LONDON, METEOROLOGICAL OFFICE.

Met.O.16 Branch Memorandum No. 3.

# Implementation of digital filters using recursive techniques.

By LEE, A.C.L.

London, Met.Off., Met.O.16 Branch Mem. No.3, 1977, 31cm. Pp.21. 5 Refs.

Met O 16 Branch Memorandum No 3





# Implementation of Digital Filters Using Recursive Techniques

by A C L LEE

# Implementation of Digital Filters Using Recursive Techniques

## 1. INTRODUCTION

The intention of this paper is to describe how filters, such as the classical Butterworth, Chebychev, Bessel etc., or else control loop elements, can be implemented in economic digital hardware or software using a recursive technique. Some emphasis is placed on low-pass filters, although the methods are applicable to all types of filter subject only to the Nyquist constraint.

The results will be appreciated, and can be applied, most quickly if the theory sections 2, 3, 4 are skimmed or even omitted at a first reading as FOUR worked examples are given in section 5. The interested reader will find much of sections 2, 3, 4 described at greater length in Tou (1959).

## 2. SAMPLING

A digital filter implementation ideally samples instantaneous values of data from the input data stream $x(t)$ regularly every T seconds. We will assume that there is no energy in $x(t)$ with frequency above the Nyquist frequency $f = 1/2T$ Hy.

The resulting time sequence $x^*(t)$ can be represented by:

$$x^*(t) = x(t) \cdot \delta_T(t) \tag{1}$$

where $\delta_T(t)$ is the ideal sampling function:

$$\delta_T(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{2}$$

consisting of a series of Dirac Delta Functions $\delta(t)$ separated by T seconds, and:

$$\delta(\tau) = \infty \;,\; \tau = 0$$

$$\delta(t) = 0 \;,\; t \neq 0$$

$$\int_{t-\epsilon}^{t+\epsilon} \delta(t) \cdot d\tau = 1$$

defines the Dirac Delta Function. (3)

Note that in this form $x^*(\tau)$ is a continuous function, and so is still amenable to analysis.

The Ideal Sampling Function is periodic in t (period T), and so can be analysed into a complex Fourier series:

$$\delta_T(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega_s t} \tag{4}$$

where $\omega_s = \frac{2\pi}{T}$ and the $c_n$ s are the Fourier coefficients. Multiplying (4) by $e^{-jn\omega_s t}$ and integrating, we have:

$$c_n = \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} \delta_T(t) \cdot e^{-jn\omega_s t}$$

$$= \frac{1}{T} \int_{0^-}^{0^+} \delta(t) \cdot d\tau = \frac{1}{T}$$

as $\delta_T(t) = \delta(t)$ from $\tau = -\frac{T}{2}$ to $+\frac{T}{2}$ ; $\delta(t) = 0$ when $t \neq 0$

$$\therefore \delta_T(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} e^{jn\omega_s t} \tag{5}$$

From (1) and (5), the sampled data stream can be expressed in terms of the original data stream by:

$$x^*(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} x(t) \cdot e^{jn\omega_s t} \tag{6}$$

Thus the analytical form of $x^*(t)$ consists of the original $x(\tau)$ (i.e. $n=0$), multiplied by a gain factor $\frac{1}{T}$ ; together with other frequency bands similar to those defined by $x(\tau)$ but shifted in frequency by multiples of $\omega_s$ , and also multiplied by $\frac{1}{T}$ . The latter property is more easily seen by taking the Laplace

Transform of (6):

$$X^*(s) = \frac{1}{T}\sum_{n=-\infty}^{\infty} L\left\{x(t)\, e^{jn\omega_s t}\right\} = \frac{1}{T}\sum_{n=-\infty}^{\infty} X(s+jn\omega_s) \qquad (7)$$

The original data can thus be recovered completely from $x^*(t)$ with a low-pass filter having a sharp cutoff at the Nyquist Frequency $\frac{\omega_s}{2}$ and flat group delay, provided that there was no energy in $x(t)$ having frequencies above $\frac{\omega_s}{2}$. (Such a filter would have a long and complex impulse response and would be difficult to realise in practice, but this does not invalidate the argument. In practice we may be prepared to keep the bandwidth of $x(t)$ below the Nyquist frequency in exchange for satisfactory realiseable filters). If this is done with a unity gain low-pass filter, the original $x(t)$ is recovered with a gain factor $\frac{1}{T}$ due entirely to the sampling process.

