60 Years of Shannon Information Theory

Dušan B. Drajić

Abstract - **In this paper an attempt is made to give a very short survey of the development of Shannon Information Theory. Also, some thoughts concerning the future of Information Theory are given.**

Keywords - **Information Theory development, future,**

I. Introduction

"Scientific theories deal with concepts, not with reality. All theoretical results are derived from certain axioms by deductive logic. In physical sciences the theories are so formulated as to correspond in some useful sense to the real world, whatever that may mean. However, this correspondence is approximate, and the physical justification of all theoretical conclusions is based on some form of inductive reasoning."

A. Papoulis: ***Probability, Random Variables and Stochastic Processes*** **(Preface)**

In Communications it is especially important to have the above citation in mind. The "real world" of communications engineers (information, messages, electrical signals etc.) is not always easy to see and models must be made for the "invisible" things.

In 19th century human race started to use more and more energy. At the end of the century the corresponding theory was developed – Statistical Mechanics with the well known Second Law of Thermodynamics (where the notion of ***entropy*** was introduced, the entropy being regarded as a quantitative measure of order against disorder).

In 20th century the "information era" started. Now, we are living in the ***Information Age***. In the middle of the century the corresponding mathematical theory of communication was brought by Claude Shannon (including also the quantity named entropy).

One more parallel: at the beginning of the 20th century Einstein gave the relationship between mass, light velocity and energy. Shannon gave the relationship between the attainable information rate, frequency band and the signal power (energy).

It should be also noted that Shannon's theory appeared practically at the beginning of the information era. So, some solutions had to wait for the corresponding technology to be used in practice.

In fact, the theory in engineering sciences is usually a little "behind" the practice, confirming the practical experience.

Dušan B. Drajić is with the Faculty of Electrical Engineering, University of Belgrade, Serbia

However, in Information Theory the theory was partially far ahead and waited to be confirmed by practice.

In this paper, an attempt to give a survey of the development of Information Theory is presented. At the end some thoughts concerning the future of Information Theory will be given.

II. Berth of Information Theory

Although the concept of mathematical model can even be found in the ancient Greece (Plato!), the paper will start only from the beginning of the 20th century. The electrical signals were used to transmit alphanumerical characters and voice (a little later pictures as well). The model used was ***deterministic*** by its nature. The Fourier analysis, invented (or discovered?) about two centuries ago to solve some other problems, was used. The signals were regarded as sine waves or as their sum (finite yielding the Fourier series, an infinite yielding the Fourier integral). In any case, deterministic approach based on Fourier analysis was used in designing classical analogue systems (needed bandwidth, power, etc.) and these systems worked quite well under the "normal" conditions. But, they operated badly in "severe" conditions, or they could not operate at all (FM is unusable for negative – in dB – signal-to-noise ratio). The severe conditions are encountered also during the war, where the security is needed, too. Therefore, a better ("more realistic") model had to be found. It was based on the ***probabilistic approach***. Only one suitable citation will be given:

"The complexity *of physical phenomena necessitates the consideration of probabilistic models. A probabilistic or stochastic description usually models the effects of causes whose origin and nature are either unknown or too complex to describe deterministically. Thus, a stochastically modelled physical phenomenon is not necessarily nondeterministic by nature; its stochastic description may merely represent the best known model for its behaviour."* [1]

Indeed, two such models appeared.

Norbert Wiener, borrowing the notion of "ensemble" from Statistical Physics [2] and generalizing harmonic analysis [3] gave the basic principles of the so-called Statistical Communication Theory. The main problem can be formulated as follows: the transmitted signals are corrupted by noise (the signals, as well as the noise are modelled as stochastic processes), the problem is how to extract signals from noise (how to improve signal-to-noise ratio). The goal was achieved by the optimum filtering and by correlation methods. For digital signal transmission, the matched filter can also be regarded as a result of this approach.

1-4244-1468-7/07/$25.00 ©2007 IEEE

Claude Shannon, on the other hand, started "from inside", i.e. from communication problems themselves and provided a brilliant and elegant solutions. He created original mathematical concepts. His fundamental paper "A Mathematical Theory of Communication" [4] (transcribed as "***The*** Mathematical Theory of Communication" by some scientists) is the basis of Information Theory.

Both models are based on the probabilistic approach. Both use the same mathematical apparatus. But, it is the only common thing.

In fact, the early beginnings of Information Theory are the works of Nyquist [5] and Hartley [6]. Nyquist found the minimum frequency band to transmit independent discrete signals at a given rate. Hartley proposed to use the logarithmic measure for information (he said that the information transmitted is proportional to the logarithm of the number of different signals we use – to the alphabet size).

