New zero-input overflow stability proofs based on Lyapunov theory

D2.7

New zero-input overflow stability proofs based on Lyapunov theory

Michiel J. Werter and John H.F. Ritzerfeld

Department of Electrical Engineering
Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

In this paper we demonstrate some new proofs of suppressing zero-input overflow oscillations in recursive digital filters. These proofs are based on the second method of Lyapunov
For second-order digital filters with complex conjugated poles the state describes a trajectory in the phase plane, spiralling towards the origin, as long as no overflow correction is applied. Following this state signal an energy function can be defined, which is a natural candidate for a Lyapunov function.
For the second-order direct form digital filter with a saturation characteristic this energy function is a Lyapunov function.
However, this function is not the only possible Lyapunov function of this filter. All energy functions with an energy matrix that is diagonally dominant, guarantee zero-input stability, if a saturation characteristic is used for overflow correction.
In this paper we determine the condition a general second-order digital filters has to fulfil so that there exists at least one energy function with a matrix, which is diagonally dominant.

1. Introduction

In this paper we demonstrate some new proofs of suppressing zero-input overflow oscillations in recursive digital filters, which have also been published in [1]. These proofs are based on the *second method of Lyapunov*, which starts with a properly chosen *energy function* $E(n)$, preferably in quadratic form:

$$E(n) = \underline{x}^T(n)\cdot P\cdot \underline{x}(n). \tag{1.1}$$

Without loss of generality, we choose P symmetrical:

$$P^T = P, \tag{1.2}$$

The Lyapunov theory demands that

$$E = \underline{x}^T\cdot P\cdot \underline{x} > 0 \quad \text{for all } \underline{x} \neq \underline{0}, \tag{1.3}$$

so that $E(n) = 0$ implies $\underline{x}(n) = \underline{0}$. From (1.3) it follows that P has to be positive definite.
Further, the system dynamics must be such that if no overflow correction is applied, according to

$$\underline{x}'(n) = A\cdot \underline{x}(n-1), \tag{1.4}$$

the energy strictly decreases with increasing time:

$$E'(n) < E(n-1) \quad \text{for all } n, \tag{1.5}$$

where $E'(n)$ is the energy pertaining to the state $\underline{x}'(n)$. This inequality is satisfied if

$$P - A^T\cdot P\cdot A \quad \text{is positive definite.} \tag{1.6}$$

Every matrix P satisfying condition (1.2), (1.3) and (1.6) defines an energy function $E(n)$, which is a candidate for a Lyapunov function. If moreover for one such a matrix a subsequent overflow correction

$$\underline{x}(n) = F\{\underline{x}'(n)\} \tag{1.7}$$

lowers the energy for all possible states,

$$E(n) \leq E'(n) \quad \text{for all } n \tag{1.8}$$

the function $E(n)$ is called a *Lyapunov function* of the nonlinear system under consideration.

The existence of a Lyapunov function in a digital filter guarantees freedom from zero-input overflow oscillations. The idealized linear filter is assumed to be stable, so the state in this filter asymptotically approaches the zero-state, for which also the energy is zero. In the actual digital filter overflow correction lowers the energy $E(n)$, implying that also here the energy asymptotically reaches zero, implying zero-state, which completes the prove of zero-input stability.

Care must be taken if "<" is replaced by "≤" in (1.5) so that the energy can remain constant. Such a situation occurs for a marginal choice of the matrix P, for which the energy function is called a *semi-Lyapunov function*. If, moreover, the equality sign in (1.8) applies it can occur that the energy remains constant, associated with the risk of zero-input oscillations.

2. Overflow stability in second-order filters

In this section we investigate the overflow stability of *second-order digital filters*. For sake of conciseness, we restrict the following discussion to sections with complex poles. Compared with real poles, they generally favour all forms of parasitic oscillations (particularly for high Q-values) and thus deserve special consideration.

