IEEE Information Theory Society Newsletter

Vol. 42, No. 3, September 1992 Editor: Nader Mehravari

Reminiscences Related to the Shannon Lecture[1]

Andrew J. Viterbi[2]

Preamble:

At Newsletter Editor Nader Mehravari's request for a summary of my lecture, I reviewed my presentation entitled "Shannon Theory from a Communication Engineer's Perspective." Unfortunately I found that it overlapped to a large extent with an article which has since been published in the *IEEE Communication Magazine*.[1]

To satisfy the spirit of the request I submit instead the following. Although it is written for a wider audience, I believe it captures the spirit of some of my personal remarks during the Shannon Lecture.

Algorithm, Competitions and applications:

I am writing these reminiscences in Torremolinos, Spain on the occasion of WARC'92,[3] trying to re-

continued on page 3

1. Editor's Note: This paper represents a summary of the Shannon Lecture delivered by the author at the 1991 IEEE International Symposium on Information Theory, Budapest, Hungary, June 1991.

2. The author is with the QUALCOMM Incorporated, San Diego, CA.

3. World Administrative Radio Conference

HDTV[1]

Murat Kunt[2]

1. Introduction

What is HDTV? Before answering this question let's discuss first a few aspects of the well-known TV, then we will introduce HDTV. Doubtlessly, the best tool man ever made for mass education is television. The main issue in TV is that of programs. Instead of putting all the necessary efforts into high-quality programs for better education, look at the mess we have done with it. We broadcast greed, violence, hate, injustice, religious crookery, Hollywood bad taste and, if there is time left, some superficial and biased information. Our good old TV system is undergoing important changes to improve its technical quality, and what may happen is to end up with much better images showing poorer quality programs. It may well be another example of improving technicality, but decreasing global quality. Some other well-known cases are fast-foods, 8 mm films which became super 8 and then became video with increased

continued on page 6

1. Editor's Note: This paper represents a summary of a Plenary Lecture delivered by the author at the 1991 IEEE International Symposium on Information Theory, Budapest, Hungary, June 1991.

2. The author is with the Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland.

Reminiscenses Related to the Shannon Lecture

continued from front cover

create events half a world away and half a career ago. It was over a quarter century ago on a quiet weekend afternoon of March 1966 in Los Angeles, California that, motivated by the curiosity of the research scientist and of the teacher who seeks a better way to explain a basic phenomenon to his class, I came upon the explanation of the performance capability of the convolutional code, and ultimately of many related signal processing schemes. I had struggled in my mind with the principles of sequential decoding, already well established by several MIT researchers, foremost among them Jack Wozencraft and Robert Fano. While the conceptual development of sequential decoding as a "tree climbing" optimization algorithm was elegant and effective, it appeared to me that there must be a simpler, more elementary, approach to the decoding of this rich and powerful class of error-correcting codes.

I approached the problem from the viewpoint of classical information theory, as established by Shannon almost two decades earlier and as refined by numerous researchers through the development of ever more detailed and revealing bounds on the error probability of error-correcting codes. From the perspective of a long period of communication practice, I now recognize how absurdly overcomplicated my approach was. A direct explanation of what has since become known as the "Viterbi Algorithm" followed immediately from recognizing the splitting and recombining nature of the convolutional code, and in fact of any numerical sequence generated by a finite-state machine, whether linear or nonlinear and whether over a finite or infinite alphabet. Today this feature is most often described by a trellis structure, and convolutional codes and their generalizations are called trellis codes. Furthermore, abundant applications of the same principle have been found throughout signal processing applications and similarly described by a trellis diagram. (I personally favor the more traditional state diagram of the finite-state machine as the more revealing description of the phenomenon and algorithm.) Yet, when in the Spring of 1966 I came upon the approach, I described it in terms of a peculiar tennis competition among members of a single family descended from a common ancestor wherein winning players were subsequently matched against distant cousins descended from the common ancestor, but with no direct ancestors in common for at least K generations (where K is the "constraint length" of the convolutional code). So concerned was I then with fundamental information theory that I conceived the algorithm for the purpose of establishing exponential upper and lower bounds on error probability and concerned myself largely with showing that the exponent was positive over the region of information rates below the channel capacity. Further, I showed it to be asymptotically exact over the upper subset of that region. The upper bound, in fact, was already known, having been shown for sequential decoding, but the lower bound and the method to attain it was new. At the same time, I lost sight of one of its valuable characteristics; namely that for memoryless channels it yields a maximum likelihood decision of the transmitted sequence. It was David Forney who, upon reading the manuscript before publication, recognized this attribute and also presented the trellis interpretation.

