BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 56, No. 3, 2008

Shannon theory. Myths and reality

J. PAWELEC*

Defence Communication Institute, Radom University of Technology, 29 Malczewskiego St., 26-600 Radom, Poland

Abstract. The significance of the famous Shannon's publication "A mathematical theory of communication" is discussed. The author states that this theory was a breakthrough for the times it was created. The present-day communications is so highly developed, that some old maxims should be up-dated, particularly the definition of the lower bound of signal reception. The author claims that this bound is no longer a constant value, ln(2), as the Shannon's theory states, but depends on many factors such, as the ratio of bandwidth-to-information transmission rate, the class of a receiver (adaptive, cognitive, MIMO[1]), the kind of reception system (on-line or off-line), and – of course – on the characteristics of noise, including entropy. Then, an absolute limit $(E_b/N_0)_{\text{abs}} = 0$ is suggested. An example of an advanced adaptive system approaching this bound is given.

Key words: communication theory, Shannon's contribution, reception bound, adaptive receiver.

1. Introduction

The 60-th anniversary of issue of the classical communications theory is just passing by. The great scientist Claude Shannon published in 1948 his famous paper [1], which caused a crucial turn in understanding the phenomena of digital communications. He disproved the prevailing views on communication errors and presented the original powerful entropy theory. His contribution survived decades and is still a milestone for the current information/communication theory and crypto.

The greatest mathematician of that period, A.N. Kolmogorov, wrote down [2], "In the ages of increasing specialization in science C. Shannon emerges as an outstanding talent combining the deep mathematical thinking with wide, but concrete reasoning of the current technology. He can be considered both as the great mathematician and the gifted engineer of the XXth century"[2].

But sixty years passed and the world has changed. The new systems have appeared, in between the DSSS, MIMO and adaptive (cognitive) ones. They have changed the old maxims. One, we challenge in this paper, is the lower reception bound (Eb,/No)$_{\text{min}}$. This bound is not ln(2), as Shannon's theory states, but zero. The architecture of the receiver and environments facilitating this bound are discussed through this paper.

2. Shannon's theory in a pill

In the Shannon's ages the scientists and engineers were convinced that there was no possibility to obtain the errorless communication as far as the channel is noisy and the rate of transmission is finite. This false view was disproved through the insertion of a new formula

$$C = \text{B}\log_2(1 + S/N). \tag{1}$$

It is derived in the original paper through 79 pages. The simplified version is given in Appendix A [3]. The formula (1) states that as far as the rate of communication does not exceed the bound C – which is called the capacity of a channel and is calculated from its bandwidth B and signal-to-noise power ratio S/N – the rate of errors can be reduced to zero. For example, $S/N = 15$ W/W, $B = 1$ MHz, hence, errorless communication can reach as much as $C = 4$ Mb/s.

At first glance it looks unbelievable, because the white Gaussian noise - being an essential assumption for (1) – manifests unlimited magnitudes and causes always some errors. Right, says Shannon, but the formula (1) reserves so high redundancy that all the errors can be corrected via adequate, non-constrained coding. And this was a weak point of his theory. Shannon didn't manage to define the postulated codes and left an open field for myths. They appear from time to time in the literature. The classical example is the case of turbo codes. The sentence: 'Turbo codes approach Shannon's bound' gives no information as we will show that such a bound is not a measure of excellence.

3. Myth I: lower bound of signal reception

The Shannon lower bound of reception is obtained through boundary transition of (1), [App. B]

$$C_0 = \lim_{B\to\infty} \text{B}\log(1 + S/N) \Rightarrow (E_b/N_0)_{\text{min}} = \ln(2). \tag{2}$$

It follows from (2) that a theoretical lower bound of signal reception $(E_b/N_0)_{\text{min}} = \ln(2)$, where E_b is energy of signal

*e-mail: j.pawelec@wil.waw.pl

[1] MIMO – multiple input – multiple output

[2] C. Shannon was born in the city Gaylord in 1916. He studied the mathematics and electro-technique at Michigan University. After graduation he joined the Bell Telephone Laboratories, where he received the doctoral degree. In 1956 he moved to Massachusetts Institute of Technology. C. Shannon was a member of the Academy of Science and Art of USA. He died in 2001

entity (bit) and N_0 – noise power density. The assumptions are: channel noise – additive white Gaussian (AWGN), bandwidth $B \to \infty$, rate of transmission $R = 1/T = C_0$, where T – signal duration.

