On Shannon and "Shannon's formula"

Lars Lundheim
Department of Telecommunication, Norwegian University of Science and Technology (NTNU)

The period between the middle of the nineteenth and the middle of the twentieth century represents a remarkable period in the history of science and technology. During this epoch, several discoveries and inventions removed many practical limitations of what individuals and societies could achieve. Especially in the field of communications, revolutionary developments took place such as high speed railroads, steam ships, aviation and telecommunications.

It is interesting to note that as practical limitations were removed, several fundamental or principal limitations were established. For instance, Carnot showed that there was a fundamental limit to how much energy could be extracted from a heat engine. Later this result was generalized to the second law of thermodynamics. As a result of Einstein's special relativity theory, the existence of an upper velocity limit was found. Other examples include Kelvin's absolute zero, Heissenberg's uncertainty principle and Gödel's incompleteness theorem in mathematics. Shannon's Channel coding theorem, which was published in 1948, seems to be the last one of such fundamental limits, and one may wonder why all of them were discovered during this limited time-span. One reason may have to do with maturity. When a field is young, researchers are eager to find out what can be done – not to identify borders they cannot pass. Since telecommunications is one of the youngest of the applied sciences, it is natural that the more fundamental laws were established at a late stage.

In the present paper we will try to shed some light on developments that led up to Shannon's information theory. When one compares the generality and power of explanation of Shannon's paper "A Mathematical Theory of Communication" [1] to alternative theories at the time, one can hardly disagree with J. R. Pierce who states that it "came as a bomb" [4]. In order to see the connection with earlier work, we will, therefore, focus on one particular case of Shannon's theory, namely the one which is sometimes referred to as "Shannon's formula". As will be shown, this result was discovered independently by several researchers, and serves as an illustration of a scientific concept whose time had come. Moreover, we will try to see how development in this field was spurred off by technological advances, rather than theoretical studies isolated from practical life.

Besides the original sources cited in the text, this paper builds on historical overviews, such as [4] and [19]-[23].

"Shannon's formula"

Sometimes a scientific result comes quite unexpected as a "stroke of genius" from an individual scientist. More often a result is gradually revealed, by several independent research groups, and at a time which is just ripe for the particular discovery. In this paper we will look at one particular concept, the channel capacity of a band-limited information transmission channel with additive white, Gaussian noise. This capacity is given by an expression often known as "Shannon's formula[1]":

$$C = W \log_2(1 + P/N) \text{ bits/second.} \qquad (1)$$

We intend to show that, on the one hand, this is an example of a result for which time was ripe exactly a few years after the end of World War II. On the other hand, the formula represents a special case of Shannon's information theory[2] presented in [1], which was clearly ahead of time with respect to the insight generally established.

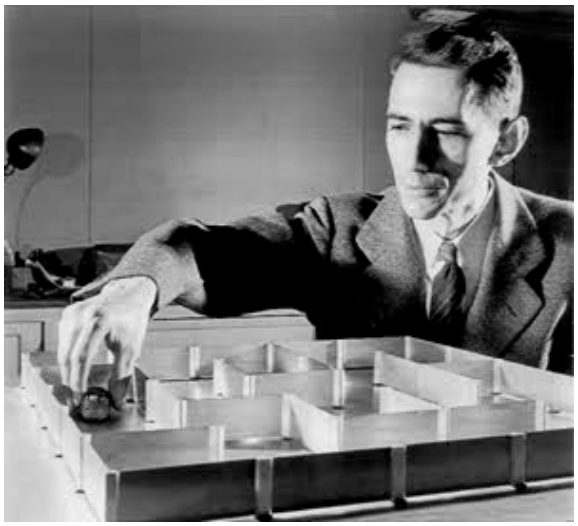
[1] Many mathematical expressions are connected with Shannon's name. The one quoted here is not the most important one, but perhaps the most well-known among communications engineers. It is also the one with the most immediately understandable significance at the time it was published.
[2] For an introduction to Shannon's work, see the paper by N. Knudtzon in this issue.

"Shannon's formula" (1) gives an expression for how many bits of information can be transmitted without error per second over a channel with a bandwidth of W Hz, when the average signal power is limited to P watt, and the signal is exposed to an additive, white (uncorrelated) noise of power N with Gaussian probability distribution. For a communications engineer of today, all the involved concepts are familiar – if not the result itself. This was not the case in 1948. Whereas bandwidth and signal power were well-established, the word *bit* was seen in print for the first time in Shannon's paper. The notion of probability distributions and stochastic processes, underlying the assumed noise model, had been used for some years in research communities, but was not part of an ordinary electrical engineer's training.

The essential elements of "Shannon's formula" are:
1. Proportionality to bandwidth W
2. Signal power S
3. Noise power P
4. A logarithmic function

The channel bandwidth sets a limit to how fast symbols can be transmitted over the channel. The signal to noise ratio (P/N) determines how much information each symbol can represent. The signal and noise power levels are, of course, expected to be measured at the receiver end of the channel. Thus, the power level is a function both of transmitted power and the attenuation of the signal over the transmission medium (channel).