The Shannon approach was totally different from the Wiener one. One should say it was at a higher level. He did not consider the signals, but the information. The information is represented (encoded) by signals, which are carriers of information. That means also that it is possible that transmitted signals do not carry any information at all (from the Information Theory point of view). Of course, these signals may be needed for the proper functioning of the communication system itself (synchronisation etc.).

Shannon defined the quantity of information emitted by information source and tried to find how to represent (encode) the information by the signals so that the information remains undistorted even if the transmitted signals are corrupted or distorted by noise. He investigated the limits of such a system having in mind the source information rate and the channel characteristics (parameters) – bandwidth and signal-to-noise ratio (SNR). A communication system from an Information Theory point of view is presented in Fig. 1 [7].

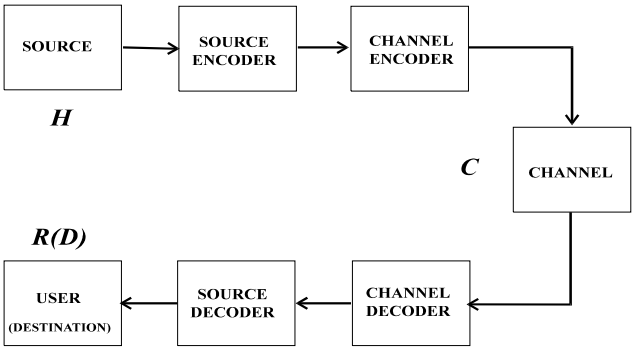

Fig. 1. Communication system as "seen" by Information Theory

The first thing when one wish to describe mathematically some process is to define a corresponding quantity as well as a unit to measure it. So Shannon had to define the quantity of information in a message. For a discrete source with finite number of messages he defined the quantity of information as a logarithm of the inverse of the message (symbol) probability. The average information rate per symbol (from the source) is obtained by averaging it over all symbols. If 2 is taken as a base for logarithm, the quantity of information is measured in *information bits* (I would prefer to use *shannons*) and information rate (suitably named entropy, the expression being the same as for the entropy of ideal gas) is measured in *information bits per symbol*. The ***entropy – H –*** can be thought as a measure of our uncertainty which message (symbol) will be chosen and emitted by the source. For a source with higher entropy this uncertainty is higher.

The next block is "source encoder". Its task is to represent (encode) the information (messages - symbols) by signals in an efficient way. Shannon showed that the number of signals needed depends on the source entropy (Source Coding Theorem).

The next one is "channel encoder" having the task to encode (represent) the information by the channel signals (symbols) is such a way that no information is lost if the signals are distorted and even if some finite error probability exists. Shannon showed that the error probability can be made as small as we wish if the information flow is smaller then the ***channel capacity – C –*** depending on bandwidth and SNR (Channel Coding Theorem).

After the corresponding decoders the last block is the "user". It was not taken into account in the beginning. About ten years later [8] Shannon brought the basis of the so-called ***Rate Distortion Theory*** where ***R(D)*** is the minimum amount of information (*shannons – information bits per symbol*) needed by the user allowing some distortion of information, quantitatively described by ***D***.

It should be noted also that Fig. 1. corresponds to "one way communication". In fact, very often the information is transmitted in both ways. Furthermore, Information Theory also considers the ***multiusers systems*** and can be connected with networking problems as we will see later.

III. Basic Results Of Information Theory

A. Source coding

The Source Coding Theorem simply states that (for binary signals - bits) the average number of bits needed to encode the source symbol can be made as small as entropy, but not smaller.

The source encoding is accomplished by giving the shorter code words to the symbols with higher probabilities (the same thing did Morse without knowing Information Theory). The algorithms for source encoding (Shannon-Fano and Huffman) were published just a few years after the basic Shannon's paper. Generally, the Huffman algorithm is the optimum one. Shannon also mentioned "arithmetic" encoding based on the cumulative probability.

One of the drawbacks of source encoding is that when one received symbol is in error, the synchronisation between encoder and decoder will be lost for some time resulting in a series of erroneously decoded symbols. The efforts have been made to find the suitable code words so as that resynchronisation is obtained as fast as possible.

One may ask oneself: Now we have disks with memory measured in gigabytes, also more gigabits per second (a few

terabits per second is the last result) can be transmitted through the fibre. Is there any need to compress? The answer is: *yes*! Every engineer knows that any system should be used efficiently. So, why not put twice (or even three times) more data on the same disk without any change in hardware? Why not transmit twice more data per second through the same channel?

It should be noted that the mentioned algorithms for source encoding are based on the complete knowledge of source statistics (i.e. symbol probabilities). But sometimes we do not know the source statistics. We have no time to store the whole incoming sequence, analyse it, and choose the corresponding code. Also, the sequence to be transmitted often is generated in "real time" and should be sent without waiting for its end.

