The Radio Receiver Saga: An Introduction to the Classic Paper by Edwin H. Armstrong

NICOLAE COTANIS, MEMBER, IEEE

Invited Paper

I. PROLOGUE

Is it just a coincidence? Many of us remember when astronaut Neil Armstrong became the first person to set foot on the Moon in 1969. All over the Earth people heard his words: “That’s one small step for man, one giant leap for mankind.” Radio stations sometimes delight us with the warm and pure trumpet sounds of Louis Armstrong. But neither of these would be possible without the work and inspiration of many contributors, among whom we must include Edwin H. Armstrong, the namesake of the above, and the inventor of the superheterodyne receiver and FM transmission.

II. THE RADIO COMMUNICATIONS DAWN

The invention of the radio is a long story. It starts with the theoretical work of J. C. Maxwell, who brought together electricity, magnetism, and light in one set of relations, preparing the way for wireless communications. Later, H. Hertz experimentally demonstrated the existence of electromagnetic waves. In 1890, E. Branly originated the coherer, a receiving device for electromagnetic waves based on the principle that, while most powdered metals are poor dc conductors, metallic powder becomes conductive when HF current is applied. The device was improved by A. S. Popoff and G. Marconi. For about ten years, the coherer was the receiving detector used in Marconi radio stations.

In 1895, Marconi achieved radio communications over more than one mile by connecting to ground one lead of the transmitter coil while the other lead was connected to a vertical metallic surface, the ancestor of the aerial. At about the same time, Popoff presented his wireless telegraph receiver, which allowed communication over 250 m. It is worthy of being remembered that, after experiments made by Marconi and Popoff, Branly did not claim to be the inventor of telegraphy, because he never thought of such an application. Thus, until 1896 all of the efforts were focused on building experiments that encouraged wireless telegraphy. To increase the communications range, further work was necessary to design new aerials, improve the transmitter oscillators, and increase sensitivity and selectivity of the receiver.

The mechanically unstable coherer gave way to the crystal detector using the semiconductor silicon or galena with a cat’s-whisker contact, as invented by Pickard and Dunwoody. At about the same time as the crystal detector, there appeared the first vacuum diode detector. But this is another story [1], one which started back in 1883 when Edison discovered current flow inside a vacuum tube.

Using the Edison thermionic effect as a basis, in 1904 Prof. A. Fleming patented the vacuum diode, the “valve.” In 1907 Lee de Forest added the grid control, resulting in the triode vacuum tube, the “audion.” The triode remained a rather unsatisfactory device until 1912, however, when H. D. Arnold showed the necessity of a very high vacuum, which was made possible by W. Gaede’s molecular pump in 1913. Wehnelt invented the oxide-coated filament, replacing the tantalum one. These early triodes made practical the cascade amplifier, the triode oscillator, and the regenerative feedback circuit.

This was the status of the radio developments in 1915, when Armstrong’s paper was published in the PROCEEDINGS OF THE IRE.

III. COMMENTS ON THE ARMSTRONG PAPER

Armstrong’s paper [2] takes us back in the pioneering age of electronics. It exploits the fundamental operating characteristic of the audion, the relationship between the anode (wing) current, and the potential of the grid with respect to the cathode.

The audion did not have an indirectly heated cathode, the potential of the negative terminal of the filament being the potential of the cathode. Having a very poor vacuum,
the first audions were highly nonuniform. As late as 1915, De Forest claimed “... the behavior of different bulbs varies in many particulars, and to an astonishing degree,” but Armstrong had foreseen the forthcoming triodes: “With ... high vacuum bulbs, the astonishing differences ... disappear to an astonishing extent.”

Starting from this point, Armstrong was involved in the development of “circuits of a new type to the actuation of the audion.” The understanding of the circuits operation is clarified through the use of oscillograms. These helped Armstrong to imagine the regenerative receiver. Shades of the superheterodyne receiver and other circuits like blocking oscillators may be discovered in the paper.

A. A Closer Look to the Audion Detector

The circuit for using the audion as a detector was already known (Fig. 3 of Armstrong’s paper). The principle of regenerative reception emerged from a simple observation, while the audion was acting as a rectifier “... oscillations in the grid circuits set up oscillations of a similar character in the wing circuit.” The next step was quite obvious “providing means ... for utilizing them (the anode oscillations) to reinforce the oscillations of the grid circuit.”