The most outstanding property of Shannon's papers from 1948 and 1949 is perhaps the unique combination of generality of results and clarity of exposition. The concept of an information source is generalized as a symbol-generating mechanism obeying a certain probability distribution. Similarly, the channel is expressed essentially as a mapping from one set of symbols to another, again with an associated probability distribution. Together, these two abstractions make the theory applicable to all kinds of communication systems, man-made or natural, electrical or mechanical.



Claude Elwood Shannon (1916-2001), the founder of information theory, had also a practical and a playful side. The photo shows him with one of his inventions: a mechanical "mouse" that could find its way through a maze. He is also known for his electronic computer working with roman numerals and a gasoline-powered pogo stick.

Independent discoveries

One indicator that the time was ripe for a fundamental theory of information transfer in the first post-war years is given in the numerous papers attempting at such theories published at that time. In particular, three sources give formulas quite similar to (1). The best known of these is the book entitled *Cybernetics* [2] published by Wiener in 1949. Norbert Wiener was a philosophically inclined and proverbially absent-minded professor of mathematics at MIT. Nonetheless, he was deeply concerned about the application of mathematics in all fields of society. This interest led him to founding the

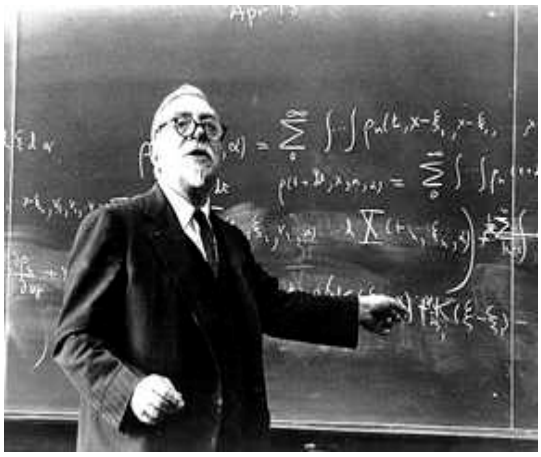
science of cybernetics. This field, which is perhaps best defined by the subtitle of [2]: "Control and Communication in the Animal and the Machine" included, among other things, a theory for information content in a signal and the transmission of this information through a channel. Wiener was, however, not a master of communicating his ideas to the technical community, and even though the relation to Shannon's formula is pointed out in [2], the notation is cumbersome, and the relevance to practical communication systems is far from obvious.

Reference to Wiener's work was done explicitly by Shannon in [1]. He also acknowledged the work by Tuller[3]. William G. Tuller was an employee at MIT's Research Laboratory for Electronics in the second half of the 1940s. In 1948 he defended a thesis at MIT on "Theoretical Limitations on the Rate of Transmission of Information[4]". In his thesis Tuller starts by referring to Nyquist's and Hartley's works (see below). Leaning on the use of sampling and quantization of a band-limited signal, and arguing that intersymbol interference introduced by a band-limited channel can in principle be eliminated, he states quite correctly that under noise-free conditions an unlimited amount of information can be transmitted over such a channel. Taking noise into account, he delivers an argument partly based on intuitive reasoning, partly on formal mathematics, arriving at his main result that the information H transmitted over a transmission link of bandwidth B during a time interval T with carrier-to-noise-ratio C/N is limited by

$$H \leq 2BT \log(1 + C/N). \qquad (2)$$

This expression has a striking resemblance to Shannon's formula, and would by most readers be considered equivalent. It is interesting to note that for the derivation of (2) Tuller assumes the use of PCM encoding.

A work not referenced by Shannon is the paper by Clavier [16][5]. In a similar fashion to Tuller, starting out with Hartley's work, and assuming the use of PCM coding, Clavier finds a formula essentially equivalent to (1) and (2). A fourth independent discovery is the one by Laplume published in 1948 [17].



Norbert Wiener (1894-1964) had been Shannon's teacher at MIT in the early 1930s. By his seminal work *Extrapolation, Interpolation and Smoothing of Stationary Time Series* made during World War II he lay the foundation for modern statistical signal processing. Although Shannon was influenced by Wiener's ideas, they had little or no contact during the years when they made their contributions to communication theory. There styles were very different. Shannon was down-to-earth in his papers, giving illustrative examples that made his concepts possible to grasp for engineers, and giving his mathematical expression a simple, crisp flavour. Wiener would rather like to use the space in-between crowded formulas for philosophical considerations and esoteric topics like Maxwell's demon.

[3] Shannon's work was in no way based on Wiener or Tuller; their then unpublished contributions had been pointed out to Shannon after the completion of [1].
[4] Later published as RLE Technical report 114 and as a journal paper [5] (both in 1949).
[5] It is, perhaps, strange that neither Shannon, nor Clavier have mutual references in their works, since both [3] and [16] were orally presented at the same meeting in New York December 12 1947, and printed more than a year afterwards.

