An Analysis of Digital Filters having MULTI~~Multiple~~ Shift Sequences in a Sampling Period

Kon Max Wong

A thesis submitted for the degree of

DOCTOR OF PHILOSOPHY

of the

UNIVERSITY OF LONDON

Department of Electrical Engineering,

Imperial College of Science and Technology,

London, 1974

亮父嘗言仁智乃立身之本
余敬力科學鑽研作此書
啟求發諸仁而達諸智以
念吾父復冀毫負其望焉

To my Father,
to whom a good life was one inspired by
love and guided by knowledge, in the
hope that this thesis may contribute
some knowledge for the benefit of mankind.

ABSTRACT

In a conventional digital filter, the digital signals in the filter are shifted forward from one register to another at the pulse repetition (sampling) rate while the output samples also appear at the rate of sampling. Also, all the filter coefficients remain unchanged throughout the filtering process. However, if the shifting is continued several times during the pulse repetition interval so that the signals stored in the registers re-circulate in the internal paths of the filter during each sample interval, then the processing leads to multiple output signals. Furthermore, if the filter multiplication coefficients are also allowed to take on different values for different shift sequences, the filter will possess some useful properties.

It is the object of this thesis to introduce such a "multi-rate" digital filter and to study its behaviour. The quantization errors for such a device are considered analytically. The investigation into these errors reveals many interesting properties of the multi-rate digital filter which are confirmed by computer simulations. Under certain circumstances, these properties render the multi-rate filter more advantageous than the conventional digital filter. Possible engineering applications of such a filter are suggested.

ACKNOWLEDGEMENTS

I wish to express my gratitude to my supervisor Mr. R.A. King of the Electrical Engineering Department, Imperial College, London for his encouragement and guidance during the course of this work. The association with Imperial College provided me the fortunate circumstances under which most helpful discussions with Dr. A.G. Constantinides, Mr. J.M. Howl and Dr. W. Saraga were available.

Many of my colleagues both at Imperial College and at Plessey Telecommunications Research Ltd. have put forth many inspiring suggestions, in particular, Drs. R.S. Couchman, V.B. Lawrence, E.M. Olcayto, M. Tomlinson and Mr. A.E. Methiwalla are gratefully acknowledged for their interest, encouragement and discussions. I am also much indebted to Mrs. N. Methiwalla for helping to type the manuscript.

Finally, this work could not have been completed without the generous and continuous financial support of Plessey Telecommunications Research Limited.

CONTENTS

VIII.3 CODA - On the Value of Scientific Research

References

CONVENTIONS AND SYMBOLS

The following system of numbering and cross-references is used in this thesis: Each chapter is labelled with a Roman numeral and is sub-divided into sections. All sections, examples, figures and equations within a chapter are numbered consecutively starting from 1. Hence "section V.3" refers to section 3 of chapter V, "fig (VII.9)" refers to figure 9 of chapter VII. Equations are generally referred to by their numbers; thus "substituting into (III.51)" means "substituting into the fifty-first equation of chapter III". At the end of the volume, there is a list of references. When such a reference is made, it is denoted by a number in the braces { }. Thus {31}, {42} refers to references 31 and 42 of the list.

The following is a list of principal symbols appearing in the thesis.

1. Boldface letters denote vectors and matrices, e.g. $\mathbf{y}$, $\mathbf{A}$, $\boldsymbol{\zeta}$

2. Greek and italic type are used for scalar-valued variables, functions and operators, e.g. α, $F(z)$, $E[u(k)]$.

3. Capital letters are used to denote the z-transforms; e.g. $H(z)$ is the z-transform of $h(n)$.

4. Capital script letters denote sets (spaces), e.g. $\mathcal{R}$

5. The operator of z-transform is denoted by $\mathcal{Z}(.)$ with the dot standing for an undesignated variable.

6. Superscript asterisk denotes the complex conjugate of a number, e.g. λ^*.

7. Superscript minus one denotes the reciprocal of a quantity, the inverse of a matrix or a z-transformation e.g. z^{-1}, $\mathbf{A}^{-1}$, $\mathbf{g}^{-1}$

8. Superscript (^{T}) denotes the transposed a matrix, e.g. $\mathbf{B}^{T}$

9. A prime over a continuous function of a single variable generally denotes the derivative w.r.t. the variable, e.g. $f'(t)$, $f^{n}(t)$ denotes the nth derivative of f w.r.t. t.

In particular, the following are some symbols and abbreviations with special meanings:

$\triangleq$ equals by definition

$\Rightarrow$ implies

$\Leftarrow$ is implied by

$\Leftrightarrow$ implies and implied by

iff if and only if

$\forall$ for all

$\approx$ is approximately equal to

$\equiv$ is equivalent to

$\langle \bullet \rangle$ matrix formed by taking the absolute values of each of its elements.

$|.|$ absolute value

$\|.\|$ norm

det determinant

tr trace

s.t. such that

w.r.t. with respect to

L.H.S. Left hand side

R.H.S. Right hand side

Q.E.D. *quod erat demonstrandum*

δ_{ij} Kronecker delta

$\delta(t)$ delta function

j denotes the square root of -1; an integer

$\mathbf{0}$ null matrix, zero vector

$\mathbf{I}$ identity matrix

Φ state-transition matrix.

CHAPTER I

INTRODUCTION - DIGITAL FILTERS

I.1 Introductory Remarks

Digital filtering techniques have been in use for some time in sampled-data control systems {28}, {39}, {44}, {53}, {60}. In the sampled-data control systems, the digital filter has been implemented with the use of a digital computer. The extension of digital filtering techniques to other areas has been limited to those where the use of a digital computer was practical.

In recent years, digital filters have been used more and more for real time signal processing. By real time, it is implied that digital processing takes place fast enough so that the output of the digital filter is available for direct control or observation in a larger system. Digital filters are constructed using digital logic computers as their basic building blocks and the rapid advance in the development of solid state devices has made such digital filters practical. The development of large scale circuit integration (LSI) promises to make these systems even more economical.

Digital filters have many advantages which recommend their application in place of passive or active filters. The most important advantage is the very accurate drift free operation which is possible. This allows the realization of stable filters with very high Q's or with extremely long time constants. There is negligible drift with temperature or time, since the filter characteristics are as stable as the digital clock source,

commonly a crystal-controlled oscillator, with stability greater than one part per million for large variation of time. Additional advantages lie in the ease with which the filter characteristic may be changed, making them particularly useful as time-varying filters with adaptive or frequency tracking requirements. With some digital filter types, a linear phase characteristic is readily obtained, resulting in improved transient response and constant delay characteristics. Filters for very low frequencies are easily constructed, with a large size reduction as compared to that of passive filters. Digital filters contain no reactive components. The elimination of the accuracy and drift problems associated with these components will be well appreciated by design engineers.

Basically, a digital filter is comprised of three units, an analogue-to-digital (A/D) converter, a digital calculator, and a digital-to-analogue (D/A) converter (Fig.I.1a).

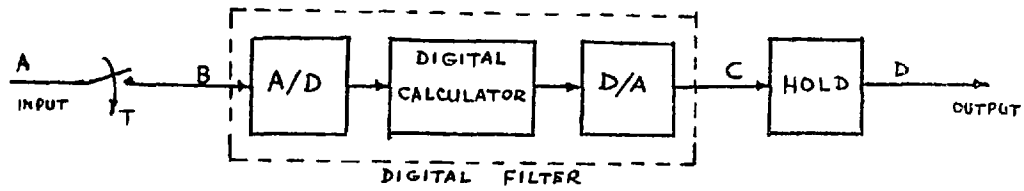


Fig I.1a Digital Filter with Input Sampler and Output Hold Circuit

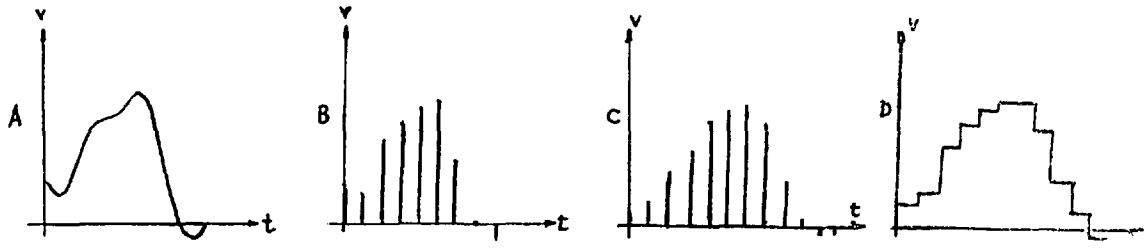


Fig I.1b Typical Voltage Waveforms at Points A,B,C,D

The input and output signals of the digital filter are narrow amplitude-modulated pulses, one pulse per sampling period T. At the time $t=nT$, the continuous input signal is momentarily sampled, and the pulse $u(nT)$ appears at the input to the digital filter. In the A/D converter, this pulse amplitude is converted into a digital word. This digital word is a coded sequence of binary digits (bits), which represents the amplitude $u(nT)$. The length of the word, i.e., the number of bits it contains determines the accuracy of the representation. The digital calculations are performed with these words, and the calculator output word is inserted into a D/A converter to produce the output pulse $y(nT)$ of the digital filter. A holding circuit follows the digital filter to convert the pulse stream to a continuous output signal as shown (fig. I.1b). Further analogue filtering may be desired to remove the signal components resulting from the step approximation which results.

I.2 Digital Filter Operation

The operation of a digital filter is defined by a difference equation. This equation defines the output pulse amplitude $y(nT)$ as a function of the present input pulse $u(nT)$ and any number of past input pulses and output pulses. The operations are performed in the digital calculator, with the words representing the required past input pulses and output pulses being stored in digital shift registers. The usual practice is to simplify the notation to $u(n)$ and $y(n)$, with the understanding that n refers to $t = nT$. This will be done in the remaining parts of this thesis. A general formula for the difference equation is

$$y(n) = \sum_{i=0}^{N} a_i u(n-i) - \sum_{i=1}^{M} b_i y(n-i) \qquad (I.1)$$

If $y(n)$ is a function of only the present and past input pulses, the filter is termed *non-recursive*. If the past output pulses are included as well, then the filter is of the *recursive* type. It is then clear that recursive filters have infinite-duration impulse responses while non-recursive filters have finite-duration impulse responses. An important difference between recursive and non-recursive digital filters exist in the range of M and N encountered in typical applications. Recursive filters usually meet the kinds of specifications arising in practice with at most 10 or 20 coefficients. Thus the computation required to produce each output, given a new input, is of the order of 10 to 20 multiplications and additions per sample point. In contrast, non-recursive filters, when used to realize complex-shaped frequency responses, may require several hundred coefficients. Here in this thesis, it is the recursive filters that are considered.

Analysis of digital filter is carried out with the use of z-transform. The z-transform, $X(z)$, of a function $x(t)$ is defined as {[29]}

$$X(z) = \mathcal{Z}\,[x(t)] = \sum_{n=0}^{\infty} x(n).z^{-n} \qquad \text{(I.2)}$$

where $z = e^{sT}$ and $x(n)$ is obtained by sampling $x(t)$. Hence taking the z-transform of eqn(I.1), one obtains

$$Y(z) = U(z)\sum_{i=0}^{N} a_i z^{-i} - Y(z)\sum_{i=1}^{M} b_i z^{-i} \qquad \text{(I.3)}$$

From eqn(I.3), the z-transform transfer function is defined as

$$H(z) = \frac{Y(z)}{U(z)} = \frac{\sum_{i=0}^{N} a_i z^{-i}}{1 + \sum_{i=1}^{M} b_i z^{-i}} \qquad \text{(I.4)}$$

The change of the variable $z = e^{sT}$ constitutes a mapping of a portion of the s plane into the z plane. This is shown in fig I.2. Since z is periodic with period T, a one-to-one s-to-z mapping is valid only for the strip in the s plane between $\pm j\omega_s/2$. Within this strip, the left half-plane maps into the unit circle ($|z| = 1$). The right half-plane lies outside the unit circle, while the unit circle itself corresponds to the $j\omega$-axis.

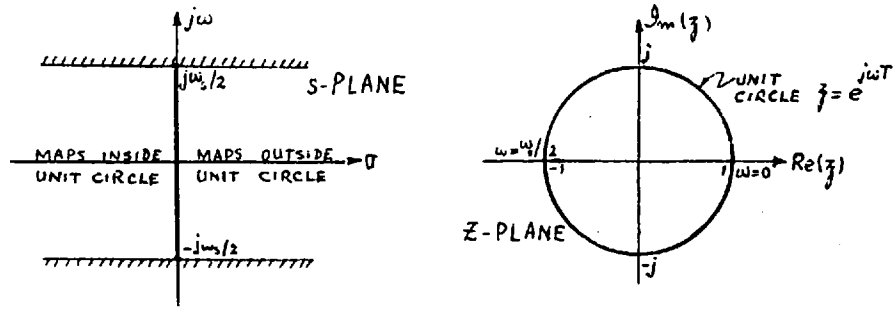


Fig I.2

I.3 Digital Filter Configurations

Assume that a digital filter has been designed in the sense that the transfer function $H(z)$ has been chosen. $H(z)$ is a ratio of polynomials in z^{-1}, and is finite outside and on the circle $|z|=1$ in the z- plane. From (I.4), $H(z)$ might be written in some other forms, for example,

$$H(z) = H_1(z) + H_2(z) + \ldots \qquad \text{(I.5)}$$

where $H_1(z)$ and $H_2(z)$, etc, are ratios of polynomials, or

$$H(z) = H_1(z) \times H_2(z) \times \ldots \qquad \text{(I.6)}$$

or some other forms[9].

For analogue filters, the realization of a given system function is moderately difficult and received considerable attention. For digital filters however, the implementation of a difference equation to realize a given $H(z)$ is almost trivial. A diagram to describe the time domain difference equation (I.1) is shown in fig. I.3. The triangle labelled with a constant represents multipication of a variable by a constant and the rectangle with z^{-1} inscribed represents a one-sample delay. The circle with a label of Σ is a summing point.

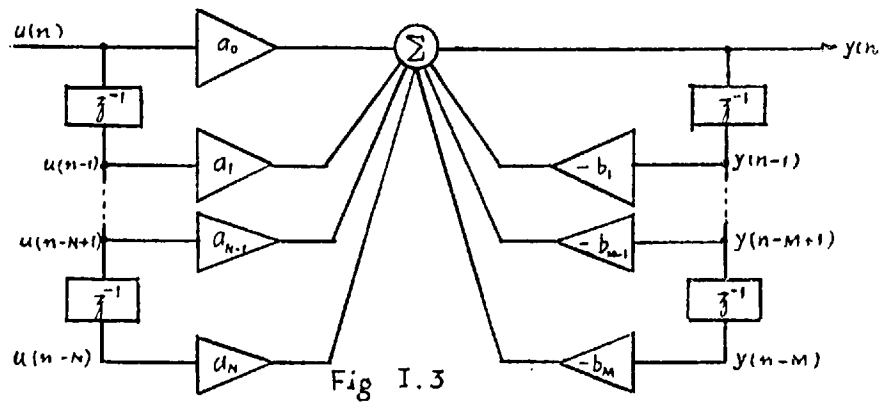


Fig I.3

If an intermediate variable $w(n)$ is introduced such that eqn. (I.1) is replaced by a pair of equations, but with no additional computation, i.e.

$$\begin{aligned} w(n) &= u(n) - \sum_{i=1}^{M} b_i w(n-i) \\ y(n) &= \sum_{i=0}^{N} a_i w(n-i) \end{aligned} \qquad (I.7)$$

then a circuit with less memory requirement than that shown in fig. I.3 is obtained (Fig.I.4). Both fig. I.3 and fig. I.4 have the same overall transfer function, and are called the

direct forms. It is to be emphasized that the coefficients a_i and b_i in the transfer function are the same as those constants in the network.

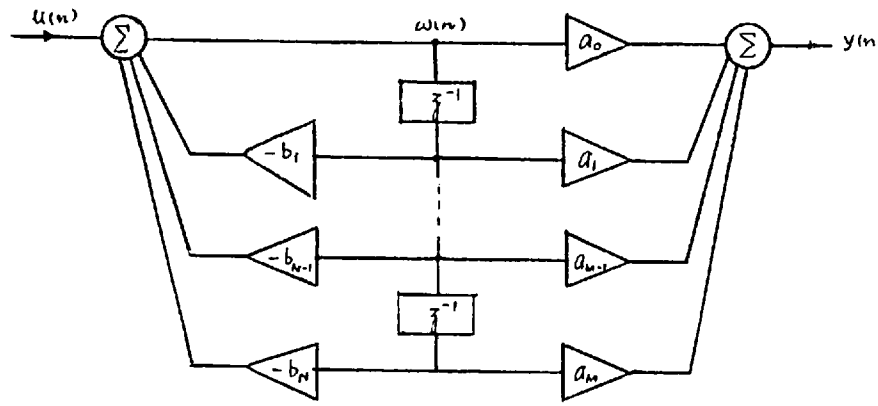


Fig I.4 The Direct (Canonic) Form for $N = M$

Despite the simplicity of the direcr forms that realize $H(z)$, they are undesirable for high-order difference equations for reasons of numerical accuracy {30}. But there are other forms. Suppose $H(z)$ is expressed in the form of eqn. (I.5), then the output $y(n)$ is the sum of the outputs of several smaller filters $H_1(z)$, $H_2(z)$, ... Each of these can be realized in either of the direct forms. Thus such representation of $H(z)$ leads to the configuration of fig I.5. In the extreme, each of the terms in eqn. (I.5) would be a ratio of first- or second-order polynomials in z^{-1}. The parallel form tends to be not nearly as sensitive to quantization effects as the direct forms {30}.

If $H(z)$ is expressed in the form of eqn (I.6), then

$$H(z) = H_1(z) \times H_2(z) \times \ldots \times H_k(z) \qquad (I.8)$$

Since these transfer functions are multiplied, the filters are in cascade. Fig. I.6 shows the realization of such a cascade configuration. The cascade form, with each of the $H_i(z)$ being

simple ratios of first- or second-order polynomials, is also preferable to the direct forms for numerical reasons.

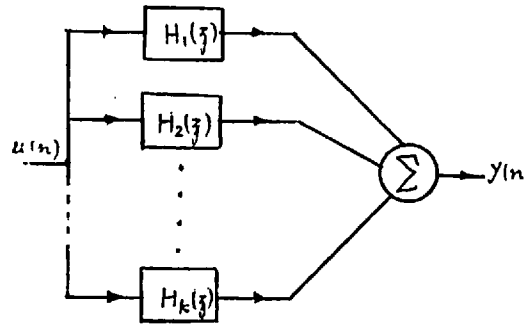


Fig I.5 The Parallel Form

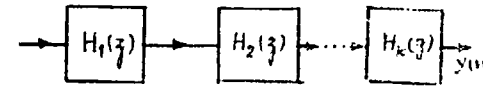


Fig I.6 The Cascade Form

There are many other ways of realizing $H(z)$. However, the direct, parallel and cascade forms are the most commonly used configurations.

The above-mentioned realizations assumed that $H(z)$ was a ratio of polynomials. These are recursive filters. For non-recursive filters, $H(z)$ is a polynomial in z^{-1} rather than a ratio of polynomials since the output of a non-recursive filter depends only on the present and past inputs. To realize a non-recursive filter, both the direct forms degenerate to a tapped delay line with a weighted sum of signals at the equally spaced taps. (Fig. I.7). This realization has also been called a transversal filter.

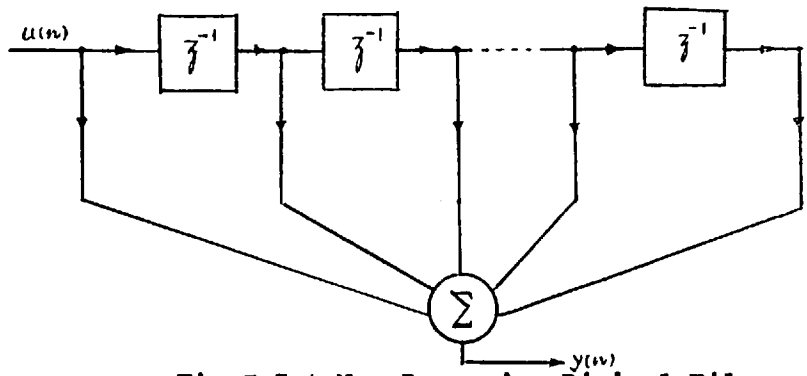


Fig I.7 A Non-Recursive Digital Filter

The parallel form has no particular meaning for a non-recursive filter; while the cascade form, although possible, is not in common use because it is usually difficult to factorize the high-order polynomials $H(z)$, moreover, there is no particular advantage to realize non-recursive digital filters in the cascade form.

I.4 Representation of Numbers in a Digital Filter

A real number in a digital filter can be represented using a finite number of bits in either the fixed-point form or the floating point form {14}, {64}. The error introduced in such a representation is discussed here. Only binary arithmetic will be discussed. The fixed-point case is first considered.

Suppose a number v which has been normalised so that $|v| \leq 1$ has the binary expansion (2's complement representation)

$$v = -v_o + \sum_{k=1}^{\infty} v_k \, 2^{-k} \qquad v_k = 1 \text{ or } 0 \quad \text{(I.9)}$$

To approximate v by a "word" of only ℓ bits rounding or truncating is used. In rounding, a 1 or 0 is first added to the ℓth bit $v_{\ell-1}$ according to whether the $(\ell+1)$th bit v_ℓ is 1 or 0. Then, only the first ℓ bits of the result are kept. In truncating, those bits beyond the most significant ℓ bits of the result are simply dropped. Since the error introduced by truncating is more serious than that introduced by rounding, truncating arithmetic is seldom used. Although modification of the analysis for truncating arithmetic is rather straightforward in most cases, only rounding arithmetic will be considered.

Let $[v]_\ell$ be the ℓ-bit representation of the number v. It is then clear that

$$-2^{-\ell} < v - [v]_\ell \leq 2^{-\ell} \qquad \text{(I.10)}$$

if rounding is used. An error of approximation ε may be defined by

$$[v]_l = v + \varepsilon \tag{I.11}$$

with $-2^{-l} \leq \varepsilon < 2^{-l}$. The approximation of v by $[v]_l$ is identical to the quantization of the number v by a quantizer with uniform step size $q = 2^{-l+1}$, and the error ε is referred to as the quantization noise {3}.

When two l-bit fixed-point numbers are added, their sum would still have l bits, provided there is no overflow. Therefore, if there is no overflow, fixed-point addition causes no error. On the other hand, the product of two l-bit numbers may have more than l-bits. Thus rounding is needed if l bits are to be kept. Let the actual computed product of two numbers v_1 and v_2 be denoted by $[v_1 v_2]_l$, then from the above discussion,

$$[v_1 v_2]_l = v_1 v_2 + \varepsilon \tag{I.12}$$

where the error ε is bounded by $-2^{-l} \leq \varepsilon < 2^{-l}$

A floating-point number is written in the form (sgn) $2^a.b$, where a is a binary integer called the exponent and b is a fraction between $\frac{1}{2}$ and 1 called the mantissa. The number of bits of the exponent determines the range of numbers that can be so represented, and the mantissa can usually take on the value 0. To represent a number v in floating-point form with only an l-bit mantissa, (it should be noted that for fixed-point numbers, l is the entire wordlength, but for floating-point numbers l is only the length of the mantissa), one first determines the smallest integer exceeding $\log_2 |v|$, denoted by $\{\log_2 |v|\}$. The binary expansion of the fraction $v/\{\log_2|v|\}$ is then rounded-off to l bits. Let $[v]_l$ denoted the l-bit mantissa floating-point approximation of the number v; then it is clear that

$$[v]_l = v(1 + \varepsilon) \tag{I.13}$$

where the relative error ε is bounded by $-2^{-l} \leq \varepsilon < 2^{-l}$.

Unlike the fixed-point case, both addition and multiplication in floating-point can introduce roundoff error. Let $[v_1 + v_2]_l$ and $[v_1 v_2]_l$ denote respectively, the actual computed sum and product of the two numbers v_1 and v_2, then {14},{64}

$$\begin{aligned} [v_1 + v_2]_l &= (v_1+v_2)(1+\varepsilon) \\ [v_1\, v_2]_l &= (v_1 \cdot v_2)(1+\delta) \end{aligned} \qquad \text{(I.14)}$$

where the relative errors ε and δ are bounded by $-2^{-l} \leq \varepsilon < 2^{-l}$, and $-2^{-l} \leq \delta < 2^{-l}$.

The roundoff error from a floating-point digital filter is usually (but not always) less than that from a fixed-point filter with the same total number of data digits because of the automatic scaling provided by floating-point arithmetic {33}, {62}. However, since floating-point arithmetic is significantly more complex and costly to implement, most digital filters have been, and will probably continue to be, constructed with fixed-point hardware. Hence, in this thesis, only fixed-point digital filters will be considered. Oppenheim {48} has proposed another interesting mode of arithmetic for digital filter implementation, called *block-floating-point*, which provides a simplified form of automatic scaling of the filter data. As would be expected, the performance of block-floating-point appears to lie somewhere between those of fixed-point and of floating-point.

I.5 Quantization Effects in Digital Filters

As discussed in the previous section, a number in the digital filter has to be rounded-off. There are three places at which

such rounding occurs. These three sources of "quantization error" are

a) analogue to digital (A/D) conversion errors

b) errors due to the finite representation of the digital filter coefficients

c) quantization errors due to rounding off the result of multiplication of data with the filter coefficients.

The first source of error, A/D conversion, is incurred when the input to the filter is quantized to a finite number of bits. This quantization creates an additive noise, which may be treated as random if the quantization is fine enough and if the signal varies sufficiently relative to the sampling rate and the number of quantization levels {3}.

The second source of error, finite representation of filter coefficients, is a deterministic effect. This effect is analogous to that encountered in continuous filters when the components called for by the design are not available. It can be taken into account by recomputing the poles and zeros of the filter with the quantized coefficients. The small changes in the filter coefficients due to finite number representation results in a corresponding change in the poles and zeros. Kaiser {30} has studied the sensitivity of the pole positions of an n^{th} order digital filter due to coefficient quantization. In his analysis, he concludes that for a direct filter realization, the sensitivity of the pole positions increases with the order n. This result has been corroborated in the work reported by Knowles and Olcayto {38}. Gold and Radar {21} have studied the coefficient quantization problem for second-order digital filters. They conclude that a realization via a pair of coupled first-order sections is less sensitive to coefficient changes than a single second-order form. Mantey {46} has studied the coefficient quantization problem by

selecting a state variable representation for the digital filter. His results, as well as the results from the other workers mentioned before, indicate that a digital filter should be realized by a parallel or cascade connection of first or second order subfilters instead of a direct n^{th} order realization. (It is for this reason that this thesis mainly considers the second-order filter).

The third source of error arises in the evaluation of the arithmetic products and their sum as indicated in eqn (I.1). For the non-recursive filter ($b_j = 0$, $j = 1, 2, \ldots, M$) the magnitude of the error incurred by using finite arithmetic can be quickly estimated by approximating the action quantizing with a noise source (which can be considered random in most cases). For the recursive filter the calculation of the errors is more difficult as a result of the feedback inherent in the b_j terms. For one thing, while there is no absolute necessity to round the product in a non-recursive filter, the sums of products that are fed back in a recursive filter must be rounded, since after a multiplication of two quantities represented by l_1 and l_2 respectively, the product contains $l_1 + l_2$ bits. If it were fed back without rounding, the next stage would generate numbers requiring yet more bits. Again, each rounding operation adds a small noise term, which can be considered to be random in most cases, and these terms are passed through a digital filter consisting of part or all of the required digital filter {20}, {22}, {36}, {37}, {38}, {42}, {52}. Obviously, in a cascade realization, the noise generated in the k^{th} stage cannot affect any of the earlier stages. A similar effect causes the noise in the direct (but not the direct canonic) realization to pass through those portions of the filter that realize the poles of $H(z)$ and not through those portions that realize the zeros.

A related effect also may occur in recursive filters as a result of round-off error when the round-off noise is highly

correlated with the signal or highly correlated with itself from iteration to iteration. This is called the *dead-band-effect*{5}, {22}, This is best illustrated by an example. Suppose the digital filter is described by

$$y(n) = 0.99\ y(n-1) + u(n) \qquad (I.15)$$

but is implemented with products rounded to the nearest integer. Then with the input zero, the output would be expected to decay to zero. However, any output in the range -50 to 50 causes the error due to quantization to exactly balance the decay per iteration, so that the erroneous output is maintained, i.e. there exist a steady output. Higher order filters have more complicated effects; the output may go through the deadband and reach the other side, or steady-state oscillations (generally known as *limit cycle oscillations*), may occur.

In considering quantization effects, it is not as necessary to compute the exact results of the effects, which may be difficult, as to estimate the bounds on them as a guide to avoiding the effects that cannot be tolerated. The theory developed in the literature so far has concentrated on rough estimates, such as upper bounds and mean square errors.

I.6 Objectives and Outline of the Thesis

So far it has been assumed that the digital signals in the filter are shifted forward from one register to another with the pulse repetition (sampling) frequency while the output samples also appear at the rate of sampling. It has also been assumed that all the filter coefficients remain unchanged throughout the filtering process. However, if the shifting is continued several times during each pulse repetition interval so that the signals stored in the registers re-circulate in the internal paths of the filter

during each sampling interval, then this processing leads to multiple output signals [13]. Furthermore, if the filter coefficients are also allowed to take on different values for the different shift sequences, the filter will possess some interesting properties.

It is the main object of this thesis to introduce such a "multirate" digital filter and to investigate its properties. The quantization errors for such a device are analysed. Because of the nature of the device, it has been found that the use of state-space method is very much more convenient in these analyses. The investigation of these errors reveals many interesting properties of the multirate digital filter, and under particular circumstances, these properties may render the multirate filter more advantageous than the conventional digital filters. Engineering applications of the multirate digital filters may be possible if such advantageous properties are utilized. To the best knowledge of the author, this investigation is novel, and the results obtained are original unless otherwise stated.

Chapter II gives a brief account of the state-space analysis which forms a basis for the mathematical analysis of the ensuing chapters. Chapter III develops a general mathematical model of the multirate digital filter from the point of view of state-space concept. From such a model, interesting properties of the ideal multirate digital filter are exposed. Several ideal multirate digital filters are designed according to the model and their performances compared to an equivalent single-rate filter. Chapter IV, V, VI and VII look into the quantization errors of the multirate digital filter. Each chapter begins with an introduction to the particular error discussed in the chapter and, where appropriate, a brief account of the usual methods used to solve the problem. Then the error associated with the multirate digital filter is discussed and analysed. Where suitable, computer simulations of the multirate digital filter are performed so as to

verify the theoretical analyses. Although limit cycle oscillations in a digital filter are phenomena which arise from the non-linear rounding of multiplication products, the subject has been treated separately in chapter VII. This is because limit cycle oscillations are, by their very nature, generated by quantization error sequences which are highly correlated, while multiplication round-off errors are treated more or less being uncorrelated in chapter VI.

The thesis concludes with an indication of those problems which remain unsolved and perhaps may be of interest for further research.

Since the commencement of this project, a number of papers have appeared [2], [27] which consider similar problems to those of this thesis and some similar results have been obtained. However, the approaches taken by these authors to the consideration of these problems are quite different from those taken in this thesis. Unless otherwise stated, the results and conclusions presented here were obtained independently. In particular, the following are considered to be the more significant contributions:

a) The mathematical modelling of a multirate digital filter with periodically varying coefficients and the investigation of the properties of its transfer functions.

b) The consideration of the processing of A/D noise through a multirate digital filter.

c) The consideration of the pole sensitivity of a multirate filter.

d) The comparison of pole sensitivities between single-rate and multirate digital filters.

e) The comparison of multiplication round-off errors between single-rate and multirate digital filters.

f) The discovery of the use of multirate digital filters to suppress limit-cycle oscillations for a deterministic input.

Chapter II

STATE-SPACE ANALYSIS OF DIGITAL FILTERS

A knowledge of certain mathematical techniques is needed to analyse discrete-time systems. This chapter is devoted to a brief description of the state-space method which is used in most parts of this thesis.

II.1 Introduction

In general, the analysis and design of linear systems may be carried out by one of two major approaches. One approach relies on the use of Laplace and z-transforms, transfer functions, block diagrams or signal flow graphs. The other method, which has gained significant importance in system theory and engineering is the state variable technique.

In a broad sense the state variable method has at least the following important advantages over the conventional transfer function method:

a) The state variable formulation is natural and convenient for computer solutions.
b) The state variable approach allows a unified representation of digital systems with various types of sampling schemes.
c) The state variable method allows a unified representation of single variable and multivariable systems.
d) The state variable method can be applied to certain types of nonlinear and time-varying systems.

In the state variable method a continuous-data system is

represented by a set of first-order differential equations. For a digital system with discrete-data components the state equations are first-order difference equations.

II.2 Reciprocal Difference Operator Systems

A discrete-time system can generally be described by a difference equation. If a system is described by a difference equation of the form

$$y(k+N) + b_1 y(k+N-1) + \ldots + b_{N-1} y(k+1) + b_N y(k) = a_N u(k) \qquad \text{(II.1)}$$

it is called a *reciprocal difference operator system*. In contradistinction to this system is the difference operator system charaterized by the difference equation

$$y(k) = a_0 u(k+M) + a_1 u(k+M-1) + \ldots + a_{M-1} u(k+1) + a_M u(k) \qquad \text{(II.2)}$$

If $E[u(k)]$ denotes the unit shift operator of $u(k)$, i.e.

$$E[u(k)] \triangleq u(k+1) \qquad \text{(II.3)}$$

and defining the shifting operators

$$\left.\begin{aligned} f(E) &= E^N + b_1 E^{N-1} + \ldots + b_{N-1} E + b_N \\ g(E) &= a_0 E^M + a_1 E^{M-1} + \ldots + a_{M-1} E + a_M \end{aligned}\right\} \qquad \text{(II.4)}$$

then (II.1) and (II.2) simplifies to

$$f(E)[y(k)] = a_N u(k) \qquad \text{(II.5)}$$

$$y(k) = g(E)[u(k)] \qquad \text{(II.6)}$$

Consider (II.1) and suppose a new set of variables (called state-

variables) are chosen such that

$$\left.\begin{aligned} x_1(k) &= y(k) \\ x_2(k) &= x_1(k+1) = y(k+1) \\ x_3(k) &= x_2(k+1) = y(k+2) \\ &\cdots\cdots\cdots\cdots\cdots\cdots \\ x_N(k) &= x_{N-1}(k+1) = y(k+N-1) \end{aligned}\right\} \quad \text{(II.7)}$$

Rearranging the set of equations(II.7), one obtains

$$\left.\begin{aligned} x_1(k+1) &= x_2(k) \\ x_2(k+1) &= x_3(k) \\ &\cdots\cdots\cdots\cdots \\ x_N(k+1) = y(k+N) &= -b_N x_1(k) - b_{N-1} x_2(k) - \cdots - b_1 x_1(k) + a_N u(k) \end{aligned}\right\} \quad \text{(II.8)}$$

or, in matrix form, (II.8) can be written as

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_{N-1}(k+1) \\ x_N(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -b_N & -b_{N-1} & -b_{N-2} & \cdots & -b_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_{N-1}(k) \\ x_N(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ a_N \end{bmatrix} u(k) \quad \text{(II.9)}$$

Eqn(II.9) is called the state equation of the system. For this choice, the output equation of the system is given by

$$y(k) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ \vdots \\ x_N(k) \end{bmatrix} \quad \text{(II.10)}$$

Equations (II.9) and (II.10), called the dynamic equations of

the system, can be written in a more compact form:

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\,\mathbf{x}(k) + \mathbf{B}\,u(k) \\ y(k) &= \mathbf{C}\,\mathbf{x}(k) \end{aligned} \tag{II.11}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -b_N & -b_{N-1} & -b_{N-2} & \cdots & -b_1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ a_N \end{bmatrix} \tag{II.12}$$

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

As can be seen, the state vector is N-dimensional, and consequently, the state space over which $\mathbf{x}(k)$ ranges is $\mathcal{R}^N$ where $\mathcal{R}^N$ denotes an N-dimensional linear space{[24]}{[57]}. Obviously, there are many other ways to choose the state vector $\mathbf{x}(k)$, however, since the order of the system is N, no matter how the state vector is defined the state space is still N-dimensional. Although it has been implicitly understood that the coefficients b_i and a_i in (II.4) are constant, they need not be. They may be functions of time and the above development of the state equation is still valid.

As previously remarked, the choice of a state vector is not unique. To illustrate this further, suppose that, instead of the state vector in (II.7), a vector $\mathbf{x}'$ is selected such that the state variables are

$$\begin{aligned} x_1'(k) &= y(k+N-1) \\ x_2'(k) &= y(k+N-2) \\ &\vdots \\ x_N'(k) &= y(k) \end{aligned} \tag{II.13}$$

In exactly the same way as shown above, the dynamic equation of the system are:-

$$\begin{aligned} \mathbf{x}'(k+1) &= \mathbf{A}'\,\mathbf{x}'(k) + \mathbf{B}'u(k) \\ y(k) &= \mathbf{C}'\,\mathbf{x}'(k) \end{aligned} \tag{II.14}$$

where

$$\mathbf{A}' = \begin{bmatrix} -b_1 & -b_2 & \cdots & -b_N \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1\ 0 \end{bmatrix}, \quad \mathbf{B}' = \begin{bmatrix} a_N \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\mathbf{C}' = [\,0 \quad \cdots \quad 0 \quad 1\,] \tag{II.15}$$

It is observed that the state vector $\mathbf{x}$ of (II.7) is related to the state vector in (II.13) through the matrix equation

$$\mathbf{x} = \mathbf{T}\,\mathbf{x}' \tag{II.16}$$

where $\mathbf{T}$ is the $N \times N$ non-singular matrix

$$\mathbf{T} = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ 0 & \cdots & 1 & 0 \\ \cdot & \cdots & \cdot & \cdot \\ 1 & \cdots & 0 & 0 \end{bmatrix} \tag{II.17}$$

Substituting $\mathbf{x} = \mathbf{T}\,\mathbf{x}'$ into the state equation (II.11) for $\mathbf{x}$, then

$$\begin{aligned} \mathbf{x}'(k+1) &= \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T}\,\mathbf{x}'(k) + \mathbf{T}^{-1}\mathbf{B}u(k) \\ y(k) &= \mathbf{C}\,\mathbf{T}\,\mathbf{x}'(k) \end{aligned} \tag{II.18}$$

Comparing the matrices in (II.18) with those of (II.14), it follows that

$$\begin{aligned} \mathbf{A}' &= \mathbf{T}^{-1}\mathbf{A}\,\mathbf{T} \\ \mathbf{B}' &= \mathbf{T}^{-1}\mathbf{B} \\ \mathbf{C}' &= \mathbf{C}\,\mathbf{T} \end{aligned} \tag{II.19}$$

What is established essentially is: There are numerous ways of associating a state vector with a system of order N. The different ways, however, amount to designating different variables to points in the space $\mathcal{R}^N$. That is to say, the state variables are connected by a one-to-one linear mapping, i.e.

$$\mathbf{x} = \mathbf{T}\,\mathbf{x}' \tag{II.20}$$

where $\mathbf{T}$ is an $N \times N$ non-singular constant matrix.

II.3 General Difference Operator Systems

The general system characterized by

$$[f(E)]\ y(k) = [g(E)]\ u(k) \tag{II.21}$$

will now be considered, where $f(E)$ and $g(E)$ are defined by (II.4). The order of $f(E)$ is N and that of $g(E)$ is M. For a physically realizable system, $M \leq N$. In terms of the z-transform transfer function

$$H(z) = \frac{g(z)}{f(z)} \tag{II.22}$$

a physically realizable system may have a finite number of poles equal to or greater than the number of zeros. Hence it does not lose any generality if (II.21) is rewritten as

$$\begin{aligned} y(k+N) + b_1 y(k+N-1) + \ldots + b_{N-1} y(k+1) + b_N y(k) \\ = a_0 u(k+N) + a_1 u(k+N-1) + \ldots + a_{N-1} u(k+1) + a_N u(k) \end{aligned} \tag{II.23}$$

where $g(z)$ and $f(z)$ are assumed to be of the same order N.

As in the case of reciprocal difference operator system, a set of state variables, $x_i(k)$, are chosen such that,

$$
\begin{aligned}
y(k) &= x_1(k) + \alpha_0 u(k) \\
x_1(k+1) &= x_2(k) + \alpha_1 u(k) \\
&\vdots \\
x_N(k+1) &= -\beta_N x_1(k) - \beta_{N-1} x_2(k) - \ldots - \beta_1 x_N(k) + \alpha_N u(k)
\end{aligned}
\qquad \text{(II.24)}
$$

where $\alpha_0, \alpha_1, \ldots \alpha_N, \beta_1, \beta_2, \ldots, \beta_N$ are to be determined.

Taking a unit shift in the first equation of the set (II.24),

$$y(k+1) = x_1(k+1) + \alpha_0 u(k+1)$$

Substituting for $x_1(k+1)$ gives

$$y(k+1) = x_2(k) + \alpha_1 u(k) + \alpha_0 u(k+1) \qquad \text{(II.25)}$$

Again taking another unit shift, one obtains

$$y(k+2) = x_2(k+1) + \alpha_1 u(k+1) + \alpha_0 u(k+2)$$

and substituting for $x_2(k+1)$, one obtains

$$y(k+2) = x_3(k) + \alpha_2 u(k) + \alpha_1 u(k+1) + \alpha_0 u(k+2) \qquad \text{(II.26)}$$

Following this procedure, the following equations are established,

$$
\begin{aligned}
y(k+N-1) &= x_N(k) + \alpha_{N-1} u(k) + \alpha_{N-2} u(k+1) + \ldots + \alpha_0 u(k+N-1) \\
y(k+N) &= -\{\beta_N x_1(k) + \beta_{N-1} x_2(k) + \ldots + \beta_1 x_N(k)\} + \alpha_N u(k) \\
&\qquad + \alpha_{N-1} u(k+1) + \ldots + \alpha_0 u(k+N)
\end{aligned}
\qquad \text{(II.27)}
$$

Substituting (II.25,26,27) into (II.23) and comparing coefficients. one finds,

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ b_1 & 1 & 0 & \cdots & 0 \\ b_2 & b_1 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ b_N & b_{N-1} & b_{N-2} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} \qquad \text{(II.28)}$$

and

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix} \qquad \text{(II.29)}$$

Hence putting these values back to eqn(II.24), the dynamic equations of the general difference operator system are

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\,\mathbf{x}(k) + \mathbf{B}\,u(k) \\ y(k) &= \mathbf{C}\,\mathbf{x}(k) + \mathbf{D}u(k) \end{aligned} \qquad \text{(II.30)}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -b_N & -b_{N-1} & -b_{N-2} & \cdots & -b_1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} \qquad \text{(II.31)}$$

$$\mathbf{C} = [\,1 \quad 0 \quad 0 \quad \cdots \quad 0\,] \qquad \mathbf{D} = a_0$$

It is noted that the matrix **A** is the same as that of the reciprocal difference operator system described in section II.2.

II.4 State-Transition Signal Flow Graphs of Discrete Systems

The state- transition signal flow graph {[41]}, or state diagram for short, may be used to portray relationships between state

variables of a system, including initial states if necessary. The state diagram of a discrete system includes elements which parallel digital computer elements.

It is assumed that a digital computer can perform, at least, the following basic computing operations:

a) Multiplying a machine variable by a positive or negative constant coefficient.
b) Time delay, involving storing a variable for a certain length of time before re-using it.
c) Producing the sum of two or more machine variables.

The mathematical description of these basic digital computations and their corresponding z-transform expressions are given by:-

Multiplication by a constant

$$\left.\begin{aligned} x_2(k) &= a\, x_1(k) \\ X_2(z) &= a\, X_1(z) \end{aligned}\right\} \quad \text{(II.32)}$$

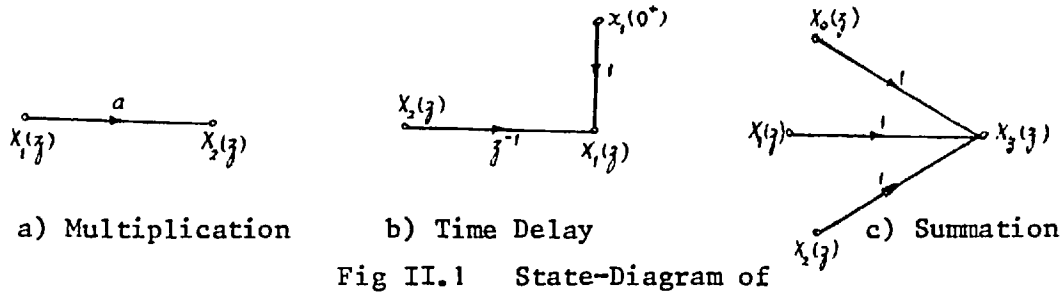
Summing

$$\left.\begin{aligned} x_3(k) &= x_o(k) + x_1(k) - x_2(k) \\ X_3(z) &= X_o(z) + X_1(z) - X_2(z) \end{aligned}\right\} \quad \text{(II.33)}$$

Storage and time delay

$$\left.\begin{aligned} x_2(k) &= x_1(k+1) \\ X_2(z) &= zX_1(z) - zx_1(0^+) \\ \text{or,} \quad X_1(z) &= z^{-1} X_2(z) + x_1(0^+) \end{aligned}\right\} \quad \text{(II.34)}$$

The state-diagram representation of the three transform equations (II.32) to (II.34) are obtained from the basic rules of signal flow graphs and are shown in fig II.1



a) Multiplication b) Time Delay c) Summation

Fig II.1 State-Diagram of the Basic Elements of a Discrete-time System

Once the state diagram of the basic discrete operations are established, the state-diagram representation of a discrete-data system may be obtained. Hence, for the system described by (II.30) and (II.31), the state diagram is as shown in fig II.2

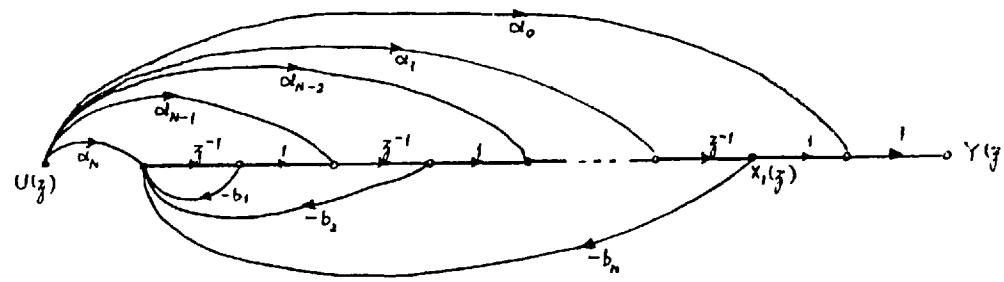


Fig II.2 State-Diagram for System (II.30) & (II.31)

II.5 State Diagram of General Difference Operator Systems by the Decomposition of Transfer Functions

In section I.3, it has been shown that a digital filter transfer function can be realized in many configurations. Here in this section, the procedure of choosing different sets of state-variables from the same transfer function, leading to different

state- diagrams (and hence different configurations of the digital filter transfer function) is described. The procedure of expressing a transfer function by a state-diagram is termed the process of decomposition. In general, the most commonly used ways of decomposition are:-

A) DIRECT DECOMPOSITION

Consider the transfer function of a general Nth order discrete system

$$H(z) = \frac{Y(z)}{U(z)} = \frac{a_0 z^N + a_1 z^{N-1} + \ldots + a_{N-1} z + a_N}{z^N + b_1 z^{N-1} + \ldots + b_{N-1} z + b_N} \qquad \text{(II.35)}$$

Multiply the numerator and the denominator of the RHS of (II.35) by $z^{-N}W(z)$, where $W(z)$ is an auxiliary variable, then

$$H(z) = \frac{Y(z)}{U(z)} = \frac{a_0 + a_1 z^{-1} + \ldots + a_{N-1} z^{-N+1} + a_N z^{-N}}{1 + b_1 z^{-1} + \ldots + b_{N-1} z^{-N+1} + b_N z^{-N}} \cdot \frac{W(z)}{W(z)} \qquad \text{(II.36)}$$

Since $W(z)$ is arbitrary, it can always be chosen such that the following relations are true

$$Y(z) = (a_0 + a_1 z^{-1} + \ldots + a_{N-1} z^{-N+1} + a_N z^{-N})\, W(z) \qquad \text{(II.37)}$$

$$U(z) = (1 + b_1 z^{-1} + \ldots + b_{N-1} z^{-N+1} + b_N z^{-N})\, W(z) \qquad \text{(II.38)}$$

Now, a quite arbitrary but convenient way of defining the state variables $x_1, x_2, \ldots, x_N$ is as follows

$$\begin{aligned} x_N(k+1) &= w(k) & X_N(z) &= z^{-1}W(z) \\ x_{N-1}(k+1) &= x_N(k) & X_{N-1}(z) &= z^{-1}X_N(z) \\ &\vdots & &\vdots \\ x_1(k+1) &= x_2(k) & X_1(z) &= z^{-1}X_2(z) \end{aligned} \qquad \text{(II.39)}$$

In this way , the state variables will turn out to be the output of the storages or time delays of the discrete system. Writing (II.38)

in the following way,

$$W(z) = U(z) - (b_N z^{-N} + b_{N-1} z^{-N+1} + b_1 z^{-1})\, W(z)$$

i.e. $$W(z) = U(z) - b_N X_1(z) - b_{N-1} X_2(z) - \ldots - b_1 X_N(z) \qquad \text{(II.40)}$$

Also from (II.37) the output is given by a combination of the state variables, hence

$$Y(z) = a_o W(z) + a_1 X_N(z) + \ldots + a_N X_1(z) \qquad \text{(II.41)}$$

The state diagram (without initial conditions) portraying (II.40) and (II.41) is shown in fig II.3

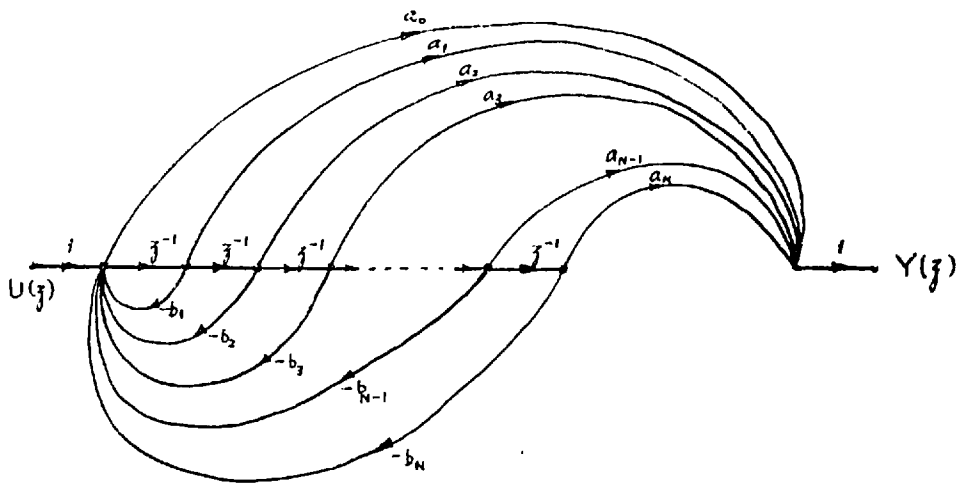


Fig II.3 Direct Decomposition of a General Discrete System

It can be seen that fig II.3 is merely a flow-graph representation of fig I.4

Defining the state variables as the output of the delay units,

the state equations are written direct from the state-diagram (without initial conditions):-

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_N(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -b_N & -b_{N-1} & -b_{N-2} & \cdots & -b_1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_N(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} u(k) \qquad \text{(II.42)}$$

Also the output equation from the diagram is

$$y(k) = \begin{bmatrix} (a_N - a_o b_N) & (a_{N-1} - a_o b_{N-1}) & \cdots & (a_1 - a_o b_1) \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_N(k) \end{bmatrix} + a_o u(k) \qquad \text{(II.43)}$$

B) PARALLEL DECOMPOSITION

Another method of decomposition relies on the partial fraction expansion of the transfer function into a sum of first- or second-order terms. Second order terms are chosen when the denominator has complex roots so that in practice, realization of complex multiplying coefficients can be avoided. To each of these first- or second-order terms, the direct decomposition is applied.

If the general transfer function shown in (II.35) is expanded by partial fraction, then it can be written as

$$H(z) = \sum_{i=1}^{M} H_i(z) \qquad \text{(II.44)}$$

where each of the terms $H_i(z)$ is given either by

$$H_i(z) = \frac{\alpha_i}{z + \beta_i} \qquad \text{(II.45)}$$

if it is of the first-order, or by

$$H_i(z) = \frac{\alpha_{1i} z + \alpha_{2i}}{z^2 + \beta_{1i} z + \beta_{2i}} \qquad \text{(II.46)}$$

if the order of the term $H_i(z)$ is two. Applying the method of direct decomposition to (II.45) and (II.46), the state-diagrams for the first- and second-order terms in the partial fraction expansion are shown in fig II.4(a) and (b) respectively.

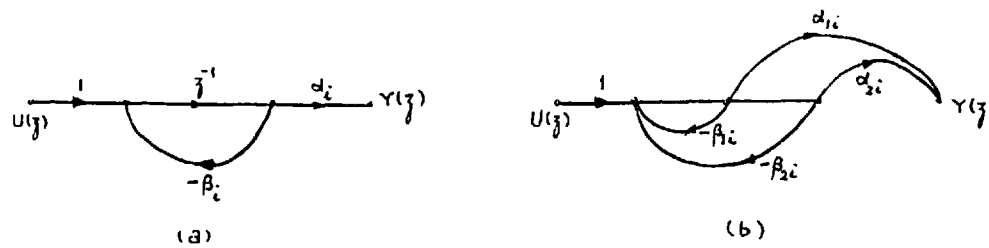


Fig II.4(a) State Diagram of 1st Order Section for Parallel Decomposition

(b) State Diagram of 2nd Order Section for Parallel Decomposition

If the transfer function of the system has only simple real poles, all $H_i(z)$ will be of the first order and hence the state matrix **A** is of the form

$$\mathbf{A} = \begin{bmatrix} -\beta_1 & & & \\ & -\beta_2 & & \\ & & \ddots & \\ & & & -\beta_M \end{bmatrix} \qquad \text{(II.47)}$$

which is a diagonal matrix. On the other hand, if all the poles of $H(z)$ are simple but in complex conjugates, then if complex multiplication are to be avoided, all $H_i(z)$ are of the second order. The state matrix **A** will then be of the block diagonal form:-

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ & & & \mathbf{A}_M \end{bmatrix} \qquad \text{(II.48)}$$

where $\mathbf{A}_i = \begin{bmatrix} 0 & 1 \\ -\beta_{2i} & -\beta_{1i} \end{bmatrix}$

However, if some of the poles of $H(z)$ are multiple, **A** is of the Jordan canonic form {[18]} (see section II.7), i.e.

$$\mathbf{A} = \begin{bmatrix} -\beta_1 & & & & \\ & -\beta_2 & 1 & & \\ & & -\beta_2 & 1 & \\ & & & -\beta_2 & \\ & & & & -\beta_3 \\ & & & & & \ddots \\ & & & & & & -\beta_n & 1 \\ & & & & & & & -\beta_n \end{bmatrix} \qquad \text{(II.49)}$$

Whether the matrix **A** is diagonal, block diagonal or of the Jordan canonic form, it will simplify the procedure of solving the state equations (section II.7).

C) CASCADE DECOMPOSITION

Still another decomposition is obtained by cascading first- or second-order transfer functions to form higher order ones, i.e.

$$H(z) = \prod_{i=1}^{M} H_i(z) \qquad \text{(II.50)}$$

where each of $H_i(z)$ is given by

$$H_i(z) = \frac{\alpha_{1i} z + \alpha_{2i}}{z + \beta_i} \quad \text{for 1st order sections} \qquad \text{(II.51)}$$

or by $$H_i(z) = \frac{\alpha_{0i}z^2 + \alpha_{1i}z + \alpha_{2i}}{z^2 + \beta_{1i}z + \beta_{2i}}$$ for second order sections (II.52)

Again applying the direct decomposition process, the state-diagrams for (II.51) and (II.52) are shown in fig II.5(a) and (b) respectively,

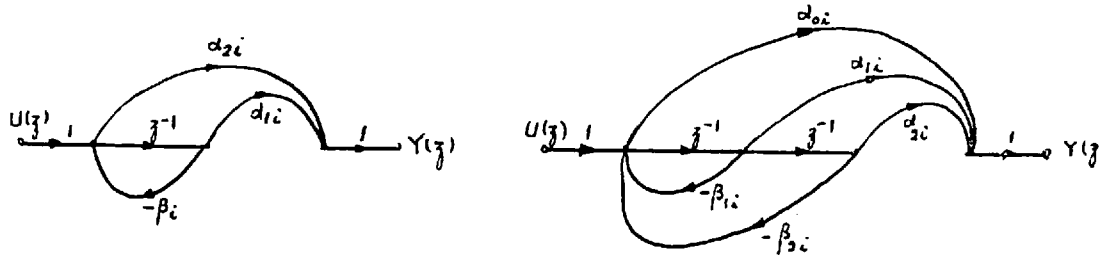


Fig II.5 (a) State-Diagram of 1st Order Section for Cascade Decomposition

(b) State-Diagram of 2nd Order Section for Cascade Decomposition

II.6 Solution of the Discrete State Equation

So far it has been demonstrated that the dynamic equations of a discrete system can be obtained by direct inspection of the difference equation or, more systematically, by decomposition of the system transfer function. Here in this section, it is shown how the solution of the state equation can be arrived at.

To begin, consider the free system

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) \qquad \text{(II.53)}$$

and let $k = 0,1,\ldots$, one obtains successively,

$$\begin{aligned}
\mathbf{x}(1) &= \mathbf{A}(0)\,\mathbf{x}(0) \\
\mathbf{x}(2) &= \mathbf{A}(1)\,\mathbf{x}(1) = \mathbf{A}(1)\,\mathbf{A}(0)\,\mathbf{x}(0) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathbf{x}(k+1) &= \left\{ \prod_{i=0}^{k} \mathbf{A}(i) \right\} \mathbf{x}(0)
\end{aligned} \tag{II.54}$$

(II.54) gives the free motion of the system, starting in the initial state $\mathbf{x}(0)$. If the system is constant and the sampling interval is constant, (II.53) has the solution

$$\mathbf{x}(k) = \mathbf{A}^{k}\,\mathbf{x}(0) \tag{II.55}$$

Now consider the forced system

$$\mathbf{x}(k+1) = \mathbf{A}\,\mathbf{x}(k) + \mathbf{B}\,u(k) \tag{II.56}$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices. Substituting $k = 0,1,\ldots$ one obtains

$$\begin{aligned}
\mathbf{x}(1) &= \mathbf{A}\,\mathbf{x}(0) + \mathbf{B}\,u(0) \\
\mathbf{x}(2) &= \mathbf{A}\,\mathbf{x}(1) + \mathbf{B}\,u(1) \\
&= \mathbf{A}\left\{ \mathbf{A}\,\mathbf{x}(0) + \mathbf{B}\,u(0) \right\} + \mathbf{B}\,u(1) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
\mathbf{x}(k) &= \mathbf{A}^{k}\,\mathbf{x}(0) + \sum_{i=0}^{k-1} \mathbf{A}^{i}\,\mathbf{B}\,u(k-i-1) \\
&= \mathbf{A}^{k}\,\mathbf{x}(0) + \sum_{i=0}^{k-1} \mathbf{A}^{k-i-1}\,\mathbf{B}\,u(i)
\end{aligned} \tag{II.57}$$

Eqn(II.57) suggests that the sequence $\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{k-1}\mathbf{B}$, may be defined as a weighting sequence. Thus, if the weighting sequence is defined by

$$\mathbf{W}(n) = \mathbf{A}^{n}\,\mathbf{B}$$

then,

$$\mathbf{x}(k) = \mathbf{A}^{k}\mathbf{x}(0) + \sum_{i=0}^{k-1} \mathbf{W}(k-i-1)\,u(i) \tag{II.58}$$

The right-hand side of eqn(II.58) represents the *convolution sum*, the analogue of the convolution integral in continuous-time systems. The matrix $\mathbf{A}^n$ so often encountered in state-space analysis of discrete-time systems is referred to as the *discrete-transition matrix*.

II.7 The Use of Similar Matrices and the Jordan Canonic Form in the Evaluation of the State-Transition Matrix

A vector $\mathbf{v}$ in an N-dimensional space can be written uniquely in terms of its basis vector $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ as

$$\mathbf{v} = \sum_{i=1}^{N} \chi_i \mathbf{w}_i \tag{II.59}$$

and the number χ_i are called the co-ordinates of $\mathbf{v}$ in the $\mathbf{w}$ basis.

Suppose the basis vector are now changed to $\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_N^*$. The $\mathbf{w}^*$-basis vector can be related to the $\mathbf{w}$-basis by a linear equation, i.e.

$$\mathbf{w}_i^* = \sum_{k=1}^{N} \gamma_{ki} \mathbf{w}_k \qquad i=1,2,\ldots,N \tag{II.60}$$

Expressing $\mathbf{v}$ in terms of $\mathbf{w}_i^*$,

$$\mathbf{v} = \sum_{i=1}^{N} \chi_i^* \mathbf{w}_i^* \tag{II.61}$$

On substituting (II.60) into (II.61),

$$\mathbf{v} = \sum_{k=1}^{N} \left(\sum_{i=1}^{N} \gamma_{ki} \chi_i^* \right) \mathbf{w}_k \tag{II.62}$$

Comparing (II.62) with (II.59),

$$\chi_i = \sum_{i=1}^{N} \gamma_{ki} \chi_i^* \qquad k=1,2,\ldots,N \tag{II.63}$$

or in matrix form, letting x be the co-ordinate vector of $\mathbf{v}$ in the $\mathbf{w}$-basis, and x^* be the co-ordinate vector of $\mathbf{v}$ in the $\mathbf{w}^*$-basis,

$$x = \Gamma x^* \tag{II.64}$$

where $\Gamma = [\gamma_{ij}]$ is the matrix of the transformation of co-ordinates which is non-singular.

Now, if the co-ordinate vector ζ is connected to the co-ordinate vector x by

$$\zeta = A x$$

in the $\mathbf{w}$-basis, then, in the $\mathbf{w}^*$-basis,

$$\zeta^* = A^* x^* \tag{II.65}$$

But $$\zeta = \Gamma \zeta^* \tag{II.66}$$

Hence $$\Gamma \zeta^* = A \Gamma x^* \tag{II.67}$$

i.e. $$\zeta^* = \Gamma^{-1} A \Gamma x^* \tag{II.68}$$

Comparing (II.68) with (II.65), it can be seen that

$$A^* = \Gamma^{-1} A \Gamma \tag{II.69}$$

Two matrices related in the form of (II.95) are said to be similar. Similar matrices are very useful in the evaluation of functions of a matrix and thus will be useful in the evaluation of the discrete transition matrix.

Similar matrices have the following properties:

(i) $\det[A - \lambda I] = \det[A^* - \lambda I]$ where $A^* = \Gamma^{-1} A \Gamma$

i.e. similar matrices have same eigenvalues.

