INFLUENCE OF THE NON-LINEAR BEHAVIOUR OF A RECORDING INSTRUMENT ON THE PROPERTIES OF A CONTROL SYSTEM

by C. H. LOOS *).

621-53.001:621.317.7.087.6

Pursuant to the articles on the application of control theory to linear systems published in the two preceding numbers of this journal, the article below deals with an element which behaves non-linearly when the input signal undergoes very rapid variations. In qualitatively analysing the stability characteristics of a control loop containing this element, use is made of a rule of thumb formulated in the second of the articles mentioned. The surprising conclusion is that the stability of the control loop in question is not a monotonic function of the speed of variation, but shows a minimum.

Introduction

For measuring and recording important variables in industrial plants, increasing use is being made of recording instruments (recording millivoltmeters) whose operation is based on automatic compensation of the measured quantity (voltage) by means of a servomechanism. Fundamentally, the circuit of these instruments stems from Poggendorf's well-known compensation method (*fig. 1a*), except that the null instrument here is replaced by an amplifier which drives the motor that moves the sliding contact of the potentiometer (see fig. 1*b*). Attached to the sliding contact is a stylus or pen. Non-electric quantities to be measured are first converted into an electrical signal.

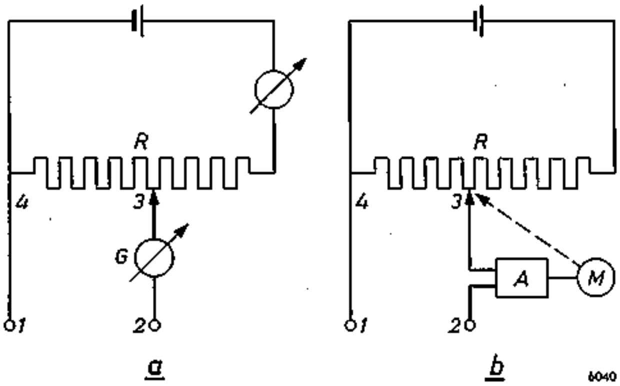

Fig. 1. *a*) Poggendorf compensation method of measuring an e.m.f. The voltage is applied to the terminals *1* and *2*. The contact *3* is then shifted until the highly sensitive meter *G* no longer shows a deflection. The potential difference between points *3* and *4* is then equal to the e.m.f. to be measured and can be calculated from the current flowing through *R* and the resistance between *3* and *4*.
b) In a recording instrument, *R* is a potentiometer whose sliding contact is moved by a motor *M*. The latter is fed by the potential difference between *2* and *3*, highly amplified by *A*. As soon as the potential difference is zero, the sliding contact remains stationary.

Frequently the recorded quantities also have to be automatically controlled, in which case the recording instrument itself can sometimes be used as part of the controller, i.e. as an amplifier. For this purpose a second potentiometer is employed, which is fed with a constant voltage much higher than the voltage across the measuring potentiometer, and whose sliding contact moves synchronously with that of the other. In this way a gain of e.g. 5000× can readily be achieved, offering a particularly simple method of effecting the control action.

A recording instrument thus modified behaves as a linear element only when the changes in the input signal are slow enough for the sliding contact (the pen) to follow. In this article we shall examine what happens when this condition is not fulfilled. It will be shown that instability effects may arise in a control loop which contains, in addition to the recording instrument, two elements having the transfer function $(1 + j\omega\tau)^{-1}$. If the recorder were an ideal amplifier, a control system of this kind ought to be stable for every value of the loop gain [1]). Remarkably enough, the extent to which the stability is endangered — we shall express this presently in a more rigorous form — does not increase monotonically with the discrepancy between the desired speed of the pen and the maximum possible speed. In fact, where this discrepancy is very large, the danger of instability decreases!

The sliding contact will no longer follow a varying signal when the amplifier *A* (fig. 1*b*) is overdriven. Irrespective of the magnitude of the potential between *2* and *3*, the amplifier

*) Research Laboratories, Eindhoven.