Up to now nobody approached this limit in common non-spread AWGN systems, but some researchers obtained even lower values for spread spectrum and non-AWGN systems [3–5]. So, during this study it is stated that the lower bound of signal reception depends on many factors such, as the ratio of bandwidth-to-transmission rate, the kind of reception system and, of course on the entropy of noise. The absolute limit is $(E_b/N_0)_{\text{abs}} = 0$ [W/W] and it refers to the spread spectrum system for $B/R \to \infty$ or to the conventional system for continuous wave interference [App. B and C, respectively].

4. Myth II: non-constrained coding and turbo codes

Non-constrained coding stands for an error correction code, which detects and corrects the maximal number of errors at the expense of minimal bandwidth loss. Through several decades nobody found such a code[3]. In 1993 the French inventors claimed that their turbo codes are just such candidates, because they permit the detection of signals very close to Shannon's bound [6]

We claim, this is misunderstanding. Turbo codes do not work on-line. If the signal and noise are analyzed (integrated) for a long time, the samples of random noise compensate each other, while the steady state samples of signal add one to another and the ratio SNR improves. Turbo codes are useful, but their higher sensitivity has a little common with the Shannon's bound. It is an artificial value refering to specific conditions and being only in some relation to the real bound.

5. Myth III: optimal reception in color noise

It is easy to show that white Gaussian noise presents maximal entropy and hence it is the most destructive for reception. Some authors judge from here that BER for AWGN is greater than for any other noise. There are, however, empirical data, which show that BER for AWGN can be a little lower than for color one, when using an optimal cross-correlation receiver [3–5]. To explain these discrepancies we use the Shannon's expanded formula [1]

$$\mathrm{B}\log_2\left(1+\frac{S}{N_e}\right) \le C \le \mathrm{B}\log_2\left(\frac{N+S}{N_e}\right) \tag{3}$$

where N – average power of noise (σ^2); N_e – its entropy power (geometric mean)

$$N_e = \exp\left[B^{-1}\int_0^B \ln F(\omega)d\omega\right] \tag{4}$$

where B – channel bandwidth; $F(\omega)$ – power spectrum of noise.

It follows from (3), that in case of $N_e = N$ (white noise) the capacity $C = C_N = \mathrm{B}\log_2(1+S/N)$, while in the opposite case, $N_e < N$ (color noise, Eq. 4) the ratio $(N+S)/N_e$ and $C = C_e$ obtain higher values than C_N. The phenomenon is known in the theory for a long time, but in practice it was unsolved till the year 2002, when the concrete operating system was developed [5]. The obtained reduction of SNR in this system for typical non-AWGN environment is 10 ÷ 20 dB in comparison with the cross-correlation receiver data [App. D].

6. Myth IV: magic wand of entropy

Shannon objected to the use of the term *'entropy'* as it defines completely another quantity in thermodynamics. The man, who inclined him towards a new term was von Neumann himself. He might say, "No one understands entropy very well, so in any discussion you will be in a position of advantage" [7]. And the proposal was accepted.

What does the entropy mean? It has many applications and definitions. In communications it stands for a measure of information contained in a random event. For example, if the probability of twin birth is $P(x) = 10^{-3}$, then the Hartley entropy $H(x) = 1/P(x) = 10^3$. The less the probability of event, the more the information contained in it and the greater $H(x)$. Because of problems with summation of entropies for sets of events Shannon introduced $\log[1/P(x)]$. Hence the mean entropy for a series of events is: $\sum P(x_i)\cdot\log[1/P(x_i)]$. There are also other entropy definitions, e.g. Kolmogorov-Sinaj and Rènyi [7].