Proof:- $$\Gamma^{-1}(\mathbf{A} - \lambda \mathbf{I})\Gamma = \Gamma^{-1}\mathbf{A}\Gamma - \lambda \mathbf{I} = (\mathbf{A}^* - \lambda \mathbf{I})$$

Taking the determinant,

$$[\det \Gamma^{-1}]\ [\det(\mathbf{A} - \lambda \mathbf{I})]\ [\det \Gamma] = \det(\mathbf{A}^* - \lambda \mathbf{I})$$

hence, $$\det(\mathbf{A} - \lambda \mathbf{I}) = \det(\mathbf{A}^* - \lambda \mathbf{I}) \qquad \text{(II.70)}$$

(ii) $\mathbf{f(A)} = \Gamma\ \mathbf{f}(\mathbf{A}^*)\ \Gamma^{-1}$

Proof:-

$$\mathbf{f}(\mathbf{A}^*) = \sum_{i=0}^{\infty} k_i (\mathbf{A}^*)^i = \sum_{j=0}^{N-1} \alpha_j (\mathbf{A}^*)^j$$

as a result of the Cayley-Hamilton theorem{18}

$$\therefore \quad \Gamma\ \mathbf{f}(\mathbf{A}^*)\ \Gamma^{-1} = \Gamma \left[\sum_{j=0}^{N-1} \alpha_j (\mathbf{A}^*)^j \right] \Gamma^{-1}$$

$$= \sum_{j=0}^{N-1} \alpha_j\ \Gamma (\mathbf{A}^*)^j\ \Gamma^{-1}$$

But, $$\mathbf{A}^{*j} = \Gamma^{-1} \mathbf{A}^j\ \Gamma$$

$$\therefore \quad \Gamma\, \mathbf{f}(\mathbf{A}^*)\, \Gamma^{-1} = \sum_{j=0}^{N-1} \alpha_j\ \mathbf{A}^j = \mathbf{f}(\mathbf{A}) \qquad \text{(II.71)}$$

(iii) Let $\mathbf{P} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \quad \mathbf{v}_N]$, i.e. a matrix with the eigenvectors of $\mathbf{A}$ as columns, then $\mathbf{P}$ is called an eigenvector matrix.

Now, if $\mathbf{A}$ has distinct eigenvalues, and if it has been chosen that the original basis should be changed to a basis formed by the eigenvectors of $\mathbf{A}$, then, the similar matrix $\mathbf{S}$ with the eigenvectors as basis is given by

$$\mathbf{S} = \mathbf{P}^{-1}\mathbf{A}\ \mathbf{P}$$

where $\mathbf{S}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ as its diagonal elements.

Proof:- Let $\mathbf{A}$ be an $N \times N$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ and eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ then

$$\left.\begin{aligned} \mathbf{A}\mathbf{v}_1 &= \lambda_1\mathbf{v}_1 \\ \mathbf{A}\mathbf{v}_2 &= \lambda_2\mathbf{v}_2 \\ &\vdots \\ \mathbf{A}\mathbf{v}_N &= \lambda_N\mathbf{v}_N \end{aligned}\right\} \Rightarrow \mathbf{A}\begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_N\end{bmatrix} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_N\end{bmatrix}\begin{bmatrix}\lambda_1 & & \mathbf{0} \\ & \lambda_2 & \\ & & \ddots \\ \mathbf{0} & & & \lambda_N\end{bmatrix}$$

But, $\quad \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_N\end{bmatrix} = \mathbf{P}$

$$\therefore \quad \mathbf{A}\,\mathbf{P} = \mathbf{P}\begin{bmatrix}\lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_N\end{bmatrix}$$

or $$\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P} = \mathbf{S} \tag{II.72}$$

(iv) $$\mathbf{f}(\mathbf{S}) = \sum_{i=0}^{N-1} \alpha_i \mathbf{S}^i$$

$$= \sum_{i=0}^{N-1} \alpha_i \begin{bmatrix}\lambda_1^i & & & \mathbf{0} \\ & \lambda_2^i & & \\ & & \ddots & \\ \mathbf{0} & & & \lambda_N^i\end{bmatrix}$$

$$= \begin{bmatrix}f(\lambda_1) & & & \mathbf{0} \\ & f(\lambda_2) & & \\ & & \ddots & \\ \mathbf{0} & & & f(\lambda_N)\end{bmatrix} \tag{II.73}$$

which is a one step process to find $\mathbf{f}(\mathbf{S})$.

Now,

$$\mathbf{f}(\mathbf{S}) = \begin{bmatrix}f(\lambda_1) & & & \mathbf{0} \\ & f(\lambda_2) & & \\ & & \ddots & \\ \mathbf{0} & & & f(\lambda_N)\end{bmatrix} \Rightarrow \text{eigenvalues of } \mathbf{f}(\mathbf{S}) \text{ are } f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_N)$$

But $\mathbf{f(S) = P^{-1} f(A) P}$

i.e. $\mathbf{f(S)}$ is similar to $\mathbf{f(A)}$

$\therefore$ eigen-values of $\mathbf{f(A)}$ are also $f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_N)$

MULTIPLE EIGENVALUES —— JORDAN CANONIC FORM

For any matrix $\mathbf{A}$, let the charateristic equation be written as

$$g(\lambda) = a_N \lambda^N + a_{N-1}\lambda^{N-1} + \ldots + a_1\lambda + a_0 \qquad \text{(II.74)}$$

Defining, $$f(\lambda,\mu) \triangleq \frac{g(\lambda) - g(\mu)}{\lambda - \mu}$$

then $$f(\lambda,\mu) = a_N \frac{(\lambda^N - \mu^N)}{(\lambda-\mu)} + a_{N-1}\frac{(\lambda^{N-1} - \mu^{N-1})}{(\lambda-\mu)} + \ldots + a_1 \qquad \text{(II.75)}$$

Replacing λ by $\lambda\mathbf{I}$ and μ by $\mathbf{A}$ in eqn(II.75), then

$$\mathbf{f(\lambda I, A) = (\lambda I - A)^{-1}\left\{g(\lambda I) - g(A)\right\}} \qquad \text{(II.76)}$$

But $\mathbf{g(A) = 0}$ as a result of the Cayley–Hamilton theorem, hence,

$$\mathbf{g(\lambda I) = (\lambda I - A)\,.\, f(\lambda I, A)} \qquad \text{(II.77)}$$

Let $\mathbf{f(\lambda I, A)}$ be written as $\Psi(\lambda)$ where $\Psi(\lambda)$ can be obtained from eqn(II.75) by substituting λ by $\lambda\mathbf{I}$ and μ by $\mathbf{A}$.

Now, $\mathbf{g(\lambda I) = g(\lambda).I}$

$\therefore$ $$g(\lambda)\,\mathbf{I} = \left[\lambda\mathbf{I} - \mathbf{A}\right].\Psi(\lambda) \qquad \text{(II.78)}$$

If λ_i is an eigenvalue of the matrix $\mathbf{A}$, then $g(\lambda_i) = 0$. Hence,

$$g(\lambda_i)\mathbf{I} = \left[\lambda_i\mathbf{I} - \mathbf{A}\right].\Psi(\lambda_i) = \mathbf{0}$$

Thus the columns of $\Psi(\lambda_i)$ must be the eigenvectors of **A**. But since to each eigenvalue λ_i, there is only one independent eigenvector, therefore the eigenvectors of the columns of $\Psi(\lambda_i)$ are linearly dependent on each other.

Suppose now the eigenvalue λ_i has multiplicity m, then $(\lambda-\lambda_i)^m$ must be a factor of $g(\lambda)$. Differentiating eqn(II.78) $(m-1)$ times with respect to λ,

$$\begin{aligned}
(\lambda\mathbf{I} - \mathbf{A})\,\Psi(\lambda) &= g(\lambda)\,\mathbf{I} \\
(\lambda\mathbf{I} - \mathbf{A})\,\Psi'(\lambda) + \Psi(\lambda) &= g'(\lambda)\,\mathbf{I} \\
&\vdots \\
(\lambda\mathbf{I} - \mathbf{A})\,\Psi^{(m-1)}(\lambda) + (m-1)\,\Psi^{(m-2)}(\lambda) &= g^{(m-1)}(\lambda)\,\mathbf{I}
\end{aligned} \tag{II.79}$$

But since $g(\lambda)$ has the factor $(\lambda-\lambda_i)^m$, then

$$g^{(k-1)}(\lambda_i) = 0 \qquad k = 1,2,\ldots,m$$

Hence substituting $\lambda = \lambda_i$ in eqn(II.79), one obtains

$$\begin{aligned}
(\mathbf{A} - \lambda_i\mathbf{I})\,\Psi(\lambda_i) &= \mathbf{0} \\
(\mathbf{A} - \lambda_i\mathbf{I})\,\Psi'(\lambda_i) &= \Psi(\lambda_i) \\
&\vdots \\
(\mathbf{A} - \lambda_i\mathbf{I})\,\Psi^{(m-1)}(\lambda_i) &= (m-1)\,\Psi^{(m-2)}(\lambda_i)
\end{aligned}$$

Rearranging and modifying these equations

$$\begin{aligned}
(\mathbf{A} - \lambda_i\mathbf{I})\,\Psi(\lambda_i) &= \mathbf{0} \\
(\mathbf{A} - \lambda_i\mathbf{I})^2\,\Psi'(\lambda_i) &= \mathbf{0} \\
&\vdots \\
(\mathbf{A} - \lambda_i\mathbf{I})^m\,\Psi^{(m-1)}(\lambda_i) &= \mathbf{0}
\end{aligned} \tag{II.80}$$

The m equations given by (II.80) generate m vectors and are called generalized eigenvectors. These are obtained from $\Psi(\lambda_i)$, $\Psi'(\lambda_i)$, ..., $\Psi^{(m-1)}(\lambda_i)$. It is sometimes possible to find independent generalized eigenvectors without actually using all the m equations. Consider a matrix having an eigenvalue of 1 with multiplicity 3. The three eigenvalues 1, 1, 1 can break in the following manner:-

$\{(1),(1),(1)\} \Rightarrow$ behaves as if they were distinct, the three linearly independent eigenvectors will be given by any one of $\Psi(1)$, $\Psi'(1)$ or $\Psi''(1)$

$\{(1,1),(1)\} \Rightarrow$ 2 split together and one separately, the eigenvectors will be given by any two of $\Psi(1)$, $\Psi'(1)$, and $\Psi''(1)$

$\{(1,1,1)\} \Rightarrow$ 3 split together, one eigenvector will be given by each of $\Psi(1)$, $\Psi'(1)$ and $\Psi''(1)$

Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ be the eigenvectors (generalized if multiple eigenvalue), then

$$\mathbf{P} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_N]$$

and

$$\mathbf{J} = \mathbf{P}^{-1}\mathbf{A}\,\mathbf{P} \tag{II.81}$$

where $\mathbf{J}$ is called the Jordan canonic form of $\mathbf{A}$.
When all eigenvalues of $\mathbf{A}$ are distinct, the form is diagonal — a special case of the Jordan canonic form.

The form of $\mathbf{J}$ depends on how the eigenvalues split. If they split separately, $\mathbf{J}$ is diagonal. Let $\mathbf{A}$ be a matrix having eigenvalues λ_1, λ_2, λ_2, λ_2, λ_3, λ_3, λ_4, λ_5, λ_5, λ_5, where λ_2 breaks as $\{(\lambda_2,\lambda_2,\lambda_2)\}$, λ_3 breaks as $\{(\lambda_3),(\lambda_3)\}$, λ_5 breaks as $\{(\lambda_5,\lambda_5),(\lambda_5)\}$, then the form of $\mathbf{J}$ will be

$$\mathbf{J} = \begin{bmatrix} \boxed{\lambda_1} & & & & \\ & \boxed{\begin{matrix}\lambda_2 & 1 & 0\\ 0 & \lambda_2 & 1\\ 0 & 0 & \lambda_2\end{matrix}} & & & \\ & & \boxed{\lambda_3} & & \\ & & & \boxed{\lambda_3} & \\ & & & & \boxed{\lambda_4} \end{bmatrix} \qquad \text{(II.82)}$$

The eigenvalues are boxed according to the way they break. In each box, 1 is placed above each λ if there are more than one λ in each box. The boxes shown in the Jordan canonic form of $\mathbf{A}$ are sometimes referred to as the Jordan boxes. If all the eigenvalues break together such as those in the box containing λ_2 in (II.82), the box is sometimes referred to as the Jordan normal form.

FUNCTIONS OF THE JORDAN NORMAL FORM

The Jordan normal form can be written as

$$\mathbf{J_n} = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix} = \lambda \mathbf{I} + \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \qquad \text{(II.83)}$$

Consider the case when the order of $\mathbf{J}$ is say 3

$$\mathbf{J_n} = \lambda \mathbf{I} + \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} = \lambda \mathbf{I} + \mathbf{Q} \qquad \text{(II.84)}$$

where $$\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

But $$\mathbf{Q}^2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\mathbf{Q}^k = \mathbf{0} \qquad \text{for} \quad k \geq 3$$

(It can be seen that if $\mathbf{Q}$ is of the order N, $\mathbf{Q}^k = \mathbf{0}$ for $k \geq N$). Hence, for the above example when the order is 3, then by Taylor's theorem,

$$\begin{aligned} \mathbf{f}(\mathbf{J}_n) &= \mathbf{f}(\lambda \mathbf{I} + \mathbf{Q}) \\ &= f(\lambda)\mathbf{I} + \mathbf{Q} f'(\lambda) + \frac{\mathbf{Q}^2}{2!} f''(\lambda) + \ldots \end{aligned} \tag{II.85}$$

where $$f'(\lambda) = \frac{\partial f}{\partial \lambda}$$

But $$\mathbf{Q}^k = \mathbf{0} \quad \text{for} \quad k \geq 3$$

$$\therefore \quad \mathbf{f}(\mathbf{J}_n) = f(\lambda)\mathbf{I} + \mathbf{Q} f'(\lambda) + \frac{\mathbf{Q}^2}{2!} f''(\lambda) \tag{II.86}$$

i.e. $$\mathbf{f}(\mathbf{J}_n) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2!} f''(\lambda) \\ 0 & f(\lambda) & f'(\lambda) \\ 0 & 0 & f(\lambda) \end{bmatrix} \tag{II.87}$$

In general if the order of the Jordan normal form is n, then

$$\mathbf{f}(\mathbf{J}_n) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2!} f''(\lambda) & \cdots & \frac{f^{n-1}(\lambda)}{(n-1)!} \\ 0 & f(\lambda) & f'(\lambda) & \cdots & \frac{f^{n-2}(\lambda)}{(n-2)!} \\ 0 & 0 & f(\lambda) & \cdots & \frac{f^{n-3}(\lambda)}{(n-3)!} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & f(\lambda) \end{bmatrix} \tag{II.88}$$

Thus if the Jordan canonic form of $\mathbf{A}$ is broken into Jordan boxes, i.e.

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & \\ & \mathbf{J}_2 & \\ & & \ddots \\ & & & \mathbf{J}_m \end{bmatrix}$$

then

$$\mathbf{f(J)} = \begin{bmatrix} \mathbf{f}(\mathbf{J}_1) & & \\ & \mathbf{f}(\mathbf{J}_2) & \\ & & \ddots \\ & & & \mathbf{f}(\mathbf{J}_m) \end{bmatrix} \qquad \text{(II.89)}$$

Hence, it can be seen that finding a function of a matrix in Jordan form is a one step process. Using the properties of similar matrices, the discrete transition matrix, $\Phi(k) = \mathbf{A}^k$, can be evaluated very simply.

II.8 *z*-Transformation

In section I.2, it has been defined that the *z*-transform of a function $x(t)$ is given by

$$X(z) = \mathcal{Z}[x(t)] = \sum_{n=0}^{\infty} x(n).z^{-n} \qquad \text{(II.90)}$$

The process of obtaining $x(n)$ from $X(z)$ is termed the *inverse z transformation*. The inverse *z* transform of $X(z)$ is denoted by

$$\mathcal{Z}^{-1}[X(z)] = \text{inverse transform of } X(z) = x(n) \qquad \text{(II.91)}$$

In general there are three ways of carrying the inverse *z*-transform operation. These are shown as follows:-

a) SERIES EXPANSION

The values of $x(k)$ for $k = 0, 1, 2, \ldots$ are obtained from $X(z)$ simply by expanding $X(z)$ into a power series of z^{-1}. Expanding (II.90),

$$X(z) = x(0) + x(1)z^{-1} + x(2)z^{-2} + \ldots \qquad \text{(II.92)}$$

Clearly, the coefficients of z^{-n} represents the values of $x(t)$ at $x(n)$. Therefore given $X(z)$, it can be expanded into a power series of z^{-1}, and $x(k)$ for $k = 0, 1, 2, \ldots$ are obtained from the coefficients of the power series.

b) PARTIAL FRACTION EXPANSION

To find the inverse z-transform of $X(z)$ by partial fraction expansion, the function $X(z)/z$ should first be expanded into the following form:

$$\frac{X(z)}{z} = \frac{A_1}{z+a_1} + \frac{A_2}{z+a_2} + \frac{A_3}{z+a_3} + \ldots \qquad \text{(II.93)}$$

and then the inverse z-transform of $X(z)$ is given by

$$x(n) = \mathcal{Z}^{-1}[X(z)] = \mathcal{Z}^{-1}\left[\frac{A_1 z}{z+a_1} + \frac{A_2 z}{z+a_2} + \frac{A_3 z}{z+a_3} + \ldots\right] \qquad \text{(II.94)}$$

where each of the individual inversion $\mathcal{Z}^{-1}\left[\frac{A_i z}{z+a_i}\right]$ is looked up from the z-transform table.

c) INVERSE FORMULA

Multiplying both sides of (II.90) by z^{k-1} and then performing a closed-line integration on both sides of the equation, one obtains

$$\oint_{\Gamma} X(z)\, z^{k-1} dz = \sum_{n=0}^{\infty} x(n) \oint_{\Gamma} z^{k-n-1} dz \qquad \text{(II.95)}$$

where Γ, the integration path, is within the regions of convergence of the infinite series of (II.90), hence enabling the summation and integration to be interchanged. By Cauchy's theorem,

$$\oint_{\Gamma} z^{k-n-1} dz = \begin{cases} 2\pi j & \text{for } k=n \\ 0 & \text{otherwise} \end{cases} \qquad \text{(II.96)}$$

Thus, substituting into (II.95)

$$x(k) = \frac{1}{2\pi j} \oint_{\Gamma} X(z)\, z^{k-1} dz \qquad \text{(II.97)}$$

Some important properties of the z-transforms are stated below without proof. The details of the proofs are available in the literature.{ 29},{ 41},{ 44},{ 53},{ 60}.

(i) $\mathcal{Z}[af(k)] = aF(z)$ (II.98)

where a is a constant and $F(z) = \mathcal{Z}[f(k)]$

(ii) $\mathcal{Z}[f_1(k) \pm f_2(k)] = F_1(z) \pm F_2(z)$ (II.99)

(iii) $\mathcal{Z}[f(k-n)] = z^{-n} F(z)$ (II.100)

where n is a positive integer.

(iv) $\mathcal{Z}[f(k+n)] = z^{n}\left[F(z) - \sum_{k=0}^{n-1} f(k) z^{-k}\right]$ (II.101)

(v) $\lim_{k\to\infty} f(k) = \lim_{z\to\infty} F(z)$ (initial value theorem) (II.102)

if the limit exists.

(vi) $\lim_{k\to\infty} f(k) = \lim_{z\to 1} (1- z^{-1})\ F(z)$ (final value theorem) (II.103)

if $(1 - z^{-1})\ F(z)$ does not have any pole which lies outside the unit circle $|z| = 1$ in the z plane

II.9 z-Transform Solution of Discrete State Equations

The discrete dynamic equations

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}u(k) \\ y(k) &= \mathbf{C}\mathbf{x}(k) + D\,u(k) \end{aligned} \tag{II.104}$$

can be solved by means of z-transform method. Taking the z-transform on both sides of the state equation yields

$$z\,\mathbf{X}(z) - z\,\mathbf{x}(0^+) = \mathbf{A}\mathbf{X}(z) + \mathbf{B}\,U(z) \tag{II.105}$$

Solving for $\mathbf{X}(z)$ from (II.105), one obtains

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{A})^{-1}\ z\mathbf{x}(0^+) + (z\mathbf{I} - \mathbf{A})^{-1}\ \mathbf{B}\,U(z) \tag{II.106}$$

which has an inverse transform,

$$\mathbf{x}(k) = \mathcal{Z}^{-1}\left[(z\mathbf{I} - \mathbf{A})^{-1}z\right]\mathbf{x}(0^+) + \mathcal{Z}^{-1}\left[(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}U(z)\right] \tag{II.107}$$

Comparing this result with that of (II.57), the following identities are established:

$$\mathbf{A}^k = \mathcal{Z}^{-1}\left[(z\mathbf{I} - \mathbf{A})^{-1}z\right] \tag{II.108}$$

and

$$\sum_{i=0}^{k-1} \mathbf{A}^{k-i-1}\mathbf{B}\,u(i) = \mathcal{Z}^{-1}\left[(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}U(z)\right] \tag{II.109}$$

To prove (II.108), take the z-transform on both sides of the equation, then premultiply both sides by $(z\mathbf{I} - \mathbf{A})$,

$$\text{LHS:} \quad (z\mathbf{I} - \mathbf{A})\,\mathcal{Z}[\mathbf{A}^k] = (z\mathbf{I} - \mathbf{A})\sum_{k=0}^{\infty}\mathbf{A}^k z^{-k}$$
$$= (z\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A}z^{-1} + \mathbf{A}^2 z^{-2} + \ldots)$$
$$= z\mathbf{I} \qquad \text{(II.110)}$$

$$\text{RHS:} \quad (z\mathbf{I} - \mathbf{A})\,\mathcal{Z}\left\{\mathcal{Z}^{-1}[(z\mathbf{I} - \mathbf{A})^{-1}z]\right\} = z\mathbf{I} \qquad \text{(II.111)}$$

Eqn(II.109) is verified also by taking the z-transform on both sides of the equation. Therefore,{40}

$$\mathcal{Z}\left[\sum_{i=0}^{k-1}\mathbf{A}^{k-i-1}\mathbf{B}u(i)\right] = \sum_{k=0}^{\infty}\sum_{i=0}^{k-1}\mathbf{A}^{k-i-1}\mathbf{B}u(i)\,z^{-k}$$
$$= \sum_{k=0}^{\infty} z^{-k+i}\sum_{i=0}^{k-1}\mathbf{A}^{k-i-1}\mathbf{B}u(i)\,z^{-i}$$
$$= \sum_{k=0}^{\infty} z^{-k+i}\,\mathbf{A}^{k-i-1}\mathbf{B}U(z) \qquad \text{(II.112)}$$

Now the exponent of $\mathbf{A}$ cannot be negative, therefore

$$\sum_{k=0}^{\infty} z^{-k+i+1}\,\mathbf{A}^{k-i-1} = \sum_{k=0}^{\infty} z^{-k}\,\mathbf{A}^k = \mathcal{Z}[\mathbf{A}^k]$$

and hence (II.112) can be written as

$$\mathcal{Z}\left[\sum_{i=0}^{k-1}\mathbf{A}^{k-i-1}\mathbf{B}u(i)\right] = z^{-1}\,\mathcal{Z}[\mathbf{A}^k]\,\mathbf{B}U(z)$$
$$= (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}U(z) \qquad \text{(II.113)}$$

where (II.108) has been made use of.

The matrix $\mathbf{A}^k$ can be written as $\Phi(k)$ and, as has been pointed out in section(II.7), is called the discrete state-transition matrix. Then the discrete state-transition equation of (II.104) becomes

$$\mathbf{x}(k) = \Phi(k)\,\mathbf{x}(0^+) + \sum_{i=0}^{k-1}\Phi(k-i-1)\,\mathbf{B}u(i) \qquad \text{(II.114)}$$

Substituting (II.114) into the output equation of (II.104), one obtains

$$y(k) = \mathbf{C}\,\Phi(k)\,\mathbf{x}(0^+) + \mathbf{C}\sum_{i=0}^{k-1}\Phi(k-i-1)\,\mathbf{B}\,u(i) + Du(k) \tag{II.115}$$

II.10 Transfer Functions and Impulse Sequences of Discrete-Data Systems

The z-transform of the dynamic equations of a discrete-data system are, ignoring initial conditions,

$$z\,\mathbf{X}(z) = \mathbf{A}\,\mathbf{X}(z) + \mathbf{B}\,U(z) \tag{II.117}$$

$$Y(z) = \mathbf{C}\,\mathbf{X}(z) + DU(z) \tag{II.118}$$

From (II.117), it can be written that

$$\mathbf{X}(z) = (z\,\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{B}\,U(z) \tag{II.11-}$$

and substituting into (II.118)

$$Y(z) = \mathbf{C}\,(z\,\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{B}\,U(z) + DU(z) \tag{II.12}$$

from which the transfer function is defined:

$$H(z) = \frac{Y(z)}{U(z)} = \mathbf{C}\,(z\,\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D \tag{II.121}$$

Here the input-output relationship is derived for a single-input, single-output system. The generalization to multiple inputs and multiple outputs is straightforward. The inverse z-transform of $H(z)$ is then

$$\mathcal{Z}^{-1}[H(z)] = \mathcal{Z}^{-1}[\mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + D]$$

$$= g(k)$$

$$= \mathbf{C}\,\Phi(k-1)\mathbf{B} + D\delta(t) \qquad \text{(II.122)}$$

and $g(k)$ is regarded as the impulse sequence of the linear process.

From (II.121), one obtains the following expression

$$H(z) = \frac{\mathbf{C}\ \mathbf{adj(A)B} + \det(z\mathbf{I}-\mathbf{A})D}{\det(z\mathbf{I} - \mathbf{A})} \qquad \text{(II.123)}$$

where $\mathbf{adj(A)}$ is the adjoint matrix of $\mathbf{A}$ and $\det(z\mathbf{I} - \mathbf{A})$ is the determinant of the matrix $(z\mathbf{I} - \mathbf{A})$. The charateristic equation of the system is defined as:

$$\det(z\mathbf{I} - \mathbf{A}) = 0 \qquad \text{(II.124)}$$

It can be seen that the roots of the characteristic equation are the eigenvalues of the matrix $\mathbf{A}$, i.e. the eigenvalues of the matrix $\mathbf{A}$ are identical to the poles of the transfer function $H(z)$.

II.11 Stability Consideration

a) EQUILIBRIUM

To study the motion of a sampled-data system, consider the autonomous system

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k)) \qquad \text{(II.125)}$$

where $k = 0, 1, 2, \ldots$

For $k = 0$, it is defined that

$$\mathbf{x}(1) \triangleq \mathbf{f}'(\mathbf{x}(0))$$

One more iteration leads to

$$\mathbf{x}(2) = \mathbf{f}(\mathbf{x}(1)) = \mathbf{f}\left[\mathbf{f}(\mathbf{x}(0))\right] \triangleq \mathbf{f}^2(\mathbf{x}(0)) \tag{II.126}$$

The notation $\mathbf{f}^2(\mathbf{x}(0))$ is not intended to indicate that the function is squared. Extending the notation,

$$\mathbf{x}(n) = \mathbf{f}^n(\mathbf{x}(0)) \tag{II.127}$$

represents the "solution".

If a control vector $\mathbf{u}$ is added to (II.125), then

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), \mathbf{u}(k)) \tag{II.128}$$

To arrive at the equilibrium state of the process, let $\mathbf{u}(k) = \mathbf{c}$, a constant vector. Any vector $\mathbf{x}_e$ which satisfies the equation

$$\mathbf{x}_e = \mathbf{f}^n(\mathbf{x}_e, \mathbf{c}) \tag{II.129}$$

is called an equilibrium state.

b) STABILITY

Loosely speaking, a system is stable if small disturbances in the system cause correspondingly small deviations in the equilibrium state. If a system returns to its equilibrium state with increasing time, then the system is *asymtotically stable*. The foregoing ideas are generalized formally in the following definitions

(i) An equilibrium state $\mathbf{x}_e$ of (II.128) is stable if, for any $\varepsilon > 0$,

there corresponds a $\delta > 0$ such that, if

$$\|\mathbf{x}(0) - \mathbf{x}_e\| \leq \delta \tag{II.130}$$

then

$$|\mathbf{x}(k) - \mathbf{x}_e| < \varepsilon \tag{II.131}$$

where $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{\frac{1}{2}}$ represents the norm of the column state vector .

(ii) The equilibrium state is asymptotically stable if it is stable and if

$$\lim_{k\to\infty} \|\mathbf{x}(k) - \mathbf{x}_e\| = 0 \tag{II.132}$$

(iii) The equilibrium state is uniformly stable if δ is independent of the initial time t_o.

(iv) If the solution is asymototically stable and if δ can be arbitrarily large, the equilibrium state is stable in the large (globally stable).

c) CONSTANT LINEAR SYSTEMS

Consider the free system represented by the following state equation:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) \qquad \mathbf{x}(t_o) = \mathbf{x}(0) \tag{II.133}$$

whose solution is given by

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) \tag{II.134}$$

For simplicity, assume $\mathbf{A}$ has no multiple eigenvalue, so that $\mathbf{A}$ is similar to a diagonal matrix Λ such that

$$\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P} = \Lambda \tag{II.135}$$

where $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_N \end{bmatrix}$

and $\mathbf{P} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_N]$

$\lambda_1, \lambda_2, \cdots \lambda_N$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ are the respective eigenvalues and eigenvectors of $\mathbf{A}$. In other words, if the following change of variable is made,

$$\mathbf{x}(k) = \mathbf{P}\,\xi(k) \tag{II.136}$$

then substituting (II.136) into (II.133) and rearranging,

$$\xi(k+1) = \mathbf{P}^{-1}\mathbf{A}\,\mathbf{P}\,\xi(k) \tag{II.137}$$

Equation(II.137) has the solution

$$\begin{aligned} \xi(k) &= (\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P})^k\,\xi(0) \\ &= \mathbf{P}^{-1}\,\mathbf{A}^k\,\mathbf{P}\,\xi(0) \end{aligned} \tag{II.138}$$

It is clear that (II.133) is stable if and only if (II.138) is stable. The stability of (II.138) is investigated by examining the kth power of (II.135),

$$\mathbf{P}^{-1}\,\mathbf{A}^k\,\mathbf{P} = \mathbf{\Lambda}^k \tag{II.139}$$

If all $|\lambda_i| < 1$. then the system is asymptotically stable; if one or more $|\lambda_i| = 1$, the system is stable; and if one or more $|\lambda_i| > 0$, the system is unstable. The following theorems are stated without proof{32}

(i) The constant system of (II.133) is stable if all eignvalues of the transition matrix $\mathbf{A}$ satisfy the condition

$$|\lambda_i| \leq 1 \qquad i = 1, 2, \ldots, N \tag{II.140}$$

If $|\lambda_i| \neq 1$ for any i, then the condition of (II.140) is both necessary and sufficient.

(ii) The zero state $\mathbf{x}(k+1) = \mathbf{0}$ of the free system

$$\mathbf{x}(k+1) = \mathbf{A}\,\mathbf{x}(k) \tag{II.141}$$

is globally stable iff every element of $\mathbf{A}^k$ tends uniformly to zero as $k \to \infty$.

II.12 Résumé

The material in this chapter gave a brief account of the method of state-space analysis, and demonstrates the versatility of the state-space approach. The idea of state variables was first introduced, the method of obtaining the state-equations from the difference equation and that from the state-diagrams were described. Then the solution of the state-equation both by the direct method and by using the z-transform were shown, and finally, the relationship between state variables and system functions, and the relationship between state variables and stability were derived.

This chapter is considered basic to ensuing work, and accordingly, much of the material will be used in many of the subsequent chapters.

CHAPTER III

MATHEMATICAL MODELS AND DESIGN OF AN IDEAL MULTIRATE DIGITAL FILTER

III.1 Introduction

In chapter I, the basic operation of a conventional digital filter has been described. Here a generalized formula giving the resulting transfer functions of a digital filter when the shifting is continued N times during each pulse repetition interval, where N is a positive integer, is derived. Several approaches to the analysis are taken and compared. Some of the interesting properties of the transfer function of such a "multirate" digital filter are shown. Computer simulations are performed to show that the multirate filters designed by using the formula derived give the desired outputs.

III.2 Derivation of the Transfer Functions using Difference Equations and the z-transform

The following derivation follows closely the method of Fjällbrant [13]. To illustrate the analysis, the derivation of the transfer functions for a second order digital filter is shown below, N being taken to be two.

If a second order digital filter realized in the direct canonic form is used in a double-rate ($N = 2$) fashion, the delay of each register is $T/2$ while the input sampling period is T. There are two separate output sequences, one sampled at $t = nT$

and is designated $y_1(n)$, while the other, sampled at $t = (n + \frac{1}{2})T$, is designated $y_2(n + \frac{1}{2})$.

Consider fig III.1(a)where $t = nT$. The input sampler $S_{I/P}$ is closed, hence the input signal $u(n)$, in the form of a binary word, comes in. At the output, the sampler S_1 is closed while S_2 is open. Hence, only the first output sequence, $y_1(n)$, exists.

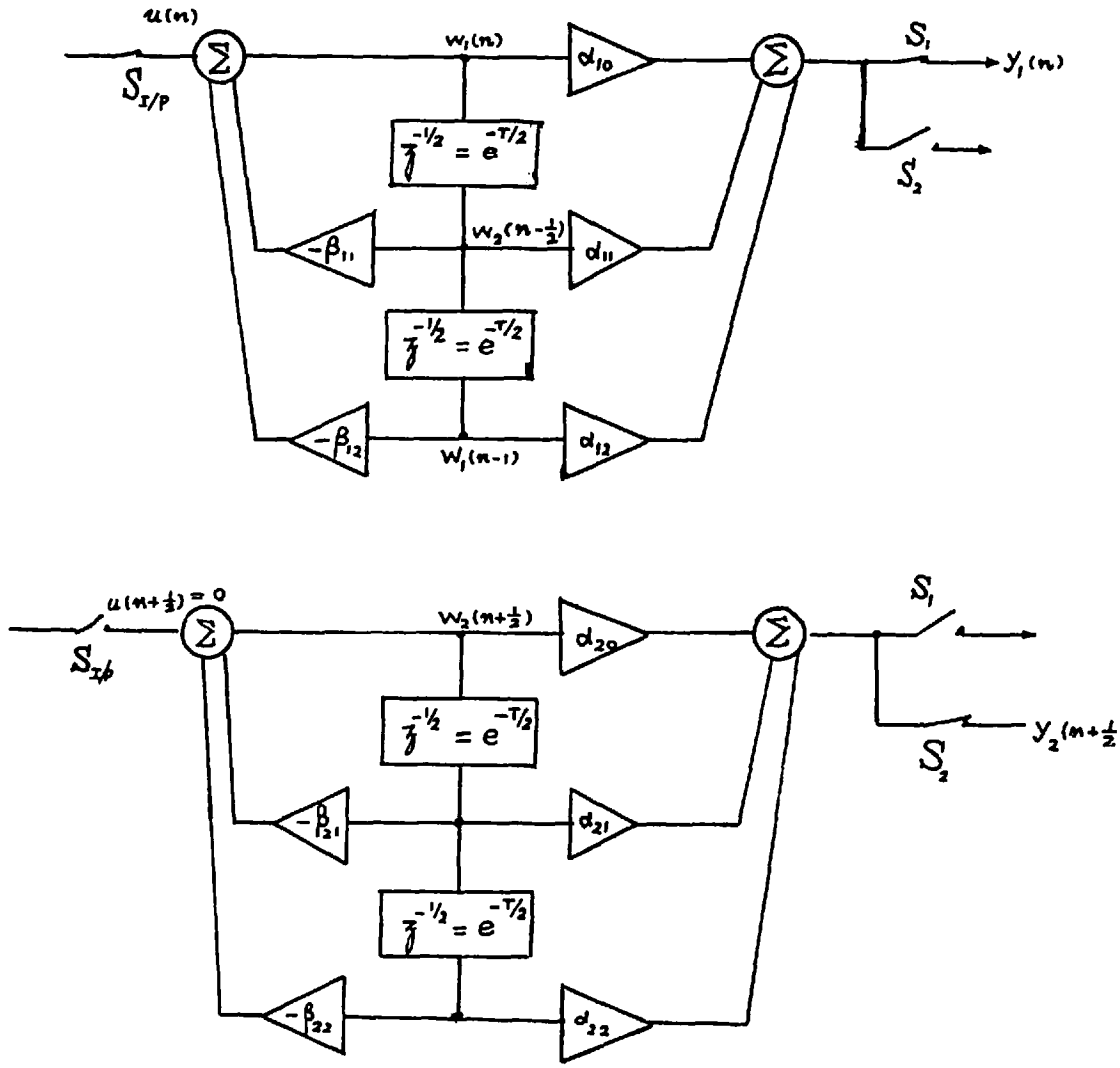


Fig III.1 A Second Order Double-Rate Digital Filter in the Direct Canonic Configuration (a) $t=nT$, (b) $t=(n+\frac{1}{2})T$

Let α_{10}, α_{11}, α_{12}, $-\beta_{11}$ and $-\beta_{12}$ be the multiplication coefficients at this instant $t = nT$ and introducing two intermediate variables $w_1(n)$ and $w_2(n + \frac{1}{2})$ to facilitate calculation, the following equations can be written,

$$w_1(n) = -\beta_{11}w_2(n-\tfrac{1}{2}) - \beta_{12}w_1(n-1) + u(n) \qquad \text{(III.1)}$$

$$y_1(n) = \alpha_{10}w_1(n) + \alpha_{11}w_2(n-\tfrac{1}{2}) + \alpha_{12}w_1(n-1) \qquad \text{(III.2)}$$

Fig III.1(b) shows the double-rate filter at the instant $t = (n + \frac{1}{2})T$. Since there is no input at this instant, $u(n + \frac{1}{2}) = 0$. The output sampler S_1 is open and S_2 is closed, hence only the output sequence $y_2(n + \frac{1}{2})$ exists. Let α_{20}, α_{21}, α_{22}, $-\beta_{21}$ and $-\beta_{22}$ be the multiplication coefficients at $t = (n + \frac{1}{2})T$, again the following difference equations can be written,

$$w_2(n+\tfrac{1}{2}) = -\beta_{21}w_1(n) - \beta_{22}w_2(n-\tfrac{1}{2}) \qquad \text{(III.3)}$$

$$y_2(n+\tfrac{1}{2}) = \alpha_{20}w_2(n+\tfrac{1}{2}) + \alpha_{21}w_1(n) + \alpha_{22}w_2(n-\tfrac{1}{2}) \qquad \text{(III.4)}$$

Taking the z-transform of equation (III.3)

$$z^{\frac{1}{2}}W_2(z) = -\beta_{21}W_1(z) - \beta_{22}z^{\frac{1}{2}}W_2(z)$$

$$\text{i.e.} \quad z^{\frac{1}{2}}W_2(z) = \frac{-\beta_{21}}{1 + \beta_{22}z^{-1}}W_1(z) \qquad \text{(III.5)}$$

Taking the z-transform of eqn (III.1) and substituting eqn (III.5),

$$W_1(z) = -\beta_{11}z^{-1}\left(\frac{-\beta_{21}}{1+\beta_{22}z^{-1}}\right)W_1(z) - \beta_{12}z^{-1}W_1(z) + U(z)$$

$$\text{i.e.} \quad W_1(z) = \left\{\frac{1 + \beta_{22}z^{-1}}{1 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z^{-1} + \beta_{12}\beta_{22}z^{-2}}\right\}U(z) \qquad \text{(III.6)}$$

From eqn (III.5) and (III.6)

$$z^{\frac{1}{2}}W_2(z) = \left\{ \frac{-\beta_{21}}{1 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z^{-1} + \beta_{12}\,\beta_{22}\,z^{-2}} \right\} U(z) \qquad \text{(III.7)}$$

Taking the z-transform of eqn (III.2), substituting eqn (III.6) and (III.7) and rearranging,

$$H_1(z) = \frac{Y_1(z)}{U(z)} = \frac{\alpha_{10}+(\alpha_{10}\beta_{22}-\alpha_{11}\beta_{21}+\alpha_{12})z^{-1}+\alpha_{12}\beta_{22}z^{-2}}{1 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z^{-1} + \beta_{12}\beta_{22}z^{-2}} \qquad \text{(III.8)}$$

Taking the z-transform of eqn (III.4), substituting in eqns (III.6) and (III.7) and rearranging, one obtains

$$H_2(z) = \frac{z^{\frac{1}{2}}Y_2(z)}{U(z)} = \frac{(-\alpha_{20}\beta_{21}+\alpha_{21})+(\alpha_{21}\beta_{22}-\alpha_{22}\beta_{21})z^{-1}}{1 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z^{-1} + \beta_{12}\beta_{22}z^{-2}} \qquad \text{(III.9)}$$

The extra $z^{\frac{1}{2}}$ in $Y_2(z)/U(z)$ simply means that the sequence $y_2(n + \frac{1}{2})$ is delayed by $T/2$. It is observed that both $H_1(z)$ and $H_2(z)$ have the same poles and that they differ only in the numerator polynomial.

Following similar procedures, the transfer functions of a "triple-rate" ($N = 3$) second order filter realized in the direct canonic form are found to be:-

$$H_1(z) = \frac{Y_1(z)}{U(z)} = \frac{\alpha_{10} + \{\alpha_{11}(\beta_{21}\beta_{31}-\beta_{32}) - \alpha_{10}\beta_{31}\beta_{32} - \alpha_{12}\beta_{21}\}z^{-1} + \alpha_{12}\beta_{22}\beta_{32}z^{-2}}{1 + (\beta_{11}\beta_{21}\beta_{31} - \beta_{11}\beta_{32} - \beta_{21}\beta_{12} - \beta_{31}\beta_{22})z^{-1} + \beta_{12}\beta_{22}\beta_{32}z^{-2}} \qquad \text{(III.10)}$$

$$H_2(z) = \frac{z^{1/3}Y_2(z)}{U(z)} = \frac{(\alpha_{21}-\alpha_{20}\beta_{21})+(\alpha_{20}\beta_{22}\beta_{32}-\alpha_{21}\beta_{31}\beta_{32}+\alpha_{22}\beta_{21}\beta_{31}-\alpha_{22}\beta_{32})z^{-1}}{1 + (\beta_{11}\beta_{21}\beta_{31} - \beta_{11}\beta_{32} - \beta_{21}\beta_{12} - \beta_{31}\beta_{22})z^{-1} + \beta_{12}\beta_{22}\beta_{32}z^{-2}} \qquad \text{(III.11)}$$

$$H_3(z) = \frac{z^{2/3}Y_3(z)}{U(z)} = \frac{(\alpha_{30}\beta_{21}\beta_{31} - \alpha_{30}\beta_{32} - \alpha_{31}\beta_{21} + \alpha_{32}) + (\alpha_{31}\beta_{22}\beta_{32} - \alpha_{32}\beta_{31}\beta_{22})z^{-1}}{1 + (\beta_{11}\beta_{21}\beta_{31} - \beta_{11}\beta_{32} - \beta_{21}\beta_{12} - \beta_{31}\beta_{22})z^{-1} + \beta_{12}\beta_{22}\beta_{32}z^{-2}} \qquad \text{(III.12)}$$

The above method fails to give a generalized formula for the transfer functions when N is a general positive integer. Also if the direct canonic configuration of fig III.1 is varied, for instance, to a transposed configuration [25] where the auxiliary variables w_1 and w_2 may vary after being passed on through the shift registers, then such an analysis will be awkwardly complicated.

III.3 Derivation of Transfer Functions of a Multirate Digital Filter using Discrete Convolution

A different approach from the above method has been taken by Ragazzini and Franklin [53] to give a generalized formula for the transfer function of a multirate digital system. The following analysis is a slight extenstion of their method.

First consider the system shown in fig III.3 where an input U is sampled with a uniform period T and applied to the continuous G. The output of G is sampled at an increased rate with period T/N to form a sequence of output samples whose transform (which will be defined shortly) is designated $Y(z_N)$. The analysis first requires the relationship between this output transform and the input transform $U(z)$.

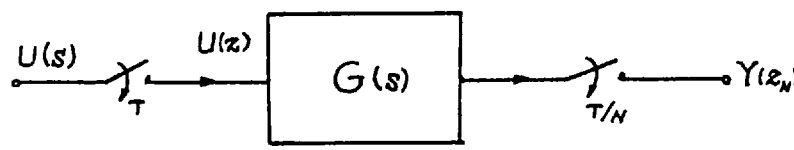


Fig III.3 A Multirate System

By the convolution theorem,

$$y(t) = \sum_{k=0}^{\infty} u(kT).g(t-kT) \qquad \text{(III.13)}$$

However, the samples which appear at the output sampler are the values of $y(t)$ at the instants $t = \frac{lT}{N}$, or

$$y(\tfrac{l}{N}) = \sum_{k=0}^{\infty} u(k).g(\tfrac{l}{N} - k) \qquad \text{(III.14)}$$

If the transform of this output is to be of use, it must obviously include all the samples in eqn (III.14); that is, the output transform must be defined on samples separated by T/N rather than the input sampling period T. To distinguish the transform variables according to the separation between successive samples which they represent, the variable z_N will be used in the pulse transform of samples separated by T/N, and the variable z retained for sequences separated by T. Hence the z_N-transform of the output in the system shown in fig III.3 is defined as

$$Y(z_N) = \sum_{l=0}^{\infty} y(\tfrac{l}{N})\ z_N^{-l}$$

$$= \sum_{l=0}^{\infty}\sum_{k=0}^{\infty} u(k).g(\tfrac{l}{N} - k).z_N^{-l} \qquad \text{(III.15)}$$

For a convergent series of eqn (III.15), summation with respect to l and k can be interchanged, thus

$$Y(z_N) = \sum_{k=0}^{\infty} u(k) \sum_{l=0}^{\infty} g(\tfrac{l}{N} - k)\ z_N^{-l} \qquad \text{(III.16)}$$

In the second sum, it is always possible to find an integer j such that $j = (l - kN)$, and the transform may then be written in terms of j as follows:

$$Y(z_N) = \sum_{k=0}^{\infty} u(k) \sum_{j=0}^{\infty} g\left(\frac{j}{N}\right) z_N^{-(j+kN)} \qquad \text{(III.17)}$$

The limits in the second sum of eqn (III.17) are from $j = 0$ to $j = \infty$ rather than from $j = -kN$ since the realisable impulse response $g(t)$ is zero for negative values of the argument. Separating out the powers of z_N, then,

$$Y(z_N) = \sum_{k=0}^{\infty} u(k) \left(z_N^N\right)^{-k} \sum_{j=0}^{\infty} g\left(\frac{j}{N}\right) z_N^{-j}$$

$$= U(z_N^N).G(z_N)$$

$$= U(z).G(z_N) \qquad \text{(III.18)}$$

That the function $U(z_N^N)$ is in fact the z-transform of $U(s)$ (based on samples separated by T) with the variable z replaced by z_N^N has been made use of in (III.18). The transform $G(z_N)$ is the ordinary pulse transfer function of the linear system, based on a sample separation of T/N. The variable z_N identifies the period of the samples used in determining $G(z_N)$.

Now that $Y(z_N)$ has been obtained; suppose this series $Y(z_N)$ is passed through a sampler that is synchronized with the input sampler, sampling with a period T (fig III.4), then one obtains an output series $Y_1(z)$ which has samples separated by a period T.

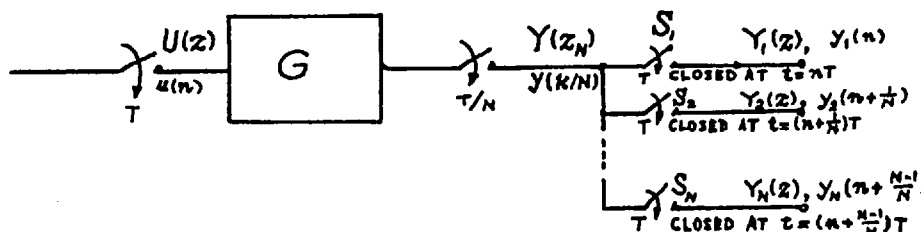


Fig III.4 A Multirate Digital Filter with Multi-Output

If a series of samplers S_1, S_2, ..., S_N are connected at the output such that S_1 closes at $t = nT$, S_2 closes at $t = (n + 1/N)T$, S_N closes at $t = (n + \frac{N-1}{N})T$, and all of them sampling with a period T, then there are N output sequences $y_1(n)$, $y_2(n + \frac{1}{N})$,.., $y_N(n + \frac{N-1}{N})$. It is desired to derive the relationship between the transform of these sequences, $Y_i(z)$ and $Y(z_N)$.

The output sequence $y(k/N)$ has a sample separation of T/N and hence its transform can be defined as

$$Y(z_N) = \sum_{k=0}^{\infty} y(\frac{k}{N}) . \; z_N^{-k} \tag{III.19}$$

Consider the ith output sequence $y_i(n + \frac{i-1}{N})$ extracted from the sequence $y(k/N)$. The sample separation is T and its transform can be defined as

$$Y_i(z) \triangleq \sum_{n=0}^{\infty} y_i(n + \frac{i-1}{N}) \; z^{-(n+\frac{i-1}{N})}$$

$$= z^{-(\frac{i-1}{N})} \sum_{n=0}^{\infty} y_i(n + \frac{i-1}{N}) \; z^{-n} \tag{III.20}$$

These transforms of equations (III.19) and (III.20) are related, since (III.20) contains only a portion of the samples of (III.19). By the inversion theorem described in section II.8(c)

$$y(\frac{k}{N}) = \frac{1}{2\pi j} \oint_{\Gamma} Y(z_N) \; z_N^{k-1} \; dz_N \tag{III.21}$$

Substituting (III.21) in (III.20) with $k = nN + i - 1$

$$Y_i(z) = z^{-(\frac{i-1}{N})} \sum_{n=0}^{\infty} \left\{ \frac{1}{2\pi j} \oint_{\Gamma} Y(z_N) \cdot z_N^{nN+i-1} \frac{dz_N}{z_N} \right\} z^{-n}$$

$$= z^{-(\frac{i-1}{N})} \frac{1}{2\pi j} \oint_{\Gamma} Y(z_N) \left(\sum_{n=0}^{\infty} z_N^{nN} z^{-N} \right) z_N^{i-1} \frac{dz_N}{z_N}$$

$$= \frac{z^{-(\frac{i-1}{N})}}{2\pi j} \oint_{\Gamma} z_N^{i-1} \cdot Y(z_N) \left(\frac{1}{1 - z_N^N z^{-1}} \right) \frac{dz_N}{z_N} \qquad \text{(III.22)}$$

The contour Γ on the z_N-plane must be so chosen that it encompasses all the poles of $[z_N^{i-1} Y(z_N) / z_N]$ but excludes the poles contributed by the factor $\{1/(1 - z_N^N z^{-1})\}$. The reason for this is that in the interchanging of summation and integration in (III.22), it is required that the infinite sum $\{\sum_{n=0}^{\infty} z_N^{nN} z^{-n}\}$ be absolutely convergent. This is assured only if $|z_N^N z^{-1}|$ is less than unity. Thus the factor $(1 - z_N^N z^{-1})$ cannot be zero in the region over which (III.22) is to be valid and the poles introduced by this factor must lie outside the contour of integration.

Substituting (III.18) into (III.22) and rearranging, it can be written that

$$H_i(z) = \frac{z^{\frac{i-1}{N}} Y_i(z)}{U(z)}$$

$$= \frac{1}{2\pi j} \oint_{\Gamma} \frac{z_N^{i-1} G(z_N)}{z_N} \left(\frac{1}{1 - z_N z^{-1}} \right) dz_N \qquad \text{(III.23)}$$

Eqn (III.23) expresses the ith transfer function relating the ith output sequence $y_i(n + \frac{i-1}{N})$ to the input $u(n)$ in terms of the pulse transfer function of a time-invariant system $G(z_N)$ which operates at a rate N times faster than that of the input.

It can be seen from (III.23) that the poles of $H_i(z)$ and $G(z_N)$ are related in a way that if $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the poles of $G(z_N)$, then the poles of $H_i(z)$ will be $\Lambda_1, \Lambda_2, \ldots, \Lambda_m$ where

$$\Lambda_i = \lambda_i^N \qquad \text{(III.24)}$$

III.4 Verification of the Derivation of $H_i(z)$ with a Second Order Double-Rate Filter

It can easily demonstrated that equation (III.23) gives the same results as eqn(III.8) and (III.9) for a double-rate second order digital filter. Consider the diagram in fig III.1. If $\alpha_{10} = \alpha_{20} = \alpha_0$, $\alpha_{11} = \alpha_{21} = \alpha_1$, $\alpha_{12} = \alpha_{22} = \alpha_2$, $\beta_{11} = \beta_{21} = \beta_1$ and $\beta_{12} = \beta_{22} = \beta_2$, then

$$G(z_N) = \frac{\alpha_0 z_N^2 + \alpha_1 z_N + \alpha_2}{z_N^2 + \beta_1 z_N + \beta_2} = \frac{\alpha_0 z_N^2 + \alpha_1 z_N + \alpha_2}{(z_N - A_1)(z_N - A_2)} \qquad \text{(III.25)}$$

where
$$A_1 = -\frac{\beta_1}{2} + \frac{\sqrt{\beta_1^2 + 4\beta_2}}{2}$$

$$A_2 = -\frac{\beta_1}{2} - \frac{\sqrt{\beta_1^2 + 4\beta_2}}{2}$$

Substituting $G(z_N)$ into eqn (III.23), with $N = 2$,

$$H_1(z) = \frac{Y_1(z)}{U(z)} = \frac{1}{2\pi j}\oint_\Gamma \frac{\alpha_o z_N^2 + \alpha_1 z_N + \alpha_2}{z_N(z_N - A_1)(z_N - A_2)}\left(\frac{1}{1 - z_N^2 z^{-1}}\right) dz_N \qquad \text{(III.26)}$$

$$H_2(z) = \frac{z^{\frac{1}{2}} Y_2(z)}{U(z)} = \frac{1}{2\pi j}\oint_\Gamma \frac{\alpha_o z_N^2 + \alpha_1 z_N + \alpha_2}{(z_N - A_1)(z_N - A_2)}\left(\frac{1}{1 - z_N z^{-1}}\right) dz_N \qquad \text{(III.27)}$$

In evaluating the integrals of eqns (III.26) and (III.27), it is possible to do so either by obtaining the residues at the poles of $(z_N^{i-1}/z_N).\,G(z_N)$ which are contained inside Γ or by obtaining the poles of $1/(1 - z_N^2 z^{-1})$ which lie outside Γ.

Evaluating the integral by obtaining the residues enclosed by Γ (III.26) becomes,

$$H_1(z) = \frac{\alpha_2}{A_1 A_2} + \frac{\alpha_o A_1^2 + \alpha_1 A_1 + \alpha_2}{A_1(A_1 - A_2)(1 - A_1^2 z^{-1})} + \frac{\alpha_o A_2^2 + \alpha_1 A_2 + \alpha_2}{A_2(A_2 - A_1)(1 - A_2^2 z^{-1})}$$

$$= \frac{\alpha_o + (\alpha_o \beta_2 - \alpha_1 \beta_1 + \alpha_2) z^{-1} + \alpha_2 \beta_2}{1 + (2\beta_2 - \beta_1^2) z^{-1} + \beta_2^2 z^{-2}} \qquad \text{(III.28)}$$

Similarly, (III.27) becomes

$$H_2(z) = \frac{\alpha_o A_1^2 + \alpha_1 A_1 + \alpha_2}{(A_1 - A_2)(1 - A_1^2 z^{-1})} + \frac{\alpha_o A_2^2 + \alpha_1 A_2 + \alpha_2}{(A_2 - A_1)(1 - A_2^2 z^{-1})}$$

$$= \frac{(\alpha_1 - \alpha_o \beta_1) + (\alpha_1 \beta_2 - \alpha_2 \beta_1) z^{-1}}{1 + (2\beta_2 - \beta_1^2) z^{-1} + \beta_2^2 z^{-2}} \qquad \text{(III.29)}$$

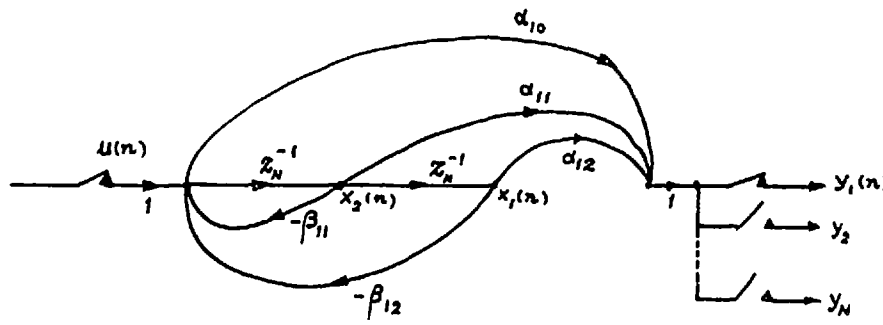
Equations (III.28) and (III.29) are the same as (III.8) and (III.9) if the multiplication coefficients of the digital filter remain unchanged throughout the sampling period.

The derivation of $H_i(z)$ using discrete convolution described in section III.3 is valid as long as $G(z_N)$ remains unchanged, i.e. a time-invariant filter. However, if the coefficients of the filter are allowed to taken on different values during the sampling period, the analysis becomes very complicated.

III.5 State-Space Analysis of a Multi-Rate Digital Filter

The following analysis makes use of the state-space method and is a modification of the method already published [58], [65]:

Consider the flow of a multi-rate digital filter with periodically varying coefficients realized in the direct form (fig III.5). Such a filter has N shift sequences during each sampling interval, while its coefficients are allowed to take on different values every T/N seconds as described before, so that α_{1j} and β_{1j} are the coefficients at nT, α_{2j} and β_{2j} are the coefficients at $(n + 1/N)T$, and α_{ij} and β_{ij} are the coefficients at $(n+\frac{i-1}{N})T$



$t = nT$

Fig III.5 (a)

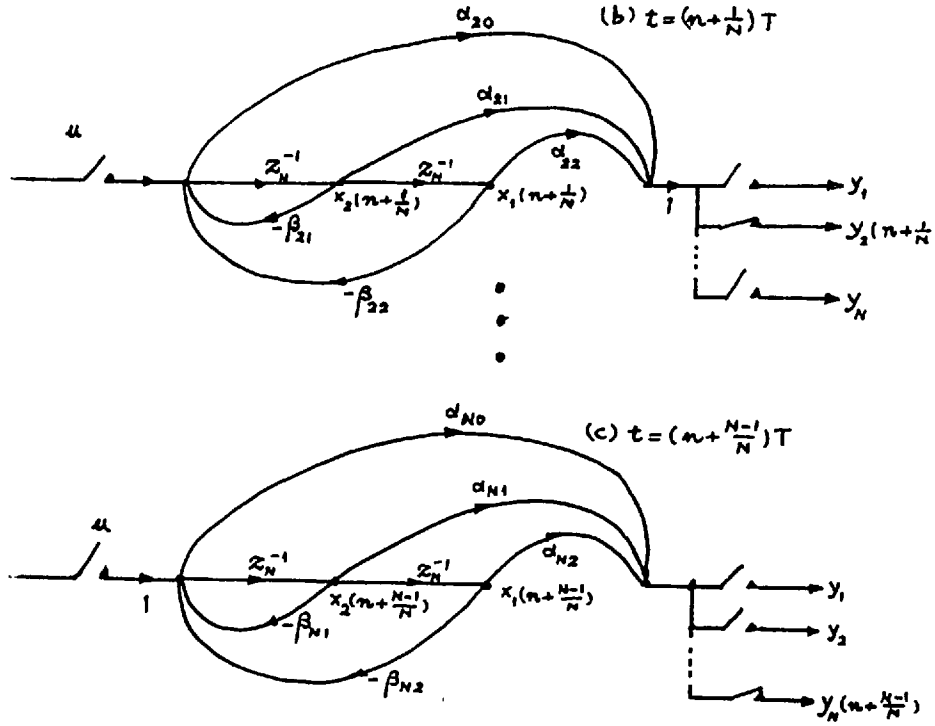


Fig III.5 Flow Graphs of a Second Order Multi-rate Digital Filter at Different Instants

Let x_1 , x_2 be the state variables at these different sampling instants as shown in the flow graphs.

