ROUND OFF NOISE MINIMIZATION IN A DIRECT FORM
DELTA OPERATOR STRUCTURE

Juha Kauraniemi1, Timo I. Laakso2, Iiro Hartimo3, and Seppo J. Ovaska3

1Laboratory of Signal Processing and Computer Technology, Helsinki University of Technology, Espoo, Finland
Juha.Kauraniemi@hut.fi
2SEMSE, University of Westminster, London, United Kingdom
3Laboratory of Telecommunications Technology, Helsinki University of Technology, Espoo, Finland

ABSTRACT

Delta operator realizations of IIR filters have recently received a lot of interest due to their good finite-wordlength properties. Here we present a detailed analysis of the computationally efficient transposed direct form II delta structure, which is also known to have excellent roundoff noise performance under fast sampling. Focus is on the roundoff noise minimization in fixed-point implementations. The free $\Delta$ parameter in the delta operator is optimized with respect to the output roundoff noise within the scaling constraints. Moreover, required internal wordlength to obtain prespecified noise performance is derived.

1. INTRODUCTION

Severe numerical problems may arise when processing signals at sampling rates much higher than the bandwidth of interest. In this case the poles of a filter cluster near to the unit circle, and make the filter implementation very noisy and sensitive to coefficient quantization. The use of the delta operator has been proposed to overcome such difficulties [1]. In our recent work, delta operator realized direct form structures were analyzed [2]. It was found out that the transposed direct form II (DFII) structure is especially well suited for the delta operator realization. However, it is not obvious how the $\Delta$ of the delta operator $\delta = (z - 1)/\Delta$ should be chosen in a fixed-point realization.

The price of the delta operator realizations is the increased implementation complexity. Besides additional arithmetic operations, the implementation complexity is increased by the need for enhanced precision $\delta^i$ operations [2-5]. In a signal processor implementation, it is not possible to obtain savings in the code length by using less than exact double precision. However, in an ASIC implementation, additional bits increase the die area and therefore it is desirable to limit their number to as few as possible.

In this paper, we focus on the minimization of roundoff noise in the DFII delta structure via optimization of the $\Delta$ parameter. The optimal $\Delta$ value is derived with the constraints for scaling to prevent overflow. It is shown that the optimal $\Delta$ can result in considerably lower output roundoff noise than the computationally efficient choice $\Delta = 1$ when the internal wordlength is minimized. Implementation aspects are also discussed, indicating that rounding the optimal $\Delta$ parameter to the closest power of two results in efficient VLSI implementations.

2. SECOND-ORDER DELTA DFII STRUCTURE

In practice, second-order sections are widely used as building blocks of high-order filters. The $z$-domain transfer function of a second-order IIR filter is

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}.$$  

(1)

It can be expressed in the delta form as

$$H_\delta(\delta) = \frac{\beta_0 + \beta_1 \delta^{-1} + \beta_2 \delta^{-2}}{1 + \alpha_1 \delta^{-1} + \alpha_2 \delta^{-2}},$$  

(2)

where $\delta^{-1} = \Delta z^{-1}(1 - z^{-1})$ is the causal inverse of the delta operator. The coefficients of the delta parametrized transfer function (2) are given in Table 1.

2.1. Scaling Transfer Functions

The delta DFII structure is shown in Fig. 1. There are three summations which are critical for scaling denoted as $s_i$, $i = 0, 1, 2$. Note that in each $\delta^i$ operation an addition is performed, but when two's complement arithmetic is used, these are allowed to overflow, if only the correct total values of the summations $s_0$ and $s_1$ do not overflow. In the specific case when $\Delta = 1$ only the summation $s_0$ has to be scaled, and the scaling transfer function of the unscaled system is

$$F_{\delta,0}(z) = \frac{S_0(z)}{X(z)} = H(z)/g(z),$$  

(3)

where $S_0(z)$ is the $z$-transform of the signal after summation $s_0$. When $\Delta < 1$, summations before the $\delta^i$ operators are not allowed to overflow. Transfer functions to these points in an unscaled system are

$$F_{\delta,1}(z) = \frac{S_1(z)}{X(z)} = \frac{1}{\Delta} \frac{d_0 + d_1 z^{-1} + d_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}},$$  

