METHODS OF JUDGING THE ANNOYANCE CAUSED BY CONGESTION

by

CONNY PALM

This posthumous article by the late Docent *Conny Palm* was prepared in connection with the five earlier papers collected in 1946 in a special number of »Tekniska Meddelanden från Kungl. Telegrafstyrelsen» devoted to telephone traffic problems. The references to »S. f. T.» in the article below refer to this special number.

The article, which was mainly completed at the time of the author's death, discusses a proposed new method for the comparison of such congestion phenomena as appear in different forms to the subscriber and which have therefore not hitherto been considered as quantitatively comparable. Since the concepts are of great theoretical and practical interest, *S. Ekberg* of the Swedish Telecommunication Administration has, at the request of the Editor of Tele, completed the article with the aid of Docent Palm's notes.

Introduction.

The investigations of traffic conditions in telephone installations pursued over a number of years in cooperation between the Board of Swedish Telecommunications and Telefonaktiebolaget L M Ericsson have had, as their main purpose, the simplification of, and development of sound foundations for the dimensioning work in the planning of telephone plant. Consequently the work has not merely been an investigation of traffic characteristics and of the functional relationship between different kinds of traffic-carrying groups of circuits and selectors, but has also been directed towards the development of suitable dimensioning standards. Thus two somewhat associated questions have been treated, the calculation of the inconvenience caused to the subscribers by congestion and the problem of how, knowing the extent of this annoyance and the amount of congestion produced in various alternative designs, the various traffic-carrying groups in the plant should be combined to give the best solution with regard to the opposing economic and quality of service aspects.

The dimensioning of the traffic-carrying groups is at present carried out in the well-known way by starting from accepted values for the grade of service and then determining the minimum number of circuits in each group for which the accepted grade of service still is obtained. The traffic density for which this calculation is carried out is the existing or expected peak traffic determined according to established rules. (For example, the average value of a number of busy hours.) A dimensioning procedure on these principles can clearly be considered to be based on the following fundamental ideas:

1. the service quality is measured uniquely by the magnitude of the congestion,
2. the best overall service is obtained if different kinds of groups provide the same service quality.

Against these assumptions several general objections may be raised, and these have resulted in certain modifications to the procedure. In the first place it is not only the magnitude of the congestion, i.e. the ratio of blocked calls to the total number of calls, which affects the quality of the service. The inconvenience produced by a blocked call can differ considerably from case to case and depends in a delay system,[1]) for example, on how long a waiting time is obtained and in a busy signal system on how many repeated attempts must be made before a call is completed. Different groups can thus give completely different service qualities even if the congestions are numerically the same. Efforts have sometimes been made to take these relationships into account by supplementing the congestion calculation procedure, for example by counting as blocked calls only those subjected to a waiting time exceeding a certain value. By these and similar methods some improvement may possibly be obtained, but an exhaustive description of the quality of service cannot result from such

[1]) A busy signal system is a telephone system in which the subscriber receives busy tone when congestion occurs in the switching system. In a delay system the subscriber does not receive busy signal at congestion in the switching system. The call is delayed, being switched when the congestion ceases.

Tele, No. 2, 1953

simple techniques. Another way to graduate the results to take account of the dependence of service quality on factors other than the magnitude of the congestion is to choose different grades of service for different types of groups. For example more blocking is allowed in register groups, since in these the waiting times are generally short. Such methods may, however, easily lead to a point where subjective judgements play so large a part that the uniform design of the installations is seriously endangered. Furthermore, it is a very doubtful procedure to compensate for the effects of certain phenomena by permitting correspondingly different limits to quantities which have little connection with the phenomena in question.

The other assumption for design mentioned above implies that the same »loss» must be allowed in all groups of an installation. In this way uniformity is expected to be obtained, but where this uniformity really is to be found and wath it means seems never to have been ascertained. The most serious argument against this uniformity principle is that it does not give the most economically advantageous plan, having regard to the total service quality of the installation. This is a relationship well known to those who work on dimensioning problems, and attempts are made to take it into account by allowing more loss in groups of circuits which are particularly expensive, such as toll circuits and trunk circuits. It is, however, very uncertain whether in this way, which is affected by subjective judgements, the most economical solutions can be reached.

We can thus see that the bases of the presently generally accepted design principles are extremely unsatisfactory and that the attempts to correct for the faults introduced involve departures from the principles laid down of an extremely arbitrary nature, with a consequent serious reduction in the possibility of objective judgement Because of this liability to uncertainty there often develops a tendency to introduce safety factors in the design which have no economic justification. On the other hand the present methods, if strictly applied, often result in designs near the lower limit with the risk that unnecessarily expensive extension must be undertaken. For the economic development and utilization of the telephone plant it must be regarded as higly desirable to replace the present planning methods by a method which is technically and economically more satisfactory. Since such a change will greatly affect work in the telephone managements and also introduces a series of effects of great importance not only economically but in subscriber relations, any alterations must clearly be preceded by thorough investigation and careful assessment. The points raised in this article make no claim to provide definite grounds for revisions of the present methods, but should rather be regarded as proposals for the lines of attack in the programme of revision work. It leads up to some general principles for the calculation of service quality, which are then applied to the conditions in full availability groups.

In this connection, the following facts must be noted. The dimensioning rules used by a telephone management are usually codified into standard practice systems which lay down how traffic measurement shall be carried out and describe how the calculation of the number of circuits shall be performed on the basis of the traffic values given by the measurements. To simplify the calculations, sets of curves are normally used so that the calculation work is of the simplest nature. Very naturally the introduction of improved design principles with greater gradation will probably result in an increase in the work of computation. In itself there is little to comment on here: the considerable economic advantages at stake would make it well worth while to devote considerably more effort than is devoted at present if a satisfactory solution could be obtained. It would, however, be advantageous from many points of view if the simple nature of the computation work could be retained without losing the possibility of introducing the wanted gradation. This matter must therefore receive some attention in putting forward proposals for new design methods. It must be noted that in this respect the outlines laid down in the following discussion should satisfy quite strict demands.

The quality of service provided by a telephone installation is determined by the sum of the inconveniences to subscribers caused by congestion effects. The magnitude of the inconvenience and the service quality are clearly inversely related quantities, so that when the inconveniences

mount the service quality falls, and vice versa. It has never been found necessary to introduce a direct measure of service quality, but instead a measure of the inconvenience has always been regarded as satisfactory. There would appear to have been no proposals for a change in this respekt. Although the concept of service quality has no direct mathematical counterpart, it may be better retained considering the psychological aspect, since it is more attractive to talk about the service quality of an installation than about its standard of inconvenience. In technical discussions, however, it is more advantageous to work with inconvenience quantities, since the necessary summations can be carried out directly, a process impossible with quantities which give a direct measure of the service quality concept.

The difficulties in choosing a satisfactory measure of the effect of the traffic disturbances due to congestion are mainly to be referred to matters of principle and are associated with the difficulty of determining an annoyance in quantitative terms. The inconvenience suffered by a subscriber as a result, for example, of a delay period, shows as a psychological reaction, and it would seem to be beyond all reason to hope that the amount of such a reaction could be measured and used in technical calculations. This leads, however, to an unduly pessimistic view. In most cases it is possible to say quite safely that one particular event causes, on an average, more annoyance than another. It is true that a 10 second delay on one occasion may be more disturbing than a 20 second delay on another occasion, but on the average a shorter delay time causes less annoyance than a longer. So far as degrees of inconvenience resulting from various events can thus be compared, it should also be possible to find a purely quantitative measure for them. Indeed it is just this which unknowingly has been done, or at least attempted, when the number of blocked calls has been used as a basis for dimensioning. This assumes that all blocked calls can be taken as producing exactly the same annoyance. It should not be impossible to obtain, by reasoning or by measurement, a sufficiently certain view of the quantitative value of the inconvenience caused by various types of traffic disturbance. However, since the psychologists have not yet determined a basic unit for the inconvenience concept, there cannot be any question of anything other than a comparison factor relating different inconveniences. For our purposes it will always be sufficient to have available such a factor.

I must be pointed out in this connection that the terminology in this field presents certain difficulties. To carry on a discussion it has been found necessary to introduce terms for a number of new concepts, and it has often been hard to find adequate and linguistically pleasing expressions.[1])

Inconvenience functions. The forbearance curve.

The inconvenience caused to a calling subscriber by congestion is of quite different kinds in busy signal systems and delay systems. The characteristic difference between the two systems in this respect is that in one case the subscriber must himself repeat the call in order to be connected, without any guarantee that congestion will not again block the call, while in the other case he is informed when the wanted call can be connected. Two forms of the delay system are considered, depending on the form of the information that the call can be completed. In one, the most common, the subscriber must wait with the telephone to his ear until a characteristic tone, or an operator, announces that the congestion has ceased. In the other case the subscriber need not wait with the microphone off the hook but is informed by being rung that the call can be obtained. An example of this is provided by normal Swedish trunk traffic. Another example is offered by a certain type of small automatic or semi-automatic exchanges, in which there is storage of blocked calls. These then receive first of all a busy signal, which consequently is not a characteristic of the busy signal system only. Such a delay system, which is associated with a ringing signal when the call can be completed, will be referred to in the following discussion as a *delay system with back signalling.* To differentiate, the usual delay system in which the subscriber must wait with the microphone off hook during congestion will be referred to as a *delay system without back signalling.*

In considering the nature of the inconvenience

[1]) In translation this difficulty is accentuated and the terminology is to be regarded as a tentative.

caused to subscribers by congestion we must thus consider three different kinds of system:

delay system without back signalling
delay system with back signalling
busy signal system

There are also some exceptional arrangements which can be regarded as combinations of these systems. We shall later consider an example of the treatment of such a *combined system.*

Delay system without back signalling.

