From T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*

A Nonlinear History of Radio

1.0 Introduction

Integrated circuit engineers have the luxury of taking for granted that the incremental cost of a transistor is essentially zero, and this has led to the high-device-count circuits that are common today. Of course, this situation is a relatively recent development; during most of the history of electronics, the economics of circuit design were the inverse of what they are today. It really wasn't all that long ago when an engineer was forced by the relatively high cost of active devices to try to get blood (or at least rectification) from a stone. And it is indeed remarkable just how much performance radio pioneers were able to squeeze out of just a handful of components. For example, we'll see how American radio genius Edwin Armstrong devised circuits in the early 1920's that trade *log* of gain for bandwidth, contrary to the conventional wisdom that gain and bandwidth should trade off more or less directly. And we'll see that at the same time Armstrong was developing those circuits, self-taught Soviet radio engineer Oleg Losev was experimenting with blue LEDs and constructing completely solid-state radios that functioned up to 5MHz, a quarter century before the transistor was invented.

These fascinating stories are rarely told because they tend to fall into the cracks between history and engineering curricula. *Somebody* ought to tell these stories, though, since in so doing, many commonly-asked questions ("why don't they do it this way?") get answered automatically ("they used to, but it caused key body parts to fall off"). This highly nonlinear history of radio touches briefly on just some of the main stories, and provides pointers to the literature for those who want to probe further.

2.0 Maxwell and Hertz

Every electrical engineer knows at least a bit about James Clerk (pronounced "clark") Maxwell; he wrote those equations that made life extra busy back in sophomore year or thereabouts. Not only did he write the electrodynamic equations[1] that bear his name, but he also published the first mathematical treatment of stability in feedback systems ("On Governors," which explained why speed controllers for steam engines could sometimes be unstable[2]).

Maxwell collected all that was then known about electromagnetic phenomena and, in a mysterious[3] and brilliant stroke, invented the displacement (capacitive) current term that

[1] Actually, Oliver Heaviside was the one who first used the notational conventions of vector calculus to cast Maxwell's equations in the form familiar to most engineers today.

[2] *Proc. Roy. Soc.*, 1868.

[3] Many E&M texts offer the logical, but historically wrong, explanation that Maxwell invented the displacement current term after realizing that there was an inconsistency between the known laws of E&M and the continuity equation for current. The truth is that Maxwell was a genius, and the inspirations of a genius often have elusive origins. This is one of those cases.

allowed him to derive an equation that led to the prediction of electromagnetic wave propagation.

Then came Heinrich Hertz, who was the first to verify experimentally Maxwell's prediction that electromagnetic waves exist, and propagate with a finite velocity. His "transmitters" worked on this simple idea: discharge a coil across a spark gap and hook up some kind of an antenna to launch a wave (unintentionally) rich in harmonics.

His setup naturally provided only the most rudimentary filtering of this dirty signal, so it took extraordinary care and persistence to verify the existence of (and to quantify) the interference nulls and peaks that are the earmarks of wave phenomena. He also managed to demonstrate such quintessential wave behavior such as refraction and polarization. And you may be surprised that the fundamental frequencies he worked with were between 50 and 500MHz. He was actually *forced* to these frequencies because his laboratory was simply too small to enclose several wavelengths of anything lower in frequency.

Because Hertz's sensor was another spark gap (integral with a loop resonator), the received signal had to be large enough to induce a visible spark. While adequate for verifying the validity of Maxwell's equations, you can appreciate the difficulties of trying to use this apparatus for wireless communication. After all, if the received signal has to be strong enough to generate a visible spark, scaling up to global proportions has rather unpleasant implications for those of us with metal dental work.

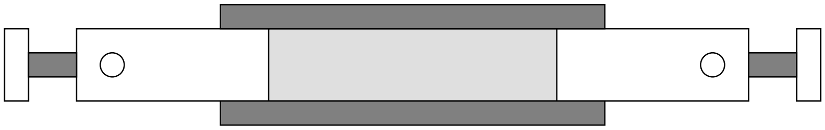
And then Hertz died. Young. Enter Marconi.

3.0 Pre-Vacuum Tube Electronics

For his radio experiments Marconi simply copied Hertz's transmitter and tinkered like crazy with the sole intent to use the system for wireless communication (and not incidentally to make a lot of money in the process). Recognizing the inherent limitations of Hertz's spark-gap detector, he instead used a bizarre creation that had been developed by Edouard Branly in 1890. As seen in Figure 1 the device, dubbed the "coherer" by Sir Oliver Lodge, consisted of a glass enclosure filled with a loosely packed, perhaps slightly oxidized metallic powder, whose resistance turned out to have interesting hysteretic behavior. Now, it must be emphasized that the detailed principles that underlay the operation of coherers have never been satisfactorily elucidated.[4] Nevertheless, we can certainly describe its behavior, even if we don't fully understand all the details of how it worked.

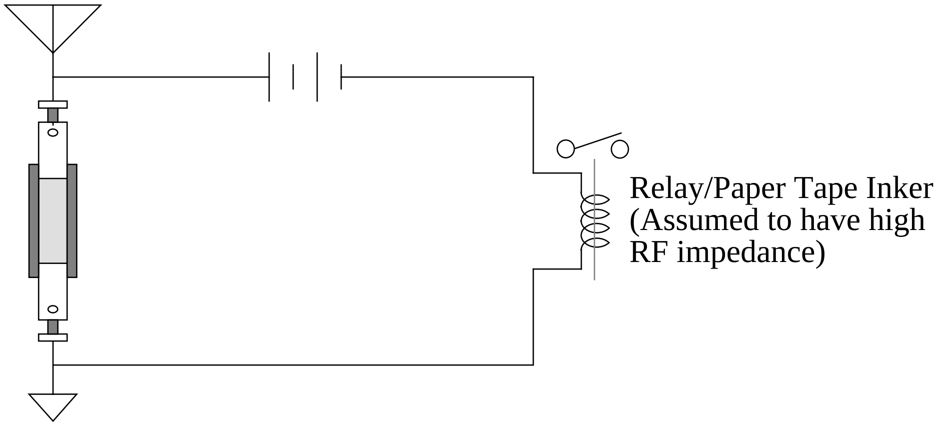
[4] Under large-signal excitation, the filings could be seen to stick together (hence the name "coherer"), and it's not hard to understand the drop in resistance in that case. However, apparently unknown to most authors, the coherer also worked with input energies so small that no such "coherence" is observed, so I assert that the detailed principles of operation remain unknown.

