LINKS FOR THE FUTURE
Science, Systems & Services for Communications
P. Dewilde and C.A. May (editors)
IEEE/Elsevier Science Publishers B.V. (North-Holland), 1984

(Proceedings, Vol. 1, IEEE Int. Conf. on Comm., May 14-17, 1984, Amsterdam)

INFORMATION THEORY, THE COPERNICIAN SYSTEM OF COMMUNICATIONS

James L. MASSEY
Swiss Federal Institute of Technology
Institute of Telecommunications
Zurich, Switzerland

It is argued that Shannon's theory of information is the scientific basis of communications in the same sense that Copernicus' heliocentric theory is the scientific basis of astronomy. Some parallels between Copernicus and Shannon are drawn. It is further argued that the fundamental thesis of information theory is that all communications is essentially digital and that the main theorem of information theory is the separation theorem for source and channel coding. Some observations about the source coding theorem and the channel coding theorem are offered.

1. INTRODUCTION

By "information theory", we shall mean the mathematical discipline that seeks to establish fundamental limits on communication processes in the manner initiated by C.E. Shannon in his 1948 paper, "A Mathematical Theory of Communication" [1]. Shannon not only founded this field, but also developed it himself to a truly remarkable degree. It is thus fully appropriate that the IEEE Transactions has coined the name "Shannon theory" to describe "information theory" in the strict sense that we have adopted here.

Shannon's paper was received with immediate and enthusiastic acclaim. Perhaps it was this response that emboldened Shannon, when his paper was reprinted a year later in book form "with no changes except the correction of minor errata and the inclusion of some additional references" [2, Preface], to amend its title to "The Mathematical Theory of Communication". This emboldens us to be equally outspoken and to assert that information theory is <u>the</u> scientific basis of communications.

We suspect in fact that most communications engineers would agree that information theory is the scientific basis of communications. But we fear that for many this would be assented to only as a pious platitude akin to "Greek and Latin are the basis of a liberal education", not something without a grain of truth but something to be observed more in the breach than in the practice. To forestall such an emasculation of our claim that information theory is the scientific basis of communications, we further assert that we mean this to have as much force as we would the claim that the heliocentric theory of Copernicus is the scientific basis of astronomy. That accounts for the title of this paper and opens the way to a few historical observations.

2. COPERNICUS AND SHANNON -- SOME PARALLELS

Sometime around the year 1530, the Polish priest Nickolaus Copernicus (1473-1543) formulated his astronomical theory, which held that the sun lies at the center of the "solar system", that the earth and the other planets move in orbits about the sun, and that the earth rotates on its own axis once each day. Copernicus' ideas quickly spread through intellectual circles although publication of his great work, <u>De Revolutionibus Orbium Caelestium</u>, was delayed on political grounds until just before his death in 1543. It is interesting to note that Copernicus' theory came well after the widely-celebrated return to Spain in 1522 of one of Magellan's ships with its 18 survivors of the first circumnavigation of the globe. Fortunately among these survivors was the expedition's chronicler, whose meticulous diaries established without doubt that Magellan's men had lost one day from the calendar during their long westward voyage [3]. This fact, totally unaccountable for within the reigning Ptolemaic theory of astronomy was simply explained by Copernicus' theory. The latter theory also gave an elegantly simple explanation of planetary motion. The Ptolemaic theory, which held that the earth is the fixed point of the universe and lies beneath the firmament in which the heavenly bodies move, ascribed fantastically-shaped orbits to the moon and planets.

Yet it should be remembered that Ptolemaic theory served mankind well enough for fourteen centuries, permitting the prediction of eclipses and other celestial phenomena. It also meshed much better with ordinary human perception than did Copernician theory. Is it not easier still today to imagine that we are standing firmly on solid ground rather than clinging precariously to the surface of a rapidly rotating sphere? The scientific superiority of the Copernician theory was undeniable. But this new theory clashed so violently with "common sense" and religious beliefs that the old theory was not abandoned until after an unreasonably long and bitter struggle. In 1633, Galileo Galilei

(1564-1642) would be forced publicly to recant his belief in the Copernician theory, more than a century after its formulation. That Galileo accepted the Copernician theory was not unusual for scientists of his day; that he advocated it openly, however, was both unusual and politically dangerous.

The history of the Copernician theory can be characterized as follows:

(i) This theory was a radical departure from a long-established and widely practiced theory.