3. FILTERING

(a) Basic Filtering

One possible reason for sampling is that we wish to modify, or filter, $x(t)$ using a digital system. This is more easily considered as a multiplication in frequency space of the input spectrum $X^*(s)$ by a filter characteristic expressed in sampled form to produce an output $Y^*(s)$.

We may require the filter to have the characteristic poles and zeroes of a continuous filter F(s), and the corresponding impulse response f(t)†. Sampling the impulse response as before, we may combine (1) and (2) (substituting f for x) to obtain:

$$f^*(t) = f(t) \sum_{n=-\infty}^{\infty} \delta(t-nT) \quad \dagger$$

or:
$$f^*(t) = \sum_{n=0}^{\infty} f(nT) \cdot \delta(t-nT) \quad \dagger \qquad (8)$$

By analogy with equation (6), this representation of F(t) gives a gain of $\frac{1}{T}$. Therefore we will realise $f(t)$ by $Tf^*(t)$:

Realise $f(t)$ by $Tf^*(t) = T \sum_{n=0}^{\infty} f(nT) \cdot \delta(t-nT)$ (9)

and the Inverse Laplace Transform:

Realise $F(s)$ by $TF^*(s) = T \sum_{n=0}^{\infty} f(nT) \cdot e^{-nsT}$ (10)

So that we have in the frequency domain

$$Y^*(s) = X^*(s) \cdot TF^*(s) = \frac{1}{T} Y(s) + \text{higher frequencies} \qquad (11)$$

where $Y(s) = X(s) \cdot F(s)$

and a similar expression involving convolution in the time domain.

(b) Cascading of Filter Stages

It is possible to cascade filter stages. In the case of a multiple low-pass filter, it is quite possible that after filtering $x(t)$ by the first stage, there may be negligible energy in $y(t)$ above some frequency very much lower than the original Nyquist frequency. A great reduction in computation can now be achieved by sampling at a lower rate. A small penalty is incurred by aliassing the energy that does exist above the second Nyquist frequency into lower frequencies, but this can be arranged to be negligible.

† See APPENDIX 2

To accommodate varying sampling rates, we will now change our notation, and represent the waveform $x(t)$ sampled by the ideal sampling function at intervals of $T_0$ as $x^{T_0*}(t)$ . To implement our first stage with sampling period $T_1$ ( $T_1 = m_1 T_0$ , $m_1 \geq 1$ ) , we need to relate $X^{T_0*}(s)$ to an input function $X^{T_1*}(s)$ to the first stage:

Consider a signal $X(s)$ sampled at the two intervals $T_0$ and $T_1$ , where in both instances the first Dirac Function occur at $t = 0$ . By analogy with equation (7) we have:

$$X^{T_0*}(s) = \frac{1}{T_0} \sum_{n=-\infty}^{\infty} X\left(s + jn\frac{2\pi}{T_0}\right)$$

$$X^{T_1*}(s) = \frac{1}{T_1} \sum_{n=-\infty}^{\infty} X\left(s + jn\frac{2\pi}{T_1}\right)$$

If we assume there is no energy present above the lower Nyquist frequency $\omega_N = \frac{\pi}{T_1}$ , then we can recover $X(s)$ from the sampled quantities by filtering off any quantities $\omega \gg \frac{\pi}{T_1}$ . This leaves only the value under the summation for n = o.

Denoting filtering by a bar, we have:

$$X(s) = T_0 \cdot \overline{X^{T_0*}(s)} = T_1 \cdot \overline{X^{T_1*}(s)}$$

$$\text{or} \qquad \overline{X^{T_1*}(s)} = \frac{1}{m_1} \overline{X^{T_0*}(s)} \qquad \ldots (12)$$

If we consider only the practically useful case where $m_1$ is an integer, we can generate $x^{T_1*}(t)$ from $x^{T_0*}(t)$ by selecting the first and thereafter every $m_1^{th}$ pulse. In this case equation (12) is what one would expect intuitively.

