IMPROVING THE NOISE PERFORMANCE OF COMMUNICATION SYSTEMS: 1930S TO EARLY 1940S

MISCHA SCHWARTZ

INTRODUCTION

The period after World War II saw an extraordinary development of activity in the field of communication theory, involving optimum reception and detection of signals in the presence of noise. This work came directly out of work done during the war on radar and control systems. Yet the radar work itself was an outgrowth of work beginning circa 1920 on improving the performance of communication systems in the presence of noise. We have previously reported on work carried out in this area during the 1920s in both radio (wireless) communications and wired telephony [1]. In this column we focus on work done in the 1930s and early 1940s, which both enlarged on and saw considerable strides ahead in these earlier studies involving noise in communication systems. We do this by presenting developments during this period of time in four interrelated and roughly chronological areas:

- Work on FM and pulse code modulation (PCM) showing, for the first time, that noise could be reduced by purposefully increasing the bandwidth (now known as the noise-bandwidth trade-off)
- Studies attempting to understand the statistical properties of noise, leading to its now well-known Gaussian amplitude characteristic
- Studies on means of reducing noise internal to receiving systems by introduction of the "noise figure" concept
- Recognition that "matched filtering" provided optimum signal detection in noise

Armstrong's path-breaking work on low-noise wideband FM in the early 1930s, which first demonstrated the noise-bandwidth trade-off, has already been described in these pages [2]. Hence we begin, in the next section, with Alec Reeves' invention, in1937, of pulse code modulation (PCM), which again provided a noise-bandwidth trade-off. Armstrong's system was successfully demonstrated and its noise-reduction properties confirmed soon after the system was announced. Reeves' invention was not implemented until years later, however, during the 1960s, when the technology had advanced enough to make its commercialization feasible.

Communication engineers working on improving the noise performance of communication systems had, until the late 1930s, used the ratio of average signal power to average noise power, or signal-to-noise ratio (SNR), as a measure of system noise performance. The period of the late 1930s to early 1940s saw a need to introduce statistical concepts as well in improving system performance. The normal or Gaussian distribution had been used for years by mathematicians and physicists in studying the statistical properties of various physical phenomena, but it was not until a paper was published in 1941, in the *Proceedings of the IRE*, that most communication engineers first became aware that noise statistics had a Gaussian distribution, and that this phenomenon could be applied to the study of noise in communication systems. We discuss this paper and experiments performed on noise prior to its publication in the section covering the period of the late 1930s to early 1940s. The section following that, on the early 1940s, provides a summary of work carried on in both the United States and Europe to provide a better and consistent measure of the noise introduced internally in a communication system, as well as the noise introduced externally, at the input to a communications receiver. This work led to the introduction of the *noise figure* concept. This work was further solidified by engineers and physicists working on radar signal detection studies all through World War II. We conclude this paper by describing the history of the "matched filter" concept, a technique designed to optimize the detection of signals in noise. A rudimentary form of this concept appears in a 1925 paper by John Carson of AT&T [1]. It was not until the period of World War II, roughly 20 years later, with an imperative need in radar to detect small signals in noise, that workers at a number of facilities engaged in the war effort almost simultaneously worked out this concept analytically and verified it experimentally. This work led directly to the amazing growth of communication theory and related fields after World War II noted at the beginning of this paper.

1930S: NOISE-BANDWIDTH TRADE-OFF

Two inventions in the 1930s changed the study of the impact of noise on system performance dramatically. These were, first, Armstrong's invention of wide-deviation (wideband) FM in 1933, and then, afterward, Reeves' independent invention in 1937 of (wideband) PCM. Armstrong's recognition that purposely widening the bandwidth of the FM transmission signal, by increasing the frequency deviation, resulted in a reduction in noise at the communication receiver output was a major achievement. Within 10 years of Armstrong's invention, with the PCM invention coming soon after, it began to be understood by the telecommunications community that both of these systems were in a class of communication systems for which one could trade off increased bandwidth for reduced noise or improved SNR. The invention of wideband FM, in particular, resulted in a flock of technical papers attempting to explain the noise improvement obtained, leading in turn to an increased understanding of the modeling of fluctuation noise in communication systems. As noted above, Armstrong's invention has already been described in these pages [2]. Hence, we focus here on Reeves' achievement.

Alex H. Reeves, a British engineer working for Standard Telecommunications Laboratories of the International Telephone and Telegraph Corporation, invented PCM in May 1937 while based in the Paris Laboratories of ITT. In PCM analog signals are converted to digital signals by first sampling the analog signals at a regular (periodic) rate, then "quantizing" each signal sample to the closest one of a set of discrete numbers or signal levels. The effect of the quantization procedure at a receiving system is to make the signal appear as if it had been received in a type of noise called "quantization noise." By choosing the set of discrete signal levels large enough, the quantization noise may be reduced to any tolerable level. The representation in digital format allows conversion of each digital number to its binary equivalent. Binary numbers, in turn, are much more easily recognized in the presence of fluctuation noise, resulting in an improvement in the detectability of the signal, a trait Reeves specifically drew on from his knowledge of telegraphy. But the conversion of each digital signal sample to its binary form reduces the time to transmit each signal, increasing the bandwidth required for transmission. There is thus the trade-off noted above, just as was the case with wide-deviation FM [2], between improving the signal-to-noise ratio and increasing the bandwidth.

