AN IMPROVED STABILITY MEASURE FOR DIGITAL FILTER IMPLEMENTATION

J.X. Hao and G. Li

School of EEE
Nanyang Technological University
Singapore 639798

ABSTRACT

In this paper, we consider the stability robustness problem of a digital filter implemented with finite word length (FWL). Based on the pole modulus sensitivities, a new stability robustness related measure is derived. This measure is less conservative than that derived with the classical pole sensitivity measure. It is shown that the normal realizations are a set of optimal realizations that maximize this proposed stability robustness measure. The stability performance of the generalized direct-form II transposed (DFIIt) structure is analysed using this measure. The optimal generalized DFIIt structure problem is defined as to identify those DFIIt structures that maximize the proposed stability measure, which is solved using exhaustive searching method. Numerical examples are given to show the design procedure.

1. INTRODUCTION

Finite Word Length (FWL) effects have been considered as one of the most important issues in actual implementation of digital filters. In fact, the actual performance of a given filter may be greatly degraded due to the FWL errors. See, e.g., [1]. It is well known that any linear system can be represented with different structures such as state-space realizations. These structures are theoretically equivalent since they yield the same system. The important point is that they have different numerical properties. The optimal structure problem is to identify those realizations that minimize the degradation of the filter performance due to the FWL effects.

A stable filter may become unstable when implemented with FWL. The stability issue in the FWL environment is usually investigated in terms of pole sensitivity measures. See, e.g., [3]. Recently, pole sensitivity based stability measures were derived for closed-loop systems [4]-[6]. In [5], a pole modulus sensitivity based stability measure was proposed and it is shown that this measure is better than those derived using the pole sensitivities.

The main objective in this paper is two-fold. The first one is to derive a pole modulus sensitivity based stability measure for digital filters and to study the corresponding optimal realization problem. The optimal structures are usually fully parametrized, which increases the implementation complexity. From a practical point of view, it is desired to implement the digital filter with such a realization that not only has a large stability robustness measure but also possesses as many trivial parameters.[1] Recently, the direct-form II transposed (DFIIt) structures have been studied by several authors [7]-[10]. A generalized DFIIt structure in delta-operator was studied in [10], where it is shown that such a structure, though simple, has very nice numerical properties. Motivated by [10], a more general DFIIt structure was derived based on the so-called polynomial operator approach in [11]. The second objective in this paper is to analyse the stability behavior of this generalized DFIIt structure with the proposed stability measure and then find the optimal DFIIt structures in terms of maximizing this measure.

2. CLASSICAL POLE SENSITIVITY BASED STABILITY MEASURE

Let $H(z)$ be the transfer function of a discrete-time linear time-invariant filter of order N, and let (A, B, C, d) be a state space realization of this filter, that is

$$\begin{aligned} x(t+1) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + du(t), \end{aligned} \tag{1}$$

with $H(z) = d + C(zI - A)^{-1}B$. It is well known that the realizations (A, B, C, d) satisfying the above equation are not unique, they form a realization set, denoted by S_H. In fact, $(T^{-1}AT,\ T^{-1}B,\ CT,\ d)$ is also a realization of $H(z)$ for any real nonsingular T.

[1] Here, trivial parameters mean those that are $0, \pm 1$, which can be implemented exactly and produce no rounding errors.

When (1) is implemented with finite precision, the well designed stable filter may become unstable since some of the eigenvalues of the FWL implemented A-matrix may go outside the unit circle. Let $A_{fwl} = A+\Delta A$ be the FWL implemented A-matrix with $\Delta A = \{\Delta a_{lm}\}$ the perturbation matrix. Denoting

$$\mu(\Delta A) \stackrel{\triangle}{=} \max_{l,m} |\Delta a_{lm}|, \tag{2}$$

one has the following classical stability robustness problem:

$$\mu_0(A) \stackrel{\triangle}{=} inf\{\mu(\Delta A) : A + \Delta A \ is \ unstable\}. \tag{3}$$

It is very hard to compute $\mu_0(A)$. This is still an open problem.