The 2 x 1 state vector $\underline{x}(n) = (x_1, x_2)^T$ in an autonomous second-order system satisfies the fundamental difference equation

$$\underline{x}(n) = F\{A\cdot \underline{x}(n-1)\}, \tag{2.1}$$

$$\text{where } A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{2.2}$$

In this paper it is understood that

$$[F\{A\cdot \underline{x}\}]_i = F\{[A\cdot \underline{x}]_i\}, \tag{2.3}$$

i.e. the individual components of $A\cdot \underline{x}$ undergo the same memoryless and local overflow correction.

CH2673-2/89/0000-0876 $1.00

The question to be analyzed is: under which circumstances (choice of A, F and $\underline{x}(0)$) does (or does not) (2.1) admit periodic solutions?

Due to the overflow bound, which is henceforth normalized to unity, the state variables conform to

$$|x_i(n)| \leq 1, \tag{2.4}$$

resulting in a state vector confined to the interior of the unit square (cf. Fig.2.1). Without overflow (as long as (2.4) holds) the solution of (2.1) is found as

$$\underline{x}(n) = \mathrm{Re}\{X(\underline{\xi}_r + j\cdot\underline{\xi}_i)\cdot e^{(\Gamma + j\theta)n + j\varphi}\}, \tag{2.5}$$

where $q_1, q_2 = e^{\Gamma \pm j\theta}$ denotes the complex eigenvalues of A and $\underline{\xi}_r \pm j\cdot\underline{\xi}_i$ denotes the pertinent eigenvectors.

It is tacitly assumed that $\Gamma < 0$, expressing linear stability. Further, the constants of integration (X, φ) are determined by the initial state $\underline{x}(0)$.

If, for the time being, time n is viewed as a continuous variable, $\underline{x}(n)$ describes a trajectory in the phase plane. For the case $\Gamma = 0$ this would be an ellipse with main axes in the direction of $\underline{\xi}_r$ and $\underline{\xi}_i$.
For $\Gamma < 0$ (corresponding to poles inside the unit circle), we obtain a non-closed, ellipse-like curve spiralling towards the origin, cf. Fig. 2.1.

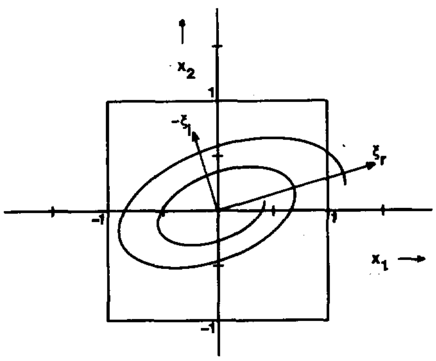

Fig. 2.1. Trajectory of the state vector $\underline{x}(n)$ in the state plane.

Of course, these results only apply to the digital filter as long as overflow does not occur. If this is not the case, the linearly determined $\underline{x}(n)$ leaves the unit square, and overflow correction has to be applied. This correction introduces one of two basic state modifications: (a) $\underline{x}(n)$ is moved towards the origin, (b) $\underline{x}(n)$ is moved away from the origin.
Case (a) is wanted because it supports the natural linear motion; no oscillation occurs if all overflows are corrected this way.
Case (b) is dangerous, because it compensates or even overcompensates the linear behaviour and, hence can (but need not) lead to oscillations. Of course, these statements ask for an unambiguous definition of "distance from the origin". Instead of the widely used Euclidean norm our definition is guided by the linear state motion, according to (2.5). Following

$$\underline{x}(n) = X(n)\cdot[\underline{\xi}_r\cdot\cos(\varphi(n)) - \underline{\xi}_i\cdot\sin(\varphi(n))] \tag{2.6}$$

two variables $X(n), \varphi(n)$ can be associated with each state $\underline{x}(n)$. Particularly, the variable $X(n)$ is determined from $\underline{x}(n)$ as

$$X^2(n) = \left[\frac{\underline{\xi}_r^T\cdot\underline{x}}{\underline{\xi}_r^T\cdot\underline{\xi}_r}\right]^2 + \left[\frac{\underline{\xi}_i^T\cdot\underline{x}}{\underline{\xi}_i^T\cdot\underline{\xi}_i}\right]^2. \tag{2.7}$$

Comparing (2.6) with (2.5) one recognizes

$$X(n) = X\cdot e^{\Gamma n}, \tag{2.8}$$

i.e. a monotonically decreasing function. Therefore, the function

$$E(n) = X^2(n) \tag{2.9}$$

is a natural candidate for a Lyapunov function. We choose $X(n)$ as the "distance from the origin".

