

# The Direct Wave Form Digital Filter Structure: an Easy Alternative for the Direct Form

Jean H.F. Ritzerfeld

Technische Universiteit Eindhoven, Fac. Elektrotechniek

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Phone: +31 (0)40 247 3252 Fax: +31 (0)40 246 6508

E-mail: [j.h.f.ritzerfeld@tue.nl](mailto:j.h.f.ritzerfeld@tue.nl)

*Abstract*—Although the Direct Form is still widely used in IIR digital filter design, this structure is known to have a high coefficient sensitivity and to produce high levels of quantization noise. Even if only second-order sections are used to design higher order filters, the sections with poles close to the unit circle and/or with angles close to 0 or  $\pi$  have unacceptably high noise levels, specifically in fixed-point applications. As an easy alternative for the second-order Direct Form we present the second-order Direct Wave Form. Comparable to the Direct Form, this structure uses only five multipliers which are directly linked to the five coefficients of a general second-order transfer function. It is shown how the Direct Wave Form can be properly scaled with an  $L_2$  scaling measure as well as with the more conservative  $L_\infty$  scaling measure. It is then shown, by looking at the second-order modes of a general second-order transfer function that the Direct Wave Form has a superior noise behaviour that is close to being optimal (within 3dB). The Direct Wave Form is also shown to be overflow stable and free from limit cycles.

*Keywords*—Digital filter design; wave digital filters; quantization noise; scaling, second-order modes.

## I. INTRODUCTION

When designing higher order IIR filters, it is common practice to implement them as a cascade or parallel connection of second-order sections with transfer functions of the general form

$$H(z) = \frac{dz^2 + ez + f}{z^2 - az - b}, \quad (1)$$

having five degrees of freedom. These sections are then implemented as second-order Direct Forms where the five degrees of freedom appear as multipliers in the signal flow graph. For those sections, however, that have poles close to the unit circle and/or with angles close to 0 or  $\pi$ , the coefficient sensitivity and the quantization noise become unacceptably high, specifically in fixed-point applications. Alternative solutions are then looked for [1], such as the Normal Form and the Optimal (i.e. minimum-noise) Form, both of which use extra degrees of freedom to achieve a better sensitivity/noise behaviour at the cost of extra mul-

tipliers and a less straightforward design, since the one-on-one relation between the values of the multipliers and the coefficients of the transfer function is lost.

In Section II the Direct Wave Form (DWF) is presented as an easy alternative for the Direct Form that is as straightforward to design and uses no extra multipliers. The DWF has a superior coefficient sensitivity and a noise behaviour that is close to that of the Optimal Form. In order to scale the structure to prevent overflow, two scaling multipliers are introduced, the values of which are given in closed form using an  $L_2$  (Section II) and an  $L_\infty$  scaling measure (Section IV). These extra multipliers can be given power-of-two values (to be implemented as shift operations), so the number of multipliers need not increase. As a quality measure for the noise behaviour the noise gain of the  $L_2$ -scaled structure is used, i.e. the scaled power gain of the quantization noise at its source ( $q^2/12$ ) to the output. In Section III a simple proof that uses the second-order modes of the general  $H(z)$  is given to show that the DWF has a noise gain that cannot exceed the minimum noise gain of the Optimal Form by more than a factor two.

## II. THE DIRECT WAVE FORM

In order to arrive at the Direct Wave Form, we start with Direct Form II (DF2) that has a state-space description

$$\underline{s}[n+1] = \mathbf{A}\underline{s}[n] + \mathbf{B}x[n], \quad y[n] = \mathbf{C}\underline{s}[n] + \mathbf{D}x[n],$$

where  $\underline{s} = (s_1, s_2)^t$  is the state vector and the four matrices are  $\mathbf{A} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix}$ ,  $\mathbf{B} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ,  $\mathbf{C} = (e + ad, f + bd)$ ,  $\mathbf{D} = (d)$ , to realize the transfer function given by (1).