We shall consider first the conditions in a delay system *without back signalling* and to begin with we apply reasoning which is perhaps not completely free from objection but which gives a guide in the assessment of the annoyance relationships. We consider a particular blocked call, that is, a subscriber who initiates a call and is subjected to delay. The waiting causes irritation to the subscriber and we can consider that the annoyance caused by a delay time t is the value of the irritation accumulated in this time. The annoyance must clearly be a function of the delay time t and we can call this the *delay time inconvenience function* $I(t)$. To find a plausible form for this function we can make use of the following reasoning. Assume that the subscriber is made to wait for a further time dt. The inconvenience increases by an amount dI. This quantity dI can, by what has just been said, be considered to give a measure of the subscriber's irritation at time t, and this must also mean that the subscriber's reaction is dependent on the quantity dI. Now we know that a subscriber who is subjected to an indefinitely prolonged delay time will always tire of waiting in the end and will replace the microtelephone. This implies that the irritation dI must have increased to the point when a reaction is provoked which clearly tends to reduce the irritation. It can be seen that dI must increase with the delay time t. A very plausible assumption is to put

$$dI = c \cdot t^{\lambda} dt \qquad (1)$$

Here c is a constant, which must obviously always be positive, and from the discussion above the exponent λ must be positive in order that dI shall increase with t. Clearly any increasing function containing suitable parameters may be considered instead of (1). However, (1) gives the simplest imaginable form which fits the preceding reasoning and there does not appear to be any reason for introducing any more complicated form. It must be noted that in spite of its simple form (1) contains two parameters, c and λ, so that it can give good flexibility for fitting to known requirements.

The core of this reasoning, which leads to the setting up of (1), is clearly the conclusion that the irritation dI must increase with delay time t, since ultimately it provokes a reaction, viz. the termination of the waiting. Furthermore, it is clear and of great importance for the reasoning which follows, that this reaction, like every physical and psychological reaction, tends to diminish the existent strain. The subscriber prefers to break off the waiting and thus reduce the irritation once it has reached a certain level.

This reasoning may appear to be very trivial but it is in fact of such importance that it must be discussed. In a later application a general discussion of possible objections is given.

By integration of (1) we obtain

$$I(t) = \frac{c}{1+\lambda} t^{1+\lambda} \qquad (2)$$

in which the integration constant is set to zero since the inconvenience caused by a zero delay time must be zero. The process of integration can be interpreted as the determination of the accumulation of irritation just discussed.

It follows from the deduction that c in (2) must be a constant independent of the delay time. It has already been pointed out, however, that there is no absolute unit for the inconvenience concept, so that c cannot be numerically determined. On the other hand it is quite safe to say that c can have very different values in different circumstances. It will not only have different values for different subscribers, but even for any one subscriber it will vary on different occasions, according to the nature and importance of the wanted call and the mood of the subscriber. We can thus say only that for each blocked call there is a constant c but that this constant cannot be assumed the same for every call. This would seem to act as a complete barrier to the possibility of any practical application of the proposed inconvenience function. Fortunately it is found to be quite easy to remove this barrier. All the c-values for the

blocked calls in a particular group of circuits are regarded as forming a sample space in other words, we introduce a probability distribution for the various c-values. A density function $g(x)$ is defined so that $g(x) \cdot dx$ gives the probability that a c-value belonging to a randomly chosen blocked call in the group will have a value between x and $x + dx$. We can say also that $g(x) \cdot dx$ gives the relative frequency of all the c-values lying between the limits x and $x + dx$. It should be noted that the parameter c can hardly be thought of as having negative values, so that $g(x)$ need only be defined for $x \geq 0$. Since the sum of all possible probabilities is always unity, the first necessary condition is

$$\int_0^\infty g(x) \cdot dx = 1 \qquad (3\,a)$$

If now γ is the mean of all possible c-values, we have, according to the usual rule for determining the first moment

$$\gamma = \int_0^\infty x \cdot g(x) dx \qquad (3\,b)$$

We can now determine quite easily the mean of the inconvenience caused by a delay of duration t. The contribution from calls which have a c-value between x and $x + dx$ is, according to (2)

$$\frac{x}{1+\lambda} t^{1+\lambda}$$

and the relative frequency of such calls is $g(x) \cdot dx$. It will easily be seen that the wanted mean is given by

$$\int_0^\infty \frac{x}{1+\lambda} t^{1+\lambda} g(x)\, dx$$

From (3 b), however, this integral must be equal to

$$I(t) = \frac{\gamma}{1+\lambda} t^{1+\lambda} \qquad (4)$$

This expression thus gives the inconvenience caused by a waiting time t to a randomly chosen blocked call in the group under consideration. This inconvenience function has exactly the same form as in the case of the inconvenience of a single call given in (2), the constant γ now being the mean of all possible c-values. The constitution of the density function thus has no effect on the inconvenience function's dependence on t.

The assumptions introduced so far are of a very general nature but in order to be able to use the results it is necessary to become more specific, since the value of λ must be determined. From the earlier reasoning for the setting up of equation (1) for the irritation it follows that λ must certainly be ≥ 0. Otherwise it would be impossible to use a form such as (1), since for negative values of λ it gives infinite irritation for $t = 0$, a somewhat unnatural assumption. It may be asked, however, whether it is not necessary to assume that for each different call there is a different λ-value in the same way as in the treatment of the c-value. It might also be hoped that by introducing a density function of the same kind as used in the c-value treatment an equally simple result might be obtained. This, however, is not the case. Every attempt to introduce a number of different λ-values into the same inconvenience function has led to the result that the calculation of the total inconvenience caused by congestion becomes practically impossible to carry through, as we shall see later. So long as we keep to an inconvenience function such as (2), which is, moreover, the most simple, we must for purely practical reasons assume that the exponent λ is the same for all cases at least for all cases in the same space under treatment, such as all blocked calls in a single group. There would be no point in proceeding from such a simple form as (2) if the immediate result was to force the use of the sum of a finite, or infinite, number of terms with different exponents, which would be the case if a density function were to be assumed for λ.

The approximation forced upon us by the assumption of the same λ-value for all cases should, in fact, not have any great significance in practice. It is unlikely that different calls will be associated with large relative deviations in respect of λ-value, and in such cases a function of the type shown in (4) gives a numerically good approximation to the sum of different powers which would otherwise be needed. In addition, it can be assumed that the unlimited flexibility of c-value gives a sufficiently wide flexibility to deal with the inconvenience relationships of in-

dividual calls so that it would be superfluous to take account also of variations in λ-value.

The limitation imposed by practical considerations to a single λ-value for each call space need not induce any apprehension. It is found in addition that for practical application of the results, consideration can only be given to some few possible values for λ, the lowest integers and zero. The expressions which are obtained for the total inconvenience caused by congestion are found to contain integrals of the product of the function $I(t)$ and the delay time density functions, and these integrals are extremely difficult to evaluate if there is no limitation to integral values of λ. Moreover, the integrals take simpler forms as lower integers are used for λ. For convenience, in fact, it is a question of choosing 0 or 1, or possibly 2. For $\lambda = 0$, (4) becomes a linear function of t, and the derivative, which is the irritation, is independent of t. This, however, is in conflict with the whole idea of the reasoning based on irritation as a cause of the phenomenon that a subscriber, subjected to an unlimited delay, sooner or later breaks off the wait, since this implies that the irritation must grow to a value which is greater than that for $t = 0$. At this point, in fact, the irritation must for natural reasons be zero, so that necessarily $\lambda > 0$.

Even if the reasoning about the connection between the irritation and the causes of the broken-off wait are not accepted there are other strong reasons which support the assumption that $\lambda > 0$. These will be expounded later. Here it will only be stated that there are many reasons for regarding $\lambda = 1$ as a very plausible value. In this case, from (4),

$$I(t) = \frac{\gamma}{2} t^2 \qquad (4\,a)$$

The derivative of $I(t)$ is then γt, which implies that the irritation is directly proportional to the elapsed delay time, which seems to be very natural.

The circumstance that for convenience λ must be chosen as an integer, and that there is really only unity to be considered, is hardly such a great disadvantage as might be believed. The λ-values which can, on the whole, be considered certainly do not lie outside the range 0.5—1.5. Now, it so happens that in a later stage of the calculations, the results obtained for $\lambda = 1$ can give an approximate idea of the results which would be obtained for all λ-values in this range. By beginning with a choice of $\lambda = 1$ all possibility of judging the effects of other values of λ is not lost.