Reeves filed three very similar patent applications on his invention: one in France in 1938, granted as patent 852,185, October 3, 1938; another in Britain, filed and granted in 1939; and a third, filed in the United States Nov. 22, 1939, and granted February 3, 1942, as U.S. patent 2,272,070. The U.S. patent has as its title "Electric Signaling System," and its first words indicate the invention is designed to reduce noise: "The present invention relates to electric signaling systems, and more particularly to systems designed to transmit complex wave forms, for example, speech, *which are practically free from any background noise*. The main object of the invention is to provide electrical signaling systems which *practically have no background noise*… " [Emphasis added]. He was well aware of the noise improvement-bandwidth trade-off, stating, further on in the patent, "The arrangements proposed necessitate a *slight* increase of the width of the transmission frequency band." (The word slight may be somewhat of an underestimate, since the bandwidth required for transmission depends on the number of binary pulses or bits required to represent the digital signal, but the important point is that he recognized the necessary increase in bandwidth.) He then goes on to describe the analog-to-digital conversion process, requiring sampling of the signal "at predetermined instants," with the signal amplitude range "divided into a finite number of predetermined amplitude values according to the fidelity required."

Writing about his invention many years later, in an unpublished 1964 paper [3], Reeves states "In 1937 I realized… that it [PCM] could be the most powerful tool so far against interference on speech — especially on long routes with many regenerative repeaters, as these devices could easily be designed and spaced in such a way as to make the noise nearly non-cumulative." (Note that this latter comment by Reeves provides another, very important, advantage of PCM over straight transmission of analog signals such as speech.) He also notes, in this paper, that PCM "is a good example of an invention that came too early… When PCM was patented in 1938 [French patent] and in 1942 [U.S. patent], I knew that no tools then existed that could make it economic for general civilian use. It is only in the last few years in this semi-conductor age that its commercial value has begun to be felt." This last sentence refers specifically to work begun by AT&T Bell Laboratories during World War II, and completed successfully in the early 1960s, in commercializing PCM. Referring to the effect of signal quantization, he notes in the same paper, "The quantizing noise was foreseen… " So Reeves knew at the outset the basic attributes of his invention: the ability of PCM to provide reduction in fluctuation noise added during transmission with an accompanying increase in transmission bandwidth, the consequent introduction of quantization noise, and the need to choose sufficient discrete signal levels to keep the quantization noise low enough and the clarity of the received signal good enough to satisfy the users for whom the transmission is intended.

Further detailed discussion of the process of the invention by Reeves of PCM appears in the 1969 book by K.W. Cattermole, *Principles of Pulse Code Modulation* [4], and is based on private discussions Prof. Cattermole had had with Reeves. In the book, Cattermole states that Reeves invented PCM in May 1937 as " a conscious attempt to realize, in the transmission of continuous signals, the noise-immunity characteristic of the telegraph." Cattermole is here referring to the fact that simple discrete signals, as exemplified at the time by telegraph signals, are more immune to noise than analog signals. ITT engineers in Paris, Reeves among them, had been studying the use of pulse methods over microwave transmission systems. Bandwidth was no problem in this case, but noise and distortion were more severe than over cable, commonly used for communication at the time. Sampling and time-division telephony were known empirically; telegraphy was well understood both in theory and in practice. The use of pulse time modulation (PTM) was first investigated. In this technique the amplitude of an analog signal is converted to variations in time of a sequence of signal pulses. Cattermole relates that Reeves told him that he had come up with the PCM invention by combining "two of his earlier concepts: PTM and a time scale defined by binary counters… [Reeves] recognized PTM as being completely digital… [but] impairment would be cumulative … with many repeaters. However, if the time position were specified in some numerical way, the signal would be truly digital like a telegraph." [4, pp. 21, 22] Reeves then apparently thought up the idea of converting from a digital time scale to one in amplitude, and the concept of PCM was born!

LATE 1930S–EARLY 1940S: STATISTICAL REPRESENTATION OF NOISE

The work on improving the noise performance of communication systems in the period prior to the late 1930s focused on maximizing the SNR, defined in terms of mean signal and mean noise powers. Interestingly, however, the communication literature throughout much of this time had nothing to say about the *statistical* characterization of noise. It is now routinely accepted throughout the communications literature, as well as in the design of communication systems, that

(Continued on page 22)

(Continued from page 20)

fluctuation noise has a Gaussian or normal probability amplitude distribution. Yet it was not until the publication of a paper by V. D. Landon of RCA in 1941 that the Gaussian representation of noise first appeared in the technical communications literature [5]. This seems somewhat surprising, considering that the normal distribution itself was well known and had been shown over the years, by many mathematicians and physicists, to arise from the "law of large numbers" or the Central Limit Theorem in a host of applications. Physicists and mathematicians, including preeminent figures such as Einstein and Laue, were well aware of the Gaussian distribution arising in many physical processes and used it in their analyses, but its application by engineers to the modeling of fluctuation noise, as noted above, took much longer.[1]

Telecommunication engineers at the time were interested in statistics beyond those of the second moment incorporated in the signal and noise mean powers. In particular, experimental studies were made by a number of engineers at various laboratories on the now obviously incorrect concept of the "crest factor" of noise, the ratio of the "highest peak value" [sic] of the noise to its root mean squared (rms) value, coming up with different measured values of this quantity [7, 8]. V. D. Landon himself is the author of a 1936 paper on the measurement of the crest factor in which he notes that measured values of this quantity by different investigators differ substantially [7]. It was not until five years later, in 1941, as noted above, that he finally came up with the Gaussian characterization of noise [5].