(ii) Its clear scientific superiority over the old theory gave it immediate intellectual appeal.

(iii) Its conclusions conflicted sharply with human intuition about the subject.

(iv) Its practical application was long and bitterly resisted by advocates of the old theory.

The history of information theory has much the same characteristics.

The "Ptolemaic" theory of communications looked upon a communication system as not essentially different from a power transmission system. The signals (which were usually considered to be sinusoids) inserted at one end were supposed to come out at the other with as little loss as possible. The losses were due to noise, which was considered to be the superposition of many sinusoids. The main concepts of the theory were bandwidth and signal-to-noise ratio. This theory served the needs of communications engineers well-enough through the decades when most of the analog modulation techniques in use today were developed; not, however, without some anomalies such as Carson's perfectly correct proof [4] of the claim that frequency modulation (FM) because of its greater bandwidth yields a worse signal-to-noise ratio than amplitude modulation (AM), which the crudest experiment would have shown to be false. The theory even survived the introduction of digital modulation, although it appears now that the difficulty of reconciling the characteristics of pulse code modulation (PCM) with the old theory was one of the factors that led Shannon to formulate his radically new theory. And, unfortunately, the old theory is still widely used and believed today. Defenders of the old faith denounce information theory as not having changed the way communications is done and therefore an irrelevant theory. Information theory has had its Copernicus, but not yet its Galileo.

3. THE THESIS OF INFORMATION THEORY

The fundamental thesis of information theory is that all technical communication is essentially digital; more precisely, that all technical communications is <u>equivalent</u> to the generation, transmission and reception of random binary digits, or "bits" to use the term first employed by Shannon at the suggestion of J.W. Tukey.

We owe to V. Aschoff [5] the observation that Sir Francis Bacon (1561-1626) was probably the first person to recognize clearly the complete generality of binary communications; in 1623, Bacon wrote that "a man may expresse and signifie the intentions of his minde, at any distance ... by ... objects ... capable of a twofold difference onely". But to recognize that binary digits suffice for all communication is not the same as to recognize that any communications transaction is characterized completely (for technical purposes) by some number of bits. This recognition is entirely due to Shannon.

Shannon's paper [1], which contains few references to prior work, gives two references at the very beginning. The first is to the 1924 paper by Nyquist [6] which showed that time is essentially discrete. If a time-continous signal has finite bandwidth, then it is completely specified by samples taken from it at discrete time instants.The second reference is to the 1928 paper by Hartley [7] which argued that the information in a message is given by the logarithm of the number of possible messages. Shannon realized that Hartley's measure of information was correct only when the messages were equally likely and must, for the general case, be amended to the entropy of the message set. Hartley had already recognized the arbitrariness of the base chosen for the logarithms in his information measure. The same arbitrariness is inherited by Shannon's measure. But when the base 2 is used, the entropy becomes the number of bits that have the same information content as the actual message set, and any technical communications transactions can be characterized by some equivalent number of bits. The crucial point is that a finite amount of information implies an essentially discrete message variable -- just as a finite bandwidth implies an essentially discrete time variable. The world of technical communications is essentially discrete or "digital". As Shannon said in [1] "the discrete case forms a foundation for the continuous and mixed cases". This is a complete reversal of the roles of digital and analog communications in the earlier theory of communications.

4. THE SEPARATION THEOREM OF INFORMATION THEORY

One of the most surprising and deepest results in Shannon's theory was his demonstration that the technical problem of transmitting an information source through some channel can be separated, with no loss of optimality, into the independent problems of representing that source by a sequence of binary digits and of transmitting a random binary sequence through that channel. This is true even if both the source and channel are analog in nature! Fano [8,p.3] is one of the few who has sufficiently stressed the importance of this result, which we shall call the <u>separation theorem</u> of information theory.

The separation theorem has vast practical implications. It means that communication transmission systems can (and we would add should) be designed to transmit random binary digits (or "bits") and can be used later without loss of optimality, to transmit any kind of information source. When one views the frantic efforts by which communication engineers now seek to transmit "data" over "telephone channels" that were tailored to transmit human speech, one sees the economic price that is now being paid for designing communication systems without benefit of the separation theorem.

That the separation theorem of information theory is often overlooked stems most likely from the fact that it appears only implicitly in Shannon's theory as a consequence of the two more celebrated theorems that we consider next. That the separation theorem was a revolutionary departure from the old theory of communications is perhaps the reason that its implications have been slow to be reflected in engineering practice.