(4)

$$F_{\delta,2}(z) = \frac{S_2(z)}{X(z)} = \frac{1}{\Delta^2} \frac{f_0 + f_1 z^{-1} + f_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}},$$  

(5)

where $d_0 = b_0 - b_2 a_1, d_1 = b_1 (a_1 - a_2) + b_2 - b_1, d_2 = b_2 a_2 - b_2 f_0 = b_1 + b_2 f_2 + b_2 f_0, f_1 = (b_0 + b_2 a_1) - b_2 (a_1 + 1), f_2 = (b_0 + b_2 a_2 - b_2 (a_1 + 1)).$ The required scaling is given by

This research work is a project of Institute of Radio Communications (IRC) and it is supported by the Technology Development Centre (TEKES).
Table 1
Relations between the coefficients of the second-order δ- and ε-polynomials

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>$b_0$</td>
<td>$\alpha_0$</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$(2b_0 + b_1)/\Delta$</td>
<td>$\alpha_1$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$(b_0 + b_1 + b_2)/\Delta^2$</td>
<td>$\alpha_2$</td>
</tr>
</tbody>
</table>

$g_i = \left( \max |F_{g_i}(z)|_p \right)^{-1}, \ i = 0, 1, 2,$ (6)
where $|F(z)|_p$ is the $L_p$ norm of the function $F(z)$ [6]. Throughout this paper the $L_p$ norm is used for scaling.

2.2. Roundoff Noise

When the roundoff noise performance of a digital filter is analyzed, the usual assumption is that the quantization noise can be modeled as a white noise process and superposition holds for the noise components [6]. If $(B + 1)$ bits are used for single precision and $(B + 1) + 1$ bits for enhanced precision, the corresponding noise variances in the case of rounding are

$$\sigma_i^2 = 2^{-2i}/12, \quad i = 0, 1, 2,$$ (7)
$$\sigma_i^2 = 2^{-2(B+1)}/12, \quad (8)$$
The usual measure of roundoff noise performance is the noise gain, which in the case of multiple noise sources with different variances is defined as

$$NG = \sigma_{\text{out}}^2 / \sigma_i^2,$$ (9)
where $\sigma_{\text{out}}^2$ is the total output noise variance. Noise gain in decibels is defined as $NG_{\text{dB}} = 10 \log NG$.

With double precision, i.e., multiplication results are not quantized before the $\delta^{-1}$ operations, and $\Delta = 1$ is used, only one noise source exists having the transfer function

$$G_{\delta,0}(z) = \frac{Y(z)}{S_0(z)} = \frac{1 - z^{-1}}{1 + a_1 z^{-1} + a_2 z^{-2}}.$$ (10)
Equation (10) is independent of $\Delta$ and because multiplication by $\Delta < 1$ introduces more noise sources, minimum roundoff noise is obtained by using $\Delta = 1$. If enhanced precision is used more noise sources result, having transfer functions

$$G_{\delta,1}(z) = \frac{Y(z)}{S_1(z)} = \frac{\Delta z^{-1}(1 - z^{-1})}{1 + a_1 z^{-1} + a_2 z^{-2}},$$ (11)
$$G_{\delta,2}(z) = \frac{Y(z)}{S_2(z)} = \frac{\Delta^2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}.$$ (12)
These transfer functions are directly proportional to $\Delta$, and $\Delta = 1$ may no longer be optimal in the roundoff noise sense. The transfer functions from the noise sources due to the multiplication by $\Delta < 1$ are obtained from (11) and (12) as

$$G_{\delta,1}(z) = G_{\delta,0}(z)/\Delta,$$ (13)
$$G_{\delta,2}(z) = G_{\delta,0}(z)/\Delta^2.$$ (14)
The noise transfer functions of the scaled system are related to those of the unscaled network as

$$x(n) \rightarrow \beta_0^s \rightarrow \varnothing \rightarrow s_0 \rightarrow \varnothing \rightarrow g_s \rightarrow y(n).$$

