Improving the Noise Performance of Communication Systems: 1930s to early 1940s

Mischa Schwartz, Charles Batchelor Professor Emeritus of Electrical Engineering
Columbia University, New York, NY 10027 USA

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 paper 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 three inter-related, and roughly chronological, areas:

1. Work by Armstrong on FM and Reeves on PCM showing, for the first time, that noise could be reduced by purposefully increasing the bandwidth (now known as the noise-bandwidth tradeoff). This work is discussed in the next section covering the period of the 1930s.
2. Studies attempting to understand the statistical properties of noise, leading to its now-well-known Gaussian amplitude characteristic. This work is described in the section covering the late 1930s to early 1940s.
3. The recognition that "matched filtering" provided optimum signal detection in noise. This work, described in the last section of this paper, arose out of the need during World War II to detect small, pulsed, radar signals in the presence of noise.

1930s: Noise–bandwidth tradeoff; Armstrong/FM and Reeves/PCM

In this section, we focus, first, on Armstrong's invention of wide-deviation (wideband) FM in the early part of that period and then, afterwards, on Reeves's independent invention of (wideband) PCM a few years later. These inventions changed the study of the impact of noise on system performance dramatically. 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 increased bandwidth off for improved signal-to-noise ratio. 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.

Edwin H. Armstrong had been experimenting for years in his laboratory at Columbia University with various methods of reducing static in radio reception [2], [3]. By 1927 he thought a noise cancellation technique might work. This technique was promptly (and properly) critiqued by John Carson of Bell Labs. By 1931 he had, however, come to the now well-known result that wide-deviation FM, in which the FM bandwidth is purposely widened, would reduce the noise. How he came up with this revolutionary idea is difficult to answer and may, in fact, never be answered exactly, since Armstrong never kept notes of his conceptions and experiments based on them. One can make a good surmise, however, based on a study of what papers are available, that he came to the conclusion that wide deviation FM would reduce noise in about September of 1931 [4]. He applied for a US patent on his wideband FM invention January 24, 1933, some 16 months later, and the patent, with the simple title *Radiosignaling*, was granted December. 26, 1933 with the number 1,941,069. His '069 wideband FM patent was one of four FM patents granted to Armstrong the same day. It was the only one dealing with noise suppression, however, The others dealt with such issues as improved means of generating FM (now called the Armstrong system), the use of a limiter in FM, and FM as a way to reduce signal fading.

The wideband FM patent is quite specific on the noise-suppression property of wide-deviation FM, indicating that this is the essence of the invention. To quote the patent, "I have discovered that by imparting greater swing to the frequency of the transmitted wave than can exist in the disturbances due to tube irregularities and providing means for selecting these large swings of frequency which are … not responsive to the lesser swings due to the tube disturbances …that a very great improvement in transmission can be produced." (It is to be noted that Armstrong was well-aware of the properties of tube noise, to which he referred specifically later in the patent. In the "jargon" of the day, tube noise referred to both shot noise and thermal noise.)

Armstrong's invention of wide-deviation/wideband FM and its ability to suppress noise was well-received by the engineering community. Very soon after being awarded patent '069, Armstrong demonstrated his wide-deviation FM system for the first time to RCA engineers [5] with impressive results. Armstrong's later demonstration of his system before an Institute of Radio Engineers (IRE) audience on November 5, 1935, and its very positive impact on the engineers attending, has been well-documented [2]. The demonstration accompanied a formal paper presented at the meeting, which was later published as Armstrong's now-classic paper "A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation"[6]. What is particularly remarkable about this paper is that Armstrong was able to analytically demonstrate the noise-reduction effect of wide-deviation FM, given a carrier-to-noise threshold had been exceeded, using a vector approach. Soon thereafter, various investigators, using a variety of mathematical approaches proved Armstrong's result. (Such proofs of the FM noise-suppression property are still going on to this day. The reception of FM signals in the presence of noise is a highly non-linear process, involving, for example, the required use of a limiter. FM analysis thus does not lend itself to a tidy mathematical approach,)

We now turn to Alec Reeves' invention of Pulse Code Modulation (PCM) as a means of reducing noise in digital communication systems. 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 tradeoff noted above, just as was the case with wide-deviation FM, between improving the signal-to-noise ratio and increasing the bandwidth.

Reeves filed three very similar patent applications on his invention: one in France, granted Oct. 3, 1938 as patent 852, 185; another in Britain, filed and granted in 1939; and a third, filed in the United States Nov. 22, 1939, and granted Feb. 3, 1942, as US patent 2,272, 070. The US 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 noi*se. 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 tradeoff, 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 under-estimate, 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 recognizes 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 [7], 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 [US 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 issues 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.

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, such as the FM and PCM inventions described in the previous section, focused on maximizing the signal-to-noise ratio, 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 fluctuation noise has a Gaussian or normal probability amplitude distribution. Yet it wasn't 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 [8]. 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 [sic] peak value" of the noise to its root-mean-squared (rms) value, coming up with different measured values of this quantity [11], [12]. 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 [11]. It wasn't until 5 years later, in 1941, as noted above, that he finally came up with the Gaussian characterization of noise [8].

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 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 [13], 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 [14] 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.

1940s: Matched Filter Concept

We conclude this paper by describing work carried out at a number of laboratories developing radar systems during World War II that showed that signal performance in the presence of noise could be optimized by using what became known as a "matched filter". Interestingly, the concept of a matched filter appears to have first been anticipated by John Carson of Bell Labs in the 1920s [1]. It wasn't 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 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 signal-to-noise ratio (S/N) 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. A practical approximation to this optimum filter characteristic turns out to be a filter whose bandwidth B is approximately the reciprocal of the pulse width T, i.e., $B=1/T$.