We can now implement the first stage of our filter:

$$Y_1^{T_1*}(s) = X_1^{T_1*}(s) \cdot T_1 F_1^{T_1*}(s) = \frac{1}{T_0} \cdot \frac{1}{m_1} Y_1(s) \text{ plus higher frequencies}$$

$$= \frac{1}{T_1} Y_1(s) \qquad \text{plus higher frequencies}$$

or for n stages, each having integrally related sampling periods $T_1 = m_1 T_0$, $T_2 = m_2 T_1$, $\ldots T_n = m_n T_{n-1}$ where the m values are all integers $\geq 1$, and the first sample occurs in every case at t=o, we have:

$$\begin{aligned} Y_1^{T_1*}(s) &= X_1^{T_1*}(s) \cdot T_1 F_1^{T_1*}(s) \\ Y_2^{T_2*}(s) &= Y_1^{T_2*}(s) \cdot T_2 \cdot F_2^{T_2*}(s) \\ &\vdots \\ Y_n^{T_n*}(s) &= Y_{n-1}^{T_n*}(s) \cdot T_n \cdot F_n^{T_n*}(s) \end{aligned} \qquad (13a)$$

where $y_{j-1}^{T_j *}(t)$ can be generated from $y_{j-1}^{T_{j-1} *}(t)$ by selecting the first, and every subsequent $m_j^{th}$ pulse, and:

$$y_n^{T_n *}(s) = \frac{1}{T_n} \cdot X(s) \cdot F_1(s) \cdot F_2(s) \cdots F_n(s) \qquad (13b)$$

Provided that there is no significant energy in $y_j(s)$ above the frequency $\omega_N = \frac{\pi}{T_{j+1}}$

(c) Mechanisation of the Recursive Filter

In practice the filter will be presented with a time series $x(t)$, so that mechanising the filter spectral characteristic is most easily performed in the time domain. This could be done by using equation(9) to generate a sampled impulse response. Direct convolution in the time domain would then give the sampled filter output. In principle, however, the impulse response of a filter whose transfer functions can be described by ratios of polynomials in $s$ is of infinite duration, and is likely to be long even if truncated when the truncated tail contribution is of the order of the expected noise levels. This is expecially true if the ratio of the Nyquist frequency to the low-pass filter cutoff frequency is large. This type of Direct Convolution digital filter requires typically 100 multiplications per timestep, which is computationally expensive. Filters of this sort are described by Craddock (1968 ), Pesaresi (1971) and Linnette (1961).

A more satisfactory solution is to make use of equation (10) which describes the sampled Transfer Function:

$$\frac{V_{OUT}^*(s)}{V_{IN}^*(s)} = T \sum_{n=0}^{\infty} f(nT) \cdot e^{-nsT} \qquad (14)$$

For filters whose continuous transfer function can be described as polynomials in s, f(nT) can be expressed in the form $\sum C \cdot e^{-(\alpha+i\beta)T}$, which means the R.H.S. of (14) can be summed as a geometric progression to a concise expression. Multiplying out the concise form of (14) and taking the Inverse Laplace Transform, we obtain a concise recursive relationship in $v_{OUT}^*$ and $v_{IN}^*$. This will be made clear from the examples shown below.

Evaluation of $v^*_{OUT}(nT)$ from previous values of $v^*_{IN}$ and $v^*_{OUT}$ will require one multiplication per pole or zero in the filter stage being considered plus one. The first stage can often be arranged to be a single pole, requiring two multiplications. It will be shown below how the multiplications can often be degenerated into accumulator shifts, giving the bulk of the multiplication as two accumulator shifts per timestep.

## 4. INTERPRETATION OF THE OUTPUT

Analytically, the output of a sampled process consists of a series of Dirac Impulses where strengths denote the required signal. In practice the digital machine presents a value at each output sampling epoch which is a code pulse modulated representation of the amplitude pulse modulated analytical signal.

To retrieve the continuous output signal (suitable, perhaps, as part of a control system) it would be possible to fabricate a device which generated a suitable approximation to the Dirac Impulse of designated strength which was then filtered by a unity gain low-pass filter. The response (delay) of this filter must be taken into account. A mathematical equivalent would be to take the Fourier Series for the discrete time series, which will have a maximum frequency of $\omega_N = \frac{\pi}{T_n}$, and then increase the number of high frequency Fourier components by adding coefficients of zero. The inverse Fourier Transform will give the output samples as before, but with extra samples at the correct amplitude between the machine generated samples. In the limit of infinitely high frequency Fourier coefficients the output is described at all points in time and effectively becomes continuous. Either operation will generate a continuous waveform of precisely the correct shape, but with a gain factor $\frac{1}{T_n}$ corresponding to the effects of sampling (equations (6), (7), (13)).