Denote $\lambda(M)$ the eigenvalue set of a matrix M, then the poles $\{\lambda_k\}$ of $H(z)$ are the eigenvalues of A, $\{\lambda_k\} = \lambda(A)$. The deviation of each pole is proportional to the pole sensitivity. In [4], under the assumption that A matrix is fully parametrized the following stability related measure was adopted for closed loop control systems

$$\mu_1(A) \stackrel{\triangle}{=} \min_k \frac{1-|\lambda_k|}{N\sqrt{\Psi_k}}, \tag{4}$$

which can be considered as a lower bound of $\mu_0(A)$, where

$$\Psi_k \stackrel{\triangle}{=} ||\frac{\partial \lambda_k}{\partial A}||_F^2 \tag{5}$$

with $||.||_F$ denoting the *Frobenius* norm:

$$||M||_F^2 = \sum_{i,j=1}^{m,n} |M(i,j)|^2 = tr(MM^{\mathcal{H}}) = tr(M^{\mathcal{H}}M),$$

where $tr(.)$ denotes the trace operation for any matrix $M \in \mathcal{C}^{m\times n}$.

Unlike μ_0, μ_1 can be computed easily with the pole sensitivity evaluated using the following well known result (see, e.g., [1], [2]):

Theorem 1 : Let $\{\lambda_k\} = \lambda(A)$ and x_k be a right eigenvector corresponding to λ_k. Denote $X = (x_1 \ \cdots \ x_k \ \cdots \ x_N)$. Assume that A has a full set of linearly independent eigenvectors and denote $Y \stackrel{\triangle}{=} X^{-\mathcal{H}} = (y_1 \ \cdots \ y_k \ \cdots \ y_N)$ with $\mathcal{H}$ denoting the transpose-conjugate operator, then

$$(\frac{\partial \lambda_k}{\partial A})^{\mathcal{T}} = x_k y_k^{\mathcal{H}}, \tag{6}$$

where $\mathcal{T}$ denotes the transpose operator.

Let (A,B,C) be obtained from the initial realization $(A_0,B_0,C_0) \in S_H$ with similarity transformation matrix T and x_k^0 be a right eigenvector and y_k^0, the reciprocal left eigenvector, of A_0, corresponding to λ_k, then

$$x_k = T^{-1}x_k^0, \quad y_k = T^{\mathcal{T}}y_k^0 \tag{7}$$

are a right eigenvector and the reciprocal left eigenvector of A, respectively. Therefore, different realizations, though having the same poles, have different pole sensitivity measures and hence different stability measures μ_1. From a stability point of view, it is desired to implement the filter using a realization that has the maximal μ_1. It was shown (see, e.g., [1], [2]) that Ψ_k is bounded from below by one and that this bound is achieved for all $k = 1,2,...,N$ if and only if A is normal (i.e., $AA^{\mathcal{T}} = A^{\mathcal{T}}A$), which is equivalently to

$$y_k = ||x_k||_F^{-2}x_k. \tag{8}$$

That means that all normal realizations achieve the maximal μ_1. In [3], an analytical expression was given for the similarity transformations T that transform A_0 to normal $A = T^{-1}A_0T$.

3. A NEW STABILITY RELATED MEASURE AND MINIMIZATION

With a first order approximation, the deviation of the modulus of λ_k can be evaluated with

$$\Delta|\lambda_k| \approx \sum_{l,m} \frac{\partial|\lambda_k|}{\partial a_{lm}}\Delta a_{lm},$$

which leads to

$$|\Delta|\lambda_k|| \le N\sqrt{\sum_{l,m}|\frac{\partial|\lambda_k|}{\partial a_{lm}}|^2}\mu(\Delta A).$$

Clearly, the kth pole is inside the unit circle if

$$N\sqrt{\sum_{l,m}|\frac{\partial|\lambda_k|}{\partial a_{lm}}|^2}\mu(\Delta A) \le 1-|\lambda_k|.$$

Therefore, we have the following lower bound of μ_0

$$\mu_2(A) \stackrel{\triangle}{=} \min_k \frac{1-|\lambda_k|}{N\sqrt{\Phi_k}}, \tag{9}$$

where

$$\Phi_k \stackrel{\triangle}{=} ||\frac{\partial|\lambda_k|}{\partial A}||_F^2 = tr[(\frac{\partial|\lambda_k|}{\partial A})^{\mathcal{T}}\frac{\partial|\lambda_k|}{\partial A}] \tag{10}$$

and it can be shown that

$$\frac{\partial|\lambda_k|}{\partial A} = \frac{1}{|\lambda_k|} Re[\lambda_k^* \frac{\partial \lambda_k}{\partial A}]. \quad (11)$$

Remark: It is easy to see that for a given realization $\Phi_k \leq \Psi_k$ and hence $\mu_2(A) \geq \mu_1(A)$, which means $\mu_2(A)$ is less conservative than $\mu_1(A)$.