The empirical determination of the exponent λ.

Before considering the inconvenience function in a delay system with back signalling we shall say something about a method which appears to give the possibility of determining λ empirically, at any rate in principle. In this connection the results of some measurements, which have a direct bearing on this procedure will be shown. The method is based on the reasoning above, according to which the termination of a waiting period is caused by the subscriber's irritation reaching a particular value. The starting point is in the measurements of how long subscribers will wait when subjected to an unlimited delay. It is thus a question of the earlier (p. 63 in »S.f.T.») mentioned fall-away distribution which is the basis for the theoretical treatment of a delay system with voluntary break-away for the waiting caller. In this treatment the distribution was assumed to be purely exponential, with the aim of obtaining a simple mathematical treatment. The fact that the fall-away distribution in reality will be somewhat flatter than a pure exponential curve does not have any appreciable effect on the accuracy of the results obtained. In the present connection, however, the shape of the fall-away distribution is of great interest, since it is found to be intimately associated with the value of the constant λ.

We shall write the fall-away distribution as $\psi(t)$ which gives the probability that a subscriber who has a delay time at least as long as t does not give up waiting during this period. We have thus to deal with a decreasing distribution function starting at $\psi(0) = 1$. The probability that a subscriber will wait for a time t and will then, during the next time interval dt, tire of waiting, is, according to the general laws for decreasing distribution functions (see the first article in »S. f. T.»),

$$-\psi'(t)\,dt$$

Now the probability that a subscriber waits for at least t is just $\psi(t)$, and we can see that the probability that a subscriber who we know has

already waited for a time t will tire during the immediately following interval dt must be

$$-\frac{\psi'(t)}{\psi(t)}dt$$

Now this probability must be proportional to the amount of irritation in the corresponding time interval, that is, to $dI(t)$. Using eq. (1) we thus have

$$-\frac{\psi'(t)}{\psi(t)} = c_0\, c \cdot t^{\lambda}$$

in which c_0 is a constant. By integration of this we obtain

$$\psi(t) = e^{-\frac{c_0 c}{1+\lambda} t^{1+\lambda}} \tag{5}$$

in which the integration constant is determined by the fact that $\psi(0) = 1$.

In the derivation of (5) a certain delay time has been considered, the inconvenience function of which contains the constant c. We must, however, take into account the fact that other waiting calls may be associated with different values of the constant c. Furthermore, it is very probable that the constant c_0 introduced above will also have different values in different cases. We must therefore proceed in a way similar to that used in the discussion of the inconveniance function $I(t)$ and introduce a density function $g(x)$, which in this case takes account of the variation of the product $c_0 c$. We define this in such a way that $g(x)dx$ is the probability that $c_0 c$ has a value between x and $x + dx$. For this function $g(x)$ the relations (3 a) and (3 b) above hold, the latter, however, now giving another mean which may be written $\gamma_0\gamma$. The distribution function $\psi(t)$, which gives the probability that a randomly chosen waiting caller will not tire of waiting in a delay of t is now no longer given by (5), but, as can easily be seen, by

$$\psi(t) = \int_0^{\infty} e^{-\frac{x}{1+\lambda} t^{1+\lambda}} g(x)\, dx \tag{6}$$

It can be convenient to have a separate term for the fall-away function in this general and natural form, and we may refer to (6) as the *forbearance distribution*. Unlike the case with the inconvenience function $I(t)$ discussed earlier, the density function $g(x)$ does not now disappear from the result but has a considerable effect on the shape of the curve. If we had $\lambda = 0$, which is certainly less than the true value, $\psi(t)$ would be an ordinary completely monotone function. The characteristics of completely monotone functions were discussed in the first article in »S. f. T.» where it was shown that a completely monotone function always has a flatter shape than the corresponding exponential function with the same mean. For $\lambda = 0$ it follows that (6) is flatter than (5) if the mean $\gamma_0\gamma$ is inserted for $c_0 c$ in the latter. A closer study of functions of the type (6) has now shown that this is a strictly general characteristic valid for all values of λ. The density function $g(x)$ thus has the effect of making (6) have a flatter shape than (5) when for comparison purposes the same mean is obtained by putting $\gamma_0\gamma$ in (5) instead *of* $c_0 c$.

Now, a question of great importance is whether (6) is reversible i.e. whether both λ and the function $g(x)$ are uniquely determined by $\psi(t)$, and how this determination can be carried out in practice. In fact, if by measurement a fall-away distribution, which must have the form of (6), is obtained, it would then be possible to derive from this a value of λ. In this connection the function $g(x)$ is of less interest; it must be pointed out, however, that it is not the same as the density function of the inconvenience coefficient c. The density function in (6) now represents the variation in $c_0 c$.

It can now be shown that the exponent λ in (6) is uniquely determined by the behaviour near the point $t = 0$ of the function $\psi(t)$. If we differentiate (6) we obtain

$$-\psi'(t) = t^{\lambda} \int_0^{\infty} x \cdot e^{-\frac{x}{1+\lambda} t^{1+\lambda}} g(x)\, dx$$

Allowing $t \to 0$ this reduces to

$$\lim_{t \to 0} \frac{-\psi'(t)}{t^{\lambda}} = \int_0^{\infty} x \cdot g(x)\, dx$$

The right hand side is the mean of $g(x)$ as defined by (3 b), and in this case, as already noted, it is $\gamma_0\gamma$ and not γ. We thus have

$$\lim_{t \to 0} \frac{-\psi'(t)}{t^{\lambda}} = \gamma_0\gamma \tag{7 a}$$

This now allows us to determine λ. If we investigate the expression

$$\lim_{t \to 0} \frac{-\psi'(t)}{t^a} \qquad (7\,b)$$

for various values of a, it follows from (7 a) that this limit goes to zero for $a < \lambda$ and goes to infinity for $a > \lambda$; only for $a = \lambda$ has it a finite value greater than zero. This finite value gives the mean $\gamma_0\gamma$ directly.

If, after the determination of λ it is also desired to determine $g(x)$ it is only necessary to make a substitution

$$z = \frac{t^{1+\lambda}}{1+\lambda}$$

when (6) is converted into a normal *Laplace* integral of the basic form shown as (7), page 6 of »S. f. T.» A number of methods have been developed for the transformation of this integral but these are difficult to apply in practice. Since we have no special interest in $g(x)$ here, these methods will not be discussed.

Limit relationships other than (7 a) can be obtained for the determination of λ. Thus after expansion in series of the exponential term in (6) we obtain

$$\lim_{t \to 0} \frac{1-\psi(t)}{t^{1+\lambda}} = \frac{\gamma_0\gamma}{1+\lambda} \qquad (8\,a)$$

and also

$$\lim_{t \to 0} \frac{\ln \psi(t)}{t^{1+\lambda}} = \frac{\gamma_0\gamma}{1+\lambda} \qquad (8\,b)$$

A common feature of all the possible methods is, however, that the shape of the curve must be analysed in the region of the point $t = 0$, and this greatly increases the difficulty of any practical application, since it is just this range which is in general hard to determine accurately by measurement.

Before we go on to show the results of some measurements we may treat in more detail the form of $\psi(t)$ for the conceivable value $\lambda = 1$. From (5) we obtain (writing here $\gamma_0\gamma$ instead of c_0c)

$$\psi(t) = e^{-\gamma_0\gamma\frac{t^2}{2}} \qquad (5\,a)$$

This expression is reminiscent of the normal *Gaussian* distribution. There are, however, a number of significant differences. Negative values of the variable t are never considered, and (5 a) is the distribution function, whereas in the normal distribution the corresponding expression is the density function.

The mean of the distribution (5 a), the *average fall-away time*, is given by

$$\int_0^\infty e^{-\gamma_0\gamma\frac{t^2}{2}}\,dt = \sqrt{\frac{\pi}{2\,\gamma_0\gamma}} \qquad (5\,b)$$

The second moment of the distribution is

$$2\int_0^\infty t \cdot e^{-\gamma_0\gamma\frac{t^2}{2}}\,dt = \frac{2}{\gamma_0\gamma} \qquad (5\,c)$$

A form factor was introduced in the first article in »S. f. T.» and defined as the ratio of the second moment to the square of the mean. In this case it is $4/\pi = 1.273$ and is thus independent of the constants γ_0 and γ. The value of the form factor shows that the distribution is appreciably steeper than the simple exponential distribution, which has a form factor of 2.

If instead we treat the more general form (6) for the special value $\lambda = 1$ it can be shown that as a result of the density function $g(x)$ the form factor will always be greater than the value $4/\pi$ just mentioned. This is related to the circumstance previously shown that (6) in general is a flatter type than (5).

Practical determination of λ.