The summary to this paper begins as follows: "The purpose of this paper is to show that fluctuation noise has a statistical distribution of amplitude versus time which follows the normal error-curve [i.e., normal distribution]." Landon goes on to note that "the term 'crest factor'… would seem to be a misconception." In his analysis Landon points out that the Law of Large Numbers applies in the case of fluctuation noise, stating "The foregoing paragraphs prove that the summation of a large number of small sinusoidal components [i.e., the model for noise] follow the normal-error law… " Later in the paper he summarizes his result, stating, "if the noise is primarily of the type called fluctuation noise or hiss then the normal-error law does apply." The normal-law or Gaussian nature of noise described in this paper was initially challenged. A brief follow-up paper by K. A. Norton in the September 1942 issue of the *Proceedings of the IRE*, as part of a discussion interchange [9], contested these results! It turned out that Norton, relying on an 1880 result by Lord Rayleigh, had mistakenly thought the Rayleigh distribution, rather than the Gaussian distribution, should apply to noise. Landon, correctly answering Norton, pointed out that Norton had confused the instantaneous amplitude of the noise with its envelope, the latter giving rise to the Rayleigh distribution. (Norton, in concluding the interchange, did concede the validity of some of Landon's results, although rather lamely, if this writer may be allowed to editorialize, since he insisted on noting some "errors" occurring in Landon's work.)

That the normal or Gaussian distribution for fluctuation noise was quickly accepted as valid is shown in a 1943 paper [10] in which the author provides an alternative thermodynamic (statistical mechanical) derivation of the distribution. He also references Einstein's earlier 1910 work in this field. Suffice it to say that the normal distribution of fluctuation noise, first enunciated for engineers by V. D. Landon in 1941, has played and continues to play a very significant role in determining the noise performance and even optimization of a multitude of communication systems.

EARLY 1940S: THE NOISE FIGURE CONCEPT

By the late 1930s, with communication engineers now well aware of the fundamental significance of fluctuation noise in limiting the performance of communication systems, efforts arose to try to reduce its impact. Much of this work, aside from the measurements of "crest factor" referenced in the previous section, dealt with maximizing the ratio of average signal power to average noise power, S/N, or, equivalently, minimizing the average noise power appearing at a communication receiver output. The S/N as a fundamental measure of system performance had already been used by communication engineers in the 1920s [1]. They recognized as well that there were two components to the noise: noise entering the communication system (i.e., the front end of the system) and noise added in the system itself. This idea of reducing internal noise added during the necessary signal processing and transmission through a communication receiver was finally generalized and applied to maximizing the output SNR in the early 1940s by the introduction of the concept of the noise figure, also frequently referred to as *noise factor*.

Most authors in the communications literature attribute the origination of the use of the term noise figure to H. T. Friis of Bell Telephone Laboratory, who introduced this concept in a paper in 1944 [11]. Friis defines the noise figure F of a network to be "the ratio of the available signal-to-noise ratio at the signal-generator terminals [at the network input] to the available signal-to-noise ratio at [the network] output terminals." All the terms listed are defined in terms of their *power*. The expression "available power" refers to the maximum power that can be delivered, measured at either the input to an electrical network or its output, as the case may be. Letting the symbol o refer to the network output and s to the input ("source") we have, using Friis' definition,

$$F = S_s/N_s/S_o/N_o.$$

Here S_s and S_o are the available signal powers at the input and output of the network in question, respectively, with N_s and N_o being the respective noise powers at the same two points. The term S_o/S_s is often called the available signal gain G of the network. We assume here that this "gain" is greater than 1, that is, that the network in question amplifies the signal, rather than attenuating it. The expression above for the noise figure may thus be rewritten as

$$F = N_o/GN_s.$$

Rearranging terms in this expression, it is apparent that the available noise power N_o at the network output is given by the expression

$$N_o = FGN_s, F \geq 1.$$

Here, as noted above, G is the available signal power gain, the ratio of the available signal power at the network output to the available signal power at its output, assumed greater than 1, and N_s is the available noise power at the input to the network. This expression shows that the output noise is due to the amplification of the input noise (the gain factor G), plus an added factor

(Continued on page 24)

[1] *S. O. Rice, of Bell Labs, in a classic 1944–1945 paper on noise [6], provides the only reference to work done by an engineer prior to 1941 on the normal distribution, in noting that Harry Nyquist, in unpublished work done at Bell Telephone Laboratories, had derived the normal distribution from the shot effect in 1932.*

(Continued from page 22)

due to noise introduced within the network itself, as indicated by the noise figure F. A noiseless network, one introducing no additional noise, would have a noise figure $F = 1$. The excess noise introduced by the network is thus given by $N_o - GN_s = (F - 1) GN_s$. The closer the noise figure is to 1, its smallest value, with the gain greater than 1, the better the network. This concept of the noise figure, showing how much noise was added by a network through which a signal plus its accompanying noise were transmitted, enabled rapid evaluation, even minimum noise operation, of networks to be made.