Early attempts at a general communication theory

Shannon and the other researchers mentioned above were not the first investigators trying to find a general communication theory. Both Shannon, Tuller and Clavier make references to the work done in the 1920s by Nyquist and Hartley.

By 1920 one can safely say that telegraphy as a practical technological discipline had reached a mature level. Basic problems related to sending and receiving apparatus, transmission lines and cables were well understood, and even wireless transmission had been routine for several years. At this stage of development, when only small increases in efficiency are gained by technological improvements, it is natural to ask whether one is close to fundamental limits, and to try to understand these limits. Harry Nyquist, in his paper "Certain Factors Affecting Telegraph Speed" [7], seems to be the first one to touch upon, if not fully identify, some of the issues that were clarified by Shannon twenty years later.

First, it is obvious to Nyquist that the "Speed of transmission of intelligence" (which he terms W) is limited by the bandwidth of the channel[6]. Without much formal mathematical argument, Nyquist derives the following approximate formula for W:

$$W = K \log m \tag{3}$$

where m is the "number of current values", which in modern terms would be called "the size of the signalling alphabet" and K is a constant.

Whereas Nyquist's paper is mostly concerned with practical issues such as choice of pulse waveform and different variants of the Morse code, a paper presented three years later by Hartley is more fundamental in its approach to the problem. The title is simply "Transmission of Information", and in the first paragraph the author says that "What I hope to accomplish (...) is to set up a quantitative measure whereby the capacities of various systems to transmit information may be compared". Even though Nyquist had given parts of the answer in his 1924 paper, this is the first time the question that was to lead up to Shannon's information theory is explicitly stated.

Compared to Nyquist, Hartley went a couple of steps further. For one thing, he stated explicitly that the amount of information that may be transmitted over a system[7] is proportional to the bandwidth of that system. Moreover, he formulated what would later be known as *Hartley's law*, that information content is proportional to the product of time and bandwidth, and that one quantity can be traded for the other. It should also be mentioned that Hartley argued that the theory for telegraph signals (or digital signals in modern terms) could be generalized to continuous-time signals such as speech or television.

$$\text{Amount of information} = const \cdot BT \cdot \log m. \tag{4}$$

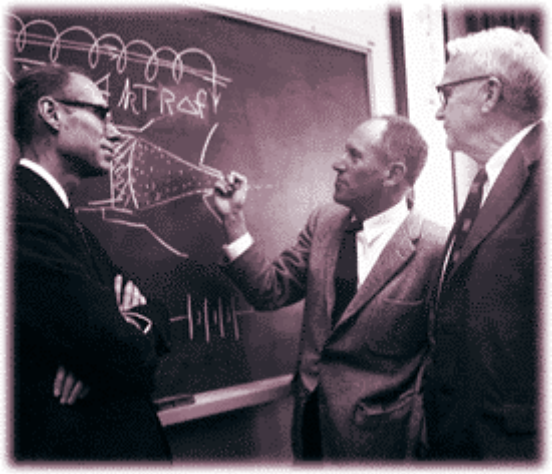
Relations between bandwidth and time similar to the one found by Nyquist was discovered simultaneously by Karl Küpfmüller in Germany [10]. A more mathematically stringent analysis of the relation was carried out in Gabor's "Theory of communication".

As was pointed out by Tuller [5], a fundamental deficiency of the Theories of Nyquist, Hartley, Küpfmuller and Gabor, is that their formulas do not include noise. The role of noise is that it sets a fundamental limit to the number of levels that can be reliably distinguished by a receiver. From expressions (3) and (4) we see that both Nyquist and Hartley were aware of the fact that the amount of information depends on the number of distinguishable signal levels (or symbols). However, they seem

[6] Proportionality to bandwidth is not explicitly stated by Nyquist in 1924. He have probably been aware of it, and includes it in his more comprehensive paper [7] four years later.

[7] Neither Nyquist, nor Hartley makes explicit distinction between source, channel and destination, as Shannon does twenty years later. This may seem like a trivial omission, but the distinction is essential for Shannon's general definition of channel capacity, which requires this separation to define quantities such as source entropy and mutual information.

content to include this number *m* in their formulas instead of deriving it from a more fundamental quantity, such as the signal-to-noise level. In a short discussion, Nyquist mentions "interference" as one of the limiting factors of the signal alphabet size. Hartley points out the inter-symbol interference due to channel distortion as the most important limiting factor. This is fundamentally wrong, as Tuller remarks, since inter-symbol interference can, in principle, be removed by an equalizer. This is precisely what Nyquist shows in his 1928 paper [7].



Harry Nyquist (right) (1889-1976) with John R. Pierce (left) and R. Kompfner. Nyquist was born in Nilsby in Wärmland, Sweden, and emigrated to U.S.A. in 1907[8]. He earned a M. S. degree in Elctrical Engineering in 1914 and a PhD in physics at Yale in 1917. The same year he was employed by AT&T where he remained until his retirement in 1954. Nyquist made contribution in very different fields, such as the modelling of thermal noise, the stability of feed-back systems, and the theory of digital communication.