[1]) Examples of control loops with linear elements will be found in the article by M. van Tol, Philips tech. Rev. **23**, 109, 1961/62 (No. 4), where the transfer function is also discussed. The relation between loop gain and stability is dealt with by M. van Tol in Philips tech. Rev. **23**, 151, 1961/62 (No. 5).

then delivers its maximum output signal and the pen consequently moves at a constant speed. Even when the amplifier is not overdriven, the pen does not of course follow a varying signal exactly: the motor can only turn when the potential between *2* and *3* differs from zero. The deviation is smaller the higher the gain factor of A; theoretically it approaches zero for an infinitely high gain factor. If a *constant* signal is applied between *1* and *2*, and the motor behaves like an ideal integrator, the deviation will of course be zero in the long run.

With the aid of two figures we shall now try to show qualitatively the way in which the output voltage of the instrument varies when the variations in the input voltage are too fast. We at once introduce the approximations necessary to simplify the theoretical treatment of the instrument's behaviour. The most general case is represented in *fig. 2*. The

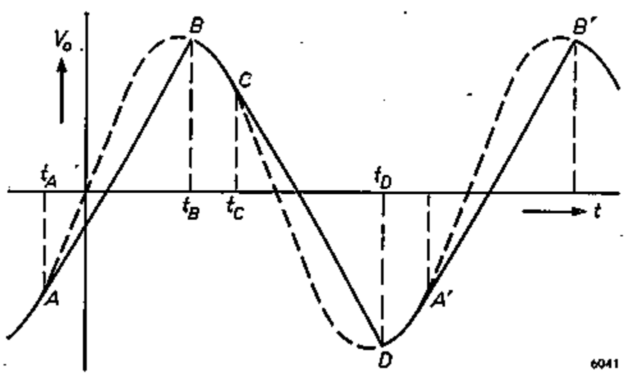

Fig. 2. Output voltage of a recording instrument adapted as a controller, when the maximum speed at which the input signal varies is too fast for the pen to follow. The pen follows the input signal only between B and C (and D and A', etc.). Outside these regions the pen moves uniformly at its maximum speed.

broken curve represents the input signal and the solid line the output signal. For convenience the gain is assumed to be unity. When the input-signal variations are too fast for the pen to follow, the pen moves at a constant speed (portion AB). This continues until the speed of variation has dropped sufficiently to allow the pen to catch up again (point B). The pen now follows the variation of the input voltage until (at point C) the signal again changes too rapidly. Thereupon the output signal again varies linearly with time (portion CD) and so on.

If we now increase the frequency or the amplitude of the output signal, points B and C etc. come closer together, until finally the signal acquires a triangular waveform (*fig. 3*).

Summarizing, then, we note that with rising frequency and/or amplitude the gain is initially linear. When a certain limit is exceeded, we obtain the case represented in fig. 2, and finally, after passing a second limit, the case in fig. 3. We shall now put this into mathematical form, after which we shall examine the behaviour of the instrument as an element in a control loop, with the aid of *describing functions*. In this method the problem is treated as if the element were a linear one, the output signal following from a sinusoidal input signal being approximated by its fundamental Fourier component. For these sinusoidal signals we can then establish a transfer function — the describing function — in the same way as for linear elements. Unlike the transfer function of a loop consisting solely of linear elements, however, this describing function may contain the amplitude as well as the frequency of the input signal.

The describing function is especially useful for analysing the stability of a control loop containing a non-linear element. Since the higher Fourier components of an output signal whose frequency is near the cut-off frequency are usually strongly attenuated in the other elements of the loop, the signal when it appears again at the input of the non-linear element, having passed once around the loop, has in fact become virtually sinusoidal, and may to a very good approximation be regarded as solely due to the fundamental Fourier component.

Calculation of the transfer function

To calculate the transfer function of the recorder we start from a sinusoidal input signal of amplitude U. Disregarding the limited speed of the recorder, we consider the instrument to be an ideal amplifier having a gain factor A. Since the magnitude of A has no effect on the behaviour of the instrument — although of course it does affect the control loop to which it belongs — we shall henceforth assume A to be equal to unity.