B. Armstrong’s Regenerative Detector

Two methods of reinforcement are proposed. In the first circuit (Fig. 8 of Armstrong’s paper) inductively coupled coils are used, L2 in the grid and L3 in the anode circuit. Some of the energy of the anode oscillations is transferred back to the grid circuit by means of the coupling M2. The second circuit (Fig. 9 of Armstrong’s paper) uses a tuned anode circuit, the energy of the anode oscillations being transferred back to the grid by means of “the condenser formed by the wing and the grid.” A general relationship between signal strength and value of the anode inductance is shown in Fig. 10 of Armstrong’s paper. The telephone current could be amplified further in exactly the same manner as the RF oscillations. The circuit presented in Fig. 14 uses a single triode and provides an amplification of about 100 times for a weak signal.

C. From Beat Receiver to Superheterodyne Receiver

It is obvious that the triode oscillator was already known and it is beyond the scope of this introductory paper to address the controversy of who is the inventor—Armstrong or de Forest [3]?

In his paper Armstrong gives a nice explanation of the mechanism that controls the amplitude of oscillation. “... Grid oscillations are continuously rectified to charge the grid condenser. ... As the negative charge builds up ... it decreases the continuous anode current and therefore limits the amplitude of oscillations. ...”

In the early days of radio communications, long-wave transoceanic telegraph signals were transmitted by very high-powered alternators in the range 10 to 20 kHz. Armstrong presents a high gain beat receiver for continuous wave (undamped oscillations). Again, Armstrong’s clarity may be appreciated: “The radio frequency beats are rectified by the audion to charge the grid and the grid condenser, and this charge varies the electron current to produce an amplifying action on the current in the telephones.”

Another improvement presented in his article is the increase in the beat receiver selectivity by tuning the telephone circuit (Fig. 21 of Armstrong’s paper). In this way, Armstrong sets the stage for the superheterodyne receiver.

D. Blocking Oscillators

Armstrong’s work on oscillators leads him haphazardly to the blocking oscillator. By chance, the telephone is still connected in the anode circuit. Increasing the coupling between the grid and anode above the point at which oscillation started “a rough note is heard in the telephone the pitch decreasing with increase of coupling.” Oscillograms helps him to see “the breaking up of the oscillations into groups.” For Armstrong, the explanation of the phenomenon is at hand. It occurs when the charge is supplied to the grid condenser “at a greater rate than that at which it can leak off. ... The result is that the grid is periodically charged to a negative potential sufficient to cut off entirely the wing current, causing a stoppage (blockage) of the local oscillations until the grid charge leaks off and the wing current re-establishes itself.”

Going ahead with his experiments, Armstrong discovered a one-audion oscillator producing two slightly different frequencies (Fig. 17 of the Armstrong paper).

E. The Fight Against Atmospheric Disturbances

For years, Armstrong had been seeking a way to reduce “static,” noise produced by severe atmospheric disturbances. Finally, he turned his attention to FM. In the wide-band FM system perfected by Armstrong in 1933 audio components up to 15 kHz were producing a frequency deviation of 75 kHz [4].

In his 1915 paper, he proposes a balanced wave heterodyne receiver (Fig. 23 of Armstrong’s paper). By proper adjustment of each oscillator, the operating point is selected to be just above the lower band in the anode current-grid voltage characteristic. Thus, the decrease in the anode current is reduced against the change in the grid potential. The balanced topology adds up the signal, eliminating the static to a considerable extent.

IV. THE ART OF RADIO RECEIVING

The prime requirements for the performance of any radio receiver are sensitivity and selectivity. Sensitivity is measured in terms of the level of a radio frequency (RF) signal applied to the input of a receiver at its nominal input frequency, which produces a specified signal-to-noise ratio at the output of the receiver. Selectivity is defined as the ability of the receiver to discriminate between the wanted input signal and the unwanted one, both of them being present simultaneously at the input. The parameters of the receiver included under this heading are adjacent channel
selectivity, cochannel rejection, blocking or desensitization, and intermodulation response.

Radio receivers might be classified by the following operating principles:

1) tuned radio-frequency (TRF);
2) regenerative;
3) superregenerative;
4) superheterodyne.

