Controlled rounding in low noise digital filter structures

Jean H.F. Ritzerfeld[1], Guergana S. Mollova[2]
[1] Technical University Eindhoven,
Department of Electrical Engineering,
P.O. Box 513, 5600 MB, Eindhoven The Netherlands
tel. +31 (0)40 247 3252 fax. +31 (0)40 244 8375
j.h.f.ritzerfeld@ele.tue.nl

[2] UACG-Sofia, Dept. Computer Aided Design,
Sofia, Bulgaria
mollov@vmei.acad.bg

Abstract— **Several IIR digital filter structures are known which exhibit freedom of limit cycles with magnitude truncation (MT) as rounding mechanism. Other well-known structures provide low noise properties with respect to quantization. Often, nonlinear stability and quantization noise reduction are incompatible. Integrator-based structures, for example, which are common for narrowband lowpass filter applications, show an excellent low noise behaviour but they are not free from limit cycles. With controlled rounding (CR), the direction of quantization is made dependent on some suitably chosen control signal within the structure. Such a quantization mechanism has been shown to stabilize the direct form structure [1] and is easy to implement: simply subtract the control signal from the signal to be quantized, truncate the difference (MT), and subsequently add the control signal. The more difficult part is to find a suitable control signal. This contribution describes how controlled rounding can be used to suppress limit cycles and how a control signal can be found, first in general, then applied to the well-known integrator-based structure by Agarwal and Burrus. The main result of the paper are two filter structures which are free from limit cycles, but which still combine low noise properties with a (near) minimum number of multipliers.**

Keywords— **Filter structures, Controlled rounding, Limit cycles, Noise.**

I. Introduction

A well-known recipe to determine if a second-order IIR digital filter structure is free from zero-input limit cycles is as follows.

- Write the state transition in matrix notation: $\underline{x}[n+1] = f\{\boldsymbol{A}\underline{x}[n]\}$, where $\underline{x} = (x_1\ x_2)^t$ is the state vector, $\boldsymbol{A}$ is the state matrix and f is the quantization nonlinearity.
- Define an energy function P of the state, i.e. a quadratic form $\underline{x}^t\boldsymbol{Q}\underline{x}$, where $\boldsymbol{Q}$ is a positive definite matrix. Note that P is constant on ellipses in the x_1, x_2-plane.
- Check if the (linear) state transition given by $\boldsymbol{A}$ decreases the energy P. This will be the case if $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$ is a positive definite matrix.
- Check if the nonlinearity f decreases energy. This can be done graphically, by inspection: quantization must map a point on an ellipse $P =$ constant to a point *inside* the ellipse, or, equivalently, to a point on an ellipse with lower energy.

Linear *and* nonlinear energy decrease is a sufficient condition for zero-input stability, i.e. freedom from limit cycles. Since P is a positive function for $\underline{x} \neq \underline{0}$, the only way it can become zero is when the zero-state is reached. Using magnitude truncation (MT) for quantization, nonlinear energy decrease is ensured, but still only for a horizontal or vertical orientation of the set of ellipses $P =$ constant, as can easily be seen by graphical inspection. This means that $\boldsymbol{Q}$ must be chosen as a (positive) *diagonal* matrix. The simultaneous condition to be met is that of positivity of the matrix $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$. The optimal choice for $\boldsymbol{Q}$ is found to be $\boldsymbol{Q} = \begin{pmatrix} |a_{21}| & 0 \\ 0 & |a_{12}| \end{pmatrix}$, where the diagonal entries denote elements of the state matrix. This choice for $\boldsymbol{Q}$ leads to the least restrictive condition for the state

matrix, reading [2]:

$$|a_{11} - a_{22}| < 1 - \det(\boldsymbol{A}). \quad (1)$$

So, second-order IIR filters meeting this condition are free from zero-input limit cycles if MT quantization is applied at both states.

Looking, for example, at the direct form structure with $\boldsymbol{A} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}$, we see that *linear* stability requires the parameters a and b to lie within the stability triangle ($|a| < 1-b$ and $b > -1$) to ensure that the zeros of the denominator polynomial $z^2 - az - b$ are within the unit circle, whereas *nonlinear* stability, as given by (1), reduces this area to the square $|a| + |b| < 1$. This means that the poles are allowed only to lie within the area bounded by two circle arcs $(|\sigma| + 1)^2 + \omega^2 < 2$, where σ and ω are real and imaginary parts of the poles. In applications where nonlinear stability is vital, the direct form is a very poor choice for a filter structure.