At $t = nT$, the dynamic equations of the filter are

$$\begin{bmatrix} x_1(n+\frac{1}{N}) \\ x_2(n+\frac{1}{N}) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\beta_{12} & -\beta_{11} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(n)$$

$$= \mathbf{A}_1 \mathbf{x}(n) + \mathbf{B}_1 u(n) \qquad \text{(III.30)}$$

$$y_1(n) = \begin{bmatrix} (\alpha_{12} - \alpha_{10}\beta_{12}) & (\alpha_{11} - \alpha_{10}\beta_{11}) \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \alpha_{10} u(n)$$

$$= \mathbf{C}_1 \mathbf{x}(n) + \mathbf{D}_1 u(n) \qquad \text{(III.31)}$$

where $$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 \\ -\beta_{12} & -\beta_{11} \end{bmatrix}, \quad \mathbf{B}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\mathbf{C}_1 = \begin{bmatrix} (\alpha_{12}-\alpha_{10}\beta_{12}) & (\alpha_{11}-\alpha_{10}\beta_{11}) \end{bmatrix}, \mathbf{D}_1 = \begin{bmatrix} \alpha_{10} \end{bmatrix}$$

At $t = (n + 1/N)T$, the equations are

$$\mathbf{x}(n+\tfrac{2}{N}) = \mathbf{A}_2\,\mathbf{x}(n+\tfrac{1}{N}) + \mathbf{B}_1 u(n+\tfrac{1}{N}) = \mathbf{A}_2\mathbf{x}(n+\tfrac{1}{N}) \qquad \text{(III.32)}$$

$$y_2(n+\tfrac{1}{N}) = \mathbf{C}_2\,\mathbf{x}(n+\tfrac{1}{N}) + \mathbf{D}_2 u(n+\tfrac{1}{N}) = \mathbf{C}_2\mathbf{x}(n+\tfrac{1}{N}) \qquad \text{(III.33)}$$

since $u(n + 1/N) = 0$

At any subsequent instant $t = (n + \frac{i-1}{N})T$, where $i = 2, 3, \ldots, N$, the dynamic equations of the multirate digital filter are

$$\mathbf{x}(n+\tfrac{i}{N}) = \mathbf{A}_i\,\mathbf{x}(n+\tfrac{i-1}{N}) \qquad \text{(III.34)}$$

$$y_i(n+\tfrac{i-1}{N}) = \mathbf{C}_i\,\mathbf{x}(n+\tfrac{i-1}{N}) \qquad \text{(III.35)}$$

where $$\mathbf{A}_i = \begin{bmatrix} 0 & 1 \\ -\beta_{i2} & -\beta_{i1} \end{bmatrix}, \mathbf{C}_i = \begin{bmatrix} (\alpha_{i2}-\alpha_{io}\beta_{i2}) & (\alpha_{i1}-\alpha_{io}\beta_{i1}) \end{bmatrix}$$

and eventually, when $t = (n + \frac{N-1}{N})T$, the equations are:-

$$\mathbf{x}(n+1) = \mathbf{A}_N\,\mathbf{x}(n+\tfrac{N-1}{N}) \qquad \text{(III.36)}$$

$$y_N(n+\tfrac{N-1}{N}) = \mathbf{C}_N\,\mathbf{x}(n+\tfrac{N-1}{N}) \qquad \text{(III.37)}$$

Equations (III.30), (III.32), (III.34) and (III.36) express the relations between the state variables at these instants within the sampling period. Eliminating the intermediate state variable vectors $\mathbf{x}(n + \frac{i-1}{N})$ where $i = 2, 3, \ldots, N$, then,

$$\mathbf{x}(n+1) = \mathbf{A}_N \mathbf{A}_{N-1} \ldots \mathbf{A}_2 \left(\mathbf{A}_1 \mathbf{x}(n) + \mathbf{B}_1 u(n)\right) \quad \text{(III.38)}$$

Also, since the state equation at $t = (n + \frac{i-2}{N})T$ can be written as

$$\mathbf{x}(n+\tfrac{i-1}{N}) = \mathbf{A}_{i-1} \mathbf{x}(n-\tfrac{i-2}{N})$$

$$= \mathbf{A}_{i-1} \mathbf{A}_{i-2} \ldots \mathbf{A}_2 \left(\mathbf{A}_1 \mathbf{x}(n) + \mathbf{B}_1 u(n)\right) \quad \text{(III.39)}$$

hence, the ith output equation is, from eqn (III.35)

$$y_i(n+\tfrac{i-1}{N}) = \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \ldots \mathbf{A}_2 \left(\mathbf{A}_1 \mathbf{x}(n) + \mathbf{B}_1 u(n)\right) + \mathbf{D}_i u(n) \quad \text{(III.40)}$$

where $$\mathbf{D}_i = \begin{cases} [\alpha_{10}] & \text{for } i = 1 \\ [0] & \text{for } 1 < i \leq N \end{cases}$$

Equations (III.38) and (III.40) represent the general dynamic equations of the system. If the coefficients of the filter remains unchanged throughout the sampling period such that

$$\mathbf{A}_1 = \mathbf{A}_2 = \cdots = \mathbf{A}_N = \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\beta_2 & -\beta_1 \end{bmatrix}$$

and $$\mathbf{C}_1 = \mathbf{C}_2 = \cdots = \mathbf{C}_N = \mathbf{C} = \begin{bmatrix} (\alpha_2 - \alpha_0\beta_2) & (\alpha_1 - \alpha_0\beta_1) \end{bmatrix}$$

then equations (III.38) and (III.40) become

$$\mathbf{x}(n+1) = \mathbf{A}^N \mathbf{x}(n) + \mathbf{A}^{N-1} \mathbf{B}_1 u(n) \quad \text{(III.41)}$$

$$y_i(n+\tfrac{i-1}{N}) = \mathbf{C} \mathbf{A}^{i-1} \mathbf{x}(n) + \left(\mathbf{C} \mathbf{A}^{i-2} \mathbf{B}_1 + \mathbf{D}_i\right) u(n) \quad \text{(III.42)}$$

To find the transfer functions of the multi-rate digital filter from its dynamic equations, the z-transform of the equations are taken. Thus the ith transfer function of a multi-rate filter with periodically varying coefficients is, by taking the z-transform of eqns (III.38) and (III.40),

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_1)^{-1}(\mathbf{A}_N\mathbf{A}_{N-1}\cdots \mathbf{A}_2\mathbf{B}_1)\ U(z) \tag{III.43}$$

and

$$z^{\frac{i-1}{N}}\ Y_i(z) = \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \cdots \mathbf{A}_2 \mathbf{A}_1 \mathbf{X}(z) + \mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots\mathbf{A}_2\mathbf{B}_1 U(z) + \mathbf{D}_i U(z) \tag{III.44}$$

Thus

$$H_i(z) = \frac{z^{\frac{i-1}{N}} Y_i(z)}{U(z)} = (\mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots\mathbf{A}\ \mathbf{A}_1)(z\mathbf{I}-\mathbf{A}_N\mathbf{A}_{N-1}\cdots\mathbf{A}_1)^{-1}(\mathbf{A}_N\mathbf{A}_{N-1}\cdots\mathbf{A}_2\mathbf{B}_1) + \mathbf{C}_i\mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots\mathbf{A}_2\mathbf{B}_1 + \mathbf{D}_i \tag{III.45}$$

and for the time-invariant multi-rate filter, the ith transform function is simply,

$$H_i(z) = \frac{z^{\frac{i-1}{N}} Y_i(z)}{U(z)} = \mathbf{C}\mathbf{A}^{i-1}(z\mathbf{I}-\mathbf{A}^N)^{-1}\mathbf{A}^{N-1}\mathbf{B} + \mathbf{C}\mathbf{A}^{i-2}\mathbf{B} + \mathbf{D}_i \tag{III.46}$$

III.6 Remarks on the State-Space Analysis of Multi-rate Digital Filter

In eqn (III.46), there are two terms $\mathbf{A}^{i-1}$ and $\mathbf{A}^{i-2}$. These terms are interpreted in a way such that

$$\mathbf{A}^0 = \mathbf{I} = \text{identity matrix}$$

and negative powers of the state matrix $\mathbf{A}$ is taken to be a null matrix. Extending this interpretation to the case when the coefficients are allowed to take on different values during the sampling period, the corresponding terms in eqn (III.45) are defined as

$$\mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots\mathbf{A}_2\mathbf{A}_1 = \begin{cases} \mathbf{I} & \text{for } i = 1 \\ \mathbf{A}_1 & \text{for } i = 2 \\ \cdots\cdots\cdots & \\ \mathbf{A}_{N-1}\mathbf{A}_{N-2}\cdots\mathbf{A}_2\mathbf{A}_1 & \text{for } i = N \end{cases}$$

$$\mathbf{A}_{i-1}\,\mathbf{A}_{i-2}\cdots\mathbf{A}_2 = \begin{cases} \mathbf{0} & \text{for } i = 1 \\ \mathbf{I} & \text{for } i = 2 \\ \vdots & \\ \mathbf{A}_{N-1}\cdots\mathbf{A}_3\mathbf{A}_2 & \text{for } i = N \end{cases}$$

It has been pointed out in section III.3 that although the discrete convolution approach gives a general formula for time-invariant multi-rate digital filters, it fails when the coefficients of the filters are allowed to take on different values during the sampling period. The state-space analysis, in a way, generalizes the difference equation approach and thus gives a general formula for both the periodically varying and the time-invariant multi-rate filters. Besides this generalization, the use of state-space method can be further justified by the fact that it also gives the relationship between the input and output in both the frequency and time domain and thus saving the process of transforming from one domain to another.

III.7 Verification of the State-Space Derivation of the Transfer Functions by a Second Order Double-Rate Digital Filter

Equation (III.45) can easily be verified by a double-rate ($N = 2$) filter. Substituting $N = 2$ into eqn (III.45),

$$H_1(z) = \mathbf{C}_1(z\mathbf{I} - \mathbf{A}_2\mathbf{A}_1)^{-1}\mathbf{A}_2\mathbf{B}_1 + \mathbf{D}_1 \qquad \text{(III.47)}$$

and, $$H_2(z) = \mathbf{C}_2\,\mathbf{A}_1(z\,\mathbf{I} - \mathbf{A}_2\mathbf{A}_1)^{-1}\mathbf{A}_2\mathbf{B}_1 + \mathbf{C}_2\mathbf{B}_1 + \mathbf{D}_2 \qquad \text{(III.48)}$$

Substituting the values of $\mathbf{A}_1, \mathbf{A}_2, \mathbf{B}_1, \mathbf{C}_1, \mathbf{C}_2, \mathbf{D}_1, \mathbf{D}_2$, then

$$\mathbf{A}_2\mathbf{A}_1 = \begin{bmatrix} -\beta_{12} & -\beta_{11} \\ \beta_{12}\beta_{21} & (\beta_{11}\beta_{21} - \beta_{22}) \end{bmatrix}$$

thus,

$$(z\mathbf{I} - \mathbf{A}_2\mathbf{A}_1)^{-1}\,\mathbf{A}_2\mathbf{B}_1 = \begin{bmatrix} z+\beta_{22} \\ -\beta_{21}z \end{bmatrix} \Big/ \{ z^2 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z + \beta_{12}\beta_{22} \}$$

Hence, equation (III.47) becomes

$$H_1(z) = \begin{bmatrix} (\alpha_{12}-\alpha_{10}\beta_{12}) & (\alpha_{11}-\alpha_{10}\beta_{11}) \end{bmatrix} \begin{bmatrix} z+\beta_{22} \\ -\beta_{21}z \end{bmatrix} \Big/ \{ z^2 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z + \beta_{12}\beta_{22} \}$$

$$= \frac{\alpha_{10}z^2 + (\alpha_{10}\beta_{22} - \alpha_{11}\beta_{21} + \alpha_{12})z + \alpha_{12}\beta_{22}}{z^2 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z + \beta_{12}\beta_{22}} \qquad \text{(III.49)}$$

Similarly, $$\mathbf{C}_2\,\mathbf{A}_1 = \begin{bmatrix} (\alpha_{22}-\alpha_{20}\beta_{22}) & (\alpha_{21}-\alpha_{20}\,\beta_{21}) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\beta_{12} & -\beta_{11} \end{bmatrix}$$

$$= \begin{bmatrix} -\beta_{12}(\alpha_{21}-\alpha_{20}\beta_{21}) & (\alpha_{22}-\alpha_{20}\beta_{22}) - \beta_{11}(\alpha_{21}-\alpha_{20}\beta_{21}) \end{bmatrix}$$

and substituting into eqn (III.48),

$$H_2(z) = \frac{(\alpha_{21} - \alpha_{20}\beta_{21})z^2 + (\alpha_{21}\beta_{22} - \alpha_{22}\beta_{21})z}{z^2 + (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z + \beta_{12}\beta_{22}} \qquad \text{(III.50)}$$

Equations (III.49) and (III.50) can be seen to be identical to eqns (III.8) and (III.9)

III.8 Application of the State-Space Approach to Other Second Order Configuration

The state-space method of analysis can be applied to any other configurations. This is demonstrated by the following example:-

Ex. III.1 Analysis of the Transposed Direct Canonic Configuration

Fig III.6 shows the transposed configuration of the direct form [25]

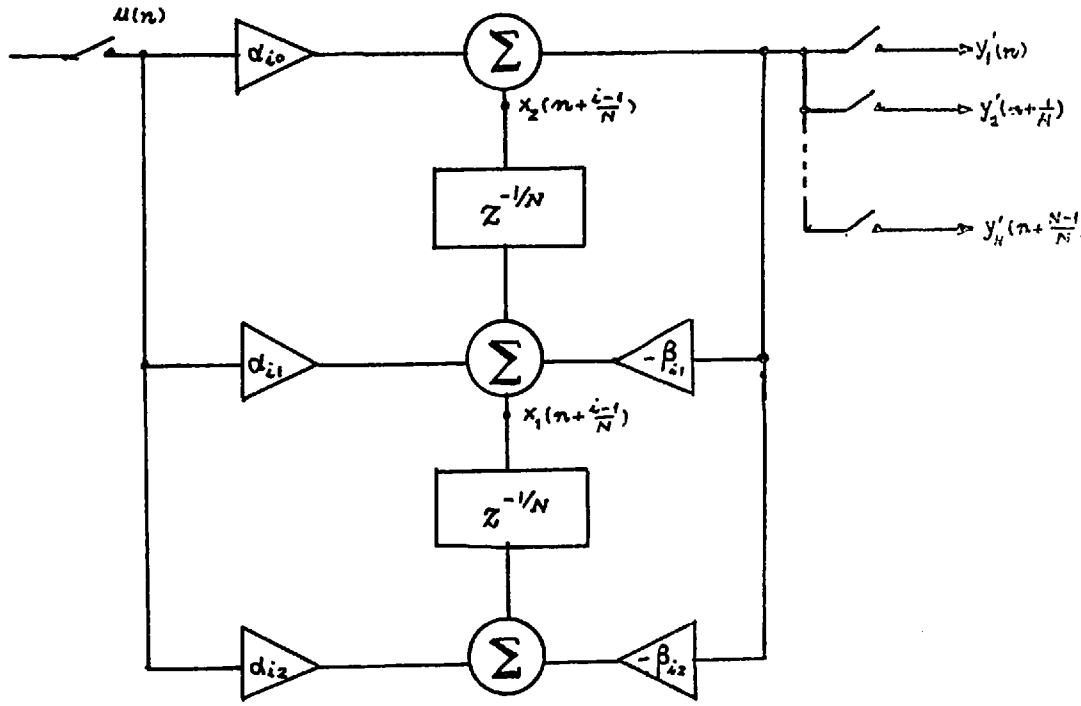


Fig III.6 A Multirate Digital Filter in the Transposed Direct Canonic Configuration

Choosing the state-variables x_1 and x_2 as indicated in fig III.6, and employing the same technique as shown in section III.5, one obtains the following equations:

$$\begin{bmatrix} x_1(n+\frac{1}{N}) \\ x_2(n+\frac{1}{N}) \end{bmatrix} = \begin{bmatrix} 0 & -\beta_{12} \\ 1 & -\beta_{11} \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \begin{bmatrix} (\alpha_{12}-\alpha_{10}\beta_{12}) \\ (\alpha_{11}-\alpha_{10}\beta_{11}) \end{bmatrix} u(n)$$

$$= \mathbf{A}_1^T \mathbf{x}(n) + \mathbf{C}_1^T u(n) \qquad \text{(III.51)}$$

$$\mathbf{x}(n+\frac{i}{N}) = \mathbf{A}_i^T \mathbf{x}(n+\frac{i-1}{N}) \qquad \text{(III.52)}$$

and $$\mathbf{x}(n+1) = \mathbf{A}_N^T \mathbf{x}(n+\frac{N-1}{N}) \qquad \text{(III.53)}$$

where $$\mathbf{A}_i^T = \begin{bmatrix} 0 & -\beta_{i2} \\ 1 & -\beta_{i1} \end{bmatrix} = \text{transposed of } \mathbf{A}_i$$

$$\mathbf{C}_1^T = \begin{bmatrix} (\alpha_{12} \; \alpha_{10}\beta_{12}) \\ (\alpha_{11} \; \alpha_{10}\beta_{11}) \end{bmatrix} = \text{transposed of } \mathbf{C}_1$$

$\mathbf{A}_i$ and $\mathbf{C}_i$ are both defined in section III.5. Solving equations (III.51) through (III.53) then,

$$\mathbf{x}(n+\frac{i}{N}) = \mathbf{A}_i^T \mathbf{A}_{i-1}^T \cdots \mathbf{A}_2^T \{\mathbf{A}_1^T \mathbf{x}(n) + \mathbf{C}_1^T u(n)\} \qquad \text{(III.54)}$$

and also $$\mathbf{x}(n+1) = \mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_2^T \{\mathbf{A}_1^T \mathbf{x}(n) + \mathbf{C}_1^T u(n)\} \qquad \text{(III.55)}$$

Again, employing the same technique as in section III.5, the output equations are:-

$$y_1^i(n) = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \alpha_{10} u(n) = \mathbf{B}_1^T \mathbf{x}(n) + \mathbf{D}_1^T u(n) \qquad \text{(III.56)}$$

$$\vdots$$

$$y_i^i(n+\frac{i-1}{N}) = \mathbf{B}_i^T \mathbf{x}(n+\frac{i-1}{N}) \qquad \text{(III.57)}$$

where $\mathbf{B}_i^T$ = (transposed of $\mathbf{B}_i$) = [0 ... 1] for $i = 1, 2, \ldots, N$

Substituting $\mathbf{x}(n + \frac{i-1}{N})$ from eqn (III.54) into eqn (III.57), one obtains

$$y_i'(n+\tfrac{i-1}{N}) = \mathbf{B}_i^T \mathbf{A}_{i-1}^T \mathbf{A}_{i-2}^T \cdots \mathbf{A}_2^T \left\{ \mathbf{A}_1^T \mathbf{x}(n) + \mathbf{C}_1^T u(n) \right\} + \mathbf{D}_i^T \qquad \text{(III.58)}$$

Taking the z-transform of eqn (III.55) and (III.58) and rearranging,

$$H_i'(z) = \frac{z^{\frac{i-1}{N}} Y_i'(z)}{U(z)} = \left(\mathbf{B}_i^T \mathbf{A}_{i-1}^T \mathbf{A}_{i-2}^T \cdots \mathbf{A}_2^T \mathbf{A}_1^T\right)\left(z\mathbf{I} - \mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_1^T\right)^{-1}\left(\mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_2^T \mathbf{C}_1^T\right) + \mathbf{B}_i^T \mathbf{A}_{i-1}^T \mathbf{A}_{i-2}^T \cdots \mathbf{A}_2^T \mathbf{C}_1^T + \mathbf{D}_i^T \qquad \text{(III.59)}$$

where $$\mathbf{D}_i^T = \begin{cases} [\alpha_{10}] & \text{for } i = 1 \\ [\,0\,] & \text{for } 1 < i \leq N \end{cases}$$

Equations (III.55) and (III.58) are the dynamic equations of the transposed direct canonic configuration used as a multirate filter. These equations are very similar to the dynamic equations, eqn (III.38) and (III.40), all matrices $\mathbf{A}_i$ are replaced by their corresponding transposed matrices $\mathbf{A}_i^T$, also $\mathbf{B}_1$ is replaced by $\mathbf{C}_1^T$ and $\mathbf{C}_i$ by $\mathbf{B}_i^T$, then the dynamic equations of the transposed direct realization is obtained.

It is observed that the transfer functions, $H_i(z)$, of the direct form and those, $H_i'(z)$, of the transposed form have the same poles, since

$$\det\left(z\mathbf{I} - \mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_1\right) \equiv \det\left(z\mathbf{I} - \mathbf{A}_N^T \mathbf{A}_{N-1}^T \cdots \mathbf{A}_1^T\right) \qquad \text{(III.60)}$$

However, except when $i = 1$, the numerators of $H_i(z)$ are not equivalent to the corresponding numerators in $H_i'(z)$ which are independent of $\mathbf{C}_i^T$ for $1 < i < N$; rather $H_i'(z)$ is dependent on

$\mathbf{B}_i^T$ which is a constant row vector for all values of i. Thus, for $1 < i \leq N$, $H_i'(z)$ of the transposed direct form is not equivalent to $H_i(z)$ of the direct canonic form. Nevertheless, in the case of a time-invariant filter where all $\mathbf{C}_i$ remain the same throughout the sampling period, $H_i'(z)$ of the transposed direct form will be equivalent to $H_i(z)$ of the direct form for all values of i.

There are many other second order configurations [25], but the procedure of obtaining their dynamic equations and transfer functions when such filters are used in a multirate fashion is the same. Also, from the point of view of quantization noise, the second order direct canonic form (or the transposed direct form) is the most commonly used [25]. Hence the other configurations will be omitted in the discussions.

III.9 Some Properties of the Transfer Function $H_i(z)$

In section III.5, it has been shown that the ith transfer function of a multi-rate digital filter in the direct canonic form is given by:-

$$H_i(z) = \left(\mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \cdots \mathbf{A}_1\right) \left(z\mathbf{I} - \mathbf{A}_m\right)^{-1} \left(\mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_2 \mathbf{B}_1\right) + \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \cdots \mathbf{A}_2 \mathbf{B}_1 + \mathbf{D}_i \qquad \text{(III.61)}$$

where for convenience, the matrix $\mathbf{A}_m$ is written instead of $\mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_1$

Here two interesting properties of this transfer function are shown in the form of the following theorems:-

a) Poles of $H_i(z)$

Lemma III.1 The poles of $H_i(z)$ are given by the equation

{15} {51}

$$\det(z\mathbf{I} - \mathbf{A}_m) = 0 \tag{III.62}$$

Proof:- The proof of the lemma follows directly from eqn (III.61) i.e.

$$H_i(z) = \left[(\mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots \mathbf{A}_2\mathbf{A}_1).\ \mathrm{adj}\,(z\mathbf{I} - \mathbf{A}_m)(\mathbf{A}_N\mathbf{A}_{N-1}\cdots\mathbf{A}_2\mathbf{B}_1)\right] \Big/ \det(z\mathbf{I} - \mathbf{A}_m) + \mathbf{C}_i\mathbf{A}_{i-1}\mathbf{A}_{i-2}\cdots\mathbf{A}_2\mathbf{B}_1 + \mathbf{D}_i$$

where adj(.) denotes the adjoint of a matrix.

Hence the poles of $H_i(z)$ are the roots of eqn (III.62)

Lemma III.2 The polynomial of the denominator of $H_i(z)$ is a quadratic function in z, i.e.

$$\det(z\mathbf{I} - \mathbf{A}_m) = z^2 + b_1 z + b_2 \tag{III.63}$$

such that

$$b_1 = -\,\mathrm{Tr}[\mathbf{A}_m] \tag{III.64}$$

and

$$b_2 = \det[\mathbf{A}_m] \tag{III.65}$$

where Tr(.) stands for the trace of a square matrix and det(.) denotes the determinant of a square matrix.

Proof:- Since

$$\mathbf{A}_m = \mathbf{A}_N\mathbf{A}_{N-1}\cdots\mathbf{A}_2\mathbf{A}_1$$

and all $\mathbf{A}_i$ are second order, then $\mathbf{A}_m$ must be of the second order. Thus, det $(z\mathbf{I} - \mathbf{A}_m)$ is a quadratic function in z.
Let the characteristic polynomial be represented by

$$f(z) = \det(z\mathbf{I} - \mathbf{A}_m) = z^2 + b_1 z + b_2 \tag{III.66}$$

then

$$f(0) = \det(-\mathbf{A}_m) = b_2$$

Hence $b_2 = \det[\mathbf{A}_m]$ since $\det[\mathbf{A}_m] = \det[-\mathbf{A}_m]$ for a second order matrix $\mathbf{A}_m$

Differentiating eqn (III.66) with respect to z at $z = 0$, then

$$\left.\frac{\partial f}{\partial z}\right|_{z=0} = b_1$$

Also
$$\left.\frac{\partial f}{\partial z}\right|_{z=0} = \left.\frac{\partial}{\partial z}[\det(z\mathbf{I} - \mathbf{A}_m)]\right|_{z=0}$$

$$= (z - a_{22} + z - a_{11})\Big|_{z=0}$$

$$= -\mathrm{Tr}(\mathbf{A}_m)$$

Hence $b_1 = -\mathrm{Tr}(\mathbf{A}_m)$ *QED*

Theorem III.1 If the polynomial of the denominator of $H_i(z)$ is expressed in the form of eqn (III.63), then

$$b_2 = \prod_{i=1}^{N} \beta_{i2} \qquad \text{(III.67)}$$

Proof:- From eqn (III.65) in lemma III.2,

$$b_2 = \det[\mathbf{A}_m] = \det[\mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_1] \qquad \text{(III.68)}$$

Now, since the multiplication rule for the nth order determinants of two square matrices $\mathbf{P}$ and $\mathbf{Q}$ are such that {15}

$$\det(\mathbf{P}).\det(\mathbf{Q}) = \det(\mathbf{R})$$

where $$r_{ij} = \sum_{k=1}^{n} p_{ik} q_{kj}$$

and p_{ij}, q_{ij} and r_{ij} being the elements of $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ respectively, then the array of the elements in det ($\mathbf{R}$) is thus

identical with that in the matrix product ($\mathbf{P}.\mathbf{Q}$). Hence, the determinant of the product of two matrices equals the product of their determinants.

Now, each matrix $\mathbf{A}_i$ is of the form

$$\mathbf{A}_i = \begin{bmatrix} 0 & 1 \\ -\beta_{i2} & -\beta_{i1} \end{bmatrix}$$

and $\det(\mathbf{A}_i) = \beta_{i2}$ (III.69)

therefore from eqn (III.68)

$$b_2 = \det(\mathbf{A}_N \mathbf{A}_{N-1} \ldots \mathbf{A}_1)$$

$$= \prod_{i=0}^{N-1} \det(\mathbf{A}_{N-i})$$

$$= \prod_{i=1}^{N} \beta_{i2} \qquad QED$$

b) Zeros of $H_i(z)$

Theorem III.2 For $1 < i \leq N$, $H_i(z)$ has a zero at the origin of the z-plane, i.e. the constant term in the numerator of $H_i(z)$ vanishes.

Proof:- Equations (III.61) can be written in the following form

$$H_i(z) = (\mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \ldots \mathbf{A}_2 \mathbf{A}_1) \frac{\mathrm{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} (\mathbf{A}_N \mathbf{A}_{N-1} \ldots \mathbf{A}_2 \mathbf{B}_1)$$

$$+ \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \ldots \mathbf{A}_2 \mathbf{B}_1 + \mathbf{D}_i \qquad \text{(III.70)}$$

where $\mathbf{A}_m = \mathbf{A}_N \mathbf{A}_{N-1} \ldots \mathbf{A}_1$ (III.71)

adj(.) denotes the adjoint of a matrix

and $\mathbf{D}_i = \mathbf{0}$

Grouping terms and rearranging, eqn (III.70) can be written as:-

$$H_i(z) = \mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2} \ldots \mathbf{A}_2 \left(\mathbf{A}_1 \frac{\text{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} \cdot \mathbf{A}_m \mathbf{A}_1^{-1} + \mathbf{I} \right) \mathbf{B}_1$$

$$= \mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2} \ldots \mathbf{A}_2\mathbf{A}_1 \left\{ \frac{\text{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} \cdot \mathbf{A}_m + \mathbf{I} \right\} \mathbf{A}_1^{-1}\mathbf{B}_1 \qquad \text{(III.72)}$$

Now, consider the term,

$$\Upsilon = \frac{\text{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} \cdot \mathbf{A}_m \qquad \text{(III.73)}$$

Recalling that

$$\det(z\mathbf{I} - \mathbf{A}_m).\mathbf{I} = \text{adj}(z\mathbf{I} - \mathbf{A}_m).(z\mathbf{I} - \mathbf{A}_m)$$

$$= \text{adj}(z\mathbf{I} - \mathbf{A}_m)z\mathbf{I} - \text{adj}(z\mathbf{I} - \mathbf{A}_m).\mathbf{A}_m \qquad \text{(III.74)}$$

Hence, $$\text{adj}(z\mathbf{I} - \mathbf{A}_m).\mathbf{A}_m = z\,\text{adj}(z\mathbf{I} - \mathbf{A}_m) - \det(z\mathbf{I} - \mathbf{A}_m).\mathbf{I} \qquad \text{(III.75)}$$

Thus substituting the value of $\{\text{adj}(z\mathbf{I} - \mathbf{A}_m).\mathbf{A}_m\}$ given by (III.75), the term in bracket in eqn (III.72) can be written as

$$\Upsilon + \mathbf{I} = \frac{z\,\text{adj}(z\mathbf{I} - \mathbf{A}_m) - \det(z\mathbf{I} - \mathbf{A}_m).}{\det(z\mathbf{I} - \mathbf{A}_m)} + \mathbf{I}$$

$$= \frac{z\,\text{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} \qquad \text{(III.76)}$$

Hence for $1 < i \leq N$, the ith transfer function of the multirate filter is given by:-

$$H_i(z) = (\mathbf{C}_i \mathbf{A}_{i-1}\mathbf{A}_{i-2} \ldots \mathbf{A}_2\mathbf{A}_1) \frac{z\,\text{adj}(z\mathbf{I} - \mathbf{A}_m)}{\det(z\mathbf{I} - \mathbf{A}_m)} \cdot \mathbf{A}_1^{-1}\mathbf{B}_1 \qquad \text{(III.77)}$$

Now, since every element of the matrix $\{z.\text{adj}(z\mathbf{I} - \mathbf{A}_m)\}$ contains terms involving z or z^2, but no constant term, it may be concluded

that, for $1 < i \leq N$, the numerators of $H_i(z)$ are always free of a constant term. *QED*

The above two theorems can be seen to be equally applicable to the transfer functions of a multirate digital filter realized in the transposed configuration.

III.10 Design of Second Order Multirate Filters

Consider the case that a second order transfer function given by

$$F(z) = \frac{a_0 z^2 + a_1 z + a_2}{z^2 + b_1 z + b_2} \qquad \text{(III.78)}$$

is to be realized by a second order multirate digital filter.

If $a_2 \neq 0$, then from the property shown in section III.8(b), eqn (III.78) can only be realized by using $H_1(z)$ of the multirate filter. However, if $a_2 = 0$, e.g. in a realization of higher order digital filters by parallel alignment of second order subfilters, then any of the $H_i(z)$ of the multirate filter can be used to realize $F(z)$. The choice $H_i(z)$ for such a design is discussed below. The two cases of time-invariant and periodically varying multirate filters are considered separately.

a) Time Invariant Multirate Digital Filter

Consider that $F(z)$ given by eqn(III.78) is to be realized by a second order time invariant multirate filter. If only complex poles are considered, then the poles of $F(z)$, Λ and Λ^* can be written as

$$\Lambda = re^{j\theta}$$
$$\Lambda^* = re^{-j\theta} \qquad \text{(III.79)}$$

It has been shown in section III.3 that if a second order time invariant filter working at single rate has poles given by λ and λ^*, then its poles when working with N shift sequences within a sampling period are λ^N and λ^{*N}. Thus to realize $F(z)$ by a time-invariant filter,

$$\lambda^N = re^{j\theta} = \Lambda \tag{III.80}$$

$$\lambda^{*N} = re^{-j\theta} = \Lambda \tag{III.81}$$

Let $$\lambda = \rho e^{j\phi} \tag{III.82}$$

$$\lambda^* = \rho e^{-j\phi} \tag{III.83}$$

then, $$\rho = r^{1/N} \tag{III.84}$$

$$\phi = \frac{\theta + 2n\pi}{N} \tag{III.85}$$

Hence for given values of r and θ, there are in general N solutions for λ, i.e.

$$\lambda = r^{1/N} e^{j(\theta+2i\pi)/N} \tag{III.86}$$

where $i = 0, 1, 2, \ldots, (N-1)$

Hence for a given transfer function $F(z)$, there are N different ways of realizing the poles of $F(z)$ using a multirate digital filter.

It can be shown [68] (see also section V.6) that the sensitivity of the resulting poles, Λ and Λ^*, from using a multirate filter is inversely proportional to $|\sin\phi|$, i.e.

$$|\delta\Lambda| \propto \frac{1}{|\sin\phi|} \tag{III.87}$$

Hence in designing $F(z)$ with a multirate filter, it is better, as far as pole sensitivity is concerned, to choose the pair of λ and λ^* having the largest value of $|\sin\phi|$. The following example will illustrate this point.

Example III.2

It has been decided to use a time-invariant triple-rate ($N = 3$) digital filter to realize a transfer function given by

$$F(z) = \frac{z^2}{z^2 - 0.9z + 0.81}$$

$$= \frac{z^2}{\{z - (0.45+j0.45\sqrt{3})\}\{z - (0.45-j0.45\sqrt{3})\}} \qquad \text{(III.88)}$$

Hence, $\Lambda = re^{j\theta}$ (III.89)

where $r = 0.9$ (III.90)

$\theta = \pi/3$ (III.91)

There are three choices of λ,

$$\lambda_1 = \rho e^{j\phi_1}$$
$$\lambda_2 = \rho e^{j\phi_2}$$
$$\lambda_3 = \rho e^{j\phi_3}$$

where $\rho = r^{1/N} = (0.9)^{1/3} = 0.9655$ (III.92)

$$\left.\begin{aligned} \phi_1 &= \tfrac{\pi}{3}/3 = 20 \text{ deg.} \\ \phi_2 &= (\tfrac{\pi}{3} + 2\pi)/3 = 140 \text{ deg} \\ \phi_3 &= (\tfrac{\pi}{3} + 4\pi)/3 = 260 \text{ deg} \end{aligned}\right\} \qquad \text{(III.93)}$$

These values of λ_i are shown on the z_N-plane in fig III.7 while their respective complex conjugates are marked with an asterik. The corresponding positions of Λ and Λ^* on the z-plane are also shown.

Of these three available choices, $|\sin \phi_3|$ has the largest value and therefore λ_3 and λ_3^* are chosen so that the resulting transfer function will have the least pole sensitivity.

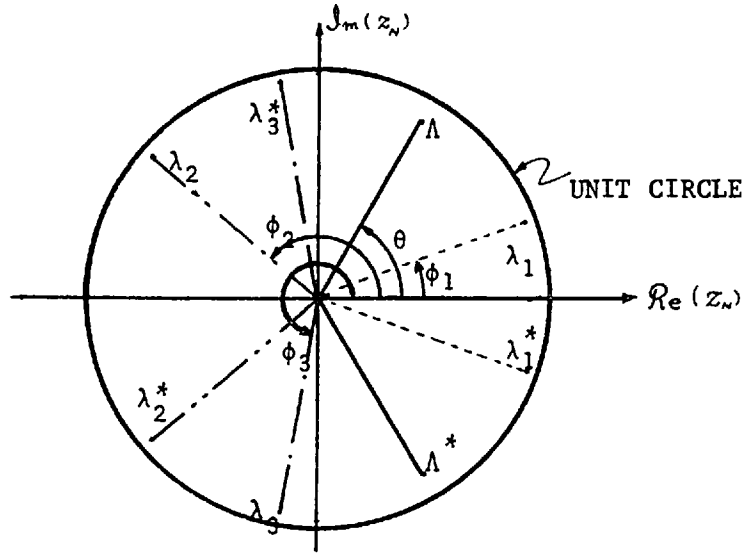


Fig III.7

Positions of the 3 Possible Pole Choices in the z_N-Plane

Choosing λ_3 and λ_3^* for the poles of the multirate filter, the feedback multipliers of the multirate filters are given by:-

$$\beta_1 = -(\lambda_3 + \lambda_3^*) = -(0.9)^{1/3}\cos 260^\circ = 0.1677 \qquad \text{(III.94)}$$

$$\beta_2 = \lambda_3\lambda_3^* = (0.9)^{2/3} = 0.9322 \qquad \text{(III.95)}$$

After determining the values of β_j, the values of α_j are considered. Since from eqns (III.10), (III.11) and (III.12), a triple-rate time-invariant filter has three different transfer functions all having the same poles. Thus we can choose any of the three to realize eqn (III.88). Rewriting eqns (III.10)

through (III.12) for the case of time-invariant filters, we have

$$H_1(z) = \left[\frac{\alpha_0 + \{\alpha_1(\beta_1^2 - \beta_2) - \alpha_0\beta_1\beta_2 - \alpha_2\beta_1\}z^{-1} + \alpha_2\beta_2^2 z^{-2}}{1 + (\beta_1^3 - 3\beta_1\beta_2)z^{-1} + \beta_2^3 z^{-2}}\right] \quad \text{(III.96)}$$

$$H_2(z) = \left[\frac{(\alpha_1 - \alpha_0\beta_1) + \{\alpha_0\beta_2^2 - \alpha_1\beta_1\beta_2 + \alpha_2(\beta_1^2 - \beta_2)\}z^{-1}}{1 + (\beta_1^3 - 3\beta_1\beta_2)z^{-1} + \beta_2^3 z^{-2}}\right] \quad \text{(III.97)}$$

$$H_3(z) = \left[\frac{\alpha_0(\beta_1^2 - \beta_2) - \alpha_1\beta_1 + \alpha_2 + (\alpha_1\beta_2^2 - \alpha_2\beta_1\beta_2)z^{-1}}{1 + (\beta_1^3 - 3\beta_1\beta_2)z^{-1} + \beta_2^3 z^{-2}}\right] \quad \text{(III.98)}$$

But considering that the numerator of $H(z)$ in eqn (III.88) has only one term z^2, and reading through eqns (III.96), (III.97) and (III.98), we find that $H_3(z)$ is the only transfer function we can use to realize $H(z)$ such that $\alpha_1 = 0$ and $\alpha_2 = 0$. The reason for choosing α_1 and α_2 to be zero is that we can save two multipliers in the implementation of the filter. For α_1 and α_2 to be zero, we have

$$\alpha_o(\beta_1^2 - \beta_2) = 1$$

i.e. $$\alpha_o = 1/(\beta_1^2 - \beta_2) = -1.1061 \quad \text{(III.99)}$$

With all values of the multipliers calculated, the final design of the triple-rate digital filter to realize eqn (III.90) is shown in fig III.8

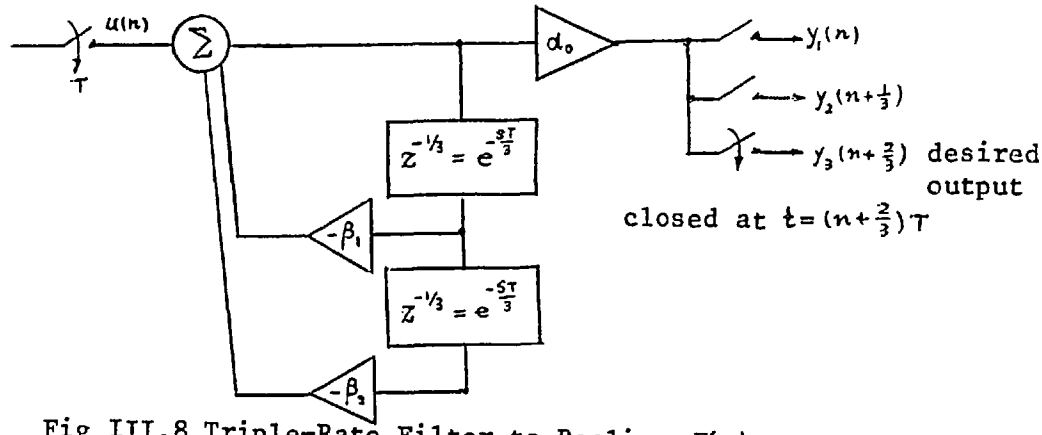


Fig III.8 Triple-Rate Filter to Realize $H(z)$

The other two transfer functions $H_1(z)$ and $H_2(z)$ also exist but they will have characteristics different from $H_3(z)$ and different from each other. It is not the concern here in this example to make use of them.

b) Multirate Digital Filters with Periodically Varying Coefficients

Consider the denominator of $F(z)$ given in eqn (III.78). It has been shown in section III.8 (a) that

$$b_1 = -\mathrm{Tr}[\mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_2 \mathbf{A}_1] \tag{III.100}$$

$$b_2 = \prod_{i=1}^{N} \det[\mathbf{A}_i] = \prod_{i=1}^{N} \beta_{i2} \tag{III.101}$$

i.e. to realize the denominator of $F(z)$ using a multirate filter, there are two equations but $2N$ unknowns. Hence there are $2(N-1)$ degrees of freedom in choosing β_{i1} and β_{i2}. However if the choice of β_{i1} and β_{i2} are stipulated so that β_{i1} and β_{i2} have to be chosen together as a set, then there are $(N-1)$ free choices of the β_{i1}— β_{i2} sets. Therefore, given b_1 and b_2, $(N-1)$ sets of β_{i1} and β_{i2} can be chosen and the remaining β_{i1} and β_{i2} set can be determined from equations (III.100) and (III.101).

Consider the numerator of $F(z)$. For a direct canonic realization of the multirate filters, after all the values of β_{ij} have been determined, the numerator of $H_i(z)$ is a function of $\mathbf{C}_i$ and $\mathbf{D}_i$ only, where $\mathbf{C}_i$ and $\mathbf{D}_i$ are given by

$$\mathbf{C}_i = [(\alpha_{i2} - \alpha_{i0}\beta_{i2}) \quad (\alpha_{i1} - \alpha_{i0}\beta_{i1})]$$

$$\mathbf{D}_i = \begin{cases} \alpha_{10} & \text{for } i = 1 \\ 0 & \text{for } 1 < i \leq N \end{cases} \tag{III.102}$$

From the properties of $H_i(z)$ shown in section III.8(b) that the numerator polynomial of $H_i(z)$ contains a constant term only when $i = 1$, then for $H_1(z)$, there are three linear equations and three unknowns, i.e.

$$\begin{aligned} a_0 &= \phi_0(\alpha_{10}, \alpha_{11}, \alpha_{12}) \\ a_1 &= \phi_1(\alpha_{10}, \alpha_{11}, \alpha_{12}) \\ a_2 &= \phi_2(\alpha_{10}, \alpha_{11}, \alpha_{12}) \end{aligned} \tag{III.103}$$

where ϕ_0, ϕ_1 and ϕ_2 are linear functions. Thus α_{10}, α_{11} and α_{12} can be determined.

However, for $1 < i \leq N$, since the constant term in the numerator of $H_i(z)$ vanishes, there are only two equations but three unknowns i.e.

$$\begin{aligned} a_0 &= f_0(\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \\ a_1 &= f_1(\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \end{aligned} \tag{III.104}$$

where f_0 and f_1 are linear functions. Hence we have one degree of freedom in the choice of α_{i0}, α_{i1} and α_{i2}, i.e. we can choose any value for one of the coefficients and solve (III.104) for the other two. A convenient value to choose is zero, in which case, one of the multipliers can be omitted. A further saving of one multiplier can be achieved if the $(N - 1)$ degree of freedom in the choice of the denominator coefficients and the freedom in choosing one of the numerator coefficients are both utilized. This is especially a useful way of reducing the number of multiplications in a sampling period if a multirate digital filter is used to realize a second order filter with a zero at the origin (e.g. in the realization of higher order filters by parallel alignment of second order subfilters). The following example may help to illustrate this point.

Example III.3

If a second order transfer function of the form

$$F(z) = \frac{a_o + a_1 z^{-1}}{1 + b_1 z^{-1} + b_2 z^{-2}} \qquad \text{(III.105)}$$

is to be realized by a single-rate digital filter, it needs four multipliers, i.e. the circuit diagram would be as shown in fig III.9

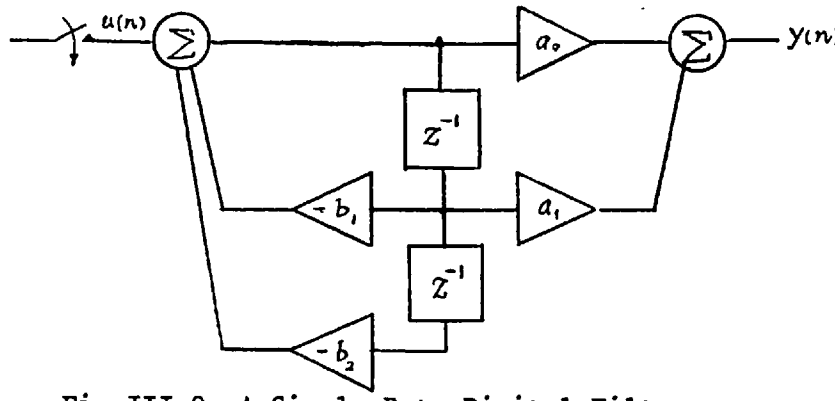


Fig III.9 A Single-Rate Digital Filter

It is shown here how, by using a time-varying double-rate filter, the number of multipliers can be reduced and at the same time saving the use of one adder.

From section III.6, it is known that there are two transfer functions associated with a double-rate filter, viz.

$$H_1(z) = \frac{\left[\alpha_{10} + (\alpha_{10}\beta_{22} - \alpha_{11}\beta_{21} + \alpha_{12})z^{-1} + \alpha_{12}\beta_{22}z^{-2}\right]}{\left[1 + (\beta_{12} + \beta_{22} - \beta_{11}\beta_{21})z^{-1} + \beta_{12}\beta_{22}z^{-2}\right]} \qquad \text{(III.106)}$$

$$H_2(z) = \frac{\left[(\alpha_{21} - \alpha_{20}\beta_{21}) + (\alpha_{21}\beta_{22} - \alpha_{22}\beta_{21})z^{-1}\right]}{\left[1 + (\beta_{12} + \beta_{22} - \beta_{11}\beta_{21})z^{-1} + \beta_{12}\beta_{22}z^{-2}\right]} \qquad \text{(III.107)}$$

Either of these two transfer functions can be used to realize eqn (III.105).

First consider the case when $H_1(z)$ of eqn (III.106) is used:- Comparing the coefficients of the numerators of $H_1(z)$ and $F(z)$, it is evident that the following equations have to be satisfied,

$$\begin{aligned} \alpha_{10} &= a_o \\ \alpha_{12} &= 0 \end{aligned} \qquad \text{(III.108)}$$

Now if α_{11} is made equal to zero as well then

$$\begin{aligned} \alpha_{10}\beta_{22} &= a_1 \\ \text{i.e.} \qquad \beta_{22} &= a_1/a_o \end{aligned} \qquad \text{(III.109)}$$

Comparing the coefficients of the denominators of $H_1(z)$ and $F(z)$, then

$$\beta_{12} = \frac{b_2}{\beta_{22}} = \frac{a_o b_2}{a_1} \qquad \text{(III.110)}$$

and if a convenient value for β_{11} is chosen, say $\beta_{11} = 1$, then

$$\begin{aligned} \beta_{21} &= (\beta_{12} + \beta_{22} - b_1)/\beta_{11} \\ &= \left(\frac{a_0 b_1}{a_1} + \frac{a_1}{a_0} - b_1 \right) / \beta_{11} \end{aligned} \qquad \text{(III.111)}$$

Hence the equivalent double rate filter has the circuit diagram as shown in fig III.10

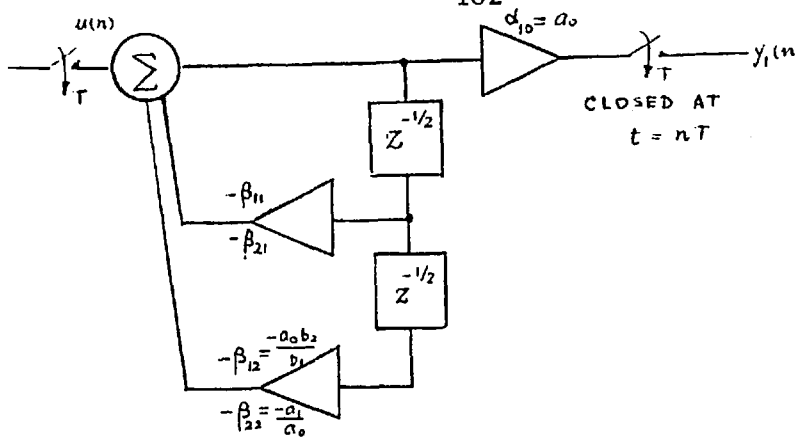


Fig III.10 A Double-Rate Digital Filter used in Place of fig III.9

On the other hand, if eqn (III.107) is used to realize $F(z)$ given by eqn (III.105) then again

$$\beta_{12}\beta_{22} = b_2 \qquad \text{(III.112)}$$

$$\beta_{12} + \beta_{22} - \beta_{11}\beta_{21} = b_1 \qquad \text{(III.113)}$$

Now setting $\alpha_{20} = \alpha_{22} = 0$

then

$$\alpha_{21} = a_o \qquad \text{(III.114)}$$

$$\alpha_{21}\,\beta_{22} = a_1 \qquad \text{(III.115)}$$

Thus it can be chosen that

$$\beta_{22} = a_1/\alpha_{21} = a_1/a_o$$

giving

$$\beta_{12} = \frac{b_2 a_0}{a_1}$$

Again choosing a convenient value for β_{11}, β_{21} can be determined from eqn (III.113). Hence the equivalent double-rate filter, if $H_2(z)$ is used to realize $F(z)$, has the following circuit diagram (fig III.11)

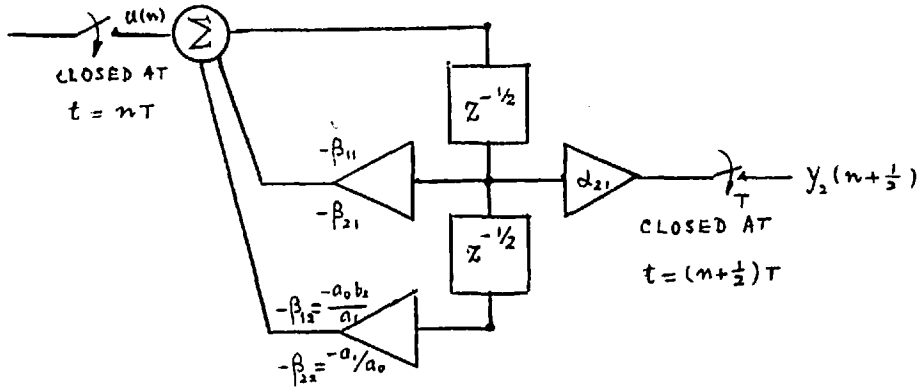


Fig III.11 An Alternative Double-Rate Digital Filter used in Place of Fig III.9

Comparing Fig III.9, III.10 and III.11, it can be seen that by making use of the fact that the zeros of a double-rate filter are related to its feedback multipliers, the number of multiplying coefficients can be reduced to four if the filter is to realize its single-rate counter part having a zero at the original.

The use of multirate digital filters to realize equivalent single rate transfer functions necessitates more multiplications in a sampling period and thus faster multiplication rates. Hence, any possibility of reducung the multiplication rate as has been demonstrated by the above example would be most welcome.

III.11 Computer Simulation Results

In order to verify that the results obtained from the state-space analysis of a multirate digital filter (section III.5) are correct, a computer program has been written. The program first designs a multi-

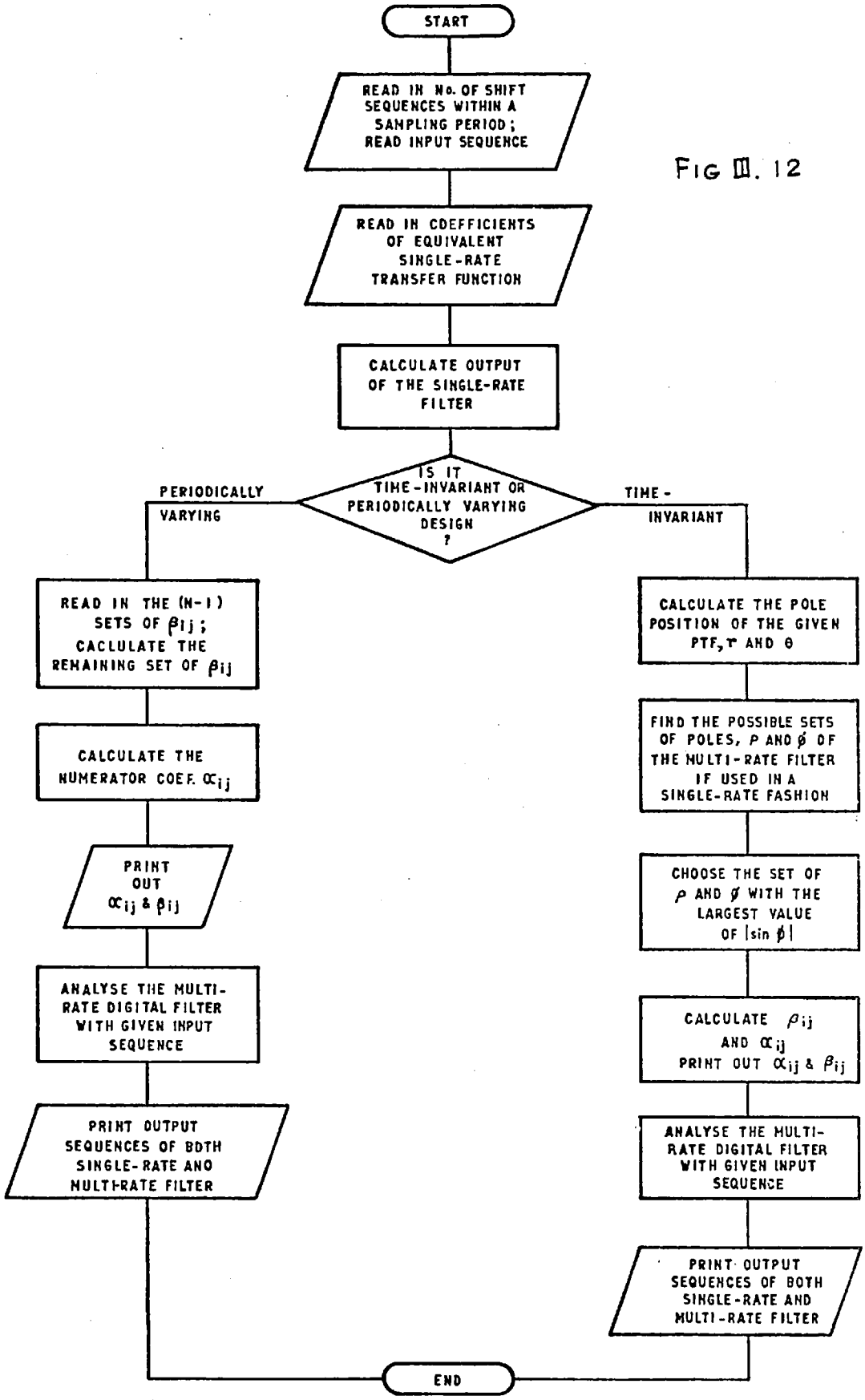

START
READ IN No. OF SHIFT SEQUENCES WITHIN A SAMPLING PERIOD; READ INPUT SEQUENCE
READ IN COEFFICIENTS OF EQUIVALENT SINGLE-RATE TRANSFER FUNCTION
CALCULATE OUTPUT OF THE SINGLE-RATE FILTER
IS IT TIME-INVARIANT OR PERIODICALLY VARYING DESIGN ?
PERIODICALLY VARYING
TIME-INVARIANT
READ IN THE (N-1) SETS OF β_{ij}; CACLULATE THE REMAINING SET OF β_{ij}
CALCULATE THE NUMERATOR COEF. α_{ij}
PRINT OUT α_{ij} & β_{ij}
ANALYSE THE MULTI-RATE DIGITAL FILTER WITH GIVEN INPUT SEQUENCE
PRINT OUTPUT SEQUENCES OF BOTH SINGLE-RATE AND MULTI-RATE FILTER
CALCULATE THE POLE POSITION OF THE GIVEN PTF, τ AND θ
FIND THE POSSIBLE SETS OF POLES, ρ AND ϕ OF THE MULTI-RATE FILTER IF USED IN A SINGLE-RATE FASHION
CHOOSE THE SET OF ρ AND ϕ WITH THE LARGEST VALUE OF $|\sin \phi|$
CALCULATE β_{ij} AND α_{ij} PRINT OUT α_{ij} & β_{ij}
ANALYSE THE MULTI-RATE DIGITAL FILTER WITH GIVEN INPUT SEQUENCE
PRINT OUTPUT SEQUENCES OF BOTH SINGLE-RATE AND MULTI-RATE FILTER
END

FIG III. 12

rate second order digital filter in the same way as described in section III.10 so that it gives the same performance as a given single-rate second order filter. Then from a given input sequence, the program evaluates and prints out the time domain output sequences of such a multirate digital filter. The time domain output sequence of the equivalent single rate digital filter is also calculated, thus enabling a comparison between the two sequences.

A flow chart of the program is shown in fig III.12.

The following examples (ExIII.4 through III.7) are taken from the computer simulation program. In each example, a single rate digital filter and its impulse response is first shown in part (a). Then a multirate digital filter is designed in the same way as described in section III.10 so that the performance of its output sequence is identical to that of the given single rate filter. The circuit diagram of the multirate digital filter as designed and its ith impulse response are shown in part (b) of each example. The designed values of the multiplying coefficients are shown on the circuit diagram.

These examples verify that the formulae developed in section III.5 are correct and that the output of a single rate digital filter is identical to one of the outputs of a multirate filter if the coefficients of the multirate filter are designed according to the analysis results.

Example III.4 a)

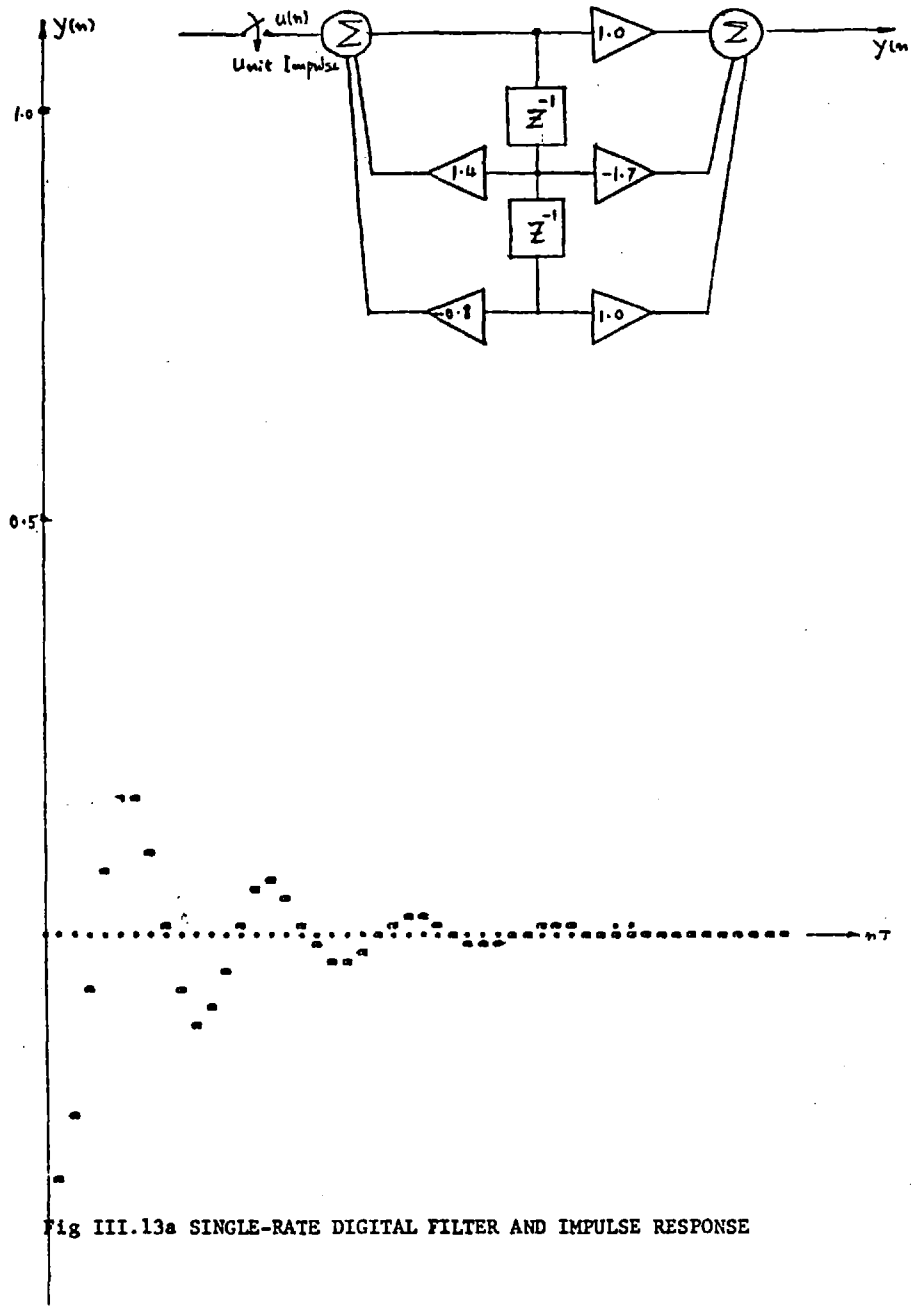


Fig III.13a SINGLE-RATE DIGITAL FILTER AND IMPULSE RESPONSE

Example III.4b)

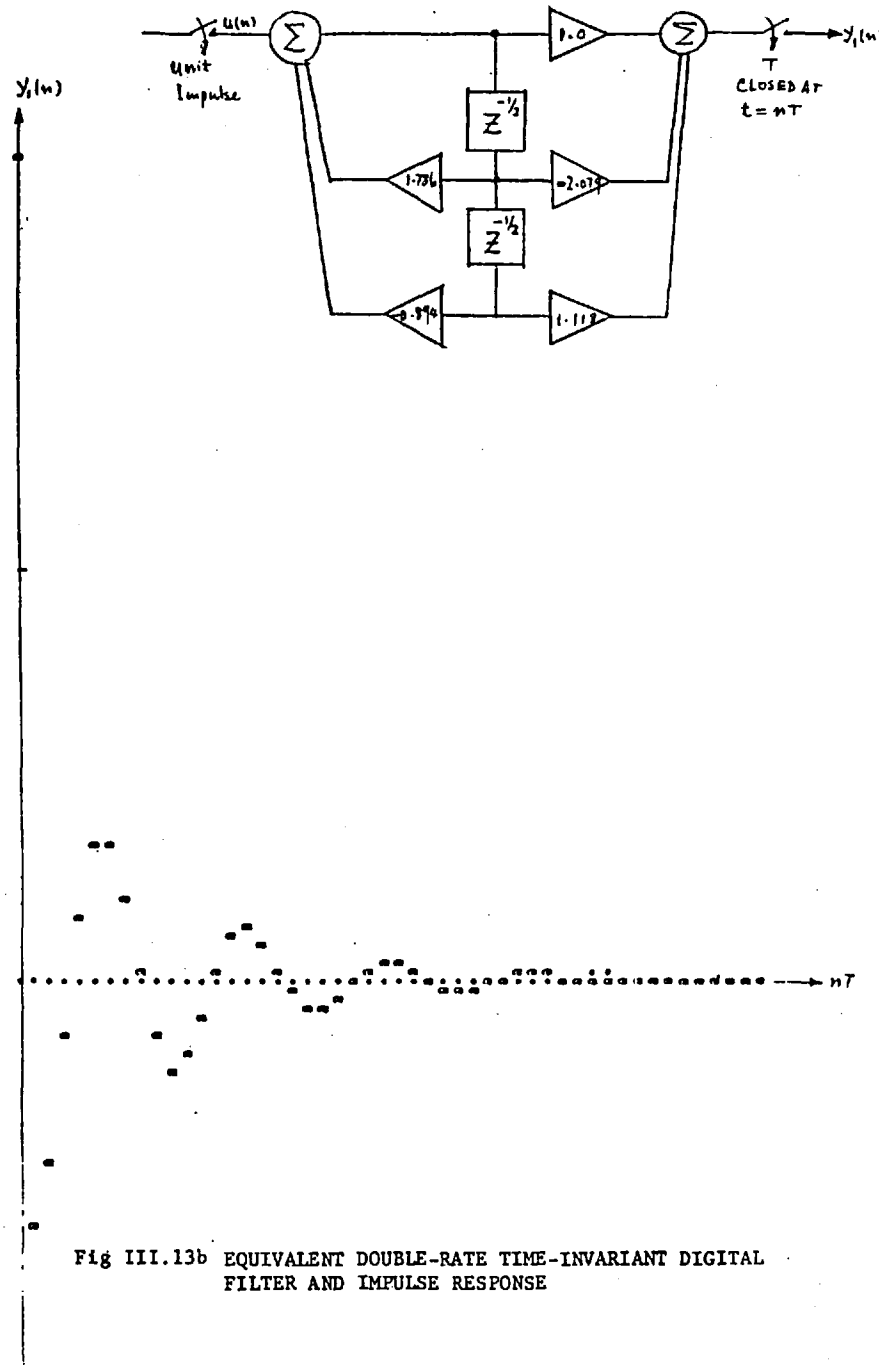


Fig III.13b EQUIVALENT DOUBLE-RATE TIME-INVARIANT DIGITAL FILTER AND IMPULSE RESPONSE

Example III.5 a)

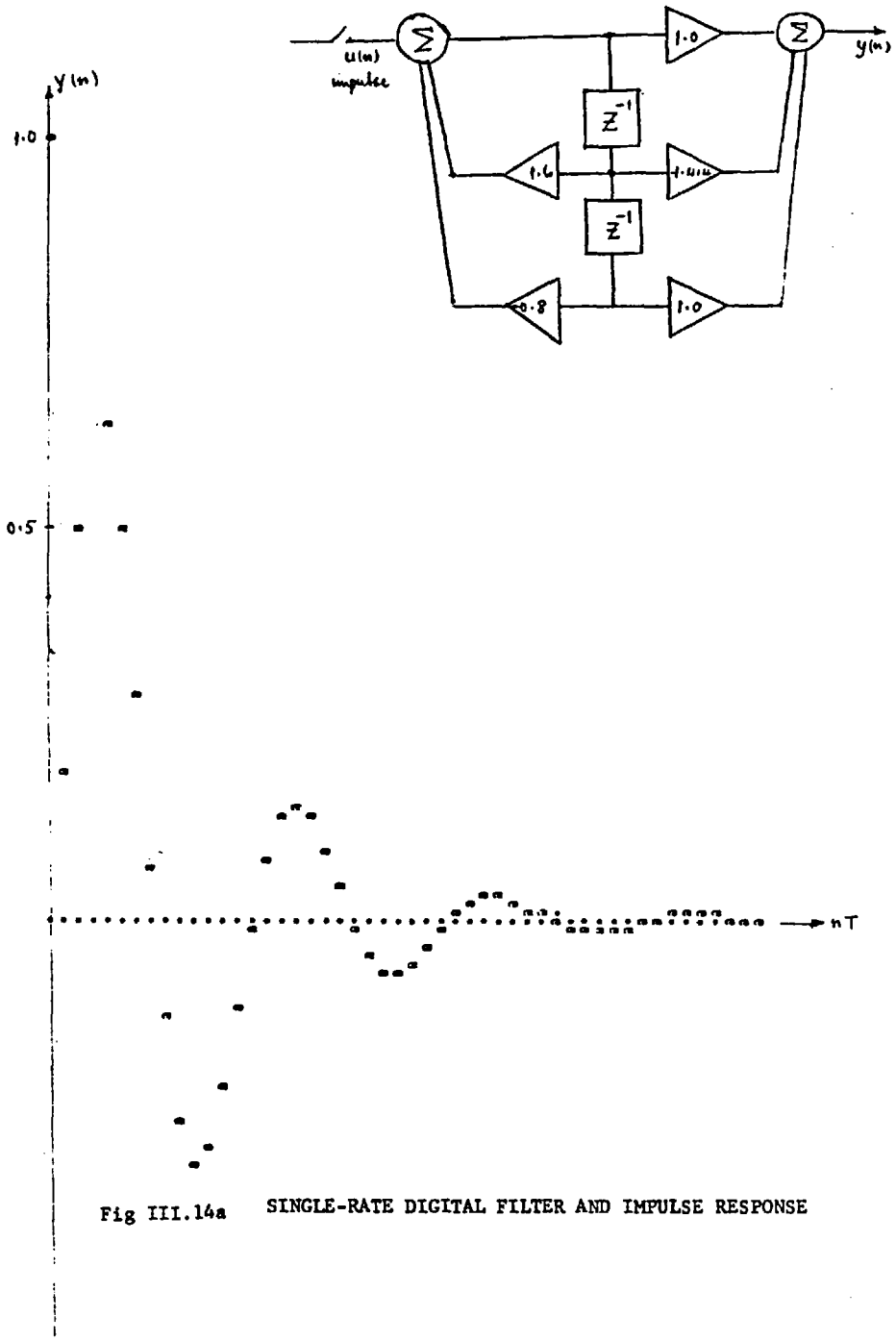


Fig III.14a SINGLE-RATE DIGITAL FILTER AND IMPULSE RESPONSE

Example III.5 b)