Then, the adaptive methods are used. The statistics is performed on the incoming part of a sequence and a corresponding adaptive encoding is performed. The adaptability is based on the fact that we know more about incoming sequence, as we have a larger part of it.

Almost all contemporary codes for data compression are made in such a way. The best known procedure is Ziv-Lempel encoding (LZ codes) with many versions. If a processed sequence is relatively long, then the adaptive method will attain the limit prescribed by Source Coding Theorem (attainable with Huffman encoder, but with theoretically infinite delay – waiting the end of a sequence to start).

It should be noted also that in the case of LZ codes and their versions there is no need for encoder and decoder to communicate before the start (by sending a list of code words as in the Huffman coding), because they begin to work on the same sequence. So decoder can draw the same conclusions as the encoder at the beginning and then to apply them for further decoding.

It is also assumed that the statistics (known or not known) of the sequence does not change with time. For such a case there is a mathematical argument that the coding will be efficient. It is Asymptotic Equipartition Property the consequence of which is that almost all (i.e. with probability approaching to 1) sequences emitted by the source will be from the "typical set", i.e., they all will have the entropy near to the source entropy as the length of a sequence approaches to infinity. For example, the number of possible binary sequences of n bits is 2^n, but if the source entropy is H, then the number of the sequences in a typical set is 2^{nH}.

For those wishing to have a better insight into LZ encoding, just a hint. Kholmogorov defined the complexity of a sequence as a length of a computer program to generate it. LZ algorithms can be thought out as an attempt to write such a program during observing the incoming sequence.

Further, there is so-called "universal coding" i.e. coding without knowing the exact source statistics (or where such a statistics is not easy to model – for example images). It should be noted the paper "Universal Compression of Memoryless Sources over Unknown Alphabets" obtained IEEE Information Society Paper Award for the year 2006.

Therefore, the Source Coding Theorem is the basis of ***non-destructive (lossless) text compression***, i.e. the original text can be reconstructed in the whole.

One of the interesting examples is the compression of the English text. Shannon wrote the paper "Prediction and Entropy of Printed English" [9]. He considered a text to be a Markov chain and tried to predict the "true" entropy of English finding it to be near to 1 *shannon per letter*, instead between 3 and 4, the result obtained when considering the text as being without memory.

It is interesting to note that Markov took also the text (the Russian one!) as an example, when formulated his theory.

Here is the very place for a little discussion about text modelling. Almost all models of English (as well as of other languages), suppose that it can be modelled as a Markov chain with constant memory, considering space sign as a part of the alphabet. In fact, the sounds (phonemes, as linguists call them) are written using letters. The sounds are generated by a human being using vocal tract. So, the sequence of letters depends on the possibility of successive sounds pronunciation in a continuous speech. If there are pauses between words, as they should be, then the pauses can not be treated in the same way as other sounds. They just ease the pronunciation. Therefore, in a better model the text should be considered as a "pulsed" Markov chain where statistical dependence between letters is disrupted at the end of a word (at least at the end of a sentence). With a new word, i.e. after a space sign, the dependence starts anew from the first letter.

It is also possible to model a text (speech) using words as units (at a syntactic level!) instead of letters. It is interesting for automatic translation.

B. Channel Coding

The Channel Coding Theorem states that the information can be transmitted with the probability of error being as small as we wish (but not zero!) until the information flow is less than the channel capacity. It is probably the most important result of Information Theory. The statement was a little surprising for the communication engineers of the époque – they thought that the noise limits the reliability of the transmission, but the noise limits in fact the rate of the reliable transmission. It means that we can transmit reliable by using an "unreliable" channel. The theorem was proved on the basis of random coding argument. So, it did not show how to find a channel code. It only proved the existence of "good" codes and gave the limit that can be approached by long coding sequences (theoretically when length approaches to infinity). Therefore, for some times there was a well known joke "all the codes are good except the ones we know". Of course, it is not the truth today! This theorem is a basis for error control codes (i.e. for error detection and correction). These codes are used for transmission as well as for the information storage (magnetic disks, CD etc.). The error control coding theory flourished for many years based sometimes on very specific mathematical apparatus (e.g. Galois Fields!) or giving sometimes the results having to wait to be put in practice with a new technology. In fact, the theory of block codes and convolutional codes developed separately, but now they have many common points (e.g. trellis for block codes or "tail-

biting" of convolutional codes to obtain the corresponding block for turbo decoding).

The first area where the communication engineers were led by Information Theory was Deep Space Communications [10]. The corresponding "space" channel with white Gaussian (cosmic) noise was very near to the theoretic model (the corresponding joke was "we have found the channel for our model"). Error correcting codes were used and the corresponding code gains were obtained. As Jim Massey appropriately said, it was a "marriage made in heaven".

