THEORY OF COMMUNICATION*

By D. GABOR, Dr. Ing., Associate Member.†

(*The paper was first received 25th November, 1944, and in revised form 24th September, 1945.*)

PREFACE

The purpose of these three studies is an inquiry into the essence of the "information" conveyed by channels of communication, and the application of the results of this inquiry to the practical problem of optimum utilization of frequency bands.

In Part 1, a new method of analysing signals is presented in which time and frequency play symmetrical parts, and which contains "time analysis" and "frequency analysis" as special cases. It is shown that the information conveyed by a frequency band in a given time-interval can be analysed in various ways into the same number of elementary "quanta of information," each quantum conveying one numerical datum.

In Part 2, this method is applied to the analysis of hearing sensations. It is shown on the basis of existing experimental material that in the band between 60 and 1 000 c/s the human ear can discriminate very nearly every second datum of information, and that this efficiency of nearly 50% is independent of the duration of the signals in a remarkably wide interval. This fact, which cannot be explained by any mechanism in the inner ear, suggests a new phenomenon in nerve conduction. At frequencies above 1 000 c/s the efficiency of discrimination falls off sharply, proving that sound reproductions which are far from faithful may be perceived by the ear as perfect, and that "condensed" methods of transmission and reproduction with improved waveband economy are possible in principle.

In Part 3, suggestions are discussed for compressed transmission and reproduction of speech or music, and the first experimental results obtained with one of these methods are described.

Part 1. THE ANALYSIS OF INFORMATION

SUMMARY

Hitherto communication theory was based on two alternative methods of signal analysis. One is the description of the signal as a function of time; the other is Fourier analysis. Both are idealizations, as the first method operates with sharply defined instants of time, the second with infinite wave-trains of rigorously defined frequencies. But our everyday experiences—especially our auditory sensations—insist on a description in terms of *both* time and frequency. In the present paper this point of view is developed in quantitative language. Signals are represented in two dimensions, with time and frequency as co-ordinates. Such two-dimensional representations can be called "information diagrams," as areas in them are proportional to the number of independent data which they can convey. This is a consequence of the fact that the frequency of a signal which is not of infinite duration can be defined only with a certain inaccuracy, which is inversely proportional to the duration, and vice versa. This "uncertainty relation" suggests a new method of description, intermediate between the two extremes of time analysis and spectral analysis. There are certain "elementary signals" which occupy the smallest possible area in the information diagram. They are harmonic oscillations modulated by a "probability pulse." Each elementary signal can be considered as conveying exactly one datum, or one "quantum of information." Any signal can be expanded in terms of these by a process which includes time analysis and Fourier analysis as extreme cases.

These new methods of analysis, which involve some of the mathematical apparatus of quantum theory, are illustrated by application to some problems of transmission theory, such as direct generation of single sidebands, signals transmitted in minimum time through limited frequency channels, frequency modulation and time-division multiplex telephony.

* Radio Section paper.

† British Thomson-Houston Co., Ltd., Research Laboratory.

(1) INTRODUCTION

The purpose of this study is to present a method, with some new features, for the analysis of information and its transmission by speech, telegraphy, telephony, radio or television. While this first part deals mainly with the fundamentals, it will be followed by applications to practical problems, in particular to the problem of the best utilization of frequency channels.

The principle that the transmission of a certain amount of information per unit time requires a certain minimum waveband width dawned gradually upon communication engineers during the third decade of this century. Similarly, as the principle of conservation of energy emerged from the slowly hardening conviction of the impossibility of a *perpetuum mobile*, this fundamental principle of communication engineering arose from the refutation of ingenious attempts to break the as yet unformulated law. When in 1922 John Carson[1.1] disproved the claim that frequency modulation could economize some of the bandwidth required by amplitude-modulation methods, he added that all such schemes "are believed to involve a fundamental fallacy." This conviction was soon cast into a more solid shape when, in 1924, Nyquist[1.2] and Küpfmüller[1.3] independently discovered an important special form of the principle, by proving that the number of telegraph signals which can be transmitted over any line is directly proportional to its waveband width. In 1928 Hartley[1.4] generalized this and other results, partly by inductive reasoning, and concluded that "the total amount of information which may be transmitted . . . is proportional to the product of frequency range which is transmitted and the time which is available for the transmission."

Even before it was announced in its general form, an applica-

tion was made of the new principle, which remains to this day probably its most important practical achievement. In 1927, Gray, Horton and Mathes[1.5] gave the first full theoretical discussion of the influence of waveband restriction on the quality of television pictures, and were able to fix the minimum waveband requirements in advance, long before the first high-definition system was realized. In fact, in this as in later discussions of the problem, the special Nyquist–Küpfmüller result appears to have been used, rather than Hartley's general but somewhat vague formulation.

The general principle was immediately accepted and recognized as a fundamental law of communication theory, as may be seen from its discussion by Lüschen[1.6] in 1932 before this Institution. Yet it appears that hitherto the mathematical basis of the principle has not been clearly recognized. Nor have certain practical conclusions been drawn, which are suggested by a more rigorous formulation.

(2) TRANSMISSION OF DATA

Let us imagine that the message to be transmitted is given in the form of a time function $s(t)$, where s stands for "signal." Unless specially stated, s will be assumed to be of the nature of a voltage, current, field strength, air pressure, or any other "linear" quantity, so that power and energy are proportional to its square. We assume that the function $s(t)$ is given in some time interval $t_2 - t_1 = \tau$, as illustrated in Fig. 1.1. Evidently

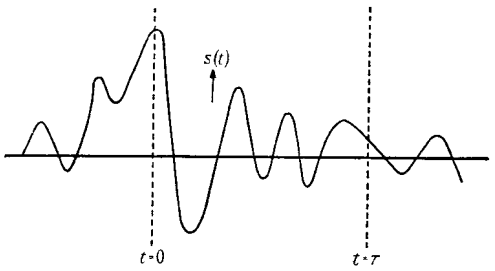

Fig. 1.1.—Signal as a function of time.

this message contains an infinity of data. We can divide τ into, say, N sub-intervals, and define, for instance, the average ordinate in each sub-interval as a "datum." If there is no limit to the sub-division, there is no limit to the number of data which could be transmitted in an *absolutely faithful* reproduction.

As this is impossible, let us see whether it is possible to transmit faithfully at least a finite number N of data. Evidently there is an infinite number of possibilities for specifying the curve $s(t)$ in the interval τ *approximately* by N data. Without knowing the specific purpose of the transmission it is impossible to decide which is the most economical system of selection and specification. Yet, certain methods will recommend themselves by reason of their analytical simplicity. One of these, division into equal sub-intervals, has been already mentioned. Another method is to replace the curve $s(t)$ in the interval τ by a polynomial of order N, to fit it as closely as possible to $s(t)$ by the method of least squares, and to take the coefficients of the polynomial as data. It is known that this method is equivalent to specifying the polynomial in such a way that its first N "moments" M_i shall be equal to those of $s(t)$:—

$$M_0 = \int_0^\tau sdt \quad M_1 = \int_0^\tau tsdt \quad M_2 = \int_0^\tau t^2 sdt \ldots M_{N-1} = \int_0^\tau t^{N-1} sdt$$

Instead of the coefficients of the polynomial, we can also consider these moments as the specified data.

A method closely related to this is the following. Expand $s(t)$, instead of in powers of time, in terms of a set of N functions $\phi_k(t)$, orthogonal in the interval $0 < t < \tau$, and consider as data the N coefficients of expansion. It is known that this is equivalent to fitting the expansion to $s(t)$ by the method of least squares.* How close the fit will be, and how well it will suit the practical purpose, depends on the set of functions selected.

One class of orthogonal functions, the simple harmonic functions sine and cosine, have always played a preferred part in communication theory. It is shown in Appendix 9.1 that there are good reasons for this preference other than their elementary character. Let us now develop the curve $s(t)$ in the interval τ into a Fourier series. This gives an infinite sequence of spectral lines, as shown in Fig. 1.2, starting with zero frequency, all equally spaced by a frequency $1/\tau$. Two data are associated with each frequency, the coefficients of the sine and cosine terms in the expansion. In a frequency range $(f_2 - f_1)$ there are therefore $(f_2 - f_1)\tau$ lines, representing $2(f_2 - f_1)\tau$ data, that is exactly *two data per unit time and unit frequency range*.

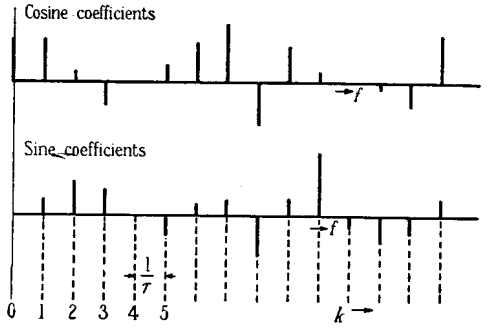

Fig. 1.2.—Fourier spectrum of signal in an interval τ.

This, in fact, proves the fundamental principle of communication. *In whatever ways we select N data to specify the signal in the interval τ, we cannot transmit more than a number $2(f_2 - f_1)\tau$ of these data, or of their independent combinations by means of the $2(f_2 - f_1)\tau$ independent Fourier coefficients.*

In spite of the extreme simplicity of this proof, it leaves a feeling of dissatisfaction. Though the proof shows clearly that the principle in question is based on a simple mathematical identity, it does not reveal this identity in a tangible form. Besides it leaves some questions unanswered: What are the effects of a physical filter? How far are we allowed to sub-divide the waveband or the time interval? What modifications would arise by departing from the rigid prescription of absolute independence of the data and allowing a limited amount of mutual interference? It therefore appears worth while to approach the problem afresh in another way, which will take considerably more space, but which, in addition to physical insight, gives an answer to the questions which have been left open.

(2.1) Time and Frequency

The greatest part of the theory of communication has been built up on the basis of Fourier's reciprocal integral relations†

$$s(t) = \int_{-\infty}^{\infty} S(f) e^{2\pi jft} df \qquad S(f) = \int_{-\infty}^{\infty} s(t) e^{-2\pi jft} dt \quad . \quad . \quad (1.1)$$

* Cf. e.g. CHURCHILL, RUEL V.: "Fourier Series and Boundary Value Problems" (McGraw Hill, 1941), p. 40. This book contains an introduction to the theory of orthogonal functions.

† The notations used will follow in the main those of Campbell and Foster.[1.7]

where $s(t)$ and $S(f)$ are a pair of Fourier transforms. We will refer to $S(f)$ also as the "spectrum" of $s(t)$.

Though mathematically this theorem is beyond reproach, even experts could not at times conceal an uneasy feeling when it came to the physical interpretation of results obtained by the Fourier method. After having for the first time obtained the spectrum of a frequency-modulated sine wave, Carson wrote:[1.1] "The foregong solutions, though unquestionably mathematically correct, are somewhat difficult to reconcile with our physical intuitions, and our physical concepts of such 'variable-frequency' mechanisms as, for example, the siren."

The reason is that the Fourier-integral method considers phenomena in an infinite interval, *sub specie aeternitatis*, and this is very far from our everyday point of view. Fourier's theorem makes of description in time and description by the spectrum, two mutually exclusive methods. If the term "frequency" is used in the strict mathematical sense which applies only to infinite wave-trains, a "changing frequency" becomes a contradiction in terms, as it is a statement involving *both* time and frequency.*

The terminology of physics has never completely adapted itself to this rigorous mathematical definition of "frequency." In optics, in radio engineering and in acoustics the word has retained much of its everyday meaning, which is in better agreement with what Carson called "our physical intuitions." For instance, speech and music have for us a definite "time pattern," as well as a frequency pattern. It is possible to leave the time pattern unchanged, and double what we generally call "frequencies" by playing a musical piece on the piano an octave higher, or conversely it can be played in the same key, but in different time. Evidently both views have their limitations, and they are complementary rather than mutually exclusive. But it appears that hitherto the fixing of the limit was largely left to common sense. It is one of the main objects of this paper to show that there are also adequate mathematical methods available for this purpose.

Let us now tentatively adopt the view that both time and frequency are legitimate references for describing a signal, and illustrate this, as in Fig. 1.3, by taking them as orthogonal co-

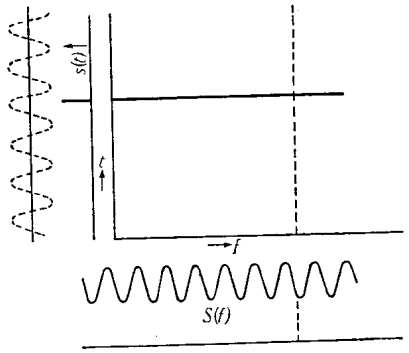

Fig. 1.3.—Unit impulse (delta function) and infinite sine wave in time/frequency diagram.

ordinates. In this diagram a harmonic oscillation is represented by a vertical line. Its frequency is exactly defined, while its epoch is entirely undefined. A sudden surge or "delta function"† (also called "unit impulse function"), on the other hand, has a sharply defined epoch, but its energy is uniformly distributed over the whole frequency spectrum. This signal is therefore represented by a horizontal line. But how are we to represent other signals, for instance a sine wave of finite duration?

In order to give this question a precise meaning we must consider the physical effects which can be produced by the signal. The physical meaning of the $s(t)$ curve, shown at the left of Fig. 1.4, is that this is the response of an ideal oscillograph

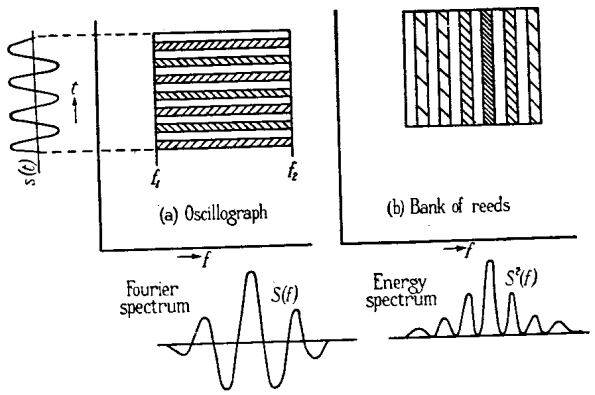

Fig. 1.4.—Time/frequency diagram of the response of physical instruments to a finite sine wave.

which has a uniform response over the whole infinite frequency range. The interpretation of the Fourier spectrum, shown at the bottom of the same figure, is somewhat less simple. It could be obtained by an infinite number of heterodyne receivers, each of which is tuned to a sharp frequency, and connected with an indicating instrument of infinite time-constant. To simplify matters we take instead a bank of reeds, or other resonators, each tuned to a narrow waveband, with equally spaced resonant frequencies. It is known that such an instrument gives only an analysis of the energy spectrum, as it cannot distinguish phases, but this will be sufficient for the purpose of discussion. Let us compare this instrument with a real oscillograph, which responds only to a certain range of frequencies $(f_2 - f_1)$. For simplicity it has been assumed in Fig. 1.4 that the bank of reeds extends over the same range, and that the time-constant of the reeds is about equal to the duration of the signal.

We know that any instrument, or combination of instruments, cannot obtain more than at most $2(f_2 - f_1)\tau$ independent data from the area $(f_2 - f_1)\tau$ in the diagram. But instead of rigorously independent data, which can be obtained in general only by calculation from the instrument readings, it will be more convenient for the moment to consider "practically" independent data, which can be obtained by direct readings. For any resonator, oscillograph or reed, a damping time can be defined, after which oscillations have decayed by, say, 10 db. Similarly one can define a tuning width as, say, the number of cycles off resonance at which the response falls off by 10 db. It is well known that in all types of resonators there is a relation between these two of the form:

$$\text{Decay time} \times \text{Tuning width} = \text{Number of the order one.}$$

This means that for every type of resonator a *characteristic rectangle* of about unit area can be defined in the time/frequency diagram, which corresponds to one "practically" independent reading of the instrument. In order to obtain their number, we must divide up the (time × frequency) area into such rectangles. This is illustrated in Figs. 1.4(*a*) and 1.4(*b*). In the case of the oscillograph the rectangles are broad horizontally and narrow vertically; for the tuned reeds the reverse. The amplitude of the readings is indicated by shading of different density. Negative amplitudes are indicated by shading of

* Carson proposed the concept of a "generalized frequency" in 1922, and in 1937 elaborated it further with T. C. Fry under the name of "instantaneous frequency" (Ref. No. 1.8). This is a useful notion for slowly-varying frequencies, but not sufficient to cover all cases in which physical feeling and the Fourier integral theorem are at variance.

† Campbell and Foster call this an $\mathfrak{S}_0$ function, but the name "delta function" as used by Dirac has now wider currency.

opposite inclination. We will return later to the question of a suitable convention for measuring these amplitudes.*

Without going into details, it is now evident that physical instruments analyse the time-frequency diagram into rectangles which have shapes dependent on the nature of the instrument and areas of the order unity, but not less than one-half. The number of these rectangles in any region is the number of independent data which the instrument can obtain from the signal, i.e. proportional to the amount of information. This justifies calling the diagram from now on the "*diagram of information.*"

We may now ask what it is that prevents any instrument from analysing the information area with an accuracy of less than a half unit. The ultimate reason for this is evident. We have made of a function of one variable—time *or* frequency—a function of two variables—time *and* frequency. This might be considered a somewhat artificial process, but it must be remembered that it corresponds very closely to our subjective interpretation of aural sensations. Indeed, Fig. 1.4(*b*) could be considered as a rough plan of analysis by the ear; rather rough, as the ear is too complicated an instrument to be replaced by a bank of tuned reeds, yet much closer than either the oscillogram or the Fourier spectrum. But as a result of this doubling of variables we have the strange feature that, although we can carry out the analysis with any degree of accuracy in the time direction or in the frequency direction, we cannot carry it out simultaneously in both beyond a certain limit. This strange character is probably the reason why the familiar subjective pattern of our aural sensations and their mathematical interpretation have hitherto differed so widely. In fact the mathematical apparatus adequate for treating this diagram in a quantitative way has become available only fairly recently to physicists, thanks to the development of quantum theory.

The linkage between the uncertainties in the definitions of "time" and "frequency" has never passed entirely unnoticed by physicists. It is the key to the problem of the "coherence length" of wave-trains, which was thoroughly discussed by Sommerfeld in 1914.† But these problems came into the focus of physical interest only with the discovery of wave mechanics, and especially by the formulation of Heisenberg's principle of indeterminacy in 1927. This discovery led to a great simplification in the mathematical apparatus of quantum theory, which was recast in a form of which use will be made in the present paper.

The essence of this method—due to a considerable part to W. Pauli‡—is a re-definition of all observable physical quantities in such a form that the physical uncertainty relations which obtain between them appear as direct consequences of a mathematical identity

$$\Delta t \Delta f \simeq 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1.2)$$

Δt and Δf are here the uncertainties inherent in the definitions of the epoch t and the frequency f of an oscillation. The identity (1.2) states that t and f cannot be simultaneously defined in an exact way, but only with a latitude of the order one in the product of uncertainties.

Though this interpretation of Heisenberg's principle is now widely known, especially thanks to popular expositions of quantum theory,* it appears that the identity (1.2) itself has received less attention than it deserves. Following a suggestion by the theoretical physicist A. Landé, in 1931 G. W. Stewart brought the relation to the notice of acousticians, in a short note†—to which we shall return in Part 2—but apparently without much response. In communication theory the intimate connection of the identity (1.2) with the fundamental principle of transmission appears to have passed unnoticed.

Perhaps it is not unnecessary to point out that it is not intended to explain the transmission of information by means of quantum theory. This could hardly be called an explanation. The foregoing references are merely an acknowledgment to the theory which has supplied us with an important part of the mathematical methods.

(3) THE COMPLEX SIGNAL

In order to apply the simple and elegant formalism of quantum mechanics, it will be convenient first to express the signal amplitude $s(t)$ in a somewhat different form.

It has long been recognized that operations with the complex exponential $e^{j\omega t}$—often called cis ωt—have distinct advantages over operations with sine or cosine functions. There are two ways of introducing the complex exponential. One is to write

$$\cos \omega t = \tfrac{1}{2}(e^{j\omega t} + e^{-j\omega t}) \quad \sin \omega t = \frac{1}{2j}(e^{j\omega t} - e^{-j\omega t}) \quad . \quad (1.3)$$

This means that the harmonic functions are replaced by the resultant of two complex vectors, rotating in opposite directions. The other way is to put

$$\cos \omega t = \mathscr{R}(e^{j\omega t}) \quad \sin \omega t = -\mathscr{R}(je^{j\omega t}) \quad . \quad . \quad (1.4)$$

In this method the harmonic functions are replaced by the real part of a single rotating vector. Both methods have great advantages against operation with real harmonic functions. Their relative merits depend on the problem to which they are applied. In modulation problems, for instance, the advantage is with the first method. On the other hand, the formalism of quantum mechanics favours the second method, which we are now going to follow. This means that we replace a real signal of the form

$$s(t) = a \cos \omega t + b \sin \omega t \quad . \quad . \quad . \quad . \quad (1.5)$$

by a complex time function

$$\psi(t) = s(t) + j\sigma(t) = (a - jb)e^{j\omega t} \quad . \quad . \quad (1.6)$$

which is formed by adding to the real signal $s(t)$ an imaginary signal $j\sigma(t)$. The function $\sigma(t)$ is formed from $s(t)$ by replacing $\cos \omega t$ by $\sin \omega t$ and $\sin \omega t$ by $-\cos \omega t$. The function $\sigma(t)$ has a simple significance. It represents the signal *in quadrature* to $s(t)$ which, added to it, transforms the oscillating into a rotating vector. If, for instance, $s(t)$ is applied to two opposite poles of a four-pole armature, $\sigma(t)$ has to be applied to the other pair in order to produce a rotating field.

If $s(t)$ is not a simple harmonic function, the process by which $\psi(t)$ has been obtained can be readily generalized. We have only to express $s(t)$ in the form of a real Fourier integral, replace every cosine in it by $e^{j\omega t}$, and every sine by $-je^{j\omega t}$. This process becomes very simple if, instead of sine and cosine Fourier integrals, the complex (cisoidal) Fourier integrals are

* *Note added 7th February*, 1946. An instrument called the "Sound Spectrograph" has been developed by the Bell Telephone Laboratories for the recording of sound patterns in two-dimensional form. The first publications have just appeared: Potter, R. K.: "Visible Patterns of Sound," *Science*, 9th November, 1945, and "Visible Speech," *Bell Laboratories Record*, January 1946.

† Sommerfeld, A.: *Annalen der Physik*, 1914, **44**, p. 177.

Another field of classical physics in which an uncertainty relation is of great importance is Brownian motion. Cf. Fürth, R.: "On Some Relations between Classical Statistics and Quantum Mechanics," *Zeitschrift für Physik*, 1933, **81**, p. 143, and Bouligand, G.: "Relations d'Incertitude en Géometrie et en Physique" (Hermann et Cie, Paris, 1934).