5. THE SOURCE AND CHANNEL CODING THEOREMS

The two great triumphs of Shannon were his source (or "noiseless") coding theorem and his channel (or "noisy") coding theorem. We consider the latter first as it was the more unexpected and has been the more controversial of the two.

The channel coding theorem states that any communication channel is completely characterized by a single number C, the channel capacity, in the manner that R random binary digits (or "bits") per second can be sent as reliably as desired through the channel if only $R < C$, but cannot be so sent if $R > C$. That the rate R of transmission need not be reduced to achieve increased reliability must have been as incredible to the communications engineer of 1948 as had been the heliocentric theory to the astronomer of 1543. Both results clashed with all the listener's previous experience and theory. For rate is inversely proportional to signal-to-noise ratio, and everyone before 1948 "knew" that to reduce errors one must increase the signal-to-noise ratio. Shannon proved, however, that what counted was not the signal-to-noise ratio (so long as it was large enough to make $R < C$) but how the information was coded. One should not transmit information one bit at a time, but rather one should code long sequences of bits into a channel input sequence so that each bit of information is spread thinly over many channel uses. This, too, was a radically new idea and gave birth to the entirely new field of coding theory.

The channel coding theorem was entirely incompatible with previous communication theory, and used concepts that had no antecedents therein. Little surprise then that it was not soon welcomed or believed by practicing communications engineers! It became the fashion (that still persists) to belittle the channel coding theorem as a mathematical curiosity, true enough perhaps "in theory" but totally impractical. This skepticism has been buttressed by the many failures of "coding" techniques to improve significantly communication systems that were designed to function without coding. That this is akin to trying to use the internal combustion engine to improve upon the horse-and-buggy by strapping the engine to the horse's legs is still generally not appreciated. Only in deep-space applications have communications systems been built from the ground up according to the lessons of the channel coding theorem -- with remarkable success! Elsewhere, the old theory still hangs on and is stubbornly clung to.

The source coding theorem of information theory asserts that every communication source is completely characterized by a single number H, the source rate, in the manner that the source can be represented as accurately as desired by R binary digits per second if only $R > H$, but cannot be so represented if $R < H$. For technical purposes, the source is equivalent to one that emits H random binary digits (or "bits") per second. This fact, coupled with the channel coding theorem, yields immediately the separation theorem discussed in the previous section.

Although it was again a result without precedent in previous theory, the source coding theorem was rather quickly welcomed and believed by practicioners. It had the virtue that it did not contradict previous theory or experience -- there had been no source coding before! This meant another new field must be developed. Unfortunately, the horse-and-buggy metaphor again applies. The usual approach has been to "digitize" a source in some fairly arbitrary way, then to use "data compression" techniques to get the bit rate R down near H. Not surprisingly, there have been many failures, which have overshadowed the occasional success.

6. CLOSING REMARKS

We have tried to make it clear that information theory is the proper scientific basis for technical communication, and that it is a revolutionary replacement for the previously existing theory. More often than not, communications systems are still designed more on the basis of the old theory than on the new. This unfortunate state of affairs is likely to persist until sufficiently many "Galileos" have demonstrated the significant practical advantages that new theory offers.

REFERENCES

[1] C.E. Shannon, "A Mathematical Theory of Communication", Bell Sys. Tech. J., vol. 27, July and Oct. 1948, 379-423 and 623-656.

[2] C.E. Shannon and W.W. Weaver, The Mathematical Theory of Communication, Univ. of Illinois Press, Urbana 1949.

[3] R.M. Fano, Transmission of Information, M.I.T. Press and Wiley, Cambridge, MA,1961.
[4] J.R. Carson, "Notes on the Theory of Modulation", Proc. IRE, vol. 10, Feb. 1922, pp. 57-64.
[5] V. Aschoff, "The Early History of the Binary Code", IEEE Comm. Magazine, vol. 21, Jan. 1983, 4-10.
[6] H. Nyquist, "Certain Factors Affecting Telegraph Speed", Bell Sys. Tech. J., vol. 3, April 1924, 324-346.
[7] R.V.L. Hartley, "Transmission of Information", Bell Sys. Tech. J., vol. 7, July 1928, 535-563.
[8] S. Zweig, Magellan, Fischer, Frankfurt,1983.