The most widely used in communications is the mean conditional entropy (equivocation) [1]

$$H(X|Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(y_j)P(x_i|y_j)\log_2 P(x_i|y_j) \tag{5}$$

where $P(y_i)$ – *a priori* probability of appearing the state j at the channel output; $P(x_i|y_j)$ – conditional probability of appearing the state i at input, while j is observed at the output Example 1.

Let us consider the binary symmetric channel with BER = 0.01 i.e. $P(x_0|y_1) = P(x_1|y_0) = 0.01$. Let *a priori* probability of states 0 and 1 be the same, $P(y = y_0) = P(y = y_1) = 0.5$. We want to know the mean conditional entropy $H(X|Y)$, which in this case defines the loss of information. After substitution of the given data into (5): $P(x_1|y_0) = 0.01$, $P(x_0|y_1) = 0.01$ and $P(x_1|y_1) = 0.99$, $P(x_0|y_0) = 0.99$, we obtain $H(X|Y) = 0.081$. Hence, although the BER is 1%, the equivocation or loss of information is more than 8%. It is interesting, that for BER = 0.5, this loss is 100%. Shannon put the equivocation $H(X|Y) = 0$ in derivation of the formula (1). This is a reason, upon which we can expect the errorless communication, whenever (1) is satisfied [App. A].

Therefore, where are the sources of myths raised at the beginning of this paper? They reside in assumptions. For ex-

[3] One of the optimal systems is the convolution code plus Viterbi decoding, but it could not get the Shannon's bound.

ample, we assume the infinite bandwidth of noise, but at the same time we put the finite density of its power (N_0). Hence, the integral can reach infinity, which can be hardly imagined ($\sigma^2 < \infty$).

The entropy, although highly deserved for information/communication theory, does not remove all the discrepancies due to the imperfect correspondence between analog medium of transmission of signals and their digital nature. Hence, some disagreements have to occur.

7. Conclusions

Shannon's theory was a breakthrough for the times it was issued. Some its notions are still valid such, as entropy measures, capacity dependence on the signal-to-noise ratio and bandwidth, the rules for ideal crypto etc. There are also some shortcomings. The most doubtful, from the present-day view, is the famous lower bound of signal reception, $(E_b/N_0)_{\min} = \ln 2$. The author shows that a real reception bound is defined by many factors such, as the ratio of bandwidth-to-transmission rate, the intelligence of the receiver and, of course, on the noise characteristics, including entropy. If this entropy is small and the bandwidth is wide enough in reference to the rate of information transmission and the receiver presents an advanced design, the bound of reception can be made much lower than ln2 [4, 5]. In the opposite case, this bound is always much higher than ln2 [8, 9]. The knowledge of noise behavior in the channel is of prime importance. The paper presents an adaptive system, which observes this behavior and adjusts receiver parameters to noise changes. The gain obtained this way for typical non-AWGN environment is 10÷20 dB in comparison with the "optimal" cross-correlation receiver [4, 5].

It should be also noted that a lower bound in off-line systems may be made arbitrary low for arbitrary noise (e.g. in space systems). Similarly, the capacity of MIMO system, composed of infinite number of antennas, can be made infinite too, nevertheless the narrowband frequency channel is used [10] Hence, the only reasonable absolute lower bound of reception is $E_b/N_0 = 0$.

Appendix A

Derivation of Shannon's formula

The fundamental aim of Shannon's work [1] was to define the maximal errorless transmission rate of digital signals through the noisy channel. As far as the entropy is a measure of information contained in a sequence of signals, then the maximal value of entropy per signal duration T can be used to asses this searched transmission rate

$$\begin{gathered} C_{\max} = \max[H(Y) - H(Y|X)]/T, \\ H(Y) = H(X) + H(N) \end{gathered} \tag{A1}$$

where $H(Y)$ – entropy of signal at the channel output; $H(Y|X)$ – equivocation; $H(X)$ – signal entropy at the input; $H(N)$ – noise entropy.