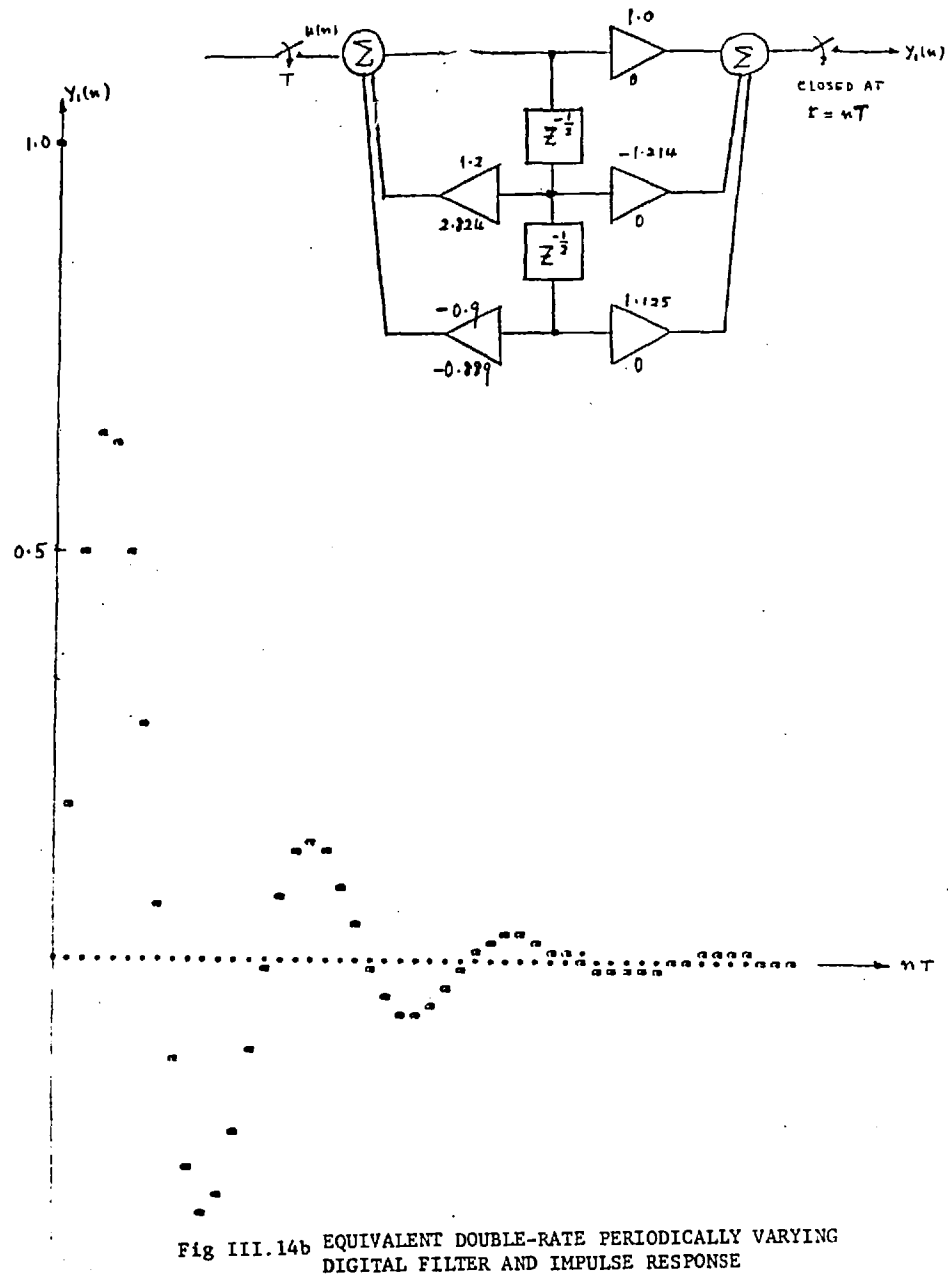


Fig III.14b EQUIVALENT DOUBLE-RATE PERIODICALLY VARYING DIGITAL FILTER AND IMPULSE RESPONSE

Example III.6 a)

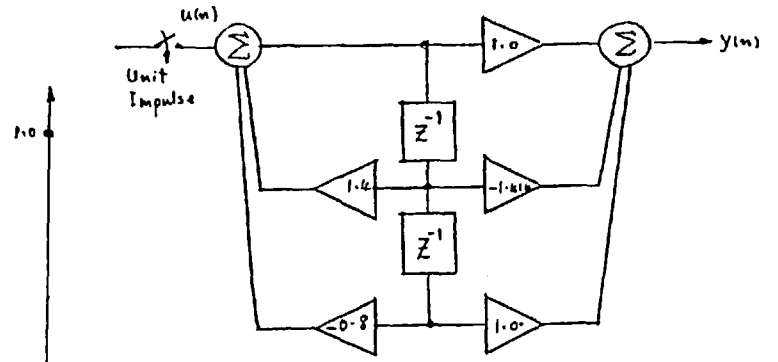


Fig III.15a SINGLE-RATE DIGITAL FILTER AND IMPULSE RESPONSE

Example III. 6 b)

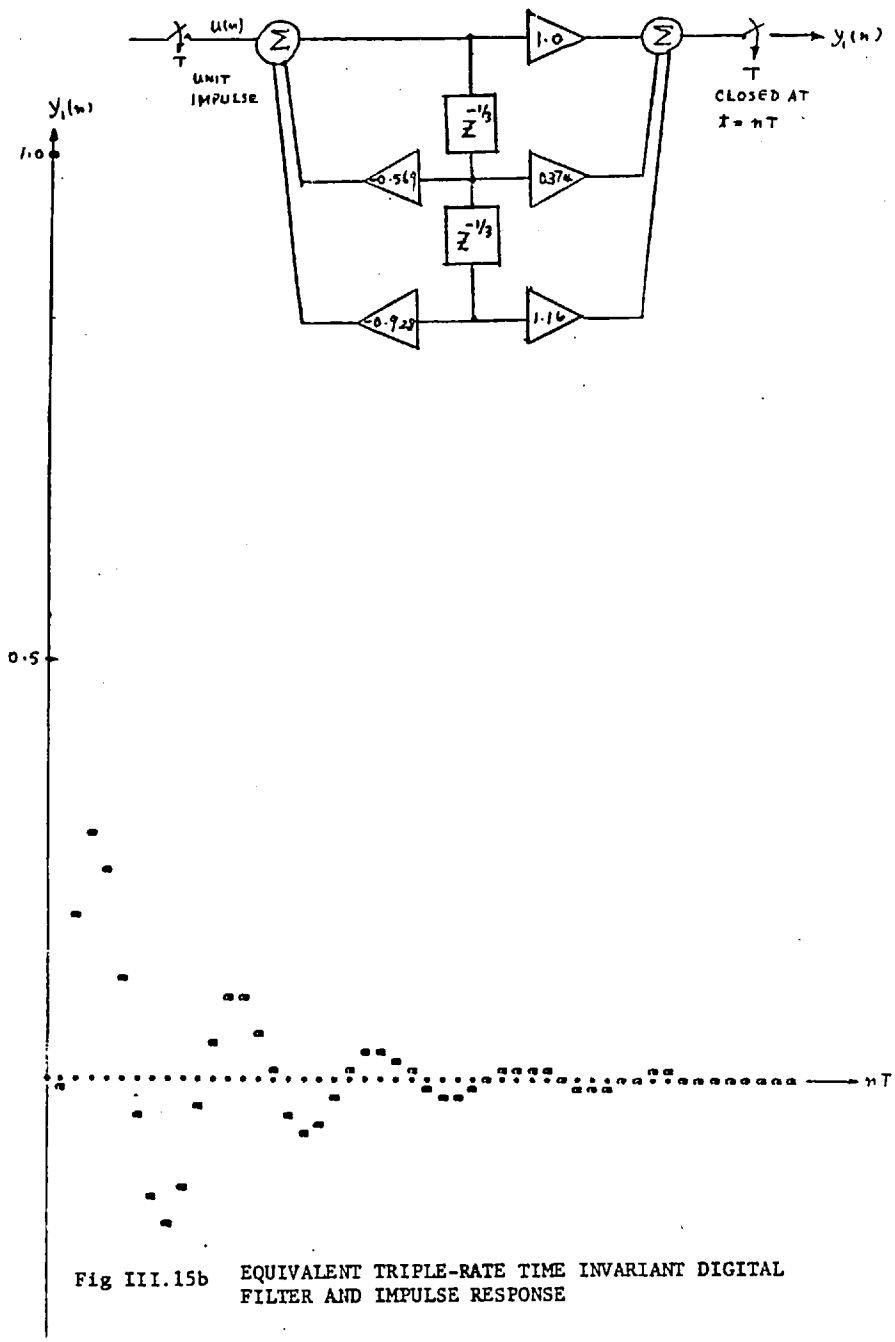


Fig III.15b EQUIVALENT TRIPLE-RATE TIME INVARIANT DIGITAL FILTER AND IMPULSE RESPONSE

Example III.7a)

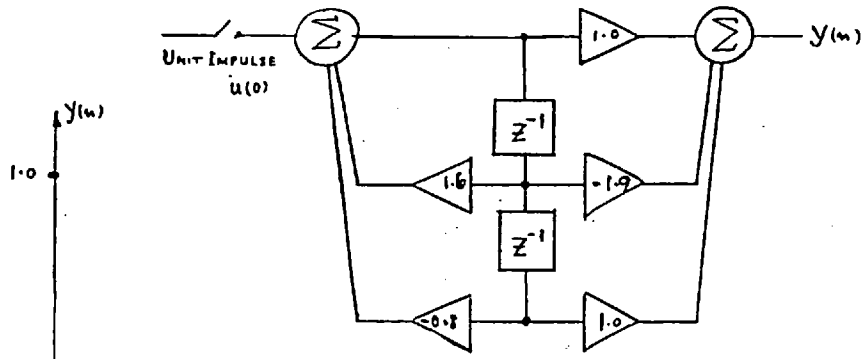


Fig III.16a SINGLE-RATE DIGITAL FILTER AND IMPULSE RESPONSE

Example (III.7 b)

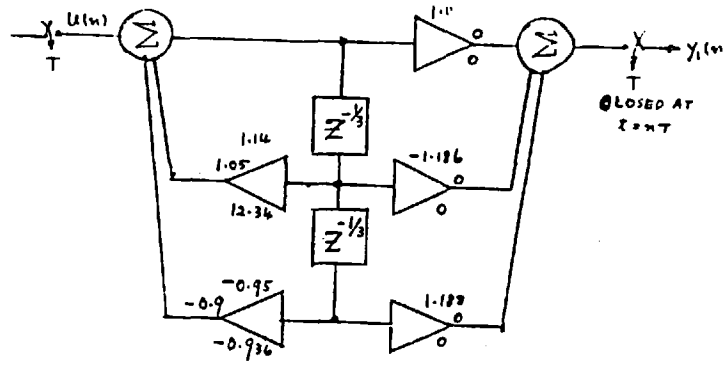


Fig III.16b. EQUIVALENT TRIPLE-RATE PERIODICALLY VARYING DIGITAL FILTER AND IMPULSE RESPONSE

III.12 Résumé

The basic action of a multi-rate digital filter has been described. A mathematical model of such a device has been established using state-space analysis and has been proved to be much more versatile than the other two existing models not only because it gives a general formula for both periodically varying and time-invariant cases, but also that it can be applied to other configurations other than the direct realization without difficulties.

Based on this model of the ideal multi-rate digital filter, some interesting properties of the device have been found, and the design of such a multi-rate filter discussed. Finally, the correctness of the model was verified by a computer simulation program, and the results shown in the examples.

CHAPTER IV

QUANTIZATION ERRORS CAUSED BY ANALOGUE-TO-DIGITAL CONVERSION IN MULTIRATE DIGITAL FILTER

IV.1 Introduction

As mentioned in chapter I, the input signal has to be quantized to a finite number of bits before entering a digital filter. Whether or not the input is considered to be quantized depends on the situation. If the input is inherently discrete, no error exists. In a great many practical cases, however, the input signals are inherently continuous, and the analogue-to-digital conversion is necessary before digital processing can be performed. Thus there is a basic source of error in this conversion. Fig IV.1 shows the action of a 15-level A-D converter, with constant level differences E_o

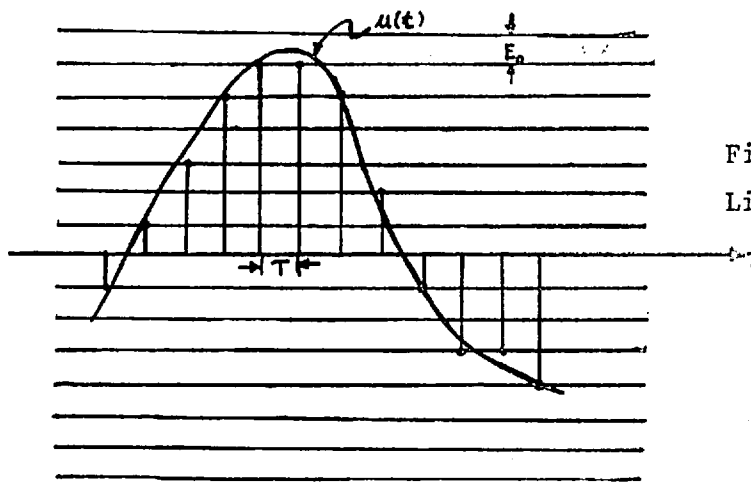


Fig IV.1
Linear Quantization of Analogue Signals

The analogue-to-digital converter in fig IV.1 effectively quantizes the signal. Such kind of quantization which approximates the signal by the nearest quantization level is called rounding,

and is the only type considered here.

It is assumed that the rounding error $e(n)$ associated with the samples is uniformly distributed, the probability density function being as shown in fig IV.2

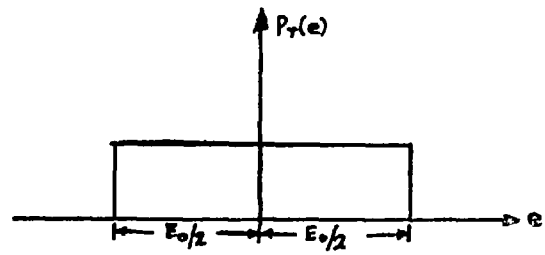


Fig IV.2

If the signal fluctuation is such that many quantization levels can be traversed from one sample to the next, it seems reasonable to expect that the error $e(n)$ at any sampling time will be statistically uncorrelated to $e(m)$, the error at any other sampling time. It is easy to give contrary examples (for instance, when the signal is constant); however, such signals are of extremely narrow bandwidth, and in practice {3}, all signals are likely to have very much richer frequency contents. Therefore this assumption holds for nearly all signals likely to be encountered in practice.

IV.2 Variance of A/D Conversion Noise

Given quantization errors, each with probability density function shown in fig IV.2, it is apparent that the effect is that of noise superimposed on the original analogue signal. The input can thus be expressed as

$$u'(n) = u(n) + e(n)$$

where $u(n)$ can be thought of as a noiseless input and $e(n)$ is the added noise.

Let μ_e and σ_e^2 be the mean and variance respectively of the A/D conversion noise. From fig IV.2, it can be seen that $\mu_e = 0$. Hence, the variance is given by

$$\sigma_e^2 = \int_{-\infty}^{\infty} (e - \mu_e)^2 . \Pr(e)\, de$$

$$= \int_{-E_o/2}^{E_o/2} e^2 . \frac{1}{E_o}\, de$$

$$= \frac{1}{E_o}\left[\frac{e^3}{3}\right]_{-E_o/2}^{E_o/2}$$

i.e. $$\sigma_e^2 = E_o^2/12 \qquad \text{(IV.2)}$$

IV.3 Errors in the Output of a Single-Rate Digital Filter Caused by A/D Conversion

The variance in the output signal, $y(n)$, of a single-rate digital filter may be computed by using linear-system noise theory, if all other errors in the filter are ignored. Since the signal and noise are independent, one can proceed with the noise computation while ignoring the signal. Let the filter be defined by the transfer function $G(z)$ and weighting function $g(n)$. Then the output $y_e(n)$, when the input consists of the noise samples $e(n)$, can be expressed by the convolutional sum {22}

$$y_e(n) = \sum_{i=0}^{n} g(i)\ e(n-i) = \sum_{i=0}^{n} g(n-i)\ e(i) \qquad \text{(IV.3)}$$

It has been assumed that the noise $e(n)$ began at $n = 0$ and was zero before; also the output $y_e(n)$ was assumed to be zero before being excited by the input.

The autocorrelation function of $y_e(k)$ is defined as

$$\phi_{y_e y_e}(n) = \lim_{M\to\infty} \frac{1}{M+1} \sum_{k=0}^{M} y_e(k)\ y_e(k+n)$$
$$= E\left[y_e(k)\ y_e(k+n)\right] \qquad \text{(IV.4)}$$

where E(.) denotes the expected value and $y_e(n)$ is assumed to be ergodic. Substituting this in eqn (IV.3), one obtains

$$\phi_{y_e y_e}(n) = E\left[\sum_{i=0}^{k} g(i)e(k-i)\sum_{j=0}^{k+n} g(j)e(k+n-j)\ \right]$$
$$= \sum_{i=0}^{k}\sum_{j=0}^{k+n} g(i)g(j)\ E\left[e(k-i)e(k+n-j)\right]$$
$$= \sum_{i=0}^{k}\sum_{j=0}^{k+n} g(i)g(j)\ E\left[e(l)e(l+n+i-j)\right] \qquad \text{(IV.5)}$$

where $l = k-i$. However, $e(l)$ is uncorrelated to $e(l+n+i-j)$. Hence

$$E\left[e(l)e(l+n+i-j)\right] = \begin{cases} E\left[e^2(l)\right] & \text{for } n+i-j=0 \\ 0 & \text{otherwise} \end{cases}$$

The variance of $y_e(k)$ is given by the autocorrelation function when $n = 0$, thus,

$$\sigma^2_{ye} = \phi_{y_e y_e}(0)$$
$$= \sum_{i=0}^{k}\sum_{j=0}^{k} g(i)g(j)\ E\left[e(l)e(l+i-j)\right]$$
$$= \sum_{i=0}^{k} g^2(i)\ E\left[e^2(l)\right] \qquad \text{(IV.6)}$$

since $i-j=0$ for $E[e(l)e(l+i-j)]$ to exist

Hence the variance of the output $y_e(k)$ is given by

$$\sigma^2_{ye} = \sigma^2_e \sum_{i=0}^{k} g^2(i)$$
$$= \frac{E_0^2}{12} \sum_{i=0}^{k} g^2(i) \qquad \text{(IV.7)}$$

Notice that σ_{ye}^2 in eqn (IV.7) is, in a sense, a time-dependent result, since it is a function of k, the number of iterations. Since $g^2(i)$ must be positive, σ_{ye}^2 must increase with k. This is reasonable, since one could not expect a large variance in the output immediately after the noise is applied. Physically, the variance of the output builds up and reaches an asymptote. A steady state is always reached if the filter is asymptotically stable. Given that a steady state is reached, it is possible to derive [22], from eqn (IV.7) another formula from which numerical results are usually more easily computed, i.e.

$$\sum_{i=0}^{\infty} g^2(i) = \frac{1}{2\pi j} \oint G(z)G(\tfrac{1}{z})\ z^{-1}\ dz \qquad \text{(IV.8)}$$

Comparing eqn (IV.7) and (IV.8), it is observed that the right hand side of eqn (IV.8) offers an alternative formula for computing the output-noise variance for the steady-state condition, i.e., only when k goes to infinity. This expression is often easier to apply to find out the variance of specific filters.

IV.4 State-Space Approach to Derive the Errors in the Output of a Single-Rate Digital Filter Caused by A/D Conversion

Consider a digital filter represented by the following dynamic equations:-

$$\begin{aligned} \mathbf{x}(n+1) &= \mathbf{A}\,\mathbf{x}(n) + \mathbf{B}u(n) \\ y(n) &= \mathbf{C}\,\mathbf{x}(n) + Du(n) \end{aligned} \qquad \text{(IV.9)}$$

It has been shown in chapter II that the output $y(n)$ of such a digital filter is given by

$$y(n) = \mathbf{C}\sum_{i=0}^{n-1} \mathbf{A}^{n-i-1}\mathbf{B}u(i) + Du(n) \qquad \text{(IV.10)}$$

However, if the input to the filter is a random sequence $e(n)$ caused by the quantization of the continuous signal $u(t)$, then the output error $y_e(n)$ is given by

$$y_e(n) = \mathbf{C}\sum_{i=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}e(i) + De(n) \qquad \text{(IV.11)}$$

Now the covariance matrix of a random vector $\mathbf{v}(n)$ is defined as {41}:-

$$\text{cov}\,[\mathbf{v}(n)] \triangleq E\Big[\{\mathbf{v}(n) - E[\mathbf{v}(n)]\}\{\mathbf{v}^T(n) - E[\mathbf{v}^T(n)]\}\Big] \qquad \text{(IV.12)}$$

where $\mathbf{v}^T$ denotes the transposed of the vector $\mathbf{v}$. The above definition of covariance matrix can be applied to the case of the scalar quantity $y_e(n)$ where $y_e^T(n)$ is, from eqn (IV.11),

$$y_e^T(n) = \sum_{i=0}^{n-1} e^T(i)\,\mathbf{B}^T\left[\mathbf{A}^T\right]^{n-i-1}\mathbf{C}^T + e^T(n)D \qquad \text{(IV.13)}$$

where obviously $y_e^T(n) \equiv y_e(n)$

and $e^T(i) \equiv e(i)$

$D^T \equiv D$

Applying the definition of covariance matrix to the scalar $y_e(n)$, then it can be written that

$$\text{cov}\,[y_e(n)] = E\Big[\{y_e(n)-E[y_e(n)]\}\{y_e^T(n)-E[y_e^T(n)]\}\Big] \qquad \text{(IV.14)}$$

However, since $e(n)$ has zero mean, the output $y_e(n)$ has zero mean, hence

$$\begin{aligned}
\text{cov}[y_e(n)] &= E[y_e(n)y_e^T(n)] \\
&= E\left[\left\{\mathbf{C}\sum_{i=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}e(i)+De(n)\right\}\left\{\sum_{j=0}^{n-1}e^T(j)\mathbf{B}^T(\mathbf{A}^T)^{n-j-1}\mathbf{C}^T+e^T(n)D^T\right\}\right] \\
&= E\Big[\Big\{\mathbf{C}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}e(i)e^T(j)\mathbf{B}^T(\mathbf{A}^T)^{n-j-1}\mathbf{C}^T\Big\}+\Big\{\mathbf{C}\sum_{i=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}e(i)e^T(n)D^T\Big\} \\
&\qquad +\Big\{De(n)\sum_{j=0}^{n-1}e^T(j)\mathbf{B}^T(\mathbf{A}^T)^{n-j-1}\mathbf{C}^T\Big\}+De(n)e^T(n)D^T\Big] \\
&= \Big\{\mathbf{C}\sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}\,E[e(i)e^T(j)]\mathbf{B}^T(\mathbf{A}^T)^{n-j-1}\mathbf{C}^T\Big\} \\
&\qquad +\Big\{\mathbf{C}\sum_{i=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}\,E[e(i)e^T(n)]\,D^T\Big\} \\
&\qquad +\Big\{D\sum_{j=0}^{n-1}E[e(n)e^T(j)]\mathbf{B}^T(\mathbf{A}^T)^{n-j-1}\mathbf{C}^T\Big\} \\
&\qquad + D^2\,E[e(n)e^T(n)]
\end{aligned} \qquad \text{(IV.15)}$$

But $e(i)$ is uncorrelated to $e(j)$ unless $i = j$, i.e.

$$E[e(i)e(j)] = \begin{cases} E_o^2/\ 12 & \text{for } i = j \\ 0 & \text{otherwise,} \end{cases}$$

Hence eqn (IV.15) can be simplified to

$$E[y_e(n)y_e^T(n)] = \mathbf{C}\sum_{i=0}^{n-1}\mathbf{A}^{n-i-1}\mathbf{B}\left(\frac{E_o^2}{12}\right)\mathbf{B}^T(\mathbf{A})^{T\,n-i-1}\mathbf{C}^T + D^2\left(\frac{E_o^2}{12}\right)$$

i.e. the variance of $y_e(n)$ is given by

$$\sigma_{ye}^2 = \frac{E_o^2}{12}\left\{\mathbf{C}\sum_{i=o}^{n-1}\left[\mathbf{A}^{n-i-1}\mathbf{B}\,\mathbf{B}^T(\mathbf{A}^T)^{n-i-1}\mathbf{C}^T\right] + D^2\right\} \qquad \text{(IV.16)}$$

Although eqn (IV.16) can be evaluated readily using a digital computer, it can be further simplified if the digital filter has only a single input and a single output. For a single-input single-output system,

$$\mathbf{B}^T(\mathbf{A}^T)^{n-i-1}\mathbf{C}^T = \mathbf{C}\,\mathbf{A}^{n-i-1}\mathbf{B} \qquad \text{(IV.17)}$$

Thus eqn (IV.16) can be also be written in the following form

$$\sigma_{ye}^2 = \frac{E_o^2}{12}\left\{\sum_{i=0}^{n-1}\left(\mathbf{C}\,\mathbf{A}^{n-i-1}\,\mathbf{B}\right)^2 + D^2\right\} \qquad \text{(IV.18)}$$

Eqn (IV.18) yet offers another expression for evaluating σ_{ye}^2. Comparing eqn (IV.18) to eqn (IV.7), it can be seen that,

$$\sum_{i=0}^{n} g^2(i) = \sum_{i=0}^{n-1}\left(\mathbf{C}\,\mathbf{A}^{n-i-1}\mathbf{B}\right)^2 + D^2 \qquad \text{(IV.19)}$$

The steady-state error for a digital filter due to A/D conversion is thus

$$\sigma_{ye}^2\Big|_{\substack{\text{steady}\\ \text{state}}} = \lim_{n\to\infty}\left\{\frac{E_o^2}{12}\sum_{i=0}^{n-1}\left(\mathbf{C}\,\mathbf{A}^{n-i-1}\mathbf{B}\right)^2 + D^2\right\}$$

$$= \frac{E_o^2}{12}\left\{\frac{1}{2\pi j}\oint G(z)G(z^{-1})\,\frac{dz}{z}\right\} \qquad \text{(IV.20)}$$

IV.5 Verification of the State Space Derivation of the Errors in the Output of a Single-Rate Digital Filter Caused by A/D Conversion

Eqn (IV.19) offers an alternative way of evaluating $\sum_{i=0}^{n} g^2(i)$, and is generally more convenient to use. This is because for a given configuration of the digital filter, the matrices **A**, **B**, **C**, and D are easily determined whereas it is not always easy to find $g(i)$, the impulse response of the filter.

To compute the expression given in eqn (IV.19), it is generally much faster, especially if n is large, to determine the eigenvalues and the eigenvector matrix of **A** first, i.e. making use of the equation

$$\mathbf{A}^n = \mathbf{P}\,\mathbf{\Lambda}^n\,\mathbf{P}^{-1} \qquad \text{(IV.21)}$$

where $\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, λ_1 and λ_2 being the eigenvalues of **A**

and $\mathbf{P} = \begin{bmatrix} \mathbf{R}_1 & \mathbf{R}_2 \end{bmatrix}$, $\mathbf{R}_1$ and $\mathbf{R}_2$ being the eigenvectors of **A** corresponding to λ_1 and λ_2

Hence eqn (IV.19) can be written as

$$g^2(i) = \left\{ \sum_{i=0}^{n-1} \left(\mathbf{C}\,\mathbf{P}\,\mathbf{\Lambda}^{n-i-1}\,\mathbf{P}^{-1}\,\mathbf{B} \right)^2 + D^2 \right\} \qquad \text{(IV.22)}$$

A computer program has been written first to determine $\mathbf{\Lambda}$ and **P**, and then to evaluate the expression in eqn (IV.22). Many different sets of state matrices **A**, **B**, **C**, D, have been used to compute eqn (IV.22) for large values of n. Comparing these values with those obtained by evaluating the integral in eqn (IV.8), it has been found that they are in perfect agreement.

The following is an example verifying that the expression in eqn (IV.19) is correct:-

Example IV.1

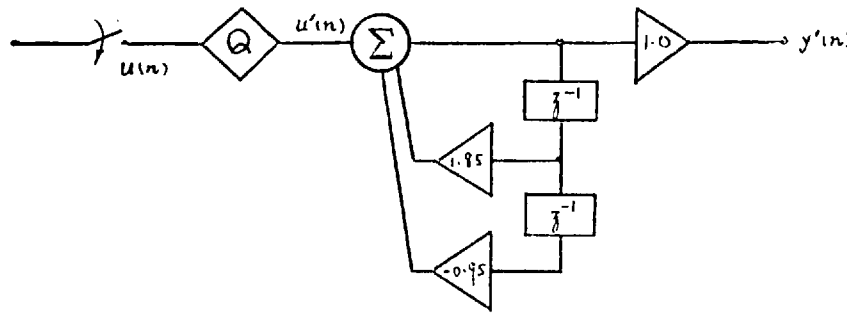


Fig IV.3 A Digital Filter with an Input Quantizer

Fig IV.3 shows a digital filter where the input sequence $u(n)$ is quantized to $u'(n)$. Find the steady state variance of the output error caused by the quantization of the input signal by (a) the state-space method (b) contour integrating over the z-plane if the quantization step at the input is E_o.

a) The state-space approach:-

Assuming that the filter is ideal, i.e. there is no round-off error inside the filter apart from the quantization of the input signal. The state matrices for the filter are:-

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.95 & 1.85 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} -0.95 & 1.85 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} 1 \end{bmatrix}$$

The eigenvalues of **A**, as computed by the program, are

$$\lambda_1 = re^{j\theta}$$
$$\lambda_2 = re^{-j\theta}$$

where $r = \sqrt{0.95}$

$\theta = 18.37$ deg

The eigenvector matrix of **A** is given by

$$\mathbf{P} = \begin{bmatrix} 0.9737 - j0.3234 & 0.9737 + j0.3234 \\ 1 & 1 \end{bmatrix}$$

and the inverse of the eigenvector matrix is

$$\mathbf{P}^{-1} = \begin{bmatrix} j1.5461981 & 0.5 - j1.505086 \\ -j1.5461981 & 0.5 + j1.505086 \end{bmatrix}$$

The value of the expression in eqn (IV.19) is then evaluated for $n = 250$ where $(|\lambda_1|)^n$ is negligible. The result is

$$\sigma^2_{ye} = \frac{E_o^2}{12} \sum_{i=0}^{n-1} \mathbf{C}\,\mathbf{P}\,\mathbf{\Lambda}^{n-i-1}\,\mathbf{P}^{-1}\,\mathbf{B} \quad + D^2 \Big|_{n=250} = \frac{102.63158 E_o^2}{12} \qquad \text{(IV.23)}$$

b) By contour integration over the z-plane:-

The transfer function of the filter is

$$G(z) = \frac{z^2}{z^2 - 1.85z + 0.95} = \frac{z^2}{(z - re^{j\theta})(z - re^{-j\theta})}$$

Hence $$G(z^{-1}) = \frac{1/r^2}{(z - \frac{1}{r} e^{j\theta})(z - \frac{1}{r} e^{j\theta})}$$

where, as given above, $r = \sqrt{0.95}$

$\theta = 18.37$ deg

Thus, we have

$$\sigma_{ye}^2 = \frac{E_o^2/12}{2\pi j} \oint_\Gamma G(z)G(z^{-1}) \frac{dz}{z}$$

$$= \frac{E_o^2/12}{2\pi j} \oint_\Gamma \frac{z/r^2}{(z-re^{j\theta})(z-re^{-j\theta})(z-\frac{1}{r}e^{j\theta})(z-\frac{1}{r}e^{-j\theta})} dz$$

$$= \frac{E_o^2/12}{r^2(r-\frac{1}{r})(e^{j\theta}-e^{-j\theta})} \left\{ \frac{1}{(re^{j\theta}-\frac{1}{r}e^{-j\theta})} + \frac{1}{(\frac{1}{r}e^{j\theta}-re^{-j\theta})} \right\}$$

i.e. $$\sigma_{ye}^2 = \left(\frac{1+r^2}{1-r^2}\right) \frac{E_o^2/12}{r^4-2r^2\cos 2\theta+1}$$

Substituting the values of r and θ, then

$$\sigma_{ye}^2 = 102.6315789 \times \frac{E_o^2}{12}$$

which agrees completely with the results shown in eqn (IV.23), using the state-space formula.

IV.6 Errors in the Output of a Multirate Digital Filter Caused by A/D Conversion

When the quantization noise is passed through a multirate digital filter whose coefficients vary periodically, one would expect the variance of the output error to be weighted by the respective weighting sequence, i.e.

$$\sigma_{ye}^2 = \frac{E_o^2}{12} \sum_{k=o}^{n} h_i^2(k+\frac{i-1}{N}) \qquad \text{(IV.24)}$$

However, to show this relationship rigorously using linear-system noise theory in the case of a multirate digital filter would be rather complicated since the weighting sequence $h_i(n + \frac{i-1}{N})$ is varying periodically. Elaborate theories have been developed for solving problems of linear time-varying system {19}{43}{50} {72}{73}. However, in this case, the state-space approach illustrated in section IV.4 shows great simplicity and renders the problem readily solvable.

A multirate digital filter realized in the direct configuration with periodically varying coefficients can, as shown in section III.5, be represented by the following dynamic equations

$$\begin{aligned} \mathbf{X}(n+1) &= \mathbf{A}_m \mathbf{X}(n) + \mathbf{B}_m u(n) \\ y_i(n+\tfrac{i-1}{N}) &= \mathbf{C}_{mi} \mathbf{X}(n) + \mathbf{D}_{mi} u(n) \end{aligned} \qquad \text{(IV.25)}$$

where

$$\mathbf{A}_m = \mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_2 \mathbf{A}_1$$

$$\mathbf{B}_m = \mathbf{A}_N \mathbf{A}_{N-1} \cdots \mathbf{A}_2 \mathbf{B}_1$$

$$\mathbf{C}_{mi} = \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \cdots \mathbf{A}_2 \mathbf{A}_1$$

$$\mathbf{D}_{mi} = \mathbf{C}_i \mathbf{A}_{i-1} \mathbf{A}_{i-2} \cdots \mathbf{A}_2 \mathbf{B}_1 + \mathbf{D}_i$$

In view of the similarity between eqns (IV.25) and (IV.9), the state space approach can be applied to a multirate digital filter in exactly the same way as illustrated in section IV.4. Thus the variance of the errors in the ith output sequence of a multirate digital filter caused by A/D conversion of the input is given by

$$\sigma^2_{yei} = \frac{E_o^2}{12} \left\{ \sum_{k=0}^{n-1} \left(\mathbf{C}_{mi} \mathbf{A}_m^{n-k-1} \mathbf{B}_m \right)^2 + D_{mi}^2 \right\} \qquad \text{(IV.26)}$$

Eqn (IV.26) is very similar to eqn (IV.18). Hence the ith weighting sequence, $h_i(k + \frac{i-1}{N})$, of the multirate digital filter has the following relationship,

$$\sum_{k=0}^{n} h_i^2(k+\tfrac{i-1}{N}) = \sum_{k=0}^{n-1} (\mathbf{C}_{mi}\mathbf{A}_m^{n-k-1}\mathbf{B}_m)^2 + D^2 \qquad \text{(IV.27)}$$

and for $n \to \infty$, in view of eqn (IV.20), the following expression can be written,

$$\frac{1}{2\pi j}\oint H_i(z)H_i(\tfrac{1}{z})\frac{dz}{z} = \lim_{n\to\infty}\sum_{k=0}^{n-1}(\mathbf{C}_{mi}\mathbf{A}_m^{n-k-1}\mathbf{B}_m)^2 + D_{mi}^2 \qquad \text{(IV.28)}$$

$$\sigma_{yei}^2 = \frac{E_o^2}{12}\sum_{k=0}^{\infty} h_i^2(k+\tfrac{i-1}{N}) = \frac{E_o^2}{12}\left[\frac{1}{2\pi j}\oint H_i(z)H_i(\tfrac{1}{z})\frac{dz}{z}\right] \qquad \text{(IV.29)}$$

If the ith transfer function, $H_i(z)$, of a multirate digital filter is designed to give the same performance as a single-rate filter, $G(z)$, the variance of the output error in both filters due to A/D conversion at the input should be equal, i.e.

$$\sigma_{yei}^2 = \sigma_{ye}^2 \qquad \text{(IV.30)}$$

This is apparent when eqn (IV.29) is compared with eqn (IV.20). Intuitively, this should be obvious since the quantized input can be represented by the ideal input sequence plus a sequence of added noise, i.e.

$$u'(n) = u(n) + e(n) \qquad \text{(IV.31)}$$

If such a sequence is passed through two systems having the same performance, the outputs of the two systems should be identical, and therefore the output errors should be the same, i.e.

$$y'(n) = y(n) + y_e(n) \qquad \text{(IV.32)}$$

where $y(n)$ corresponds to the output of the system due to $u(n)$ and $y_e(n)$ to the output due to $e(n)$.

IV.7 Accuracy of the Statistical Estimation of the Output Errors due to A/D Conversion

A program has been written to test the accuracy of the statistical estimation of the output errors due to A/D conversion. The simulation has been carried out for both the single-rate and multi-rate filters. Fig IV.4 shows a flow chart of the simulation program. An example is given later to show that the statistical estimation of the output error is reasonably accurate.

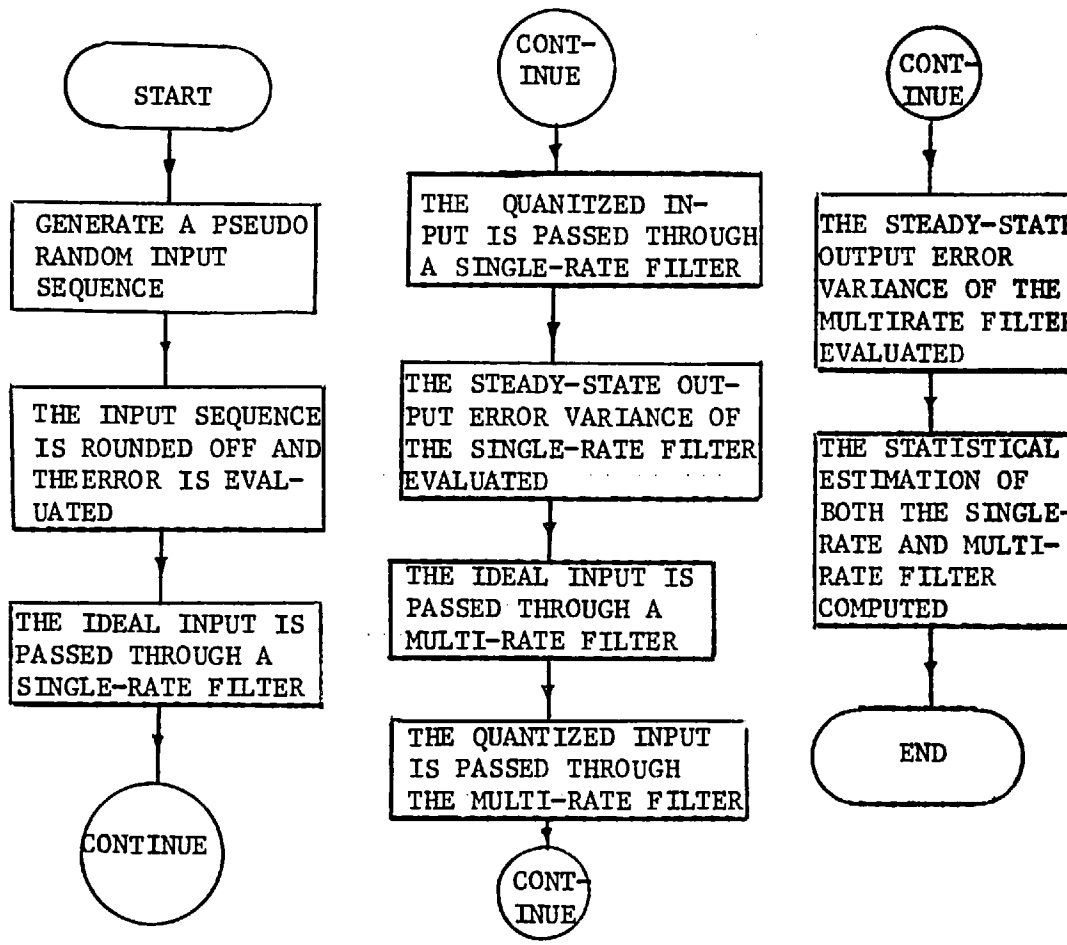
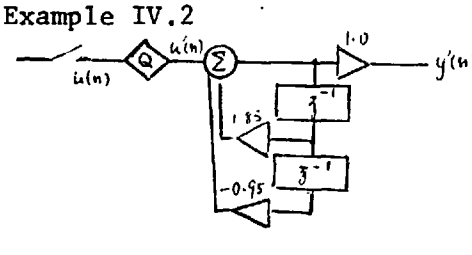
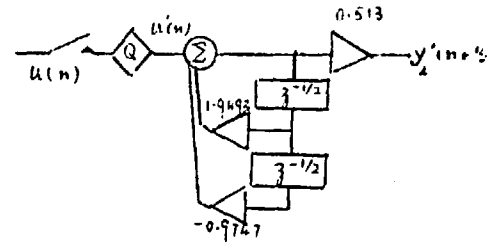


Fig IV.4 Flow-Chart of the Simulation Program

Example IV.2



(a) Single-rate Filter



(b) Equivalent Double-rate Filter

Fig IV.5

Fig IV.5 (a) shows a single-rate digital filter which has a transfer function given by

$$G(z) = \frac{z^2}{z^2 - 1.85z + 0.95}$$

Fig IV.5 (b) shows a time-invariant double-rate digital filter the coefficients of which are designed such that the second output sequence, $y_2(n + \frac{1}{2})$, is identical to the output $y(n)$ of the single-rate filter.

Now a "pseudo-random" sequence $u(n)$, having 350 samples, is quantized to 2 places after decimal and then passed into both the single-rate and double-rate filters separately. The output of both filters are observed and are found to be identical as expected, i.e. $y'(n) = y'(n+\frac{1}{2})$

Now the same "pseudo-random" input sequence without being quantized is passed through either the single-rate or the double-rate filter and the output sequence $y(n)$ recorded. The difference between the last 150 samples of both $y(n)$ and $y_2'(n + \frac{1}{2})$ are taken and the variance (which represents with sufficient accuracy the steady state error variance) calculated.

Let $y_e(n)$ represents the difference between $y_2'(n+\frac{1}{2})$ and $y(n)$. Then the variance of $y_e(n)$ when steady state is reached is, from the above simulation experiment, found to be

$$\sigma^2_{ye}\Big|_{\substack{\text{steady}\\ \text{state}}} = \frac{1}{(150-1)} \sum_{i=201}^{350} \left[y_2(i+\tfrac{1}{2}) - y(i) \right]^2$$
$$= 0.000792 \qquad \text{(IV.33)}$$

The factor (150-1) is taken although there are 150 sample. This is because an unbiased estimate {35}{61} is desired.

This result of $\sigma^2_{ye}\Big|_{\substack{\text{steady}\\ \text{state}}}$ shown in eqn (IV.33) is compared with the statistical estimation of the output error due to the A/D conversion:-

The estimated error can be evaluated in two ways, firstly by contour integral as shown in example IV.1, and secondly by evaluating the state-space expression for a double-rate (or its equivalent single-rate) digital filter as shown in eqn (IV.26)

a) Contour Integral Approach

As shown in example IV.1, the estimated output error due to A/D conversion has a variance given by

$$\sigma^2_{ye} = 102.6315789 \times \frac{E_o^2}{12}$$

where in this case $E_o = 0.01$. Hence we have

$$\sigma^2_{ye} = 102.6315789 \times \frac{(0.01)^2}{12} = 0.000855 \qquad \text{(IV.34)}$$

b) Evaluating the State-Space Expression

From eqn (IV.26) the estimated error in the output is

$$\sigma^2_{yei} = \frac{E_o^2}{12} \left\{ \sum_{k=0}^{n-1} \left(\mathbf{C}_{mi}\, \mathbf{A}_m^{n-k-1}\, \mathbf{B}_m \right)^2 + \mathbf{D}^2_{mi} \right\}$$

where $n = 350 \quad i = 2$

and for this time-invariant double-rate filter

$$\mathbf{A}_m = \begin{bmatrix} 0 & 1 \\ -0.9747 & 1.9492 \end{bmatrix}^2$$

$$\mathbf{B}_m = \begin{bmatrix} 0 & 1 \\ -0.9747 & 1.9492 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\mathbf{C}_{mi} = \begin{bmatrix} -(0.9747\times0.513) & (1.9492\times0.513) \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -0.9747 & 1.9492 \end{bmatrix}$$

$$\mathbf{D}_{mi} = \begin{bmatrix} -(0.9747\times0.513) & (1.9492\times0.513) \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Evaluating the expression with these values, it has been found that

$$\sigma_{ye}^2 = 102.63156 \frac{E_o^2}{12}$$

$$= 0.000855 \qquad \text{(IV.35)}$$

which agrees perfectly with the result evaluated by contour integration shown in eqn (IV.34) as has been expected.

Comparing the statistically estimated result (eqn (IV.35)) with that obtained from computer simulation (eqn (IV.33)), it could be seen that they are in reasonable agreement. The possible source of this small discrepancy is that the variance of the input error due to A/D conversion is equal to $E_o^2/12$ only if an infinite number of samples are taken.

IV.8 Résumé

The error caused by the analogue-to-digital conversion of the signal before entering the digital filter has been introduced

and its properties discussed. The error in the output of a single-rate digital filter due to such A/D conversion is first analysed by linear system noise theory, and then analysed, from a different point of view, using the state-space method. The validity of the state-space approach has been verified by actually evaluating the expression and comparing it with the result obtained by contour integration.

The main advantage of the state-space approach is that it can be applied to the case of periodically varying multirate digital filter without any further elaboration or modification of the theory. Although the result obtained for the error in the output of a multi-rate digital filter caused by A/D conversion using state-space approach is merely a confirmation of what is expected, the method demonstrates the usefulness of state-space analysis.

Finally, the statistically estimated output errors due to A/D conversion in both single-rate and multirate filters have been compared to those obtained from computer simulations of such filters. While the output errors of both single-rate and double-rate filters are identical if the filters are designed to give same outputs, the estimated output error and the experimental output error are in close agreement.

CHAPTER V

POLE SENSITIVITY OF A MULTIRATE DIGITAL FILTER TO THE QUANTIZATION OF THE COEFFICIENTS

V.1 Introduction - Effects of Coefficient Inaccuracy

As a result of the finite word length used in a digital filter, each coefficient is replaced by its l-bit representation. That is, if fixed-point arithmetic is used, the coefficient a_k is replaced by $[a_k]_l$ which equals $(a_k + \Delta a_k)$, with Δa_k bounded in absolute value by 2^{-l}. Similarly, each b_k is replaced by $[b_k]_l$ which is $(b_k + \Delta b_k)$. Therefore, the filter characteristics are changed. This problem can be approached in a number of ways.

Firstly, one can simply compute the frequency response of the actual filter with l-bit rounded coefficients, that is, by using the actual transfer function

$$[H(z)]_l = \frac{\sum_{k=0}^{M} [a_k]_l \cdot z^{-k}}{1 + \sum_{k=1}^{M} [b_k]_l \cdot z^{-k}} \qquad (V.1)$$

The result can then be compared with the ideal response for the original design. For a certain bandstop filter, calculations {38} show that in addition to a greatly increased transition-region width between stop and passbands, the minimum inband rejection deteriorates from 75 dB to less than 50 dB when the wordlength is reduced from 40 bits to 12 bits.

Secondly, if a single number as a measure of the change is desired, an integrated squared deviation of the frequency response

may be used such as

$$\frac{1}{2\pi j}\oint |H(z) - [H(z)]_l|^2 \frac{dz}{z} \qquad (V.2)$$

where $H(z)$ is the ideal transfer function and $[H(z)]_l$ is the transfer function where each coefficient is replaced by its l-bit approximation. By regarding the filter coefficient errors Δa_k and Δb_k as independent random variables, the statistical average of the integrated squared frequency response error as defined by eqn (V.2) has been calculated {38}. However, since the error in each coefficient is fixed throughout the operation of the filter, the validity of the assumption of random coefficient error can be doubted when the order of the filter is low.

Finally, one can also calculate the movements of the poles and zeros of the transfer function due to coefficient rounding and then apply network sensitivity theory to study the changes in the filter response {30}{46}. From these movements, the change in the overall filter response can be studied. This approach has been adopted here in this chapter because of its simplicity and because of the deterministic nature of coefficient round-off errors.

The infinitesimal pole sensitivity is first discussed from the point of view of state space. A comparison of the pole sensitivities of both the time-invariant and periodically varying multirate filter is then made. The idea of "sensitivity ellipse" is then introduced which is established as a criterion for comparing pole sensitivities of second order single rate and time-invariant multirate digital filters. The theoretical analyses are then verified by computer simulations.

V.2 State Space Representation and Infinitesimal Eigenvalue Sensitivity

The dynamic equations of a digital filter, as shown in chapter II, are given by:-

$$\begin{aligned} \mathbf{x}(n+1) &= \mathbf{A}\mathbf{x}(n) + \mathbf{B}u(n) \\ y(n) &= \mathbf{C}\mathbf{x}(n) + \mathbf{D}u(n) \end{aligned} \tag{V.3}$$

and the transfer function of the filter is

$$H(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

$$= \mathbf{C}\frac{\mathrm{Adj}(z\mathbf{I} - \mathbf{A})}{\det(z\mathbf{I} - \mathbf{A})}\mathbf{B} + \mathbf{D} \tag{V.4}$$

Hence the poles of the transfer function are given by the roots of the characteristic equation

$$\det(z\mathbf{I} - \mathbf{A}) = 0 \tag{V.5}$$

that is, the poles of the transfer function are the eigenvalues of the state matrix **A**. Thus the pole sensitivity of $H(z)$ is simply the sensitivity of the eigenvalues of **A**.

Eigenvalue sensitivity is defined as the expected change in the location of an eigenvalue of **A** for a change in a parameter of **A**. Due to the rounding of the coefficients in a digital filter, the parameter of **A** varies. An infinitesimal approximation to this sensitivity, valid for small parameter inaccuracies, is given by

$$\frac{\Delta\lambda_k}{\Delta A_{ij}} \simeq \frac{\partial\lambda_k}{\partial A_{ij}} \tag{V.6}$$

where A_{ij} is a parameter of **A** and λ_k is the k^{th} eigenvalue of **A** satisfying the characteristic equation

$$f(\lambda, A_{ij}) = \det(\lambda\mathbf{I} - \mathbf{A}) = 0 \tag{V.7}$$

This characteristic equation can be written in other forms such as

$$f(\lambda) = \prod_{k=1}^{M} (\lambda - \lambda_k) = \lambda^M + \sum_{l=1}^{M} b_l \lambda^{M-l} \tag{V.8}$$

where of course, λ_k and b_l are functions of A_{ij}

For any form of the state matrix **A**, the infinitesimal sensitivity of the k^{th} eigenvalue to a parameter A_{ij} of **A** is given by

$$\frac{\partial \lambda_k}{\partial A_{ij}} = \left. \frac{\partial f / \partial A_{ij}}{\partial f / \partial \lambda} \right|_{\lambda=\lambda_k} \tag{V.9}$$

Using equation (V.8), this becomes

$$\frac{\partial \lambda_k}{\partial A_{ij}} = \frac{\sum_{l=1}^{M} \frac{\partial b_l}{\partial A_{ij}} \cdot \lambda_k^{M-l}}{\prod_{\substack{m=1 \\ m \neq k}}^{M} (\lambda_k - \lambda_m)} \tag{V.10}$$

which is an estimate of the change in the kth eigenvalue due to a change in A_{ij}. The movement of the kth eigenvalue is thus

$$\Delta\lambda_k \simeq \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\frac{\partial \lambda_k}{\partial A_{ij}} \cdot \Delta A_{ij}\right) = \sum_{i=1}^{M}\sum_{j=1}^{M}\left[\frac{\sum_{l=1}^{M} \frac{\partial b_l}{\partial A_{ij}} \cdot \lambda_k^{M-l}}{\prod_{\substack{m=1 \\ m \neq k}}^{M} (\lambda_k - \lambda_m)} \cdot \Delta A_{ij}\right] \tag{V.11}$$

It should be noted that $\partial\lambda / \partial A_{ij} = 0$ if A_{ij} is unity or zero because these parameters can be realized exactly with digital hardware.

If the digital filter is realized in the direct configuration, the state matrix **A** will be of the companion form, i.e.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ . & . & . & \dots & . \\ 0 & 0 & 0 & \dots & 1 \\ -b_M & -b_{M-1} & -b_{M-2} & \dots & -b_1 \end{bmatrix} \qquad (V.12)$$

and the characteristic equation is then,

$$f(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = \lambda^M + b_1\lambda^{M-1} + \dots + b_{M-1}\lambda + b_M = 0 \quad (V.13)$$

If the digital filter has real coefficients, then the parameter of **A** are real, because the parameters of **A** are simply the coefficients of the digital filter. Since our attention is focused on second order digital filters realized in the direct configuration, a consideration on the eigenvalue sensitivity of the second order companion matrix form will be sufficient.

V.3 Eigenvalue Sensitivity of a Multirate Digital Filter with Periodically Varying Coefficients

It has been shown in chapter III that if a second order digital filter realized in the direct configuration has N shift sequences during each sampling interval while its coefficients are allowed to take on different values every T/N seconds, so that α_{1j} and β_{1j} are the coefficients at nT, α_{2j} and β_{2j} are the coefficients at $(n + 1/N)\,T$ and α_{ij} and β_{ij} are the coefficients at $(n + \frac{i-1}{N})T$, then its state equation is given by,

$$\begin{bmatrix} x_1(n+1) \\ x_2(n+1) \end{bmatrix} = \mathbf{A}_N \cdot \mathbf{A}_{N-1} \cdot \dots \cdot \mathbf{A}_2 \left(\mathbf{A}_1 \begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} + \mathbf{B}_1 u(n) \right) \qquad (V.14)$$

where $\mathbf{A}_i = \begin{bmatrix} 0 & 1 \\ -\beta_{i2} & -\beta_{i1} \end{bmatrix}$ $\qquad \mathbf{B}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

i.e. its state matrix $\mathbf{A}_m$ is given by

$$\mathbf{A}_m = \mathbf{A}_N \cdot \mathbf{A}_{N-1} \cdot \dots \cdot \mathbf{A}_1 \qquad (V.15)$$

It is the aim of this section to find the sensitivity of the eigenvalues of $\mathbf{A}_m$, and to find the condition under which such eigenvalue sensitivity is minimum. However since a completely general approach is rather complicated, it seems sensible to start with the case of a double-rate digital filter, i.e. when $N = 2$.

Consider a double-rate digital filter. If such a filter is used to realize a fixed second order single-rate filter whose state matrix is $\mathbf{A}_S$, where

$$\mathbf{A}_S = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix} \tag{V.16}$$

then the eigenvalues of $\mathbf{A}_S$ and $(\mathbf{A}_2\,\mathbf{A}_1)$ are identical, i.e.

$$z^2 + (\beta_{12} + \beta_{22} - \beta_{11}\beta_{21})z + \beta_{12}\,\beta_{22} \equiv z^2 + b_1 z + b_2 \tag{V.17}$$

Equating the coefficients in eqn (V.17), one obtains

$$\beta_{12} + \beta_{22} - \beta_{11}\beta_{21} = b_1 \tag{V.18}$$

$$\beta_{12}\,\beta_{22} = b_2 \tag{V.19}$$

Equations (V.18) and (V.19) represent the constraints on the choices of β_{11}, β_{12}, β_{21} and β_{22}. If only complex eigenvalues of $\mathbf{A}_S$ are considered and if Λ and Λ^* are the eigenvalues of $\mathbf{A}_S$, then

$$\begin{aligned} \Lambda &= re^{j\theta} \\ \Lambda^* &= re^{-j\theta} \end{aligned} \tag{V.20}$$

where

$$r = \sqrt{b_2} = \sqrt{\beta_{12}\beta_{22}} \tag{V.21}$$

$$\begin{aligned} \theta &= \tan^{-1}\{\sqrt{4b_2 - b_1^2}/(-b_1)\} \\ &= \tan^{-1}\left\{\sqrt{4\beta_{12}\beta_{22} - (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2}\Big/(\beta_{11}\beta_{21} - \beta_{12} - \beta_{22})\right\} \end{aligned} \tag{V.22}$$

and for complex poles $b_1^2 < 4b_2$ and b_2 is positive.

Now for small changes of the coefficients β_{11}, β_{12}, β_{21} and β_{22},

$$\Delta r = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{\partial r}{\partial \beta_{ij}}\right) \cdot \Delta\beta_{ij} \qquad (V.23)$$

$$\Delta\theta = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\frac{\partial \theta}{\partial \beta_{ij}}\right) \cdot \Delta\beta_{ij} \qquad (V.24)$$

Since β_{11}, β_{12}, β_{21} and β_{22} can take on any values provided the constraints of eqn (V.18) and (V.19) are satisfied, the general values of $\Delta\beta_{ij}$ will not be known; thus $\Delta\beta_{ij}$ can be regarded as random variables with zero mean and bounded by $\pm E_o/2$ where E_o is the quantization step for the coefficients. Assuming the quantization steps are equal for all coefficients, then the variance of $\Delta\beta_{ij}$ is given by (section IV.2)

$$\text{var}(\Delta\beta_{ij}) = \frac{E_o^2}{12} \qquad (V.25)$$

Now, the value $\partial r/\partial\beta_{ij}$ is independent of the random variable $\Delta\beta_{ij}$, and $\Delta\beta_{ij}$ are independent of each other, thus the variance of Δr is given by

$$\begin{aligned}\text{var}(\Delta r) &= \text{var}\left(\sum_{i,j}\frac{\partial r}{\partial\beta_{ij}} \cdot \Delta\beta_{ij}\right)\\ &= \sum_{i,j}\left(\frac{\partial r}{\partial\beta_{ij}}\right)^2 \cdot \text{var}(\Delta\beta_{ij})\\ &= \frac{E_o^2}{12}\sum_{i,j}\left(\frac{\partial r}{\partial\beta_{ij}}\right)^2 \qquad (V.26)\end{aligned}$$

Similarly,

$$\text{var}(\Delta\theta) = \frac{E_o^2}{12}\sum_{i,j}\left(\frac{\partial \theta}{\partial \beta_{ij}}\right)^2 \qquad (V.27)$$

It is desired to find the condition under which Var (Δr) and Var ($\Delta\theta$) are at their relative minima. To do this it seems convenient to look at the two quantities separately.

a) Consideration of Var (Δr):-

Now since

$$r = \sqrt{\beta_{12}\,\beta_{22}}$$

then
$$\frac{\partial r}{\partial \beta_{11}} = 0 \qquad \frac{\partial r}{\partial \beta_{12}} = \tfrac{1}{2}\sqrt{\beta_{22}/\beta_{12}}$$

$$\frac{\partial r}{\partial \beta_{21}} = 0 \qquad \frac{\partial r}{\partial \beta_{22}} = \tfrac{1}{2}\sqrt{\beta_{12}/\beta_{22}}$$

Hence the variance of Δr is given by

$$\mathrm{Var}(\Delta r) = \frac{E_o^2}{48}\left(\frac{\beta_{22}}{\beta_{12}} + \frac{\beta_{12}}{\beta_{22}}\right) \qquad (V.28)$$

To find the condition for minimum Var (Δr), the stationary points are first found.

$$\frac{\partial}{\partial \beta_{12}}\left[\mathrm{var}(\Delta r)\right] = \frac{E_o^2}{48}\left(\frac{-\beta_{22}}{\beta_{12}^2} + \frac{1}{\beta_{22}}\right) \qquad (V.29)$$

Equating $\frac{\partial}{\partial \beta_{12}}\left[\mathrm{var}(\Delta r)\right]$ to be zero for stationary points, then

$$\beta_{12} = \beta_{22} \qquad (V.30)$$

(The case when $\beta_{12} = -\beta_{22}$ can be ignored since for complex eigenvalues of $\mathbf{A}_S$, $\beta_{12}.\beta_{22} > 0$). The same condition that $\beta_{12} = \beta_{22}$ for stationary points can be arrived at if Var (Δr) is differentiated with respect to β_{22}.

To test whether the stationary points are relative minima, the second derivatives are calculated {34}:-

$$\left.\begin{aligned}
\frac{\partial^2}{\partial \beta_{12}^2}\left[\mathrm{var}(\Delta r)\right] &= \frac{E_o^2}{48}\left(\frac{2\beta_{22}}{\beta_{12}^3}\right) \\
\frac{\partial^2}{\partial \beta_{22}^2}\left[\mathrm{var}(\Delta r)\right] &= \frac{E_o^2}{48}\left(\frac{2\beta_{12}}{\beta_{22}^3}\right) \\
\frac{\partial^2}{\partial \beta_{12}\,\partial\,\beta_{22}}\left[\mathrm{var}(\Delta r)\right] &= \frac{-E_o^2}{48}\left(\frac{1}{\beta_{12}^2}+\frac{1}{\beta_{22}^2}\right)
\end{aligned}\right\} \qquad (V.31)$$

At $\beta_{12} = \beta_{22}$

$$\left\{\frac{\partial^2}{\partial\beta_{12}\partial\beta_{22}}\left[\mathrm{var}(\Delta r)\right]\right\}^2\Bigg|_{\beta_{12}=\beta_{22}} - \left\{\frac{\partial^2}{\partial\beta_{12}^2}\left[\mathrm{var}(\Delta r)\right]\cdot\frac{\partial^2}{\partial\beta_{22}^2}\left[\mathrm{var}(\Delta r)\right]\right\}\Bigg|_{\beta_{12}=\beta_{22}}$$

$$= \left(\frac{E_o^2}{48}\right)^2\left\{\left(\frac{1}{\beta_{12}^2}+\frac{1}{\beta_{22}^2}\right)^2 - \frac{4\beta_{12}\,\beta_{22}}{\beta_{12}^3\beta_{22}^3}\right\}\Bigg|_{\beta_{12}=\beta_{22}}$$

$$= \left(\frac{E_o^2}{48}\right)^2\left\{\frac{4}{\beta_{12}^2}-\frac{4}{\beta_{12}^2}\right\} = 0 \qquad (V.32)$$

Hence further investigation is needed to understand the nature of the stationary points. To carry out the investigation, a transformation of axis on the $\beta_{12}-\beta_{22}$ plane is performed.