FIGURE 1. Branly's coherer



A coherer's resistance generally had a large value (say, megohms) in its quiescent state, and then dropped orders of magnitude (to kilohms or less) after an EM wave impinged on it. This large resistance change was usually used to trigger a solenoid to produce an audible click, as well as to ink a paper tape for a permanent record of the received signal. To prepare the coherer for the next EM pulse, it had to be shaken or whacked to restore the "incoherent" high resistance state. Figure 2 shows how a coherer was actually used in a receiver:

FIGURE 2. Typical receiver with coherer

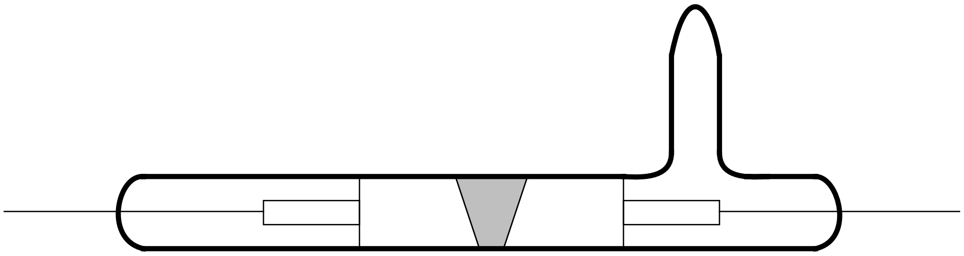


As can be seen, the coherer activated a relay (for audible clicks) or paper tape inker (for a permanent record) when a received signal triggered the transition to a low resistance state. It is evident that the coherer was basically a digital device, and therefore unsuitable for uses other than radiotelegraphy.

Marconi spent a great deal of time improving what was inherently a terrible detector and finally settled on the configuration shown in Figure 3. He greatly reduced the spacing between the end plugs (to a minimum of 2mm), filled the intervening space with a particular mixture of nickel and silver filings (in 19:1 ratio) of carefully selected size, and sealed the entire assembly in a partially evacuated tube. As an additional refinement in the receiver, a solenoid provided an audible indication in the process of automatically whacking the detector back into its initial state after each received pulse.[5]

[5] The coherer was most recently used in a radio-controlled toy truck in the late-1950's.

FIGURE 3. Marconi's coherer



As you can imagine, many EM events other than the desired signal could trigger a coherer, resulting in some difficult-to-read messages. Even so, Marconi was able to refine his apparatus to the point of achieving transatlantic wireless communications by 1901, with much of his success attributable to more powerful transmitters and large, elevated antennas that used the earth as one terminal (as did his transmitter), as well as to his improved coherer.

It shouldn't surprise you, though, that the coherer, even at its best, performed quite poorly. Frustration with the coherer's erratic nature impelled an aggressive search for better detectors. Without a suitable theoretical framework as a guide, however, this search sometimes took macabre turns. In one case, a human brain from a fresh cadaver was even used as a coherer, with the experimenter claiming remarkable sensitivity for his apparatus.[6] Let us all be thankful that this particular type of coherer never quite caught on.

Most research was guided by the vague intuitive notion that the coherer's operation depended on some mysterious property of imperfect contacts, and a variety of experimenters stumbled, virtually simultaneously, on the point-contact crystal detector (Figure 4). The first patent for such a device was awarded in 1904 (filed in 1901) to J.C. Bose for a detector that used galena (lead sulfide).[7] This appears to be the first patent awarded for a semiconductor detector, although it was not recognized as such (indeed, the word semiconductor had not yet been coined). Work along these lines continued, and General Henry Harrison Chase Dunwoody received a patent in late 1906 for a detector using carborundum (silicon carbide), followed in early 1907 by a patent to Greenleaf Whittier Pickard (an MIT graduate whose great-uncle was the poet John Greenleaf Whittier) for a silicon (!) detector. As shown in the figure, one connection to this type of detector consisted of a small wire (whimsically known as a catwhisker) that made a point contact to the crystal surface. The other connection was a large area contact typically formed by a low-melting-point alloy (usually a mixture of lead, tin, bismuth and cadmium known as Wood's metal

[6] A.F. Collins, *Electrical World and Engineer*, 39, 1902; he started out with brains of other species and worked his way up to humans.

[7] J.C. Bose, U.S. Patent #755,840, granted 19 March 1904. Actually, Ferdinand Braun had reported asymmetrical conduction in galena and copper pyrites (among others) back in 1874, in "Ueber die Stromleitung durch Schwefelmetalle ("On Current Flow through Metallic Sulfides"), *Poggendorff's Annalen der Physik und Chemie*, v. 153, pp. 556-563. The large-area contact was made through partial immersion in mercury, and the other with copper, platinum and silver wires. None of the samples showed more than a 2:1 forward/reverse current ratio. Braun later shared a Nobel Prize with Marconi for contributions to the radio art.

that has a melting temperature of under 80° C), that surrounded the crystal. One might call a device made this way a point-contact Schottky diode, although measurements are not always easily reconciled with such a description. In any event, we can see how the modern symbol for the diode evolved from a depiction of this physical arrangement, with the arrow representing the catwhisker point contact, as seen in the figure.

