Suppression of parasitic oscillations due to overflow and quantization in recursive digital filters

SUPPRESSION OF PARASITIC OSCILLATIONS DUE TO OVERFLOW AND QUANTIZATION IN RECURSIVE DIGITAL FILTERS

M.J. WERTER

SUPPRESSION OF PARASITIC OSCILLATIONS

DUE TO OVERFLOW AND QUANTIZATION

IN RECURSIVE DIGITAL FILTERS

SUPPRESSION OF PARASITIC OSCILLATIONS

DUE TO OVERFLOW AND QUANTIZATION

IN RECURSIVE DIGITAL FILTERS

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de Rector Magnificus, prof. ir. M. Tels, voor een commissie aangewezen door het College van Dekanen in het openbaar te verdedigen op vrijdag 10 februari 1989 te 16.00 uur

door

MICHAEL JOZEF WERTER
geboren te Eindhoven

druk: wibro dissertatiedrukkerij, helmond

DIT PROEFSCHRIFT IS GOEDGEKEURD
DOOR DE PROMOTOREN

Prof. Dr. -Ing. H.J. Butterweck
en
Prof. Dr. Ir. W.M.G. van Bokhoven

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Werter, Michael Jozef

Suppression of parasitic oscillations due to overflow
and quantization in recursive digital filters /
Michael Jozef Werter. - [S.l. : s.n.], - Fig., tab.
Proefschrift Eindhoven. - Met lit. opg., reg.
ISBN 90-9002583-9
SISO 663.12 UDC 621.372.54.037.37.018.783 NUGI 832
Trefw.: digitale filters.

Contents

Voorwoord

Het tot stand komen van dit proefschrift en het uitvoeren van het hierin beschreven onderzoek is slechts mogelijk geweest door de inbreng van verscheidene mensen, die ik hierbij dan ook van harte wil bedanken voor hun bijdrage aan dit werk.
Met name wil ik mijn promotor prof.dr.-ing. H.J. Butterweck noemen, die mij in de gelegenheid stelde dit onderzoek te verrichten en met wie ik vele uren zeer levendige en boeiende discussies heb gehad.
Verder noem ik mijn tweede promotor, prof.dr.ir. W.M.G. van Bokhoven, en de overige leden van de promotiecommissie die ik wil danken voor de leerzame gesprekken en hun nuttig commentaar op dit proefschrift.
Mijn collega's van de vakgroep Theoretische Elektrotechniek wil ik heel hartelijk bedanken voor de uitstekende sfeer en plezierige samenwerking, die ik de afgelopen jaren ervaren heb.
Tenslotte wil ik mijn ouders bedanken voor hun enthousiaste stimulans, hetgeen voor mij een blijvende steun bij mijn werkzaamheden is geweest.

Summary

In every digital signal processor arithmetical operations are applied to signals, which are represented with a finite wordlength. These operations generally lead to an increase of wordlength. Therefore precaution measures have to be taken for signal wordlength reduction, namely quantizations and overflow corrections. This wordlength reduction causes a filter to operate deviating from the linear behaviour. In a recursive digital filter wordlength reduction results in a number of characteristic instabilities such as limit cycles, subharmonics and overflow oscillations. The aim of this thesis is to analyse these instabilities and to search for measures for suppressing the unwanted phenomena.

The effects of quantization and overflow are treated independently. The overflow stability is analysed in first-order, second-order, direct-form, wave, normal and state-space digital filters. Freedom from overflow oscillations in these filters is proved with the second method of Lyapunov, starting with a properly chosen quadratic energy function, which is positive definite and time-decreasing without wordlength reduction. If, moreover, wordlength reduction lowers the energy for all possible states, this energy function is a Lyapunov function, which guarantees freedom from overflow oscillations.

The quantization stability is analysed for the same type of filters. Freedom from zero-input limit cycles in these filters is proved with the same Lyapunov theory. But forced-response stability cannot be guaranteed. A new type of digital filter has been introduced, which is free from subharmonics for discrete-time periodic input signals with a given period N. Using the theory of cyclotomic polynomials this filter structure is so extended that it is also free from subharmonics for input signals which are periodic with a divisor of N.

Suppression of subharmonics for all possible periodic input signals appears to be impossible. Therefore we have analysed the subharmonics in digital filters in more detail. A computer program has been developed, performing a search for all possible subharmonics in a given filter structure. The region of filter coefficients for which a subharmonic can occur has been derived and subharmonics are described by an effective value model and by a linear time-variant system. Finally they have been analysed by a new decomposition method.

Samenvatting

In iedere digitale signaalprocessor vinden rekenkundige bewerkingen plaats op signalen die met een eindige woordlengte gerepresenteerd zijn. Deze bewerkingen leiden in het algemeen tot een toename van de benodigde woordlengte. In zo'n geval moet de signaalwoordlengte verminderd worden: kwantisatie en overflow correctie. Deze woordlengte-reductie veroorzaakt een afwijking in de filterwerking ten opzichte van het lineaire gedrag. In een recursief digitaal filter resulteert woordlengte-reductie in een aantal karakteristieke instabiliteiten, zoals limit cycles, overflow-oscillaties en subharmonischen. Het doel van dit proefschrift is het analyseren van deze instabiliteiten en het mogelijk onderdrukken ervan. De gevolgen van kwantisatie en overflow worden onafhankelijk van elkaar beschouwd. De overflow-stabiliteit is geanalyseerd in eerste-orde, tweede-orde, directe-vorm, golf, normale en toestands digitale filters. Vrijheid van overflow-oscillaties wordt met behulp van de tweede methode van Lyapunov bewezen. Deze methode gaat uit van een geschikt gekozen kwadratische energiefunctie, die positief definiet is en zonder woordlengte-reductie vermindert in de loop van de tijd. Als, bovendien, overflowcorrectie energieverlagend werkt, dan is stabiliteit gegarandeerd. De kwantisatie-stabiliteit is voor dezelfde typen filters geanalyseerd. Vrijheid van zero-input limit cycles wordt ook met de Lyapunov-theorie bewezen. Maar, forced-response stabiliteit kan niet worden gegarandeerd. Er is een nieuw type filter ontwikkeld dat vrij is van subharmonischen voor tijddiskreet-periodieke ingangssignalen met een gegeven periode N. Onder gebruikmaking van de theorie der cyclotomische polynomen kan deze filter structuur worden uitgebreid, zodat de schakeling ook stabiel wordt voor ingangssignalen die periodiek zijn met een deler van N. Onderdrukking van subharmonischen voor alle mogelijke ingangssignalen schijnt onmogelijk te zijn. Daarom zijn ze nader geanalyseerd. Er is een computerprogramma geschreven, dat zoekt naar alle subharmonischen in een gegeven filterstructuur.Het filterparametergebied waarvoor ze kunnen optreden is afgeleid en ze zijn beschreven met een effectieve waardemodel en met een linear tijd-varierend systeem. Tenslotte zijn ze geanalyseerd met behulp van een nieuwe decompositiemethode.

1. Finite wordlength effects in digital signal processing

1.1. Introduction

This thesis is concerned with the processing of *digital signals*. Such signals have a discrete character in time and in amplitude, i.e. they occur at distinct (mostly equidistant) times and can only assume distinct values. In the commonly used fixed-point representation these values are integer multiples of an elementary quantum so that such signals are throughout referred to as "quantized". In many applications, a digital signal is derived from an analog signal through an AD-conversion, in which the discreteness in time is obtained by the process of sampling, while the discreteness in amplitude occurs through applying an appropriate quantization (resulting e.g. in the signal on a compact disc record). However, there are also natural digital signals, like the balance of a bank-account, in which the successive account statements have no direct relation to time.

Digital signals are often desired to be *processed*, in order to extract or separate or modify characteristic information carried by such signals. In one application the removal of additional noise may be required, in another one the separation of two frequency bands may be desired, while in a third one the distortion set up by a preceding system may be wished to be compensated. Such processing of digital signals can be performed with suitable software in general-purpose computers or with special-purpose hardware.

In idealized form, digital signal processing studies the processing of discrete-time signals, which are data sequences of real or complex numbers. The signal processing in digital filters is intended to be performed in the form of *linear operations*, which for the important class of time-invariant systems are of the convolution type. On the other hand in a practical digital processor the signals are represented with a *finite wordlength* as a consequence of the encoding of the signals in a particular, mostly binary, format and due to the fact that the signal must be stored in the (limited) memory of the processor.

In every digital signal processor arithmetical operations take place, such as the multiplication of a signal with a constant factor and the addition of two or more signals. Irrespective of the encoding of the signals (binary or decimal, fixed-point or floating-point) these arithmetical operations generally lead to an increase in the required wordlength. Therefore precaution measures have often to be taken for *signal wordlength reduction*, namely *quantizations* after multiplications with non-integer factors and *overflow corrections* after additions of signals and multiplications with factors larger than unity. This wordlength reduction causes a filter to operate deviating from the linear behaviour. Fortunately, this deviation can be made arbitrarily small through choosing sufficiently long binary words. Yet there remain typical finite-wordlength effects that cause an actual digital filter to behave as a (weakly) *nonlinear system*.

Contrary to the finite wordlength of the signals to be processed the finite wordlength of the *filter coefficients* in a system with infinite signal wordlength does not affect the linearity of the filter behaviour. This effect only amounts to restrictions on the linear filter characteristics, resulting in discrete grids of pole-zero patterns (see [97]). Once a filter design with some combination of permitted coefficients meets the required specifications (with regard to amplitude and phase characteristics) the actual filter performance differs from that predicted by linear theory only as to the previously mentioned nonlinear finite-wordlength effects. These effects form the subject matter of this thesis.

In a *non-recursive* (*finite impulse response*) *filter* a wordlength reduction is not strictly required. It is in fact only applied for practical reasons in order to avoid too long words. Furthermore the only effect due to the wordlength reduction is the addition of an error signal, which is correlated with the main signal and can result in such effects as quantization noise and crosstalk.

The situation is basically different for *recursive* (*infinite impulse response*) *structures*. In the first place quantization and overflow

correction are absolutely necessary in every feedback loop of the processor to prevent an unlimited increase of the wordlength.
Secondly the errors introduced by the nonlinear operation corresponding to this wordlength reduction are fed back in the filter resulting in a number of characteristic instabilities, such as limit cycles, subharmonics and overflow oscillations.
The instabilities due to quantization (limit cycles, subharmonics) have relatively small amplitudes. Together with quantization noise and crosstalk they are considered as the most serious deviation from linear behaviour under normal operating conditions of a digital filter.
The oscillations associated with overflow correction have large amplitudes; because of their disastrous effects on the filter behaviour they have to be absolutely avoided.

The aim of this thesis is to *analyse* these instabilities and to search for measures for *reducing* or possibly *suppressing* the unwanted phenomena.
The results of this paper mainly concern first- and second-order filter sections, since higher-order digital filters can be constructed as cascades or parallel configurations of these sections. In due course, we summarize more general results for higher-order filter sections.

We conclude this section with a few remarks on the content of this thesis. In Chapter 1 some basic theory of digital signal processing is reviewed. In Section 1.2 we analyse the linear response of the idealized digital filter, in which there is no finite wordlength reduction. Section 1.3 contains some alternative approaches for wordlength reduction, for quantization as well as overflow correction. In Section 1.4 nonlinear wordlength reduction effects are investigated for two different input conditions: zero input (leading to limit cycles) and non-zero periodic input signals (leading to subharmonics and jump phenomena).

In Chapter 2 we demonstrate some methods for suppressing overflow oscillations in elementary digital filters.

In Section 2.1 the common approximation is stated that quantization and overflow are not only conceptually decoupled, but also analytically treatable as independent effects. Overflow stability is analysed in first-order filters (2.2), second-order filters (2.3), direct form digital filters (2.4), wave digital and normal filters (2.5) and state-space digital filters (2.6). Freedom from overflow oscillations is proved with the second method of Lyapunov. In all these digital filters a quadratic energy function has to be found which is positive definite and decreases with increasing time. If furthermore for one of such functions overflow correction lowers the energy for all possible states, freedom from overflow oscillations is guaranteed.

In Chapter 3 the quantization stability is analysed for the same type of filters (Sections 3.1 to 3.4). Absence of zero-input limit cycles is again proved with an adequate Lyapunov function, but forced-input stability cannot be realized in these digital filters. In Section 3.5 we introduce a new type of digital filter which is free from subharmonics for discrete-time periodic input signals with a given period N. Using the theory of cyclotomic polynomials we can extend the new digital filter structure in order to make it also free from subharmonics for periodic input signals with other periods (3.6). In Section 3.7 we mention a general principle to convert a stable autonomous state-space digital filter into a digital filter which is free from subharmonics for a periodic input signal with a given period N.

Suppression of subharmonics in digital filters for all possible input signals appears to be impossible. Therefore in Chapter 4 we present some investigations into the properties of subharmonics in digital filters. In Section 4.1 a computer program is presented, performing a search procedure for all possible subharmonics in a given structure. The region of filter coefficients for which a subharmonic can occur in a digital filter is calculated (4.2) and the results are in agreement with computer simulations. Subharmonics in digital filters can be described by an effective value model (4.3) and by a linear time-variant system (4.4). In Section 4.5 we analyse these periodic oscillations by a new decomposition method. Finally in Chapter 5 some conclusions are formulated.

1.2. Linear analysis of digital filters

Before analysing the nonlinear effects due to the finite wordlength representation of the signals in a digital signal processor the *linear response* of the *idealized filter* will be determined in this section. This linear filter theory is well described in many textbooks on digital signal processing (see [13, 19, 22, 41, 78, 79, 108, 132, 164, 166, 167, 191, 200, 218, 273, 274, 286, 290, 297, 305]).

In linear form, a single-input-single-output time-invariant recursive digital filter with a total of J time-delay elements can be described by the *state equations*

$$\underline{x}(n+1) = A\cdot\underline{x}(n) + \underline{b}\cdot u(n)$$

$$y(n) = \underline{c}^{T}\cdot\underline{x}(n) + d\cdot u(n), \qquad (1.2.1)$$

where $u(n)$ denotes the input signal ($u(n) \in R;\ n \in Z$)
$\underline{x}(n)$ denotes the state signal ($\underline{x}(n) \in R^{J};\ n \in Z$)
$y(n)$ denotes the output signal ($y(n) \in R;\ n \in Z$)
A denotes the system matrix ($A \in R^{J} \times R^{J}$)
$\underline{b}$ denotes the input scaling vector ($\underline{b} \in R^{J}$)
$\underline{c}$ denotes the output scaling vector ($\underline{c} \in R^{J}$)
d denotes the direct input-output scalar ($d \in R$).

Strictly speaking, $u(n)$ is a single number from the sequence for a given index n, while $\{u(n)\}$ denotes the entire finite, or countably infinite sequence. However, we will follow the general practice of using $u(n)$ to represent the entire sequence as well as a number from the sequence, depending on whether n is assumed to be a running or fixed variable. The same holds for the state signal $\underline{x}(n)$ and the output signal $y(n)$ [164].

A linear time-invariant digital filter can also be described by a *convolution equation*

$$y(n) = \sum_{m=-\infty}^{\infty} h(m)\cdot u(n-m), \qquad (1.2.2)$$

where $h(m)$ is the *impulse response* of the filter, which is the response on an input $u(n) = \delta(n)$,
where $\delta(n)$ is the unit-sample or impulse sequence:

$$\begin{aligned} \delta(n) &= 1 && \text{for } n = 0 \\ &= 0 && \text{for } n \neq 0. \end{aligned} \qquad (1.2.3)$$

Substitution of (1.2.3) in the state equations (1.2.1) shows that for a system with $\underline{x}(n) = \underline{0}$ for $n \leq 0$:

$$\begin{aligned} h(n) &= 0 && \text{for } n < 0 \\ &= d && \text{for } n = 0 \\ &= \underline{c}^T\cdot A^{n-1}\cdot \underline{b} && \text{for } n > 0. \end{aligned} \qquad (1.2.4)$$

In this thesis we only consider *causal systems* for which

$$h(n) = 0 \qquad \text{for } n < 0. \qquad (1.2.5)$$

Then the convolution equation (1.2.2) can be reduced to

$$y(n) = \sum_{m=0}^{\infty} h(m)\cdot u(n-m). \qquad (1.2.6)$$

The z-transform $U(z)$ of a sequence $u(n)$, $(U(z) \in C;\ z \in C)$, is defined according to

$$U(z) = \sum_{n=-\infty}^{\infty} u(n)\cdot z^{-n}, \qquad (1.2.7)$$

with the inverse z-transform given by the complex contour integral

$$u(n) = \frac{1}{2\pi j} \oint_C U(z) \cdot z^{n-1} \, dz, \tag{1.2.8}$$

where C is a counter clockwise closed contour in the region of convergence of $U(z)$ around the origin of the complex plane.

The z-transform of the impulse response $h(n)$ is the *transfer function* $H(z)$, which is given by

$$H(z) = \frac{Y(z)}{U(z)} = \sum_{n=0}^{\infty} h(n) \cdot z^{-n}$$

$$= \underline{c}^T \cdot (z \cdot I - A)^{-1} \cdot \underline{b} + d, \tag{1.2.9}$$

with I the J x J unit matrix.

For a controllable and observable filter, the *eigenvalues* q_μ of the system matrix A are the poles of the transfer function $H(z)$:

$$\mathrm{Det}[A - q_\mu \cdot I] = 0. \tag{1.2.10}$$

The filter is *stable* if and only if all poles q_μ are within the unit circle:

$$|q_\mu| < 1 \quad \text{for all poles } q_\mu, \tag{1.2.11}$$

corresponding with a bounded-input-bounded-output system, where

$$\sum_{n=0}^{\infty} |h(n)| < \infty. \tag{1.2.12}$$

The general second-order digital filter has a system matrix A of the form

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \tag{1.2.13}$$

with a determinant and a trace defined by

$$\begin{aligned} \mathrm{Det}[A] &= a_{11} \cdot a_{22} - a_{12} \cdot a_{21} \\ \mathrm{Tr}[A] &= a_{11} + a_{22}. \end{aligned} \tag{1.2.14}$$

For the second-order filter the stability condition (1.2.11) is equivalent to

$$|\mathrm{Det}[A]| < 1$$

and

$$|\mathrm{Tr}[A]| < 1 + \mathrm{Det}[A]. \tag{1.2.15}$$

As an example, we consider the *second-order digital filter in "direct form"*, shown in Fig. 1.2.1.

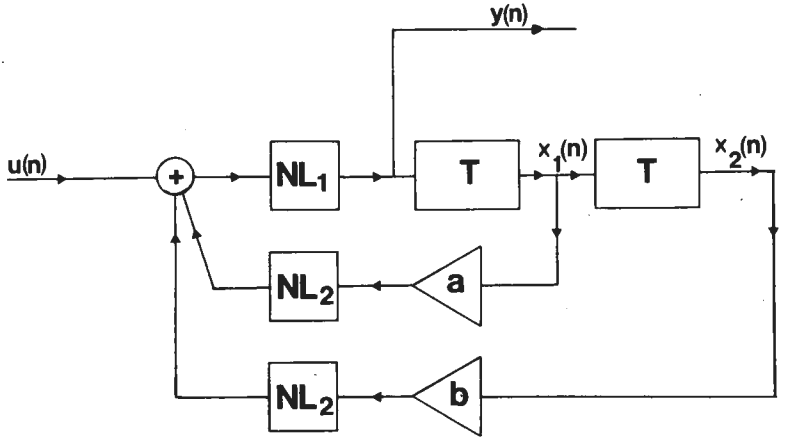


Fig. 1.2.1. Second-order digital filter in direct form.

With the wordlength corrections included, it is described by the non-linear state equations

$$\begin{aligned} x_1(n+1) &= NL_1\{NL_2\{a\cdot x_1(n)\} + NL_2\{b\cdot x_2(n)\} + u(n)\} \\ x_2(n+1) &= x_1(n) \\ y(n) &= x_1(n+1), \end{aligned} \tag{1.2.16}$$

where $NL_1\{\cdot\}$ and $NL_2\{\cdot\}$ are two wordlength reduction operators. In the idealized linear filter there is no wordlength reduction so

$$\begin{aligned} NL_1\{x\} &= x \\ \text{and} \quad NL_2\{x\} &= x. \end{aligned} \tag{1.2.17}$$

The impulse response $h(n)$ of this second-order direct form filter, satisfies the relation

$$h(n) = a\cdot h(n-1) + b\cdot h(n-2) + \delta(n) \qquad \text{for } n \geq 0, \tag{1.2.18}$$

so the transfer function $H(z)$ becomes

$$H(z) = \frac{z^2}{z^2 - a\cdot z - b}, \tag{1.2.19}$$

with poles

$$q_{1,2} = \frac{a \pm \sqrt{a^2 + 4b}}{2}. \tag{1.2.20}$$

The linear filter is stable for filter coefficients a and b in the interior of the "*stability triangle*", shown in Fig. 1.2.2, as described by the inequalities

$$\begin{aligned} 1 - |a| - b &> 0 \\ 1 + b &> 0. \end{aligned} \tag{1.2.21}$$

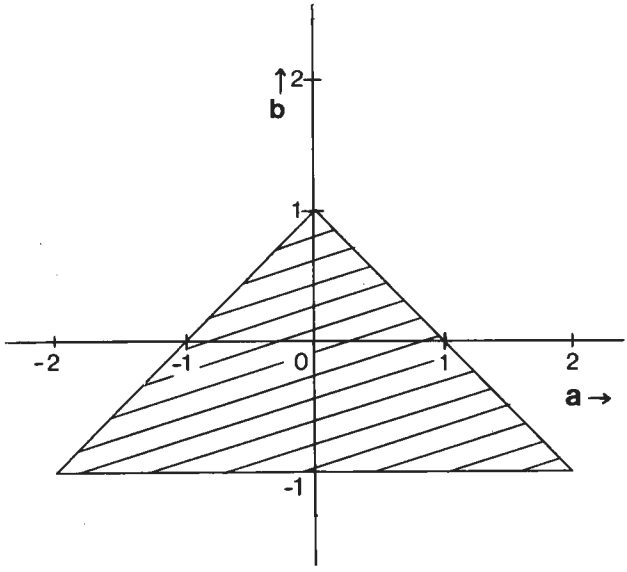


Fig. 1.2.2. Stability region of the second-order direct form filter.

The impulse response $h(n)$ of this second-order filter section is

$$h(n) = \frac{q_1^{n+1} - q_2^{n+1}}{q_1 - q_2} \quad \text{for} \quad n \geq 0. \tag{1.2.22}$$

For complex conjugated poles

$$q_{1,2} = \rho \cdot \exp(\pm j\theta), \tag{1.2.23}$$

we have

$$\begin{aligned} a &= 2\rho \cdot \cos(\theta) \\ b &= -\rho^2, \end{aligned} \tag{1.2.24}$$

and the impulse response is

$$h(n) = \frac{\rho^n \cdot \sin((n+1)\theta)}{\sin(\theta)} \quad \text{for} \quad n \geq 0. \tag{1.2.25}$$

Intermezzo

For the pair of filter coefficients $a = b = 1$, we have an unstable filter. The impulse response $h(n)$ of this filter forms the sequence

$$h(n) = 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \cdots \qquad (1.2.26)$$

which (at a first glance) can be recognized as the *Fibonacci numbers*.

This guess is correct, since $h(n)$ satisfies the well-known recurrence relation [287]

$$h(n+2) = h(n+1) + h(n) \quad \text{for } n \geq 0$$
$$\text{with} \quad h(0) = 1 \quad \text{and} \quad h(1) = 1, \qquad (1.2.27)$$

according to equation (1.2.18).

In the z-domain, this filter is described by a transfer function $H(z)$ which has poles $q_1 = \tau$ and $q_2 = 1/\tau$, with

$$\tau = \frac{1}{2} + \frac{1}{2}\sqrt{5} = \text{the } \textit{Golden section ratio}. \qquad (1.2.28)$$

According to equation (1.2.22) we can write the impulse response $h(n)$ in the form

$$h(n) = \frac{(\frac{1}{2} + \frac{1}{2}\sqrt{5})^{n+1} - (\frac{1}{2} - \frac{1}{2}\sqrt{5})^{n+1}}{\sqrt{5}}, \qquad (1.2.29)$$

which is a closed form to calculate the individual Fibonacci numbers. Further we confirm one of the special properties of the Fibonacci numbers:

$$\lim_{n \to \infty} \frac{h(n+1)}{h(n)} = \tau = \frac{1}{2} + \frac{1}{2}\sqrt{5}. \qquad (1.2.30)$$

□

In the further course of this thesis, we often consider input signals $u(n)$, which are *discrete-time periodic* with a given period N, according to

$$u(n+N) = u(n) \quad \text{for all } n. \tag{1.2.31}$$

In a stable linear time-invariant digital filter such an input signal $u(n)$ is a eigenfunction of the system. So the state $\underline{x}(n)$ and the output $y(n)$ are periodic with the same period N as the input signal, which can also be concluded from (1.2.32):

$$\begin{aligned} y(n) &= \sum_{m=-\infty}^{\infty} h(m)\cdot u(n-m) \\ &= \sum_{m=0}^{N-1} u(n-m)\cdot \sum_{i=-\infty}^{\infty} h(m+i\cdot N) \\ &= \sum_{m=0}^{N-1} u(n-m)\cdot \tilde{h}(m), \end{aligned} \tag{1.2.32}$$

where

$$\tilde{h}(n) = \sum_{i=-\infty}^{\infty} h(n+i\cdot N). \tag{1.2.33}$$

Equation (1.2.32) shows that the output $y(n)$ is determined by the circular convolution of the periodic input signal $u(n)$ with the periodic impulse response $\tilde{h}(n)$. This equation can be translated to the frequency domain.

The *Discrete Fourier Transform* $\tilde{H}(k)$ of the periodic impulse response $\tilde{h}(n)$, $(\tilde{H}(k) \in C;\ k \in Z)$, is defined according to

$$\tilde{H}(k) = \sum_{n=0}^{N-1} \tilde{h}(n)\cdot \exp\left[-j\frac{2\pi kn}{N}\right], \tag{1.2.34}$$

and the *Inverse Discrete Fourier Transform* by

$$\tilde{h}(n) = \frac{1}{N} \cdot \sum_{k=0}^{N-1} \tilde{H}(k) \cdot \exp\left[j\frac{2\pi kn}{N}\right]. \tag{1.2.35}$$

With $Y(k)$ and $U(k)$ denoting the Discrete Fourier Transform of $y(n)$ and $u(n)$ respectively, equation (1.2.32) is equivalent to

$$Y(k) = \tilde{H}(k) \cdot U(k). \tag{1.2.36}$$

Substitution of equation (1.2.33) in (1.2.34) shows that

$$\tilde{H}(k) = H(z) \qquad \text{for} \quad z = \exp\left[j\frac{2\pi k}{N}\right]. \tag{1.2.37}$$

For the second-order direct form filter we have

$$\tilde{h}(n) = \frac{1}{q_1 - q_2} \cdot \left[\frac{q_1^{n+1}}{1 - q_1^N} - \frac{q_2^{n+1}}{1 - q_2^N}\right] \tag{1.2.38}$$

and

$$\tilde{H}(k) = \frac{1}{\left[q_1 \cdot \exp\left[-j\frac{2\pi k}{N}\right] - 1\right] \cdot \left[q_2 \cdot \exp\left[-j\frac{2\pi k}{N}\right] - 1\right]}. \tag{1.2.39}$$

For complex conjugated poles the periodic impulse response $\tilde{h}(n)$ of the second-order direct form filter is

$$\tilde{h}(n) = \frac{\rho^n \cdot \sin((n+1)\theta) + \rho^{n+N} \cdot \sin((N-n-1)\theta)}{\left[1 - 2 \cdot \rho^N \cdot \cos(N\theta) + \rho^{2N}\right] \cdot \sin(\theta)}. \tag{1.2.40}$$

1.3. Types of wordlength reduction

In normal filter operation the wordlength reduction is usually performed by affecting the least significant bits only, which takes place for quantization after multiplication of a signal with a constant non-integer value. Sometimes the result of an arithmetical operation exceeds the largest machine-representable number so that an overflow correction is necesarry, which requires a change of the most significant bits. For both types of wordlength reduction there exists a number of alternative methodes (see the many review articles [21, 60, 66, 67, 92, 93, 94, 96, 127, 131, 154, 169, 170, 272, 281, 285, 289]).

For *quantization* we mainly choose from three quantization schemes with specific individual merits: (a) *rounding* RO, (b) *magnitude truncation* MT and (c) *value truncation* VT (see Fig. 1.3.1, where q denotes the quantization step size). Each method is characterized by a peculiar instruction rule concerning the direction of quantization (upwards or downwards): (a) for RO towards the nearest machine-representable number, (b) for MT towards zero and (c) for VT always downwards. Let x and $Q\{x\}$ denote the unquantized and quantized number, respectively, and let further

$$\epsilon(x) = Q\{x\} - x \tag{1.3.1}$$

denote the "*quantization error*", and q the quantization step size, then we have

$$\begin{aligned} |\epsilon_{RO}(x)| &\leq q/2 \\ |\epsilon_{MT}(x)| &< q \\ |\epsilon_{VT}(x)| &< q, \end{aligned} \tag{1.3.2}$$

which admits the conclusion that rounding is the most attractive form of quantization with regard to the average error signal amplitude.

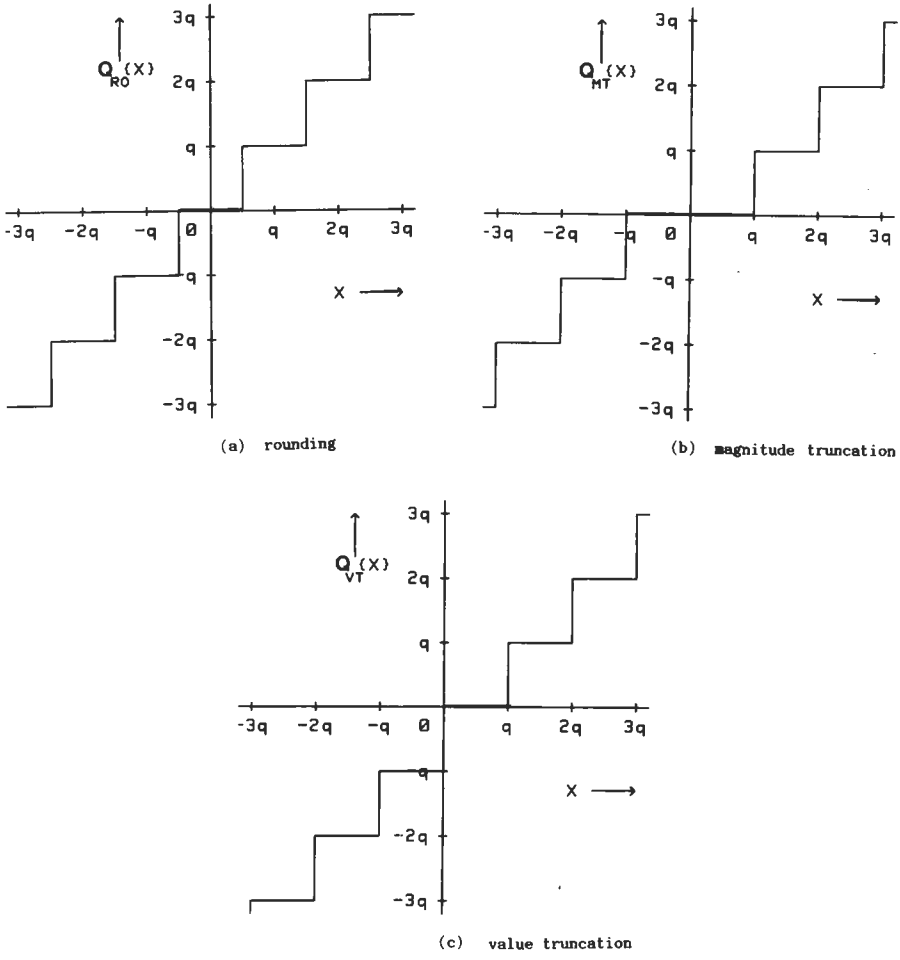


Fig. 1.3.1. Quantization characteristic for (a) rounding, (b) magnitude truncation and (c) value truncation.

The specific advantage of magnitude truncation lies in its inherent capability of limit cycle suppression, that follows from energy considerations in connection with the basic MT property

$$|Q_{MT}(x)| \leq |x|. \tag{1.3.3}$$

Finally, value truncation is the natural quantization method for a two's-complement arithmetic. Its formal treatment is similar to that of rounding due to the simple relation

$$Q_{VT}\{x\} = Q_{RO}\{x - q/2\} \qquad (1.3.4)$$

stating that VT yields the same result as RO after adding the constant signal $q/2$ to the unquantized signal.

Comparing the two main quantization schemes "rounding" and "magnitude truncation" we observe fundamental differences in their nonlinear signal processing behaviour, which follow from their error characteristics $\epsilon(x)$ (cf. Fig.1.3.2).

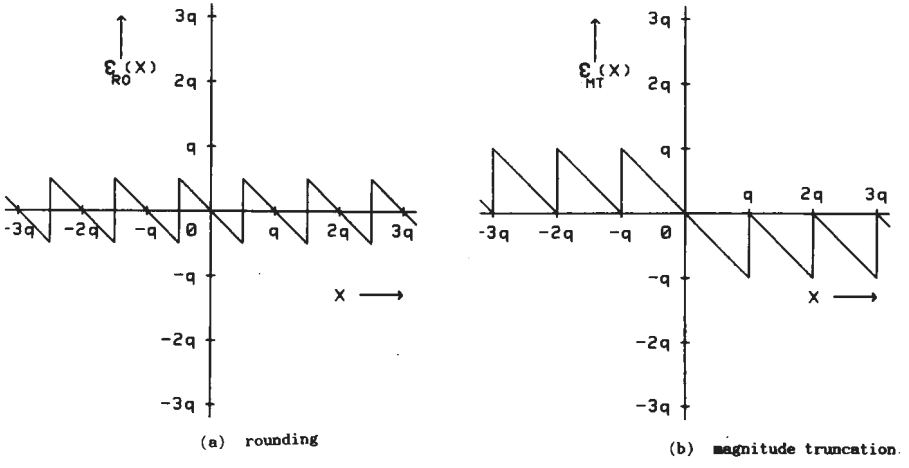


Fig. 1.3.2. Quantization error characteristic $\epsilon(x)$ for (a) rounding and (b) magnitude truncation.

It is true that both characteristics are strictly deterministic, i.e. with every x a unique signal $\epsilon(x)$ is associated. Nevertheless, we are inclined to attribute "quasi-random" features to the rounding characteristics in the following sense.

Let $x(n)$ represent a stationary random process characterized by a probability density function $P(x)$ and an autocorrelation function

$$\phi_{xx}(m) = E\{x(n)\cdot x(n+m)\}. \tag{1.3.5}$$

This process is transformed by the rounding error characteristic into another process $\epsilon(n)$ which "almost always" has *white-noise character* with

$$\phi_{\epsilon\epsilon}(m) = \frac{q^2}{12}\cdot\delta(m) \tag{1.3.6}$$

as well as a uniform probability distribution in the interval

$$-\frac{q}{2} \leq \epsilon \leq \frac{q}{2} \tag{1.3.7}$$

(see [157, 275, 320]).

This property is the basis for the well-established white-noise model of the rounding error, which we also adopt in this thesis. The reliability of this model improves with increasing level of the signal $x(n)$ and with increasing spread of its power spectrum [38]. It fails completely if $x(n)$ varies periodically, associated with a line power spectrum. Then also $\epsilon(n)$ is periodic and, hence, not noisy. Such a periodicity applies e.g. when a recursive filter oscillates in a limit cycle mode.

To analyse the corresponding error characteristic for magnitude truncation we first split it into two parts according to Fig. 1.3.3. The first part resembles the $\epsilon_{RO}(x)$ characteristic and will henceforth be referred to as the "*quasi-rounding*" component $\epsilon_{QR}(x)$ of magnitude truncation. The white-noise model of rounding error likewise applies to quasi-rounding, so that RO and MT essentially differ in the second part of the MT error characteristic, the so-called *sign-part* $\epsilon_{SGN}(x)$ (cf. Fig. 1.3.3).

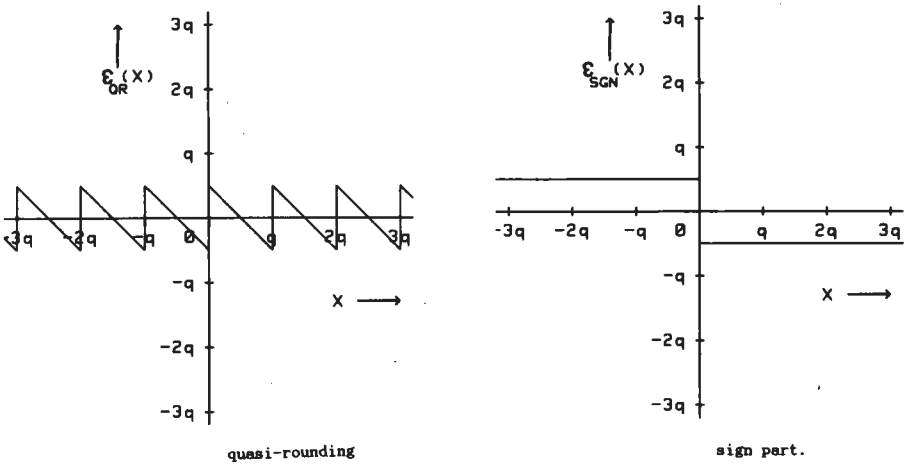


Fig. 1.3.3. Decomposition of the quantization error $\epsilon_{MT}(x)$ for magnitude truncation in a quasi-rounding and a sign part.

As to their signal-processing behaviour, the quasi-rounding and the sign part are basically dissimilar. While the former part lends itself to a modelling as an additive (white-) noise source, the latter remains an essentially nonlinear component whose output is strongly correlated with the input signal. In some applications a straight line through the origin with an appropriate negative slope can be advantageously split off from $\epsilon_{SGN}(x)$, resulting in slight modifications of the filter coefficients and, ultimately, in effects of detuning (including Q-factor modifications). Apparently such detuning is level-dependent and decreases with increasing signal amplitude. What remains is a pure nonlinear signal degradation, that leads to a number of interference phenomena (including crosstalk) and that has to be interpreted as ordinary distortion in the audio region [59, 89, 90, 91, 105, 106, 138, 216, 217, 345, 346].

While quantization has to be accepted as an unavoidable concomitant of any digital signal processing, the situation is less inevitable with respect to *overflow*. Obviously, overflow can be reduced by using floating-point representation. However, we will restrict ourselves to systems with fixed-point representation. Even in these systems, overflow can be completely avoided through using sufficiently small input signals: for a given impulse response (considered between the input terminal and a node of potential overflow) and for a prescribed overflow level an upper bound for the input signal can easily be derived. Nevertheless, it is common practice to accept a small risk of overflow, occurring for very unfavourably chosen excitations. Thus the dynamic range of a filter is better exploited, ultimately resulting in a lower quantization noise level. This mild "*scaling policy*" consciously tolerates a small nonzero probability of overflow. So, infrequent overflows and accompanying interruptions of normal operation are accepted under the obvious tacit assumption that after each overflow the normal operation recovers, preferably with high speed [156, 225].

For *overflow correction* we mainly choose from three schemes: (a) *saturation*, (b) *zeroing arithmetic* and (c) *two's-complement overflow correction* (see Fig. 1.3.4, where p denotes the overflow level). Saturation yields the smallest deviation from the normal operation, although during overflow the filter becomes more or less inoperative. It has also the best stability properties. Zeroing means that the output is set to zero, if the input exceeds the overflow threshold; it can easily be generalized to reset all states, when one state exhibits overflow. Two's complement overflow correction amounts to a periodic continuation of the 45^{o}-straight line; in fact it is not a correction because this happens automatically in two's-complement signal representation. With regard to stability it is the least favourable of these three overflow correction mechanisms so that the choice of the linear circuit is more restricted than for the other characteristics.

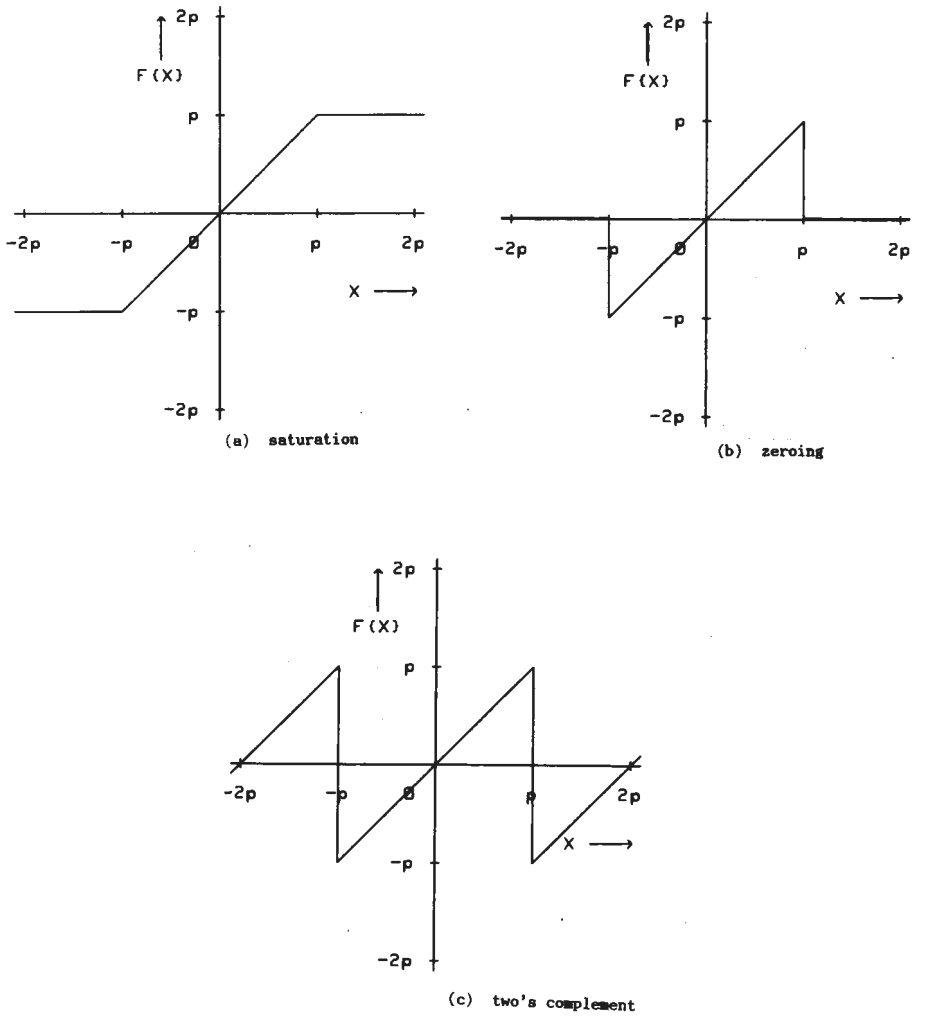


Fig. 1.3.4. Overflow characteristics for (a) saturation, (b) zeroing and (c) two's complement.

Many authors have tried to find the optimal structure, i.e. the structure achieving the lowest possible output noise power in a properly scaled filter. This problem has been solved using a state-space description of a filter [36, 147, 148, 149, 161, 180, 258, 259], while a great number of papers has been denoted to sub-optimal filters [12, 17, 35, 37, 45, 46, 54, 55, 72, 101, 119, 121, 146, 153, 159, 160, 215, 239, 240, 241, 257].

1.4. Nonlinear wordlength reduction effects

Wordlength reduction of the signals in a digital filter is a nonlinear operation, causing the filter properties to deviate from the desired specifications. The response of the nonlinear digital filter will be investigated from time $n = 0$ for two different types of inputs:

1) zero input
2) non-zero discrete-time periodic input signals.

Former input signals $u(n)$ before $n = 0$ result in an initial state $\underline{x}(0)$. Together with the input signal $u(n)$ for $n \geq 0$ this initial condition $\underline{x}(0)$ uniquely determines the evolution of the state vector $\underline{x}(n)$ for $n > 0$.
In the stable linear filter (without wordlength reduction) the state will asymptotically tend to a *steady state* $\tilde{\underline{x}}(n)$, which is independent of the initial condition $\underline{x}(0)$ and is determined by methods discussed in Section 1.2. In general, the state $\underline{x}(n)$ of the actual digital filter will asymptotically ($n \to \infty$) differ from the steady state $\tilde{\underline{x}}(n)$ of the idealized linear filter. The difference $\underline{e}(n)$ will be referred to as the *error vector*:

$$\underline{e}(n) = \underline{x}(n) - \tilde{\underline{x}}(n). \tag{1.4.1}$$

In case 1) of a *zero-input signal* the steady state vector $\tilde{\underline{x}}(n)$ of the stable linear filter equals zero:

$$\tilde{\underline{x}}(n) = \underline{0} \qquad \text{for all } n \geq 0. \tag{1.4.2}$$

The nonlinear digital filter is called *zero-input stable* if the state $\underline{x}(n)$ reaches $\underline{0}$ for $n \to \infty$, independent of the initial condition $\underline{x}(0)$.

Due to quantization and overflow, each component of $\underline{x}(n)$ can only assume a finite number of values. Therefore, if the state $\underline{x}(n)$ does not converge to zero, it must, at some finite time $n = N_o$, come across a previous value $\underline{x}(N_o\text{-}S)$. The filter will exhibit a periodic oscillation of period S, which is called a *zero-input limit cycle*. This includes

It is possible that there exists more than one limit cycle in a digital filter. Which of these limit cycles actually appears in the filter depends upon the initial state $\underline{x}(0)$ (see [155, 276, 279, 302, 356]).

