# IMPROVED ROUNDOFF NOISE PERFORMANCE IN A DIRECT-FORM IIR FILTER USING A MODIFIED DELTA OPERATOR

*Ngai Wong, Student Member, IEEE, and Tung-Sang Ng, Senior Member, IEEE*

Department of Electrical and Electronic Engineering,
The University of Hong Kong,
Pokfulam Road, Hong Kong.
nwong@eee.hku.hk, tsng@eee.hku.hk
Tel.: ++ 852 + 2857 8406, Fax: ++ 852 + 2559 8738

## ABSTRACT

Among various direct-form delta operator filters, the delta direct-form II transposed ($\delta$DFIIt) has been shown to produce the lowest roundoff noise in finite-word-length implementations. Recent analyses focus on the optimization of the free parameter $\Delta$ of the delta operator, with scaling of the structure to prevent arithmetic overflow. This paper proposes a modified $\delta$DFIIt second-order section in which the $\Delta$s at different branches are separately optimized to further suppress roundoff noise gain. Noise variance plots against pole locations are presented. Closed-form expressions for the optimal filter coefficients are derived and reduction of noise gain is confirmed by numerical examples.

## 1. INTRODUCTION

Delta operator realized digital filter structures have attracted increasing attention in this decade due to their good numerical properties when compared to traditional delay structures [1]-[8]. This is especially true when the sampling rate is much higher than the underlying signal bandwidth in which the $z$-plane poles cluster towards unity. By replacing the conventional $z^{-1}$ operator with the inverse delta operator $\delta^{-1} = \Delta z^{-1} / (1 - z^{-1})$, certain ill-conditioned numerical issues can be overcome. Moreover, delta operator filters are generally accompanied with better roundoff noise performance and more robust coefficient and frequency sensitivities [1], [2]. Although an $\delta^{-1}$ operator is more complicated to implement in terms of hardware, its excellent numerical properties allow the use of shorter word-length, which results in moderate complexity or even gross savings in silicon area [6].

Extensive study of different direct-form delta structures has been carried out in [4]. It was found that the delta direct-form II transposed ($\delta$DFIIt) exhibits the best roundoff noise properties among various delta structures. Focus has been put on the optimization of the free parameter $\Delta$ of the delta operator to achieve minimum roundoff noise gain at the output. The basic second-order $\delta$DFIIt section was studied in detail [3], [4].

In this paper, instead of limiting to optimizing a single $\Delta$ within the second-order $\delta$DFIIt section, the concept of separately optimizing the $\Delta$s in each $\delta^{-1}$ operator is introduced[1]. It will be shown that this approach enables further reduction of roundoff noise gain. Simple and readily computable expressions for the optimal filter coefficients are derived.

To begin with, Section 2 of this paper describes the transformation of a conventional second-order direct-form II transposed (DFIIt) filter into its corresponding $\delta$DFIIt counterpart. Signal and noise transfer functions are then given. The graphical approach in Section 3 shows how the noise variance in a $\delta$DFIIt structure behaves as filter poles approach unity in the complex $z$-plane. Section 4 presents the derivation of the optimal $\Delta$ in each branch. Numerical examples are given in Section 5 and Section 6 concludes this paper.

Before proceeding, it should be noted that the $L_p$-norm [9] of a transfer function $F$ is defined as

$$\|F\|_p = \left[ \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| F(e^{j\omega}) \right|^p d\omega \right]^{\frac{1}{p}}. \qquad (1)$$

In fixed-point implementations, the noise variance for $(B+1)$-bit quantization with rounding is [10]

$$\sigma_e^2 = 2^{-2B}/12. \qquad (2)$$

Assuming additive uncorrelated white noise process, the noise variance $\sigma_o^2$ seen at the output of a linear system $F$ is related to the $L_2$-norm and $\sigma_e^2$ by

$$\sigma_o^2 = \sigma_e^2 \|F\|_2^2. \qquad (3)$$

## 2. STRUCTURAL TRANSFORMATION

Suppose a transfer function in the $z$-domain, represented by (4), is obtained under certain sampling conditions, it can then be transformed into an equivalent delta structure by substituting $z = 1 + \delta\Delta$ where $\Delta$ is a positive constant.