Although Friis is usually given credit for originating and formalizing the term noise figure, a number of other investigators had been developing related concepts at about the same time, and even earlier. In fact, in his paper Friis references earlier work of 1942 by D. O. North [12] involving the sensitivity of networks to noise, introducing a term related to noise figure North had called "noise factor." A vigorous exchange between the two ensued [13], with Friis finally noting that there was no substantial difference between the two approaches.

This exchange between North and Friis indicates that the problem of reducing noise or, equivalently, improving the SNR at the output of communication receivers, first introduced in the 1920s, had taken on increased urgency by this period of the early 1940s. Investigators other than North and Friis had been coping with methods of better characterizing the different sources of noise in receivers in an attempt to improve the receiver SNR. Among these workers were E. W. Herold, a colleague of North at RCA, who, in a companion paper to North's 1942 paper, reported on an "analysis of the effect of various sources of noise on the signal-to-noise ratio of radio receivers…" [14]. D. K. C. MacDonald, then an instructor at a British military school, published a brief paper in 1944 on the noise figure concept [15]. He begins the paper by stating "There are known to the writer two definitions of noise figure (or factor of merit as regards signal to noise performance)," one of which he claims is "more fundamental," but both leading to the same result. The "fundamental" definition of noise figure turns out to be the same as that adopted by Friis, leading, of course, to the same results as Friis. There is, however, no reference at all to Friis's work or publication of the same year! In fact, MacDonald's only reference is to the paper by Herold [14] referenced above. It is thus not clear what MacDonald means by the words "There is known to the author…" cited above. Was this his own definition and subsequent analysis, or was he referring to Friis' and possibly other investigators' work, without explicitly acknowledging them? His reference to Herold's paper meant that he must have been aware of D. O. North's paper, cited above, as well, since it was a companion paper to that of Herold.

Additional papers along similar lines, but appearing in the German technical literature of 1939 to 1942, were cited by a Dutch engineer, M. J. O. Strutt, some years later in a1946 Proceedings of the IRE paper [16]. Strutt includes MacDonald's paper among a number of references on the definition of noise figure, but simply says he, Strutt, will "make use of [the one] proposed" by Friis. He offers no indication as to the similarity of the two measures of noise figure. He does indicate, however, that the noise figure measure introduced by Friis is closely related to a measure proposed and used earlier by two German engineers, K. Franz and W. Kleen, whose papers are referenced by Strutt, as noted above. In particular, one paper by Franz is cited by Strutt as having appeared in 1939, well before the papers of Friis and North.

Some comments are in order here. We note, first, that the period under discussion here, 1939–1945, was the period of World War II. Communication among engineers of different nations, even between allies such as the United States and Britain, was difficult, if at all possible. Engineers of the allied nations had presumably little knowledge of the engineering work and literature at the time in Germany, and vice versa. This may account for the lack of cross-referencing noted above. A second point to be noted is that almost all of the papers cited on noise figure and optimization of SNR referred specifically to operation at short-wave radio frequencies and above. This movement into the higher frequencies was influenced heavily by the advent of World War II and a specific focus on radar systems operating at much higher frequencies than had hitherto been the case. This point is emphasized in a brief paper by Matthew T. Lebenbaum appearing years later on the history of receiver noise measurement [17]. In discussing the history of the concept of noise figure/noise factor, Lebenbaum makes the point "But it was not until 1942 that there was sufficient usage of the 'short-wave' bands that it became imperative that some standardization of terms be accomplished, and that we should decide what it was we were trying to measure." He notes that this movement into the short-wave bands was influenced heavily by World War II, with the need to develop high-frequency radars capable of searching for, and detecting, enemy aircraft playing the key role in this move to higher frequencies. (Transmission wavelengths at the higher frequencies become much smaller, with focusing antennas many wavelengths in diameter thereby becoming practical and economical to design, build, and deploy.) Radio transmission engineers at this period of time quickly realized the need to control receiver "absolute sensitivity," this quantity at the ultra-high frequency (UHF) band and microwave band, still higher in frequency, being principally determined by the first receiver stages rather than by man-made noise or "atmospherics." (Radio engineers in the 1920s, dealing with frequencies below that of short wave, tended to be either unaware of, or not concerned with, fluctuation noise, i.e., thermal and shot noise [1]. They were concerned principally with atmospherics or static due to the Sun and other natural processes.) In particular, radar engineers, designing radars to operate at a fixed frequency, and particularly interested in detecting aircraft targets as far away as possible, were very concerned about being able to accurately detect very small signals in the presence of noise. It was therefore imperative to reduce internally generated receiver noise as much as possible. The concept of noise figure and its use in maximizing the receiver output SNR, which determined the detectability of a particular radar target, thus played significant roles in radar receiver design.