Some filters structures can be guaranteed to be free from zero-input limit cycles. This freedom is usually proved with the *second method of Lyapunov* (see [107]). This method starts with a properly chosen *energy function* $E(n)$, preferably in quadratic form:

$$E(n) = \underline{x}^T(n) \cdot P \cdot \underline{x}(n). \tag{1.4.3}$$

Without loss of generality, the matrix P can be chosen symmetrical:

$$P^T = P. \tag{1.4.4}$$

The Lyapunov theory demands that

$$E = \underline{x}^T \cdot P \cdot \underline{x} > 0 \quad \text{for all } \underline{x} \neq \underline{0}, \tag{1.4.5}$$

so that $E(n) = 0$ implies $\underline{x}(n) = \underline{0}$. From (1.4.5) it follows that P has to be positive definite. Further, the system dynamics must be such that if no wordlength reduction is applied, according to

$$\underline{x}_o(n) = A \cdot \underline{x}(n-1), \tag{1.4.6}$$

the energy strictly decreases with increasing time n:

$$E_o(n) < E(n-1) \quad \text{for all } n > 0, \tag{1.4.7}$$

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

$$P - A^T \cdot P \cdot A \text{ is positive definite.} \tag{1.4.8}$$

Every matrix P satisfying condition (1.4.4), (1.4.5) and (1.4.8) defines an energy function $E(n)$, which is a candidate for a Lyapunov function. If moreover for one such a matrix a subsequent wordlength reduction

$$\underline{x}(n) \;=\; \mathrm{NL}\{\underline{x}_o(n)\} \;^{\dagger} \tag{1.4.9}$$

lowers the energy for all possible states, the function $E(n)$ is called a *Lyapunov function* of the nonlinear system under consideration. The above condition

$$E(n) \;\leq\; E_o(n) \qquad \text{for all } n \geq 0 \tag{1.4.10}$$

reads in terms of the nonlinear (wordlength reduction) operation $\mathrm{NL}\{\cdot\}$ as

$$[\mathrm{NL}\{\underline{x}_o\}]^T \cdot P \cdot \mathrm{NL}\{\underline{x}_o\} \;\leq\; \underline{x}_o{}^T \cdot P \cdot \underline{x}_o \qquad \text{for all } \underline{x}_o. \tag{1.4.11}$$

The existence of a Lyapunov function in a digital filter (with wordlength reduction) guarantees freedom from limit cycles. The idealized linear filter is assumed to be stable, so the state in the filter without wordlength reduction asymptotically approaches the zero-state, for which also the energy is zero. In the actual digital filter the necessary wordlength reduction lowers the energy $E(n)$, implying that also here the energy asymptotically reaches zero, implying zero-state. Therefore a filter with a Lyapunov function is free from zero-input limit cycles.

---

† At a first glance the splitting of the transformation $\underline{x}(n\text{-}1) \rightarrow \underline{x}(n)$ into a wanted linear part (1.4.6) and a nonlinear correction (1.4.9) seems to be arbitrary and inadequate to describe the most general transformation. However, if x_o is eliminated from (1.4.6) and (1.4.9), the last equation reads $\underline{x}(n) = \mathrm{NL}\{A \cdot \underline{x}(n\text{-}1)\}$, which for nonsingular A is equivalent to the general transformation $\underline{x}(n)$ = function of $\underline{x}(n\text{-}1)$. Notice that in many structures $\mathrm{NL}\{\underline{x}\}$ is a true multidimensional function in the sense that it contains couplings between the various components of $\underline{x}$. It is only in the (albeit important) case of wordlength reduction of the individual state variables (before storing them in a memory) that $\mathrm{NL}\{\underline{x}\}$ is a scalar function independently acting upon the individual components of $\underline{x}$.

Care must be taken if "<" is replaced by "≤" in (1.4.7) so that the energy can remain constant. Such a situation occurs for a marginal choice of the matrix P, for which in the filter without wordlength reduction the energy function is called a *semi-Lyapunov function*. If, moreover, the equality sign in (1.4.11) applies it can occur that the energy remains constant, associated with the risk of a zero-input limit cycle.

In case 2) the input signal $u(n)$ of the digital filter is chosen to be *discrete-time periodic*, with a given period N. For $N = 1$ the input signal is constant for all $n \geq 0$. The idealized linear digital filter, satisfying the conditions for stability, has an asymptotic steady state $\tilde{\underline{x}}(n)$, which is periodic with the same period N. This need not to be true for the digital filter with nonlinearities. The state $\underline{x}(n)$ in the actual digital filter, with the same filter coefficients, can exhibit an asymptotic oscillation with a different period.

For a finite state machine like a digital filter it is easily recognized that after a finite number of periods N_o the state $\underline{x}(n)$ will always enter a periodic function, but in general with a different period. Such a state satisfies the periodicity condition

$$\underline{x}(n+S\cdot N) = \underline{x}(n) \text{ for all } n \geq N_o \cdot N, \tag{1.4.12}$$

where the integer value of S is chosen as small as possible.
(Note that the requirement that S be an integer can imply that the elementary period of $\underline{x}(n)$ is a fraction of $S\cdot N$.)

The validity of the above statement is due to the fact that during a step by step increase of n with steps of N the state vector $\underline{x}(n)$ can only assume one of the finite number of values afforded by the machine. Maximally after all states have been covered (but in general far earlier) it returns to a previous state. But then not only the state but also the input signal assumes the former value. Because $\underline{x}(n+1)$ is uniquely determined by $\underline{x}(n)$ and the input signal $u(n)$, the whole cycle is repeated and the state has entered periodicity (see [61, 64, 65]).

In passing we note that even if overflow is neglected the digital filter in fact behaves like a finite-state machine. This stems from the fact that under linear stability and for a given initial state $\underline{x}(0)$ and input $u(n)$ only a finite region of the state space is reachable. The size of this region can be estimated with the aid of methods similar to those commonly used for the estimation of the maximum amplitude of zero-input limit cycles, as will be shown in Section 4.1.

If in the periodic state sequence, satisfying (1.4.12), the value of S is larger than unity, so that the period of the state is an integer multiple of the period of the input signal the filter produces a S^{th}-*order subharmonic*. If $S = 1$, that is if the period of the state equals that of the input signal, which is the "normal type" response with preserved period, no subharmonic is produced.

Moreover, for a given input signal a whole set of periodic solutions can occur. Which of these oscillation actually appears in the digital filter depends upon the initial state $\underline{x}(0)$ (see [151, 322]). These periodic state sequences differ in form and can also have different periods. However, not every initial state $\underline{x}(0)$ is associated with a periodic state sequence of its own. Rather, a group of initial conditions generally leads to the same oscillation. This situation is much the same as for zero-input limit cycles. There we call a filter *stable* if it is free from limit cycles so that any initial condition leads to the zero state. For periodically excited digital filters the logical extension of *stability* requires that there is only one asymptotic oscillation, which is reached from all initial states $\underline{x}(0)$.

Since associated with every subharmonic we can formally distinguish S distinct waveforms, which are shifted replicas of each other with a mutual shift equal to the period N of the input signal, a filter that produces subharmonics is apparently unstable. Stability in this sense implies preservation of period.

Summarizing we have three major differences between the filter responses in the idealized situation, where no wordlength reduction is applied, and the practical digital filter, with quantization and overflow.

In a filter without wordlength reduction starting from an initial state $\underline{x}(0)$ the *linear filter* theory predicts in a stable filter:

- the steady state is reached asymptotically after an infinite number of timesteps n
- the resulting waveform is unique
- the period of the response equals that of the excitation.

In a practical digital filter with *finite-wordlength reduction* the following effects are observed:

- the steady state is entered after a finite number of periods N_o
- the resulting waveform is non-unique but depends upon the initial state $\underline{x}(0)$
- the signal period need not be preserved; the occurrence of sub-harmonics is possible.

If for a given input signal the state is in a periodic oscillation a small change of the state vector $\underline{x}(n)$, caused f.e. by an additional pulse on the input, can let the output jump from one oscillation to another. This effect is called the *jump phenomenon*. It will only occur for input signals for which the digital filter has two or more periodic state solutions, implying instability [187, 188]. A digital filter excited by two or more sinusoids can generate all kind of inter-modulation products [11, 47, 48, 256, 312].

For periodically excited digital filters stability requires that there exists only one asymptotic oscillation, which is reached from all initial conditions.

If we only consider overflow the steady state $\tilde{\underline{x}}(n)$ of the idealized linear filter is also an asymptotic oscillation of the actual digital filter, entered directly from some suitable initial state $\underline{x}(0)$. (It is assumed that the components of $\tilde{\underline{x}}(n)$ do not exceed the overflow level p). Stability requires that the steady state $\tilde{\underline{x}}(n)$ is the only asymptotic oscillation. Freedom from overflow instabilities in the forced response can then again be proved with the *second method of Lyapunov*. Just as for zero-input signals we define a properly chosen quadratic energy function $E(n)$ characterized by a symmetric and positive definite matrix P, according to

$$E(n) = \underline{e}^{T}(n) \cdot P \cdot \underline{e}(n), \tag{1.4.13}$$

where $\underline{e}(n)$ is the error vector (see (1.4.1)).
The system dynamics must be such that if no wordlength reduction is applied the energy $E(n)$ decreases with increasing time n and asymptotically approaches zero. The nonlinear operation NL{·} must lower the energy even more to guarantee freedom from forced response instabilities, so that also in the actual digital filter the energy asymptotically reaches zero. This implies that the state $\underline{x}(n)$ reaches the steady state $\tilde{\underline{x}}(n)$ of the idealized linear digital filter.
If also quantization is considered the steady state $\tilde{\underline{x}}(n)$ of the idealized filter will in general be not an asymptotic oscillation of the actual digital filter, since this can only be true if all the values of the components of $\tilde{\underline{x}}(n)$ are integer multiples of the quantization step q. Freedom from quantization instabilities in the forced response can therefore rarely be proved with the method of Lyapunov, which makes the stability test rather difficult.

2. Overflow stability in recursive digital filters

2.1. Introduction

In a common approximation, quantization and overflow are not only conceptually *decoupled*, but also analytically treated as *independent effects*. This implies that for large signals the fine quantization structure is neglected. For a filter with overflow level p this means that the nonlinearity is approximated as follows:

$$\begin{aligned} F(x) &= x \quad \text{for} \quad |x| \le p \\ |F(x)| &\le p \quad \text{for} \quad |x| > p, \end{aligned} \tag{2.1.1}$$

where $F(x)$ is the nonlinear overflow characteristic.

Apparently this approximation can only be justified if the total number of quantization steps is large enough or, in other words, the binary words are sufficiently long. Even for this extreme case several authors have queried the validity of the decoupling approximation [302, 308, 361, 362, 363].

Indeed, there are overflow effects that can only be properly understood in connection with the quantization fine structure. As an example, consider a filter initially in the zero state and then excited by a short, strong pulse such that overflow occurs at some point inside the filter. Assume that the idealized (quantization-free) filter asymptotically returns to equilibrium (zero state), which implies "*overflow stability*". Apparently, the filter has "forgotten" the overflow after a sufficiently long time. With quantization, the situation is not as simple: before excitation, the filter might (necessarily) oscillate in a limit cycle mode, while after overflow the filter does not recover to the zero state but again enters a limit cycle. The mode of oscillation can, however, be completely different from the former one. Because the filter never forgets the overflow, it has apparently to be considered as unstable.

Recently, *chaotic overflow oscillations* have been observed [80, 213, 247, 249]. Also in that case the quantization has been neglected in the first instance. Taking the fine structure of the finite wordlength characteristic into account, the filters under consideration become finite-state machines with strictly periodic (non-chaotic) oscillations. These examples belong to a small group of exceptional phenomena where the decoupling assumption fails even for a large dynamic range (long binary words). For most effects to be treated in this thesis it is valid with sufficient accuracy.

In this chapter we demonstrate some methods for suppressing overflow oscillations in elementary digital filters. To discuss the item of overflow stability we assume that the underlying idealized, linear system is stable and that quantization can be neglected (decoupling assumption). Then the stability problem is attacked in two situations, (a) under zero-input conditions, (b) under nonzero-input conditions. Stability according to (a) is defined as absence of spontaneous oscillations, particularly of periodic nature. A system stable in this sense is asymptotically (from a certain time instant N_o) overflow-free. Then it behaves linearly and exponentially approaches the equilibrium point in which all state variables become zero. Stability according to (b), the so-called "*forced-response stability*" is defined for a certain class U_o of input signals $u(n)$. Such signals are defined with the aid of the idealized linear system and characterized by the property that for at least one initial condition the overflow threshold p is never reached. For periodic input signals $u(n)$ this definition requires that no component of the steady state $\tilde{\underline{x}}(n)$ of the idealized linear filter does exceed the overflow level p.

The filter with overflow correction is then called "*forced-response stable*" if for any $u(n) \in U_o$ and any initial condition $\underline{x}(0)$ the response asymptotically ($n \to \infty$) approaches the waveform of the linear counterpart. If the response of the filter does not asymptotically approach this waveform the filter has a *forced-response overflow oscillation* (see [86, 87, 88, 298, 299, 300, 361, 366]).

So for the given class of input signals forced-response stability implies that the actual filter eventually "forgets" former overflows and becomes overflow-free. Clearly, forced-response stability is a stronger condition than zero-input stability and includes the latter. If the system is excited with a rather irregular waveform, zero-input stability will often suffice; only for periodic waveforms the stronger condition is strictly required.

In recursive filters, quantization and overflow can lead to instabilities, even if the underlying linear filter is designed to behave stable. Instabilities due to quantization ("limit cycles") lead to relatively small deviations from the linear behaviour. While these effects will be treated in the next chapter, we now deal with those instabilities that are related to register overflow. The associated oscillations have large amplitudes; because of their disastrous effects on the filter behaviour they have to be absolutely avoided. One of the main factors determining their occurrence is the "overflow characteristic" (i.e. the way overflow is corrected), of which we treat the three commonly used types (a) saturation, (b) zeroing and (c) two's complement. These characteristics are used in first-order, second-order, direct form, wave, normal and state-space digital filters.

2.2. Overflow stability in first-order digital filters

In this section we investigate the zero-input and forced-response stability of the *first-order recursive digital filter* with overflow correction. This filter is shown in Fig. 2.2.1 and described by the state equations

$$\begin{aligned} x(n+1) &= NL_1\{a\cdot x(n) + u(n)\} \\ y(n) &= x(n+1), \end{aligned} \tag{2.2.1}$$

where $NL_1\{x\} = F\{x\}$ denotes the overflow correction function (henceforth the overflow level p is normalized to unity).

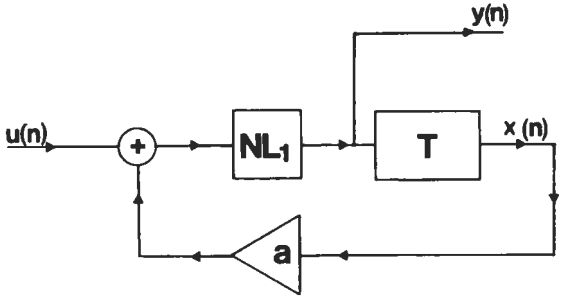


Fig. 2.2.1. First-order recursive digital filter.

The idealized linear filter, in which there is no wordlength reduction, is stable for values of the filter parameter a satisfying the condition

$$|a| < 1. \tag{2.2.2}$$

The nonlinear filter is free from zero-input oscillations for all possible overflow characteristics. This zero-input stability can be proved with the second method of Lyapunov. To this end we define the generalized energy function

$$E(n) = x^2(n). \tag{2.2.3}$$

Then, due to equation (2.1.1)

$$E(n) = [F\{a \cdot x(n-1)\}]^2$$

$$\leq a^2 \cdot x^2(n-1) = a^2 \cdot E(n-1) \qquad (2.2.4)$$

So

$$E(n) \leq a^{2n} \cdot E(0) \qquad (2.2.5)$$

and

$$\lim_{n \to \infty} E(n) = 0. \qquad (2.2.6)$$

Independent of the method used for overflow correction, the energy $E(n)$ and herewith the state $\underline{x}(n)$ asymptotically reach zero and no overflow oscillation can occur in the first-order recursive digital filter for zero-input signals.

On the other hand, the forced response of this digital filter can exhibit overflow oscillations. This occurs e.g., for two's-complement overflow correction, as can be concluded from the following example.

Example

For a filter with a filter coefficient $a = -1/3$ and excited by a constant-input signal $u(n) = 1.2$ the steady state $\tilde{x}(n)$ of the idealized linear filter is constant with a value $\tilde{x}(n) = 0.9$. In this filter a forced-response overflow oscillation of period 1 is possible with a state value $x(n) = -0.6$. For an initial state $x(0) > 0.6$ the filter tends to the linear response, for $x(0) < 0.6$ the overflow oscillation is asymptotically reached.

□

No overflow oscillations are possible in the first-order digital filter if saturation is used for overflow correction.

This means that the actual state signal $x(n)$ in this filter asymptotically reaches the steady state $\tilde{x}(n)$ of the idealized linear filter, independent of the initial condition x(0). This forced-response stability is only defined for a certain class of input signals U_o for which the asymptotic state of the linear filter remains below the overflow threshold

$$|\tilde{x}(n)| \leq 1 \qquad \text{for all } n. \tag{2.2.7}$$

The forced response stability can be proved with the method of Lyapunov. Define the generalized energy function

$$E(n) = e^2(n) \tag{2.2.8}$$

with

$$e(n) = x(n) - \tilde{x}(n). \tag{2.2.9}$$

Define $x_o(n)$ as the arithmetical result before overflow correction is applied:

$$x_o(n) = a \cdot x(n-1) + u(n-1) \tag{2.2.10}$$

and $E_o(n)$ as the energy of this uncorrected signal:

$$E_o(n) = [x_o(n) - \tilde{x}(n)]^2 = a^2 \cdot E(n-1). \tag{2.2.11}$$

If overflow correction is needed because $x_o(n) > 1$ then after saturation $x(n) = 1$ and

$$E(n) = [1 - \tilde{x}(n)]^2$$

$$< [x_o(n) - \tilde{x}(n)]^2 = E_o(n) \tag{2.2.12}$$

The same conclusion can be drawn for $x_o(n) < -1$.

We conclude that

$$E(n) \leq E_o(n) = a^2 \cdot E(n-1) \leq a^{2n} \cdot E(0) \qquad (2.2.13)$$

and

$$\lim_{n \to \infty} E(n) = 0, \qquad (2.2.14)$$

implying

$$\lim_{n \to \infty} e(n) = 0. \qquad (2.2.15)$$

This completes the proof for the forced-response stability of the first-order recursive digital filter with saturation as overflow correction.

2.3. Overflow stability in second-order digital filters

In this section we investigate the overflow stability of *second-order digital filters*. We begin with a study of overflow oscillations in the original sense, i.e. for an otherwise unexcited digital system. In addition to this "zero-input" condition we assume in this section that (a) overflow and quantization can be treated independently ("decoupling assumption") and (b) overflow correction is only required for signals entering a delay element. The latter assumption forbids intermediate overflows. For sake of conciseness, we restrict the following discussion to second-order sections with complex poles. Compared with real poles, complex conjugated pole-pairs generally favour all forms of parasitic oscillations (particularly for high Q-values) and thus deserve special consideration. In due course, we summarize more general results without reference to complex pole pairs.

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

$$\underline{x}(n+1) = F\{A \cdot \underline{x}(n)\} \tag{2.3.1}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{2.3.2}$$

denotes the system matrix, while $F\{\cdot\}$ stands for the overflow characteristic. In this section it is understood that

$$[F\{A \cdot \underline{x}\}]_i = F\{[A \cdot \underline{x}]_i\}, \tag{2.3.3}$$

i.e. the individual components of $A \cdot \underline{x}$ undergo the same memoryless and local (i.e. not controlled by other signals) overflow correction.

The question to be analyzed is: under which circumstances (choice of A, $F\{\cdot\}$ and $\underline{x}(0)$) does (or does not) (2.3.1) admit periodic solutions?

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

$$|x_i(n)| \leq 1, \tag{2.3.4}$$

resulting in a state vector confined to the interior of the unit square (cf. Fig. 2.3.1). Without overflow (i.e. as long as (2.3.4) satisfies) the solution of (2.3.1) is found as

$$\begin{aligned}\underline{x}(n) &= \mathrm{Re}\{X_o(\underline{\xi}_r+j\cdot\underline{\xi}_i)\cdot e^{[(\Gamma+j\theta)n+j\varphi_o]}\} \\ &= X_o\rho^n\cdot[\underline{\xi}_r\cdot\cos(\theta n+\varphi_o) - \underline{\xi}_i\cdot\sin(\theta n+\varphi_o)],\end{aligned} \tag{2.3.5}$$

where $q_{1,2} = e^{\Gamma\pm j\theta} = \rho\cdot e^{\pm j\theta}$ 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, without loss of generality, the real and imaginary parts $\underline{\xi}_r$, $\underline{\xi}_i$ of the suitably normalized eigenvector are assumed to be orthogonal,

$$\underline{\xi}_r^T\cdot\underline{\xi}_i = 0. \tag{2.3.6}$$

(This freedom is provided by the indeterminancy of the complex magnitude of any eigenvector.) Finally, the constants of integration (X_o, φ_o) 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 (unrealizable) case $\Gamma = 0$ this would be an ellipse with main axes in the direction of the eigenvectors $\underline{\xi}_r$ and $\underline{\xi}_i$.

For $\Gamma < 0$ (corresponding to poles inside the unit circle), we obtain a nonclosed, ellipse-like curve spiralling towards the origin, cf. Fig. 2.3.1.

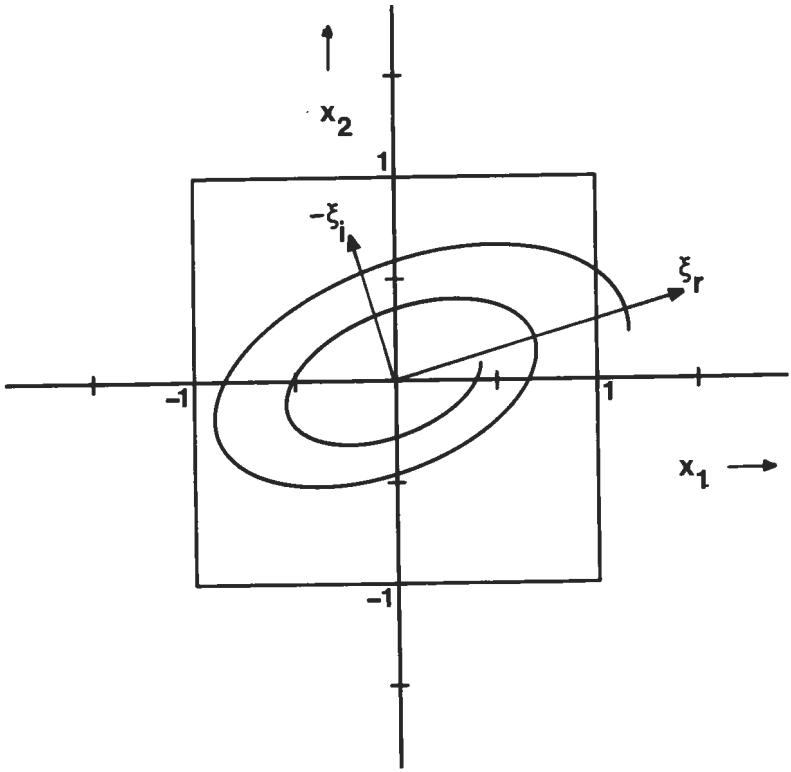


Fig. 2.3.1. Trajectory of the state vector $\underline{x}(n)$ in the state plane.

Of course, these results only apply to the digital filter as long as overflow does not occur ($|x_i(n)| \leq 1$). In general, this condition is not met for all initial conditions $\underline{x}(0)$ inside the unit square. Only the initial vectors $\underline{x}(0)$ of the region R of Fig. 2.3.2 lead to "allowed" $\underline{x}(n)$ for all (continuous) values of n.
What occurs if $\underline{x}(0)$ is outside R? Then, at some time instant n, the linearly determined $\underline{x}(n)$ might leave the unit square, and overflow correction has to be applied.

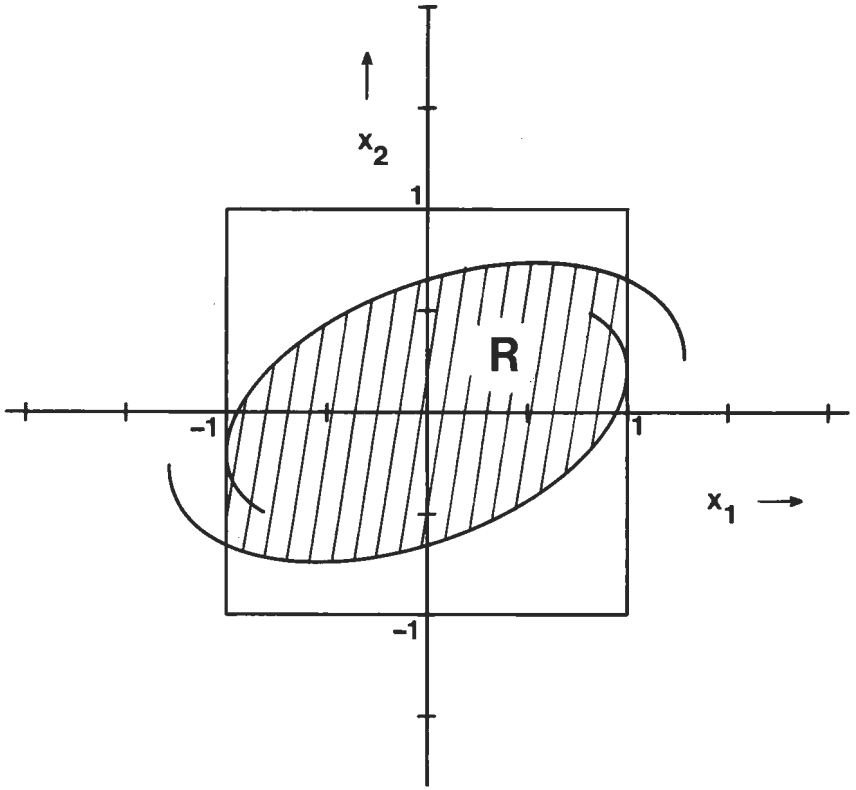


Fig. 2.3.2. Region R of initial states that never lead to overflow.

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 (2.3.5).

Following

$$\underline{x} = X\cdot[\underline{\xi}_r\cdot\cos(\varphi) - \underline{\xi}_i\cdot\sin(\varphi)] \qquad (2.3.7)$$

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

$$X^2 = \left[\frac{\underline{\xi}_r^T \cdot \underline{x}}{\underline{\xi}_r^T \cdot \underline{\xi}_r}\right]^2 + \left[\frac{\underline{\xi}_i^T \cdot \underline{x}}{\underline{\xi}_i^T \cdot \underline{\xi}_i}\right]^2 . \tag{2.3.8}$$

Comparing (2.3.7) with the linear motion as described in (2.3.5) one recognizes

$$X(n) = X_o \cdot e^{\Gamma n} = X_o \cdot \rho^n , \tag{2.3.9}$$

i.e. a monotonically decreasing function. Combined with the fact that X^2 is a quadratic form in x_1, x_2 as formulated by (2.3.8), this quantity is a natural candidate for a Lyapunov function:

$$E(n) = X^2(n) . \tag{2.3.10}$$

Observe that the curves X = constant constitute a family of "concentric" ellipses (with axes along $\underline{\xi}_r$ and $\underline{\xi}_i$) and that low X ellipses are enclosed by high X ellipses. Naturally, we choose X as the "distance from the origin".

Overflow correction is now visualized in Fig. 2.3.3. 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 X 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.

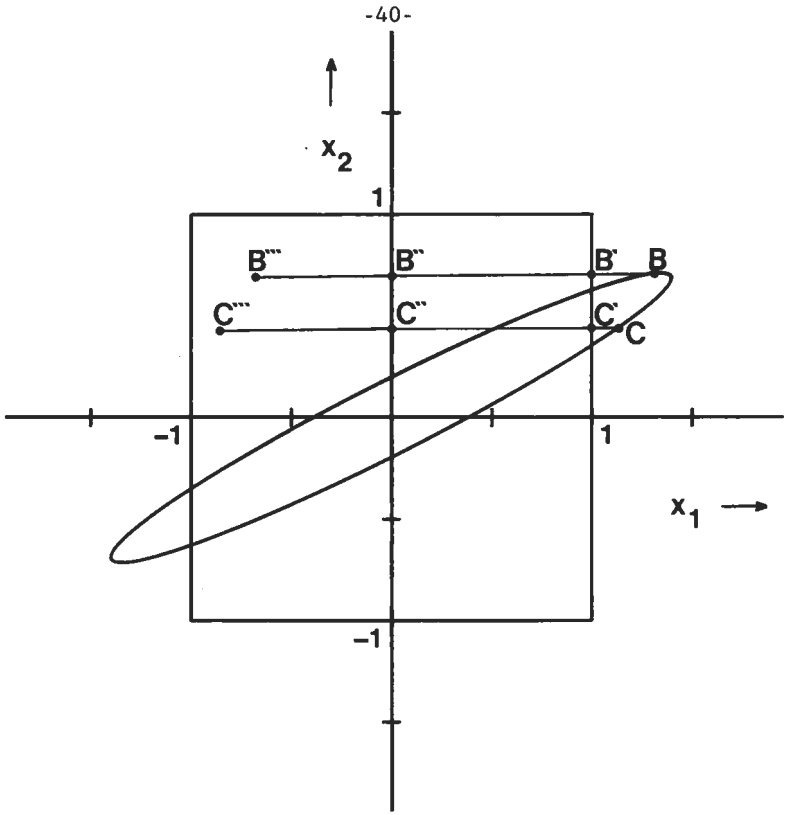


Fig. 2.3.3. Ellipse X = constant in the state plane.

Obviously this is not the case for the arbitrarily oriented ellipse of Fig. 2.3.3. 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.

It should be noted that in this picture the potentiality for stabilizing overflow is determined by the eigenvectors of A and not by the eigenvalues. While the latter determine the speed with which trajectories are traversed, the eigenvectors determine the appearance of the ellipse, i.e. the orientation of the axes and their length ratio. These parameters are essentially determined by the filter structure, examples of which are the direct form filter, the wave digital filter and the normal filter.

Until now we have discussed sufficient conditions guaranteeing that no zero-input overflow oscillations occur. The non-existence of such oscillations was viewed as an absolute design requirement that every usable filter has to meet.

An ill-designed filter can exhibit autonomous oscillations under suitable initial conditions. Physically, these are e.g. determined through connecting the digital circuit to a power supply or as a residue of former (meanwhile terminated) input signals. Such an initial condition need not immediately cause overflow but can lead to it after a number of time steps. Thereafter overflow becomes periodic or asymptotically periodic or irregular (chaotic). All these instabilities are characterized by the non-existence of a time instant, after which overflow ceases to occur.

On the other hand, stability implies that such a time instant does exist. This requirement is also the starting point for the *forced response stability* to be discussed in the remainder of this section [83, 86, 87, 88, 133, 134, 361, 362, 363, 366]. Occasional overflows are allowed, but there has to exist a last overflow, after which the system behaves linearly and thus recovers from potential former overflows. Asymptotically ($n \to \infty$) there remains the "forced response", which is independent of the initial conditions and, as such, not affected by all former overflows.

Stability in this sense depends upon the excitation. For each digital filter a (possibly empty) set of input signals exists for which stability holds. An apparent minimum requirement is that only such input signals $u(n)$ are admitted for which the associated linear filter (without overflow correction) does not exceed the overflow level p after some time N_o. The ensemble of all such signals (with N_o unspecified) is said to form the class U_o (definition "A"). Besides this definition "A" an alternative definition "B" is in current use which examines $u(n)$ only for $n \geq N_o$.

Following "B" we have $u(n) \in U_o$ if and only if there exists an initial condition at $n = N_o$ such that the linear filter does not exceed the overflow level for all $n \geq N_o$. Apparently, the past history of $u(n)$ in the "A" interpretation is condensed in the initial condition according to "B" so that the "tails" of the "A"-signals form the class U_o in the "B" sense [86, 87, 88]. †

A stable filter with overflow correction always exhibits a finite number of overflows (which may be zero or one in special cases) after $n = N_o$, which number depends upon the initial condition at N_o. Assuming that $u(n) \in U_o$, there is at least one initial condition (mostly a set of neighbouring initial conditions) with no overflow after $n = N_o$.

If stability in the above sense holds for all $u(n) \in U_o$, the filter is called "*forced-response stable*" with respect to U_o [88]. Since excitations $u(n) \notin U_o$ are meaningless in the context of stability, the addition "with respect to U_o" is often omitted. Weaker forms of stability are found with respect to subsets of U_o such as U_o^c with a scale factor c satisfying $0 \leq c < 1$. Compared with U_o the signal amplitudes are reduced by a factor c such that $u(n)/c \in U_o$. In this notation $c = 0$ corresponds to "zero-input stability" being the weakest form of stability.

It is somewhat surprising that systems whose stability is guaranteed only for zero input also behave stable for most excitations of practical importance. In fact, only periodic or almost-periodic signals [298, 299, 300] appear to produce forced-response instabilities (with commensurate periods) in such systems.

---

† In an uncontrollable system it can occur that not all initial conditions can be generated with the aid of suitable input signals. In such an exceptional case, the "B" definition is more general. This definition was already introduced in Section 2.1

Concerning the analytical investigations of forced-response stability it is a lucky circumstance that the non-zero input problem can be transformed into a zero-input problem with time-varying nonlinearities [88]. Let $\underline{x}(n)$ and $\tilde{\underline{x}}(n)$ denote the state vectors of the actual and the idealized filter with excitation $u(n)$ such that

$$\underline{x}(n+1) = F\{A\cdot\underline{x}(n) + \underline{b}\cdot u(n)\}$$

$$\tilde{\underline{x}}(n+1) = A\cdot\tilde{\underline{x}}(n) + \underline{b}\cdot u(n) \qquad (2.3.11)$$

then the difference

$$\underline{e}(n) = \underline{x}(n) - \tilde{\underline{x}}(n) \qquad (2.3.12)$$

satisfies the difference equation

$$\underline{e}(n+1) = F\{A\cdot\underline{e}(n) + \tilde{\underline{x}}(n+1)\} - \tilde{\underline{x}}(n+1). \qquad (2.3.13)$$

Let us consider a certain component of $\tilde{\underline{x}}(n+1)$ and $A\cdot\underline{e}(n)$ and denote it provisionally by μ and ν, respectively. Then the same component of the right-hand term of (2.3.13) reads as $F\{\nu + \mu\} - \mu$, i.e. a time-varying (due to $\mu = \mu(n)$) nonlinear function of ν. With a linearly determined $\mu(n)$ the function $F\{\nu + \mu\} - \mu$ is a shifted replica of $F\{\nu\}$, with equal horizontal and vertical μ-shifts of the F-plot. Fig. 2.3.4 shows the result for the three basic overflow characteristics.

With the knowledge that for many structures (e.g. normal and wave-digital filters c.f. Section 2.5) the condition $|F\{\nu\}| \leq |\nu|$ ensures zero-input stability we can likewise conclude that (2.3.13) has a stable solution (with $\underline{e}(n) \to \underline{0}$ for $n \to \infty$) if $|F\{\nu+\mu\}-\mu| \leq |\nu|$. From Fig. 2.3.4 we conclude that this is true for saturation if $|\mu| < p$, for zeroing if $|\mu| < 0.5\cdot p$, and for two's complement if $\mu = 0$, i.e. for excitations that are elements of the classes U_o, $U_o^{0.5}$, U_o^0, respectively, (in the sense of definition "A" as given above).

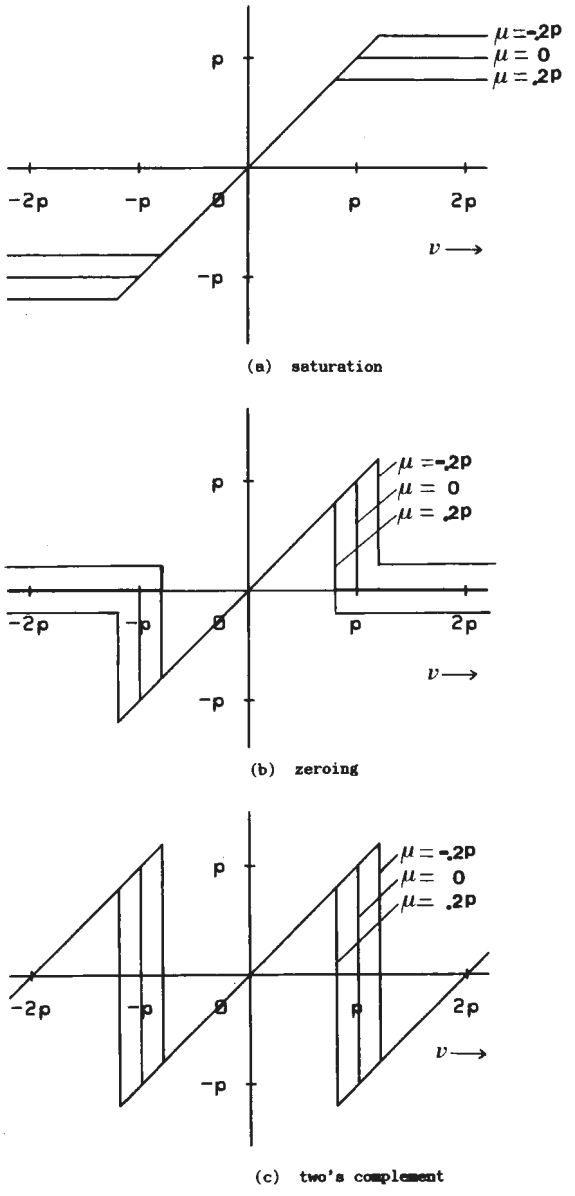


Fig. 2.3.4. Plots of $F\{\nu + \mu\} - \mu$ for (a) saturation (b) zeroing and (c) two's complement.

We conclude this section with some phenomena occurring in an unstable filter. For a given $u(n) \in U_o$ there exists a set of initial conditions, for which no overflow occurs. In general, there exists another set of initial conditions, which leads to a finite, nonzero number of overflows. Finally, due to the assumed instability, a third set of initial conditions gives rise to an infinite number of overflows. It is only in this situation that the instability becomes manifest. For a periodic excitation, the response, too, becomes asymptotically periodic, but the period need not be the same. Subharmonics can occur, but also completely different periods are observed [61, 64, 65, 83]. For an input signal, which consists of two sinusoids intermodulation frequencies have been observed [11, 47, 48, 256, 312].

In general, the asymptotic response is not unique, even if the periods of excitation and response are equal. Additional pulse excitations can lead to jump phenomena from one response to another [187, 188].

2.4. Overflow stability in direct form digital filters

In this section we investigate the zero-input and forced-response stability of *direct form digital filters* with overflow correction. In these filters the coefficients of the transfer function $H(z)$ are realized directly in the filter structure [278]. The major part of this section concerns the second-order filter, depicted in Fig. 1.2.1 and described by the state equations (1.2.16). The nonlinear operation $NL_1\{\cdot\} = F\{\cdot\}$ denotes the overflow characteristic.

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

$$\begin{aligned} &1 - |a| - b > 0 \\ &1 + b > 0. \end{aligned} \tag{2.4.1}$$

The analytic proof of this statement has been reported by Ebert e.a. [103]. 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 $q_{1,2} = \rho \cdot e^{\pm j\theta}$ we define an energy function $E(n)$, according to (2.3.10):

$$E(n) = x_1^2(n) - a \cdot x_1(n) \cdot x_2(n) - b \cdot x_2^2(n), \tag{2.4.2}$$

where $a = 2\rho \cdot \cos(\theta)$

$b = -\rho^2$.

This energy function $E(n)$ is characterized by the symmetrical matrix

$$P = \begin{bmatrix} 1 & -a/2 \\ -a/2 & -b \end{bmatrix}, \tag{2.4.3}$$

which is positive definite, due to

$$\mathrm{Det}[P] \;=\; -b - a^2/4 \;=\; \rho^2 \cdot \sin^2(\theta) \;>\; 0 \qquad (2.4.4)$$

and

$$\mathrm{Tr}[P] \;=\; 1 - b \;=\; 1 + \rho^2 \;>\; 0. \qquad (2.4.5)$$

The system dynamics is such that if no wordlength reduction is applied, the energy strictly decreases:

$$E_o(n) \;=\; \rho^2 \cdot E(n\text{-}1) \;<\; E(n\text{-}1). \qquad (2.4.6)$$

In the nonlinear system, saturation causes an additional energy reduction. If before correction we have $x_{o1}(n) > 1$, then after saturation $x_1(n) = 1$. Component $x_{o2}(n)$ can never overflow since $x_{o2}(n) = x_1(n\text{-}1)$. So

$$E(n) \;=\; 1 - a \cdot x_{o2}(n) - b \cdot x_{o2}^{\,2}(n) \qquad (2.4.7)$$

and

$$E(n) - E_o(n) \;=\; -\Big[x_{o1}(n)\text{-}1\Big] \cdot \Big[1 + x_{o1}(n) - a \cdot x_{o2}(n)\Big]$$

$$< \; 0 \qquad \text{for all } n \geq 0. \qquad (2.4.8)$$

The latter inequality is due to the stability requirement $|a| < 2$. The same conclusion can be drawn for values of $x_{o1}(n) < -1$. 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, namely

$$E(n) \;=\; (1\text{-}b) \cdot x_1^{\,2}(n) - 2 \cdot a \cdot x_1(n) \cdot x_2(n) + (1\text{-}b) \cdot x_2^{\,2}(n) \qquad (2.4.9)$$

which is characterized by the symmetrical matrix

$$P = \begin{bmatrix} 1-b & -a \\ -a & 1-b \end{bmatrix}, \tag{2.4.10}$$

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

$$\mathrm{Det}[P] = (1+a-b)\cdot(1-a-b) > 0 \tag{2.4.11}$$

and

$$\mathrm{Tr}[P] = 2\cdot(1-b) > 0. \tag{2.4.12}$$

Without overflow correction the energy cannot increase, since

$$E_o(n) - E(n-1) = -(1+b)\cdot[a\cdot x_1(n-1) + (b-1)\cdot x_2(n-1)]^2$$

$$\leq 0 \qquad \text{for all } n > 0. \tag{2.4.13}$$

In the nonlinear system saturation causes a reduction of the energy. If $x_{o1}(n) > 1$ then after saturation $x_1(n) = 1$ and with $x_2(n) = x_{o2}(n)$,

$$E(n) - E_o(n) = -\left[x_{o1}(n) - 1\right]\cdot\left[(1-b)\cdot(1+x_{o1}(n)) - 2a\cdot x_{o2}(n)\right]$$

$$< 0 \qquad \text{for all } n \geq 0. \tag{2.4.14}$$

The last inequality is a consequence of the stability condition $1-|a|-b > 0$. The same conclusion can be drawn for $x_{o1}(n) < -1$.
In a strict sense function $E(n)$ is not a Lyapunov function since the energy can remain constant. This situation can, however, 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 has now been proved for all pairs of filter coefficients a and b in the stability triangle.

The forced response of the second-order direct form digital filter is only guaranteed to be free from overflow oscillations if saturation is used for overflow correction and the filter coefficients a and b satisfy the condition

$$|a| + |b| < 1. \tag{2.4.15}$$

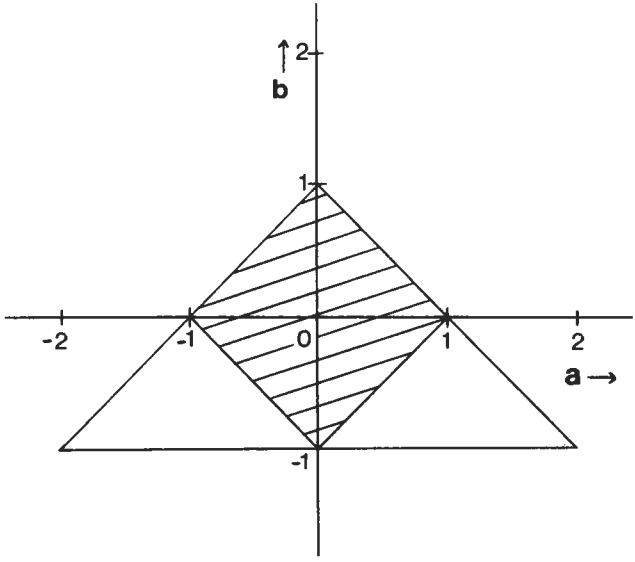


Fig. 2.4.1. Forced-response stability region of the second-order direct form digital filter.

The forced response stability of this filter has been proved with frequency domain criteria [86, 87, 88]. We prove this statement with Lyapunov methods. To this end we define an energy function

$$E(n) = \underline{e}^{T}(n) \cdot D \cdot \underline{e}(n), \tag{2.4.16}$$

where D denotes a positive diagonal matrix ($d_1 > 0$, $d_2 > 0$):

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}. \tag{2.4.17}$$

and $\underline{e}(n)$ is the error vector.

If no overflow correction is applied, resulting in

$$\underline{e}_o(n) = \underline{x}_o(n) - \tilde{\underline{x}}(n)$$

$$= A\cdot\underline{x}(n-1)+\underline{b}\cdot u(n) - [A\cdot\tilde{\underline{x}}(n-1)+\underline{b}\cdot u(n)] = A\cdot\underline{e}(n-1), \quad (2.4.18)$$

the energy strictly decreases if $D - A^T\cdot D\cdot A$ is positive definite. This condition is satisfied if

$$\mathrm{Det}[D - A^T\cdot D\cdot A] > 0 \quad (2.4.19)$$

and

$$\mathrm{Tr}[D - A^T\cdot D\cdot A] > 0. \quad (2.4.20)$$

So there must exist some pair of positive values d_1 and d_2 for which

$$\mathrm{Det}[\cdot] = -b^2\cdot d_1^{\,2} + (1-a^2+b^2)\cdot d_1 d_2 - d_2^{\,2} > 0 \quad (2.4.21)$$

and

$$\mathrm{Tr}[\cdot] = d_1\cdot[1-a^2-b^2] > 0. \quad (2.4.22)$$

The inequality (2.4.21) can have real solutions if the discriminant of the quadratic function Det[·] is positive:

$$\left[1 - a^2 + b^2\right]^2 - 4\cdot b^2 > 0 \quad (2.4.23)$$

or

$$|a| + |b| < 1. \quad (2.4.24)$$

It is easy to check that for $d_1 = 1$ and $d_2 = |b|$ the inequalities (2.4.21) and (2.4.22) are now satisfied.
In the nonlinear system, in which

$$\underline{e}(n) = F\{\underline{x}_o(n)\} - \tilde{\underline{x}}(n), \quad (2.4.25)$$

overflow correction with saturation lowers the energy. If $x_{o1}(n) > 1$ then

$$|e_1(n)| = |1 - \tilde{x}_1(n)|$$

$$< |x_{o1}(n) - \tilde{x}_1(n)| = |e_{o1}(n)|. \qquad (2.4.26)$$

Since matrix D has a diagonal form the energy $E(n)$ will decrease, too. The same conclusions can be drawn for $x_{o1}(n) < -1$. Therefore the energy $E(n)$ is a Lyapunov function, which guarantees forced-response stability; no forced-response overflow oscillations can occur in the second-order direct form filter with saturation and filter coefficients a and b satisfying (2.4.15).