There are generally serious difficulties in carrying out measurements on the fall-away time of subscribers. In normal groups congestion is a rare phenomenon, and measuring equipment can only be applied to a limited number of subscribers, so that it can take a considerable amount of time before sufficient data are accumulated. On the other hand it is possible to introduce experimentally an unlimited amount of congestion and in this way to obtain more data. This is, however, a rather unattractive method because of the resulting disturbances. In the measurements referred to below, however, specially favourable conditions could be used which are no longer available to the same extent. The measurements were made during 1938 in Stockholm by *G. A. Ankarberg*. At this time there was a large number of non-automatic suburban exchanges in the Stockholm area. Among the group selector levels

for outgoing junctions of the automatic junction traffic exchange there were thus some levels which had no connection to any succeeding selector stage. Relatively often, however, subscribers would be connected to these levels through errors in dialling, and obviously they would not receive any ringing tone, so that they were presented with a delay time of unlimited duration. The time during which a subscriber making such a false call continued to wait before hanging up was determined from the number of times the wiper of the selector which in this instance was a 500-point Ericsson selector went into the empty levels. The measured times are thus approximate; the object of the measurements at that time did not require higher accuracy. For the application of the fall-away curve now under discussion greater accuracy would clearly have been advantageous, especially as regards the shorter waiting times. Despite this defect the data have proved to be of great interest.

One circumstance which must be indicated in connection with these measurements is that the fall-away times may obviously be different depending on which stage of the connection is the source of the congestion or stops the progress of the call. In an early stage congestion may have the effect of preventing the subscriber receiving dialling tone. In the case to which the measurements refer the subscribers received dialling tone and dialled the wanted number, but then did not obtain the normal ringing tone at the earpiece. It is therefore possible that the average of the time a subscriber will wait would be found to be different if the corresponding measurements were carried out in an earlier connecting stage. On the other hand it is highly probable that the shape of the fall-away curve will be similar in all cases, and it is only the shape which is of interest in determining λ.

The measurement covered 2140 cases and gave 29.0 seconds as average fall-away time. The curves obtained are shown in Fig. 1. Curve 1 in the upper diagram shows the distribution function obtained from the measurements. If we compare this with curve 3, which shows the exponential function for the same mean, $e^{-t/m}$, where m is the average value above, it is seen immediately that the measured distribution is steeper. This shows again that λ must be greater than zero since for $\lambda = 0$, according to what has already been said, we must have a curve which is flatter than the corresponding exponential function. The same condition is shown by the form factor, which is 1.62 for the measured distribution, while it is always greater than 2 for a flat distribution.

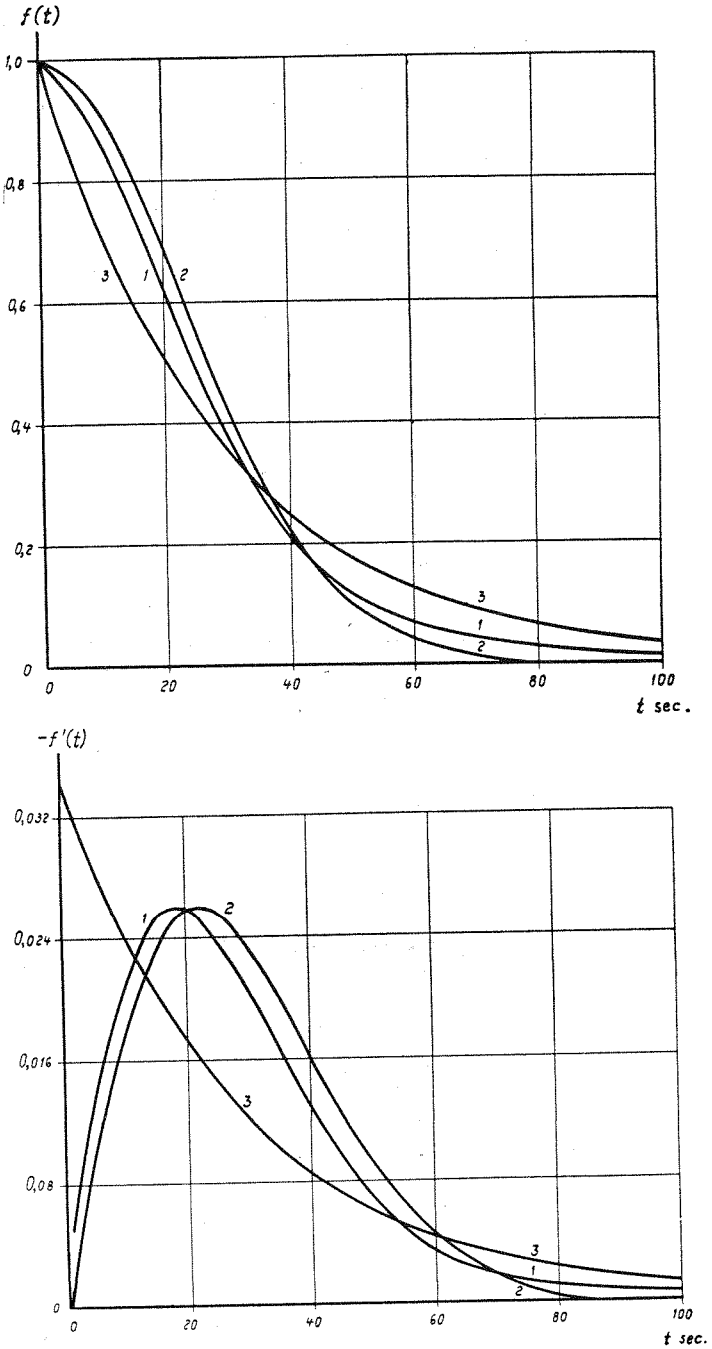

Fig. 1. Different distribution curves with the same mean value, 29 sec. *1* measured curve, *2* curve according to eq. (5a), *3* exponential distribution. The upper part of the figure shows the distribution functions and the lower part the density functions.

Curve 2 in the upper diagram shows the theo-

retical distribution function (5 a), which using the average value $m = 29$ sec. has the form

$$e^{-\pi\left(\frac{t}{2m}\right)^2}$$

As will be seen, this distribution is still steeper than the measured one. The form factors are respectively 1.27 och 1.62. However, as we have already shown, the effect of the density function $g(x)$ is always to make the general distribution (6) flatter than the basic distribution (5). The measured curve thus shows just precisely the type of shape which should be obtained for $\lambda=1$. It is also of interest to note that the measured curve is nearer the theoretical (5 a) than to the simple exponential function.

From the preceding discussion the criterion that a fall-away distribution should be of the type given by (6) with a certain λ-value is that its shape in the neighbourhood of $t = 0$ should be the same as for the function (5 a). For reasons already mentioned the measurements in this region are unfortunately not sufficiently accurate to permit any satisfactory application of the limit conditions of (7) or (8). We can, however, obtain valuable results by a study of the frequency curves of the distributions. These are shown in the lower diagram of Fig. 1, and have been constructed from the differences between successive values of the distribution curves. It will be seen that curve 1 obtained from the measurements and the frequency curve given from (5 a) are of similar types, and the difference is just of the kind which should be introduced by the effect of the density function $g(x)$. In contrast, the frequency curve 3 obtained from the simple exponential function is of a completely different type. From these results and from the approximate shape of the measured curve in the neighbourhood of $t = 0$ the following conclusions can be drawn:

λ is certainly greater than zero,

λ is very probably not appreciably greater than unity,

λ is probably very nearly unity, possibly slightly less than one.

The measurements referred to have thus given a result which can be taken as good support for the assumption of a quadratic inconvenience function of the form (4 a), with $\lambda = 1$. It should, however, be noted that the whole result is built on the assumption that the magnitude of the irritation is the deciding factor in settling the time at which a waiting subscriber decides to stop his waiting. If the reasoning which leads to this assumption is not considered to carry any weight, measurements of the fall-away distributions can never give any indication of the magnitude of λ. Unfortunately in such a case it would be impossible by any experimental method to arrive at a value of λ.

Delay systems with back signalling.

The conditions discussed so far have related to delay systems without back signalling, in which the waiting subscriber must wait continuously to receive in the receiver the information that the congestion has ceased. We shall now investigate the corresponding conditions in a delay system *with back signalling*. In this case it is not necessary to listen in the receiver while waiting, since an announcement that the congestion has ceased is given by a ringing signal. The subscriber can therefore occupy himself with other tasks while waiting so that the total inconvenience caused by the wait should generally be less in this case than in a delay system without back signalling. This is one of the reasons which justify the introduction of the system with back signalling. It is clear, however, that for very short delays, e.g. 1—2 seconds, it is less disturbing to wait at the microtelephone than to hang up and then take up the telephone immediately when the ringing signal is received. For short delays therefore, a delay system without back signalling is preferable to a system with back signalling. This discussion shows that the inconvenience function for a delay system with back signalling, which will be written here as $I_0(t)$, should be greater than $I(t)$ for small values of t, and less than $I(t)$ for large values of t.