That a focus on noise and its influence on radar system performance played a key role in radar design and development during World War II is documented by papers and books published by participants in radar activities during that time. In particular, a number of volumes in the Radiation Laboratory series, a series of books published at the end of the war documenting the work on radar development at the MIT Radiation Laboratory during World War II, provide a vivid record of the specific work on radar receiver noise reduction carried out by engineers and

(Continued on page 26)

(Continued from page 24)

physicists working there during that period of time. These books incorporate references to the papers on noise figure noted above, as well as describing work on that subject and related topics carried out at the Radiation Laboratory itself. These volumes collectively document related work carried on as well in many other organizations, including, in the United States, Bell Labs, RCA (both organizations already mentioned), Sperry Gyroscope Company, and the Harvard Radio Research Laboratory.

This work on noise reduction techniques carried out at radar research and development laboratories worldwide was only a portion of the work done on signal and noise differentiation procedures. The urgent need to improve the detection of the pulsed radar signals in the presence of noise, and the recognition that this detection process was inherently statistical in nature, led radar engineers to incorporate and extend statistical optimization techniques well known to statisticians in the design of the radar systems. These new signal detection procedures in turn fed back on the statistical literature, leading to very significant advances in the field of statistics as well. This work also led to advances in the understanding of the noise processes underlying the problem of detecting signals in noise. The classic paper by S. O. Rice on noise processes referenced earlier [6] incorporates both the new work on noise processes and the older work carried out years before. Even more important, the work on radar led to the development, and subsequent flowering after the war, of a completely new field of communications called *statistical communication theory*. It played an important role in the development of improved communication systems both during the war and shortly afterward. Textbooks began to be written shortly after the war based on these new statistical approaches to the design of communication systems. The field of telecommunications, in general, was literally revolutionized by these developments made during World War II.

We do not go beyond the period of the early 1940s in this column. We leave out related work carried out in the design of control systems during the period of World War II (see, e.g., [18]). We leave out as well significant work in the Soviet Union and elsewhere carried out during the period of World War II on optimum signal detection in the presence of noise (see, e.g., [19]). Instead, we conclude this column by returning, in the next and final section, to a concept first noted by John Carson in his early work of the 1920s on improving the noise performance of communication systems. This is the concept of the *matched filter*, discovered independently, as we shall see, by a number of investigators in different radar laboratories during World War II.

1940S: THE MATCHED FILTER CONCEPT

It was noted in the introduction to this column that John Carson of Bell Labs had, in the 1920s, come up with the concept of "matched filtering." It was not until the period of World War II, roughly 20 years later, with an emphasis on the development of radar systems and a consequent imperative need to detect small-amplitude pulse signals in the presence of noise, as noted in the previous section, that matched filter design was shown to be an optimum way of detecting signals in the presence of noise. Carson was thus quite prescient for his time.

The term "matched filter" refers to the fact that the optimum receiver filter frequency characteristic, in the sense of maximizing the SNR at the receiver output, should be one that is "matched" to the frequency spectrum of the signal pulse to be detected. What this basically means is that the filter characteristic should be emphasized in frequency ranges where the pulse signal energy is high compared to the noise energy, and de-emphasized in ranges where the noise energy dominates. The most common case occurs where the noise energy is equally distributed over all frequencies. Noise of this type is referred to as "white noise." In this case, with the signal pulse frequency spectrum written as $S(\omega)$, with ω the frequency in units of radians per second, the best filter characteristic $H(\omega)$ for maximizing SNR is found to be given by

$$H(\omega) = KS^*(\omega),$$

the * meaning complex conjugate and K a constant. The filter is then said to be "matched" to the signal pulse frequency characteristic. Where the noise spectrum is not white, a term involving the square root of the noise spectrum appears in the denominator of this equation, further enhancing the filter characteristic at frequencies where the noise spectrum is small, or de-emphasizing the filter effect where the noise spectrum is large. (Signal and filter frequency characteristics are commonly written in terms of complex numbers. If the characteristics may be written as real numbers, the * symbol disappears.) A practical approximation to this optimum filter characteristic turns out to be a filter whose bandwidth is approximately the reciprocal of the pulse width. Thus, letting B be the filter bandwidth in Hertz, the filter bandwidth that maximizes the SNR is simply

$$B = 1/T$$

with T the width of the pulse. As an example, if the signal pulse width is 1 μs, the optimum bandwidth is 1 MHz. It turns out that the critical design of the filter is related to incorporating this bandwidth requirement. The exact form of the "matching" filter is of secondary importance in maximizing the receiver SNR.

A corollary to this result shows that the "detectability" of a pulsed signal, defined as the ability, with some probability of being correct, to detect the signal in the presence of noise, assuming the use of a matched filter, depends solely on the energy of the signal: this is independent of the specific pulse shape, be it narrow with a high peak value or wide with a smaller peak value. (This result is so commonly known in modern communication systems that it has, for years, been textbook material.) George Turin, in a comprehensive 1960 paper, summarized the properties of the matched filter and its applications in many areas of communication theory post-World War I [20]. In this section we focus on the early development of the matched filter concept during World War II, as its applicability arose in the development of radar signal detection in the presence of noise. (Turin does include a reference to two of the papers and reports cited below.) As is often the case in science and engineering, a number of investigators working on radar in a variety of government laboratories and development organizations during this period came up with these results independently and at about the same time. (Recall, from the previous section, a similar situation occurring with the introduction of the noise figure concept.)