For use in our noise calculations, we also look at the controllability matrix  $\mathbf{K} = \sum_{k=0}^{\infty} (\mathbf{A}^k \mathbf{B})(\mathbf{A}^k \mathbf{B})^t$  and the observability matrix  $\mathbf{W} = \sum_{k=0}^{\infty} (\mathbf{C} \mathbf{A}^k)^t (\mathbf{C} \mathbf{A}^k)$ . On the principal diagonals of these matrices are the power gains from input to state and from state to output, respectively (and the cross-power gains on the off-diagonals) [2]. The scaled noise gain, to be used as a quality measure, is then

$$G = K_{11}W_{11} + K_{22}W_{22}. \quad (2)$$

For DF2,  $\mathbf{K}$  and  $\mathbf{W}$  can be given in closed form [1]:

$$\mathbf{K} = \frac{\begin{pmatrix} 1-b & a \\ a & 1-b \end{pmatrix}}{(1+b)(1-a-b)(1+a-b)} \quad (3)$$

and  $\mathbf{W}$  has entries

$$W_{11} = \frac{(1-b)(c_1^2 + c_2^2) + 2ac_1c_2}{(1+b)\{(1-b)^2 - a^2\}}, \quad W_{22} = b^2W_{11} + c_2^2$$

$$W_{12} = W_{21} = \frac{ab(c_1^2 + c_2^2) + (1-a^2-b^2)c_1c_2}{(1+b)(1-a-b)(1+a-b)}, \quad (4)$$

where  $c_1$  and  $c_2$  are the elements of  $\mathbf{C} = (e + ad, f + bd)$ . Note, incidentally, that  $W_{11}$  is just equal to the quadratic form  $\mathbf{C}\mathbf{K}\mathbf{C}^t$ , and that  $W_{22}$  can also be written as  $W_{11} - \{(1-b)c_1 + ac_2\}^2 / \{(1-b)^2 - a^2\}$ , so  $W_{22} \leq W_{11}$ .

The Direct Wave Form arises from a state transformation (a rotation in the state plane) applied to DF2. It has a recursive part that is reminiscent of a wave digital filter [3], hence the designation. The state description follows from the matrices  $\mathbf{A}$ ,  $\mathbf{B}$ ,  $\mathbf{C}$ ,  $\mathbf{D}$  of DF2 via a similarity transformation  $\mathbf{A}_w = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ ,  $\mathbf{B}_w = \mathbf{T}^{-1}\mathbf{B}$ ,  $\mathbf{C}_w = \mathbf{C}\mathbf{T}$  and  $\mathbf{D}_w = \mathbf{D}$ , where  $\mathbf{T}$  is taken as  $\frac{1}{2} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ . We find:

$$\mathbf{A}_w = \begin{pmatrix} 1-\gamma_1 & -\gamma_2 \\ \gamma_1 & -1+\gamma_2 \end{pmatrix}, \quad \mathbf{B}_w = \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

$$\mathbf{C}_w = (\eta_1 - \gamma_1 d, \eta_2 - \gamma_2 d) \quad \text{and} \quad \mathbf{D}_w = (d), \quad (5)$$

where  $\gamma_1 = \frac{1}{2}(1-a-b)$ ,  $\gamma_2 = \frac{1}{2}(1+a-b)$ , and where  $\eta_1 = \frac{1}{2}(d+e+f)$ ,  $\eta_2 = \frac{1}{2}(d-e+f)$ . Fig. 1 depicts the Direct Wave Form, which is as yet unscaled and uses only five multipliers directly related to the five coefficients of  $H(z)$ , as does the Direct Form. Incidentally, apart from a factor  $\frac{1}{2}$ , the  $\gamma$ 's are simply found from evaluating the denominator of  $H(z)$  at  $z = \pm 1$ , whereas the  $\eta$ 's are found from the numerator at  $z = \pm 1$ .