A. TRF Receivers

A TRF receiver is composed of a narrow-band RF filter that is tunable over the frequency range of the receiver. The TRF principle has the advantages of simplicity and freedom from the multiple and spurious responses that can occur with the superheterodyne receiver. Its disadvantages are poor sensitivity and selectivity compared with those of the superheterodyne. The neutrodyne circuit, invented in 1918 by Hazeltine, extended the life of TRF receivers until about 1930. A current obtained from the anode circuit was fed back into the grid circuit in the proper magnitude and phase to balance out, or neutralize, the effect of the grid-to-plate (anode) capacitance inside the tube, and thus it achieved stability and prevented oscillations [5].

The advent of electrically tunable yttrium–iron–garnet (YIG) filters [6] has made this principle useful nowadays in TRF spectrum analyzers. Some typical parameters for such equipment are: 1–18-GHz frequency range, 20-MHz bandwidth, and about −45-dBm sensitivity.

B. Regenerative Receivers

Regenerative receivers marked a step forward in providing greatly increased sensitivity. They inherently manifest large amplifications of small signals and small amplifications of large signals.

This principle was independently discovered in 1912 by four inventors (de Forest, Armstrong, and Langmuir in the United States and Meissner in Germany). The patent litigation that followed was the most controversial in the annals of electronics. (The passion is noticeable in the discussion accompanying the Armstrong paper.) The case dragged on for about 20 years. In 1917, Armstrong received the first IRE Medal of Honor for his work on regeneration and production of oscillations while in 1934, the U.S. Supreme Court decided in de Forest’s favor. By this time the regenerative receiver was obsolete.

In a regenerative detector, RF energy is fed back from the anode circuit to the grid circuit to give positive feedback at the carrier frequency. This raises the quality factor Q of the input resonant circuit, thereby increasing the sensitivity of the circuit.

As expected, the regenerative detector could easily be made to oscillate if the grid and anode coupling (regenerative control) was turned up too far. When the receiver was in an oscillating condition it became a threat, producing heterodyne whistles in nearby receivers. The regenerative detectors were also hard to adjust, and it was almost impossible to receive the same station at the same position of the tuning dial.

Regeneration is still used occasionally in Q-multiplier circuits.

C. Superregenerative Receivers

Superregeneration was discovered by Armstrong during the defense of his patent for the regenerative receiver [7]. Basically, such a receiver looks like a beat receiver, but two modifications are implemented: the oscillator is tuned on the same frequency as the received signal and a periodic signal is applied to the grid of the superregenerative tube (oscillator) to prevent sustained oscillations. The quenching frequency is usually between 20 000 and 100 000 Hz. When the superregenerative tube is enabled to oscillate, oscillations start up depending on the received signal level. For instance, if the amplitude of the received signal is high, the oscillations reach the steady-state amplitude quickly. The resulting RF pulses are duration modulated, and after demodulation and low-pass filtering the modulating signal is obtained.

Superregeneration has been applied to parametric amplifier detectors. Their relatively broad bandwidth with stable and high gain are desirable attributes in S-band experiments.

D. Superheterodyne Receivers

It might seem that receiver evolution could have progressed from the TRF circuit to the superheterodyne. Although Prof. R. Fessenden touched on a basic element with his invention of the heterodyne oscillator in 1912, the contemporary invention of the regenerative receiver focused attention on a simpler means of increasing sensitivity.

The basic idea of superheterodyne reception seemed to appear from J. H. Hammond, A. Meissner, L. Levy, E. F. Alexanderson, and Armstrong at about the same time. The combination of the received signal and a local oscillation resulted in an audio beat-note, the difference of the frequency between the two waves. The next step involved amplification of the beat note or intermediate frequency. Armstrong fully appreciated this concept and obtained a patent in 1920.

During World War I, Major Armstrong tried to devise a very sensitive receiver to detect enemy aircraft by short-wave ignition noise. It then occurred to him to use the heterodyne principle [8] to lower the incoming frequency to a value that could readily be amplified, but which was still high enough to use selective circuits, at what we now call the intermediate frequency. Other voices contend that the superheterodyne is the direct outcome of Armstrong’s work on the problem of the generation of superaudible frequencies by means of heterodyning [9].

The superheterodyne obtains only a relatively small part of its sensitivity and selectivity in the RF circuits (front end); by far, the larger part is obtained at a lower frequency in the IF amplifier that is tuned on a fixed frequency. This way, a more desirable bandpass selectivity characteristic...
can be achieved. Extremely small and efficient quartz-crystal or ceramic IF bandpass filters have been used to reduce the number of conventional tuned circuits. Where triple IF conversion was sometimes required it is now feasible to use simple conversion. The IF frequency of consumer AM receivers was raised from 42 kHz to about 175 kHz, and after 1938 it was set at 455 kHz by the Radio Manufacturers Associations ([RMA], now Electronic Industries Associations [EIA]). By 1930, superheterodynes were in universal use, as they are today.