‡ Pauli, W.: "Handbuch der Physik," vol. 24/1, 2nd ed. (Berlin, 1933). A very lucid exposition of quantum mechanics on these lines is given by Tolman, R. C.: "The Principles of Statistical Mechanics" (Oxford, 1938), pp. 189–276. In Dirac's system Pauli's postulates appear as results, derived from another set of postulates. Cf. Dirac, P. A. M.: "Quantum Mechanics," 2nd ed. (Oxford, 1938), p. 103.

* Schrödinger, E.: "Science and the Human Temperament" (Allen and Unwin, 1935), pp. 126–129. Lindemann, F. A.: "The Physical Significance of Quantum Theory" (Oxford, 1932), pp. 126–127. Darwin, C. G.: "The New Conceptions of Matter" (G. Bell and Sons, 1931), pp. 78–102.

† Stewart, G. W.: Ref. No. 1.9.

A. Landé has made use of acoustical examples to illustrate the uncertainty relation in his "Vorlesungen über Wellenmechanik" Akademische Verlagsges (Leipzig, 1930), pp. 17–20.

used according to equation (1.1). In this case the passage from $s(t)$ to $\psi(t)$ is equivalent to the instruction: *Suppress the amplitudes belonging to negative frequencies, and multiply the amplitudes of positive frequencies by two.* This can be readily understood by comparing equations (1.3) and (1.4).

Though the Fourier transform of $\psi(t)$ is thus immediately obtained from the Fourier transform of $s(t)$, to obtain $\psi(t)$ itself requires an integration. It can be easily verified that the signal $\sigma(t)$ associated with $s(t)$ is given by the integral

$$\sigma(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} s(\tau)\frac{d\tau}{\tau - t} \quad . \quad . \quad . \quad . \quad (1.7)$$

This is an improper integral, and is to be understood as an abbreviation of the following limit

$$\int_{-\infty}^{\infty} = \lim_{\varepsilon=0}\left[\int_{-\infty}^{t-\varepsilon} + \int_{t+\varepsilon}^{\infty}\right]$$

which is called "Cauchy's principal value" of an improper integral.* To verify equation (1.7) it is sufficient to show that it converts $\cos \omega t$ into $\sin \omega t$ and $\sin \omega t$ into $-\cos \omega t$. Conversely $s(t)$ can be expressed by $\sigma(t)$ as follows:—

$$s(t) = -\frac{1}{\pi}\int_{-\infty}^{\infty} \sigma(\tau)\frac{d\tau}{\tau - t} \quad . \quad . \quad . \quad . \quad (1.8)$$

Associated functions $s(t)$ and $\sigma(t)$ which satisfy the reciprocal relations (1.7) and (1.8) are known as a pair of "*Hilbert transforms.*"†

Pairs of signals in quadrature with one another can be generated by taking an analytical function $f(z)$ of the complex variable $z = x + jy$, which can be expressed in the form $f(z) = u(x,y) + jv(x,y)$. Provided that there are no poles at one side of the x-axis (and if certain other singularities are excluded), $u(x,0)$ and $v(x,0)$ will be in quadrature. The function e^{jz} is an example which gives $u(x,0) = \cos x$ and $v(x,0) = \sin x$. It follows that, as the real axis is in no way distinguished in the theory of analytical functions of a complex variable, we can draw any straight line in the complex plane which leaves all the poles at one side, and the values of the two conjugate functions along this line will give a pair of functions in quadrature.

An example of two functions in quadrature is shown in Fig. 1.5. In spite of their very different forms they contain the same spectral components. If these functions were to represent amplitudes of sound waves, the ear could not distinguish one from the other.‡

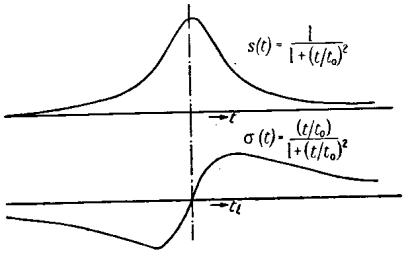

Fig. 1.5.—Example of signals in quadrature.

A mechanical device for generating the associated signal $\sigma(t)$ to a given signal $s(t)$ is described in Appendix 9.2, which contains also a discussion of the problem of single-sideband generation.

* WHITTAKER, E. T., and WATSON, G. N.: "Modern Analysis," 4th ed. (Cambridge), p. 75.

† Cf. TITCHMARSH, E. C.: "Introduction to the Theory of Fourier Integrals" (Oxford, 1937).

‡ Provided that Ohm's law of hearing holds with sufficient accuracy. Such associated signals could be used for testing the limits of validity of Ohm's law.

(4) EXACT FORMULATION OF THE UNCERTAINTY RELATION

By means of the complex signal $\psi(t)$ it is now easy to frame the uncertainty relation in a quantitative manner, using the formalism of quantum mechanics. In order to emphasize the analogy, the same symbol ψ has been chosen for the complex signal as is used in that theory for the "wave" or "probability" amplitudes.

$\psi(t)$ is the time description of the signal. We can associate with this its frequency description by means of its Fourier transform $\phi(f)$, which will also be called the "spectrum" of $\psi(t)$. The two descriptions are connected by the reciprocal Fourier relations

$$\psi(t) = \int_{-\infty}^{\infty} \phi(f)e^{2\pi jft}df \quad . \quad . \quad . \quad . \quad (1.9)$$

$$\phi(f) = \int_{-\infty}^{\infty} \psi(t)e^{-2\pi jft}dt \quad . \quad . \quad . \quad . \quad (1.10)$$

In order to emphasize the symmetry, the first integral has been also written with limits $-\infty$ and ∞, although we have specified $\psi(t)$ in such a way that $\phi(f) = 0$ for negative frequencies; hence we could have taken zero as the lower limit. As in the following all integrals will be taken in the limits $-\infty$ to ∞, the limits will not be indicated in the formulae.

In Section 1 several methods have been discussed for specifying a signal by an infinite set of denumerable (countable) data. One of these was specification by moments, M_0, M_1 . . . This method, with some modifications, will be the best suited for quantitative discussion. The first modification is that it will be more convenient to introduce instead of $s(t)$ the following "weight function":—

$$\psi^*(t)\psi(t) = [s(t)]^2 + [\sigma(t)]^2 \quad . \quad . \quad . \quad (1.11)$$

The asterisk denotes the conjugate complex value. The new weight function is therefore the square of the absolute value of ψ. This can be considered as the "power" of the signal, and will be referred to by this name in what follows. A second convenient modification is that, instead of with the moments themselves, we shall operate with their values divided by M_0, i.e. with the following quotients:—

$$\bar{t} = \frac{\int\psi^* t\psi dt}{\int\psi^*\psi dt} \quad \bar{t}^2 = \frac{\int\psi^* t^2\psi dt}{\int\psi^*\psi dt} \quad . \quad . \quad \bar{t}^n = \frac{\int\psi^* t^n\psi dt}{\int\psi^*\psi dt} \quad . \quad . \quad (1.12)$$

These are the mean values of the "epoch" t of the signal of orders 1, 2 . . . n . . . The factor t^n has been placed between the two amplitude factors to emphasize the symmetry of the formulas with later ones. By a theorem of Stieltjes, if all mean values are known, the weight function $\psi^*\psi = |\psi|^2$ is also determined, apart from a constant factor. The signal ψ itself is determined only as regards absolute value; its phase remains arbitrary. This makes the method particularly suitable, for instance, for acoustical problems. In others, where the phase is observable, it will not be difficult to supplement the specification, as will be shown later.

Similarly we define mean frequencies f^n of the signal as follows:—

$$\bar{f} = \frac{\int\phi^* f\phi df}{\int\phi^*\phi df} \quad \bar{f}^2 = \frac{\int\phi^* f^2\phi df}{\int\phi^*\phi df} \quad . \quad . \quad \bar{f}^n = \frac{\int\phi^* f^n\phi df}{\int\phi^*\phi df} \quad . \quad . \quad (1.13)$$

It now becomes evident why we had to introduce a complex signal in the previous Section. If we had operated with the real signal $s(t)$ instead, the weight function would have been even, and the mean frequency f always zero. This is one of the

points on which physical feeling and the usual Fourier methods are not in perfect agreement. But we could eliminate the negative frequencies, only at the price of introducing a complex signal.

As by equations (1.9) and (1.10), ψ and ϕ mutually determine one another, it must be possible to express the mean frequencies by ψ, and, conversely, the mean epochs by ϕ. This can be done indeed very simply by means of the following elegant reciprocal relations:—

$$\int \psi^* \psi dt = \int \phi^* \phi df \quad . \quad . \quad . \quad . \quad . \quad (1.14)$$

$$\int \phi^* f^n \phi df = \left(\frac{1}{2\pi j}\right)^n \int \psi^* \frac{d^n}{dt^n} \psi dt \quad . \quad . \quad . \quad (1.15)$$

$$\int \psi^* t^n \psi dt = \left(\frac{-1}{2\pi j}\right)^n \int \phi^* \frac{d^n}{df^n} \phi df \quad . \quad . \quad . \quad (1.16)$$

The first of these, (1.14), is well known as the "Fourier energy theorem" (Rayleigh, 1889). The other relations can be derived from the identity†

$$\int \psi_1(t) \psi_2(t) dt = \int \phi_1(f) \phi_2(-f) df \quad . \quad . \quad . \quad (1.17)$$

by partial integration, assuming that ψ, ϕ and all their derivatives vanish at infinity.

These very useful reciprocal relations can be summed up in the following simple instructions. When it is desired to express one of the mean values (1.12) by integrals over frequency, replace ψ by ϕ, and the *quantity* t by the *operator* $-\frac{1}{2\pi j}\frac{d}{df}$. This can be called "translation from time language into frequency language." Conversely, when doing the inverse translation, replace ϕ by ψ and the frequency f by the operator $\frac{1}{2\pi j}\frac{d}{dt}$. This corresponds to the somewhat mysterious rule of quantum mechanics: Replace in classical equations the momentum p_x by the operator $\frac{h}{2\pi j}\frac{\partial}{\partial x}$, where x is the co-ordinate conjugate to the momentum p_x. Actually it is no more mysterious than Heaviside's instruction: "Replace the operator d/dt by p," which has long been familiar to electrical engineers.

Applying the rule

$$\bar{f} = \frac{1}{2\pi j} \frac{\int \psi^* \frac{d}{dt} \psi dt}{\int \psi^* \psi dt} \quad . \quad . \quad . \quad . \quad . \quad (1.18)$$

to a simple cisoidal function $\psi = \text{cis}\, 2\pi f_0 t$, we obtain the value f_0 for the mean frequency $\bar{f}$, and similarly $\bar{f}^n = f_0^n$. The mean epochs $\bar{t}^n$, on the other hand, are zero for odd powers, and infinite for even powers $n > 1$. The cisoidal function is to be considered as a limiting case, as the theory is correctly applicable only to signals of finite duration, and with frequency spectra which do not extend to infinity, a condition which is fulfilled by all real, physical signals.

These definitions and rules enable us to formulate the uncertainty relation quantitatively. Let us consider a finite signal, such as is shown, for example, in Fig. 1.6. Let us first fix the mean epoch and the mean frequency of the signal, by means of equations (1.12) and (1.13) or (1.18). These, however, do not count as data, as in a continuous transmission there will be some signal strength at any instant, and at any frequency. We consider $\bar{t}$ and $\bar{f}$ as references, not as data. The first two data will be therefore determined by the mean-square values of epoch and frequency, i.e.

$$\bar{t}^2 = \frac{\int \psi^* t^2 \psi dt}{\int \psi^* \psi dt} \quad . \quad . \quad . \quad . \quad . \quad (1.19)$$

† Cf. CAMPBELL and FOSTER: Reference 1.7, p. 39.

$$\bar{f}^2 = \frac{\int \phi^* f^2 \phi df}{\int \phi^* \phi df} = -\frac{1}{(2\pi)^2} \frac{\int \psi^* \frac{d^2}{dt^2} \psi dt}{\int \psi^* \psi dt} = \frac{1}{(2\pi)^2} \frac{\int \frac{d\psi^*}{dt} \frac{d\psi}{dt} dt}{\int \psi^* \psi dt} \quad . \quad . \quad . \quad (1.20)$$

The second of these has been first translated into "time language," as explained, and transformed by partial integration to put its essentially positive character into evidence.

It may be noted that $\bar{t}^2$ and $\bar{f}^2$, and in general all mean values of even order, remain unaltered if the real signal $s(t)$ or its associate, $\sigma(t)$, is substituted in the place of $\psi(t) = s(t) + j\sigma(t)$. Hence in the following we could again use the real instead of the complex signal, but ψ will be retained in order to simplify some of the analytical expressions and to emphasize the similarity with the formulas of quantum mechanics.

We now define what will be called "the effective duration" Δt and the "effective frequency width" Δf of a signal by the following equations

$$\Delta t = [2\pi \overline{(t - \bar{t})^2}]^{\frac{1}{2}} \quad . \quad . \quad . \quad . \quad (1.21)$$

$$\Delta f = [2\pi \overline{(f - \bar{f})^2}]^{\frac{1}{2}} \quad . \quad . \quad . \quad . \quad (1.22)$$

In words, the effective duration is defined as $\sqrt{(2\pi)}$ times the r.m.s. deviation of the signal from the mean epoch t, and the effective frequency width similarly as $\sqrt{(2\pi)}$ times the r.m.s. deviation from $\bar{f}$. The choice of the numerical factor $\sqrt{(2\pi)}$ will be justified later.

Using the identities

$$\overline{(t - \bar{t})^2} = \bar{t}^2 - (\bar{t})^2 \qquad \overline{(f - \bar{f})^2} = \bar{f}^2 - (\bar{f})^2$$

Δt and Δf can be expressed by means of (1.19) and (1.20). The expressions are greatly simplified if the origin of the time scale is shifted to $\bar{t}$, and the origin of the frequency scale to $\bar{f}$. Both transformations are effected by introducing a new time scale

$$\tau = t - \bar{t} \quad . \quad . \quad . \quad . \quad . \quad . \quad (1.23)$$

and a new signal amplitude

$$\Psi(\tau) = \psi(t) e^{-2\pi j \bar{f} \tau} \quad . \quad . \quad . \quad . \quad (1.24)$$

Expressing t and ψ by the new quantities τ and Ψ, it is found that, apart from a numerical factor 2π, $(\Delta t)^2$ and $(\Delta f)^2$ assume the same form as equations (1.19) and (1.20) for $\bar{t}^2$ and $\bar{f}^2$. Multiplying the two equations we obtain

$$(\Delta t \Delta f)^2 = \frac{1}{4}\left[\frac{\int \Psi^* \tau^2 \Psi d\tau \int \frac{d\Psi^*}{d\tau} \frac{d\Psi}{d\tau} d\tau}{[\int \Psi^* \Psi d\tau]^2}\right] \quad . \quad (1.25)$$

But, by a mathematical identity, a form of the "Schwarz inequality" due to Weyl and Pauli,† the expression in brackets is always larger than unity for any function Ψ for which the integrals exist. We obtain, therefore, the uncertainty relation in the rigorous form

$$\Delta t \Delta f \geqslant \tfrac{1}{2} \quad . \quad . \quad . \quad . \quad . \quad (1.26)$$

This is the mathematical identity which is at the root of the fundamental principle of communication. We see that the r.m.s. duration of a signal, and its r.m.s. frequency-width define a minimum area in the information diagram. How large we assume this minimum area depends on the convention for the numerical factor. By choosing it as $\sqrt{(2\pi)} = 2{\cdot}506$ we have made the number of elementary areas in any large rectangular

† WEYL, H.: "The Theory of Groups and Quantum Mechanics" (Methuen, London, 1931), pp. 77 and 393. Cf. also TOLMAN, R. C.: *loc. cit.*, p. 235, and Appendix 9.3 of this paper.

region of the information diagram equal to the number of independent data which that region can transmit, according to the result obtained in Section 1.

Relation (1.26) is symmetrical in time and frequency, and it suggests that a new representation of signals might be found in which t and f played interchangeable parts. Moreover, it suggests that it might be possible to give a more concrete interpretation to the information diagram by dividing it up into "cells" of size one half, and associating each cell with an "elementary signal" which transmitted exactly one datum of information. This programme will be carried out in the next Section.

(5) THE ELEMENTARY SIGNAL

The mathematical developments up to this point have run rather closely on the lines of quantum mechanics. In fact our results could have been formally obtained by replacing a coordinate x by t, the momentum p by f, and Planck's constant h by unity. But now the ways part, as questions arise in the theory of information which are rather different from those which quantum theory sets out to answer.

The first problem arises directly from the inequality (1.26). What is the shape of the signal for which the product $\Delta t \Delta f$ actually assumes the smallest possible value, i.e. for which the inequality turns into an equality?

The derivation of this signal form is contained in Appendix 9.3; only the result will be given here, which is very simple. *The signal which occupies the minimum area $\Delta t \Delta f = \frac{1}{2}$ is the modulation product of a harmonic oscillation of any frequency with a pulse of the form of a probability function.* In complex form

$$\psi(t) = e^{-\alpha^2(t-t_0)^2} \operatorname{cis}(2\pi f_0 t + \phi) \quad . \quad . \quad (1.27)$$

α, t_0, f_0 and ϕ are constants, which can be interpreted as the "sharpness" of the pulse, the epoch of its peak, and the frequency and phase constant of the modulating oscillation. The constant α is connected with Δt and Δf by the relations

$$\Delta t = \sqrt{\left(\frac{\pi}{2}\right)}\frac{1}{\alpha} \qquad \Delta f = \frac{1}{\sqrt{(2\pi)}}\alpha$$

As might be expected from the symmetrical form of the condition from which it has been derived, the spectrum is of the same analytical form

$$\phi(f) = e^{-\left(\frac{\pi}{\alpha}\right)^2(f-f_0)^2} \operatorname{cis}[-2\pi t_0(f-f_0) + \phi] \quad . \quad (1.28)$$

The envelopes of both the signal and its spectrum, or their absolute values, have the shape of probability curves, as illustrated in Fig. 1.6. Their sharpnesses are reciprocal.

Because of its self-reciprocal character, the probability signal has always played an important part in the theory of Fourier transforms. In three recent papers, Roberts and Simmonds have called attention to some of its analytical advantages.[1.11, 1.12, 1.13] But its minimum property does not appear to have been recognized. It is this property which makes the modulated probability pulse the natural basis on which to build up an analysis of signals in which both time and frequency are recognized as references.

It may be proposed, therefore, to call a pulse according to equation (1.27) an *elementary signal.* In the information diagram it may be represented by a rectangle with sides Δt and Δf, and area one-half, centring on the point (t_0, f_0). It will be shown below that any signal can be expanded into elementary signals in such a way that their representative rectangles cover the whole time-frequency area, as indicated in Fig. 1.7. Their amplitudes can be indicated by a number written into the rectangle, or by shading. Each of these areas, with its associated datum, represents, as it were, one elementary

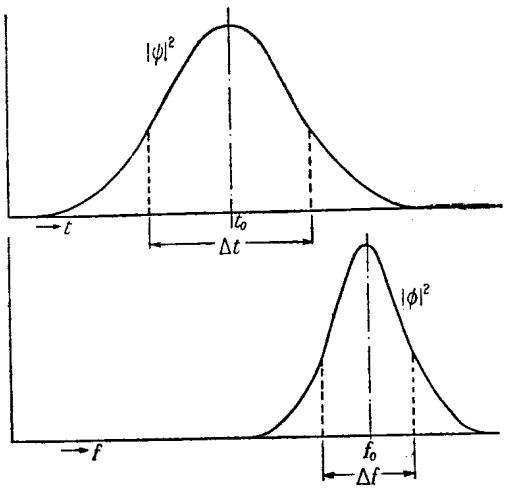

Fig. 1.6.—Envelope of the elementary signal.

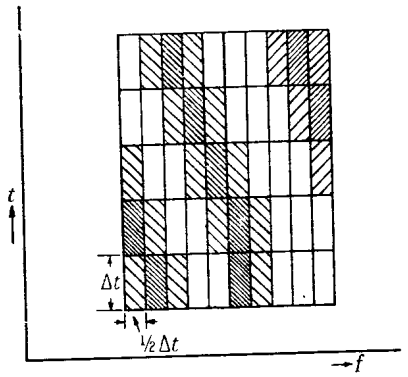

Fig. 1.7.—Representation of signal by logons.

quantum of information, and it is proposed to call it a *logon.* Expansion into elementary signals is a process of which Fourier analysis and time description are special cases. The first is obtained at $\alpha = 0$, in which case the elementary signal becomes a sine wave of infinite length; the second at $\alpha \to \infty$, when it passes into a "delta function."

It will be convenient to explain the expansion into elementary signals in two steps. The first step leads to elementary areas of size unity, with two associated data, but it is simpler and more symmetrical than the second step, which takes us to the limit of sub-division.

This first step corresponds to division of the information area by a network of lines with distances Δt and $1/\Delta t$ respectively, as illustrated in Fig. 1.8.* The elementary areas have suffixes n

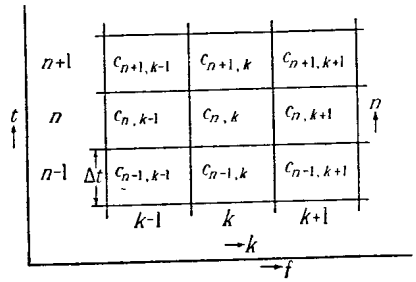

Fig. 1.8.—Representation of signal by a matrix of complex amplitudes.

* For perfect symmetry the spacings in the network ought to have been taken as $(\sqrt{2})\Delta t$ and $1/(\sqrt{2})\Delta t = (\sqrt{2})\Delta f$ respectively.

in the time direction, and k in the frequency direction. The centre lines (horizontally) may be at $t_n = n\,\Delta t$, assuming for convenience that we measure time from the "zero"-th of these lines: The expansion is given by the following formula

$$\psi(t) = \sum_{-\infty}^{\infty} n \sum_{-\infty}^{\infty} k\; c_{nk} \exp - \pi \frac{(t - n\Delta t)^2}{2(\Delta t)^2} \operatorname{cis} (2\pi k t/\Delta t) \qquad (1.29)$$

The matrix of the complex coefficients c_{nk} represents the signal in a symmetrical way, as it is easy to see that if the expansion exists we arrive—apart from a constant factor—at the same coefficients if we expand $\phi(f)$ instead of $\psi(t)$.