In order to  $L_2$ -scale the DWF, we need to calculate its controllability matrix, which proves to be diagonal:

$$\mathbf{K}_w = \mathbf{T}^{-1}\mathbf{K}\mathbf{T}^{-t} = \frac{\begin{pmatrix} 4\gamma_2 & 0 \\ 0 & 4\gamma_1 \end{pmatrix}}{4\gamma_1\gamma_2(2-\gamma_1-\gamma_2)}, \quad (6)$$

where the denominator is the same as that of  $\mathbf{K}$  in (3), as governed by the stability triangle of a Direct Form in the  $(a, b)$ -plane:  $1+b > 0$ ,  $1-a-b > 0$ ,  $1+a-b > 0$ . For the DWF, the region of linear stability is also a triangle in the  $(\gamma_1, \gamma_2)$ -plane given by:  $\gamma_1 > 0$ ,  $\gamma_2 > 0$ ,  $2-\gamma_1-\gamma_2 > 0$ . In Fig. 2 the  $L_2$ -scaled DWF is drawn, where two extra multipliers  $1/\sqrt{K_{w11}}$  and  $1/\sqrt{K_{w22}}$  are introduced.

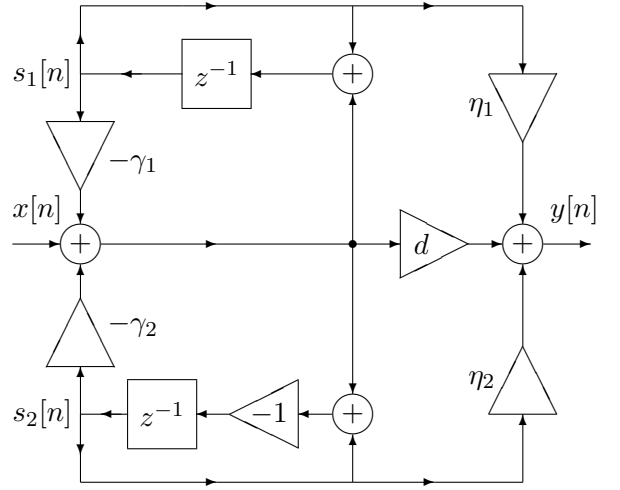


Fig. 1. Unscaled Direct Wave Form, where the multipliers are given by  $\gamma_1 = \frac{1}{2}(1-a-b)$ ,  $\gamma_2 = \frac{1}{2}(1+a-b)$ ,  $\eta_1 = \frac{1}{2}(d+e+f)$ ,  $\eta_2 = \frac{1}{2}(d-e+f)$ , realizing the general transfer function (1).

In essence, the scaled DWF of Fig. 2 is the same as the Wave Digital Form derived in [1]. The difference is that in the DWF the non-state node  $x[n] - \gamma_1 s_1[n] - \gamma_2 s_2[n]$  is used explicitly to create the output  $y[n]$  via the coefficient  $d$ , much in the same way as the non-state node  $s_1[n+1]$  in DF2 is connected to the multiplier  $d$  to compute  $y[n]$ . In this sense, the DWF is much more analogous to the Direct Form, hence the designation. Also, the  $\eta$ 's in Figs. 1 and 2 are more directly related to the coefficients of  $H(z)$  than the elements of  $\mathbf{C}_w$ , which would appear as multipliers if  $y[n]$  were computed in the normal 'state-space' way with  $d$  directly connecting  $x[n]$  and  $y[n]$ .

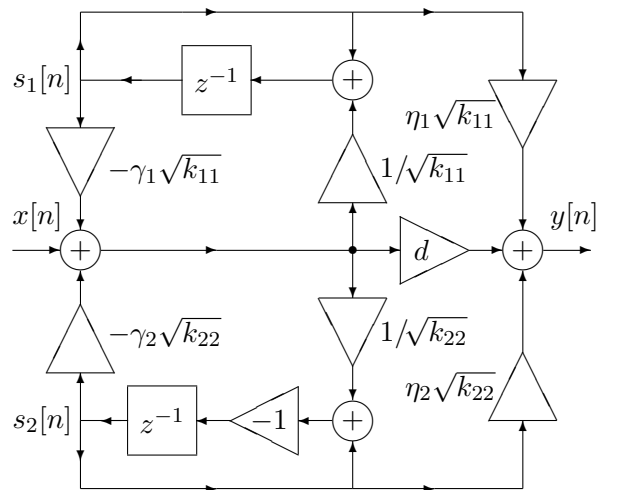


Fig. 2.  $L_2$ -scaled Direct Wave Form, where the extra scaling multipliers are given by  $1/\sqrt{K_{w11}} = \sqrt{\gamma_1(2-\gamma_1-\gamma_2)}$  and  $1/\sqrt{K_{w22}} = \sqrt{\gamma_2(2-\gamma_1-\gamma_2)}$ , as determined by  $\mathbf{K}_w$  in (6).