Historically, the stability problem was solved with the introduction of alternative structures such as the *normal* form with $\boldsymbol{A} = \begin{pmatrix} \sigma & -\omega \\ \omega & \sigma \end{pmatrix}$,and the *wave digital* form with $\boldsymbol{A} = \frac{1}{2}\begin{pmatrix} 1+a+b & -1+a+b \\ 1+a-b & -1+a-b \end{pmatrix}$.

In filter design, however, other considerations may come into play, such as demands on the number of multiplications and the level of quantization noise. In this respect, the normal form is non-canonical (it uses four instead of the minimum number of two multipliers), whereas the wave digital form, though having far better noise properties than the direct form, leaves room for improvement for extreme pole locations near the point $z = 1$. This led to the introduction of special structures for narrow-band lowpass filters, used for example as decimation filters in multirate digital signal processing. These structures are based on the concept of translating the origin of the z-plane to the point $z = 1$. In other words, the basic element z^{-1} is replaced by the integrator $(z-1)^{-1}$. Doing so for the direct form structure, the parameters a and b must be replaced by $a-2$ and $-(1-a-b)$ in order not to change the denominator $z^2 - az - b$ of the transfer function [3]. The resulting form, which we will designate *Agarwal-Burrus* structure, exhibits low quantization noise for extreme pole locations near $z = 1$. The state matrix of this form is

$$\boldsymbol{A} = \begin{pmatrix} a-1 & -(1-a-b) \\ 1 & 1 \end{pmatrix}, \quad (2)$$

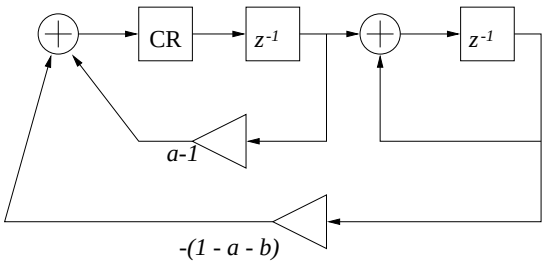

Fig. 1. Agarwal-Burrus structure.

where the entry $a-1$ must be read as $1+(a-2)$, if we want to distinguish between the part of the integrator on state x_1 and the feedback coefficient from state x_1. The problem with this form lies in its nonlinear stability. Since $a < 2$, the condition given by (1) reads $2-a < 1+b$, whereas within the stability triangle we have $2-a > 1+b$ (or $1-a-b > 0$). So, the Agarwal-Burrus structure is instable in the nonlinear sense for any point (a, b) within the triangle of linear stability. Clearly, we need some mechanism other than MT to quantize state x_1 in such a way that a suitably chosen energy function will always decrease. (Note that x_2 does not need to be quantized at all, since $x_2[n+1]$ is the sum of $x_1[n]$ and $x_2[n]$.) With controlled rounding (CR), the direction of quantization is made dependent on some signal within the structure. CR has been shown to (almost) stabilize the direct form structure for all points (a, b) within the stability triangle [1]. In the next section we will see how CR can be used to stabilize the Agarwal-Burrus structure (Fig. 1).

II. Controlled rounding

First, we demonstrate the use of CR for the direct form. We know that a diagonal matrix for $\boldsymbol{Q}$ (which is necessary in case of MT) restricts the state matrix too much if positivity of $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$ is to be met. An orientation of the set of ellipses P = constant that closely resembles that of the natural response of the direct form structure is along a 45 degree axis. Such an orientation is found when $\boldsymbol{Q}$ is non-diagonal with equal elements q_{11} and q_{22}. For instance, if we choose

$$\boldsymbol{Q} = \begin{pmatrix} 1-b & -a \\ -a & 1-b \end{pmatrix}, \quad (3)$$

we find that

$$\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A} = (1+b)\begin{pmatrix} a^2 & -a(1-b) \\ -a(1-b) & (1-b)^2 \end{pmatrix}. \quad (4)$$

