On the transmission capacity of the 'ether' and of cables in electrical communications

Proceedings of the first All-Union Conference on the technological reconstruction of the communications sector and the development of low-current engineering. Moscow, 1933.

V A Kotelnikov

Translated by C C Bissell (Open University, UK) and V E Katsnelson (Weizmann Institute, Israel)

Introduction to the Translation

In 1933 the young Russian communications engineer Vladimir Aleksandrovich Kotelnikov published a paper in which he formulated the sampling theorem for lowpass and bandpass signals, and also considered the bandwidth requirements of discrete signal transmission for telegraphy and images. Although Kotelnikov's name later became known in the West as a result of his subsequent work, particularly that on optimal detection, his pioneering 1933 results received little attention at the time outside Russian-speaking areas.

Disputes about priority in science and technology are rarely productive. Different societies quite naturally associate seminal work with pioneers from their own history, particularly when such work is carried out independently within a few years in various geographical locations. The sampling theorem for telecommunications applications is thus associated with Claude Shannon (and, somewhat erroneously, with Harry Nyquist[1]) in the West, and with Vladimir Kotelnikov in Russia. However, Kotelnikov's work is much less well known globally than that of the American engineers[2]**.**

Vladimir Aleksandrovich Kotelnikov was born in 1908 in Kazan. He studied radio engineering as an undergraduate at the Moscow Energy Institute (MEI), and remained there to do postgraduate work. Following wartime research and development work in Ufa (one of the temporary locations of scientific institutions evacuated from endangered cities) he returned to MEI in 1944, where he became professor and Dean of the Radioengineering Faculty. He gained his *Doctorate of Sciences*[3] in 1947 on the topic of optimal detection and became a full Academician in 1953 – unusually, for the time, without the intervening stage of 'corresponding member'. His book *The Theory of Optimum Noise Immunity*, a translation of the 1956 Russian monograph (essentially the author's 1947 dissertation), appeared in the US in 1959. Kotelnikov's distinguished later career included a leading rôle in the Soviet space programme, directing work on planetary observation and mapping, for which he was awarded the prestigious Lenin Prize in 1964. He was Vice-President of the Russian Academy of Sciences from 1975 to 1988. Other awards include the German Eduard Rhein Prize in 1999, and the IEEE Alexander Graham Bell Medal in 2000. In 2003 Kotelnikov celebrated

[1] Nyquist's classic 1928 paper considered the maximum signalling rate over a bandlimited channel. He also noted the necessary and sufficient conditions on the number of coefficients of a Fourier series representing a bandlimited signal. Because of this, and because the Nyquist rate is equal numerically to the minimum sampling rate for a bandlimited signal, the contributions of Nyquist and Shannon to the notion of sampling have often become confused. Nyquist did not consider explicitly the question of sampling a signal in the time domain.

[2] An earlier English-language version was made by one of the present translators (Katsnelson) as Chapter 2 of Benedetto & Ferreira (2001), a highly specialised collection of papers on the state-of-the-art of sampling theory. The current version is a substantial revision as regards style, together with the correction of one or two minor errors.

[3] The Russian *Doktor Nauk* degree is a higher research degree awarded to distinguished researchers on the basis of a thesis, not to be confused with the *kandidat* degree more akin to a Western PhD or equivalent, which Kotelnikov gained in 1933.

his 95th birthday; greetings from Vladimir Putin, President of the Russian Federation, and congratulatory remarks from the IEEE can be found in Lantsberg (2004). He died in February 2005.

Kotelnikov begins his 1933 paper by stating the general problem of spectrum management, and raising the issue of whether it is possible, even theoretically, to do better than single-sideband modulation. Interestingly, he raises the question of the possibility of "some way of separating channels whose frequencies overlap, perhaps even employing a method based not on frequency, as has been the case until now, but by some other means" – something routine now using spread-spectrum techniques.

The next eight pages of the paper derive theoretical results relating to lowpass and bandpass signals and their transmission. The most significant are the following two theorems, although Kotelnikov is then at pains to work out the detailed context, including a discussion of bandpass rather than lowpass sources:

> *Theorem I*
>
> Any function F(t) consisting of frequencies from 0 to f_1 cycles per second can be represented by the series:
>
> $$F(t) = \sum_{-\infty}^{+\infty} D_k \frac{\sin \omega_1 (t - \frac{k}{2f_1})}{t - \frac{k}{2f_1}} \qquad (1)$$
>
> where k is an integer, $\omega_1 = 2\pi f_1$, and D_k are constants depending on F(t).
>
> Conversely, any function F(t) represented by the series (1) consists only of frequencies from 0 to f_1.
>
> *Theorem II*
>
> Any function F(t) consisting of frequencies from 0 to f_1 can be transmitted continuously, with arbitrary accuracy, by means of numbers sent at intervals of $1/2f_1$ seconds.
>
> Indeed, measuring the value of F(t) at times $t = n/2f_1$, where n is an integer, we obtain
>
> $$F(\frac{n}{2f_1}) = D_n \omega_1$$
>
> All the terms of series (1) are zero for this value of t, except for the term with k = n which [...] will equal $D_n\omega_1$. Hence after each interval of $1/2f_1$ seconds we can determine the next D_k. Having transmitted these D_k in turn at intervals of $1/2f_1$ seconds we can reconstruct F(t) to any degree of accuracy according to Equation (1).

At this point it is worth noting Shannon's remarks, published sixteen years later, about the origin of the sampling theorem: "This is a fact which is common knowledge in the communication art. [...] but in spite of its evident importance [it] seems not to have appeared explicitly in the literature of communication theory" (Shannon, 1949). Indeed, mathematicians had studied the problem from a function-theoretic point of view from the mid 19th century or even earlier (Lüke, 1999; Benedetto & Ferreira, 2001).

Kotelnikov's theoretical development is followed by four pages of discussion of the engineering context. Particularly noteworthy is his consideration of the time-bandwidth trade-off (p. 15 of this document), and what we would now call M-ary signalling (pp. 16-17). Interestingly, he does not address the issue of noise, except in so far as he discusses power requirements for multi-level discrete signalling (implicitly to overcome noise). Indeed, he even states that for discrete signalling "the necessary frequency range can be reduced as much as desired", which is not quite how we would put it today. A decade or so later, of course, Kotelnikov took up all these matters in detail, with his work on optimal detection.