In practice few would attempt the ideal filtering situation above. In particular the Dirac Impulse generation might be approximated by holding the output value as the equivalent analogue voltage $V^*_{OUT}$ for a vanishingly small time $\tau$ secs before low-pass filtering. The integrated area under this impulse would be $\tau \cdot V^*_{OUT}$ rather than the $V^*_{OUT}$ of the Dirac Impulse, and the overall gain factor becomes $\tau / T_n$.

Increasing the value of $\tau$ will bring the gain closer to unity, but now the "holding circuit" itself becomes a filter. The "ideal zero-order hold circuit" holds the voltage constant between output samples. It has the advantage of being easily approximated physically, and has the transfer function:

$$G_{HO} = \frac{1 - e^{-sT_n}}{s} \qquad (14)$$

If we consider frequencies low compared to the reciprocal of $T_n$, then

$$sT_n \ll 1 \quad , \quad \text{and:}$$

$$G_{HO}(s \to 0) \simeq T_n \left(1 - \frac{sT_n}{2}\right) \qquad (15)$$

Thus for sufficiently low frequencies the combination of sampling gain and hold-circuit gain gives an overall gain of unity. The factor in brackets is the same as the first order expansion of $\exp(-\frac{sT_n}{2})$, indicating that to first order the signal is delayed by $\frac{T_n}{2}$ seconds.

For many applications we do not need to re-generate a continuous time sequence. Sometimes the object of filtering was to reduce the bandwidth of the input time-series so that it could be sampled (say) every 10 minutes and the value recorded in order to best describe the low-frequency content of the original time series. By analogy with equation (8) we have:

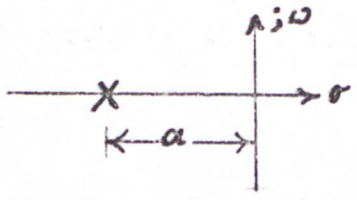
$$\psi_n^{T_n *}(t) = \sum_{t=0}^{\infty} \psi_n(+\tau_t) \cdot \delta(t - +\tau_t) \qquad (16)$$

so that we have the expected result that the machine presented values are the sampled points on the output waveform with unity gain.

## 5. FOUR EXAMPLES

FOUR worked examples are described. The first is for a single pole on the negative real axis in the Complex Plane (corresponding to a single time-constant). The second is for a complex conjugate pair of poles in the left hand half of the Complex Plane (corresponding to a simple resonant damped low-pass circuit, such as a single low-pass Sallan and Key Element). As examples the first is analytically trivial, and so illustrates the application to digital filters very clearly. The second illustrates the more general analytical approach. The two solutions are all that is required to implement low-pass polynomial denominator filters of any complexity.

(a) Single Pole on the Negative Real Axis.




The filter transfer function is

$$F(s) = \frac{Ka}{(s+a)} \qquad \text{(K arbitrary)}$$

and $$f(t) = Ka\,e^{-at}$$

Applying equation (10), implement $F(s)$ by $TF^*(s)$:

$$TF^*(s) = T\sum_{n=0}^{\infty} f(nT)\cdot e^{-nsT} = KaT\sum_{n=0}^{\infty} e^{-(a+s)nT}$$

$$= \frac{KaT}{1-e^{-(a+s)T}} = \frac{V^*_{OUT}(s)}{V^*_{IN}(s)}$$

or: $$KaT\cdot V^*_{IN}(s) = V^*_{OUT}(s) - e^{-aT}\cdot e^{-sT}\cdot V^*_{OUT}(s)$$

Taking the Inverse Laplace Transform:

$$KaT\,v^*_{IN}(n) = v^*_{OUT}(n) - e^{-aT}\cdot v^*_{OUT}(n-1)$$

or

$$\boxed{v^*_{OUT}(n) = KaT\,v^*_{IN}(n) + e^{-aT}\,v^*_{OUT}(n-1)} \qquad (17)^{\dagger}$$