Note that $\boldsymbol{Q}$ is positive definite for all (a, b) within the stability triangle, whereas $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$ is only positive semi-definite since its determinant is zero. So, albeit only marginally, the first three items on our recipe of the previous section are dealt with. All that remains is to ensure that quantization lowers energy. In order to do so, we need to round x_1 in the direction of the 45 degree axis of the ellipses of constant energy. (Note that, again, x_2 does not need to be quantized.) In other words, given the unquantized new state $(ax_1[n] + bx_2[n], x_1[n])$, we round x_1 such that the difference $x_1 - x_2$ decreases. This means that $x_2[n+1]$ can serve as a control signal to determine the direction of quantization, hence the designation *controlled* rounding. To implement this rounding mechanism, we truncate the (new) difference $x_1 - x_2$ and subsequently add the control signal. Note that we need to decrease the absolute difference, so we use MT (Fig. 2). In the figure the control signal is designated s. We see that s is first subtracted before truncation, hence the parameter $a - 1$ instead of a, and immediately added again after the quantization point. Unfortunately, the resulting structure is not fully stable, since it can still exhibit limit cycles of period 1 (i.e. constants). The reason behind this is somewhat involved and is beyond the scope of this short paper. Suffice it to say, that rounding x_1 in the direction of x_2 is literally inviting limit cycles of period 1 to occur in a structure where the new x_2 is the current x_1. The *transpose* version of this direct form with state matrix $\begin{pmatrix} a & 1 \\ b & 0 \end{pmatrix}$, however, does not have this problem and can indeed be fully stabilized.

Generalizing the idea behind controlled rounding, we may take the following steps.

- Find an energy function $P = \underline{x}^t\boldsymbol{Q}\underline{x}$ that minimally restricts the state matrix to meet the condition of positivity of $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$. A good candidate for $\boldsymbol{Q}$ is found when the orientation of the ellipse-set P = constant closely resembles that of the natural response of the structure. How to find such a candidate for a general second-order structure is well-documented [4].
- Determine the slope r of the line of orientation: $x_2 = rx_1$. Now we have two choices: if CR is to be applied to state x_2, use rx_1 as a control signal, i.e. round x_2 in the direction of rx_1; if CR is to be applied to state x_1, use x_2/r as a control signal. In both cases, the other state is simply truncated (MT), or, as we have seen, may not need to be quantized at all.
- Implement the CR mechanism with control signal s by subtracting s from the signal to be rounded, applying MT to the difference and subsequently adding s to form the quantized state. Of course this implies that s in itself need not be quantized, or, in other words, either r or $1/r$ has to be integer. If this is not the case, the control signal must first be rounded.

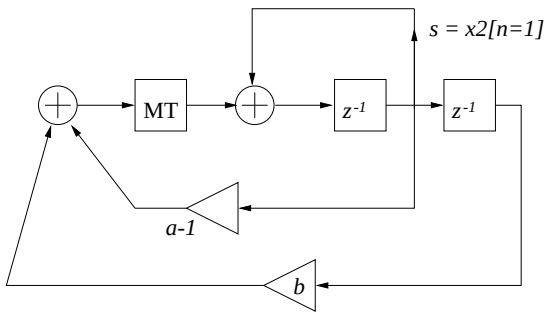

Fig. 2. Direct form structure with controlled rounding.

Next, we arrive at the use of CR for the Agarwal-Burrus structure. Suitable candidates for $\boldsymbol{Q}$ are:

$$\begin{pmatrix} 2 & 2-a \\ 2-a & 2(1-a-b) \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} -b & 1-a-b \\ 1-a-b & 1-a-b \end{pmatrix} \tag{5}$$

The first matrix is positive definite for all *complex* pole locations ($-4b - a^2 > 0$), whereas the second matrix is positive definite in the region $a > 1$ of the stability triangle. Either area is valid, since this structure is meant for narrow-band lowpass filters ($a > 1$ corresponds to poles with a real part greater than $\frac{1}{2}$.) Both candidates for $\boldsymbol{Q}$ meet the condition of positivity of $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$ in the areas mentioned. Determining the orientation of the ellipse-set, we find a slope r that is very close to $-1/(1-a-b)$ for both matrices $\boldsymbol{Q}$, the approximation getting better the closer the poles are to $z = 1$. For such extreme pole locations, r is a large negative number. Since we want to apply CR to x_1, we have a small control signal $s = x_2/r = -(1-a-b)x_2$. If we round s in order to be able to implement the CR mechanism, we must take care that small, but vital values of s are not lost, i.e. rounded off to zero. Simulations show that we even need to round upwards, i.e. *increase* the magnitude of s to $t = \text{sign}(s)\ \text{ceiling}(|s|)$ (Fig. 3).

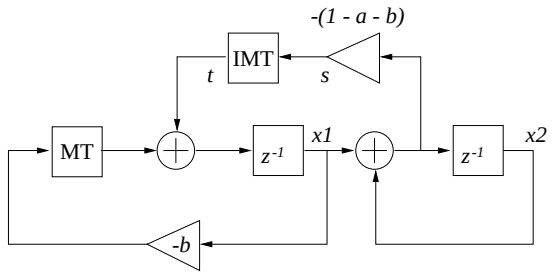

Fig. 3. Unscaled Agarwal-Burrus structure with CR.