Among the questions I was asked to address in this article, presumably as a model for a new generation of researchers, was "On what occasion and under what circumstances did the idea present itself?" The question is easily answered but the reply is hardly revealing of any profound approach to research. After having struggled with the issue for several days (or weeks), the peculiar competition analogy came to me as I was sitting quietly watching the judging of a costume competition in which my small children were participating (and in which they won prizes).[4]

To return to matters of substance, I suspected then, as I do now, that the concept is such a natural outgrowth of the finite-state machine structure of convolutional codes and many other "filtering" processes that others had previously pursued the same thought processes but had discarded the approach as being too obvious, and more significantly, too impractical. This is much like a prospector for gold who stumbles across a bright nugget but upon closer examination decides it is too well shaped to be of any real or practical value.

Having approached the "nugget" with a theoretical and even a pedagogical goal, when I examined the result from the context of pure information theory, I was unsure as to whether it was sufficiently important to merit publication; true, the lower bound was new, but the fact that information rates up to channel capacity could be achieved with a positive error exponent

4. Westerners may recognize this festive event originating two and a half millennia ago as the redemption of the Jews of Persia from the threat of a genocide of an earlier era. It is known as the Festival of Purim and is of particular importance to children.

was well known. This is not to say that I was not excited at defining the algorithm and particularly by the simplicity of the proofs of the asymptotic theorems. But it was my friends and colleagues who encouraged me on its importance and the value of publishing all my results. Jim Massey, then at Notre Dame, wrote the first congratulatory letter, having read the manuscript carefully and catching a non-critical but significant error. Bob Gallager of MIT also wrote of its potential significance, and, as noted, Dave Forney of Codex was the first to build on and clarify the result.

Still, in mid 1966, while some recognized value in the simplicity and natural evolution of the approach from elementary principles, none foresaw any practical value. At an early presentation to a small group in a respected southern California electronic company, then on the forefront of satellite communication technology, I quickly lost the senior members of my audience with the statement that the decoder might require on the order of a thousand memory shift registers to implement.

The state of 1960's technology aside, the practicality of the algorithm could hardly be judged given that it had been presented in terms of asymptotically long constraint lengths. My young colleague, Jerry Heller, having completed his doctoral dissertation on sequential decoding at MIT under Irwin Jacobs, ran the first computer simulations of the algorithm in early 1969 at the Jet Propulsion Laboratory using short constraint lengths: below K=10. His conclusions were most surprising - that a coding gain of almost 6 dB could be achieved with a constraint length as small as K=7. Since the number of states, or memory registers, of the decoder equals 2 raised to the power K-1, this meant that with only 64 registers, a given performance level (i.e., one error in one million bits) could be achieved with a received signal-to-noise ratio which is almost one-quarter that required by systems not employing coding. The implications of this realization were immediate: space vehicles, the focus of JPL, could travel nearly twice as far from earth without quadrupling their transmitter power or reducing data rate proportionally. Space missions began to employ convolutional codes in the seventies, some using sequential decoding but eventually most with the shorter codes and the new algorithm. Military communications, long plagued by very noisy environments usually caused by intentional interferes or "jammers," very quickly applied the technique to military satellites, and gradually the international and domestic communication satellite community incorporated the algorithm into its digital communication systems.

The pioneering role in creating a de facto standar based on the K=7 Viterbi decoder was played by small group of very talented communication engineer in a small company, LINKABIT, of which Irwin Jacob and I were co- founders in the late sixties. The earl struggle to render the algorithm practical, using th only slightly integrated circuits of the early seventie was led by Jacobs, Heller, Andrew Cohen and Klei Gilhousen. The competition for the earliest militar and NASA contracts for the technology was truly "David and Goliath" confrontation, pitting tiny LINK ABIT against some telecommunication companie more than one thousand times larger. Luckily for th U.S. government and the future of the industry, LINK ABIT's proposals showed that clever design could prc duce better results with about 10% as much hardwar as the competition. Out of this came the LV7017 (KY80 in U.S. military nomenclature), a constraint length K= decoder which operated at data rates up to 1 Mbits/sec. The design principles employed in this pio neering implementation are still employed today; bu the approximately 250 medium-scale integrated cir cuits of the mid seventies have been replaced by a single VLSI circuit which can operate several time faster. Early military and NASA applications involvec only hundreds of modems equipped with our decod ers. The proliferation of VSAT terminals and mor recently mobile satellite terminals has increased thei usage into the hundreds of thousands.