Let the axes of reference on the $\beta_{12}-\beta_{22}$ plane be transformed to the lines

$$\beta_{12} = \beta_{22}$$
$$\beta_{12} = -\beta_{22} \qquad (V.33)$$

and let these two lines of new reference axes be denoted by β'_{12} β'_{22}. Then, the following equations can be written (fig V.1),

$$\left.\begin{aligned} \beta_{12} &= \beta'_{12}\cos\gamma - \beta'_{22}\sin\gamma \\ \beta_{22} &= \beta'_{12}\sin\gamma + \beta'_{22}\cos\gamma \end{aligned}\right\} \qquad (V.34)$$

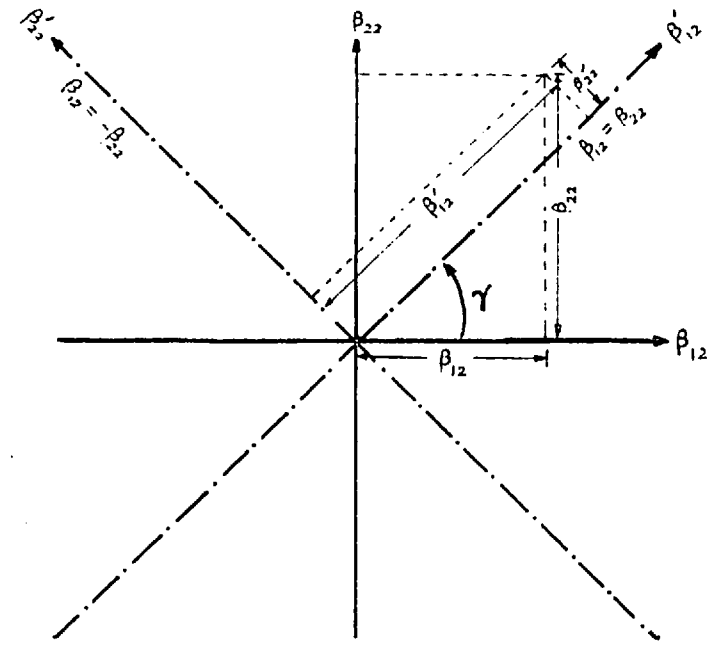


Fig. V.1 Rotation of Reference Axes on the $\beta_{12}-\beta_{22}$ Plane

Now $\gamma = \frac{\pi}{4}$, $\quad \sin\gamma = \cos\gamma = 1/\sqrt{2}$

Hence substituting eqn (V.34) into (V.28), one obtains

$$\mathrm{var}(\Delta r) = \frac{E_o^2}{24}\left\{\frac{\beta'^2_{12}+\beta'^2_{22}}{\beta'^2_{12}-\beta'^2_{22}}\right\} \qquad (V.35)$$

and differentiating with respect to β'_{12} and β'_{22}

$$\frac{\partial\,\mathrm{var}(\Delta r)}{\partial\,\beta'_{12}} = \frac{E_o^2}{24}\left\{\frac{(\beta'^2_{12}-\beta'^2_{22}).2\beta'_{12} - (\beta'^2_{12}+\beta'^2_{22})\,2\beta'_{12}}{(\beta'^2_{12}-\beta'^2_{22})^2}\right\}$$

$$= \frac{E_o^2}{12}\left\{\frac{-2\beta'_{12}\beta'^2_{22}}{(\beta'^2_{12}-\beta'^2_{22})^2}\right\} \qquad (V.36)$$

$$\frac{\partial\,\mathrm{var}(\Delta r)}{\partial\,\beta'_{22}} = \frac{E_o^2}{12}\left\{\frac{2\beta'^2_{12}\beta'_{22}}{(\beta'^2_{12}-\beta'^2_{22})^2}\right\} \qquad (V.37)$$

Equating both equations (V.36) and (V.37) to zero for stationary points, then

$$\beta'_{12}\,\beta'_{22} = 0$$

i.e. $\beta'_{12} = 0$ or $\beta'_{22} = 0$. But for $\beta'_{12} = 0$, then Var $(\Delta r) < 0$ which contradicts the definition of variance, hence it can be concluded that stationary points occur only when $\beta'_{22} = 0$.

To study the nature of these stationary points, the second derivatives of Var (Δr) with respect to β'_{12} and β'_{22} are found i.e.

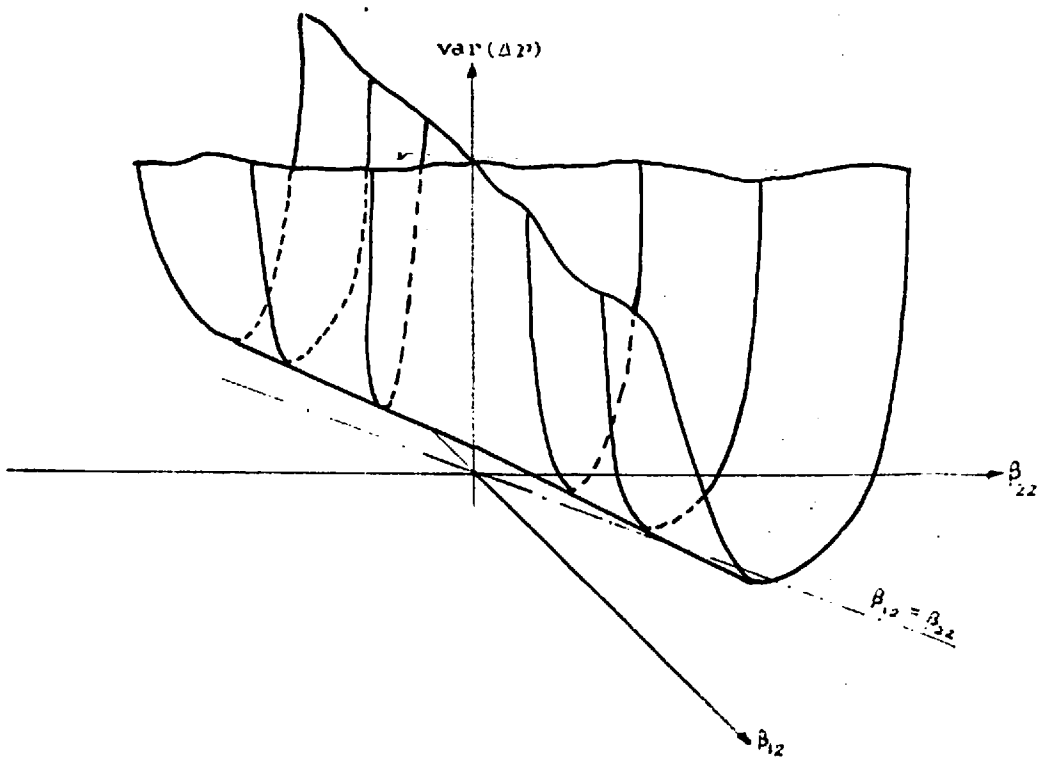
$$\left.\frac{\partial^2 \text{var}(\Delta r)}{\partial \beta'^2_{12}}\right|_{\beta'_{22}=0} = \frac{E_o^2}{12}\left\{\frac{4\beta'_{12}\beta'_{22}(\beta'^2_{12}-\beta'^2_{22})^2 - 8\beta'^3_{12}\beta'_{22}(\beta'^2_{12}-\beta'^2_{22})}{(\beta'^2_{12}-\beta'^2_{22})^4}\right\}\Bigg|_{\beta'_{22}=0}$$

$$= 0 \qquad \text{(V.38)}$$

$$\left.\frac{\partial^2 \text{var}(\Delta r)}{\partial \beta'^2_{22}}\right|_{\beta'_{22}=0} = \frac{E_o^2}{12}\left\{\frac{2\beta'^2_{12}(\beta'^2_{12}-\beta'^2_{22})^2 + 8\beta'^2_{12}\beta'^2_{22}(\beta'^2_{12}-\beta'^2_{22})}{(\beta'^2_{12}-\beta'^2_{22})^4}\right\}\Bigg|_{\beta'_{22}=0}$$

$$= \frac{E_o^2}{12}\,\frac{2}{\beta'^2_{12}} > 0 \qquad \text{(V.39)}$$

Eqns (V.36) and (V.38) show that the gradient of Var (Δr) at $\beta'_{22} = 0$ is constantly zero, i.e. the value of Var (Δr) is a constant along the line $\beta'_{22} = 0$. However eqns (V.37) and (V.39) show that at $\beta'_{22} = 0$ the value of Var (Δr) is at its minimum with respect to the change of β'_{22}. In fact the shape of the function Var (Δr) is of the form as shown in fig (V.2)



It can be seen from the above argument and from fig (V.2) that Var (Δr) is at its relative minimum when $\beta_{12} = \beta_{22}$. Thus for a double-rate digital filter, in general, the change in the moduli of the resultant poles is a minimum when $\beta_{12} = \beta_{22}$.

b) Consideration of Var ($\Delta\theta$):-

Eqn (V.22) gives the expression for θ in terms of β_{11}, β_{12}, β_{21} and β_{22}, i.e.

$$\theta = \tan^{-1}\left\{\sqrt{4\beta_{12}\beta_{22} - (\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2}\,/\,(\beta_{11}\beta_{21}-\beta_{12}-\beta_{21})\right\} \qquad \text{(V.40)}$$

Differentiating with respect to β_{ij}, we have

$$\left.\begin{aligned}\left(\frac{\partial\theta}{\partial\beta_{11}}\right)^2 &= \beta_{21}^2 \,/\, \left\{4\beta_{12}\beta_{22}-(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2\right\} \\ \left(\frac{\partial\theta}{\partial\beta_{21}}\right)^2 &= \beta_{11}^2 \,/\, \left\{4\beta_{12}\beta_{22}-(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2\right\}\end{aligned}\right\} \qquad \text{(V.41a)}$$

$$\left.\begin{aligned}\left(\frac{\partial\theta}{\partial\beta_{12}}\right)^2 &= \frac{1}{4\beta_{12}^2}\left\{\frac{(\beta_{12}-\beta_{22}+\beta_{11}\beta_{21})^2}{4\beta_{12}\beta_{22}-(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2}\right\}\\ \left(\frac{\partial\theta}{\partial\beta_{22}}\right)^2 &= \frac{1}{4\beta_{22}^2}\left\{\frac{(-\beta_{12}+\beta_{22}+\beta_{11}\beta_{21})^2}{4\beta_{12}\beta_{22}-(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2}\right\}\end{aligned}\right\} \quad (V.41b)$$

Hence substituting eqns (V.41) into eqn (V.27), we have

$$\text{var}(\Delta\theta) = \frac{E_o^2}{12}\left[\frac{4\beta_{12}^2\beta_{22}^2\beta_{21}^2+4\beta_{12}^2\beta_{22}^2\beta_{11}^2+\beta_{22}^2(\beta_{12}-\beta_{22}+\beta_{11}\beta_{21})^2+\beta_{12}^2(-\beta_{12}+\beta_{22}+\beta_{11}\beta_{21})^2}{4\beta_{12}^2\beta_{22}^2\left[4\beta_{12}\beta_{22}-(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})^2\right]}\right] \quad (V.42)$$

However, it is known that Var (Δr) is a minimum for $\beta_{12} = \beta_{22}$, if this condition is kept in the consideration of Var ($\Delta\theta$), then eqn (V.42) can be written as:-

$$\text{var}(\Delta\theta) = \frac{E_o^2}{12}\left\{\frac{2\beta_{12}^2\beta_{21}^2 + 2\beta_{12}^2\beta_{11}^2 + \beta_{11}^2\beta_{21}^2}{2\beta_{12}^2\left[4\beta_{12}^2 - (2\beta_{12} - \beta_{11}\beta_{21})^2\right]}\right\} \quad (V.43)$$

Applying the constraints of eqns (V.18) and (V.19), such that

$$2\beta_{12} - \beta_{11}\beta_{21} = b_1 \quad (V.44)$$

$$\beta_{12}^2 = b2 \quad (V.45)$$

then,

$$\text{var}(\Delta\theta) = \frac{E_o^2}{12}\left(\frac{2b_2\beta_{21}^2 + 2b_2\beta_{11}^2 + \beta_{11}^2\beta_{21}^2}{2b_2\left[4b_2 - b_1^2\right]}\right) \quad (V.46)$$

But from eqns (V.44) and (V.45),

$$\beta_{11}\beta_{21} = \pm 2\sqrt{b_2} - b_1$$

i.e. $$\beta_{21} = K/\beta_{11} \quad (V.47)$$

where $$K = \pm 2\sqrt{b_2} - b_1 \quad (V.48)$$

Hence eqn (V.47) simplifies to,

$$\text{var}(\Delta\theta) = \frac{E_o^2}{12}\left\{\frac{2b_2(\beta_{11}^2+\frac{K^2}{\beta_{11}^2})+K^2}{2b_2(4b_2-b_1^2)}\right\} \qquad (V.49)$$

Differentiating with respect to β_{11} and equating to zero for stationary points, we have

$$\frac{d\ \text{var}(\Delta\theta)}{d\beta_{11}} = \frac{E_o^2}{12}\ \frac{2}{(4b_2-b_1)}(\beta_{11}-\frac{K^2}{\beta_{11}^3}) = 0$$

i.e. for stationary points,

$$\beta_{11} = \pm\sqrt{K} \qquad (V.50)$$

and since β_{11} is real, the negative sign before $2\sqrt{b_2}$ in eqn (V.48) can be ignored. Substituting eqn (V.50) into eqn (V.47), we have the final condition for stationary points in Var ($\Delta\theta$) as

$$\beta_{11} = \beta_{21} = \pm\sqrt{K} = \pm\sqrt{2\sqrt{b_2}-b_1} \qquad (V.51)$$

It should be noticed that the value $(2\sqrt{b_2}-b_1) > 0$ for complex poles. Evaluating the second derivative to investigate the nature of the stationary points, it is found that

$$\left.\frac{d^2\ \text{var}(\Delta\theta)}{d\ \beta_{11}^2}\right|_{\beta_{11}=\pm\sqrt{K}} = \frac{E_o^2}{12}\cdot\frac{2}{(4b_2-b_1^2)}\left(1+\frac{3K^2}{\beta_{11}^4}\right)\Bigg|_{\beta_{11}=\pm\sqrt{K}} = \frac{E_o^2}{12}\cdot\frac{8}{(4b_2-b_1^2)} > 0 \qquad (V.52)$$

Thus the stationary points at $\beta_{11} = \beta_{21}$ are relative minima of Var ($\Delta\theta$). In fact, the shape of the curve when Var ($\Delta\theta$) is plotted against β_{11} is shown in fig (V.3),

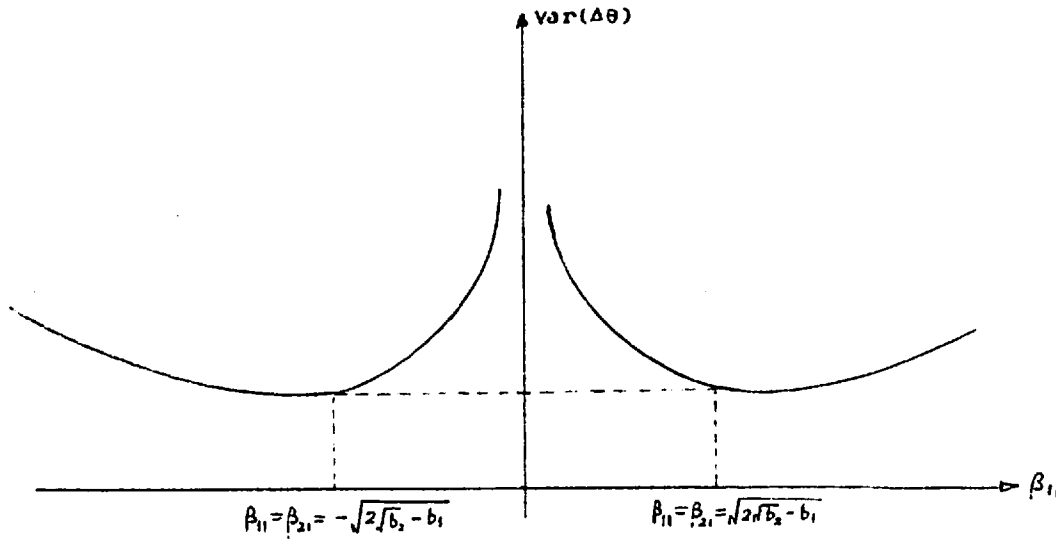


Fig V.3 The Points of Minimum Values of var(Δθ) with Respect to β_{11}

From the above considerations, it can be concluded that, in general, for a double-rate second order digital filter, a time-invariant realization has the least sensitive poles. However, the above considerations are based on the fact that $\Delta\beta_{ij}$ are unknown random variables bounded by $\pm E_o/2$. If it can be chosen that two of the $\Delta\beta_{ij}$ are zero, the resultant double-rate filter with periodically varying coefficients may have less sensitive poles than the time-invariant filter.

A complete and rigorous generalisation of the above theory is complicated and difficult (see {11}). However, judging from the similarity in the expressions of γ and θ for $N > 2$, a reasonable conjecture can be made that a general multi-rate digital filter would have the least pole sensitivity if it is time-invariant. Furthermore, extensive computer simulation seems to confirm such a conjecture. The following section consists of some of the examples extracted from the simulation.

V.4 Comparison of Pole Sensitivities between Time-Invariant and Periodically Varying Multi-rate Digital Filters

The considerations in the previous section indicate that the eigenvalue sensitivity of a general multi-rate digital filter is least when the filter is time-invariant. A computer program is written to simulate the time-invariant and periodically varying multi-rate filters so that the effects of quantization of the filter coefficients can be observed and compared.

According to a given transfer function, the program first designs two multi-rate digital filters (both having the same rate, i.e. equal N), one having fixed coefficients and the other having periodically varying coefficients. Both filters are designed to give the same performance as required by the given transfer function. Then the coefficients of these filters so designed are rounded-off to the same finite accuracy. The outputs of these filters with quantized coefficients are then plotted on the same diagram with the outputs of the ideal filters. The discrepancies between the outputs of the filter with quantized coefficients and that of the ideal filters are examined.

It has been observed that, in general, the discrepancy is larger in the case of periodically varying multi-rate filter than in the case of time-invariant multi-rate filter. However, if, in particular, the coefficients of the periodically varying filter are designed such that some of its coefficients present no quantization error, then, in that case, the discrepancy between the outputs of the periodically varying multi-rate filter and that of the ideal filter may be lower.

The following are a few examples taken from the simulation. These results, in general, support the analysis developed in section V.3.

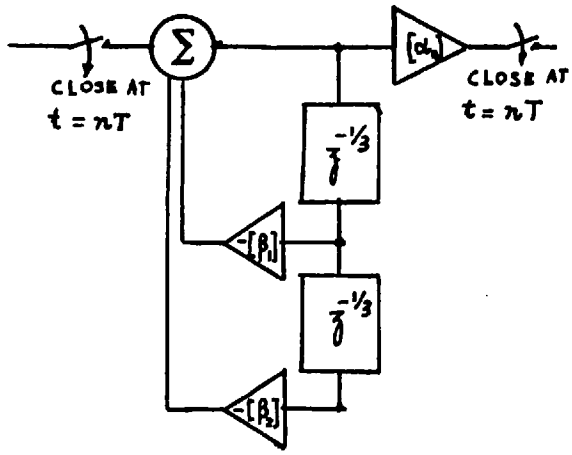
Example V.1

To design two triple-rate digital filters, one with fixed coefficients the other with periodically varying coefficients, so that each will have an ideal L.P. characteristic according to the following transfer function

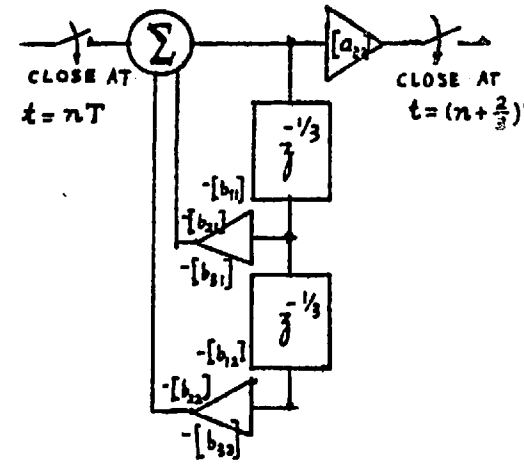
$$G(z) = \frac{0.27996}{1 - 1.8355179z^{-1} + 0.8464z^{-2}} \qquad (V.53)$$

Each of these two triple-rate filters are to have coefficients rounded off to two places after decimal, (decimal arithmetic is used throughout), and their performances compared.

The triple-rate filter designs are shown in the following diagrams (fig V.4(a) and (b)). α_o, $-\beta_1$ and $-\beta_2$ are the ideal coefficients while $[\alpha_o]$, $[-\beta_1]$ and $[-\beta_2]$ are the coefficients rounded-off to two places after decimal. Similarly, a_{30}, $-b_{11}$, $-b_{12}$, $-b_{21}$, $-b_{22}$, $-b_{31}$, $-b_{32}$ and $[a_{30}]$, $[-b_{11}]$, $[-b_{12}]$, $[-b_{21}]$, $[-b_{22}]$, $[-b_{31}]$, $[-b_{32}]$ are respectively the ideal and quantized coefficients of the triple-rate digital filter with periodically varying coefficients. The third output sequence (i.e. $y(n + 2/3)$) of each of the triple-rate filter is used to realize eqn (V.53) because, as mentioned in section III.10, less multipliers are used.



(a) Time-Invariant Triple-Rate Digital Filter



(b) Triple-Rate Digital Filter with Periodically Varying Coefficients

Fig V.4

The design gives the following values for the coefficients.

Time invariant filter:-

α_o = - 3.7226802	$[\alpha_0]$ = - 3.72
β_1 = 0.93312654	$[\beta_1]$ = 0.93
β_2 = 0.94592902	$[\beta_2]$ = 0.95

Periodically varying filter:-

a_{30} = 0.5573384	$[a_{30}]$ = 0.56
b_{11} = 0.8532	$[b_{11}]$ = 0.85
b_{12} = 0.885	$[b_{12}]$ = 0.89
b_{21} = 1.4531	$[b_{21}]$ = 1.45
b_{22} = 0.9255	$[b_{22}]$ = 0.93
b_{31} = 1.0568345	$[b_{31}]$ = 1.06
b_{32} = 1.0333703	$[b_{32}]$ = 1.03

It should be noted that in designing the triple-rate filter with periodically varying coefficients, two of the sets of coefficients (i.e. b_{i1}, b_{i2}) have to be specified. Here in this example, b_{11}, b_{12}, b_{21} and b_{22} are specified and chosen more or less randomly within reasonable limits (not having tremendous differences between values).

The output sequences of these filters are shown in fig(V.5(a) and (b)). On these diagrams, the ideal filter response is plotted in juxtaposition with the response of the filter with quantized coefficients. It can be seen from these diagrams that the time-invariant triple-rate filter deviates less from the ideal response than the periodically varying triple-rate filter, and hence confirming the analysis in section V.3 that a time-invariant multi-rate digital filter is, in general, less in pole sensitivity than a periodically varying filter.

Fig V.5(a) Impulse Response of a Triple-Rate Digital LP Filter with Fixed Coefficients

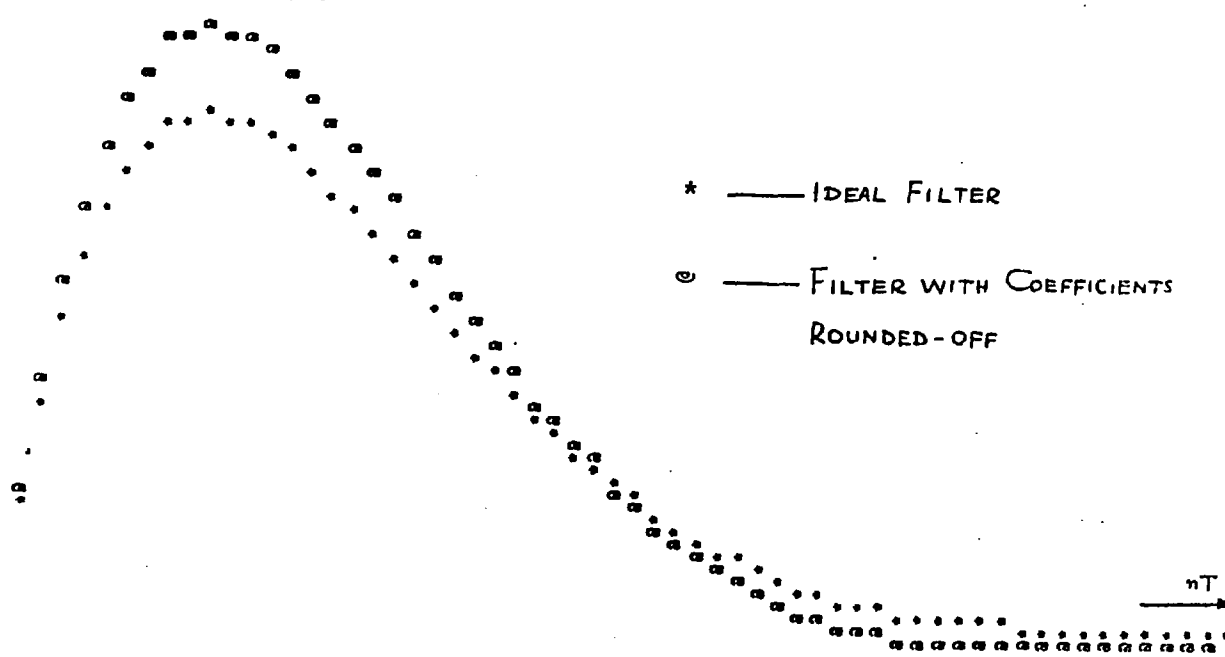
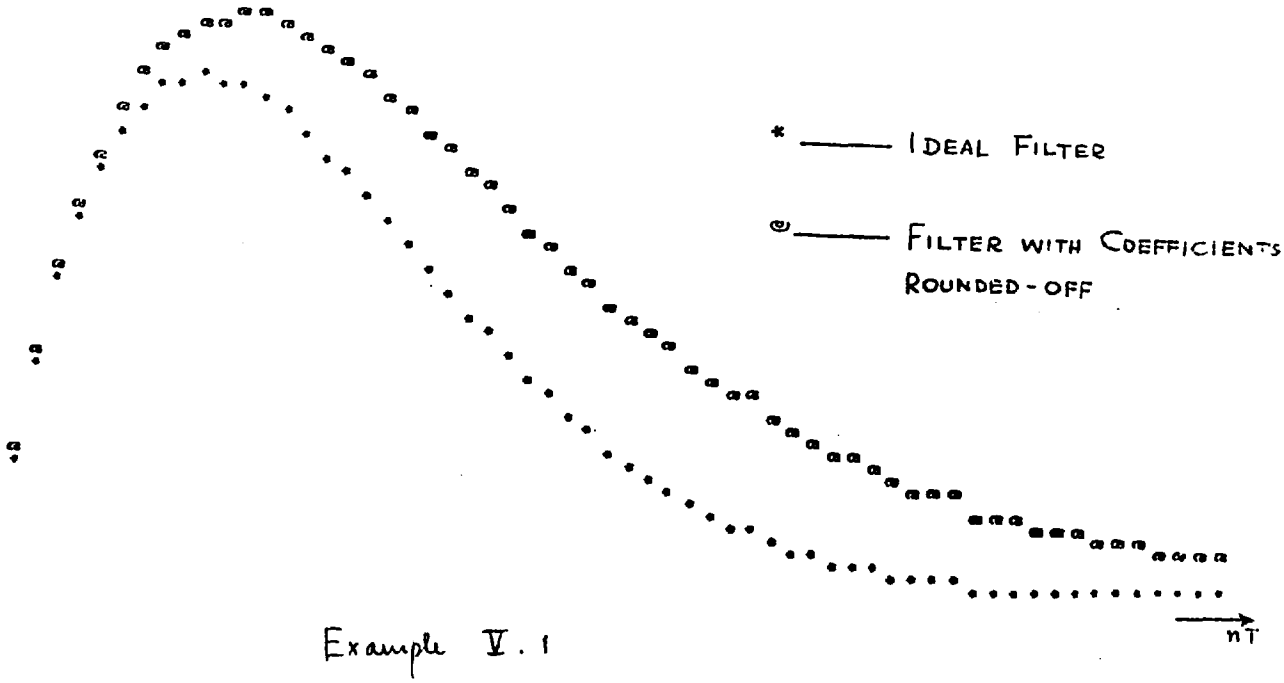


Fig V.5(b) Impulse Response of a Triple-Rate Digital LP Filter with Periodically Varying Coefficients



Example V.2

As in example V.1, two triple-rate digital filters are designed; but, this time, to give a HP characteristic according to the following transfer function,

$$G(z) = \frac{0.27996}{1 + 1.8355179z^{-1} + 0.8464z^{-2}} \qquad (V.54)$$

Again both cases of time-invariant and periodically varying triple-rate filters are considered, each coefficient being rounded-off to two places after decimal.

The design gives the following values for the coefficients.

Time-invariant filter:-

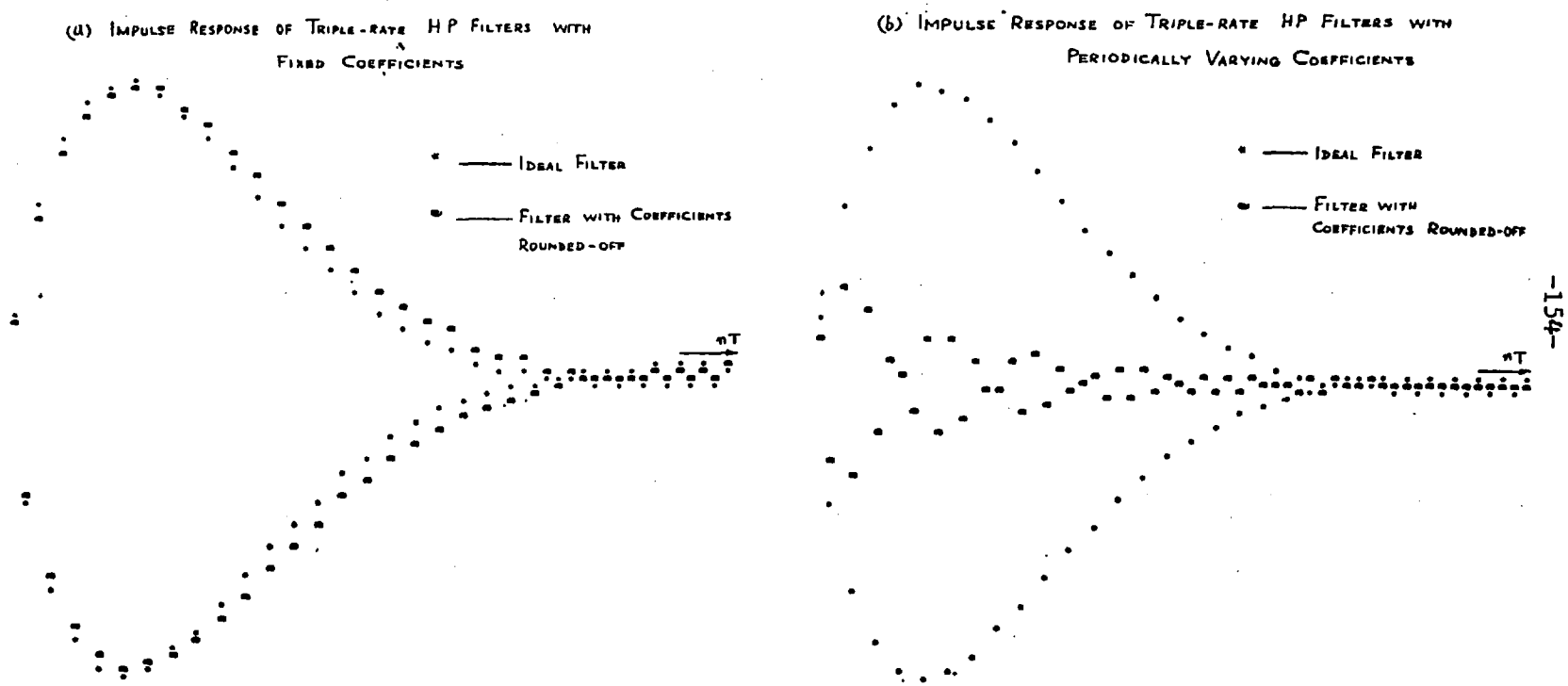
$\alpha_o = 3.6239862$	$[\alpha_o] = 3.62$
$\beta_1 = -1.0115241$	$[\beta_1] = -1.01$
$\beta_2 = 0.94592902$	$[\beta_2] = 0.95$

Periodically varying filter:-

$a_{30} = 0.014449962$	$[a_{30}] = 0.01$
$b_{11} = 0.832$	$[b_{11}] = 0.83$
$b_{12} = 0.885$	$[b_{12}] = 0.89$
$b_{21} = 1.4531$	$[b_{21}] = 1.45$
$b_{22} = 0.9255$	$[b_{22}] = 0.93$
$b_{31} = 14.04433$	$[b_{31}] = 14.04$
$b_{32} = 1.0333703$	$[b_{32}] = 1.03$

Again the impulse response of the filters are plotted (figV.6(a), (b)). This time, the periodically varying triple-rate filter deviates very much from the ideal response. This probably is due to the fact that some of its coefficients are so vastly different. Again this example confirms the result of the analysis in section V.3.

FIG V. 6 for Example V.2



Example V.3.

To design two double-rate digital filters so that both would give the following H.P transfer function

$$G(z) = \frac{0.8235}{1 + 1.78\ z^{-1} + 0.8012 z^{-2}} \qquad (V.55)$$

Again one of the double-rate filters has fixed coefficients while the other has periodically varying coefficients. b_{11} and b_{12} are chosen more or less randomly in designing the periodically varying double-rate digital filter. The second output sequences of both double-rate filters (i.e. $y(n + \frac{1}{2})$) are used to realize eqn (V.55).

The following values for the coefficients are obtained from the design of the filters:-

Time-invariant double-rate filter:-

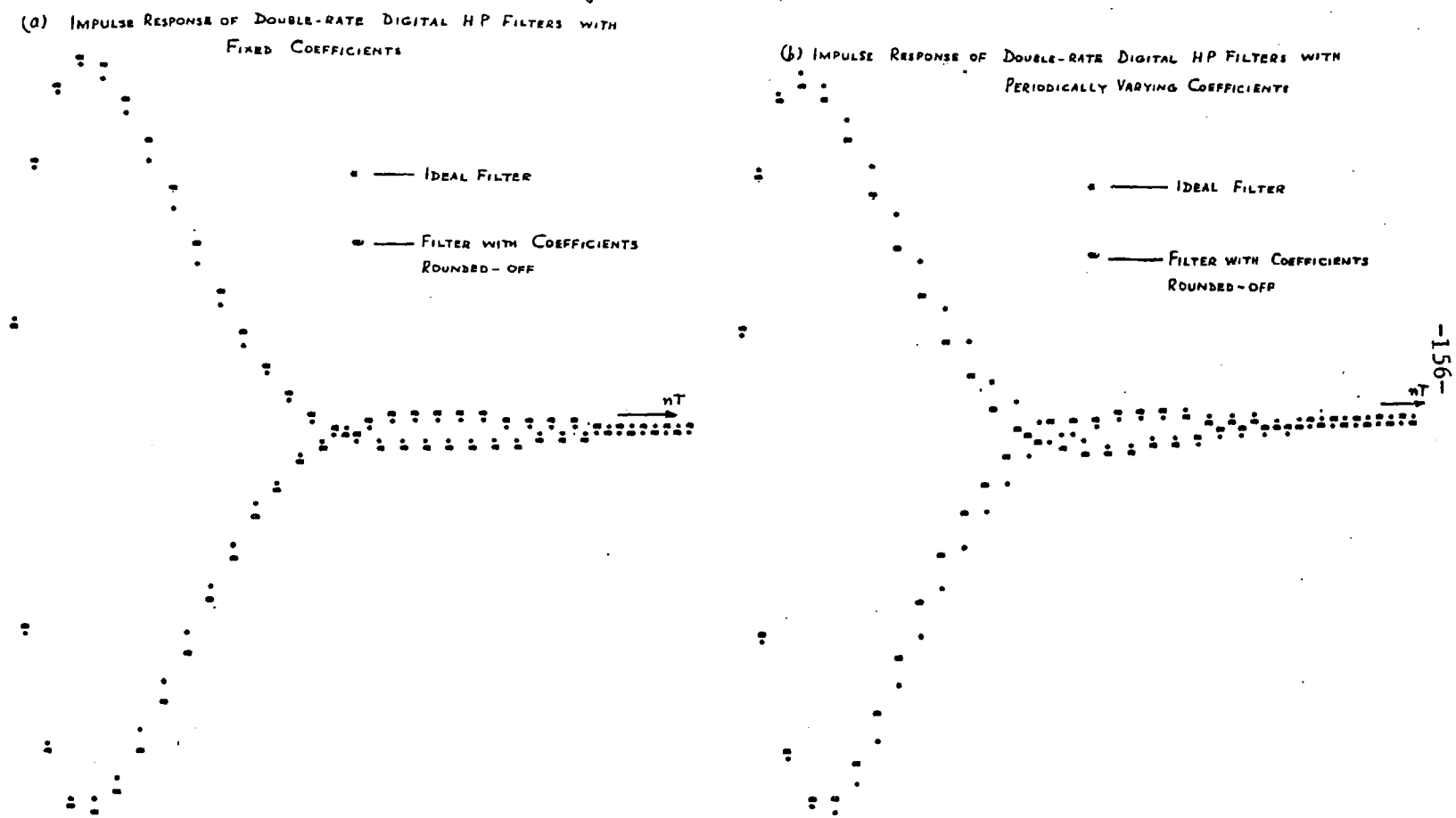
$\alpha_o = 8.1556562$	$[\alpha_o] = 8.16$
$\beta_1 = -0.10097287$	$[\beta_1] = -0.10$
$\beta_2 = 0.89509776$	$[\beta_2] = 0.90$

Double-rate filter with periodically varying coefficients:-

$\alpha_{20} = -1.0290287$	$[\alpha_{20}] = -1.03$
$\beta_{11} = 0.0254$	$[\beta_{11}] = 0.03$
$\beta_{12} = 0.8048$	$[\beta_{12}] = 0.80$
$\beta_{21} = 0.80026925$	$[\beta_{21}] = 0.80$
$\beta_{22} = 0.99552684$	$[\beta_{22}] = 1.00$

The impulse responses are plotted in fig (V.7). Again the time-invariant filter shows smaller deviation from the ideal output.

Fig V.7 for Example V.3



Example V.4

As in the previous example, two double-rate digital filters are designed to give the HP transfer function given by eqn (V.55). But this time the coefficients of the periodically varying double-rate filters are so specified that two of the coefficients β_{11} and β_{12} present no error after quantization. Again, the performances of both the time-invariant and periodically varying double-rate filters are compared.

The following values for the coefficients are obtained.

Time-invariant filter:-

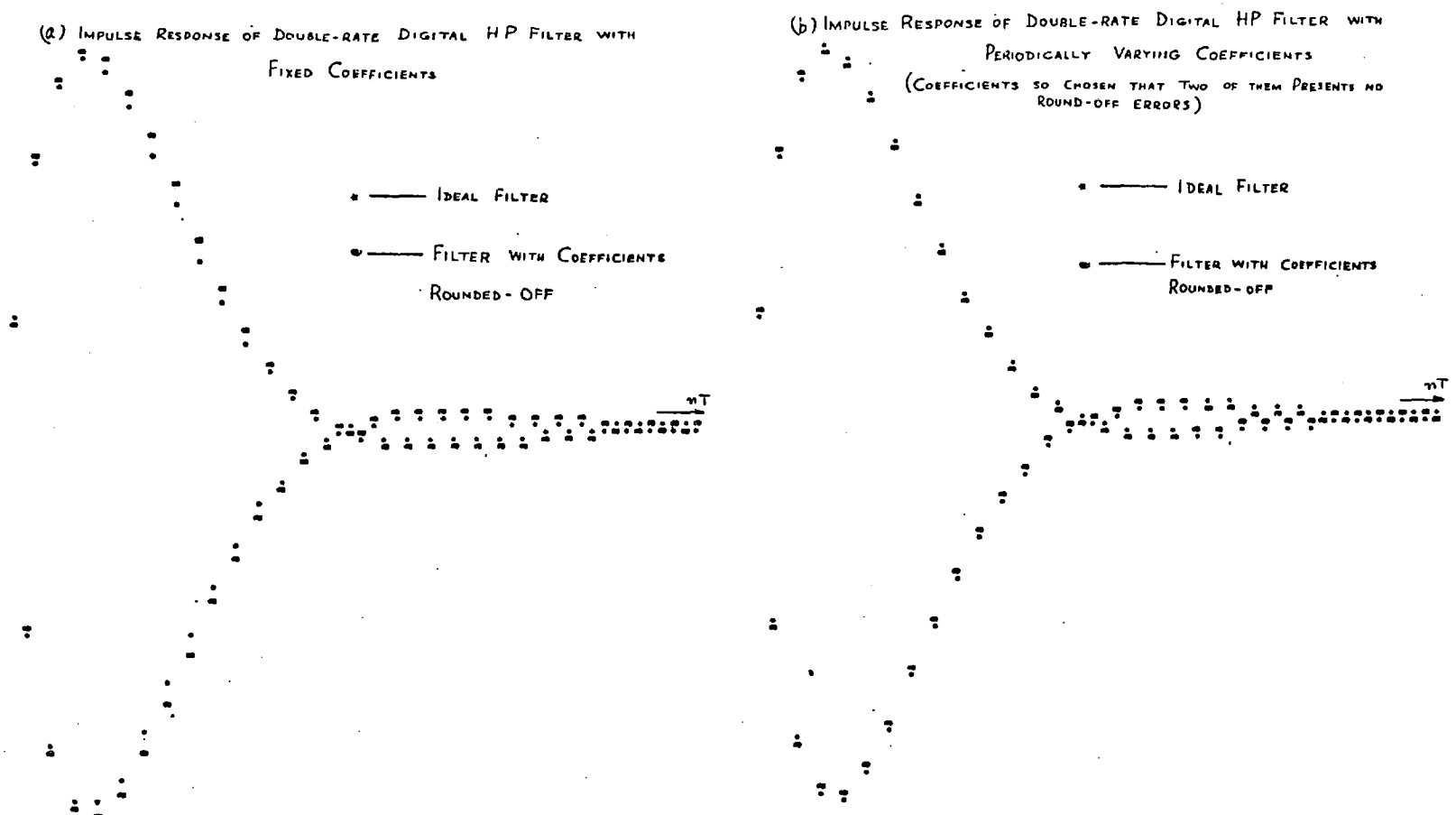
$$\alpha_o = 8.1556562 \qquad [\alpha_o] = 8.16$$
$$\beta_1 = -0.10097287 \qquad [\beta_1] = -0.10$$
$$\beta_2 = 0.89509776 \qquad [\beta_2] = 0.90$$

Double-rate filter with periodically varying coefficients:-

$$\alpha_{20} = -1.1490698 \qquad [\alpha_{20}] = -1.15$$
$$\beta_{11} = 0.03 \qquad [\beta_{11}] = 0.03$$
$$\beta_{12} = 0.80 \qquad [\beta_{12}] = 0.80$$
$$\beta_{21} = 0.7166667 \qquad [\beta_{21}] = 0.72$$
$$\beta_{22} = 1.0015 \qquad [\beta_{22}] = 1.00$$

As can be seen, $\Delta\beta_{11} = \Delta\beta_{12} = 0$. The impulse responses are shown in fig V.4. This time the deviation from the ideal output is smaller in the case of the periodically varying double-rate filter. The reason is, quite obviously, that two of the coefficients β_{11} and β_{12} are exactly realized and thus reduced two sources of error.

Fig V.8 for Example V.4



The above examples all support the result of the analysis given in section V.3. From these examples, it can be seen that in general, the poles of a multi-rate digital filter is least sensitive to coefficient quantization errors when the filter is time-invariant, its sensitivity can still be further reduced if careful choices are made such that some of its coefficients give no quantization error. Hence, the general rule seems to be: the coefficients should be chosen so that one set of coefficients is as nearly equal as possible to the other sets, but at the same time full use should be made of the freedom in choosing these sets of coefficients so that the maximum number of coefficients are specified to be exactly realizable by the filter hardware.

V.5. Sensitivity Ellipse - a Criterion for Measuring Pole Sensitivities

This section and the next contain the development of some results already published {68}. The object of this section is to examine the effect of quantization of the coefficients on both the single-rate and multi-rate digital filters. Since, in general, the pole sensitivity of a periodically varying multi-rate filter is unknown and is, as shown in section V.3., greater than that of the corresponding time-invariant multi-rate filter, attention has been focused on the time-invariant case. Examination of the effect of coefficient quantization on both the single-rate and the time-invariant multi-rate filter leads to the concept of "sensitivity ellipse" {46} {68} which is used as a criterion for measuring pole sensitivity of digital filters. The cases of the single-rate and the time-invariant multi-rate filters are considered separately:-

a) Sensitivity Ellipse of a Single-rate Digital Filter

Consider the second order transfer function

$$H(z^{-1}) = \frac{a_o + a_1 z^{-1} + a_2 z^{-2}}{1 + b_1 z^{-1} + b_2 z^{-2}} \tag{V.56}$$

and let such a second order system be represented by the following dynamic equations:-

$$\begin{aligned} \mathbf{x}_s(n+1) &= \mathbf{A}_s \mathbf{x}_s(n) + \mathbf{B}_s u(n) \\ y_s(n) &= \mathbf{C}_s \mathbf{x}_s(n) + D_s u(n) \end{aligned} \tag{V.57}$$

where, if the transfer function is realized in the direct form

$$\mathbf{A}_s = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix}, \qquad \mathbf{B}_s = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\mathbf{C}_s = \begin{bmatrix} (a_2 - a_o b_2) & (a_1 - a_o b_1) \end{bmatrix}, D_s = a_o \tag{V.58}$$

Let Λ and Λ^* be the complex eigenvalues of $\mathbf{A}_s$, then one obtains the following relationships,

$$\begin{aligned} \Lambda &= re^{j\theta} \\ \Lambda^* &= re^{-j\theta} \end{aligned} \tag{V.59}$$

where

$$r = \sqrt{b_2} \tag{V.60}$$

$$\tan\theta = \frac{-1}{b_1}\sqrt{4b_2 - b_1^2} \tag{V.61}$$

For small changes of b_1 and b_2, the change of the pole position is given by

$$\Delta\Lambda \simeq \frac{\partial\Lambda}{\partial b} \cdot \Delta b_1 + \frac{\partial\Lambda}{\partial b} \cdot \Delta b_2 \tag{V.62}$$

However, from eqn (V.59),

$$\begin{aligned} \frac{\partial\Lambda}{\partial b_i} &= \left(\frac{\partial\Lambda}{\partial r} \cdot \frac{\partial r}{\partial b_1}\right) + \left(\frac{\partial\Lambda}{\partial\theta} \cdot \frac{\partial\theta}{\partial b_1}\right) \\ &= \left(e^{j\theta} \cdot \frac{\partial r}{\partial b_1}\right) + \left(jre^{j\theta} \cdot \frac{\partial\theta}{\partial b_1}\right) \end{aligned} \tag{V.63}$$

and similarly,

$$\frac{\partial \Lambda}{\partial b_2} = e^{j\theta} \cdot \frac{\partial r}{\partial b_2} + jre^{j\theta} \cdot \frac{\partial \theta}{\partial b_2} \tag{V.64}$$

Now, writing eqns (V.60) and (V.61) in the following way

$$f(r,\theta,b_1,b_2) = r - \sqrt{b_2} = 0 \tag{V.65}$$

$$g(r,\theta,b_1,b_2) = \tan\theta + \frac{1}{b_1}\sqrt{4b_2 - b_1^2} = 0 \tag{V.66}$$

then,

$$\frac{\partial r}{\partial b_1} = -\frac{\partial(f,g)}{\partial(b_1,\theta)} \Big/ \frac{\partial(f,g)}{\partial(r,\theta)} = -\begin{vmatrix} f_{b1} & f_\theta \\ g_{b1} & g_\theta \end{vmatrix} \Big/ \begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix} = 0$$

$$\frac{\partial \theta}{\partial b_1} = -\frac{\partial(f,g)}{\partial(r,b_1)} \Big/ \frac{\partial(f,g)}{\partial(r,\theta)} = \frac{1}{2r\sin\theta}$$

$$\frac{\partial r}{\partial b_2} = -\frac{\partial(f,g)}{\partial(b_2,\theta)} \Big/ \frac{\partial(f,g)}{\partial(r,\theta)} = \frac{1}{2r} \tag{V.67}$$

$$\frac{\partial \theta}{\partial b_2} = -\frac{\partial(f,g)}{\partial(r,b_2)} \Big/ \frac{\partial(f,g)}{\partial(r,\theta)} = \frac{\cos\theta}{2r^2\sin\theta}$$

Substituting these equations of (V.67) into (V.63) and (V.64), then one obtains

$$\frac{\partial \Lambda}{\partial b_1} = -\tfrac{1}{2}\left(1 - j\,\frac{\cos\theta}{\sin\theta}\right) \tag{V.68}$$

$$\frac{\partial \Lambda}{\partial b_2} = j\frac{1}{2r\sin\theta} \tag{V.69}$$

Thus the change of the pole due to small changes of b_1 and b_2 is given by

$$\Delta\Lambda \simeq -\tfrac{1}{2}\Delta b_1 + j\frac{1}{2\sin\theta}\left(\cos\theta \cdot \Delta b_1 + \frac{\Delta b_2}{r}\right) \tag{V.70}$$

For complex conjugate poles

$$\Delta(\Lambda^*) = (\Delta\Lambda)^* \simeq -\frac{\Delta b_1}{2} - j\frac{1}{2\sin\theta}\left(\cos\theta \cdot \Delta b_1 + \frac{\Delta b_2}{r}\right) \qquad (V.71)$$

If a condition is imposed on the movement of the poles such that for small changes of b_1 and b_2, each pole would not move out of a small circle of radius σ, then

$$|\Delta\Lambda| \leq \sigma \qquad (V.72)$$

and from eqn (V.70), this condition can be rewritten as

$$\tfrac{1}{4}(1+\cot^2\theta)(\Delta b_1)^2 + 2\left(\frac{\cos\theta}{4r\sin^2\theta}\right)(\Delta b_1)(\Delta b_2) + \frac{1}{4r^2\sin^2\theta}(\Delta b_2) \leq \sigma^2 \qquad (V.73)$$

Taking the equality sign, it can be seen that the equation of the condition represents a general conic on the (Δb_1) — (Δb_2) plane. To test the nature of the conic, the quantity

$$\left\{\tfrac{1}{4}(1+\cot^2\theta)\cdot\frac{1}{4r^2\sin^2\theta} - \left(\frac{\cos\theta}{4r\sin^2\theta}\right)^2\right\}$$

is examined[6]

Now,
$$\tfrac{1}{4}(1+\cot^2\theta)\frac{1}{4r^2\sin^2\theta} - \left(\frac{\cos\theta}{4r\sin^2\theta}\right)^2 = \frac{1}{16r^2\sin^2\theta} > 0 \qquad (V.74)$$

then, equation (V.73) is an ellipse on the (Δb_1) — (Δb_2) plane. Thus, if (Δb_1) and (Δb_2) are both within the ellipse, then Λ(and Λ*) will not move out of the small circle of radius σ. This ellipse which stipulates the magnitude of Δb_1 and Δb_2 for a restricted movement of Λ is designated the "pole sensitivity ellipse" of a second order digital filter.

b) Sensitivity Ellipse of Time-invariant Multirate Digital Filter

As shown in chapter III, if a second order digital filter realized in the direct form (fig V.9) is used in a multi-rate fashion, then its dynamic equation will be

$$\begin{aligned} \mathbf{x}_m(n+1) &= \mathbf{A}^N \mathbf{x}_m(n) + \mathbf{A}^{N-1}\mathbf{B}\,u(n) \\ y_{mi}\left(n+\frac{i-1}{N}\right) &= \mathbf{C}\,\mathbf{A}^{i-1}\,\mathbf{x}_m(n) + \left(\mathbf{C}\,\mathbf{A}^{i-2}\,\mathbf{B} + D_i\right)u(n) \end{aligned} \qquad (V.75)$$

where N is the number of shift of signal in the filter within one sampling period,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\beta_2 & -\beta_1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} (\alpha_2 - \alpha_0\beta_2) & (\alpha_1 - \alpha_0\beta_1) \end{bmatrix}, D_i = \begin{cases} \alpha_0 & \text{for } i = 1 \\ 0 & \text{for } 1 < i \leq N \end{cases}$$

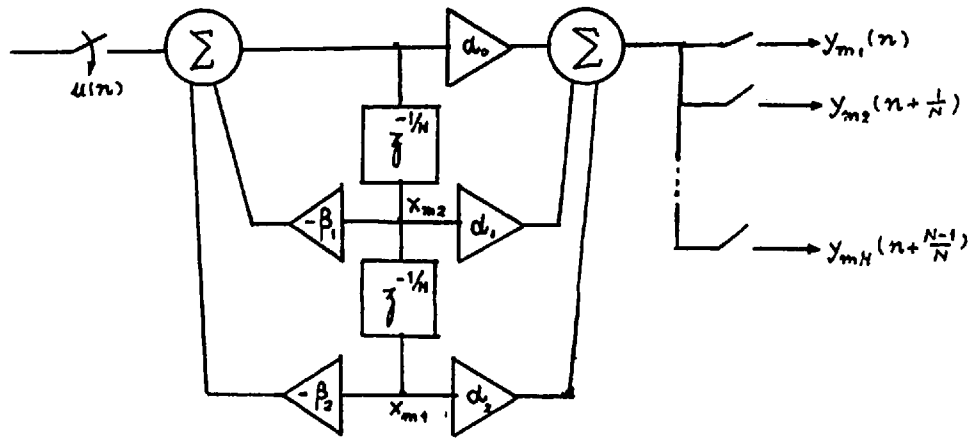


Fig V.9 A Time-Invariant Multi-Rate Digital Filter

Let λ and λ^* be the complex eigenvalues of $\mathbf{A}$, then

$$\begin{aligned} \lambda &= \rho e^{j\phi} \\ \lambda^* &= \rho e^{-j\phi} \end{aligned} \tag{V.76}$$

where
$$\rho = \sqrt{\beta_2} \tag{V.77}$$

$$\tan\phi = \frac{-1}{\beta_1}\sqrt{4\beta_2 - \beta_1^2} \tag{V.78}$$

If this multi-rate digital filter is to realize the second-order transfer function shown in eqn (V.56), then the eigenvalues of $\mathbf{A}^N$ are identical to those of $\mathbf{A}_s$, i.e.

$$\begin{aligned} \Lambda &= \lambda^N \\ \Lambda^* &= (\lambda^*)^N \end{aligned} \tag{V.79}$$

It is desired to see the effect of quantization of β_1 and β_2 on the change of Λ and Λ^*. Now, for small changes of β_1 and β_2, then

$$\Delta\Lambda \simeq \frac{\partial\Lambda}{\partial\beta_1}\cdot\Delta\beta_1 + \frac{\partial\Lambda}{\partial\beta_2}\cdot\Delta\beta_2$$

$$= \frac{\partial\Lambda}{\partial\lambda}\left(\frac{\partial\lambda}{\partial\beta_1}\cdot\Delta\beta_1 + \frac{\partial\lambda}{\partial\beta_2}\cdot\Delta\beta_2\right)$$

$$= N\lambda^{N-1}\left[\frac{\partial\lambda}{\partial\beta_1}\cdot\Delta\beta_1 + \frac{\partial\lambda}{\partial\beta_2}\cdot\Delta\beta_2\right] \qquad (V.80)$$

The terms inside the bracket of eqn (V.80) is similar to RHS of eqn (V.80), thus, the same procedure in evaluating these quantities can be followed. Using the same method as in the case of the single-rate filter, one can write,

$$\Delta\Lambda \simeq N\lambda^{N-1}\left[\sum_{i=1}^{2} e^{j\phi}\frac{\partial\rho}{\partial\beta_i} + j\rho e^{j\phi}\frac{\partial\phi}{\partial\beta_i}\right] \qquad (V.81)$$

where the values of the various partial differentials can be evaluated by rewriting eqns (V.77) and (V.78) and calculating their respective Jacobians.

Hence one obtains,

$$\Delta\Lambda \simeq N\rho^{N-1}e^{j(N-1)\phi}\left[-\frac{1}{2}\Delta\beta_1 + j\frac{1}{2\sin\phi}\left(\cos\phi\cdot\Delta\beta_1 + \frac{\Delta\beta_2}{\rho}\right)\right] \qquad (V.82)$$

$$\Delta\Lambda^* \simeq N\rho^{N-1}e^{-j(N-1)\phi}\left[-\frac{1}{2}\Delta\beta_1 - j\frac{1}{2\sin\phi}\left(\cos\phi\cdot\Delta\beta_1 + \frac{\Delta\beta_2}{\rho}\right)\right] \qquad (V.83)$$

Again imposing a constraint to the movement of the pole Λ, such that it does not move out of a circle of radius σ, then

$$|\Delta\Lambda| \leq \sigma \qquad (V.84)$$

i.e.

$$N^2\rho^{2(N-1)}\left\{\frac{1}{4}(1+\cot^2\phi)(\Delta\beta_1)^2 + 2\left(\frac{\cos\phi}{4\rho\sin^2\phi}\right)(\Delta\beta_1)(\Delta\beta_2) + \frac{1}{4\rho^2\sin^2\phi}(\Delta\beta_2)^2\right\}$$

$$\leqslant \sigma^2 \qquad (V.85)$$

which, in view of its similarity to eqn (V.73) represents the equation of an ellipse on the $(\Delta\beta_1)$ - $(\Delta\beta_2)$ plane. Again, if $\Delta\beta_1$ and $\Delta\beta_2$ are both within the ellipse, then Λ and Λ^* will not move out of the circle of radius σ

Example V.5.

This example illustrates the idea of the "sensitivity ellipse" of a digital filter:-

A second order digital filter is to be designed such that it has the following transfer function

$$G(z^{-1}) = \frac{1}{1 + 1.62z^{-1} + 0.725z^{-2}} \qquad (V.86)$$

It is desired that the movements of the poles due to the quantization of the coefficients be confined to a circle of radius $\sigma = 0.01$ on the z-plane. If the transfer function is realized as

a) a single-rate digital filter
b) a double-rate time-invariant filter
c) a triple-rate time-invariant filter,

find the equations and plot the graphs of the ellipses of the three filters, such that if the coefficient quantization errors lie inside the ellipse, the above condition is not violated.

a) If the transfer function of eqn (V.85) is realized as a single-rate filter, then the coefficients of the filter are simply the coefficients of the transfer function. The radius and the angle of the complex poles are given by:-

$$r = \sqrt{b_2} = 0.851469$$
$$\theta = 162.0445 \text{ deg}$$

$$2.63026(\Delta b_1)^2 - 5.87808(\Delta b_1)(\Delta b_2) + 3.62845(\Delta b_2)^2 = 0.0001 \quad (V.87)$$

b) If the transfer function is realized as a double-rate time-invariant filter, then the feedback multipliers of the filter are given by:-

$$\beta_2 = \sqrt{b_2}$$
$$\beta_1 = \pm \sqrt{2\beta_2 - b_1}$$

Hence

$$\rho = r^{\frac{1}{2}} = 0.922751$$
$$\phi = \theta/2 = 81.02225 \text{ deg}$$

and the equation of the ellipse on the $(\Delta\beta_1) - (\Delta\beta_2)$ plane is, from eqn (V.85),

$$0.25624(\Delta\beta_1)^2 + .0866674(\Delta\beta_2)(\Delta\beta_2) + 0.300938(\Delta\beta_2)^2 = 0.00002936 \quad (V.88)$$

c) Similar to a double-rate filter, a triple-rate filter has the following parameters,

$$\rho = r^{1/3} = 0.947814$$

$$\phi = \frac{2\pi - \theta}{3} = 65.98526 \text{ deg}$$

and the equation of the ellipse is,

$$0.299626(\Delta\beta_1)^2 + 0.257208(\Delta\beta_1)(\Delta\beta_2) + 0.333529(\Delta\beta_2)^2 = 0.00001377 \quad (V.89)$$

All the three ellipses of eqns (V.87), (V.88) and (V.89) are plotted in fig V.10.

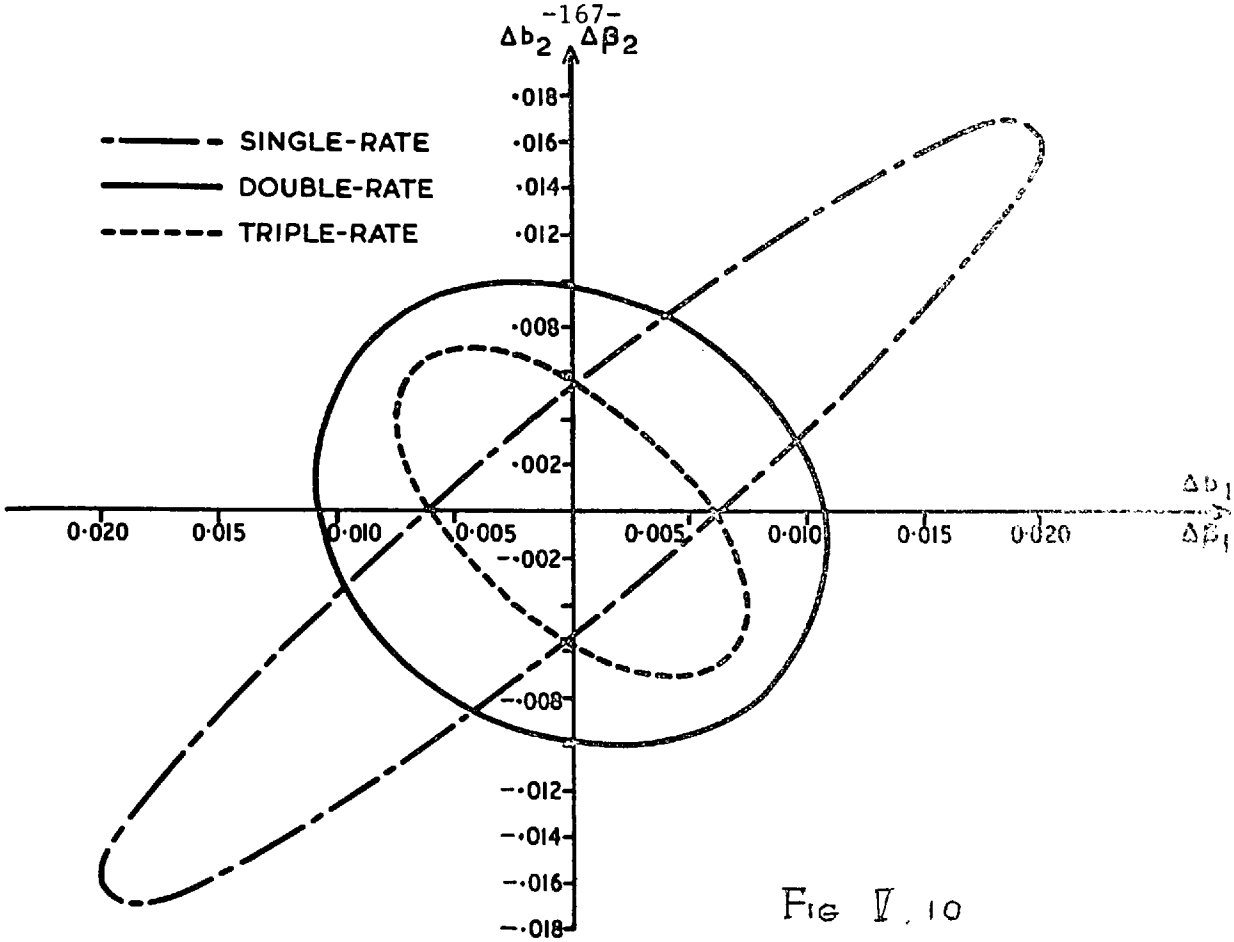


Fig V.10

V.6 Single-rate and Time Invariant Multi-rate Digital Filters -- a Comparison of Pole Sensitivities

The previous section introduces the concept of "sensitivity ellipse". Here in this section, the concept is utilized to compare the pole sensitivities of the single-rate and the multi-rate time-invariant filters.