Developments in the understanding of bandwidth

As we have seen, the work of Nyquist, Hartley and Küpfmüller in the 1920s represent an important step towards the fully developed channel capacity formulas expressed twenty years later. A key insight was the realization that information transmission rate was limited by system bandwidth. We have argued that this understanding came when the maturity of telegraph technology made it natural to ask if fundamental limitations were in reach. Another, related factor was that the concept of bandwidth, so essential in the cited works, was now clearly understood by the involved researcher. Today, when students of electrical engineering are exposed to Fourier analysis and frequency domain concepts from early on in their education, it seems strange that such a fundamental signal property as bandwidth was not fully grasped until around 1920. We will therefore take a look at how frequency domain thinking was gradually established as communications technology evolved.

If one should assign a birth date to practical (electrical) telecommunications technology, it would be natural to connect it to one of the first successful demonstrations of telegraphy by Veil and Morse in the 1840s. It took only a few years before all developed countries had established systems for telegraph transmission. These systems were expensive, both to set up and operate, and from the start it was important to make the transmission as cost-effective as possible. This concern is already reflected in the Morse code[9] alphabet which is designed according to the relative frequencies of characters in written English.

---

[8] More about Harry Nyquist's early years can be found at Lars-Göran Nylén's homepage URL: http://members.tripod.com/~lgn75/

[9] The *Vail Code* would be a more proper term, since it was Morse's assistant Alfred Vail who in 1837 visited a print shop at Morris-town to learn from the contents of the type cases which letter were more frequent in use. He then advised Morse to abandon his plan of using a *word code* involving the construction of a dictionary assigning a number to all English words, and use the more practical *character code* with unique dash-dot combinations for each letter [19].

It was soon evident that, for transmission on overhead wires, the transmission speed was limited by the telegraph operator and, possibly, the inertia of the receiving apparatus, not by any properties of the transmission medium. The only problem connected to what we today would call the channel was signal attenuation. This was, however, a relatively small problem, since signal retransmission was easily accomplished, either manually or with automatic electromagnetic repeater relays.

For the crossing of rivers or stretches of ocean, the telegraph signals were transmitted using cables. This transmission medium soon showed to be far more problematic than overhead lines. For long spans repeaters were not a practical solution, meaning that the attenuation problem became serious. It was also found that operators would have to restrain themselves and use a lower speed than normal to obtain a distinguishable message at the receiver end. Both these problems were of concern in the planning of the first transatlantic telegraph cable. Expert opinions were divided from the start, but the mathematical analysis of William Thomson (later Lord Kelvin) showed that even though the attenuation would be large, practical telegraphy would be possible by use of sensitive receiving equipment. In particular, Thomson's analysis explained how the dispersion of the cable sets a limit to the possible signalling speed.

In our connection, Thomson's work is interesting because it was the first attempt of mathematical analysis of a communication channel. We see that two of the four elements of Shannon's formula were indirectly taken into account: signal power reduced by the attenuation and bandwidth limiting the signalling speed. Bandwidth was not explicitly incorporated in Thomson's theory. This is quite natural, since the relevant relationships were expressible in physical cable constants such as resistance and capacitance. These were parameters that were easily understood and that could be measured or calculated. Bandwidth, on the other hand, was simply not a relevant notion, since engineers of the time had not yet learnt to express themselves in frequency domain terms.

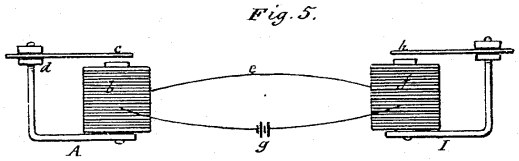
During the nineteenth century topics such as oscillation, wavelength and frequency were thoroughly studied in fields such as acoustics, optics and mechanics. For electrical and telegraph engineers, however, these concepts had little interest from the start.

Resonance was a well-known phenomenon in acoustics. It was therefore an important conceptual break-through when Maxwell showed mathematically how a circuit containing both capacitance and inductance would respond significantly different when connected to generators producing alternating current of different frequencies. The phenomenon of electrical resonance was then demonstrated in practical experiments by Hertz[10].

It is interesting to see how, even before electrical resonance was commonly understood, *acoustical* resonance was suggested as a means of enhancing the capacity of telegraph systems. As noted above, the telegraph operator represented the bottleneck in transmission by overhead wires. Thus, many ingenious schemes were suggested by which two or more operators could use the same line at a time. As early as 1853 the American inventor M. B. Farmer is reported to have suggested the first system for *time division multiplex (TDM)* telegraphy. The idea, which was independently set forth several times, was perfected and made practical by the Frenchman J. M. E. Baudot[11] around 1878. In parallel with the TDM experiments, several inventors were working with *frequency division multiplex (FDM)* schemes. These were based on vibrating reeds, kept in oscillation by electromagnets. By assigning a specific frequency to each telegraph operator, and using tuned receivers, independent connections could be established over a single telegraph line. One of the most famous patents is the "harmonic telegraph" by A. G. Bell from the 1870s. It was during experiments with this idea that Bell more or less accidentally discovered a way of making a practical telephone.