As is often the case in science and engineering, a number of investigators working on radar came up with these results independently and at about the same time. Thus, Andrew V. Haeff of the Naval Research Laboratory (NRL) in Washington D.C., 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 above was met, with the detectability then depending solely on the signal energy [15]. This work was followed by, and utilized in, the work of Kenneth A. Norton of the US National Bureau of Standards (NBS) and Arthur C. Omberg of Bendix Corporation who published a classified report in February 1943 [16] on a study carried out of the maximum range of a radar system. 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 signal-to-noise ratio (S/N) 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.

These two 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, who was among the first to demonstrate the full matched filter result including the optimum filter characteristic in addition to the inverse bandwidth-time relationship, as part of a theoretical analysis of pulsed signal detection in noise [17]. [2]

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 post-script to the paper references Haeff's NRL report, [15], "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 RRL, independently proved at about the same time as North that the matched filter maximized a signal-to-noise ratio (S/N) [19]. Their technique of proof was somewhat different than that of North; their choice of S/N, based on a model of the human visual detection process, being somewhat different as well: They too came up with the inverse bandwidth-time result noted above.

Note

[1] S. O. Rice, of Bell Labs, in a classic 1944-45 paper on noise [9], provides the only reference this author has been able to find to work done by an engineer prior to 1941 on the normal distribution for noise. He notes that Harry Nyquist, in unpublished work done at Bell Telephone Laboratories, had derived the normal distribution from the shot effect in 1932 [10].

[2] Norbert Wiener had, in a 1942 report [18], obtained the matched filter result for the special case of small signal-to-noise ratio. He didn't explicitly use the phrase "matched filter", simply stating the "best" filter is one which has the signal characteristic itself.

References

[1] M. Schwartz, Improving the Noise Performance of Communication Systems: 1920s to early 1930s, presented, Symposium on the History of Communication Technologies, Smithsonian National Postal Museum, Washington, DC, Oct. 17, 2007.

[2] Lawrence Lessing, Man of High Fidelity, J.B. Lippincott Co., Philadelphia., 1956; Bantam Paperback, 1969.

[3] Report from Edwin H. Armstrong to Michael Pupin, April 3, 1933, Box 115, Armstrong papers, Rare Book Library, Columbia University.
[4] M. Schwartz, Armstrong's Invention of Noise-Suppressing FM, published in *Antenna*, Newsletter of the Mercurians, Special Interest Group, Society for the History of Technology, Vol. 20, no.1, Oct. 2007, 2-6.
[5] Gary Lewis Frost, The Evolution of Frequency Modulation Radio, 1902-1940, Ph. D. dissertation, University of North Carolina, Chapel Hill, 2004.
[6] E. H. Armstrong, A Method of Reducing Disturbances in Radio Signaling by a System of Frequency Modulation, Proc. IRE, vol. 24, no. 5, May 1936, 689-740.
[7] Alex H. Reeves, The Past, Present, and Future of Pulse-Code Modulation, unpublished paper, 1964, prepared for IEE. .
[8] V. D. Landon, The Distribution of Amplitude with Time in Fluctuation Noise, Proc. IRE, vol.29, no. 2, Feb. 1941, 50-55.
[9] S. O. Rice, Mathematical Analysis of Random Noise, Bell System Technical Journal, vol. 23, July 1944, 283-332; vol. 24, Jan. 1945, 46-156.
[10] H. Nyquist, Fluctuations in Vacuum Tube Noise and the Like, unpublished memorandum, Bell Laboratories, March 7, 1932. Cited in reference [9]: footnote, section 2.8, and footnote 9, section 1.9,
[11] V. D. Landon, A Study of the Characteristics of Noise, Proc. IRE, vol. 24, no. 11, Nov. 1936, 1514-1521.
[12] Karl G. Jansky, An Experimental Investigation of the Characteristics of Certain Types of Noise, Proc. IRE, vol. 27, no. 12, Dec.1939, 763-768.
[13] 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, 425-429.
[14] M. Surdin, Distribution in Time of Spontaneous Fluctuation Noise, Philosophical Mag., no. 34, Oct. 1943, 716-722.
[15] Andrew V. Haeff, Minimum Detectable Radar Signal and its Dependence upon Parameters of Radar Systems, Proc. IRE, vol. 34, no. 11, Nov. 1946, 857-861. (Reprint of a confidential report first issued 1943.)
[16] Kenneth A. Norton and Arthur C. Omberg, The Maximum Range of a Radar Set, Proc. IRE, vol. 35, no. 1, Jan. 1947, 4-24. (Originally published as a classified report , ORG P-9-1, Operations Research Group Report of Office of Chief Signal Officer, US War Dept., Feb. 1943.)
[17] D. O. North, An Analysis of the Factors which Determine Signal/Noise Discrimination in Pulsed Carrier Systems, RCA Labs., Princeton, N. J. Technical Report PTR-6C, June 25, 1943. Reprinted, after declassification, Proc. IRE, vol. 51, no. 7, July 1963, 1016-1027.
[18] Norbert 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 John Wiley and Sons, NY, 1949.
[19] J. H. Van Vleck and David 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, 940-971. (Based on Harvard RRL Report 411-86, May 1944.)