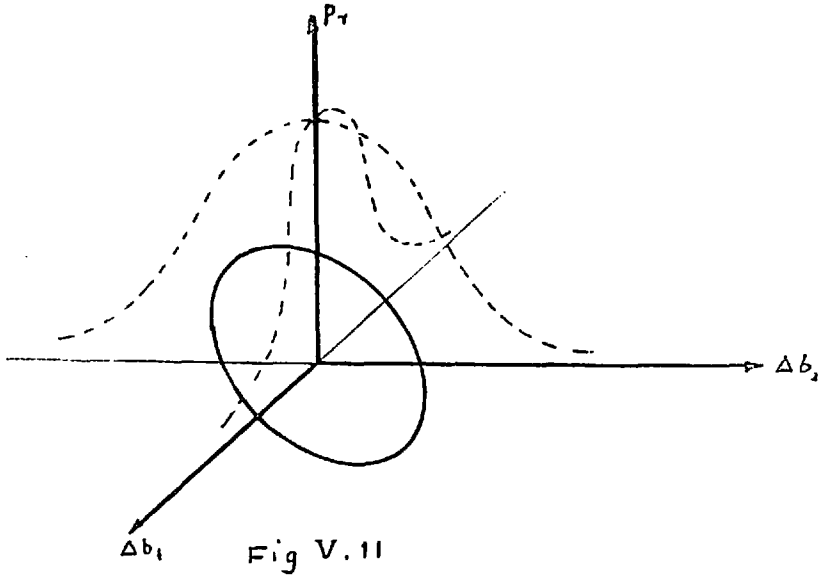


Fig V.11

Fig V.11 shows the sensitivity ellipse of a second order single-rate digital filter on the Δb_1 - Δb_2 plane. Associated with the ellipse are the arbitrary probability density functions $Pr(\Delta b_1)$ and $Pr(\Delta b_2)$

For any value of Δb_1, the probability that Δb_1 will fall within $d(\Delta b_1)$ is $Pr(\Delta b_1) . d(\Delta b_1)$. But for this value of Δb_1, the extreme values of Δb_2 are bounded by the ellipse in order that the movements of the poles are confined within a circle of radius σ. If these bounded values are denoted by $(\Delta b_2)'$ and $(\Delta b_2)''$ respectively, then the probability that Δb_2 will fall within this strip of the ellipse is $\int_{(\Delta b_2)'}^{(\Delta b_2)''} Pr(\Delta b_2) . d(\Delta b_2)$.

Now, Δb_1 and Δb_2 are independent of each other, so the total probability that Δb_1 and Δb_2 will fall within the ellipse is simply the product of the individual probabilities, i.e.

$$(Pr)_s = \iint_{\mathcal{R}_s} Pr(\Delta b_1) . Pr(\Delta b_2) . d(\Delta b_1) d(\Delta b_2) \qquad (V.90)$$

where $\mathcal{R}_s$ is the region enclosed by the ellipse.

Similarly, if $\Delta\beta_1$ and $\Delta\beta_2$ are the errors in the coefficients in a multi-rate time-invariant digital filter, then the total probability that $\Delta\beta_1$ and $\Delta\beta_2$ will fall within the sensitivity ellipse of the multi-rate filter is

$$(Pr)_m = \iint_{\mathcal{R}_m} Pr(\Delta\beta_1).Pr(\Delta\beta_2).\, d(\Delta\beta_1)d(\Delta\beta_2) \qquad (V.90)$$

By comparing the probabilities $(Pr)_s$ and $(Pr)_m$, the condition under which one filter is less sensitive to coefficient quantization than the other can be found.

The comparison depends considerably on the probability density functions $Pr(\Delta b_1)$ and $Pr(\Delta b_2)$, and similarly depends on $Pr(\Delta\beta_1)$ and $Pr(\Delta\beta_2)$. For a digital multiplier, if the word is rounded to an accuracy of E_o, it can be assumed that, in general, the round-off error is distributed evenly between $-\frac{E_o}{2}$ and $\frac{E_o}{2}$, i.e. if both b_1 and b_2 (or, in the case of time-invariant multirate filter, β_1 and β_2) are rounded off to the same accuracy E_o, their probability density functions will be as shown in fig V.12

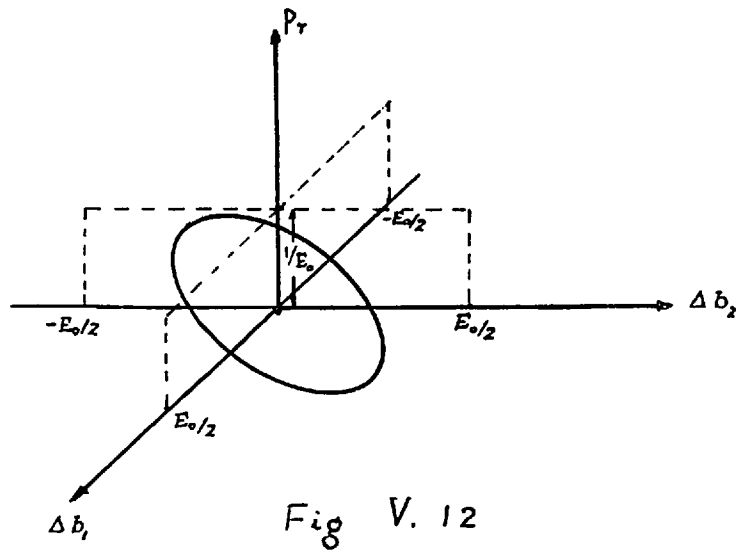


Fig V. 12

Now, if the radius, σ, of the circle to which the movements of the poles are confined is specified arbitrarily small, then the sensitivity ellipses of both the single-rate and multi-rate filters will lie within the probability density functions of Δb_1 and Δb_2 (or $\Delta\beta_1$ and $\Delta\beta_2$ for multirate filters). In other words, the ellipses will lie within the square of area E_o^2 centred at the origin of the Δb_1— Δb_2 (or $\Delta\beta_1$— $\Delta\beta_2$) plane. Hence for the single-rate digital filter, the total probability that Δb_1 and Δb_2 will fall within the ellipse is, from eqn (V.89), given by

$$(\Pr)_s = \iint_{\mathcal{R}_s} \frac{1}{E_o} \cdot \frac{1}{E_o} \, d(\Delta b_1) d(\Delta b_2)$$

$$= \frac{1}{E_o^2} \cdot \text{(area of the single-rate sensitivity ellipse)} \qquad \text{(V.91)}$$

Similarly, for the multi-rate digital filter, the total probability that $\Delta\beta_1$ and $\Delta\beta_2$ will fall within its sensitivity ellipse is given by

$$(\Pr)_m = \frac{1}{E_o^2} \iint_{\mathcal{R}_m} d(\Delta\beta_1) d(\Delta\beta_2)$$

$$= \frac{1}{E_o^2} \cdot \text{(area of the multi-rate sensitivity ellipse)} \qquad \text{(V.92)}$$

Hence, from eqn (V.91) and (V.92), the comparison of the probabilities $(\Pr)_s$ and $(\Pr)_m$ becomes the comparison between the areas of the respective sensitivity ellipses.

Now, for a general conic equation represented by

$$Ax^2 + 2Hxy + By^2 = C \qquad \text{(V.93)}$$

to be an ellipse on the $x - y$ plane, then {6}

$$AB - H^2 > 0$$

and the area of the ellipse is

$$Area = \frac{\pi C}{\sqrt{AB - H^2}} \tag{V.94}$$

Comparing equation (V.73) to equation (V.93), the area of the single-rate sensitivity ellipse is given by

$$(Area)_s = \pi\sigma^2 \Big/ \sqrt{\frac{1}{4} \cdot \frac{(1+\cot^2\theta)}{4r^2\sin^2\theta} - \frac{\cos^2\theta}{16r^2\sin^4\theta}}$$

$$= 4\pi\sigma^2 r|\sin\theta| \tag{V.95}$$

Similarly, the area of the multi-rate sensitivity ellipse is given by

$$(Area)_m = \frac{4\pi\sigma^2|\sin\phi|}{N^2\ \rho^{2N-2}} \tag{V.96}$$

The poles of the multi-rate filter will be less sensitive to coefficient quantization error if $(area)_m > (area)_s$, i.e. if

$$\frac{|\sin\phi|}{N^2\ \rho^{2N-2}} > r\ |\sin\theta| \tag{V.97}$$

or,

$$\frac{|\sin\phi|}{N^2\ \rho^{3(N-1)}} > |\sin N\phi| \tag{V.98}$$

where the fact that $r = \rho^N$ and $\theta = N\phi$ has been used.

The corresponding values of r and θ in eqn (V.98) for $N = 2$ and $N = 3$ have been plotted in fig (V.13 a and b). The shaded area on the z-plane represents the region in which the poles of a second-order time-invariant multirate digital filter are less sensitive to coefficient quantization errors than those of a single-rate filter.

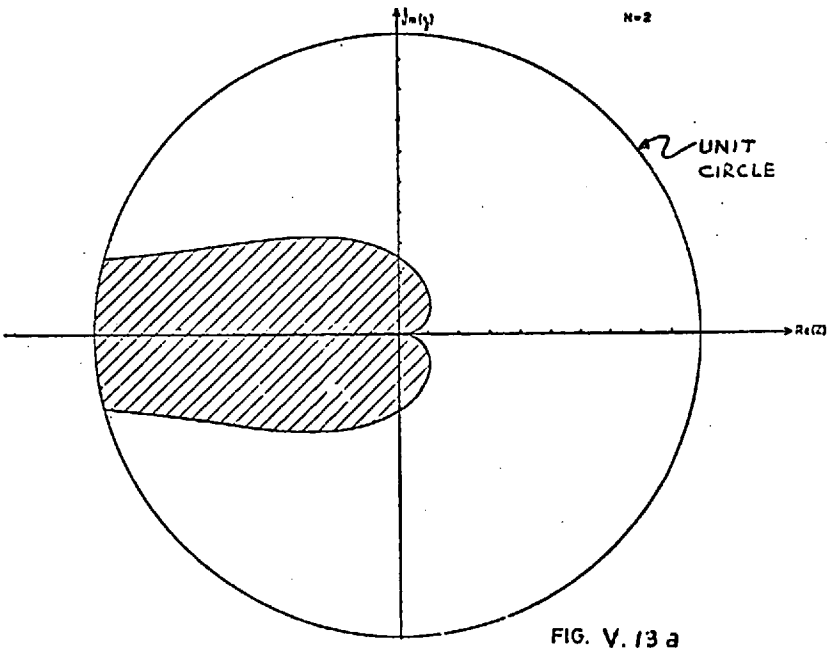


FIG. V. 13 a

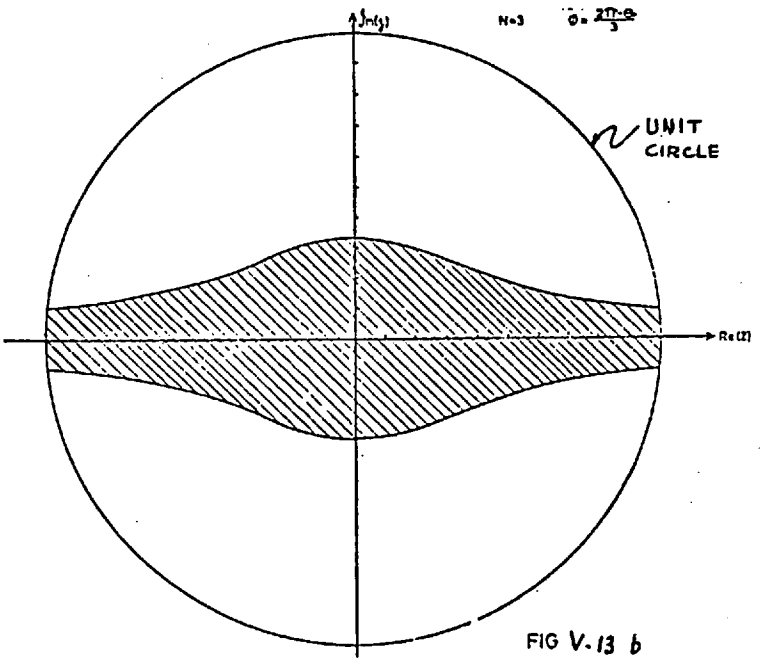


FIG V. 13 b

From the inequality of (V.97), it can be seen that the larger is the value of $|\sin\phi|$, the larger would be the area in which the multi-rate time-invariant filter is less sensitive to coefficient quantization error than the single-rate filter. But, as has been shown in section III.10, there are, in general, N different ways of chosing λ, i.e.

$$\lambda = \rho e^{j\phi} = r^{1/N} e^{j(\frac{\theta}{N} + \frac{2i\pi}{N})} \qquad (V.99)$$

where $i = 0, 1, 2, \ldots, N-1$

Hence, in designing a second order time-invariant digital filter to perform the function of a single-rate filter, it is best, in general, to chose from the N different solutions, the eigenvalue λ which has the largest value of $|\sin\phi|$.

V.7 Computer Simulation Results Comparing the Pole-Sensitivity of Single-rate Filters to that of Time-Invariant Multi-rate Filters

The analysis in the last section shows that if the specified second order transfer function has its poles situated inside certain regions on the z-plane, it is generally more advantageous to realize the transfer function as a multi-rate digital filter. These regions are given by the inequality of (V.98) and are plotted in fig V.13 a) and b) for $N = 2$ and $N = 3$.

A computer program has been written to simulate both the second order single-rate filter and the second order time-invariant multi-rate filters so that the effects of quantization of the filter coefficients in both cases can be observed and compared. According to a given transfer function, the computer program locates the position of the poles. It designs a single-rate digital filter according to the given transfer function. Then the

program examines the position of the poles on the z-plane and tests if it falls into one of the regions given by the inequality of (V.98) from which the number of shift sequences (N) within one sampling period is determined. Then, the multi-rate digital filter is designed by the same method as given in section III.10 so that it gives the same performance as specified by the transfer function. Again, the eigenvalue with the largest value of $|\sin \phi|$ is chosen for the design. Then the coefficients of these filters (both single- and multi-rate) so designed are rounded-off to the same finite accuracy. The responses of these filters with quantized coefficients are then plotted on the same diagram together with the responses of the ideal filters. The discrepancies between the ideal responses and the responses of the filters with quantized coefficients are examined.

The following are some examples extracted from the simulation:-

Example V.6

Design a single-rate and a multi-rate digital filter such that both filters have the following transfer function

$$F(z^{-1}) = \frac{0.8235}{1 - 1.78z^{-1} + 0.803125z^{-2}}$$

Each coefficient of these filters is rounded-off to an accuracy of two places after the decimal. Their performances are compared.

The single-rate digital filter so designed has the following coefficients:-

a_o = 0.8235 rounded-off to 0.82

b_1 = -1.78 rounded-off to - 1.78

b_2 = 0.803125 rounded-off to 0.80

It is found that the poles of the transfer function are

$$\Lambda = 0.89 \pm j0.105$$

and are within the region in which a triple-rate time-invariant filter is less sensitive to quantization errors. The desired triple-rate filter has the following coefficients.

α_o = - 6.6913456 rounded-off to -6.69

β_1 = 0.89802869 rounded-off to 0.90

β_2 = 0.92952494 rounded-off to 0.93

The step responses of these filters are shown in fig V.14 (a) and (b). It could be seen that the triple-rate filter gives a much closer response to the ideal performance than the single-rate filter.

STEP RESPONSES

FIG V. 14(a)

SINGLE-RATE FILTER

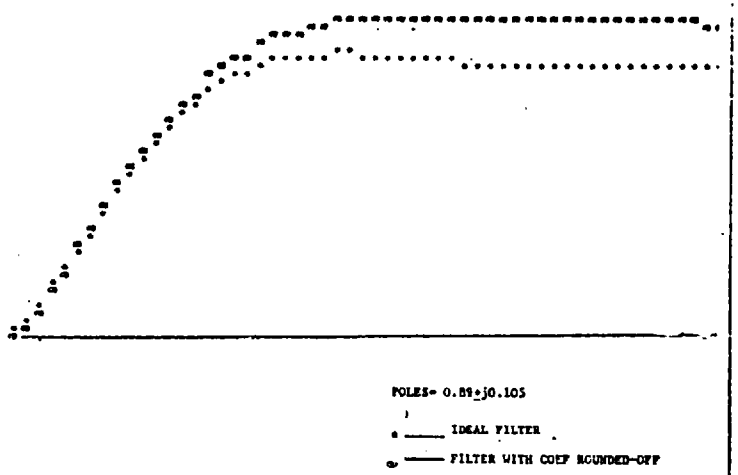
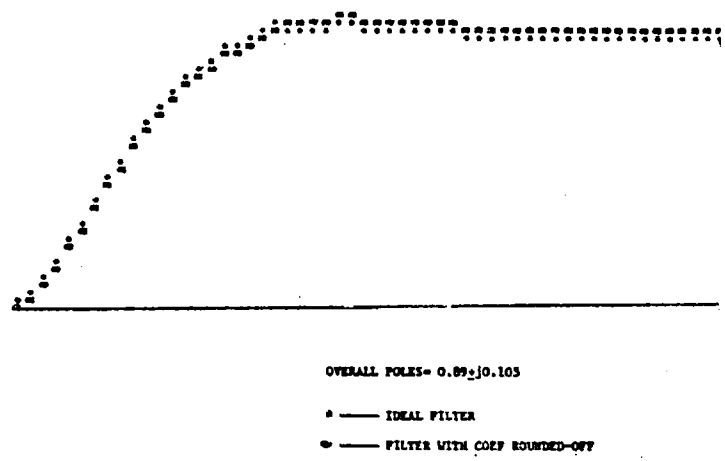


FIG V. 14 (b)

TRIPLE-RATE FILTER



Example V.7

Design a single-rate and a multi-rate digital filter such that both filters have the following transfer function

$$F(z^{-1}) = \frac{0.8235}{1 + 1.78 z^{-1} + 0.803125 z^{-2}}$$

Each coefficient of the filters is to be rounded-off to an accuracy of two places after the decimal. Their performances are to be compared.

The single-rate filter so designed has the following coefficients

a_0 = 0.8235 rounded-off to 0.82
b_1 = 1.78 rounded-off to 1.78
b_2 = 0.803125 rounded-off to 0.80

The poles of the transfer function are

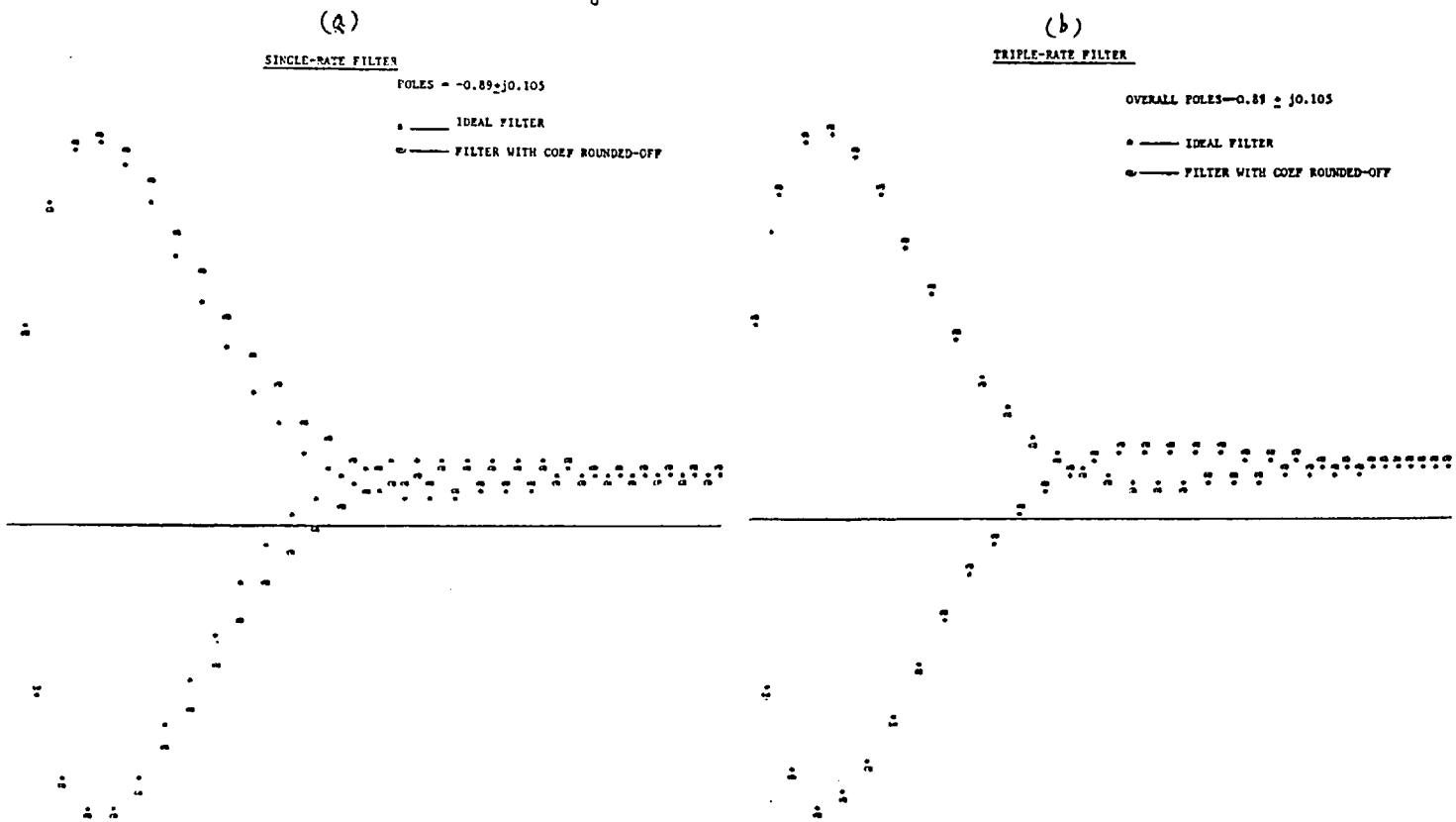
$$\Lambda = -0.89 \pm j\, 0.105$$

These poles are found to be in the region where both the double-rate or the triple-rate filter will be less sensitive than the single-rate filter to coefficient quantization errors. Hence either $N = 2$ or $N = 3$ will be a better design than $N = 1$. A triple-rate time-invariant filter is chosen, and has the following coefficients

α_0 = 6.3954292 rounded-off to 6.40
β_1 = - 1.0287316 rounded-off to - 1.03
β_2 = 0.9295494 rounded-off to 0.93

The step responses of these filters are shown in fig V.15 (a) and (b). Again it can be seen that the triple-rate filter has a response much closer to the ideal performance than the single-rate filter.

Fig V.15 STEP RESPONSES



Example V.8

Design a single-rate and a multi-rate (time-invariant) digital filter such that both filters have the following transfer function

$$F(z^{-1}) = \frac{0.8235}{1+1.78z^{-1}+0.8041z^{-2}}$$

The coefficients of the filters are rounded-off to two places after the decimal and their performances compared.

The single-rate filter has the following coefficients:-

$a_0 = 0.8235$	rounded-off to	0.82
$b_1 = 1.78$	rounded-off to	1.78
$b_2 = 0.8041$	rounded-off to	0.80

It is found that the poles of the transfer functions are

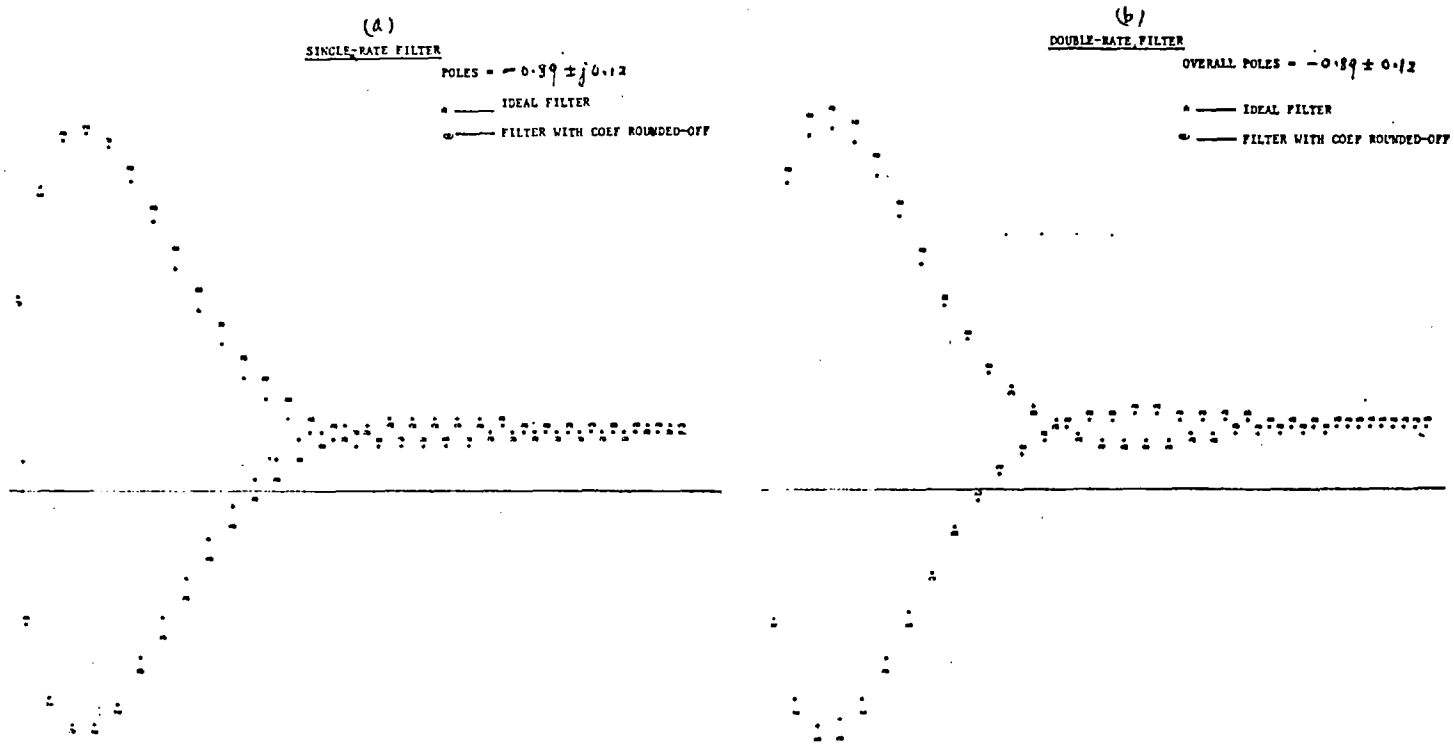
$$\Lambda = -0.89 \pm j\,0.12$$

The poles are situated just outside the region where the triple-rate filter will be less sensitive, but are situated well within the region in which the double-rate filter is less sensitive to coefficient quantization errors than the single-rate filter. So a time-invariant double-rate filter is chosen. It has the following coefficients.

$\alpha_0 = 7.7810954$	rounded-off to	7.78
$\beta_1 = -0.10583343$	rounded-off to	-0.11
$\beta_2 = 0.89560036$	rounded-off to	0.90

The step responses of these filters are shown in fig V.16 (a) and (b). Again it can be seen that the double-rate filter has a response closer to the ideal one than the single-rate filter.

Fig V.16 STEP RESPONSES



All the above examples confirm the theory put forward in section V.6, that if the poles lie inside the cross-hatched regions shown in fig V.13 (a) and (b), a double-rate or a triple-rate time-invariant filter will be less sensitive to coefficient quantization errors. However, as can be seen from the examples above, as the poles move closer to the boundaries of the cross-hatched regions, the differences between the performance of the quantized multi-rate filters and that of the quantized single-rate filters becomes less and less. Thus, the validity of the analysis in section V.6 is verified.

V.8 Résumé

A deterministic approach has been chosen to analyse the effect of coefficient quantization on the movements of the poles. Using this approach, it has been shown that in general, the pole movement of a multi-rate digital filter with periodically varying coefficients is more sensitive to coefficient quantization errors than that of a time-invariant multi-rate filter.

Examination of the effects of coefficient quantization errors leads to the concept of "sensitivity ellipse" which has been used as a criterion to compare the pole sensitivities of the single-rate and the time invariant multi-rate digital filters. The inequality of (V.97) states the condition under which the pole movement of a time-invariant digital filter is less sensitive than that of a single-rate digital filter. The validity of this condition has been confirmed by extensive computer simulations.

In the design of a digital filter, the errors introduced by quantization of multiplying coefficients remains a considerable problem. The performance of the conventional single-rate digital

filter is especially sensitive to the errors of coefficient quantization when its poles are in the vacinity of the real axis on the z-plane. With the introduction of the multi-rate digital filter, such sensitivity would be grossly reduced.

CHAPTER VI

MULTIPLICATION ROUND-OFF ERRORS IN A MULTI-RATE DIGITAL FILTER

VI.1 Introduction

This chapter is devoted to the most intricate manifestation of quantization errors, namely, errors caused by rounding off the computations used in the execution of the actual digital filter program. As mentioned in Chapter I, it is not necessary to compute the exact results of the effects of rounding off the multiplication products. In this chapter, two of the existing methods estimating the effects of rounding off the multiplication products are described.

The first [70] employs the state-space method and estimates the upper bound of the multiplication round-off errors. However, multiplication round-off errors would be intrinsically statistical if the input signal to the digital filter is sufficiently rich in the variation of its frequency contents, and the evaluation of an upper bound would then seem pessimistic. The second makes use of the spectral density of the round-off error and estimates the mean square error introduced by rounding off the multiplication products.

Since the analysis of multi-rate digital filters is much facilitated using the state-space approach, the two methods are combined [66] so that a statistical estimation of the effect of multiplication round-off errors is achieved using the state-space method. Such an estimation is applied to the case of multi-rate filters.

VI.2 Estimation of the Upper Bound for Multiplication Round-Off Errors in Digital Filters

The following method is due to Yakowitz and Parker {70}. In their paper, an upper bound is developed for the multiplication round-off errors in the state-variables of a digital filter. Here is an outline of their method.

Let an ideal filter be represented by the following dynamic equations:-

$$\begin{aligned} \mathbf{x}(n+1) &= \mathbf{A}\,\mathbf{x}(n) + \mathbf{B}\,\mathbf{u}(n) \\ \mathbf{y}(n) &= \mathbf{C}\,\mathbf{x}(n) + \mathbf{D}\mathbf{u}(n) \end{aligned} \tag{VI.1}$$

where $\mathbf{x}(n)$ = state variable vector at $t = nT$
$\mathbf{u}(n)$ = input vector at $t=nT$
$\mathbf{y}(n)$ = output vector at $t=nT$

$\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are constant matrices for a time-invariant digital filter.

Let the rounded product be denoted by $[.]_{E_o}$ where E_o is the quantization step; also let the erroneous values arising from these rounded values be denoted by dashes. Hence the actual filter with round-off errors can be represented by

$$\begin{aligned} \mathbf{x}'(n+1) &= \left[\mathbf{A}\,\mathbf{x}'(n)\right]_{E_o} + \left[\mathbf{B}\,\mathbf{u}(n)\right]_{E_o} \\ \mathbf{y}'(n) &= \left[\mathbf{C}\,\mathbf{x}'(n)\right]_{E_o} + \left[\mathbf{D}\,\mathbf{u}(n)\right]_{E_o} \end{aligned} \tag{VI.2}$$

It can be assumed that $\mathbf{x}'(0) = \mathbf{x}(0)$.

Define an error vector $\mathbf{e}(n)$ such that

$$\mathbf{e}(n) \triangleq \mathbf{A}\,\mathbf{x}'(n) - \left[\mathbf{A}\,\mathbf{x}'(n)\right]_{E_o} + \mathbf{B}\mathbf{u}(n) - \left[\mathbf{B}\,\mathbf{u}(n)\right]_{E_o} \tag{VI.3}$$

Also let $\Delta\mathbf{x}(n)$ be the vector denoting the errors in the state variables at $t = nT$, then

$$\begin{aligned}\Delta\mathbf{x}(n+1) &\triangleq \mathbf{x}(n+1) - \mathbf{x}'(n+1)\\ &= \mathbf{A}\mathbf{x}(n) + \mathbf{B}\mathbf{u}(n) - \left[\mathbf{A}\mathbf{x}'(n)\right]_{E_o} - \left[\mathbf{B}\,\mathbf{u}(n)\right]_{E_o}\\ &= \mathbf{A}\{\mathbf{x}(n) - \mathbf{x}'(n)\} + \mathbf{e}(n) \qquad \text{(VI.4)}\end{aligned}$$

i.e. $$\Delta\mathbf{x}(n+1) = \mathbf{A}\,\Delta\mathbf{x}(n) + \mathbf{e}(n) \tag{VI.5}$$

where $\mathbf{e}(n)$ is defined in equation(VI.3)

Similarly, for the output equation,

$$\mathbf{y}(n) = \mathbf{C}\,\mathbf{x}(n) + \mathbf{D}\,\mathbf{u}(n) \tag{VI.6}$$

and the actual output vector $\mathbf{y}'(n)$ with round-off errors is

$$\mathbf{y}'(n) = \left[\mathbf{C}\,\mathbf{x}'(n)\right]_{E_o} + \left[\mathbf{D}\,\mathbf{u}(n)\right]_{E_o} \tag{VI.7}$$

Again, an error vector $\varepsilon(n)$ can be defined s.t.

$$\varepsilon(n) = \mathbf{C}\,\mathbf{x}'(n) - \left[\mathbf{C}\,\mathbf{x}'(n)\right]_{E_o} + \mathbf{D}\,\mathbf{u}(n) - \left[\mathbf{D}\,\mathbf{u}(n)\right]_{E_o} \tag{VI.8}$$

and therefore,

$$\Delta\mathbf{y}(n) = \mathbf{C}\,\Delta\mathbf{x}(n) + \varepsilon(n) \tag{VI.9}$$

The solution of the equations (VI.5) and (VI.9) are, from sectionII.6 given by

$$\Delta \mathbf{x}(n) = \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1} \mathbf{e}(i) \tag{VI.10}$$

and

$$\Delta \mathbf{y}(n) = \mathbf{C} \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1} \mathbf{e}(i) + \boldsymbol{\varepsilon}(n) \tag{VI.11}$$

Now, let us examine the bounds of the error vector $\mathbf{e}(n)$ and $\boldsymbol{\varepsilon}(n)$. Firstly, let $\langle . \rangle$ denotes the matrix (or vector) determined by taking the magnitude of each element. We shall write $\langle \mathbf{A} \rangle < \langle \mathbf{B} \rangle$ if (i) the magnitude of each element of $\mathbf{A}$ is less than the magnitude of the corresponding element of $\mathbf{B}$, or if (ii) the magnitude of at least one of the elements of $\mathbf{A}$ are less than the magnitude of the corresponding elements of $\mathbf{B}$ while the rest of the elements of $\mathbf{A}$ and $\mathbf{B}$ are equal in magnitude correspondingly. Thus,

$$\left\langle \begin{bmatrix} -1 & 0 \\ 2 & 1 \end{bmatrix} \right\rangle < \left\langle \begin{bmatrix} 2 & 2 \\ 4 & 3 \end{bmatrix} \right\rangle$$

and

$$\left\langle \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \right\rangle < \left\langle \begin{bmatrix} 1 & 3 \\ 2 & 2 \end{bmatrix} \right\rangle$$

Extending the above definition further, $\langle \mathbf{A} \rangle \leq \langle \mathbf{B} \rangle$ is written if either of the above condition (i) and (ii) holds or if $\langle \mathbf{A} \rangle = \langle \mathbf{B} \rangle$, where the equality sign has the same meaning as is generally used in matrix equality.

Let μ_{Aj} be the number of non-zero and non-unity elements in the jth row of $\mathbf{A}$, and ν_{Bj} be the number of non-zero and non-unity elements in the jth row of $\mathbf{B}$. Hence, if the quantization step is E_o, the jth element of $\mathbf{e}(n)$ is bounded by

$$|e_j(n)| \leq (\mu_{Aj} + \nu_{Bj}) E_o/2 \tag{VI.12}$$

With the above definition in mind, the error vector for the state variables is bounded in magnitude by,

$$\langle \mathbf{e}(n) \rangle \leq \hat{\mathbf{e}} = \begin{bmatrix} \mu_{A1} + \nu_{B1} \\ \mu_{A2} + \nu_{B2} \\ \vdots \\ \mu_{AM} + \nu_{BM} \end{bmatrix} E_o/2 \qquad \text{(VI.13)}$$

where M is the order of the filter. Similarly if μ_{Cj} and ν_{Dj} are the number of non-zero and non-unity elements in the jth row of $\mathbf{C}$ and $\mathbf{D}$ respectively, then the error vector for the output is bounded in magnitude by,

$$\langle \boldsymbol{\varepsilon}(n) \rangle \leq \hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \mu_{C1} + \nu_{D1} \\ \mu_{C2} + \nu_{D2} \\ \vdots \\ \mu_{CK} + \nu_{DK} \end{bmatrix} E_o/2 \qquad \text{(VI.14)}$$

where K is the number of elements in the output vector.

Having assessed the bounds of $\langle \mathbf{e}(n) \rangle$ and $\langle \boldsymbol{\varepsilon}(n) \rangle$, the bounds for $\langle \Delta\mathbf{x}(n) \rangle$ and $\langle \Delta\mathbf{y}(n) \rangle$ can be estimated. Thus, from (VI.10),

$$\begin{aligned} \langle \Delta\mathbf{x}(n) \rangle &= \Big\langle \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1} \mathbf{e}(i) \Big\rangle \\ &\leq \sum_{i=1}^{n-1} \langle \mathbf{A}^{n-i-1} \rangle \, \langle \mathbf{e}(i) \rangle \qquad \text{Schwarz Inequality} \\ &\leq \sum_{i=1}^{n-1} \langle \mathbf{A}^{n-i-1} \rangle \, \hat{\mathbf{e}} \qquad \text{(VI.15)} \end{aligned}$$

The lower limit of the summation in the inequality of (VI.15) has been changed from 0 to 1 since $\mathbf{e}(0) = \mathbf{0}$. (VI.15) can be rewritten as

follows

$$\langle \Delta \mathbf{x}(n) \rangle \leq \sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \, \hat{\mathbf{e}} \qquad \text{(VI.16)}$$

The inequality of (VI.16) expresses an upper bound for the error in the state-variable vector due to the quantization of multiplication products. However, the summing of the infinite series in(VI.16) is awkward and a closed form for the expression is desirable.

In general, a closed form for the bound shown in the inequality of (VI.16) can be obtained in either of the following ways:-

(a) By the Schwarz Inequality,

$$\langle \mathbf{A}^k \rangle \leq \langle \mathbf{A} \rangle^k \qquad \text{(VI.17)}$$

Thus, (VI.16) can be written as

$$\langle \Delta \mathbf{x}(n) \rangle \leq \sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \, \hat{\mathbf{e}} \leq \sum_{k=0}^{\infty} \langle \mathbf{A} \rangle^k \, \hat{\mathbf{e}} \qquad \text{(VI.18)}$$

If $\langle \mathbf{A} \rangle$ is stable, i.e. if

$$\lim_{k \to \infty} \langle \mathbf{A} \rangle^k = \mathbf{0} \qquad \text{(VI.19)}$$

then, a closed form of (VI.18) can be obtained. Now,

$$\sum_{k=0}^{m} \langle \mathbf{A} \rangle^k = \mathbf{I} + \langle \mathbf{A} \rangle + \langle \mathbf{A} \rangle^2 + \ldots + \langle \mathbf{A} \rangle^m \qquad \text{(VI.20)}$$

Premultiplying both sides of (VI.20) by $(\mathbf{I} - \langle \mathbf{A} \rangle)$ and simplifying, one obtains,

$$(\mathbf{I} - \langle \mathbf{A} \rangle) \sum_{k=0}^{m} \langle \mathbf{A} \rangle^k = \mathbf{I} - \langle \mathbf{A} \rangle^{m+1} \qquad \text{(VI.21)}$$

Hence,

$$\sum_{k=0}^{\infty} \langle \mathbf{A} \rangle^k = \lim_{m \to \infty} (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} (\mathbf{I} - \langle \mathbf{A} \rangle^{m+1}) = (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \qquad \text{(VI.22)}$$

and thus the bound for the state-variable error vector in (VI.16) is reduced to,

$$\langle \Delta \mathbf{x}(n) \rangle = \sum_{k=0}^{\infty} \langle \mathbf{A} \rangle^k \hat{\mathbf{e}} = (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \hat{\mathbf{e}} \qquad \text{(VI.23)}$$

Substituting this closed form expression in (VI.9), the output error is bounded in magnitude by

$$\langle \Delta \mathbf{y}(n) \rangle \leq \langle \mathbf{C} \rangle \langle \Delta \mathbf{x}(n) \rangle + \epsilon(n)$$

$$\leq \langle \mathbf{C} \rangle (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \hat{\mathbf{e}} + \hat{\epsilon} \qquad \text{(VI.24)}$$

where $\hat{\epsilon}$ is defined in (VI.14).

(b) The second method to express the bound of (VI.16) in a closed form involves the use of the Jordan canonic form of the matrix $\mathbf{A}$ (section II.7). Let $\mathbf{P}$ be the generalized eigenvector matrix of $\mathbf{A}$ and $\mathbf{J}$ be the Jordan canonic form of $\mathbf{A}$, then

$$\sum_{k=0}^{\infty} \mathbf{A}^k = \mathbf{P} \left(\sum_{k=0}^{\infty} \mathbf{J}^k \right) \mathbf{P}^{-1} \qquad \text{(VI.25)}$$

Hence we can write,

$$\sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \leq \sum_{k=0}^{\infty} \langle \mathbf{P} \rangle \langle \mathbf{J}^k \rangle \langle \mathbf{P}^{-1} \rangle$$

$$= \langle \mathbf{P} \rangle \left(\sum_{k=0}^{\infty} \langle \mathbf{J}^k \rangle \right) \langle \mathbf{P}^{-1} \rangle \qquad \text{(VI.26)}$$

It should be noted that $\langle \mathbf{J}^k \rangle = \langle \mathbf{J} \rangle^k$ since $\mathbf{J}$ is block diaganal, i.e.

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & \bigcirc \\ & \mathbf{J}_2 & \\ \bigcirc & & \ddots \\ & & & \mathbf{J}_m \end{bmatrix} \qquad \text{(VI.27)}$$

where $\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_m$ are the Jordan blocks constituting $\mathbf{J}$. If Λ_i is the eigenvalue of the ith Jordan block $\mathbf{J}_i$ and Λ_i has multiplicity l, then, from section II.7,

$$\mathbf{f}(\langle \mathbf{J}_i \rangle) = \sum_{k=0}^{\infty} \langle \mathbf{J}_i^k \rangle = \begin{bmatrix} f(|\Lambda|) & f'(|\Lambda|) & f''(|\Lambda|)/2! & \ldots & f^{(l-1)}(|\Lambda|)/(l-1)! \\ 0 & f(|\Lambda|) & f'(|\Lambda|) & \ldots & f^{(l-2)}(|\Lambda|)/(l-2)! \\ 0 & 0 & f(|\Lambda|) & \ldots & f^{(l-3)}(|\Lambda|)/(l-3)! \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \ldots & f(|\Lambda|) \end{bmatrix}_{\Lambda=\Lambda_i}$$

(VI.28)

For a stable filter, $|\Lambda_i| < 1$ and

$$f(|\Lambda|)\Big|_{\Lambda=\Lambda_i} = \sum_{k=0}^{\infty} |\Lambda|^k \Big|_{\Lambda=\Lambda_i} = (1 - |\Lambda_i|)^{-1} \qquad \text{(VI.29)}$$

Hence, (VI.28) can be written as

$$\mathbf{f}(\langle \mathbf{J}_i \rangle) = \sum_{k=0}^{\infty} \langle \mathbf{J}_i^k \rangle = \begin{bmatrix} (1-|\Lambda_i|)^{-1} & (1-|\Lambda_i|)^{-2} & \ldots & (1-|\Lambda_i|)^{-l} \\ 0 & (1-|\Lambda_i|)^{-1} & \ldots & (1-|\Lambda_i|)^{-l+1} \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \ldots & (1-|\Lambda_i|)^{-1} \end{bmatrix}$$

(VI.30)

and thus for the Jordan canonic form of $\mathbf{A}$,

$$\mathbf{f}(\langle \mathbf{J} \rangle) = \sum_{k=0}^{\infty} \langle \mathbf{J}^k \rangle = \begin{bmatrix} \mathbf{f}(\langle \mathbf{J}_1 \rangle) & & & \\ & \mathbf{f}(\langle \mathbf{J}_2 \rangle) & & \\ & & \ddots & \\ & & & \mathbf{f}(\langle \mathbf{J}_m \rangle) \end{bmatrix} \qquad \text{(VI.31)}$$

where $\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_m$ are the m Jordan blocks of $\mathbf{A}$ having $\Lambda_1, \Lambda_2, \ldots, \Lambda_m$ as their respective eigenvalues. Substituting (VI.31) into (VI.26), the bound for the state-variable error vector becomes,

$$\begin{aligned} \langle \Delta \mathbf{x}(n) \rangle &\leqq \sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \quad \hat{\mathbf{e}} \\ &\leqq \langle \mathbf{P} \rangle \sum_{k=0}^{\infty} \langle \mathbf{J}^k \rangle \langle \mathbf{P}^{-1} \rangle \hat{\mathbf{e}} \\ &= \langle \mathbf{P} \rangle \mathbf{f}(\langle \mathbf{J} \rangle) \langle \mathbf{P}^{-1} \rangle \hat{\mathbf{e}} \end{aligned} \qquad \text{(VI.32)}$$

Eqn(VI.32) represents another closed form expression for the bound to the state-variable error vector due to multiplication round-off. Again, substituting this expression into (VI.9), the output error is bounded in magnitude by

$$\begin{aligned} \langle \Delta \mathbf{y}(n) \rangle &\leqq \langle \mathbf{C} \rangle \langle \Delta \mathbf{x}(n) \rangle + \varepsilon(n) \\ &\leqq \langle \mathbf{C} \rangle \langle \mathbf{P} \rangle \mathbf{f}(\langle \mathbf{J} \rangle) \langle \mathbf{P}^{-1} \rangle \hat{\mathbf{e}} + \hat{\varepsilon} \end{aligned} \qquad \text{(VI.33)}$$

The two pairs of inequalities (VI.23), (VI.24) and (VI.32), (VI.33) represent valid closed forms of the bounds on the state-variable error vector and the output error vector when the filter is stable. (VI.23) and (VI.24) further necessitate that $\langle \mathbf{A} \rangle$ is stable, i.e. $\lim_{k \to \infty} \langle \mathbf{A} \rangle^k = \mathbf{0}$. Whichever is the tighter bound depends on the form of $\mathbf{A}$ as the next section will show.

VI.3 Remarks on the Evaluation of the Upper Error Bound Using State-Space Approach

The previous section has shown that the bound for the state-variable error vector is given by

$$\langle \Delta \mathbf{x}(n) \rangle \leq \sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \, \hat{\mathbf{e}} \qquad \text{(VI.34)}$$

Two closed forms to evaluate this bound have been developed. Here, the choice of these two forms is discussed:-

It has been mentioned in the previous section that the bound of (VI.23) is not valid unless $\sum_{k=0}^{\infty} \langle \mathbf{A} \rangle^k$ is convergent. The following example may serve to substantiate this statement.

Example VI.1

Let $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.8 & 1.4 \end{bmatrix}$,

then the eigenvalues of $\mathbf{A}$ are given by

$$\Lambda_{1\atop 2} = 0.7 \pm j0.5568$$

i.e. $$|\Lambda| = 0.8$$

Hence $\mathbf{A}$ is stable, and thus $\sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle$ is convergent.

However,

$$\langle \mathbf{A} \rangle = \begin{bmatrix} 0 & 1 \\ 0.8 & 1.4 \end{bmatrix}$$

and the eigenvalues of $\langle \mathbf{A} \rangle$ are,

$$\Lambda'_{1\atop 2} = 0.7 \pm 1.10198$$

Now, $$\lim_{k \to \infty} \langle \mathbf{A} \rangle^k = \mathbf{P} \begin{bmatrix} (1.802)^k & 0 \\ 0 & (0.402)^k \end{bmatrix} \mathbf{P}^{-1} \longrightarrow \infty$$

which means that $\langle \mathbf{A} \rangle$ is not stable any more, and the bound given by (VI.23) is thus not valid.

Now, if $\langle \mathbf{A} \rangle$ is stable, then the bound (VI.23) is valid. Consider the case when $\mathbf{A} \geqq \mathbf{0}$, i.e. each element of $\mathbf{A}$ is real and positive, then,

$$\langle \mathbf{A}^k \rangle = \langle \mathbf{A} \rangle^k \tag{VI.35}$$

But, $$\langle \mathbf{A}^k \rangle = \langle \mathbf{P}\, \mathbf{J}^k\, \mathbf{P}^{-1} \rangle \leqq \langle \mathbf{P} \rangle \langle \mathbf{J}^k \rangle \langle \mathbf{P}^{-1} \rangle \tag{VI.36}$$

therefore $$\langle \mathbf{A} \rangle^k \leqq \langle \mathbf{P} \rangle \langle \mathbf{J}^k \rangle \langle \mathbf{P}^{-1} \rangle \tag{VI.37}$$

$$\Longrightarrow \quad \sum_{k=0}^{\infty} \langle \mathbf{A} \rangle^k \leqq \langle \mathbf{P} \rangle \sum_{k=0}^{\infty} \langle \mathbf{J}^k \rangle \langle \mathbf{P}^{-1} \rangle \tag{VI.38}$$

Thus for a convergent series,

$$(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \leqq \langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle \tag{VI.39}$$

and (VI.23) is a tighter bound than (VI.32) for $\mathbf{A} \geqq \mathbf{0}$.

On the other hand, consider the case when $\mathbf{A} \leqq \mathbf{0}$, i.e. each element of $\mathbf{A}$ is real and negative. Let

$$\mathbf{A} = -\langle \mathbf{A} \rangle \tag{VI.40}$$

then, $$\langle \mathbf{A}^k \rangle = \langle (-1)^k \langle \mathbf{A} \rangle^k \rangle = \langle \mathbf{A} \rangle^k \tag{VI.41}$$

Hence, follwing the previous argument

$$(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \leqq \langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle \quad \text{for } \mathbf{A} \leqq \mathbf{0} \tag{VI.42}$$

From the inequalities of (VI.39) and (VI.42), it can be seen that for $\mathbf{A} \lneqq \mathbf{0}$, the closed form of (VI.23) is a tighter bound than that of (VI.32). Similarly, for $\mathbf{A} \lneqq \mathbf{0}$, the closed form of the output error vector shown in (VI.24) is tighter than that shown in (VI.33). The following example may help to illustrate this point. (It should be

noted that if $\langle \mathbf{A} \rangle$ is stable, $(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \geq \mathbf{0}$ since this is the limit of the sum of an infinite series of positive matrices).

Example VI.2

Let $$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0.16 & 0.6 \end{bmatrix}$$

then,
$$(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} = \begin{bmatrix} 1.667 & 4.1667 \\ 0.667 & 4.1667 \end{bmatrix}$$

Now the eigenvalues of $\mathbf{A}$ are given by

$$\Lambda_{\substack{1 \\ 2}} = \begin{cases} -0.2 \\ 0.8 \end{cases}$$

Hence $$\mathbf{J} = \begin{bmatrix} -0.2 & 0 \\ 0 & 0.8 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ -0.2 & 0.8 \end{bmatrix}$$

$$\mathbf{P}^{-1} = \begin{bmatrix} 0.8 & -1 \\ 0.2 & 1 \end{bmatrix}$$

$$\langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle = \begin{bmatrix} 2 & 6.25 \\ 1 & 4.25 \end{bmatrix} > (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1}$$

Unfortunately, no clear-cut conclusion of the type shown in (VI.39) and (VI.42) can be drawn when $\mathbf{A}$ is neither positive nor negative, i.e. when some elements of $\mathbf{A}$ are positive while some are negative. Since we are mainly interested in 2nd order filters in the direct canonic form, attention is focused on such a configuration. For a second order filter realized in the direct canonic form, the state matrix $\mathbf{A}$ is of the form

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix} \qquad \text{(VI.43)}$$

and if Λ_1 and Λ_2 are the two distinct eigenvalues (may be complex) of **A**, then the Jordan canonic form of **A** is

$$\mathbf{J} = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \tag{VI.44}$$

and its eigenvector matrix **P** is given by

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ \Lambda_1 & \Lambda_2 \end{bmatrix} \tag{VI.45}$$

For such a configuration, the maximun error vector is given by

$$\hat{\mathbf{e}} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \frac{E_o}{2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} E_o \tag{VI.46}$$

where E_o is the quantization step. This is because there are two non-zero and non-unity member in the second row of **A**.

When **A** is neither positive nor negative, even if $\langle \mathbf{A} \rangle$ is stable so that a closed form $(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1}$ can be obtained for the bound, there is, in general, no definite conclusion that can be drawn for the comparison of the two closed forms. It has been found that sometimes one closed form is tighter, sometimes the other is tighter, while in some cases the two forms are not comparable (i.e. some elements of the matrix obtained from one closed form are smaller than those of the other closed form, whereas the other elements are larger), as the following examples show.

Example VI.3

If $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0.16 & -0.6 \end{bmatrix}$, then $\Lambda_{1,2} = \begin{cases} 0.2 \\ -0.8 \end{cases}$

Thus, $(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} = \begin{bmatrix} 1.667 & 4.1667 \\ 0.667 & 4.1667 \end{bmatrix}$,

hence, $\langle \Delta \mathbf{x}(n) \rangle \overset{\leq}{=} (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \cdot \hat{\mathbf{e}} = \begin{bmatrix} 4.1667 \\ 4.1667 \end{bmatrix} E_o$ from(VI.23)

On the other hand,

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 0.2 & -0.8 \end{bmatrix},$$

$$\mathbf{J} = \begin{bmatrix} 0.2 & 0 \\ 0 & -0.8 \end{bmatrix}$$

and $\langle \mathbf{P} \rangle \ (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle = \begin{bmatrix} 2 & 6.25 \\ 1 & 4.25 \end{bmatrix}$

which means, from (VI.32)

$$\langle \Delta \mathbf{x}(n) \rangle \overset{\leq}{=} \langle \mathbf{P} \rangle \ (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle \ \hat{\mathbf{e}} = \begin{bmatrix} 6.25 \\ 4.25 \end{bmatrix} E_o$$

Hence, the closed form of $(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1}$ is tighter in this case.

Example VI.4

If $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.15 & 0.8 \end{bmatrix}$, then $\Lambda_{\frac{1}{2}} = \begin{cases} 0.5 \\ 0.3 \end{cases}$

In this case $\langle \mathbf{A} \rangle$ is stable, thus both closed forms are valid. Now,

$$(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} = \begin{bmatrix} 4 & 20 \\ 3 & 20 \end{bmatrix}$$

$$\Longrightarrow \quad \langle \Delta \mathbf{x}(n) \rangle \overset{\leq}{=} (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \hat{\mathbf{e}} = \begin{bmatrix} 20 \\ 20 \end{bmatrix} E_o$$

On the other hand,

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ 0.5 & 0.3 \end{bmatrix}$$

$$\mathbf{J} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.3 \end{bmatrix},$$

$$\langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle = \begin{bmatrix} 6.5725 & 17.145 \\ 4 & 10 \end{bmatrix}$$

$$\Rightarrow \langle \Delta \mathbf{x}(n) \rangle \leqq \langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle \hat{\mathbf{e}} = \begin{bmatrix} 17.145 \\ 10 \end{bmatrix} E_o$$

Hence the bound obtained from the closed form of $\langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle$ is tighter.

Example VI.5

If $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.9 & 0.05 \end{bmatrix}$, then $\Lambda_{\substack{1 \\ 2}} = 0.025 \pm j0.9484$

and $\mathbf{A}$ is stable since $|\Lambda| = 0.9$

$\langle \mathbf{A} \rangle$ is also stable in this case since both its eigenvalues are less than unity. Now,

$$(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} = \begin{bmatrix} 19 & 20 \\ 18 & 20 \end{bmatrix}$$

i.e. $\langle \Delta \mathbf{x}(n) \rangle \leqq (\mathbf{I} - \langle \mathbf{A} \rangle)^{-1} \hat{\cdot} = \begin{bmatrix} 20 \\ 20 \end{bmatrix} E_o$

On the other hand,

$$\mathbf{P} = \begin{bmatrix} 1 & 1 \\ \Lambda_1 & \Lambda_2 \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$$

$$\langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle = \begin{bmatrix} 19.494 & 20.548 \\ 18.493 & 19.494 \end{bmatrix}$$

i.e. $\langle \Delta \mathbf{x}(n) \rangle \leqq \langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle \hat{\mathbf{e}} = \begin{bmatrix} 20.548 \\ 19.494 \end{bmatrix} E_o$

In this case, the bound obtained from the closed form of $(\mathbf{I} - \langle \mathbf{A} \rangle)^{-1}$

is not comparable to that obtained from $\langle \mathbf{P} \rangle (\mathbf{I} - \langle \mathbf{J} \rangle)^{-1} \langle \mathbf{P}^{-1} \rangle$.

In general, there is no rule as to which ever closed form should be used. Yakowitz and Parker {70} have developed expressions for the bounds of multiplication round-off errors in the state-variables for a second order filter realized in the direct canonic configuration. However, in their development of those expressions, there is no direct application of either of the two closed forms discussed above. Instead, the bound was evaluated direct from (VI.16) as follows:-

From (VI.16)

$$\begin{aligned}
\langle \Delta \mathbf{x}(n) \rangle &\leq \sum_{k=0}^{\infty} \langle \mathbf{A}^k \rangle \, \hat{\mathbf{e}} \\
&= \sum_{k=0}^{\infty} \langle \mathbf{P} \, \mathbf{J}^k \, \mathbf{P}^{-1} \rangle \begin{bmatrix} 0 \\ 1 \end{bmatrix} E_o \\
&= \sum_{k=0}^{\infty} \frac{E_o}{|\Lambda_2 - \Lambda_1|} \left\langle \begin{bmatrix} 1 & 1 \\ \Lambda_1 & \Lambda_2 \end{bmatrix} \begin{bmatrix} \Lambda_1^k & 0 \\ 0 & \Lambda_2^k \end{bmatrix} \begin{bmatrix} \Lambda_2 & -1 \\ -\Lambda_1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\rangle \\
&= \frac{E_o}{|\Lambda_2 - \Lambda_1|} \sum_{k=0}^{\infty} \left\langle \begin{bmatrix} (\Lambda_2^k - \Lambda_1^k) \\ (\Lambda_2^{k+1} - \Lambda_1^{k+1}) \end{bmatrix} \right\rangle \\
&= \frac{E_o}{|\Lambda_2 - \Lambda_1|} \sum_{k=1}^{\infty} |\Lambda_2^k - \Lambda_1^k| \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\end{aligned} \tag{VI.47}$$

Now,

$$|\Lambda_2^k - \Lambda_1^k| = |\Lambda_2 - \Lambda_1| \sum_{k=1}^{\infty} \left| \sum_{i=0}^{k-1} \Lambda_2^{k-1-i} \, \Lambda_1^i \right| \tag{VI.48}$$

so that (VI.47) becomes

$$\langle \Delta \mathbf{x}(n) \rangle \leq \sum_{k=0}^{\infty} \left| \sum_{i=0}^{k-1} \Lambda_2^{k-i-1} \, \Lambda_1^i \right| \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{VI.49}$$

When Λ_1 and Λ_2 are both real and with the same sign, i.e. $b_1^2 > 4b_2 > 0$, it follows that

$$\sum_{k=1}^{\infty}\left|\sum_{i=0}^{k-1}\Lambda_2^{k-1-i}\,\Lambda_1^{i}\right| = \sum_{i=0}^{\infty}|\Lambda_1|^{i}\sum_{k=0}^{\infty}|\Lambda_2|^{k}$$
$$= 1/(1 - |\Lambda_1|)(1 - |\Lambda_2|)$$
$$= 1/(1 - |b_1| + b_2) \qquad \text{(VI.50)}$$

Hence, for this case,

$$\langle \Delta \mathbf{x}(n) \rangle \leq \frac{E_o}{(1 - |b_1| + b_2)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{(VI.51)}$$

When Λ_1 and Λ_2 are real with opposite signs ($b_2 < 0$), then let $\Lambda_1 = |\Lambda_1|$ and $\Lambda_2 = -|\Lambda_2|$ where $|\Lambda_2| > |\Lambda_1|$, then,

$$\sum_{k=1}^{\infty}\left|\sum_{i=0}^{\infty}\Lambda_2^{k-1-i}\Lambda_1^{i}\right| = \sum_{i=0}^{\infty}(-1)^{i}|\Lambda_2|^{i}\sum_{k=0}^{\infty}|\Lambda_1|^{k}$$
$$= 1/(1 + |\Lambda_2|)(1 - |\Lambda_1|)$$
$$= 1/(1 - |b_1| + b_2) \qquad \text{(VI.52)}$$

and the bound for the state-variable vector is again

$$\langle \Delta \mathbf{x}(n) \rangle \leq \frac{E_o}{(1 - |b_1| + b_2)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{(VI.53)}$$

For complex poles, $\Lambda_2 = \Lambda_1^*$, i.e. $|\Lambda_1| = |\Lambda_2| = \sqrt{b_2}$, then

$$\sum_{k=1}^{\infty}\left|\sum_{i=0}^{k-1}\Lambda_2^{k-1-i}\Lambda_1^{i}\right| < \sum_{k=0}^{\infty}(\sqrt{b_2})^{k}\sum_{k=0}^{\infty}(\sqrt{b_2})^{k} = (1 - \sqrt{b_2})^{-2} \qquad \text{(VI.54)}$$

thus,

$$\langle \Delta \mathbf{x}(n) \rangle < \frac{E_o}{(1 - \sqrt{b_2})^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{(VI.55)}$$

However, for complex poles, the bound can be further lowered as follows. Let $\Lambda_1 = re^{j\theta}$, $\Lambda_2 = re^{-j\theta}$. From (VI.49), it follows that

$$\langle \Delta x(n) \rangle \leq \frac{E_o}{|\Lambda_2 - \Lambda_1|} \sum_{k=1}^{\infty} r^k \left| (e^{jk\theta} - e^{-jk\theta}) \right| \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{2E_o}{|\Lambda_2 - \Lambda_1|} \sum_{k=1}^{\infty} r^k |\sin k\theta| \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\leq \frac{2E_o}{|\Lambda_2 - \Lambda_1|} \frac{r}{(1 - r)} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \frac{2\sqrt{b_2}\ E_o}{(1 - \sqrt{b_2})\sqrt{(4b_2 - b_1^2)}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{(VI.56)}$$

This way of evaluation is in general superior to either of the two closed forms developed in the previous section since the limit is not imposed on the evaluation until the end and thus saving the estimation of the limits in the intermediate stages.

VI.4 Statistical Estimation of the Multiplication RoundOff Errors in a Digital Filter [36]

If the input signal is sufficiently rich in frequency content, the variation of its multiplication round-off errors can reasonably be assumed to be random. The probability density function of the rounding error is then uniformly distributed (see section IV.1 and fig IV.2); moreover, the error $e(n)$ at any sampling time will be statistically uncorrelated to $e(m)$, the error at any other sampling instant. Under such conditions, the evaluation of an upper bound seems pessimistic. A more realistic approach is

thus to estimate the round-off error statistically.