For a second-order direct form filter with a pair of filter coefficients which does not satisfy equation (2.4.15) it is possible to construct an input signal and find some initial conditions for which overflow oscillations actually appear in the digital filter (see [86]). So for the second-order direct form digital filter with saturation inequality (2.4.15) is not only a sufficient condition for forced-response stability, it is also necessary for stability under *all* excitations $u(n) \in U_o$. No overflow correction method has been found which guarantees this filter to be free from forced-response overflow oscillations for all parameters a and b within the stability triangle.

There exists also a *second type* of direct form digital filters (see [278]). These filters have a matrix A which is the transposed form of the previous discussed system matrix. The second-order digital filter structure of this type is shown in Fig. 2.4.2 and described by the state equations

$$\begin{aligned} x_1(n+1) &= NL_1\{a\cdot[x_1(n) + u(n)] + x_2(n)\} \\ x_2(n+1) &= NL_2\{b\cdot[x_1(n) + u(n)]\} \\ y(n) &= x_1(n) + u(n). \end{aligned} \qquad (2.4.27)$$

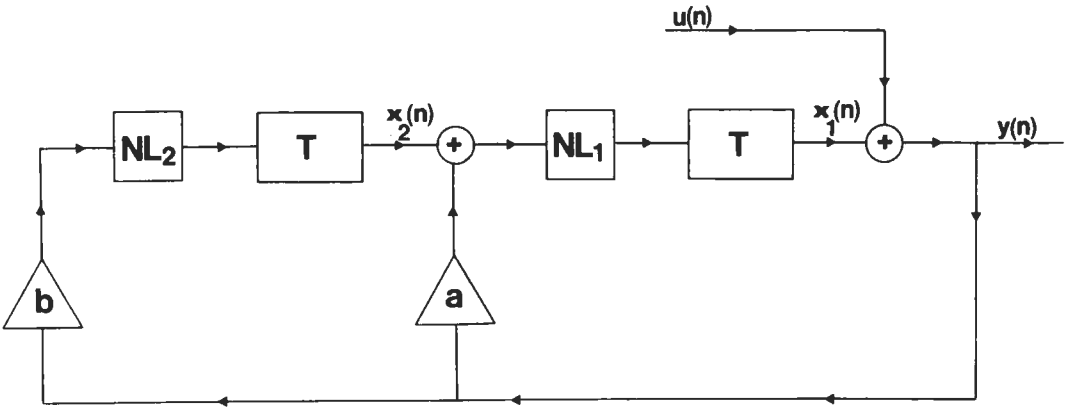


Fig 2.4.2. Second-order direct form digital filter of type II.

With $NL_1(\cdot)$ and $NL_2(\cdot)$ describing a saturation characteristic it is easy to prove (using the energy function of (2.4.9)) that also this second-order direct form filter is zero-input stable for all filter coefficients a and b in the stability triangle.

Using the energy function of (2.4.16) we can also prove that this filter is forced response stable for filter coefficients satisfying the condition

$$|a| + |b| < 1. \tag{2.4.28}$$

So the direct form digital filter of type II has the same overflow stability properties as the second-order direct form digital filter of type I.

Higher-order direct form digital filters are in general unstable with respect to overflow; high period and chaotic oscillations have been observed in such structures [231, 242, 243, 245, 247, 249, 304, 306]. For these direct form filters no overflow characteristic has been found for which the filter is zero-input stable for all allowed filter coefficients.

2.5. Overflow stability in wave digital and normal filters

Another class of filter structures is formed by the *wave digital filters*, which are structures derived from classical LC-twoports [117, 118, 127, 128]. In this section we investigate the zero-input and forced response stability of these filters and we especially focus on the second-order structures. There are two types of second-order wave digital filters as presented in Fig. 2.5.1 and Fig. 2.5.2. The second-order wave digital filter of type I is described by the state equations

$$\begin{aligned} x_1(n+1) &= NL_1\{(1-\gamma_1)\cdot x_1(n) + \gamma_1\cdot x_2(n) - \gamma_1\cdot u(n)\} \\ x_2(n+1) &= NL_2\{-\gamma_2\cdot x_1(n) + (-1+\gamma_2)\cdot x_2(n) - \gamma_2\cdot u(n)\}, \end{aligned} \qquad (2.5.1)$$

where $NL_1\{\cdot\}$ and $NL_2\{\cdot\}$ denote two overflow correction functions.

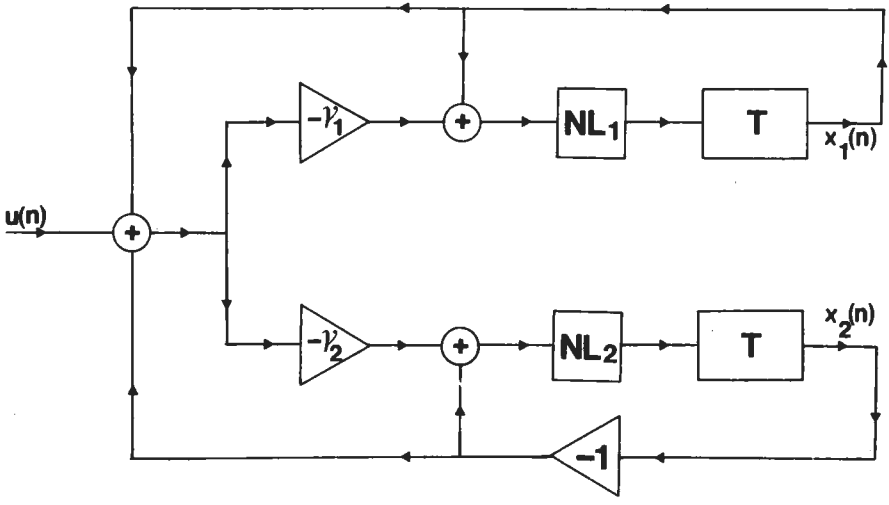


Fig. 2.5.1. Second-order wave digital filter of type I.

The second-order wave digital filter of type II has a system matrix A which is the transposed version of that of type I. This filter is described by the state equations

$$\begin{aligned} x_1(n+1) &= NL_1\{x_1(n) + NL_3\{-\gamma_1 \cdot x_1(n) - \gamma_2 \cdot x_2(n) + u(n)\}\} \\ x_2(n+1) &= NL_2\{-x_2(n) - NL_3\{-\gamma_1 \cdot x_1(n) - \gamma_2 \cdot x_2(n) + u(n)\}\}. \end{aligned} \quad (2.5.2)$$

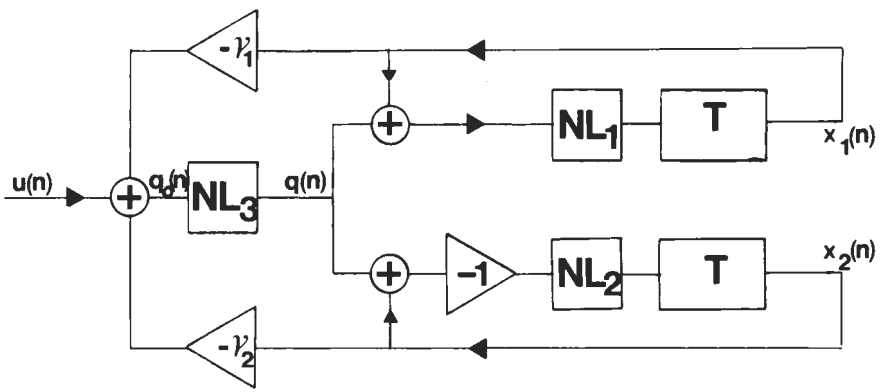


Fig. 2.5.2. Second-order wave digital filter of type II.

We again assume that quantization can be neglected and that overflow correction is only applied to the state signals, so $NL_1\{\cdot\} = NL_2\{\cdot\} = F\{\cdot\}$ stands for the overflow correction function and $NL_3\{\cdot\}$ is a through connection.

The eigenfrequencies (eigenvalues of the system matrix A) $q_{1,2}$ are equal to those of the direct form filter if the filter coefficients γ_1 and γ_2 are chosen as follows [343, 344]:

$$\begin{aligned} \gamma_1 &= (1-a-b)/2 \\ \gamma_2 &= (1+a-b)/2. \end{aligned} \quad (2.5.3)$$

The linear stability conditions (1.2.21) then translate into:

$$\gamma_1 > 0, \quad \gamma_2 > 0 \quad \text{and} \quad \gamma_1 + \gamma_2 < 2. \tag{2.5.4}$$

The second-order wave digital filters are zero-input stable for all overflow characteristics. This statement cannot be proved with the energy function of (2.3.10), but can be concluded with the aid of another function $E(n)$ [118, 122, 123, 124]. According to [118] we define for wave digital filter I:

$$E(n) = x_1^2(n)/\gamma_1 + x_2^2(n)/\gamma_2. \tag{2.5.5}$$

In the idealized system, the energy cannot increase:

$$\begin{aligned} E_o(n) - E(n-1) &= -(2-\gamma_1-\gamma_2)\cdot[x_1(n-1) - x_2(n-1)]^2 \\ &\leq 0 \qquad \text{for all } n > 0. \end{aligned} \tag{2.5.6}$$

In the nonlinear system a subsequent overflow correction implies $|x_1(n)| \leq |x_{o1}(n)|$ and $|x_2(n)| \leq |x_{o2}(n)|$, where the equality sign only applies if no overflow correction is required. So if overflow correction actually is needed the energy is reduced:

$$E(n) < E_o(n). \tag{2.5.7}$$

The energy function $E(n)$ is monotonically non-increasing and therefore a semi-Lyapunov function. The energy remains constant in the linear system for $x_1(n-1) = x_2(n-1)$. In this case

$$\begin{aligned} x_1(n) &= x_1(n-1) \\ x_2(n) &= -x_2(n-1), \end{aligned} \tag{2.5.8}$$

so that $E(n)$ can only remain constant during two successive time steps if

$$x_1(n) = x_2(n) = 0. \tag{2.5.9}$$

This implies that no overflow oscillation can occur: the second-order wave digital filter of type I is zero-input stable.

The same statement is valid for the second-order wave digital filter of type II with the energy function

$$E(n) = \gamma_1 \cdot x_1^2(n) + \gamma_2 \cdot x_2^2(n). \tag{2.5.10}$$

The pertinent proof is similar to that of wave digital filter I.

For nonzero-input signals the wave digital filters can become unstable. If, however, saturation is used for overflow correction forced-response overflow oscillations cannot appear. The proof of this statement is equivalent to that of the zero-input stability and has also been presented in [118, 122, 123, 124]. For wave digital filter I we define an energy function

$$E(n) = e_1^2(n)/\gamma_1 + e_2^2(n)/\gamma_2, \tag{2.5.11}$$

and for wave digital filter II

$$E(n) = \gamma_1 \cdot e_1^2(n) + \gamma_2 \cdot e_2^2(n), \tag{2.5.12}$$

where $\underline{e}(n)$ is the error vector (see(1.4.1)).

According to (2.4.26) overflow correction with a saturation characteristic yields $|e_1(n)| \leq |e_{o1}(n)|$ and $|e_2(n)| \leq |e_{o2}(n)|$ which proves the asserted stability in the actual digital filter.

A third class of digital filter structures comprises the *normal or coupled form digital filters* [30, 31, 268, 334]. The second-order digital filter structure of this class is shown in Fig. 2.5.3 and described by the state equations (2.5.13). This filter cannot have real eigenfrequencies but only complex conjugated values $q_{1,2} = \rho \cdot \exp(\pm j\theta)$.

$$\begin{aligned} x_1(n+1) &= NL_1\{\sigma \cdot x_1(n) - \omega \cdot x_2(n) + b_1 \cdot u(n)\} \\ x_2(n+1) &= NL_2\{\omega \cdot x_1(n) + \sigma \cdot x_2(n) + b_2 \cdot u(n)\} \\ y(n) &= c_1 \cdot x_1(n) + c_2 \cdot x_2(n) + d \cdot u(n) \end{aligned} \qquad (2.5.13)$$

where

$$\sigma = \rho \cdot \cos(\theta) \quad \text{and} \quad \omega = \rho \cdot \sin(\theta). \qquad (2.5.14)$$

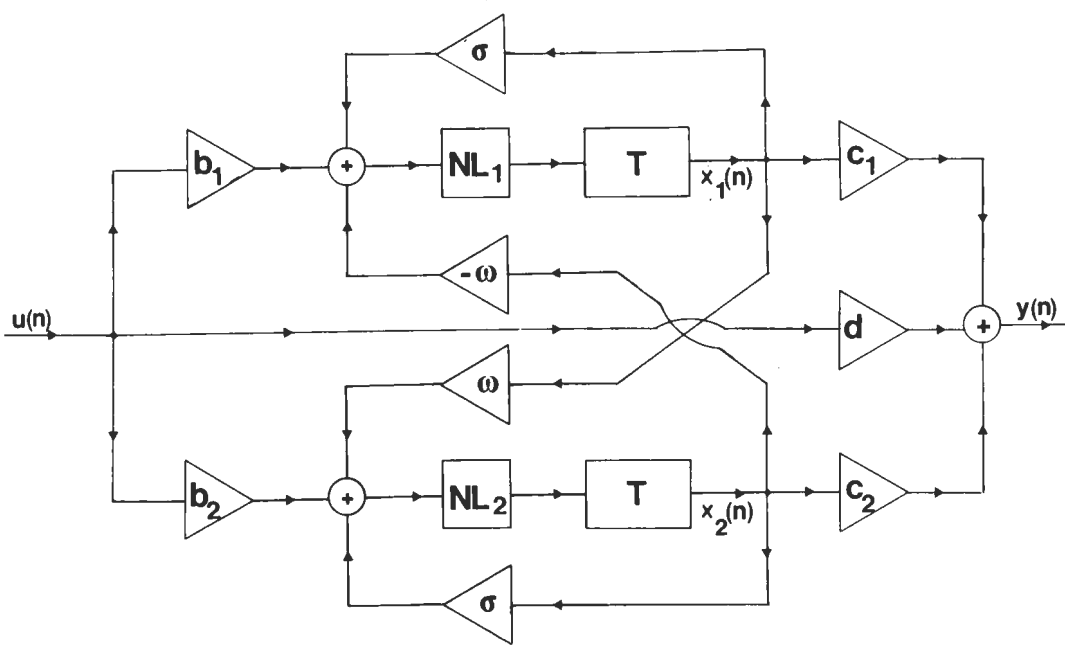


Fig. 2.5.3. Second-order normal form digital filter.

With the energy function $E(n)$ defined according to equation (2.3.10)

$$E(n) = \underline{x}^T(n) \cdot \underline{x}(n) = x_1^2(n) + x_2^2(n), \tag{2.5.15}$$

it is easy to prove that, like the wave digital filters, the normal filter is zero-input stable for all overflow characteristics and forced-response stable for a saturation characteristic.

Normal filters and wave digital filters of orders higher than two can likewise be stabilized for zero-input with all types of overflow characteristics and for forced-input with a saturation overflow characteristic [30, 31, 38, 186, 236, 334].

2.6. Overflow stability in state-space digital filters

The *state-space digital filters* are characterized by the fact that all coefficients of the state equations are realized directly in the filter structure [238]. The *second-order filter* is shown in Fig. 2.6.1 and decribed by the state equations

$$\begin{aligned} x_1(n+1) &= NL_1\{a_{11}\cdot x_1(n) + a_{12}\cdot x_2(n) + b_1\cdot u(n)\} \\ x_2(n+1) &= NL_2\{a_{21}\cdot x_1(n) + a_{22}\cdot x_2(n) + b_2\cdot u(n)\} \\ y(n) &= c_1\cdot x_1(n) + c_2\cdot x_2(n) + d\cdot u(n). \end{aligned} \qquad (2.6.1)$$

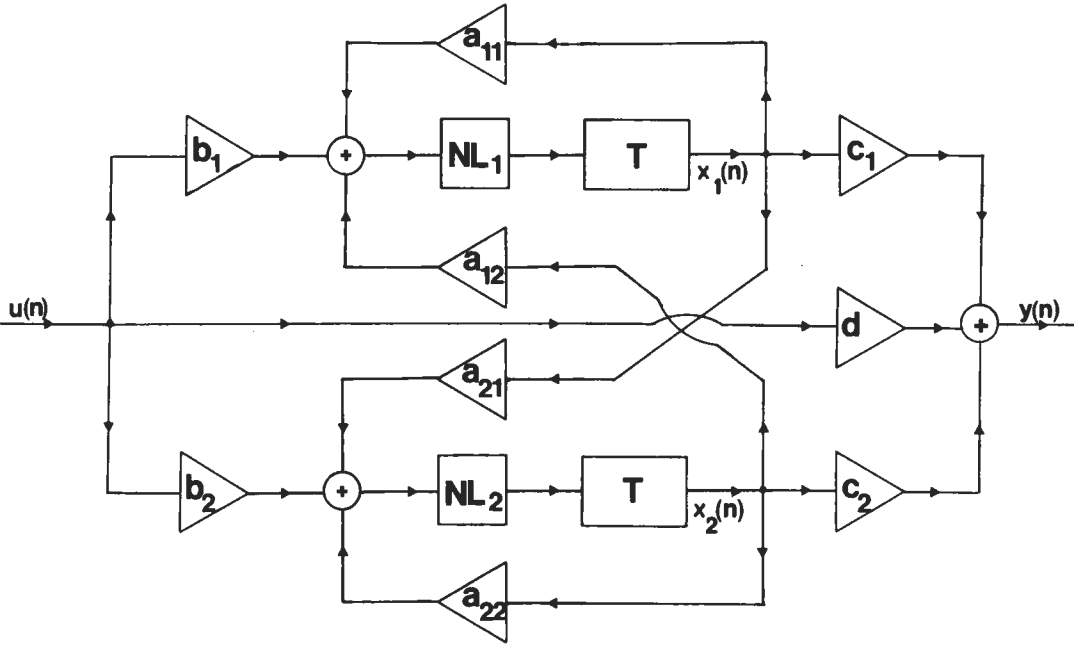


Fig. 2.6.1. Second-order state space digital filter.

It is a common property of normal, wave digital filters and lattice filters [136, 269, 284, 323] 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 overflow oscillations [252, 255, 308, 310, 313, 314, 315, 316, 339, 340].

The question arises: Which A matrices admit a diagonal "energy matrix" so that (1.4.3) passes into

$$E(n) = \underline{x}^T(n) \cdot D \cdot \underline{x}(n), \tag{2.6.2}$$

where D is a positive diagonal matrix? This question has been solved in [237, 238] for the second-order filter, where D is of the form

$$D = \begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}. \tag{2.6.3}$$

In appendix I we show that this is only possible if the system matrix A fulfils the condition:

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A]. \tag{2.6.4}$$

We conclude that a second-order state-space filter satisfying equation (2.6.4) is zero-input stable for all overflow characteristics.

Do the three basic filter types satisfy (2.6.4)? The answer is "yes" for the normal form, "yes" if $|a| - b < 1$ for the direct form and "almost yes" for the wave digital filters with the inequality sign replaced by an equality sign (reflecting the semi-Lyapunov character of the energy function).

The same diagonal matrix D can be used in the Lyapunov function

$$E(n) = \underline{e}^T(n) \cdot D \cdot \underline{e}(n), \tag{2.6.5}$$

(where $\underline{e}(n)$ denotes the error vector), to prove that a filter is *forced-response stable* under saturation. The proof is essentially based upon (2.4.26) stating that

$$|e_1(n)| \leq |e_{o1}(n)| \quad \text{and} \quad |e_2(n)| \leq |e_{o2}(n)|. \tag{2.6.6}$$

Therefore, (2.6.4) is also a sufficient condition for freedom from overflow oscillations in the forced response of a second-order state-space filter with a saturation characteristic [88].

If a *saturation characteristic* is used in a *second-order state-space digital filter* the condition (2.6.4) for zero-input stablility (valid for all overflow characteristics) can be relaxed. We shall prove in Appendix II the following theorem:

If

$$E(n) = \underline{x}^T(n) \cdot P \cdot \underline{x}(n), \tag{2.6.7}$$

satisfies all conditions for an energy function (see Section 1.4) and if

$$|p_{12}| \leq p_{11} \quad \text{and} \quad |p_{12}| \leq p_{22}, \tag{2.6.8}$$

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

The question arises: Does a given matrix A possess at least one energy function $E(n)$ with an energy matrix P that satisfies (2.6.8)?

The energy strictly decreases in the filter without overflow correction if $P - A^T \cdot P \cdot A$ is positive definite, or

$$\mathrm{Det}[P - A^T \cdot P \cdot A] > 0 \tag{2.6.9}$$

and

$$\mathrm{Tr}[P - A^T \cdot P \cdot A] > 0. \tag{2.6.10}$$

In Appendix II we show that it is only possible to fulfil these inequalities, for some matrix P, which fulfils (2.6.8), if the system matrix A satisfies the condition:

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A] + 2 \cdot \min(|a_{12}|, |a_{21}|). \tag{2.6.11}$$

These results will also be published in [296, 352].
As it should be, this condition is weaker than (2.6.4). Contrary to the former condition, all stable direct form filters satisfy (2.6.11).

Observe that every stability requirement yields sufficient conditions and not a necessarry one. These sufficient conditions can often be weakened with various analytic measures [145, 175] or with computer generated Lyapunov functions [109, 110]. Attempts have also been reported with unconventional overflow characteristics [104, 364] and overflow signalling schemes [26, 43, 53], while experimental results have been reported in [199].

3. Quantization stability in recursive digital filters

3.1. Introduction

Besides the large-amplitude overflow oscillations treated in the previous chapter, still other parasitic oscillations are observed in recursive digital structures, which have their origin in the quantization fine structure and as a result have relatively small amplitudes. These oscillations can occur under zero- and under nonzero periodic-input conditions and are generally called "*limit cycles*" and "*subharmonics*" respectively. Together with quantization noise limit cycles and subharmonics are considered as the most serious deviation from linear behaviour under normal operating conditions of a digital filter. In contrast with quantization noise, they can, however, be completely avoided. Unfortunately the involved techniques complicate the noise analysis such that a systematic noise minimization cannot be achieved with analytic tools. Thus in current literature we observe almost independent studies of limit cycle suppression and of noise optimization. The first problem mostly deals with magnitude truncation MT (or related methods) while the second is completely based on rounding RO [38, 178, 179, 336].

The main factor determining the occurrence of limit cycles is the quantization characteristic. In this chapter we mainly consider RO and MT quantization; modified quantizations like controlled rounding CR and stochastic quantization require additional signals and, as such, more complicated descriptions than a simple characteristic.

The analytical treatment of limit cycles resembles that of overflow oscillations. This implies an organization of the present chapter similar to that of the previous chapter.

Whereas in the overflow problem $\underline{x}(n)$ denotes a continuous set of variables, quantization in a fixed-point digital filter implies a discrete-amplitude character of $\underline{x}(n)$ with all components x_i being integer multiples of the *quantum* q. Without any restriction we may normalize the quantization step by choosing q equal to unity, which means that the components of the state $\underline{x}(n)$ have integer values only. Reckoning with the fact that all signals are bounded, any filter with quantized state variables can consequently be viewed as a finite-state machine.

While for any arbitrary initial condition $\underline{x}(0) \neq \underline{0}$ the state $\underline{x}(n)$ in an autonomous (excitation-free) linear filter asymptotically approaches the origin ($\underline{x}(n) \to \underline{0}$ for $n \to \infty$), this is not the rule for the nonlinear filter. Instead, at some $n = N_0$ the state $\underline{x}(n)$ might enter a limit cycle. This is a periodic motion characterized by S state points which are cyclically occupied by $\underline{x}(n)$. "*Accessible*" limit cycles can be entered from points outside the cycle which together with all their predecessors form a (mostly immense) set of state points to be assigned to such a cycle (see [82, 84, 263, 266]). On the other hand, "*inaccessible*" cycles have to be started on the limit cycle itself. Limit cycles of period $S = 1$ consist of one point, which can be accessible or inaccessible. If and only if the origin $\underline{x} = \underline{0}$ is ultimately reached from any initial condition $\underline{x}(0)$ (implying accessibility of the origin) the filter is limit cycle free.

Without any quantization (corresponding to the ideal, linear filter), the trajectory of $\underline{x}(n)$ in a second-order filter would follow an ellipse-like curve spiralling towards the origin, as shown in Fig. 2.3.1. In the actual filter, quantization introduces a slight modification of the state, such that the quantized $\underline{x}(n)$ becomes a point in the quantization grid, located in the close vicinity of the state before quantization. Like the overflow correction discussed in Chapter 2, quantization can be associated with a state motion towards the origin or away from it. The first motion supports the linear motion and ensures freedom of limit cycles, if quantization is always performed this way. Clearly, this rule provides a sufficient condition. Conversely, quantization correction away from the origin does not admit any conclusion: limit cycles can, but need not occur.

The above statements ask for a definition of "distance from the origin". Equivalent with the definition of Chapter I we identify any "*Lyapunov energy function*" with the squared distance from the origin. Such a function $E(n)$ is defined by

$$E(n) = \underline{x}^T(n) \cdot P \cdot \underline{x}(n), \tag{3.1.1}$$

where P is a symmetrical and positive definite matrix. The system dynamics must be so that if no quantization is applied, according to $\underline{x}_o(n) = A \cdot \underline{x}(n-1)$, the energy strictly decreases with increasing time n. Furthermore the quantization operation $\underline{x}(n) = Q\{\underline{x}_o(n)\}$ must not lead to an energy increase:

$$[Q\{\underline{x}_o(n)\}]^T \cdot P \cdot Q\{\underline{x}_o(n)\} \leq \underline{x}_o^T(n) \cdot P \cdot \underline{x}_o(n) \quad \text{for all } \underline{x}_o(n). \tag{3.1.2}$$

Most limit-cycle suppressing mechanisms, which will be treated in the next sections of this chapter, make use of magnitude truncation (including the related controlled rounding described in Section 3.3), utilizing its ability to reduce energy as required by (3.1.2). On the other hand, *rounding* RO can amplify the signal magnitude by a factor $c \leq 2$, where the maximum factor $c = 2$ occurs for a signal magnitude equal to half a quantization step. In an attempt to achieve freedom from limit cycles also for RO quantization, the potential nonlinear energy increase has to be compensated by an equal energy decrease associated with the linear filter operation. In concrete terms, the necessary damping finds expression in the condition $\|A\| < 1/2$, where $\|A\|$ denotes the norm of the system matrix. If this condition is not met by the design requirements, the matrix A can be transformed into some power A^L by means of a "block-state realization" or a "matrix-power feedback" such that $\|A^L\| < 1/2$ and RO quantization can be applied without risk of limit cycles [32, 33, 115, 116]. For high-Q filters this method, however, requires a great amount of hardware.

Another way to eliminate the various types of limit cycles is to control quantization through an external *random signal*.

This way potential conditions favourable to the occurrence of parasitic oscillations are irregularly disturbed, which results in an asymptotic, albeit noisy approach of the zero state.

The disadvantages of such stochastic methods are evident: they require additional random sources (preferably independent sources for all quantizers) and, at a first glance, yield additional quantization noise. The latter point is, however, compensated by the flatness of the noise spectrum that contrasts with the (mostly) narrow bandwidth of the noise generated by MT quantization. Particularly in high-Q filters the ultimate noise contributions at the output terminals can thus be considerably smaller than those occurring with deterministic stabilization [59, 89, 90, 91, 138].

The simplest method is *random rounding*, where the decision concerning the handling of the least significant bit is exclusively left to the exterior random signal [70, 71].
A variation on this strategy is found when the unquantized signal is supplemented with *random dither*, whose spectral distribution is flat and whose amplitude distribution is uniform in the interval $[-q/2, q/2]$. The total signal is subsequently subject to RO quantization [52, 294, 318]. In contrast with this "*uniform random dither*", the above random rounding is occasionally referred to as "*binary random dither*". In rough terms, uniform random dither has a better noise performance, whereas binary random dither is superior with respect to limit cycle suppression.
In a variant, uniform random dither is subject to spectral shaping, particularly with a bandstop characteristic. The resulting "*bandstop dither*" has an improved noise performance, to be sure, but is costly to implement [14, 223].

Guided by the inherent properties of RO and MT quantization, one can combine their respective merits into "*random quantization*", in which an external generator randomly switches between RO and MT quantization, with comparatively short MT operating times. This way the excellent RO noise properties are coupled with the stabilizing capability of MT quantization, which has to be paid for with a prolonged limit cycle expiration time [182, 183, 184, 185, 202, 204].

3.2. Quantization stability in first-order digital filters

The first-order recursive digital filter with a general wordlength reduction $NL\{\cdot\}$, as shown in Fig. 2.2.1, was described by the state equations (2.2.1)

$$\begin{aligned} x(n+1) &= NL\{a\cdot x(n) + u(n)\} \\ y(n) &= x(n+1). \end{aligned} \tag{3.2.1}$$

In our present context $NL\{x\} = Q\{x\}$ denotes the quantization characteristic (with the quantization step q normalized to unity). The idealized linear filter is stable for filter coefficients a satisfying the condition: $|a| < 1$.

The stability of this filter is first investigated for the zero-input situation, with $Q\{\cdot\}$ representing the rounding, magnitude truncation and value truncation characteristic.

For *rounding* the quantization error $\epsilon(n)$, defined according to (1.3.1) by $\epsilon(n) = Q\{a\cdot x(n)\} - a\cdot x(n)$, is bounded by $|\epsilon_{RO}(n)| \leq 1/2$.
So

$$\begin{aligned} |x(n+1)| &= |Q\{a\cdot x(n)\}| \\ &= |a\cdot x(n) + \epsilon(n)| \\ &< |x(n)| + 1/2. \end{aligned} \tag{3.2.2}$$

Since $x(n)$ and $x(n+1)$ have integer values only, we conclude that

$$|x(n+1)| \leq |x(n)|. \tag{3.2.3}$$

If at some time n the state $x(n) \neq 0$ reaches a previous state $x(n-S)$, a zero-input limit cycle is entered. Such a limit cycle must have a constant amplitude value, since a decrease of the state signal can never be restored, according to (3.2.3).

There exist two types of constant amplitude limit cycles, those with $x(n+1) = x(n)$ and those with $x(n+1) = -x(n)$. The first type is periodic with period 1, with a constant value $x(n) = X$. The amplitude is bounded by

$$|X| \leq \frac{1}{2(1-a)}, \tag{3.2.4}$$

as can be concluded from (3.2.2).
Since the smallest limit cycle has an amplitude $|X| = 1$, these limit cycles can only occur for filter coefficients $a \geq 1/2$. Moreover, for $a \geq 1/2$ the zero-state is inaccessible with the consequence that limit cycles do occur for all initial conditions.
The second type of limit cycles is of the form $x(n) = (-1)^n \cdot X'$, which is periodic with period 2. These limit cycles are amplitude-bounded by

$$|X'| \leq \frac{1}{2(1+a)}. \tag{3.2.5}$$

They occur if and only if $a \leq -1/2$.

For the first-order digital filter with *magnitude truncation* we have

$$\begin{aligned} |x(n+1)| &= |Q\{a \cdot x(n)\}| \\ &\leq |a \cdot x(n)| \\ &< |x(n)| \qquad \text{for } x(n) \neq 0. \end{aligned} \tag{3.2.6}$$

Since $|x(n)|$ is monotonically decreasing and the state signal can assume only integer values the state signal $x(n)$ must become zero after a finite number of time steps. Therefore no zero-input limit cycle can occur in the first-order recursive digital filter with magnitude truncation.

If *value truncation* is used for quantization we have

$$|x(n+1)| = |a \cdot x(n) + \epsilon(n)| < |x(n)| + 1. \qquad (3.2.7)$$

Since $x(n)$ has integer values, we conclude that

$$|x(n+1)| \leq |x(n)| \qquad (3.2.8)$$

and, just as for rounding, only limit cycles with constant amplitude can occur. The limit cycles with period 1, $x(n) = X$, have an amplitude bounded by

$$\frac{-1}{1 - a} < X \leq 0, \qquad (3.2.9)$$

and thus can appear for $a > 0$.
The limit cycles with period 2, $x(n) = (-1)^n \cdot X'$, cannot occur if value truncation is used because they have to fulfil the set of equations

$$\begin{array}{lll} X' = -a \cdot X' + \epsilon(1) & \text{with} & -1 < \epsilon(1) \leq 0 \\ -X' = a \cdot X' + \epsilon(2) & \text{with} & -1 < \epsilon(2) \leq 0 \\ \hline 0 = \epsilon(1) + \epsilon(2). & & \end{array} + \qquad (3.2.10)$$

So only limit cycles with period 1 are possible in the first-order recursive digital filter with value truncation, and they can occur only for $a > 0$. For filters with a coefficient $a \leq 0$ there are no limit cycles in this filter; the zero-state will be reached from any initial state $x(0)$.

The investigation of the stability of this filter for the zero-input situation will now be generalized to all quantization characteristics with a quantization error $\epsilon(n)$, which is bounded by 1, (the normalized quantization step q). Then

$$|x(n+1)| = |a \cdot x(n) + \epsilon(n)| < |x(n)| + 1, \qquad (3.2.11)$$

so that limit cycles (if they occur) have a constant amplitude:

$$x(n) = X \qquad (3.2.12)$$

or

$$x(n) = (-1)^n \cdot X'. \qquad (3.2.13)$$

The limit cycles of the first type are bounded by $|X| < \frac{1}{1 - a}$ and can only occur for $a > 0$. The limit cycles of the second type are bounded by $|X'| < \frac{1}{1 + a}$ and can only occur for $a < 0$.

Next we analyse the forced response of a first-order recursive digital filter, excited with a *periodic input signal* with period N. In the idealized filter the steady state $\tilde{x}(n)$ is periodic with the same period N as the input signal; but in the filter with quantization the period of the state can differ from N, which amounts to the generation of subharmonics.

We start with the analysis of the first-order digital filter with *rounding*. Define the difference signal

$$d(n) = x(n) - x(n+N). \qquad (3.2.14)$$

If we can prove that $d(n)$ becomes zero after some finite time no subharmonic occurs in this filter. With

$$x(n+1) = a \cdot x(n) + u(n) + \epsilon(n) \qquad (3.2.15)$$

we have

$$d(n+1) = a \cdot d(n) + \epsilon(n) - \epsilon(n+N). \tag{3.2.16}$$

So

$$|d(n+1)| < |d(n)| + 1. \tag{3.2.17}$$

Like $x(n)$ also $d(n)$ has integer values resulting in

$$|d(n+1)| \leq |d(n)|. \tag{3.2.18}$$

If there occurs a periodic state sequence $x(n)$ in this filter, then also $d(n)$ becomes periodic and it must satisfy the condition:

$$|d(n)| = \text{constant} \qquad \text{for all n.} \tag{3.2.19}$$

From (3.2.16) we conclude that for $a > 0$ we have

$$d(n) = D \qquad \text{for all n,} \tag{3.2.20}$$

and for $a < 0$

$$d(n) = (-1)^n \cdot D' \qquad \text{for all n.} \tag{3.2.21}$$

In a finite state machine like a digital filter the state $x(n)$ will ultimately enter a periodic state sequence with

$$x(n + S \cdot N) = x(n) \qquad \text{for sufficiently high } n \tag{3.2.22}$$

and for some integer value of S. Since in such a periodic state sequence

$$\sum_{i=0}^{S-1} d(n+i \cdot N) = \sum_{i=0}^{S-1} [x(n+i \cdot N) - x(n+(i+1) \cdot N)] = 0, \tag{3.2.23}$$

substitution of (3.2.20) in (3.2.23) yields $D = 0$, while substitution of (3.2.21) in (3.2.23) yields $D' = 0$ for N is even. So subharmonics can only appear in this filter for $a < 0$ and N is odd. Such a subharmonic satisfies

$$x(n) - x(n+N) = d(n) = -d(n+N) = -x(n+N) + x(n+2N), \qquad (3.2.24)$$

resulting in $S = 2$, and, according to (3.2.16), is amplitude bounded by

$$|d(n)| < \frac{1}{1+a}. \qquad (3.2.25)$$

Next we analyse the situation for *magnitude truncation*. With

$$\begin{aligned} d(n+1) &= x(n+1) - x(n+N+1) \\ &= a \cdot d(n) + \epsilon(n) - \epsilon(n+N), \end{aligned} \qquad (3.2.26)$$

it is possible to have $|d(n+1) - a \cdot d(n)| \geq 1$. But, it is not possible to have $\epsilon(n) > 0$ and $\epsilon(n+N) < 0$ for $d(n+1) > 0$ or $\epsilon(n) < 0$ and $\epsilon(n+N) > 0$ for $d(n+1) < 0$, because with magnitude truncation $\epsilon(n) \cdot x(n+1) \leq 0$ for all n. This implies that

$$|d(n+1)| < |d(n)| + 1. \qquad (3.2.27)$$

Inequality (3.2.27) is identical with (3.2.17) so that the same conclusions apply as for rounding.

In the case of *value truncation* we have

$$|d(n+1)| = |a \cdot d(n) + \epsilon(n) - \epsilon(n+N)| < |d(n)| + 1, \qquad (3.2.28)$$

so again the same conclusions hold as for rounding.

If the quantization method is not one of the above three operations RO, MT or VT it is possible to have subharmonics of order S larger than 2. Furthermore, subharmonics can occur for periodic input signals with even N. This is shown in the following examples.

Example 1

The quantization method used in this example is "*anti-rounding*", as defined by:

$$\begin{aligned} Q(x) &= x && \text{for } x \text{ integer} \\ &= x + \epsilon \quad \text{with } 1/2 \le |\epsilon| < 1 && \text{for } x \text{ non-integer.} \end{aligned} \qquad (3.2.29)$$

The pertinent quantization characteristic is shown in Fig. 3.2.1. For a filter with a filter coefficient $a = -.72$ and excited by a constant input signal $u(n) = 14$, we can get a subharmonic of order $S = 3$ with the successive states $x(0) = 7$, $x(1) = 8$ and $x(2) = 9$. For $u(n) = 10$ we can get a subharmonic of order $S = 4$ with $x(0) = 4$, $x(1) = 8$, $x(2) = 5$ and $x(3) = 7$.

□

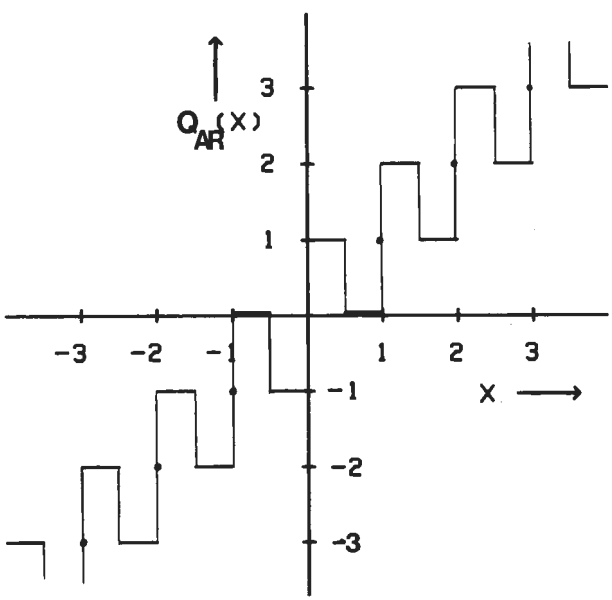


Fig. 3.2.1. Quantization characteristic for "anti-rounding".

Example 2

In this example we choose the same quantization method as used in Example 1 but with a filter coefficient $a = -.4$. For the periodic input signal of period 2 with $u(2n) = 2$ and $u(2n+1) = 4$ we can get a subharmonic of order $S = 2$ with the successive states $x(0) = 0$, $x(1) = 4$, $x(2) = 1$ and $x(3) = 3$.

□

For a quantization function $Q\{\cdot\}$ which satisfies the quantized Lipschitz condition

$$|\alpha - \beta| < 1 \quad \rightarrow \quad |Q\{\alpha\} - Q\{\beta\}| \leq 1, \tag{3.2.30}$$

it is easy to prove that, with $d(n) = x(n) - x(n+N)$,

$$|d(n+1)| \leq |d(n)| \quad \text{for all } n. \tag{3.2.31}$$

So we can conclude that, just as in the filter with rounding, only subharmonics of order $S = 2$ are possible, which can occur for odd values of N and negative values of a. Rounding, magnitude truncation and value truncation belong to the category under consideration.

3.3. Quantization stability in direct form digital filters

In this section we investigate the quantization stability of *direct form digital filters*. The major part of this section concerns the second-order filter of Fig. 1.2.1, which is described according to (1.2.16) by

$$\begin{aligned} x_1(n+1) &= NL_1\{NL_2\{a\cdot x_1(n)\} + NL_2\{b\cdot x_2(n)\} + u(n)\} \\ x_2(n+1) &= x_1(n) \\ y(n) &= x_1(n+1). \end{aligned} \tag{3.3.1}$$

The nonlinear functions $NL_1\{\cdot\}$ and $NL_2\{\cdot\}$ now represent quantization operations. Function $NL_1\{\cdot\}$ denotes the quantization of the signal before entering a time-delay element and $NL_2\{\cdot\}$ denotes this operation after a multiplication with a constant factor. It is unnecessary to perform both quantizations. If the results of the multipliers are added to the input signal with full precision the result is the so-called 1-*quantizer direct form filter*. On the other hand if $NL_2\{\cdot\}$ is performed and input signal $u(n)$ is a quantized signal then $NL_1\{\cdot\}$ is superfluous, resulting in the so-called 2-*quantizer direct form filter*.

The three well-known quantization methods: RO, MT and VT are unable to suppress zero-input limit cycles for all pairs of permitted filter coefficients a and b in the stability triangle. However, for each of these quantization methods and for both the 2-quantizer and the 1-quantizer configuration, regions in the stability triangle have been derived for which limit cycles cannot occur. The pertinent proofs make use of the frequency domain methods [77, 81, 84, 85, 309, 319], derived in more general contexts [29, 130, 133, 134, 135, 330, 331], and of Lyapunov theory [109, 110, 171, 173, 208, 210, 211, 212].

The best results (in the sense of the largest region in the plane of the filter coefficients) are found for the 1-quantizer direct form digital filter with magnitude truncation for quantization [81, 85].

A new type of quantization has been presented in [58], in which the direction of quantization is determined by a control signal.

With this "*controlled rounding*" the second-order direct form filter can be made free from zero-input limit cycles with the exception of those with period 1 or 2. The proof of this statement is given with a semi-Lyapunov function. According to [58] we define an energy function

$$E(n) = \underline{x}^T(n) \cdot P \cdot \underline{x}(n), \tag{3.3.2}$$

where

$$P = \begin{bmatrix} 1-b & -a \\ -a & 1-b \end{bmatrix}. \tag{3.3.3}$$

Matrix P is positive definite for all filter coefficients a and b in the stability triangle, because

$$\mathrm{Det}[P] = (1-a-b) \cdot (1+a-b) > 0 \tag{3.3.4}$$

and

$$\mathrm{Tr}[P] = 2 \cdot (1-b) > 0. \tag{3.3.5}$$

Without wordlength reduction the energy cannot increase, since

$$E_o(n+1) - E(n) = \underline{x}^T(n) \cdot [A^T \cdot P \cdot A - P] \cdot \underline{x}(n)$$

$$= -(1+b) \cdot [a \cdot x_1(n) + (b-1) \cdot x_2(n)]^2 \leq 0. \tag{3.3.6}$$

In the system with quantization,

$$\begin{aligned} x_1(n+1) &= a \cdot x_1(n) + b \cdot x_2(n) + \epsilon(n) \\ x_2(n+1) &= x_1(n), \end{aligned} \tag{3.3.7}$$

we have

$$\Delta E(n) = E(n+1) - E(n)$$

$$= -(1+b) \cdot [x_1(n+1) - x_2(n)]^2 + 2\epsilon(n) \cdot [x_1(n+1) - x_2(n)]. \tag{3.3.8}$$

The energy function $E(n)$ is indeed a semi-Lyapunov function of the non-linear system if

$$\epsilon(n)\cdot[x_1(n+1) - x_2(n)] \leq 0. \qquad (3.3.9)$$

So, if $x_2(n) > x_1(n+1)$ the quantization has to be performed upwards and if $x_2(n) < x_1(n+1)$ then the direction of quantization is downwards. This means that with controlled rounding quantization is performed "in the direction" of the "control signal" $x_2(n)$.

Zero-input limit cycles can still occur if the energy $E(n)$ remains constant, which implies

$$x_1(n+1) = x_2(n) = x_1(n-1). \qquad (3.3.10)$$

So only limit cycles with periods 1 or 2 can occur [58].

Stabilization by controlled rounding can also be *visualized* in the state plane. This alternative approach of the stability problem is depicted in Fig. 3.3.1. With (3.3.3) the curves $E(n)$ = constant are ellipses with axes under $\pm 45^{\circ}$. First recall that $x_2(n+1) = x_1(n)$ so that some grid point M in the state plane is always linearly mapped on a horizontal straight line through the mirror point M^* with respect to the 45° -line, (cf. Fig. 3.3.1). Let M′ denote the result of this linear mapping of M, and let M′′ denote the result of the subsequent quantization. Then M′ lies on the line segment $x_2(n+1) = x_1(n)$ inside or on the Lyapunov ellipse due to (3.3.6), while M′′ is desired to lie there too.