The base for the application of Information Theory was well known notion of channel capacity. For the additive white noise Gaussian channel with signal power P, noise power N and frequency band B, the capacity is:

$$C = B \cdot \log_2\left(1 + \frac{P}{N}\right). \quad (1)$$

It should noticed that, for higher signal-to-noise ratios, the doubling of the signal (transmitted) power (3 dB) increases capacity for only one bit per second per-Hertz of the bandwidth. By introducing the single-sided noise power spectral density N_0 [W/Hz] $\Rightarrow N=B\cdot N_0$ [W], the capacity of the channel having un unlimited frequency band (often called "the space channel") can be found as follows

$$C_{\mathrm{SC}} = \lim_{B\to\infty} B \cdot \log_2\left(1 + \frac{P}{B \cdot N_0}\right) = \frac{P}{N_0} \cdot \log_2 \mathrm{e}. \quad (2)$$

Therefore, the bandwidth is "infinite", but the capacity is finite, because the noise power is also "infinite". Here, the capacity increases linearly with the signal power, but an infinite bandwidth is needed.

To evaluate the efficiency of error control coding, the quantity "coding gain" can be defined giving the difference in signal-to-noise ratios (in dB) achieving the same error probability without and with the error control coding (Figure 2.[11]). Instead of the signal-to-noise ratio (P/N) it is customary for comparison to introduce E_b/N_0 ratio, where E_b is the "energy per bit." For the comparison sake the "uncoded system" is usually BPSK. To have a "fair" comparison of various systems, E_b is calculated (*not* measured) as the energy per information bit (and *not* per channel bit). The coding gain can be "negative" as well, especially for smaller values of E_b/N_0. In the figure the "Shannon limit", obtained from expression (2) is also indicated (calculated as $10\cdot\log(1/\log_2 e) = -1.59$dB). The point is that even when the band is infinite, allowing the application of the most powerful codes, the curve can not be "left" from the limit. In fact, this limit can be achieved when the ratio of number of information bits and the number of corresponding channel bits approaches to zero. For the "real world" coding, this ratio is "finite" being usually between 1/2 and 1/3. For such a case the limit is a few dB higher. In our case (Deep Space) the coding gain of more than 3 dB was achieved (from about 9 dB achievable by theory for the error probability 10^{-5}). Now, in that area, we are much nearer to the corresponding limit.

Error control is achieved by including the properly designed redundancy into the message before transmission (the redundancy possibly contained by the message itself can also be used!). The redundancy can be added either by inserting the extra bits (digits) – pariti-check symbols (*conventional coding*) or by expanding the used channel signal-set and, finally, by combination of both (*coded modulation*). The conventional coding lowers the bandwidth efficiency (i.e. either expands the bandwidth used or reduces the data rate). It is suitable for the channel where the power is limited (*power limited systems*), not bandwidth (e.g. the space channel).

The coded modulation neither expands the bandwidth used, nor reduces the data rate, but increases the efficiency at the expense of the increased power. Therefore, it is suitable for the *bandwidth limited systems* (e.g. the telephone channel). When bandwidth is limited, the *spectral efficiency* must be taken into account. It is the ratio of equivalent binary rate and the used bandwidth (the dimension is information bits per second per Hz).

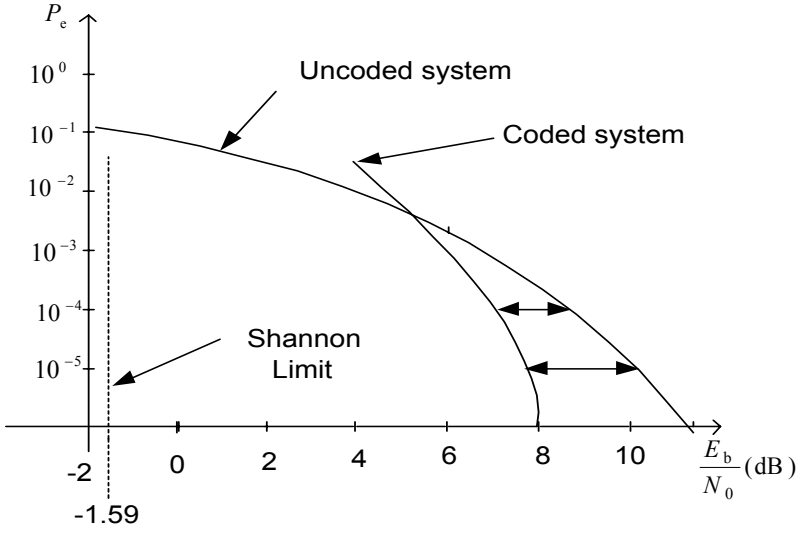

Fig. 2. Coding gain definition

In Figure 3. [11] the spectral efficiency vs. E_b/N_0 ratio is shown, for some fixed value of the error probability. The curve shown corresponds to the channel capacity. Therefore, the system cannot work reliably (i.e. keeping the error probability under control) in the area over the curve.