There is little doubt that Kotelnikov's 1933 paper was the first to address the problem of sampling a continuous, bandlimited signal in an engineering context, even though the mathematical basis of sampling had been considered earlier by a number of mathematicians. Russian work in the 1930s and 1940s was not known at the time in the West; indeed, only after the Second World War, and in the context of the Cold War, did translations of public-domain Russian scientific and engineering research become widely available[4]. Because of this (and also because of the tendency under Stalin for spurious claims to be made for Russian priority in science and technology) English-language histories of 20th century technology do not always recognise the significance of Russian contributions. This translation has attempted to redress the balance for one of the most important theoretical results of information engineering.

Kotelnikov in later years

References

Benedetto, J J and Ferreira, P J S G (2001), eds, *Modern Sampling Theory*, Birkhäuser, Boston/Basel/Berlin

Kotelnikov, V A (1933), On the capacity of the 'ether' and of cables in electrical communication. In: *Procs. of the first All-Union Conference on the technological reconstruction of the communications sector and low-current engineering*, Moscow.

Kotelnikov, V A (1959), *The Theory of Optimum Noise Immunity*, McGraw-Hill, New York/Toronto/London

[4] Before, and just after, the Russian Revolution, scientists often published their work in French and German (but rarely English) journals in addition to Russian periodicals. By the late 1930s this was severely discouraged by the Soviet state.

Lantsberg, H (2004), IEEE Life Fellow honoured by Russia's President Putin, *IEEE Region 8 News*, Vol. 7 No. 1 **(**February/March) http://www.ieee.org/r8. For more details see also the *Global Communications Newsletter*, December 2003, at http://www.comsoc.org/pubs/gcn/gcn1203.html (Both websites accessed 26 June 2005.)

Lüke, H D (1999), The origins of the sampling theorem, *IEEE Communications Magazine*, April, pp. 106-8

Nyquist, H (1928), Certain topics in telegraph transmission theory, *Trans. AIEE*, Vol. 47, pp. 363-390

Shannon, C E (1948), A mathematical theory of communication, *Bell System Technical Journal*, Vol. 27, pp. 379-423, 623-56

Shannon, C E (1949), Communication in the presence of noise, *Procs. IRE*, Vol. 37, pp. 10-21

A short biography of Kotelnikov (in Russian) can be found at http://www.biograph.comstar.ru/bank/kotelnik.htm (Accessed 26 June 2005.)

All-Union Power Engineering Committee

Proceedings of the first All-Union Conference on the technological reconstruction of the communications sector and the development of low-current engineering. Moscow, January 1933.

Radio Section

On the transmission capacity of the 'ether' and of cables in electrical communications

V A Kotelnikov

In both wired and radio engineering, any transmission requires the use of not simply a single frequency, but a whole range of frequencies. As a result, only a limited number of radio stations (broadcasting different programmes) can operate simultaneously. Neither is it possible to convey more than a given number of channels at any one time over a pair of wires, since the frequency band of one channel may not overlap that of another: such an overlap would lead to mutual interference.

In order to extend the capacity of the 'airwaves' or a cable (something that would be of enormous practical importance, particularly in connection with rapid developments in radio engineering, television transmission, etc), it is necessary either to reduce in some way the range of frequencies needed for a given transmission (without adversely affecting its quality), or to devise some way of separating channels whose frequencies overlap – perhaps even employing a method based not on frequency, as has been the case until now, but by some other means[5].

At the present time no technique along these lines permits, even theoretically, the capacity of the 'airwaves' or a cable to be increased any further than that corresponding to 'single side-band' transmission. So the question arises: is it possible, in general, to do this? Or will all attempts be tantamount to efforts to build a perpetual motion machine?

This question is currently very pressing in radio engineering, since each year sees an increase in the 'crowding of the airwaves'. It is particularly important to investigate it now in connection with the planning of scientific research, since in order to plan, it is important to know what is possible, and what is completely impossible, in order to direct efforts in the required manner.

In the present paper this question is investigated, and it is demonstrated that for television, and for the transmission of images with a full range of half tones, and also for telephony, there exists a fully determined minimum necessary frequency band, which cannot be reduced by any means without adversely affecting quality or speed of transmission. It is further demonstrated that for such transmissions it is impossible to increase either wireless or wired capacity by any means not based on frequency bands – or, indeed, any other method (except, of course, by the use of directional antennas for separate channels). The maximum possible capacity for these transmissions can be achieved through 'single sideband transmission', something fully achievable in principle at the present time.

For transmissions such as telegraphy, or for the transmission of images or television pictures without half tones, where the source may not change continuously, but is limited to specific, pre-determined values, it is demonstrated that the required bandwidth can be reduced as much as

[5] Indeed, it is sometimes possible to do this by means of directional antennas, but we shall consider here only the situation when this is not the case.

desired, without adversely affecting the quality or speed of transmission, but at the expense of increasing the power and the complexity of the equipment. One such method of bandwidth reduction is outlined in the present paper, with a discussion of the necessary power increase.

There is thus no theoretical limit to the capacity of either the ‘airwaves’ or a cable for transmissions of this kind; it is simply a matter of technical implementation.

The proof of these propositions in the present paper is independent of the method of transmission on the following grounds: in electrical communication any transmitter can transmit, and any receiver receive, only some function of time that cannot be completely arbitrary, since the frequencies of which it consists, or into which it may be decomposed, must lie within defined limits. In radio transmission such a function is converted into current strength in the transmitting antenna, which is interpreted more or less exactly by the receiver; in cable transmission it is the electromotive force at the transmitting end of the line. In both cases, the function to be transmitted will consist of a limited range of frequencies: since firstly very high and very low frequencies will not reach the receiver on account of propagation conditions; and secondly, frequencies outside a defined narrow band are normally eliminated deliberately, so as not to interfere with other channels.

This need to transmit using functions of time with limited bandwidth leads to a completely determined limitation to channel capacity, as will be shown below.

In order to prove the above assertions, let us consider functions of a fixed bandwidth.

Functions consisting of frequencies between 0 and f_1.

Theorem I

Any function F(t) consisting of frequencies between 0 and f_1 cycles per second can be represented by the series:

$$F(t) = \sum_{-\infty}^{\infty} D_k \frac{\sin \omega_1 \left(t - \frac{k}{2f_1}\right)}{t - \frac{k}{2f_1}}, \qquad (1)$$

where k is an integer, $\omega_1 = 2\pi f_1$, and D_k are constants depending on F(t).

Conversely, any function F(t) represented by the series (1) consists only of frequencies between 0 and f_1.