Overflow correction is now visualized in Fig. 2.2. An uncorrected state point B is mapped into B', B'', or B''' after applying saturation, zeroing and two's complement, respectively. For this example all types lead to an increase of $E(n)$ and, hence, to a movement away from the origin. On the other hand, for point C this is only true for zeroing and two's complement overflow correction.

For some ellipse geometries it is possible to use appropriate overflow characteristics such that the state always moves towards the origin and oscillations are suppressed.
Obviously this is not the case for the arbitrarily oriented ellipse of Fig. 2.2. However, it is easily *recognized that for an ellipse whose axes coincide* with the x_1-x_2-axes, each of the three overflow corrections satisfies the stability condition, while for an ellipse with a 45° inclination stabilization can be obtained at least with a saturation characteristic.

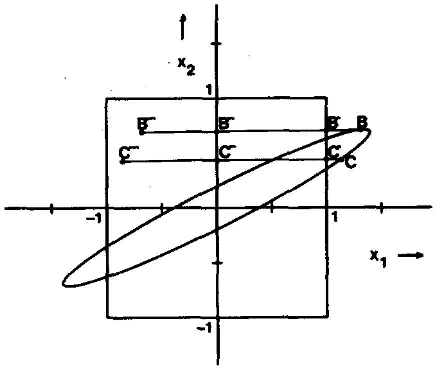

Fig. 2.2. Ellipse X = constant in the state plane.

3. Overflow stability in direct form filters

In this section we investigate the zero-input stability of the second-order *direct form digital filter* with overflow correction. This filter is described by the system matrix

$$A = \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix}. \tag{3.1}$$

It turns out to be free from zero-input overflow oscillations, if saturation is used. This is true for all pairs of filter coefficients a and b in the "stability triangle", described by

$$\begin{aligned} 1 - |a| - b &> 0 \\ 1 + b &> 0. \end{aligned} \quad (3.2)$$

The analytic proof of this statement has been reported by Ebert e.a. [2]. In this section we present two novel proofs using Lyapunov theory, one proof only for complex conjugated poles, the other for all pairs a and b in the stability triangle.

For complex conjugated pole pairs $q_1, q_2 = e^{r \pm j\theta}$ we define an energy function $E(n)$, according to (2.10):

$$E(n) = x_1^2(n) - a \cdot x_1(n) \cdot x_2(n) - b \cdot x_2^2(n), \quad (3.3)$$

where
$$\begin{aligned} a &= 2e^{r} \cdot \cos(\theta) \\ b &= -e^{2r}. \end{aligned} \quad (3.4)$$

This energy function is characterized by the matrix

$$P = \begin{bmatrix} 1 & -a/2 \\ -a/2 & -b \end{bmatrix}, \quad (3.5)$$

which is symmetrical. It is positive definite, due to

$$\mathrm{Det}[P] = -b - a^2/4 = e^{2r} \cdot \sin^2(\theta) > 0 \quad (3.6)$$

and
$$\mathrm{Tr}[P] = 1 - b = 1 + e^{2r} > 0. \quad (3.7)$$

The system dynamics is such that if no overflow correction is applied, the energy decreases:

$$E'(n) = e^{2r} \cdot E(n-1) < E(n-1). \quad (3.8)$$

In the nonlinear system, saturation causes an additional energy reduction. If before correction we have $x_1'(n) > 1$, then after saturation $x_1(n) = 1$. The component $x_2'(n)$ can never overflow since $x_2'(n) = x_1(n-1)$.