Fig. 1(a) and 1(b) show how a DFIIt structure is converted into a $\delta$DFIIt implementation. Fig. 1(c) incorporates the scaling constant $g$ to prevent arithmetic overflow. This is necessary since finite-word-length arithmetic is used in practice. This basic second-order $\delta$DFIIt building block was studied extensively for roundoff noise minimization in previous work [3], [4], in which the same $\Delta$ occurs in both $\delta^{-1}$ operators.

[1] The notion of using different $\Delta$s also appeared in [5] but optimization with respect to them has not been done.

0-7803-6685-9/01/$10.00©2001 IEEE

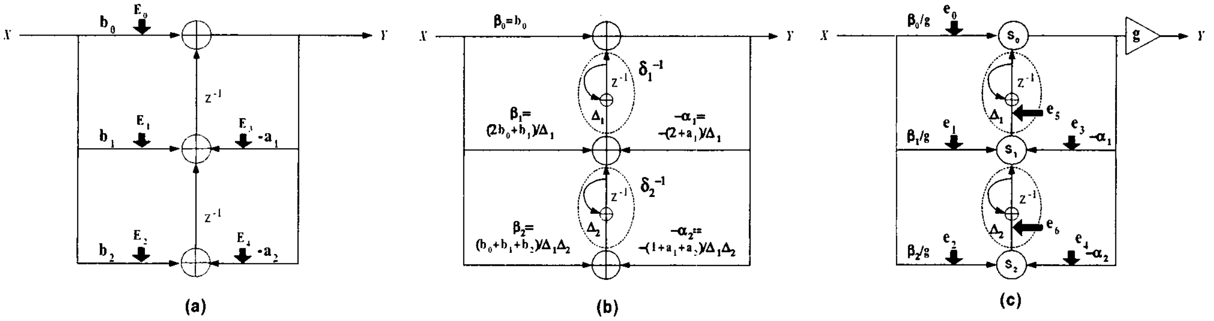

**Figure 1.** **(a)** Conventional DFIIt structure with possible quantization noise sources. **(b)** Transformation into a $\delta$DFIIt structure. **(c)** $\delta$DFIIt structure with overflow prevention scaling and possible quantization noise sources.

In our analysis, two separate $\Delta$ s, $\Delta_1$ and $\Delta_2$ as in Fig. 1(b) and 1(c), are used to allow more degree of freedom for optimization. If less-than-double-precision fixed-point arithmetic is used after each coefficient multiplication (therefore the roundoff quantization error sources $e_0$ to $e_6$), the summation nodes $S_0$, $S_1$ and $S_2$ (called branch nodes [9]) have to be scaled to prevent overflow. Their signal transfer functions without scaling [i.e., $g=1$ in Fig. 1(c)] are given in (4)-(6). A common overflow prevention strategy is to use $L_\infty$-norm scaling. Using the convention (also throughout this paper) that a tilde-topped transfer function represents its $z$-dependent part after any prefixing constants, the scaling factor will be

$$g=\max\left(\|F_0\|_\infty,\|F_1\|_\infty,\|F_2\|_\infty\right)$$
$$=\max\left(\|F_0\|_\infty,\|\tilde{F}_1\|_\infty\Delta_1^{-1},\|\tilde{F}_2\|_\infty\Delta_1^{-1}\Delta_2^{-1}\right). \qquad (7)$$

Assuming noise from the back scaling by $g$ before the output $Y$ [see Fig. 1(c)] is absorbed into the next section, transfer functions from different roundoff quantization noise sources due to coefficient multiplication are given as follows,

$$G_0=\frac{Y}{e_0}=g\frac{(1-z^{-1})^2}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (8)$$

$$G_{1,3}=\frac{Y}{e_1}=\frac{Y}{e_3}=g\Delta_1\frac{z^{-1}(1-z^{-1})}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (9)$$

$$G_{2,4}=\frac{Y}{e_2}=\frac{Y}{e_4}=g\Delta_1\Delta_2\frac{z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (10)$$

$$G_5=\frac{Y}{e_5}=g\frac{z^{-1}(1-z^{-1})}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (11)$$

$$G_6=\frac{Y}{e_6}=g\Delta_1\frac{z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}. \qquad (12)$$