Proof

Any function F(t) satisfying the Dirichlet conditions[6] (finite number of maxima, minima and discontinuities in a given finite interval), and integrable from $-\infty$ to $+\infty$, which is always the case in electrical engineering, can be represented by the Fourier integral[7]:

[6] In this paper we shall consider only functions satisfying the Dirichlet conditions

[7] See, for example, Smirnov, *Course of Higher Mathematics*, Vol. II, 1931, p. 427

$$F(t) = \int_0^\infty C(\omega)\,\mathrm{Cos}\,\omega t\,d\omega + \int_0^\infty S(\omega)\,\mathrm{Sin}\,\omega t\,d\omega \qquad (2)$$

that is, as an infinite sum of sinusoidal oscillations with frequencies from 0 to ∞ and amplitudes C(ω) dω and S(ω) dω dependent on frequency.

Moreover

$$\left.\begin{aligned} C(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} F(t)\ \mathrm{Cos}\,\omega t\,dt \\ S(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} F(t)\ \mathrm{Sin}\,\omega t\,dt \end{aligned}\right\}. \qquad (3)$$

In our case, where F(t) consists only of frequencies between 0 and f_1 it is clear that C(ω) and S(ω) = 0 for $\omega > \omega_1 = 2\pi f_1$, and for this reason F(t) can be represented according to Equation 2 as:

$$F(t) = \int_0^{\omega_1} C(\omega)\,\mathrm{Cos}\,\omega t\,d\omega + \int_0^{\omega_1} S(\omega)\,\mathrm{Sin}\,\omega t\,d\omega \qquad (4)$$

Now, the functions C(ω) and S(ω), like any other, can always be represented by Fourier series over the interval $0 < \omega < \omega_1$. Moreover, if we so desire, these series will consist only of sine or only of cosine terms if we take double the interval – that is, $2\,\omega_1$ – as the period[8]. Then

$$C(\omega) = \sum_0^\infty A_k\,\mathrm{Cos}\,\frac{2\pi}{2\omega_1}\,k\omega \qquad (5a)$$

$$S(\omega) = \sum_0^\infty B_k\,\mathrm{Sin}\,\frac{2\pi}{2\omega_1}\,k\omega$$

(5b)

Introducing the following notation

$$\left.\begin{aligned} D_k &= \frac{A_k + B_k}{2} \\ D_{-k} &= \frac{A_k - B_k}{2} \end{aligned}\right\}. \qquad (6)$$

we can rewrite 5a and 5b as:

[8] See Smirnov, *Course of Higher Mathematics*, Vol. II, 1931, p. 385

$$\left.\begin{aligned} C(\omega) &= \sum_{-\infty}^{\infty} D_k \operatorname{Cos} \frac{\pi}{\omega_1} k\omega \\ S(\omega) &= \sum_{-\infty}^{\infty} D_k \operatorname{Sin} \frac{\pi}{\omega_1} k\omega \end{aligned}\right\} . \qquad (7)$$

Substituting expressions (7) into equation (4) leads, after some manipulation (see Appendix I), to equation (1), thus proving the first part of Theorem 1.

In order to prove the second part of the theorem, consider the special case of a function F(t) with a frequency spectrum contained within the interval between 0 and f_1, and with all but one of the coefficients of D_k equal to zero. Such a function clearly consists of only one term of series (1). Conversely, if F(t) consists of any one term of series (1), its frequency spectrum is confined to the range 0 to f_1. Hence any sum of arbitrary terms of the series, and thus the sum of all the terms, will always consist of frequencies between 0 and f_1.

All the terms of series (1) are of similar form, differing only in magnitude and displacement in time. The k^th term is shown in Figure 1; it has a maximum at time $t = k/2f_1$ and gradually reduces in magnitude on either side.

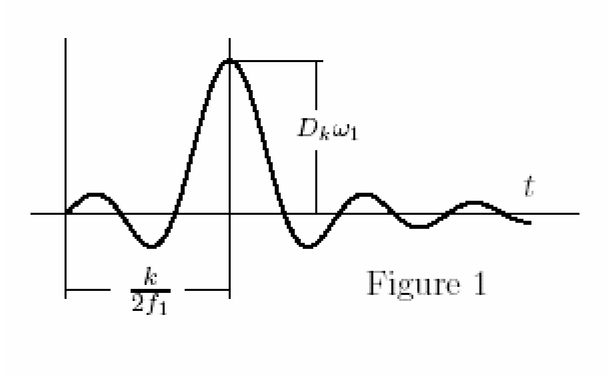


Figure 1

Theorem II

Any function F(t) consisting of frequencies between 0 and f_1 can be transmitted continuously, with arbitrary accuracy, by means of numbers sent at intervals of $1/2f_1$ seconds. Indeed, measuring the value of F(t) at times $t = n/2f_1$, where n is an integer, we obtain

$$F\left(\frac{n}{2f_1}\right) = D_n\,\omega_1 \qquad (8)$$

All the terms of series (1) are zero for this value of t, except for the term with k = n which, as may easily be obtained by expansion, will equal $D_n\omega_1$. Hence after each interval of $1/2f_1$ seconds we can determine the next D_k. Having transmitted these D_k in turn at intervals of $1/2f_1$ seconds we can, according to equation (1), reconstruct F(t) to any degree of accuracy.

Theorem III

It is possible to transmit, continuously and uniformly, arbitrary values D_k at a rate of N per second using a function F(t) with vanishingly small frequency components greater than $f_1 = N/2$.

So, on obtaining each given value, we construct a function $F_k(t)$ such that:

$$\left.\begin{array}{ll}
t < \frac{k}{2f_1} - T & F_k(t) = 0 \\[2ex]
\frac{k}{2f_1} - T < t < \frac{k}{2f_1} + T & F_k(t) = D_k \dfrac{\mathrm{Sin}\,\omega_1\left(t - \frac{k}{2f_1}\right)}{t - \frac{k}{2f_1}} \\[2ex]
t > \frac{k}{2f_1} + T & F_k(t) = 0
\end{array}\right\} \quad (9)$$

and transmit their sum F(t). If T = ∞, then the function F(t) consists exclusively of frequencies less than f_1, corresponding to series (1). Unfortunately, however, it is not possible to construct such an infinite sequence of terms, so we must restrict ourselves to finite T. Let us now prove, therefore, that the greater the value of T, the smaller the amplitudes of those components $f > f_1$, and that such amplitudes can be made arbitrarily small. To do this, we find the amplitudes C(ω) and S(ω) for the function (9), by substituting it into equation (3). We obtain

$$C(\omega) = \frac{1}{\pi} \int\limits_{\frac{k}{2f_1} - T}^{\frac{k}{2f_1} + T} D_k \frac{\mathrm{Sin}\,\omega_1\left(t - \frac{k}{2f_1}\right)}{t - \frac{k}{2f_1}} \,\mathrm{Cos}\,\omega t\, dt$$

$$S(\omega) = \frac{1}{\pi} \int\limits_{\frac{k}{2f_1} - T}^{\frac{k}{2f_1} + T} D_k \frac{\mathrm{Sin}\,\omega_1\left(t - \frac{k}{2f_1}\right)}{t - \frac{k}{2f_1}} \,\mathrm{Sin}\,\omega t\, dt$$