Let $\lambda_k = \lambda_k^r + j\lambda_k^i$, $x_k = x_k^r + jx_k^i$ and $y_k = y_k^r + jy_k^i$. With some manipulations, one can show

$$\begin{aligned}(\frac{\partial|\lambda_k|}{\partial A})^{\mathcal{T}} &= |\lambda_k|^{-1}\{\lambda_k^r[x_k^r y_k^{r\mathcal{T}} + x_k^i y_k^{i\mathcal{T}}] \\ &\quad +\lambda_k^i[x_k^i y_k^{r\mathcal{T}} - x_k^r y_k^{i\mathcal{T}}]\} \triangleq Q_k. \quad (12)\end{aligned}$$

It follows from (7) that $Q_k = T^{-1}Q_k^0 T$, where Q_k^0 is given by (12) but corresponding to x_k^0 and y_k^0.

It turns out that

$$\Phi_k = tr(P^{-1}Q_k^0 P Q_k^{0\mathcal{T}}),\ P = TT^{\mathcal{T}}. \quad (13)$$

Like μ_1, μ_2 is a function of T or equivalently of P. Therefore, we have the following optimal stability realization problem:

$$\max_{P>0} \mu_2. \quad (14)$$

First of all, let us consider the following optimal pole sensitivity minimization problem.

$$\min_{P=TT^{\mathcal{T}}>0} \Phi_k. \quad (15)$$

The necessary condition for solutions to (15) is $\frac{\partial \Phi_k}{\partial P} = 0$. Noting $\frac{\partial M^{-1}}{\partial w} = -M^{-1}\frac{\partial M}{\partial w}M^{-1}$, where the nonsingular matrix M is a function of variable w, it can be shown with some manipulations that

$$\frac{\partial \Phi_k}{\partial P} = Q_k^{0\mathcal{T}} P^{-1} Q_k^0 - P^{-1} Q_k^0 P Q_k^{0\mathcal{T}} P^{-1}. \quad (16)$$

Based on the above, we can present one of our main results to be given in Theorem 2. To prove this theorem, we need the following technical lemma:

Lemma 1 : Let X be a right eigenvector matrix of $A \in \mathcal{R}^{N\times N}$, and $Y = X^{-\mathcal{H}}$, the reciprocal left eigenvector matrix. If λ_k is complex, then the corresponding $x_k = x_k^r + jx_k^i$ and $y_k = y_k^r + jy_k^i$ satisfy

$$\begin{aligned}x_k^{r\mathcal{T}} y_k^r = x_k^{i\mathcal{T}} y_k^i &= \frac{1}{2} \\ x_k^{r\mathcal{T}} y_k^i = x_k^{i\mathcal{T}} y_k^r &= 0. \quad (17)\end{aligned}$$

Proof: From $X^{\mathcal{H}}Y = I$, one has $x_k^{\mathcal{H}} y_k = 1$, which leads to

$$\begin{aligned}x_k^{r\mathcal{T}} y_k^r + x_k^{i\mathcal{T}} y_k^i &= 1 \\ x_k^{r\mathcal{T}} y_k^i - x_k^{i\mathcal{T}} y_k^r &= 0 \quad (18)\end{aligned}$$

Since A is real and λ_k is complex, there exists $k_0 \neq k$ such that $x_{k_0} = x_k^* = x_k^r - jx_k^i$. It then follows from $x_{k0}^{\mathcal{H}} y_k = 0$ that

$$\begin{aligned}x_k^{r\mathcal{T}} y_k^r - x_k^{i\mathcal{T}} y_k^i &= 0 \\ x_k^{r\mathcal{T}} y_k^i + x_k^{i\mathcal{T}} y_k^r &= 0 \quad (19)\end{aligned}$$

The lemma follows from (18) and (19). ■

Now, we can present our main result.