Modeling rounding quantization noise as an additive uncorrelated white noise process, superposition holds and the output noise variance is expressed as a sum of noise powers,

$$\sigma_o^2=\sigma_e^2g^2\Big(\|\tilde{G}_0\|_2^2+\|\tilde{G}_5\|_2^2+$$
$$\left(\|\tilde{G}_6\|_2^2+2\|\tilde{G}_{1,3}\|_2^2\right)\Delta_1^2+2\|\tilde{G}_{2,4}\|_2^2\Delta_1^2\Delta_2^2\Big). \qquad (13)$$

A word-length-independent noise gain term can therefore be defined as the ratio $\sigma_o^2/\sigma_e^2$. Minimization of this term is carried out in Section 4.

## 3. GRAPHICAL NOISE ANALYSIS

As noted in Section 1, the noise gain is proportional to the squared $L_2$-norm of the noise transfer function, e.g. the noise power produced by $e_1$ is $\sigma_e^2g^2\Delta_1^2\|\tilde{G}_{1,3}\|_2^2$.

Form (8)-(12), it is seen that the noise transfer functions are dependent on the poles. As the filter coefficients are real, the poles are complex conjugates, say, $re^{j\theta}$ and $re^{-j\theta}$. Ignoring the effects of the noise gain prefixing constants $g$, $\Delta_1$ and $\Delta_2$ which are filter-dependent, it is possible to derive closed-form solutions

$$F_0=\frac{Y}{X}=\frac{S_0}{X}=\frac{b_0+b_1z^{-1}+b_2z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (4)$$

$$F_1=\frac{S_1}{X}=\Delta_1^{-1}\frac{(b_1-b_0a_1)+\big(b_0(a_1-a_2)+b_2-b_1\big)z^{-1}+(b_0a_2-b_2)z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}, \qquad (5)$$

$$F_2=\frac{S_2}{X}=\Delta_1^{-1}\Delta_2^{-1}\frac{(b_1+b_2-b_0(a_1+a_2))+(a_1(b_0+b_2)-b_1(1+a_2))z^{-1}+(a_2(b_0+b_1)-b_2(1+a_1))z^{-2}}{1+a_1z^{-1}+a_2z^{-2}}. \qquad (6)$$

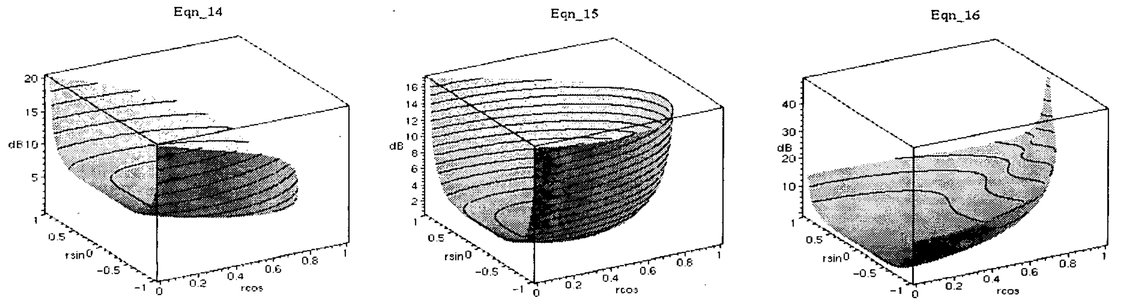

**Figure 2.** Plots of equations (14), (15) and (16) on the right half of the unit circle up to a radius of 0.99.

for the norms as functions of poles. Application of Parseval's theorem and Cauchy's residue theorem [10] gives

$$\left\|\tilde{G}_0\right\|_2^2 = \frac{2(r^2 + 2r\cos\theta - 3)}{(r^2 - 1)(1 + 2r\cos\theta + r^2)}, \qquad (14)$$

$$\left\|\tilde{G}_{1,3}\right\|_2^2 = \left\|\tilde{G}_5\right\|_2^2 = \frac{2}{(1 - r^2)(1 + 2r\cos\theta + r^2)}, \qquad (15)$$

$$\left\|\tilde{G}_{2,4}\right\|_2^2 = \left\|\tilde{G}_6\right\|_2^2 = \frac{1 + r^2}{(1 - r^2)(1 - 2r^2\cos 2\theta + r^4)}. \qquad (16)$$