As the elementary signals in (1.29) are not orthogonal, the coefficients c_{nk} are best obtained by successive approximations. In the first approximation we consider each horizontal strip with suffix n by itself, and expand the function $\psi(t)$ as if the other strips did not exist, in the interval $(t_n - \frac{1}{2}\Delta t)$ to $(t_n + \frac{1}{2}\Delta t)$, by putting

$$\psi(t) \exp \pi \frac{(t - n\Delta t)^2}{2(\Delta t)^2} = \sum_{0}^{\infty} k\; c_{nk} \operatorname{cis} (2\pi k t/\Delta t)$$

In this formula the exponential function, which is independent of k, has been brought over to the left. We have now a known function on the left, and a Fourier series on the right, which by known methods gives immediately the first approximation for the coefficients c_{nk}. This represents $\psi(t)$ correctly in the intervals for which the series are valid, but not outside them. If the first approximations are added up with summation indices n, there will be a certain error due to their overlap. A second approximation can be obtained by subtracting this error from $\psi(t)$ in eqn. (1.29) and repeating the procedure. It can be expected to converge rapidly, as the exponential factor decays so fast that only neighbouring strips n influence each other perceptibly.

This expansion gives ultimately one complex number c_{nk} for every two elementary areas of size one-half. The real and imaginary parts can be interpreted as giving the amplitudes of the following two real elementary signals

$$\begin{matrix} s_c(t) \\ s_s(t) \end{matrix} = \exp - \alpha^2 (t - t_0)^2 \begin{matrix} \cos \\ \sin \end{matrix} 2\pi f_0 (t - t_0) \quad . \qquad (1.30)$$

where $\alpha^2 = \frac{1}{2}\pi/(\Delta t)^2$. These can be called the "cosine-type" and "sine-type" elementary signals. They are illustrated in Fig. 1.9. We can use them to obtain a real expansion, allocating

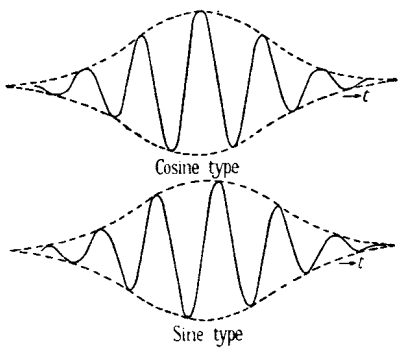

Fig. 1.9.—Real parts of elementary signal.

one datum to every cell of one-half area. But it may be noted that this will have to be necessarily a more special and less symmetrical expansion than the previous one, as the transform of a cosine-type elementary signal, for example, will not in general be of the same type. As always in communication theory, a description by complex numbers is formally simpler than by real data.

We now divide up the information plane as shown in Fig. 1.10

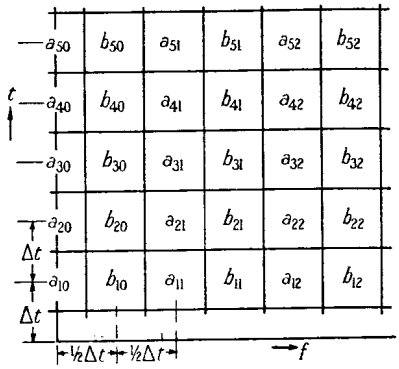

Fig. 1.10.—Expansion of arbitrary signal in cosine-type and sine-type elementary signals.

into cells of size one-half, measuring Δt in the time, and $\frac{1}{2}\Delta t$ in the frequency, direction. Starting from the line of zero frequency, we allocate to these areas in every strip alternately a cosine-type and a sine-type elementary signal. Evidently we must start with a cosine signal at $f = 0$, as the sine-type signal would be zero. This leads us to the following expansion of the real signal $s(t)$:—

$$s(t) = \sum_{-\infty}^{\infty} n \exp - \pi \frac{(t - n\Delta t)^2}{2(\Delta t)^2} \sum_{0}^{\infty} k \left[a_{nk} \cos 2\pi k (t - n\Delta t)/\Delta t + b_{nk} \sin 2\pi (k + \tfrac{1}{2})(t - n\Delta t)/\Delta t \right] . \qquad (1.31)$$

In order to find the coefficients a_{nk} and b_{nk} we can carry out the same process of approximation as explained in connection with expansion (1.30), but with a difference. At the first step we arrive at an equation of a form

$$f_n(x) = \sum_{0}^{\infty} k\; a_{nk} \cos kx + b_{nk} \sin (k + \tfrac{1}{2})x$$

with the abbreviations $x = 2\pi(t - n\Delta t)/\Delta t$, and $f(x) = s(t) \exp \frac{1}{2}\pi(t - n\Delta t)^2/(\Delta t)^2$. But the trigonometric series on the right is not a Fourier series. It is of a somewhat unusual type, in which the sine terms have frequencies mid-way between the cosine terms. It will be necessary to show briefly that this series can be used also for the representation of arbitrary functions. First we separate the even and odd parts on both sides of the equation, by putting

$$\tfrac{1}{2}[f_n(x) + f_n(-x)] = \sum_{0}^{\infty} k\; a_{nk} \cos kx$$

$$\tfrac{1}{2}[f_n(x) - f_n(-x)] = \sum_{0}^{\infty} k\; b_{nk} \sin (k + \tfrac{1}{2})x$$

The first is a Fourier series, but not the second. We have seen, however, in Section 3, how all the frequencies contained in a function can be raised by a constant amount by means of a process which involves calculating the function in quadrature with it. Applying this operation to both sides of the last equation we can add $\frac{1}{2}$ to $k + \frac{1}{2}$, and obtain the ordinary Fourier sine series, which enables the coefficients to be calculated.

The expansion into logons is, in general, a rather inconvenient

process, as the elementary signals are not orthogonal. If only approximate results are required, it may be permitted to neglect the effect of their interference. This becomes plausible if we consider that an elementary signal has 76·8% of its energy inside the band Δt or Δf, and only 11·6% on either side. Approximately correct physical analysis could be carried out by means of a bank of resonators with resonance curves of probability shape. It can be shown that if the energy collected by a resonator tuned to $\bar{f}$ is taken as 100%, the resonators on the right and left of it, tuned to $\bar{f}+\Delta f$ and $\bar{f}-\Delta f$, would collect only 0·65% each. Roberts and Simmonds[1.11, 1.12, 1.13] have given consideration to the problem of realizing circuits with responses of probability shape.

Though the overlapping of the elementary signals may be of small practical consequence, it raises a question of considerable theoretical interest. The principle of causality requires that any quantity at an epoch t can depend only on data belonging to epochs earlier than t. But we have seen that we could not carry out the expansion into elementary signals exactly without taking into consideration also the "overlap of the future." In fact, strict causality exists only in the "time language"; as soon as we use frequency as an additional reference the sort of uncertainty occurs which in modern physics has often been called the "breakdown of causality." But rigorous time-analysis is possible only with ideal oscillographs, not with any real physical instrument; hence strict causality never applies in practice. A limitation of this concept ought not to cause difficulties to electrical engineers who are used to the Fourier integral, i.e. to an entirely non-causal method of description.

(6) SIGNALS TRANSMITTED IN MINIMUM TIME

The elementary signals which have been discussed in the last Section assure the best utilization of the information area in the sense that they possess the smallest product of effective duration by effective frequency width. It follows that, if we prescribe the effective width Δf of a frequency channel, the signal transmitted through it in minimum time will have an envelope

$$\Psi(t) = \exp -(2\pi)(\Delta f)^2(t-\bar{t})^2 \quad . \quad . \quad (1.32)$$

and, apart from a cisoidal factor, a Fourier transform

$$\Phi(f) = \exp -\frac{\pi}{2}\left(\frac{f-\bar{f}}{\Delta f}\right)^2 \quad . \quad . \quad . \quad (1.33)$$

But the problem which most frequently arises in practice is somewhat different. Not the effective spectral width is prescribed, but the total width; i.e. a frequency band (f_2-f_1) is given, outside which the spectral amplitude must be zero. What is the signal shape which can be transmitted through this channel in the shortest effective time, and what is its effective duration?

Mathematically the problem can be reduced to finding the spectrum $\phi(f)$ of a signal which makes

$$\Delta t = \frac{1}{(2\pi)^2}\int_{f_1}^{f_2}\frac{d\phi^*}{df}\frac{d\phi}{df}df \Big/ \int_{f_1}^{f_2}\phi^*\phi df \quad . \quad . \quad (1.34)$$

a minimum, with the condition that $\phi(f)$ is zero outside the range f_1-f_2. But this is equivalent to the condition that $\phi(f)$ vanishes at the limits f_1 and f_2. Otherwise, if $\psi(f)$ had a finite value at the limits but vanished outside, the discontinuity at the limits would make the numerator of equation (1.34) divergent. (This is the converse of the well-known fact that a signal with an abrupt break contains frequencies up to infinity, which decay only hyperbolically, not fast enough to make f^2 finite.)

The problem is one of the calculus of variations, and is solved in Appendix 9.4, where it is shown that the signals transmitted in minimum time must be among the solutions of a differential equation

$$\frac{d^2\phi}{df^2}+\Lambda\phi = 0 \quad . \quad . \quad . \quad . \quad . \quad (1.35)$$

where Λ is an undetermined constant. But the possible values of Λ are defined by the auxiliary condition that $\phi(f)$ must vanish at the limits of the waveband.* Hence all admissible solutions are of the form

$$\phi(f) = \sin k\pi\frac{f-f_1}{f_2-f_1} \quad . \quad . \quad . \quad . \quad (1.36)$$

where k is an integer. We can call this the kth characteristic function of transmission through an ideal band-pass filter. Its effective duration is

$$\Delta t = \sqrt{\left(\frac{\pi}{2}\right)}\frac{k}{f_2-f_1} \quad . \quad . \quad . \quad . \quad (1.37)$$

and its effective frequency width

$$\Delta f = (f_2-f_1)\sqrt{\left(\frac{\pi}{6}-\frac{1}{\pi k^2}\right)} \quad . \quad . \quad . \quad (1.38)$$

The shortest duration Δt belongs to $k=1$, i.e. to the fundamental characteristic function, which is illustrated in Fig. 1.11.

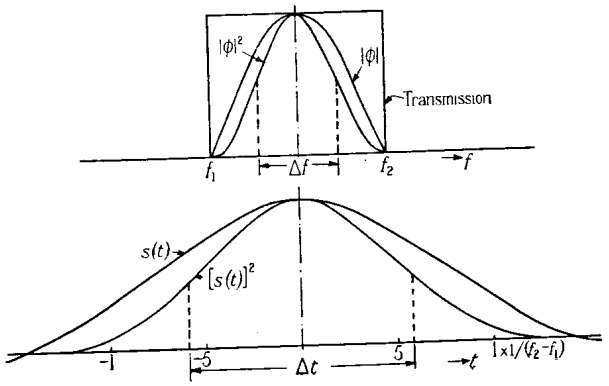

Fig. 1.11.—Spectrum of signal which can be transmitted in minimum time through an ideal band-pass filter, and the signal itself.

The product $\Delta t\Delta f$ is also smallest for $k=1$; its value is 0·571. Though this is not much more than the absolute minimum, 0·5, the transmission channel is poorly utilized, as the effective frequency width is only 0·456 of (f_2-f_1). Practice has found a way to overcome this difficulty by means of asymmetric, vestigial or single-sideband transmission. In these methods the spectrum is cut off at or near the centre more or less abruptly. This produces a "splash," a spreading out of the signal in time, but this effect is compensated in the reception, when the other sideband is reconstituted and added to the received signal.

The advantages of a signal of sine shape, as shown in Fig. 1.11, have already been noticed, as it were, empirically by Wheeler and Loughren† in their thorough study of television images. As in television the signals transmitted represent light intensities, i.e. energies, our definitions must be applied here with a modification. Either the square root of the light intensity must be substituted for ψ, or the square root of the Fourier transform

* Problems of this kind are known in mathematics and theoretical physics as Sturm–Liouville "proper value" problems. Cf. COURANT, R., and HILBERT, D.: "Methoden der mathematischen Physik," Vol. 1 (Springer, Berlin, 1931), or "Interscience" (New York, 1943), p. 249, or any textbook on wave mechanics.

† Ref. No. 14. In comparing the above results with theirs it may be borne in mind that their "nominal cut-off frequency" is one-half of a sideband, and one-quarter of the total channel width.

of the signal for ϕ. The practical difference between these two possible definitions becomes very small in minimum problems. If we adopt the second, we obtain the same "cosine-squared" law for the optimum spectral distribution of energy which Wheeler and Loughren have considered as the "most attractive compromise."

Fig. 1.11 shows also the signal $s(t)$ which is transmitted in minimum time by a band-pass filter. It can be seen that it differs in shape very little indeed from its spectrum. It may be noted that the total time interval in which the signal is appreciably different from zero is $2/(f_2 - f_1)$.

It can be seen from Fig. 1.11, that the optimum signal utilizes the edges of the waveband—in single-sideband television, the upper edge—rather poorly. But this is made even worse in television by the convention of making the electromagnetic amplitudes proportional to the light intensities, so that the electromagnetic energy spectrum in the optimum case has the shape of a $\cos^4$ curve. This means that the higher frequencies will be easily drowned by atmospherics. Conditions can be improved by "compression-expansion" methods, in which, for example, the square root of the light intensity is transmitted, and squared in the receiver.

(7) DISCUSSION OF COMMUNICATION PROBLEMS BY MEANS OF THE INFORMATION DIAGRAM

As the foregoing explanations might appear somewhat abstract, it appears appropriate to return to the information diagram and to demonstrate its usefulness by means of a few examples.

Let us take frequency modulation as a first example. Fig. 1.12

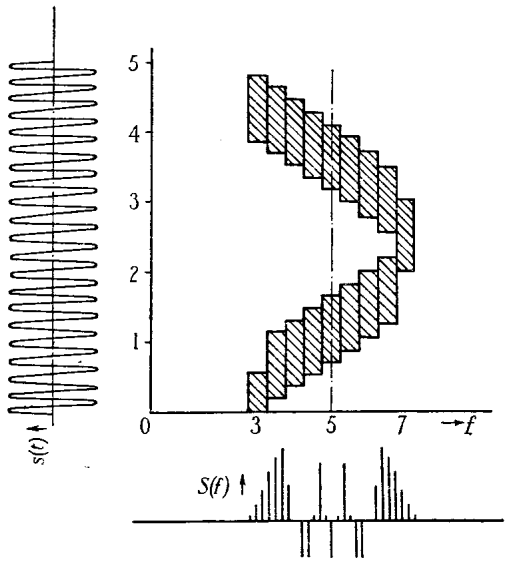

Fig. 1.12.—Three representations of frequency modulation.

contains three different illustrations of the same slowly modulated carrier: the time representation, the spectrum and its picture in the information diagram. It can be seen that the third illustration corresponds very closely to our familiar idea of a variable frequency. The only departure from the naïve expectation that its pictorial representation would be an undulating curve is that the curve has to be thick and blurred. But it appears preferable not to show the blurring, not only because it is difficult to draw, but also because it might give rise to the idea that the picture could be replaced by a definite density distribution. Instead we have represented it by logons of area one-half. The shape of the rectangles, i.e. the ratio $\Delta t/\Delta f$, is entirely arbitrary and depends on the conventions of the analysis. If Δt is taken equal to the damping time of, say, a bank of reeds, the picture gives an approximate description of the response of the instrument. It gives also a rough picture of our aural impression of a siren. How this rough picture can be perfected will be shown in Part 2.

A second example is time-division multiplex telephony, a problem which almost forces on us the simultaneous consideration of time and frequency. Bennett[1.15] has discussed it very thoroughly by an irreproachable method, but, as is often the case with results obtained by Fourier analysis, the physical origin of the results remains somewhat obscure. An attempt will now be made to give them a simple interpretation.

In time-division multiplex telephony, synchronized switches at both ends of a line connect the line in cyclic alternation to a number N of channels. Let f_s be the switching frequency, i.e. the number of contacts made per second. What is the optimum switching frequency if N conversations, each occupying a frequency band w are to be transmitted without loss of information and without crosstalk—i.e. mutual interference between channels—and what is the total frequency-band requirement W?

The information diagram is shown in Fig. 1.13. The fre-

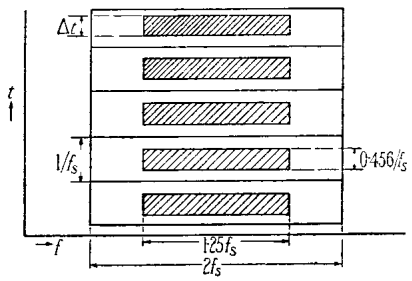

Fig. 1.13.—Information diagram of time-division multiplex-telephony system.

quency band W is sub-divided in the time direction into rectangles of a duration $1/f_s$, i.e. f_s rectangles per sec. If these are to transmit independent data they cannot transmit less than one datum at a time. But one datum, or logon, at a time is also the optimum, as otherwise the receivers would have to discriminate between two or more data in the short time of contact, and distribute them somehow over the long waiting time between two contacts. Hence, if no information is to be lost, the number of contacts per second must be equal to the data of N conversations each of width w, i.e. $f_s = 2Nw$. This is also Bennett's result.

We now consider the condition of crosstalk. This is the exact counterpart of the problem of minimum transmission time in a fixed-frequency channel, considered in the last Section, except that time and frequency are interchanged. Thus we can say at once that the optimum signal form will be the sine shape of Fig. 1.11, and the frequency requirement will be very nearly $2f_s$. The characteristic rectangle $\Delta t \Delta f$ of this signal is shown in every switching period, with the dimensions as obtained in the last Section. The total frequency band requirement becomes $W = 2f_s = 4Nw$. This can be at once halved by single-sideband transmission, i.e. transmitting only one-half of W. But even this does not represent the limit of economy, as the signal is symmetrical not only in frequency, but also in time. In the case of the example treated in the previous Section this was of no use, as the epoch of the signal was unknown. But in time-division multiplex the epoch of each signal is accurately known; hence it must be possible to halve the waveband once more and reduce W to the minimum requirement $W = Nw$. An ingenious,

though rather complicated, method of achieving this, by means of special filters associated with the receiving channels, has been described by Bennett.[1.15]

(8) REFERENCES

(1.1) CARSON, J. R.: "Notes on the Theory of Modulation," *Proceedings of the Institute of Radio Engineers*, 1922, **10**, p. 57.

(1.2) NYQUIST, H.: "Certain Factors affecting Telegraph Speed," *Bell System Technical Journal*, 1924, **3**, p. 324.

(1.3) KÜPFMÜLLER, K.: "Transient Phenomena in Wave Filters," *Elektrische Nachrichten-Technik*, 1924, **1**, p. 141.

(1.4) HARTLEY, R. V. L.: "Transmission of Information," *Bell System Technical Journal*, 1928, **7**, p. 535.

(1.5) GRAY, F., HORTON, J. W., and MATHES, C. R.: "The Production and Utilization of Television Signals," *ibid.*, 1927, **6**, p. 560.

(1.6) LÜSCHEN, F.: "Modern Communication Systems," *Journal I.E.E.*, 1932, **71**, p. 776.

(1.7) CAMPBELL, G. A., and FOSTER, R. M.: "Fourier Integrals for Practical Applications," *Bell Telephone System Monograph* B 584, 1931.

(1.8) CARSON, J. R., and FRY, T. C.: "Variable-frequency Electric Circuit Theory," *Bell System Technical Journal*, 1937, **16**, p. 513.

(1.9) STEWART, G. W.: "Problems Suggested by an Uncertainty Principle in Acoustics," *Journal of the Acoustical Society of America*, 1931, **2**, p. 325.

(1.10) GOLDMARK, P. C., and HENDRICKS, P. S.: "Synthetic Reverberation," *Proceedings of the Institute of Radio Engineers*, 1939, **27**, p. 747.

(1.11) ROBERTS, F. F., and SIMMONDS, J. C.: "Some Properties of a Special Type of Electrical Pulse," *Philosophical Magazine*, (VII), 1943, **34**, p. 822.

(1.12) ROBERTS, F. F., and SIMMONDS, J. C.: "Further Properties of Recurrent Exponential and Probability Waveforms," *ibid.*, (VII), 1944, **35**, p. 459.

(1.13) ROBERTS, F. F., and SIMMONDS, J. C.: "The Physical Realizability of Electrical Networks having Prescribed Characteristics," *ibid.*, (VII), 1944, **35**, p. 778.

(1.14) WHEELER, H. A., and LOUGHREN, A. V.: "The Fine Structure of Television Images," *Proceedings of the Institute of Radio Engineers*, 1938, **26**, p. 540.

(1.15) BENNETT, W. R.: "Time-division Multiplex Systems," *Bell System Technical Journal*, 1941, **20**, p. 199.

(9) APPENDICES

(9.1) Analysis in Terms of Other than Simple Periodic Functions

The discussion in Section 1 suggests a question: Why are we doing our analysis in terms of sine waves, and why do we limit our communication channels by fixed frequencies? Why not choose other orthogonal functions? In fact we could have taken, for example, the orthogonalized Bessel functions

$$\sqrt{(t)}J_n(r_k t/\tau)$$

as the basis of expansion. J_n is a Bessel function of fixed but arbitrary order n; r_k is the kth root of $J_n(x) = 0$; k is the expansion index. These functions are orthogonal in the interval $0 < t < \tau$. The factors r_k/τ have the dimension of a frequency. We could now think of limiting the transmission channel by two "Bessel frequencies," say μ_1 and μ_2. Here the first difference arises. The number of spectral lines between these limits will be the number of the roots of $J_n(x) = 0$ between the limits $\mu_1\tau$ and $\mu_2\tau$. But this number is not proportional to τ. Hence a Bessel channel, or a channel based on any function other than simple harmonic functions, would not transmit the same amount of information in equal time intervals.

In principle it would be possible to construct circuits which transmitted without distortion any member of a selected set of orthogonal functions. But only harmonic functions satisfy linear differential equations in which time does not figure explicitly; hence these are the only ones which can be transmitted by circuits built up of constant elements. Every other system requires variable circuit components, and as there will be a distinguished epoch of time it will also require some sort of synchronization between transmitter and receiver. In competition with fixed-waveband systems any such method will have the disadvantage that wider wavebands will be required to avoid interference with other transmissions. Though this disadvantage —as in the case of frequency modulation—might be outweighed by other advantages, investigation of such systems is outside the scope of the present study, which is mainly devoted to the problem of waveband economy.

(9.2) Mechanical Generation of Associated Signals, and the Problem of Direct Production of Single Sidebands

In order to gain a more vivid picture of signals in quadrature than the mathematical explanations of Section 3 can convey, it may be useful to discuss a method of generating them mechanically. It is obvious from equations (1.7) and (1.8) that, in order to generate the signal $\sigma(t)$ associated with a given signal $s(t)$, it is necessary *to know not only the past but also the future*. Though formally the whole future is involved, the "relevant future" in transmission problems is usually only a fraction of a second. This means that we can produce $\sigma(t)$ with sufficient accuracy if we convert, say, 0·1 sec of the future into the past; in other words, if we *delay* the transmission of $s(t)$ by about this interval. Fig. 1.14 shows a device which might accomplish this.