FIGURE 4. Typical crystal detector

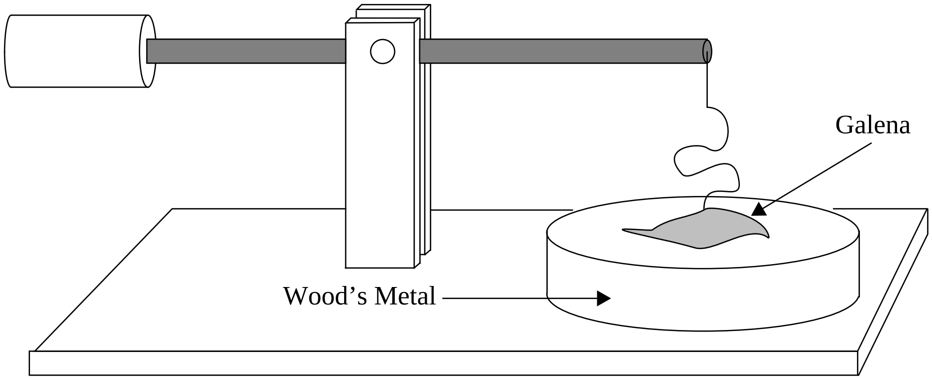
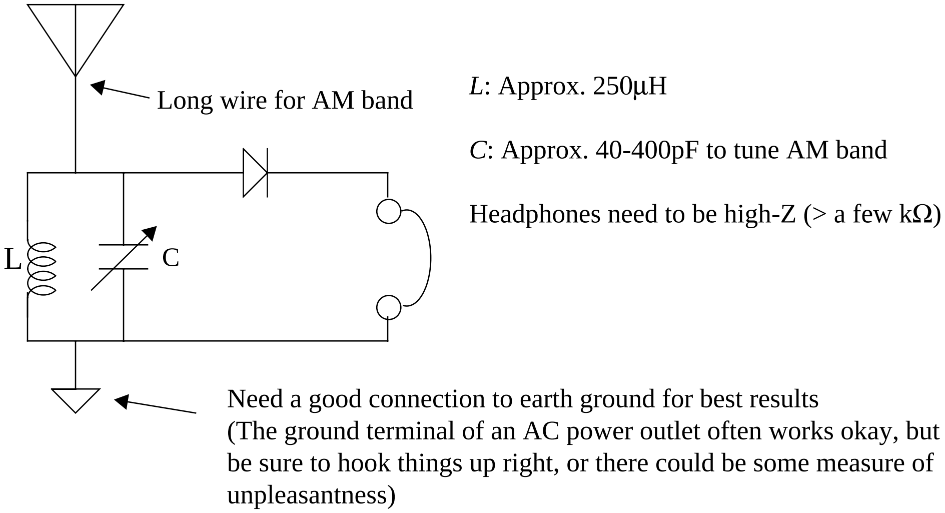


Figure 5 shows a simple crystal[8] radio made with these devices.[9] An *LC* circuit tunes the desired signal, which the crystal then rectifies, leaving the demodulated audio to drive the headphones. A bias source is not needed with some detectors (such as galena), so it is possible to make a "free-energy" radio![10]

FIGURE 5. Simple crystal radio


Long wire for AM band
L
C
L: Approx. 250μH
C: Approx. 40-400pF to tune AM band
Headphones need to be high-Z (> a few kΩ)
Need a good connection to earth ground for best results
(The ground terminal of an AC power outlet often works okay, but
be sure to hook things up right, or there could be some measure of
unpleasantness)

[8] In modern electronics, "crystal" usually refers to quartz resonators used, for example, as frequency determining elements in oscillators; these bear absolutely no relationship to the crystals used in crystal radios.

[9] A 1N34A germanium diode works fine and is more readily available, but lacks the charm of galena, Wood's metal and a catwhisker to fiddle with.

[10] Perhaps we should give a little credit to the human auditory system: the threshold of hearing corresponds to an eardrum displacement of about the diameter of a hydrogen atom!

Pickard worked harder than anyone else to develop crystal detectors, eventually trying over 30,000 combinations of wires and crystals. Among these were iron pyrites (fool's gold), and rusty scissors, in addition to silicon. Galena detectors became quite popular because they were inexpensive and needed no bias. Unfortunately proper adjustment of the catwhisker wire contact was difficult to maintain because anything other than the lightest pressure on galena destroyed the rectification. Plus, you had to hunt around the crystal surface for a sensitive spot in the first place. On the other hand, although carborundum detectors needed a bias of a couple of volts, they were more mechanically stable (a relatively high contact pressure was all right), and found wide use on ships as a consequence.[11]

At about the same time that these crude semiconductors were first coming into use, radio engineers began to struggle with a problem that was assuming greater and greater prominence: interference.

The broad spectrum of a spark signal made it impractical to attempt much other than Morse code types of transmissions (although some intrepid engineers did attempt AM transmissions with spark gap equipment, with little success). This broadband nature fit well with coherer technology, since the varying impedance of the latter made it difficult to realize tuned circuits anyhow. However, the inability to provide any useful degree of selectivity became increasingly vexing as the number of transmitters multiplied.

Marconi had made headlines in 1899 by contracting with the *New York Herald* and the *Evening Telegram* to provide up-to-the-minute coverage of the America's Cup yacht race, and was so successful that two additional groups were encouraged to try the same thing in 1901. One of these was led by Lee de Forest, whom we'll meet later, and the other by an unexpected interloper (who turned out to be none other than Pickard) from American Wireless Telephone and Telegraph. Unfortunately with *three* groups simultaneously sparking away that year, *no one* was able to receive intelligible signals, and race results had to be reported the old way, by semaphore. A thoroughly disgusted de Forest threw his transmitter overboard, and news-starved relay stations on shore resorted to making up much of what they reported.