We begin with Andrew V. Haeff of the Naval Research Laboratory (NRL) in Washington, DC, who, while carrying out experimental studies with human observers of radar signal detection beginning in early 1942, found that a signal pulse in the presence of noise was detectable if the inverse bandwidth-pulse width condition $B = 1/T$ noted

(Continued on page 28)

(Continued from page 26)

above was met, with the detectability then depending solely on the signal energy, also as noted above [21].[2] This work was followed by, and utilized in, the work of Kenneth A. Norton of the U.S. National Bureau of Standards (NBS) and Arthur C. Omberg of Bendix Corporation who published a classified report in February 1943 [22] on a study carried out of the maximum range of a radar system. (This is the same Norton who was involved with the exchange over the use of the Gaussian distribution for noise noted earlier.) They found the maximum range depended on a quantity called the *visibility factor*, defined as the ratio of minimum signal pulse energy to the receiver fluctuation noise energy, as referred to the input circuit of the receiver, required to detect a signal, minimum signal referring to a "barely visible pulse." The visibility factor is thus essentially the SNR we have introduced a number of times, except that it refers specifically to an *energy* rather than *power* ratio. They then went on to show, using an empirical formula for the minimum signal peak power derived from the results of Haeff's work, that in maximizing the visibility factor, there exists an optimum receiver bandwidth, exactly the term quantity $1/T$, T the signal pulse width, as indicated above. Investigators at the MIT Radiation Laboratory doing extensive work on radar signal detection in the presence of noise also recognized the inverse time-bandwidth relationship required to get optimum signal detection in noise [23].[3]

These various groups of investigators came up with the inverse bandwidth-signal pulse-width relationship required for optimum signal pulse detectability. It appears to be D. O. North of RCA, however, whom we encountered in the previous section on noise figure, who was among the first to demonstrate the full matched filter result expressed in the equation above, as part of a theoretical analysis of pulsed signal detection in noise. His work, appearing in a 1943 classified RCA report and reprinted 20 years later in the *Proceedings of the IRE* [24], was done at about the same time as the work just cited.[4] In addition to his analysis leading to matched filtering, North, in his paper, includes such topics as signal detection improvement through appropriate summation (integration) of successive signal pulses received in noise, the optimum detector law for radio signals, pre- vs. post-detector integration, and a discussion of "envelope detection" in pulse-carrier systems. Under this last item, he derives the now well-known Rician distribution for the envelope of a pulsed signal plus Gaussian noise implicitly noted in the exchange over the Gaussian distribution for noise described earlier in the section on the statistical representation of noise and referenced in [9].

To quote the introduction to the North report and paper, the "Object of [the paper] is to formulate the smallest signal discernible through background noise in terms of the pulse energy, the receiver design, and the choice of integrating and indicating means." After deriving the matched filter result in a section of his paper, he goes on to show that for a practical filter the optimum bandwidth, again in the sense of maximizing the SNR, is close to $1/T$, just the bandwidth result we quoted above. (It is to be noted that a postscript to the paper references Haeff's NRL report [21], "just received [which provides] close functional agreement" with the theoretical analysis of the paper.)

In addition to North, J. H. Van Vleck and David Middleton, working at the Harvard Radio Radiation Laboratory, independently proved at about the same time as North that the matched filter maximized an SNR [25]. Their technique of proof was somewhat different than that of North, their choice of SNR being somewhat different as well: They were dealing with the visual detection of pulsed signals, those appearing on a radar screen. Implicit in this detection was the effective "integration" or summing of multiple received signals due to successive repetitions of a pulsed transmitted signal common in the radar systems whose performance in noise they were engaged in studying. Their choice of SNR, based on a model of the human visual detection process, used as the signal term in the numerator the difference ("excess") between the average amplitude on the screen at the time at which the signal plus noise had its largest value and the average amplitude in the absence of signal (i.e., due to noise only). The noise term in the denominator was the rms value of the noise alone. As an example, they show that for a Gaussian (bell)-shaped pulse, the matched filter frequency characteristic is similarly Gaussian-shaped. The product BT of pulse width times bandwidth is then precisely 1, the expression noted above. In this case B and T are each defined as the rms deviation about the corresponding peak value.

CONCLUSION

We have briefly described in this column the development of an understanding of the impact of noise on communication system performance during the 1930s and early 1940s. This followed on earlier work in the 1920s [1]. We described the invention of wideband FM and PCM as being quite pivotal in showing how the proper design of communication systems could lead to a considerable decrease in the impact of noise. In these two cases engineers became aware of the concept that purposely increasing bandwidth could, with certain designs, result in a considerable reduction of noise.