Realizing that, according to their definitions, M′′ and M^* are grid points, we conclude that M^* is a suitable candidate for M′′, but also any (if existent) intermediate grid point between M′ and M^*. To achieve minimum error, we choose for M′′ the grid point nearest to M′, in the direction of M^*. This construction yields the quantization rule that $x_1(n+1)$ is quantized "in the direction" of $x_2(n) = x_1(n-1)$.

Unfortunately, this "controlled rounding" CR does admit constant energy limit cycles of periods 2 or 1, in which the state jumps between M and M^* or, for $M = M^*$, remains at M.

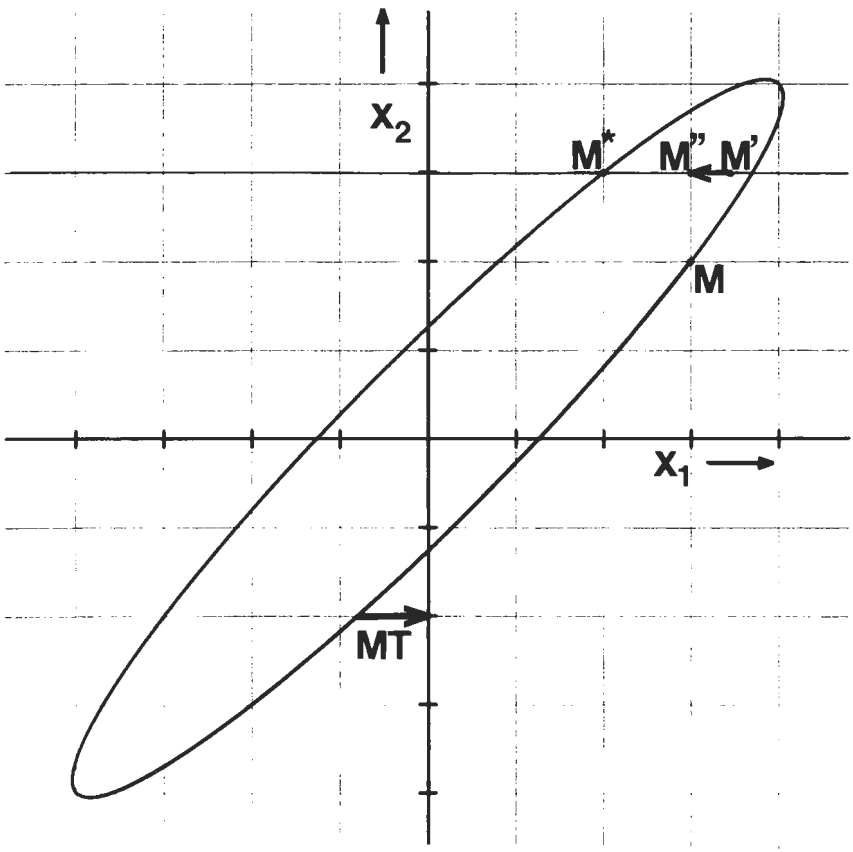


Fig. 3.3.1. State-plane of the direct form digital filter with controlled rounding.

It is an advantage of CR that not only zero-input stability (with the exceptions mentioned) is achieved, but also stability under any constant input condition, because the quantization rule only involves signal differences (see [58]).

With

$$E(n) = \underline{e}^{T}(n) \cdot P \cdot \underline{e}(n) \qquad (3.3.11)$$

and

$$\underline{e}(n) = \underline{x}(n) - \tilde{\underline{x}}(n), \qquad (3.3.12)$$

which is the error vector, the difference between the actual state $\underline{x}(n)$ and the asymptotic steady state $\tilde{\underline{x}}(n)$ of the idealized linear filter, we have

$$\Delta E(n) = E(n+1) - E(n)$$

$$= -(1+b)\cdot[a\cdot e_1(n) + (b-1)\cdot e_2(n) + \epsilon(n)]^2 + 2\epsilon(n)\cdot[a\cdot e_1(n) + (b-1).e_2(n) + \epsilon(n)]. \qquad (3.3.13)$$

According to (3.3.13) the energy $E(n)$ is guaranteed to be non-increasing for

$$\epsilon(n)\cdot[e_1(n+1) - e_2(n)] \leq 0. \qquad (3.3.14)$$

So the component $x_1(n+1)$ must be quantized in the direction of $[x_2(n) + \tilde{x}_1(n+1) - \tilde{x}_2(n)]$. For input signals which are periodic with period 1 or 2 we have

$$\tilde{x}_1(n+1) = \tilde{x}_2(n) = \tilde{x}_1(n-1). \qquad (3.3.15)$$

So controlled rounding in the original sense guarantees freedom from quantization oscillations with the exception of the periods 1 and 2.

With some additional hardware, zero-input limit cycles of periods 1 and 2 can be supressed, too, but at the expense of the general constant- and period 2 -input stability [248, 251].
For input signals which are periodic with a period $N \geq 3$ the controlled rounding mechanism does not lead to generalized stability: subharmonics have been observed in our simulations.

Recently, a new idea has been published to suppress zero-input limit cycles in the second-order direct form filter. This idea is based on a change of the filter structure so that the values of the multipliers are less than unity:

$$
\begin{aligned}
x_1(n+1) &= Q\{a\cdot x_1(n)\} + Q\{b\cdot x_2(n)\} && \text{for } |a| \le 1\\
&= Q\{(a+b)\cdot x_1(n)\} + Q\{b\cdot(x_2(n)-x_1(n))\} && \text{for } a > 1\\
&= Q\{(a-b)\cdot x_1(n)\} + Q\{b\cdot(x_2(n)+x_1(n))\} && \text{for } a < -1\\
x_2(n+1) &= x_1(n)\\
y(n) &= x_1(n+1). && (3.3.16)
\end{aligned}
$$

All quantizers in this filter are magnitude truncators. In computer simulations no zero-input limit cycles have been found in this filter structure, but a stability proof based on Lyapunov theory or other methods has not yet been derived [27].

For the second-order direct form digital filter of type II, as described in Section 2.4.1, no *controlled rounding mechanism* can be devised to suppress zero-input limit cycles.
In this filter both state signals have to be quantized:

$$
\begin{aligned}
x_1(n+1) &= a\cdot x_1(n) + x_2(n) + \epsilon_1(n)\\
x_2(n+1) &= b\cdot x_1(n) + \epsilon_2(n). && (3.3.17)
\end{aligned}
$$

In Appendix III we prove that suppression of limit cycles with controlled rounding cannot be performed with control signals which are integer multiples of the state signals, as was the case for the direct form filter of type I. But, only for integer multiples the quantization direction can be determined uniquely. For non-integer values it is possible that both quantization upwards and downwards cause an increase of energy.
Therefore no controlled rounding mechanism can be devised to suppress zero-input limit cycles in the second-order direct form digital filter of type II.

Higher-order direct form digital filters are in general unstable with respect to quantization. For these filters no controlled rounding mechanism has been found for which the filter is zero-input stable for all allowed filter coefficients.

3.4. Quantization stability in wave digital filters

Historically, Lyapunov theory (with appropriate modifications) was first applied to *wave digital filters* [122, 123, 124]. For the second-order structure of Fig. 2.5.1, with state equations presented in (2.5.1), the energy function $E(n)$, given in (2.5.5) is a proper candidate for a Lyapunov function. With this choice the ellipses $E(n)$ = constant have axes parallel to the coordinate axes in the state plane. Applying MT quantization on the individual state variables reduces $|x_1(n)|$ and $|x_2(n)|$ and, consequently the energy $E(n)$, so that zero-input limit cycles are forbidden [234, 235, 341, 342, 343, 344, 347].

Under any constant-input condition† ($N = 1$), with a "controlled rounding" applied to $x_1(n)$, so that $x_1(n)$ is quantized in the direction of $-u(n)$, and MT applied to $x_2(n)$, freedom from subharmonics can be proved, according to [341, 342, 343, 344], with the energy function of (2.5.11), where

$$\underline{e}(n) = \underline{x}(n) - \tilde{\underline{x}}(n), \tag{3.4.1}$$

and

$$\tilde{\underline{x}}(n) = (-u(n), 0)^T. \tag{3.4.2}$$

This quantization mechanism of controlled rounding can be reduced to MT by first subtracting the control signal, then applying MT and finally again adding the control signal. Structures thus derived from wave digital filters are often presented in multi-output form with lowpass, highpass, bandpass, bandstop and allpass outputs [111, 193, 194, 195, 220, 222, 336].

In wave digital filters of orders higher than two some of the MT-quantizers can be replaced by RO without affecting the passivity of the complete filter [20, 228].

---

† Subharmonics of order S, which occur for discrete-time periodic input signals with period $N = 1$ are generally called constant-input limit cycles of period S.

For alternating input signals of the form

$$u(n) = (-1)^n \cdot U \quad \text{for} \quad n \geq 0, \tag{3.4.3}$$

freedom from quantization oscillations can be guaranteed by controlled rounding of $x_2(n)$ in the direction of $u(n)$, and MT applied to $x_1(n)$. Again, the energy function (2.5.11) can be used to prove this stability, where now

$$\tilde{\underline{x}}(n) = (0,\ u(n))^T. \tag{3.4.4}$$

For input signals with a period $N \geq 2$ it is not possible to guarantee stability of the filter; subharmonics have been observed in our simulations.

There are two strategies for suppressing zero-input limit cycles in *wave digital filter* of type II.

(a) We can use MT on both state signals, implying no intermediate word-length reduction: $NL_1\{x\} = NL_2\{x\} = MT\{x\}$ and $NL_3\{x\} = x$. This results in a filter which is equivalent to the wave digital filter of type I. The zero-input stability can be proved with the energy function $E(n)$ of (2.5.10).

(b) We can use a magnitude truncator on the summation point Q, implying $NL_3\{x\} = MT\{x\}$, so signal $q(n)$ has an integer value and quantization on the states does not have any effect: $NL_1\{x\} = NL_2\{x\} = x$. The zero-input stability is guaranteed; we follow the proof presented in [62, 63, 221].

The state equations of this second-order wave digital filter now become

$$\begin{aligned} q_o(n) &= -\gamma_1 \cdot x_1(n) - \gamma_2 \cdot x_2(n) + u(n) \\ q(n) &= MT\{q_o(n)\} \\ x_1(n+1) &= x_1(n) + q(n) \\ x_2(n+1) &= -x_2(n) - q(n). \end{aligned} \tag{3.4.5}$$

In the autonomous filter the signal $q(n)$ becomes zero after a finite time N_o. This statement is proved with the energy function $E(n)$ of (2.5.10):

$$\begin{aligned}\Delta E(n) &= E(n+1) - E(n) \\ &= \gamma_1 \cdot [x_1(n)+q(n)]^2 + \gamma_2 \cdot [-x_2(n)-q(n)]^2 - \gamma_1 \cdot x_1^2(n) - \gamma_2 \cdot x_2^2(n) \\ &= (\gamma_1+\gamma_2) \cdot q^2(n) - 2 \cdot q(n) \cdot q_o(n) \\ &\leq -(2-\gamma_1-\gamma_2) \cdot q^2(n). \end{aligned} \tag{3.4.6}$$

The last inequality is a consequence of magnitude truncation, for which we have

$$|q(n)| \leq |q_o(n)| \tag{3.4.7}$$

and

$$\text{sign}\{q(n)\} = \text{sign}\{q_o(n)\} \qquad \text{for } q(n) \neq 0. \tag{3.4.8}$$

With $2 - \gamma_1 - \gamma_2 > 0$ due to the linear stability condition (2.5.4) and the fact that $q(n)$ has integer values only, the energy function $E(n)$ is a monotonically decreasing function for $q(n) \neq 0$, which results in $q(n) = 0$ after a finite time N_o. Signal $q(n) = 0$ does, however, not imply that also the energy $E(n)$ is decreased to zero. If this is not the case, at least one of the two variables $x_1(n)$, $x_2(n)$ differs from zero, representing some form of limit cycle. Due to $q(n) = 0$ these limit cycles are local oscillations in the small loops of Fig. 2.5.2, the upper with period 1 and the lower with period 2. Their amplitudes can be estimated via the requirement that the absolute value of the signal $q_o(n)$ must not exceed the unit threshold of the MT characteristic.
So

$$|x_1(n)| < 1/\gamma_1$$

and

$$|x_2(n)| < 1/\gamma_2. \tag{3.4.9}$$

These oscillations can be made invisible by using a pair of *threshold detectors*, with characteristics

$$\begin{aligned} TH(x) &= x \qquad \text{for } |x| \geq T \\ &= 0 \qquad \text{for } |x| < T, \end{aligned} \tag{3.4.10}$$

with $T = 1/\gamma_1$ in the signal flow from $x_1(n)$ to the output $y(n)$
and $T = 1/\gamma_2$ in the flow from $x_2(n)$ to $y(n)$.
This way the output signal $y(n)$ becomes free from zero-input limit cycles.

Under constant-input conditions this filter remains free from parasitic oscillations, which can be proved, according to [62, 63], with the energy function of (2.5.12), where

$$\tilde{\underline{x}}(n) = (u(n)/\gamma_1,\ 0). \tag{3.4.11}$$

Just as for zero-input, we can show that the signal $q(n)$ becomes zero after a finite time N_o resulting in $|x_2(n)| < 1/\gamma_2$, so that these oscillations can be trapped with a threshold detector. The component $x_1(n)$ can, however, exceed the threshold value T, but it remains constant for $n \geq N_o$ implying that also the output signal $y(n)$ is periodic with period 1. Therefore this filter is free from constant-input limit cycles.

For periodic input signals with period $N = 2$, we can again prove that $q(n)$ becomes zero with the same energy function (2.5.12), where now

$$\tilde{x}_1(n) = \frac{u(n) + u(n\text{-}1)}{2\cdot\gamma_1}, \qquad \tilde{x}_2(n) = \frac{u(n) - u(n\text{-}1)}{2\cdot\gamma_2}. \tag{3.4.12}$$

For $n \geq N_o$ the state component $x_1(n)$ becomes periodic with period 1 and $x_2(n)$ with period 2. Therefore the output $y(n)$ is periodic with the same period as the input signal and no subharmonic can occur.

For input signals which are periodic with a period $N \geq 3$ the stabilization mechanism fails, so that subharmonics can appear again.
We conclude that the second-order wave digital filter of type II with magnitude truncation on the summation point Q is free from subharmonics for all input signals with periods 1 or 2.

Suppression of subharmonics for periods larger than two is a far more difficult task. While so far no general stabilization mechanism has been reported that works for all periods N, we develop such mechanisms for fixed values or at most a set of fixed values of N in the following sections.

3.5. Subharmonic-free filter for input signals with period N

In this section we present a new digital filter structure which is *free from subharmonics* for a discrete-time periodic input signal with an even period $N = 2M$. This filter is so designed that it also suppresses subharmonics for periodic input signals of period M. As an example, we choose $M = 3$, so we present a filter structure which is free from subharmonics for all input signals with periods 3 and 6.

This filter structure forms an extension of the wave digital filter II, as described in the previous section. The recursive part of this filter is shown in Fig. 3.5.1.

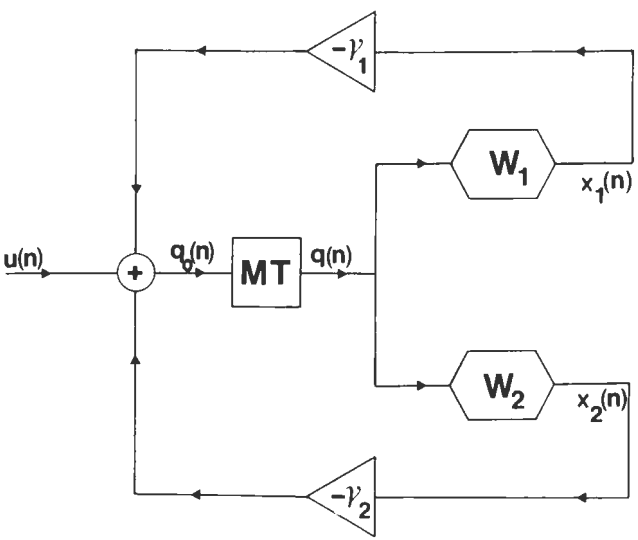


Fig. 3.5.1. Subharmonic-free filter for input signals with $N=M$ or $N=2M$.

In this filter the quantizer is a magnitude truncator MT with quantization step unity. The blocks W_1 and W_2 are two linear systems with transfer functions:

$$W_1(z) = \frac{X_1(z)}{Q(z)} = \frac{-M \cdot z^{-M}}{z^{-M} - 1} = \frac{M}{z^{M} - 1} \tag{3.5.1}$$

$$W_2(z) = \frac{X_2(z)}{Q(z)} = \frac{-M \cdot z^{-M}}{z^{-M} + 1} = \frac{-M}{z^{M} + 1}. \tag{3.5.2}$$

The output signal $y(n)$ (not shown in Fig. 3.5.1) is formed by a linear combination of $x_1(n)$, $x_2(n)$ and the internal state signals $x_3(n)$ to $x_{2M}(n)$ inside the boxes W_1 and W_2, according to

$$y(n) = \sum_{i=1}^{2M} c_i \cdot x_i(n) + d \cdot u(n). \tag{3.5.3}$$

The circuit of Fig. 3.5.1 is intended to function as a second-order filter section with transfer function

$$H(z) = \frac{(z - z_1) \cdot (z - z_2)}{(z - q_1) \cdot (z - q_2)}. \tag{3.5.4}$$

For a general choice of the coefficients c_i the system function is however of order $2M$. Reduction to the required second-order is obtained through cancellation of all unwanted poles by zeros through an appropriate choice of the coefficients c_i.

After some manipulations, the multipliers γ_1 and γ_2 are derived from the remaining poles q_1 and q_2:

$$\begin{aligned} \gamma_1 &= (1 - q_1^{M}) \cdot (1 - q_2^{M})/2M \\ \gamma_2 &= (1 + q_1^{M}) \cdot (1 + q_2^{M})/2M. \end{aligned} \tag{3.5.5}$$

Since the filter without wordlength reduction is assumed to be stable, expressed by

$$|q_1| < 1 \quad \text{and} \quad |q_2| < 1, \tag{3.5.6}$$

we have

$$\gamma_1 > 0, \quad \gamma_2 > 0 \quad \text{and} \quad \gamma_1 + \gamma_2 < \frac{2}{M}. \tag{3.5.7}$$

The transfer function from the input terminal to the quantizer is

$$F(z) = \frac{Q(z)}{U(z)} = \frac{z^{2M} - 1}{(z^M - q_1^M)\cdot(z^M - q_2^M)}. \tag{3.5.8}$$

The blocks W_1 and W_2 can be realized with a delay-line of M time-delay elements and a feedback multiplier with factor ± 1. For block W_1 we have:

$$x_1(n+M) = M\cdot q(n) + x_1(n), \tag{3.5.9}$$

and for block W_2:

$$x_2(n+M) = -M\cdot q(n) - x_2(n). \tag{3.5.10}$$

All signals in the blocks W_1 and W_2 have integer values only, so there is no need for quantization within these blocks. The only quantizer in the complete filter structure is the magnitude truncator MT.
In Fig. 3.5.2 the recursive part of this filter is shown in detailed form for $M = 3$.

For zero-input we prove with the function

$$E(n) = \gamma_1 \cdot \sum_{i=0}^{M-1} x_1^2(n+i) + \gamma_2 \cdot \sum_{i=0}^{M-1} x_2^2(n+i) \tag{3.5.11}$$

that signal $q(n)$ becomes zero after a finite time N_o.
Due to (3.5.7) this function is an energy function, with

$$E(n) > 0 \qquad \text{for all } \underline{x}(n) \neq \underline{0}. \tag{3.5.12}$$

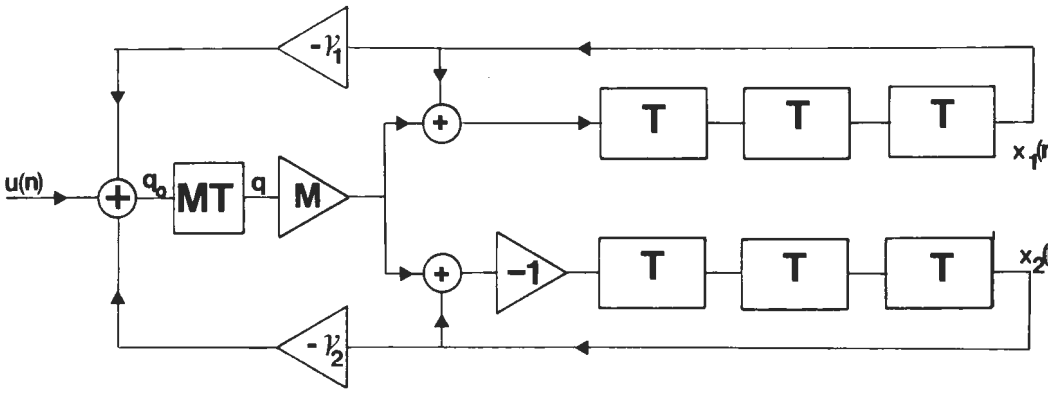


Fig. 3.5.2. Subharmonic-free filter for input signals with period 3, 6.

The energy function $E(n)$ is a semi-Lyapunov function, since

$$\begin{aligned}
\Delta E(n) &= E(n+1) - E(n) \\
&= \gamma_1 \cdot \left[x_1^{\,2}(n+M) - x_1^{\,2}(n) \right] + \gamma_2 \cdot \left[x_2^{\,2}(n+M) - x_2^{\,2}(n) \right] \\
&= M^2 \cdot (\gamma_1 + \gamma_2) \cdot q^2(n) - 2M \cdot q(n) \cdot q_o(n) \\
&\leq -M \cdot [2 - M \cdot (\gamma_1 + \gamma_2)] \cdot q^2(n). \qquad (3.5.13)
\end{aligned}$$

The last inequality follows from $q(n) \cdot q_o(n) \geq q^2(n)$, which holds for magnitude truncation. Further we know, according to (3.5.7) that

$$[2 - M\cdot(\gamma_1+\gamma_2)] > 0, \tag{3.5.14}$$

so

$$\Delta E(n) < 0 \qquad \text{for } q(n) \neq 0. \tag{3.5.15}$$

Since $q(n)$ has integer values only this signal must become zero after a finite time N_o. Signal $q(n) = 0$ does, however, not imply that also the energy $E(n)$ has decreased to zero. If this is not the case some form of limit cycle appears in the filter. Due to $q(n) = 0$ these limit cycles are local oscillations in the blocks W_1 and W_2 of Fig. 3.5.1, the first one with period M and the second one with period $2M$. Their amplitudes are bounded by the values

$$|x_1(n)| < 1/\gamma_1 \tag{3.5.16}$$

and

$$|x_2(n)| < 1/\gamma_2. \tag{3.5.17}$$

These oscillations can be made invisible by using a threshold dectector with $T = 1/\gamma_1$ in the signal lines from $x_1(n)$ and its delayed versions to $y(n)$ and $T = 1/\gamma_2$ in the lines from $x_2(n)$ and its delayed versions to $y(n)$. This way the output signal $y(n)$ is free from zero-input limit cycles.

For input signals with period $N = M$ or $N = 2M$ it can be proved that the filter of Fig. 3.5.1 is free from subharmonics with the energy function

$$E(n) = \gamma_1 \cdot \sum_{i=0}^{M-1} e_1^{\,2}(n+i) + \gamma_2 \cdot \sum_{i=0}^{M-1} e_2^{\,2}(n+i), \tag{3.5.18}$$

where error vector $\underline{e}(n)$ is the difference between the actual state $\underline{x}(n)$ and the idealized linear response $\tilde{\underline{x}}(n)$, which in this filters is

$$\begin{aligned} \tilde{x}_1(n) &= [u(n) + u(n+M)]/2\gamma_1 \\ \tilde{x}_2(n) &= [u(n) - u(n+M)]/2\gamma_2. \end{aligned} \tag{3.5.19}$$

The energy function $E(n)$ is a semi-Lyapunov function, since

$$\begin{aligned}
\Delta E(n) &= E(n+1) - E(n) \\
&= \gamma_1 \cdot \left[e_1^{\,2}(n+M) - e_1^{\,2}(n) \right] + \gamma_2 \cdot \left[e_2^{\,2}(n+M) - e_2^{\,2}(n) \right] \\
&= M^2 \cdot (\gamma_1 + \gamma_2) \cdot q^2(n) - 2M \cdot q(n) \cdot q_o(n) \\
&\leq -M \cdot [2 - M \cdot (\gamma_1 + \gamma_2)] \cdot q^2(n). \qquad (3.5.20)
\end{aligned}$$

So, the signal $q(n)$ will become zero after a finite time N_o resulting in local oscillations with period M in block W_1 and with period $2M$ in block W_2. For input signals with period $N = M$ the oscillations in block W_2 can be trapped with a threshold detector so they are invisible in the output signal. The signal in block W_1 can exceed the threshold value, but it is periodic with period M, implying that also the output signal is periodic with this period and no subharmonics can occur.
For input signals with period $N = 2M$ all the states become periodic with period $2M$, thus no subharmonic can occur.

For periodic input signals with other periods the occurrence of subharmonics cannot be excluded. For example, the appearance of constant-input limit cycles is always possible in this filter for $M \neq 1$.

In the next section we will change the structure of this filter in order to get a filter which is free from subharmonics for all input signals which are discrete-time periodic with a period N, being a divisor of $2M$.

3.6. Subharmonic-free filter for input signals with N a divisor of $2M$

In this section we present a new digital filter structure, which is free from subharmonics for *all* input signals which are discrete-time periodic with a *period* N being a *divisor of* $2M$. We demonstrate this filter for $M = 3$, so we present a filter structure which is free from subharmonics for all input signals with periods 1, 2, 3 and 6.
The basic idea of this filter is to separate the periodic state sequence which appears in the circuit of the previous section, in components which are periodic with a divisor of $2M$. Subharmonic components can then be suppressed by a threshold detector.

The new filter structure forms an *extension* of the circuit described in Section 3.5. In that structure (cf. Fig. 3.5.1) we split the block W_1 into some blocks G_D connected in parallel

$$W_1(z) = \sum_{D \mid M} G_D(z), \tag{3.6.1}$$

where $D|M$ stands for all divisors D of M, including $D = 1$ and $D = M$.
In the same way the block W_2 is split up into some blocks G_D, where D is a divisor of $2M$, which is <u>not</u> also a divisor of M:

$$W_2(z) = \sum_{\substack{D \mid 2M \\ D/M}} G_D(z). \tag{3.6.2}$$

The poles of the functions $W_1(z)$ and $W_2(z)$, which are the roots of the polynomial $z^{2M}-1$, are zeros of the denominators of the functions $G_D(z)$.

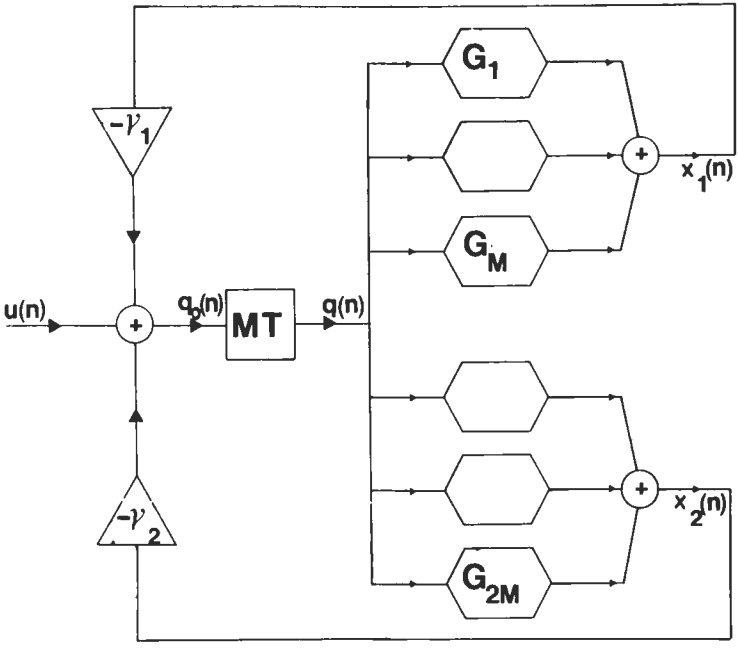


Fig. 3.6.1. Subharmonic-free filter for input signals which are periodic with a period N which is a divisor of $2M$.

In Appendix IV we show that the denominators of $G_D(z)$ can be performed by cyclotomic polynomials $Q_D(z)$, which have very simple coefficients. In more explicit terms, the functions $G_D(z)$ are described by

$$G_D(z) = \frac{z \cdot \frac{d}{dz}(Q_D(z^{-1}))}{Q_D(z^{-1})} = z \cdot \frac{d}{dz}\left[\log(Q_D(z^{-1}))\right], \qquad (3.6.3)$$

where $Q_D(z)$ is the cyclotomic polynomial of order D.

Insertion of (3.6.3) into (3.6.1) indeed satisfies (3.5.1), since according to Appendix IV

$$\sum_{D|M} G_D(z) = z\cdot\frac{d}{dz}\left[\log\{\prod_{D|M} Q_D(z^{-1})\}\right]$$

$$= z\cdot\frac{d}{dz}\left[\log\{z^{-M} - 1\}\right] = W_1(z). \qquad (3.6.4)$$

The same reasoning leads to a verification of (3.6.2).

We have split the functions $W_1(z)$ and $W_2(z)$ into the sum of a number of functions so that for every divisor D of $2M$ there is exactly one function $G_D(z)$ (as shown in Fig. 3.6.1).

According to property 9) of Appendix IV the functions $G_D(z)$ have integer coefficients only, implying that they can be realized with integer multipliers, without requiring any quantization in the blocks G_D.

Moreover, the coefficients of the most cyclotomic polynomials have the values -1, 0 or 1 only (see property 1) to 8) in Appendix IV) so that the integer multipliers can be replaced by simple adders.
The complete realization for the case $M = 3$ is shown in Fig. 3.6.2.

The number of time-delay elements in a block G_D equals the degree of the cyclotomic polynomial $Q_D(z)$ which, according to Appendix IV, is equal to $\phi(D)$, Eulers phi function. The total number of time delay elements in the complete circuit is

$$\sum_{D|2M} \phi(D) = 2M, \qquad (3.6.5)$$

which is the same as in the structure in Fig. 3.5.1.

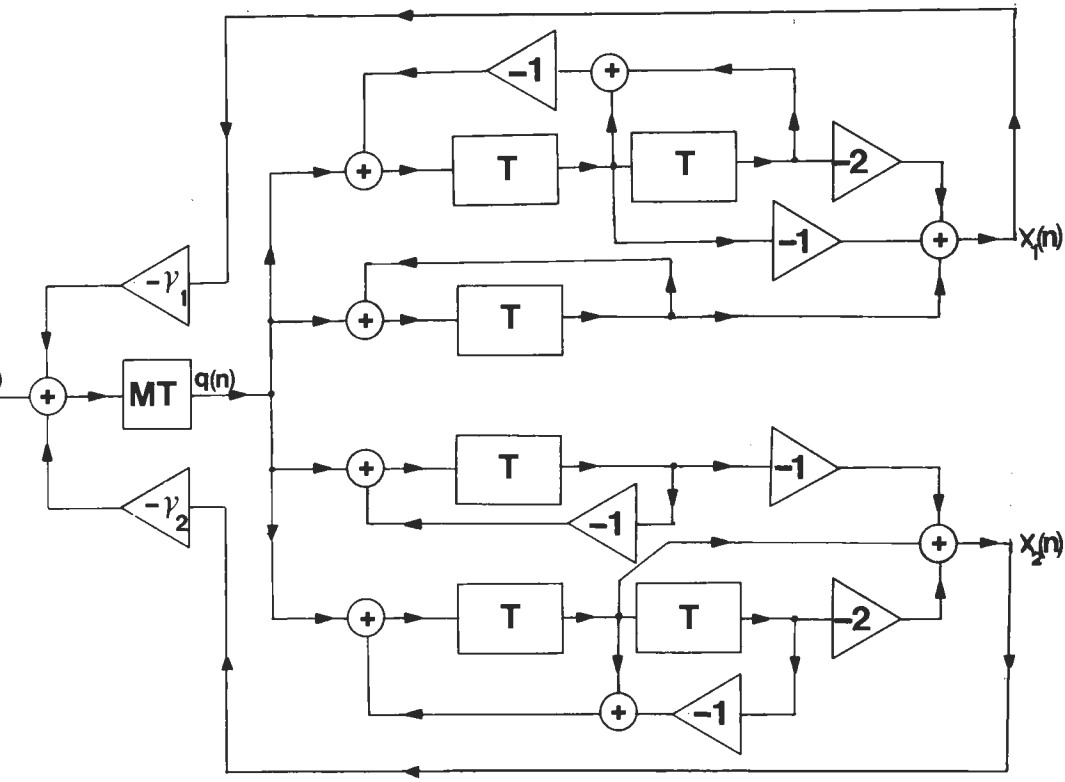


Fig. 3.6.2. Subharmonic-free filter for periodic input signals with periods 1, 2, 3 and 6.

Comparing the circuit of Fig. 3.6.1 with that of Fig. 3.5.1 we observe that the transfer function $F(z)$ between the input signal $u(n)$ and the quantizer $q(n)$ remains exactly the same (see (3.5.8)). So we can still prove that for zero-input the value of $q(n)$ becomes zero after a finite time N_o.

From this moment, limit cycles are locally oscillating within the blocks G_1 to G_{2M}, each with a period equal to the index D of G_D. These limit cycles are amplitude bounded and can be made invisible in the output by using threshold detectors in the signal lines from the state signals to output signal $y(n)$.

For a periodic input signal, with a period N equal to a divisor of $2M$, we can define an energy function equivalent to (3.5.18). This energy function is a semi-Lyapunov function and thus guarantees that the value of $q(n)$ becomes zero after a finite time N_o. This results in local oscillations within the blocks G_1 to G_{2M}, each with a period equal to the index D of G_D. For a periodic input signal with period N, the oscillations within the blocks G_D for values of D which are <u>not</u> a divisor of N do <u>not</u> exceed the threshold value and thus are invisible in the output signal. This means that the output $y(n)$, which is a weighted sum of state signals $x_D(n)$ with D a divisor of N, also is periodic with period N. The input and the output signal have the same period, so $S = 1$ and no subharmonic can occur in this filter.

This filter is designed in such a way that it not only suppresses subharmonics for discrete-time periodic signals with period $N = M$ or $N = 2M$, but that the same is true for periods which are divisors of $2M$. The price to pay for this stabilization is an extension of the circuit with extra adders, integer-multipliers and time-delay elements. For suppressing the previously mentioned subharmonics this circuit requires $2M$ time-delay elements and at most an equal number of adders and integer multipliers.

3.7. Quantization stability in state-space digital filters

In *state-space digital filters* (presented in Section 2.6) only the state variables undergo quantization, while intermediate results are represented with full precision.

If magnitude truncation is used in a such a digital filter, it is guaranteed to be zero-input stable if in the idealized filter (without wordlength reduction) there exists a monotonically decreasing energy function of the form

$$E(n) = \underline{x}^T(n) \cdot D \cdot \underline{x}(n), \tag{3.7.1}$$

where D a positive diagonal matrix [339, 340].

For the second-order state-space filter this condition corresponds to that of (2.6.3) and therefore, just as in (2.6.4), we find that it can be fulfilled for all system matrices A whose coefficients satisfy the condition

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A]. \tag{3.7.2}$$

In such state-space filters, the ellipses $E(n)$ = constant have axes parallel to the coordinate axes of the state plane, so that magnitude truncation applied to the individual state signals reduces energy, implying that $E(n)$ is a Lyapunov function of the nonlinear system and the filter is free from zero-input limit cycles.

Concerning stability with respect to *constant-* (*nonzero-*) *inputs*, we mention a general principle to convert a stable autonomous state-space digital filter into a system with input terminals stable under any constant excitation (see [27, 100, 101, 102, 112, 176, 177, 178, 179, 270, 335, 336, 337]).
In more explicit terms, let the solution of the autonomous system become zero after a finite time (expressing freedom of zero-input limit cycles)

then through suitably supplying such a system with an input terminal, a constant-input stable system can be created as follows.
At each quantization point i, some signal v_i is added after quantization, while the same signal is subtracted after the sum signal has passed the subsequent time-delay element (see Fig. 3.7.1 for the second-order state-space digital filter).

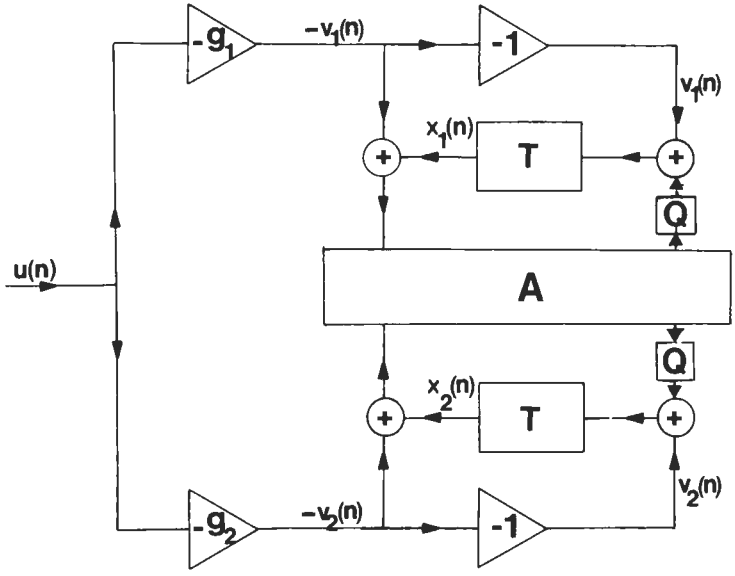


Fig. 3.7.1. Constant-input limit cycle free state-space digital filter.

The pair of injected signals v_i is proportional to the input signal,

$$v_i(n) = g_i \cdot u(n), \tag{3.7.3}$$

where the coefficient g_i is chosen to have an integer value.

The state in the modified system satisfies

$$\underline{x}(n+1) = NL\{A\cdot[\underline{x}(n) - \underline{g}\cdot u(n)]\} + \underline{g}\cdot u(n). \tag{3.7.4}$$

The output is formed by a linear combination of the state signals and the input signal

$$y(n) = \underline{c}^T\cdot\underline{x}(n) + d\cdot u(n). \tag{3.7.5}$$

In the idealized linear system with no quantization the output and the input are related by the transfer function

$$H(z) = \frac{Y(z)}{U(z)} = \underline{c}^T\cdot(z\cdot I - A)^{-1}\cdot\underline{b} + d, \tag{3.7.6}$$

where I denotes the unit matrix and $\underline{b} = (I - A)\cdot\underline{g}$.
The poles of the transfer function $H(z)$ are determined by the system matrix A. The zeros of this function can be chosen arbitrarily (for any vector $\underline{g} \neq \underline{0}$) by choosing the correct values of the coefficients of $\underline{c}$ and d, because the components of the state signal $\underline{x}(n)$ and the input signal $u(n)$ are mutually independent: this freedom in choice of $\underline{g}$, $\underline{c}$ and d provides the filter to be controllable and observable.

For a constant excitation $u(n) = U$ the difference vector

$$\underline{e}(n) = \underline{x}(n) - \underline{g}\cdot u(n) \tag{3.7.7}$$

satisfies the homogeneous nonlinear equation

$$\underline{e}(n+1) = NL\{A\cdot\underline{e}(n)\}, \tag{3.7.8}$$

whose solution becomes zero after a finite time, in accordance with our assumptions concerning the unexcited system.

So state vector $\underline{x}(n)$ asymptotically approaches the stationary solution $\underline{g}\cdot U$ without superimposed oscillations: the filter is free from constant-input limit cycles.

This principle to convert a stable autonomous state-space digital filter into a filter stable for constant-input excitations, will now be extended to form a *subharmonic-free filter for periodic input signals with period N*.

First a signal $v_i(n)$, which is proportional to the input signal

$$v_i(n) = g_i \cdot u(n), \tag{3.7.9}$$

where g_i has an integer value, is subtracted from the state component $x_i(n)$. Then, a future value of the signal $v_i(n+1) = g_i \cdot u(n+1)$ is added after the quantizer, just before the time-delay element. For a periodic input signal with period N it is easy to determine the value of $v_i(n+1)$ in a causal way by choosing

$$v_i(n+1) = g_i \cdot u(n-N+1). \tag{3.7.10}$$

So the state equation of the modified system becomes

$$\underline{x}(n+1) = \mathrm{NL}\{A\cdot[\underline{x}(n) - \underline{g}\cdot u(n)]\} + \underline{g}\cdot u(n-N+1) \tag{3.7.11}$$

(see Fig. 3.7.2 for the second-order state-space digital filter).

The output $y(n)$ of this filter (not shown in Fig. 3.7.2) is formed by a linear combination of the state signals and N samples of the input signal, the latter to compensate additional zeros introduced in the transfer function by the delayed input value $u(n-N+1)$, according to

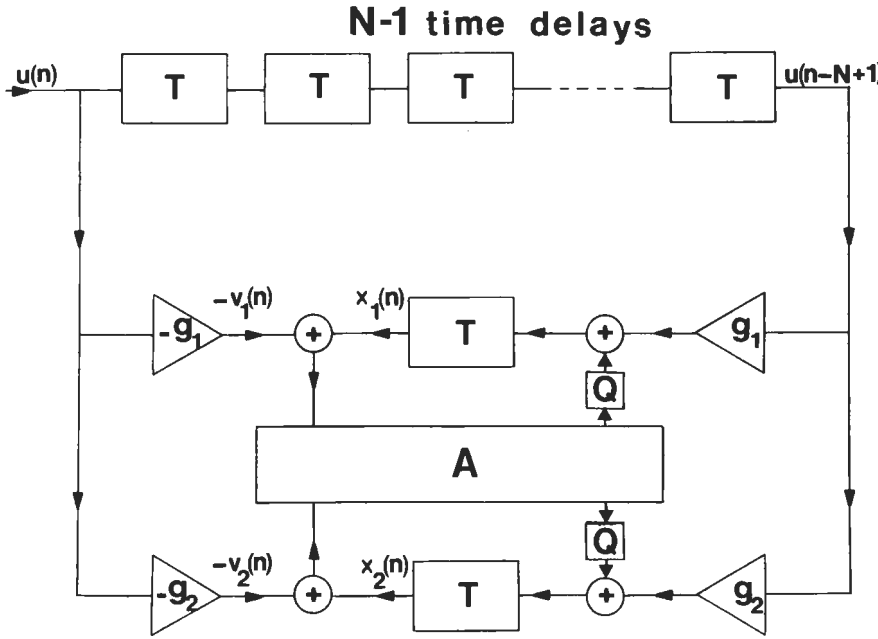


Fig. 3.7.2. Subharmonic-free filter for input signals with period N.

$$y(n) = \sum_{i=1}^{2} c_i \cdot x_i(n) + \sum_{i=0}^{N-1} d_i \cdot u(n-i). \tag{3.7.12}$$

The circuit of Fig. 3.7.2 is intended to function as a second-order filter section with a transfer function

$$H(z) = \frac{(z - z_1)\cdot(z - z_2)}{(z - q_1)\cdot(z - q_2)}. \tag{3.7.13}$$

This can be achieved by an appropriate choice of the values of the coefficients c_i and d_i.

For a periodic input signal $u(n)$ with period N the difference vector

$$\underline{e}(n) = \underline{x}(n) - \underline{g}\cdot u(n) \tag{3.7.14}$$

satisfies the homogeneous equation (3.7.8) so the state vector $\underline{x}(n)$ asymptotically approaches the stationary solution $\underline{g}\cdot u(n)$, which is periodic with the same period as the input signal. According to (3.7.12) also the output $y(n)$ will become periodic with period N, implying $S = 1$, which expresses freedom from subharmonics.
Moreover, for all periodic input signals with a period D which is a divisor of N the state $\underline{x}(n)$ will asymptotically approache the stationary solution $\underline{g}\cdot u(n)$, which is periodic with the same period D. According to (3.7.12) also the output $y(n)$ will become periodic with this period, so this filter is free from subharmonics for all input signals which have a period which is a divisor of N.

According to (3.7.12) the suppression of subharmonics for a periodic input signal with period N requires a total number of $(N\text{-}1)$ extra adders, multipliers and time-delay elements, the latter to store the values of $u(n\text{-}i)$ for $i = 1$ to $N\text{-}1$.

At a first glance, it seems to be possible to get a subharmonic-free filter by adjusting a zero-input stable filter according to

$$\underline{x}(n+1) = \mathrm{NL}\{A\cdot[\underline{x}(n) - \underline{g}\cdot u(n)]\} + \underline{g}\cdot u(n+1). \tag{3.7.15}$$

For any input $u(n)$ the difference vector $\underline{e}(n) = \underline{x}(n) - \underline{g}\cdot u(n)$ satisfies the homogeneous equation (3.7.8), resulting in an asymptotically stationary solution which is periodic with the same period as the input signal and therefore subharmonic free.
But in this system, with $\underline{x}(n) = \underline{g}\cdot u(n)$, all the state signals $x_i(n)$ are linearly dependent upon the excitation $u(n)$ so that this filter is worthless in the sense that no filtering can take place; the transfer function becomes

$$H(z) = \underline{c}^{\mathrm{T}}\cdot\underline{g} + d = \text{constant}. \tag{3.7.16}$$

For a purely *sinusoidal input signal* of the form

$$u(n) = U\cdot\cos(\Omega\cdot n + \phi) \tag{3.7.17}$$

the value of the input signal on time $n+1$ can be determined from the present and previous input signal value according to

$$u(n+1) = 2\cdot\cos(\Omega)\cdot u(n) - u(n-1). \tag{3.7.18}$$

This value can be used to modify a zero-input stable filter into a system which suppresses subharmonics for sinusoidal input signals with frequency Ω (see Fig. 3.7.3 for the second-order state-space filter).