The evolution of the algorithm as a key component ir digital satellite communications was predictable sinc the late 60's. The surprise came with its application ir narrowband wireline telecommunications, magnetic recording and ultimately terrestrial digital cellular anc wireless personal communications as well. The firs step toward this much broader application came from Forney's recognition that the intersymbol interference resulting from linear filtering, ever-present in the com- munication transmitter and channel, can be viewed as the real number equivalent of a convolutional encoder and described by a trellis diagram. Thus the optimum equalizer for the channel utilizes the same algorithm. The next step was due to Ungerboeck, who recognized that trellis coding using large symbol alphabets, geometrically mapped onto 2-dimensional signal space, could produce notable coding gains for highly bandwidth-limited channels characteristic of wireline digital transmission. Hence, since the mid eighties all wireline data modems above 9600 baud have employed trellis codes, although simpler, more practical channel equalization has been combined with trellis decoding. This significantly increased the potential

usage of the algorithm as FAX machines and ultimately PC modems begin to employ higher speed wireline modems.

Another major application with possibly even greater potential is in the magnetic recording industry. The channel in this case, though nonlinear, possesses the attributes of a finite-state machine and hence lends itself to trellis decoding. The approach is called "partial response maximum likelihood (PRML) detection" in the magnetic recording field. Considerable research over the last several years has established the underlying signal processing technology on a firm foundation. An excellent recent review article by Siegel and Wolf [2] describes the evolution from peak detector to PRML detector technology for magnetic recording and its relation to communication technology. It seems likely that millions of devices employing the algorithm will eventually be in common usage.

The most recent convert to digital communication techniques is radio telephony, now known more commonly as cellular or wireless personal communication. Many of the emerging standards are derivatives of communication satellite technology. Yet, while the algorithm is employed for both convolutional decoding and maximum likelihood equalization in two of the early proposed digital cellular standards, the coding is not integrated effectively into the multiple access terrestrial propagation environment, which limits its value. The difficulty in communication through such environments, plagued by multipath fading, other-user and other-cell interference, is inherent in narrowband channelization. With wideband continuous-power waveforms, such impairments are much more effectively mitigated and powerful error-correcting coding becomes much more productive. Such wideband "spread spectrum" techniques often referred to as code-division multiple access (CDMA) also have their origins in Shannon's information theory, as recently described in detail elsewhere. Without elaborating further here, we note only our strong belief that such information theoretically inspired, properly matched modulation and coding techniques will make possible the goal of the ubiquitous wireless personal communication device, portable and equally useful on the street, in the office and in the home, providing universal access to the public switched telephone network.

This direct wireless access to the local loop in urban areas for mobile as well as for stationary individuals is achievable for most of the urban inhabitants of the developed nations within this decade. For rural inhabitants, as well as many of the developing nations, the infrastructure for this service is lacking. For such usage, the solution may be a network of low earth-orbit satellites with simple "bent-pipe" transponders serving, through a gateway earth station in each region, as the infrastructure for the universal personal wireless system. Here again, wideband technology offers the simplest and most frequency-efficient approach. This is a major current topic of discussion which has particular impact on the deliberations at this year's WARC.

Let me return to the basic theme of this article and address the algorithm from the perspective of my overall career. While I cannot deny my good fortune in being in the right place (university professor with strong ties to industrial and government R&D laboratories) at the right time (mid sixties) to recognize the obvious (in retrospect), which many of my colleagues could equally well have done, I do not regard this algorithm as an isolated event in my career. I was privileged to have begun in the late fifties with applications of coherent digital communication to military telemetry links employing spread spectrum techniques, crude in implementation by today's standards but sophisticated in system concepts. From phase locked loops for tracking weak signals from missiles and later from NASA's deep space vehicles, I migrated to digital communication systems and particularly error-correcting codes. After publication of the algorithm and related theoretical papers, I turned back toward applications of digital communication, which in the early seventies were mostly for space and military systems. The first effective introduction of coding techniques into military satellite communication systems was quickly followed by their adoption by the commercial satellite industry, all following the lead of our team at LINKABIT, thus establishing an ad hoc standard without requiring the burdensome proceedings inherent in "standards committees."