All systems can work (without or with error control coding) in the area under the curve. But, this area can be divided into two subareas. The systems having $R_b/B>1$ (R_b is the information bit rate) are mutlilevel systems (the baseband binary ransmission, where $R_b=2\cdot B$, is an exception) – they can be considered as the bandwidth limited systems, while the systems having $R_b/B<1$ are power limited systems.

For a fixed value of the probability of the error (P_e), the system is represented in a diagram by a point whose coordinates are E_b/N_0 and the corresponding achieved spectral efficiency. The change in P_e would only move the corresponding point in the diagram. The systems "closer" to the curve are more efficient spectrally than the systems being "more distant" from the curve. The "classical" telephone channel can be regarded as bandwidth limited Gaussian noise channel. The Ungerboeck's invention of trellis-coded modulation (TCM) [12], was practically the first case where we approached the theoretical limit given by Information Theory. In fact, even with commercial modems (with data rates 28800 b/s, and even 33600 b/s) we are very near now to the capacity of the telephone channel! The point corresponding to the *state-of-the-art* for a telephone channel (R_b=33600 b/s, P/N = 33 dB $\Rightarrow B$=3064 Hz, ρ =10.97, E_b/N_0= 22.61 dB) is denoted by an asterisk.

A new breakthrough in efforts to approach in praxis the ultimate performance limits (the "Promised Land") was done in 1993, when so-called *turbo codes* [13] were invented. Their name was given by analogy with the turbo engine and its efficient use of the feedback. However, it took several years to obtain a satisfactory understanding of the concept used [14]. The principle of iterative decoding with SISO (soft-in/soft-out) decoders using *maximum a posteriori* (MAP) algorithm, combined with interleaving, gave high coding gains and the obtained results were very near to the Shannon limit. This principle is also applicable to some other codes, and especially to LDPC (Low-Density Parity-Check) codes proposed by Gallager [15] almost half a century ago, who also considered some kind of iterative decoding. Furthermore, the "turbo principle" can also be implemented in the other communication fields as well (channel equalization, interference cancellation, multi-user detection). The main drawback is the delay inherent to the decoding with iteration.

It should be noted that capacity was defined for point-to-point transmission (channel). Now, the Network Information Theory is developing. It is a system with many senders and receivers (multi-user) and many new elements as well (delay, interference, co-operation, feedback etc.). The general problem is: Given many senders and receivers and the channel transition matrix (describing the effects of interference and noise), decide whether or not the sources can be transmitted over the channel (for example TDMA, CDMA). In fact we do not talk about channel capacity but about the whole medium capacity. The general problem has not yet been solved, but only for some special cases.

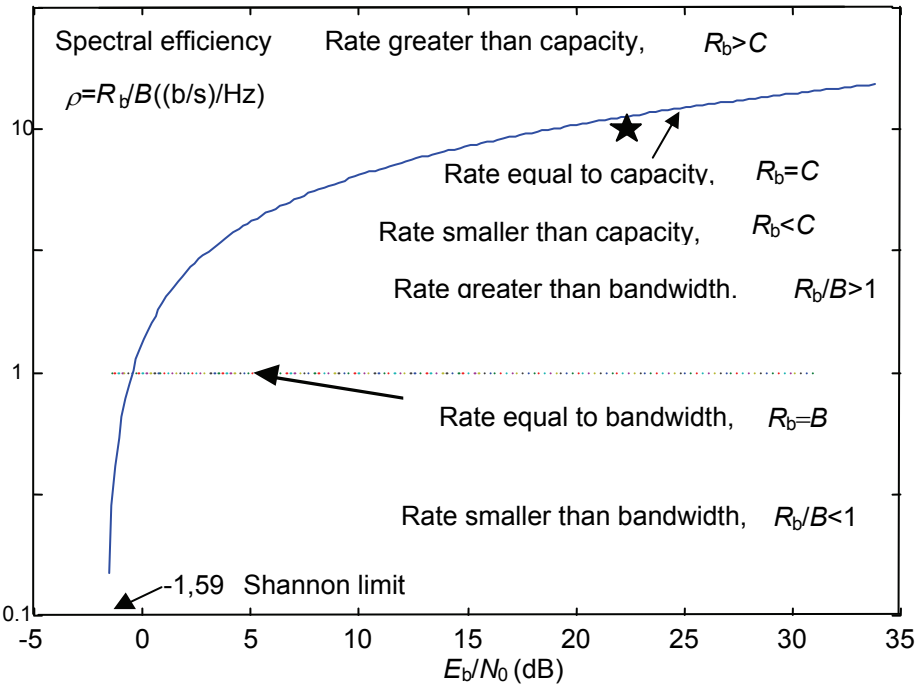

Fig. 3. Spectral efficiency vs. signal-to-noise ratio

In June 2006 the special issue on *Networking and Information Theory* was jointly issued by *IEEE Transactions on Information Theory* and *IEEE/ACM Transactions on Networking* showing the strong connection between Information Theory and Networking. The areas of interest were scaling laws in networks, network coding, multiusers systems, wireless network design, data dissemination algorithms, network utility maximization framework, queuing and delay issues in information-theoretic capacity settings.