From (14)-(16), the pole location dependence of the noise gain can be easily visualized. This is useful for providing a qualitative perspective for narrow-band lowpass filters design, as the poles will cluster towards unity in the complex $z$-plane when the sampling rate is raised. Fig. 2 shows the plots of (14)-(16) in a domain on the right half of the complex-plane unit circle up to a radius of 0.99. Now the effect of filter structure on the noise gain can easily be seen. Plot of (16) reveals that $e_2$, $e_4$ and $e_6$ see the maximum gain when the poles move towards unity due to the all-pole transfer functions (10) and (12). To maintain a desired signal-to-noise ratio, the word-length must be increased [see (2) and (3)]. Plot of (15) shows that $e_1$, $e_3$ and $e_5$ still see increasing gain around unity but the increase is much smaller due to the single zero in (9) and (11). Plot of (14) in fact shows a decreasing trend and a limiting gain of 3 dB at unity, thereby indicating that $e_0$ sees the minimum gain. This is accounted for by the double zeros in (8).

For the DFIIt structure in Fig. 1(a), all quantization noise sources $E_0$ through $E_4$ experience a gain with the form of (16) [10]. Even with two more sources $e_5$ and $e_6$ in the $\delta$DFIIt structure, its noise variance sum is still much lower than that of the DFIIt structure. By using double-precision in the $\delta^{-1}$ operators, $e_5$ and $e_6$ can further be eliminated. Therefore it is qualitatively verified that $\delta$DFIIt is superior to DFIIt in terms of roundoff noise gain.

## 4. NOISE MINIMIZATION

This section performs the minimization of (13) under the scaling constraint of (7). First, because (7) contains two free variables, $\Delta_1$ and $\Delta_2$, a trick is made to transform them into one common variable $\Delta$ by putting

$$\Delta_1 = \Delta k_1^{-1}\,,\ \Delta_2 = \Delta k_2^{-1}. \qquad (17)$$

The scaling factor in (7) becomes

$$g = \max\left(\left\|F_0\right\|_\infty, \left\|\tilde{F}_1\right\|_\infty k_1\Delta^{-1}, \left\|\tilde{F}_2\right\|_\infty k_1k_2\Delta^{-2}\right). \qquad (18)$$

This expression is a max-function dependent on $\Delta$ and there are three possible regions of $\Delta$ corresponding to the three possible maxima. Using a similar approach as in [4], each argument in (18) is set to be the maximum and solved for its valid region, three regions are obtained, namely

- Region 1 where $g = \left\|F_0\right\|_\infty$

$$\infty \geq \Delta \geq \max\left(k_1\left\|\tilde{F}_1\right\|_\infty / \left\|F_0\right\|_\infty, \sqrt{k_1k_2\left\|\tilde{F}_2\right\|_\infty / \left\|F_0\right\|_\infty}\right). \qquad (19)$$

Substituting $g$ into (13) results in an expression with positive powers of $\Delta$, thus noise is minimized by choosing $\Delta$ to be the lower bound.

- Region 2 where $g = \left\|\tilde{F}_1\right\|_\infty k_1\Delta^{-1}$

$$k_1\left\|\tilde{F}_1\right\|_\infty / \left\|F_0\right\|_\infty \geq \Delta \geq k_2\left\|\tilde{F}_2\right\|_\infty / \left\|\tilde{F}_1\right\|_\infty. \qquad (20)$$

In this case, the optimum $\Delta$ that minimizes the gain ratio lies somewhere within the region.