Thus to calculate $v^*_{OUT}(n)$ we require two multiplications per timestep. However the factor K is present purely to determine the gain of the filter; with K=1 the gain = 1. We could choose K such that $KaT = 1$, and so avoid this multiplication completely, allowing for the extra gain at a later filter stage where the timestep can be much larger. However, for the common case where aT is small (and especially on cascaded filters) this practice can lead to a situation where values of $v^*_{OUT}$ exceed the

† SEE APPENDIX 2

dynamic range of integer arithmetic. A compromise solution is to implement KaT mainly by an accumulator-shift-right by a suitable number of places, and allow for the exact gain adjustment in the final cascaded stage.

This leaves the factor multiplying $V^{*}_{out}(n)$. For many applications the precise value of a may be unimportant, and a suitable value for either a $aT$ can be selected so that this multiplication can be implemented accurately by an accumulator shift right and subtract from the previous value.

In implementing equation (17) the machine dynamic range may be important. Suppose we have a fast sensor delivering a signal with energy almost up to 5Hz, so that T = 0.1 secs, and we require a time constant of 10 minutes. Then $aT = 1/6000$

$$e^{-aT} = 0.9998333$$

This number can barely be represented as different from unity in 12 bit arithmtic; in 16 bit arithmetic the accuracy of representing its difference from unity (and hence the accuracy of representing aT in this single calculation) is about 10%. Clearly when aT is small, multiple precision arithmetic may be required.

This situation can in practice be dodged quite neatly. For the above example we can pre-filter the signal with an extra filter stage having a single high frequency pole at $s = -a_1$. The precise value of $\alpha_1$ is unimportant so that it can probably be implemented by an accumulator shift. As $\alpha_1 T$ is chosen not to be too small we do not require high resolution, we may be able to choose K so that $K\alpha_1 T$ is unity. Thus the pre-filter requires only one accumulator shift per timestep in lieu of multiplications. The function of the pre-filter is to reduce the signal bandwidth so that a longer timestep can be used for the main filter, making more time available for multiplications. For the main filter we use a pole at $s = -\alpha_2$ where $\frac{1}{\alpha_1} + \frac{1}{\alpha_2} = \frac{1}{\alpha}$

The overall filter response will be

$$F'(s) = \frac{\alpha_1 \alpha_2}{(s+\alpha_1)(s+\alpha_2)} = \frac{1}{1 + \frac{\alpha_1+\alpha_2}{\alpha_1\alpha_2}s + \frac{s^2}{\alpha_1\alpha_2}} \quad \frac{1}{1 + \frac{s}{\alpha} + \frac{s^2}{\alpha_1\alpha_2}}$$

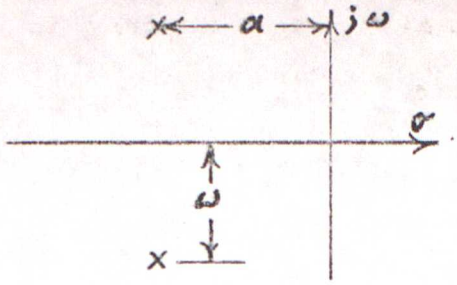
rather than the required

$$F(s) = \frac{\alpha}{s+\alpha} = \frac{1}{1+\frac{s}{\alpha}}$$

but these are equal for low frequencies where $\frac{s^2}{\alpha_1\alpha_2} \ll 1$ ,

and the attenuation for high frequencies (especially for a cascaded filter where accuracy is important) is so high that the error in attenuation is unimportant.