So
$$E(n) = 1 - a \cdot x_2'(n) - b \cdot x_2'^2(n) \quad (3.9)$$

and

$$E(n) - E'(n) = -[x_1'(n) - 1] \cdot [1 + x_1'(n) - a \cdot x_2'(n)] < 0 \quad \text{for all } n. \quad (3.10)$$

The latter inequality is due to the stability requirement $|a| < 2$. The same conclusion can be drawn for values of $x_1'(n) < -1$. Hence the energy function $E(n)$ is a Lyapunov function, which guarantees zero-input stability in the second-order direct form filter with saturation and complex conjugated poles.

The zero-input stability of this filter can also be proved with another Lyapunov function $E(n)$, which is characterized by the symmetrical matrix

$$P = \begin{bmatrix} 1-b & -a \\ -a & 1-b \end{bmatrix}. \quad (3.11)$$

Matrix P is positive definite for all pairs of coefficients a and b in the stability triangle, due to

$$\mathrm{Det}[P] = (1+a-b) \cdot (1-a-b) > 0 \quad (3.12)$$

and

$$\mathrm{Tr}[P] = 2 \cdot (1-b) > 0. \quad (3.13)$$

Without overflow correction the energy cannot increase, since

$$E'(n) - E(n-1) = -(1+b) \cdot [a \cdot x_1(n-1) + (b-1) \cdot x_2(n-1)]^2 \le 0 \quad \text{for all } n. \quad (3.14)$$

In the nonlinear system saturation causes a reduction of the energy. If $x_1'(n) > 1$ then after saturation $x_1(n) = 1$ and with $x_2(n) = x_2'(n)$, we have

$$E(n) - E'(n) = -[x_1'(n) - 1] \cdot [(1-b) \cdot (1 + x_1'(n)) - 2a \cdot x_2'(n)] < 0 \quad \text{for all } n. \quad (3.15)$$

The last inequality is a consequence of the stability condition $1 - |a| - b > 0$. The same conclusion can be drawn for $x_1'(n) < -1$.
In a strict sense function $E(n)$ is not a Lyapunov function since the energy can remain constant. This situation can only appear if no overflow correction is applied. But then the filter responds linearly and the state will asymptotically reach zero; no zero-input overflow oscillation is possible in the second-order direct form filter with saturation. This result is now proved for all pairs of filter coefficients a and b in the stability triangle.

4. Overflow stability in filters with saturation

It is a common property of normal, wave digital and lattice filters of second and higher order that the "energy matrix" P is *diagonal* and that $E(n)$=constant are ellipses oriented parallel to the coordinate axes. Only this ellipse geometry allows for all overflow characteristics applied to the individual state variables, without risk of zero-input overflow oscillations.

The question arises: Which A matrices admit a diagonal "energy matrix"? This question has been solved in [3] for the second-order filter, where it is shown that this is only possible if the system matrix A fulfils the condition:

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A]. \quad (4.1)$$

If a *saturation characteristic* is used in a *second-order digital filter* the condition (4.1) for zero-input stability (valid for all overflow characteristics) can be relaxed. We shall prove the following theorem (see also [4]):

If
$$E(n) = \underline{x}^T(n) \cdot P \cdot \underline{x}(n), \quad (4.2)$$

satisfies all conditions for an energy function (see Section 1) and if matrix P is diagonally dominant:

$$|p_{12}| \le p_{11} \quad \text{and} \quad |p_{12}| \le p_{22}, \quad (4.3)$$

then the function $E(n)$ is a Lyapunov function of the filter with saturation.