[10] For more details of the history of electrical resonance, including an early contribution by Thomson, see the paper by Blanchard [20].

[11] Baudot is today remembered by the unit *baud* for measuring the number of symbols per second transmitted through a communication channel.

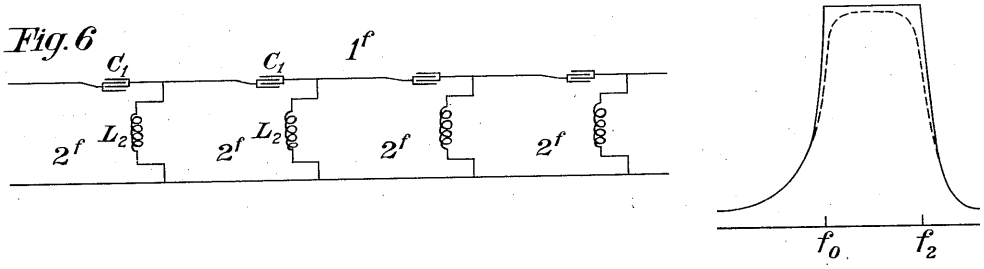


Detail from Alexander Graham Bell's patent for the "Harmonic Telegraph". An alternating current is generated in the left coil with a frequency given by the vibrating reed *c*. The reed *h* to the right will resonate and give a sound only if it is tuned to the same frequency. By this mechanism several users could transmit telegraph signal over the same line by using equipment tuned to different frequencies.

By 1890 electrical resonance phenomena were generally understood by scientists and well-informed engineers. Consequently, practical patents on how to use electrical resonance in FDM telegraphy began to be filed. These attempts, which also included some ideas of FDM telephony (not practical at the time) were, however, overshadowed by a new invention: *the wireless*.

After the first few experiments with wireless telegraphy at the start of the twentieth century, it became clear that sharp resonance circuits, tuned to specific frequencies or wavelengths were necessary to avoid disturbance among different users. This requirement made it necessary for radio engineers to have a good understanding of frequency and the behaviour of electrical circuits when connected to sources of varying frequency content. It is important to note that the concept of *bandwidth* was much more elusive than that of frequency. Today, the decomposition of an information-bearing signal into its Fourier components is a routine operation to any electrical or communications engineer. This was not the case 80-100 years ago. At that time, one would commonly assume that a radio transmitter was tuned to one – and only one – frequency. The bandwidth of a telegraph signal was small compared to both the carrier frequencies used and to the width of the resonance circuits employed. A general awareness of bandwidth did not develop until some experience had been gained with telephone transmission.

In parallel with the development of wireless telegraphy, and, gradually, telephony, work continued on FDM in telecommunications systems or "carrier current telephony and telegraphy"[12] which was the term used at the time. In these systems, first intended for enhancing the capacity of long-distance wire-bound connections, it became important to minimize the spacing between carrier frequencies and at the same time prevent cross-talk between the channels. To obtain this, traditional resonance circuits were no longer adequate for transmitter or for receiver tuning. What was needed was band-pass filters, sufficiently broad to accept the necessary bandwidth of the modulated speech signal, flat enough to avoid distortion, and with sufficient stop-band attenuation to avoid interference with neighbour channels. This kind of device was developed during the years prior to World War I, and was first patented by G. A. Campbell of Bell Systems in 1917.



Example of Campbell's bandpass filter designs with transfer function. (Original figures from U. S. Patent No 1 227 113.)

In hindsight it is curious to note that before Campbell's invention, band-limited channels in a strict and well-defined way did not exist! Earlier transmission channels were surely band-limited in the sense that they could only be used in practice for a limited frequency range. However, the frequency response

[12] A paper [21] with this title by Colpitts and Blackwell was published in three instalments in *Journal of the American Institute of Electrical Engineers* in 1921, giving a comprehensive overview both of the history of the subject and the state-of-the-art around 1920.

tended to roll-off gradually so as to make the definition of bandwidth, such as is found in Shannon's formula questionable.

So, around 1920 it was evident that an information-bearing signal needed a certain bandwidth, whether it was to be transmitted in original or modulated (by a carrier) form. There was, however, some discussion as to how large this bandwidth had to be. The discussion seems to have ceased after Carson's "Notes on the Theory of Modulation" in 1922. By this time it had been theoretically shown (by Carson) and practically demonstrated that by using so-called single sideband modulation (SSB) a modulated signal can be transmitted in a bandwidth insignificantly larger than the bandwidth of the original (unmodulated) signal. The aim of Carson's paper was to refute claims that further bandwidth reduction could be obtained by modulating the frequency of the carrier wave instead of the amplitude[13]. An often quoted remark from the introduction is that, according to Carson, "all such schemes [directed towards narrowing the bandwidth] are believed to involve a fundamental fallacy". Among others, Gabor [10] takes this statement as a first step towards the understanding that bandwidth limitation sets a fundamental limit to the possible information transfer rate of a system.