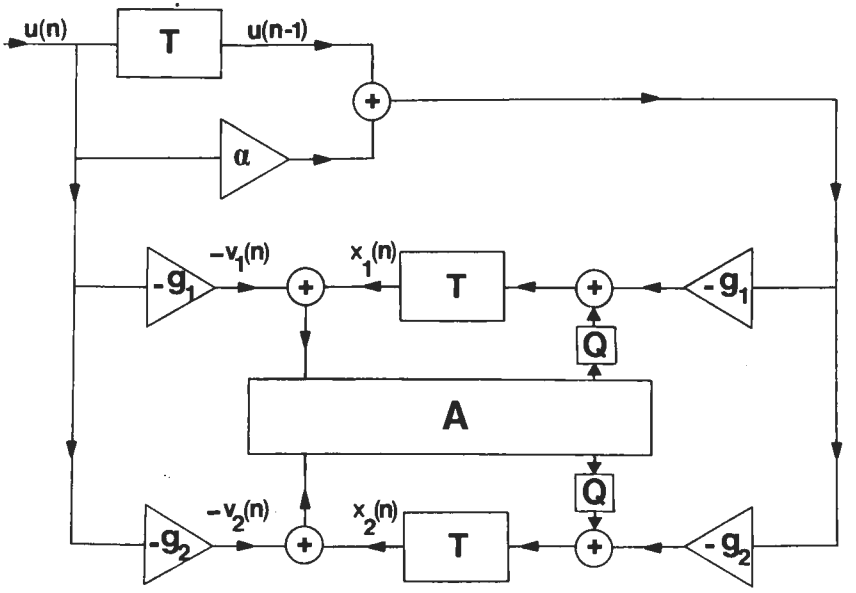


Fig. 3.7.3. Subharmonic-free filter for pure sinusoid input signals with frequency Ω, where $\alpha = -2\cdot\cos(\Omega)$.

The modified system is described by the state equation

$$\underline{x}(n+1) = NL\{A\cdot[\underline{x}(n)-\underline{g}\cdot u(n)]\} + \underline{g}\cdot[2\cos(\Omega)\cdot u(n)-u(n-1)]. \quad (3.7.19)$$

For a sinusoidal input signal with frequency Ω the difference vector $\underline{e}(n)$ satisfies the homogeneous state equation (3.7.8), which asymptotically reaches zero, so the state vector $\underline{x}(n)$ is subharmonic free. Moreover this state vector is just as the input signal a pure sinusoid with frequency Ω, so also harmonical distortion of the signal is absent.

The suppression of subharmonics for a purely sinusoidal input signal requires only 1 extra multiplier and 1 time-delay element.

Besides the basic zero-input limit-cycle suppressing concept in state-space digital filters, a vast amount of ideas has been published dealing with special structures and more complicated (deterministic) stabilization methods. In summarized form we mention: multirate filters [95, 317, 369], error-feedback filters [77, 295, 319], digital incremental computers [368] and other special structures [204, 205, 332, 333]. Special investigations concern coupled-form filters [268], cascade sections [171, 181, 182], sections with non-uniform internal wordlength [135, 253] and with small input signals [152].

The principle to convert an above mentioned stable autonomous digital filter into a system with input terminals stable under any periodic input signal with period N is equivalent to that of the state-space digital filters.
So we have develloped a mechanism to suppress subharmonics for a fixed value or at most a set of fixed values of N. Suppression of subharmonics in digital filters for all possible input signals appears to be impossible. Therefore we present some investigation into the properties of subharmonics in digital filters in the next chapter.

4. Analysis of subharmonics in recursive digital filters

4.1. Computer search for subharmonics

A periodically excited recursive digital filter can respond with a variety of periodic output signals. A given filter with a given excitation can generate a whole set of subharmonics of different order and different structure. There can also occur one or more periodic output signals with the same period as the input signal. Finally, pure harmonics and even output signals with periods completely unrelated to the input period can belong to the set of periodic responses. Which of these responses is actually appearing in the digital filter is determined by the initial conditions.

In this section we present a computer program searching for the complete set of periodic output signals responding to a certain periodic input signal, with period N. In order to generate a subharmonic we could start from an initial state $\underline{x}(0)$ and run the filter until a periodic state sequence is reached, in which the value of the state signal $\underline{x}(n)$ equals a previous state value $\underline{x}(n-S\cdot N)$ for some integer value of S, and where S is chosen as small as possible. Therefore starting from $\underline{x}(0)$ all the states $\underline{x}(k\cdot N)$ will be labelled with the index k. If for some value k_1 we reach a state $\underline{x}(k_1\cdot N)$ which has already been labelled with an index k_o we have entered a periodic state sequence with period $S\cdot N = (k_1-k_o)N$. If $S \neq 1$ this periodic state sequence forms a subharmonic. This procedure can be repeated for all initial states $\underline{x}(0)$ in order to get a complete catalogue of subharmonics.
For zero-input we can generate the complete set of limit cycles by repeating this procedure. Starting from an initial state $\underline{x}(0)$ a zero-input limit cycle is reached if the state $\underline{x}(k_1)$ equals a previous state $\underline{x}(k_o) \neq \underline{0}$, with period $S = k_1-k_o$.

The set of states which can form part of a periodic state sequence is restricted. In order to find some necessary conditions for this set we will determine an amplitude bound for the error vector $\underline{e}(n)$, which is the difference of the periodic state sequence $\underline{x}(n)$ and the asymptotic steady state sequence $\tilde{\underline{x}}(n)$ of the idealized linear filter.

This error vector $\underline{e}(n)$ satisfies the relation

$$\begin{aligned}\underline{e}(n+1) &= \underline{x}(n+1) - \tilde{\underline{x}}(n+1) \\ &= Q\{A\cdot[\tilde{\underline{x}}(n) + \underline{e}(n)] + \underline{b}\cdot u(n)\} - A\cdot\tilde{\underline{x}}(n) - \underline{b}\cdot u(n) \\ &= A\cdot\underline{e}(n) + \underline{\epsilon}(n), \end{aligned} \tag{4.1.1}$$

where we have substituted quantization error $\underline{\epsilon}(n)$, defined according to (1.3.1) by $\underline{\epsilon} = Q(\underline{x}) - \underline{x}$. Equation (4.1.1) is equivalent to that of a zero-input limit cycle, so that the amplitude bounds derived in this section for the error vector $\epsilon(n)$ are the same as those for zero-input limit cycles.

The transfer function from quantizer j to the state component x_i is determined by

$$F_{ij}(z) = \underline{c}^T\cdot(z\cdot I - A)^{-1}\cdot\underline{b}, \tag{4.1.2}$$

where $c_i = 1$, $b_j = 1$ and all other components of $\underline{b}$ and $\underline{c}$ are zero.

With quantization error $\epsilon_j(n)$ bounded by q_j we see that

$$|e_i(n)| \le \sum_j q_j \cdot \sum_{m=0}^{\infty} |f_{ij}(m)| = L_i. \tag{4.1.3}$$

For the second-order direct form filter of Fig. 1.2.1 with one quantizer we have

$$\begin{aligned} F_{11}(z) &= z^{-1}\cdot H(z) \\ \text{and}\quad F_{21}(z) &= z^{-2}\cdot H(z), \end{aligned} \tag{4.1.4}$$

so for both $i = 1$ and $i = 2$ we find

$$|e_i(n)| \le q\cdot \sum_{m=0}^{\infty} |h(m)| = L. \tag{4.1.5}$$

If the transfer function $H(z)$ has two different real poles the bound L can exactly be calculated and equals

$$L = \frac{q}{1 - |a| - b}. \tag{4.1.6}$$

For two complex conjugated poles $q_{1,2} = \rho \cdot e^{\pm j\theta}$ we have

$$L = \frac{q}{\sin(\theta)} \cdot \sum_{m=0}^{\infty} \rho^m \cdot |\sin((m+1)\cdot\theta)|, \tag{4.1.7}$$

for which so far no closed form expression has been reported. This fact has motivated many authors to derive simpler upper bounds more or less higher than L (see e.g. [8, 23, 25, 40, 57, 74, 129, 137, 143, 155, 158, 189, 190, 226, 227, 279, 280, 321, 338, 367, 370].

Besides the amplitude of the error vector also its power has an upper bound [244, 246, 303], the same holds for other norms of this signal [144, 224]. Many papers deal with amplitude bounds in special digital structures, such as coupled form digital filters [162, 163, 165, 209], digital incremental computers [1, 2, 5, 253, 328, 329, 368], filters with residue numbers [28, 113, 114, 293], filters with error feedback [3, 4, 6, 7, 9, 10, 34, 75, 76, 83, 219, 260, 261, 262, 264, 295, 357, 358, 359, 360, 365], multi-rate digital filters [95, 280, 317, 369] and sampled data systems [24].

Amplitude bounds have been determined for wave digital filters with internal oscillations [282], filters with floating-point arithmetic [51, 98, 99, 172, 196, 197, 198, 214, 250, 301] and cascade sections [68, 69, 181, 229].

Special attention has been devoted to limit cycles that are almost sinusoidal [150, 152, 155, 162, 163, 233, 254, 271, 277, 328, 329, 353], limit cycles with period 1 and 2 [15, 16, 57, 73, 82, 173, 210, 212, 292, 324, 328, 329], and rolling-pin limit cycles [201, 203]. Amplitude bounds have also been derived with Lyapunov functions [276, 292, 353], with mathematical models [25, 139, 140, 283], and with computer simulations [265, 267, 288].

Determining the amplitude bounds is, in fact, the classical technique to cope with zero-input limit cycles. This enables a user to apply a course requantization such that a potential limit cycle remains smaller than a quantization step [174, 207]. Usually, the requantization is only applied at the lowest signal levels (thus forming a threshold detector) although it then fails to work under constant (nonzero-) input conditions. In view of the modern suppression methods the "screening" method introduces a relatively high degree of signal distortion.

The search of subharmonics starts from an initial state $\underline{x}(0)$ and runs the filter until a periodic state sequence is reached, so that state signal $\underline{x}(n)$ equals a previous state $\underline{x}(n-S\cdot N)$ for some integer value of S. If $S \neq 1$ this sequence forms a subharmonic. This procedure is repeated for all initial states $\underline{x}(0)$ in order to get a complete catalogue of all possible subharmonics. However, since the error vector is bounded according to (4.1.3) we only have to search from a starting point $\underline{x}(0)$ in the neighbourhood of the steady state $\tilde{\underline{x}}(0)$ with a difference in component i absolutely bounded by the value L_i.

For the second-order direct form filter with two complex conjugated poles $q_{1,2} = \rho\cdot\exp(\pm j\theta)$ we can restrict this area even more by the two tighter bounds

$$|e_1(n) - \rho\cdot\cos(\theta)\cdot e_2(n)| \leq q\cdot \sum_{m=0}^{\infty} \rho^m \cdot |\cos(m\theta)| \qquad (4.1.8)$$

and

$$|e_1(n) - 2\rho\cdot\cos(\theta)\cdot e_2(n)| \leq q\cdot\left[1 + \rho^2\cdot\sum_{m=0}^{\infty} \rho^m \cdot \frac{|\sin((m+1)\theta)|}{|\sin(\theta)|}\right]. \qquad (4.1.9)$$

Proof:

Starting form the relation

$$e_1(n) = \sum_{m=0}^{n} \epsilon(m)\cdot h(n-m)$$

$$= \sum_{m=0}^{n} \epsilon(m)\cdot \rho^{n-m}\cdot \frac{\sin(n-m+1)\theta}{\sin(\theta)}$$

$$= \rho\cdot\cos(\theta)\cdot \sum_{m=0}^{n} \epsilon(m)\cdot h(n-1-m) + \sum_{m=0}^{n} \epsilon(m)\cdot \rho^{n-m}\cdot\cos(n-m)\theta$$

$$= \rho\cdot\cos(\theta)\cdot e_2(n) + \sum_{m=0}^{n} \epsilon(n-m)\cdot \rho^{m}\cdot\cos(m\theta), \qquad (4.1.10)$$

the bound of equation (4.1.8) can be derived.

Further, with the expression

$$e_1(n) = Q\{a\cdot e_1(n-1) + b\cdot e_2(n-1)\}$$

$$= 2\rho\cdot\cos(\theta)\cdot e_2(n) - \rho^2\cdot e_2(n-1) + \epsilon(n) \qquad (4.1.11)$$

the second bound can be found.

□

To illustrate the possible occurrence of subharmonics we consider the second-order direct form digital filter of Fig. 1.2.1 with one magnitude truncator.

For filter coefficients

$$a = 1.336$$
$$\text{and} \quad b = -0.893, \tag{4.1.12}$$

resulting in complex conjugated poles

$$q_{1,2} = 0.945 \cdot e^{\pm j \cdot \pi/4}, \tag{4.1.13}$$

we obtain the numerical value

$$L = 15.9474. \tag{4.1.14}$$

Including (4.1.8) and (4.1.9), the error vector is bounded by the inequalities

$$\begin{aligned} |e_1(n)| &< 15.9474 \\ |e_2(n)| &< 15.9474 \\ |e_1(n) - 0.668 \cdot e_2(n)| &< 11.1845 \\ |e_1(n) - 1.336 \cdot e_2(n)| &< 15.2414 \end{aligned} \tag{4.1.15}$$

The filter is assumed to be excited by the periodic sequence

$$u(n) = 8, 8, -8, -8, \ldots . \tag{4.1.16}$$

which has period $N = 4$. The asymptotic linear response can directly be started from the initial state values

$$\begin{aligned} \tilde{x}_1(0) &= -6.4264 \\ \tilde{x}_2(0) &= 5.4734. \end{aligned} \tag{4.1.17}$$

So subharmonics can only start from initial conditions within the region shown in Fig. 4.1.1.

The following periodic responses turn out to be possible in this filter, depending upon the initial state $\underline{x}(0)$ (see [64]).

$$
\begin{aligned}
y(n) &= -4,\ 8,\ 6,\ -7,\ -6,\ 6,\ 5,\ -6,\ \ldots\ . \\
y(n) &= -3,\ 9,\ 6,\ -8,\ -8,\ 4,\ 4,\ -6,\ \ldots\ . \\
y(n) &= -5,\ 6,\ 4,\ -8,\ -6,\ 7,\ 6,\ -6,\ \ldots\ . \\
y(n) &= -6,\ 8,\ 8,\ -4,\ -4,\ 6,\ 3,\ -9,\ \ldots\ . \\
y(n) &= -5,\ 7,\ 5,\ -7,\ \ldots\ .
\end{aligned}
\tag{4.1.18}
$$

The system thus responds with four different types of second-order subharmonics and one output signal with preserved period. The output signal with period 4 is inaccessible, which means that it can be found only by starting with the initial state $\underline{x}(0)$ of the sequence itself. The remaining subharmonic output signals are all accessible sequences.

If we use rounding instead of magnitude truncation the value of bound L becomes

$$L = 7.9737. \tag{4.1.19}$$

Now we only have to search starting from an initial state $\underline{x}(0)$ in the neighbourhood of $\tilde{\underline{x}}(0)$ with an error bounded by

$$
\begin{aligned}
|e_1(n)| &\leq 7.9737 \\
|e_2(n)| &\leq 7.9737 \\
|e_1(n) - 0.668 \cdot e_2(n)| &< 5.5923 \\
|e_1(n) - 1.336 \cdot e_2(n)| &< 7.6207
\end{aligned}
\tag{4.1.20}
$$

(see Fig. 4.1.1)

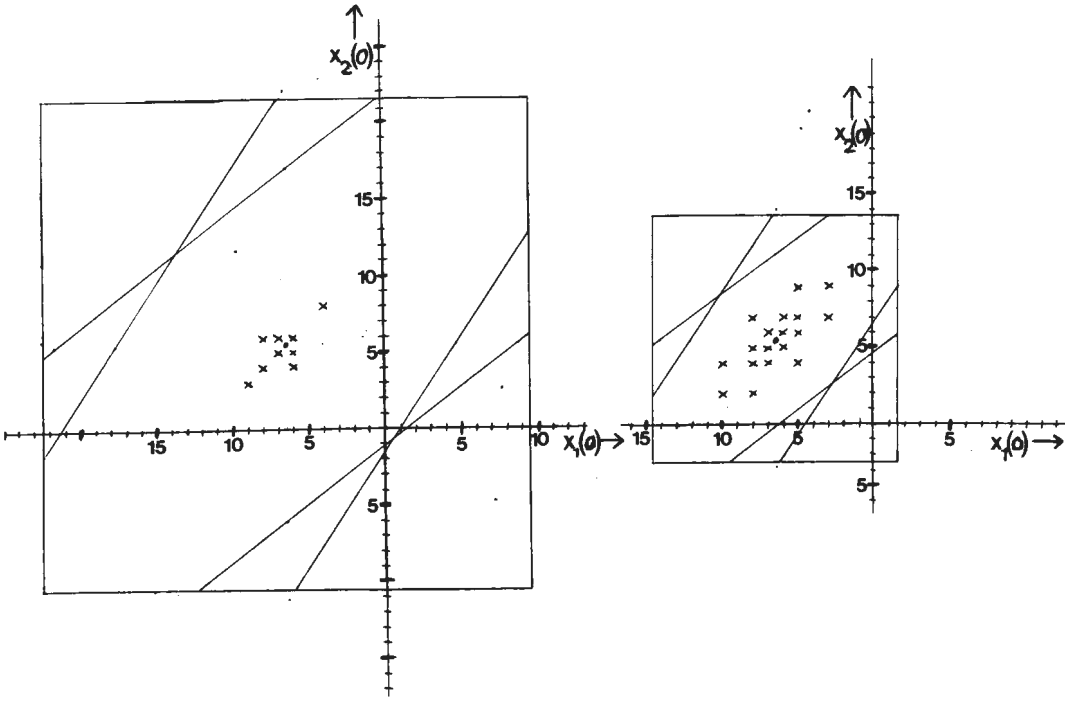


Fig. 4.1.1. Region of initial conditions $\underline{x}(0)$, which can lead to subharmonics in the filter described in the text for (a) magnitude truncation and (b) rounding. The crosses x denote initial states which actually lead to subharmonics.

The periodic state sequences which have been found in this filter are

$y(n)$ = -4, 10, 9, -5, -7, 3, 2, -8, ...
$y(n)$ = -4, 8, 6, -7, -7, 5, 5, -6, ...
$y(n)$ = -5, 8, 7, -6, -6, 5, 4, -7, ...
$y(n)$ = -4, 7, 5, -8, -7, 6, 6, -5, ...

$$
\begin{aligned}
y(n) &= -2,\ 10,\ 7,\ -8,\ -9,\ 3,\ 4,\ -5,\ \ldots\ldots \\
y(n) &= -7,\ 8,\ 9,\ -3,\ -4,\ 5,\ 2,\ -10,\ \ldots\ldots \\
y(n) &= -5,\ 6,\ 4,\ -8,\ -6,\ 7,\ 7,\ -5,\ \ldots\ldots \\
y(n) &= -2,\ 8,\ 4,\ -10,\ -9,\ 5,\ 7,\ -3,\ \ldots\ldots \\
y(n) &= -6,\ 6,\ 5,\ -7,\ \ldots\ldots \\
y(n) &= -5,\ 7,\ 6,\ -6,\ \ldots\ldots
\end{aligned}
\qquad (4.1.21)
$$

So here the system can respond with 8 different types of second-order subharmonics and two types of output signals which have the same period as the input signal. The frequency spectrum of the input signal and of the first in the above list of output signals are shown in Fig. 4.1.2. This way the appearance of a subharmonic component is elucidated.

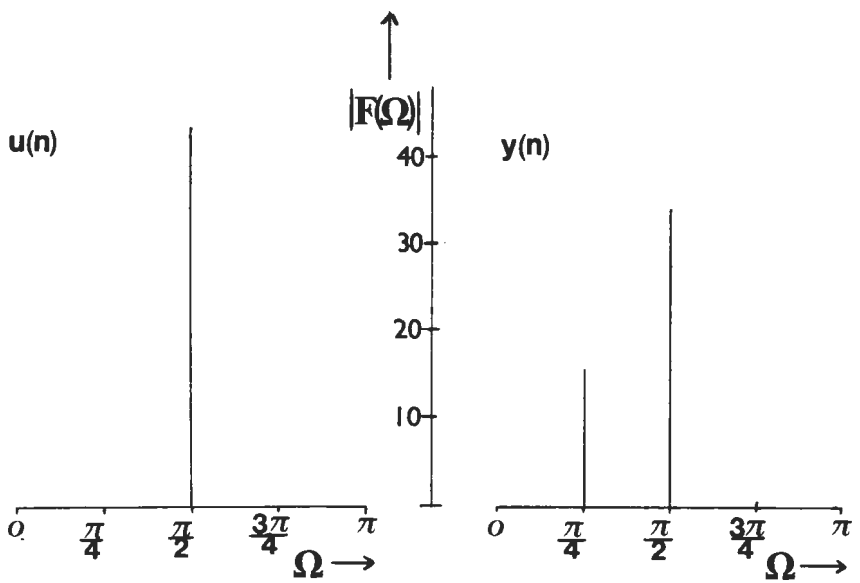


Fig. 4.1.2. Frequency spectrum $|F(\Omega)|$ of the input signal $u(n)$ and the first output signal $y(n)$.

4.2. Regions of filter coefficients for which subharmonics can occur

Let a recursive digital filter be excited by a discrete-time periodic input signal with a period N. The filter can respond with a subharmonic of order S only for certain combinations of filter coefficients. To be more precise, we will now calculate the *region of filter coefficients* which potentially admits a subharmonic of order S for a second-order state-space digital filter with rounding for quantization and excited by an input signal with period N. The conditions thus found only indicate situations favourable for the occurrence of subharmonics and do not predict their actual appearance. On the other hand, if the conditions are not met, subharmonics cannot occur.

The state equations of a general second-order digital filter with rounding are

$$\begin{aligned} x_1(n+1) &= a_{11}\cdot x_1(n) + a_{12}\cdot x_2(n) + b_1\cdot u(n) + \epsilon_1(n) \\ x_2(n+1) &= a_{21}\cdot x_1(n) + a_{22}\cdot x_2(n) + b_2\cdot u(n) + \epsilon_2(n), \end{aligned} \qquad (4.2.1)$$

where $|\epsilon_i(n)| \le 1/2$.

If a periodic state sequence $\underline{x}(n)$ of period $S\cdot N$ is found for some input signal $u(n)$ of period N this sequence has to fulfil the conditions

$$\begin{aligned} &|x_1(n+1) - a_{11}\cdot x_1(n) - a_{12}\cdot x_2(n) - b_1\cdot u(n)| \le \tfrac{1}{2} \\ &|x_2(n+1) - a_{21}\cdot x_2(n) - a_{22}\cdot x_2(n) - b_2\cdot u(n)| \le \tfrac{1}{2} \\ &\text{for all } n \in \{0, 1, 2, \cdots, S\cdot N-1\}, \end{aligned} \qquad (4.2.2)$$

which form a set of $2\cdot N\cdot S$ inequalities.
We define the difference vector $\underline{d}(n,K)$ according to

$$\underline{d}(n,K) = \underline{x}(n) - \underline{x}(n + K\cdot N). \qquad (4.2.3)$$

If there exists a periodic state sequence of period $S\cdot N$ in this filter the signal $\underline{d}(n,K)$ has to fulfil a total number of $2\cdot N\cdot S\cdot(S-1)$ equations:

$$
\begin{aligned}
&|d_1(n+1,K) - a_{11}\cdot d_1(n,K) - a_{12}\cdot d_2(n,K)| \le 1\\
&|d_2(n+1,K) - a_{21}\cdot d_1(n,K) - a_{22}\cdot d_2(n,K)| \le 1\\
&\text{for all} \quad n \in \{0, 1, 2, \cdot\cdot\cdot, S\cdot N\text{-}1\}\\
&\qquad\quad K \in \{1, 2, \cdot\cdot\cdot, S\text{-}1\}. \qquad (4.2.4)
\end{aligned}
$$

Assume that a sequence $\underline{d}(n,K)$ satisfies the conditions of (4.2.4) for all $n \in \{0, 1, \cdot\cdot, S_o\cdot N_o\text{-}1\}$ and $K \in \{1, 2, \cdot\cdot, S_o\text{-}1\}$. Now form the sequence $\underline{d}'(n,K)$ which is a periodic continuation of $\underline{d}(n,K)$ according to

$$\underline{d}'(n+i\cdot S_o\cdot N_o,K) = \underline{d}(n,K) \quad \text{for all } i \in \{0, 1, \cdot\cdot, J\text{-}1\}. \qquad (4.2.5)$$

Then the sequence $\underline{d}'(n,K)$ satisfies the conditions of (4.2.4) for all $n \in \{0, 1, \cdot\cdot, J\cdot S_o\cdot N_o\text{-}1\}$ and $K \in \{1, 2, \cdot\cdot, S_o\text{-}1\}$ for the same region of filter coefficients as $\underline{d}(n,K)$ did. The sequence $\underline{d}'(n,K)$ corresponds with a periodic response of the filter excited with a period $N = J\cdot N_o$ and it is periodic with a period $S_o\cdot J\cdot N_o$. This implies that if J and S_o have no common factor ($\gcd(J,S_o) = 1$) we now have found a subharmonic of order $S = S_o$ for an input sequence with period $N = J\cdot N_o$. So a subharmonic of order S_o, which does appear for an input signal with period N_o, can appear also for some input signals with period $J\cdot N_o$ for the same region of filter coefficients.

But for the latter class of input signals also periodic sequences can occur which did not appear for input signals with period $N = N_o$, resulting in a possible new region of filter coefficients where subharmonics of order S_o can occur. Therefore the region of filter coefficients which can lead to subharmonics of order S_o is for the class of input signals with period $N = J\cdot N_o$ larger than for those with $N = N_o$.

On the other hand, the probability that a special subharmonic $\underline{d}(n)$ fits the complete set of equations (4.2.4) is smaller. This is possibly the reason why subharmonics of high orders are rarely observed in digital filters excited by input signals with large periods N.

As an example, we investigate the second-order direct form digital filter of Fig. 1.2.1 with one rounder, in which

$$d_2(n+1,K) = d_1(n,K) = d(n,K). \tag{4.2.6}$$

A subharmonic state sequence of order S has to fulfil the conditions

$$\begin{aligned} &\left|d(n+1,K) - a\cdot d(n,K) - b\cdot d(n-1,K)\right| \leq 1 \\ &\text{for all} \quad n \in \{0, 1, \cdots, S\cdot N-1\} \\ &\qquad\quad K \in \{1, 2, \cdots, S-1\}. \end{aligned} \tag{4.2.7}$$

A constant-input limit cycle of period 2, ($N = 1$ and $S = 2$), has $d(n+1,K) = -d(n,K)$ for $K = 1$ and so it has to fulfil the condition

$$\begin{aligned} &\left|d(n,K)\right|\cdot(1 + a - b) \leq 1 \\ &\text{for } n \in \{0, 1\} \quad \text{and} \quad K \in \{1\}. \end{aligned} \tag{4.2.8}$$

Since in a limit cycle $\left|d(n,K)\right| \geq 1$ for some value of n and K, this condition can only be fulfilled for filter coefficients satisfying the relation

$$a \leq b. \tag{4.2.9}$$

A constant-input limit cycle of period 3, ($N = 1$ and $S = 3$), has $d(n-1,K) = -d(n,K) - d(n+1,K)$ for $K = 1, 2$, so it has to fulfil the conditions:

$$\begin{aligned} &\left|(1+b)\cdot d(n+1,K) + (b-a)\cdot d(n,K)\right| \leq 1 \\ &\left|(a-b)\cdot d(n+1,K) + (1+a)\cdot d(n,K)\right| \leq 1 \\ &\left|(1+a)\cdot d(n+1,K) + (1+b)\cdot d(n,K)\right| \leq 1. \end{aligned} \tag{4.2.10}$$

These three conditions can only be fulfilled for

$$a \leq 0 \quad \text{and} \quad b \leq 0. \tag{4.2.11}$$

Proof:

There exists some values of n, K for which $d(n+1,K) \geq 0$ and $d(n,K) \geq 0$. If $d(n+1,K) = 0$ then with $d(n,K) \geq 1$ inequality (4.2.10) can only be fulfilled for $a \leq 0$ and $b \leq 0$. If $d(n,K) = 0$ we find the same solution for $d(n+1,K) \geq 1$. If $d(n+1,K) \geq 1$ and $d(n,K) \geq 1$ then for $a > 0$ and $a \geq b$ the second inequality of (4.2.10) is not satisfied, and for $b > 0$ and $b \geq a$ the same holds for the first condition of (4.2.10). So constant-input limit cycles with period $S=3$ can only appear for $a \leq 0$ and $b \leq 0$. □

In the same way the region of filter coefficients for which subharmonics can occur are calculated for other values of N and S. Some of these results are shown in table 4.2.1.

An extensive computer search has been performed for subharmonics in this second-order direct form digital filter. This search has been executed for a grid of filter coefficients a and b over a range of periodic input signals with the same period N. The results are depicted in the a-b-plane, where subharmonics of order S have been found for input signals with period N; cf. Fig. 4.2.1 for

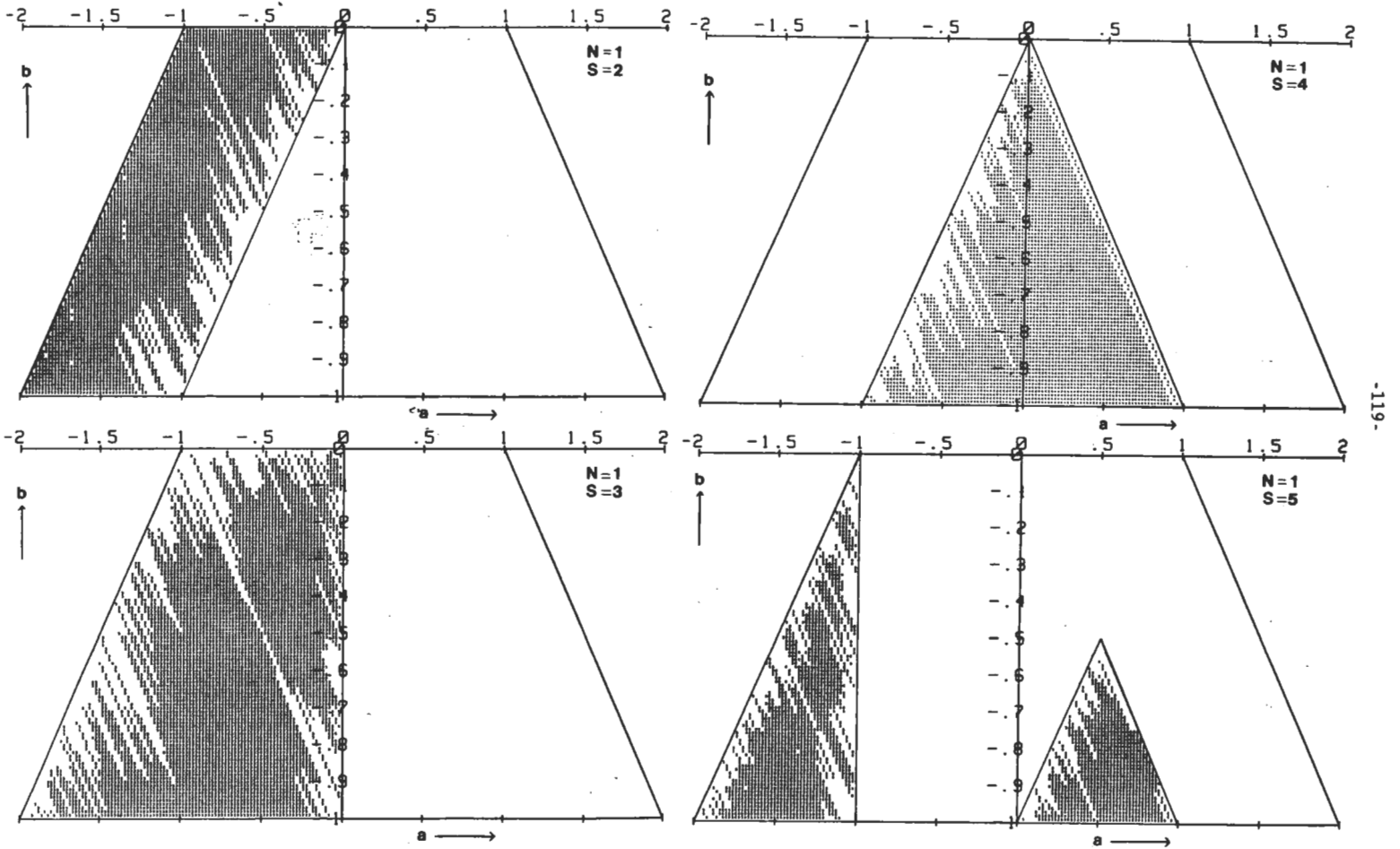
$N = 1$, $S = 2$ to 9
$N = 2$, $S = 2$ to 5
$N = 3$, $S = 2$ to 5

In Fig. 4.2.1 also the theoretical bounds are presented, as calculated in this section. From this figure we conclude that the calculated bounds are rather tight, i.e. in good agreement with the regions where subharmonics actually appear.

N=1,	S = 2,	$a \le b$,
	S = 3,	$a \le 0$ and $b \le 0$,
	S = 4,	$\lvert a \rvert + b \le 0$,
	S = 5,	$1 + b \le a \le -b$,
	or	$a \le -1$,
	S = 6,	$a + b \le 1/2$ and $a \ge 1/2$ and $b \le -1/2$,
	S = 7,	$2 + b \le a \le -b - 1/2$ and $5 \cdot a + 4 \cdot b \le 3$,
	or	$1 + 2 \cdot b \le a \le 1 + b$,
	or	$2 \cdot a - b \le -2$,
	S = 8,	$2 + b \le a \le \frac{2}{3} - b$,
	or	$-\frac{2}{3} + b \le a \le -2 - b$ and $1 + 3 \cdot b \le a \le -\frac{3}{2} - \frac{1}{2} \cdot b$.
N=2,	S = 2,	$\lvert a \rvert \le 1$ and $b \le 0$,
	S = 3,	$b \le 0$,
	S = 4,	$1 \le a \le 1 - b/2$,
	or	$-1 + b/2 \le a \le -1$,
	or	$-\frac{2}{3} + b \le a \le -2 - b$,
	or	$2 + b \le a \le \frac{2}{3} - b$,
	S = 5,	$b \le 0$ and $a \ge 1$,
	or	$b \le 0$ and $a \le -1$,
	or	$b \le a \le -1 - b$,
	or	$1 + b \le a \le -b$.
N=3,	S = 2,	$a \le b$,
	or	$b \le 0$ and $a \ge 0$,
	S = 3,	$1 \le a \le -2b$,
	or	$1 \le a \le 1 - b/2$,
	or	$0 \le a \le -b$,
	or	$a \le -1 + b/2$ and $a \le -\frac{2}{3} + b$.

table 4.2.1. Area of filter coefficients a and b for which there can possibly appear subharmonics of order S for an input signal with period N.

Fig 4.2.1.a Region of filter coefficients a and b where subharmonics have been found for $N = 1$ and $S = 2$ to 5.



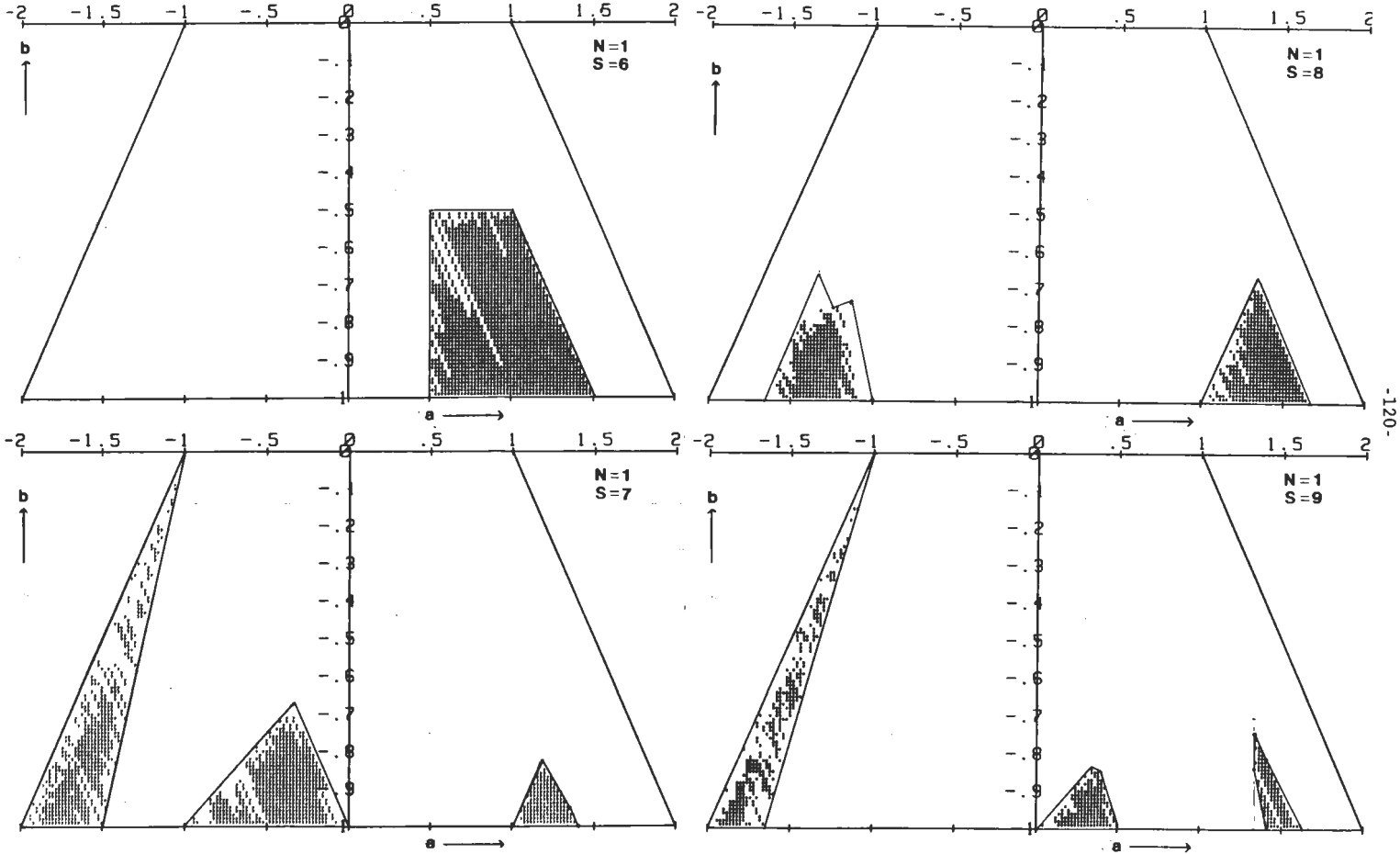
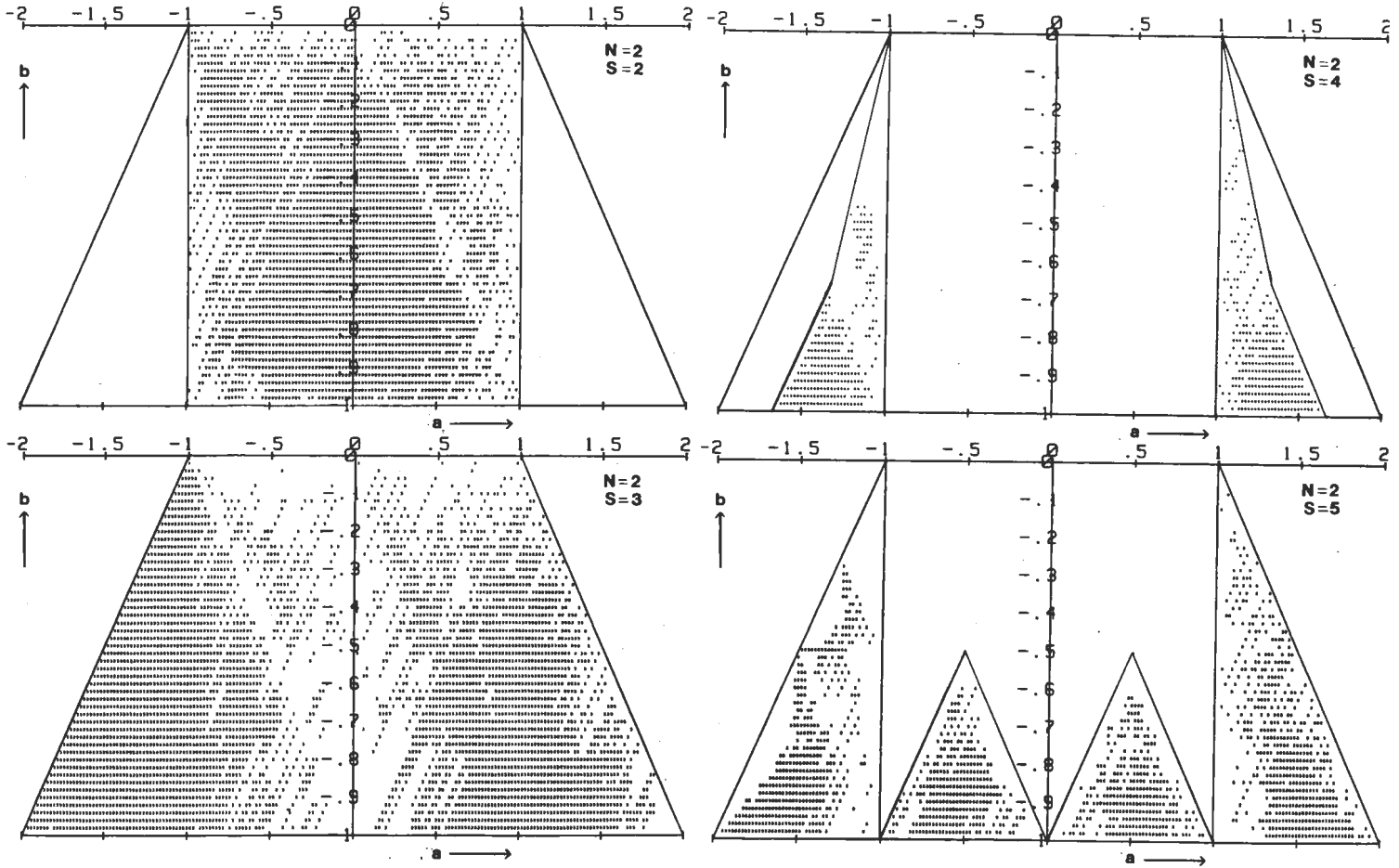


Fig 4.2.1.b Region of filter coefficients a and b where subharmonics have been found for $N = 1$ and $S = 6$ to 9.

Fig 4.2.1.c Region of filter coefficients a and b where subharmonics have been found for $N = 2$ and $S = 2$ to 5.



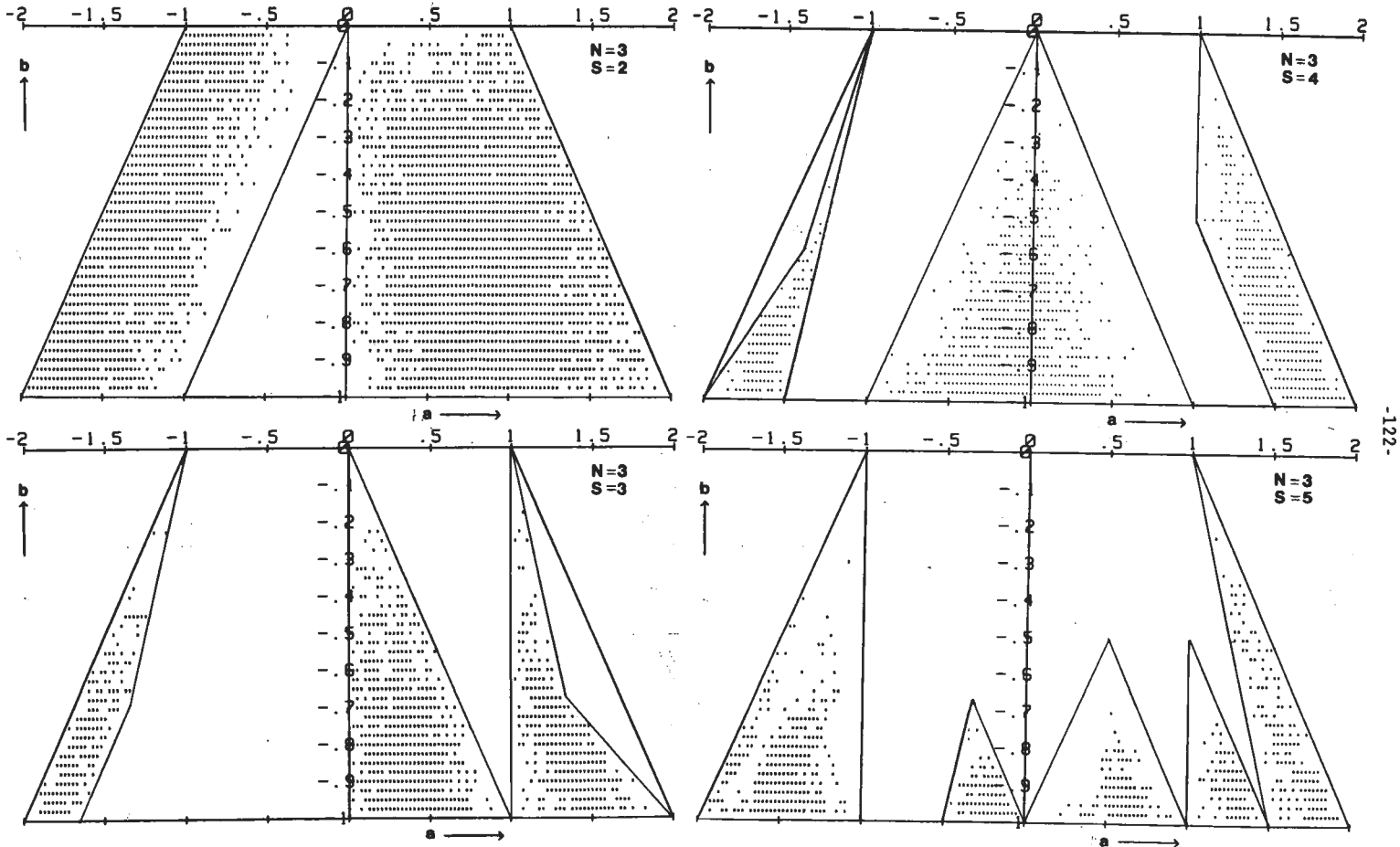


Fig 4.2.1.d Region of filter coefficients a and b where subharmonics have been found for $N = 3$ and $S = 2$ to 5.