Proof:
The difference in energy between the uncorrected signal $\underline{x}'(n)$ and the actual state $\underline{x}(n)$ is

$$\begin{aligned} E'(n) - E(n) &= \\ &= p_{11} \cdot (x_1'^2 - x_1^2) + p_{22} \cdot (x_2'^2 - x_2^2) + 2p_{12} \cdot (x_1'x_2' - x_1x_2) \\ &\ge p_{11}(x_1'^2 - x_1^2) + p_{22}(x_2'^2 - x_2^2) - 2|p_{12}|(|x_1'x_2'| - |x_1x_2|) \\ &= (p_{11} - |p_{12}|) \cdot (x_1'^2 - x_1^2) + (p_{22} - |p_{12}|) \cdot (x_2'^2 - x_2^2) \\ &\quad + |p_{12}|\{(|x_1'| - |x_2'|)^2 - (|x_1| - |x_2|)^2\}. \end{aligned} \quad (4.4)$$

The inequality is valid due to $\mathrm{sgn}(x_1')=\mathrm{sgn}(x_1)$ and $\mathrm{sgn}(x_2')=\mathrm{sgn}(x_2)$. If only one of the components overflows, f.e. $x_1' > 1$ then $x_1 = 1$, $x_2 = x_2'$ and

$$E'(n)-E(n) \geq (p_{11}-|p_{12}|)\cdot(x_1'^2-1)+|p_{12}|\cdot([x_1'-x_2]^2-[1-x_2]^2) \geq 0. \quad (4.5)$$

If both components overflow then $|x_1| = |x_2| = 1$ and

$$E'(n)-E(n) \geq (p_{11}-|p_{12}|)(x_1'^2-1)+(p_{22}-|p_{12}|)(x_2'^2-1) + |p_{12}|(|x_1'|-|x_2'|)^2 \geq 0. \quad (4.6)$$

So in a system with an energy $E(n)$ satisfying (4.3) saturation causes an additional energy reduction.

Now we solve the question: Which second-order A matrices possess at least one energy function $E(n)$ whose energy matrix P is diagonally dominant?

Therefore the energy in the idealized system must strictly decrease, which will be satisfied if

$$P - A^T\cdot P\cdot A \text{ is positive definite.} \quad (4.7)$$

For a second-order system this condition is equivalent to the inequalities

$$\mathrm{Det}[P - A^T\cdot P\cdot A] > 0 \quad (4.8)$$

$$\text{and } \mathrm{Tr}[P - A^T\cdot P\cdot A] > 0. \quad (4.9)$$

Expression (4.8) is satisfied if there exists some real values p_{11}, p_{12} and p_{22} with

$$\mathrm{a}p_{11}^2+\mathrm{b}p_{11}p_{22}+\mathrm{c}p_{22}^2+\mathrm{d}p_{11}p_{12}+\mathrm{e}p_{12}p_{22}+\mathrm{f}p_{12}^2 < 0 \quad (4.10)$$

where
$$\begin{aligned} \mathrm{a} &= a_{12}^2 \\ \mathrm{b} &= a_{11}^2 + a_{22}^2 - 1 - \mathrm{Det}^2[A] \\ \mathrm{c} &= a_{21}^2 \\ \mathrm{d} &= -2\cdot a_{12}\cdot(a_{11}-a_{22}) \\ \mathrm{e} &= 2\cdot a_{21}\cdot(a_{11}-a_{22}) \\ \mathrm{f} &= (1 + \mathrm{Det}[A])^2 - 4\cdot a_{11}\cdot a_{22}. \end{aligned}$$

Equation (4.10) *can be written in the form*

$$\mathrm{a}(p_{11}-\alpha p_{12})^2+\mathrm{b}(p_{11}-\alpha p_{12})(p_{22}-\beta p_{12})+\mathrm{c}(p_{22}-\beta p_{12})^2<\mathrm{K}p_{12}^2 \quad (4.11)$$

where
$$\alpha = \frac{-2\cdot a_{21}\cdot(a_{11}-a_{22})}{(a_{11}-a_{22})^2-(1-\mathrm{Det}[A])^2}$$

$$\beta = \frac{2\cdot a_{12}\cdot(a_{11}-a_{22})}{(a_{11}-a_{22})^2-(1-\mathrm{Det}[A])^2}$$

$$\mathrm{K} = \frac{(1-\mathrm{Det}[A])^2\cdot\{(1+\mathrm{Det}[A])^2-(a_{11}+a_{22})^2\}}{(a_{11}-a_{22})^2-(1-\mathrm{Det}[A])^2}.$$