In general, the following relationship for entropies can be used for the binary symmetric channel

$$\begin{gathered} H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \\ \Rightarrow H(Y|X) = H(Y) + H(X|Y) - H(X). \end{gathered} \tag{A2}$$

If the transmission has to be errorless, then $H(X|Y)$. Hence

$$\begin{gathered} H(Y|X) = H(Y) - H(X) \\ = [H(X) + H(N)] - H(X) = H(N) = 0. \end{gathered} \tag{A3}$$

Equation (A3) expresses the peculiarity that only factor, which differs output Y from its input equivalent X, is noise N. Hence $H(Y|X)$ is replaced by $H(N)$. Substitution of (A3) into (A1) gives

$$C_{\max} = \max[H(Y) - H(N)]/T. \tag{A4}$$

Entropy of an individual sample of noise is

$$\begin{gathered} H_i(n) = -\int_{-\infty}^{\infty} \left(\sigma\sqrt{2\pi}\right)^{-1} \exp(-n^2/2\sigma^2). \\ \cdot \ln\left(\sigma\sqrt{2\pi}\right)^{-1} \exp(-n^2/2\sigma^2)dn = \log_2\left(\sigma\sqrt{2\pi e}\right). \end{gathered} \tag{A5}$$

Taking into account that AWGN samples are independent each other and that as many as $2BT$ signals can be fitted in bandwidth B (Nyquist theorem), then the overall entropy is

$$H(N) = \sum_{i=1}^{2BT} H_i(n) = BT \log_2 \sigma^2 2\pi e. \tag{A6}$$

The output entropy of noise plus signal – per analogy to (A6) – assuming its Gaussianity is

$$H(Y) = BT \log_2[(2\pi e)(\sigma^2 + S)] \tag{A7}$$

where S is the useful signal power.

Substituting (A6) and (A7) into (A4) we finally obtain

$$C_{\max} = B \log_2(1 + S/N) \tag{A8}$$

where N represent the noise power (σ^2).

Appendix B

Shannon's bound and its discussion

The capacity C in formula (1) for $B \to \infty$ is

$$\begin{gathered} C_0 = \lim_{B\to\infty} \{B \log_2[1 + S/N]\} \\ = \lim_{B\to\infty} \left\{ B \log_2[1 + \left[\frac{E_b/T}{N_0 B}\right] \right\} \\ = \lim_{x\to 0} \left\{ \frac{\log_2[1 + (E_b/N_0 T)x}{x} \right\} \end{gathered} \tag{B1}$$

where $x = 1/B$, E_b – energy of signal bit, T – its time duration.

Using the L'Hospital rule for the expression in curly bracket we obtain [8]

$$\begin{gathered} C_0 = \lim_{x\to 0} \left\{ \frac{1}{1 + (E_b/N_0 T)x} \left[\frac{E_b}{N_0 T}\right] \log_2 e \right\} \\ = \frac{E_b}{N_0 T \ln 2}. \end{gathered} \tag{B2}$$

Assuming further that $E_b = ST$ and $C_0 = 1/T$, the searched capacity and lower bound are, respectively

$$C_0 = \frac{S}{N_0 \ln 2} \quad \text{and} \quad \left(\frac{E_b}{N_0}\right)_{\min} = \ln 2. \tag{B3}$$

Please note, that despite of infinite channel bandwidth, the capacity C_0 is finite and depends only on signal power S (assuming $N_0 = \text{const}$). This looks a little suspicious.

Let us consider the more realistic case of $B >> C$ (instead of $B \to \infty$).

Example 2.

Let $S = 1$ W, $N_0 = 10^{-9}$ W/Hz and $B = 10^9$ Hz.