Theorem 2 : The normal realizations are solutions to (15), for which

$$\Phi_k = \begin{cases} 1 & if\ \lambda_k\ is\ real \\ \frac{1}{2} & if\ \lambda_k\ is\ complex \end{cases} \quad (20)$$

Proof: First of all, let us take a normal realization as the initial realization. According to (8), we have $y_k^0 = ||x_k^0||_F^{-2} x_k^0$. In order to avoid complicated notations, in the proof we drop the superscript "0", which indicates the initial realization.

If λ_k is real, there exists x_k which is also real. Then with (11) we have $Q_k = \frac{\lambda_k}{|\lambda_k|}||x_k||_F^{-2} x_k x_k^{\mathcal{T}}$. Clearly, $Q_k Q_k^{\mathcal{T}} = Q_k^{\mathcal{T}} Q_k$, which means $\frac{\partial \Phi_k}{\partial P}|_{P=I} = 0$, that is, the normal realizations are solutions to (15). And $\Phi_k = tr(Q_k Q_k^{\mathcal{T}}) = 1$.

If λ_k is complex, it follows from (12), (17) and $y_k = ||x_k||_F^{-2} x_k$ that

$$\begin{aligned}Q_k &= |\lambda_k|^{-1}||x_k||_F^{-2}\{\lambda_k^r[x_k^r x_k^{r\mathcal{T}} + x_k^i x_k^{i\mathcal{T}}] \\ &\quad +\lambda_k^i[x_k^i x_k^{r\mathcal{T}} - x_k^r x_k^{i\mathcal{T}}]\}\end{aligned}$$

and

$$x_k^{r\mathcal{T}} x_k^r = x^{i\mathcal{T}} x_k^i = \frac{1}{2}||x_k||_F^2,\ x_k^{r\mathcal{T}} x_k^i = 0,$$

from which one obtains with direct computation

$$Q_k Q_k^{\mathcal{T}} = Q_k^{\mathcal{T}} Q_k = \frac{1}{2}||x_k||_F^{-2}[x_k^r x_k^{r\mathcal{T}} + x_k^i x_k^{i\mathcal{T}}]$$

and $\Phi_k = \frac{1}{2}$. The former implies $\frac{\partial \Phi_k}{\partial P}|_{P=I} = 0$, that is, the normal realizations are solutions to (15). ■

Therefore, normal realizations are solutions to (14). It should be pointed out that the optimal realizations are usually fully parametrized. In the next section, we will analyse the stability behavior of the generalized direct-form II transposed structure that was proposed in [7]-[11].

4. STABILITY ANALYSIS OF THE GENERALIZED DFIIT STRUCTURE

The generalized DFIIt structure with polynomial operators was studied in [11], where the kth shift operator z^{-1} in the traditional DFIIt structure was replaced by $\frac{\Delta_k}{z-\gamma_k}$ with $\{\Delta_k\}$ to achieve the l_2-scaling to avoid overflow oscillation and γ_k taking values from $\{-1, 0, 1\}$. This generalized DFIIt is the δDFIIt structure in [10] when $\gamma_k = 1,\ \forall k$.

It can be shown that the equivalent state-space realization of this structure is sparse and the elements of the A-matrix, denoted as A_ρ, are all zeros except $A_\rho(1,1) = \gamma_1 - \Delta_1\alpha_1$ and $A_\rho(m,1) = -\Delta_1\alpha_m,\ A_\rho(m, m+1) = \Delta_{m+1},\ A_\rho(m,m) = \gamma_m$ for $m = 2, \cdots, N$.