(10)

and after integrating (see Appendix II) we obtain:

$$\left.\begin{array}{l}
C(\omega) = \frac{D_k}{\pi}\,\mathrm{Cos}\,\omega\,\frac{k}{2f_1}\,\left[\mathrm{Si}\,T\,(\omega+\omega_1) - \mathrm{Si}\,T\,(\omega-\omega_1)\right] \\
S(\omega) = \frac{D_k}{\pi}\,\mathrm{Sin}\,\omega\,\frac{k}{2f_1}\,\left[\mathrm{Si}\,T\,(\omega+\omega_1) - \mathrm{Si}\,T\,(\omega-\omega_1)\right]
\end{array}\right\} \quad (11)$$

In this expression Si denotes the sine integral – that is, the function:

$$\mathrm{Si}\,x = \int\limits_0^x \frac{\mathrm{Sin}\,y}{y}\,dy \qquad (12)$$

The value of this function can be calculated and has been tabulated[9]; it is plotted in Figure 2.

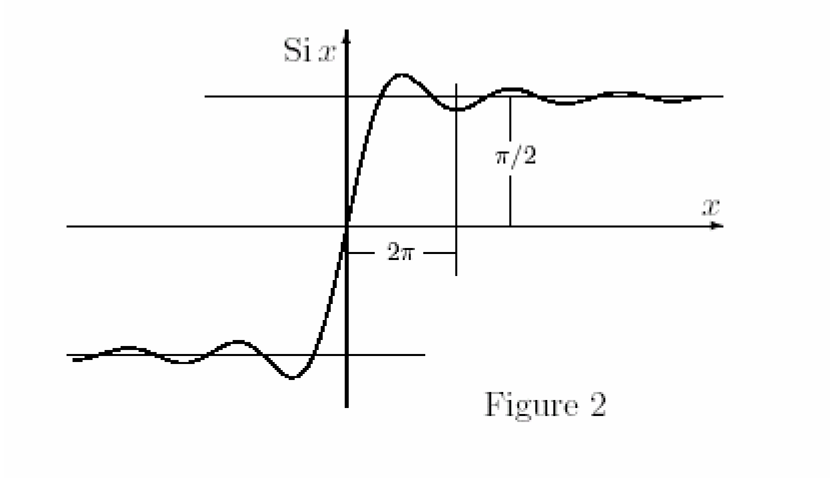
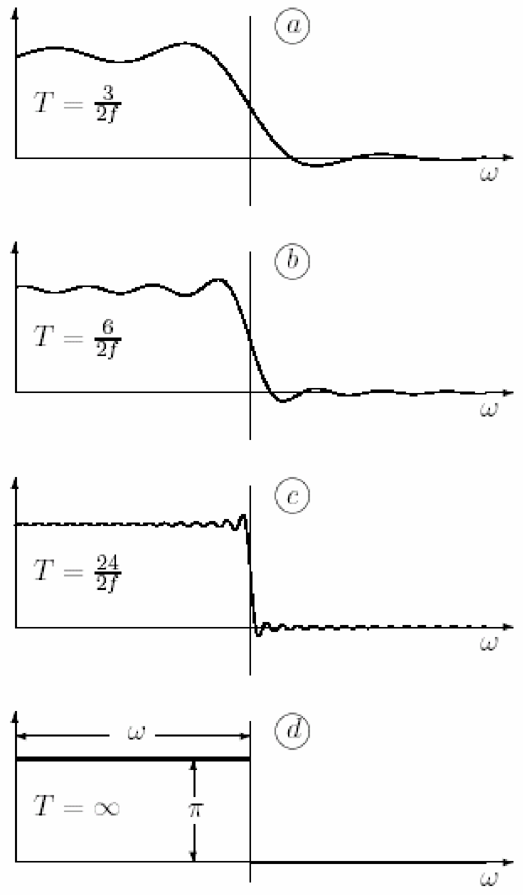


Figure 2

As can be seen from the figure, Si x tends to k [sic] ± π/2 as x tends to ± ∞.

Let us consider the value of the expression in square brackets in Equation 11. Figure 3a shows the plot of this for $T = 3/2f_1$; Figure 3b for $T = 6/2f_1$; Figure 3c for $T = 24/2f_1$; and Figure 3d for $T = \infty$.

Figure 3



[9] See, for example, E. Jah[n]ke and F. Emde, *Funktiontafeln mit Formeln und Kurven*.

As can be seen from these figures, the value in square brackets in Equation 11 tends to that shown in Figure 3d as T increases. That is, for $\omega > \omega_1$, [ ] = 0; for $\omega < \omega_1$, [ ] = π. This is also evident directly from Equation 11, since increasing T corresponds to 'scaling' ω, and [the oscillations] of Si die rapidly away.

Hence the resulting sum of $F_k(t)$ will have arbitrary small amplitudes for frequencies $f > f_1$, providing T is sufficiently large. It is easy to recover the transmitted numbers D_k from the received function F(t), since at time $t = n/2f_1$ all the terms are zero with the exception of that for which k = n, which is equal to $D_n\omega$. Hence

$F(n/2f_1) = D_n\omega$

In this manner, by measuring the value of our function at times $t = k/2f_1$ we are able to obtain the value of a new D_k after each $1/2f_1$ second. In other words, we receive $N = 2f_1$ numbers per second, and the theorem is proved.

Functions consisting of frequencies from f_1 to f_2.