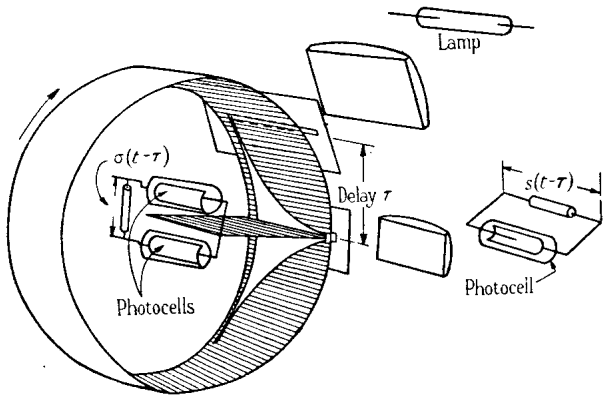

Fig. 1.14.—Device for mechanical generation of a signal in quadrature with a given signal.

The light of a lamp, the intensity of which is modulated by the signal $s(t)$, is thrown through a slit on a transparent rotating drum, coated with phosphorescent powder. The drum therefore carries a record of the signal with it, which decays slowly. After turning through a certain angle the record passes a slit, and here the light is picked up by a photocell, which transmits $s(t)$ with a delay corresponding to the angle.* On the inside of the drum two hyperbolically-shaped apertures are arranged at both sides of the slit opposite to the first photocell. The light from the two hyperbolic windows is collected by two photocells, which are connected in opposition. By comparing this arrangement

* A somewhat similar device (for another purpose) has been described by Goldmark and Hendricks (Ref. No. 1.10).

with equation (1.7) it is easy to see that the difference of the two photocell currents will be proportional to the function in quadrature with $s(t)$.

The complex signal has been discussed at some length as it helps one to understand certain problems of communication engineering. One of these is the problem of single-sideband transmission. It is well known that it is not possible to produce a single sideband directly. The method employed is to produce both sidebands and to suppress one. Equation (1.7) explains the reason. *Direct single-sideband production involves knowledge of the future.* The conventional modulation methods always add *and* subtract frequencies simultaneously. With mechanisms like the one shown in Fig. 1.14 it becomes possible to add *or* subtract them. This means forming the following expression

$$\mathscr{R}[\psi(t)\,\exp j\omega_c t] = s(t)\cos\omega_c t - \sigma(t)\sin\omega_c t$$

where ω_c is the angular carrier frequency. By substituting a harmonic oscillation for $s(t)$ is is easy to verify that ω_c has been added to every frequency present in the signal. Direct production of single sidebands involves, therefore, the following operations: Modulate the signal with the carrier wave, and subtract from the product the modulation product of the signal in quadrature with the carrier wave in quadrature. It is not, of course, suggested that this might become a practical method; the intention was merely to throw some light on the root of a well-known impossibility.

(9.3) The Schwarz Inequality and Elementary Signals

The inequality

$$(\int\Psi^*\Psi d\tau)^2 \leqslant 4(\int\Psi^*\tau^2\Psi d\tau)\left(\int\frac{d\Psi^*}{d\tau}\,\frac{d\Psi}{d\tau}d\tau\right) \qquad (1.39)$$

is valid for any real or complex function Ψ which is continuous and differentiable and vanishes at the integration limits. The following is a modification of a proof given by H. Weyl.†

If a_1, b_1 are two sets of n real or complex numbers, a theorem due to H. A. Schwarz states that

$$|a_1b_1 + \ldots + a_nb_n|^2 \leqslant (a_1a_1^* + \ldots + a_na_n^*)(b_1b_1^* + \ldots + b_nb_n^*) \quad . \quad (1.40)$$

If a's and b's are all real numbers, this can be interpreted as expressing the fact that the cosine of the angle of two vectors with components $a_1 \ldots a_n$ and $b_1 \ldots b_n$ in an n-dimensional Euclidian space is smaller than unity. This can be easily understood, as in a Euclidian space of any number of dimensions a two-dimensional plane can be made to pass through any two vectors issuing from the origin; hence the angle between them has the same significance as in plane geometry. Equation (1.40) is a generalization of this for "Hermitian" space, in which the components or co-ordinates of the vectors are themselves complex numbers.

By a passage to the limit the sums in (1.40) may be replaced by integrals, so that

$$\Sigma a_1b_1 \to \int f(\tau)g(\tau)d\tau$$

and similarly for the other two sums. The real variable τ now takes the place of the summation index. The Schwarz inequality now becomes

$$|\int fg\,d\tau|^2 \leqslant (\int ff^*d\tau)(\int gg^*d\tau) \quad . \quad . \quad . \quad (1.41)$$

This remains valid if we replace f and g by their conjugates

$$|f^*g^*d\tau|^2 \leqslant (\int ff^*d\tau)(\int gg^*d\tau) \quad . \quad . \quad . \quad (1.42)$$

† Weyl, H.: "The Theory of Groups and Quantum Mechanics" (Methuen, 1931), p. 393.

Adding (1.41) and (1.42) we obtain

$$2(\int ff^*d\tau)(\int gg^*d\tau) \geqslant |\int fgd\tau|^2 + |\int f^*g^*d\tau|^2 \geqslant \tfrac{1}{2}[\int(fg + f^*g^*)d\tau]^2 \quad . \quad (1.43)$$

The second part of this inequality states the fact that the sum of the absolute squares of two conjugate complex numbers is never less than half the square of their sums.

We now put

$$f = \tau\Psi \qquad g = \frac{d\Psi^*}{d\tau} \qquad . \quad . \quad . \quad (1.44)$$

Substitution in (1.43) gives

$$4(\int ff^*d\tau)(\int gg^*d\tau) \geqslant \left[\int\left(\Psi\frac{d\Psi^*}{d\tau} + \Psi^*\frac{d\Psi}{d\tau}\right)\tau d\tau\right]^2 \qquad (1.45)$$

The right-hand side can be transformed by partial integration into

$$\int\left(\Psi\frac{d\Psi^*}{d\tau} + \psi^*\frac{d\Psi}{d\tau}\right)\tau d\tau = \int\tau\frac{d}{d\tau}(\Psi^*\Psi)d\tau = -\int\Psi^*\Psi d\tau \qquad (1.46)$$

where it has been assumed that Ψ vanishes at the integration limits. Substituting this in (1.45) we obtain the inequality (1.39).

In order to obtain the elementary signals we must investigate when this inequality changes into an equality. From the geometrical interpretation of Schwarz's inequality (1.40), it can be concluded at once that the equality sign will obtain if, and only if, the two vectors a, b have the same direction, i.e.

$$b_1 = Ca_1$$

In Hermitian space the direction is not changed by multiplication by a complex number, hence C need not be real.

This condition can be applied also to the inequality (1.39), but with a difference. (1.39) will become an equation only if *both* the conditions (1.41) and (1.42) become equalities; i.e. if the following two equations are fulfilled

$$f = Cg \text{ and } f^* = C'g^* \qquad . \quad . \quad . \quad (1.47)$$

where C and C' are real or complex constants. But these two equations are compatible if, and only if,

$$C' = C^* \qquad . \quad . \quad . \quad . \quad . \quad . \quad (1.48)$$

in which case the two equations (1.47) become identical. On substituting f and g from (1.44) they give the two equivalent equations

$$\frac{d\psi^*}{d\tau} = C\tau\Psi \text{ and } \frac{d\Psi}{d\tau} = C^*\tau\Psi^* \quad . \quad . \quad (1.49)$$

From either of these we can eliminate Ψ or its conjugate Ψ^* and are led to the second-order differential equation

$$\frac{d}{d\tau}\left(\frac{1}{\tau}\frac{d\Psi}{d\tau}\right) = CC^*\tau\Psi \quad . \quad . \quad . \quad . \quad (1.50)$$

Multiplying both sides by $(d\Psi/d\tau)/\tau$, this becomes integrable and gives

$$\left(\frac{1}{\tau}\frac{d\Psi}{d\tau}\right)^2 = CC^*\Psi^2 + \text{const.} \quad . \quad . \quad . \quad (1.51)$$

But the constant is zero, as at infinity both Ψ and $d\Psi/d\tau$ must vanish. We thus obtain the first-order equation

$$\frac{d\Psi}{d\tau} = \pm\,(CC^*)^{\frac{1}{2}}\tau\Psi \quad . \quad . \quad . \quad . \quad (1.52)$$

th the solution (apart from a constant factor)

$$\Psi = \exp \pm \tfrac{1}{2}|C|\tau^2 \quad . \quad . \quad . \quad . \quad . \quad (1.53)$$

the two signs we can retain only the negative one, as other- e the signal would not vanish at infinity. Putting $\frac{1}{2}|C| = \alpha^2$ obtain the envelope of the elementary signal. The signal ψ lf results from this by multiplying by cis $2\pi f'(t - \bar{t})$ and is cussed in Section 5.

It will be useful to sketch briefly the difference between the alysis based on elementary signals and the method of wave chanics. In the foregoing we have answered the question: hat functions Ψ make the product $\Delta f \Delta t$ assume its smallest ssible value, i.e. one-half? The question posed by wave chanics is more general: What functions Ψ makes $\Delta f \Delta t$ a inimum, while fulfilling the condition of vanishing at infinity? his is a problem of the calculus of variations, which leads, stead of to eqn. (1.50), to a more general equation, called the ave equation of the harmonic oscillator":

$$\frac{d^2\Psi}{d\tau^2} + (\lambda - \alpha^2\tau^2)\Psi = 0$$

here λ and α are real constants. This equation, which con- ins (1.50) as a special case, has solutions which are finite every- here and vanish at infinity only if

$$\lambda = \alpha(2n + 1)$$

here n is a positive integer. These "proper" or "characteristic" lutions of the wave equation are (apart from a constant factor)

$$\Psi_n = e^{-\frac{1}{2}\omega^2\tau^2}\frac{d^n}{d\tau^n}e^{-\alpha^2\tau^2}$$

hey are known as orthogonal Hermite functions* and form e basis of wave mechanical analysis of the problem of the near oscillator. They share with the probability function— hich can be considered as the Hermite function of zero order— e property that their Fourier transforms are of identical type. he product $\Delta f \Delta t$ for the nth Hermite function is

$$\Delta t \Delta f = \tfrac{1}{2}(2n + 1)$$

That is to say that the Hermite functions occupy in the informa- on diagram areas of size $\frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ . . . Because of their ortho- onality Hermite functions readily lend themselves to the expan- on of arbitrary signals; hence their importance in wave echanics. But they are less suitable for the analysis of con- nuously emitted signals, as they presuppose a distinguished poch of time $t = 0$, and they do not permit the sub-division of he information area into non-overlapping elementary cells.†

* Also known as parabolical cylinder functions and Weber–Hermite functions f. WHITTAKER and WATSON: "Modern Analysis," pp. 231, 347. They are discussed all textbooks on wave mechanics. Cf. also the study by BABER, T. D. H., and RSKY, L.: "Note of Certain Integrals involving Hermite's Polynomials," *Philo- phical Magazine* (VII), 1944, **35**, p. 532.

† The derivations in this Appendix can be considerably shortened if use is made the symbolic operator method of quantum mechanics. Cf. MAX BORN: "Atomic hysics" (Blackie, 1935), Appendix XXI, pp. 309–313.

(9.4) Signals Transmitted in Minimum Time through a Given Frequency Channel

It will be convenient to use "frequency language," i.e. to express the signal by its Fourier transform $\phi(f)$. The problem is to make the effective duration Δt of a signal a minimum, with the condition that $\phi(f) = 0$ outside an interval $f_1 - f_2$. Thus

$$\Delta t = \frac{1}{(2\pi)^2 M_0}\int_{f_1}^{f_2}\frac{d\phi^*}{df}\frac{d\phi}{df}df \quad . \quad . \quad . \quad (1.54)$$

must be a minimum, where

$$M_0 = \int_{f_1}^{f_2}\phi^*\phi df$$

This is equivalent to making the numerator in (1.54) a minimum with the auxiliary condition $M_0 = \text{constant}$, and this in turn can be formulated by Lagrange's method in the form

$$\delta\int\left(\frac{d\phi^*}{df}\frac{d\phi}{df} + \Lambda\phi^*\phi\right)df = 0 \quad . \quad . \quad . \quad (1.55)$$

where Λ is an undetermined multiplier. The variation of the first term is

$$\begin{aligned}\delta\int\frac{d\phi^*}{df}\frac{d\phi}{df}df &= \int\left(\frac{d\phi^*}{df}\delta\frac{d\phi}{df} + \frac{d\phi}{df}\delta\frac{d\phi^*}{df}\right)df = \int\left(\frac{d\phi^*}{df}\frac{d\delta\phi}{df} + \frac{d\phi}{df}\frac{d\delta\phi^*}{df}\right)df \\ &= \left[\frac{d\phi^*}{df}\delta\phi + \frac{d\phi}{df}\delta\phi^*\right]_{f_1}^{f_2} - \int\left(\frac{d^2\phi^*}{df^2}\delta\phi + \frac{d^2\phi}{df^2}\delta\phi^*\right)df \quad (1.56)\end{aligned}$$

But at the limits ϕ must vanish, as it is zero outside the interval and must be continuous at the limit, as otherwise the integral (1.54) would not converge. Hence we have here $\delta\phi = \delta\phi^* = 0$, and the first term vanishes. The variation of the second term in (1.55) is

$$\Lambda\int(\phi^*\delta\phi + \phi\delta\phi^*)df \quad . \quad . \quad . \quad . \quad (1.57)$$

The condition (1.55) thus gives

$$\int\left[\left(\frac{d^2\phi^*}{df^2} + \Lambda\phi^*\right)\delta\phi + \left(\frac{d^2\phi}{df^2} + \Lambda\phi\right)\delta\phi^*\right]df = 0 \quad (1.58)$$

and this can be identically fulfilled for arbitrary variations $\delta\phi$ if, and only if,

$$\frac{d^2\phi}{df^2} + \Lambda\phi = 0 \quad . \quad . \quad . \quad . \quad (1.59)$$

This is the differential equation which has to be satisfied by the signal transmitted in minimum time. Its solution is discussed in Section 6.

Part 2. THE ANALYSIS OF HEARING

SUMMARY

The methods developed in Part 1 are applied to the analysis of hearing sensations, in particular to experiments by Shower and Biddulph, and by Bürck, Kotowski and Lichte on the discrimination of frequency and time by the human ear. It is shown that experiments of widely different character lead to well-defined threshold "areas of discrimination" in the information diagram. At the best, in the interval 60–1 000 c/s the human ear can discriminate very nearly every second datum of information; i.e. the ear is almost as perfect as any instrument can be which is not responsive to phase. Over the whole auditory range the efficiency is much less than 50%, as the discrimination falls off sharply at higher frequencies.

The threshold area of discrimination appears to be independent of the duration of the signals between about 20 and 250 millisec. This remarkably wide interval cannot be explained by any mechanism in the inner ear, but may be explained by a new hypothetical effect in nerve conduction, i.e. the mutual influence of adjacent nerve fibres.

(1) ANALYSIS OF HEARING

In relation to the ear, two rather distinct questions will have to be answered. The first is: How many logons must be transmitted per second for intelligible speech? The second is the corresponding question for the reproduction of speech or music which the ear cannot distinguish from the original.

A precise answer to the first question will not be attempted, but some important data must be mentioned. Qrdinarily it is assumed that the full range between about 100 and 3 000 c/s is necessary for satisfactory speech transmission. But Dudley Homer's ingenious speech-analysing and synthetizing machine, the Vocoder,[2.1] has achieved the transmission of intelligible speech by means of 11 channels of 25 c/s each, 275 c/s in all. This means a condensation, or compression, ratio of about 10.

Another datum is an estimate by Küpfmüller* of the product of time-interval by frequency-width required for the transmission of a single letter in telephony, and in the best system of telegraphy, as used in submarine cables. The ratio is about 40. This suggests that the Vocoder has probably almost reached the admissible limit of condensation.

The transmission which the ear would consider as indistinguishable from the original presents a more exactly defined and intrinsically simpler problem, as none of the higher functions of intelligence come into play which make distorted speech intelligible. G. W. Stewart in 1931 was the first to ask whether the limit of aural sensation is not given by an uncertainty relation, which he wrote in the form $\Delta t \Delta f = 1$, without, however, defining Δt and Δf precisely. He found the experimental material insufficient to decide the question, though he concluded that there was some evidence of agreement. New experimental results, which have become available since Stewart's note, and a more precise formulation of the question, will allow us to give a more definite answer.

In Section 5 of Part 1, methods were described for the expansion of an arbitrary signal into elementary signals, allocated to cells of a lattice. Fig. 2.1 is an example of a somewhat different method of analysis, in which the elementary areas have fixed shape but no fixed position, and are shifted so as to give a good representation with a minimum number of elementary signals. We now go a step further, and adjust not only the position but also the shape of the elementary areas to the signal, in such a way that it will be approximately represented by a minimum number of logons. This may be called "black-and-white" representation, and it is suggested that—within certain limits—it is rather close to our subjective interpretation of aural

* Quoted by Lüschen, Reference 1.6.

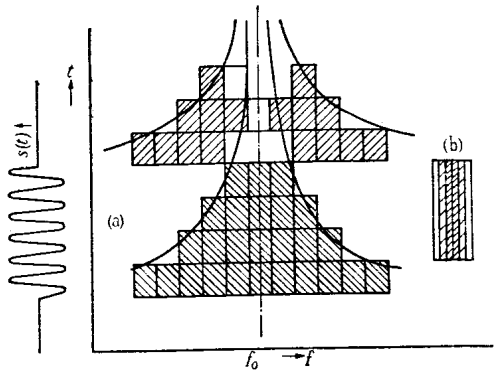

Fig. 2.1.—Sine wave of finite length.

(*a*) Response of a bank of resonators.
(*b*) Approximate response of the ear.

sensations. Fig. 2.1 illustrates this. If a sine wave of finite duration strikes a series of resonators, say a bank of reeds, with a time-constant which is a fraction of the duration, their response will be approximately as shown by (*a*). But, as the ear hardly hears the two noises or "clicks" at the beginning and end of the tone, its sensations can be better described by Fig. 2.1(*b*). We shall find later more evidence for what may be called the "adjustable time-constant" of the ear. It appears that, in general, the ear tends to simplify its sensations in a similar way to the eye, and the analogy becomes very evident in the two-dimensional representation.

It will be shown below that there is good evidence for what may be called a "threshold information sensitivity" of the ear, i.e. a certain minimum area in the information diagram, which must be exceeded if the ear is to appreciate more than one datum. The usefulness of this concept depends on how far this threshold value will be independent of the shape of the area. We must therefore test it by analysing experiments with tone signals of different duration.

It has been known for a long time (Mach, 1871) that a very short sinusoidal oscillation will be perceived as a noise, but beyond a certain minimum duration as a tone of ascertainable pitch. The most recent and most accurate experiments on this subject have been carried out by Bürck, Kotowski and Lichte.[2.2, 2.3] They found that both at 500 and 1 000 c/s the minimum duration after which the pitch could be correctly ascertained was about 10 millisec for the best observers. In a second series of experiments they doubled the intensity of the tone after a certain time, and measured the minimum duration necessary for hearing the step. For shorter intervals the stepped tone could not be distinguished from one which started with double intensity.

These two series of tests enable us to estimate the threshold area for very short durations. Fig. 2.2 explains the method for a frequency of 500 c/s. After 10 millisec the signal was just recognizable as a tone. But unless it lasted for at least 21 millisec, the ear was not ready to register a second datum, independent of and distinguishable from the first. We conclude, therefore, that the threshold area is determined by the frequency width of the first signal and the duration of the second. It is not necessary to approximate the chopped sine waves by elementary signals, as the ratio of the durations would remain the same. This was 2·1 for 500 c/s, and 3·0 for 1 000 c/s. We conclude that in these regions it takes 2·1 and 3 elementary areas respectively to convey more than one datum to the ear.

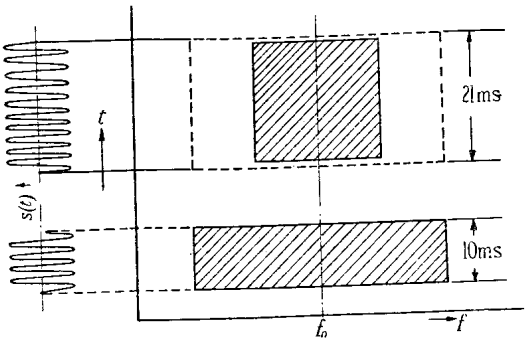

Fig. 2.2.—Experiments of Bürck, Kotowski and Lichte.

Let us now consider another series of tests, the experiments Shower and Biddulph on the pitch sensitivity of the ear[2,4]. these tests the frequency of a note was varied almost sinu-idally between a lower and an upper limit. The actual law variation was not exactly sinusoidal, as the top of the wave as flattened and rather difficult to analyse in an exact manner. the following approximate analysis we will replace it by nusoidal frequency modulation with a total swing δf, equal the maximum swing in the experiments. By this we are likely

It is well known[1,5] that the spectrum of a frequency-modulated wave with mean frequency f_0, total swing δf and modulation frequency f_m can be expressed by the following series

$$\operatorname{cis}\left(2\pi f_0 t + \frac{\delta f}{2f_m}\sin 2\pi f_m t\right) = \sum_{-\infty}^{\infty} J_n(\delta f/2f_m)\operatorname{cis} 2\pi(f_0 + nf_m)t \quad . \quad (2.1)$$

J_n is the Bessel function of nth order. The amplitudes of the side lines, spaced by the repetition frequency, are therefore proportional to $J_n(\delta f/2f_m)$. Their absolute values are shown at the bottom of Fig. 2.3 for four tests of Shower and Biddulph. On the other hand, the absolute amplitudes of the side lines in the spectrum of the two alternating sequences of elementary signals are given by the following formulae

$$I_n = \exp - \left(\frac{\pi}{\alpha}\right)^2 (nf)^2 \frac{\cosh}{\sinh}\left(\frac{\pi}{\alpha}\right)^2 nf_m f_s \quad . \quad . \quad (2.2)$$

The upper formula is valid for even, the lower for odd, orders n. With the help of equations (2.1) and (2.2) the available constants α and f_s have been fitted so as to represent exactly the ratio of the first two side lines to the central one. The result is shown in Fig. 2.3, in which the elementary signals are represented by

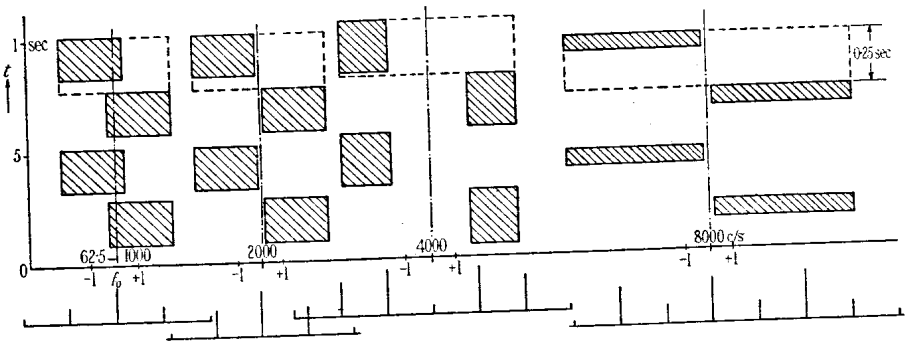

Fig. 2.3.—Experiments of Shower and Biddulph.