To find a plausible form for $I_0(t)$ we can consider the irritation by a similar reasoning process to that used for (1) and take it as proportional to a power of t:

$$dI_0 = \omega_0 t^{\lambda_0} dt \qquad (9)$$

in which ω_0 is a constant (the notation is chosen for reasons which will be discussed later). By integration of (9) we now have, assuming the integration constant to be determined by $I_0(0)=0$

$$I_0(t) = \frac{\omega_0}{1+\lambda_0} t^{1+\lambda_0} \qquad (10)$$

Comparing this with (2) it is seen that we must have $\lambda_0 < \lambda$ if the inequalities between $I_0 t$ and $I(t)$ discussed above are to hold. Since we found earlier that $\lambda = 1$ was a plausible value, the discussion of suitable values of λ_0 can be limited to the range below unity. On the other hand it seems hardly likely that λ_0 should be negative, for in such a case the increase of inconvenience per unit time would diminish as the delay time increased. We can find no justification for such a view. It is thus possible to fix, a priori, the limit $0 \leq \lambda_0 < 1$. There appear, unfortunately, to be no possibilities of measurement in the manner used for λ to decide which value inside this range should be used for λ_0. It is possible to get some guidance by considering the irritation dI_0. We have indicated earlier as one of the reasons why λ must be greater than zero that the irritation in a delay system without back signalling must increase with the waiting time, since sooner or later it always produces a reaction in the subscriber, who hangs up. So far as λ_0 is concerned a corresponding reason can hardly be produced and there does not seem to be any justification for assuming that the irritation increases with time. This leads us to assume that $\lambda_0 = 0$, which from (9) indicates that the irritation dI_0 is constant during the waiting time. A further reason for this choice is that in this case, as in the previous one, convenience in practical application must be considered, and non-integer values of λ_0 make the mathematical work extremely complicated. Furthermore, in this case as in the preceding one, any consideration of the deviation of the value of λ_0 from the assumed value can to some extent be carried through approximately at a later stage of the calculations.

As a plausible form of the inconvenience function in delay systems with back signalling we thus have

$$I_0(t) = \omega_0 t \qquad (10\,a)$$

which thus provides a measure of the inconvenience caused by a wait of length t.

It is clear that the constant ω_0 must, like the constant c in (2) be capable of having different values in different circumstances. If, then, as was previously made in the case of $I(t)$, a density function $g(x)$ is introduced which in the present case gives the probability of different ω_0-values, we find through a corresponding integration a result corresponding to (4) and having the same form as (10 a) although ω_0 in this result represents the mean of the constants for the individual delays. If the constant ω_0 is defined as such a mean then the expressions (10) and (10 a) are also valid for the whole collective of waiting calls.

Busy signal systems.

We shall finally consider the conditions in *busy signal systems*. With these, congestion makes it necessary for the subscribers themselves to take action in the form of renewed calls to obtain a connection. The delay time for a subscriber subjected to congestion is here the interval between the initiation of a call which suffers congestion and the moment when a renewed call finds that the congestion has ceased. This is, therefore, as in a delay system, a question of the time between demand and completed call. A fundamental difference from the delay system is that in a busy signal system the subscriber can affect the duration of the delay by his own actions.[1]) If after meeting congestion he calls repeatedly with a short delay between tries he has a high probability of getting through immediately the congestion ceases. If on the contrary he allows a longer time to elapse between each repetition it may happen that his waiting time is unnecessarily long, since the congestion has already ceased. It may also happen that congestion has developed again and the subscriber has missed the intervening clear period by not repeating his call sufficiently often. The retardation time should thus, on the average, diminish with reduced spacing between renewed calls. On the other hand the subscriber's trouble increases with a reduction of the spacing, since each renewed call causes a certain amount of trouble. It may therefore be expected that during congestion the subscribers will alter the spacing of renewed calls according to the importance of rapid completion of the wanted connection.

The inconvenience experienced by a subscriber subjected to congestion in a busy signal system is clearly composed of two terms: the annoyance

[1]) For this reason there will be used in the following discussion of conditions in busy signal systems the terms *retardation* and *retardation time* as distinguished from *delay* and *delay time* in delay systems.

caused by the retardation time and the work caused by the call renewal. During the retardation time the subscriber need not listen at the receiver so that the inconvenience caused by the retardation itself should be of the same nature as that in a delay system with back signalling, for which a linear form (10 a) of annoyance function was found to be reasonable. When we come to the work caused by renewal of calls it is convenient to assume this to be the same for each such renewal, independently of how many renewals the subscriber has made in the case in question. If, therefore, we consider a subscriber who receives busy tone because of congestion and then makes n fresh attempts during time t, all receiving busy tone, except the last which is completed at time t, the total inconvenience is expressed as

$$\omega_0 t + a \cdot n$$

The first term is the inconvenience caused by the retardation time t itself and is derived from (10 a). The second term is the inconvenience of the n repeated calls, in which a is the inconvenience per call initiation. It will be clear that not only can the constants ω_0 and a have different values for different blocked calls, but also the number of repeated calls n can vary widely even for the same retardation time. Among the total set of congested calls we consider first only those having certain common values of ω_0 and a. Among these there will be found, with varying probabilities, different values of a. The object will then be to determine the mean value of n as a function of t. To do this it can hardly be assumed that every subscriber makes the necessary call renewals with the same constant intervals. It may rather be assumed that the renewed calls are made at random with a particular average spacing. Let us take a particular blocked call and write the average number of call renewals per unit time as y. We have now assumed that at time t there is an effective call renewal (this is actually the call which ends the retardation. The probability that in a retardation time t a total of n call renewals occurs is thus equal to the probability that in time t there are $n-1$ ineffective call renewals (the probability of the last call is unity, since this call must occur at time t in order that the retardation time shall be just t). If now the call renewals occur at random with an average of y per unit time, the probability for $n-1$ ineffective calls in time t is expressed, as is known, by

$$\frac{(yt)^{n-1}}{(n-1)!} e^{-yt}$$

This is also the probability for a total of n call renewals, since the terminating call must always occur. The mean value of the number of call renewals is thus

$$\sum_{n=1}^{\infty} n \frac{(yt)^{n-1}}{(n-1)!} e^{-yt}$$

This sum is easily calculated and is $yt + 1$. The average inconvenience for a subscriber who is blocked for a time t is thus

$$\omega_0 t + ayt + a$$

This expression now is deduced for given constant values of ω_0 and a but it must be pointed out that as before ω_0 and a can vary for different blocked calls. To avoid making the assumptions introduced above too narrow we must, moreover, consider also the possibility that the mean value y varies for different blocked calls. These variations do not introduce any new complications. We need only introduce a density function for each of the three quantities ω_0, a and y, and then proceed as before to determine, the mean value of the inconvenience expression. The density functions do not occur in the result and the inconvenience expression becomes invariant if ω_0, a and y are now taken to be the mean values of the respective quantities.

If we introduce

$$\omega = \omega_0 + ay \tag{11}$$

the resulting inconvenience function takes the form

$$I_1(t) = a + \omega t \tag{12}$$

The inconvenience function of a busy signal system is thus a linear function of time like $I_0(t)$. From the definition of ω it follows that it must always be $> \omega_0$, which is obvious, in any case.

The constant term a in $I_1(t)$ represents the inconvenience associated with the last call initiation, which leads to the completion of the call. It might alternatively be said that it represents the inconvenience of the first call, since a call which is completed directly is a normal event, which can hardly be assumed to cause any inconvenience. In fact, acceptance of this principle should

lead to the introduction also in $I_0(t)$ of a constant term a representing one of the two calling operations needed when congestion occurs in a delay system with back signalling. It is however, very doubtful whether such a constant term should be introduced in either case. Neither the first call, which was blocked, nor the last, which was completed, actually occurs during the retardation time itself. Neither of these calls is thus so disturbing to the subscriber as fruitless renewed calls during the retardation time itself, which to some considerable extent hinder the subscriber in carrying out other tasks during the retardation period. The value of a in the constant term in (12) should therefore be less than the value of a in the definition of ω. The question is whether it would not be safe to drop the constant term in $I_1(t)$ completely. This would also offer considerable practical advantages. The retention of the constant term in $I_1(t)$ also necessitates a knowledge of at least the ratio a/ω which would be difficult even to estimate roughly. Furthermore the computations needed for design work will be appreciably simpler if the constant terms are not present in the inconvenience functions. There are thus strong reasons why for further discussion the following simpler form should be taken for $I_1(t)$:

$$I_1(t) = \omega t \qquad (12\,a)$$

It is of interest to compare the three inconvenience functions in the forms (4 a), (10 a) and (12 a). Fig. 2 shows their general character. So far as the three coefficients γ, ω and ω_0 are concerned we still do not know their relative orders of magnitude except that, as we have just seen, $\omega > \omega_0$. As a result of this $I_1(t)$ is always above $I_0(t)$ as shown in the figure. From Fig. 2 some general conclusions can be drawn regarding the advantages and disadvantages of the various systems. If a group is so designed that the delay times or congestion times are in general less than the cross-over point shown on the figure as t_0 we can clearly expect that the usual delay system without back signalling will cause the least inconvenience. If the delay times are

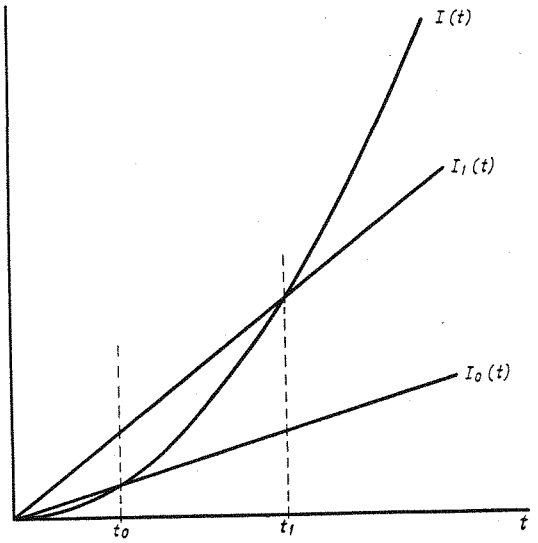

Fig. 2. Inconvenience functions: $I(t)$ for delay system without back signalling, $I_0(t)$ for delay system with back signalling and $I_1(t)$ for busy signal system.

mainly greater than t_0 a delay system with back signalling should give the least inconvenience. For all delay times less than the value at the other cross-over point in the figure t_1, the ordinary delay system without back signalling is advantageous compared with the busy signal system. If however, the delay times average more than t_1 the busy signal system is preferable to a delay system without back signalling. Finally it should be noted that a delay system with back signalling always gives less inconvenience than a busy signal system.