This failure was all the more discouraging because Marconi, Lodge and that erratic genius Nikola Tesla had actually already patented circuits for tuning, and Marconi's apparatus had employed bandpass filters to reduce the possibility of interference.[12]

The problem was that, even though adding tuned circuits to spark transmitters and receivers certainly helped to filter the signal, no practical amount of filtering could ever really convert a spark train into a sinewave. Recognizing this fundamental truth, a number of

[11] Carborundum detectors were typically packaged in cartridges and often adjusted through the delicate procedure of slamming them against a hard surface.

[12] Marconi was the only one backed by strong financial interests (essentially the British government), and his British patent (no. 7777, the famous "four sevens" patent, granted 26 April 1900) was the dominant tuning patent of the early radio days. It was also involved in some of the lengthiest and most intense litigation in the history of technology. The U.S. Supreme Court finally ruled in 1943 that Marconi had been anticipated by Lodge, Tesla and others.

engineers sought ways of generating continuous sinewaves at radio frequencies. One group, which included Danish engineer Valdemar Poulsen[13] (who had also invented a crude magnetic recording device called the telegraphone) and Australian-American engineer (and Stanford graduate) Cyril Elwell, used the negative resistance associated with a glowing DC arc to keep an *LC* circuit in constant oscillation[14] to provide a sinewave RF carrier. Engineers quickly discovered that this approach could be scaled up to impressive power levels: an arc transmitter of over 1 *mega*watt was in use shortly after WWI!

Pursuing a somewhat different approach, Ernst F.W. Alexanderson of G.E. acted on Reginald Fessenden's request to produce RF sinewaves at large power levels with huge alternators (*really* big, high-speed versions of the thing that charges your car battery as you drive). This dead-end technology culminated in the construction of an alternator that put out 200kW at 100kHz! It was completed just as WWI ended, and was already obsolescent by the time it became operational.[15]

The superiority of the continuous wave over spark signals was immediately evident, and spurred the development of better receiving equipment. Thankfully, the coherer was gradually supplanted by a number of improved devices, including the semiconductor devices described earlier, and was well on its way to extinction by 1910 (although as late as the 1950's there was at least that one radio-controlled toy that used a coherer).

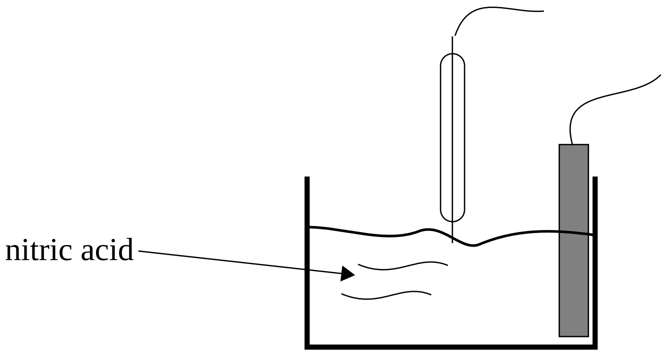
One such improvement, invented by Fessenden, was the "liquid barretter" shown in Figure 6. This detector consisted of a thin silver-coated platinum wire (a "Wollaston wire") encased in a glass rod. A tiny bit of the wire protruded from the rod and made contact with a small pool of nitric acid. This arrangement had a quasi-quadratic *V-I* characteristic near the origin and therefore could actually demodulate RF signals. The barretter was widely used in a number of incarnations since it was a "self-restoring" device (unlike typical coherers), and required no adjustments (unlike crystal detectors). Except for the hazards associated with the acid, the barretter was apparently a satisfactory detector, judging from the many infringements (including an infamous one by de Forest) of Fessenden's patent.

[13] Some sources persistently render his name incorrectly as "Vladimir," a highly un-Danish name!

[14] Arc technology for industrial illumination was a well developed art by this time. In his Stanford Ph.D. thesis, Leonard Fuller provided the theoretical advances that allowed arc power to break through a 30kW "brick wall" that had stymied others. Thanks to Fuller, 1,000kW arc transmitters were possible by 1919.

[15] Such advanced rotating machinery severely stretched the metallurgical state of art.

FIGURE 6. Fessenden’s liquid barretter



Enough rectifying detectors were in use by late 1906 to allow shipboard operators on the east coast of the U.S. to hear, much to their amazement (despite a forewarning by radiotelegraph three days before), the first AM broadcast by Fessenden himself on Christmas Eve.[16] Delighted listeners were treated to a program of poetry, Fessenden’s violin playing of Christmas carols, and some singing. He used a water-cooled carbon microphone *in series with the antenna* to modulate a 5kW (approximately), 50kHz (also approximate) carrier generated by a prototype Alexanderson alternator located at Brant Rock, Massachusetts. Those unfortunate enough to use coherers missed out on the historic event, since coherers as typically used are completely unsuited to AM demodulation. Fessenden repeated his feat a week later, on New Year’s Eve, to give more people a chance to get in on the fun.

The next year, 1907, was a significant one for electronics. Aside from following on the heels of the first AM broadcast (which marked the transition from radiotelegraphy to radiotelephony), it saw the emergence of important semiconductors. In addition to the patenting of the silicon detector, the LED was also discovered that year! In a brief article in Wireless World titled “A Note on Carborundum,” Henry J. Round of Great Britain reported the puzzling emission of a cold, blue[17] light from carborundum detectors under certain conditions (usually when the catwhisker potential was very negative relative to that of the crystal). The effect was largely ignored and ultimately forgotten as there were just so many more pressing problems in radio at the time. Today, however, carborundum is in fact used in blue LED’s,[18] and has been investigated by some to make transistors that can operate at elevated temperatures. And as for silicon, well, we all know how that turned out.

[16] Aitken (see references) erroneously gives the date as Christmas day.

[17] He saw orange and yellow, too. He may have been drinking.