The foregoing ideas led to a number of mainly theoretical papers concerning the "new" ultimate limits in wireless. The Information Theory performed its well-known role being the "conscience" of Communications. For systems with a single antenna link advances in coding and signal processing made it feasible to approach the Shannon limit very closely [16]. The next step was to increase the number of antennas at both the transmitter and the receiver (MIMO – *Multiple-Input Multiple-Output*). The corresponding theoretical framework was developed by Foshini and Gans [17] and separately by Telatar [18]. Practically at the same time the fundamental paper appeared concerning the space-time codes for wireless communications [19].

Under idealized conditions (yielding an upper bound – the "normal case" in Information Theory!) for the same number of transmitting and receiving antennas (n), the following formula for capacity is obtained [18]

$$C = nB \cdot \log_2\left(1 + \frac{P}{nN}\right). \tag{3}$$

The total transmitted power is the same as for a single antenna link, but here it is equally distributed over all the transmitting antennas. Further, the transmitted symbols are zero-mean independent identically distributed Gaussian variables (an old trick from the Shannon's Second Theorem!). Letting the number of antennas go to infinity, the following limit is obtained [15]

$$C_{\text{MIMO}}\big|_{\max} = \lim_{n\to\infty} C = B\frac{P}{N} \cdot \log_2 \text{e}. \tag{4}$$

Accordingly, a very important conclusion can be drawn. While, for a single antenna link the capacity increases logarithmically with the signal-to-noise ratio (i.e. for only 1 bit per second when SNR increases for 3 dB, for higher SNR), for MIMO systems the capacity increases much faster – linearly with SNR! Introducing $N=BN_0$ in (4) the same expression for capacity as in (2) is obtained, but there is a great difference. In (4) the bandwidth (B) does not change, while in (2) the bandwidth is infinite!

Of course, for real conditions, this increase will be smaller, but still much faster than that for a single antenna link case. With some "space-time coding" techniques spectral efficiencies from about 20 up to 42 b/s/Hz were obtained in laboratory conditions (indoor slow-fading environment). The notion of *space-time (ST) coding* was born with the appearance of MIMO systems. Here, the encoding is performed in the both domains (space and time). Now, we can talk about ST block codes, ST trellis codes, ST turbo trellis codes etc.). In October 2003 the special issue on *Networking and Information Theory* appeared concerning *Space-Time Transmission, Reception, Coding and Signal Processing*.

C. Rate Distortion Theory

While the Source Coding Theorem concerns to so-called ***non-destructive data compression*** where the original information must not be distorted, here the "wishes" of the user, concerning the "distortion" of the received information are taken into account. In fact, the user dictates the "fidelity criterion" and the average allowable distortion per symbol (***D***). The Rate Distortion Function ***R(D)*** is defined as giving the minimum mutual information between the source and the user needed for the average distortion being smaller than allowed (***D***).

Therefore, it is, in fact, ***destructive*** (in good will, of course!) ***(lossy) data compression***. This part of Information Theory is the basis for finding the limits when quantising the analogue signals (speech, picture (image) etc.). The theory can be applied to the discrete sources as well.

D. General impact of Information Theory on Communications

Firstly, it should be noted that all important quantities are obtained by statistical averaging. So, all results should be taken "on the average" – over long symbol sequences. But our basic aim is the same – our system should work efficiently on the long run.

Further, the ultimate limits in Communications are given by Information Theory, sometimes without the clear algorithm how to approach them. So, we know what we can do and what we cannot do. But often we have ourselves to find how to do that what we can do according to theory.

Last, but not least, let us discuss the capacity of the "human channel", i.e. what are our limits when we communicate (consciously) with the outside world. The information rate (flow) of the language is smaller than 50 shannons per second (according to the Information Theory it is near to 10 shannons (Sh) per second). So, we need only 50 bits (or less) per second to transmit speech. In practice, there are commercial vocoders using 1200 b/s or less. In laboratories we are under 300 b/s. For the intelligibility of the received speech, the capacity of our ear (hearing system) should not be greater than 50 Sh/s. In fact it is shown that the human being cannot consciously communicate (taking into account all five senses) with the surrounding faster then at the rate of 300 Sh/s. It is the "human channel capacity". So, there is still much room, especially to compress the pictures. Many standards were created or are created now for efficient speech and picture transmission (JPEG, MPEG1, MPEG2 etc.). We are now in picture (image) compression under a hundred kb/s and try to transmit a picture by a modem over the telephone channel having videoconferencing and a multimedia in view). From the Information Theory point of view it is possible to approach the rate of 300 b/s for picture transmission, but we still do not know how to do it. Of course, this limit can be approached only asymptotically.