The frequency-modulated signals are replaced by two alternating series of elementary signals which produce very nearly the same spectrum.

o commit an error in the sense of overrating the ear sensitivity, ut this will give us a safe basis for estimating the chances of eceiving the ear. The modulation frequency in Shower and iddulph's experiments was 2 c/s, and the sensation level was ept constant at 40 db above the threshold of audibility. Their esults for the minimum variation δf at which the trill could ust be distinguished from a steady tone are as follows:—

f_0	62·5	125	250	500	1 000	2 000	4 000	8 000 c/s
$\delta f/f_0$	0·043	0·025	0·012	0·005	0·003	0·0023	0·00225	0·0037 c/s
δf	2·7	3·1	2·9	2·5	3·0	4·6	9·0	29·5 c/s

t will be seen that δf remains almost constant up to 1 000 c/s; rom about 1 000 c/s it is the ratio $\delta f/f_0$ which is nearly constant.

We now replace the signals used in these experiments by two periodic sequences of elementary signals with frequencies $f_0 \pm \frac{1}{2}f_s$, staggered in relation to one another, so that pulses with higher and lower frequency alternate at intervals of 0·25 sec. In order to approximate the actual signal as well as possible, we must use the available constants f_s and α (the "sharpness" of the elementary signals) so as to produce nearly the same spectrum.

their rectangles of area one-half. The agreement of the spectra even for higher orders n is very good up to 2 000 c/s, but less satisfactory at 4 000 and 8 000 c/s. But it would be useless to try better approximations, for example by adding one or two further sequences of elementary signals. More accurate information could be obtained only from experiments based on elementary signals. It may be hoped that such tests will be undertaken, especially as Roberts and Simmonds have suggested easy methods for producing such signals.

For a first orientation the results derived from the tests of Shower and Biddulph appear quite satisfactory. It can be seen from Fig. 2.3 how rectangles can be constructed in the information diagram which mark the limit at which the ear can just begin to appreciate a second datum. In this case the meaning of the threshold is that the trill can just be distinguished from a steady tone. Measured in units of elementary areas of one-half, their values are as follows:—

Frequency	62·5–1 000	2 000	4 000	8 000 c/s
(Threshold area)/0·5	2·34	2·88	3·92	6·9

The reciprocals of these figures can be considered as performance figures of the ear as compared with an ideal instrument. In fact, the performance figure of an ideal instrument would be unity, as it would begin to appreciate a second datum as soon as the minimum information area of one-half was exceeded by

an amount, however small. The performance figure derived from the experiments of Shower and Biddulph between 62 and 8 000 c/s is shown in Fig. 2.4. The diagram also contains two

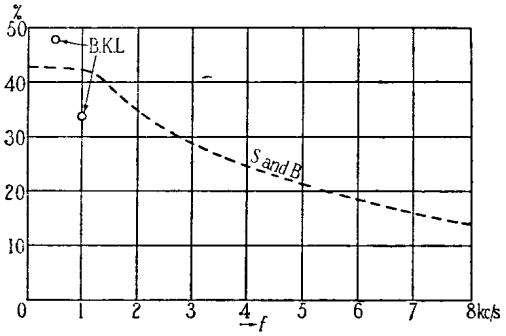

Fig. 2.4.—Performance figure of the ear.
B.K.L.—Bürck, Kotowski and Lichte. S. and B.—Shower and Biddulph.

points derived from the experiments of Bürck, Kotowski and Lichte, which fit in as well as can be expected. It is very remarkable that up to about 1 000 c/s the performance figure is almost 50%, which is the ideal for an instrument like the ear which cannot distinguish the phase of oscillations, i.e. rejects one-half of the data. At higher frequencies, however, the efficiency is much less.

The good fit of the figures obtained from the experiments of Bürck, Kotowski and Lichte, which were carried out with durations of 10–20 millisec, with those of Shower and Biddulph, in which the threshold area measured 250 millisec in the time direction, indicates two facts. One is that, at least up to about 1 000 c/s, and for durations at least in the limits 20–250 millisec, the threshold information area is a characteristic of the ear. Evidently the performance figure must go to zero both for extremely short and for extremely long elementary signals, but within these wide and very important limits it appears to have an almost constant value.

The other fact which arises from the first is that the ear appears to have a time-constant *adjustable at least between* 20 *and* 250 *millisec*, and that the ear adjusts it to the content of the information which it receives. But there can be little doubt that, whatever resonators there are in the ear, they are very strongly damped, and that their decay time is of the order of 20 millisec or rather less. This is borne out by the experiments of Wegel and Lane on the amplitudes of the oscillations of the basilar membrane in the inner ear.* A pure tone excites such a broad region to oscillations that R. S. Hunt,[2.7] who has recently made a thorough investigation of Wegel and Lane's data, infers from them a decay by 1 bel in only 2 cycles, i.e. in only 2 millisec at 1 000 c/s! Though this estimate might be too low, there can be no doubt that the decay time of the ear resonators cannot substantially exceed 10 millisec, and it is impossible to imagine that they would keep on vibrating for as much as a quarter of a second. Hence, even if the duration of a pure tone is considerably prolonged beyond the 10 millisec approximately required for pitch perception, the ear resonators will still display the same broad distribution of amplitude. This is illustrated in Fig. 2.5. In order to explain the high pitch sensitivity of the ear, as shown, for example, by the experiments of Shower and Biddulph, *it is therefore necessary to assume a second mechanism which locates the centre of the resonance region with a precision increasing with the duration of the stimulus.* Its effect is indicated in Fig. 2.5. The second mechanism acts as if there were

* Reference 2.6. Also HARVEY FLETCHER: "Speech and Hearing" (Macmillan, 1929), p. 184.

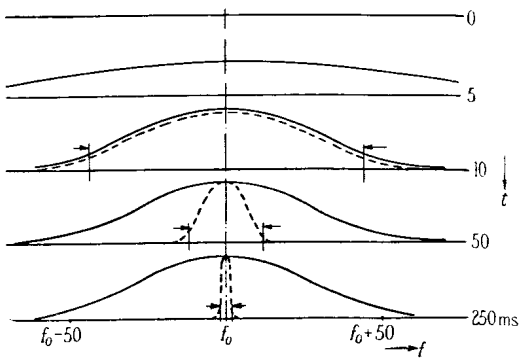

Fig. 2.5.—The two mechanisms of pitch determination.

a second resonance curve, of a non-mechanical nature, which after about 10 millisec detaches itself from the mechanical resonance curve and continues to contract until, after about 250 millisec, it covers only a few cycles per second.

Both mechanisms are essential for our hearing. The first by itself would probably enable us to understand speech, but only the second makes it possible to appreciate music. One might be tempted to locate this second function in the brain, but mechanisms of nerve conduction can be imagined which might achieve the same effect. Perhaps the simplest assumption is that the conduction of stimuli in adjacent nerve fibres is to some extent unstable, so that in an adjacent pair the more strongly stimulated fibre will gradually suppress the conduction in its less excited neighbour. The available evidence would not justify the suggestion that this is the actual mechanism; the intention is only to show that what manifests itself as the "adjustable time-constant" of the ear is not necessarily a consequence of some higher function of intelligence.

In the light of these results we can now approach the question of a condensed transmission which entirely deceives the ear. The performance figure as shown in Fig. 2.4 appears to indicate that considerable economy might be possible, especially in the range of higher frequencies. This is brought into evidence even more clearly in Fig. 2.6, which contains the integrals over frequency of the performance figures for the ear and for an ideal instrument. Between zero and 8 000 c/s, for instance, the maximum number of data which the ear can appreciate is only about one-quarter of the data which can be transmitted in a band of 8 000 c/s. It is even likely that further investigations might substantially reduce this figure. It may be remembered that the experiments on which Fig. 2.6 is based have all been carried out with sharp or rather angular waveforms; it is not

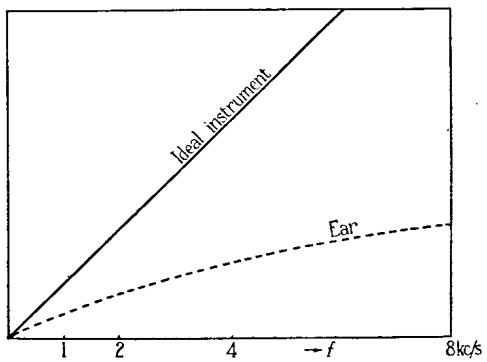

Fig. 2.6.—Utilization of information area.

likely that the threshold was essentially determined by logons outside the area considered in our analysis. But it must also be remembered that the "adjustable time-constant" makes it very difficult to deceive the ear entirely. It will be shown in Part 3 that methods are possible which could deceive any non-ideal instrument with *fixed* time-constant. But the ear has the remarkable property that it can submit the material presented to it not only to one test, but, as it were, to several. Ultimately only direct tests can decide whether any such scheme will work satisfactorily.

(2) REFERENCES

(2.1) HOMER, DUDLEY: "Re-making Speech," *Journal of the Acoustical Society of America*, 1939, **11**, p. 169.

(2.2) BÜRCK, W., KOTOWSKI, P., and LICHTE, H.: "Development of Pitch Sensations," *Elektrische Nachrichten-Technik*, 1935, **12**, p. 326.

(2.3) BÜRCK, W., KOTOWSKI, P., and LICHTE, H.: "Audibility of Delays," *ibid.*, 1935, **12**, p. 355.

(2.4) SHOWER, E. G., and BIDDULPH, R.: "Differential Pitch Sensitivity of the Ear," *Journal of the Acoustical Society of America*, 1931, **3**, p. 275.

(2.5) BLOCH, A.: "Modulation Theory," *Journal I.E.E.*, 1944, **91**, Part III, p. 31.

(2.6) WEGEL, R. L., and LANE, C. E.: "Auditory Masking and the Dynamics of the Inner Ear," *Physical Review*, 1924, **23**, p. 266.

(2.7) HUNT, R. S.: "Damping and Selectivity of the Inner Ear," *Journal of the Acoustical Society of America*, 1942, **14**, p. 50.

Part 3. FREQUENCY COMPRESSION AND EXPANSION

SUMMARY

It is suggested that it may be possible to transmit speech and music in much narrower wavebands than was hitherto thought necessary, not by clipping the ends of the waveband, but by condensing the information. Two possibilities of more economical transmission are discussed. Both have in common that the original waveband is compressed in transmission and re-expanded to the original width in reception. In the first or "kinematical" method a temporary or permanent record is scanned by moving slits or their equivalents, which replace one another in continuous succession before a "window." Mathematical analysis is simplest if the transmission of the window is graded according to a probability function. A simple harmonic oscillation is reproduced as a group of spectral lines with frequencies which have an approximately constant ratio to the original frequency. The average departure from the law of proportional conversion is in inverse ratio to the time interval in which the record passes before the window. Experiments carried out with simple apparatus indicate that speech can be compressed into a frequency band of 800 or even 500 c/s without losing much of its intelligibility. There are various possibilities for utilizing frequency compression in telephony by means of the "kinematical" method.

In a second method the compression and expansion are carried out electrically, without mechanical motion. This method consists essentially in using non-sinusoidal carriers, such as repeated probability pulses, and local oscillators producing waves of the same type. It is shown that one variety of the electrical method is mathematically equivalent to the kinematical method of frequency conversion.

(1) INTRODUCTION

High-fidelity reproduction of speech or music by current methods requires a waveband of about 8 000 c/s. It has been shown in Part 1 that this band-width is sufficient for the transmission of 16 000 exact and independent numerical data per second. This high figure naturally suggests the question whether all of this is really needed for the human ear to create an illusion of perfection. In Part 2 it was shown that, even in the frequency range in which it is most sensitive, the human ear can appreciate only one datum in two at the best, and not more than one in four as an average over the whole a.f. range. Moreover, it must be taken into consideration that, in the experiments which gave these limits of aural discrimination, attention was fixed on a very simple phenomenon. It appears highly probable that for complex sound patterns the discriminating power of the ear is very much less. This evidence suggests that methods of transmitting and reproducing sound may be found which are much more economical than those used at present, in which the original signal shape is carefully conserved through all the links of transmission or reproduction. In an economical method the information content must be condensed to a minimum before transmission or before recording, and the reconstruction need not take place before some stage in the receiver or reproducer. There is no need for the signal to be intelligible at any intermediate stage. Economical methods must therefore comprise some stage of "condensing" or "coding" and some stage of "expanding" or "decoding."

Dudley Homer's ingenious Vocoder,[3.1] which transmits intelligible speech through 11 channels of only 25 c/s each, is a well-known example of such a system. It operates with a method of spectral analysis and synthesis. The spectrum of speech is roughly analysed into 10 bands of 250 c/s each, and the aggregate intensity in each band is transmitted through a separate channel of 25 c/s. The transmitted intensity is used for modulating a buzzer at the receiving end, which roughly reproduces the original spectrum. The eleventh channel is used for transmitting the "pitch," which is, broadly speaking, the frequency of the vocal cords. The Vocoder in its present form has probably very nearly reached the limit of tolerable compression.

In this Part new methods will be discussed in which the coding of the message consists essentially in compression, i.e. in a proportional reduction of the original frequencies, and the decoding in expansion to the original range. It is evident that neither compression nor expansion can be exact if economy is to be effected. If, for instance, *all* frequencies were exactly halved, this would mean that it would take twice the time for transmitting the same message and there would be no saving. Compression and expansion—in general, "conversion"—of frequencies must be rather understood in an approximate sense. There will be unavoidable departures from the simple linear law, and hence there will be some unavoidable distortion. But it appears that these can be kept within tolerable limits while still effecting appreciable waveband economy.

Two compression-expansion systems will be described. The first, which operates with mechanically moving parts, will be called the "kinematical" method,* while the second does not require mechanical motion and will be called the "electrical" method. So far, experiments have been carried out only with the kinematical method, and for this reason it will occupy most of this Part.

(2) THE KINEMATICAL METHOD OF FREQUENCY CONVERSION

It will be convenient to explain this method by means of a particular example before generalizing the underlying principle. Assume that the message to be condensed or expanded is recorded

* British Patent Application No. 24624/44.

as a sound track on a film. For simplicity, assume that the original signal is a simple harmonic oscillation, that is to say a frequency f_0—to be called the "original frequency"—is produced if the record moves with standard speed v past a stationary slit. Imagine now that the slit itself is moving with some speed u, so that its speed relative to the film is $v - u$. The photocell behind the film now collects fluctuations of light of frequency

$$f_1 = \frac{v-u}{v} f_0 \quad . \quad . \quad . \quad . \quad . \quad (3.1)$$

This means that all frequencies in the record are converted in a constant ratio $(v - u)/v$. There is evidently no gain, as it would take the moving slit $v/(v - u)$ times longer to explore a certain length of the film than if it were stationary. But let us now imagine that the film moves across a fixed window, so that the moving slit is effective only during the time in which it traverses the window. In order to get a continuous record let a second slit appear at or before the instant at which the first slit moves out of the window, after which a third slit would appear, and so on. The device is still not practicable, as evidently every slit would produce a loud crack at the instant at which it appeared before the window and when it left it. But now assume that the window has continuously graded transmission, full in the middle and fading out at both sides to. total opacity. In this arrangement the slits are faded in and out gradually, so that abrupt cracks can be avoided. This is the prototype of a kinematical frequency convertor, schematically illustrated in Fig. 3.1, which will be investigated below. Though

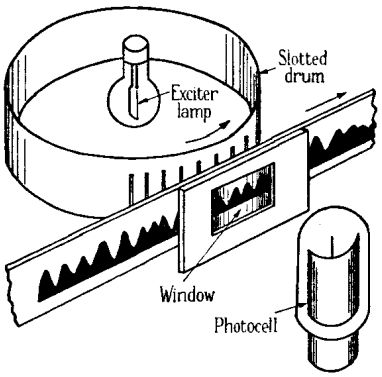

Fig. 3.1.—Frequency convertor with sound film.

the nomenclature will be taken from this special example, the mathematical theory can be transferred bodily to any other realization of the same principle.

In Fig. 3.1 the film is supposed to move in close contact with the slotted drum, but at different speed. A photocell collects the sum of the light transmitted by the individual slits and by the window. To obtain its response we must first write down the contribution of one slit and sum over the slits. All slits will be assumed to have negligible width. For simplicity let us measure all distances x from the middle of the window and all times t from the instant in which a slit, to be called the "zero"-th slit, passes through $x = 0$. The other slits will be distinguished by suffixes k, which increase in the direction in which the film is moving. Their position at the time t will be called x_k. The nomenclature is explained in Fig. 3.2.

Let v be the speed of the film, while the velocity of the slits will be called

$$u = (1 - \kappa)v \quad . \quad . \quad . \quad . \quad . \quad (3.2)$$

The reason for this notation is that eqn. (3.1) now simplifies to $f_1 = \kappa f_0$, i.e. κ has the meaning of a frequency-conversion ratio.

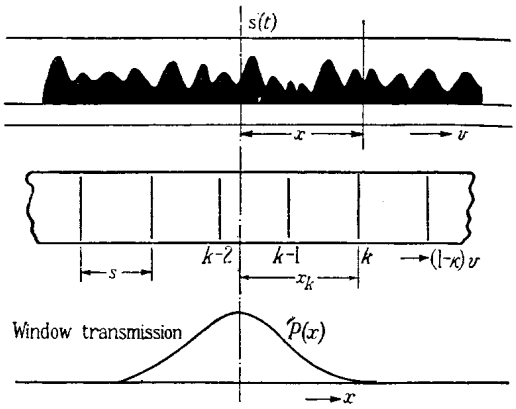

Fig. 3.2.—Explanation of notations.

If the spacing of two slits is s the position of the kth slit at time t is given by

$$x_k = (1 - \kappa)vt + ks \quad . \quad . \quad . \quad . \quad (3.3)$$

The record will be characterized by the signal $s_1(t)$ which it would produce if it were scanned in the ordinary way by a stationary slit in the position $x = 0$. Hence, if the window were fully transparent, the signal due to the kth slit at time t would be

$$s_1(t - x_k/v) = s_1(\kappa t - ks/v) \quad . \quad . \quad . \quad (3.4)$$

The total reproduced signal, i.e. the light sum collected by the photocell, is obtained from this by multiplying by the transmission coefficient $P(x)$ of the window and summing over k.

In all the following calculations we will assume that this transmission follows a probability law. This law has unique properties in Fourier analysis and will immensely simplify our investigations. Other laws which appear equally simple *a priori*, and which may have even some practical advantages—such as triangular or trapezoidal windows—lead to expressions which are too complicated for anything but numerical discussion. Hence we assume

$$P(x) = \exp - (x/Ns)^2 \quad . \quad . \quad . \quad . \quad (3.5)$$

N is a number, to be called the "slit number," which characterizes the reproduction process. It is the number of slits in the length over which the transmission of the window falls from unity to $1/e$. The total length of the window in which the transmission exceeds 1% is $4 \cdot 3Ns$. Thus we can say broadly that the total number of slits simultaneously before the window is $4 \cdot 3N$.

The reproduced signal, i.e. the total light collected by the photocell, at time t, is

$$s(t) = \sum_{-\infty}^{\infty} k \exp - (x_k/Ns)^2 s_1(t - x_k/v) \quad . \quad . \quad (3.6)$$

This, in combination with eqn. 3.3, is a complete description of the operation of the frequency convertor. It will now be illustrated in the special case in which s_1 is a simple harmonic oscillation

$$s_1(t) = e^{2\pi j f_0 t} = \text{cis } 2\pi f_0 t \quad . \quad . \quad . \quad . \quad (3.7)$$

The complex form will be used, with the understanding that the real part constitutes the physical signal. Simple harmonic oscillations are suitable for the analysis, as their spectrum will consist of a few lines. But it may be mentioned that analysis in terms of the elementary signals discussed in Part 1 (harmonic oscillations with probability envelope) can be carried out almost equally simply, as the reproduction of an elementary signal

onsists also in the sum of a few elementary signals. This s carried out in Appendix 7.1, but in the text only the more miliar method of Fourier analysis will be employed.

Substituting the signal (3.7) in eqn. (3.6) and using eqn. (3.3), e obtain

$$(t) = \sum_{-\infty}^{\infty} k \exp - [(1-\kappa)vt + ks]^2/(Ns)^2 \operatorname{cis} 2\pi f_0(\kappa t - ks/v) \quad . \quad . \quad . \quad (3.8)$$

he meaning of this somewhat complicated expression is explained in Fig. 3.3. Each slit, as it passes before the window,

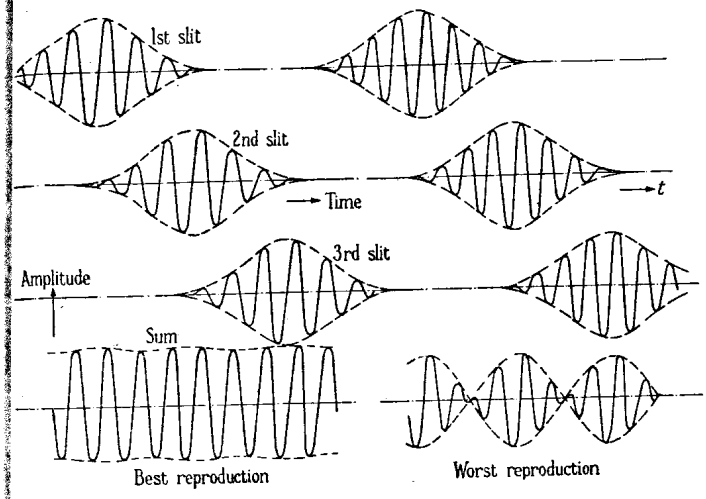

Fig. 3.3.—The contributions of individual slits and the resulting light output.

transforms the sine wave into an elementary signal. By adding up the contributions of the individual slits we obtain for some frequencies a very nearly faithful reproduction, i.e. an almost pure tone but of different frequency from the original. For other frequencies we obtain strong beats.

A more convenient and complete description of the frequency conversion process is obtained by Fourier analysis. It will now be convenient to measure distances in time intervals, and to introduce, instead of the slit spacing s, the time interval τ between the passage of two consecutive slits before a fixed point

$$\tau = s/(1-\kappa)v \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.9)$$

With this notation the Fourier transform, i.e. the spectrum of the signal $s(t)$, becomes, by known rules,*

$$S(f) = \sqrt{(\pi)}N\tau \exp - (\pi N\tau)^2(f - \kappa f_0)^2 \sum_{-\infty}^{\infty} k \operatorname{cis} 2\pi k\tau(f - f_0) \quad . \quad . \quad . \quad (3.10a)$$

This expression allows of a simple interpretation. The second factor

$$\Sigma k \operatorname{cis} 2\pi k\tau(f - f_0)$$

is the sum of an infinite number of complex vectors of unit length, with an angle of $2\pi\tau(f - f_0)$ between two consecutive vectors. This series, though not convergent, is summable,† and its sum is zero for all values of f except those for which

$$\tau(f - f_0) = \text{an integer} \quad . \quad . \quad . \quad . \quad (3.11)$$

Physically this means that the spectrum consists of sharp *lines* which differ from one another by multiples of $1/\tau$. In other words, the spectrum consists of all combination notes of the original frequency f_0 with the repetition frequency $1/\tau$.