Fig. 1 Second-order delta operator realized DF1I section. Quantizations to enhanced precision are drawn with dotted line and $\beta_1^s = g_0^s \beta_1$ and $g_s = g_0/g_s$.

$$G_{\delta,0}^s(z) = g_0 \max_n \left[ F_{\delta,0}(z) \right]_l G_{\delta,0}(z),$$ (15)
where $\times \in \delta \{ (10) - (12) \} \text{ or } \Delta \{ (13) \text{ or } (14) \}$. If $\Delta = 1$ and double precision is used, the total output noise variance is

$$\sigma_{\text{out}}^2 = \left\| G_{\delta,0}^s(z) \right\|_2^2 \sigma_i^2.$$ (16)
If $\Delta < 1$ and enhanced precision is used, the total output noise variance is

$$\sigma_{\text{out}}^2 = \left\| G_{\delta,0}^s(z) \right\|_2 \left( \sigma_{\text{out}}^2 + \sum_{i=1}^2 \left\| G_{\delta,1}^s(z) \right\|_2^2 + 2 \left\| G_{\delta,2}^s(z) \right\|_2^2 \right) \sigma_i^2.$$ (17)
From (3) - (5) and (10) - (15) it is seen that the total output noise variance is a function of $\Delta$, and the optimal value for the $\Delta$ parameter can be derived.

3. PARAMETER OPTIMIZATION

3.1. Enhanced Precision

If $D$ times increase in the output noise power (when compared to the double precision case) is allowed, the following formula for the number of the additional bits $B_d$ is obtained:

$$B_d = \left\lfloor \frac{1}{2 \log_2 \left( \frac{\left\| G_{\delta,0}(z) \right\|_2 + \sum_{i=1}^2 \left\| G_{\delta,1}(z) \right\|_2^2 + 2 \left\| G_{\delta,2}(z) \right\|_2^2}{(D-1) \left\| G_{\delta,0}(z) \right\|_2^2} \right) \right\rfloor,$$ (18)
and $D > 1$. By the choice $D = 2$ the number of bits to maximum of 3 dB increase in the noise variance is obtained. It was noted earlier, that when double precision in the $\delta^{-1}$ line is used, minimum noise gain is obtained using $\Delta = 1$. However, when signal values have to be quantized also in the $\delta^{-1}$ line, it may be advantageous to use a value smaller than unity for $\Delta$.

3.2. Optimization of the Delta Parameter

Minimization of (17) with respect to $\Delta$ is carried out in the Appendix. It was observed that the optimum can be found from one of the three different regions ($A, 3$), ($A, 5$) or ($A, 7$), but, according to our experience, when a narrowband lowpass filter is designed and the $L_p$-scaling strategy is used the optimum is
Fig. 2 Noise gain of delta DFIIT as a function of the pole angle and parameter $\Delta$. The pole radius $r = 0.99$, the zeros of the transfer function are at $z = -1$ and $B_0 = 3$.

$\Delta = \max \left\{ \frac{F_{k,i}(z)}{F_{k,0}(z)}, \frac{F_{k,0}(z)}{F_{k,i}(z)} \right\}$, \hspace{1cm} \text{(19)}$

where $F_{k,i}(z) = \Delta^i F_{k,0}(z)$, $i = 1, 2$. To simplify the hardware implementation, $\Delta$ can be rounded to the nearest power of two and enhanced precision multiplications can be replaced by right shifts. Noise gain of the delta DFIIT section as a function of the pole angle and parameter $\Delta$ is shown in Fig. 2.

4. EXAMPLES

To illustrate the properties of the delta structures, two sixth-order lowpass filters are designed. The test filters are implemented as cascaded second-order sections. The individual sections are scaled and scaling is embedded into the numerator coefficients. Delta realizations are compared to DFIIT and DFIIT delay structures.

Three different delta realizations are considered. In the first one, internal double precision is used and in this case the optimal $\Delta = 1$ in each section. In the second one, enhanced precision is used and the optimal $\Delta$ is chosen independently for each section and rounded to the nearest power of two. The number of additional bits in enhanced precision is chosen so that the increase in the output noise level is less than 3 dB. It is denoted as $\Delta$DFIITX in Tables 2-3, where X is the number of additional bits. Also in the third realization X additional bits are used but $\Delta = 1$ is used instead of the optimal value ($\Delta$DFIITXb in tables).