4.3. An effective value model for describing subharmonics

In a finite impulse response filter (FIR-filter) of order J the output signal $y(n)$ is formed as a linear combination of the input signal and its delayed versions, according to

$$y(n) = \mathrm{NL}\{ \sum_{i=0}^{J} a_i \cdot u(n-i) \}. \qquad (4.3.1)$$

This implies that if $u(n) = 0$ for $n \geq N_o$ also the output will become zero:

$$y(n) = 0 \qquad \text{for } n \geq N_o + J, \qquad (4.3.2)$$

so no zero-input limit cycle is possible in this filter.
Further, if $u(n)$ is periodic with period N also the output $y(n)$ will become periodic with this period, since

$$y(n+N) = \mathrm{NL}\{ \sum_{i=0}^{J} a_i \cdot u(n+N-i) \}$$

$$= \mathrm{NL}\{ \sum_{i=0}^{J} a_i \cdot u(n-i) \} = y(n). \qquad (4.3.3)$$

So no subharmonic can occur in a FIR-filter. The appearance of limit cycles and subharmonics is restricted to filters with feedback-loops (with time-delay elements), and such a feedback-loop must contain a non-linear element.

It is remarkable that zero-input limit cycles in a digital filter are often nearly sinusoidal, with a frequency close to some resonance frequency of the filter. This gives us the idea that limit cycles are the result of an effective complex-conjugated pair of poles on the unit circle in the z-plane, and, as such, are undamped eigen oscillations of the system. The effective value model of Jackson is based on this hypothesis [155].

Also the subharmonic frequency in the response of a periodically excited digital filter is often close to some resonance frequency of the filter, which provides support to an equivalent effective value model.
In this model the output signal $y(n)$ basically contains the two frequencies $\Omega_{input} = 2\pi/N$ and $\Omega_o = \Omega_{resonance}$ and all their intermodulation products.

In the second-order direct form digital filter of Fig. 1.2.1 with one quantizer we have

$$\begin{aligned} x_1(n+1) &= Q\{a\cdot x_1(n) + b\cdot x_2(n) + u(n)\} \\ &= a\cdot x_1(n) + b\cdot x_2(n) + u(n) + \epsilon(n) \\ x_2(n+1) &= x_1(n) \\ y(n) &= x_1(n+1). \end{aligned} \tag{4.3.4}$$

With $\tilde{\underline{x}}(n)$ denoting the steady state solution of the idealized linear system and $\underline{e}(n) = \underline{x}(n) - \tilde{\underline{x}}(n)$ denoting the error vector, we have

$$\begin{aligned} e_1(n+1) &= a\cdot e_1(n) + b\cdot e_2(n) + \epsilon(n) \\ e_2(n+1) &= e_1(n) \\ y(n) &= NL\{a\cdot e_1(n) + b\cdot e_2(n) + \tilde{x}_1(n+1)\}. \end{aligned} \tag{4.3.5}$$

Applying an effective value model, we get

$$\begin{aligned} e_1(n+1) &= a'\cdot e_1(n) + b'\cdot e_2(n) \\ e_2(n+1) &= e_1(n) \\ y(n) &= NL\{a'\cdot e_1(n) + b'\cdot e_2(n) + \tilde{x}_1(n+1)\}. \end{aligned} \tag{4.3.6}$$

$$\text{So} \quad H'(z) = \frac{z^2}{z^2 - a'\cdot z - b'} \tag{4.3.7}$$

denotes the new transfer function, which has an effective complex-conjugated pole pair $q'_{1,2} = \exp(\pm j\cdot\Omega_o)$ on the unit circle in the z-plane, for

$$a' = 2\cdot\cos(\Omega_o)$$
$$b' = -1. \qquad (4.3.8)$$

The equations of (4.3.6) result in an autonomous oscillation

$$e_1(n) = E\cdot\cos(\Omega_o\cdot n + \varphi), \qquad (4.3.9)$$

where Ω_o is the resonance frequency of the effective value model.

Unfortunately this model is not always correct. For example, for filter coefficients

$$a = 1.602$$
$$b = -0.98, \qquad (4.3.10)$$

resulting in complex conjugated poles

$$q_{1,2} = 0.99\cdot\exp(\pm j\frac{2\pi}{10}), \qquad (4.3.11)$$

the direct form filter excited by

$$u(n) = \sin(\frac{2\pi}{5}n) \qquad (4.3.12)$$

produces a subharmonic of order $S = 6$ with the succeeding values

$$\begin{aligned} y(n) = {} & -9, -8, -5, -1, 3, 7, 9, 7, 1, -5, \\ & -8, -7, -4, 0, 4, 7, 8, 5, -1, -7, \\ & -9, -7, -3, 1, 5, 8, 9, 6, 0, -6, \ldots . \end{aligned} \qquad (4.3.13)$$

In this filter we have $\Omega_o = \frac{2\pi}{10}$ so the effective value model predicts a signal which is periodic with period 10. This way the subharmonic of order $S = 6$ can never be found.

4.4. A linear time-variant system description of subharmonics

In this section we present a description of a *linear time-variant system* in the frequency domain. We investigate the output $y(n)$ of such a system, excited by an input signal $u(n)$, which is assumed to be discrete-time periodic with a certain period N_u. These investigations can be used for the analysis of subharmonics in digital filters, in which we have replaced the nonlinear quantization operator by a linear time-variant multiplicator. It can also be used for the analysis of periodic excited adaptive filters [142, 354, 355].

A periodically excited linear *time-invariant system* is described (according to the theory of Section 1.2) by its "periodic impulse response" $\tilde{h}(n)$ or its discrete transfer function $\tilde{H}(k)$, according to

$$y(n) = \sum_{m=0}^{N_u-1} \tilde{h}(n-m)\cdot u(m) = \sum_{m=0}^{N_u-1} \tilde{h}(m)\cdot u(n-m) \qquad (4.4.1)$$

and

$$Y(k) = \tilde{H}(k)\cdot U(k). \qquad (4.4.2)$$

Now, we will start with the analysis of recursive time-variant systems. The first-order time-variant system is described by the equation

$$y(n) = a(n)\cdot y(n-1) + u(n). \qquad (4.4.3)$$

The filter coefficient a is a function of the time n. We define the *time-variant impulse response* $h_m(n)$ as the signal responding to a Dirac input signal $u(n) = \delta(n-m)$ occurring at the moment m; so

$$y(n) = h_m(n-m) \quad \text{for this input signal.} \qquad (4.4.4)$$

In the first-order recursive time-variant system we have

$$\begin{aligned} h_m(n) &= 0 && \text{for } n < 0 \\ &= 1 && \text{for } n = 0 \\ &= a(m+1)\cdots a(m+n) && \text{for } n > 0, \end{aligned} \tag{4.4.5}$$

which will be denoted by $\prod_{i=1}^{n} a(m+i)$.
The response to an input signal $u(n)$ becomes

$$y(n) = \sum_{m=-\infty}^{\infty} h_m(n-m)\cdot u(m) = \sum_{m=-\infty}^{\infty} h_{n-m}(m)\cdot u(n-m). \tag{4.4.6}$$

It is assumed that the coefficient $a(n)$ is discrete-time periodic with a certain period N_a. For the time being we further assume that N_a and N_u have no common divisor, so the greatest common divisor of N_a and N_u is

$$\gcd(N_a, N_u) = 1. \tag{4.4.7}$$

In a stable filter the output signal $y(n)$ will be periodic with a period

$$N_{tot} = N_a \cdot N_u. \tag{4.4.8}$$

The N_{tot}-point Discrete Fourier Transform $Y(k)$ of the periodic output signal $y(n)$ is defined according to

$$Y(k) = \sum_{n=0}^{N_{tot}-1} y(n)\cdot \exp\left[-j\frac{2\pi kn}{N_{tot}}\right]. \tag{4.4.9}$$

In this section we try to find a relation between the N_{tot}-point DFT $Y(k)$ and the N_u-point DFT $U(k)$. Whereas the input signal $u(n)$ is periodic with period N_u the output signal $y(n)$ will have subharmonic components of the order N_a.

For a periodic input signal $u(n)$ we get a steady-state output signal $y(n)$, just as in Section 1.2, with

$$y(n) = \sum_{m=0}^{N_{tot}-1} \tilde{h}_m(n-m)\cdot u(m) = \sum_{m=0}^{N_{tot}-1} \tilde{h}_{n-m}(m)\cdot u(n-m), \qquad (4.4.10)$$

where

$$\tilde{h}_m(n) = \sum_{i=-\infty}^{\infty} h_m(n + i\cdot N_{tot}). \qquad (4.4.11)$$

The time-variant discrete transfer function $\tilde{H}_m(k)$ is defined according to

$$\tilde{H}_m(k) = \sum_{n=0}^{N_{tot}-1} \tilde{h}_m(n)\cdot \exp\left[-j\frac{2\pi kn}{N_{tot}}\right]. \qquad (4.4.12)$$

The DFT of the output signal can be written in the form

$$Y(k) = \sum_{n=0}^{N_{tot}-1} \sum_{m=0}^{N_{tot}-1} \tilde{h}_{n-m}(m)\cdot u(n-m)\cdot \left[-j\frac{2\pi kn}{N_{tot}}\right]. \qquad (4.4.13)$$

The value of $n-m \pmod{N_{tot}}$ can be replaced by

$$n-m \pmod{N_{tot}} = \alpha\cdot N_a+\beta, \qquad (4.4.14)$$

where

$$0 \le \alpha \le N_u-1$$
$$0 \le \beta \le N_a-1.$$

Since $\tilde{h}_n(m) = \tilde{h}_{n+N_a}(m)$ for all n equation (4.4.13) becomes

$$Y(k) = \sum_{\alpha=0}^{N_u-1} \sum_{\beta=0}^{N_a-1} \sum_{m=0}^{N_{tot}-1} \tilde{h}_\beta(m) \cdot u(\alpha \cdot N_a + \beta) \cdot \exp\left[-j\frac{2\pi k(\alpha \cdot N_a + \beta + m)}{N_{tot}}\right]$$

$$= \sum_{\beta=0}^{N_a-1} \tilde{H}_\beta(k) \cdot \sum_{\alpha=0}^{N_u-1} u(\alpha \cdot N_a + \beta) \cdot \exp\left[-j\frac{2\pi k(\alpha \cdot N_a + \beta)}{N_{tot}}\right]. \qquad (4.4.15)$$

Now we will use the *Chinese Remainder Theorem* (see [232]): if N_a and N_u have no common divisor, so $\gcd(N_a, N_u) = 1$, then there exist a unique set of integer values r and s for which

$$r \cdot N_u + s \cdot N_a = 1 \pmod{N_{tot}} \qquad (4.4.16)$$

and

$$0 \le r \le N_a - 1$$
$$0 \le s \le N_u - 1.$$

The values of r and s can be constructed by *Euclid's algorithm*.

We will substitute this result in the power of the exponent of (4.4.15):

$$\sum_{\alpha=0}^{N_u-1} u(\alpha \cdot N_a + \beta) \cdot \exp\left[-j\frac{2\pi k(\alpha \cdot N_a + \beta)(r \cdot N_u + s \cdot N_a)}{N_{tot}}\right] =$$

$$\sum_{\alpha=0}^{N_u-1} u(\alpha \cdot N_a + \beta) \cdot \exp\left[-j\frac{2\pi ks(\alpha \cdot N_a + \beta)}{N_u}\right] \cdot \exp\left[-j\frac{2\pi k\beta r}{N_a}\right] = U(sk) \cdot \exp\left[-j\frac{2\pi k\beta r}{N_a}\right]. \qquad (4.4.17)$$

So

$$Y(k) = U(sk) \cdot \sum_{m=0}^{N_a-1} \tilde{H}_m(k) \cdot \exp\left[-j\frac{2\pi krm}{N_a}\right]. \qquad (4.4.18)$$

One important conclusion we can draw from this formula is that there is a one-to-one relation between a component of the output spectrum and one of the components of the input spectrum.

In the first-order time-variant filter we have

$$\tilde{h}_m(n) = \frac{\prod_{i=1}^{n} a(m+i)}{1 - \left[\prod_{i=0}^{N_a-1} a(i)\right]^{N_u}} \tag{4.4.19}$$

and

$$\tilde{H}_m(k) = \frac{\sum_{n=0}^{N_a-1} \left[\prod_{i=1}^{n} a(m+i)\right] \cdot \exp\left[-j\frac{2\pi kn}{N_{tot}}\right]}{1 - \left[\prod_{i=0}^{N_a-1} a(i)\right] \cdot \exp\left[-j\cdot\frac{2\pi k}{N_u}\right]}. \tag{4.4.20}$$

Now we will investigate the situation where the greatest common divisor of N_a and N_u is

$$\gcd(N_a, N_u) = g \neq 1. \tag{4.4.21}$$

The Chinese remainder theorem concludes for $g \neq 1$ that

$$r \cdot N_u + s \cdot N_a = g \pmod{N_a \cdot N_u}, \tag{4.4.22}$$

where

$$0 \leq r \leq N_a - 1$$
$$0 \leq s \leq N_u - 1.$$

Equivalent to equation (4.4.18) we can now deduce the equation:

$$Y(k) = \sum_{m=0}^{N_a-1} \tilde{H}_m(k) \cdot \exp\left[-j\frac{2\pi krm}{N_a}\right] \cdot \frac{1}{g} \sum_{i=0}^{g-1} U(sk+\frac{iN_u}{g}) \cdot \exp\left[j\frac{2\pi im}{g}\right]. \tag{4.4.23}$$

So $Y(k)$ is the linear combination of g components of the input spectrum $U(k)$. In this case we have crosstalk of the frequency components.

The subharmonics which occur in digital filters can be described by a timevariant system. But, unfortunately, the multipliers of this system usually vary with a period $N_a = S \cdot N_u$, where S is the order of the subharmonic. This implies that $\gcd(N_a, N_u) = N_u$.
With $r = 1$ and $s = 0$ equation (4.4.23) becomes

$$Y(k) = \sum_{m=0}^{N_a-1} \tilde{H}_m(k) \cdot \exp\left[-j\frac{2\pi km}{N_a}\right] \cdot \frac{1}{\tilde{N}_u} \cdot \sum_{i=0}^{N_u-1} U(i) \cdot \exp\left[j\frac{2\pi mi}{N_u}\right] \tag{4.4.24}$$

So every component $Y(k)$ is a linear combination of all components of the input signal, which makes this analysis rather complicated.

4.5. A new decomposition of discrete-time periodic signals

In this section a *new decomposition* of a discrete-time periodic signal $f(n)$ with period N is presented. This decomposition is characterized by the property that each component $f_k(n)$ is periodic with a particular period k, which is a member of the set V_N of all divisors of N.
This decomposition might be useful for e.g. the analysis of limit cycles. In the context of subharmonics generation in nonlinear systems it can serve to describe the degree of signal distortion, as will be shown in this section.

The classical mathematical tool for the representation of a discrete-time periodic signal $f(n)$ with period N is the Discrete Fourier Transformation:

$$f(n) = \sum_{\ell=1}^{N} f_\ell(n) = \frac{1}{N} \cdot \sum_{\ell=1}^{N} F(\ell) \cdot \exp\left[j\frac{2\pi \ell n}{N}\right], \qquad (4.5.1)$$

where

$$F(\ell) = \sum_{n=1}^{N} f(n) \cdot \exp\left[-j\frac{2\pi \ell n}{N}\right]. \qquad (4.5.2)$$

While all components $f_\ell(n)$ in this representation are repetitive after N time steps, some of the $f_\ell(n)$ may have smaller elementary periods, viz. the divisors of N. Thus a periodic signal can "contain" components with periods smaller than N.

The starting point in the new decomposition of a periodic signal is the general question: "how much" signal with smaller periods can be extracted from the given signal $f(n)$?

With

$$f(n) = \sum_{k \in V_N} f_k(n) \qquad (4.5.3)$$

we start with writing $f(n)$ as a sum of components $f_k(n)$ with periods k that are divisors of N (including 1 and N). The periods k are arranged in ascending order and form the set V_N. As an example we have

$$V_{12} = \{1, 2, 3, 4, 6, 12\}. \qquad (4.5.4)$$

Notice that according to the decomposition (4.5.3), the structure of the signal components $f_k(n)$ is not fixed, as was the case for the discrete Fourier transformation, but depends upon the signal $f(n)$. This follows immediately from the observation that the number of components $f_k(n)$ is smaller than N. This number is extremely small in case N is prime, where we have only $k = 1$ and $k = N$, so $f(n) = f_1(n) + f_N(n)$. It will be shown that $f_1(n)$ is the DC-component of $f(n)$ and so in this case $f_N(n)$ is the AC-component of this signal.

Without further assumptions, the decomposition (4.5.3) is not unique. In accordance with our premise, we therefore require that in a first step $f_{k1}(n) = f_1(n)$ is so chosen that the power of the remaining signal $f(n) - f_1(n)$ is minimized. Then $f_{k2}(n)$, with the next $k = k_2$ out of V_N, is so chosen that the power of the remaining signal $f(n) - f_{k1}(n) - f_{k2}(n)$ is again minimized etc.
Finally, after all but one f_k's have been determined and subtracted, there remains the last $f_k(n)$ whose period equals N. This remainder will be referred to as "*essentially periodic*" with period N, because it does not contain lower period components in the above sense. From the definition it follows that all components $f_k(n)$ are essentially periodic with their own period.

The minimization procedure described here leads to a recursive algorithm. Let $f_{k1}(n)$, $f_{k2}(n)$, . . denote the successive signal components in (4.5.3), and let $\hat{f}_{k_1}(n)$, $\hat{f}_{k_2}(n)$, . . denote the "*remainders*"

$$\hat{f}_{k_i}(n) = f(n) - \sum_{j=1}^{i-1} f_{k_j}(n) = \sum_{j \geq i} f_{k_j}(n), \qquad (4.5.5)$$

where

$$\hat{f}_{k_1}(n) = \hat{f}_1(n) = f(n)$$

and $\hat{f}_{k_{max}}(n) = \hat{f}_N(n) = f_N(n)$.

Then we obtain

$$f_k(n) = \frac{k}{N} \cdot \sum_{m=1}^{N/k} \hat{f}_k(n-km), \qquad (4.5.6)$$

where the index i of k_i has been dropped.
Thus a certain component $f_k(n)$ is determined as the average of the pertinent remainder $\hat{f}_k(n)$ taken at N/k equidistant time instants.

For example we consider the sequence $f(n)$ which is periodic with period $N = 12$

$$f(n) = 4, 4, 2, -2, 4, 5, 4, -8, 4, -6, 0, 1, . \qquad (4.5.7)$$

First $f_1(n)$ is determined as the average of $\hat{f}_1(n) = f(n)$ over all 12 time instants, which yields $f_1(n) = 1$, and leaves the remainder

$$\hat{f}_2(n) = 3, 3, 1, -3, 3, 4, 3, -9, 3, -7, -1, 0, . \qquad (4.5.8)$$

Next $f_2(n)$ is found as the average of $\hat{f}_2(n)$ taken at the time instants $n-2$, $n-4$, . . , $n-12$, which yields the sequence

$$f_2(n) = 2, -2, 2, -2, 2, -2, 2, -2, 2, -2, 2, -2, . \qquad (4.5.9)$$

By continuing this procedure we find

$$
\begin{aligned}
f_3(n) &= -1, -1, 2, -1, -1, 2, -1, -1, 2, -1, -1, 2, \ldots \\
f_4(n) &= 1, 2, -1, -2, 1, 2, -1, -2, 1, 2, -1, -2, \ldots \\
f_6(n) &= 2, 0, -2, -2, 0, 2, 2, 0, -2, -2, 0, 2, \ldots \\
f_{12}(n) &= -1, 4, 0, 4, 1, 0, 1, -4, 0, -4, -1, 0, \ldots
\end{aligned} \qquad (4.5.1)
$$

The proof of (4.5.6) proceeds in three steps. First it is recognized that $f_k(n)$ has period k; then $f_{k_i}(n)$ and $\hat{f}_{k_{i+1}}(n)$ are shown to be orthogonal; finally as required any variation $\delta f_{k_i}(n)$ added to $f_{k_i}(n)$ increases the power of the remaining signal $\hat{f}_{k_{i+1}}(n)$. The details are omitted here.

The number of components into which a discrete-time periodic signal with period N is decomposed in (4.5.3) is equal to the number of divisors of N, including 1 and N.

For $N = \prod_{i=1}^{t} p_i^{q_i}$ with p_i = prime, this number of divisors turns out to be equal to $\prod_{i=1}^{t} (q_i+1)$. (4.5.11)

The proof of this statement can be given by induction.

For $N = p_1^{q_1}$ the divisors of N are p_1^{j} with $0 \le j \le q_1$, so the number of divisors equals (q_1+1).

Let $N' = \prod_{i=1}^{t-1} p_i^{q_i}$ with a total number of divisors $\prod_{i=1}^{t-1} (q_i+1)$, then the new divisors of N are of the form $p_t^{j} \cdot$divisor of N', in which $0 \le j \le q_t$. So the total number of divisors equals $(q_t+1) \cdot \prod_{i=1}^{t-1} (q_i+1)$.

The N-point discrete Fourier transform of a periodic signal $f(n)$ is found by (4.5.1). In the same way, the N-point DFT of component $f_k(n)$ is

$$F_k(\ell) = \sum_{n=1}^{N} f_k(n) \cdot \exp\left[-j\frac{2\pi \ell n}{N}\right]. \tag{4.5.12}$$

According to the algorithm of equation (4.5.6) we obtain

$$\begin{aligned} F_k(\ell) &= \frac{1}{N} \cdot \sum_{n=1}^{N} \sum_{m=1}^{N} \hat{f}_k(n-km) \cdot \exp\left[-j\frac{2\pi \ell n}{N}\right] \\ &= \frac{1}{N} \cdot \sum_{i=1}^{N} \hat{f}_k(i) \cdot \exp\left[-j\frac{2\pi \ell i}{N}\right] \cdot \sum_{m=1}^{N} \exp\left[-j\frac{2\pi \ell km}{N}\right] \\ &= 0 \qquad \text{for} \quad \ell \bmod (N/k) \neq 0 \\ &= \hat{F}_k(\ell) \quad \text{for} \quad \ell \bmod (N/k) = 0. \end{aligned} \tag{4.5.13}$$

Here we have substituted $i = n-km$ and rearranged the summation sequence.

By evaluating equation (4.5.13) in ascending order over all elements k of the set V_N we can conclude that

$$\begin{aligned} F_k(\ell) &= F(\ell) \qquad \text{for} \quad \ell/N \in U_k \\ &= 0 \qquad \text{otherwise}, \end{aligned} \tag{4.5.14}$$

where

$$U_k = \{ \tfrac{i}{k} \mid 1 \leq i \leq k \quad \text{and} \quad \gcd(i, k) = 1\}. \tag{4.5.15}$$

For example

$$\begin{aligned} U_1 &= \{1\} & U_2 &= \{\tfrac{1}{2}\} \\ U_3 &= \{\tfrac{1}{3}, \tfrac{2}{3}\} & U_4 &= \{\tfrac{1}{4}, \tfrac{3}{4}\} \\ U_6 &= \{\tfrac{1}{6}, \tfrac{5}{6}\} & U_{12} &= \{\tfrac{1}{12}, \tfrac{5}{12}, \tfrac{7}{12}, \tfrac{11}{12}\} \end{aligned} \tag{4.5.16}$$

□

Equation (4.5.14) determines how the values of the N-point discrete Fourier transform $F(\ell)$ are distributed over the DFT of the components. This distribution can be visualized on the unit circle (see Fig. 4.5.1).

For a signal with period $N = 12$, we get

$$
\begin{aligned}
&F_1(12) = F(12),\\
&F_2(6) = F(6),\\
&F_3(4) = F(4) \text{ and } F_3(8) = F(8),\\
&F_4(3) = F(3) \text{ and } F_4(9) = F(9),\\
&F_6(2) = F(2) \text{ and } F_6(10) = F(10),\\
&F_{12}(1) = F(1),\ F_{12}(5) = F(5),\ F_{12}(7) = F(7) \text{ and } F_{12}(11) = F(11),\\
&\text{and all other values of the DFT-components are zero.} \qquad (4.5.17)
\end{aligned}
$$

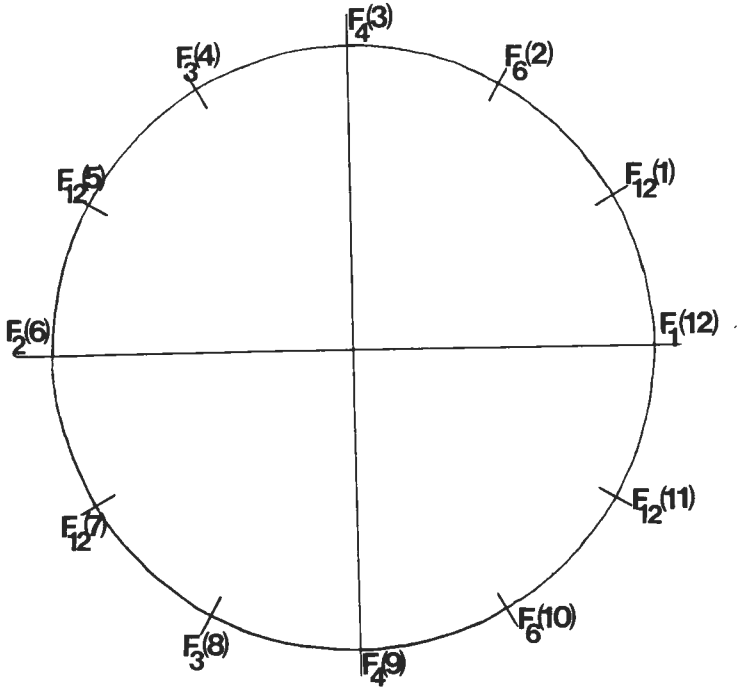


Fig. 4.5.1. Distribution of the DFT-values over the components for N=12.

The component $f_k(n)$ can be calcualted by the inverse discrete Fourier transformation of $F_k(\ell)$ resulting in

$$f_k(n) = \frac{1}{\tilde{N}} \cdot \sum_{\ell=1}^{N} F_k(\ell) \cdot \exp\left[j\frac{2\pi\ell n}{N}\right]$$

$$= \frac{1}{\tilde{N}} \cdot \sum_{\frac{\ell}{\tilde{N}} \in U_k} F(\ell) \cdot \exp\left[j\frac{2\pi\ell n}{N}\right]. \qquad (4.5.18)$$

For example

$$f_{12}(n) = \frac{1}{12} \cdot \left[F(1) \cdot e^{j\frac{2\pi n}{12}} + F(5) \cdot e^{j\frac{10\pi n}{12}} + F(7) \cdot e^{-j\frac{10\pi n}{12}} + F(11) \cdot e^{-j\frac{2\pi n}{12}}\right] \qquad (4.5.19)$$

□

Here we see that each component $f_k(n)$ is the combination of some appropriate terms of the discrete Fourier representation of the original signal. In fact the components $f_k(n)$ form a symbolic calculation of the discrete Fourier transformation to a different basis (see [18, 44, 230, 232]). It is important to recognize that the set of functions $f_k(n)$ are mutual orthogonal which implies the minimization of the remaining signal powers.

Every component $f_k(n)$ can be derived from the original signal by weighting the values of $f(n)$ with proper coefficients, according to

$$f_k(n) = \frac{1}{\tilde{N}} \cdot \sum_{i=1}^{N} a_{ki} \cdot f(n-i). \qquad (4.5.20)$$

The coefficients a_{ki} can be derived from the algorithm (4.5.6).
They can however also be derived from (4.5.18).

$$f_k(n) = \frac{1}{N} \cdot \sum_{\frac{\ell}{N} \in U_k} \sum_{m=1}^{N} f(m) \cdot \exp\left[-j\frac{2\pi\ell m}{N}\right] \cdot \exp\left[j\frac{2\pi\ell n}{N}\right]$$

$$= \frac{1}{N} \cdot \sum_{i=1}^{N} \sum_{q \in U_k} f(n-i) \cdot \exp\left[j2\pi qi\right], \qquad (4.5.21)$$

so

$$a_{ki} = \sum_{q \in U_k} \exp[j2\pi qi]. \qquad (4.5.22)$$

Obviously the coefficients a_{ki} are independent of N. Further they have integer values only, as can be concluded from the algorithm (4.5.6). Therefore the hardware realization of this decomposition is very simple and the components can be calculated very fast and in full precision.

For example

$$\begin{aligned}
a_{1i} &= 1 && \text{for all } i \\
a_{2i} &= (-1)^i && \text{for all } i \\
a_{3i} &= 2 && \text{for } i \bmod 3 = 0 \\
&= -1 && \text{for } i \bmod 3 \neq 0 \\
a_{4i} &= 2 && \text{for } i \bmod 4 = 0 \\
&= -2 && \text{for } i \bmod 4 = 2 \\
&= 0 && \text{otherwise.}
\end{aligned} \qquad (4.5.23)$$

□

We now propose a recursive structure which realizes the components $f_k(n)$. To this end, the discrete Fourier transform of (4.5.14) is extended to the entire complex z-plane. This way a transfer function from $F(z)$ to $F_k(z)$ is determined.

We define

$$F_k(z) = \sum_{n=1}^{N} f_k(n) \cdot z^{-n}. \qquad (4.5.24)$$

From equation (4.5.14) we can conclude that

$$F_k(z) = F(z) \quad \text{for } z = \exp\left[-j\frac{2\pi\ell}{N}\right] \text{ and } \frac{\ell}{N} \in U_k$$

$$= 0 \quad \text{for } z = \exp\left[-j\frac{2\pi\ell}{N}\right] \text{ and } \frac{\ell}{N} \notin U_k. \qquad (4.5.25)$$

The points $z = \exp\left[-j\frac{2\pi\ell}{N}\right]$ and $\frac{\ell}{N} \in U_k$ form the roots of cyclotomic polynomial $Q_k(z)$ of order k (see Appendix IV). The points where $F_k(z) = 0$, according to (4.5.25), are the roots of the polynomial $(z^N-1)/Q_k(z)$. Extending (4.5.25) to the entire complex z-plane, we obtain

$$F_k(z) = \frac{\frac{d}{dz}\{Q_k(z)\}}{Q_k(z)} \cdot \frac{z^N - 1}{N \cdot z^{N-1}} \cdot F(z). \qquad (4.5.26)$$

Equation (4.5.26) describes a transfer function which can be realized as shown in Fig. 4.5.2, where the transfer function within the block G_k is

$$G_k(z) = \frac{z \cdot \frac{d}{dz}\{Q_k(z^{-1})\}}{Q_k(z^{-1})}. \qquad (4.5.27)$$

In Fig. 4.5.2 we recognize the subharmonic-free filter of Section 3.6, where the feedback multipliers γ_1 and γ_2 have the value $-1/N$.

The cyclotomic polynomial $Q_k(z)$ has integer coefficients only, as has been proved in Appendix IV. So the blocks G_k in Fig. 4.5.2 can be realized with integer multipliers. This means that the signals can be calculated extremely fast and represented in full precision: no quantization is needed in this circuit.
Moreover the coefficients of the lower-order polynomials $Q_k(z)$ ($k < 105$) have only the values -1, 0, or 1. In this case the multiplier hardware reduces to pure adders.

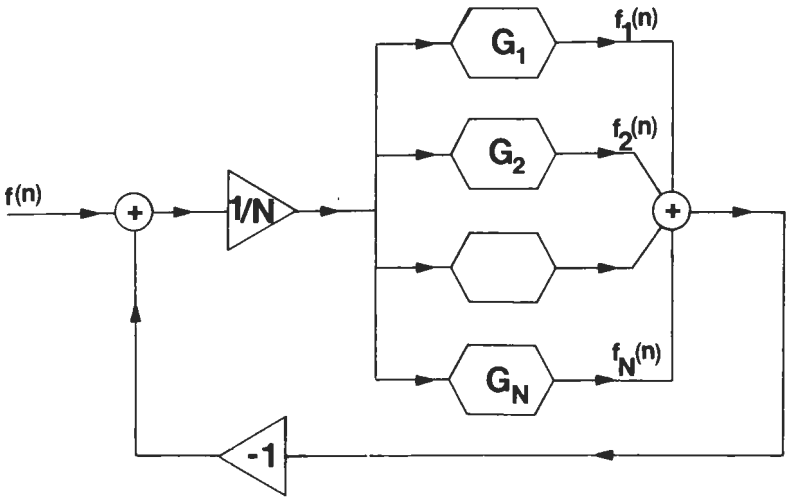


Fig. 4.5.2. Recursive structure for realizing the new decomposition.

Now we have realized two methods to decompose a discrete-time periodic signal with period N in components $f_k(n)$, which are periodic with a particular period k. Both realizations have a very simple hardware implementation with integer multipliers only. The original signal $f(n)$ is synthesized from the components $f_k(n)$ by a simple addition of all components.

The decomposition as described in this section is useful for the quantitative description of subharmonics generated in a digital filter.
In fact, it yields a unique measure for the energy content of such subharmonics. Here we split the periodic sequence $f(n)$, with period $S \cdot N$ in a harmonic and a subharmonic part [64].

The harmonic part $f_{har}(n)$ is formed by all components, which are periodic with a period, which is a divisor of N.
So

$$f_{har}(n) = \sum_{k \in V_N} f_k(n). \tag{4.5.28}$$

The subharmonic part $f_{sub}(n)$ contains all other components of $f(n)$;

$$f_{sub}(n) = f(n) - f_{har}(n). \tag{4.5.29}$$

According to (4.5.18) the harmonic part can be calculated by

$$f_{har}(n) = \frac{1}{S} \cdot \sum_{m=1}^{S} f(n-m \cdot N). \tag{4.5.30}$$

For example if the signal $f(n)$ of (4.5.7) is the output of a digital filter excited by an input signal with period $N = 4$, we have

$$f_{har}(n) = 4, \quad 1, \quad 2, \quad -3, \quad 4, \quad 1, \quad 2, \quad -3, \quad 4, \quad 1, \quad 2, \quad -3, \quad . \; . \tag{4.5.31}$$

and

$$f_{sub}(n) = 0, \quad 3, \quad 0, \quad 1, \quad 0, \quad 4, \quad 2, \quad -5, \quad 0, \quad -7, \quad -2, \quad 4, \quad . \; . \tag{4.5.32}$$

□

5. Final Conclusions

In this thesis we have developed various methods to suppress zero-input and forced-response parasitic oscillations, which occur due to overflow and quantization in recursive digital filters. The stability requirements have been investigated in first-order, second-order, direct-form, wave, normal and state-space digital filters. Freedom from parasitic oscillations is proved with the second method of Lyapunov. This method starts with a properly chosen quadratic energy function, which is positive definite and time-decreasing without wordlength reduction. If, moreover, wordlength reduction lowers the energy for all possible states this energy function is a Lyapunov function, which guarantees freedom from parasitic oscillations.

In this thesis the effects of quantization and overflow are treated independently. The pertinent results will first be discussed for overflow correction.

The first-order recursive digital filter is zero-input stable for all overflow characteristics and forced-response stable only if saturation is used for overflow correction.

The state in an autonomous second-order digital filter describes an ellipse-like curve spiralling towards the origin of the state-plane. Guided by this linear state motion an energy function can be defined, which is a natural candidate for a Lyapunov function. The forced-response stability problem can be transformed into a zero-input problem with time-varying nonlinearities.

The second-order direct form filter is zero-input stable if saturation is used. Two novel proofs using Lyapunov theory are presented; one proof applies only for complex conjugated poles, based on the previously mentioned natural candidate for a Lyapunov function, the other proof is valid for all pairs a and b in the stability triangle. The forced-response of the second-order direct form filter is only guaranteed to be free from overflow oscillations if saturation is used and the filter coefficients a and b satisfy the condition

$$|a| + |b| < 1. \tag{5.1}$$

Higher-order direct form digital filters are in general unstable with respect to overflow.
The wave and normal digital filters are zero-input stable for all overflow characteristics and forced-response stable with a saturation characteristic. This property is shared by all second-order state-space digital filters, whose system matrix A fulfils the condition

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A]. \qquad (5.2)$$

Zero-input stability in a second-order state-space digital filter with saturation has been proved for a system matrix A, which fulfils the condition

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A] + 2 \cdot \min(|a_{12}|, |a_{21}|). \qquad (5.3)$$

Next, we discuss the results for wordlength reduction with quantization. The first-order recursive digital filter is zero-input stable if magnitude truncation (MT) is used for quantization. It can have limit cycles of period 1 or 2 if rounding (RO) or value truncation (VT) is used. In the forced-response, this filter with RO, MT or VT can generate subharmonics of order $S = 2$, which can occur for an input signal with a odd period N and a negative filter coefficient a.
The second-order direct form digital filter of type I is free from zero-input limit cycles, with the exception of those with period 1 or 2, if controlled rounding is used for quantization. We have visualized this statement in the state-plane, which was formerly proved with the Lyapunov theory. For the second-order direct form digital filter of type II no controlled rounding mechanism can be devised to suppress zero-input limit cycles.
The wave digital filters are zero-input stable if magnitude truncation is applied to the individual state variables. Under any constant-input condition a controlled rounding mechanism provides freedom of parasitic oscillations. This statement has also been proved for periodic input signals with $N = 2$ in the wave digital filter of type II. But for input signals with a period $N \geq 3$ the stabilization mechanism fails and subharmonics have been found in our simulations.

A new type of digital filter structure has been introduced, which is free from subharmonics for all periodic input signals with a period $N = M$ or $N = 2M$. This filter structure forms an extension of wave digital filter II. The price to pay for the stabilization is a total number of $2M$ additional time-delay elements and two threshold detectors, the latter to make the parasitic oscillations invisible in the output signal. For input signals with other periods the occurrence of subharmonics cannot be excluded. For example, the appearance of constant-input limit cycles is possible in this filter for $M \neq 1$.
Using the theory of cyclotomic polynomials the previously mentioned subharmonic-free filter was changed in order to make it free from subharmonics for all input signals which are periodic with a period N, being a divisor of $2M$. The basic idea of this filter is to separate the periodic state sequence, which appears in the subharmonic-free filter of period $2M$, in components which are periodic with a divisor of $2M$. Subharmonic components can then be suppressed by a threshold detector.
The second-order state-space digital filter is zero-input stable if MT is used and the system satisfies the condition (5.2). Concerning stability with respect to constant inputs a general principle was mentioned to convert a stable autonomous state-space filter into a system with input terminals stable under any constant excitation. This principle has been extended to form a subharmonic-free filter for discrete-time periodic input signals with a given period N. The suppression of subharmonics requires a total number of N-1 extra time-delay elements. For a pure sinusoidal input signal this can be achieved with only 1 extra multiplier and 1 time-delay element.

Suppression of subharmonics in digital filters for all possible periodic input signals appears to be impossible. Therefore these subharmonics have been analysed in more detail. A computer program has been developed, performing a search for all possible subharmonics in a given filter structure. The region of filter coefficients for which subharmonics can occur has been calculated. These results are in good agreement with an extensive computer search for subharmonics over a grid of filter coefficients and over a range of periodic input signals with the same period N.

Subharmonics in recursive digital filters can be described by an effective value model. Unfortunately this model is not always correct.
They can also be described by a linear time-variant system. Using a time-variant impulse response and the Chinese remainder theorem we have demonstrated that if the period of the input signal and the period of the variation of the filter coefficients have no common factor then a component of the output spectrum is related with only one of the components of the input spectrum. But the multipliers in a digital filter with subharmonics usually vary with a period which is a multiple of that of the input signal, so every component of the output spectrum is a linear combination of all components of the input spectrum, which makes this analysis rather difficult.
A new decomposition of a discrete-time periodic signal $f(n)$, with period N has been presented. This decomposition is characterized by the property that each component $f_k(n)$ is periodic with a particular perod k, which is a member of the set of all divisors of N. All these components are "essentially periodic", because they do not contain energy of lower period. The components can be generated by a recursive algorithm, but can also be found by a combination of some appropriate terms of the discrete Fourier representation of the original signal. In fact, the components form a symbolic calculation of the discrete Fourier transform to a different basis. The hardware realization of this decomposition can be implemented directly with integer multipliers or by a recursive structure, where the coefficients usually have the values -1, 0, or 1 only. This decomposition is useful for the analysis of subharmonics in recursive digital filters.

Appendix I

If energy matrix P is diagonal all overflow characteristics are allowed without risk of overflow oscillations. In this appendix we solve the question: Which second-order A matrices admit a Lyapunov function with a *diagonal* "energy matrix" so that (1.4.3) passes into

$$E(n) = \underline{x}^T(n)\cdot D\cdot\underline{x}(n), \tag{I.1}$$

where D is a positive diagonal matrix? This question has been solved in [237, 238], and is presented here before solving the more difficult question: Which second-order A matrices belong to a state-space digital filter which is guaranteed to be free from overflow oscillations if saturation is used for overflow correction? (see Appendix II)

The energy in the system without overflow correction must strictly decrease, which is satisfied if

$$D - A^T\cdot D\cdot A \text{ is positive definite} \tag{I.2}$$

This condition is equivalent with the two inequalities:

$$\mathrm{Det}[D - A^T\cdot D\cdot A] > 0 \tag{I.3}$$

and

$$\mathrm{Tr}[D - A^T\cdot D\cdot A] > 0. \tag{I.4}$$

So, there must exist some pair of positive values d_1 and d_2 for which

$$-a_{12}^2 d_1^2 - a_{21}^2 d_2^2 + \left[1 + \mathrm{Det}^2[A] - a_{11}^2 - a_{22}^2\right]\cdot d_1 d_2 > 0 \tag{I.5}$$

and

$$\left[1 - a_{11}^2 - a_{12}^2\right]\cdot d_1 + \left[1 - a_{21}^2 - a_{22}^2\right]\cdot d_2 > 0. \tag{I.6}$$

Equation (I.5) can be satisfied for some pair of real values d_1 and d_2 if the discriminant of the quadratic function is positive:

$$\left[1 + \mathrm{Det}^2[A] - a_{11}^2 - a_{22}^2\right]^2 \geq 4 \cdot a_{12}^2 a_{21}^2, \qquad \text{(I.7)}$$

or

$$\left[(1+\mathrm{Det}[A])^2-(a_{11}+a_{22})^2\right]\cdot\left[(1-\mathrm{Det}[A])^2-(a_{11}-a_{22})^2\right] \geq 0. \qquad \text{(I.8)}$$

The first factor is positive due to the stability condition

$$|a_{11}+a_{22}| < 1 + \mathrm{Det}[A]. \qquad \text{(I.9)}$$

So there remains the condition

$$|a_{11}-a_{22}| < 1 - \mathrm{Det}[A]. \qquad \text{(I.10)}$$

It should be noted that according to (I.9) and (I.10) we conclude that

$$|a_{11}| < 1 \quad \text{and} \quad |a_{22}| < 1. \qquad \text{(I.11)}$$

One pair of positive values d_1 and d_2 that satisfies (I.5) and (I.6) for condition (I.10) is:

$$d_1 = |a_{21}| \quad \text{and} \quad d_2 = |a_{12}|. \qquad \text{(I.12)}$$

We shall check this now.

$$\mathrm{Det}[D - A^T \cdot D \cdot A] = |a_{12}\cdot a_{21}|\cdot\left[1+\mathrm{Det}^2[A]-a_{11}^2-a_{22}^2-2\cdot|a_{12}\cdot a_{21}|\right]$$

$$> 0. \qquad \text{(I.13)}$$

The latter inequality is a consequence of (I.7).

$$\mathrm{Tr}[D - A^T.D.A] = |a_{21}|\cdot\left[1 - a_{11}^2 - |a_{12}\cdot a_{21}|\right]$$

$$+ |a_{12}|\cdot\left[1 - a_{22}^2 - |a_{12}\cdot a_{21}|\right] \qquad \text{(I.14)}$$

For $a_{12} \cdot a_{21} \geq 0$ condition (I.14) is satisfied if

$$\begin{aligned} 1 + \mathrm{Det}[A] &> a_{11} \cdot (a_{11} + a_{22}) \\ \text{and} \quad 1 + \mathrm{Det}[A] &> a_{22} \cdot (a_{11} + a_{22}), \end{aligned} \tag{I.15}$$

which is correct according to (I.9) and (I.11).
For $a_{12} \cdot a_{21} < 0$ condition (I.14) is satisfied if

$$\begin{aligned} 1 - \mathrm{Det}[A] &> a_{11} \cdot (a_{11} - a_{22}) \\ \text{and} \quad 1 - \mathrm{Det}[A] &> a_{22} \cdot (a_{22} - a_{11}), \end{aligned} \tag{I.16}$$

which is correct according to (I.10) and (I.11).

We conclude that a second-order state-space filter satisfying (I.10) admits a Lyapunov function with a diagonal energy matrix.