And finally, more so than the Wave Digital Form in [1], the DWF is a universal ‘biquad’, i.e. an IIR second-order section to create the standard filter types in a generic way. For example, taking  $\eta_1 = \eta_2 = 0$  creates a bandpass filter (zeros at  $\pm 1$ ), while high- and lowpass filters result from either  $\eta_1 = 0$ , or  $\eta_2 = 0$ . Next, if we take  $d = -\frac{1}{2}(1+b)$  in the bandpass solution and add  $x[n]$  to the output, a bandstop filter is created with transfer function

$$-\frac{1}{2}(1+b)\frac{z^2 - 1}{z^2 - az - b} + 1 = \frac{\frac{1}{2}(1-b)z^2 - az + \frac{1}{2}(1-b)}{z^2 - az - b}, \quad (7)$$

which has unity gain at the frequency edges  $\vartheta = 0$  and  $\vartheta = \pi$  and a stop frequency  $\vartheta = \arccos\{a/(1-b)\}$ . Likewise, taking  $d = -(1+b)$  and adding  $x[n]$  to the output results in a unity gain allpass filter:

$$-(1+b)\frac{z^2 - 1}{z^2 - az - b} + 1 = \frac{-bz^2 - az + 1}{z^2 - az - b}. \quad (8)$$

So, the filter types bandpass, bandstop and allpass use only three multipliers in the unscaled structure and five in the scaled version (since  $\eta_1 = \eta_2 = 0$ ).

### III. NOISE GAIN AND SECOND-ORDER MODES

The noise gain  $G_w = K_{w11}W_{w11} + K_{w22}W_{w22}$  of the DWF can be shown to be near-optimal without actually calculating it. Due to the fact that  $\mathbf{K}_w$  is diagonal, we can also write  $G_w = \text{tr}(\mathbf{K}_w \mathbf{W}_w)$ , where  $\text{tr}(\cdot)$  denotes the trace of a matrix, i.e. the sum of the elements on its principal diagonal. Since under a state transformation with any (regular) matrix  $\mathbf{T}$  the product matrix  $\mathbf{K}\mathbf{W}$  is subject to a similarity transformation  $\mathbf{T}^{-1}(\mathbf{K}\mathbf{W})\mathbf{T}$ , its eigenvalues (denoted  $\mu_1^2$  and  $\mu_2^2$ ), trace and determinant are invariant. The square roots  $\mu_1$  and  $\mu_2$  of these positive, real eigenvalues are called the second-order modes of  $H(z)$ , since they are determined only by the transfer function and do not change with any specific state-space realization. From [2] we know that the optimal gain is  $\frac{1}{2}(\mu_1 + \mu_2)^2$ , whereas  $G_w = \mu_1^2 + \mu_2^2$ , since the trace of a matrix is the sum of its eigenvalues. So  $G_w$  cannot exceed the optimal gain by more than a factor two, the worst case of 3dB occurring if  $\mu_1\mu_2 \downarrow 0$ , whereas  $G_w$  is optimal if  $\mu_1 = \mu_2$ .