(b) Complex Conjugate Pole Pair in the Left Hand Half of the Complex Plane.




$$F(s) = \frac{K(a^2+\omega^2)}{(s+a)^2+\omega^2}$$

$$f(t) = \frac{K(a^2+\omega^2)}{\omega} \cdot e^{-at} \cdot \sin(\omega t)$$

K is arbitrary

Applying (10):

$$TF^*(s) = T\sum_{n=0}^{\infty} f(nT)\cdot e^{-nsT}$$

$$= KT\left(\frac{a^2+\omega^2}{\omega}\right)\left[\sum_{n=0}^{\infty} e^{-anT}\cdot e^{-snT}\cdot \sin(\omega nT)\right]$$

We wish to sum this infinite series. Let

$$S \equiv \sum_{n=0}^{\infty} e^{-anT}\cdot e^{-snT}\cdot \sin(\omega nT)$$

$$C \equiv \sum_{n=0}^{\infty} e^{-anT}\cdot e^{-snT}\cdot \cos(\omega nT)$$

Then $$C+iS = \sum_{n=0}^{\infty} e^{-(a+s-i\omega)nT} = \frac{1}{1-e^{-(a+s-i\omega)T}}$$

$$= \frac{1-e^{-(a+s+i\omega)T}}{1-2\cos(\omega T)\cdot e^{-(a+s)T}+e^{-(a+s)2T}}$$

We are interested only in the imaginary part:

$$S = \frac{e^{-(a+s)T}\cdot \sin(\omega T)}{1-2\cos(\omega T)\cdot e^{-(a+s)T}+e^{-(a+s)2T}}$$

Thus

$$TF^*(s) = KT\left(\frac{a^2+\omega^2}{\omega}\right)\left(\frac{e^{-aT}\cdot \sin(\omega T)\cdot e^{-sT}}{1-2\cos(\omega T)\cdot e^{-aT}\cdot e^{-sT}+e^{-2aT}\cdot e^{-2sT}}\right)$$

$$= \frac{V_{OUT}^*(s)}{V_{IN}^*(s)}$$

Cross-multiplying and performing an Inverse Laplace Transform to the Time Domain, we have:

$$v^*_{out}(n) = 2\cos(\omega T)\cdot e^{-aT}\cdot v^*_{out}(n-1) - e^{-2aT} v^*_{out}(n-2) + TK\left(\frac{a^2+\omega^2}{\omega}\right) e^{-aT}\cdot \sin(\omega T)\cdot v^*_{IN}(n-1) \quad (18)$$

or

$$v^*_{out}(n) = C_1\cdot v^*_{out}(n-1) + C_2\cdot v^*_{out}(n-2) + C_3\cdot v^*_{IN}(n-1) \quad (19)^{\dagger}$$

where the constants are:

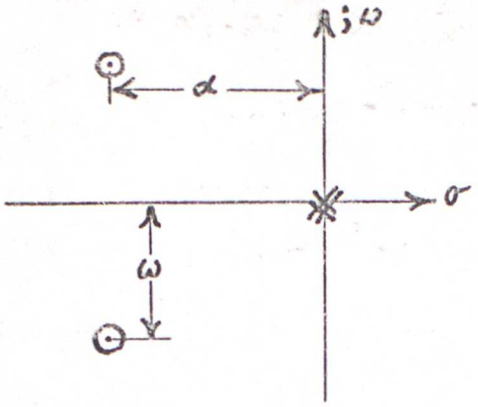
$$C_1 = 2\cos(\omega T)\cdot e^{-aT}$$
$$C_2 = -e^{-2aT}$$
$$C_3 = KT\left(\frac{a^2+\omega^2}{\omega}\right) e^{-aT}\cdot \sin(\omega T)$$

As before $C_3$ is a gain factor. K can be chosen to make $C_3$ unity, or else the equivalent of an accumulator-shift-right. $C_1$ and $C_2$ define the pole locations. It may be possible to choose a, $\omega$ or T to make one of these an accumulator shift, but in general full multiplications must be performed for an accurate cascaded filter. It will often be possible to operate a complex conjugate pole pair at a sampling interval much longer than the basic data sampling interval because a previous stage (negative real axis pole or pre-filter) has reduced the input bandwidth.

Using this type of analytical technique other pole-zero patterns useful in high-pass and bandpass filters, or in control circuits, can be implemented.

