This probability can be expressed as follows [8]:

\[
P(V_{in}) = \frac{1}{2} + \frac{1}{2} \text{erf} \left( \frac{V_{in} - V_{ref}}{\sqrt{2} \sigma_{off}} \right) \approx \frac{n_H}{N_{comp}},
\]

(5)

where \( V_{in} \) is an input signal, \( V_{ref} \) is a reference voltage, and \( n_H \) is the number of comparators for which the output is one (high). As shown in Eq. (5), this probability is approximately equal to the proportion of the comparator counts for which the output is one, i.e., \( n_H/N_{comp} \). The SF-ADC determines its digital outputs by \( n_H \). The SF-ADC uses comparator offsets as reference levels and does not require any offset cancellation or calibration. The comparator is designed with minimum sized transistors in Ref. [4] and any offset cancellation is not used. This leads to decreasing the area occupation. The offset calibration techniques can be used even in an SF-ADC. In that case, the input range is decreased and sensitivity is increased.

The SF-ADC uses the linear range of the cumulative distribution, as shown in **Fig. 4**. Therefore, the input range of the SF-ADC is determined by the standard deviation of the comparator offsets \( \sigma_{off} \).

The ones adder shown in **Fig. 2** counts the number of comparator...
tors for which the output is one \( (nH) \) and output the \( nH \) in binary code. The architecture shown in Fig. 5 is hardware-efficient and called Wallace-tree ones adder. The output number of bits \( k \) is determined by the total comparator counts \( N_{\text{comp}} \). The \( k \) needs to meet the inequality as follows,

\[
  k \geq \log_2(N_{\text{comp}} + 1).
\]

The Wallace-tree ones adder consists of \( (k - 1) \) stages of adders. Therefore, it becomes difficult to meet timing requirements when \( N_{\text{comp}} \) increases or sampling frequency increases. The timing requirement can be relaxed by placing D-flip-flops between each stage to pipeline the adder tree [9]. However, this creates many cycles of latency, in exchange for relaxed timing requirements.

The SF-ADC is a new architecture that uses device mismatches, so its advantages are more effective for advanced CMOS processes. Reducing the area of the quantizer and detecting faint signals are possible even when using an SF-ADC as a quantizer of a multi-bit DSM.

3. DAC-Error Correction

3.1 Multi-bit Delta-Sigma ADC

Multi-bit delta-sigma ADCs can increase stability and dynamic range compared to single-bit ones [5]. However, multi-bit delta-sigma ADCs require multi-bit internal DACs. The non-linearity of internal DACs directly affects the linearity of ADCs. Therefore, good linearity of the internal DAC or a DAC error-correction technique is needed for multi-bit delta-sigma ADCs. Dynamic element matching has been used as DAC error-correction technique. Dynamic element randomization and DWA are primarily used for DEM. Dynamic element randomization randomly selects analog elements in each clock cycle, and so converts the DAC non-linearity into random noise. DWA selects analog elements cyclically and shapes the DAC errors. Dynamic element randomization increases the noise floor in exchange for reducing distortion. DWA has the problem of distortion due to cyclic selection. Various methods have been proposed in order to address this problem [10], [11]. DEM depends on OSR and hence become ineffective for very low values of the OSR, which are sometimes required in wide band data converters [5]. Moreover, DEM is not favorable for low power consumption because DEM corrects DAC errors dynamically.

3.2 Digitally-Controlled Comparator

In order to calibrate DAC errors in delta-sigma ADCs, first we propose a programmable quantizer utilizing SF-ADCs. Figure 6 shows the block diagram of a comparator circuit of which threshold can digitally be controlled. This comparator circuit consists of a SF-ADC and digital comparator. SF-ADC outputs correspond to the number of comparators that output one. As shown in Fig. 4 that describes the characteristic of SF-ADCs, the effective threshold voltage of this comparator circuit can be controlled by quantizing the SF-ADC output to 1-bit. From Eq. (5), when the threshold of a digital comparator is \( nH \), the effective threshold of the digitally-controlled comparator \( V_{\text{comp,eff}} \) becomes as follows,

\[
  V_{\text{comp,eff}} \approx \sqrt{2 \sigma_{\text{off}} \text{erf}^{-1}\left(2\left(\frac{nH}{N_{\text{comp}} - 1/2}\right)\right)} + V_{\text{ref}},
\]

where \( \text{erf}^{-1}(x) \) is an inverse error function.