The rate of change of the input signal $U \sin \omega t$ is $\omega U \cos \omega t$. As long as ωU is smaller than the

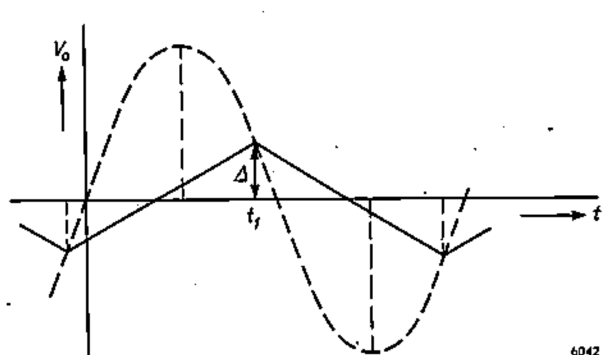

Fig. 3. When the maximum speed at which the input signal changes substantially exceeds the maximum writing speed, the output signal has a triangular waveform.

value corresponding to the maximum speed B of the pen, the behaviour of the instrument is linear.

We shall first consider the case represented in fig. 3, where the pen is no longer able to follow the input signal at all, and the output signal has a triangular waveform. The slope of the straight lines is $+B$ and $-B$ respectively, and the amplitude Δ of the signal (half the peak-to-peak value) is $\pi B/2\omega$. The first Fourier component (first harmonic) of a triangular signal of amplitude Δ (see appendix) is:

$$\frac{8\Delta}{\pi^2}\sin(\omega t+\varphi). \quad \ldots\ldots \quad (1)$$

The amplitude ratio of the first harmonics of output and input signals is thus:

$$\frac{4}{\pi}\times\frac{B}{\omega U}. \quad \ldots\ldots\ldots \quad (2)$$

The phase angle φ is quickly found when it is remembered that the peak value of the output signal, and hence of its first harmonic, occurs at the moment at which the input signal (in the second quadrant) also has the value $\pi B/2\omega$, and that the peak value of the input signal occurs when $\omega t = \pi/2$. It follows from this that:

$$\varphi=\frac{\pi}{2}-\left(\pi-\arcsin\frac{\pi B}{2\omega U}\right)=-\frac{\pi}{2}+\arcsin\frac{\pi B}{2\omega U}. \quad \ldots\ldots \quad (3)$$

Before considering the case of fig. 2, we shall consider what are the "limits" mentioned above. We have seen that the behaviour of the instrument is linear if $\omega U < B$, i.e. if $B/\omega U > 1$.

The case of fig. 3 — triangular output voltage — occurs where $-\omega U\cos\omega t_1 > B$, that is where $\omega U\cos\arcsin\frac{\pi B}{2\omega U} > B$, i.e. where

$$\sqrt{1-\left(\frac{\pi B}{2\omega U}\right)^2} > \frac{B}{\omega U}.$$

This is the case when

$$\frac{B}{\omega U} < \frac{1}{\sqrt{1+\pi^2/4}},$$

i.e. when

$$\frac{B}{\omega U} < 0.538.$$

The case of fig. 2 thus occurs in the region

$$0.538 < \frac{B}{\omega U} < 1. \quad \ldots\ldots \quad (4)$$

If we now calculate the first Fourier component of the output signal when the latter is partly sinusoidal and partly linear with respect to time (see appendix), we find for the amplitude of the first sine term:

$$b_1=\frac{U}{\pi}\{\pi-\arccos p-k(p)-\sin k(p)\cos k(p)+ + p\sqrt{1-p^2}+2p\sin k(p)\}, \quad \ldots \quad (5)$$

and for that of the first cosine term:

$$a_1=\frac{U}{\pi}\{\cos k(p)-p\}^2, \quad \ldots \quad (6)$$

where $p = B/\omega U$ and $k = \omega t_B$ (cf. fig. 2). From this we can directly derive the amplitude ratio $(\sqrt{a_1^2+b_1^2}/U)$ and the phase shift ($\arctan a_1/b_1$) for the relevant range of p values. Combining the result with those for the regions $p > 1$ and $p < 0.538$, and plotting a Bode diagram — with the quantity $\omega U/B$ or $1/p$ as the abscissa — we arrive at *fig. 4*. As can be seen, the characteristics closely resemble those of a linear element having one time constant τ_R of magnitude $\pi U/4B$. This time constant is thus proportional to the amplitude of the input signal and inversely proportional to the maximum writing speed.