Appendix II

In this appendix we first prove theorem (2.6.7):

If

$$E(n) = \underline{x}^T(n)\cdot P\cdot \underline{x}(n), \tag{II.1}$$

satisfies all conditions for an energy function (see Section 1.4) and if

$$|p_{12}| \leq p_{11} \quad \text{and} \quad |p_{12}| \leq p_{22}, \tag{II.2}$$

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

Proof:

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

$$E_o(n)-E(n) =$$

$$= p_{11}\cdot\left[x_{o1}^2-x_1^2\right] + p_{22}\cdot\left[x_{o2}^2-x_2^2\right] + 2p_{12}\cdot\left[x_{o1}\cdot x_{o2} - x_1\cdot x_2\right]$$

$$\geq p_{11}\cdot\left[x_{o1}^2-x_1^2\right]+p_{22}\cdot\left[x_{o2}^2-x_2^2\right]-2|p_{12}|\cdot\left[|x_{o1}|\cdot|x_{o2}|-|x_1|\cdot|x_2|\right]$$

$$= \left[p_{11}-|p_{12}|\right]\cdot\left[x_{o1}^2-x_1^2\right] + \left[p_{22}-|p_{12}|\right]\cdot\left[x_{o2}^2-x_2^2\right] +$$

$$+ |p_{12}|\cdot\left[[|x_{o1}|-|x_{o2}|]^2-[|x_1|-|x_2|]^2\right]. \tag{II.3}$$

The inequality is valid due to $\mathrm{sgn}(x_{o1})=\mathrm{sgn}(x_1)$ and $\mathrm{sgn}(x_{o2})=\mathrm{sgn}(x_2)$. If only one of the components overflows, f.e. $|x_{o1}| > 1$ then

$$E_o(n)-E(n) \geq \left[p_{11}-|p_{12}|\right]\cdot\left[x_{o1}^2-1\right] + |p_{12}|\cdot\left[[|x_{o1}|-x_2]^2-[1-x_2]^2\right]$$

$$\geq 0. \qquad \text{(II.4)}$$

If both components overflow then

$$E_o(n)-E(n) \geq \left[p_{11}-|p_{12}|\right]\cdot\left[x_{o1}^2-1\right] + \left[p_{22}-|p_{12}|\right]\cdot\left[x_{o2}^2-1\right]$$

$$+ |p_{12}|\cdot[|x_{o1}| - |x_{o2}|]^2 \qquad \geq 0. \qquad \text{(II.5)}$$

So in a system with an energy function $E(n)$ satisfying (II.2) saturation causes an additional energy reduction.

□

Next we solve the question: Which second-order A matrices possess at least one energy function $E(n)$ with satisfies (II.2), for which $E(n)$ is also a Lyapunov function of the system under consideration?

Therefore the energy in the idealized system, without overflow correction, must strictly decrease, which will be satisfied if

$$P - A^T\cdot P\cdot A \text{ is positive definite.} \qquad \text{(II.6)}$$

This condition is for a second-order system equivalent with the inequalities

$$\text{Det}[P - A^T\cdot P\cdot A] > 0 \qquad \text{(II.7)}$$

and

$$\text{Tr}[P - A^T\cdot P\cdot A] > 0. \qquad \text{(II.8)}$$

Expression (II.7) is valid if there exists some real values p_{11}, p_{12} and p_{22} with

$$a \cdot p_{11}^2 + b \cdot p_{11}p_{22} + c \cdot p_{22}^2 + d \cdot p_{11}p_{12} + e \cdot p_{12}p_{22} + f \cdot p_{12}^2 < 0 \qquad \text{(II.9)}$$

where

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

Equation (II.9) can be written in the form

$$a \cdot (p_{11} - \alpha \cdot p_{12})^2 + b \cdot (p_{11} - \alpha \cdot p_{12}) \cdot (p_{22} - \beta \cdot p_{12}) + c \cdot (p_{22} - \beta \cdot p_{12})^2 < K \cdot p_{12}^2 \qquad \text{(II.10)}$$

where

$$\begin{aligned} \alpha &= \frac{-2 \cdot a_{21} \cdot (a_{11} - a_{22})}{(a_{11} - a_{22})^2 - (1 - \mathrm{Det}[A])^2} \\ \beta &= \frac{2 \cdot a_{12} \cdot (a_{11} - a_{22})}{(a_{11} - a_{22})^2 - (1 - \mathrm{Det}[A])^2} \\ K &= \frac{(1 - \mathrm{Det}[A])^2 \cdot \{(1 + \mathrm{Det}[A])^2 - (a_{11} + a_{22})^2\}}{(a_{11} - a_{22})^2 - (1 - \mathrm{Det}[A])^2} \end{aligned}$$

The discriminant of this quadratic form is

$$b^2 - 4ac = -\left[(1 + \mathrm{Det}[A])^2 - (a_{11} + a_{22})^2\right] \cdot \left[(a_{11} - a_{22})^2 - (1 - \mathrm{Det}[A])^2\right] \qquad \text{(II.11)}$$

For $p_{12} = 0$ equation (II.10) is equivalent with (2.6.4), which has a solution with $p_{11} > 0$ and $p_{22} > 0$ for a system matrix A satisfying the condition

$$|a_{11} - a_{22}| < 1 - \mathrm{Det}[A]. \qquad \text{(II.12)}$$

If condition (II.12) is <u>not</u> satisfied, which can only be true for filters with $a_{12}\cdot a_{21} < 0$, we have according to (II.11) a discriminant which is negative, with as result that (II.10) describes the inner part of an elliptical curve in the p_{11}-p_{22}-plane, which is non-empty, since $K > 0$.

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 $|a_{12}| < |a_{21}|$, we have $|\beta| < |\alpha|$, so the centre point of the ellipse lies underneath the 45^{o}-axes in the p_{11}-p_{22}-plane (see Fig. II.1)

If $|\beta| \geq 1$ then the centre point itself satisfies condition (II.2) and therefore we have found a suitable energy function which is a Lyapunov function of the system. This condition $|\beta| \geq 1$ is fulfilled for

$$|a_{11}-a_{22}| \ < \ |a_{12}| + \sqrt{|a_{12}|^2 + (1-\mathrm{Det}[A])^2}. \qquad \text{(II.13)}$$

If $|\beta| < 1$ there only exists some point (p_{11},p_{22}) within the ellipse (II.10) satisfying condition (II.2) if the ellipse curves the line $p_{22} = |p_{12}|$.
There exists such a point of intersection if

$$a\cdot p_{11}^{\,2}+[b\cdot \mathrm{sgn}(p_{12})+d]\cdot p_{11}p_{12}+[c+e\cdot \mathrm{sgn}(p_{12})+f]\cdot p_{12}^{\,2} < 0 \qquad \text{(II.14)}$$

has real solutions. So

$$[b\cdot \mathrm{sgn}(p_{12})+d]^2 \ \geq \ 4\cdot a\cdot[c+e\cdot \mathrm{sgn}(p_{12})+f], \qquad \text{(II.15)}$$

or

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

$$\left[(1-\mathrm{Det}[A])^2-(a_{11}-a_{22}-2\cdot a_{12}\cdot \mathrm{sgn}(p_{12}))^2\right] \geq 0. \qquad \text{(II.16)}$$

The first factor of (II.16) is positive due to stability condition (1.2.6). The second factor shows the remaining condition which can be written in the form

$$|a_{11}-a_{22}| \;<\; 2\cdot|a_{12}| + 1 - \mathrm{Det}[A]. \qquad \text{(II.17)}$$

In the points of intersection, we have $p_{11} > |p_{12}|$ so that (II.2) is satisfied and we have found a suitable energy function which is a Lyapunov function of the system.
All filters with a system matrix A, whose coefficients satisfy equation (II.13) also satisfy (II.17), so the latter one is sufficient. Moreover, all matrices P that satisfy (II.16) and therefore (II.7) does also satisfy (II.8).
For example, one of such a matrix P has the values

$$\begin{aligned} p_{11} &= -[b\cdot\mathrm{sgn}(p_{12})+d] \\ &= 1 + \mathrm{Det}^2[A] + 2\cdot|a_{12}|\cdot|a_{11}-a_{22}| - a_{11}^{\,2} - a_{22}^{\,2} \\ p_{12} &= 2\cdot a_{12}^{\,2} \\ p_{22} &= 2\cdot a_{12}^{\,2} \end{aligned} \qquad \text{(II.18)}$$

Substitution of these values in (II.7) and (II.8) shows that the results are correct.

The case $|a_{21}| < |a_{12}|$ is equivalent to the previous case resulting in the condition

$$|a_{11}-a_{22}| \;<\; 2\cdot|a_{21}| + 1 - \mathrm{Det}[A]. \qquad \text{(II.19)}$$

The conditions (II.17) and (II.19) form together the result of this appendix: all second-order state-space digital filters with a system matrix A satisfying:

$$|a_{11}-a_{22}| \leq 2\cdot\min(|a_{12}|,|a_{21}|) + 1 - \mathrm{Det}[A], \qquad \text{(II.20)}$$

possess at least one energy function $E(n)$, with a matrix P satsifying (II.5), which is a Lyapunov function of the system under consideration.

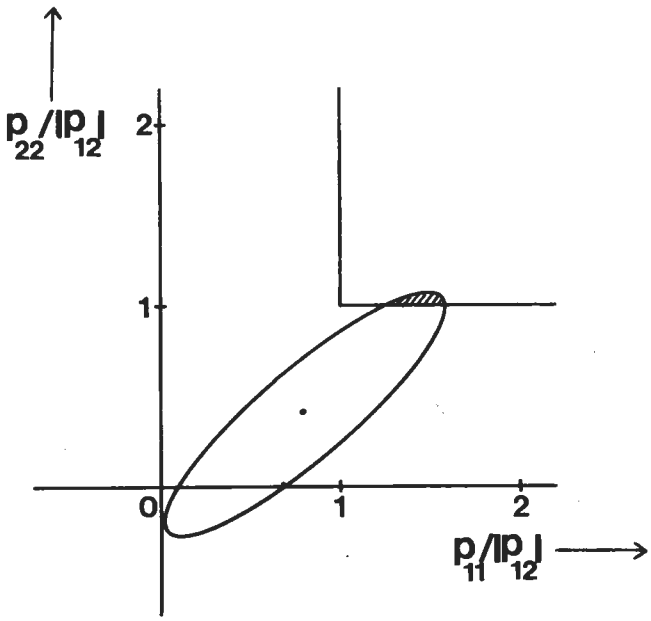


Fig. II.1. Region of parameters satisfying equation (II.10).

Appendix III

In this appendix we prove that there exists no controlled rounding mechanism to suppress zero-input limit cycles in the second-order direct form digital filter of type II.
In this filter both state signals have to be quantized:

$$\begin{aligned} x_1(n+1) &= a\cdot x_1(n) + x_2(n) + \epsilon_1(n) \\ x_2(n+1) &= b\cdot x_1(n) + \epsilon_2(n). \end{aligned} \qquad \text{(III.1)}$$

The general energy function of a second-order system is

$$E(n) = \underline{x}^T(n)\cdot P\cdot \underline{x}(n), \qquad \text{(III.2)}$$

where matrix P is positive definite for

$$p_{11} > 0, \quad p_{22} > 0 \quad \text{and} \quad p_{11}\cdot p_{22} - p_{12}^2 > 0. \qquad \text{(III.3)}$$

Without wordlength reduction the energy cannot increase with increasing time if

$$P - A^T\cdot P\cdot A \quad \text{is positive definite.} \qquad \text{(III.4)}$$

For $b \to -1$ the complex conjugated poles of this system lie on the unit circle. Then the idealized linear system is oscillating without any reduction of the energy. So

$$E(n+1) = E(n) \qquad \text{for all } n, \qquad \text{(III.5)}$$

or

$$\underline{x}^T(n+1)\cdot P\cdot \underline{x}(n+1) = \underline{x}^T(n)\cdot P\cdot \underline{x}(n) \quad \text{for all } \underline{x}(n). \qquad \text{(III.6)}$$

This means that

$$P - A^T\cdot P\cdot A = 0, \qquad \text{(III.7)}$$

and we have no freedom in choice of the P matrix anymore.

The energy function $E(n)$ becomes

$$E(n) = p_{11}\cdot[x_1^2(n) + a\cdot x_1(n)\cdot x_2(n) + x_2^2(n)], \qquad \text{(III.8)}$$

with

$$\Delta E(n) = E(n+1) - E(n)$$

$$= p_{11}\cdot\epsilon_1(n)\cdot[a\cdot x_1(n) + 2x_2(n) + \epsilon_1(n)] + a\cdot p_{11}\cdot\epsilon_1(n)\cdot\epsilon_2(n)$$

$$+ p_{11}\cdot\epsilon_2(n)\cdot[(a^2-2)\cdot x_1(n) + a\cdot x_2(n) + \epsilon_2(n)]. \qquad \text{(III.9)}$$

We can use the principle of controlled rounding if $\Delta E(n)$ can be written in the form

$$\Delta E(n) = c_1\cdot\epsilon_1(n)\cdot[x_1(n+1)+i_{11}\cdot x_1(n)+i_{12}\cdot x_2(n)+i_{13}\cdot x_2(n+1)]$$
$$+ c_2\cdot\epsilon_2(n)\cdot[x_2(n+1)+i_{21}\cdot x_1(n)+i_{22}\cdot x_2(n)+i_{23}\cdot x_1(n+1)]. \qquad \text{(III.10)}$$

Only for integer values of the i parameters the quantization direction can be determined uniquely. For non-integer values of the parameters i it is possible that both quantization upwards and downwards cause an increase of the energy $E(n)$.

The first part of expression (III.10) can only be a part of (III.9) for $c_1 = p_{11}$ and $i_{11} = 0$, $i_{12} = 1$ and $i_{13} = 0$. The second part of expression (III.10), however, cannot form the other part of (III.9) with integer values of the parameters i for non-integer values of coefficient a. Therefore no controlled rounding mechanism can be devised to suppress zero-input limit cycles in the second-order direct form digital filter of type II.

Appendix IV

In this appendix we present the *cyclotomic polynomials* [42, 49, 232]. The function $z^M - 1$ has M complex conjugated roots z_k:

$$z_k = \exp(j\frac{2\pi k}{M}) \qquad \text{(IV.1)}$$

All these roots z_k satisfy the condition $(z_k)^M = 1$. But some of these roots also satisfy the condition

$$(z_k)^D = 1 \qquad \text{for } 1 \leq D < M, \qquad \text{(IV.2)}$$

where D is a divisor of M. The smallest value of D for which eq (IV.2) is valid is called the *order of the root* z_k. The order D of some root z_k is independent of the choice of M. All z_k of the same order D form a set of points which together just form the roots of a polynomial $Q_D(z)$. This function is called the *cyclotomic polynomial of order D*. Some of the cyclotomic polynomials are shown in table IV.1.

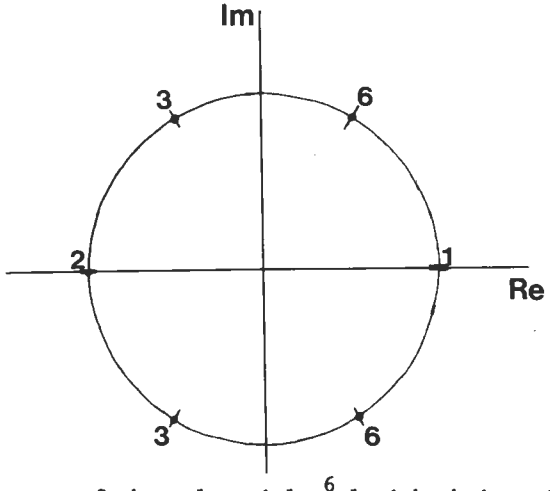


Fig. IV.1. The roots of the polynomial z^6-1 with their order.

For example, the roots of z^6-1 are shown in Fig. IV.1. In this figure the order of each root is indicated. There is one root of order 1 at $z = 1$, which implies that $Q_1(z) = z - 1$. There is one root of order 2 at

$z = -1$, so that $Q_2(z) = z + 1$. In Fig. III.1 we see two roots of order 3, which form the polynomial $Q_3(z) = z^2 + z + 1$. The two remaining roots have order 6 and belong to the polynomial $Q_6(z) = z^2 - z + 1$.

One of the properties of the cyclotomic polynomials is

$$\prod_{D|M} Q_D(z) = z^M - 1, \qquad \text{(IV.3)}$$

where $D|M$ stands for all divisors D of M, including $D = 1$ and $D = M$. This property can immediately be understand from the definition of the cyclotomic polynomials, since the union of all sets of roots of cyclotomic polynomials $Q_D(z)$ form the set of all roots of the function z^M-1.

According to the definition of the cyclotomic polynomials they can directly be calculated with the formula

$$Q_D(z) = \prod_{d|D} [z^{\frac{D}{d}} - 1]^{\mu(d)} \qquad \text{(IV.4)}$$

where

$\mu(d)$ is the *Moebius function*, defined by

$$\begin{aligned} \mu(d) &= 1 && \text{for } d = 1 \\ &= (-1)^k && \text{for } d \text{ is a product of } k \text{ distinct primes} \\ &= 0 && \text{otherwise.} \end{aligned} \qquad \text{(IV.5)}$$

For example

$$Q_6(z) = \frac{(z^6 - 1)\cdot(z - 1)}{(z^3 - 1)\cdot(z^2 - 1)} = z^2 - z + 1 \qquad \text{(IV.6)}$$

The roots of the cyclotomic polynomial $Q_6(z)$ are roots of the function z^6-1 which are not roots of z^3-1 or z^2-1. The term z-1 is added in the numerator of this quotient because the root $z = 1$ is twice subtracted, both by z^3-1 and z^2-1.

□

$$Q_1(z) = z - 1$$
$$Q_2(z) = z + 1$$
$$Q_3(z) = z^2 + z + 1$$
$$Q_4(z) = z^2 + 1$$
$$Q_5(z) = z^4 + z^3 + z^2 + z + 1$$
$$Q_6(z) = z^2 - z + 1$$
$$Q_7(z) = z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$
$$Q_8(z) = z^4 + 1$$
$$Q_9(z) = z^6 + z^3 + 1$$
$$Q_{10}(z) = z^4 - z^3 + z^2 - z + 1$$
$$Q_{11}(z) = z^{10} + z^9 + z^8 + z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$
$$Q_{12}(z) = z^4 - z^2 + 1$$
$$Q_{13}(z) = z^{12} + z^{11} + z^{10} + z^9 + z^8 + z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$
$$Q_{14}(z) = z^6 - z^5 + z^4 - z^3 + z^2 - z + 1$$
$$Q_{15}(z) = z^8 - z^7 + z^5 - z^4 + z^3 - z + 1$$
$$Q_{16}(z) = z^8 + 1$$
$$Q_{17}(z) = z^{16} + z^{15} + z^{14} + z^{13} + z^{12} + z^{11} + z^{10} + z^9 + z^8 + z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$
$$Q_{18}(z) = z^6 - z^3 + 1$$
$$Q_{19}(z) = z^{18} + z^{17} + z^{16} + z^{15} + z^{14} + z^{13} + z^{12} + z^{11} + z^{10} + z^9 + z^8 + z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$
$$Q_{20}(z) = z^8 - z^6 + z^4 - z^2 + 1$$
$$Q_{21}(z) = z^{12} - z^{11} + z^9 - z^8 + z^6 - z^4 + z^3 - z + 1$$
$$Q_{22}(z) = z^{10} - z^9 + z^8 - z^7 + z^6 - z^5 + z^4 - z^3 + z^2 - z + 1$$
$$Q_{23}(z) = z^{22} + z^{21} + z^{20} + z^{19} + z^{18} + z^{17} + z^{16} + z^{15} + z^{14} + z^{13} + z^{12} + z^{11} + z^{10} + z^9 + z^8 + z^7 + z^6 + z^5 + z^4 + z^3 + z^2 + z + 1$$

Table IV.1. Cyclotomic polynomials

Some properties of the cyclotomic polynomials are listed here for quick reference. Most of these properties are proved in [42, 49, 232].

1) For p = prime is $Q_p(z) = \sum_{i=0}^{p-1} z^i$. (IV.7)

2) For p = prime is $Q_{p^k}(z) = Q_p(z^{p^{k-1}})$. (IV.8)

3) For p = prime and m has no factor p is $Q_{m\cdot p^k}(z) = \frac{Q_m(z^{p^k})}{Q_m(z^{p^{k-1}})}$. (IV.9)

4) For m = odd and $m \geq 3$ is $Q_{2m}(z) = Q_m(-z)$. (IV.10)

5) The degree of cyclotomic polynomial $Q_m(z)$ equals the *Eulers phi function* $\phi(m)$

For $m = \prod_{i=1}^{t} p_i^{e_i}$ with p_i = prime and $e_i \geq 1$ $\phi(m) = \prod_{i=1}^{t} (p_i-1)\cdot p_i^{(e_i-1)}$ (IV.11)

6) For $m \geq 2$ is $Q_m(z^{-1}) = Q_m(z)\cdot z^{-\phi(m)}$. (IV.12)

with other words the cyclotomic polynomials have symmetric coefficients.

7) For cyclotomic polynomials $Q_m(z)$ with an order m, which contains two prime factors all the coefficients have values -1, 0 or 1.
The same statement accounts for an even order m, which has three prime factors as can be concluded from property 5) (see [39]).

8) By calculation of all the cyclotomic polynomials upto $Q_{105}(z)$, this is the first polynomial which has a coefficient which does not equal one of the values -1, 0 or 1. The particular coefficient has a value 2.
$m = 105$ is the smallest product of three distinct odd prime factors, so it is the smallest value that does not satisfy 7); $m = 105 = 3\cdot5\cdot7$.

9) The coefficients of cyclotomic polynomial are integers only.
These coefficients are not bounded by some value [50, 168, 206].

References

[1] A.I. Abu-El-Haija, K. Shenoi and A.M. Peterson, "Digital filter structures having low errors and simple hardware implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 593-599, 1978.

[2] A.I. Abu-El-Haija and A.M. Peterson, "Limit cycle oscillations in digital incremental computers," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 902-908, 1978.

[3] A.I. Abu-El-Haija and A.M. Peterson, "An approach to eliminate roundoff errors in digital filters," *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, Tulsa, Oklahoma, USA (New York, USA: IEEE 1978), pp. 75-78, 1978.

[4] A.I. Abu-El-Haija and A.M. Peterson, "An approach to eliminate roundoff errors in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 195-198, 1979.

[5] A.I. Abu-El-Haija, "Correction to 'Limit cycle oscillations in digital incremental computers'," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 898, 1979.

[6] A.I. Abu-El-Haija and A.M. Peterson, "On limit cycle amplitudes in errorfeedback digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, Georgia, USA (New York, USA: IEEE 1981), pp. 1227-1230, 1981.

[7] A.I. Abu-El-Haija, "On implementing error-feedback digital filters," *25th Midwest Symp. Circuits Syst.*, Houghton, Michigan, USA, ed. R.E. Stuffle, P.H. Lewis (North-Hollywood, California, USA: Western Periodicals Comp. 1982), pp. 140-147, 1982.

[8] A.I. Abu-El-Haija, "A tight bound on $\Sigma |h(n)|$ for general second-order H(z)," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 492-497, 1982.

[9] A.I. Abu-El-Haija, "Error-feedback digital filters with minimum limit cycle oscillations," *Proc. EUSIPCO-83, 2nd European Signal Processing Conf.*, Erlangen, Germany, ed. H.W. Schüssler (Amsterdam, The Netherlands: North-Holland 1983), pp. 111-114, 1983.

[10] A.I. Abu-El-Haija, "Determining coefficients of error-feedback digital filters to obtain minimum roundoff errors with minimum complexity," *Proc. IEEE Int. Symp. Circuits Syst.*, Newport Beach, California, USA (New York, USA: IEEE 1983), pp. 819-822, 1983.

[11] M.T. Abu-El-Ma'atti, "The intermodulation due to multicarrier quantization," *IEEE Trans. Communications*, vol. COM-32, pp. 1211-1214, 1984.

[12] J.I. Acha and J. Payan, "Low-noise structures for narrow-band recursive digital filters without overflow oscillations," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 96-99, 1987.

[13] M.H. Ackroyd, *Digital filters*, London, England: Butterworths, 1973.

[14] M.H. Ackroyd and H.M. Liu, "Limit cycle suppression in digital filters," *Saraga Memorial Coll. Electronic Filters*, London, England (London, England: IEE 1982), pp. 5/1-6, 1982.

[15] K.M. Adams, "Elementary limit cycles with applications to the testing of digital circuits," *Proc. IEEE Int. Symp. Circuits Syst.*, Rome, Italy (New York, USA: IEEE 1982), pp. 1237-1240, 1982.

[16] K.M. Adams, "Elementary limit cycles with applications to the testing of digital circuits," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 809814, 1983.

[17] 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.

[18] R.C. Agarwal and J.W. Cooley, "New algoritms for digital convolution," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 392-409, 1979.

[19] J.K. Aggarwal, *Digital signal processing*, North-Hollywood, California, USA: Western Periodicals Comp., 1979.

[20] M.J.J.C. Annegarn, "Chopping operations in wave digital filters," *Electronics Letters*, vol. 11, pp. 378-380, 1975.

[21] M.J.J.C. Annegram, A.H.H.J. Nilesen and J.G. Raven, "Digitale signaalbewerking in TV-ontvangers (in dutch)," *Philips Research Reports*, vol. 42, pp. 191-209, 1985.

[22] A. Antoniou, *Digital filters: analysis and design*, New York, USA: Mc Graw-Hill Book Comp., 1979.

[23] U. Appel, "Bounds on second-order digital filter limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 630-632, 1975.

[24] D.P. Atherton, "Limit cycles in sampled data and digital control systems," *Proc. IEE Coll.*, London, England (London, England: IEE 1984), pp. 2/1-4, 1984.

[25] E. Auer, "A method for the determination of all limit cycles," *Proc. 6th European Conf. Circuit Theory Design*, Stuttgart, Germany, ed. E. Lueder (Berlin, Germany: VDE-Verlag 1983), pp. 154-156, 1983.

[26] E. Auer, "New methods to guarantee overflow-stability for the canonic second-order digital filter structure," *Proc. 7th European Conf. Circuit Theory Design*, Prague, Czechoslovakia, ed. V. Zina and J. Krasil (Amsterdam, The Netherlands: North-Holland 1985), pp. 477-480, 1985.

[27] E. Auer, "Digital filter structures free of limit cycles," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, Texas, USA (New York, USA: IEEE 1987), pp. 904-907, 1987.

[28] A.Z. Baraniecki and G.A. Julien, "Quantization error and limit cycle analysis in residue number system coded recursive filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Paris, France (New York, USA: IEEE 1982), pp. 52-55, 1982.

[29] A.I. Barkin, "Sufficient conditions for the absence of auto-oscillations in pulse systems," *Automation and Remote Control*, vol. 31, pp. 942-946, 1970.

[30] C.W. Barnes and A.T. Fam, "Minimum norm recursive digital filters that are free of overflow limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 569-574, 1977.

[31] C.W. Barnes, "Roundoff noise and overflow in normal digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 154-159, 1979.

[32] C.W. Barnes and S. Shinnaka, "Finite word effects in block-state realizations of fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 345-349, 1980.

[33] C.W. Barnes and S. Shinnaka, "Stability domains for second-order recursive digital filters in normal form with 'matrix power' feedback," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 841-843, 1980.

[34] C.W. Barnes, "Error-feedback in normal realization of recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 72-75, 1981.

[35] C.W. Barnes and T. Miyawaki, "Roundoff noise invariants in normal digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 251-256, 1982.

[36] C.W. Barnes, "On the design of optimal state-space realizations of second-order digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 602-608, 1984.

[37] C.W. Barnes, "A computationally efficient second-order digital filter sections with low roundoff noise gain," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 841-847, 1984.

[38] C.W. Barnes, "A parametric approach to the realization of second-order digital filter sections," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 530-539, 1985.

[39] S.M. Beiter, "Magnitude of the coefficients of the cyclotomic polynomial Fpqr(x)," *American Mathematical Monthly*, vol. 75, pp. 370-372, 1968.

[40] A.A. Belal, "On the quantization error bounds in second-order digital filters with complex conjugate poles," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 45, 1977.

[41] M. Bellanger, *Digital processing of signals, theory and practice*, New York, USA: John Wiley & Sons, Inc., 1984.

[42] E.R. Berlekamp, *Algebraic coding theory*, New York, USA: Mc Graw-Hill Book Comp., 1968.

[43] U. Bernhardt, H. Lubenow and H. Unger, "On avoiding limit cycles in digital filters (in German)," *Nachrichtentechn. Elektr.*, vol. 27, pp. 433-435, 1977.

[44] T. Beth and W. Fumy, "Hardware oriented algorithms for the fast symbolic calculation of the DFT," *Electronics Letters*, vol. 19, pp. 901-902, 1983.

[45] D.V. Bhaskar-Rao, "A study of coefficient quantization errors in state-space digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, USA (New York, USA: IEEE 1985), pp. 1715-1718, 1985.

[46] D.V. Bhaskar-Rao, "Analysis of coefficient quantization errors in state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 131-139, 1986.

[47] N.M. Blachman, "Third-order intermodulation due to quantization," *IEEE Trans. Communications*, vol. COM-29, pp. 1386-1389, 1981.

[48] N.M. Blachman, "The intermodulation and distortion due to quantization of sinusoids," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1417-1426, 1985.

[49] I.F. Blake and R.C. Mullin, *An introduction to algebraic and combinatorial coding theory*, New York, USA: Academic Press, 1976.

[50] D.M. Bloom, "On the coefficients of the cyclotomic polynomials," *American Mathematical Monthly*, vol. 75, pp. 372-377, 1968.

[51] R. Boite, X.L. He and J.P. Renard, "On the floating-point realizations of digital filters," *Proc. EUSIPCO-88, 4th European Signal Processing Conf.*, Grenoble, France, ed. J.L. Lacoume et al. (Amsterdam, The Netherlands: North-Holland 1988), pp.1501-1504, 1988.

[52] A.G. Bolton, "Analysis of digital filters with randomly dithered coefficients," *IEE Proceedings G*, vol. 128, pp. 61-66, 1981.

[53] A.G. Bolton, "A two's complement overflow limit cycle free digital filter structure," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 1045-1046, 1984.

[54] B.W. Bomar, "New second-order state-space structures for realizing low roundoff noise digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 106-110, 1985.

[55] B.W. Bomar, "Computationally efficient low roundoff noise second-order state-space structures," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 35-41, 1986.

[56] R.K. Brayton and C.H. Tong, "Stability of dynamical systems: a constructive approach," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 224-234, 1979.

[57] T.A. Brubaker and J.N. Gowdy, "Limit cycles in digital filters," *IEEE Trans. Automatic Control*, vol. AC-17, pp. 675-677, 1972.

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

[59] H.J. Butterweck, "On the quantization noise contributions in digital filters which are uncorrelated with the output signal," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 901-910, 1979.

[60] H.J. Butterweck, "Quantization effects in various second-order digital filters: a comparitive study," *Proc. 5th European Conf. Circuit Theory Design*, The Hague, The Netherlands, ed. R. Boite, P. Dewilde (Delft, The Netherlands: Delft University Press 1981), pp. 863-864, 1981.

[61] H.J. Butterweck, F.H.R. Lucassen and G. Verkroost, "Subharmonics and related quantization effects in periodically excited recursive digital filters," *Proc. EUSIPCO-83, 2nd European Signal Processing Conf.*, Erlangen, Germany, ed. H.W. Schüssler (Amsterdam, The Netherlands: North-Holland 1983), pp. 57-59, 1983.

[62] H.J. Butterweck, A.C.P. van Meer and G. Verkroost, "New second-order digital filter sections without limit cycles," *Proc. EUSIPCO 83, 2nd European Signal Processing Conf.*, Erlangen, Germany, ed. H.W. Schüssler (Amsterdam, The Netherlands: North-Holland 1983), pp. 97-98, 1983.

[63] H.J. Butterweck, A.C.P. van Meer and G. Verkroost, "New second-order digital filter sections without limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 141-146, 1984.

[64] H.J. Butterweck, F.H.R. Lucassen and G. Verkroost, "Subharmonics and other quantization effects in periodically excited recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 958-964, 1986.

[65] H.J. Butterweck, "Subharmonics and other quantization effects in periodically excited recursive digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, San Jose, California, USA (New York, USA: IEEE 1986), pp. 861-862, 1986.

[66] H.J. Butterweck, J.H.F. Ritzerfeld and M.J. Werter, "Finite word-length effects in digital filters," *Arch. Elektron. Uebertragung. (to be published)*, 1989.

[67] H.J. Butterweck, J.H.F. Ritzerfeld and M.J. Werter,"*Finite word-length effects in digital filters - a review*," Eindhoven, The Netherlands: Eindhoven University of Technology, EUT Report 88-E-205, 1988.

[68] M. Büttner, "Some experimental results concerning random-like noise and limit cycles in recursive digital filters," *Nachrichtentechn. Z.*, vol. NTZ-28, pp. 402-406, 1975.

[69] M. Büttner and J. Schloss, "On the experimental investigations of bounds of the amplitude of limit cycles," *Proc. 2nd European Conf. Circuit Theory Design*, Genova, Italy, pp. 709-716, 1976.

[70] M. Büttner, "A novel approach to eliminate limit cycles in digital filters with a minimum increase in the quantization noise," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 291-294, 1976.

[71] M. Büttner, "Elimination of limit cycles in digital filters with very low increase in the quantization noise," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 300-304, 1977.

[72] D.S.K. Chan, "Constrained minimization of roundoff noise in fixed-point digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Washington, D.C., USA (New York, USA: IEEE 1979), pp. 335-339, 1979.

[73] T.L. Chang and C.S. Burrus, "Oscillations caused by quantization in digital filters," *Proc. IEEE Int. Symp. Circuit Theory*, North-Hollywood, California, USA (New York, USA: IEEE 1972), pp. 228-232, 1972.

[74] T.L. Chang, "A note on upper bounds on limit cycles in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 99-100, 1976.

[75] T.L. Chang, "Error-feedback digital filters," *Electronics Letters*, vol. 15, pp. 348-349, 1979.

[76] T.L. Chang, "Comments on 'An approach to eliminate roundoff errors in digital filters' (Author's reply: A.I. Abu-El-Haija)," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 244-245, 1980.

[77] T.L. Chang, "Suppression of limit cycles in digital filters designed with one magnitude-truncation quantizer," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 107-111, 1981.

[78] C. Chen, *One-dimensional digital signal processing*, New York, USA: Marcel Dekker, Inc., 1979.

[79] D. Childers and A. Durling, *Digital filtering and signal processing*, New York, USA: Western Publ. Comp., 1975.

[80] L.O. Chua and T. Lin, "Chaos in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 648-658, 1988.

[81] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Second-order digital filters with only one magnitude-truncation quantizer and having practically no limit cycles," *Electronics Letters*, vol. 9, pp. 531-532, 1973.

[82] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Some remarks on the classification of limit cycles in digital filters," *Philips Research Reports*, vol. 28, pp. 297-305, 1973.

[83] T.A.C.M. Claasen and L.O.G. Kristiansson, "Improvement of overflow behaviour of second-order digital filters by means of error feedback," *Electronics Letters*, vol. 10, pp. 240-241, 1974.

[84] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "A comparison between the stability of second-order filters with various arithmetics," *Proc. 1st European Conf. Circuit Theory Design*, London, England (London, England: IEE 1974), pp. 354-358, 1974.

[85] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Remarks on the zero-input behaviour of second-order digital filters designed with one magnitude-truncation quantizer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 240-242, 1975.

[86] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Necessary and sufficient conditions for the absence of overflow phenomena in a second-order recursive digital filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 509-515, 1975.

[87] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Frequency domain criteria for the absence of zero-input limit cycles in nonlinear discrete-time systems, with applications to digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 232-239, 1975.

[88] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "On the stability of the forced response of digital filters with overflow nonlinearities," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 692-696, 1975.

[89] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Quantization noise analysis for fixed point digital filters using magnitude truncation for quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 887-895, 1975.

[90] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Quantization errors in digital filters using magnitude truncation quantization," *Proc. IEEE Int. Conf. Comm.*, San Francisco, California, USA (New York, USA: IEEE 1975), pp. 31/17-21, 1975.

[91] T.A.C.M. Claasen, "Quantization noise analysis of digital filters with controlled quantization," *Electronics Letters*, vol. 12, pp. 46-48, 1976.

[92] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Effects of quantization and overflow in recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 517-529, 1976.

[93] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "A survey of quantization and overflow effects in recursive digital filters" *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 620-624, 1976.

[94] T.A.C.M. Claasen, W.F.G. Mecklenbräuker and J.B.H. Peek, "Een overzicht van niet-lineaire effekten in rekursieve digitale filters (in dutch)," *Nederl. Electron. Radio Genootschap, NERG*, vol. 42, pp. 105-109, 1977.

[95] L. Claesen, J. Vandewalle and H. De Man, "General bounds on parasitic oscillations in arbitrary digital filters and their application in CAD," *Proc. IEEE Int. Symp. Circuits Syst.*, Montreal, Canada (New York, USA: IEEE 1984), pp. 747-750, 1984.

[96] R. Czarnach, "*Antwortstabile rekursive digitalfilter in festkommaarithmetik* (in german)," Erlangen, Germany: Institut für Nachrichtentechnik, Universität Erlangen-Nürnberg, 1984.

[97] E.D. De Luca and G.O. Martens, "A new coefficient quantization scheme to suppress limit cycles in state-variable digital filters" *Proc. 6th European Conf. Circuit Theory Design*, Stuttgart, Germany ed. E. Lueder (Berlin, Germany: VDE-Verlag 1983), pp.157-159, 1983

[98] C.D.R. De Vaal and R. Nouta, "Suppression of parasitic oscillations in floating point wave digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, New York, USA (New York, USA: IEEE 1978), pp. 1018-1022, 1978.

[99] C.D.R. De Vaal and R. Nouta, "On the suppression of zero-input parasitic oscillations in floating point wave digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 144-145, 1980.

[100] P.S.R. Diniz and A. Antoniou, "On the elimination of constant-input limit cycles in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 670-671, 1984.

[101] P.S.R. Diniz and A. Antoniou, "New improved state-space digital filter structures," *Proc. IEEE Int. Symp. Circuits Syst.*, Kyoto, Japan (New York, USA: IEEE 1985), pp. 1599-1602, 1985.

[102] P.S.R. Diniz and A. Antoniou, "More economical state-space digital filter structures which are free of constant-input limit cycles," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 807-815, 1986.

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

[104] E. Eckhardt and W. Winkelnkemper, "Implementation of a second-order digital filter section with stable overflow behaviour," *Nachrichtentechn. Z.*, vol. NTZ-26, pp. 282-284, 1973.

[105] B. Eckhardt, "On the roundoff error of a multiplier," *Arch. Elektron. Uebertragung.*, vol. AEU-29, pp. 162-164, 1975.

[106] B. Eckhardt and H.W. Schüssler, "On the quantization error of a multi-plier," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 634-637, 1976.

[107] O.I. Elgerd, *Control systems theory*, Tokyo, Japan: Mc Graw-Hill Kogakusha, Ltd., 1967.

[108] A.W.M.vd Enden and N.A.M. Verhoeckx, *Digitale signaalbewerking (in dutch)*, Overberg, The Netherlands: Delta Press bv., 1987.

[109] K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions Part I direct form and coupled form filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 113-132, 1985.

[110] K.T. Erickson and A.N. Michel, "Stability analysis of fixed-point digital filters using computer generated Lyapunov functions Part II: wave digital and lattice digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 132-139, 1985.

[111] C. Eswaran and A. Antoniou, "Wave digital biquads that are free of limit cycles under zero- and constant-input conditions," *Proc. IEEE Int. Symp. Circuits Syst.*, Montreal, Canada (New York, USA: IEEE 1984), pp. 723-726, 1984.

[112] C. Eswaran, A. Antoniou and K. Manivannan, "Universal digital biquads which are free of limit cycles," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 1243-1248, 1987.

[113] M.H. Etzel and W.K. Jenkins, "Error correction and suppression properties of RRNS digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Houston, Texas, USA (New York, USA: IEEE 1980), pp. 1117-1120, 1980.

[114] M.H. Etzel and W.K. Jenkins, "The design of specialized classes for efficient recursive filter realizations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 52-55, 1982.

[115] A.T. Fam, "Multiplexing preserving filters," *Proc. 12th Asilomar Conf. Circuits, Syst., Computers*, Pacific Grove, California, USA (North-Hollywood, California, USA: Western Periodicals Comp. 1978) pp. 257-259, 1978.

[116] A.T. Fam and C.W. Barnes, "Nonminimal realizations of fixed-point digital filters that are free of all finite wordlength limit cycles," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 149-153, 1979.

[117] A. Fettweis, "Digital filter structures related to classical filter networks," *Arch. Elektron. Uebertragung.*, vol. AEU-25, pp. 79-89, 1971.

[118] A. Fettweis, "Pseudopassivity, sensitivity and stability of wave digital filters," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 668-673, 1972.

[119] A. Fettweis, "Roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 174-175, 1973.

[120] A. Fettweis, "On the properties of floating-point roundoff noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 149-151, 1974.

[121] A. Fettweis, "On sensitivity and roundoff noise in wave digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 383-384, 1974.

[122] A. Fettweis and K. Meerkötter, "Suppression of parasitic oscillations in wave digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, San Fransisco, California, USA (New York, USA: IEEE 1974), pp. 682-686, 1974.

[123] A. Fettweis and K. Meerkötter, "Suppression of parasitic oscillations in wave digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 239-246, 1975.

[124] A. Fettweis and K. Meerkötter, "Correction to 'Suppression of parasitic oscillations in wave digital filters'," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 575, 1975.

[125] A. Fettweis and K. Meerkötter, "Suppression of parasitic oscillations in half-synchronic wave digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 125-126, 1976.

[126] A. Fettweis and K. Meerkötter, "On parasitic oscillations in digital filters under looped conditions," *IEEE Trns. Circuits Syst.*, vol. CAS-24, pp. 475-481, 1977.

[127] A. Fettweis, "Digital circuits and systems," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 31-48, 1984.

[128] A. Fettweis, "Wave digital filters: Theory and practice," *Proc. of the IEEE*, vol. 74, pp. 270-327, 1986.

[129] A.M. Fink, "A bound on quantization errors in second-order digital filters with complex poles that is tight for small theta," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 325-326, 1976.

[130] E.D. Garber, "Frequency criteria for the absence of periodic modes," *Automation and Remote Control*, vol. 28, pp. 1776-1780, 1967.

[131] J.H.T. Geerlings, *Limit cycles in digital filters: a bibliography*, 1975-1984, Eindhoven, The Netherlands: Eindhoven University of Technology, 1985.

[132] B. Gold and C.M. Rader, *Digital processing of signals*, New York, USA: Mc Graw-Hill Book Comp., 1969.

[133] L.M. Gol'denberg, "Stability of recursive digital filters," *Avtomatika i Telemekhanika*, vol. 7, pp. 22-27, 1977.

[134] L.M. Gol'denberg and B.D. Matyushkin, "Stability of recursive second-order digital filters," *Avtomatika i Telemekhanika*, vol. 12, pp. 54-58, 1978.

[135] L.M. Gol'denberg and B.D. Matyushkin, "Conditions for the absolute stability of processes in recursive digital filters," *Automation and Remote Control*, vol. 40, pp. 1161-1169, 1979.

[136] A.H. Gray, jr., "Passive cascaded lattice digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 337-344, 1980.

[137] B.D. Green and L.E. Turner, "New limit cycles bounds for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 365-374, 1988.

[138] R.W.C. Groen, A.W.M. v.d. Enden and J.H.F. Ritzerfeld, "On the calculation of quantization noise in digital filters caused by magntude truncation," *Proc. 8th European Conf. Circuit Theory Design*, Paris, France, ed. R. Gerber (Amsterdam, The Netherlands: Elseviers Science Publishers 1987) pp. 717-722, 1987.

[139] P. Gruber, "The determination of limit cycles in nonlinear sampled data systems," *Proc. 1976 Joint Automatic Control Conf.*, W-Lafayette, Indiana, USA (New York, USA: ASME 1976), pp. 424-433, 1976.

[140] P. Gruber, *Beitrag zum problem der bestimmung von grenzzyklen in nicht-linearen abgetasteten regelsystemen (in german)*, Zurich, Switzerland: Zurich University of Technology, Doctor's Thesis, 1978.

[141] C. Hayashi, *Nonlinear oscillations in physical systems*, New York, USA: Mc Graw-Hill Book Comp., 1964.

[142] S. Haykin, *Adaptive filter theory*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1986.

[143] D.G. He and S.K. Han, "Novel approximations to $\Sigma\,|h(n)|$ for second order digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, USA (New York, USA: IEEE 1985), pp. 1731-1734, 1985.

[144] J.A. Heinen, "A bound on the norm of the individual state variables of a digital filter," *IEEE Trans. Circuits Syst.*, vol.CAS-32, pp. 1073-1074, 1985.

[145] T. Higuchi and H. Takeo, "A state-space approach for the elimination of limit cycles in digital filters with arbitrary structures, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Washington, D.C., USA (New York, USA: IEEE 1979), pp. 355-358, 1979.

[146] S.Y. Hwang, "Dynamic range constraint in state-space digital filtering," *IEEE Trans. Acoust.,Speech, Signal Processing*, vol. ASSP-23, pp. 591-593, 1975.

[147] S.Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.ASSP-24, pp. 256-262, 1976.

[148] S.Y. Hwang, "Roundoff noise minimization in state-space digital filtering," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Phila-delphia, Pennsylvania, USA (New York, USA: IEEE 1976), pp. 498-500, 1976.

[149] S.Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filtering," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273-281, 1977.

[150] R. Ishii, Y. Mitome and S. Kawasaki, "Limit cycles in a second-order all-pole digital system (in Japanese; English abstract)," *Trans. Inst. Electron. Commun. Eng. Japan, sect. A, vol. A-63 pp. 51-58. English abstract: ibid. sect. E*, vol. E-63, pp. 120, 1980.

[151] R. Ishii and M. Kato, "An effect of a limit cycle on an output signal," *Proc. IEEE Int. Symp. Circuits Syst.*, Rome, Italy (New York, USA: IEEE 1982), pp. 1041-1044, 1982.

[152] Y.S. Itskovich, "Roundoff noise in digital filters with small input signal (transl. of: Radiotekhnika, vol. 37, no. 5, p. 39, May 1982)," *Telecommun. Radio Eng., part 2, no.5 (May 1982)*, vol. 36/37, pp. 74-77, 1982.