While it is good to know that the DWF is near-optimal, it is still nice to have explicit expressions for its noise gain, as well as for the second-order modes. To that end we can use  $\mathbf{K}$  and  $\mathbf{W}$  of DF2 as given by (3) and (4). Starting with the noise gain,  $G_w$  is given by

$$\text{tr}(\mathbf{K}\mathbf{W}) = \mu_1^2 + \mu_2^2 = \frac{g_1c_1^2 + g_2c_2^2 + g_3c_1c_2}{(1+b)^2(1-a-b)^2(1+a-b)^2},$$

where

$$\begin{aligned} g_1 &= (1+b^2)\{(1-b)^2 - a^2\} + a^2(1+b)^2, \\ g_2 &= 2\{(1-b)^2 - a^2\} + a^2(1+b)^2, \\ g_3 &= 2a\{(1-b)^2 - a^2\} + 2a(1-b)(1+b)^2. \end{aligned} \quad (9)$$

Next, knowing the sum of the eigenvalues, we calculate their product in order to determine  $\mu_1$  and  $\mu_2$  separately:

$$\det(\mathbf{K}\mathbf{W}) = \mu_1^2\mu_2^2 = \frac{(-bc_1^2 + c_2^2 + ac_1c_2)^2}{(1+b)^4(1-a-b)^2(1+a-b)^2}, \quad (10)$$

where the algebra is not as cumbersome as it might seem, since  $\det(\mathbf{K}\mathbf{W}) = \det(\mathbf{K})\det(\mathbf{W})$  and  $\det(\mathbf{K})$  from (3) is simply  $(1+b)^{-2}(1-a-b)^{-1}(1+a-b)^{-1}$ . The surprisingly nice square numerator of (10) also allows us to take its square root for a compact expression for  $\mu_1\mu_2$ . The optimal gain  $\frac{1}{2}(\mu_1 + \mu_2)^2$  then is  $\frac{1}{2}G_w + \mu_1\mu_2$ , whereas  $\frac{1}{2}(\mu_1 - \mu_2)^2$  is  $\frac{1}{2}G_w - \mu_1\mu_2$ . Combining one with the other leads to closed-form expressions for  $\mu_1$  and  $\mu_2$ :

$$\mu_{1,2} = \frac{1}{2}\sqrt{G_w + 2\mu_1\mu_2} \pm \frac{1}{2}\sqrt{G_w - 2\mu_1\mu_2}. \quad (11)$$

To conclude this section, let us look at the case  $\mu_1 = \mu_2$  more closely, since the DWF is optimal in this case. Developing  $\mu_1 - \mu_2 = \sqrt{G_w - 2\mu_1\mu_2}$  leads to the strikingly simple result

$$\mu_1 - \mu_2 = \frac{|(1-b)c_1 + ac_2|}{(1-a-b)(1+a-b)}, \quad (12)$$

so  $\mu_1 = \mu_2$  if  $(1-b)c_1 + ac_2 = (1-b)e + a(d+f) = 0$ . The most useful way to do this, is to take  $e = 0$  and  $f = -d$ , or equivalently,  $\eta_1 = \eta_2 = 0$ . So, the DWF is optimal in case of a bandpass, a bandstop, or an allpass filter. Stated more generally, the DWF is optimal in case of a bandpass transfer function plus an arbitrary constant, which includes the bandstop case ( $d = f = \frac{1}{2}(1-b)$ ,  $e = -a$ ) and the allpass case ( $d = -b$ ,  $e = -a$ ,  $f = 1$ ). Incidentally, (12) holds only if  $-bc_1^2 + c_2^2 + ac_1c_2 > 0$ , which is true for any practical filter. First of all, for complex poles we have  $-b > a^2/4$ , so  $-bc_1^2 + c_2^2 + ac_1c_2$  is always positive and secondly, in the real-pole case, the value zero is crossed only if one of the zeros of  $H(z)$  crosses a pole on the real axis (so  $\mu_1\mu_2 \neq 0$  as long as the system is second-order). The resulting system has alternating real poles and zeros, and its second-order modes can no longer be equal.

If  $\mu = \mu_1 = \mu_2$ , the product matrix  $\mathbf{K}\mathbf{W}$  is diagonal, and so  $\mathbf{K}\mathbf{W} = \mu^2\mathbf{I}$ , where  $\mu = |d|/(1+b)$ , since with  $e = 0$  and  $f = -d$ , the matrix  $\mathbf{W}$  simplifies to

$$\mathbf{W} = \frac{d^2}{1+b} \begin{pmatrix} 1-b & -a \\ -a & 1-b \end{pmatrix} \quad (13)$$