A characteristic difference is that for p values greater than unity the gain is exactly constant and the phase shift exactly zero. In this region the

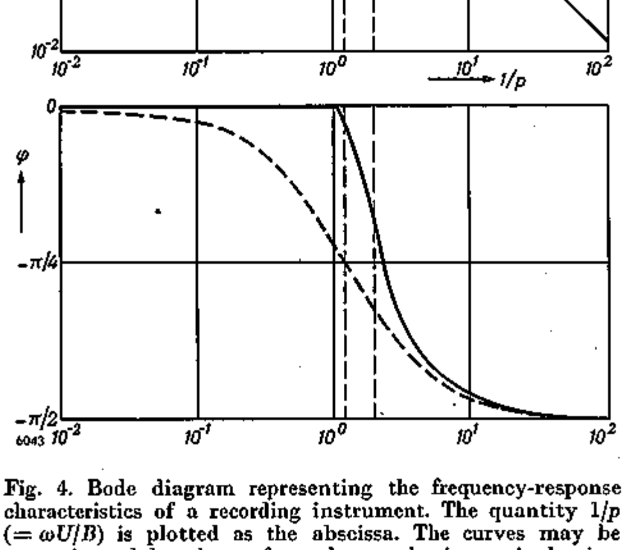

Fig. 4. Bode diagram representing the frequency-response characteristics of a recording instrument. The quantity $1/p$ ($=\omega U/B$) is plotted as the abscissa. The curves may be approximated by those of an element having a single time constant τ_R of the value $\pi U/4B$ (broken lines).

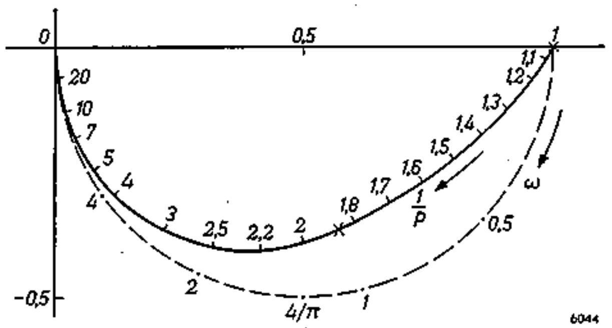

Fig. 5. Nyquist diagram appertaining to fig. 4. The quantity $1/p$ is again chosen as the variable in order to draw one curve instead of a set of curves. The crosses mark the points $p = 1$ and $p = 0.538$. As in fig. 4, the broken curve relates to an element having the transfer function $(1 + j\omega\pi U/4B)^{-1}$. The figures given beside this curve relate, of course, to ω.

instrument does not behave like an element with a single time constant but in fact like an ideal amplifier. The relevant Nyquist diagram is shown in *fig. 5.* This too is represented in such a way that a point on the curve is not applicable to a certain value of ω but to $1/p$. It should be noted that all points for which $0 < 1/p < 1$ coincide on the real axis.

A recording instrument in a control loop with two time constants

We shall now examine the characteristics of a control loop as shown in *fig. 6*. Here R and A together constitute the recorder; R represents the behaviour of the instrument as such and A the gain independent of R, which is solely determined by the voltage applied to the second potentiometer.