In the figure we designated this operation as IMT for 'inverse' MT. Note that, again, the *new* state x_2 must be used in the control signal, since we want to control the rounding of the *new* state x_1. Since $x_2[n+1] = x_1[n] + x_2[n]$, this means that the feedback coefficient $-(1-a-b)$ from x_2 (cf. Fig. 1) is cancelled at the subtraction point of the control signal, and that the feedback coefficient $a-1$ from x_1 changes into $-b$. The resulting structure is very simple indeed, and it still uses only two multipliers.

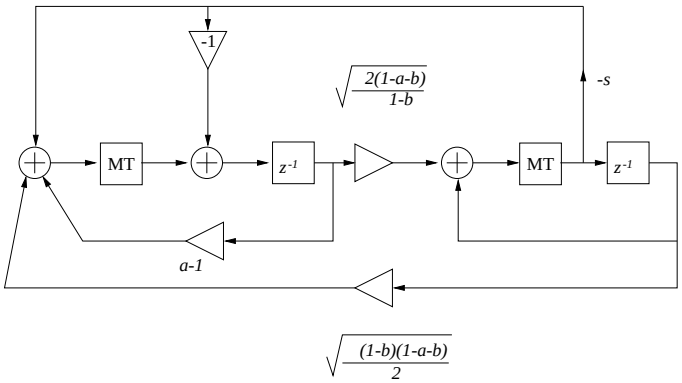

Fig. 4. Scaled Agarwal-Burrus structure with controlled rounding.

III. Scaling

When we look at the natural response of the Agarwal-Burrus structure in the (x_1, x_2)-plane, we see that the dynamic range of state x_1 is much smaller than that of state x_2, the ratio being approximately $\sqrt{1-a-b}$. This means that we can improve on the noise behaviour even more by proper scaling of the structure. In practice, scaling is applied to the system in the presence of an input signal. If the input $u[n]$ is supplied to $x_1[n+1]$, we have the state equation

$$\underline{x}[n+1] = \boldsymbol{A}\underline{x}[n] + \boldsymbol{B}u[n], \qquad (6)$$

with $\boldsymbol{B} = (1 \;\; 0)^t$. Then, if we were to take x_1 or x_2 as output signal, we would have the transfer functions

$$H_1(z) = \frac{z-1}{z^2-az-b}$$

and

$$H_2(z) = \frac{1}{z^2-az-b},$$

respectively. So, by suitable linear combination of u, x_1 and x_2, the numerator of a general second-order transfer function can be made. An l_2-scaling matrix $\boldsymbol{T}$ can now be found if we determine the controllability matrix (or covariance matrix of the state), which is defined recursively as $\boldsymbol{K} = \boldsymbol{A}\boldsymbol{K}\boldsymbol{A}^t + \boldsymbol{B}\boldsymbol{B}^t$. Solving for $\boldsymbol{K}$ we find

$$\boldsymbol{K} = \gamma \begin{pmatrix} 2(1-a-b) & -(1-a-b) \\ -(1-a-b) & 1-b \end{pmatrix}, \qquad (7)$$

where $\gamma = \{(1+b)(1-a-b)(1+a-b)\}^{-1}$. Note that $\gamma = \infty$ on the boundary of the stability triangle. The scaling matrix $\boldsymbol{T} = \mathrm{diag}\{\sqrt{k_{11}}, \sqrt{k_{22}}\}$ transforms the state matrix to its scaled form $\boldsymbol{A'} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$:

$$\boldsymbol{A'} = \begin{pmatrix} a-1 & -\sqrt{\frac{(1-b)(1-a-b)}{2}} \\ \sqrt{\frac{2(1-a-b)}{1-b}} & 1 \end{pmatrix}. \qquad (8)$$

Next, we want to apply CR to this scaled version of the Agarwal-Burrus structure. First, we note that we need an extra (non-integer) multiplier for the state matrix. This means that both states have to be quantized, and that we can choose whether to apply CR to x_1 or x_2. Indeed, both choices are feasible. Here, we will take x_1 as a candidate for CR and apply MT to x_2. In order to determine the control signal, we need to find a suitable energy function and the orientation of the curves of constant energy. Since scaling should equalize the dynamic ranges of the two state variables, we expect a -45 degree orientation of the natural response. Indeed, proper scaling should always produce either circles of constant energy or ellipses along a 45 or -45 degree axis. In [4] a closed form expression for $\boldsymbol{Q} = \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix}$ is derived, such that $\boldsymbol{Q}$ and $\boldsymbol{Q} - \boldsymbol{A}^t\boldsymbol{Q}\boldsymbol{A}$ are positive definite on the whole area of the stability triangle for a general second-order structure that is properly scaled. The parameter q is given by