These comparisons naturally apply only to groups which are otherwise similar, with the same number of circuits and the same applied traffic load.

Finally it must be pointed out that in the forms of the inconvenience function discussed here the coefficients must have the following dimensions:

γ has dimension inconvenience/(unit time)^2,

ω and ω_0 have dimensions inconvenience/unit time.

It is with these dimensional forms in mind that the symbols γ and ω were chosen, since they are widely used in other fields for quantities with similar time dimensions.

Disturbance calculations.

With the aid of the expressions for the inconvenience caused by various delay times, or more generally various lengths of time between initiation and completion of a connection, which were proposed in the immediately preceding sections, it should now be possible to calculate the *aver-*

age inconvenience per call in groups of various kinds. It has been thought convenient to introduce a special expression for the amount of inconvenience to which the subscribers are subjected by the influence of the unavoidable congestion. A convenient term which is proposed is *disturbance,* because a state of congestion and its consequences can be regarded as a disturbance of operation, or a traffic disturbance, which the calling subscribers encounter during the congestion period. The disturbance concept may be defined more closely in the following way: the disturbance in a group of circuits for a particular traffic is equal to the sum of the inconveniences to which the subscribers are subjected by calls which are part of the traffic under consideration being blocked within the group. The dimensional unit for disturbance is clearly the same as for inconvenience and is suitably linked with the inconvenience function used. It should be noted that the disturbance is defined above as an absolute quantity, roughly in analogy with an absolute traffic unit, although there is nothing to prevent the reference of a disturbance to a particular traffic time or to a certain total number of calls. The term *relative disturbance* will now be taken to mean the average disturbance per call in the traffic under consideration. All calls must be counted here, whether they were blocked or not. A more precise definition is: the relative disturbance in a particular group subjected to a particular traffic, and during a certain time, is obtained by dividing the total disturbance for the same group and same traffic at the same time by the total number of calls in the traffic during the time.

A group of circuits incorporated in a delay system will now, during a particular period of time T be subjected to a total of yT calls constituting a particular traffic for consideration and having the call frequency y. The group may in other respects, be of arbitrary type; there may, therefore, be other traffic present applied to the group, which means that there is grading. We write R for the proportion of delayed calls in the group out of the traffic under consideration, which means that R is the fraction of its total calls, yT, during time T, which are delayed in the group and which can thus be regarded as blocked. We now write $F(t)$ as the fraction of the blocked calls which must wait for at least time t. As is easily seen, this gives $-dF(t)$ as the fraction of the blocked calls which suffer a delay time between t and $t + dt$. Assume now, as before, that $I(t)$ is the inconvenience caused by a delay time of this duration. The total number of calls originated during the time T in the traffic under consideration which are delayed for between t and $t + dt$ will thus be

$$-yTR \cdot dF(t)$$

and every such call provides a disturbance of magnitude $I(t)$. The total disturbance for calls originated during time T is thus expressed by

$$-yTR \int_0^\infty I(t) \cdot dF(t)$$

To obtain the relative disturbance we must divide this by the total number of calls, yT. Thus

$$-R \int_0^\infty I(t) \cdot dF(t) \qquad (13)$$

is the relative disturbance in the group for the traffic considered during the time T.

We now assume the form (2) for the inconvenience function $I(t)$ with the proviso that c is an individual inconvenience coefficient for each blocked call. It is easily seen that by forming the mean of the values of c in the same way as before, the unchanged form (4) is reached, in which γ now symbolizes the mean of all possible values of c. For the relative disturbance we thus obtain

$$-\frac{\gamma}{1+\lambda} R \int_0^\infty t^{1+\lambda} dF(t) \qquad (13\,a)$$

The function $F(t)$ in this integral is the distribution function of the delay time. With the exception of the special case of random traffic in a full availability group with strictly queued serving of the waiting calls and exponential holding-time distribution, this function has a very complicated form and cannot in general be expressed explicitly. With the exception of cases with certain special values of λ the integral in (13 a) is thus very difficult to handle. Even in this exceptionally simple case a gamma-function is obtained, which can hardly be assumed to be among the paraphernalia of the ordinary telephone engineer. If reasoning about inconvenience functions is to lead to equations which can

be used for practical applications a start must be made using values of λ in (13 a) which make the integral in (13 a) easily calculable. This now means only $\lambda = 0$ and $\lambda = 1$. We exclude here $\lambda = -1$, a trivial case of no importance in this connection. Now, we have shown in the preceding sections that it is just these values of λ which can be considered plausible, $\lambda = 0$ in delay systems with back signalling and 1 in delay systems without back signalling. If we then introduce the mean delay time m, defined by

$$m = -\int_0^\infty t \cdot dF(t) \qquad (14\,a)$$

and the delay time distribution function form factor ε, defined in accordance with eq. 3, p. 3 of »S. f. T.» by

$$\varepsilon = -\frac{1}{m^2}\int_0^\infty t^2 \cdot dF(t) \qquad (14\,b)$$

we find, using form (13) and assuming the inconvenience function (4 a), the relative disturbance in delay systems without back signalling to be

$$\gamma R m^2 \frac{\varepsilon}{2} \qquad (15\,a)$$

and analogously, assuming the inconvenience function (10 a), for the relative disturbance in delay systems with back signalling

$$\omega_0 R m \qquad (15\,b)$$

The possibility was indicated earlier that from the results obtained for $\lambda = 0$ and 1 approximations might be derived for non-integer values of λ. For this purpose we must replace the integral

$$-\int_0^\infty t^{1+\lambda}\, dF(t)$$

by some simple function of λ which will give a good approximation at least over the range $0<\lambda<1$ and which at the boundaries is equal to the exact expressions. These latter are, for $\lambda = 0$ the linear or first moment and for $\lambda = 1$ the second moment. A simple function of the wanted kind is thus

$$\text{first moment} \times \left\{\frac{\text{second moment}}{\text{first moment}}\right\}^\lambda$$

For $\lambda = 0$ this is equal to the first moment, and for $\lambda = 1$ it equals the second moment. Using the mean delay time and the form factor we thus have as an approximation to the relative disturbance (13 a)

$$\frac{\gamma}{1+\lambda} R m^{1+\lambda}\varepsilon^\lambda \qquad (13\,b)$$

A study of how closely this expression fits to (13 a) cannot be carried through here. For the forms of the function $F(t)$ known hitherto, however, the deviation seems to be of little importance, at any rate in the range of λ which is conceivable in practice, say from 0 to 1.5. If, therefore, the simple equations (15 a) and (15 b) cannot be accepted for design calculations it is always possible to go over to (13 b) with some other plausible value of λ inserted. Since, however, the computation work will be more tedious, this should not be done without urgent necessity. It must be noted here that the expressions obtained for m and ε in some cases of traffic investigated theoretically so far are generally very simple, even for numerical calculations. The mean delay time m can always be calculated without closer knowledge of $F(t)$ directly from the traffic conditions in the group under consideration. For the calculation of the form factor ε it is sufficient to know a determining equation for $F(t)$, which need not, however, be explicitly determinable.

It is of interest to examine more closely the expressions to which the equations above for relative disturbance lead in full availability groups in delay systems. The traffic conditions resulting from random traffic in such groups are now relatively thoroughly investigated. For this we refer to the articles in »S. f. T.»

We shall consider a full availability group with n circuits in a delay system. The traffic offered to the group is assumed to be random with an average number of calls per unit time y and an average holding time s. The traffic loading may be written $A = sy$. The congestion in such a group, which is the relative number of calls which must wait, is expressed by the Erlang formula $E_{2,n}(A)$ in eq. 5, p. 40 of »S. f. T.», which, according to the results of article 3 of the same paper, applies exactly with exponential holding time distribution functions and with close approximation in all other cases.

It is assumed, however, that none of the waiting callers falls away before obtaining the wanted connection.