† SEE APPENDIX 2

c. Complex Conjugate Zero Pair in the Left-Hand-Half of the Complex Plane (with two Poles at the Origin).




$$F(s) = K\left\{\frac{(s+\alpha)^2+\omega^2}{s^2}\right\}$$

(K arbitrary)

$$= K\left\{1 + \frac{2\alpha}{s} + \frac{\alpha^2+\omega^2}{s^2}\right\}$$

$$f(t) = K\left\{\delta(t) + 2\alpha + (\alpha^2+\omega^2)t\right\}$$

Applying (10):

$$TF^*(s) = T\sum_{n=0}^{\infty} f(nT)\,\varepsilon^{-nsT}$$

$$= K + KT\sum_{n=0}^{\infty} 2\alpha\,\varepsilon^{-nsT} + KT(\alpha^2+\omega^2)\sum_{n=0}^{\infty} nT\,\varepsilon^{-nsT}$$

$$= K + \frac{2\alpha KT}{1-\varepsilon^{-sT}} + KT^2(\alpha^2+\omega^2)\varepsilon^{-sT}\left\{1+2\varepsilon^{-sT}+3\varepsilon^{-2sT}+\cdots\right\}$$

$$= K + \frac{2\alpha KT}{1-\varepsilon^{-sT}} + \frac{KT^2(\alpha^2+\omega^2)\,\varepsilon^{-sT}}{(1-\varepsilon^{-sT})^2}$$

$$= \frac{V^*_{OUT}(s)}{V^*_{IN}(s)}$$

Or:

$$V^*_{OUT}(s)\left\{1 - 2\varepsilon^{-sT} + \varepsilon^{-2sT}\right\}$$

$$= V^*_{IN}(s)\left\{K - 2K\varepsilon^{-sT} + K\varepsilon^{-2sT} + 2\alpha KT(1-\varepsilon^{-sT}) + KT^2(\alpha^2+\omega^2)\varepsilon^{-sT}\right\}$$

Taking the Inverse Laplace Transform:

$$v^*_{OUT}(n) = 2\,v^*_{OUT}(n-1) - v^*_{OUT}(n-2) + C_1 K\, v^*_{IN}(n) + C_2 K\, v^*_{IN}(n-1) + K\, v^*_{IN}(n-2) \qquad (20)$$

where:

$$C_1 = 1 + 2\alpha T$$

$$C_2 = -\left(2 - 2\alpha T - T^2(\alpha^2 + \omega^2)\right)$$

This particular Pole-Zero pattern is useful in implementing band-reject notches.

d. Integrator

$$F(s) = \frac{K}{s} \qquad f(t) = K$$

$$T F^*(s) = T \sum_{n=0}^{\infty} f(nT)\, e^{-nsT}$$

$$= KT \left\{ 1 + e^{-sT} + e^{-2sT} + \cdots \right\}$$

$$= \frac{KT}{(1 - e^{-sT})} = \frac{V^*_{OUT}(s)}{V^*_{IN}(s)}$$

Cross-Multiplying, and Taking the Inverse Laplace Transform:

$$v^*_{OUT}(n) = v^*_{OUT}(n-1) + KT\, v^*_{IN}(n) \qquad (21)$$

## 6. REFERENCES


Tou J.T. 1959 "Digital and Sampled-Data Control Systems". McGraw-Hill Book Company, New York, Toronto, London.

Craddock J.M. 1968 "Statistics and the Computer Age" Unibooks. English Universities Press.

Pesaresi R. 1971 "Numerical Filtering Techniques for the Time-Series Analysis of Oceanographic and Meteorological Data" N.A.T.O. Tech.Memo.166 Saclant ASW Research Centre; Viale San Bartolomeo 400; I 19026 - La Spezia, Italy.

Linette H.M. 1961 "Statistical Filters for Smoothing and Filtering Equally Spaced Data". U.S. Navy Electronics Laboratory, Report No. 1049, San Diego, California.



Meteorological Office

Ministry of Defence


February 1977

Appendix 1.

## A Related Topic - The Digital Differential Analyser

Many functions that require computation can be expressed to first order in an incremental form. Thus if a previous value of the function is known, and the variables are incremented, the new value of the function can be calculated. Specifically, if the solution to the function is calculated sufficiently slowly so that variable increments can only take on the values -1, 0, or +1 (least significant bit) then evaluating the updated function requires only additions, and no multiplications.

e.g. for the function $Z = X \cdot Y$ ; $\delta Z = X \cdot \delta Y + Y \cdot \delta X$

or to calculate varying values of $\cos\theta$ or $\sin\theta$ :

$$\sin(\theta + \delta\theta) = \sin\theta + \delta\theta \cdot \cos\theta$$

$$\cos(\theta + \delta\theta) = \cos\theta - \delta\theta \cdot \sin\theta$$

To build up a system, the output $\delta Z$ of a "Digital Integrator" which implements $\delta Z = Y \cdot \delta X$ must also be in incremental form. This can be done without loss of accuracy by taking the increment $\delta Z$ as the overflow from an integer counter having the same bit length as the variables.