The absolute sharpness of the spectral lines is a consequence of the assumption that the slits pass before the window at mathematically exact equal intervals. In each spectral line $S(f)$ is a "delta function," i.e. a sharp peak of infinite height but finite area. But as in what follows we shall always have to deal with line spectra, it is more convenient to re-interpret $S(f)$ as a function which is zero except at certain discrete values of f, where it assumes finite values, proportional to the amplitude of the spectral lines. In the same sense, we write the second factor of eqn. (3.10a) somewhat more simply as

$$\Sigma \operatorname{cis} 2\pi k\tau(f - f_0) = \Sigma\delta(f - f_0 - k/\tau) \quad . \quad (3.12)$$

and interpret this as a "selecting factor" which has zero value everywhere except for those values of f which fulfil condition

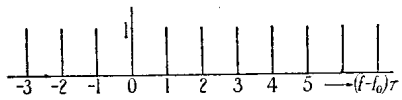

Fig. 3.4.—The selection factor.

(3.11), where it assumes the value unity (see Fig. 3.4). Thus we write eqn. (3.10)

$$S(f) = \exp - (\pi N\tau)^2(f - \kappa f_0)^2 \Sigma\delta(f - f_0 - k/\tau) \quad (3.10b)$$

The first factor is independent of the summation index k and represents an attenuation function of probability shape, which has its maximum at

$$f = \kappa f_0$$

i.e. at frequencies which have been converted in the correct ratio κ. The sharpness of this attenuation curve is reciprocal to the sharpness of the transmission curve of the window, measured in units of time. Thus, if the window were infinitely broad we should obtain exact conversion of all frequencies. But this would have the disadvantage that short signals occurring at some definite time would be reproduced at completely indefinite times (with an infinite number of repetitions). Conversely, if the window were infinitely short the attenuation would be zero and the frequencies scattered evenly over all possible values defined by eqn. (3.11). Thus we meet again the fundamental uncertainty relation between frequency and time (or rather, "epoch") which was discussed in some detail in Part 1. It follows immediately from previously obtained results that the probability window is ideal in the sense that it produces the smallest possible product of the linked uncertainties of frequency and epoch, as defined in Part 1.* Nevertheless the probability window is not necessarily the best from a practical point of view. Some possible improvements will be discussed later.

Equation (3.10a) or (3.10b) allows also a simple graphical interpretation, which is explained in Fig. 3.5 in a numerical example. The original frequency f_0 is the ordinate; the reproduced frequencies f are the abscissae. Both are conveniently measured in units $1/\tau$, i.e. as multiples of the repetition frequency. All points (f, f_0) which satisfy condition (3.11) lie on lines at 45° to the two axes, and intersect the horizontal axis at integral values of $f\tau$. The attenuation curve

$$\exp - (\pi N\tau)^2(f - \kappa f_0)^2$$

need be drawn only once, though in the Figure it has been done twice in order to give a clear visual impression of the way in

* The calculation is carried out in Appendix 7.1. The rules for Fourier transforms may be found in Reference 3.2, and, particularly for signals of the type (3.8), in References 3.3 and 3.4.

† Summation is to be understood in the sense of Cesàro. Cf. WHITTAKER-WATSON: "Modern Analysis," 1935, p. 155.

* In Part 1 uncertainties were defined, apart from a constant factor, as the r.m.s. deviations from the average value.

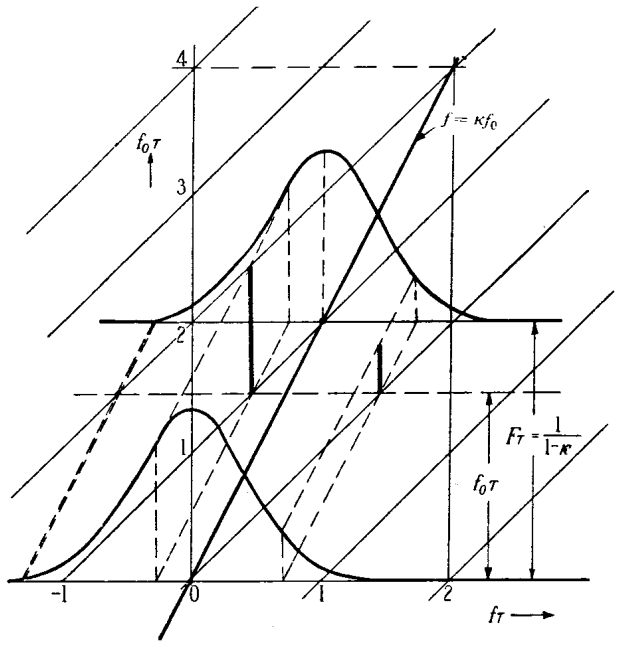

Fig. 3.5.—Diagram of frequency compression.
$N = \frac{1}{2}$; $\kappa = \frac{1}{2}$

which the amplitude is distributed over the (f, f_0) plane. The spectral lines are given by the height of the attenuation curve above the points in which a line f_0 = constant crosses the lines $(f - f_0)\tau$ = integer, as shown in an example.

This Figure shows the action of the frequency convertor at one glance. The correctly converted frequency $f = \kappa f_0$ appears in the reproduction only where a line $(f - f_0)\tau =$ *an integer* intersects the line $f = \kappa f_0$. This condition is always fulfilled for $f_0 = 0$, and for all frequencies which are multiples of

$$F = 1/\tau(1 - \kappa) \quad . \quad . \quad . \quad . \quad (3.13)$$

This may be called the length of the "cycle of reproduction", as the quality of reproduction varies cyclically with this period. If f_0 is an integral multiple of F the reproduction can be made almost perfect, as the side lines can be almost entirely suppressed if the slit number N is made sufficiently large. As can be seen in Fig. 3.6, $N = 1$ is sufficient to achieve this. But this improvement in the reproduction of certain tones is made at the cost of others. If N is large, not only the side lines but almost all amplitudes near the middle of a cycle of reproduction will be suppressed, i.e. certain notes will be missing. It is evident that a compromise must be struck between the purity of reproduction at the ends and at the middle of every cycle length F.

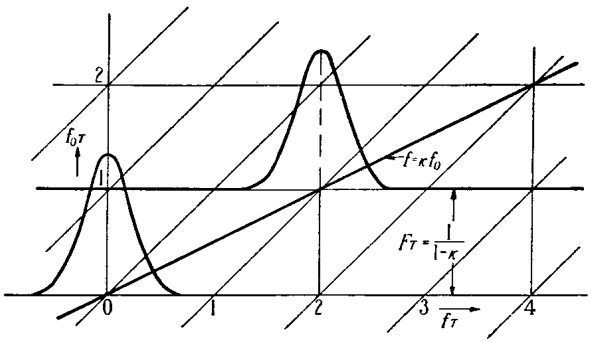

Fig. 3.6.—Diagram of frequency expansion.
$N = 1$; $\kappa = 2$

The effect of the slit number N on the quality of the reproduction is shown in Fig. 3.7. Three cases are illustrated, all for an expansion ratio $\kappa = 2$, and for $N = 0{\cdot}25$, $0{\cdot}5$ and 1. It may be recalled that the average number of slits before that part of the window in which the transmission exceeds 1% is $4{\cdot}3N$. In each case a full cycle of reproduction is shown, with ten equally-spaced original frequencies.

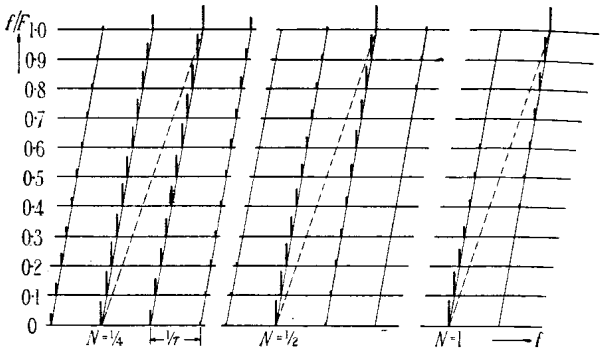

Fig. 3.7.—Influence of slit number on quality of reproduction.
$\kappa = 2$

At the left, $N = 0{\cdot}25$, the Figure shows the effect of too small slit numbers. The reproduction is very "noisy," no frequency being reproduced as an approximately pure tone. There is little difference between the spectra of frequencies near the middle or ends of the cycle; they are all of uniformly poor quality.

At the right, $N = 1$, the Figure shows the effect of a too large number of slits (cf. Fig. 3.6). The frequencies at the ends of the cycle are reproduced nearly ideally, as practically pure tones, but the frequencies in the middle of the cycle are almost entirely missing in the reproduction.

The best compromise appears to be $N = 0{\cdot}5$, shown in the middle of the Figure. The end frequencies are still reproduced as almost pure tones, and the intensity falls off little towards the middle of the cycle. (The intensity is obtained by squaring the amplitudes shown in the Figure and finding their sum. It falls, in the middle of the cycle, to $0{\cdot}56$ of the maximum.) The spectra of the intermediate tones consists mostly of only two lines; i.e. these will be vibrating tones, vibrating with a beat frequency of $1/\tau$. The beats are strongest in the middle, where the two spectral components have equal amplitudes.

It might appear at first sight that, by reducing the beat frequency below any limit, the reproduction could be made perfect to any desired degree. But there are limits to the increase of τ. As N is fixed more or less at $0{\cdot}4$–$0{\cdot}5$, τ can be increased only by making the window longer. The length of the window may be now defined as the length of time T in which a point of the film passes through the part of the window in which the transmission exceeds 1%. This is

$$T = 4{\cdot}3Ns/v = 4{\cdot}3N\tau(1 - \kappa) \quad . \quad . \quad . \quad (3.14)$$

Hence, for the optimum, $N = 0{\cdot}5$

$$\tau = 0{\cdot}47T/(1 - \kappa)$$

If the time T is too long, the time resolution in the reproduction will be poor. Determining the best compromise between time resolution and frequency reproduction is a matter for experiment. On general grounds one would expect that the window length T must be kept below the limit at which the ear could begin to separate the contribution of the two or more slits which are simultaneously before the window. For speech the optimum of T is probably about 100 millisec; for music

probably about 250 millisec.* With $\kappa = 2$ this would make the beat frequency 21 c/s for speech, and about 8 c/s for music.

It may be noted by comparing eqns. (3.13) and (3.14) that a simple reciprocity relation obtains between the cycle length F and the window length T, of the form

$$FT = 4{\cdot}3\,N \quad . \quad . \quad . \quad . \quad . \quad (3.15)$$

With optimum choice of N the value of this is about 2. Thus for a window length of 100 millisec the optimally-reproduced frequencies are spaced by about 20 c/s; for $T = 250$ millisec by about 8 c/s. In the reproduction the spacing will be κ times more.

The theory so far discussed was based on the assumption of a probability window, which not only has the advantage of mathematical simplicity, but also gives the most advantageous reciprocity relation between time resolution and frequency resolution. But the optimum number N was found to be only about 0·5, which means that there are on the average only about two slits before the window. This might produce a slight but noticeable noise in the optimally-reproduced frequencies, in particular for $f_0 = 0$ (background). Hence it may be advantageous to depart somewhat from the probability shape in order to suppress the noise. Fig. 3.8 shows window transmission

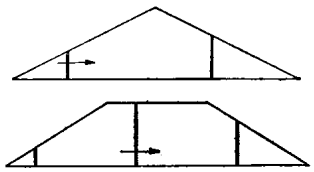

Fig. 3.8.—Window shapes with zero noise for two and three slits.

shapes for two and three slits which produce no noise when passing before an even background, as the light sum is constant in any position. Though the mathematical theory of such windows is very much more complicated, it is not to be expected that they would produce essentially different results from probability windows of comparable effective width.

(3) DISTORTIONS RESULTING FROM THE COMPRESSION-EXPANSION CYCLE

A full cycle of condensed transmission of the kind discussed consists in compression by a factor $\kappa < 1$, followed by expansion in the ratio $1/\kappa$. In general, if two conversion processes are applied in succession to a simple harmonic oscillation of frequency f_0, the resulting spectrum is given by

$$S(f) = \sum_{-\infty}^{\infty}{}_k \sum_{-\infty}^{\infty}{}_m \exp - \{(\pi N_1\tau_1)^2[f_0(1-\kappa_1) + k/\tau_1]^2 + (\pi N_2\tau_2)^2[f - \kappa_2(f_0 + k/\tau_1)]^2\} \delta(f - f_0 - k/\tau_1 - m/\tau_2) \quad . \quad (3.16)$$

The derivation is given in Appendix 7.2. All data N, τ, κ of the first conversion have been given a suffix 1, those of the second conversion the suffix 2. k and m are summation indexes which run over all integral values.

The second factor is again a selection operator, which is zero for all values of f with the exception of those where

$$f = f_0 + k/\tau_1 + m/\tau_2 \quad . \quad . \quad . \quad . \quad (3.17)$$

This means that only those frequencies will appear in the spectrum which correspond to combination tones of the original frequency with one or the other or both of the repetition frequencies $1/\tau_1$ and $1/\tau_2$. These form, in general, a double series, which in the particularly important practical examples to be considered reduces to a simple series.

In what follows we will consider only pairs of conversion processes which, on the average, reconstruct the original frequencies. The condition for this is

$$\kappa_1\kappa_2 = \pm 1 \quad . \quad . \quad . \quad . \quad . \quad (3.18)$$

The ambiguity of sign expresses the fact that positive and negative frequencies are equivalent. But only the plus sign will be considered, and it will be assumed, moreover, that both κ_1 and κ_2 are positive. Negative conversion ratios are less advantageous, as for a given window length they require higher repetition frequencies [Eqn. (3.14)]. The whole compression-expansion cycle will be characterized by the compression ratio κ, $0 < \kappa < 1$, and the expansion ratio will be assumed as $1/\kappa$.

To simplify the discussion it will be assumed that the window length T is the same in the transmitter and in the receiver. This corresponds to optimum conditions, as it will evidently be best to operate at both ends with the longest permissible T, which may have different values for speech and for music. This means

$$T/4{\cdot}3 = N_1\tau_1(1-\kappa) = N_2\tau_2(1-\kappa)/\kappa \quad . \quad (3.19)$$

or

$$\tau_1/\tau_2 = N_2/\kappa N_1 \quad . \quad . \quad . \quad . \quad (3.20)$$

A second simplifying assumption will be

$$\tau_1/\tau_2 = p = \text{an integer} \quad . \quad . \quad . \quad . \quad (3.21)$$

This again is an assumption which is fulfilled in the most important practical cases. In the interest of optimum transmission the slit number will be used, in both the transmitter and the receiver, which gives the best results in simple conversion ($N = 0{\cdot}4 - 0{\cdot}65$), and if κ is the reciprocal of an integer $\frac{1}{2}$, $\frac{1}{3}$, $\frac{1}{4}$ . . . the condition (3.21) will be fulfilled.

Mathematically this has the advantage that the double series of frequencies in the reproduced spectrum

$$k/\tau_1 + m/\tau_2$$

now becomes a simple series, with period $1/\tau_1$, as in simple conversion. We write

$$k/\tau_1 + m/\tau_2 = (k + pm)/\tau_1 = n/\tau_1 \quad . \quad . \quad (3.22)$$

so that the spectral lines are now characterized by the single suffix n, which can be called the "order number." As $S(f)$ will be different from zero for integer values of n, and for these only, we can now omit the selection operator δ in eqn. (3.16), on the understanding that we consider only integral values of n. Eqn. (3.17) now becomes

$$f = f_0 + n/\tau_1 \quad . \quad . \quad . \quad . \quad (3.23)$$

Eliminating f by means of eqn. (3.23) and introducing the assumptions (3.20) and (3.21) into eqn. (3.16) we now obtain the simplified formula

$$S(f_0,n) = \sum k \exp - (\pi N_1)^2\{[f_0\tau_1(1-\kappa) + k]^2 + [f_0\tau_1(1-\kappa) + k - n\kappa]^2\} \quad . \quad (3.24)$$

In this sum, however, not all integral values of k are included, but only those which are compatible with the given value of the order n. If there are two values k_0, m_0 which satisfy the equation

$$n = k_0 + m_0 p$$

all other values which satisfy it must be of the form

$$k = k_0 + \nu p \qquad m = m_0 - \nu$$

* *Note added 15th June*, 1946.—Recent experiments with perfected apparatus have confirmed the expectations as regards the optimum value of T for speech, but not for music.

where ν is any integer. It will therefore be convenient to introduce ν as the summation index, and to make the convention that k_0 is the smallest positive number in the sequence of k's. In other words, let k be the residue of n divided by p, or, in the notation of the elementary theory of numbers,

$$n \equiv k_0 (\mathrm{mod}\, p) \quad . \quad . \quad . \quad . \quad . \quad (3.25)$$

As a further simplification we note that $S(f_0,n)$ is a periodic function of f_0, with a cycle length

$$F = \frac{p}{\tau_1(1-\kappa)} = \frac{1}{\tau_2(1-\kappa)} \quad . \quad . \quad . \quad (3.26)$$

and obtain

$$S(f_0,n) = \sum_{-\infty}^{\infty} \nu \exp - \left(\frac{\pi N_2}{\kappa}\right)^2 \left[\left(\frac{f_0}{F} + \frac{k_0}{p} + \nu\right)^2 + \left(\frac{f_0}{F} + \frac{k_0}{p} + \nu - \frac{n\kappa}{p}\right)^2\right] \quad . \quad (3.27)$$

By rearranging the terms in the exponent this can be written, finally,

$$S(f_0,n) = \exp - \tfrac{1}{2}\left(\frac{\pi N_2}{p}\right)^2 n^2 \sum_{-\infty}^{\infty} \nu \exp - 2\left(\frac{\pi N_2}{\kappa}\right)^2 \left(\frac{f_0}{F} + \frac{k_0}{p} + \nu + \frac{n\kappa}{2p}\right)^2 \quad . \quad (3.28)$$

This formula lends itself well to graphical interpretation. In Fig. 3.9 the ordinate is again the original frequency f_0, measured

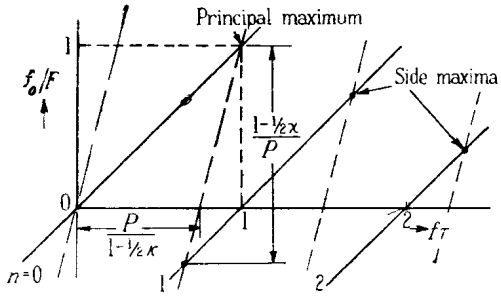

Fig. 3.9.—Explanation of frequency-conversion diagrams.

in units F, and the abscissae are the reproduced frequencies f. A line at 45° through the origin represents the correct reconversion law, $f = f_0$. This is the line of zero order, $n = 0$. Parallel to this we draw lines through all multiples of $1/\tau_1$ on the f-axis. These are the loci of all non-zero intensities. If we imagine the amplitude $S(f_0,n)$ as a surface above the (f_0,f) plane, this surface consists of a number of profiled planes, projecting above the lines n = constant.

On the line $n = 0$ we have evidently a maximum of $S(f_0,0)$ for every integral value of f_0/F. These may be called the "principal maxima." At the side lines of higher order there will also be maxima, but because of the probability function in front of the sum these will be smaller. We can draw lines connecting these maxima of different orders. We obtain a set of straight lines connecting the points where

$$f_0/F + k_0/p - n\kappa/2p = \text{an integer} \quad . \quad . \quad . \quad (3.29)$$

If the order n increases by one, by eqn. (3.24) k_0 also increases by unity, and f_0/F changes by

$$-(1 - \tfrac{1}{2}\kappa)/p \quad . \quad . \quad . \quad . \quad (3.30)$$

as shown in Fig. 3.9. It can be shown from the geometry of Fig. 3.9 that these lines will intersect the horizontal axis at multiples of

$$p/(1 - \tfrac{1}{2}\kappa) \quad . \quad . \quad . \quad . \quad . \quad . \quad (3.31)$$

These lines, together with the lines n = integer, form a network with intersections at every maximum of the spectral function $S(f_0,n)$.

Along each line n = constant, the spectral amplitude is the same function of f_0/F, apart from the shift (3.30) and the factor

$$\exp\left[-\tfrac{1}{2}(\pi N_2/p)^2 n^2\right]$$

which varies with n but is a constant along each line. Thus it is sufficient to compute the amplitude function once, for $n = 0$, where the shift is zero and the exponential factor unity. This function is

$$S(f_0,0) = \Sigma\nu \exp - 2(\pi N_2/\kappa)^2 (f_0/F + \nu)^2 \quad . \quad (3.32)$$

This, as a function of f_0/F, is the sum of probability functions, recurring at unit distance. It is shown in Appendix 7.3 that it can be reduced to a recognized transcendental function of analysis, the theta function θ_{00}. Fig. 3.10 shows this function

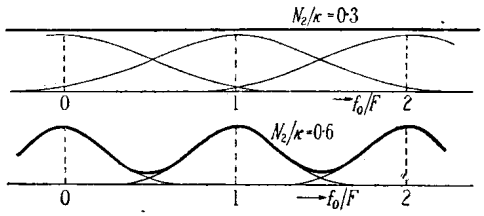

Fig. 3.10.—The function $S(f_0, 0)$.

for two values of the parameter N_2/κ. In the cases which are of practical interest N_2/κ is equal to or larger than unity, and the probability functions become so sharp that their overlap is negligible, and (3.32) consists of recurring peaks of probability shape.

It is now possible to construct diagrams, which may be called frequency reconversion diagrams, which show the reproduced spectrum of any pure original tone in the same way as the previous simple conversion diagrams. Fig. 3.11 is a first example of such a diagram, with $\kappa = \frac{1}{2}$; i.e. the cycle consists in compression to one-half, followed by expansion to the original range. The slit numbers are assumed as $N_1 = N_2 = \frac{1}{2}$, which was previously found to represent the most advantageous compromise. The diagram can be considered as three-dimensional, with the profiles of the S-function at right angles to the (f_0,f) plane. The amplitudes are plotted in the direction f_0, so that the spectrum corresponding to any original frequency f_0 can be immediately constructed by drawing a horizontal line and plotting the heights of the S-function at the intersections with the lines of constant order.