If we compare the two expressions (15 a) and (15 b), for delay systems without back signalling and for delay systems with back signalling respectively, we see that the latter is dependent on the delay time distribution form factor. Now, this form factor is extremely sensitive to the order in which waiting calls are served if congestion gives rise to a plurality of simultaneously waiting calls. The average delay time, on the other hand is not at all affected by this order. This implies that the disturbance in a delay system with back signalling is the same, no matter what order the serving of waiting calls follows. This clearly applies only, however, if it is assumed initially that the inconvenience function has the form (10 a), so that $\lambda = 0$. In a delay system without back signalling, for which we took the inconvenience function (4 a), the disturbance (15 a) will vary depending on the order in which waiting calls are served. We have studied two different cases in this respect earlier, queued service and random service and assumed that all practical cases should be classifiable under one of these two heads. In setting up the expressions for relative disturbance in full availability groups in delay systems we must therefore distinguish the following three different cases:

delay system with back signalling,

delay system without back signalling and with queued service,

delay system without back signalling and with random service.

The average delay time is given in all three kinds of delay system by equation (19) p. 47 in »S. f. T.» as

$$m = \frac{s}{n - A} = \frac{s}{n} \cdot \frac{1}{1 - \alpha} \qquad (16\,a)$$

Here the symbol $\alpha = A/n$ is introduced for the average occupancy per device in the group. So far as the form factor is concerned we have shown in article 3 of »S. f. T.» that the delay time distribution function for strictly queued serving is a pure exponential, so that the form factor in this case is 2. With random service, on the other hand, the form factor for the delay time distribution is, from eq. (23), p. 79 of »S. f. T.»,

$$\varepsilon = \frac{4}{2 - \alpha} \qquad (16\,b)$$

If we introduce these expressions in eq. (15) we obtain finally for the relative disturbance in full availability groups in:

delay systems with back signalling

$$W_1 = \omega_0 \cdot E_{2,n}(A) \frac{s}{n} \cdot \frac{1}{1 - \alpha} \qquad (17\,a)$$

delay systems without back signalling and with queued serving

$$W_2 = \gamma \cdot E_{2,n}(A) \left(\frac{s}{n}\right)^2 \frac{1}{(1 - \alpha)^2} \qquad (17\,b)$$

delay systems without back signalling and with random service

$$W_3 = \gamma \cdot E_{2,n}(A) \left(\frac{s}{n}\right)^2 \frac{1}{(1 - \alpha)^2} \cdot \frac{2}{2 - \alpha} \qquad (17\,c)$$

It is not possible to make any comparison between the disturbances in delay systems with and without back signalling unless the connection between ω_0 and γ is known. It is, on the other hand, possible to compare directly the conditions in delay systems with and without queued service, since the corresponding equations contain the same coefficient γ. We find, thus

$$\frac{W_2}{W_3} = 1 - \frac{\alpha}{2} \qquad (18\,a)$$

Now α is always between 0 and 1, and with properly designed groups it is between 0.3 and 0.8. By changing from random service to queued service the disturbance can thus be reduced by from 15 % to 40 %. With very fully utilized groups having α close to 1 the disturbance can be reduced by almost one half, but a greater reduction cannot be obtained in this way.

It is easy to see what changes must occur in the expressions (17) if the general expression (13 b) is used for the disturbance. It need only be pointed out here that in place of (18 a) we obtain

$$\frac{W_2}{W_3} = \left(1 - \frac{\alpha}{2}\right)^{\lambda} \qquad (18\,b)$$

For example if $\lambda = 0.5$, the reduction of the disturbance on changing to queued serving would only be about one half the value cited above.

At present in the designing process the theoretically calculated loss for a busy signal system is used as a measure of the relative disturbance, independently of whether the groups under consideration are arranged in busy signal or delay systems, and in view of that fact it may be of interest to express the disturbance (17) in terms of this loss. In such a case we can use the approximate relations, given on p. 46 of »S. f. T.», between the proportion of waiting calls $E_{2,n}$ in a delay system and the loss $E_{2,n}$ in a busy signal system. This is tolerably accurate in normally designed groups, where the congestion in a busy signal system does not exceed 5—10 %. If we use this relation we obtain instead of (17), for:

delay systems with back signalling

$$W_1 = \omega_0 \cdot E_{1,n}(A)\frac{s}{n}\cdot\frac{1}{(1-\alpha)^2} \qquad (19\,a)$$

delay systems without back signalling and with queved service

$$W_2 = \gamma \cdot E_{1,n}(A)\left(\frac{s}{n}\right)^2 \frac{1}{(1-\alpha)^3} \qquad (19\,b)$$

delay systems without back signalling and with random service

$$W_3 = \gamma \cdot E_{1,n}(A)\left(\frac{s}{n}\right)^2 \frac{1}{(1-\alpha)^3}\cdot\frac{2}{(2-a)} \qquad (19\,c)$$

The objections which can be raised against the expressions for the disturbance in delay systems without back signalling derived so far are that both the congestion and the delay time conditions are actually influenced by the fact that the waiting subscribers in some cases abandon their calls before they are completed. We should thus start with the equations for congestion and delay time given in article 3 of »S. f. T.», which apply when there is falling away of waiting callers. In such a case both the congestion and the average delay time and consequently also the disturbance are less than in the cases assumed above. A disadvantage is that the equations for the disturbance with fall-away of waiting callers are appreciably more complicated and the numerical calculations are thus more tedious than when no account is taken of fall-away. Furthermore we are still far from having enough experience to determine to what extent such falling away actually occurs in practice. Finally account must be taken of the fact that some of those waiting callers who break off the waiting period return shortly afterwards with fresh calls and thus produce a *congestion reaction* which is obviously difficult to calculate.

The difficulties indicated are not perhaps impossible to master if in reality an urgent need exists to take account of the fall-away of waiting callers. Other difficulties also arise, however, involving the definition of the inconvenience, and these are questions of principle. If a waiting caller breaks off his waiting at the instrument before the call is completed, and after a short period makes a fresh attempt, the inconvenience during the interlude can be assumed to grow in the same way as in a busy signal system, linearly with time. What is more difficult to decide, however, is how, after the renewed call, the inconvenience must be assumed to grow during the continued waiting period at the instrument. Is a quadratic function to be assumed here, and if so, what is its initial slope? How should cases be treated from the point of view of inconvenience when the subscriber delays so long before renewing the call that it is most conveniently regarded as a new call initiation? These questions seem to be intricate. Similar difficulties arise moreover in the treatment of the disturbance in a busy signal system and occasion there a compromise solution which is discussed below. In the delay systems it is fortunate that strong reasons exist why in design work the conditions always used for calculation are those in which there is no fall-away, irrespective of the actual circumstances. That the delay times at all produce sufficient inconvenience to make a number of subscribers break off their calls and perhaps try later is a minus mark for the service quality, so that it seems essentially wrong to accept as a credit factor in design a decrease in inconvenience resulting from the subscribers' reaction to operational disturbances. It is therefore with reason that in designing a delay system the inconvenience produced if all blocked calls are assumed to wait for connection should be used for calculation.

The reasoning given applies mainly to delay systems without back signalling. The delay times in such systems must always be kept relatively small for natural reasons, since otherwise the conditions obtained would be such that the subscribers could rightly assert that there was no

longer any telephone service. It is never necessary to consider cases in which $A > n$, which would make the equations derived above useless for comparison purposes. In delay systems with back signalling, however, the conditions shape themselves differently in important respects. In this case the question discussed above of the fall-away of waiting callers is of less significance, since here it must be relatively rare that a subscriber abandons his call because of excessive delay, an event which should occur only if he leaves the place where the telephone is situated. Moreover, it is possible to provide a unique definition of the inconvenience independently of whether the subscriber answers the back signal or not, since the inconvenience depends only on the time interval between the call and the back signal.

We shall now investigate the busy signal system. Here some difficulties arise because research has no yet solved all the problems arising in connection with congestion. It is known that the Erlang formula applies only on the assumption that any calls which are blocked do not lead to renewed calls within too short a space of time. In practice, however, it must be expected that such calls occur to a large extent, so that in reality the proportion of blocked calls is greater than given by $E_{1,n}(A)$. If it is assumed, on the other hand, that all blocked subscribers repeat their calls perpetually until they are completed, the fraction of disturbed calls becomes almost the same as in a delay system, $E_{2,n}(A)$. Now in practice such extreme cases are obviously rare, so that it may be concluded that this fraction lies between $E_{1,n}(A)$ and $E_{2,n}(A)$.

For the busy signal system we have already used a linear annoyance function $I_1(t)$, but the question is to which retardation time this shall be applied. The time which elapses for a blocked subscriber after the first blocking to the completion of the call depends partly on how long the congestion lasts and partly on how often the subscriber renews his attempt to complete the call. It appears appropriate to calculate the retardation time only by the time the congestion lasts, preventing the subscriber from completing the call. Although this leads to a retardation time which is too low, there is some compensation since we are calculating with a congestion value $E_{2,n}(A)$ which is somewhat too high.