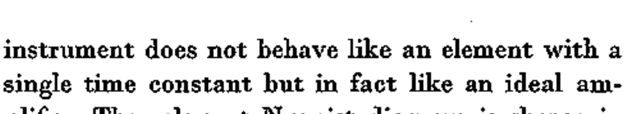

Fig. 6. Block diagram of a control loop consisting of a recorder R, an ideal amplifier with gain factor A, and two elements whose transfer function is $G(j\omega) = (1 + j\omega\tau)^{-1}$.

Blocks I and II both have a transfer function of the form $(1 + j\omega\tau)^{-1}$, but markedly different values of τ. Where $p > 1$, the recorder behaves like an ideal amplifier and the control loop is stable. Where $p < 1$, however, R also contributes to the phase shift, which may therefore in principle be greater than 180°, so that instability may occur. We shall examine this point presently.

First, however, we shall emphasize that, for the purposes of stability considerations, the situation in a loop containing a nonlinear element differs somewhat from that of a loop containing nothing but linear elements. For in this case the Nyquist diagram does not contain simply one curve, which may or may not enclose the point (—1,0), but a set of curves whose parameter is an amplitude — in our case the amplitude of the signal (which may consist only of noise) appearing at the input of the recorder. To be sure of stability, the loop gain should be chosen such that that curve in the complete set of curves which cuts off the largest section of the negative real axis does not enclose the point (—1,0). If that section has no finite value, then stability is out of the question.

A good qualitative insight into the behaviour of the control loop in fig. 6 can be obtained by using the rule of thumb arrived at in the above-mentioned article [2]). This stated that in a control loop which, besides an ideal amplifier, contained solely elements having the transfer function $G(j\omega) = (1 + j\omega\tau)^{-1}$, the maximum permissible loop gain A_{max} is equal to τ_1/τ_2. Here τ_1 is the longest and τ_2 the next longest time constant. If $A = \tau_1/\tau_2$, then the gain drops to unity at the second break in the double-logarithmic amplitude characteristic, i.e. at a phase shift of 135° (90° due to the block with τ_1 and 45° due to that with τ_2). Although the break in the case of the recorder corresponds to a phase shift somewhat smaller than 45°, that does not affect the validity of the argument.

If τ_R is initially the longest of the time constants ($\tau_1 = \tau_R$), then τ_R determines the position of the *first* break (*fig. 7a*) and an increase in U — we call the initial value U_1 — leads to greater stability, and a decrease to reduced stability. If τ_R is the next largest time constant (fig. 7*b*), then τ_R determines the position of the *second* break in the curve, and the stability reacts in precisely the opposite way to variations in U.

If we now let the amplitude U pass through a range of values such that τ_R begins with the next largest time constant and ends with the largest, and if we start from a stable state ($A = \tau_1/\tau_R$), we then see that as U increases the stability decreases — and may finally result in instability — but that the stability of the system increases again as soon as τ_R has become the largest time constant. The stability is therefore not a monotonic function of U, but shows a minimum when τ_R is roughly equal to the longer of the two fixed time constants.

An important consequence of this effect is that when a control loop of the type in fig. 6 becomes unstable it does not start to oscillate with ever-increasing amplitude, but enters into a *stationary* state (see below). By measuring the amplitude and frequency occurring in this state for various values of the two fixed time constants we have been able to verify experimentally the theory described above.

We shall work this out quantitatively for the case where the two fixed time constants are identical. Let the transfer function of the recording instrument be approximately $(1 + j\omega\tau_R)^{-1}$ where $\tau_R = \pi U/4B$ (see above), then the transfer function KG of the entire (open) control loop is given by:

$$KG = \frac{A}{(1 + j\omega\tau)^2\,(1 + j\omega\tau_R)} =$$

$$= \frac{A}{(1-\omega^2\tau^2-2\omega^2\tau\tau_R) + j(2\omega\tau + \omega\tau_R-\omega^3\tau^2\tau_R)}.$$

The Nyquist diagram for this function, for the case where $A = 10$ and both fixed time constants are equal to one second, is shown in *fig. 8*. Here again, there is not just one curve but a set of curves with τ_R as parameter. The three curves shown relate to cases where τ_R is equal to 0.1, 1.0 and 10 seconds,