3.3 Programmable Quantizer with SF-ADCs

Programmable quantizer can be realized using SF-ADCs. The block diagram is shown in Fig. 7. The proposed programmable quantizer consists of the digitally-controlled comparator circuits in Fig. 6. “Programmable” means that its characteristic can be configured by digital data. The reference ladder coarsely determines the quantization levels and digitally-controlled comparators finely determines the quantization levels. Therefore, quantization characteristic can finely be adjusted by digital threshold data for each comparator unit. The programmable quantizer can also be realized by one SF-ADC and
digital quantizer, but there are some problems. First, the input range is determined by the offset distribution ($\sigma_{\text{off}}$). Next, the number of comparators $N_{\text{comp}}$ increases when we use one SF-ADC as a quantizer, and then latency will be created to relax timing requirements of a ones adder. This latency becomes extra-loop delay for the DSM loop and induces instability in delta-sigma ADCs [12]. On the other hand, the input range is determined by reference voltages generated by a resistor ladder in the quantizer shown in Fig. 7. The number of comparators for each digitally-controlled comparator can be reduced because each SF-ADC output is quantized to only 1-bit. Therefore, the timing requirements of ones adder can be relaxed without any pipelined architecture and the DSM loop can be kept stable.

3.4 Proposed DAC-Error Correction Technique

In the present paper, a novel DAC error-correction technique using a programmable quantizer with SF-ADCs is proposed. Figure 8 shows a block diagram of the proposed architecture, although a decimation filter and a digital low pass filter are missing from the figure. The DAC error-correction block consists of a programmable quantizer and a programmable encoder. The programmable quantizer adjusts the quantization levels in order to cancel DAC errors. The programmable encoder corrects the non-linear quantizer output. The programmable quantizer cancels the influence of DAC errors by adjusting the quantization characteristic to DAC non-linearity.

4. Output Encoder

The proposed DAC error-correction technique can reduce the influence of DAC errors on output signals. However, the quantization characteristic becomes non-linear, and so the output digital code corresponds to a non-linear analog value. Therefore, the quantizer outputs still include DAC non-linearity. The programmable encoder shown in Fig. 8 adds a fractional part to output codes in order to correct the non-linear codes, as shown in Fig. 9. The programmable encoder can be implemented by a register table. The output of the programmable quantizer is thermometer code and this thermometer output will be used as address signals of the register table. The readout data is used as a corrected output codes. This mapping table (the correspondence of encoder inputs to output data) can easily changed by rewriting the data of the register table. The number of bits of the encoder output must be larger than the required resolution of delta-sigma ADCs.

5. Optimization with Genetic Algorithm

DAC errors and offset distribution of SF-ADCs cannot be estimated before production, so configuration data for a programmable quantizer and programmable encoder need to be determined after production. The DSM is a non-linear system as well as dynamic system due to the quantization and memory function of integrators [5]. Therefore, the relationships between configuration parameters and DSM performances are difficult to formulate with simple mathematical equations, so it is effective to employ metaheuristics. Genetic algorithm (GA) [13] can be adopted as an effective optimization method for this problem. The GA is based on the evolution process. The GA obtains a globally optimized solution using the iterative calculation toward populations of candidate solutions (individuals). The population for one iteration is called generation in the GA. The GA consists of the steps as follows.

- Step 1: Generate an initial population. Generation $g$ is set to 1.
- Step 2: Generate a new population by applying cross-over, mutation and selection method.
- Step 3: Terminate calculations if termination conditions are met, and output the optimum solution. If the termination
conditions are not met, set the generation $g$ to $g + 1$ and back to the step 2.

Evaluation value is called fitness and parameter settings to be optimized are called chromosome in the GA. In this optimization problem for the proposed DSM, we adopt ENOB (Effective Number of Bits) as follows,

$$\text{ENOB} = \frac{\text{SNDR [dB]} - 1.76 \text{dB} + \text{(Level of Signal Below FS)}}{6.02 \text{ dB}}.$$  \quad (8)

where SNDR is signal-to-noise-and-distortion ratio. The ENOB will be calculated for multiple levels of input and minimum value of ENOB will be maximized. The objective function can be described as follows,

$$\text{max : min(ENOB).}$$  \quad (9)

The chromosome is the parameter set of quantizer threshold levels and the value of the encoder mapping table (configuration data in Fig. 8).

There are several precedents to use the GA for design parameter optimization for a DSM [14] and parameter optimization for a configurable circuit [15]. The GA is also effective to complex optimization problem for the configurable system proposed in this paper.

6. Simulation Results

System-level verification of the proposed architecture was carried out. The eight-level, fourth-order, feed-forward DSM shown in Fig. 10 was used for the simulations as an example. The $I(z)$ shown in Fig. 10 represents the transfer function of a discrete-time delayed integrator expressed as follows,

$$I(z) = \frac{z^{-1}}{1 - z^{-1}}.$$  \quad (10)

The gain parameters are shown in Table 1. Since the feed-forward DSM has one internal DAC, the proposed DAC error-correction technique can be applied.