We follow the discussion of PCM with a brief description of the realization by engineers that noise statistics obeyed the well-known Gaussian or bell-shaped distribution. This was then followed by a discussion of the increasing need to establish a definition for the effect of noise on systems, leading to the introduction of the concept of the noise figure. This concept immediately focused attention on the first stages of communication systems and the need to reduce the noise introduced there as much as possible. It served to quantitatively allow the impact of noise introduced in stages following to be determined. We conclude this history by summarizing the work in the early 1940s by a number of investigators coming up with the concept of a matched filter to provide the "best" detection of a pulsed signal in noise. This work arose out of the exigencies of developing radar detection schemes during World War II.

We have left out much important work done by various laboratories worldwide on other aspects of the understanding of the properties of noise and its impact on communication sys-

(Continued on page 30)

[2] In the 1946 declassified journal version of Haeff's paper, the noise energy is written in terms of the noise figure, with reference made to some of the papers discussed here in the section on noise figure.

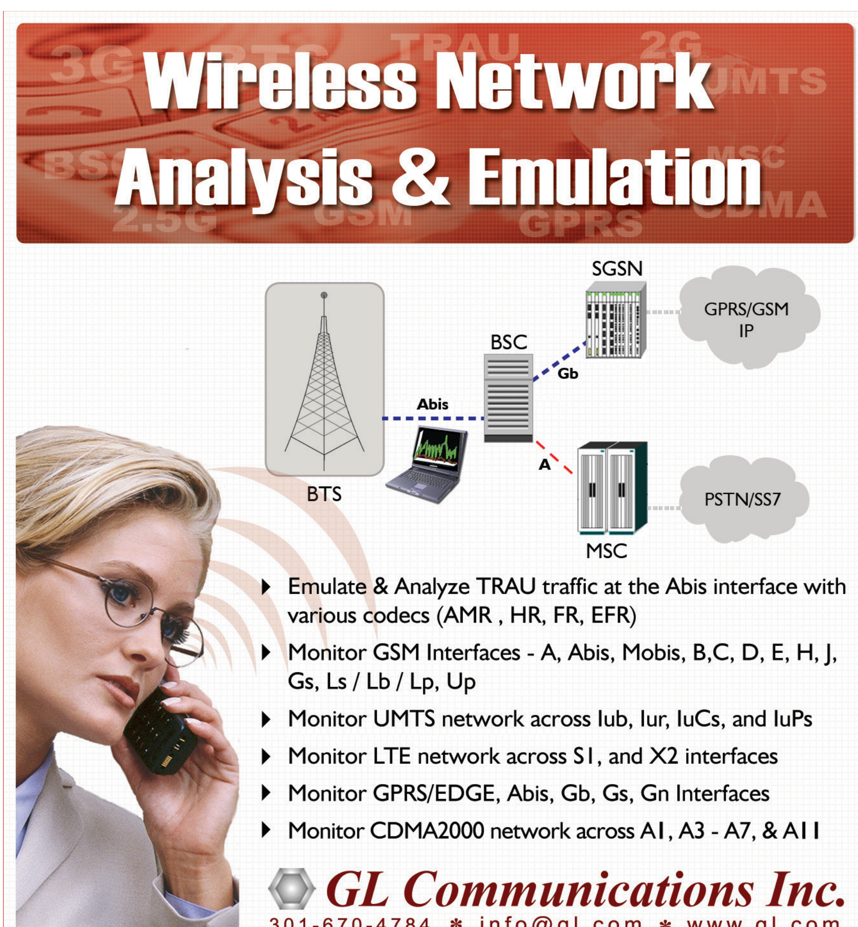
[3] Section 8-6 of [23] describes a number of experiments carried out at the MIT Radiation Laboratory that showed the best SNR was obtained with $BT = 1$*, but no indication is given as to the year in which these were done.*

[4] Norbert Wiener had, in his 1942 report cited earlier [18], obtained the matched filter result for the special case of small SNR. He did not explicitly use the phrase "matched filter," simply stating the "best" filter as one that has the signal characteristic itself.

(Continued from page 28)

tems. Most important, we have not traced, in this brief presentation, the effect of system nonlinearities on the reception of communication signals in noise. Some of this work during the 1930s is cited in the work by S. O. Rice referenced earlier [6], as well as in many books on communication theory written in the post-World War II period.

As is often the case in both science and engineering, progress can be hesitant and sometimes, seen in retrospect, painfully slow. We note that all through the 1920s, radio engineers paid little attention to fluctuation noise and focused on static or atmospherics [1]. During the same period the validity of the concept of noise spectral density (variation of noise power with frequency) was at first challenged by some workers in the field [1]. In the period under discussion in this article, it took time to conclude that noise statistics did not have a so-called crest factor and did obey the Gaussian distribution. The noise figure concept was, in retrospect, late in being incorporated in communication systems design. But there *was* progress in the understanding of noise and its quantitative effect on communications systems, and clearly, the development of radar systems during World War II expedited the understanding of the impact of noise on communication systems immeasurably.



REFERENCES

[1] M. Schwartz, "Improving the Noise Performance of Communication Systems: Radio and Telephony Developments of the 1920s," *IEEE Commun. Mag.*, vol. 47, no. 12, Dec. 2009, pp. 16–20.