V. DIGITAL TECHNOLOGY

A digital receiver is a receiver in which most of the receiver functions, such as signal amplification, filtering, frequency conversion, demodulation, and gain control are carried out by digital signal processing (DSP). Analog counterparts such as local oscillators, mixers, amplifiers, and filters tend to be performing close to their limits.

In a genuine digital receiver, the whole radio spectrum allowed by the input low-pass filter connected immediately after the antenna is digitized, and channel selection is achieved by changing the parameters of a digital filter according to the required carrier and bandwidth. Amplification, AGC, and demodulation are performed digitally. Such a simple architecture has many advantages: well-defined transfer characteristics, better selectivity, lack of careful alignment, no variable frequency components, more accurate digital demodulation, and so on. The data input to the DSP module of such a receiver, however, would depend on the upper frequency limit of the receiver [10]. Thus, a HF digital receiver for up to 30 MHz will require a signal processing capacity of 1.68 Gb/s resulting from a 2:1 shape factor for the input low-pass filter, a sampling frequency of 120 MHz associated with a 14-b word length.

For most of the communications band, a hybrid analog/digital receiver remains the only possibility for the time being. The main purpose is to convert the RF spectrum to a fixed IF which can be digitized using available ADC technology. The front end is a conventional analog superheterodyne with sound performance in terms of image and IF rejection, RF selectivity, and so on. The main advantages to be gained from digital filtering and demodulation, however, are still attainable. This architecture has yielded the best results to date and is widely used for wireless telephony mobile receivers.

VI. SOME CONCLUDING REMARKS

The basic obstacles to perfect radio reception lie in the propagation medium, interference problems, and in the receiver itself. Although present receivers have narrow IF bandwidth, much unnecessary noise and cross-modulation can result from the wide-open front end [11]. It is suggested that the development of sharp, tunable receiver input filters would hold great possibilities for improving radio-receiver performance. Parametric amplifiers, low-temperature, and atomic-excitation techniques have extended the range and reliability of electromagnetic-wave reception from the VLF bands through higher bands. A challenge remains. Will a better principle for radio reception be replacing the superheterodyne receiver? After more than 50 years, Armstrong’s results are still with us today, although the technology is different.

A sparkling imagination combined with a good engineering understanding was necessary to discover so many basic circuits. The unique referee for such a work is TIME and Marriott’s remark is here to stay: “Mr. Armstrong deserves much praise in carrying out his highly interesting investigation . . .” The Armstrong Medal is awarded by the Radio Club of America in his honor.

The discussion that follows the paper reveals that human nature, like the basic principles, have not changed; very few people can easily accept a different point of view than their own.

REFERENCES


Nicolae Cotanis (Member, IEEE) was born in 1952 in Constanta, Romania. He received the electronic engineer (wireless communication) degree in 1971 and the Ph.D. degree in 1991 from Politehnica University, Bucharest, Romania.

After serving several years in the Romanian Broadcast Company, he joined the Electronics and Telecommunication Faculty, Bucharest, Romania, in 1981, where he was a Lecturer and is currently a Professor. He developed the Electronic Measurement Laboratory and introduced the study of mobile communication. Since 1981, he has been one of the designers of the Automatic Vehicle location system used by the Regie Autonome des Transports, Bucharest, where he has acted as a consultant. His research interests focus on various fields of telecommunications, namely land mobile propagation, diversity techniques, handover algorithms, statistical modelling of multipath channel, random variables and processes, digital modulation and demodulation, and electronic measurements. He also has more than 15 years of research experience in automatic vehicle location and vehicular electronics. He is the author or co-author of nine books and more than 40 scientific papers and holds several patents.

Dr. Cotanis is a member of several professional societies.
Is it just a coincidence? Many of us remember when astronaut Neil Armstrong became the first person to set foot on the Moon in 1969. All over the Earth people heard his words: “That’s one small step… (More). Listeners receive the station they want by tuning their... to the station’s frequency. (to receive). The... were always correct provided the necessary instruments were used. (to measure). The sky wave from a very... transmitter can be reflected several times between the ionosphere and the Earth. (power). In radio telephones such as cellular mobile phones, voice signals are sent across town or over long distances by high-frequency radio signals called microwaves.