[18] It should be mentioned that GaN-based LEDs offer much higher efficiency, but it was only very recently that people figured out how to dope the stuff without introducing serious defects. GaN blue LEDs are much more efficient than SiC ones.

4.0 Birth of the Vacuum Tube

The year 1907 also saw the patenting, by Lee de Forest, of the first electronic device capable of amplification: the triode vacuum tube. Unfortunately, de Forest didn't understand how his invention actually worked, having stumbled upon it by way of a circuitous (and occasionally unethical) route.

The vacuum tube actually traces its ancestry to the lowly incandescent light bulb of Thomas Edison. Edison's bulbs had a problem with progressive darkening caused by the accumulation of soot (given off by the carbon filaments) on the inner surface of the bulb. In an attempt to cure the problem, he inserted a metal electrode, hoping somehow to attract the soot to this plate rather than to the glass. Ever the experimentalist, he applied both positive and negative voltages (relative to one of the filament connections) to this plate, and noted in 1883 that a current mysteriously flowed when the plate was positive, but none flowed when the plate was negative. Furthermore, the current that flowed depended on how hot he made the filament. He had no theory to explain these observations (remember, the word electron wasn't even coined until 1891, and the particle itself wasn't unambiguously identified until J.J. Thomson's experiments of 1897), but Edison went ahead and patented in 1884 the first electronic (as opposed to electrical) device, one that exploited the dependence of plate current on filament temperature to measure line voltage indirectly. This Rube Goldberg instrument never made it into production since it was inferior to a standard voltmeter; Edison just wanted another patent, that's all (that's one way he ended up with over 1000 of them).

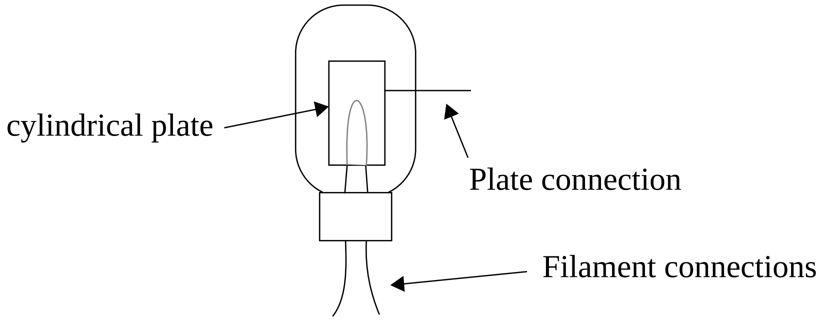
The funny thing about this episode is that Edison arguably had never invented anything in the fundamental sense of the term, and here he had stumbled across an electronic rectifier but nevertheless failed to recognize the implications of what he had found. Part of this blindness was no doubt related to his emotional (and financial) fixation on the DC transmission of power, where a rectifier had no role.

At about this time a consultant to the British Edison Company named John Ambrose Fleming happened to attend a conference in Canada. He dropped down to the U.S. to visit his brother in New Jersey and also stopped by Edison's lab. He was greatly intrigued by the "Edison effect" (much more so than Edison, who found it difficult to understand Fleming's excitement over something that had no obvious promise of practical application), and eventually published papers on the Edison effect from 1890 to 1896. Although his experiments created an initial stir, Röntgen's announcement in January 1896 of the discovery of X-rays as well as the discovery of natural radioactivity later that same year soon dominated the interest of the physics community, and the Edison effect quickly lapsed into obscurity.

Several years later, though, Fleming became a consultant to British Marconi and joined in the search for improved detectors. Recalling the Edison effect, he tested some bulbs, found out that they worked all right as RF rectifiers, and patented the Fleming valve (vacuum tubes are thus still known as valves in the U.K.) in 1905 (Figure 7). The nearly-deaf Fleming used a mirror galvanometer to provide a visual indication of the received signal, and included this feature as part of his patent.

©1998 Cambridge University Press

FIGURE 7. Fleming valve



While not particularly sensitive, the Fleming valve was at least continually responsive, and required no mechanical adjustments. Various Marconi installations used them (largely out of contractual obligations), but the Fleming valve never was popular (contrary to the assertions of some poorly researched histories) -- it needed too much power, filament life was poor, the thing was expensive, and it was a remarkably insensitive detector compared with, say, Fessenden's barretter, and well-made crystal detectors.

De Forest, meanwhile, was busy in America setting up shady wireless companies whose sole purpose was to earn money via the sale of stock. "Soon, we believe, the suckers will begin to bite," he wrote in his journal in early 1902. As soon as the stock in one wireless installation was sold, he and his cronies picked up stakes (whether or not the station was actually completed), and moved on to the next town. In another demonstration of his sterling character, he just outright stole Fessenden's barretter (simply reforming the Wollaston wire into the shape of a spade) after visiting Fessenden's laboratory, and even had the audacity to claim a prize for its invention. In this case however, justice did prevail and Fessenden won an infringement suit against de Forest.

Fortunately for de Forest, Dunwoody invented the carborundum detector just in time to save him from bankruptcy. Not content to develop this legitimate invention,[19] though, de Forest proceeded to steal Fleming's vacuum tube diode, and actually received a patent for it in 1905. He simply replaced the mirror galvanometer with a headphone, and added a huge forward bias (thus reducing the sensitivity of an already insensitive detector). De Forest repeatedly and unconvincingly denied throughout his life that he was aware of Fleming's prior work (even though Fleming published in professional journals that de Forest habitually and assiduously scanned), and to bolster his claims, de Forest pointed to his use of bias, where Fleming used none.[20] Conclusive evidence that de Forest had lied outright finally came to light when historian Gerald Tyne obtained the business records of W.