Let us now prove the following:

Theorem IV

Any function F(t) consisting of frequencies from f_1 to f_2 can be represented as:

$$F(t) = F_1(t)\,\mathrm{Cos}\,\frac{\omega_1+\omega_2}{2}\,t + F_2(t)\,\mathrm{Sin}\,\frac{\omega_1+\omega_2}{2}\,t \qquad (13)$$

where $\omega_1 = 2\pi f_1$, $\omega_2 = 2\pi f_2$, and $F_1(t)$ and $F_2(t)$ are some functions consisting of frequencies between 0 and $f = (f_2 - f_1)/2$. Conversely, if in Equation 13 $F_1(t)$ and $F_2(t)$ are some functions consisting of frequencies between 0 and $f = (f_2 - f_1)/2$, then F(t) consists of frequencies from f_1 to f_2.

If F(t) consists only of frequencies from f_1 to f_2, then evidently $C(\omega)$ and $S(\omega)$ for such a function may be represented as:

$C(\omega) = S(\omega) = 0$ for $\omega > \omega_2$ or $\omega < \omega_1$,

$$\left.\begin{aligned} C(\omega) &= \sum_0^\infty A_k\,\mathrm{Cos}\frac{\pi k}{2(\omega_2-\omega_1)}(\omega-\omega_1) \\ S(\omega) &= \sum_0^\infty B_k\,\mathrm{Sin}\frac{\pi k}{2(\omega_2-\omega_1)}(\omega-\omega_1) \end{aligned}\right\}\ \text{при}\ \omega_1 < \omega < \omega_2$$

or, reintroducing the convention

$$\left.\begin{aligned} D_k &= \frac{A_k+B_k}{2} \\ D_{-k} &= \frac{A_k-B_k}{2} \end{aligned}\right\}. \qquad (6)$$

we obtain

$$\left.\begin{aligned} C(\omega) &= \sum_{-\infty}^{\infty} D_k \,\mathrm{Cos}\, \frac{\pi}{\omega_2-\omega_1}\, k\,(\omega-\omega_1) \\ S(\omega) &= \sum_{-\infty}^{\infty} D_k \,\mathrm{Sin}\, \frac{\pi}{\omega_2-\omega_1}\, k\,(\omega-\omega_1) \end{aligned}\right\} \qquad (14)$$

for $\omega_1 < \omega < \omega_2$ and $C(\omega) = S(\omega) = 0$ for $\omega > \omega_2$ or $\omega < \omega_1$.

Substituting (14) into (2) we obtain, after integration and some manipulation (see Appendix III):

$$F(t) = \left[2\sum_{-\infty}^{\infty} (-1)^n D_{2n} \frac{\mathrm{Sin}\dfrac{\omega_2-\omega_1}{2}\left(t - \dfrac{k}{f_2-f_1}\right)}{t - \dfrac{k}{f_2-f_1}} \right] \mathrm{Cos}\,\frac{\omega_1+\omega_2}{2}\,t \;+$$

$$+ \left[2\sum_{-\infty}^{\infty} (-1)^n D_{2n+1} \frac{\mathrm{Sin}\dfrac{\omega_2-\omega_1}{2}\left(t - \dfrac{k+\frac{1}{2}}{f_2-f_1}\right)}{t - \dfrac{k+\frac{1}{2}}{f_2-f_1}} \right] \mathrm{Sin}\,\frac{\omega_1+\omega_2}{2}\,t \qquad (15)$$

That is,

$$F_1(t) = 2\sum_{-\infty}^{\infty} (-1)^n D_{2n} \frac{\mathrm{Sin}\dfrac{\omega_2-\omega_1}{2}\left(t - \dfrac{k}{f_2-f_1}\right)}{t - \dfrac{k}{f_2-f_1}} \qquad (16)$$

$$F_2(t) = 2\sum_{-\infty}^{\infty} (-1)^n D_{2n+1} \frac{\mathrm{Sin}\dfrac{\omega_2-\omega_1}{2}\left(t - \dfrac{k+\frac{1}{2}}{f_2-f_1}\right)}{t - \dfrac{k+\frac{1}{2}}{f_2-f_1}} \qquad (17)$$

Recalling that, according to Theorem I, $F_1(t)$ and $F_2(t)$ must possess a frequency spectrum between 0 and $f = (f_2 - f_1)/2$, since the series (16) and (17) differ from series (1) only in notation, we can consider the first part of Theorem IV as proved.

Since, according to Theorem I, it is possible to represent any functions $F_1(t)$ and $F_2(t)$ consisting of frquencies between 0 and $f = (f_2 - f_1)/2$ by means of series (16) and (17), and since no conditions are imposed on the coefficients D_k appearing in these series, then it is clear that the second part of Theorem IV is also true.

Let us now prove two theorems which are generalizations of Theorems I and II.

Theorem V

Any function F(t) consisting of frequencies from f_l to f_2 can be transmitted continuously with any accuracy by means of numbers sent at intervals of $1/[2(f_2-f_l)]$ seconds.

At time $t = k/(f_l + f_2)$ (where k is an integer) we obtain, according to Equation 13

$$F\left(\frac{k}{f_2+f_1}\right) = F_1\left(\frac{k}{f_2+f_1}\right) \tag{18}$$

since at this value of t the cosine term equals 1 and the sine term 0. When

$$t = \frac{k+\frac{1}{2}}{f_2+f_1}$$

we obtain

$$F\left(\frac{k+\frac{1}{2}}{f_2+f_1}\right) = F_2\left(\frac{k+\frac{1}{2}}{f_2+f_1}\right)$$

for analogous reasons.

Hence, after each $1/(f_2 + f_l)$ seconds we shall be in a position to determine the individual values of $F_1(t)$ and $F_2(t)$. From these values we can recover the functions $F_1(t)$ and $F_2(t)$ themselves, since according to Theorem II this rate allows us to recover functions consisting of frequencies between 0 and $(f_2 + f_l)/2$, whereas $F_1(t)$ and $F_2(t)$ consist only of frequencies between 0 and $(f_2 – f_l)/2$.

According to Theorem II, each of the functions obtained in this way, with frequencies between 0 and $(f_2 – f_l)/2$, can be transmitted by numbers sent at intervals of $1/(f_2-f_l)$ seconds. The two functions can thus be transmitted simultaneously by means of numbers sent at intervals of $1/2(f_2-f_l)$ seconds; from these numbers, once $F_1(t)$ and $F_2(t)$ have been recovered we can recover F(t) itself according to Equation 13.