Upon moving on to found QUALCOMM in 1985, we improved on these techniques to create cost effective mobile satellite systems and ultimately terrestrial mobile and wireless personal communication networks. All this, of course, was made possible only through the remarkable evolution of solid state devices, which reduced multiple racks of electronics equipment down to a single integrated circuit. All the communication system concepts I have worked on for 35 years would be a mere unachievable academic curiosity were it not for this phenomenal evolution of solid state circuit integration, which has reduced the cost of digital electronic devices by many orders of magnitude, even while proportionately reducing their size and com-

plexity. It is both encouraging and challenging that solid state integration reduces size and increases speed of devices by a factor of two approximately every two years. Thus one decade's limit on feasibility is revised upward by more than an order of magnitude in the subsequent decade.

The exciting journey in the evolution of telecommunication systems through digital technology continues. Fueled by physically-based solid state technology but steered by mathematically-based information and communication theory, the end is not in sight. What is fairly clear is that all transmission will be digital by the first quarter of the next century. It will be integrated with smart terminals offering many services and, while the fiber network backbone will reach to within a few tens of meters to a few kilometers of the user's personal terminal, the final direct access to the individual will be mostly wireless. Error-correcting coding is certain to play a key role in the realization of the entire network.

I feel privileged to have played a part, along with many other communication engineers and scientists, in bringing this about. I have been particularly fortunate to be able to influence the field through industrial and entrepreneurial ventures and through professional activities, as well as by teaching and research. While the early years of my career were filled by the alternating exhilaration and disappointment inherent in the research process, I occasionally recapture those emotions, though highly attenuated by time and the pressure of other concerns and duties. Nevertheless, research and its handmaiden, teaching, has always added spice to my already full and satisfying existence. The excitement at finally getting the right answer to a puzzling and initially not well formulated question exceeds even the satisfaction afforded by recognition for one's achievements.

Postamble:

To which I would now add my introductory comment in Budapest that the Shannon Lecture was a great challenge for me in that this recognition by my peers, the most knowledgeable critics of my work, was the highest honor that I could receive.

References:

[1] A.J. Viterbi, "Wirelsss Digital Communication: A View Based on Three Lessons Learned," *IEEE Communications Magazine*, pp. 33-36, September 1991.

[2] P.H. Siegel and J.K. Wolf "Modulation and Coding for Information Storage," *IEEE Communications Magazine*, Vol. 29, No. 12, pp. 68-86, Dec. 1991.

HDTV

continued from front cover

practicality but poorer quality, without forgetting pu lishing and book writing. Today everybody write nobody reads, and the same is true for scientific pape as well. "Cut & paste" helps proliferation and kills qua ity. Nowadays, films are still edited by mechanical c and paste, which is rather time consuming. If this ev becomes electronic and more efficient, imagine ho many cheap and bad-taste films one can produce p week to pollute all the TV channels around the worl

2. Short history of TV

The first attempt to try to convert a visual scene intc electrical signal was done by Nipkow in 1884. Fro improvement to perfection, we move to the iconosco of Zworkin in 1923, and then to the black and white T tubes in 1941, and finally to color in 1950. Let us sur marize in a few words how a TV system works. It is typical communication system including a transmitt a channel and a receiver. To be more specific, consid Figure 1, where such a system is shown. The first fa is that the 4-D space we are living in (three spa variables and one time variable) is compressed in t system to a 1-D signal which is transmitted in t channel, and then expended to a 3-D signal which displayed in front of the viewer. One dimension is lc by projecting the real world into an image plane, a the remaining 3-D space is sampled in time (thir images per second in the US or 25 images per seco in Europe). 2-D image space is scanned line by li producing a signal called a video signal. If line by li scanning and time sampling is repeated fast enoug one creates the illusion of a continuity to fool the ey

3. Video signal

Since we are transmitting the video signal over t channel, it is interesting to determine its bandwid Assuming a square screen to display the image of si N x N points refreshed every F second, the high frequency we can have is $N^2F/2$. For example, for frames per second and 625 lines per screen, the ba bandwidth of the video signal is around 5 MHz. Not that so far there is no NTSC, PAL or SECAM. T previous bandwidth represents a black and white i age. Color was introduced through 3 primaries. The are either additive primaries like red, green and bl (RGB), or subtractive primaries, like yellow, magen and cyan. A combination of either group of primar may reproduce any desired color. Furthermore, it is r necessary to work with any of these primaries. A reversible three by three transformation is acceptat For example, from the three additive primaries, R,