[153] M. Iwatsuki, M. Kawamata and T. Higuchi, "Synthesis of minimum sensitivity structures in linear systems using controllability and observability measures," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokoyo, Japan (New York, USA: IEEE 1986), pp. 501-504, 1986.

[154] L.B. Jackson, J.F. Kaiser and H.S. Mc Donald, "An approach to the implementation of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 413-421, 1968.

[155] L.B. Jackson, "An analysis of limit cycles due to multiplication rounding in recursive digital (sub-) filters," *Proc. 7th Annual Allerton Conf. Circuit Syst. Theory*, Monticello, Illinois, USA, ed. A.H. Hadad et al. (Urbana, Illinois, USA: Dept. of Electrical Eng., Univ. of Illinois 1969), pp. 69-78, 1969.

[156] L.B. Jackson, "On the interaction of roundoff noise and dynamic range in digital filters," *Bell System Techn. J.*, vol. 49, pp. 159-184, 1970.

[157] L.B. Jackson, "Roundoff noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 107-122, 1970.

[158] L.B. Jackson, "Comments on 'Quantizer-induced digital controller limit cycles'," *IEEE Trans. Automatic Control*, vol. AC-15, pp. 614-615, 1970.

[159] L.B. Jackson, "Roundoff noise bounds derived from coefficient sensitivities for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 481-485, 1976.

[160] L.B. Jackson, A.G. Lindgren and Y. Kim, "Synthesis of state-space digital filters with low round-off noise and coeffcient sensitivity," *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, Arizona, USA (New York, USA: IEEE 1977), pp. 41-44, 1977.

[161] L.B. Jackson, A.G. Lindgren and Y. Kim, "Optimal synthesis of second-order state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 149-153, 1979.

[162] L.B. Jackson, "Limit cycles in state-space structures for digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp.67-68, 1979.

[163] L.B. Jackson and N.H.K. Judell, "Addendum to 'Limit cycles in state-space structures for digital filters'," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 320, 1980.

[164] L.B. Jackson, *Digital filters and signal processing*, Dordrecht, The Netherlands: Kluwer Academic Publ., 1986.

[165] G.U. Jatnieks and B.A. Shenoi, "Zero-input limit cycles in coupled digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 146-149, 1974.

[166] N.B. Jones, *Digital signal processing (IEE Control engineering; 22)*, Exeter, England: Short Runpress, Ltd., 1982.

[167] E.I. Jury, *Theory and application of the z-transform method*, New York, USA: John Wiley & Sons, Inc., 1964.

[168] J. Justin, "Bornes des coefficients du polynome cyclotomique et de certains autres polynomes (in french)," *C.R. Acad. Sc. Paris*, vol. 268, pp. A995-997, 1969.

[169] J.F. Kaiser, "Quantization effects in digital filters," *Proc. IEEE Int. Symp. Circuit Theory*, Toronto, Canada (New York, USA: IEEE 1973), pp. 415-417, 1973.

[170] J.F. Kaiser, "On the limit cycle problem," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 642-645, 1976.

[171] E.P.F. Kan and J.K. Aggarwal, "Minimum-deadband design of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 292-296, 1971.

[172] T. Kaneko, "Limit cycle oscillations in floating-point digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-12, pp. 100-106, 1973.

[173] C.Y. Kao, "An analysis of limit cycles due to sign-magnitude truncation in multiplication in recursive digital filters," *Proc. 5th Asilomar Conf. Circuits, Syst., Computers*, Pacific Grove, California, USA (North-Hollywood, California: Western Periodicals Comp. 1971), pp. 349-353, 1971.

[174] C.Y. Kao, "Apparatus for suppressing limit cycles due to quantization in digital filters," *U.S. Patent no. 3749895*, 1973.

[175] M. Kawamata and T. Higuchi, "A sufficient condition for the absence of overflow oscillations in arbitrary digital filters based on the element equations," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, Colorado, USA (New York, USA: IEEE 1980), pp. 85-88, 1980.

[176] M. Kawamata, "Synthesis of limit cycle free digital filters based on the state equations (in Japanese; English abstract)," *Trans. Inst. Electron. Commun. Eng. Japan, sect. A, vol. A-63 pp. 870-877. English abstract: ibid. sect. E*, vol. E-63, pp. 870, 1980.

[177] M. Kawamata and T. Higuchi, "A systematic approach to synthesis of limit cycle free digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 212-214, 1983.

[178] M. Kawamata and T. Higuchi, "Synthesis of limit cycle free state-space digital filters with minimum coefficient quantization error," *Proc. IEEE Int. Symp. Circuits Syst.*, Newport Beach, California, USA (New York, USA: IEEE 1983), pp. 827-830, 1983.

[179] M. Kawamata and T. Higuchi, "On the absence of limit cycles in a class of state-space digital filters which contains minimum noise realizations," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 928-930, 1984.

[180] M. Kawamata and T. Higuchi, "A unified approach to the optimal synthesis of fixed-point state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 911-920, 1985.

[181] R.B. Kieburtz, "An experimental study of roundoff effects in a tenth-order recursive digital filter," *IEEE Trans. Communications*, vol. COM-21, pp. 757-763, 1973.

[182] R.B. Kieburtz, "Rounding and truncation limit cycles in a recursive digital filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 73, 1974.

[183] R.B. Kieburtz and K.V. Mina, "Digital filter circuit," *U.S. Patent no. 3906199*, 1975.

[184] R.B. Kieburtz, K.V. Mina and V.B. Lawrence, "Control of limit cycles in recursive digital filters by randomized quantization," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 624-627, 1976.

[185] R.B. Kieburtz, K.V. Mina and V.B. Lawrence, "Control of limit cycles in recursive digital filters by randomized quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 291-299, 1977.

[186] U. Kleine and T.G. Noll, "On the forced-response stability of wave digital filters using carry-save arithmetic," *Arch. Elektron. Uebertragung.*, vol. AEU-41, pp. 321-324, 1987.

[187] A. Krishnan and R. Subramanian, "Jump phenomena in digital filters," *J. Inst. Electron. Telecomm. Eng. (India)*, vol. 25, pp. 373-376, 1979.

[188] L.O.G. Kristiansson, "Jump phenomena in digital filters," *Electronics Letters*, vol. 10, pp. 14-15, 1974.

[189] H. Kubota and S. Tsuji, "An upper bound on the RMS value of limit cycles in digital filters and a reduction method (in Japanese; English abstract)," *Trans. Inst. Electron. Commun. Eng. Japan, sect. A, vol. A-64 pp. 63-70. English abstract: ibid. sect. E*, vol. E-64, pp. 38, 1981.

[190] H. Kubota, T. Yoshida and S. Tsuji, "A consideration on limit cycles due to valuetruncation errors in digital filters (in Japanese; English abstract)," *Trans. Inst. Electron. Commun. Eng. Japan, sect. A, vol. A-64 pp. 371-377. English abstract: ibid. sect. E*, vol. E-64, pp. 365, 1981.

[191] M. Kunt, *Digital signal processing*, Norwood, Massachusetts, USA: Artech House, Inc., 1986.

[192] K. Kurosawa, "Limit cycle and overflow free digital filters (in Japanese; English abstract)," *Trans. Inst. Electron. Commun. Eng. Japan, sect. A, vol. A-65 pp. 263-264. English abstract: ibid. sect. E*, vol. E-65, pp. 180, 1982.

[193] H.K. Kwan, "A multi-output second-order digital filter structure for VLSI implementation," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 108-109, 1985.

[194] H.K. Kwan, "A multi-output second-order digital filter without limit cycle oscillations," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 974-975, 1985.

[195] H.K. Kwan, "A multi-output wave digital biquad using magnitude truncation instead of controlled rounding," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1185-1187, 1985.

[196] A. Lacroix, "Limit cycles in floating point digital filters," *18th Midwest Symp. Circuits Syst.*, Montreal, Canada, ed. M.N.S. Swamy (North-Hollywood, California, USA: Western Periodicals Comp. 1975), pp. 475-479, 1975.

[197] A. Lacroix, "Underflow limit cycles in floating point digital filters," *Proc. Florence Conf. Digital Signal Processing*, Florence, Italy, pp. 75-84, 1975.

[198] A. Lacroix, "Limit cycles in floating point digital filters," *Arch. Elektron. Uebertragung.*, vol. AEU-30, pp. 277-284, 1976.

[199] A. Lacroix and N. Hoptner, "Simulation of digital filters with the aid of a universal program system," *Frequenz*, vol. 33, pp. 14-24, 1979.

[200] A. Lacroix, *Digitale filter, eine einfuhrung in zeitdiskrete signale und systeme (in German)*, Munich, Germany: R. Oldenbourg Verlag, 1980.

[201] V.B. Lawrence and K.V. Mina, "A new and interesting class of limit cycles," *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, Arizona, USA (New York, USA: IEEE 1977), pp. 191-194, 1977.

[202] V.B. Lawrence and K.V. Mina, "Control of limit cycles oscillations in second-order recursive digital filters using constrained random quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 127-134, 1978.

[203] V.B. Lawrence and K.V. Mina, "A new and interesting class of limit cycles in recursive digital filters," *Bell System Techn. J.*, vol. 58, pp. 379-408, 1979.

[204] V.B. Lawrence and D. Mitra, "Digital filters with control of limit cycles," *U.S. Patent no. 4213187*, 1980.

[205] V.B. Lawrence and E.A. Lee,"Quantization schemes for recursive digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Rome, Italy (New York, USA: IEEE 1982), pp. 690-694, 1982.

[206] E. Lehmer, "On the magnitude of the coefficients of the cyclotomic polynomials," *Bull. American Mathematical Society*, vol. 42, pp. 389-392, 1936.

[207] A. Lepschy, G.A. Mian and U. Viaro, "Parameter space quantization in fixed-point digital filters," *Electronics Letters*, vol. 22, pp. 384-386, 1986.

[208] A. Lepschy, G.A. Mian and U. Viaro, "Stability analysis of second-order direct form digital filters with two roundoff quantizers," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 824-826, 1986.

[209] A. Lepschy, G.A. Main and U. Viaro, "Stability of coupled-form digital filters with roundoff quantization," *Alta Frequenza*, vol. 56, pp. 357-360, 1987.

[210] A. Lepschy, G.A. Mian and U. Viaro, "A contribution to stability analysis of second-order direct-form digital filters with magnitude truncation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1207-1210, 1987.

[211] A. Lepschy, G.A. Mian and U. Viaro, "Effects of quantization in second-order fixed-point digital filters with two's complement truncation quantizers," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 461-466, 1988.

[212] A. Lepschy, G.A. Mian and U. Viaro, "Parameter plane quantisation induced by the signal quantisation in second-order fixed-point digital filters with one quantiser," *Signal Processing*, vol. 14, pp. 103-106, 1988.

[213] T.Y. Li and J.A. Yorke, "Period three implies chaos," *American Mathematical Monthly*, vol. 82, pp. 985-992, 1975.

[214] B. Liu and T. Kaneko, "Error analysis of digtal filters realized with floating-point arithmetic," *Proc. of the IEEE*, vol. 57, pp. 1735-1747, 1969.

[215] B. Liu, "Effects of finite wordlength on the accuracy of digital filters - a review," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 670-677, 1971.

[216] B. Liu and M.E. van Valkenburg, "On roundoff error of fixed-point digital filters using sign-magnitude truncation," *Proc. IEEE Int. Symp. Electrical Network Theory*, London, England, pp. 68-69, 1971.

[217] B. Liu and M.E. van Valkenburg, "On roundoff error of fixed-point digital filters using sign-magnitude truncation," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 536-537, 1972.

[218] B. Liu, *Digital filters and the fast fourier transform*, Stroudsburg, Pennsylvania, USA: Dowden, Hutchinson & Ross, Inc., 1975.

[219] B. Liu and M.R. Bateman, "Limit cycle bounds for digital filters with error spectrum shaping," *Proc. 14th Asilomar Conf. Circuits, Syst., Computers*, Pacific Grove, California, USA, ed. D.E. Kirk (North-Hollywood, California, USA: Western Periodicals Comp. 1981), pp. 215-218, 1980.

[220] E.S.K. Liu and L.E. Turner, "Stability, dynamic range and roundoff noise in a new second-order recursive digital filter," *Proc. IEEE Int. Symp. Circuits Syst.*, Rome, Italy (New York, USA: IEEE 1982), pp. 1045-1048, 1982.

[221] E.S.K. Liu and L.E. Turner, "Quantisation effects in second-order wave digital filters," *Electronics Letters*, vol. 19, pp. 487-488, 1983.

[222] E.S.K. Liu and L.E. Turner, "Stability, dynamic range and roundoff noise in a new second-order recursive digital filter," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 815-821, 1983.

[223] H.M. Liu and M.H. Ackroyd, "Suppression of limit cycles in digital filters by random dither," *Radio Electron. Eng.*, vol. 53, pp. 235-240, 1983.

[224] P.H. Lo and Y.C. Jenq, "An ℓm-norm bound for state variables in second-order recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 1170-1171, 1981.

[225] P.H. Lo and Y.C. Jenq, "On the overflow problem in a second-order digital filter," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, Georgia, USA (New York, USA: IEEE 1981), pp. 1221-1226, 1981.

[226] J.L. Long and T.N. Trick, "An absolute bound on limit cycles due to roundoff errors in digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 27-30, 1973.

[227] J.L. Long and T.N. Trick, "A note on absolute bounds on quantization errors in fixed-point implementatons of digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 567-570, 1975.

[228] G. Lucioni, "Alternative method to magnitude truncation in wave digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 106-107, 1987.

[229] G.A. Maria and M.M. Fahmy, "Limit cycle oscillations in a cascade of first- and second-order digital filter sections," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 131-134, 1975.

[230] J. Martens, "Recursive cyclotomic factorization - a new algorithm for calculating the discrete fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 750-761, 1984.

[231] J.E. Mazo, "On the stability of higher-order digital filters which use saturation arithmetic," *Bell System Techn. J.*, vol. 57, pp. 747-763, 1978.

[232] J.H. McClellan and C.M. Rader, *Number theory in digital signal pro cessing*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1979.

[233] D.C. McLernon and R.A. King, "Additional properties of one-dimensional limit cycles," *IEE Proceedings G*, vol. 133, pp. 140-144, 1986.

[234] K. Meerkötter and W. Wegener, "A new second-order digital filter without parasitic oscillations," *Arch. Elektron. Uebertragung.*, vol. AEU-29, pp. 312-314, 1975.

[235] K. Meerkötter, "Realization of limit cycle-free second-order digital filters," *Proc. IEEE Int. Symp. CircuitsSyst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 295-298, 1976.

[236] G. Meyer, "Limit cycles in digital filters with fixed-point arithmetic (in German)," *Nachrichtentechn. Elektr*, vol. 26, pp. 267-273, 1976.

[237] 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.

[238] W.L. Mills, C.T. Mullis and R.A. Roberts, "Digital filter realizations without overflow oscillations," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, Oklahoma, USA (New York, USA: IEEE 1978), pp. 71-74, 1978.

[239] W.L. Mills, C.T. Mullis and R.A. Roberts, "Normal realizations of IIR digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Washington, D.C., USA (New York, USA: IEEE 1979), pp. 340-343, 1979.

[240] W.L. Mills, C.T. Mullis and R.A. Roberts, "Low roundoff noise and normal realizations of fixed-point IIR digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 893-903, 1981.

[241] S.K. Mitra and K. Mondal, "A novel approach to recursive digital filter realization with low roundoff noise," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 299-302, 1976.

[242] D. Mitra, "Criteria for determining if a high-order digital filter using saturation arithmetic is free of overflow oscillations," *Bell System Techn. J.*, vol. 56, pp. 1679-1699, 1977.

[243] D. Mitra, "Large amplitude self-sustained oscillations in difference equations which describe digital filter sections using saturation arithmetic," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 134-143, 1977.

[244] D. Mitra, "A bound on limit cycles in digital filters which exploits a particular structural property of the quantization," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 581-589, 1977.

[245] D. Mitra, "Summary of some results on large amplitude, self-sustained oscillations in high-order digital filter sections using saturation arithmetic," *Proc. IEEE Int. Symp. Circuits Syst.*, Phoe nix, Arizona, USA (New York, USA: IEEE 1977), pp. 195-198, 1977.

[246] D. Mitra, "A bound on limit cycles in digital filters which exploits a particular structural property of the quantization," *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, Arizona, USA (New York, USA: IEEE 1977), pp. 605-610, 1977.

[247] D. Mitra, "The absolute stability of high-order discrete-time systems utilizing the saturation nonlinearity," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 365-371, 1978.

[248] D. Mitra and V.B. Lawrence, "Summary of results on controlled rounding arithmetics, for direct-form digital filters, that eliminate all self-sustained oscillations," *Proc. IEEE Int. Symp. Circuits Syst.*, New York, USA (New York, USA: IEEE 1978), pp. 1023-1028, 1978.

[249] D. Mitra, "Summary of results on the absolute stability of high-order discrete-time systems utilizing the saturation nonlinearity" *Proc. IEEE Int. Symp. Circuits Syst.*, New York, USA (New York, USA: IEEE 1978), pp. 1029-1033, 1978.

[250] D. Mitra and J.R. Boddie, "Limit cycles in floating point digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Tokyo, Japan (New York, USA: IEEE 1979), pp. 374-377, 1979.

[251] D. Mitra and V.B. Lawrence, "Controlled rounding arithmetics for second-order direct-form digital filters, that eliminate all self-sustained oscillations," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 894-905, 1981.

[252] M. Miyata, "Roundoff noise control in time domain for digital filters and oscillators," *Electron. Comm. Japan*, vol. 63-A, no.10, pp. 1-8, 1980.

[253] O. Monkewich and W. Steenaart, "Deadband effects and limit cycles in stored-product digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Chicago, Illinois, USA (New York, USA: IEEE 1981), pp. 813-816, 1981.

[254] H.D. Montgomery, "A non-linear digital oscillator," *Proc. IEEE Int. Conf. Comm.*, Philadelphia, Pennsylvania, USA (New York, USA: IEEE 1972), pp. 33.3-33.8, 1972.

[255] P.R. Moon, "Limit cycle suppression by diagonally dominant Lyapunov functions in state-space digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Montreal, Canada (New York, USA: IEEE 1984), pp. 1082-1085, 1984.

[256] D.R. Morgan and A. Aridgides, "Discrete-time distortion analysis of quantized sinusoids," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 323-326, 1985.

[257] C.T. Mullis and R.A. Roberts, "Roundoff noise in digital filters: Frequency transformations and invariants," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 539-550, 1976.

[258] C.T. Mullis and R.A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551-561, 1976.

[259] C.T. Mullis and R.A. Roberts, "Filter structures which minimize round-off noise in fixed-point digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, Pennsylvania, USA (New York, USA: IEEE 1976), pp. 505-508, 1976.

[260] D.C. Munson, jr. and B. Liu, "Narrow-band recursive filters with error spectrum shaping," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Washington, D.C., USA (New York, USA: IEEE 1979), pp. 367-370, 1979.

[261] D.C. Munson, jr. and B. Liu, "Low-noise realizations for narrow-band recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 41-54, 1980.

[262] D.C. Munson, jr. and B. Liu, "ROM/ACC realization of digital filters for poles near the unit circle," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 147-151, 1980.

[263] D.C. Munson, jr., "Accessibility of zero-input limit cycles," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1027-1032, 1981.

[264] D.C. Munson, jr. and B. Liu, "Narrow-band recursive filters with error spectrum shaping," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 160-163, 1981.

[265] D.C. Munson, jr., "Determining exact maximum amplitude limit cycles in digital filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, Georgia, USA (New York, USA: IEEE 1981), pp. 1231-1234, 1981.

[266] D.C. Munson, jr., "Accessibility of zero-input limit cycles," *Proc. IEEE Int. Symp. Circuits Syst.*, Chicago, Illinois, USA (New York, USA: IEEE 1981), pp. 821-824, 1981.

[267] D.C. Munson, jr., J.H. Strickland and T.P. Walker, "Maximum amplitude zero-input limit cycles in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 266-275, 1984.

[268] B.H. Nam and N.H. Kim, "Stability analysis of modified coupled-form digital filter using a constructive algorithm," *Trans. Korean IEE*, vol. 34, pp. 430-435, 1985.

[269] D.T. Nguyen, "Overflow oscillations in digital lattice filters," *IEE Proceedings G*, vol. 128, pp. 269-272, 1981.

[270] A. Nishihara, "Design of limit cycle-free digital biquad filters," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japan (New York, USA: IEEE 1986), pp. 517-520, 1986.

[271] H.A. Ojongbede, "Limit cycle constraints for recursive digital filter design," *Electronics Letters*, vol. 6, pp. 698-700, 1970.

[272] A.V. Oppenheim and C.J. Weinstein, "Effects of finite register length in digital filtering and the Fast Fourier Transform," *Proc. of the IEEE*, vol. 60, pp. 957-976, 1972.

[273] A.V. Oppenheim and R.W. Schafer, *Digital signal processing*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1975.

[274] A.V. Oppenheim, *Applications of digital signal processing*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1978.

[275] E.L. O'Neill, *Introduction to statistical optics*, London, England: Addison-Wesley Publ. Comp., Inc., 1963.

[276] S.R. Parker and S.F. Hess, "Limit cycle oscillations in digital filters," *IEEE Trans.Circuit Theory*, vol. CT-18, pp.687-697, 1971.

[277] S.R. Parker and S.F. Hess, "Heuristic bands for the frequency of digital oscillators due to quantization noise," *Electronics Letters*, vol. 8, pp. 86-87, 1972.

[278] S.R. Parker and S.F. Hess, "Canonic realizations of second-order digital filters due to finite precision arithmetic," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 410-413, 1972.

[279] S.R. Parker, "The phenomena of quantization error and limit cycles in fixed point digital filters," *Proc. 28th Nat. Electr. Conf.*, Chicago, USA, ed. R.E. Horton (Oak Brook, Illinois, USA: Nat. Electr. Conf. 1972), pp. 38-42, 1972.

[280] S.R. Parker and S. Yakowitz, "A general method for calculating quantization error bounds in fixed-point multivariable digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 570-572, 1975.

[281] S.R. Parker, "Limit cycles and correlated noise in digital filters," *Digital Signal Processing*, ed. J.K. Aggarwal (North-Hollywood California, USA: Western Periodicals Comp. 1979), pp. 117-179, 1979.

[282] S.R. Parker and F.A. Perry, "Hidden limit cycles and error bounds in wave digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Tokyo, Japan (New York, USA: IEEE 1979), pp. 372-373, 1979.

[283] A.M.B. Pavani and J. Szczupak, "A mathematical model for digital filters under limit cycle condition," *Proc. IEEE Int. Symp. Circuits Syst.*, Kyoto, Japan (New York, USA: IEEE 1985), pp. 1615-1616, 1985.

[284] J. Payan and J.I. Acha, "Parasitic oscillations in normalized digital structures," *Int. J. Electronics*, vol. 60, pp.591-595, 1986.

[285] G. Peceli, "Finite wordlength effects in digital filters," *Periodica Polytechn. Electr. Eng.*, vol. 28, pp. 191-200, 1984.

[286] A. Peled and B. Liu, *Digital signal processing, theory , design and implementation*, New York, USA: John Wiley & Sons, Inc., 1976.

[287] A.N. Philippou, G.E. Bergum and A.F. Horadam, *Fibonacci numbers and their applications*, Dordrecht, The Netherlands: D. Reidel Publ. Comp., 1986.

[288] K.P. Prasad and P.S. Reddy, "Limit cycles in second-order digital filters," *J. Inst. Electron. Telecomm. Eng. (India)*, vol. 26, pp. 85-86, 1980.

[289] L.R. Rabiner, "Terminology in digital signal processing," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 322-337, 1972.

[290] L.R. Rabiner and B. Gold, *Theory and applications of digital signal processing*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1975.

[291] T. Rado, *Subharmonic functions*, New York, USA: Chelsa Publ. Comp., 1949.

[292] M.H. Rahman, G.A. Maria and M.M. Fahmy, "Bounds on zero-input limit cycles in all-pole digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 189-192, 1976.

[293] R. Ramnarayan and F. Taylor, "Limit cycles in large moduli residue number system digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 912-916, 1986.

[294] P. Rashidi and R.E. Bogner, "Suppression of limit cycles oscillations in second-order recursive digital filters," *Aust. Telecomm. Res.*, vol. 12, pp. 8-16, 1978.

[295] M. Renfors, B. Sikström and L. Wanhammer, "LSI implementation of limit cycle free digital filters using error-feedback techniques," *Proc. EUSIPCO-83, 2nd European Signal Processing Conf.*, Erlangen, Germany, ed. H.W. Schüssler (Amsterdam, The Netherlands: North-Holland 1983), pp. 107-110, 1983.

[296] J.H.F. Ritzerfeld, "A condition for the overflow stability of second-order digital filters that is satisfied by all scaled state-space structures," submitted to *Trans. IEEE Circuits Syst.*.

[297] R.A. Roberts and C.T. Mullis, *Digital signal processing*, London, England: Addison-Wesley Publ. Comp., Inc., 1987.

[298] H. Samueli and A.N. Willson jr., "Almost periodic forced overflow oscillations in digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Houston, Texas, USA (New York, USA: IEEE 1980), pp. 1108-1112, 1980.

[299] H. Samueli and A.N. Willson jr., "Almost period P sequences and the analysis of forced overflow oscillations in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-29, pp. 510-515, 1982.

[300] H. Samueli and A.N. Willson jr., "Nonperiodic forced overflow oscillations in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-30, pp. 709-722, 1983.

[301] I.W. Sandberg, "Floating-point roundoff accumulation in digital filter realizations," *Bell System Techn. J.*, vol. 46, pp. 1775-1791, 1967.

[302] I.W. Sandberg, "A theorem concerning limit cycles in digital filters," *Proc. 7th Annual Allerton Conf. Circuit Syst. Theory*, Monticello, Illinois, USA, ed. A.H. Hadad et al. (Urbana, Illinois, Dept. of Electrical Eng., Univ. of Illinois 1969), pp.63-68, 1969.

[303] I.W. Sandberg and J.F. Kaiser, "A bound on limit cycles in fixed-point implementations of digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 110-112, 1972.

[304] I.W. Sandberg, "The zero-input response of digital filters using saturation arithmetics," *IEEE Trans. Circuits Syst.*, vol. CAS-26, pp. 911-915, 1979.

[305] H.W. Schüssler, *Digitale Systeme zur Signalverarbeitung (in german)*, Berlin, Germany: Springer-Verlag, 1973.

[306] P.K. Sim and K.K. Pang, "On the asymptotic stability of a N-th order nonlinear recursive digital filter," *IREECON Int. Sydney '83: Digest of papers 19th Int. Electr. Conv. & Exhib.*, Sydney, Australia (Sydney, Australia: Inst. Radio & Electron. Eng., Australia 1983), pp. 108-110, 1983.

[307] P.K. Sim and K.K. Pang, "Effects of input-scaling on the asymptotic overflow stability properties of second-order recursive digital filters," *IEEE Trans. Circuits Syst.*, vol.CAS-32, pp. 1008-1015, 1985.

[308] P.K. Sim and K.K. Pang, "Design criterion for zero-input asymptotic overflow stability of recursive digital filters in the presence of quantization," *Proc. IEEE Int. Symp. Circuits Syst.*, Kyoto, Japan (New York, USA: IEEE 1985), pp. 1607-1611, 1985.

[309] P.K. Sim and K.K. Pang, "Quantization phenomena in a class of complex biquad recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 892-899, 1986.

[310] P.K. Sim and K.K. Pang, "Conditions for overflow stability of a class of complex biquad digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 471-479, 1987.

[311] P.K. Sim, "Relationship between input-scaling and stability of the forced-response of recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 506-511, 1988.

[312] P.J.M. Simons and M.P.G. Otten, "Intermodulation due to magnitude truncation in digital filters," *Proc. 8th European Conf. Circuit Theory Design*, Paris, France ed. R. Gerber (Amsterdam, The Netherlands: Elseviers Science Publishers 1987) pp. 723-728, 1987.

[313] V. Singh, "Formulation of a criterion for the absence of limit cycles in digital filters designed with one quantizer," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1062-1064, 1985.

[314] V. Singh, "A new realizability condition for limit cycle free state-space digital filters employing saturation arithmetic," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1070-1071, 1985.

[315] V. Singh, "Realization of two's complement overflow limit cycle free state-space digital filters: a frequency-domain viewpoint," *IEEE Trans. Circuits Syst.*, vol. CAS-33, pp. 1042-1044, 1986.

[316] V. Singh, "On the realization of two's complment overflow limit cycle free state-space digital filters," *Proc. of the IEEE*, vol. 74, pp. 1287-1288, 1986.

[317] J.O. Smith, "Elimination of limit cycles in time-varying lattice filters," *Proc. IEEE Int. Symp. Circuits Syst.*, San Jose, California, USA (New York, USA: IEEE 1986), pp. 197-200, 1986.

[318] G. Spahlinger, "Suppression of limit cycles in digital filters by statistical rounding," *Proc. IEEE Int. Symp. Circuits Syst.*, Kyoto, Japan (New York, USA: IEEE 1985), pp. 1603-1606, 1985.

[319] S. Sridharan and D. Williamson, "Comments on 'Suppression of limit cycles in digital filters designed with one magnitude-truncation quantizer'," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 235-236, 1984.

[320] A. Sripad and D.L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 442-448, 1977.

[321] B. Steinle and P. Gruber, "Comments on upper bounds on limit cycles in digital filters of second-order," *Signal Processing*, vol. 8, pp. 415-422, 1985.

[322] W.H. Storzbach, "Forced oscillations in recursive digital filters" *Proc. IEEE Int. Symp. Circuit Theory*, North-Hollywood, California, USA (New York, USA: IEEE 1972), pp. 233-236, 1972.

[323] M. Takizawa, H. Kishi and N. Hamada, "Synthesis of lattice digital filters by the state-variable method," *Electron. Comm. Japan*, vol. 65-A, no.4, pp. 27-36, 1982.

[324] E.C. Tan, "Limit cycles and similarity transformation of simpel digital circuits," *Int. J. Electronics*, vol. 55, pp.767-773, 1983.

[325] V. Tavsanoglu and L. Thiele, "Simultaneous minimization of round-off noise and sensitivity in state-space digital filters," *Proc. IEEE Int. Symp. Circuits Syst.*, Newport Beach, California, USA (New York, USA: IEEE 1983), pp. 815-818, 1983.

[326] V. Tavsanoglu and L. Thiele, "Optimal design of state-space digital filters by simultaneous minimization of sensitivity and round-off noise," *IEEE Trans. Circuits Syst.*, vol. CAS-31, pp. 884-888, 1984.

[327] L. Thiele, "Design of sensitivity and roundoff noise optimal state space discrete systems," *Int. J. Circuit Theory Appl.*, vol. 12, pp. 39-46, 1984.

[328] T. Thong and B. Liu, "Limit cycles in combinatorial filters using two's complement truncation arithmetic," *Proc. 8th Asilomar Conf. Circuits, Syst., Computers*, Pacific Grove, California, USA (North-Hollywood, California, USA: Western Periodicals Comp. 1975), pp. 51-55, 1974.

[329] T. Thong and B. Liu, "Limit cycles in combinatorial implementation of digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 248-256, 1976.

[330] Y.Z. Tsypkin, "Frequency criteria for the absolute stability of non-linear sampled-data systems," *Automation and Remote Control*, vol. 25, pp. 261-267, 1964.

[331] Y.Z. Tsypkin, "A criterion for absolute stability of automatic pulse systems with monotonic characteristics of the nonlinear element," *Soviet Physics-Doklady*, vol. 9, pp. 263-266, 1964.

[332] L.E. Turner and L.T. Bruton, "Elimination of zero-input limit cycles by bounding the state transition matrix," *Proc. IEEE Int. Symp. Circuits Syst.*, Phoenix, Arizona, USA (New York, USA: IEEE 1977), pp. 199-202, 1977.

[333] L.E. Turner and L.T. Bruton, "Elimination of zero-input limit cycles by bounding the state transition matrix," *Int. J. Circuit Theory Appl.*, vol. 7, pp. 97-111, 1979.

[334] L.E. Turner and L.T. Bruton, "Elimination of granularity and overflow limit cycles in minimum norm recursive digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 50-53, 1980.

[335] L.E. Turner and L.T. Bruton, "Elimination of constant-input limit cycles in recursive digital filters using a generalised minimum norm," *Proc. IEEE Int. Symp. Circuits Syst.*, Chicago, Illinois, USA (New York, USA: IEEE 1981), pp. 817-820, 1981.

[336] L.E. Turner, "Second-order recursive digital filter that is free from all constant-input limit cycles," *Electronics Letters*, vol. 18, pp. 743-745, 1982.

[337] L.E. Turner, "Elimination of constant-input limit cycles in recursive digital filters using a generalised minimum norm," *IEE Proceedings G*, vol. 130, pp. 69-77, 1983.

[338] Z. Ünver and K. Abdullah, "A tighter practical bound on quantization errors in second-order digital filters with complex conjugated poles," *IEEE Trans. Circuits Syst.*, vol. CAS-22, pp. 632-633, 1975.

[339] P.P. Vaidyanathan, "The discrete-time bounded-real lemma in digital filtering," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 918-924, 1985.

[340] P.P. Vaidyanathan and V. Liu, "An improved sufficient condition for absence of limit cycles in digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-34, pp. 319-322, 1987.

[341] G. Verkroost and H.J. Butterweck, "Suppression of parasitic oscillations in wave digital filters and related structures by means of controlled rounding," *Arch. Elektron. Uebertragung.*, vol. AEU-30, pp. 181-186, 1976.

[342] G. Verkroost and H.J. Butterweck, "Suppression of parasitic oscillations in wave digital filters and related structures by means of controlled rounding," *Proc. IEEE Int. Symp. Circuits Syst.*, Munich, Germany (New York, USA: IEEE 1976), pp. 628-629, 1976.

[343] G. Verkroost, "A general second-order digital filter with controlled rounding to exclude limit cycles for constant-input signals," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 428-431, 1977.

[344] G. Verkroost, "Een tweede-orde digitaal filter waarin door middel van gestuurde kwantisering 'limit cycles' voorkomen worden (in dutch)," *Nederl. Elektron. Radio Genootschap, NERG*, vol. 42, pp. 111-114, 1977.

[345] G. Verkroost and G.J. Bosscha, "On the measurement of quantization noise in digital filters," *IEEE Trans. Circuits Syst.*,vol. CAS-31, pp. 222-223, 1984.

[346] G. Verkroost, "Noise caused by sampling-time jitter with applications to sampling-frequency conversion," *Proc. IASTED Int. Symp. Appl. Signal Processing Digital filtering*, Paris, France, ed M.H. Hamza (Anaheim, California, USA: Acta PRESS 1985), pp. 275-276, 1985.

[347] W. Wegener, *Entwurf von wellen digital filters mit minimalem reali sierungsaufwand (in german)*, Bochum, Germany: Bochum University of Technology, Doctor's Thesis, 1980.

[348] M.J. Werter, "Elimination of subharmonics in periodically excited recursive digital filters," *Proc. IASTED Int. Symp. Appl. Signal Processing Digital filtering*, Paris, France, ed M.H. Hamza (Anaheim, California, USA: Acta PRESS 1985), pp. 108-112, 1985.

[349] M.J. Werter, "Suppression of subharmonics in digital filters for discrete-time periodic input signals with period P or a divisor of P," *Proc. EUSIPCO-86, 3rd European Signal Processing Conf.*, The Hague, The Netherlands, ed. I.T. Young et al. (Amsterdam, The Netherlands: North-Holland 1986), pp. 179-182, 1986.

[350] M.J. Werter, "Digital filter sections which suppress subharmonics for discrete-time periodic input signals with period P or a divisor of P," *Proc. IEEE Int. Symp. Circuits Syst.*, San Jose, California, USA (New York, USA: IEEE 1986), pp. 863-866, 1986.

[351] M.J. Werter, "A new decomposition of discrete-time periodic signals," *Proc. 8th European Conf. Circuit Theory Design*, Paris, France, ed. R. Gerber (Amsterdam, The Netherlands: Elseviers Science Publishers 1987) pp. 139-144, 1987.

[352] M.J. Werter, "New zero-input overflow stability proofs based on Lyapunov theory," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glascow, Scotland (to be published), 1989.

[353] S.A. White, "Quantizer-induced digital controller limit cycles," *IEEE Trans. Automatic Control*, vol. AC-14, pp. 430-432, 1969.

[354] B. Widrow, "Adaptive noise cancelling: principles and applications," *Proc. of the IEEE*, vol. 63, pp. 1692-1719, 1975.

[355] B. Widrow and S.D. Stearns, *Adaptive signal processing*, Englewood Cliffs, New Jersey, USA: Prentice Hall, Inc., 1985.

[356] R.L. Wigington, "A new concept in computing," *Proc. of the IRE*, vol. 47, pp. 516-523, 1959.

[357] D. Williamson, P.G. McCrea and S. Sridharan, "Residue feedback in digital filters using fractional coefficients and block floating point arithmetic," *Proc. IEEE Int. Symp. Circuits Syst.*, Newport Beach, California, USA (New York, USA: IEEE 1983), pp. 831-834, 1983.

[358] D. Williamson and S. Sridharan, "Residue feedback in digital filters using fractional feedback coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 477-483, 1985.

[359] D. Williamson and S. Sridharan, "Residue feedback in ladder and lattice filter structures," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, USA (New York, USA: IEEE 1985), pp. 53-56, 1985.

[360] D. Williamson and S. Sridharan, "Error-feedback in a class of orthogonal polynomial digital filter structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 1013-1016, 1986.

[361] A.N. Willson, jr., "Some effects of quantization and adder overflow on the forced response of digital filters," *Bell System Techn. J.*, vol. 51, pp. 863-887, 1972.

[362] A.N. Willson, jr., "Limit cycles due to adder overflow in digital filters," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 342-346, 1972.

[363] A.N. Willson, jr., "Limit cycles due to adder overflow in digital filters," *Proc. IEEE Int. Symp. Circuit Theory*, North-Hollywood, California, USA (New York, USA: IEEE 1972), pp. 223-227, 1972.

[364] A.N. Willson, jr., "A stability criterion for non-autonomous difference equations with applications to the design of a digital FSK oscillator," *IEEE Trans. Circuits Syst.*, vol. CAS-21, pp. 124-130, 1974.

[365] A.N. Willson, jr., "Error-feedback circuits for digital filters," *Electronics Letters*, vol. 12, pp. 450-452, 1976.

[366] A.N. Willson, jr., "Computation of the periods of forced overflow oscillations in digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 89-97, 1976.

[367] J. Wisliceny, "Estimation of the maximum amplitude of quantisation conditioned limit cycles in linear recursive digital filters (in German)," *Zeitschr. Elektron.*, Inf., Energietechn., vol. 10, pp. 330-338, 1980.

[368] I.H. Witten and P.G. McCrea, "Suppressing limit cycles in digital incremental computers," *IEEE Trans. Circuits Syst.*, vol. CAS-28, pp. 723-730, 1981.

[369] K.M. Wong and R.A. King, "Method to suppress limit cycle oscillations in digital filter," *Electronics Letters*, vol. 10, pp. 55-57, 1974.

[370] S. Yakowitz and S.R. Parker, "Computation of bounds for digital filter quantization errors," *IEEE Trans. Circuit Theory*, vol. CT-20, pp. 391-396, 1973.

Curriculum Vitae

The author of this doctoral dissertation, ir. Michael J. Werter, was born in Eindhoven, The Netherlands, on January 17, 1960. He received the Ingenieur degree in electrical engineering (cum laude) from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1984. In that same year he joined the network group of the Department of Electrical Engineering at this university, where he carried out the research reported in this thesis in the project group Digital Signal Processing under supervision of prof. dr. -ing. H.J. Butterweck.

List of publications by M.J. Werter

1. M.J. Werter, "*Minimal RLC-impedance synthesis with the theory on Hurwitz polynomials (in dutch)*," Master's Thesis, Report ET-7-84, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, 1984.
2. S. Tirtoprodjo and M.J. Werter, "On the generation of a complete catalogue of higher-order Hurwitz filter functions," *Proc. 27th Midwest Symp. Circuits and Systems*, Morgantown, West Virginia, USA, ed R.E. Swartwout (North-Hollywood, California, USA: Western Periodicals Company 1984), pp. 595-598, 1984.
3. M.J. Werter, "Elimination of subharmonics in periodically excited recursive digital filters," *Proc. IASTED Int. Symp. Applied Signal Processing and Digital filtering*, Paris, France, ed. M.H. Hamza, ISBN 0-88986-084-X (Anaheim, California, USA: Acta Press 1985), pp. 108-112, 1985.
4. M.J. Werter, "Suppression of subharmonics in digital filters for discrete-time periodic input signals with period P or a divisor of P," *Proc. EUSIPCO-86, 3rd European Signal Processing Conference.*, The Hague, The Netherlands, ed I.T. Young et al. (Amsterdam, The Netherlands: North Holland 1986), pp. 179-182, 1986.
5. M.J. Werter, "Digital filter sections which suppress subharmonics for discrete-time periodic input signals with period P or a divisor of P," *Proc. IEEE Int. Symp. on Circuits and Systems*, San Jose, California, USA (New York, USA: IEEE 1986), pp. 863-866, 1986.
6. M.J. Werter, "A new decomposition of discrete-time periodic signals," *Proc. 8th European Conf. on Circuit Theory and Design*, Paris, France, ed. R. Gerber (Amsterdam, The Netherlands: Elsevier Science publishers 1987) pp. 139-144, 1987.
7. H.J. Butterweck, J.H.F. Ritzerfeld and M.J. Werter, "*Finite word-length effects in digital filters - a review*," EUT Report 88-E-205, ISBN 90-6144-205-2, Eindhoven University of Technology, Eindhoven, The Netherlands, 1988.
8. M.J. Werter, "*Suppression of parasitic oscillations due to overflow and quantization in recursive digital filters*," ISBN 90-9002583-9, Doctor's Thesis, Eindhoven University of Technology, Eindhoven, The Netherlands, 1989.
9. H.J. Butterweck, J.H.F. Ritzerfeld and M.J. Werter, "Finite word-length effects in digital filters," *Arch. Elektron Uebertragungstechn.*, (to be published), Feb. 1989.
10. M.J. Werter and J.H.F. Ritzerfeld, "New zero-input overflow stability proofs based on Lyapunov theory," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Glascow, Scotland (to be published), May 1989.

STELLINGEN

behorende bij het proefschrift van

M.J. Werter

Eindhoven, 10 februari 1989

1. Er bestaat geen universele methode voor het volledig onderdrukken van alle parasitaire oscillaties, die ontstaan ten gevolge van overflow en kwantisatie in recursieve digitale filters.

   dit proefschrift

2. In tegenstelling tot de gebruikelijke zienswijze, kunnen de niet-lineaire verschijnselen, die ontstaan ten gevolge van overflow en kwantisatie in recursieve digitale filters, strikt genomen analytisch niet onafhankelijk van elkaar beschouwd worden.

   A.N. Willson, jr., "Some effects on quantization and adder overflow on the forced response of digital filters," Bell System Techn. J., vol 51, pp. 863-872, 1972

3. Het wijdverbreide principe om de kwantisatieruis van een recursief digitaal filter dat volledig vrij is van limit cycles te berekenen met een model dat gebruik maakt van een witte ruisbron is onjuist, indien het in deze filters toegepaste kwantisatie-mechanisme is gebaseerd op magnitude truncation of een aanverwant energie-reducerend afrondmechanisme.

   E.S.K. Liu and L.E. Turner, "Quantisation effects in second-order digital filters," Electronics Letters, vol. 19, pp. 487-488, 1983

4. Chaos ten gevolge van overflow correctie in een digitaal filter kan in tegenstelling tot de vermelding in recente literatuur nooit in een werkelijke signaalprocessor optreden.

   L.O. Chua and T. Lin, "Chaos in digital filters," IEEE Trans. Circuits Syst., vol. CAS-35, pp. 648-658, 1988

5. De bewering dat bij een gegeven willekeurige energiefunctie in een recursief digitaal filter er altijd een geschikt gekozen gestuurde kwantisatie mechanisme bestaat zodat de energiefunctie een Lyapunov-functie wordt, is onwaar.

   M. Miyata, "Roundoff noise control in time domain for digital filters and oscillators," Electron. Comm. Japan, vol. 63-A, no. 10, pp. 1-8, 1980.

6. Het gepatenteerde kwantisatie-mechanisme, dat via een uitgebreide versie van het "controlled-rounding" principe een schakeling oplevert dat niet alleen zonder excitatie maar ook bij constante en alternerende excitatie vrij is van parasitaire oscillaties, werkt niet. Integendeel, zelfs de stabiliteit bij het ontbreken van een excitatie gaat ten gevolge van deze uitbreiding verloren.

   V.B. Lawrence and K.V. Mina, "Digital filters with control of limit cycles," U.S. Patent no. 4213187, 1980

7. De met de invoering van de z-transformatie gevonden gesloten uitdrukking voor het berekenen van een willekeurig element uit de rij van Fibonacci-getallen heeft de magie rond deze getallen niet kunnen breken; de Fibonacci-rij blijft vele wetenschappers in zijn ban houden.

   The Fibonacci Quarterly, Official publication of the Fibonacci Association, Editor G.E. Bergum, South Dakota State University, Brookings, USA

8. In de huidige situatie worden verkeerslichten "stop"-lichten genoemd, in de betekenis dat rood een verplichte wachtperiode aangeeft. Het zou de verkeers-vriendelijkheid veel verbeteren indien men deze lichten zou vervangen door "voorrangs"-lichten: een tijdperiodieke wisseling van de voorrangsregeling.

9. Het kaartspel Bridge wijkt in zoverre af van andere gerenommeerde denk-sporten als dammen en schaken, dat een regelrechte blunder nogal eens tot een absolute topscore kan leiden.