This is carried out for a full cycle of reproduction in Fig. 3.12, which may be compared with Fig. 3.7 (central figure) illustrating the result of the expansion, starting from an undistorted record. It must be noted that τ in Fig. 3.7 corresponds to τ_2 in Fig. 3.12, and as $\tau_1 = 2\tau_2$ the minimum interval between two frequencies in the spectrum in Fig. 3.12 is half of that in Fig. 3.7. If this is borne in mind, it can be seen immediately that the difference between the two cases is mainly that the two side-lines in Fig. 3.7 have now split up into two lines each (with some insignificant satellites), and the centre of gravity of these two lines follows very nearly the same course as in Fig. 3.7. But it has been shown before that with $\kappa = \frac{1}{2}$, $1/\tau_1$ can be made so small that the ear can hardly, if at all, distinguish between the two tones.

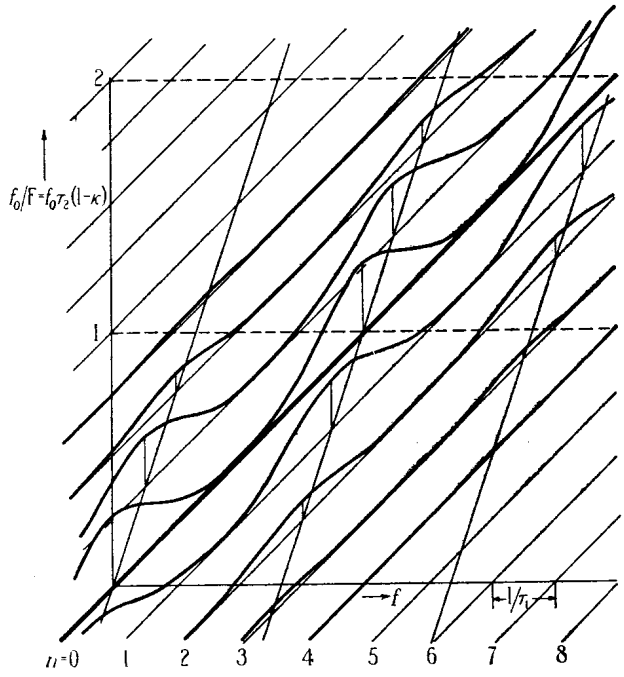

Fig. 3.11.—Frequency reconversion diagram.
$N_1 = N_2 = \frac{1}{2}$; $\kappa = \frac{1}{2}$

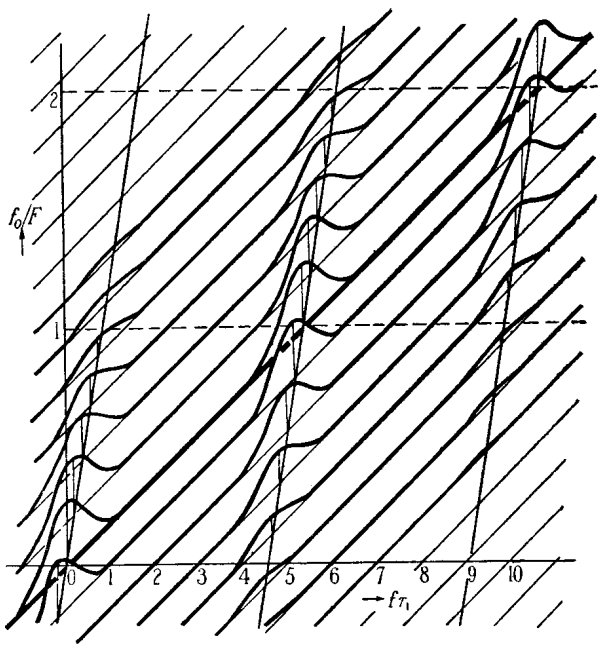

Fig. 3.13.—Frequency reconversion diagram.
$N_1 = N_2 = \frac{1}{2}$; $\kappa = \frac{1}{4}$

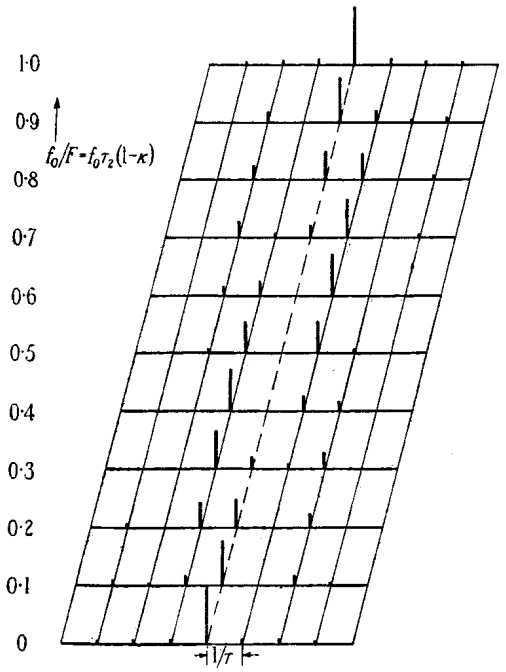

Fig. 3.12.—Re-expanded spectrum of ten frequencies (full cycle of reproduction).
$N_1 = N_2 = \frac{1}{2}$; $\kappa = \frac{1}{2}$

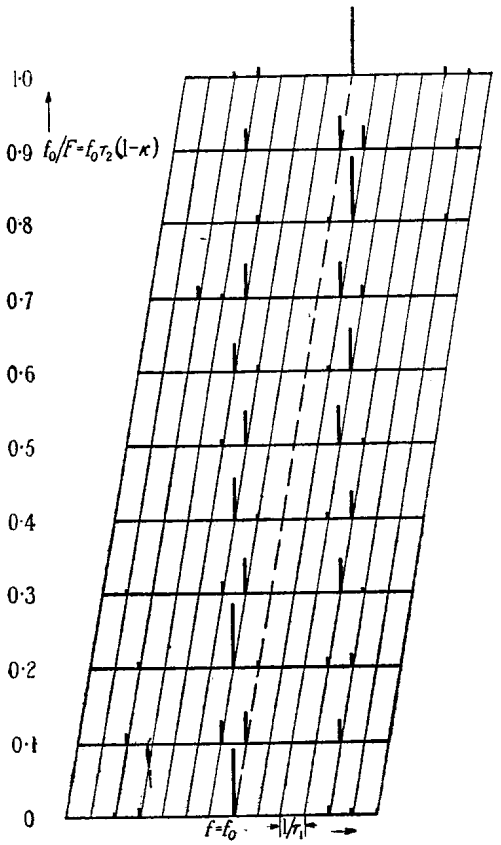

Fig. 3.14.—Re-expanded spectrum of ten frequencies (full cycle of reproduction).
$N_1 = N_2 = \frac{1}{2}$; $\kappa = \frac{1}{4}$

($1/\tau_1$ can be made about 7–10·5 c/s for speech, and 4 c/s for music.) Thus the practical difference between Figs. 3.7 and 3.12 is almost negligible, and we can say that the distortions arise almost entirely in the expansion process.

Fig. 3.13 is a reconversion diagram for a transmission cycle with $\kappa = \frac{1}{4}$, with the same slit numbers as before. Fig. 3.14 contains the reproduced spectra. This diagram approximates even more closely to Fig. 3.7, as the separation in the doublets at either side of the correct reproduction has become even smaller as compared with the frequency interval between the doublets. Thus in this case the distortions arise even more exclusively in the expansion process. The only essential differ-

ence as compared with the case $\kappa = \frac{1}{2}$ is quantitative. The beat frequency between the doublets is now about $4/\tau_1$, twice as large as before.

If in Fig. 3.14 the doublets are imagined as merged into one, the lines connecting them will be almost vertical. Thus we can interpret the operation of the frequency reconvertor in a somewhat different way. It acts very nearly like a musical instrument with a discrete set of frequencies, which tries to imitate speech or music as closely as possible with a limited number of tones. It is well known that if a vowel is sung into an open piano with the loud pedal depressed it will echo the vowel very clearly. The frequency reconvertor performs a similar imitation, but with the difference that its fixed frequencies are set at equal arithmetical, not geometrical, intervals. Hence the reproduction will tend to become more perfect at higher frequencies. At lower frequencies there must necessarily be departures from perfect reproduction. This becomes evident if it is remembered that the frequency convertor does not change the rhythm or "time-pattern" of speech or music. In frequency language this means that frequencies well below the audible range are reproduced almost with the original value, whatever the value of κ.

Summing up, we can say that a frequency compressor and an expander operating in succession produce as close a reproduction of the original as is compatible with the uncertainty relation, and the limit is set almost entirely by the expansion, the errors introduced by the compression being relatively small.

(4) PROVISIONAL REPORT ON EXPERIMENTAL WORK

Theory can give a complete description of the operation of the frequency convertor either in time language, or in frequency language, or in the more general representation discussed in previous communications, but it does not enable us to draw conclusions on the quality of the reproduction.

In order to subject the theory to a first rough test, a 16-mm sound-film projector was converted by a few simple modifications into a frequency convertor. Fig. 3.15 is a photograph of the essential parts, and Fig. 3.16 is a schematic illustration of the optical arrangement.

The usual single, stationary slit of the sound head was replaced by a slotted drum which rotated round an axis passing through the filament of the exciter lamp. The drum was of 0·005-in steel tape, and the width of the slits was also about 0·005 in. The condenser lens was replaced by as large a lens as the fitting would take, with a free diameter of about 1 inch. Immediately in front of the slotted drum a frame was arranged for the "window." In the case of films with variable-area sound tracks this was a film with graded transmission, produced by a photographic process or sprayed with an airbrush. For variable-density films the window was cut out of black film or paper to the desired shape. The window and the slits behind it were imaged on the film by the same microscope objective as used in ordinary operation, which reduced their image to about one-quarter. Thus, allowing for optical errors, the effective slit-width was 0·0015–0·002 in. The maximum length, T, of the window which could be utilized was limited both by the diameter of the condenser lens and by the collecting system which guides the collected light to the photocell. Measured on the film it was about 6 mm. Sound-film moves at the standard speed of 183 mm/sec; thus the maximum T was about 32 millisec. By running the film on the "silent" setting, at about 125 mm/sec, this could be increased to about 48 millisec. The shortness of these times was a severe limitation of the apparatus. The improvement between 32 and 48 millisec was so marked that it appears to confirm the expectation that the optimum T is considerably longer, probably 100 millisec, perhaps even more.

The slotted drum had a stepped pulley attached to it which could be driven at different speeds by means of a spring belt from another stepped pulley attached to a sprocket of the projector. By crossing the belt the motion could be reversed. The following values of κ were tried:—

$$\kappa = 0{\cdot}25 \quad 0{\cdot}33 \quad 0{\cdot}42 \quad 1{\cdot}5 \quad 1{\cdot}75 \quad 2{\cdot}0 \quad 3{\cdot}0 \quad 3{\cdot}33$$

It became evident in the first experiments that the window length of 32 millisec was insufficient for the reproduction of music, hence the later tests were mostly restricted to the reproduction of speech. The uneven rotation of the drum due to the elasticity of the spring belt was also much less objectionable with speech than with music.

Male speech remains completely intelligible with $\kappa = 1{\cdot}5$, i.e. if the frequencies are raised by 50%, though a baritone changes into a high tenor. The intelligibility falls appreciably with $\kappa = 1{\cdot}75$, when the voice changes into a mezzo-soprano, though even with $\kappa = 2$ almost half of the words were intelligible.

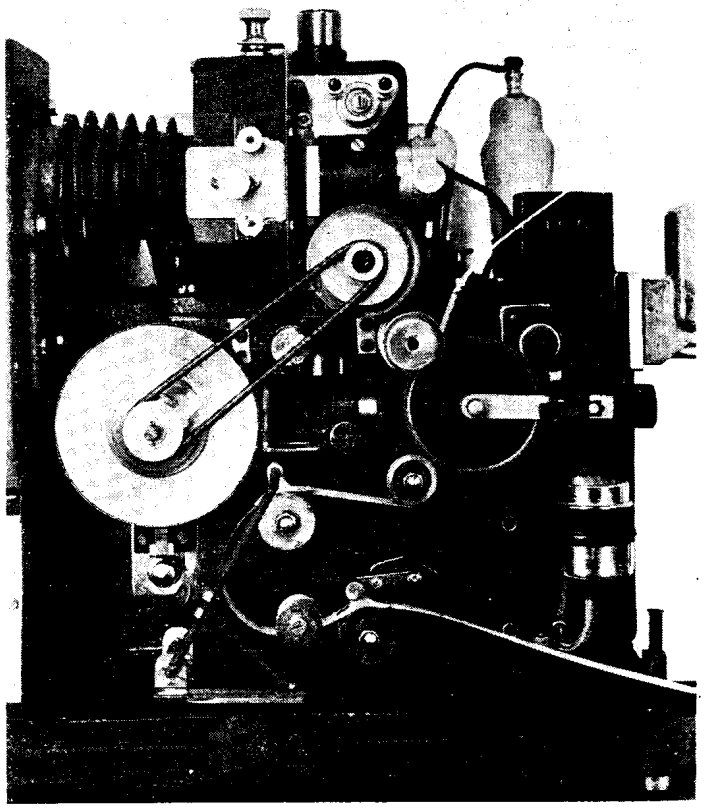

Fig. 3.15.—16-mm sound-film projector converted into an experimental frequency convertor.

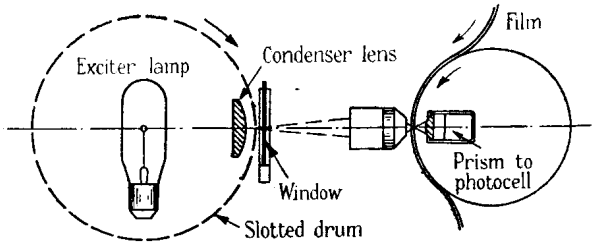

Fig. 3.16.—Optical arrangement in frequency convertor.

This changes a baritone into a soprano. Reduction by the available compression ratios of 0·42 or less, on the other hand, changed male speech into a deep growling, entirely unintelligible.

Such conversion experiments, in which the voice becomes unnatural by frequency transposition, do not, of course, give a test of intelligibility after reconversion to the original frequency range. But two tests could be carried out immediately which allow a first rough estimate of these effects to be made. One test was to run the sound film at "silent" speed, i.e. about $\frac{3}{4}$ standard speed, and apply expansion with $\kappa = \frac{3}{2}$. Speech restored in this way sounded almost entirely natural, and the intelligibility was appreciably better than if the record was run at $\frac{3}{4}$ speed before a stationary slit.

A second reconversion test is based on the fact that positive and negative frequencies are indistinguishable, so that $\kappa = +1$ and $\kappa = -1$ both reproduce the original frequencies of the record. But while $+1$ can be realized with a stationary slit, -1 means that the slits have to run in the same direction as the record, with double speed, so that the relative speed of the film against the slits is $-v$ instead of $+v$, i.e. the same in absolute value. This experiment was tried with different slit numbers, $N = 0{\cdot}5$, 0·75, 1 and 2. The beat frequencies $1/\tau$ were 60, 90, 120 and 240 c/s. $N = 0{\cdot}5$ was easily the best, in full agreement with the theoretical expectations. It gave perfectly intelligible, though not quite natural, reproduction. The larger slit numbers produced strong "rrr" sounds, which decreased the intelligibility, but it is remarkable that even with a beat frequency of 240 c/s about half the words were intelligible. It may be seen from eqn. (3.14) that the beat frequencies at $\kappa = -1$ are the same as for $\kappa = +3$. Thus this test corresponded roughly to a reconversion with $\kappa = \frac{1}{3}$, at a window length of 32 millisec. As it appears highly probable that the best window length will be about three times as much, perhaps even more, it appears that ultimately even sevenfold compression and re-expansion can be realized without essential loss in intelligibility, though with noticeable distortion.

(5) DEVICES FOR KINEMATICAL FREQUENCY CONVERSION

So far the theory has been explained and illustrated only in the case of a sound film, i.e. with a permanent optical record, but evidently there are many more possibilities for realizing the underlying general principle.

The essential features of the kinematical method are as follows. A permanent or temporary record moves past a fixed window with suitably graded attenuation, and inside this window the record is scanned by pick-ups which are themselves moving with some speed different from that of the record. Hence we can use any sort of record which persists long enough to pass across the window, and any sort of pick-up which does not damage the record. The last condition excludes gramophone records with needle pick-ups, but there are many more promising possibilities.

Phosphorescence, wave motion and magnetization are well-known physical processes with "memory." The last of these is suitable for permanent as well as for temporary records, and will be discussed later. The first two are suitable for condensed transmission in communication channels.

Phosphorescent records can be used in very much the same way as the permanent optical records previously discussed. The film is replaced by a loop of film coated with phosphorescent material, or by a coated rotating drum. This is excited by a suitable recorder, such as a variable light source or an oscillograph, after which it passes immediately into the window, where it is scanned by moving slits or their optical equivalents. The exponential decay of the phosphorescence can be compensated by a suitable exponential wedge. Behind the window the phosphorescence can be removed by heating or by infra-red irradiation. A similar apparatus can be used at the receiving end.*

Wave motion in fluids is an interesting substitute for a moving record. It has been used in the Scophony system of television in order to preserve the picture of a whole line for about 10^{-4} sec. The Scophony trough contains a piezo-electric crystal at one end and an absorber at the other. The pressure waves running along the trough produce differences in the refractive index of the liquid and form an equivalent of a film running at extraordinary speed. It is well known that such a trough can also imitate a succession of running slits if the crystal is operated with a series of sharp pulses.† Thus a system of two Scophony troughs, in combination with a suitable optical system, appears to be a practicable form of frequency convertor. But it is not very suitable for the conversion of sound, where the window width required is of the order of 0·1 sec, whereas Scophony troughs, unless they are made very large, conserve the record for only about 10^{-4} sec. They might perhaps be suitable for compressed television transmission, if such a scheme should prove practicable. This subject, however, is outside the scope of the present paper.

The most convenient method of condensed transmission will probably use magnetic tape or wire recorders at both ends of the communication channel. Fig. 3.17 shows the schematic

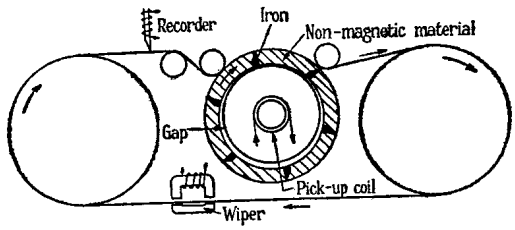

Fig. 3.17.—Frequency convertor with magnetic tape.

arrangement. A loop of the tape or wire runs continuously over two pulleys. Before reaching the recorder the previous record is wiped out, by demagnetization by saturation, or—as in some modern systems—by demagnetization with high frequency. After passing under the recording edge the tape runs over a wheel which has a number of sharp, wedge-shaped iron spokes. To avoid scraping, these are embedded in non-magnetic material; friction may be prevented by an oil film. The spoked wheel rotates with some speed different from that of the film, according to the κ of the conversion. It forms the equivalent of the rotating slits in the film scanner. The equivalent of a window with graded transmission is formed by a suitably shaped magnetic gap between the annular wheel and a central iron pole-piece which carries the pick-up coil. The current induced in the pick-up coil is amplified and transmitted through the communication channel. At the receiving end the current is applied to the recorder of a similar instrument, the only difference being that the wheel rotates here with a different speed relative to the film. The window length can be varied by changing the position of the two pulleys which determine the arc of contact, or—more advantageously—by running the motor at different speeds. This may be necessary if it is desired to transmit both speech and music under optimum conditions.

All systems of this kind necessarily produce a certain delay between transmission and reception. The average delay cannot be less than the width T of one window, plus twice the time

* A somewhat similar arrangement has been used for other purposes by Goldmark and Hendricks, Ref. No. 3.5.

† First suggested by F. Okolicsányi.

interval between the recorder and the near edge of the window. In the transmission of speech this can probably be kept below 200 millisec.

The device shown in Fig. 3.17 could be used also for long-playing magnetic gramophones, dictaphones and the like. The only change is that a permanent instead of a temporary record is used and the "wiper" is eliminated. But it may be mentioned that in gramophones, sound-film apparatus and the like, in which the aim is as high a quality of reproduction as possible, and which must be ready to reproduce speech or music without any change of adjustment, it does not appear practicable to apply compression to the whole range of audible frequencies. In such cases it may be better to divide the audio range into two parts, say 25–1 500 c/s and 1 500–7 500 c/s. A track may be provided for each, of which the first is an ordinary record, whereas the second is compressed fourfold. Thus with a double-track record it may be possible to reproduce a waveband of 7 500 c/s, at film speeds which would be normally sufficient only for about 1 500 c/s. This application may perhaps be of interest in sub-standard sound-film projectors.

(6) ELECTRICAL METHODS OF CONDENSED TRANSMISSION

It may be surmised *a priori* that mechanical motion is not an indispensable part of condensed transmission schemes. Mathematically speaking, the essence of the methods previously discussed was to apply certain linear but time-dependent operators to an original signal $s_0(t)$, and it appears very likely that these can be produced also by suitable circuits. It will be shown that these, and even more general operators, can be produced electrically if suitable signal generators are available.

Mechanical motion in the schemes previously described had the general function of producing new frequencies from one given original frequency. Mathematical analysis has shown that this consists essentially in the repeated addition of the "repetition frequencies" of the device to the original frequency. But it is well known that addition and subtraction of frequencies can be produced without mechanical means, by the technique of "mixing." Hence in order to devise an electrical equivalent of the kinematical method we must search in the first place for a suitable method of modulation. Evidently modulation with other than simple sine-wave carriers is necessary, as multiplication with a simple carrier produces only a shifting and duplication of wavebands.

The other essential feature of the kinematical method was a permanent or temporary record, or more generally "memory" of some sort. Can ordinary electrical circuits have memory? The answer to this is that *every* tuned electrical system, i.e. every system which has no unlimited flat response, has a sort of memory, because an instantaneous impulse has a certain after-effect. A particularly interesting special case is a system with sharp resonance peaks which are at multiples of some fundamental frequency, approximating to the "selection factor" shown in Fig. 3.4. Such a system would incessantly repeat the same waveform. If the damping were appreciable, the repetitions would become gradually less and less like the original. This repetition is something rather close to the everyday concept of memory.

It might appear that the simplest method of transmission with non-constant carrier frequency is modulation with a carrier of constant amplitude, but with a frequency which varies between two limits sinusoidally, or according to a saw-tooth curve. If the local oscillator of the receiver varies its frequency according to the same law, a signal similar to the original can be expected. This system is known as "re-entrant modulation." A certain amount of saving in frequency band may be obtained with this system without prohibitive distortions, if the transmission channel is made smaller than the total frequency sweep. But, though this system may be the simplest to realize, its mathematical treatment leads to considerable complications. Therefore the following investigation will be based on a system of modulation which may not be easy to realize, but which allows comparatively simple and general mathematical discussion. This will be achieved by making use once more of the unique properties of certain signal shapes with probability envelope.