Hence, using (1) and putting $N_0 B$ instead of N we obtain

$$\begin{aligned} C_{09} &= 10^9 \log_2(1 + 1/10^{-9} \cdot 10^9) \\ &= 10^9 \log_2(2) = 10^9\ b/s. \end{aligned} \tag{B4}$$

Now, let $B >> C$, e.g. $B = 10^{11}$ Hz. Hence

$$\begin{aligned} C_{11} &= 10^{11} \log_2(1 + 1/10^{-9} 10^{11}) \\ &= 10^{11} \cdot 0.0144 = 1.44 \cdot 10^9\ b/s. \end{aligned} \tag{B5}$$

We can see, that capacity C_{11} has increased a little in reference to C_{09}, but the SNR has decreased significantly, from 1 to 0.01⇔ −20 dB.

Let B be further increased to 10^{13} Hz

$$\begin{aligned} C_{13} &= 10^{13} \log_2(1 + 1/10^{-9} \cdot 10^{13}) \\ &= 10^{13} \cdot 0.0001442 \approx 1.44 \cdot 10^9\ b/s. \end{aligned} \tag{B6}$$

We see, that capacity is nearly the same ($C_{13} \approx C_{11} \approx 1.44$ Gb/s), while signal-to-noise ratio has been further decreased by 20 dB and its final value is SNR = −40 dB. So, we can deduce, that for $B >>> C$, SNR ⇒ 0 [W/W]. Because of constant $C = 1/T$, we can also write $E_b/N_0 \to 0$ for $B \to \infty$.

This is in contrary to (B3b). The source of discrepancy resides in the silent assumption of the transition $(B1) \to (B2)$. It is supposed that capacity C follows B (according Nyquist theorem 2BT = const). However, if C is constant and B increases to infinity ($B \to \infty$) the signal-to-noise ratio decreases to 0 and this is an absolute bound $(E_b/N_0)_{\text{abs}} = 0$. This phenomenon is commonly known in spread spectrum systems, where the bandwidth is exchanged for the SNR.

Appendix C

Signal reception in Dirac noise

Let the noise is expressed by the two-component function [3]

$$F(\omega) = \alpha(\omega) + \beta(\omega) \quad \text{for} \quad 0 < \omega < B. \tag{C1}$$

Let $\alpha(\omega) = \alpha = \text{const}$ for $\omega_0 < \omega < \omega_0 + \delta\omega$, $\beta(\omega) = \beta = \text{const}$ for $\omega = B \backslash \delta\omega$, $\alpha >> \beta$ and

$$\int_0^B F(\omega) d\omega = \sigma^2. \tag{C2}$$

After substitution of (C1) into (4) the formula for the entropy power takes the form

$$N_e = \exp\{B^{-1}[(\delta\omega) \ln \alpha + (\ln \beta)(B - \delta\omega)]\}. \tag{C3}$$

In the boundary case for $\delta\omega \to 0$, $\beta \to 0$ and $\alpha \to \infty$

$$\lim N_e = \exp\{B^{-1}[\ln \alpha^0 + B \ln \beta]\} = e^{\ln \beta} = \beta = 0. \tag{C4}$$

It follows from (C4) that Dirac noise possess an entropy power equal to 0. Hence – in accordance with (3) – both the signal-to-noise ratio and the channel capacity approach infinity. Such an interference is produced by the continuous sine wave (CW) of the constant amplitude, constant frequency and endless duration. It is easy to conclude that CW is not able to interfere reception even at its infinite growth. Hence, we can write down: $(S/A^2/2)_{\min} \Leftrightarrow (E_b/N_0)_{\text{abs}} = 0$.

At this moment the question arises, whether fighting against myths we do not create the new ones? The response is, no. We have strong support from experiment. If the sine interference is inserted into the channel, we can remove it via infinitely narrowband filter, which makes no harm to useful signal. One can ask, is it possible to build such a filter? It is difficult within analog technology, but in digital technology it is feasible, at least up to the small error depending on the number of the filter cells. Then, an asymptotic bound Eb/No exists and it is zero (not ln2).