Theorem VI

It is possible to transmit arbitrary numbers D_k continuously and uniformly at a rate of N numbers per second by means of a function F(t) with arbitrarily small frequency components $f > f_2$ and $f < f_l$ (that is, in practice they can be neglected) providing that

$$N = 2(f_2 - f_1) \tag{19}$$

According to Theorem III we can transmit N numbers per second by means of two functions $F_1(t)$ and $F_2(t)$, each having arbitrarily small components with frequencies greater than $(f_2 – f_l)/2$.

These functions may also be transmitted uniformly by means of the function F(t), having arbitrarily small components for $f > f_2$ and $f < f_l$. Indeed, according to Equation 13 we obtain F(t) from $F_1(t)$

and $F_2(t)$. By transmitting $F(t)$ we can, as shown above, recover $F_1(t)$ and $F_2(t)$ and thereby the transmitted numbers.

In order to prove the final theorem, which states that it is impossible to transmit an indefinite amount by means of a function with a restricted bandwidth, let us prove the following lemma:

Lemma

It is impossible to transmit N arbitrary numbers by means of M numbers if

$$M < N \tag{20}$$

Assume that it can be done. Then, clearly, the M numbers $m_1 \ldots m_M$ are some functions of the N numbers $n_1 \ldots n_N$. That is

$$\left.\begin{array}{lcl} m_1 & = & \varphi_1(n_1 \ldots n_N) \\ m_2 & = & \varphi_2(n_1 \ldots n_N) \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots\ldots\ldots} \\ m_M & = & \varphi_M(n_1 \ldots n_N) \end{array}\right\} \tag{21}$$

and we have to recover the numbers $n_1 \ldots n_N$ from a knowledge only of the M numbers $m_1 \ldots m_M$ and, of course, the functions $\varphi_1 \ldots \varphi_M$. But this is equivalent to solving M equations with N unknowns, which is impossible if the number of equations is less than the number of unknowns – that is, if the inequality (20) holds.

Theorem VII

It is possible continuously to transmit arbitrary numbers, in a uniform sequence, at a rate of N per second, and M arbitrary functions $F_1(t) \ldots F_M(t)$ with bandwidths $\Delta f_1 \ldots \Delta f_M$, by means of a continuous sequence of numbers at a rate N′ per second and M′ functions $F'_1(t) \ldots F'_M(t)$ with bandwidths $\Delta f'_1 \ldots \Delta f'_M$, if

$$N + 2\sum_{1}^{M} \Delta f_k \leq N' + 2\sum_{1}^{M'} \Delta f'_k \tag{22}$$

It is impossible to do this in any way if

$$N + 2\sum_{1}^{M} \Delta f_k > N' + 2\sum_{1}^{M'} \Delta f'_k \tag{23}$$

The first part of this theorem can be proved on the basis of theorems V and VI. According to Theorem V we can transmit our N numbers per second and M curves by means of P numbers per second if

$$P = N + 2\sum_{1}^{M} \Delta f_k \tag{24}$$

These P numbers per second can be transmitted partly by means of N′ numbers per second and partly, by Theorem VI, using the curves $F'_1(t)$ ... $F'_M(t)$ providing the equality (22) holds.

The second part of the theorem we prove by *reductio ad absurdum* from a lemma. Suppose we need to transmit P arbitrary numbers per second. From Theorem VI this is possible by transmitting N numbers per second and the functions $F_1(t)$ ... $F_M(t)$ with bandwidths Δf_1 ... Δf_M providing the equality (24) holds. But these numbers and functions, if the second part of the theorem is incorrect, could have been transmitted by means of the functions $F'_1(t)$ ... $F'_M(t)$ and N′ numbers per second. The latter numbers and functions, according to Theorem V, can be transmitted by P′ numbers per second if

$$P' = N' + 2\sum_{1}^{M'} \Delta f'_k \qquad (25)$$

In other words, we could transmit continuously P numbers per second using P′ numbers per second even though, from equalities (24) and (25) and inequality (23), $P > P'$.

Hence the assumption that the second part of the theorem is false leads to an inadmissible, contradictory result.

Channel capacity for telephone transmission

Speech, music, and other objects of telephone transmission are arbitrary functions of time, having a frequency spectrum whose width is completely defined and which depends on how well we wish to represent the [original] sound.

When transmitting this function by wire or radio, we convert it into another function of time, and it is the latter, strictly speaking, that we transmit. Moreover, for continuous transmission, this latter function may not, according to Theorem VII, have a spectrum of frequencies narrower than that of the audible frequencies that we wish to transmit. Thus, a continuous telephone transmission may not occupy a frequency band in the ether or on a wire narrower than the width of the spectrum of the audible frequencies required for the given transmission. This is true independently of the method of transmission, and it is impossible to invent any method that would allow a narrower bandwidth to be used for continuous transmission. Such a minimum spectral width can be achieved at present by means of single sideband transmission, as is well known.

The proviso “for continuous transmission” is of great importance, since by means of transmission with interruptions any sounds - music say – can occupy less bandwidth than that of the original audible spectrum. To do this, it is sufficient first to record the music to be transmitted on gramophone records, and then to transmit from them but playing back at half the speed, say, of the recording. Then every frequency will take on half its normal value, and the transmission will require half the bandwidth. The original can similarly be recovered by gramophone. It is clear that such transmission cannot increase channel capacity, since the “ether” or the cable will be occupied the whole time, while communication will proceed with interruptions.

This does not contradict Theorem VII, since the latter states “it is impossible continuously to transmit an arbitrary function”, and by using this kind of transmission we can transmit either an arbitrary function (with interruptions), or a not entirely arbitrary function having known breaks (continuously).

From Theorem VII it also follows that it is impossible to increase channel capacity by using any selection of a non-frequency character (apart from directional antennas). If it were possible to do this, such a method would mean that, for example, it would be possible to transmit simultaneously from one place to another n telephone channels each requiring bandwidth Δf, using a total frequency band of less than n Δf. But in this case the field (or, in a wire, current) intensities from different channels would be mixed into a single function of time with a bandwidth less than n Δf which would appear at the receivers. We should have transmitted n functions of time with bandwidth Δf with the aid of a single function of bandwidth less than n Δf, which according to Theorem VII is completely impossible.

From what has been said, it is clear that for telephony the only way to increase the capacity of the "airwaves" is to use directional antennas or to extend the exploitation of the frequency spectrum to the ultra-short wave region.