The Technique can be implemented in hardware or software, the hardware configuration involving basically adder and storage elements. Early aircraft simulators used the technique in the form of the "Digital Differential Analyser" for solving differential equations, which is one way of describing a filter.

The basic advantage of the method is that multiplications are avoided. The important disadvantage is that computation steps must be made sufficiently rapidly that variables can change by only one least significant bit, which may be very much faster than the rate required to satisfy the Nyquist condition for resolving the signal. This may even be useful if a non delayed "continuous" signal is required. However, it is probably true that the method is now of little relevance to most systems. The exception is in some very simple systems

where it can provide fast software, or require less hardware than a fast Von-Neumann (computer) hardware structure. The interested reader is referred to Sizer (1967).

REFERENCES

Sizer T.R.H. 1967 "The Digital Differential Analyser" Chapman and Hall.

## Appendix 2

### POSTSCRIPT

1. Since the bulk of this paper was written, some experience has been gained in the use of these filters, and this has shown up deficiencies in the analytical expressions given. These deficiencies are not sufficient to invalidate the basic arguments used, but appear to arise out of a basic non-linearity present in time-sampling. They are comparable with the truncation errors experienced in digitising analogue signals. The user should be aware of the problems, although in practice they do not appear to be serious.

2. The simplest example can be seen by applying the expression for a single pole on the negative real axis, eqn 17:

$$v^*_{out}(n) = K a T \, v^*_{IN}(n) + e^{-aT} v^*_{out}(n-1) \qquad (17)$$

If the gain of the filter were unity, and we apply a steady signal v for a sufficiently long time, then

$$1 = K a T + e^{-aT}$$

or

$$K = \frac{1 - e^{-aT}}{aT} \; ; \qquad (22)$$

But this is inconsistent with the statement in 5(a) that the filter gain is unity when K = 1. As $aT \to 0$, then $K \to 1$, but this near equality is often insufficient in low-pass filters where the absolute gain near zero frequency can be critically important. In this situation, the expression $KaT$ in (17) should be replaced by $K(1 - e^{-aT})$. The impulse response will now be slightly wrong, but in this situation the exact high frequency response may well be less critical.

3. It is useful to find the source of the inconsistency described above. I believe that it first appears in eqn (8) where the impulse response of a continuous filter f(t) is to be represented in sampled form f*(t), where

$$f^*(t) = f(t) \sum_{n=-\infty}^{\infty} \delta(t - n\tau)$$

The problem is that a continuous impulse response f(t) will contain high frequencies that are above the Nyquist frequency 1/2T for any finite sampling rate. To this extent f*(t) cannot be an exact representation of f(t). The problem can be reduced if the Nyquist frequency $f_N = 1/2T$ is such that:

$f_N \gg$ the 3 dB "corner frequency" of a low-pass filter (23)

Eqn 23 is the same condition as $aT \to 0$.

$f_N$ can approach the "corner frequency" more closely if F(s) is such that attenuation increases rapidly above the "corner frequency", as there will be less high-frequency energy available. Thus it may be advantageous to implement several poles in one stage.

In a strict mathematical sense, therefore, the results of this paper are invalid. However a similar non-linear effect occurs whenever linear processes are applied to digitised data. In practice an engineering compromise can be reached - sufficient digitisation levels must be taken to represent the data, and the Nyquist frequency must be sufficiently higher than any significant frequencies in the filter to be implemented. Any particular implementation must be tested to ensure that these conditions are met.

4. When implementing Low-Pass filters, the worst feature of the above effect lies in the slight change in absolute filter gain at low frequencies. As indicated pragmatically in para 2 above, this can be avoided by using alternative expressions:

eqn (17): replace $Ka^T$ by $K(1 - e^{-aT})$

eqn (19): use $C_3 = K(1 - 2\cos(\omega T)e^{-aT} + e^{-2aT})$

Equations(20) and (21) have no finite D.C. response.