Appendix D

Optimal reception under arbitrary noise

The term ‘arbitrary noise’ stands for the stationary noise (interference), at least of the second order, and of arbitrary correlation function (power spectrum). The term ‘optimal’ refers to the lowest signal-to-noise ratio. The appropriate scheme is shown in Fig. 1.

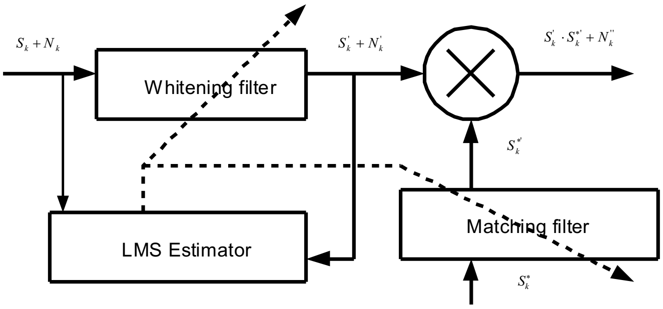

Fig. 1. The scheme of optimal adaptive receiver for arbitrary noise

It operates as follows. The consecutive samples of signal and noise, S_k and N_k enter the input whitening filter. Noise dominates over the signal, $\sigma_N^2 >> \sigma_S^2$. Parameters of the input filter are controlled by an estimator LMS (least mean square). This estimator observes output samples and adjusts filter weights so, as the difference between the estimate of n-th sample[4] output and its real value $(S_n + N_n)$ is minimal. This way the resulting output process ($\sim N'_k$) becomes white.

[4] This estimate is: $S'_n + N'_n = \sum\limits_{i=n-N}^{n-1} (S_i + N_i) \cdot h_{n-i}$ where h_i are weights of the filter and N is its order

The whitening filter changes also the shape of useful signal. To match its new shape to the replica, the last one is transferred via similar filter as the signal does. So, the replica undergoes the same filtration process as the signal undergoes (see dotted lines in Fig. 1). Hence, both conditions of optimality are satisfied: the signals entering the multiplier, S'_k and $S^{*\prime}_k$ are matched one to another, and the noise N'_k is white.

The receiver can be directly applied in spread spectrum systems. In other systems the more sophisticated estimator should be thought up, e.g. one basing on the difference of known useful spectrum and the estimated signal plus noise spectrum.

REFERENCES

[1] C. Shannon, “A mathematical theory of communication”, *The Bell System Techn. J.* 1–77 (1948).

[2] C. Shannon, *Works on Information Theory and Cybernetics*, IIŁ, Moscow, 1963, (in Russian).

[3] J. Pawelec, “Optimum nonwhite detection. a double-matched digital filter approach”, *Bull. Pol. Ac.: Tech.* 51 (1), 1–12 (2002).

[4] J. Pawelec and R. Piotrowski, “Optimum filtration of NB interference in DS-CDMA satellite systems”, *Int. J. Sat. Com.& Networking* 21, 119–125 (2003).

[5] J. Pawelec, “An adaptive non-AWGN spread spectrum multiple access receiver”, *IEEE Communications Magazine*, 126–127 (2002).

[6] C. Berrou, A. Glavieux, and P. Thitimajshima, ”Near Shannon limit error correcting coding and decoding: turbo codes”, *Int. Communications Conf. (ICC)*, 1064–1068 (1993).

[7] M. Leśniewicz, *Private Communications*, 2007.

[8] L. Couch, *Digital and Analog Communication Systems*, Macmillan, London,1987.

[9] B. Sklar, *Digital Communications*, Prentice Hall, New Jersey, 2002.

[10] J. Pawelec, “An information-communication system based on MIMO technology”, *J. Communications* 17, (2008), (invited paper – to be published).