First, the optimization of the configuration parameters was carried out using GA. The input amplitude was swept from $-34$ dBFS to $-4.4$ dBFS and other simulation conditions and parameters are shown in Table 2. The encoder output was 15-bit (12-bit for fractional part). The number of comparators of a unit SF-ADC was 31 and the total number of comparators was 217. The $N_{\text{comp}}$ should be as many as possible for higher resolution, but the large number of $N_{\text{comp}}$ will lead the extra loop delay as described in Section 3. So, the $N_{\text{comp}} = 31$ was determined as many as possible within the range that does not create any latency (extra loop delay) with 0.13 $\mu$m CMOS process, where 0.13 $\mu$m CMOS process is just an example for fabrication. The analog element mismatch for a reference ladder and DAC was 1% in standard deviation. For GA, population size was 30 and generation was 50. Figure 11 shows the learning curve of optimization with GA. Final fitness was 12.3 bits and optimized configuration parameters are shown in Table 3 and Table 4. Figure 12 shows the output spectrum (power spectral density, PSD) of the DSM without any calibration. The quantizer was flash ADC with ideal comparators and DAC errors were not calibrated. Figure 13 shows the output spectrum of the proposed DSM with parameter optimization using GA. The input amplitude was $-4.4$ dBFS. These results
Table 3 Optimized thresholds for programmable quantizer.

<table>
<thead>
<tr>
<th>Level</th>
<th>Threshold Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>26</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 4 Optimized encoder mapping table.

<table>
<thead>
<tr>
<th>DSM Output</th>
<th>Encoder Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000000</td>
<td>0000000000100100</td>
</tr>
<tr>
<td>0000001</td>
<td>0010000000101000</td>
</tr>
<tr>
<td>0000011</td>
<td>010000000011111</td>
</tr>
<tr>
<td>0000111</td>
<td>011000000000011</td>
</tr>
<tr>
<td>0001111</td>
<td>100000000001000</td>
</tr>
<tr>
<td>0011111</td>
<td>101000000010101</td>
</tr>
<tr>
<td>0111111</td>
<td>110000000101010</td>
</tr>
<tr>
<td>1111111</td>
<td>111000000011111</td>
</tr>
</tbody>
</table>

Fig. 12 Power spectral density of the DSM output without any calibration.

Fig. 13 Power spectral density of the DSM output with optimization.

Fig. 14 SNDR versus input signal level. The SNDR of the proposed system was maximally improved by 15 dB compared to the DSM without any calibration. The results also show that the proposed DAC-error correction technique is valid as a DAC-error correction technique for multi-bit DSMs. The residual DAC error limits the performance of the proposed DSM. The residual error can be reduced by increasing the number of comparators or the number of bits of the encoder. In addition, the resolution can also be increased by increasing the OSR, the order of loop filter or quantizer resolution for higher resolution.

Fig. 15 SNDR curves versus the input level for DWA and the proposed technique (OSR = 32).
posed technique can achieve the same resolution without any dynamic selections. The same simulation was carried out with lower value of OSR (OSR = 8). The GA optimization was carried out for the input amplitude from −34 dBFS to −20 dBFS. The result is shown in Fig. 16. This result shows that the proposed calibration technique is more effective than DWA for low OSR and small input signal. The proposed DSM can also be used as a non-linear DSM [16] by adjusting the input range.

7. Conclusion

The application of stochastic A/D conversion to delta-sigma ADCs was discussed, and a novel DAC-error correction technique using the SF-ADC was proposed. The proposed DAC-error correction technique can reduce the influence of DAC errors without DEM. The quantizer and output encoder characteristics are optimized using GA. The proposed technique was shown to be valid as a DAC-error correction technique through simulations using the eight-level, fourth-order, feed-forward DSM. Moreover, the proposed technique was demonstrated to be advantageous for small OSR values and small input amplitude.

Acknowledgments The authors would like to thank the Semiconductor Technology Academic Research Center (STARC) for financial support. The present study was supported in part by the Japan Society for the Promotion of Science (JSPS) through a Grant-in-Aid for Scientific Research (B) (#23360155).

References


Yusaku Hirai received his B.S. and M.S. degrees in electronic engineering from Osaka University, Osaka, Japan, in 2013, and 2015, respectively. He is now working for THine Electronics Inc., where he is developing high-speed serial interfaces.

Shinya Yano received his B.S. and M.S. degrees in electronic engineering from Osaka University, Osaka, Japan, in 2011, and 2013, respectively. He is now working for Mitsubishi Electric Inc., where he is carrying out the research of power modules. He is a member of IEICE.
Toshimasa Matsuoka received his B.S., M.S. and Ph.D. degrees in electronic engineering from Osaka University, Osaka, Japan, in 1989, 1991, and 1996, respectively. During 1991–1998, he worked for Central Research Laboratories, Sharp Corporation, Nara, Japan, where he was engaged in the research and development of deep-submicron CMOS devices and ultra-thin gate oxides. Since 1999, he has been working at Osaka University, where he is an Associate Professor. His current research includes CMOS analog/RF circuits and device modeling. Dr. Matsuoka is a member of the Japan Society of Applied Physics, IEICE, and IEEJ.

(Recommended by Associate Editor: Akio Hirata)