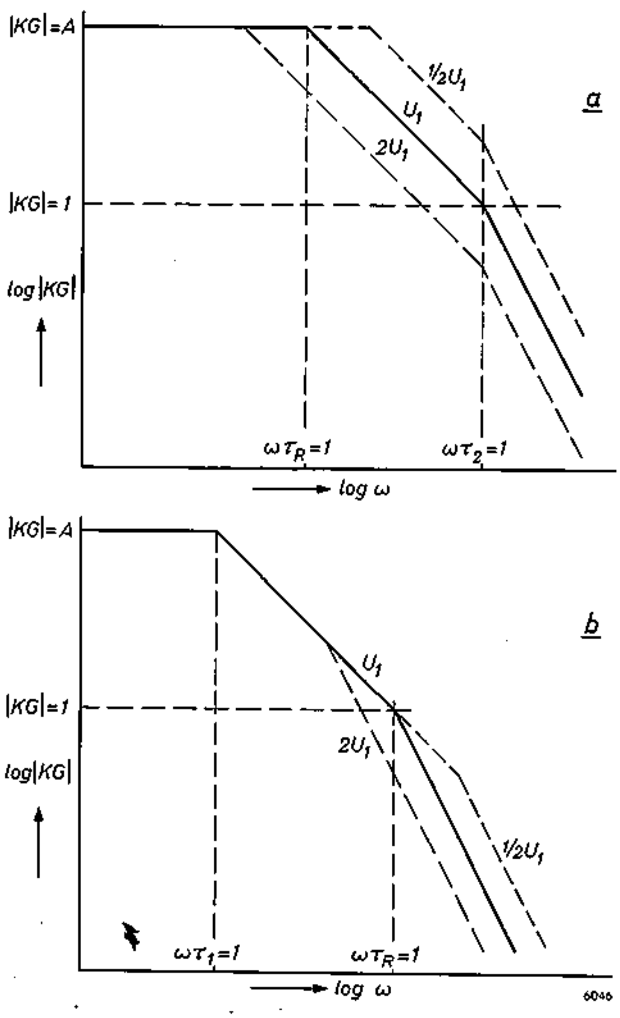

Fig. 7. Bode diagram of the control loop in fig. 6, approximated by straight lines. *a*) The time constant τ_R of *R* is the longest of the three time constants. When the amplitude *U* increases (the other parameters remaining constant) the first break shifts to the left and the gain at $\omega = 1/\tau_2$ decreases, as a result of which the stability increases. *b*) τ_R is the next largest time constant and determines the position of the second break in the curve. With increasing amplitude the stability decreases.

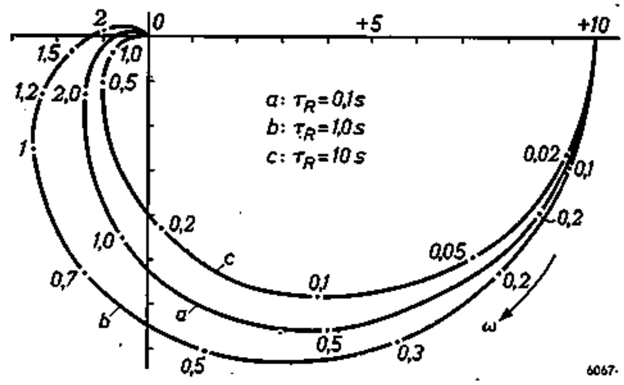

Fig. 8. Nyquist diagram of the control loop in fig. 6, for the case where the two fixed time constants are both 1 second and *A* is 10. A set of curves is found whose parameter is the time constant τ_R of the recorder. For $\tau_R = 0.1$ sec and 10 sec the closed loop is stable; for $\tau_R = 1.0$ sec it is unstable. The figures beside the curves again give the relevant values of ω.

respectively. As can be seen, the closed loop *is* stable in the two extreme cases, but not when τ_R is 1 second.