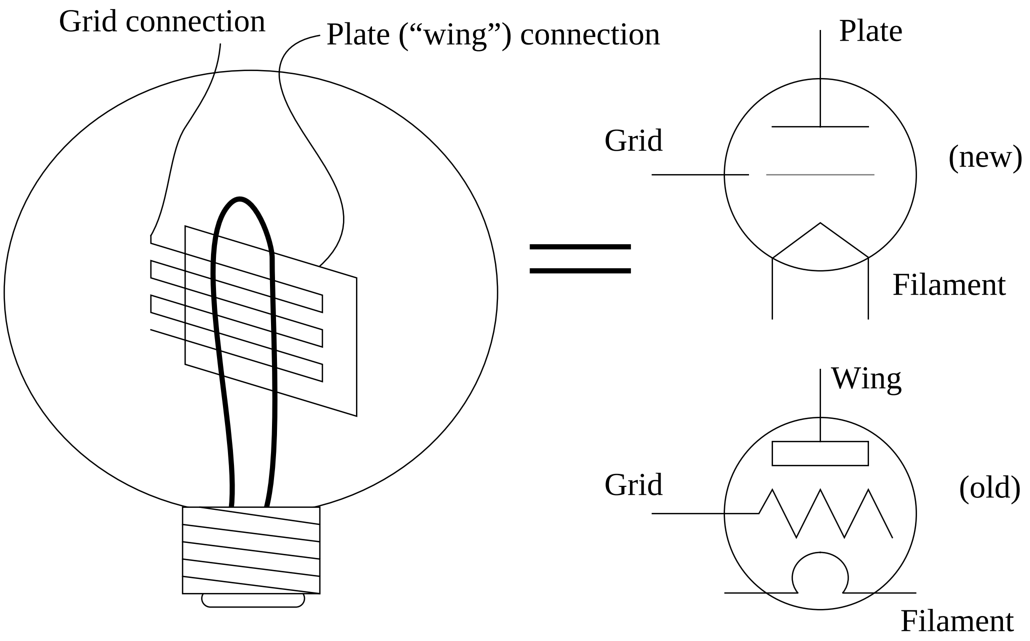
[19] Dunwoody had performed this work as a consultant to de Forest. He was unsuccessful in his efforts to get de Forest to pay him for it.

[20] In his efforts to establish that he had worked independently of Fleming, he repeatedly and stridently stated that it was his researches into the conductivity properties of flames that informed his work in vacuum tubes, arguing that ionic conduction was the key to the operation of his tubes. As a consequence, he boxed himself into a corner that he found difficult to escape later, after others developed the superior high-vacuum tubes that were essentially free of ions.

McCandless, the man who made all of de Forest's first vacuum tubes (de Forest called them *audions*). The records clearly show that de Forest had asked McCandless to duplicate some Fleming valves months before he filed his patent. There is thus no room for a charitable interpretation that de Forest independently invented the vacuum tube diode.

His crowning achievement came soon after, however. He added a zigzag wire electrode, which he called the grid, between the filament and wing electrode (later known as the plate), and thus the triode was born (see Figure 8). This three-element audion was capable of amplification, but de Forest did not realize this fact until years later. In fact, his patent application only mentioned the triode audion as a detector, not as an amplifier.[21] Motivation for the addition of the grid is thus still curiously unclear. He certainly did not add the grid as the consequence of careful reasoning, as some histories claim. The fact is that he added electrodes all over the place. He even tried "control electrodes" outside of the plate! We must therefore regard his addition of the grid as merely the result of haphazard but persistent tinkering in his search for a detector to call his own. It would not be inaccurate to say that he stumbled onto the triode, and it is certainly true that others had to explain its operation to him.[22]

FIGURE 8. De Forest triode Audion and symbols



From the available evidence, neither de Forest nor anyone else thought much of the audion for a number of years (1906-1909 saw essentially no activity on the audion). In fact, when de Forest barely escaped conviction and a jail sentence for stock fraud after the collapse of one of his companies, he had to relinquish interest in all of his inventions as a condition of

[21] Curiously enough, though, his patent for the two-element audion *does* talk about amplification.

[22] Aitken (see references at end of this chapter) argues that de Forest has been unfairly accused of not understanding his own invention. However, the bulk of the evidence contradicts Aitken's generous view.

the subsequent reorganization of his companies, with just one exception: the lawyers let him keep the patent for the audion, thinking it worthless.[23]

He intermittently puttered around with the audion and eventually discovered its amplifying potential, as did others almost simultaneously (including rocket pioneer Robert Goddard).[24] He managed to sell the device to AT&T in 1912 as a telephone repeater amplifier, but initially had a tough time because of the erratic behavior of the audion. Reproducibility of device characteristics was rather poor and the tube had a limited dynamic range. It functioned well for small signals, but behaved badly upon overload (the residual gas in the tube would ionize, resulting in a blue glow and a frying noise in the output signal). To top things off, the audion filaments (made of tantalum) had a life of only about 100-200 hours. It would be a while before the vacuum tube could take over the world.

5.0 Armstrong and the Regenerative Amplifier/Detector/ Oscillator

Fortunately, some gifted people finally became interested in the audion. Irving Langmuir at GE Labs in Schenectady worked to achieve high vacua, thus eliminating the erratic behavior caused by the presence of (easily ionized) residual gases. De Forest never thought to do this (in fact, warned against it, believing that it would reduce the sensitivity) because he never really believed in thermionic emission of electrons (indeed, it isn’t clear he even believed in electrons at the time), asserting instead that the audion *depended* fundamentally on ionized gas for its operation.