The significance of noise

We have seen that the work towards a general theory of communication had two major breakthroughs where several investigator made similar but independent discoveries. The first one came in the 1920s by the discovery of the relation between bandwidth, time and information rate. The second one came 20 years later. An important difference between the theories published during these two stages is that in the 1920s the concept of noise was completely lacking.

Why did it take twenty years to fill the gap between Hartley's law and Shannon's formula? The only necessary step was to substitute $1+C/N$ for m in (4). Why, all of a sudden, did three or more people independently "see the light" almost at the same time? Why did neither Nyquist, nor Hartley or Küpfmüller realize that noise, or more precisely the signal-to-noise ratio play as significant a role for the information transfer capacity of a system as does the bandwidth?

One answer might be that they lacked the necessary mathematical tools for an adequate description of noise. At this time Wiener had just completed a series of papers on Brownian motion (1920-24) which would become a major contribution to what was later known as stochastic processes, the standard models for description of noise and other unpredictable signals, and one of Shannon's favourite tools. These ideas, based on probabilistic concepts, were, however, far from mature to be used by even the most sophisticated electrical engineers of the time[14]. Against this explanation, it may be argued that when Shannon's formula was discovered, only two[15] of the independent researchers used a formal probabilistically based argument. The others based their reasoning on more common-sense reasoning, not resorting to other mathematical techniques than what were tools of the trade in the 1920s.

Another explanation could be that the problem of noise was rather new at the time. Noise is never a problem as long as it is sufficiently small compared to the signal amplitude. Therefore it usually arises in situations where a signal has been attenuated during transmission over a channel. As we have seen, such attenuation had been a problem from the early days of telegraphy. From the beginning, the problem of attenuation was not that noise then became troublesome, but rather that the signal disappeared altogether. (More precisely, it became too weak to activate the receiving apparatus in the case of telegraphy, or to weak to be heard by the human ear in case of telephony.) This situation changed radically around 1910, when the first practical *amplifiers* using vacuum tubes were devised. By use of these, long-distance telephony could for the first time be achieved, and, ten years later the first commercial radio broadcastings could begin. But, alas, an electronic amplifier is not able to distinguish between signal and noise. So, as a by-product, interference, thermal noise[16], always present

[13] This was an intuitively appealing idea at the time, but sounds almost absurd today, when the bandwidth expansion of FM is a well-known fact.
[14] One of the first papers with a rudimentary mathematical treatment of noise was published by Carson in 1925 [12]. It should also be mentioned that Harry Nyquist derived a mathematical model of thermal noise in 1928 [13]. This model was, however, derived without use of probabilistic methods.
[15] Shannon and Wiener, of course.
[16] Not to mention the "shot noise" generated by travelling electrons in the tubes themselves.

in both the transmission lines and the components of the amplifiers, would be amplified – and made audible – together with the signal. Early amplifiers were not able to amplify the signal very much, due to stability problems. When the feed-back principle, patented by H. S. Black in 1927, came in use, gains in the order of hundreds or thousands became possible by cascading several amplifier stages. This made noise a limiting factor to transmission systems, important to control, and by the 1930s signal-to-noise ratio had become a common term among communications engineers.

Although the researchers of the 1920s were aware of the practical problem represented by noise and interference, it seems that they did not regard it as a *fundamental* property of the transmission system, but rather as one of the many imperfections of practical systems that should be disregarded when searching for principal limits of what could be accomplished.

Thus, the fact that noise had just begun to play an active role in communications systems, might partly explain why it was not given sufficient attention as a limitation to transmission capacity. However, when one looks at the reasoning used by both Tuller and Clavier (and to some degree Shannon), one will find that their arguments are inspired by two practical ideas, both invented during the 1930s, namely *frequency modulation (FM)* and *pulse code modulation (PCM).*

Two important inventions

When reading textbooks and taking university courses in engineering, one may get the idea that new products are based on the results of engineering science, which again rely on a thorough understanding of more basic sciences such as physics and chemistry, which finally lean on mathematics as the most basic of all exact knowledge. One can also get the impression that the development in these fields follow the same pattern: technology must wait for physics to explain new phenomena by mathematics made ready for the purpose in advance. Reality is quite different. Time and time again inventors with only coarse knowledge of the physical phenomena they exploit, have come up with ingenious problem solutions. Similarly, engineers and physicists have discovered several mathematical results, which they have not been able to prove satisfactorily, leaving it to the established mathematicians to "tie up the ends" and provide the necessary comprehensive theory afterwards.

With this in mind, it is interesting to note that the understanding of noise in a theory of channel capacity had to wait for two practical invention. These inventions illustrated how signal-to-noise ratio (SNR) and bandwidth of a transmission system could actually be traded one against the other.