The behaviour of the stability as a function of τ_R — i.e. as a function of U/B — can be derived from the displacement of the point where the KG curve intersects the negative real axis. For this purpose we equate the imaginary part with zero:

$$\omega(2\tau + \tau_R - \omega^2\tau^2\tau_R) = 0\,. \qquad (7)$$

The curve therefore intersects the negative real axis ($\omega \neq 0$) when:

$$\omega^2 = \frac{2\tau + \tau_R}{\tau^2\tau_R}. \qquad (8)$$

The coordinate of the point of intersection is:

$$\frac{A}{1 - \{\tau^2 + 2\,\tau\tau_R\}\left\{\dfrac{2\tau + \tau_R}{\tau^2\tau_R}\right\}} = \frac{A\tau\tau_R}{-2(\tau+\tau_R)^2}. \qquad (9)$$

It follows directly from eq. (9) that the point of intersection tends to the origin when τ_R is very small or very large. The absolute value of the real coordinate is maximum when $\tau_R = \tau$. Substituting this in eq. (9) we find that this maximum value is equal to $A/8$. Where $A > 8$, as in the example of fig. 8, there is therefore a region of τ_R values at which the system is unstable. The limits τ_R' and τ_R'' of that region can be calculated with the aid of eq. (9). We find

$$\tau_R' = \tfrac{1}{4}\tau\,\{(A-4) - \sqrt{A(A-8)}\}, \qquad (10a)$$

and

$$\tau_R'' = \tfrac{1}{4}\tau\,\{(A-4) + \sqrt{A(A-8)}\}. \qquad (10b)$$

If we let τ_R — or the amplitude U, where B is fixed — increase from a low value, the system begins to oscillate as soon as τ_R exceeds the value τ_R'. As a result, the amplitude goes on increasing of its own accord to a value corresponding to τ_R''. The system then remains oscillating at this amplitude with a frequency given by:

$$\omega^2 = \frac{2\tau + \tau_R''}{\tau^2 \tau_R''} . \quad . \quad . \quad . \quad . \quad (11)$$

Experiments have shown the measured oscillation frequencies and amplitudes to be in good agreement with the relations derived theoretically.

Appendix: Calculation of the first Fourier components of the output signal

The amplitudes b_n and a_n of the nth sine and cosine terms in the Fourier expansion of the periodic function $f(x)$ are given by the equations:

$$b_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(\xi) \sin n\xi \, d\xi, \quad . \quad . \quad . \quad . \quad . \quad (12a)$$

and

$$a_n = \frac{1}{\pi}\int_{-\pi}^{+\pi} f(\xi) \cos n\xi \, d\xi. \quad . \quad . \quad . \quad . \quad . \quad (12b)$$

Writing the nth harmonic in the form

$$\sqrt{a_n^2 + b_n^2} \sin(\omega t + \varphi_n),$$

we find that the phase angle φ_n is equal to arc tan a_n/b_n. If $f(x)$ cannot be described by one analytical function in the whole region from $-\pi$ to $+\pi$, the integrals in (12) must be split into separate integrals whose limits are those within which the relevant expression for $f(x)$ is applicable. In calculating the first Fourier component of the triangular output voltage (fig. 3) we shall disregard the phase — which has already been found by other means — and choose the zero point of the time axis so as to enable us to use a sine series. We then find

$$b_1 = \frac{2}{\pi}\int_0^{\pi/2} \frac{2\Delta}{\pi} x \sin x \, dx + \frac{2}{\pi}\int_{\pi/2}^{\pi} \frac{2\Delta}{\pi}(\pi - x) \sin x \, dx = 8\Delta/\pi^2. \quad (13)$$

In order to calculate the first Fourier component of the partly sinusoidal and partly linear output signal (the case of fig. 2) we have to ascertain the moments at which a particular part changes to the next. We shall first consider the transition from a sinusoidal to a linear part (fig. 2, point A). For the relevant moment of time t_A we can write:

$$\omega U \cos \omega t_A = B,$$

or

$$\omega t_A = -\text{arc cos}\, B/\omega U. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