- Region 3 where $g = \left\|\tilde{F}_2\right\|_\infty k_1k_2\Delta^{-2}$

$$\min\left(\sqrt{k_1k_2\left\|\tilde{F}_2\right\|_\infty / \left\|F_0\right\|_\infty}, k_2\left\|\tilde{F}_2\right\|_\infty / \left\|\tilde{F}_1\right\|_\infty\right) \geq \Delta \geq 0. \qquad (21)$$

Using similar argument, the optimum $\Delta$ in this region occurs at the upper bound.

Previous analyses [3], [4] implicitly set $k_1 = k_2 = 1$, and the optimum was chosen by comparing the effect of the local minima in different regions on the noise gain. In [3] and [4], the optimal $\Delta$ is given by

$$\Delta_{op} = \max\left(\left\|\tilde{F}_1\right\|_\infty / \left\|F_0\right\|_\infty, \sqrt{\left\|\tilde{F}_2\right\|_\infty / \left\|F_0\right\|_\infty}\right). \qquad (22)$$

In our approach, the sub-optimals in the three regions are brought to converge to a global optimal on the real number line by setting

| | | |
|---|---|---|
| $g=\left\|F_0\right\|_\infty$ | $\alpha_1=(2+a_1)\left\|F_0\right\|_\infty/\left\|\tilde{F}_1\right\|_\infty$ | $\alpha_2=(1+a_1+a_2)\left\|F_0\right\|_\infty/\left\|\tilde{F}_2\right\|_\infty$ |
| $\beta_0 g^{-1}=b_0/\left\|F_0\right\|_\infty$ | $\beta_1 g^{-1}=(2b_0+b_1)/\left\|\tilde{F}_1\right\|_\infty$ | $\beta_2 g^{-1}=(b_0+b_1+b_2)/\left\|\tilde{F}_2\right\|_\infty$ |
| $\sigma_o^2/\sigma_e^2=\left(\left\|\tilde{G}_0\right\|_2^2+\left\|\tilde{G}_5\right\|_2^2\right)\left\|F_0\right\|_\infty^2+\left(\left\|\tilde{G}_6\right\|_2^2+2\left\|\tilde{G}_{1,3}\right\|_2^2\right)\left\|\tilde{F}_1\right\|_\infty^2+2\left\|\tilde{G}_{2,4}\right\|_2^2\left\|\tilde{F}_2\right\|_\infty^2$ | | |

**Table 1.** Various expressions for the $\delta$DFIIt structure.

| Polynomial | Section A1 | Section A2 | Section A3 | Section B1 | Section B2 | Section B3 |
|---|---|---|---|---|---|---|
| Denominator $a_1,a_2$ | -1.93504729<br>0.96471582 | -1.86611453<br>0.88788503 | -1.80612859<br>0.81824041 | -1.99512547<br>0.99610130 | -1.98883573<br>0.98938327 | -1.98540165<br>0.98552386 |
| Numerator $b_1$ ($b_0$=$b_2$=1) | -1.25901348 | -1.87112896 | -1.92379959 | 2 | 2 | 2 |
| Noise gain in dB using (22) | 15.0978 | 10.7450 | 8.7220 | 24.1235 | 19.9887 | 19.0695 |
| Noise gain in dB using (24) | 15.0430 | 9.9691 | 7.0330 | 24.1109 | 19.7033 | 17.4691 |
| Improvement in dB | 0.05483 | 0.7759 | 1.6890 | 0.01263 | 0.2854 | 1.6005 |

**Table 2.** Noise gain comparison for two example sixth-order lowpass filters.

$$\Delta_{op}^{*}=k_1\left\|\tilde{F}_1\right\|_\infty/\left\|F_0\right\|_\infty=k_2\left\|\tilde{F}_2\right\|_\infty/\left\|\tilde{F}_1\right\|_\infty . \qquad (23)$$

From (17), the optimal $\Delta_1$ and $\Delta_2$ (denoted by asterisks) are

$$\Delta_1^{*}=\left\|\tilde{F}_1\right\|_\infty/\left\|F_0\right\|_\infty \text{ , } \Delta_2^{*}=\left\|\tilde{F}_2\right\|_\infty/\left\|\tilde{F}_1\right\|_\infty . \qquad (24)$$

With (24), optimal filter coefficient expressions and the noise variance in the proposed $\delta$DFIIt section [Fig. 1(c)] can be obtained (see Table 1). For efficient hardware implementation, the optimized $\Delta$s can be rounded to the nearest powers of two such that scaling can be accomplished by simple bit shifting.