We have already mentioned Caron's 1922 paper, where he showed that frequency modulation (FM) would result in a signal bandwidth considerably larger than what would result by SSB, or even traditional AM. This is undeniably true, and it was therefore natural that most researchers also accepted Carson's rejection in the same article that FM would be more robust with respect to noise. Among the few who continued working seriously with FM was Edmund Armstrong. After several years of experimentation, and after introducing an *amplitude limiter* in the receiver, he was able to demonstrate that it was possible to significantly increase the SNR of a radio communication system by using FM at the cost of expanded bandwidth[17] [13].

What Armstrong's results showed was that a trade-off between bandwidth and SNR could in principle be possible. The next step, to realize that the information transfer capacity of a system depended both on bandwidth and SNR took some time, and needed another invention.

PCM – Pulse Code Modulation consists of the sampling and quantizing of a continuous waveform. The use of sampling had been suggested as early as 1903 by W. M. Miner.[18] Miner's motivation was to enhance the capacity of transmission lines by time division multiplex, as had already been done for telegraphhy (see above). Miner's concept contained no form of quantizing, and should not be considered a PCM system. This important addition was first introduced by A. H. Reeves in 1937.[19]

---

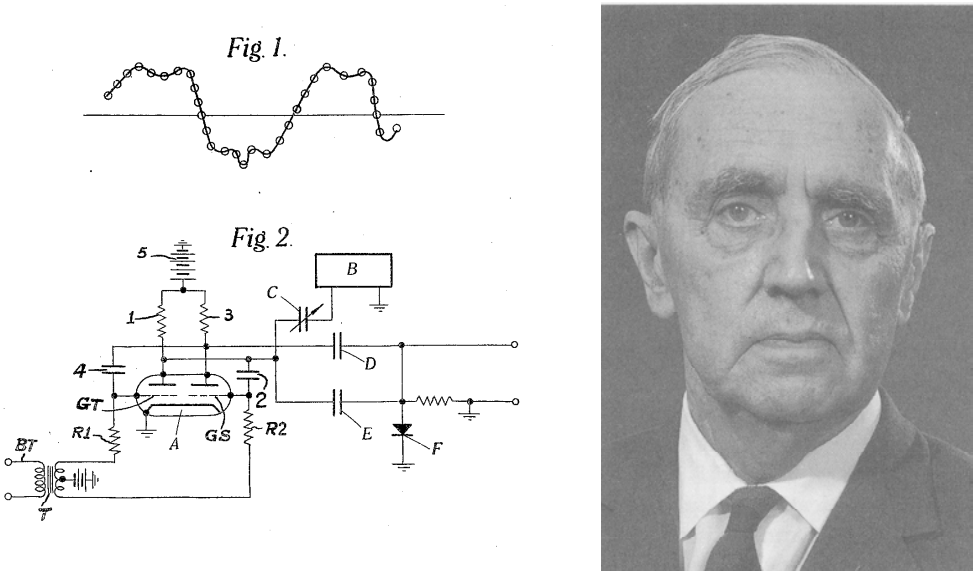
[17] Carson immediately accepted this as a fact, and very soon afterwards presented a paper together with C. Fry [15] showing mathematically how this mechanism worked, and that this was due to the amplitude limiter not included in the early proposals refuted by him in 1922.
[18] U. S. Patent 745,734.
[19] French Patent 852,183.

Reeves realized that his system would need more bandwidth than traditional modulation methods. His rationale was the same as Armstrong's: the combat of noise. Reeves's radical insight was that by inserting repeaters at suitable intervals along the transmission line, no additional noise would be added during transmission together with the quantizing noise introduced by the encoding (modulation) process at the transmitter. The quantizing noise could be made arbitrarily small by using a sufficiently high number of quantization levels.

Implicit in Reeves's patent lies two important principles:

1. An analog signal, such as speech, can be represented with arbitrary accuracy by use of sufficiently frequent sampling, and by quantizing each sample to one of a sufficiently large number of predefined levels.
2. Each quantized sample can be transmitted on a channel with arbitrarily small probability of error, provided the SNR is sufficiently large.

A result from this, not explicitly stated by Reeves, is that on a noise-free channel, an infinite amount of information can be transmitted in an arbitrarily small bandwidth. This is in sharp contrast to the results of the 1920s, and should be considered as a major reason why "Shannon's formula" was discovered by so many just at a time when PCM was starting to become well-known.



Alec Harley Reeves (1902-1971) with two figures from his PCM patent. Although aware that his invention was far ahead of what was possible with the technology of his time, the patent included several circuit solutions to some of the involved functionalities.

Concluding remarks

In this paper we have not written much about Shannon's general information theory. On the other hand, we have tried to show how the ideas leading up to "Shannon's formula" gradually emerged, as practical technological inventions made such ideas relevant. We have also seen that after the end of World War II, the subject was sufficiently mature, so that several independent researchers could complete what had been only partially explained in the 1920s.