The discriminant of this quadratic form is $\mathrm{b}^2-4\mathrm{ac}$ =

$$= [(1+\mathrm{Det}[A])^2-(a_{11}+a_{22})^2]\cdot[(1-\mathrm{Det}[A])^2-(a_{11}-a_{22})^2] \quad (4.12)$$

For $p_{12} = 0$ the energy matrix P is diagonal which has a solution for a system matrix A satisfying condition (4.1). If condition (4.1) is <u>not</u> satisfied, which can only be true for filters with $a_{12}\cdot a_{21} < 0$, we have according to (4.12) a discriminant which is negative, with as result that (4.11) describes the inner part of an elliptical curve in the p_{11}-p_{22}-plane, which is non-empty, since $\mathrm{K} > 0$.

Since $a_{12}\cdot a_{21} < 0$, it is possible to choose parameter p_{12} in such a way that the centre point of the ellipse $(\alpha\cdot p_{12}, \beta\cdot p_{12})$ lies in the first quadrant of the p_{11}-p_{22}-plane.

For $|a_{12}| < |a_{21}|$, there exists some point (p_{11},p_{22}) within the ellipse (4.11) satisfying condition (4.3) if the ellipse curves the line $p_{22} = |p_{12}|$. Such a point of intersection is found if

$$\mathrm{a}p_{11}^2+[\mathrm{b}\cdot\mathrm{sgn}(p_{12})+\mathrm{d}]p_{11}p_{12}+[\mathrm{c}+\mathrm{e}\cdot\mathrm{sgn}(p_{12})+\mathrm{f}]p_{12}^2 < 0 \quad (4.13)$$

has real solutions. So

$$[\mathrm{b}\cdot\mathrm{sgn}(p_{12})+\mathrm{d}]^2 \geq 4\mathrm{a}[\mathrm{c}+\mathrm{e}\cdot\mathrm{sgn}(p_{12})+\mathrm{f}], \quad (4.14)$$

or

$$[(1+\mathrm{Det}[A])^2-(a_{11}+a_{22})^2]\cdot[(1-\mathrm{Det}[A])^2-(a_{11}-a_{22}-2a_{12}\mathrm{sgn}(p_{12}))^2] \geq 0. \quad (4.15)$$

The first factor of (4.15) is positive due to stability conditions. The second factor shows the remaining condition which can be written in the form

$$|a_{11}-a_{22}| < 2\cdot|a_{12}| + 1 - \mathrm{Det}[A]. \quad (4.16)$$

All matrices P that satisfy (4.15) and therefore (4.8) does also satisfy (4.9).
The case $|a_{21}| < |a_{12}|$ is equivalent to the previous case resulting in the condition

$$|a_{11}-a_{22}| < 2\cdot|a_{21}| + 1 - \mathrm{Det}[A]. \quad (4.17)$$

The conditions (4.16) and (4.17) form together the result of this section: All second-order digital filters with a system matrix A satisfying:

$$|a_{11}-a_{22}| \leq 2\cdot\min(|a_{12}|,|a_{21}|)+1-\mathrm{Det}[A], \quad (4.18)$$

possess at least one energy function $E(n)$, with a matrix P satisfying (4.3), and thus are free from zero-input overflow oscillations for saturation.

As it should be, this condition is less restrictive than (4.1). Contrary to the former condition, all stable direct form filter satisfy (4.18).

REFERENCES

[1] M.J. Werter, "*Suppression of parasitic oscillations due to overflow and quantization in recursive digital filters*," Ph.D.-Thesis, Eindhoven University of Technology, The Netherlands, 1989.

[2] P.M. Ebert, J.E. Mazo and M.G. Taylor, "Overflow oscillations in recursive digital filters," *Bell Syst. Tech. J.*, vol. 48, pp. 2999-3020, 1969.

[3] 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.

[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," submitted to *IEEE Trans. Circuits Syst.*