Since the multiplication roundoff errors occur for each iteration of the difference equation, the effect is that of a set of random noise samples superimposed on the signal; in this sense, it is similar to A/D conversion noise. However, the precise location at which this noise is injected in the digital filter depends on the particular configuration of the filter. Since we are mainly interested in second order filters realized in the direct canonic form, the analysis here focuses its attention on such filters(fig VI.1)

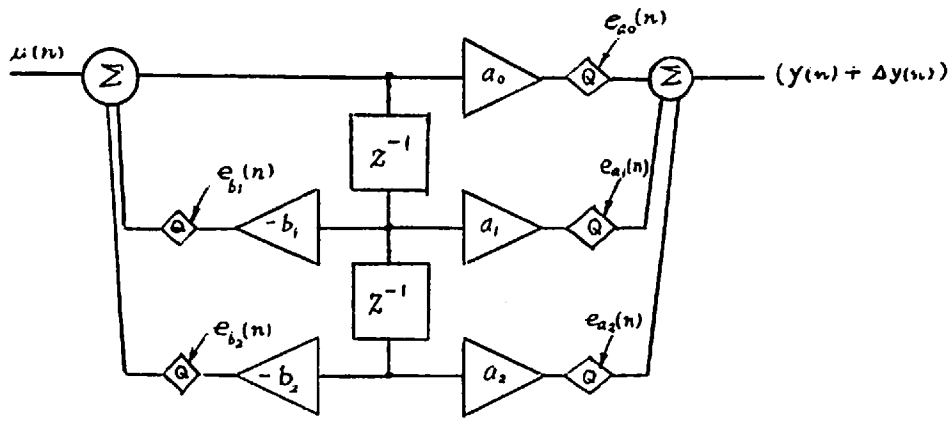


Fig VI.1 A Second Order Digital Filter with Multiplication Quantization

Fig VI.1 represents a second order digital filter realized in the direct canonic form. Each of the multipliers is followed by a quantizer which is effectively a noise generator. The effect of these noise generators can be combined to form two noise sources, one at the input adder and the other at the output adder (fig VI.2) where the noise source at the input $E_b(n)$ is the sum of the noise sources from the feedback multipliers and the noise source at the output adder $E_a(n)$ is the sum of the noise sources from the feed-forward multipliers. Since these two noise sources are both independent of each other, their effects can be considered separately.

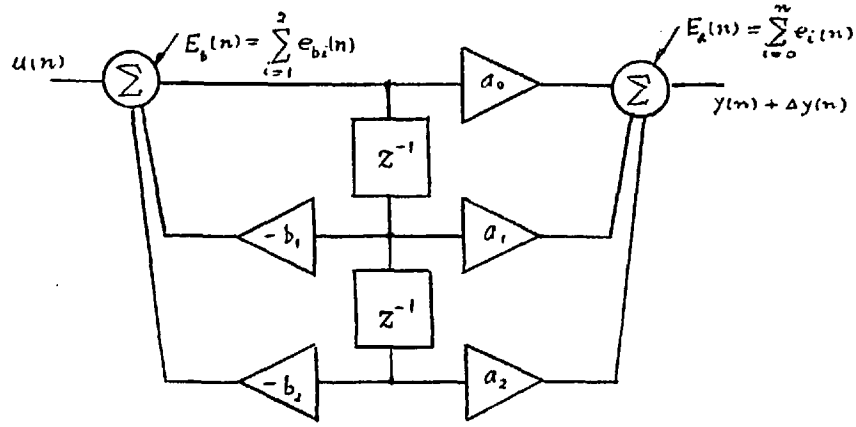


Fig VI.2 Equivalent Noise Model

Let $\Delta y_a(n)$ be the output of the filter due to the effect of the noise source $E_a(n)$. Now since $E_a(n)$ is the sum of three independent noise sources each with variance $E_o^2/12$, where E_o is the quantization step, then the variance of $E_a(n)$ is given by

$$\sigma^2_{Ea} = \sum_{i=0}^{2} \sigma^2_{e_{ai}}(n) = 3E_o^2/12 \qquad \text{(VI.57)}$$

Since the effect of this noise is merely adding a noise to the output, the variance of $\Delta y_a(n)$ is given by,

$$\sigma^2_{\Delta y_a} = 3E_o^2/12 \qquad \text{(VI.58)}$$

Now, let $\Delta y_b(n)$ be the output of the filter due to the effect of the noise source $E_b(n)$ at the input. Since this noise source is a combination of two independent noise sources each of variance $E_o^2/12$, then the variance of this combined noise source is simply $2E_o^2/12$. The effect of this noise source is similar to that described in section IV.3. Thus the variance of $y_{eb}(n)$ is given by

$$\sigma^2_{\Delta y_b} = 2E_o^2/12 \sum_{i=0}^{N} g^2(i) \qquad \text{(VI.59)}$$

where $g(i)$ is the impulse response of the filter. Given that a

steady state is reached, the variance of $\Delta y_b(n)$ is given by

$$\sigma^2_{\Delta y_b} = 2E_o^2/12 \sum_{i=0}^{\infty} g^2(i)$$

$$= 2E_o^2/12 \left\{ \frac{1}{2\pi j} \oint_\Gamma G(z)G(\frac{1}{z})z^{-1}\, dz \right\} \qquad \text{(VI.60)}$$

Hence the variance of the total noise in the output due to multiplication round-off error is given by

$$\sigma^2_{\Delta y} = \frac{E_o}{12} \left\{ \frac{2}{2\pi j} \oint_\Gamma G(z)G(z^{-1})z^{-1}\, dz + 3 \right\} \qquad \text{(VI.61)}$$

The same principle can be applied to digital filters with various configurations. It is obvious that different configurations will give rise to different values of multiplication quantization errors. Extension of this basic principle of analysis to other configurations can be found in other works of reference {1},{22}, {36}.

VI.5 State-Space Approach to the Statistical Estimation of Multiplication Errors in a Digital Filter {66}

As mentioned in the previous section, a statistical estimation of the multiplication round-off error is more realistic than the evaluation of an upper bound, thus the view taken here is again statistical. However, since the state-space method is generally very much more convenient to use in the analysis of multi-rate digital filters, the statistical estimation of multiplication errors here employs the state-space method.

Following the argument in section VI.2, the error in the state-variable vector is given by

$$\Delta \mathbf{x}(n) = \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1}\, \mathbf{e}(i) \tag{VI.62}$$

and that the error in the output vector is given by

$$\Delta \mathbf{y}(n) = \mathbf{C} \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1}\, \mathbf{e}(i) + \boldsymbol{\varepsilon}(n) \tag{VI.63}$$

The variance of the error at each step of quantization is given by

$$\sigma^2 = E_0^2/12 \tag{VI.64}$$

where E_o is the quantization step. Hence, the variance of the vector $\mathbf{e}$ is given by

$$\sigma_{\mathbf{e}}^2 = \overline{\mathbf{e}^2} = \begin{bmatrix} \mu_{A1}+\nu_{B1} \\ \mu_{A2}+\nu_{B2} \\ \vdots \\ \mu_{AM}+\nu_{BM} \end{bmatrix} \cdot \frac{E_0^2}{12} \tag{VI.65}$$

where M is the order of the filter.

and the variance of $\boldsymbol{\varepsilon}$ is

$$\sigma_{\boldsymbol{\varepsilon}}^2 = \overline{\boldsymbol{\varepsilon}^2} = \begin{bmatrix} \mu_{C1}+\nu_{D1} \\ \mu_{C2}+\nu_{D2} \\ \vdots \\ \mu_{CK}+\nu_{DK} \end{bmatrix} \cdot \frac{E_0^2}{12} \tag{VI.66}$$

provided that all the roundoff errors are uncorrelated.

Following the definition of covariance matrix as mentioned in chapter IV, the covariance matrix for the state variable error vector is given by,

$$\mathrm{cov}[\Delta \mathbf{x}(n)] = \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1} \cdot \mathrm{cov}[\mathbf{e}(i)] \cdot (\mathbf{A}^{n-i-1})^{T} \tag{VI.67}$$

But the covariance matrix, $\mathrm{cov}[\mathbf{e}(i)]$, of the quantization error vector is given by,

$$\text{cov}[\mathbf{e}(i)] = \begin{bmatrix} E(e_1^2) & E(e_1e_2) & \cdots & E(e_1e_M) \\ E(e_2e_1) & E(e_2^2) & \cdots & E(e_2e_M) \\ \cdot & \cdot & \cdots & \cdot \\ E(e_Me_1) & E(e_Me_2) & \cdots & E(e_M^2) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{e1}^2 & & \bigcirc \\ & \sigma_{e2}^2 \cdot & \\ & & \cdot \; \sigma_{eM}^2 \end{bmatrix} \qquad \text{since } e_i \text{ and } e_j \text{ are uncorrelated when } i \neq j$$

i.e.

$$\text{cov}[\mathbf{e}(i)] = \frac{E_o^2}{12} \begin{bmatrix} (\mu_{A1} + \nu_{B1}) & & \bigcirc \\ & (\mu_{A2} + \nu_{B2}) & \\ & & \cdot \; \cdot \\ \bigcirc & & (\mu_{AM} + \nu_{BM}) \end{bmatrix} \qquad \text{(VI.68)}$$

where $E(.)$ denotes the expected value.

Hence substituting the expression of $\text{cov}[\mathbf{e}(i)]$ into the covariance matrix of the quantization error vector, one obtains

$$\text{cov}[\Delta\mathbf{x}(n)] = \frac{E_o^2}{12} \sum_{i=0}^{n-1} \mathbf{A}^{n-i-1} \begin{bmatrix} (\mu_{A1}+\nu_{B1}) & & \bigcirc \\ & (\mu_{A2}+\nu_{B2}) & \\ & & \cdot \; \cdot \\ \bigcirc & & (\mu_{AM}+\nu_{BM}) \end{bmatrix} (\mathbf{A}^{n-i-1})^T \qquad \text{(VI.69)}$$

The solution of (VI.69) is more easily facilitated by premultiplying (VI.69) by $\mathbf{A}$ and post-multiplying by $\mathbf{A}^T$; and after subtraction, one obtains,

$$\mathrm{cov}[\Delta \mathbf{x}(n)] - \mathbf{A}.\mathrm{cov}[\Delta \mathbf{x}(n)].\mathbf{A}^{T}$$

$$= \frac{E_o^2}{12}\left\{\begin{bmatrix}(\mu_{A1}+\nu_{B1}) & & \bigcirc \\ & (\mu_{A2}+\nu_{B2}) & \\ & & \ddots \\ \bigcirc & & (\mu_{AM}+\nu_{BM})\end{bmatrix} - \mathbf{A}^{n}\begin{bmatrix}(\mu_{A1}+\nu_{B1}) & & \bigcirc \\ & (\mu_{A2}+\nu_{B2}) & \\ & & \ddots \\ \bigcirc & & (\mu_{AM}+\nu_{BM})\end{bmatrix}(\mathbf{A}^{n})^{T}\right\}$$

(VI.70)

For a stable filter, in the steady state, $\mathbf{A}^{n} \longrightarrow \mathbf{0}$. Then (VI.70) (VI.70) can be further simplified to

$$\mathrm{cov}[\Delta \mathbf{x}(n)] - \mathbf{A}.\mathrm{cov}[\Delta \mathbf{x}(n)].\mathbf{A}^{T} = \frac{E_o^2}{12}\begin{bmatrix}(\mu_{A1}+\nu_{B1}) & & \bigcirc \\ & (\mu_{A2}+\nu_{B2}) & \\ & & \ddots \\ \bigcirc & & (\mu_{AM}+\nu_{BM})\end{bmatrix}$$

(VI.71)

The solution of (VI.71) involves the solution of M^2 linear equations, which yields the values of the M^2 elements of $\mathrm{cov}[\Delta \mathbf{x}(n)]$. The diagonal elements of $\mathrm{cov}[\Delta \mathbf{x}(n)]$ are the variance of $\Delta x_i(n)$ in the steady state.

After solving for $\mathrm{cov}[\Delta \mathbf{x}(n)]$, we can proceed on to solve for the variance of the output, i.e.

$$\mathrm{cov}[\Delta \mathbf{y}(n)] = \mathbf{C}.\ \mathrm{cov}[\Delta \mathbf{x}(n)]\ .\ \mathbf{C}^{T} + \begin{bmatrix}\mu_{C1}+\nu_{D1} \\ \mu_{C2}+\nu_{D2} \\ \vdots \\ \mu_{CK}+\nu_{DK}\end{bmatrix}\frac{E_0^2}{12} \quad \text{(VI.72)}$$

VI.6 Application of the State-Space Statistical Estimation to a Second Order Filter Realized in the Direct Canonic Form

For a second order digital filter realized in the direct canonic form, the state matrix is given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix} \tag{VI.73}$$

and

$$\mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{VI.74}$$

Hence, from (VI.68), the covariance matrix of the error vector is

$$\begin{aligned} \mathrm{cov}\left[\mathbf{e}(i)\right] &= \frac{E_Q}{12}\begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \\ &= \frac{2E_Q^{\,2}}{12}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{VI.75}$$

Sucstituting the expression of (VI.75) in (VI.71), in the steady state, then,

$$\mathrm{cov}\left[\Delta\mathbf{x}(n)\right] - \mathbf{A}\,.\,\mathrm{cov}\left[\Delta\mathbf{x}(n)\right].\mathbf{A}^{T} = \frac{2E_Q^{\,2}}{12}\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \tag{VI.76}$$

which gives four linear equations in terms of the four unknown elements of $\mathrm{cov}[\Delta \mathbf{x}(n)]$. Now let these elements be represented by

$$\mathrm{cov}[\Delta \mathbf{x}(n)] = \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}$$

Multiplying out the expression on the LHS of (VI.76) and simplifying, the following four equations are obtained

$$v_{11} - v_{22} = 0 \qquad \text{(VI.77)}$$

$$v_{12} + b_2 v_{21} + b_1 v_{22} = 0 \qquad \text{(VI.78)}$$

$$b_2 v_{12} + v_{21} + b_1 v_{22} = 0 \qquad \text{(VI.79)}$$

$$-b_2^2 v_{11} - b_1 b_2 v_{12} - b_1 b_2 v_{12} + (1-b_1^2) v_{22} = \frac{2E_0^{\,2}}{12} \qquad \text{(VI.80)}$$

Equation(VI.77) gives

$$v_{11} = v_{22} \qquad \text{(VI.81)}$$

Also, from equations (VI.78) and (VI,79), one obtains

$$v_{12} = v_{21} \qquad \text{(VI.82)}$$

Substituting (VI.81) and (VI.82) into (VI.79) and (VI.80) and solving, one obtains

$$v_{11} = v_{22} = \frac{(1 + b_2)}{(1-b_2)(1-b_1^2 + 2b_2 + b_2^2)} \frac{2E_0^2}{12}$$

$$v_{12} = v_{21} = \frac{-b_1}{(1-b_2)(1-b_1^2 + 2b_2 + b_2^2)} \frac{2E_0^2}{12} \qquad \text{(VI.83)}$$

i.e. the variance of the error in the state-variables x_1 and x_2 are given by

$$\sigma^2_{\Delta x_1} = \sigma^2_{\Delta x_2} = v_{11} = \frac{(1 + b_2)}{(1-b_2)(1-b_1^2+2b_2+b_2^2)} \cdot \frac{2E_0^2}{12} \qquad \text{(VI.84)}$$

After solving for the elements of $\text{cov}[\Delta \mathbf{x}(n)]$, the variance of the steady-state error in the output is given by, from (VI.72)

$$\sigma^2_{\Delta y} = \left[(\alpha_2-\alpha_0\beta_2) \quad (\alpha_1-\alpha_0\beta_1)\right] \cdot \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} \begin{bmatrix} (\alpha_2-\alpha_0\beta_2) \\ (\alpha_1-\alpha_0\beta_1) \end{bmatrix} + \frac{3E_0^2}{12}$$

(VI.85)

If only the variance of the output error is desired, a similar method to that shown in section IV.4 can be employed direct to (VI.63) Hence,

$$\begin{aligned} \sigma^2_{\Delta y} &= \sum_{i=0}^{n-1} \mathbf{C}\mathbf{A}^{n-i-1} \text{cov}[\mathbf{e}(i)] (\mathbf{A}^{n-i-1})^T \cdot \mathbf{C}^T + \text{cov}[\mathbf{\varepsilon}(n)] \\ &= \frac{2E_0^2}{12} \sum_{i=0}^{n-1} \mathbf{C}\mathbf{A}^{n-i-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} (\mathbf{A}^{n-i-1})^T \cdot \mathbf{C}^T + \frac{3E_0^2}{12} \\ &= \frac{2E_0^2}{12} \sum_{i=0}^{n-1} \mathbf{C}\mathbf{A}^{n-i-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{A}^{n-i-1})^T \mathbf{C}^T + \frac{3E_0^2}{12} \end{aligned} \qquad \text{(VI.86)}$$

Now since

$$\begin{bmatrix} 0 & 1 \end{bmatrix} (\mathbf{A}^{n-i-1})^T \mathbf{C}^T = \mathbf{C}(\mathbf{A}^{n-i-1}) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \text{(VI.87)}$$

then (VI.86) can be rewritten as

$$\sigma^2_{\Delta y} = \frac{2E_0^2}{12} \sum_{i=0}^{n-1} \left(\mathbf{C}\mathbf{A}^{n-i-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right)^2 + \frac{3E_0^2}{12} \qquad \text{(VI.88)}$$

As mentioned in section IV.5, (VI.88) can be more conveniently evaluated, especially when n is large, by determining the eigenvalues and eigenvectors of $\mathbf{A}$, i.e. by making use of the equation

$$\mathbf{A}^n = \mathbf{P}\,\Lambda^n\,\mathbf{P}^{-1} \tag{VI.89}$$

Thus, if Λ_1 and Λ_2 are the two eigenvalues of $\mathbf{A}$, then (VI.88) is reduced to

$$\sigma^2_{\Delta y} = \frac{2E_0^2}{12}\sum_{i=0}^{n-1}\left(\mathbf{C}\,\mathbf{P}\begin{bmatrix}\Lambda^{n-i-1} & 0\\ 0 & \Lambda^{n-i-1}\end{bmatrix}\mathbf{P}^{-1}\begin{bmatrix}0\\ 1\end{bmatrix}\right)^2 + \frac{3E_0^2}{12} \tag{VI.90}$$

A word of caution must be added here that in counting the number of non-zero and non-unity elements in the matrix $\mathbf{C}$, the matrix $\mathbf{C}$ is of the form

$$\mathbf{C} = \begin{bmatrix}(a_2 - a_0 b_2) & (a_1 - a_0 b_1)\end{bmatrix} \tag{VI.91}$$

If $a_1 = a_2 = 0$, then the number of non-zero or non-unity elements in $\mathbf{C}$ is zero. This is rather obvious from the diagram of a second order filter realized in the direct canonic form.

Consider the second order filter shown in fig VI.3

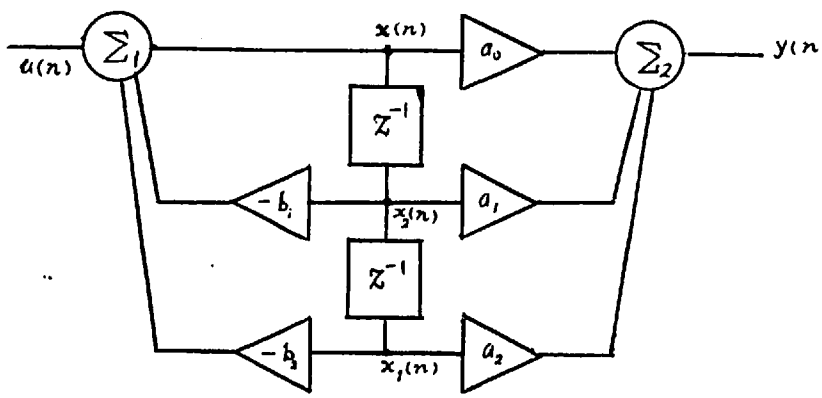


Fig VI.3

The noise source immediately after the multiplication with $(-b_1)$ and $(-b_2)$ has been taken into account in calculating $\mathbf{x}(n+1)$ since the state equation is given by

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{B}u(n) \qquad \text{(VI.92)}$$

Assuming no further error is introduced in the input summer Σ_1, the only subsequent multiplication error on $x(n)$ is from the multiplication with a_o. Thus the noise introduced by the parts $(-a_o b_2)$ and $(-a_o b_1)$ in the elements of $\mathbf{C}$ has been taken into account in the consideration of $\Delta x_1(n)$ and $\Delta x_2(n)$.

VI.7 Multiplication Round-Off Errors in a Multirate Digital Filter

Consider a time-invariant multirate digital filter in which the product after each multiplication has to be rounded-off to an accuracy E_o (fig VI.4)

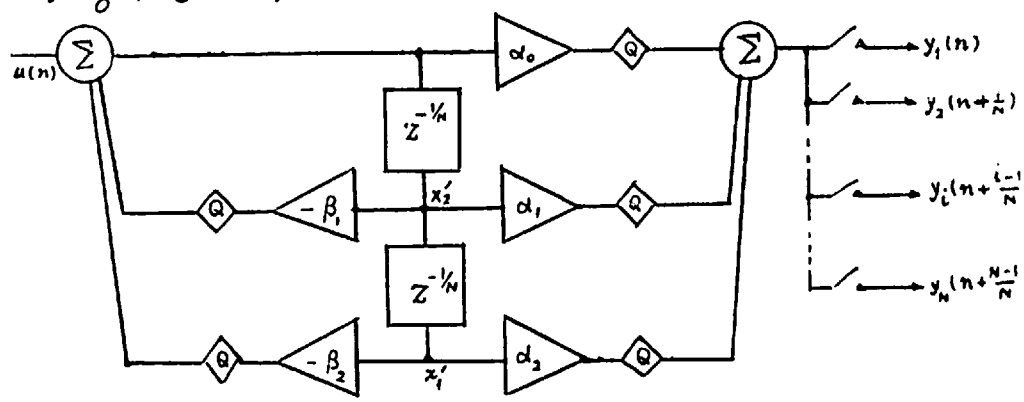


Fig VI.4 A Time-Invariant Multirate Digital Filter

For an ideal filter the following sets of dynamic equations can be be written:-

$$\begin{aligned}
\mathbf{x}(n+\tfrac{1}{N}) &= \mathbf{A}\,\mathbf{x}(n) + \mathbf{B}u(n)\ ; & y_1(n) &= \mathbf{C}\,\mathbf{x}(n) + Du(n)\ ;\\
\mathbf{x}(n+\tfrac{2}{N}) &= \mathbf{A}\,\mathbf{x}(n+\tfrac{1}{N}) & ;\quad y_2(n+\tfrac{1}{N}) &= \mathbf{C}\,\mathbf{x}(n+\tfrac{1}{N}) \quad ;\\
&\vdots & &\vdots \qquad\qquad (VI.93)\\
\mathbf{x}(n+1) &= \mathbf{A}\,\mathbf{x}(n+\tfrac{N-1}{N}) & ;\quad y_N(n+\tfrac{N-1}{N}) &= \mathbf{C}\,\mathbf{x}(n+\tfrac{N-1}{N}) \quad ;
\end{aligned}$$

However, for a non-ideal filter shown in fig VI.4, the dynamic equations becomes:-

$$\begin{aligned}
\mathbf{x}'(n+\tfrac{1}{N}) &= [\mathbf{A}\,\mathbf{x}'(n)]_{E_o} + [\mathbf{B}u(n)]_{E_o}; & y_1'(n) &= [\mathbf{C}\,\mathbf{x}'(n)]_{E_o} + [Du(n)]_{E_o};\\
\mathbf{x}'(n+\tfrac{2}{N}) &= [\mathbf{A}\,\mathbf{x}'(n+\tfrac{1}{N})]_{E_o} & ;\quad y_2'(n+\tfrac{1}{N}) &= [\mathbf{C}\,\mathbf{x}'(n+\tfrac{1}{N})]_{E_o};\\
&\vdots & &\vdots \qquad\qquad (VI.94)\\
\mathbf{x}'(n+1) &= [\mathbf{A}\,\mathbf{x}'(n+\tfrac{N-1}{N})]_{E_o}; & y_N'(n+\tfrac{N-1}{N}) &= [\mathbf{C}\,\mathbf{x}'(n+\tfrac{N-1}{N})]_{E_o};
\end{aligned}$$

where $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\beta_2 & -\beta_1 \end{bmatrix}$; $\mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$;

$\mathbf{C} = [(\alpha_2 - \alpha_0\beta_2) \quad (\alpha_1 - \alpha_0\beta_1)]$; $D = \alpha_0$

and $[.]_{E_o}$ again as in section VI.2 denotes the quantization of a multiplication product to an accuracy of E_o

Defining the error vectors in the following way:-

$$
\begin{aligned}
\mathbf{e}(n) &= \mathbf{A}\mathbf{x}'(n) - [\mathbf{A}\mathbf{x}'(n)]_{E_o} + \mathbf{B}u(n) - [\mathbf{B}u(n)]_{E_o};\\
\mathbf{e}(n+\tfrac{1}{N}) &= \mathbf{A}\mathbf{x}'(n+\tfrac{1}{N}) - [\mathbf{A}\mathbf{x}'(n+\tfrac{1}{N})]_{E_o};\\
&\vdots\\
\mathbf{e}(n+\tfrac{N-1}{N}) &= \mathbf{A}\mathbf{x}'(n+\tfrac{N-1}{N}) - [\mathbf{A}\mathbf{x}'(n+\tfrac{N-1}{N})]_{E_o};
\end{aligned}
\qquad \text{(VI.95)}
$$

$$
\begin{aligned}
\varepsilon(n) &= \mathbf{C}\mathbf{x}'(n) - [\mathbf{C}\mathbf{x}'(n)]_{E_o} + Du(n) - [Du(n)]_{E_o};\\
\varepsilon(n+\tfrac{1}{N}) &= \mathbf{C}\mathbf{x}'(n+\tfrac{1}{N}) - [\mathbf{C}\mathbf{x}'(n+\tfrac{1}{N})]_{E_o};\\
&\vdots\\
\varepsilon(n+\tfrac{N-1}{N}) &= \mathbf{C}\mathbf{x}'(n+\tfrac{N-1}{N}) - [\mathbf{C}\mathbf{x}'(n+\tfrac{N-1}{N})]_{E_o};
\end{aligned}
$$

Then the error in the state vectors and the output vectors are,

$$
\begin{aligned}
\Delta\mathbf{x}(n+\tfrac{1}{N}) &= \mathbf{x}(n+\tfrac{1}{N}) - \mathbf{x}'(n+\tfrac{1}{N}) & \Delta y_1(n) &= y_1(n) - y_1'(n)\\
&= \mathbf{A}\Delta\mathbf{x}(n) + \mathbf{e}(n); & &= \mathbf{C}\Delta\mathbf{x}(n) + \varepsilon(n);\\
\Delta\mathbf{x}(n+\tfrac{2}{N}) &= \mathbf{A}\Delta\mathbf{x}(n+\tfrac{1}{N}) + \mathbf{e}(n+\tfrac{1}{N}); & \Delta y_2(n) &= \mathbf{C}\Delta\mathbf{x}(n+\tfrac{1}{N}) + \varepsilon(n+\tfrac{1}{N});\\
&\vdots & &\vdots\\
\Delta\mathbf{x}(n+1) &= \mathbf{A}\Delta\mathbf{x}(n+\tfrac{N-1}{N}) + \mathbf{e}(n+\tfrac{N-1}{N}); & \Delta y_N(n) &= \mathbf{C}\Delta\mathbf{x}(n+\tfrac{N-1}{N}) + \varepsilon(n+\tfrac{N-1}{N});
\end{aligned}
\qquad \text{(VI.96)}
$$

Hence,

$$
\Delta\mathbf{x}(\tfrac{nN+i}{N}) = \mathbf{C}\sum_{K=0}^{nN+i-1}\mathbf{A}^{(nN+i)-k-1}\begin{bmatrix} e_1(k/N)\\ e_2(k/N)\end{bmatrix}
\qquad \text{(VI.97)}
$$

and $$\Delta y_i\left(n+\frac{i-1}{N}\right) = \mathbf{C}\sum_{k=0}^{nN+i-2}\mathbf{A}^{(nN+i-1-k-1)}\begin{bmatrix}e_1(k/N)\\ e_2(k/N)\end{bmatrix} + \varepsilon\left(n+\frac{i-1}{N}\right) \qquad \text{(VI.98)}$$

For $k = nN$, the variance of the vector $\mathbf{e}$ is given by

$$\sigma_e^2 = \overline{\mathbf{e}^2(n)} = \begin{bmatrix}\mu_{A1}+\nu_{B1}\\ \mu_{A2}+\nu_{B2}\end{bmatrix}\frac{E_o^2}{12} \qquad \text{(VI.99)}$$

and the variance of ε is,

$$\sigma_\varepsilon^2 = \overline{\varepsilon^2(n)} = \left[\mu_{C1}+\nu_{D1}\right]\frac{E_0^2}{12} \qquad \text{(VI.100)}$$

For other values of k, the variance of $\mathbf{e}$ and ε are respectively given by,

$$\sigma_e^2 = \begin{bmatrix}\mu_{A1}\\ \mu_{A2}\end{bmatrix}\frac{E_0^2}{12} \qquad \text{(VI.101)}$$

$$\sigma_\varepsilon^2 = \left[\mu'_{C1}\right]\frac{E_0^2}{12} \qquad \text{(VI.102)}$$

However, for a direct canonic configuration, $\nu_{B1} = \nu_{B2} = 0$, and $\mu'_{C1} = \mu_{C1} + \nu_{D1}$; thus the variance of $\mathbf{e}$ and ε are simply given by equations (VI.101) and (VI.102). Hence, from (VI.67), the covariance matrix of $\Delta\mathbf{x}(n)$is given by,

$$\begin{aligned}\mathrm{cov}\left[\Delta\mathbf{x}\left(n+\frac{i}{N}\right)\right] &= \sum_{k=0}^{nN+i-1}\mathbf{A}^{(nN+i)-k-1}\cdot\mathrm{cov}\left[\mathbf{e}(n)\right]\cdot\left\{\mathbf{A}^{(nN+i)-k-1}\right\}^T\\ &= \frac{2E_0^2}{12}\sum_{k=0}^{nN+i-1}\mathbf{A}^{nN+i-k-1}\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix}\left[\mathbf{A}^{nN+i-k-1}\right]^T \qquad \text{(VI.103)}\end{aligned}$$

and the variance of the ith output error is

$$\sigma^2_{\Delta y_i} = \frac{2E_0^2}{12}\sum_{k=0}^{nN+i-2} \mathbf{C}\,\mathbf{A}^{(nN+i-1)-k-1}\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix}\left\{\mathbf{A}^{nN+i-1-k-1}\right\}^{\mathsf{T}}\mathbf{C}^{\mathsf{T}} + \mu'_{C1}\cdot\frac{E_0^2}{12} \tag{VI.104}$$

$i = 0, 1, 2, \ldots N$. μ'_{C1} = number of non-zero and non-unity multipliers in the feed-forward paths of the digital filter.

If the multirate digital filter is periodically varying, i.e. if $\mathbf{A}_i$, $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{D}_i$ are the matrices of its dynamic equations at $t=(n+\frac{i}{N})$, a slightly more complicated expression for the multiplication round-off error can be derived. Following a similar argument as in the case of a time-invariant filter, the errors in the state vector are

$$\begin{aligned}
\Delta\mathbf{x}(n+\tfrac{1}{N}) &= \mathbf{A}_1\,\Delta\mathbf{x}(n) + \mathbf{e}(n)\\
\Delta\mathbf{x}(n+\tfrac{2}{N}) &= \mathbf{A}_2\,\Delta\mathbf{x}(n+\tfrac{1}{N}) + \mathbf{e}(n+\tfrac{1}{N})\\
&\vdots\\
\Delta\mathbf{x}(n+1) &= \mathbf{A}_N\,\Delta\mathbf{x}(n+\tfrac{N-1}{N}) + \mathbf{e}(n+\tfrac{N-1}{N})
\end{aligned} \tag{VI.105}$$

The solution of (VI.105) is given by

$$\Delta\mathbf{x}(n+\tfrac{i}{N}) = \sum_{j=0}^{n}(\mathbf{A}_N\cdots\mathbf{A}_2\,\mathbf{A}_1)^{n-j}\sum_{k=0}^{i-2}(\mathbf{A}_i\,\mathbf{A}_{i-1}\cdots\mathbf{A}_{k+2})\,\mathbf{e}(j+\tfrac{k}{N}) + \mathbf{e}(j+\tfrac{i-1}{N}) \tag{VI.106}$$

and hence the error in the ith output is

$$\Delta y_i(n+\tfrac{i-1}{N}) = \mathbf{C}_i\,\Delta\mathbf{x}(n+\tfrac{i}{N}) + \varepsilon(n+\tfrac{i-1}{N}) \tag{VI.107}$$

The covariance matrix of the errors in the state variable vector is

given by

$$\operatorname{cov}[\Delta \mathbf{x}(n+\tfrac{i}{N})] = \frac{2E_0^2}{12}\left\{\sum_{j=0}^{n}\left(\mathbf{A}_N \cdots \mathbf{A}_2 \mathbf{A}_1\right)^{n-j} \sum_{k=0}^{i-2}(\mathbf{A}_i \mathbf{A}_{i-1} \cdots \mathbf{A}_{k+2})\left(\begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix}\right)\cdot\right.$$
$$\left.\cdot(\mathbf{A}_i \mathbf{A}_{i-1} \cdots \mathbf{A}_{k+2})^{T} \left(\mathbf{A}_N \cdots \mathbf{A}_2 \mathbf{A}_1\right)^{T^{n-j}} + \begin{bmatrix}0\\1\end{bmatrix}\begin{bmatrix}0 & 1\end{bmatrix}\right\} \qquad \text{(VI.108)}$$

and substituting this into the variance of the ith output error

$$\sigma^2_{\Delta y_i} = \mathbf{C}_i \cdot \operatorname{cov}\left[\Delta \mathbf{x}(n+\tfrac{i-1}{N})\right] \cdot \mathbf{C}_i^{T} + \mu'_{Ci} \cdot \frac{E_0^2}{12} \qquad \text{(VI.109)}$$

where μ'_{Ci} = number of non-zero and non-unity multipliers in the feed-forward paths.

VI.8 Comparison of Multiplication Round-Off Errors between Single-Rate and Multirate Time-Invariant Digital Filters

If a second order multirate digital filter is periodically varying, i.e. $\mathbf{A}_i$, $\mathbf{B}_i$, $\mathbf{C}_i$ and $\mathbf{D}_i$ are the matrices at $t = (n+\frac{i}{N})$ where $i = 0,1,\ldots,N-1$, then there are $(N-1)$ degrees of freedom in the choice of the coefficients β_{i1} and β_{i2}. Hence it would be difficult to compare the multiplication errors of such a filter with its equivalent single-rate filter. Here in this section, attention is focused on the time-invariant multirate digital filter and its single-rate counterpart.

Consider a single-rate filter, from (VI.84), the steady state variance of the errors in the state variables x_1 and x_2 are

$$\sigma^2_{\Delta x_s} = \frac{1 + b_2}{(1 - b_2)(1 - b_1^2 + 2b_2 + b_2^2)} \cdot \frac{2E_0^2}{12} \qquad \text{(VI.110)}$$

Now, if the complex poles of this second order filter are given by

$$\Lambda = re^{\pm j\theta} \tag{VI.111}$$

then (VI.110) becomes

$$\sigma^2_{\Delta x_s} = \frac{(1 + r^2)}{(1 - r^2)(1 - 2r^2\cos 2\theta + r^4)} \cdot \frac{2E_0^2}{12} \tag{VI.112}$$

Most digital filters contains poles very close to the unit circle. This is certainly true for the design of highly selective filters. For these cases (VI.112) can be greatly simplified.

Let $r = 1 - \varepsilon$ and ignoring terms with quadratic and higher exponents in ε, eqn(VI.112) becomes

$$\sigma^2_{\Delta x_s} = \frac{2E_0^2}{12} \cdot \frac{1 + (1-\varepsilon)^2}{\{1 - (1-\varepsilon)^2\}\{1 + (1-\varepsilon)^4 - 2(1-\varepsilon)^2\cos 2\theta\}}$$

$$\simeq \frac{2E_0^2}{12} \cdot \frac{(1 - \varepsilon)}{2\varepsilon(1 - 2\varepsilon)(1 - \cos 2\theta)}$$

$$= \frac{2E_0^2}{12} \cdot \frac{(1 - \varepsilon)}{4\varepsilon(1 - 2\varepsilon)\sin^2\theta} \tag{VI.113}$$

Now, $(1 - 2\varepsilon) \simeq (1 - \varepsilon)^2$, thus again ignoring the terms with quadratic exponents in ε, (VI.113) can be reduced to

$$\sigma^2_{\Delta x_s} \simeq \frac{2E_0^2}{12} \cdot \frac{1}{4\varepsilon(1 - \varepsilon)\sin^2\theta}$$

$$\simeq \frac{2E_0^2}{12} \cdot \frac{1}{4\varepsilon \sin^2\theta} \tag{VI.114}$$

Now, consider a multirate time-invariant second order filter with state matrix given by

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\beta_2 & -\beta_1 \end{bmatrix}$$

Let the complex eigenvalues of $\mathbf{A}$ be given by

$$\lambda = \rho e^{\pm j\phi} \tag{VI.115}$$

If the time-invariant multirate filter is to give the same performance as the single-rate digital filter, from section III.10,

$$\begin{aligned} r &= \rho^N \\ \theta &= N\phi \end{aligned} \tag{VI.116}$$

Following similar developments of argument as in the case of single rate filters, the steady state variance of the error in the state variables x_1 and x_2 of the multirate filter is given by

$$\begin{aligned} \sigma^2_{\Delta x_m} &= \frac{2E_0^2}{12} \cdot \frac{1 + \beta_2}{(1 - \beta_2)(1 - \beta_1^2 + 2\beta_2 + \beta_2^2)} \\ &= \frac{2E_0^2}{12} \cdot \frac{1 + \rho^2}{(1 - \rho^2)(1 - 2\rho^2\cos 2\phi + \rho^4)} \end{aligned} \tag{VI.117}$$

Again, for highly selective filters, let $r = \rho^N = (1 - \varepsilon)$ and ignoring terms with quadratic and higher powers of ε, (VI.117) can be simplified to

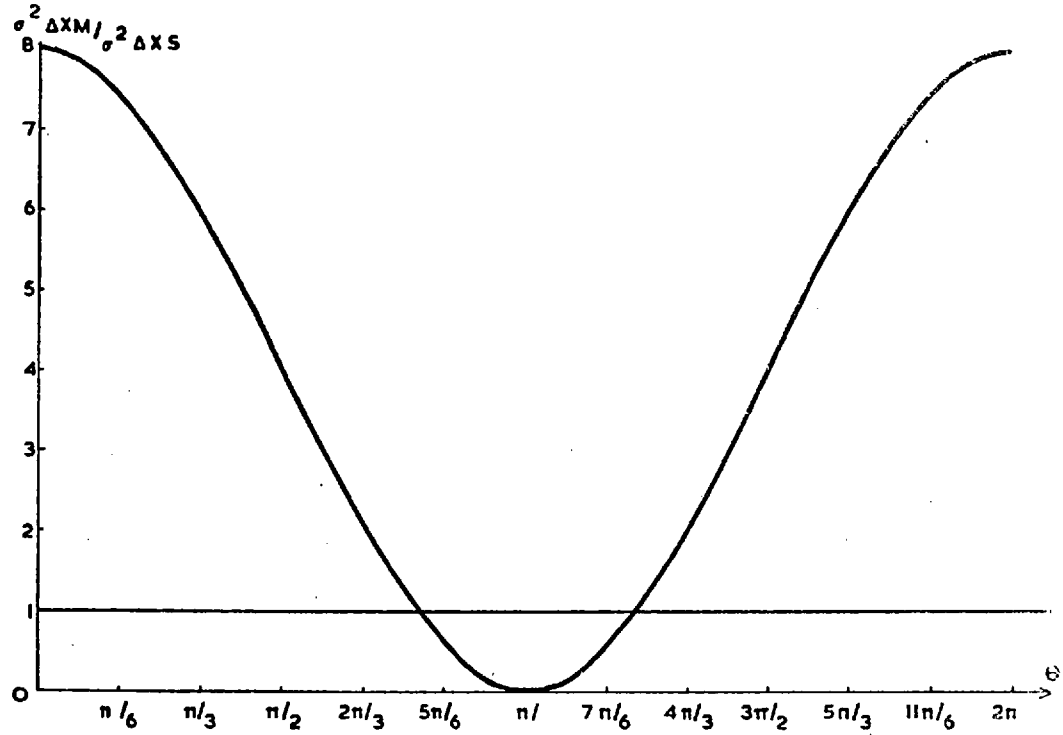
$$\begin{aligned} \sigma^2_{\Delta x_m} &\simeq \frac{2E_0^2}{12} \cdot \frac{1}{4\,\frac{\varepsilon}{N}\sin^2\frac{\theta}{N}} \\ &= \frac{2E_0^2}{12} \cdot \frac{N}{4\varepsilon . \sin^2\frac{\theta}{N}} \end{aligned} \tag{VI.118}$$

Hence the ratio $\sigma^2_{\Delta xm} / \sigma^2_{\Delta xs}$ can be written as

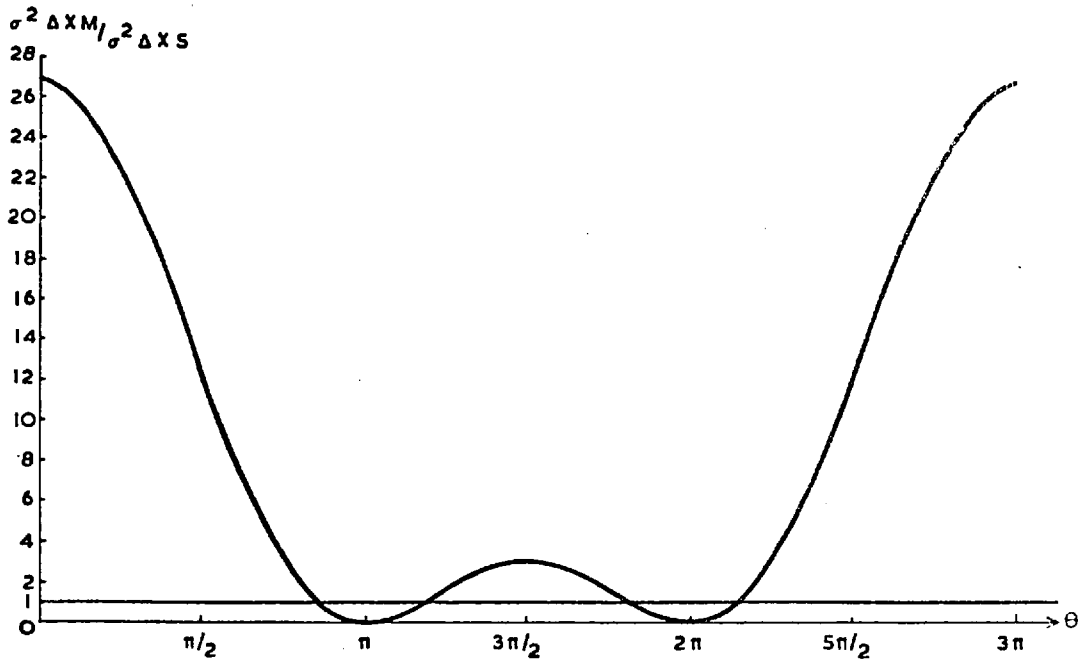
$$\frac{\sigma^2_{\Delta xm}}{\sigma^2_{\Delta xs}} \simeq \frac{N \sin^2\theta}{\left(\sin^2 \frac{\theta}{N}\right)} \qquad \text{(VI.119)}$$

Eqn(VI.119) expresses the ratio of the variance of the state variable errors in the multirate time-invariant digital filter to that in the single-rate filter. For a given N, if the filter is highly selective, it is easy to compute from eqn (VI.119) the value of θ for which the ratio $\sigma^2_{\Delta xm} / \sigma^2_{\Delta xs} \leq 1$, i.e. the position of the poles for which the multiplication round-off errors in a time-invariant multirate digital filter is lower than that in a single-rate filter.

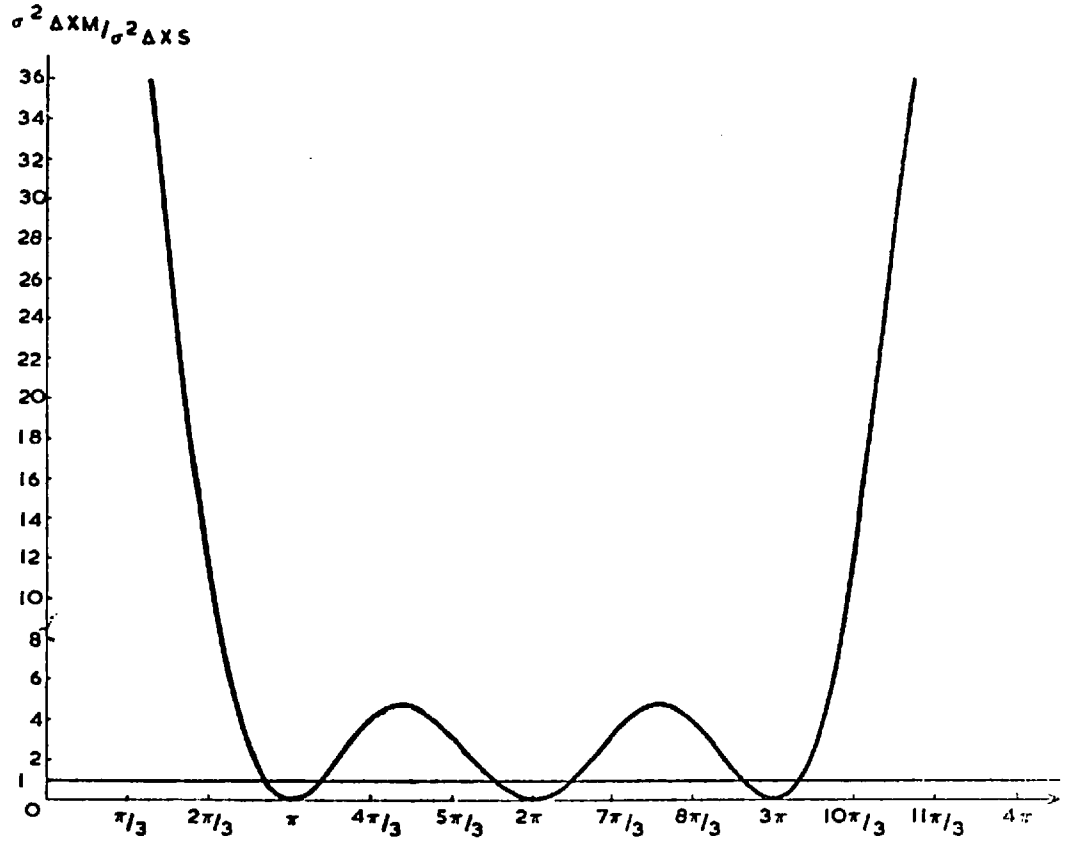
The graphs of equation (VI.119) for double-rate ($N = 2$), triple-rate ($N = 3$) and quadriple-rate ($N = 4$) time invariant digital filters are shown in figs (VI.5), (VI.6) and (VI.7) respectively. From these graphs, if the pole angle θ of the desired digital filter falls in the region where the state variable noise ratio $\sigma^2_{\Delta xm} / \sigma^2_{\Delta xs}$ is below the line of unity, one would expect the performance of the time-invariant multirate digital filter to be better than its equivalent single-rate filter.



POLE ANGLE ~ STATE VARIABLE NOISE RATIO
FOR A DOUBLE-RATE FILTER (N = 2) FIG. VI·5



POLE ANGLE ~ STATE VARIABLE NOISE RATIO
FOR A TRIPLE-RATE FILTER (N = 3) FIG. VI·6



POLE ANGLE ~ STATE VARIABLE NOISE RATIO
FOR A QUADRIPLE-RATE FILTER (N = 4)

FIG. VI·7

VI.9 Résumé

The multiplication round-off errors in a digital filter have been considered. In general, there are two ways of tackling the problem, viz, the evaluation of the upper bound of the errors and the evaluation of a statistical bound. Both methods have been discussed. The statistical approach is, in general, more realistic due to the fact that the input to the filter is commonly stochastic. A method to evaluate the statistical estimation of the multiplication has been derived using the state-space method since the state-space method can be readily applied to periodically varying multirate digital filters. Using this state-space method, expressions have been derived for the statistical bound of the multiplication round-off errors in both the time-invariant and periodically varying multirate digital filters.

Finally, a comparison of the multiplication errors in a time-invariant multirate filter and that in a single-rate filter has been performed. A simple expression has been derived giving the ratio of the error variances in the two cases. Such an expression is based on the assumption that the desired filters are highly selective. From such an expression of the ratio, the pole position in which a time-invariant filter is more desirable than its equivalent single-rate filter can be determined.

CHAPTER VII

LIMIT CYCLE OSCILLATIONS IN A MULTI-RATE DIGITAL FILTER

VII.1 Introduction

In the previous chapter, the multiplication round-off errors in a second order digital filter have been discussed. It has been assumed that the input signal to the filter is stochastic so that the error in the output can be treated statistically. However, if the input to the filter is deterministic, say an impulse or a step function, then, since the round-off errors in the state-variables would be highly correlated, the problem has to be treated differently.

If the input to a stable and ideal digital filter is set to zero, the output will decay asymtotically to zero. However, when rounding of intermediate products is performed in the filter implementation, it is possible that the output may sustain a non-zero level or oscillate indefinitely about zero. Similarly, for a constant nonzero input, the ideal filter output should asymtotically approach a steady-state level determined by the filter transfer function. Due to rounding, however, the output may oscillate or it may maintain a constant level different from the ideal output. When these phenomena occur, the filter is said to exhibit a limit cycle. The range of output values that occur for a particular limit cycle about the desired response is sometimes called a deadband.

Limit cycle oscillations remain an undesirable property of the digital filter for most engineering applications. To suppress

these oscillations, it has been proposed {5}, that a small random noise should be added to the input of the filter; this is called dithering. However, although this method breaks up the regular pattern of the limit cycle oscillation (LCO), it introduces a new random error and sometimes may not be very effective. Here in this chapter, the properties of LCO in a digital filter are briefly discussed, and it is demonstrated that LCO may be completely absent in some cases of a multi-rate digital filter. Thus this type of multi-rate digital filters will be more suitable for engineering applications where LCO is an embarrassing problem.

VII.2 Classification and Existence of Limit Cycle Oscillations

Limit cycle oscillations in a digital filter are caused by the non-linear feedback within the filter. Their existence is not affected by the presence of the zeros of the filter transfer function. Hence it will be sufficient to study the second-order digital filter with the transfer function

$$H(z) = \frac{1}{1 + b_1 z^{-1} + b_2 z^{-2}} \qquad \text{(VII.1)}$$

The transfer function can be implemented in the form as shown in fig VII.1

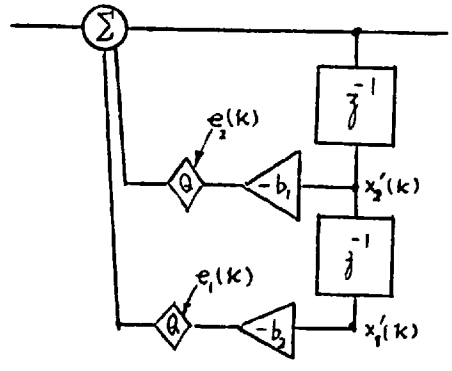


Fig VII.1
Second Order Digital Filter with Single Precision Adder

The filter in fig VII.1 can be represented by the following state equation:

$$\begin{bmatrix} x_1'(k+1) \\ x_2'(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix} \begin{bmatrix} x_1'(k) \\ x_2'(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \{e_1(k)+e_2(k)\} \quad \text{(VII.2)}$$

where $e_1(k)$ and $e_2(k)$ are the errors introduced by quantizing the multiplication products at $t = kT$.

If the linear part of eqn (VII.2) is stable, then the state vector $\mathbf{x}'(k)$ is bounded, i.e. $\mathbf{x}'(k)$ will actually either enter a limit cycle, or enter the origin of the "pseudo phase plane". (The pseudo phase plane {49} is the plane with x_1 and x_2 as the coordinates)

Now, a limit cycle is defined as a finite sequence of state vectors that satisfy a difference equation, i.e.

$$\Psi_p = \{\mathbf{x}'(1), \mathbf{x}'(2), \ldots, \mathbf{x}'(p)\} \text{ is a limit cycle}$$

iff

$$\left.\begin{aligned} &\mathbf{x}'(k+1) = \Phi\{\mathbf{x}'(k), k\} \\ \text{and} \quad &\mathbf{x}'(k+p) = \mathbf{x}'(k) \qquad \forall \ k \end{aligned}\right\} \quad \text{(VII.3)}$$

The positive integer p is called the period of the limit cycle. If the limit cycle is a constant state vector, then it is called a limit point. If the limit cycle consists of only two alternating state vectors, i.e.

$$\left.\begin{aligned} &\Psi_2 = \{\mathbf{x}'(1), \mathbf{x}'(2)\} \\ \text{and} \quad &\mathbf{x}'(1) = -\mathbf{x}'(2) \end{aligned}\right\} \quad \text{(VII.4)}$$

then it is called an alternate limit point. Any other type of limit cycle that exists in the system of (VII.2) is called a second order limit cycle.

The existence conditions for limit cycles in eqn (VII.2)

are now considered. It is now shown that if $|b_2| > 0.5$, then all solutions of eqn (VII.2) (except the trivial solution) are limit cycles. This existence condition was first observed by Jackson {26}, but the rigorous proof of it was first produced by Parker and Hess {40}. The following proof is another way of formulating the argument:-

Suppose that the state vector does enter the origin of the pseudo phase plane. Let $\mathbf{x}'(n)$ be the last vector before the system enters the origin, then

$$\begin{bmatrix} x_1'(n+1) \\ x_2'(n+1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -b_2 & -b_1 \end{bmatrix} \begin{bmatrix} x_1'(n) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \{e_1(n)+e_2(n)\} \quad \text{(VII.5)}$$

However, $e_2(n) = 0$ since $x_2'(n) = 0$. Thus, we have

$$-b_2 x_1'(n) + e_1(n) = 0 \quad \text{(VII.6)}$$

i.e.

$$-E_o/2 < b_2 x_1'(n) < E_o/2 \quad \text{(VII.7)}$$

where E_o=rounding step size. But $|x_1'(n)| \geq E_o$, thus the inequality of (VII.7) implies that

$$|b_2| < 0.5 \quad \text{(VII.8)}$$

if equation (VII.5) is possible. It follows that the zero state cannot be reached if $|b_2| > 0.5$, i.e. limit cycles always exist in a second order digital filter with complex poles if

$$\sqrt{\tfrac{1}{2}} < r < 1.0$$

where r is the magnitude of the complex pole.

It has been observed {25} that if $|b_2| < 0.5$, then the limit

cycles that can exist in the digital filter will be either limit points or alternate limit points. Also if $|b_2| < 0.5$ and if $\begin{bmatrix} 2 \\ \pm 2 \end{bmatrix} E_o$ and $\begin{bmatrix} 1 \\ \pm 1 \end{bmatrix} E_o$ enter the origin of the pseudo plane, then there exists no limit point or alternate limit point, i.e. there is no LCO.

VII.3 Bounds on the Amplitude of LCO

Various authors have derived upper bounds for the amplitudes of LCO that occur in a second order digital filter. The output of the system shown in eqn (VII.1) can be written as

$$y(n) = u(n) - [b_1 y(n-1)]_{E_o} - [b_2 y(n-2)]_{E_o} \qquad \text{(VII.9)}$$

where E_o again denotes the quantization step.

Jackson {26} estimated a bound on the magnitude of limit cycles for roundoff by equating (VII.9) to the output of an equivalent linear digital filter with $b_2 = 1$ (the condition for an oscillatory response), and using the following equation

$$[b_2 y(n-2)]_{E_o} = b_2 y(n-2) \pm \{0.5 \pm \delta(n-2)\} E_o \qquad \text{(VII.10)}$$

where $\delta(n)$ is any number between zero and unity. The resulting estimate for the magnitude of the limit cycle is given by

$$|y(n)| \leq 0.5E_o/(1-b_2) \qquad \text{VII.11}$$

Doubt is reserved on such a bound since it is independent of b_1. In fact, it has been found that the bound is too low for certain cases {49}. Experimental results by Parker and Hess {49} have indicated that the following bound seems to be a safer approximation than (VII.11)

$$|y(n)| \leq 1.5E_o/(1-b_2) \qquad \text{(VII.12)}$$

Sandberg and Kaiser {56} have arrived at the following formulae for the r.m.s. value, σ, of limit cycles in a second order section

$$\sigma = (1 - b_2)^{-1} \left(1 - \frac{b_1^2}{4b_2}\right)^{-\frac{1}{2}} \|e\| \qquad \text{(VII.13)}$$

for $b_2 > 0$ and $|b_1| \leq 4b_2/(1 + b_2)$

$$\sigma = (1 - |b_1| + b_2)^{-1} \|e\| \qquad \text{(VII.14)}$$

$b_2 \leq 0$, $b_2 > 0$ and $|b_1| \geq 4b_2/(1 + b_2)$

where $\|e\| \triangleq \left[\frac{1}{k+1} \sum_{n=0}^{k} e^2(n)\right]^{\frac{1}{2}}$,

$(k + 1)$ being the period of the limit cycle, and $e(n)$ is the roundoff error at $t = nT$ when limit cycle is reached, i.e.

$$y'(n) = -b_1 y'(n-1) - b_2 y'(n-2) + e(n) \qquad \text{(VII.15)}$$

For a general approach to the upper bound of LCO, the method by Yakowitz and Parker{70} described in section VI.2 and VI.3 is just as good as any other method. The advantage of this approach is that it also yields a bound on quantization errors during the transient period. Limit cycles, which are steady-state conditions, are included in this bound. As a recapitulation the bounds are written below:

For a filter with real poles, the bound of the state vector is

$$\Delta \mathbf{x}(n) \leq (1-|b_1|+b_2)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} E_o \qquad \text{(VII.16)}$$

and that for a filter with complex poles is

$$\Delta \mathbf{x}(n) \leq \frac{2\sqrt{b_2}\, E_o}{(1-\sqrt{b_2})\sqrt{4b_2 - b_1^2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{(1+\sqrt{b_2})E_o}{(1-b_2)\sqrt{1 - \frac{b_1^2}{4b_2}}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{(VII.17)}$$

It is interesting to compare the expressions of (VII.16) and (VII.17) with those of (VII.13) and (VII.14). The bound of (VII.16) is identical to that of (VII.14) for real poles. The bound of (VII.17), which holds for all n, including transients

and the limit cycle condition, is greater by a factor of $(1 + \sqrt{b_2})/\sqrt{2}$ than that of (VII.13). This factor varies from unity to $\sqrt{2}$ when $b_2 = 1$

VII.4 LCO in a Multi-rate Digital Filter - Computer Simulation

As discussed in the previous chapter, the multiplication roundoff error in a time-invariant multirate digital filter can sometimes be less than the corresponding single-rate filter. Since limit cycles are multiplication roundoff errors in the steady state, one would expect that the bounds of the LCO in a time-invariant multirate filter be lower than those in the corresponding single-rate filter if the condition that (VI.119) < 1 is satisfied.

It is not the object of this chapter to develop a bound for the LCO in a multi-rate digital filter (since the bound for multiplication roundoff errors with zero input equally applies to LCO), but rather to demonstrate by computer simulations that a single-rate digital filter suffering from LCO can, in general, be replaced by a periodically varying multi-rate digital filter free from limit cycles provided that the periodically varying coefficients are suitably chosen. This section describes the computer simulations of such a filter and the process by which a suitable multirate filter with zero LCO can be found {69}:-

In a digital filter, LCO is caused only by the poles of the transfer function. The presence of the zeros does not affect the existence of LCO. Therefore it would be sufficient to consider an all-pole filter with a transfer function of the form

$$H(z) = \frac{1}{z^2 + b_1 z + b_2} \qquad \text{(VII.18)}$$

In order to simplify the problem, attention has been focused on realizing eqn (VII.18) by double-rate ($N = 2$) digital filters with periodically varying coefficients. The use of higher rate ($N > 2$) filters has been ignored since it is sufficient to demonstrate the principle using a double-rate filter.

Consider a double-rate filter (fig VII.2).

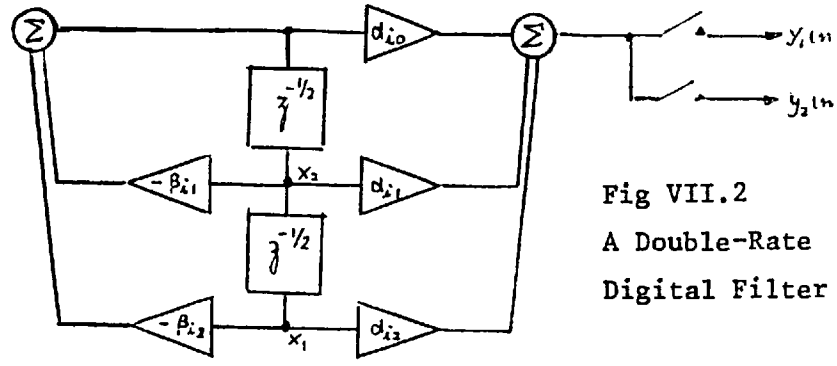


Fig VII.2
A Double-Rate
Digital Filter

There are two transfer functions (see Chapter III), i.e.

$$H_1(z) = \frac{\alpha_{10}z^2+(\alpha_{10}\beta_{22}-\alpha_{11}\beta_{21}+\alpha_{12})z+\alpha_{12}\beta_{22}}{z^2+(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z+\beta_{12}\beta_{22}} \qquad \text{(VII.19)}$$

and

$$H_2(z) = \frac{(-\alpha_{20}\beta_{21}+\alpha_{21})z^2+(\alpha_{21}\beta_{22}-\alpha_{22}\beta_{21})z}{z^2+(\beta_{12}+\beta_{22}-\beta_{11}\beta_{21})z+\beta_{12}\beta_{22}} \qquad \text{(VII.20)}$$

Either of these two transfer functions can be used to realize the all-pole filter of eqn (VII.18). However, if the double-rate filter is to be implemented with hardware, it will be more economical to use eqn (VII.20) since, in this case, both α_{21} and α_{22} can be made equal to zero, thus saving two multipliers. This is the reason why $H_2(z)$ has been used in all the simulations. Now, using eqn (VII.20) to realize $H(z)$ in eqn (VII.18), one obtains the following equations:

$$\beta_{12}+\beta_{22}-\beta_{11}\beta_{21} = b_1 \qquad \text{(VII.21)}$$

$$\beta_{12}\beta_{22} = b_2 \qquad \text{(VII.22)}$$

$$\alpha_{20} = -1/\beta_{21} \qquad \text{(VII.23)}$$

It means that we can choose any set of values for β_{11} and β_{12} and determine the values of β_{21} and β_{22} from eqns (VII.21) and (VII.22), or vice versa. If β_{11} and β_{12} are chosen and, if for each chosen set of β_{11} and β_{12}, a unit impulse input is injected, it has been found from the computer simulations of a double-rate filter that some of these sets of β_{11} and β_{12} give no LCO in the output. On the other hand, if β_{21} and β_{22} are chosen and varied, there are again sets of β_{21} and β_{22} that give no LCO. Thus if the set β_{11} and β_{12} is chosen and varied by one quantization step each time over the "triangle of stability" {25} (fig VII.3), and each time analysed with a unit impulse input, then all the values of β_{11} and β_{12} that give no LCO can be found. Again, if β_{21} and β_{22} are varied and analysed in the same way, then all the values of β_{21} and β_{22} giving no LCO can be recorded. Note that even if β_{11} and β_{12} lie outside the triangle, the resultant transfer function of eqn (VII.20) is not necessarily unstable. Values of β_{11} and β_{12} that lie outside the triangle are taken care of if β_{21} and β_{22} are the chosen coefficients and if their values are sufficiently small.