Transmission of images and television with a full range of half tones

For the transmission of images and television we need to transmit the level of blackness of some N elements per second, which is equivalent to transmitting arbitrary numbers at the speed of N per second. If we wish to do this using a function of time, as is always the case in practice, then according to Theorem VII a frequency band of not less than N/2 cps is required. Thus it is obvious immediately that here, too, it is impossible to reduce the bandwidth to less than that required for single sideband transmission. Indeed, the implementation even of this may encounter great technical difficulties owing to possible phase distortion during transmission.

Neither is it possible to reduce the bandwidth by means of any "group scanning" of the image – that is, not scanning the elements individually – since with such scanning one still has to transmit, even if by another method, the blackness level of the same N elements per second, and hence N arbitrary numbers per second, which cannot be achieved in any way in a reduced frequency band. Selection methods not based on frequency cannot help here (not including directional antennas) for exactly the same reason as in telephone transmission.

Telegraph transmission and the transmission of images without half tones or with a limited number of them

In telegraph transmission, and also in the transmission of images without half tones or with a limited number of half tones specified in advance, we again have to deal with the transmission of some N elements per second, equivalent to the transmission of N numbers per second. In this case, however, the magnitudes of these elements, and thus the numbers, are not completely arbitrary, but must take on completely determined values known in advance. It is therefore not possible to apply the theorems derived above directly, because these are concerned with the transmission of arbitrary numbers completely unknown in advance.

In fact, for such transmission the necessary frequency range can be reduced as much as desired and hence, at least in theory, the channel capacity can be increased indefinitely. We can proceed in the following manner. Suppose we wish to transmit, at a rate of N per second, elements that can take on the values 0 and 1 only, and to use a bandwidth of only N/4 rather than the N/2 of Theorem VII. In order to do this we shall transmit two such elements by means of a single element or number, as given in the following table. Here column I holds the value of the first element, column II the second, and column III the value of the element by means of which we wish to transmit the other two.

I	*II*	*III*
0	0	0
1	0	1
1	1	2
0	1	3

In this way we can transmit N elements per second, each of which can take on one of two values, by means of N/2 elements per second, each of which can take on four values; the latter, according to Theorem VII, can be transmitted using a bandwidth N/4.

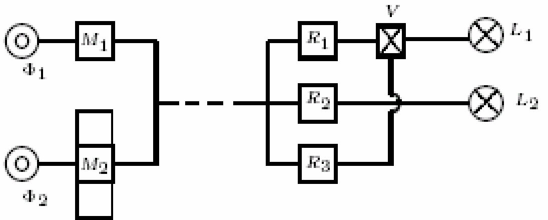


Fig.4

In practice such a replacement of two elements by one can be realised as illustrated in Figure 4, where φ_1 and φ_2 are two photoelectric cells or telegraph transmitters: φ_1 activates modulator M_1 which sends to the line an amplitude of value 1; φ_2 drives modulator M_2 with an amplitude 3. When acting together φ_1 and φ_2, which are connected in an opposing sense, result in an amplitude of 2 being sent. The received signal is fed to three receivers: the first begins to operate at amplitude 1, the second at amplitude 2 and the third at amplitude 3. The first receiver drives output L_1, the second L_2, while the third disconnects receiver 1 from output L_1 on receiving amplitude 3. With the aid of such a scheme we achieve the reduction in bandwidth discussed above.

In such transmission, in view of the fact that it is necessary to distinguish between four instead of two levels in the received signal, it is obvious that the power of the transmitter has to be increased by a factor of $3^2 = 9$ in comparison with the usual transmission.

In an analogue fashion it is possible to reduce the frequency band by a factor of n, transmitting n elements, each of which can take on one of two values, by a single element taking on one of 2^n values (the number of combinations of n elements, each taking on one of two values). But for this it is necessary to increase the power by a factor of $(2^n -1)^2$.

For the transmission of images with a predefined number of half tones each element will take on several values, m say (in this case m > 2). In order to decrease the bandwidth by n times in such a transmission it is possible to replace n transmitted elements by one with m^n possible values (the number of possible combinations of n elements each having m possible values). Then the power must clearly be increased by a factor

$$\frac{(m^n - 1)^2}{(m - 1)^2}$$

As an be seen, reducing the bandwidth in this way requires an enormous increase in power. Furthermore, such methods are very bad for shortwave transmission owing to fading. For wired communication, however, such bandwidth reduction may be of current practical importance since the powers are necessarily small and received strengths do not vary rapidly.

Appendix I

Substituting expressions (7) into Equation 4 we obtain

$$F(t) = \int_{0}^{\omega_1} \sum_{-\infty}^{\infty} D_k \operatorname{Cos} \frac{\pi}{\omega_1} k\omega \cdot \operatorname{Cos} \omega t \, d\omega \; +$$

$$+ \int_{0}^{\omega_1} \sum_{-\infty}^{\infty} D_k \operatorname{Sin} \frac{\pi}{\omega_1} k\omega \cdot \operatorname{Sin} \omega t \, d\omega \; =$$

$$= \sum_{-\infty}^{\infty} D_k \int_{0}^{\omega_1} \left(\operatorname{Cos} \frac{\pi}{\omega_1} k\omega \cdot \operatorname{Cos} \omega t + \operatorname{Sin} \frac{\pi}{\omega_1} k\omega \cdot \operatorname{Sin} \omega t \right) d\omega \; =$$

$$= \sum_{-\infty}^{\infty} D_k \int_{0}^{\omega_1} \operatorname{Cos} \, \omega \left(t - \frac{\pi}{\omega_1} k \right) d\omega$$

or, integrating and substituting $2\pi f_1$ for ω_1 in the brackets:

$$F(t) = \sum_{-\infty}^{\infty} D_k \frac{\operatorname{Sin} \omega_1 \left(t - \dfrac{k}{2f_1} \right)}{t - \dfrac{k}{2f_1}}$$

Appendix II

In the expression:

$$C(\omega) = \frac{1}{\pi} \int\limits_{\frac{k}{2f_1} - T}^{\frac{k}{2f_1} + T} D_k \, \frac{\mathrm{Sin}\ \omega_1 \left(t - \dfrac{k}{2f_1} \right)}{t - \dfrac{k}{2f_1}} \ \mathrm{Cos}\ \omega t \ dt$$

let us substitute

$$t = u + \frac{k}{2f_1}, \qquad dt = du,$$

Then

$$\begin{aligned} C(\omega) \ &= \ \frac{1}{\pi} \int\limits_{-T}^{T} D_k \ \frac{\mathrm{Sin}\ \omega_1 u}{u} \ \mathrm{Cos}\ \omega \left(u + \frac{k}{2f_1} \right) du \ = \\ &= \ \frac{1}{\pi} \int\limits_{-T}^{T} D_k \ \frac{\mathrm{Sin}\ \omega_1 u \ \cdot \ \mathrm{Cos}\ \omega u}{u} \ \mathrm{Cos}\ \omega \frac{k}{2f_1} \ du \ + \\ &+ \ \frac{1}{\pi} \int\limits_{-T}^{T} D_k \ \frac{\mathrm{Sin}\ \omega_1 u \ \cdot \ \mathrm{Sin}\ \omega u}{u} \ \mathrm{Sin}\ \omega \frac{k}{2f_1} \ du. \end{aligned}$$