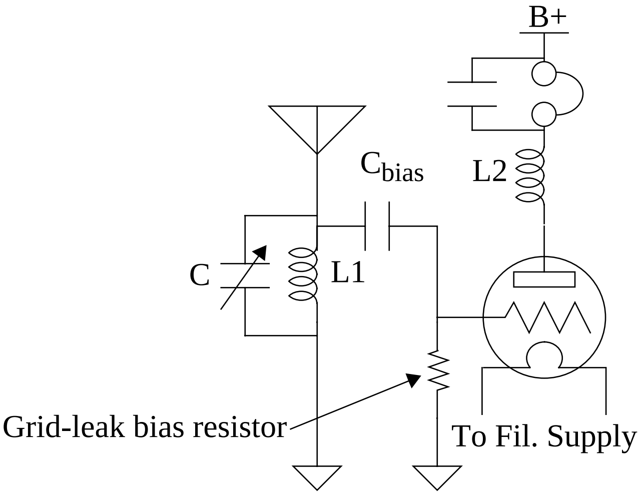
After Langmuir’s achievement, the way was paved for a bright engineer to devise useful circuits to exploit the audion’s potential. That bright engineer was Edwin Howard Armstrong who invented the regenerative amplifier/detector[25] in 1912 at the tender age of 21. This circuit (a modern version of which is shown in Figure 9) employed positive feedback (via a “tickler coil” that coupled some of the output energy back to the input with the right phase) to boost the gain and Q of the system simultaneously. Thus high gain (for good sensitivity) and narrow bandwidth (for good selectivity) could be obtained rather simply from one tube. Additionally, the nonlinearity of the tube demodulated the signal. Furthermore, overcoupling the output to the input turned the thing into a wonderfully compact RF oscillator.

[23] The recently unemployed de Forest then went to work for Elwell at Federal Telephone and Telegraph in Palo Alto.

[24] His U.S. Patent #1,159,209, filed 1 August 1912 and granted 2 November 1915, describes an audion oscillator, and thus actually predates even Armstrong’s documented work.

[25] His notarized notebook entry is actually dated 31 January 1913.

FIGURE 9. Armstrong regenerative receiver



The amount of positive feedback is controlled by the coupling between L1 and L2.

In a 1914 paper titled “Operating Features of the Audion,”[26] Armstrong published the first correct explanation for how the triode worked, and provided experimental evidence to support his claims. He followed this paper with another (“Some Recent Developments in the Audion Receiver”)[27] in which he additionally explained the operation of the regenerative amplifier/detector, and showed how to make an oscillator out of it. The paper is a model of clarity and quite readable even to modern audiences. De Forest, however, was quite upset at Armstrong’s presumptuousness. In a published discussion section following the paper, de Forest repeatedly attacked Armstrong. It is clear from the published exchange that, in sharp contrast with Armstrong, de Forest had difficulty with certain basic concepts (e.g., that the average value of a sinewave is zero), and didn’t even understand how the triode, his own invention (more of a discovery, really), actually worked.

The bitter enmity that arose between these two men never waned.

Armstrong went on to develop circuits that continue to dominate communications systems to this day. While a member of the U.S. Army Signal Corps during World War I, Armstrong became involved with the problem of detecting enemy planes from a distance, and pursued the idea of trying to home in on the signals naturally generated by their ignition systems (spark transmitters again). Unfortunately, little useful radiation was found below about 1MHz, and it was exceedingly difficult with the tubes available at that time to get much amplification above that frequency. In fact, it was only with extraordinary care that H.J. Round (of blue LED fame) achieved useful gain at 2MHz in 1917, so Armstrong had his work cut out for him.

He solved the problem by employing a principle originally used by Poulsen and later elucidated by Fessenden. When demodulating a CW signal, the resultant DC pulse train

[26] *Electrical World*, 12 December 1914.

[27] *IRE Proceedings*, v.3, 1915, pp 215-247.

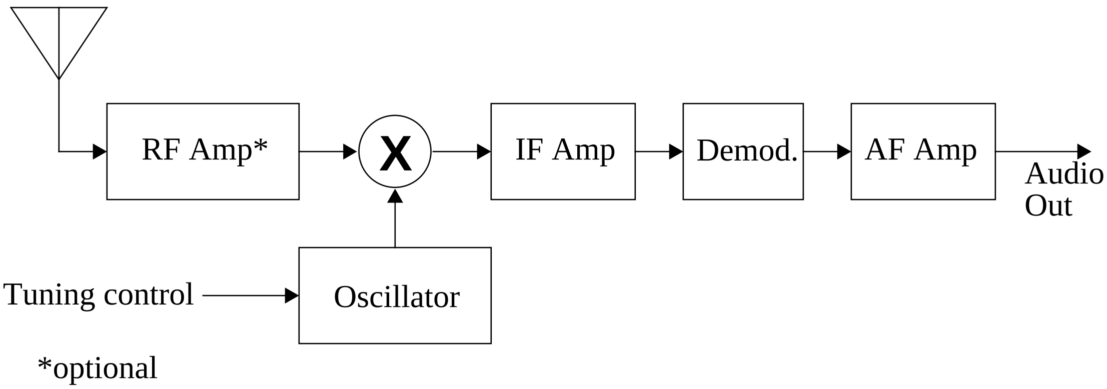
could be hard to make out. Valdemar Poulsen offered an improvement by inserting a rapidly driven interrupter in series with the headphones. This way, a steady DC level is chopped into an audible waveform. The “Poulsen Tikker” made CW signals easier to copy as a consequence.

Fessenden, whose fondness for rotating machines was well known, used much the same idea, but derived his signal from a high speed alternator that could heterodyne signals to any desired audible frequency, allowing the user to select a tone that cut through the interference.

Armstrong decided to employ Fessenden’s heterodyne principle in a different way. Rather than using it to demodulate CW directly, he used the heterodyne method to convert an incoming high frequency RF signal into one at a lower frequency, where high gain and selectivity could be obtained with relative ease. This signal, known as the intermediate frequency (IF), was then demodulated after much filtering and amplification at the IF had been achieved. The receiver could easily possess enough sensitivity so that the limiting factor was actually atmospheric noise (which is quite large in the AM broadcast band). Furthermore, a single tuning control was made possible, since the IF amplifier works at a fixed frequency.