[2] M. Schwartz, "Armstrong's Invention of Noise-Suppressing FM," *IEEE Commun. Mag.*, vol. 47, no. 4, Apr. 2009, pp. 20–22.

[3] A. H. Reeves, "The Past, Present, and Future of Pulse-Code Modulation," unpublished paper, prepared for IEE.

[4] K. W. Cattermole, *Principles of Pulse Code Modulation*, London: Iliffe Books; New York: Elsevier, 1969.

[5] V. D. Landon, "The Distribution of Amplitude with Time in Fluctuation Noise," *Proc. IRE*, vol., 29, no. 2, Feb. 1941, pp. 50–55.

[6] S. O. Rice, "Mathematical Analysis of Random Noise," *Bell Sys. Tech. J.*, vol. 23, July 1944, pp. 283–332; vol. 24, Jan. 1945, pp. 46–156 ,See footnote, section 2.8, and footnote 9, section 1.9. This paper in two parts by Rice has become a classic in the literature of noise and its impact on communication systems. Both parts are reprinted in the book *Selected Papers on Noise and Stochastic Processes*, N. Wax, Ed., Dover, 1954.

[7] V. D. Landon, "A Study of the Characteristics of Noise," *Proc. IRE*, vol. 24, no. 11, Nov. 1936, pp. 1514–21.

[8] K. G. Jansky, "An Experimental Investigation of the Characteristics of Certain Types of Noise," *Proc. IRE*, vol. 27, no. 12, Dec.1939, pp. 763–68.

[9] K.A. Norton and V. D. Landon, Discussion on "The Distribution of Amplitude with Time in Fluctuation Noise," *Proc. IRE*, vol. 30, no. 9, Sept. 1942, pp. 425–29.

[10] M. Surdin, "Distribution in Time of Spontaneous Fluctuation Noise," *Philosophical Mag.*, no. 34, Oct. 1943, pp. 716–22.

[11] H. T. Friis, "Noise Figures of Radio Receivers," *Proc. IRE*, vol. 32, no. 7, July 1944, pp. 419–22.

[12] D. O. North, "The Absolute Sensitivity of Radio Receivers," *RCA Review*, vol. 6, no. 3, Jan. 1942, pp. 332–43.

[13] D. O. North and H. T. Friis, Discussion on "Noise Figures of Radio Receivers," *Proc. IRE*, vol. 33, no.2, Feb. 1945, pp. 125–27.

[14] E. W. Herold, "An Analysis of the Signal-to-Noise Ratio of Ultra-High Frequency Receivers," *RCA Review*, vol. 6, no. 3, Jan. 1942, pp. 302–31.

[15] D. K. C. MacDonald, "A Note on Two Definitions of Noise Figure in Radio Receivers," *Philosophical Mag.*, vol. 35, 194, pp. 386–95.

[16] M. J. O. Strutt, "Noise Figure Reduction in Mixer Stages," *Proc. IRE*, vol. 34, no. 12, Dec. 1946, pp. 942–47.

[17] M. T. Lebenbaum, "Some Perspectives on Receiver Noise Measurement and Its Performance," *Microwave J.*, vol. 16, Jan. 1973, pp. 31–32, 34.

[18] N. Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series," MIT Cambridge, Report of the Services 19, Research Project DIC-6037, Feb. 1942; reprinted by Wiley, 1949.

[19] V.A. Kotelnikov, "The Theory of Optimum Noise Immunity," translated by R. A. Silverman, McGraw-Hill, 1960. (Originally prepared as Kotelnikov's 1947 doctoral dissertation.)

[20] G. L. Turin, "An Introduction to Matched Filters," *IRE Trans. Info. Theory*, vol. 7, no. 4, June 1960, pp. 311–29.

[21] A. V. Haeff, "Minimum Detectable Radar Signal and its Dependence upon Parameters of Radar Systems," *Proc. IRE*, vol. 34, no. 11, Nov. 1946, pp. 857–61, (Reprint of a confidential report first issued 1943.)

[22] K. A. Norton and A. C. Omberg, "The Maximum Range of a Radar Set," *Proc. IRE*, vol. 35, no. 1, Jan. 1947, pp. 4–24, (Originally published as a classified report , ORG P-9-1, Operations Research Group Report of Office of Chief Signal Officer, U.S. War Dept., Feb. 1943.)

[23] J. L. Lawson and G. E. Uhlenbeck, Eds., *Threshold Signals*, vol. 24, MIT Rad. Lab. Series, McGraw-Hill, NY, 1950.

[24] D. O. North, "An Analysis of the Factors which Determine Signal/Noise Discrimination in Pulsed Carrier Systems," RCA Labs, Princeton, NJ. Tech. rep. PTR-6C, June 25, 1943, teprinted, after declassification, *Proc. IRE*, vol. 51, no. 7, July 1963, pp. 1016–27.

[25] J. H. Van Vleck and D. Middleton, "A Theoretical Comparison of the Visual, Aural, and Meter Reception of Pulsed Signals in the Presence of Noise," *J. Appl. Physics*, vol. 17, Nov. 1946, pp. 940–71 (based on Harvard RRL Report 411-86, May 1944).