$$q = \mathrm{sign}\{a_{12}(a_{11}-a_{22})\} \frac{|a_{12}-a_{21}||a_{11}-a_{22}|}{\{1-\det(\boldsymbol{A})\}^2 - 4a_{12}a_{21}}, \qquad (9)$$

yielding a ± 45 degree orientation of the ellipse-set $P = \underline{x}^t \boldsymbol{Q} \underline{x} =$ constant, depending on the sign of q. For the Agarwal-Burrus structure we find a positive q, hence a -45 degree axis as expected, since both a_{12} and $a_{11} - a_{22}$ are negative. This means that we can use $-x_2[n+1]$ as a control signal for the rounding of x_1. Fig. 4 shows the resulting scaled version of the structure.

The beauty of scaling is that we can apply CR to any second-order structure without actually having to determine a suitable energy function. We only have to know if its line of orientation has a slope of $+1$ or -1 by checking the sign of $a_{12}(a_{11} - a_{22})$. If $r = 1$ we take $x_2[n+1]$ or $x_1[n+1]$ as a control signal, depending on whether we want to apply CR to x_1 or x_2, resp. If $r = -1$ we have negative (new) states as control signals. Sometimes, we do not need to perform scaling, because the dynamic ranges of the two state variables are inherently equal (like in the direct form), or are close to being equal (like in the transpose version of the direct form). A simple criterion to check whether an energy function of slope ± 1 is valid is given in [4]:

$$1 - \det(\boldsymbol{A}) \geq |a_{12} + a_{21}|. \tag{10}$$

For example, since the transpose direct form meets this condition with equality, a control signal $\pm x_1[n+1]$ for CR on x_2 should work. Indeed, further analysis shows that the choice

$$\boldsymbol{Q} = \begin{pmatrix} 1-b & a \\ a & 1-b \end{pmatrix},$$

consistent with (9), leads to

$$\boldsymbol{Q} - \boldsymbol{A}^t \boldsymbol{Q} \boldsymbol{A} = (1+b)(1-a-b) \cdot (1+a-b) \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

which is positive semi-definite. For $a > 0$ we need to use $-x_1[n+1]$ as control signal for the rounding of x_2. Fig. 5 shows this alternative direct form with CR applied to x_2. In contrast with the original solution of Fig. 2, this structure is fully stable.

IV. Conclusion

With controlled rounding we have stabilized the Agarwal-Burrus structure, thereby retaining its low noise and low sensitivity properties. We have given two new structures, one scaled and one unscaled, that are free from zero-input limit cycles. We have shown how CR can be used in general to stabilize any second-order IIR filter section. Specifically, we have seen how scaling greatly simplifies the finding of a suitable control signal and the implementation of the CR mechanism. Determining whether scaling is necessary is done by checking a simple condition on the state matrix. In fact, any structure meeting this condition can very easily be stabilized, with the exception of a few pathological cases. The direct form II, for example, cannot be fully stabilized, whereas its transpose version (direct form I) can. This stable direct form structure is quite unique and has, to the knowledge of the authors, never been published before.

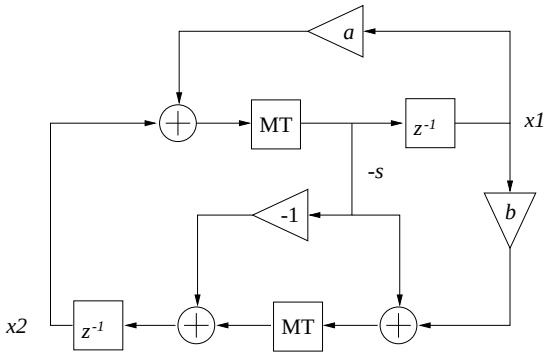

Fig. 5. Transpose direct form structure with CR.

References

[1] H.J. Butterweck, "Suppression of parasitic oscillations in second-order digital filters by means of a controlled-rounding arithmetic," *Arch. El. Übertragungstechnik*, vol. AEÜ-29, pp. 371-374, 1975.

[2] W.L. Mills, C.T. Mullis and R.A. Roberts, "Digital filter realizations without overflow oscillations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 334-338, 1978.

[3] R.C. Agarwal and C.S. Burrus, "New recursive digital filter structures having very low sensitivity and roundoff noise," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 921-927, 1975.

[4] J.H.F. Ritzerfeld, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures using saturation," *IEEE Trans. Circuits Syst.*, vol. CAS-36, pp. 1049-1057, 1989.