We assume a carrier of the form

$$\sum_{-\infty}^{\infty} k \exp - \lambda(t - k\tau)^2 \quad . \quad . \quad . \quad (3.33)$$

If the constant λ is real and positive this represents a recurrent probability pulse. But the discussion is just as simple if we make the more general assumption that λ is a complex constant with a positive real part

$$\lambda = \alpha^2 + j\beta^2 \quad . \quad . \quad . \quad . \quad (3.34)$$

The real part of (3.33) is, apart from a phase constant, the sum of pulses of the form

$$e^{-(\alpha t)^2} \cos (\beta t)^2 \quad . \quad . \quad . \quad . \quad (3.35)$$

An example of such a pulse is shown in Fig. 3.18. It represents a sine wave with a linearly-varying frequency, modulated

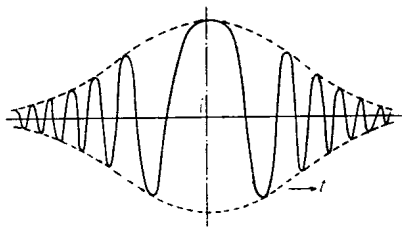

Fig. 3.18.—Modulating pulse.

by a probability pulse. An interesting feature of these waves is that, by choosing the recurrent frequency conveniently, their superposition can result in a waveshape which closely approximates to a wave of constant amplitude with a frequency varying according to a saw-tooth curve; hence by suitable choice of the constants it is possible to cover "re-entrant modulation" without its mathematical complications.

The great advantage of the waveform (3.33) or (3.35) is that its Fourier transform is of the same type as the signal. This allows us to evade the danger of the formulae growing more and more complicated with every step of the analysis.

The signal $s_0(t)$ may again be a pure harmonic oscillation, which may be written in complex form as

$$s_0(t) = \text{cis } 2\pi f_0 t \quad . \quad . \quad . \quad . \quad . \quad (3.36)$$

Only the complex modulation product of (3.36) and (3.33) will be considered. It is well known that the real product can be obtained from this by adding to it the product with the sign of f_0 reversed, and adding to the sum its complex conjugate. But it will not be necessary to carry out this process in order to recognize the essential features of this method of transmission. The complex modulation product is

$$s_m(t) = \sum_{-\infty}^{\infty} k \exp [- \lambda(t - k\tau)^2 + 2\pi j f_0 t] \quad . \quad (3.37)$$

The Fourier transform of this is

$$S_m(f) = \sqrt{\left(\frac{\pi}{\lambda}\right)} \exp \left[- \frac{\pi^2}{\lambda}(f - f_0)^2 \sum_{-\infty}^{\infty} k\, \delta(f - f_0 - k/\tau) \right] \quad . \quad . \quad . \quad (3.38)$$

Thus by modulation with the carrier (3.33) the spectrum has been spread out according to a probability law on both sides of the original frequency, while the result of the recurrence is to split up the spectrum into sharp lines with constant frequency interval $1/\tau$.

We now assume that the modulated signal is passed through a filter with a transfer admittance

$$\exp - \pi^2(f - f_c)^2/\sigma \quad . \quad . \quad . \quad . \quad (3.39)$$

where σ is a complex constant with positive real part. If σ is real this is a "probability filter." The filter transmission centres on f_c, but this will not be the centre of the transmitted wave. As illustrated in Fig. 3.19, the product of two probability functions is again a probability function, with a centre somewhere between the centre of the two factors. Hence the filtered spectrum $S(f)$ can again be expressed in a mathematical form similar to (3.38) but with changed constants:—

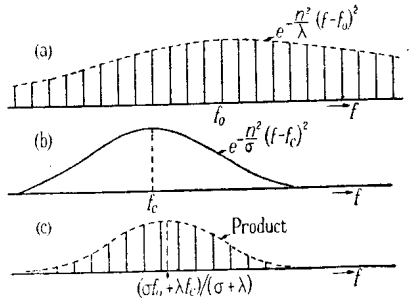

Fig. 3.19.—Electrical frequency conversion.

$$S(f) = \sqrt{\left(\frac{\pi}{\lambda}\right)} \exp\left[-\frac{\pi^2}{\sigma + \lambda}(f_0 - f_c)^2\right] \exp\left[-\pi^2\left(\frac{1}{\sigma} - \frac{1}{\lambda}\right)\left(f - \frac{\sigma f_0 + \lambda f_c}{\sigma + \lambda}\right)^2\right] \sum_{-\infty}^{\infty} k\,\delta(f - f_0 - k/\tau) \quad . \quad (3.40)$$

We write now

$$\sigma/(\sigma + \lambda) = \kappa \quad . \quad . \quad . \quad . \quad (3.41)$$

and obtain the spectrum in the form

$$S(f) = \sqrt{\left(\frac{\pi}{\lambda}\right)} \exp\left[-\frac{\pi^2}{\lambda}(1 - \kappa)(f_0 - f_c)^2\right] \exp\left\{-\frac{\pi^2}{\kappa\lambda}[f - \kappa f_0 - (1 - \kappa)f_c]^2\right\} \sum_{-\infty}^{\infty} k\,\delta(f - f_0 - k/\tau) \quad . \quad (3.42)$$

This is a formula very similar to that obtained in the case of kinematical compression, but with some differences, the most important of which is that σ, λ and κ need not be real. It is interesting, however, to consider the special case in which σ, λ and consequently also κ are real and positive. In this case eqn. (3.42) differs from eqn. (3.10*a*) or (3.10*b*) only in two points. One is that the maximum of the amplitudes is not at $f = \kappa f_0$, but at

$$f = \kappa f_0 + (1 - \kappa)f_c$$

i.e. the spectrum is not only compressed, but also shifted by a certain constant amount, depending on the position of maximum filter transmission, f_c. The other new feature is the factor

$$\exp\left[-\pi^2(1 - \kappa)(f_0 - f_c)^2/\lambda\right] \quad . \quad . \quad (3.43)$$

which is independent of f but dependent on the original frequency f_0. Hence different frequencies are not reproduced with equal intensity. This effect can be reduced or eliminated by boosting the original amplitudes in a ratio inverse to the factor (3.43) before modulation.

We see now that by applying in succession the operations of boosting, modulation with repeated probability pulses, and filtering, we can produce by purely electrical means a *compressed* spectrum identical to that obtainable by mechanical methods. But it is important to note that only compression can be achieved in this special case, not expansion, as κ, given by eqn. (3.41), is necessarily smaller than unity.

By a rather complicated calculation, which may be omitted, it can be shown that by a second modulation—in the receiver—with a modulating wave of the type (3.33) it is possible to restore the original frequency, with very much the same distortions as in kinematical reconversion. But it is essential that both λ and σ should have imaginary components, i.e. both the modulating pulses and the filter characteristic must be of the type as shown in Fig. 3.19. Simple probability pulses and probability filters can achieve only part of the reconversion cycle. Hence the electrical method is better described as "condensation-dilution" than as "compression-expansion." The transmitted signal spectrum is entirely dissimilar to the original, as the spectrum corresponding to a single original frequency is spread out over the whole transmitted range.

At the present stage it is impossible to overlook the possibilities of electrical methods of condensed transmission, which in principle appear almost unlimited. Progress is likely to be slow and difficult, as the mathematical treatment of pulses different from those considered here is liable to become excessively complicated, and experiments unguided by theory do not appear very promising. But the economy which may ultimately be achieved is likely to be large enough to encourage efforts in this direction.

(7) REFERENCES

(3.1) HOMER, DUDLEY: "Re-making Speech," *Journal of the Acoustical Society of America*, 1939, **11**, p. 169.

(3.2) CAMPBELL, G. M., and FORSTER, R. M.: "Fourier Integrals for Practical Applications," Bell Telephone System Monograph B.584, 1931.

(3.3) ROBERTS, F. F., and SIMMONDS, J. C.: "Some Properties of a Special Form of Electrical Pulse," *Philosophical Magazine*, 1943, **34**, p. 822.

(3.4) ROBERTS, F. F., and SIMMONDS, J. C.: *ibid.*, VII, 1944, **35**, p. 459.

(3.5) GOLDMARK, P. C., and HENDRICKS, P. S.: "Synthetic Reverberation," *Proceedings of the Institute of Radio Engineers*, 1939, **27**, p. 747.

(8) APPENDICES

(8.1) The Response of Frequency Convertors to Elementary Signals

It has been shown in Parts 1 and 2 that signal analysis in terms of certain "elementary signals" has particular advantages, especially in problems of physiological acoustics. These elementary signals are simple harmonic oscillations, modulated with a probability pulse. Analysis in terms of these functions contains the representation of a signal as a time function $s(t)$ and as a frequency function $S(f)$ as limiting special cases.

Elementary signals are also very suitable for describing the operation of a frequency convertor with a probability window, as a convertor reproduces any function of this type as the sum of functions of the same type.

The frequency convertor transforms an "original" signal $s_1(t)$ into

$$s(t) = \sum_{-\infty}^{\infty} k \exp\left\{-\left(\frac{t+k\tau}{N\tau}\right)^2 s_1[\kappa t - (1-\kappa)k\tau]\right\} \quad (3.44)$$

This formula is obtained from eqn. (3.6) if x_k is substituted from eqn. (3.3) and the repetition interval τ from eqn. (3.9). Substitute for $s_1(t)$ a general elementary signal

$$s_1(t) = \exp - \frac{\epsilon^2(t-t_0)^2}{(N\tau)^2} \operatorname{cis} 2\pi f_0(t-t_0) \quad . \quad (3.45)$$

The dimensionless parameter ϵ characterizes the sharpness of the signal. In Part 1 the effective duration of a signal has been defined as $\sqrt{(2\pi)}$ times its r.m.s. duration. In the present case this is

$$(\Delta t)_1 = \sqrt{\frac{\pi}{2}} \frac{N\tau}{\epsilon} \quad . \quad . \quad . \quad . \quad . \quad (3.46)$$

The effective spectral width is, by the same definition,

$$(\Delta f)_1 = \frac{1}{\sqrt{(2\pi)}} \frac{\epsilon}{N\tau} \quad . \quad . \quad . \quad . \quad (3.47)$$

The relation of the time interval $N\tau$ to the window width T is given by eqn. (3.14), which combined with (3.46) gives

$$\epsilon = \frac{0{\cdot}41}{1-\kappa} \frac{T}{(\Delta t)_1} \quad . \quad . \quad . \quad . \quad . \quad (3.48)$$

E.g. for $\kappa = 2\epsilon$ (in absolute value) is 0·41 for signals with an effective duration equal to the window length. It is larger for sharper signals, smaller for longer ones.

Substitution of (3.45) in eqn. (3.44) gives

$$s(t) = \sum k \exp\left[-\left(\frac{1}{N\tau}\right)^2 \left\{(t+k\tau)^2 + \epsilon^2[\kappa t - k(1-\kappa)\tau - t_0]^2\right\}\right] \times \operatorname{cis} 2\pi f_0[\kappa t - (1-\kappa)k\tau - t_0] \quad (3.49)$$

This can be written in the simpler form

$$s(t) = \Sigma k \exp\left[-\Omega^2(t-\beta_k)^2 + \gamma_k\right] \quad . \quad . \quad (3.50)$$

where the constants have the following values:—

$$\Omega^2 = (1+\epsilon^2\kappa^2)/(N\tau)^2$$

$$\beta_k = -\left\{k\tau - \epsilon^2\kappa[k\tau(1-\kappa) + t_0] - j\pi f_0(N\tau)^2\right\}/(1+\epsilon^2\kappa^2)$$

$$\gamma_k = -\left\{(1+\epsilon^2\kappa^2)\beta_k^2 - (k\tau)^2 - \epsilon^2[(1-\kappa)k\tau + t_0]\right\}/(N\tau)^2 - 2\pi j f_0[(1-\kappa)k\tau + t_0] \quad (3.51)$$

The Fourier transform of the kth term of (3.50) is

$$\frac{\sqrt{\pi}}{\Omega} \exp\left[-\left(\frac{\pi f}{\Omega}\right)^2 - 2\pi j\beta_k f + \gamma_k\right] \quad . \quad (3.52)$$

Applying this to (3.50) a somewhat lengthy calculation leads to the following expression for the spectrum of the reproduced signal:—

$$S(f) = \frac{\sqrt{(\pi)}N\tau}{\sqrt{(1+\epsilon^2\kappa^2)}} \exp\left[-\frac{(\pi N\tau)^2}{1+\epsilon^2\kappa^2}(f-\kappa f_0)^2\right] \times \operatorname{cis}\frac{2\pi}{1+\epsilon^2\kappa^2}(\epsilon^2\kappa f - f_0)t_0 \times \sum_{-\infty}^{\infty} k \exp\left[-\frac{\epsilon^2(k\tau+t_0)^2}{(1+\epsilon^2\kappa^2)(N\tau)^2}\right] \times \operatorname{cis}\frac{2\pi k\tau}{1+\epsilon^2\kappa^2} \left\{[1-\epsilon^2\kappa(1-\kappa)]f - f_0\right\} \quad (3.53)$$

The first factor, in the first line, can be called the attenuation factor, the second the phase factor, and the third the spectral separation factor. If $\epsilon = 0$ and $t_0 = 0$, eqn. (3.53) simplifies to eqns. (3.10*a*) and (3.10*b*), discussed in the text. In this special case of infinite wave trains the separation factor becomes a "selection factor" and the spectrum becomes a line spectrum.

In the general case the attenuation factor has the effect that the effective spectral width of the reproduced signal becomes

$$\Delta f = \frac{1}{\sqrt{(2\pi)}} \frac{\sqrt{(1+\epsilon^2\kappa^2)}}{N\tau} \quad . \quad . \quad . \quad (3.54)$$

which is $\sqrt{(1+\epsilon^2\kappa^2)}/\epsilon$ times the original value (3.47). If ϵ is very large, i.e. if the signal is very sharp, this ratio approaches κ, the conversion ratio. This means that with very short signals the spectral envelope is reproduced accurately, on a scale κ times the original. The reproduced signal $s(t)$ itself consists in this case of an accurate reproduction of the original, but on a time scale $1/\kappa$ times extended, and of similar but weaker "echoes," produced by repeated passage of slits across the record of the short signal.

The opposite case arises if the signal is of long duration. Here the spectral width, which in the original is very small, is expanded to a value $1/\sqrt{(2\pi)}N\tau$, whereas the envelope of the reproduced signal $s(t)$ approaches the original very closely. This is the reason why the frequency convertor can reproduce without much distortion the articulation of speech or the time pattern of music.

(8.2) Combination of Two Conversion Processes in Succession

Consider the conversion as described by eqn. (3.10*b*) as a first operation on the frequency f_0, with suffix "1," which produces a certain spectrum S_i on an intermediate frequency scale f_i

$$S_i(f_i) = \exp\left[-(\pi N_1\tau_1)^2(f_i - \kappa_1 f_0)^2\right]\Sigma k\, \delta(f_i - f_0 - k/\tau_1) \quad (3.55)$$

This is different from zero only if

$$f_i = f_0 + k/\tau_1 \quad . \quad . \quad . \quad . \quad (3.56)$$

where k is any integer. Apply now a second similar operation, with suffix "2," to the result of the first operation. This splits every spectral line (3.56) into an infinity of equidistant lines, given by

$$f = f_i + m/\tau_2 \quad . \quad . \quad . \quad . \quad . \quad (3.57)$$

Eliminating the intermediate frequency f_i from the two last equations, we see that non-zero amplitudes in the final spectrum will appear at

$$f = f_0 + k/\tau_1 + m/\tau_2 \quad . \quad . \quad . \quad (3.58)$$

The reduction of the spectrum to discrete lines can be conveniently expressed by the selection operator

$$\Sigma m \Sigma k\, \delta(f - f_0 - k/\tau_1 - m/\tau_2) \quad . \quad . \quad . \quad (3.59)$$

using again the "delta function," which is zero everywhere except at argument zero. We can now write the result of the two operations as follows:—

$$S(f) = \Sigma m \Sigma k \exp\left[-(\pi N_1\tau_1)^2(f_i - \kappa_1 f_0)^2\right] \times \exp\left[-(\pi N_2\tau_2)^2(f - \kappa_2 f_i)^2\right] \times \delta(f - f_0 - k/\tau_1 - m/\tau_2) \quad . \quad (3.60)$$

Eqn. (3.16) is obtained from this by substituting the values of f_i and f from eqns. (3.56) and (3.57).

Any spectral line as given by eqn. (3.58) can be characterized by two integers k_0 and m_0

$$f = f_0 + k_0/\tau_1 + m_0/\tau_2 \quad . \quad . \quad . \quad (3.61)$$

If τ_1 and τ_2 are incommensurable there will be no other integral values which satisfy this equation; hence only a single term of the sum (3.60) will contribute to the amplitude of this frequency. But if τ_1 and τ_2 are in a rational relation

$$\tau_1/\tau_2 = p/q \quad . \quad . \quad . \quad . \quad . \quad (3.62)$$

where p and q are relative primes, there will be an infinity of integer solutions of (3.61) of the form

$$\begin{aligned} k &= k_0 + \nu_p \\ m &= m_0 - \nu_q \end{aligned} \quad . \quad . \quad . \quad . \quad . \quad (3.63)$$

where ν is any integer. But if the same window width is used in both conversion processes, and if the speed ratios are produced by toothed wheels, by eqn. (3.20) τ_1/τ_2 is bound to be rational. This means that the line spectrum (3.58) will repeat itself with a period

$$p/\tau_1 = q/\tau_2 \quad . \quad . \quad . \quad . \quad . \quad (3.64)$$

To avoid unessential complications the discussion in the text is restricted to the case $q = 1$.

(8.3) Reduction of the Recurrent Exponential Pulse to Theta Functions

In eqn. (3.28) for $S(f, n)$ introduce the following notations:—

$$2\left(\frac{\pi N_2}{\kappa}\right)^2 = \alpha^2; \quad \frac{n\kappa}{2p} = \mu; \quad \frac{f_0}{F} + \frac{k_0}{p} = y \quad . \quad (3.65)$$

This enables us to write it in the form

$$S(f, n) = S(y, \mu) = e^{-\alpha^2\mu^2}\sum_\nu e^{-\alpha^2(y+\nu-\mu)^2} \quad . \quad (3.66)$$

The theta function θ_{00} as defined in analysis* is

$$\theta_{00}(z, \tau) = \sum_{-\infty}^{\infty}{}_\nu e^{j\pi(\nu^2\tau + 2\nu z)} \quad . \quad . \quad . \quad (3.67)$$

or, with imaginary arguments,

$$\theta_{00}(jz, j\tau) = e^{\pi z^2/\tau}\sum_{-\infty}^{\infty}{}_\nu e^{-\pi\tau(\nu + z/\tau)^2} \quad . \quad . \quad (3.68)$$

Now put $\pi\tau = \alpha^2; \quad z/\tau = y - \mu \quad . \quad . \quad . \quad (3.69)$

This gives

$$\theta_{00}\left[j\frac{\alpha^2}{\pi}(y-\mu), j\frac{\alpha^2}{\pi}\right] = e^{\alpha^2(y-\mu)^2}\sum_{-\infty}^{\infty}{}_\nu e^{-\alpha^2(\nu+y-\mu)^2} \quad (3.70)$$

Dividing this by eqn. (3.66) we obtain

$$S(y, \mu) = e^{-\alpha^2[(y-\mu)^2+\mu^2]}\theta_{00}\left[j\frac{\alpha^2}{\pi}(y-\mu), j\frac{\alpha}{\pi}\right] \quad . \quad (3.71)$$

and finally, substituting the original values for α, μ and y,

$$S(f_0, n) = e^{-2\left(\frac{\pi N_2}{\kappa}\right)^2\left[\left(\frac{f_0}{F} + \frac{k_0}{p} - \frac{n\kappa}{2p}\right)^2 + \left(\frac{n\kappa}{2p}\right)^2\right]}$$
$$\theta_{00}\left[j2\pi\left(\frac{N_2}{\kappa}\right)^2\left(\frac{f_0}{F} + \frac{k_0}{p} - \frac{n\kappa}{2p}\right), j2\pi\left(\frac{N_2}{\kappa}\right)^2\right] \quad (3.72)$$

Tables of theta functions may be found in Jahnke and Emde, "Tables of Functions" (Dover Publications, New York, 1945) and in other works.

ACKNOWLEDGMENTS

The author desires to thank Professor Max Born for criticism of Part 1 of the paper; Messrs. B. Tuppen and I. Williams for their valuable help in the experimental work in connection with Part 3; and the Directors of the British Thomson-Houston Co., Ltd., for permission to publish the paper.

* Also called θ_0. Cf. COURANT and HILBERT: "Methoden der mathematischen Physik," vol. I (Interscience, New York, 1943), p. 331. The notations employed by Whittaker and Watson in "Modern Analysis" (4th ed., pp. 462–490), are somewhat different.

DISCUSSION ON

"RADIO MEASUREMENTS IN THE DECIMETRE AND CENTIMETRE WAVEBANDS"*

NORTH-WESTERN RADIO GROUP, AT MANCHESTER, 18TH JANUARY, 1946

Mr. R. Cooper: The authors use the terms (*a*) accuracy, (*b*) absolute accuracy, (*c*) reading accuracy, (*d*) setting accuracy. The accuracy of a measurement is determined by the deviation of the measurement from the true value of the quantity measured. This deviation is due to errors which occur in making the various instrument settings, readings and calibrations necessary to make the measurement. Will the authors state the sources of error considered in defining each of the above variations of instrument accuracy? In the case of the calorimeter method of measuring high powers the authors state that the "absolute accuracy" of the method is of the order 5%. I presume this value pertains to the equipment used by the authors and is not a statement of the limit of accuracy of the method. I am particularly interested in this system and would appreciate a statement of the sources and magnitudes of the various errors which contribute to the 5% "absolute accuracy."

In considering the measurement of high powers the authors do not mention that it is sometimes necessary to feed the energy into the calorimeter in the form of recurrent impulses. Under these conditions high electrostatic stresses may be set up in the system. To what extent does this consideration influence the design of the resonant-chamber calorimeter (Fig. 14)?

A point having bearing on the design of calorimeters for operation below 10 cm wavelength is the fact that water exhibits an absorption band in this region and its dielectric constant is a function of temperature. Consequently mismatches may be caused by excessive temperature rises. This effect is likely to be pronounced in calorimeters containing a considerable volume of water such as that shown in Fig. 15. I have found a calorimeter of the type shown in Fig. 16 to be free from the effect.

I agree in general with the authors' remarks concerning the design of standing-wave detectors. However, I prefer to limit the length of the slot to about three-quarters of a wavelength, and I judge the performance of standing-wave detectors from curves obtained with an approximately correct termination and with a highly reflecting short-circuit termination. Can the authors give

* Paper by R. J. CLAYTON, J. E. HOULDEN, H. R. L. LAMONT, and W. E. WILLSHAW (see 1946, 93, Part III, p. 97).