Filter #1 is an elliptic design having magnitude response as shown in Fig. 3. Theoretical noise gain values of various section orderings for different filter structures are presented in Table 2. Pole-zero pairing is not explicitly optimized and it is the same in all cases. From this table, it is seen that the delta realizations perform much better than the delay realizations. Moreover, only three additional bits are required to keep the noise increase under 3 dB. In this configuration, significant reduction in the noise gain is obtained by optimizing the $\Delta$ parameters.

Filter #2 is a very narrowband Chebyshev type I filter (Fig. 4). With this filter the delta realizations are superior to delay realization (Table 3). However, more bits (five) are required to the enhanced precision than in the test Filter #1 when noise increase is to stay below 3 dB. Importance of the optimized $\Delta$ parameters is also greater than in the previous case.

5. CONCLUSIONS

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Theoretical noise gain values (dB) for Filter #1 with various section orderings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordering</td>
<td>1-2-3</td>
</tr>
<tr>
<td>DFIIT</td>
<td>36.08</td>
</tr>
<tr>
<td>DFIIT</td>
<td>39.30</td>
</tr>
<tr>
<td>8DFIIT</td>
<td>2.98</td>
</tr>
<tr>
<td>8DFIIT3</td>
<td>7.46</td>
</tr>
<tr>
<td>8DFIIT3b</td>
<td>15.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Theoretical noise gain values (dB) for Filter #2 with various section orderings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordering</td>
<td>1-2-3</td>
</tr>
<tr>
<td>DFIIT</td>
<td>63.11</td>
</tr>
<tr>
<td>DFIIT</td>
<td>67.87</td>
</tr>
<tr>
<td>8DFIIT</td>
<td>0.86</td>
</tr>
<tr>
<td>8DFIIT5</td>
<td>9.78</td>
</tr>
<tr>
<td>8DFIIT3b</td>
<td>36.01</td>
</tr>
</tbody>
</table>
portant when a filter is implemented as an ASIC. It is also clear
that significant improvement in the roundoff noise performance
and other parameters (11 - 28 dB
with our example filters).

Due to a quantization nonlinearity in the feedback loop a
digital filter may produce autonomous oscillations or limit cycles
when the input is zero. Furthermore, overflows in summations
can cause large amplitude periodic oscillations. These both are
important topics for the further study.

6. APPENDIX

In this appendix, the roundoff noise optimal Δ parameter is de-

rived. Optimization is performed for a single second-order sec-
tion. Each section of a cascade structure can be optimized inde-

pendently and the derivation is similar, only the transfer func-
tions are changed to those of the cascaded sections. Moreover,

optimization is performed for a fixed section ordering and pole

tzero pairing which have to be optimized separately.

It is assumed that the filter coefficients are constants and the

only variable affecting on the roundoff noise minimization is Δ.
The result that the parameter can be taken at the front of the

norm because it is a real-valued positive constant, Δ ∈ (0, 1).

The noise sources are assumed to be independent of all the

other noise sources so that they can be summed in powerwise.
The output noise variance can be expressed as

\[ \sigma_{out}^2 = \sigma_1^2 \lVert F_{p,0} \rVert_p \lVert G_{r,0} \rVert_p (\sigma_1^2 + \sigma_2^2) + \lVert G_{r,1} \rVert_p \sigma_2^2 + \left( \lVert G_{r,5} \rVert_p^2 + 2 \lVert G_{r,6} \rVert_p \sigma_1^2 \Delta^2 + 2 \lVert G_{r,9} \rVert_p^2 \sigma_1^2 \Delta^4 \right) \]  

(6.1)

where the scaling transfer function is

\[ \|F_p\|_p = \max \left( \|F_{p,0}\|_p, \|F_{p,1}\|_p \Delta^{-1}, \|F_{p,2}\|_p \Delta^{-2} \right). \]  

(6.2)

The transfer functions with hyphens (−) are the Δ-independent

parts of the corresponding transfer functions \( F_{p,0}(z) = Δ F_{p}(z), \)

\( G_{r,0}(z) = Δ^{-1} G_{r}(z), G_{r,2}(z) = Δ^{-1} G_{r,2}(z). \) Because arguments

of the max-function depend on different powers of Δ, the minimization

has to be performed in three parts, corresponding to the three

possible maxima in (6.2).