E. Information Theory and other fields

Information Theory was born inside the Communications and primarily for Communications. Still, it sometimes pays to apply the fresh ideas from one field into another field.

Shannon, himself, formulated a theory of cryptography (secrecy systems) in terms of the concepts of Information Theory [20]. For some scientists the cryptology, after Shannon, "from an art became a science". He showed that the entropy of the language (in fact its complement – the redundancy) is related to the possibility for solving cryptograms in this language. For simple substitution cipher he calculated the minimum length of the cryptogram (the number of letters in it) needed to break the cipher The result was confirmed in practice. He also defined a "perfect secrecy" using entropy concept. The same approach can be used to obtain the average number of frames needed to obtain the frame synchronisation. In fact, this paragraph we could put in the previous section.

In Statistics, regarded as the science of extracting information from data, one has to expect the application of ideas of Information Theory. We will mention only the corresponding information measures as the base for hypotheses testing and maximum entropy principle for inferring the unknown distribution.

After some, not always very successful tries to apply Information Theory in Biology, a few years ago appeared T. Berger's paper concerning the Information Theory of living systems [21]. Besides the application of well known notions (and expressions) from Information Theory he gave as well "Time Discrete Model of Neural Coalition" and "Block Diagram of Neural Sensory Processing".

There were many attempts to find the optimal gambling systems as well as the growth rate optimal policy.

There were also many other attempts to apply the concepts of Information Theory in other fields (genetics, psychology, linguistics, etc.). Some of them gave results, some did not. Why?

In fact, this is the question of the model. The basic results of Information Theory are aimed at a very specific direction – a direction that may not be necessarily relevant to all fields. So, everyone trying to apply concepts of Information Theory in some other (his!) field should know also the mathematical foundations of Information Theory as well as its communication application. That would help him to evaluate the applicability of Information Theory concepts in his own field.

IV. Conclusion

At the end, a few thoughts about the future of Information Theory. It is better not to be a prophet, but some trends can be seen now.

Firstly, it is a further development of multi-user (Network) Information Theory.

Also, there will be always some new error control codes, as well as some new decoding algorithms for a known codes – the algorithms more suitable for technology to come. For example, some non-linear codes (Preparata) can be regarded as linear in some higher mathematical structures [22] facilitating the corresponding coding and decoding procedures.

Turbo codes, better to say iterative decoding algorithms, were first discovered and later properly understood. They were obtained by concatenation (parallel, serial or hybrid) of two or more convolutional codes and decoded by iterative decoding. They are very efficient at a very low SNR. In fact, these codes approach the channel capacity. There is still more room to try various codes as well as to implement more efficient decoding algorithms. LDPC codes are also very important, having in view their specific decoding process.

The possibility of "soft-decision" decoding for some well known and frequently used codes (e.g. Reed-Solomon (RS) codes [23]) should also be mentioned. In the book [29] dedicated to RS codes only, one can find the following sentence: "With applications ranging from digital audio disc player to the *Voyager* Spacecraft, Reed-Solomon codes are the most frequently used error control codes in this corner of the galaxy". The similar ideas with soft-decision can be tried for some other codes, too.

An interesting unsolved theoretical problem is the "zero-error channel capacity", i.e. the capacity of channel without errors.

At the end (?) Quantum Information Theory is now flourishing. One citation will be sufficient: "*It has become clear that an information theory based on quantum principles extends and completes classical information theory, somewhat as complex numbers extend and complete the reals. The new theory includes quantum generalizations of classical notions such as sources, channels and codes, and two, complementary, quantifiable kinds of information – classical information and quantum entanglement*" [10]. Without going further in this "green" (quantum!) field, the first quantum error-correcting (block) code (nine-qubit single-error-correcting) was discovered (invented?) in 1995 [25]. Now, the theory of quantum codes greatly evolved including convolutional codes as well [26]. The crux of the matter is that such a code "protect quantum states from unwanted perturbations, allowing the implementation of robust quantum computing and communication systems".