Noting that $\{\gamma_k\}$ can be implemented exactly, the poles are affected by FWL errors of the parameters $\{\alpha_k, \Delta_k\}$. So, the corresponding stability measure is given by

$$\mu_2 = \min_k \frac{1 - |\lambda_k|}{\sqrt{2N\sum_{m=1}^{N}[|\frac{\partial|\lambda_k|}{\partial\alpha_m}|^2 + |\frac{\partial|\lambda_k|}{\partial\Delta_m}|^2]}}. \qquad (21)$$

To evaluate the above μ_2, one has to compute $\frac{\partial|\lambda_k|}{\partial\alpha_m}$ and $\frac{\partial|\lambda_k|}{\partial\Delta_m}$. Denote e_k as the kth elementary (column) vector whose elements are all zero except the kth one which is 1. It is easy to see that

$$\begin{aligned}
\frac{\partial\lambda_k}{\partial\alpha_m} &= -e_m^{\mathcal{T}}\frac{\partial\lambda_k}{\partial A_\rho}e_1\Delta_1,\ m = 1, \cdots, N \\
\frac{\partial\lambda_k}{\partial\Delta_m} &= e_{m-1}^{\mathcal{T}}\frac{\partial\lambda_k}{\partial A_\rho}e_m, m = 2, ..., N \\
\frac{\partial\lambda_k}{\partial\Delta_1} &= -(\alpha_1\ \cdots\ \alpha_N)\frac{\partial\lambda_k}{\partial A_\rho}e_1. \qquad (22)
\end{aligned}$$

For a given digital filter and a fixed set $\{\gamma_k\}$ with $\gamma_k \in \{-1, 0, 1\}$, one can compute $\{\alpha_k\}$ and hence the corresponding l_2-scaling factors $\{\Delta_k\}$ with the procedure in [11]. With $\frac{\partial\lambda_k}{\partial A_\rho}$ evaluated with (6), one can compute μ_2 using (21). The interesting problem is to maximize μ_2 with respect to $\{\gamma_k\}$:

$$\max_{\{\gamma_k\}} \mu_2. \qquad (23)$$

This problem can be solved using exhaustive searching method.

5. NUMERICAL EXAMPLES AND SIMULATIONS

Several numerical examples and simulations are performed to illustrate the design procedure. Due to the limited space, they will be presented on the conference.

6. REFERENCES

[1] Michel Gevers and Gang Li, Parametrizations in Control, Estimation and Filtering Problems: Accuracy Aspects, Springer Verlag London, Communication and Control Engineering Series, 1993.

[2] R. E. Skelton and D. A. Wagie, 'Minimal root sensitivity in linear systems,' J. Guidance Contr., Vol. 7, pp. 570-574, September-October 1984.

[3] G. Li, "On Pole and Zero Sensitivity of Linear Systems," IEEE Trans. on Circuits and Systems I, vol. 44, no. 7, pp. 583-590, July, 1997.

[4] G. Li, "On The Structure of Digital Controllers with Finite Word Length Consideration," IEEE Trans. on Automatic Control, vol. 43, no. 5, pp. 689-693, May, 1998.

[5] Jun Wu, Sheng Chen, G. Li, R.H. Istepanian, and Jian Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations," IEEE Trans. on Automatic Control, vol. 46, no. 7, pp. 1162-1166, Jul., 2001.

[6] S. Chen, J. Wu, R.H. Istepanian and J. Chu, "Optimizing stability bounds of finite-precision PID controller structures," IEEE Trans. on Automatic Control, vol. 44, no. 11, pp. 2149-2153, Nov. 1999.

[7] N. Wong and T.S. Ng, "Roundoff noise minimization in a modified direct form delta operator IIR structure," IEEE Trans. on Circuits and Systems - II, Vol. 47, pp. 1533-1536, Dec. 2000.

[8] J. Kauraniemi, T.I. Laakso, I. Harttimo, and S.J. Ovaska, "Roundoff noise minimization in a direct form delta operator structure," in Proc. IEEE Int. Acoust., Speech, and Signal Processing, Atlanta, GA, May, 1996.

[9] J. Kauraniemi, T.I. Laakso, I. Harttimo, and S.J. Ovaska , "Delta operator realizations of direct-form IIR filters," IEEE Trans. Circuits and Systems-II, Vol. 45, pp. 41-45, Jan. 1998.

[10] N. Wong and T.S. Ng, "A generalized direct-form delta operator-based IIR filter with minimum noise gain and sensitivity," IEEE Trans. on Circuits and Systems - II, Vol. 48, pp. 425-431, April 2001.

[11] G. Li, "A polynomial operator based DFIIt structure for IIR filters," submitted to IEEE Trans. on Circuits and Systems - II, 2002.