In the first region, the Δ parameter is limited by:

\[ \max \left( \|F_{p,0}\|_p, \|F_{p,1}\|_p, \|F_{p,2}\|_p \Delta^{-2} \right) \leq Δ \leq 1. \]  

(6.3)

Using \( \|F_{p,0}\|_p \) as the largest norm in (6.2), (6.1) becomes

\[ \sigma_{out}^2 = \sigma_1^2 \lVert F_{p,0} \rVert_p \lVert G_{r,0} \rVert_p (\sigma_1^2 + \sigma_2^2) + \lVert G_{r,1} \rVert_p \sigma_2^2 + \left( \lVert G_{r,5} \rVert_p^2 + 2 \lVert G_{r,6} \rVert_p \sigma_1^2 \Delta^2 + 2 \lVert G_{r,9} \rVert_p^2 \sigma_1^2 \Delta^4 \right) \]  

(6.4)

Solving for the zero of the derivative of (6.4) results in a com-

plex valued Δ or Δ = 0, which are out of interest. The minimum

is then found from one of the end points of the interval (6.3). In

the formula (6.4), there exist terms which are independent or di-

rectly proportional to the Δ. As a consequence, it is minimized by

choosing the lower limit for the parameter Δ from (6.3).

In the second region,

\[ \|F_{p,2}\|_p / \|F_{p,0}\|_p \leq Δ \leq \|F_{p,1}\|_p / \|F_{p,0}\|_p. \]  

(6.5)

The zero point of the derivative of the noise variance is found to

be

\[ Δ = \left( \lVert G_{r,0} \rVert_p \left( 2 \Delta + 1 \right) + 1 \right) / \lVert G_{r,1} \rVert_p^2 \lVert G_{r,2} \rVert_p^2 \Delta^4, \]  

(6.6)

where \( B_1 \) is the number of additional bits in the enhanced

precision. To ensure that (6.6) is the minimum, the second deriv-

ative of the noise variance equation has to be positive at (6.6).

It follows that it is positive and (6.6) is a minimum if it satisfies

(A.5). The region defined by (A.5) can also be empty.

In the third case, the following region is obtained:

\[ 0 < Δ ≤ \min \left( \|F_{p,1}\|_p / \|F_{p,0}\|_p, \|F_{p,2}\|_p / \|F_{p,0}\|_p \right). \]  

(6.7)

Solving for the zeros of the derivative of the resulting noise vari-

ance formula, it is found that three of the zeros are at the infinity

and therefore are not the desired minimums. Two of the zeros

again result in complex-valued Δ which is not of interest to us.

Because the scaling transfer function is inversely proportional to

the second power of Δ the minimum is found from the upper

limit of the interval (6.7).

If the second region is empty or consists of only one point it is

straightforward to show that the global minimum is:

\[ Δ = \left( \|F_{p,2}\|_p / \|F_{p,0}\|_p \right)^2. \]  

(6.8)

With the Filter #1 the second region is always empty and the
global minimum is (6.8). With the Filter #2 there are two section

orderings (1-2-3 and 2-1-3) where the second region is not empty

in the second section. However, in both cases (6.6) is outside the

interval (6.5) and when compared it is found that the minimum

is given by upper limit of (A.5), which is equal to lower limit of

(A.3) when the second region is not empty.

7. REFERENCES

[1] G. C. Goodwin, R. H. Middleton, and H. V. Poor, "High

speed digital signal processing and control," Proc. IEEE,


"Delta operator realizations of recursive digital direct form

filters," in Proc. 12th European Conference on Circuit

Theory and Design, Vol. 2, pp. 667-670, Istanbul, Turkey,


effects in shift and delta operator parametrizations," IEEE


Trans. Acoust., Speech, and Signal Processing,