Putting $B/\omega U = p$, the value V_A of the output voltage V_0 that occurs at $t = t_A$, and which is equal to $U \sin \omega t_A$, can be reduced using eq. (14) to:

$$V_A = -U\sqrt{1-p^2}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

The equation of the line section AB is then:

$$V_0 = -U\sqrt{1-p^2} + B(t + \frac{1}{\omega}\text{arc cos}\, p). \quad . \quad . \quad . \quad . \quad (16)$$

The point B where the output voltage again becomes sinusoidal is found by ascertaining the value of t at which the line defined by (16) intersects the sinusoidal line:

$$U \sin \omega t_B = -U\sqrt{1-p^2} + B(t_B + \frac{1}{\omega}\text{arc cos}\, p). \quad (17)$$

Putting $\omega t_B = k$, equation (17) transposes to:

$$\sin k + \sqrt{1-p^2} = +kp + p \text{ arc cos } p. \quad . \quad . \quad (18)$$

This equation cannot be solved analytically, and therefore no formula can be derived from it for t_B. We have therefore adopted a graphic solution. In *fig. 9* can be seen how k ($=\omega t_B$)

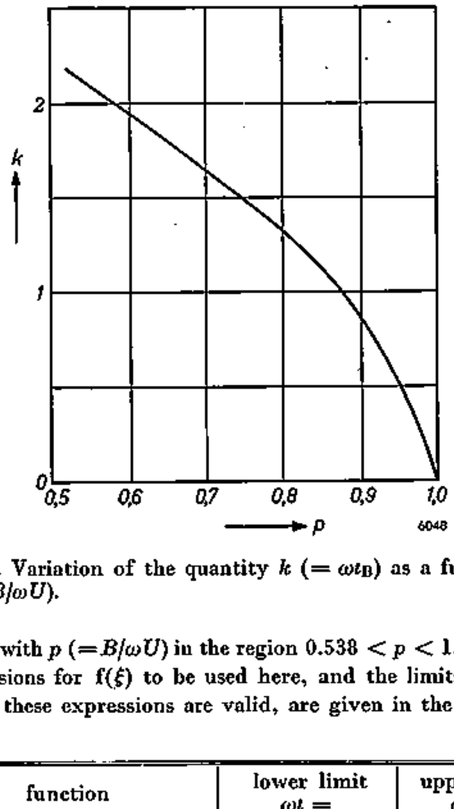

Fig. 9. Variation of the quantity k ($= \omega t_B$) as a function of p ($=B/\omega U$).

varies with p ($=B/\omega U$) in the region $0.538 < p < 1$. The four expressions for $f(\xi)$ to be used here, and the limits between which these expressions are valid, are given in the following table.

function	lower limit $\omega t =$	upper limit $\omega t =$
$Bt + Up \text{ arc cos } p - U\sqrt{1-p^2}$	$-$arc cos p	k
$U \sin \omega t$	k	$\pi -$ arc cos p
$-Bt + Up\pi - Up \text{ arc cos } p + U\sqrt{1-p^2}$	$\pi -$ arc cos p	$\pi + k$
$U \sin \omega t$	$\pi + k$	$2\pi -$ arc cos p

The fact that k can only be calculated numerically does not make it impossible to carry out the integrations analytically (see (5) and (6)). Numerical calculation is required only when it is necessary to determine the variation of the coefficients a_1 and b_1 with p.

Summary. When the speed at which the input signal varies exceeds a certain value, the pen of a recorder is no longer able to follow the signal, but moves uniformly at its maximum speed B. In such a case the recording instrument may no longer be regarded as a linear and lag-free element. The frequency-response characteristics found when the output signal is approximated by its first Fourier component are found to resemble closely those of an element having a single time constant. The value of this time constant, however, is here proportional to the amplitude U of the input signal and inversely proportional to B. When this element is included in a control loop having a further two time constants, the stability of the loop is a function of U. The maximum stability is found at very small *and* at very large values of U.