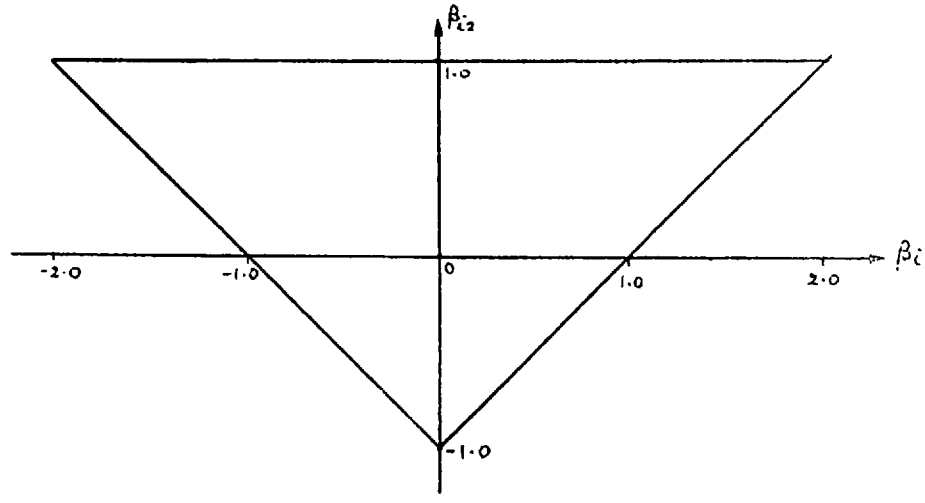
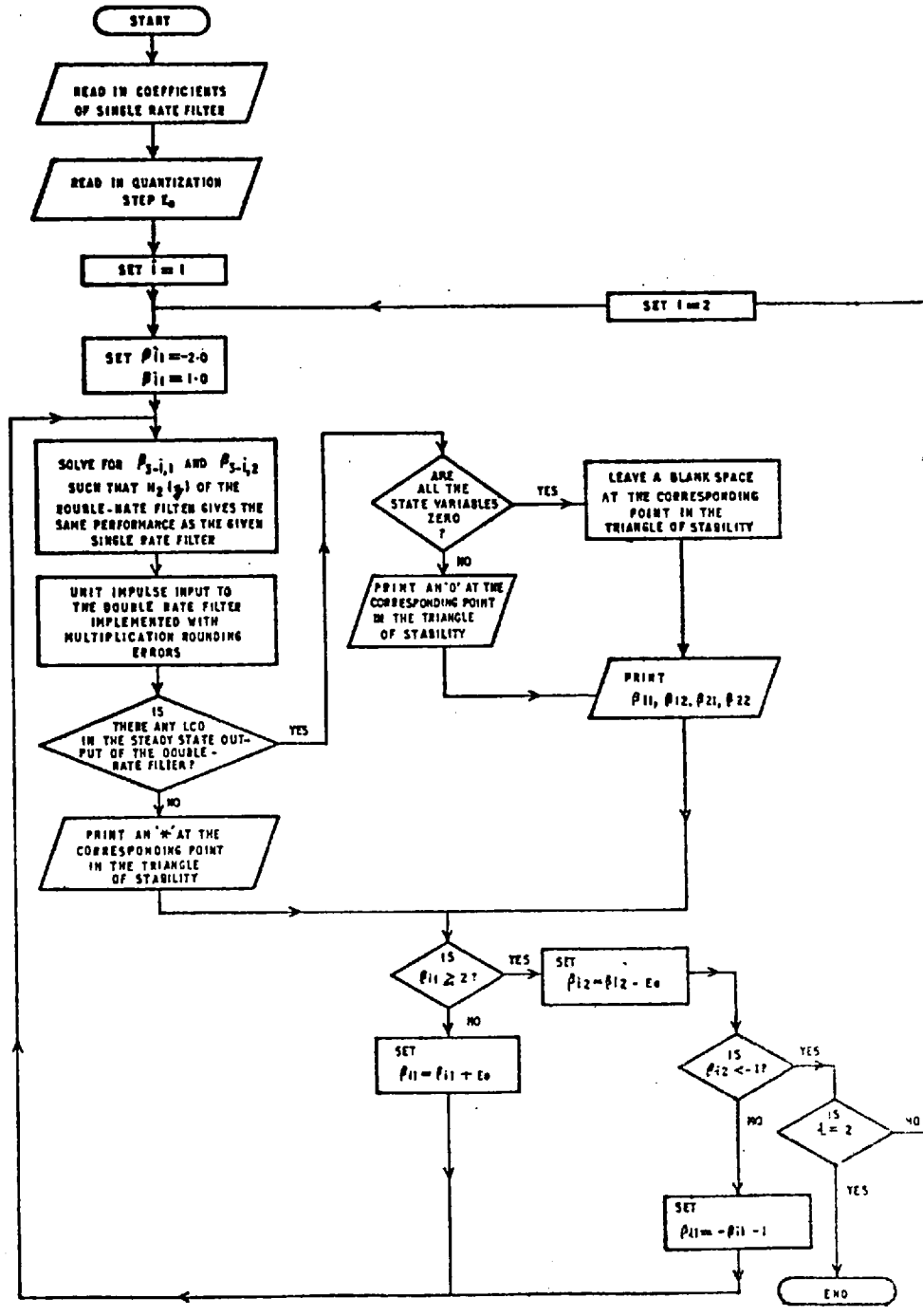


Fig VII.3 "Triangle of Stability"

Similarly, values of β_{21} and β_{22} lying outside the triangle are taken care of if β_{11} and β_{12} are chosen and are

sufficiently small. The flow-diagram of the computer simulation program is shown in fig VII.4.



FLOW CHART FOR THE SEARCH OF ZERO LCO POINTS ON THE $\beta_{i1} - \beta_{i2}$ PLANE

FIG. VII 4

X97-3006-1092

VII.5 Results and Observations from the Computer Simulations

Many all-pole second order filter simulations have been performed, and the process of searching for β_{ij} which give no LCO repeated. The following general observations have been obtained:-

(1) It has been observed that the suppression of LCO occurs not only when all the state variables x_1 and x_2 (fig VII.2) are zero, i.e.

$$\begin{bmatrix} x_1(n) \\ x_2(n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{(VII.24)}$$

for sufficiently large n, but also that when not all the state variables vanish yet the LCO in the output can be zero.

(2) For zero state variables, it is observed that

$$|\beta_{i2}| < 0.5 \qquad \text{(VII.25)}$$

This is similar to the conclusion drawn in section VII.2 for a single-rate digital filter.

(3) For zero LCO in the output but non-zero state variables,

$$\alpha_{20} < 0.5 \qquad \text{(VII.26)}$$

(4) The points on the β_{11}— β_{12} plane giving zero state-variables correspond approximately in position to those on the β_{21}— β_{22} plane. These points on both planes are approximately symmetrical about the β_{i2} axis.

(5) Both the zero-state-variable points and the non-zero-state variable points that give no LCO do not seem to bear any simple relationship to each other.

(6) A change in the input, say, from a unit impulse to a step function, or even a change in the input magnitude may change the position of the zero LCO points on the β_{i1}— β_{i2} plane.

(7) For a different given single-rate filter, generally, a different set of zero LCO point is to be found on the β_{i1}— β_{i2} plane. However, these different sets of zero-LCO points apparently bear no simple relationship to each other.

(8) If the quantization errors in the coefficients are severe, then even if we start with a double-rate filter with no LCO, the resultant filter may possess LCO in its output.

The following is an example showing the process of searching for the equivalent double-rate filters with zero-LCO:-

Example VII.1

It is desired to have a digital filter the transfer function of which is given by

$$H(z) = \frac{1}{1-1.4z^{-1}+0.8z^{-2}} \qquad \text{(VII.27)}$$

The quantization step is 0.01.
Using the simulation program, search for the equivalent double-rate filters which offer no LCO in the output.

It can be seen that the coefficients of the periodically varying double-rate filter must satisfy the following equations:-

$$\begin{aligned} \beta_{12} + \beta_{22} - \beta_{11}\beta_{21} &= -1.4 \\ \beta_{12}\beta_{22} &= 0.8 \\ \alpha_{20} &= -1/\beta_{21} \end{aligned} \qquad \text{(VII.28)}$$

The information of these coefficients, a_o, b_1 and b_2, and the quantization error E_o = 0.01 is fed into the computer. Fig.VII.5 shows the print out of the contour map on part of the triangle of stability on the β_{11}— β_{12} plane.

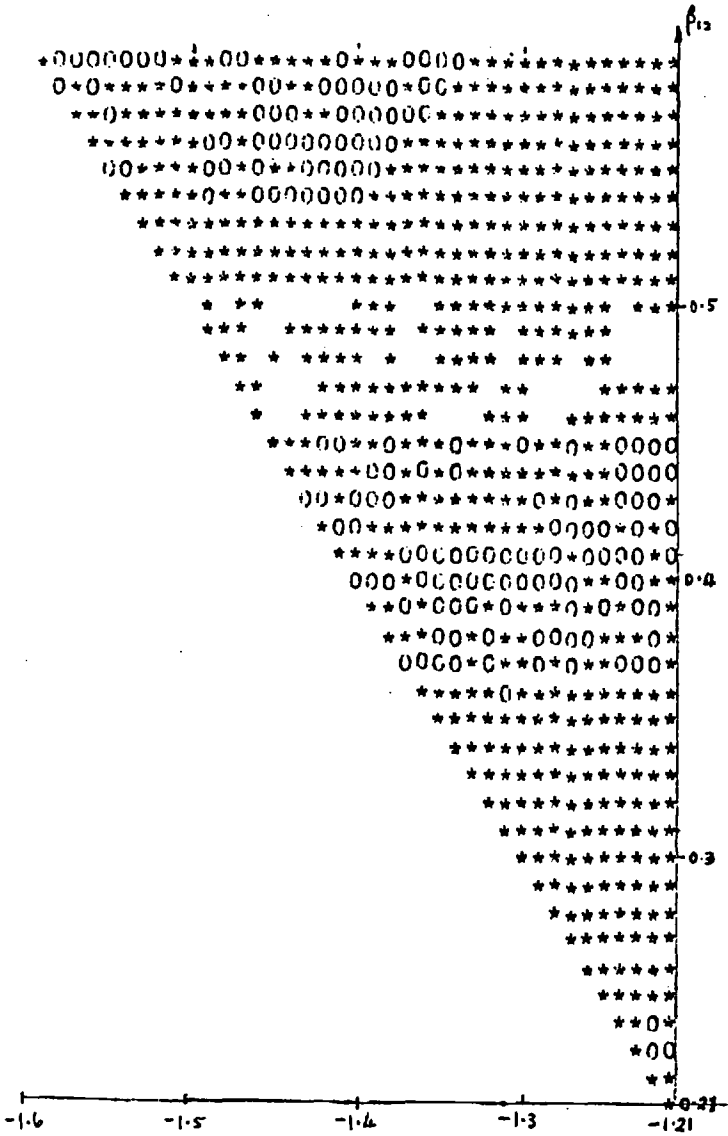


FIG VII.5

COMPUTER PRINT-OUT OF PART OF THE MAP OF SEARCHING FOR ZERO LCO REALIZATIONS ON THE β_{11}- β_{12} PLANE

* — LCO EXISTS

0 — NO LCO IN O/P OF FILTER YET STATE-VARIABLE NON-ZERO

BLANK — ZERO STATE-VARIABLES THUS ZERO LCO

Due to the difficulty in reproducing the computer print-out of the whole plane, the map of the "stability triangles" on the β_{11}— β_{12} and β_{21}— β_{22} planes are redrawn in fig VII.6(a) and (b),

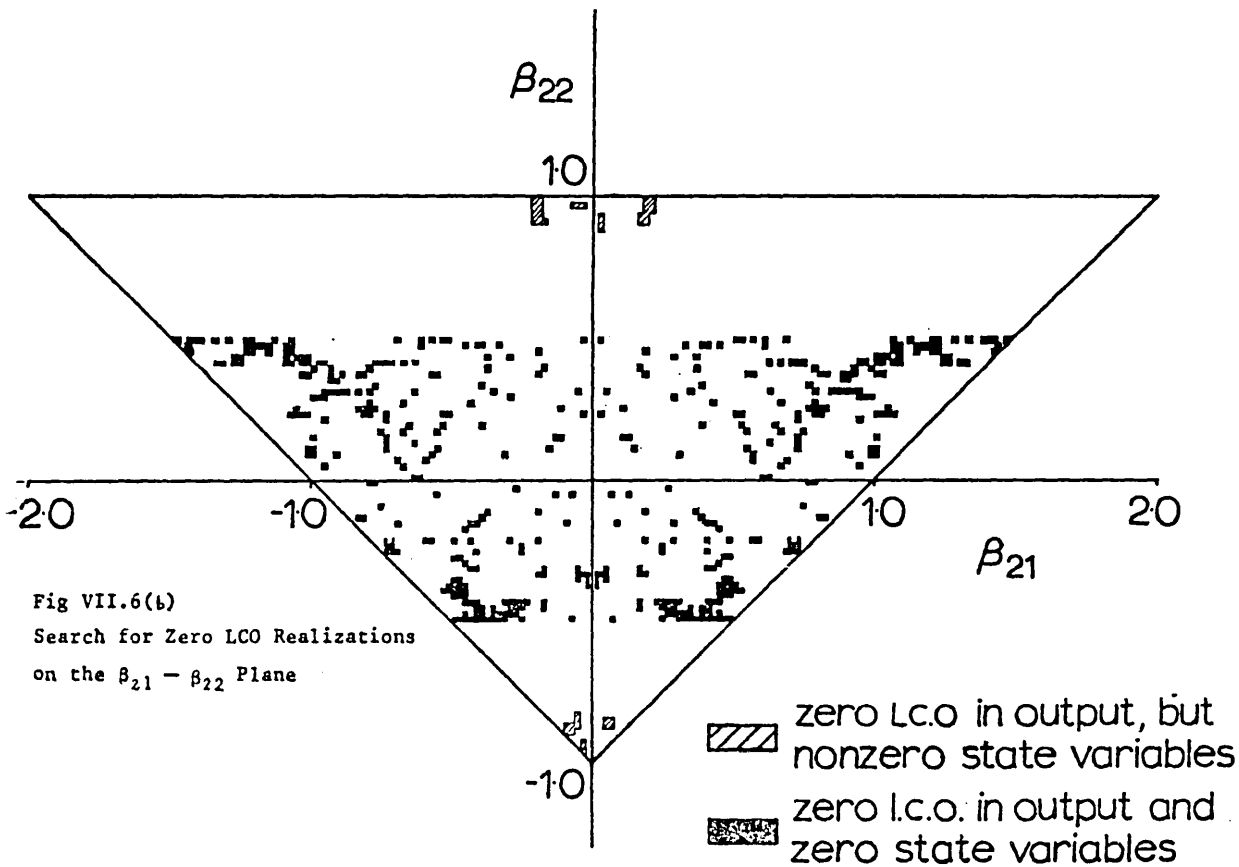


Fig VII.6(b)
Search for Zero LCO Realizations
on the $\beta_{21} - \beta_{22}$ Plane

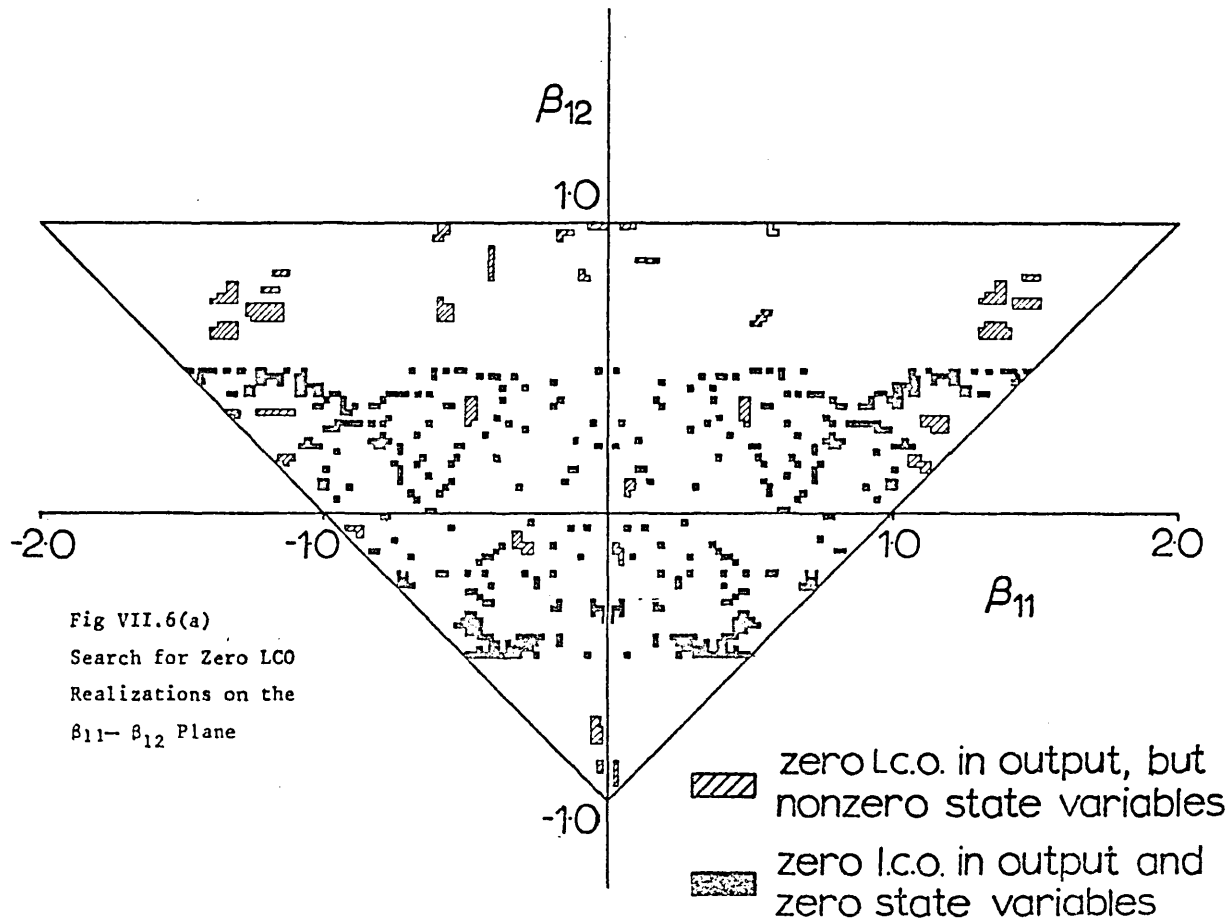


Fig VII.6(a)
Search for Zero LCO Realizations on the β_{11}– β_{12} Plane

As can be observed from the maps, there are thousands of equivalent double-rate digital filters with periodically varying coefficients that possess no LCO in its impulse response. Fig VII.7 shows the response of such a filter. Together on the graph is the original single-rate digital filter. It can be seen that LCO is completely absent in the output of the double-rate filter. The coefficients of the single-rate filter and the zero-LCO double-rate filter are printed on the diagram.

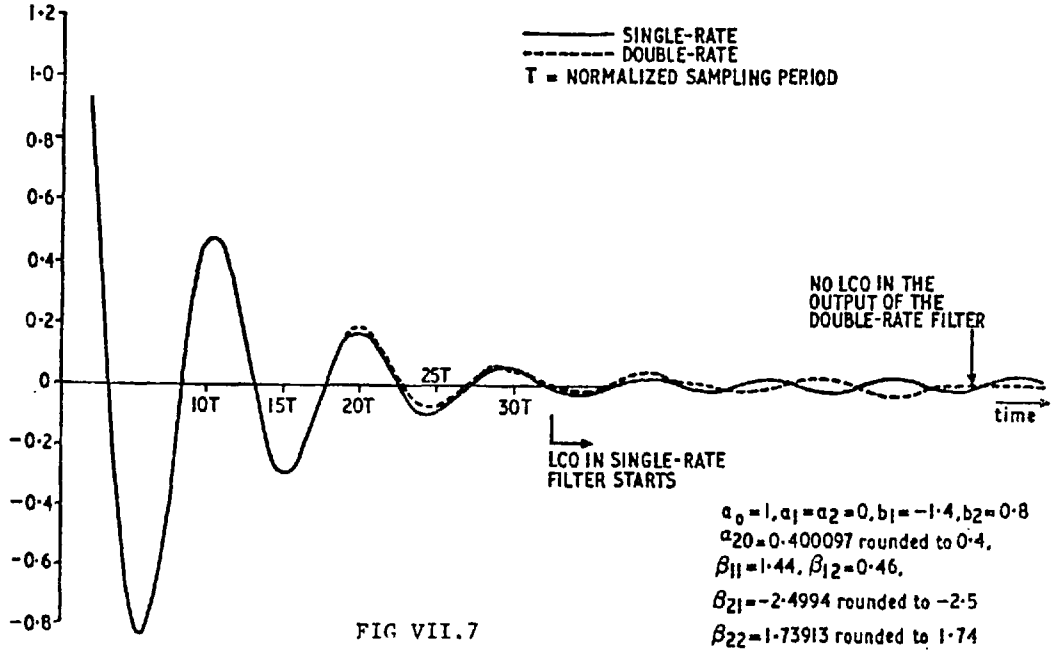


FIG VII.7

IMPULSE RESPONSES OF A DIGITAL FILTER AND ITS EQUIVALENT ZERO-LCO DOUBLE RATE FILTER

Obviously among these thousands of equivalent double-rate filters which give the same impulse response as the given single-rate filter and yet there is no LCO in the output, some must be more preferred than others as far as realization is concerned. Firstly, the periodically varying coefficients should be of the

same order. Looking through the list of all the possible double-rate filters, one finds some of the filters have vast differences between its coefficients, for example

$$\alpha_{20} = -0.003$$
$$\beta_{11} = -0.20$$
$$\beta_{12} = -0.01$$
$$\beta_{21} = 393.05$$
$$\beta_{22} = -80.00$$

Comparing the values of these coefficients to those of the filter chosen in the previous example (fig VII.7), i.e.

$$\alpha_{20} = 0.4$$
$$\beta_{11} = 1.44$$
$$\beta_{12} = 0.46$$
$$\beta_{21} = -2.5$$
$$\beta_{22} = 1.74$$

although both double-rate filters give no LCO in their impulse responses, the latter is certainly preferred from the point of view of fixed-point hardware implementation, and from the point of view of sensitivity. (Section V.3, figs V.2 and V.3).

Secondly, although β_{11} and β_{12} are in increments of one quantization step, i.e. there is no quantization error in these coefficients, the values of β_{21}, β_{22} and α_{20} calculated from eqns (VII.21) (VII.22) (VII.23) may have to be rounded off. If the round-off errors of these coefficients are severe, the resultant filter may have LCO in the output again. The following example may help to illustrate this point.

Example VII.2

The transfer function

$$H(z) = \frac{1}{1-1.4\,z^{-1}+0.8z^{-2}} \qquad \text{(VII.29)}$$

can be realized by the following double-rate filter which gives no LCO in its impulse response.

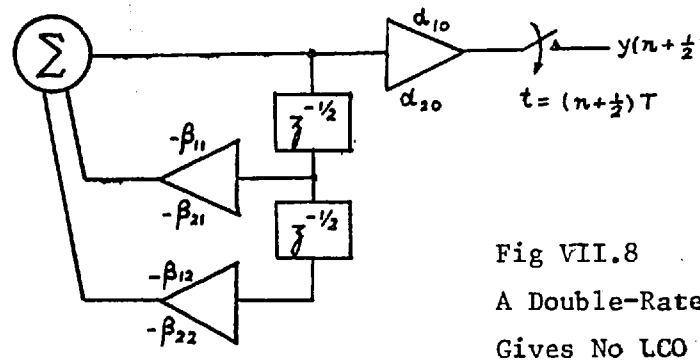


Fig VII.8
A Double-Rate Filter that Gives No LCO in its Impulse Response

where $\alpha_{10} = 0$ $\quad \beta_{11} = -0.03$ $\quad \beta_{12} = -0.38$

$\alpha_{20} = -0.0276431$ $\quad \beta_{21} = 36.1754$ $\quad \beta_{22} = -2.10526$

The impulse response of this filter is shown by the full line in fig VII.9. However, the coefficients α_{20}, β_{21} and β_{22} have to be rounded-off to two places of decimal if they were to be implemented. Thus the final filter would have the following coefficients

$\alpha_{10} = 0$ $\quad \beta_{11} = -0.03$ $\quad \beta_{12} = -0.38$

$\alpha_{20} = -0.03$ $\quad \beta_{21} = 36.18$ $\quad \beta_{22} = -2.11$

and the impulse response of this filter is shown by the broken line in fig VII.9. It can be seen that LCO exists in this filter if the coefficients are quantized.

It is observed that, in general, if the coefficients of the double-rate digital filter are not greatly different in their values, the suppression of LCO in the impulse response is not so easily disturbed by the quantization of the coefficients.

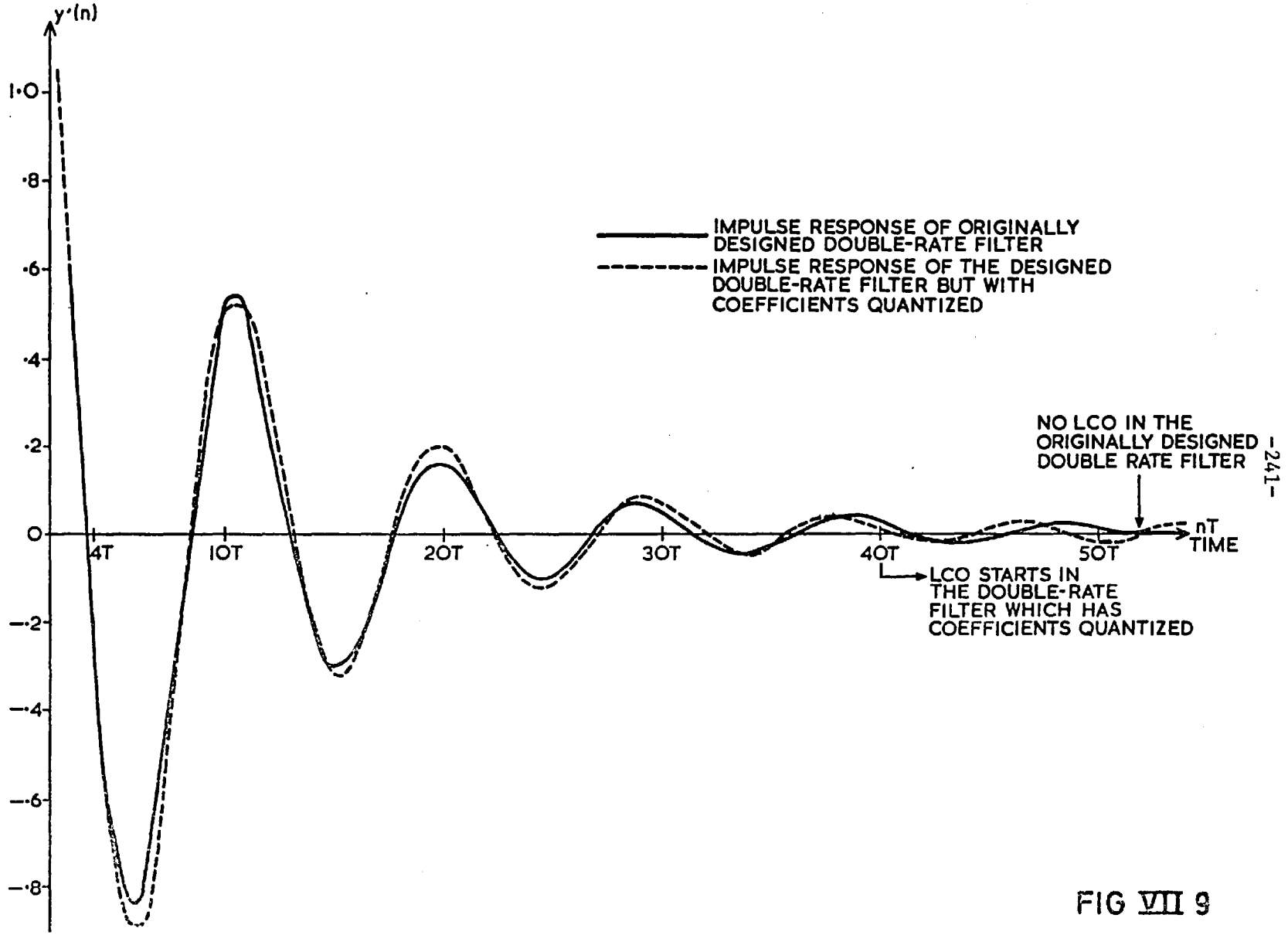

y'(n)
1·0
·8
·6
·4
·2
0
-·2
-·4
-·6
-·8
4T
10T
20T
30T
40T
50T
nT
TIME
IMPULSE RESPONSE OF ORIGINALLY DESIGNED DOUBLE-RATE FILTER
IMPULSE RESPONSE OF THE DESIGNED DOUBLE-RATE FILTER BUT WITH COEFFICIENTS QUANTIZED
NO LCO IN THE ORIGINALLY DESIGNED DOUBLE RATE FILTER
LCO STARTS IN THE DOUBLE-RATE FILTER WHICH HAS COEFFICIENTS QUANTIZED

FIG VII 9

It has been mentioned before that a double-rate filter that suppresses LCO for a unit impulse input may not suppress LCO for a step input. The following example illustrates this point.

Example VII.3

The double-rate filter shown in fig VII.8 will realize the transfer function

$$H(z) = \frac{1}{1 - 1.4\, z^{-1} + 0.8\, z^{-2}}$$

and yet gives no LCO in the impulse response if the following values are chosen for the periodically varying coefficients:-

$\alpha_{10} = 0$ $\quad \beta_{11} = -0.92$ $\quad \beta_{12} = 0.18$

$\alpha_{20} = 0.1527$ $\quad \beta_{21} = -6.548$ $\quad \beta_{22} = 4.444$

The impulse response of the filter is shown in fig VII.10(a) and can be seen, there is no LCO in the output. However, if the input is a step function, the response (fig VII.10(b)) will have LCO.

VII.6 Comparison of the two methods of Suppressing LCO

It has been mentioned in the beginning of this chapter that in trying to break up LCO in a digital filter, a small random noise, usually of the magnitude of the least significant digit, is added to the input of the filter. This section describes a computer program which simulate such a method when applied to a single-rate filter. The resultant output is compared with that obtained by an equivalent double-rate filter giving no LCO.

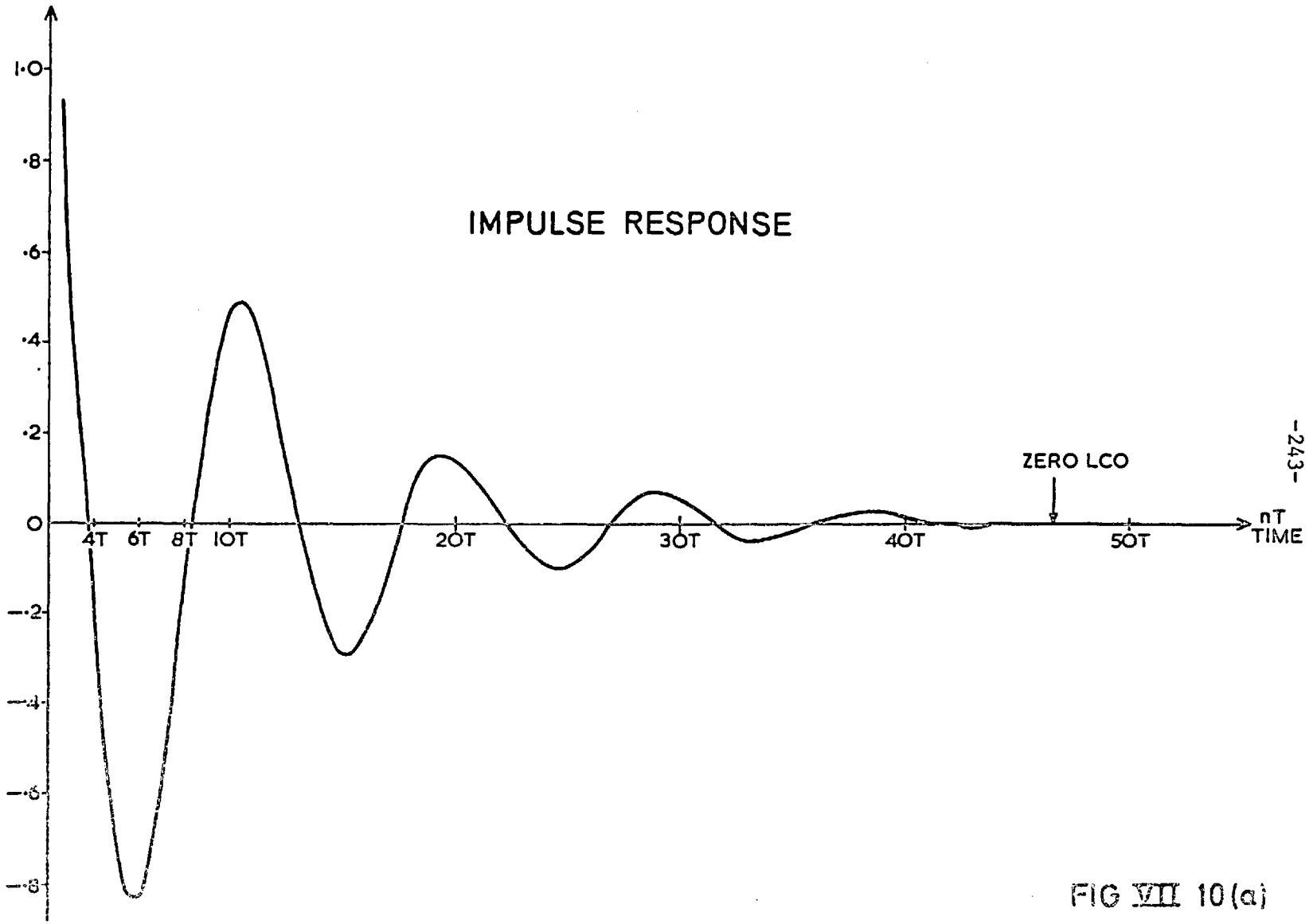

IMPULSE RESPONSE
1·0
·8
·6
·4
·2
0
—·2
—·4
—·6
—·8
4T
6T
8T
10T
20T
30T
40T
50T
nT
TIME
ZERO LCO

FIG VII 10(a)

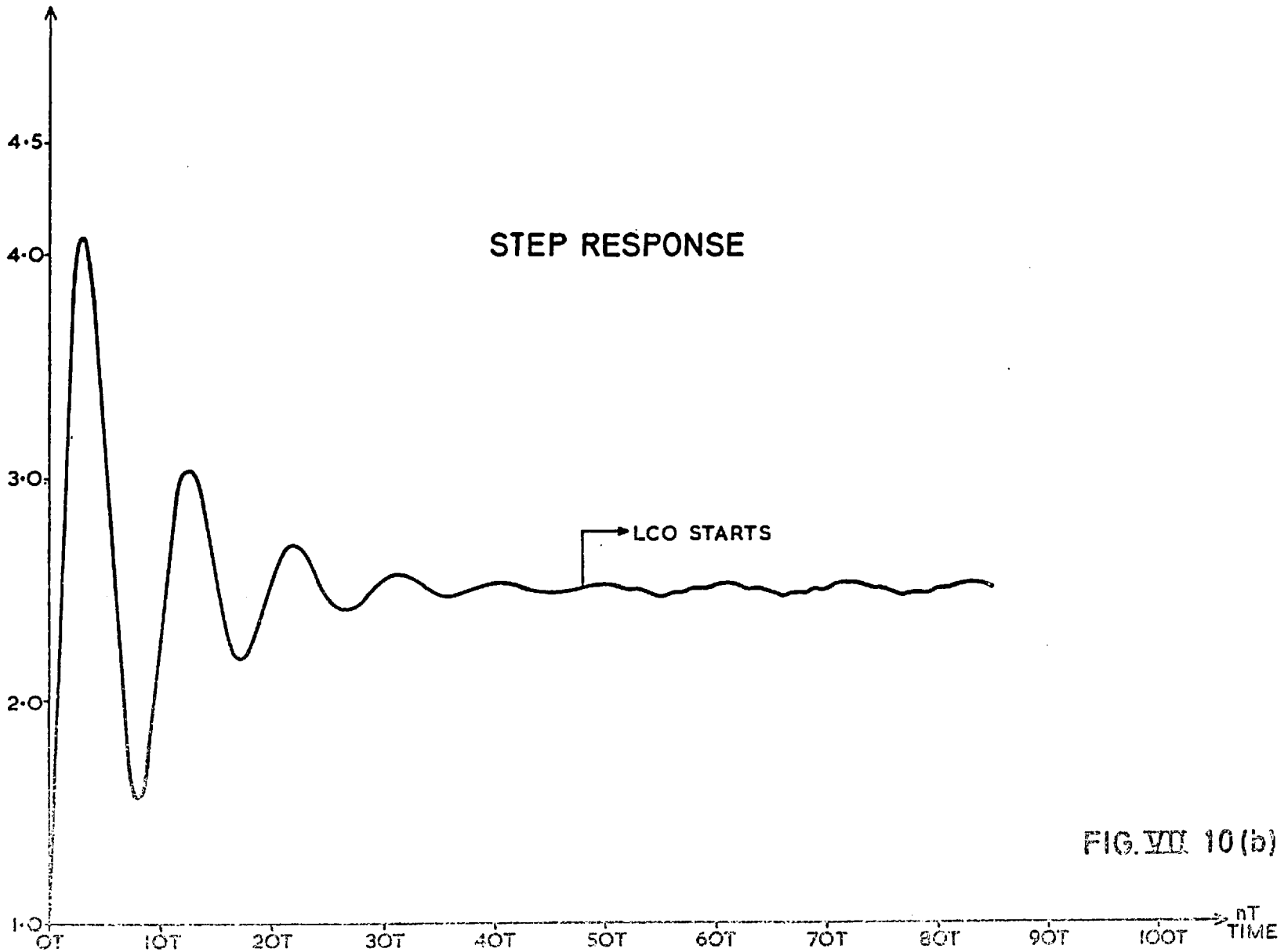

STEP RESPONSE
4·5
4·0
3·0
2·0
1·0
LCO STARTS
0T
10T
20T
30T
40T
50T
60T
70T
80T
90T
100T
nT
TIME

FIG. VII 10(b)

The program performs the following:-

1) A unit impulse is injected into an ideal single-rate digital filter and the output is recorded.
2) A unit impulse is injected into a single-rate digital filter with multiplication round-off errors; the output (up to a steady state) is recorded.
3) A unit impulse with added noise (the magnitude of which is confined to the least significant digit) is passed into the non-ideal single-rate filter, and after a while, the added noise is stopped. The response of the filter is recorded.
4) A unit impulse is injected into a non-ideal double-rate filter designed to give the same performance as the single-rate filter but without LCO. Again the response is recorded.

It has been observed that, in general, the added noise in the input breaks up the regular pattern of the limit cycle oscillations that would have existed in the impulse response of a single-rate digital filter with multiplication round-off errors. However, this method of dithering usually gives rise to additional noise in the response of the filter. Also if this added noise is stopped, LCO (which is generally different in magnitude to the LCO when there is no added noise in the input) will start again. The following example may help to show this.

Example VII.4

A single-rate digital filter having the transfer function

$$H(z) = \frac{1}{1 - 1.4z^{-1} + 0.8z^{-2}}$$

is implemented with multiplication round-off errors. If a unit impulse is passed into the filter, LCO would occur in the output. Suppression of such LCO is carried out by dithering. Compare the output of the single-rate filter when dithering is applied to the output of the double-rate filter designed to give no LCO.

The double-rate filter which has the same transfer function as $H(z)$ but gives no LCO has been chosen to have the following coefficients

$$\alpha_{10} = 0 \qquad \beta_{11} = 1.44 \qquad \beta_{12} = 0.46$$
$$\alpha_{20} = 0.4 \qquad \beta_{21} = -2.5 \qquad \beta_{22} = 1.74$$

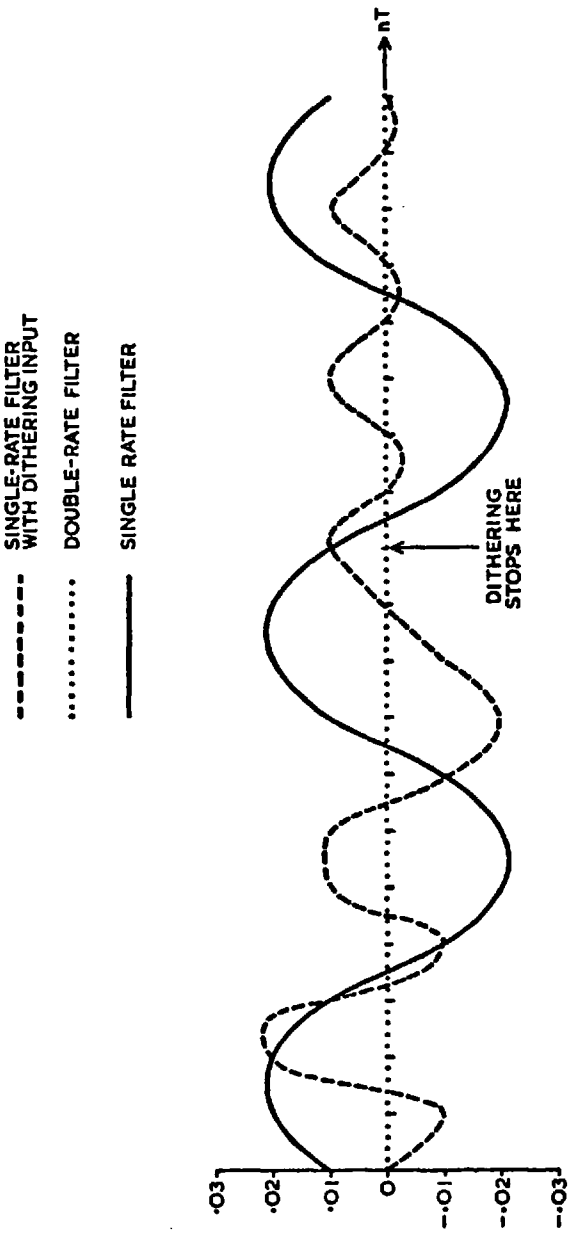
The steady state responses of the three cases are shown in fig VII.11. It can be seen that LCO can be totally suppressed in the case of the double-rate filter while the dithering method can only break up the regular pattern. LCO starts again in the single-rate filter immediately after the dithering stops.

Hence it can be concluded that the method of using double-rate filters to suppress LCO is very much more effective than the method of dithering.

VII.7 Résumé

The nature of steady state limit cycle oscillations in a digital filter was briefly described. The conditions for the existence of LCO were stated and the bound discussed.

It has been suggested that a small random noise should be added to the input to break up the LCO in a digital filter. However, it has been found that a multi-rate digital filter (here,



STEADY-STATE OF THE FILTERS

FIG VII 11

a double-rate filter has been solely considered) would suppress LCO completely. In general, it has been found that for a given transfer function, there are many equivalent double-rate filters that suppress LCO. The filter should be chosen such that the values of the coefficients should not differ too greatly, nor should the quantization of the coefficients be large enough to affect the suppression of LCO.

If a proper choice of such double-rate dilters have been made it has been found that this method of suppressing LCO is very much more effective than the dithering method.

CHAPTER VIII

CONCLUSIONS

VIII.1 General Summary

The main object of this thesis is to investigate the principal properties of multirate digital filters. Due to the nature of the device, more especially in the case of multirate filters with periodically varying coefficients, it has been found that the analysis is considerably facilitated by the use of state-space methods.

After an introduction to digital filters in general, and a brief account of the state-space method of analysis, a mathematical model of the multirate digital filter is developed. The method of developing this model has been shown to be more versatile than those using conventional methods since it could be applied to time-varying or time-invariant filters, and to filters of different configurations without any modification of the method. Using this model, some interesting properties of the transfer functions of the multirate filter can be derived. The realization of an equivalent multirate filter from an original single-rate filter is straightforward and almost trivial. But since there exist various possible designs, the choice has been discussed from the point of view of economy and performance.

Quantization errors are the main factors affecting the performance of a digital filter. It is on the basis of these errors that the multirate filter is compared to its equivalent single-rate filter. A rigorous mathematical analysis has shown that the A/D

conversion errors are identical in both the single-rate and its equivalent multirate filters. Simulation results also support this analysis. However, the errors due to the quantization of coefficients and rounding of multiplication products are different in the two devices.

Mathematical analysis, confirmed by computer simulations,shows that in general, the poles of a second order multirate filter are least sensitive to the quantization of coefficients when the filter is time-invariant. The pole sensitivity of a time-invariant multirate filter is compared to that of a single-rate filter based on a novel criterion — the sensitivity ellipse, and it is found that in some regions where the single-rate filter performance is vulnerable to coefficient quantization, its equivalent multirate filter can be used giving rise to a much less sensitive realization. Again, the superior performance of the multirate filter in these regions has been confirmed by computer simulations.

The effect of the rounding of multiplication products can be treated in two ways. If the input signal is stochastic, it is generally more realistic to evaluate the error statistically. Using state-space methods, the statistical estimation of the errors due to multiplication roundoff have been evaluated in both single-rate and multirate filters, and it has been found that in some regions, the multirate filter is superior to its single-rate counterpart. On the other hand, if the input is deterministic, the multiplication roundoff error usually leads to a steady-state limit cycle oscillation which is detrimental to most engineering applications of digital filters. However, using an equivalent multirate digital filter with periodically varying coefficients, such oscillations can generally be suppressed completely. This method of suppressing LCO in a digital filter by using its equivalent multirate realization has been confirmed to be successful and shown to be more ef-

fective than the existing method of dithering.

With these properties, the advantages of applying multirate digital filters are apparent. When an ordinary single-rate filter is found too sensitive to the quantization of coeeficients, or when it is found to yield too high a noise due to multiplication roundoff errors, a multirate digital filter, which has a much greater degree of freedom in the choice of its coefficients, can be considered as an alternative, and it can be assured to give a more satisfactory performance if the conditions given in Chapters V and VI are fulfilled. Perhaps the greatest use of multirate filters lies in that they can suppress limit cycle oscillations completely provided that the coefficients are properly chosen. Since the conventional single-rate filters that are likely to be encountered will almost certainly give LCO in the output for a deterministic input, the use of the equivalent multirate filter would be most welcome if LCO give an undesirable effect.

Finally, since the poles and zeros of a multirate digital filter are interrelated, it can be applied to the construction of variable filters the characteristics of which are controlled by one single multiplier. This idea was first put forth by Fjällbrant and has been explained in detail {13}.

VIII.2 Some Open Questions and Suggestions for Further Research

Although many properties of the multirate digital filters have been revealed through analyses and simulations, the research work is far from being complete. There remain many question yet unanswered and some may be interesting and challenging enough to stimulate further research:

(1) In dealing with the effects of quantizing the multipliers in Chapter V, only the pole sensitivities of the filters were considered. But what would be the effects of the zeros of the transfer function? The non-linear and complicated relationship between the zeros and the multipliers of a multirate filter (see section III.5) renders the problem very difficult, if at all fruitful, for analysis. But perhaps with the aid of computer simulations, one may be able to estimate the movements of the zeros caused by the multiplier quantizations. If the relative movements of the poles and the zeros are known, it should be helpful to estimate the change in the sharpness of the cut-off and in the translation of the resonant frequency.

(2) The comparison in Chapter VI of the statistical errors due to multiplication quantization in a multirate and in a single-rate filter was based on the assumption that the filters are highly selective, i.e. the poles are very closed to the unit circle. Clearly, if this is not the case, the analysis would be different. Can a clear-cut comparison be possible if the poles are not so closed to the unit circle such that ε^2 cannot be ignored?

(3) The problem of limit cycle oscillations in a multirate digital filter leaves many unanswered questions and perhaps even opens a wide field of research. Chapter VII has only demonstrated that LCO can be totally suppressed by periodically varying multirate filters for one type of deterministic input of a paricular magnitude, viz. a unit impulse. It has been found that the set of double-rate filters which suppresses LCO for a unit impulse is different from the set that suppresses LCO for a unit step. What is the relationship between the two sets, or indeed, between the different sets for different types of deterministic inputs?

On the other hand, if a set of double-rate filters give no LCO

for a particular deterministic input, is there any relationship between these filters within the set?

For a unit impulse, a set of double-rate filters would suppress LCO. But if the magnitude of the impulse changes, most members of the set would still suppress LCO, but for a few of them, LCO would arise again. What is the reason for this? Would the same happen if the input is some deterministic signal other than the unit impulse?

What would happen if the number of shift sequences within a sampling period is greater than two, i.e. $N > 2$? Would the same observations that are discovered in the double-rate filter still hold? Are there any advantages over the double-rate filter if $N > 2$ as far as suppression of LCO is concerned?

In general, if a single-rate filter exhibits LCO for a particular input, its equivalent time-invariant multirate filter would exhibit LCO as well. Is there any relationship between these oscillations, say, in their amplitudes, frequencies or harmonic contents?

(4) Throughout the whole thesis, attention has been focused on second order filters realized in the direct canonic form. Recently, many other configurations have been suggested by research workers in the field of digital filtering [9],[12]. In section III.8, it has been shown that the same technique of developing a mathematical model can be applied to filters of other configurations. But can the same advantages in pole sensitivity, multiplication quantization errors and suppression of LCO offered by the second order direct canonic form be found in other configurations?

(5) It has been tacitly assumed throughout the whole thesis that the number of shift sequences, N, in a sampling period is a positive integer. If the view is widened, and if N is taken to be a fraction or even an irrational number, what would the properties of such a "multirate" filter be?

The above questions suggest a few fields for further research; some, like the first two, are short-term and specific, while the others are of longer terms and more general. Until these questions are answered, the work on multirate digital filters is still far from complete, by which time, I am sure, other questions will arise and the frontier of research on the subject will be pushed still further.

Digital filters are not yet widely used in industry at the present moment, the main reason probably being the cost of the device. With the rapid advances in integrated circuit technology, it appears that in the not too distant future digital filters will be economically possible for implementation. However, while it is certainly true that a lot of significant work has been done in the area of digital filtering, it is my opinion that the field is still very much in an embryonic state in the sense that several basic questions have remained yet unanswered, and new areas, of which the present work may be regarded as but one of numerous, yet unexplored. Nevertheless, with so much research effort being exerted in this important field, each contributing some new ideas and discoveries in the vast realm of knowledge upon which the progress of science is based, our knowledge in this area will soon be much more sophisticated and refined.

VIII.3 CODA — On the Value of Scientific Research

Newton once remarked, "*Hypothesis, whether metaphysical or physical, whether of occult qualities or mechanical, has no place in experimental philosophy*". This is certainly a very high standard. But is it the least that every man must break through to avoid mediocrity? Should the value of scientific research be merely measured by how big a step of progress the work has carried forward? Lao Tse, who flourished some twenty-two centuries before Newton, put forth a far more convincing and encouraging philosophy, "..., *a mountain is built of individual grains, a journey of a thousand miles is made up of small steps, ..., so regard not your house too empty and your room too poor*"†. Should this be a more suitable outlook? I wonder.

Furthermore, in estimating the value of scientific research, it is my opinion that too much importance has been attached to the acquisition of power by using the new knowledge — power of an individual over another, power of one group over another, power of one nation over another. But scientific research is not itself at fault. Knowledge is good and ignorance is evil; to this principle the lover of the world can admit no exception. Nor is it power in and for itself that is the source of danger. What is dangerous is power wielded for the sake of power, not power wielded for the sake of genuine good. Power is not one of the ends of life, but merely a means to other ends, and until men remember the ends that power should subserve, science will not do what it might to minister to "the good life". Science may bestow joys and beauties of life upon more people than could otherwise enjoy them. If so, its power will be wisely used. But when it takes out of life

†Author's own translation

the moments to which life owes its value, science will not deserve admiration, however cleverly and however elaborately it leads men along the road to despair. The sphere of values lies outside science except in so far as science consists in the pursuit of knowledge. Science as the pursuit of power must not obtrude upon the sphere of values, and scientific technique, if it is to enrich human life, must not outweigh the ends which it should serve. The new powers that science has given to man can only be utilized safely by those who, whether through the study of history or through their own experience of life, have acquired some reverence of human feelings and some tenderness towards the emotions that give colour to the daily existence of men and women. This to me, is a subtle and often elusive state of aesthetical and spiritual values that are embedded in the studies and professions of science, and it is in this spirit that this research project has been performed and that future works are hoped to be carried out.

In conclusion, I would like to re-echo by quoting Bertrand Russell {55}:

"*Knowledge and feeling are equally essential ingredients both in the life of the individual and in that of the community. Knowledge, if it is wide and intimate, brings with it a realization of distant times and places, an awareness that the individual is not omni-potent or all-important and a perspective in which values are seen more clearly than by those to whom a distant view is impossible. Even more important than knowledge is the life of the emotions. A world without delight and without affection is a world destitute of value. These things the scientific manipulator must remember, and if he does his manipulation may be wholly beneficial. All that is needed is that men should not be so intoxicated by new power as to forget the truths that were familiar to every previous generation. Not all wisdom is new, nor is all folly out of date.*

Man has been disciplined hitherto by his subjection to nature. Having emancipated himself from this subjection, he is showing something of the defects of slave-turned-master. A new moral outlook is called for in which submission to the powers of nature is replaced by respect for what is best in man. It is where this respect is lacking that scientific technique is dangerous. So long as it is present, science, having delivered man from bondage of nature, can proceed to deliver him from bondage to the slavish part of himself. The dangers exist, but they are not inevitable, and hope for the future is at least as rational as fear."

REFERENCES

1. Aggarwal, J.K., "Input Quantization and Arithmetic Roundoff in Digital Filters - a Review", in Skwirzinski, J.K., and J.O. Scanlan (Eds). "Network and Signal Theory" Peter Peregrinus, London 1973.

2. Atiya, F.S., A.Y. Bilal and I.E. Abdou, "Multiple Shift Sequences in Digital Filter Design" *IEEE Trans Circuits and Systems*, vol CT-21, January 1974.

3. Bennette, W.R., "Spectra of Quantized Signals" *Bell Syst Tech J*, vol 27, July 1948.

4. Bertram, J.E., "The Concept of State in the Analysis of Discrete-time Control System", *Proc JACC*, June 1962.

5. Blackman, R.B., "Linear Data-Smoothing and Prediction in Theory and Practice", Addison-Wesley, Reading, Mass., 1965.

6. Blakey, J., "University Mathematics", Blackie and Son Limited London 1961.

7. Camrass, R.J., "Implementation of a Second Order Digital Filter", *Report No. 97/73/16/TR, Plessey Telecommunications Research Ltd.* Taplow, Bucks.

8. Chang, T.L., and C.S. Burrus, "Oscillations Caused by Quantization in Digital Filters", *Proc. IEEE International Symposium on Circuit Theory*, April 1972.

9. Constantinides, A.G., "Some New Digital Filter Structures Based on Continued Fraction Expansion", in Skwirzinski, J.K.

and J.O. Scanlan (Eds.). "Network and Signal Theory", Peter Peregrinus, London 1973.

10. Dorf, R.C., "Time-Domain Analysis and Design of Control Systems", Addison-Wesley, 1965.

11. Duffin, R.J., E.L. Peterson and C. Zener, "Geometric Programming", Wiley, New York, 1967.

12. Fettweis, A., "Some Principles of Designing Digital Filters Imitating Classical Filter Structures", *IEEE Trans on Circuit Theory*, vol CT-18, March 1971.

13. Fjallbrant, T., "Digital Filters with a Number of Shift Sequences in Each Pulse Repitition Interval" *IEEE Trans on Circuit Theory*, vol CT-17, August 1970.

14. Forsythe, G., and C.B. Moler, "Computer Solution of Linear Algebraic Systems", Prentice-Hall, Englewood Cliffs, N.J. 1967.

15. Frazer, R.A., W.J. Duncan and A.R. Collar, "Elementary Matrices", Cambridge University Press, England, 1963.

16. Freeman, H., "Discrete-Time Systems", Wiley, New York, 1965.

17. G-AE Concepts Subcommittee, "On Digital Filtering", *IEEE Trans on Audio and Electroacoustics*, vol AU-16, No. 3, September 1968.

18. Gantmacher, F.R., "Matrix Theory", vol 1, Chelsea, N.Y., 1965.

19. Gerlach, A.A., "A Time-Variable Transform and its Application to Spectral Analysis", *IRE Trans on Circuit Theory CT-2*, March 1955.

20. Gold, B., and C.M. Radar, "Effects of Quantization Noise in Digital Filters", *AFIPS Proc*, vol 28, Spring Joint Computer Conference, 1966.

21. Gold, B., and C.M. Radar, "Effects of Parameter Quantization on the Poles of a Digital Filter", *Proc IEEE*, vol 55, May 1967.

22. Gold, B., and C.M. Radar, "Digital Processing of Signals". McGraw-Hill, New York, 1969.

23. Gupta, S.C., "Transform and State-Variable Methods in Linear Systems", John Wiley and Son, N.Y., 1966.

24. Halmos, P.R., "Finite Dimensional Vector Spaces", D van Nostrand, Princeton, N.J., 1958.

25. Jackson, L.B., "An Analysis of Roundoff Noise in Digital Filters" *Sc.D. Dissertation, Stevens Institute of Technology*, Hoboken, N.J., 1969.

26. Jackson, L.B., "An Analysis of Limit Cycles due to Multiplication Rounding in Recursive Digital (Sub) Filters", *Proc. 7th Annual Allerton Conference on System and Circuit Theory*, 1969.

27. Jullien, G.A., "A General Analysis for the Time-Invariant Multirate Digital Filters", *IEEE Trans on Circuit Theory*, vol CT-20, July 1973.

28. Jury, E.I., "Sampled-Data Control Systems", Wiley, N.Y. 1958.

29. Jury, E.I., "Theory and Application of the Z-Transform Method", Wiley, N.Y., 1964.

30. Kaiser, J.F., "Some Practical Considerations in the Realization of Linear Digital Filters", *Proc 3rd Annual Allerton Conference on Circuit and System Theory*, 1965.

31. Kalman, R.E., and J.E. Bertram, "A Unified Approach to the Theory of Sampling Systems", *J. Franklin Inst.* vol 267, 1959.

32. Kalman, R.E., and J.E. Bertram, "Control System Analysis and Design via the Second Method of Liapunov: Discrete-Time Systems", *J. Basic Eng.* ser.D. vol 82, 1960.

33. Kaneko, T., and B. Liu, "Round-off Error of Floating-Point Digital Filters", *Proc 6th Annual Allerton Conference on Circuit and System Theory*, 1968.

34. Kaplan, W., "Advanced Calculus", Addison-Wesley, Mass. 1952.

35. Kendall, M.G., and A. Stuart, "The Advanced Theory of Statistics", vol 2 Charles Griffin & Co. Ltd. London 1961.

36. Knowles, J.B.,and R. Edwards, "Effect of a Finite Word Length Computer in a Sampled-Data Feedback System", *Proc IEEE* vol 112, June 1965.

37. Knowles, J.B., and R. Edwards, "Complex Cascade Programming and Associated Computational Errors", *Electronic Letters*, vol 1, August 1965.

38. Knowles, J.B., and E.M. Olcayto, "Coefficient Accuracy and Digital Filter Response", *IEEE Trans on Circuit Theory*, vol CT-15, March 1968.

39. Kuo, B.C., "Analysis and Synthesis of Sampled-Data Control Systems", Prentice-Hall, Englewood Cliffs, N.J., 1963.

40. Kuo, B.C., "Linear Networks and Systems", McGraw-Hill, N.Y., 1967.

41. Kuo, B.C., "Discrete-Data Control Systems", Prentice-Hall, Englewood Cliffs, N.J., 1970.

42. Kuo, F.F., and J.F. Kaiser, "System Analysis by Digital Computers", Wiley, N.Y., 1969.

43. Lampard, D.G., "The Response of Linear Networks to Suddenly Applied Stationary Random Noise", *IRE Trans on Circuit Theory*, vol CT-2, March 1955.

44. Lindorff, D.P., "Theory of Sampled-Data Control Systems", Wiley, N.Y., 1965.

45. Liu, B., "Effect of Finite Word Length on the Accuracy of Digital Filters", *IEEE Trans on Circuit Theory*, vol CT-18, November 1971.

46. Mantey, P.E., "Eigen-value Sensitivity and State-Variable Selection", *IEEE Trans on Automatic Control*, vol AC-13, June 1968.

47. Novak, D.J., and P.E. Schmid, "Introduction to Digital Filters", *IEEE Trans on Electromagnetic Compatibility*, vol EMC-10, June 1968.

48. Oppenheim, A.V., "Realization of Digital Filters using Block-Floating-Point Arithmetic", *IEEE Trans on Audio and Electro-acoustics*, vol AU-18, June 1970.

49. Parker, S.R., and S.F. Hess, "Limit-Cycle Oscillations in Digital Filters", *IEEE Trans on Circuit Theory*, vol CT-18, November 1971.

50. Pipes, L.A., "Four Methods for the Analysis of Time-Variable Circuits", *IRE Trans on Circuit Theory*, vol CT-2, March 1955.

51. Pipes, L.A., "Matrix Methods for Engineering", Prentice-Hall, Englewood Cliffs, N.J., 1963.

52. Radar, C.M., and B. Gold, "Digital Filter Design Techniques in the Frequency Domain", *Proc IEEE*, vol 55, February 1967.

53. Ragazzini, J.R., and G.F. Franklin, "Sampled-Data Control Systems", McGraw-Hill, N.Y., 1958.

54. Rice, S.O., "Mathematical Analysis of Random Noise", *Bell Syst. Tech J*, vol 23, July 1944.

55. Russell, B., "The Scientific Outlook", George Allen and Unwin, London, 1962.

56. Sandberg, I.W., and J.F. Kaiser, "A Bound on Limit Cycles in Fixed Point Implementations of Digital Filters", *IEEE Trans on Audio and Electroacoustics*, vol AU-2, June 1972.

57. Shilov, G.E., "Theory of Linear Spaces", Prentice-Hall, N.J., 1961.

58. Thomson, W.E., Private Communications, Telecom HQ., GPO, Dollis Hill, London NW2, February 1973.

59. Timothy, L.K., and B.E. Bona, "State-Space Analysis: An Introduction", McGraw-Hill, N.Y., 1968.

60. Tou, J.T., "Digital and Sampled-Data Control Systems", McGraw-Hill, N.Y., 1959.

61. Weatherburn, C.E., "Mathematical Statistics", Cambridge University Press, England, 1962.

62. Weinstein, C., and A.V. Oppenheim, "A Comparison of Roundoff Noise in Floating Point and Fixed Point Digital Filter Realizations", *Proc IEEE*, vol 57, June 1969.

63. Widrow, B., "Statistical Analysis of Amplitude-Quantized Sampled-Data Systems", *AIEE Trans Appl Ind*, vol 79, January 1961.

64. Wilkinson, J.H., "Rounding Errors in Algebraic Processes", Prentice-Hall, Englewood Cliffs, N.J., 1963.

65. Wong, K.M., and R.A. King, "State-Space Description of a Multirate Digital Filter", *Electronic Letters*, vol 9, January 1973, and *Errata*, ibid, vol 9, January 1973.

66. Wong, K.M., "State-Space Approach to the Multiplication Error in a Digital Filter", *Electronic Letters*, vol 9, May 1973.

67. Wong, K.M., "Introduction to State-Space Analysis", *Report No. 97/73/29/TR, Plessey Telecommunications Research Ltd.* Taplow, 1973.

68. Wong, K.M., and R.A. King, "Pole Sensitivity of a Digital Filter with Multishift Sequences in Each Sampling Interval", *Electronic Letters*, vol 9, October 1973.

69. Wong, K.M., and R.A. King, "A Method to Suppress Limit-Cycle Oscillations in Digital Filters", *Electronic Letters*, vol 10, March 1974.

70. Yakowitz, S., and S.R. Parker, "Computation of Bounds for Digital Filter Quantization Errors", in Skwirzynski, J.K. and J.O. Scanlan (Eds). "Network and Signal Theory", Peter Peregrinus, London 1973.

71. Zadeh, L.A., "Frequency Analysis of Variable Networks", *Proc IRE*, vol 38, March 1950.

72. Zadeh, L.A., "Correlation Functions and Power Spectra in Variable Networks", *Proc IRE*, vol 38, November 1950.

73. Zadeh, L.A., and C.A. Desoer, "Linear System Theory — A State-Space Approach", McGraw-Hill, N.Y., 1963.