The term under the second integral has the same magnitude but opposite sign for positive and negative u and for this reason the second integral is zero. The value of the term under the first integral does not change if u is replaced by –u so the integral can be taken from 0 to T and multiplied by 2. Hence

$$C(\omega) \ = \ \frac{2D_k}{\pi} \ \ \mathrm{Cos}\frac{k}{2f_1} \int\limits_{0}^{T} \frac{\mathrm{Sin}\ \omega_1 u\ \mathrm{Cos}\,\omega u}{u} \ du$$

or

$$C(\omega) = \frac{D_k}{\pi} \ \mathrm{Cos}\,\frac{k}{2f_1} \left[\int_0^T \frac{\mathrm{Sin}\,(\omega_1 + \omega)\,u}{u}\,du - \int_0^T \frac{\mathrm{Sin}\,(\omega - \omega_1)\,u}{u}\,du \right]$$

Substituting in the first integral

$$(\omega_1 + \omega)\,u = y,$$

and in the second

$$(\omega - \omega_1)\,u = y_1$$

we have

$$C(\omega) = \frac{D_k}{\pi} \ \mathrm{Cos}\,\frac{k}{2f_1} \left[\int_0^{(\omega+\omega_1)T} \frac{\mathrm{Sin}\ y}{y}\,dy - \int_0^{(\omega-\omega_1)T} \frac{\mathrm{Sin}\ y_1}{y_1}\,dy_1 \right]$$

The integrals in brackets cannot be evaluated analytically. Clearly, they are some functions of the upper limits. They are denoted integral sine functions, and introducing this notation we have

$$C(\omega) = \frac{D_k}{\pi}\ \mathrm{Cos}\,\frac{\omega k}{2f_1} \left[\mathrm{Si}\ T\,(\omega + \omega_1) - \mathrm{Si}\ T\,(\omega - \omega_1) \right].$$

Proceeding similarly with S(ω) we obtain Equation 11.

Appendix III

Substituting Equation 14 into Equation 2 we have

$$F(t) \;=\; \int_{\omega_1}^{\omega_2} \sum_{-\infty}^{\infty} D_k \,\mathrm{Cos}\, \frac{\pi k}{\omega_2-\omega_1}(\omega-\omega_1)\;\;\mathrm{Cos}\,\omega t\; d\omega \;+$$

$$+\int_{\omega_1}^{\omega_2} \sum_{-\infty}^{\infty} D_k\;\mathrm{Sin}\,\frac{\pi k}{\omega_2-\omega_1}(\omega-\omega_1)\;\;\mathrm{Sin}\,\omega t\; d\omega.$$

The limits are from ω_1 to ω_2 because $C(\omega) = S(\omega) = 0$ for $\omega < \omega_1$ or $\omega > \omega_2$. After trigonometrical manipulation we have:

$$F(t) \;=\; \sum_{-\infty}^{\infty} D_k \int_{\omega_1}^{\omega_2} \mathrm{Cos}\left[\omega\left(t-\frac{\pi k}{\omega_2-\omega_1}\right)+\frac{\pi k\omega_1}{\omega_2-\omega_1}\right] d\omega \;=$$

$$= \sum_{-\infty}^{\infty} D_k\; \frac{\mathrm{Sin}\left[\omega_2\left(t-\frac{\pi k}{\omega_2-\omega_1}\right)+\frac{\pi k\omega_1}{\omega_2-\omega_1}\right]}{t-\frac{\pi k}{\omega_2-\omega_1}} \;-$$

$$\frac{-\;\mathrm{Sin}\left[\omega_1\left(t-\frac{\pi k}{\omega_2-\omega_1}\right)+\frac{\pi k\omega_1}{\omega_2-\omega_1}\right]}{t-\frac{\pi k}{\omega_2-\omega_1}}$$

Replacing the difference of the sines by a product, and simplifying, we have

$$F(t) \;=\; 2\sum_{-\infty}^{\infty} D_k\;\mathrm{Cos}\left(\frac{\omega_2+\omega_1}{2}\,t-\frac{\pi}{2}\,k\right)\;\frac{\mathrm{Sin}\left[\frac{\omega_2-\omega_1}{2}\left(t-\frac{k}{2(f_2-f_1)}\right)\right]}{t-\frac{k}{2(f_2-f_1)}}$$

or, grouping even and odd k together, we obtain Equation 15.

Conclusions

1. In the light of the ‘crowding of the airwaves’ that is already occurring, and in connection with the further rapid development of radio engineering, particularly the development of shortwave telephony and image transmission, the question of methods for increasing the channel capacity of the ‘ether’ must be addressed with great urgency by scientific research institutes. The question of increasing the channel capacity of wire communications is also of great economic importance, and should also be studied.

2. Since it is impossible to increase the capacity of either the airwaves or cables for telephone or image transmission beyond that required for single sideband transmission (for example, by overlaying the frequency bands of separate channels and then separating them), all attempts of this nature should be abandoned as unrealizable.

3. For telegraphy and the transmission of images without half tones or with a limited number of them, channel capacity may, in theory, be increased indefinitely, although only at the expense of greatly increased power and equipment complexity. Because of this it appears that such methods of bandwidth reduction will find application in the immediate future only in wire communications; effort should be devoted to this area.

4. For telephony and image transmission with a full range of half tones every effort should be directed towards the development of methods for single-sideband transmission and reception, since these allow the most efficient use of ‘ether’ and cable. The aim of research and development should be the improvement and simplification of equipment, which is currently very complex.

5. It is necessary to study the question of increasing the channel capacity of the ‘ether’ by means of directional antennae for both transmission and reception.

6. It is necessary to increase the range of frequencies exploited in the ultra shortwave region, as well as to research this frequency band.

7. It is necessary to study the question of improving the stability of radio station frequencies, since this allows more effective use of the ‘ether’.