## 5. NUMERICAL EXAMPLES

Several second-order sections are taken from [4] to test against the proposed approach. Sections $A\,i$ and sections $B\,i$ ($i$=1,2,3) are cascade sections for two sixth-order narrow-band lowpass filters. Each section is pre-scaled using its $L_\infty$-norm, with scaling embedded into the numerator coefficients. To eliminate the effect of word precision on the output noise variance, the word-length-independent noise gain $\sigma_o^2/\sigma_e^2$ is evaluated for each section. Two approaches, namely the $\Delta_{op}$ in (22) and the proposed $\Delta_1^*$ and $\Delta_2^*$ in (24), are tested. From Table 2, it is clear that there is always reduction in roundoff noise gain in the proposed structure. When each three sections are cascaded, in whichever order, the total reduction in noise gain will further be improved. In fact, as noted in Section 3, even bigger improvement can be observed when the pole angles in a second-order section come closer to zero (e.g., in the case of very narrow-band filters). From the viewpoint of circuit implementation, a fixed $\Delta$ may save complexity in DSP coding. However, for ASIC design, the advantage of separate $\Delta$ optimization over a fixed $\Delta$ is certainly worthwhile.

## 6. CONCLUSION

In this paper the concept of optimizing different $\Delta$s in a second-order $\delta$DFIIt section has been introduced. This approach leads to further minimization of output roundoff noise gain as compared to using a single optimal $\Delta$ only. The theoretical minimum noise gain and expressions for the optimal filter coefficients have been derived analytically. Qualitative graphical noise analysis has been presented. Using numerical examples, roundoff noise performance of this modified structure has been demonstrated to be better than the best results obtained so far.

## 7. REFERENCES

[1] R. H. Middleton and G. C. Goodwin , *Digital Control and Estimation: A Unified Approach.* Englewood Cliffs, NJ: Prentice Hall, 1990.

[2] G. C. Goodwin, R. H. Middleton, and H. V. Poor, "High speed digital signal processing and control", *Proc. IEEE*, Vol. 80, pp. 240-259, Feb. 1992.

[3] J. Kauraniemi, T. I. Laakso, I. Hartimo, and S. J. Ovaska, "Roundoff Noise Minimization in a Direct Form Delta Operator Structure", in *Proceedings of 1996 International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, Georgia, USA, May 1996.

[4] —, "Delta Operator Realizations of Direct-Form IIR Filters", *IEEE Transactions on Circuits and Systems-II.*, Vol. 45, No. 1, pp. 41-52, Jan 1998.

[5] J. Kauraniemi and T. I. Laakso, "Roundoff Noise Analysis of Modified Delta Operator Direct Form Structures", *IEEE International Symposium on Circuits and Systems*, 1997.

[6] M. Eraluoto, J. Kautraniemi, I. Hartimo, "VLSI Implementation of High Speed Digital Filters Using Direct Form Delta Structures", *IEEE 39th Midwest symposium on Circuits and Systems*, 1996.

[7] G. Li and M. Gevers, "Roundoff Noise Minimization Using Delta Operator Realizations", *IEEE Trans. Signal Processing*, Vol 41, pp. 629-637, Feb. 1993.

[8] —, "Comparative Study of Finite Wordlength Effects in Shift and Delta Operator Parameterizations", *IEEE Trans. Autom. Contr.*, Vol. 38, pp. 803-807, May 1993.

[9] L. B. Jackson, *Digital Filters and Signal Processing*, 2nd ed., Boston, MA: Kluwer Academic Publishers, 1989.

[10] A. Oppenheim, R. Shaffer, *Discrete-Time Signal Processing*, Prentice-Hall, New Jersey, 1989.