In any case, we should remind ourselves that on the cover of *IEEE Transactions on Information Theory* there was once written that "the boundaries of these transactions are deliberately not sharply defined". Now, it is "A Journal Devoted to the Theoretical and Experimental Aspects of Information Transmission, Processing, and Utilization". In the last numbers of *IEEE Transactions on Information Theory* the papers from the following domains were published: Channel Coding; Communications, Detection and Estimation (and Inference); Communications and Shannon Theory; Shannon Theory and Source Coding; Communication Networks; Quantum Information Theory. The last volume of the transactions (Vol. 52, 2006) reached the enormous number of pages (4444+92(Index)), showing that Information Theory is very much alive and suggesting possibly the need for some further branching of the domains.

REFERENCES

[1] D. Kazakos, P. Papantoni-Kazakos, *Detection and Estimation*, Computer Science Press, New York, 1990.

[2] N. Wiener, *The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications*, John Wiley & Sons, Inc. New York, 1949.

[3] N. Wiener, "Generalized Harmonic Analysis", *Acta Mathematica*, Vol. 55 (1930), pp. 117-258

[4] C. E. Shannon, "A Mathematical Theory of Communication", *BSTJ*, Vol. 27, pp. 379-423 (July 1948), 623-656, October 1948.

[5] H. Nyquist, "Certain Topics in Telegraph Transmission Theory", *Trans. of the AIEE*, Vol. 47. pp. 617-644, April 1928.

[6] R. V. L. Hartley, "Transmission of Information", *BSTJ*, Vol. VII, pp. 535-563, July, 1928.

[7] Dusan Drajic, Dragana Bajic: "Information theory after 50 years", Plenary invited lecture, *TELSIKS'97*, Nis, pp. 5-12, October 1997.

[8] C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion", *IRE Nat. Convention Record*, Part 4, pp. 142-163, 1959.

[9] C. E. Shannon, "Prediction and Entropy of Printed English", *BSTJ*, Vol. 30, pp. 50-64, January 1951.

[10] J. Hagenauer, "The Impact of Information Theory on Communications", *IEEE Inf. Theory Society Newsletter*, Special Golden Jubilee Issue, Summer 1998, pp. 6-8.

[11] D. Drajic, D. Bajic, "Communication System Performance: Achieving the Ultimate Information-Theoretic Limits?", *IEEE Comm. Mag.*, Vol. 40 (2002), No. 6, pp 124-129.

[12] G. Ungerboek, "Channel coding with multilevel/phase signals", *IEEE Trans. on Inf. Theory*, Vol. IT-28, pp. 55-67, Jan. 1982.

[13] C. Berrou, A. Clavieux, P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding Turbo- Codes", *Proc of ICC'93*, Geneve, pp. 1064-1070, May 1993.

[14] M. Fossorier, S. Olcer, "Capacity Approaching Codes, Iterative Decoding Algorithms, and Their Applications", *IEEE Comm. Mag.*, Vol. 41 (2003), No. 8, pp 100-101.

[15] R. Gallager, "Low-Density Parity-Check Codes" , *IRE Trans. Inf. Theory*, Vol. 7 (1962), pp. 21-28.

[16] B. Vucetic, J. Yuan, *Turbo Codes – Principles and Applications*, Kluwer Academic Publishers, Boston 2000.

[17] G. Foschini, J. Gans, "On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas", *Wireless Personal Communications*, Vol. 6. (1998), pp. 311-335.

[18] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels", *European Transactions on Telecomm.*, Vol. 10.(1999), No. 6., pp. 585-595.

[19] V. Tarokh, N. Seshardi, A. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction", *IEEE Trans. Inf. Theory*, Vol. 44 (1998), No. 2, pp. 744-765.

[20] C. E. Shannon, "Communication Theory of Secrecy Systems", *BSTJ*, Vol. 28 (1949), October, pp. 656-715.

[21] T. Berger, "Living Information Theory", *IEEE Inf. Theory Society Newsletter*, Vol. 53 (2003), No. 1, March pp. 1,6-19.

[22] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, P. Sole, "The Z_4-linearity of Kerdock, Preparata, Goethals, and Related Codes", *IEEE Trans. Inf. Theory*, Vol. 40, pp. 301-319, March, 1994.

[23] J. S. Vuckovic, B. S. Vucetic, "Maximum-Likelihood Decoding of Reed-Solomon Codes", in *1997 IEEE Intern. Symp. on Inf. Theory*, Ulm, June 29-July 4, 1997, p. 400.

[24] S. B. Wicker, V. K. Bhargava (Eds.), *Reed-Solomon Codes and Their Applications*, IEEE Press, New York 1994.

[25] P. Shor, "Scheme for Reducing Decoherence in Quantum Computer Memory", *Phys. Rev. A,* Vol. 52 (1995), pp. 2493-2496.

[26] G. D. Forney Jr., M. Grassl, S. Guha, "Convolutional and Tail-Biting Quantum Error-Correcting Codes", *IEEE Trans. Inf. Theory*, Vol. 53 (2007), pp. 865-880.