The distribution function for the duration of the congestion is clearly involved also in delay systems, and is written, as in article 5 of »S.f.T.», as $F_s(t)$. This distribution function is defined as the probability that there will still be congestion at a time t from a randomly chosen instant of time at which there was congestion.

The average disturbance per call is expressed by

$$-E_{2,n}(A)\int_0^\infty I_1(t)\,F'_s(t)\,dt$$

which can be written, using (12 a), as

$$-\omega\cdot E_{2,n}(A)\int_0^\infty t\cdot F'_s(t)\,dt$$

The first moment of the distribution $F_s(t)$ is

$$\frac{\alpha}{(1-\alpha^2)}$$

and gives the average time between a randomly chosen moment at which there is congestion and the termination of the state of congestion. For the mean duration of the total congestion condition we have, as before.

$$m=\frac{s}{n}\cdot\frac{1}{1-\alpha}$$

The relative disturbance for full availability groups in a busy signal system will be

$$W_4=\omega\cdot E_{2,n}(A)\frac{s}{n}\cdot\frac{1}{(1-\alpha)^2} \qquad (17\,\mathrm{d})$$

In this equation we can replace $E_{2,n}(A)$ by $E_{2,n}(A)$ under the conditions earlier discussed in going from equations (17) to (19).

$$W_4=\omega\cdot E_{1,n}(A)\frac{s}{n}\cdot\frac{1}{(1-\alpha)^3} \qquad (19\,\mathrm{d})$$

To get an approximate idea of the relationship between the various inconvenience coefficients we consider Fig. 2. After a time t, the inconvenience will be greater in a delay system without back signalling than in a busy signal system. When a subscriber confronted by congestion in a delay system hangs up, it implies that he will not accept the system, but substitutes his own busy signal system instead. It can be supposed that this change-over takes place, on

the average, at the time at which both systems are equal in satisfaction, i.e. at time t_1. This means that

$$\frac{\gamma}{2} t_1^2 = \omega t_1$$

If we use here the results of the fall-away investigation referred to earlier, with an average waiting time of 29 seconds, we must have,

$$\omega = 14{,}5\ \gamma \text{ sec.}$$

The results of the investigation can, however, probably not be applied to those congestion cases in which dialling tone is not obtained.

It may, however, be thought that subscribers stop waiting when the inconvenience in a delay system is growing more rapidly than with a busy signal system. This would mean that the derivatives should be equal at the average fall-away time, which gives

$$\gamma t = \omega$$

and in the case discussed this will give

$$\omega = 29\ \gamma \text{ sec.}$$

The combined system.

A third way of obtaining a relation between γ and ω is by a study of the *combined system.* This is based on the reasoning that it should be advantageous to have a delay system for short delay times and to have a busy signal system for long delay times. In a combined system the arrangements are such that a subscriber is allowed to wait at most for a time t_0 and if the connection has not then been completed the busy tone is transmitted and the subscriber must renew his call. This arrangement is assumed to combine the advantages of the delay system and the busy signal system.

If, in calculating the disturbance in this combined system, we use the complicated distribution functions derived in article 5 of »S. f. T.», the problem becomes very difficult of solution. Having in mind the uncertainty inherent in the determination of our constants it is justifiable to simplify very considerably the calculations by introducing the assumption that the waiting calls are handled in order, this, clearly, so far as they are not cancelled after time t_0. The assumption, which in our case involves only a very slight approximation, makes it possible to use the distribution function of simple exponential form, which is given in »S. f. T.», eq. (23) on page 49.

The average number of subscribers who must wait for a time $t<t_0$ will be

$$-y \cdot E_{2,n}(A) \cdot F'(t)\, dt$$

The average disturbance per unit time for all blocked calls which have a delay time less than t_0 is

$$-y \cdot E_{2,n}(A) \int_0^{t_0} I(t) \cdot F'(t)\, dt$$

With $I(t)$ taken from (4 a), this gives, after integration

$$y\, \gamma\, m^2\, E_{2,n}(A) \left\{ 1 - \left(1 + \frac{t_0}{m} + \frac{t_0^2}{2m^2} \right) e^{-\frac{t_0}{m}} \right\} \quad (20)$$

where

$$m = \frac{s}{n} \cdot \frac{1}{1-\alpha}$$

as obtained earlier.

The disturbance to a subscriber receives busy tone at time t_0 is calculated in the same way as as in a busy signal system.

The average number of calls per unit time which are cancelled after t_0 is

$$y \cdot E_{2,n}(A)\, e^{-\frac{t_0}{m}}$$

For each such call the delay for a time t_0 contributes an inconvenience $I(t_0)$. To this must be added the inconvenience occasioned by the renewal of the call which becomes necessary. This has been assumed to equal that in a busy signal system and the average value per cancelled call will be, from (17 d)

$$\frac{\omega m}{1-\alpha}$$

The total disturbance for a call which is not completed in a time t_0 will be

$$y \cdot E_{2,n}(A)\, e^{-\frac{t_0}{m}} \left\{ \frac{\gamma}{2} t_0^2 + \frac{\omega m}{1-\alpha} \right\} \quad (21)$$

By adding the two disturbance terms (20) and (21) and division by y we obtain the relative disturbance in the combined system:

$$W_5 = m \cdot E_{2,n}(A) \left\{ \gamma m + \right.$$
$$\left. + e^{-\frac{t_0}{m}} \left(\frac{\omega}{1-\alpha} - \gamma m - \gamma t_0 \right) \right\} \quad (17\,e)$$

Regarded as a function of t_0, W_5 has at

$$t_0 = \frac{\omega}{\gamma\,(1-\alpha)} \qquad (22)$$

a minimum value

$$W_5 \text{ min.} = \gamma \cdot E_{2,n}\,(A) \left(\frac{s}{n}\right)^2 \frac{1 - e^{\frac{-\omega n}{\gamma s}}}{(1-\alpha)^2}$$

A comparison with the disturbance formula for a delay system without back signalling and with queued service according to (17 b), gives

$$W_5 \text{ min} = W_2 \left(1 - e^{-\frac{\omega n}{\gamma s}}\right)$$

The gain with the combined system will thus be bigger as $\frac{\omega n}{\gamma s}$ is made smaller, so that it is with small groups that the system shows itself most advantageous.

The voluntary fall-away of blocked calls with long delay times in normal delay systems makes it probable that the traffic conditions in such systems approach those which we have assumed to rule in a combined system for it is reasonable to assume that those subscribers, who are subjected to congestion, react automatically to satisfy the conditions for minimum disturbance. The average fall-away time will then be t_0, and this gives us a possibility of determining the relationship between ω and γ.

Since α also appears in eq. (22), the average occupancy of the circuits of the group must be included as a parameter in the relationship between ω and γ. Using the same traffic measurements as before, where we had $t_0 = 29$ sec., we obatin in a normal group with $\alpha = 0.6$ the equation $\omega = 11.6\ \gamma$ sec.

Table 1. $s = 120$ seconds.

n	A	E_1	E_2	α	m	W_1	W_2	W_3	W_4
5	1,0	0,0031	0,0038	0,2000	30,00	0,57	3,42	3,80	2,14
10	3,5	0,0023	0,0035	0,3500	18,46	0,32	1,18	1,43	1,48
20	10,0	0,0019	0,0037	0,5000	12,00	0,22	0,53	0,71	1,32
36	22,0	0,0016	0,0041	0,6111	8,57	0,18	0,31	0,45	1,39
5	1,5	0,0142	0,0201	0,3000	34,29	3,45	23,7	27,8	14,8
10	4,5	0,0105	0,0189	0,4500	21,82	2,06	9,0	11,6	11,2
20	12,0	0,0098	0,0241	0,6000	15,00	1,81	5,4	7,8	13,6
36	26,0	0.0122	0,0427	0,7222	12,00	2,56	6,1	9,6	27,6
5	2,0	0,0367	0,0597	0,4000	40,00	11,9	96	119	60
10	6,0	0,0431	0,1013	0,6000	30,00	15,2	91	130	114
20	15,0	0,0456	0,1607	0,7500	24,00	19,3	93	148	231
36	30,0	0,0429	0,2119	0,8333	20,00	21,2	85	145	381

Table 2. $s = 12$ seconds.

36	22,0	0,0016	0,0041	0,6111	0,857	0,02	0,003	0,005	0,14
36	26,0	0,0122	0,0427	0,7222	1,200	0,26	0,061	0,096	2,76
36	30,0	0,0429	0,2119	0,8333	2,000	2,12	0,85	1,45	38,1

Tables 1 and 2 give some disturbance values calculated for various groups and systems. To give some possibility of comparison we have assumed $\omega = 15\ \gamma$ sec. and $\omega = 3\omega_0$, values which appear reasonable. All disturbance values are given in γ.

The first table is calculated with an average holding time $s = 120$ sec., corresponding to the duration of a normal call. The second table applies to an average holding time of 12 sec., corresponding to a normal holding time for registers.

The first section of the first table corresponds to about 2 ‰ loss according to the Erlang loss formula. The second section corresponds to about 1 % loss and the third section to about 5 %.