On this background, one might be led to conclude that Shannon's work was only one among the others, and, by a stroke of luck, he was the first one to publish his results. To avoid such a misunderstanding, we will briefly indicate how Shannon's work is clearly superior to the others. First, we should make a distinction between the works of Shannon and Wiener ([1]-[3]) and the others ([5] [16][17]). Both Shannon and Wiener delivered a general information measure based on the probabilistic behaviour of information sources, which they both designate by *entropy* due to the likeness with similar expressions in statistical mechanics. Shannon, furthermore, uses this concept in his general definition of channel capacity:

$$C = \max[\, H(x) - H_y(x)].$$

This expression can be interpreted as the maximum of the difference of the uncertainty about the message before and after reception. The result is given in bit/second and gives an upper bound of how much information can be transmitted *without error* on a channel. The most astonishing with Shannon's result, which was not even hinted at by Wiener, was perhaps not so much the quantitative expression as the fact that completely error-free information exchange was possible at *any* channel, as long as the rate was below a certain value (the channel capacity).

The entropy concept is absent in all the presentations of the other group, which deal explicitly with a channel with additive noise. All reasoning is based on this special case of a transmission channel. The genius of Shannon was to see that the role of noise (or any other disturbances, being additive or affecting the signal in any other way) was to introduce an element of *uncertainty* in the transmission of symbols from source to destination. This uncertainty is adequately modelled by a probability distribution. This understanding was shared by Wiener, but his attention was turned in other directions than Shannon's. According to some, Wiener "under the misapprehension that he already knew what Shannon had done, never actually found out" [4].

References

[1] C. E. Shannon, "A Mathematical Theory of Communication", *Bell Syst. Techn. J.*, Vol. 27, pp. 379-423, 623-656, July, October, 1948.

[2] N. Wiener: "Cybernetics: or Control and Communication in the Animal and the Machine", MIT press 1948.

[3] C. E. Shannon, "Communication in the Presence of Noise", Proc. IRE, Vo. 37 1949, pp. 10-21.

[4] J. R. Pierce, "The Early Days of Information Theory", IEEE Trans. on Information Theory, Vol. IT-19, No. 1, January 1973.

[5] W. G. Tuller, "Theoretical Limitations on the Rate of Information", Proc. IRE, v. 37, No 5, May 1949, pp. 468-78.

[6] J. R. Carson: "Notes on the Theory of Modulation", Proc. IRE, 1922, 10, p. 57.

[7] H. Nyquist, "Certain factors affecting telegraph speed", *Bell Syst. Tech. J.*, vol. 3, pp. 324-352, April 1924.

[8] H. Nyquist, "Certain topics in telegraph transmission theory", *AIEE Trans.*, vol. 47, pp. 617-644, April. 1928.

[9] R. V. L. Hartley, "Transmission of information", *Bell Syst. Techn. J.* vol. 7, pp. 535-563, July 1928.

[10] K. Küpfmuller: "Über Einschwingvorgange in Wellen Filtern", Elektrische Nachrichten-Technik, vol. 1, pp. 141-152, November 1924.

[11] D. Gabor, "Theory of communication", *J. IEE*¸ vol. 93, pt. 3, pp. 429-457, September 1946.

[12] J. R. Carson, "Selective Circuits and Static Interference", *Bell Syst. Techn. J.* vol 4, p. 265, April, 1925.

[13] H. Nyquist: "Thermal Agitation of Electric Charge in Conductors", *Phys. Rev.*, 32, July 1928.

[14] E. H. Armstrong: "A Method of Reducing Disturbances in Radio Signaling by a System of Frequency-Modulation," Proc. IRE, 24, pp. 689-740, May, 1936.

[15] J. R. Carson and T. C. Fry, "Variable Frequency Circuit Theory with Application to the Theory of Frequency-Modulation", *Bell Syst. Thech. J.* Vol 16, pp. 513-540, 1937.

[16] A. G. Clavier, "Evaluation of transmission efficiency according to Hartley's expression of information content," *Elec. Commun.: ITT Tech. J.* vol., 25, pp. 414-420, June 1948.

[17] J. Laplume, "Sur le nombre de signaux discernables en présence du bruit erratique dans un système de transmission à bande passante limitée," Comp. Rend. Adac. Sci. Paris, 226, pp 1348-1349, 1948.

[18] J. Carson: "The statistical energy-frequency system spectrum of random distrubances", Bell Syst. Tech. J. 10; 1931, 3:374 ff.

[19] G. P. Oslin: "The Story of Telecommunications", Macon, Georiga, 1992.

[20] J. Blanchard, "The History of Electrical Resonance", *Bell Syst. Techn. J.* vol. 23, pp. 415-433.

[21] Colpitts and Blackwell: "Carrier Wave Telephony and Telegraphy", *J. AIEE*, April 1921.

[22] J. Bray: "The Communications Miracle", New York, 1995.

[23] F. W. Hagemeyer: "Die Entstehung von Informationskonzepten in der Nachrichtentechnik: eine Fallstudie zur Theoriebildung in der Technik in Industrie- und Kriegsforschung," Ph.D. dissertation, Freie Universität Berlin, 1979.