He called this system the “superheterodyne” and patented it in 1917 (see Figure 10). Although the war ended before Armstrong could use the superhet to detect German planes, he continued to develop it with the aid of several talented engineers, finally reducing the number of tubes to five from an original complement of ten (good thing, too: the prototype had a total filament current requirement of ten amps). David Sarnoff of RCA eventually negotiated the purchase of the superhet rights, and RCA came to dominate the radio market by 1930 as a consequence

FIGURE 10. Superheterodyne Receiver Block Diagram



The great sensitivity enabled by the invention of the vacuum tube allowed transmitter power reductions of orders of magnitude while simultaneously increasing useful communications distances. Today, 50kW is considered a large amount of power, while ten times this amount was the norm right after WWI.

The 1920's saw greatly accelerated development of radio electronics. The war had spurred the refinement of vacuum tubes to an astonishing degree, with the appearance of improved filaments (longer life, higher emissivity, lower power requirements), lower interelectrode capacitances, higher transconductance and greater power handling capability. These developments set the stage for the invention of many clever circuits, some designed to challenge the dominance of Armstrong's regenerative receiver.

6.0 Other Radio Circuits

6.1 The TRF and the Neutrodyne

One wildly popular type of radio in the early days was the tuned radio-frequency (TRF) receiver. The basic TRF circuit typically had three RF bandpass stages, each tuned separately, and then a stage or two of audio after demodulation (the latter sometimes accomplished with a crystal diode). The user thus had to adjust three or more knobs to tune in each station. While this array of controls may have appealed to the tinkering-disposed technophile, it was rather unsuited to the average consumer.

Oscillation of the TRF stages was also a big problem, caused by the parasitic feedback path provided by the grid-plate capacitance C_{gp}.[28] While limiting the gain per stage was one way to reduce the tendency to oscillate, the attendant degradation in sensitivity was usually unacceptable.

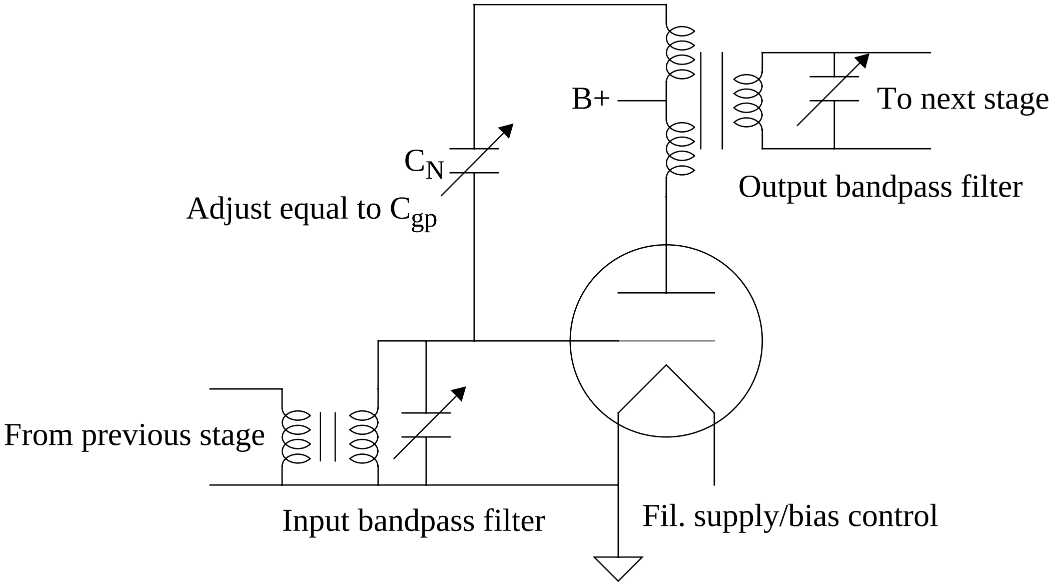
The problem caused by C_{gp} was largely eliminated by Harold Wheeler's invention[29] of the Neutrodyne circuit (see Figure 11).[30]

[28] It is left as "an exercise to the reader" to show that the real part of the input impedance of an inductively loaded common-cathode amplifier can be less than zero because of the feedback through C_{gp}, and that this negative resistance therefore can cause instability.

[29] He did this work for Louis Hazeltine, who is frequently given credit for the circuit.

[30] Of course, it should be noted that Armstrong's superheterodyne neatly solves the problem by obtaining gain at a number of different frequencies: RF, IF and AF. This approach also reduces greatly the danger of oscillation from parasitic input-output coupling.

FIGURE 11. Basic Neutrodyne amplifier



Recognizing the cause of the problem, he inserted a compensating capacitance (C_N), termed the neutralizing capacitor (actually, *condenser* was the term back then). When properly adjusted, the condenser fed back a current exactly equal in magnitude but opposite in phase with that of the plate-to-grid capacitance, so that no input current was required to charge the capacitances. The net result was the suppression of C_{gp}'s effects, permitting a large increase in gain per stage without oscillation.[31] After the War, Westinghouse acquired the rights to Armstrong's regeneration patent, negotiated licensing agreements with a limited number of radio manufacturers, then aggressively prosecuted those who infringed (which was just about everybody). To protect themselves, those "on the outside" organized into the Independent Radio Manufacturers Association, and bought the rights to Hazeltine's circuit. Tens of thousands of Neutrodyne kits and assembled consoles were sold in the 1920's by members of IRMA, all in an attempt to compete with Armstrong's regenerative circuit.

Meanwhile, de Forest was up to his old tricks. He bought a company that had a license to make Armstrong's regenerative circuits. Although he knew that the license was non-transferable, he nonetheless started to sell regenerative radios until he was caught and threatened with lawsuits. He eventually skirted the law by selling a radio that *could* be hooked up as a regenerator by the customer simply by reconnecting a few wires between binding posts that had been conveniently provided for this purpose.[32]

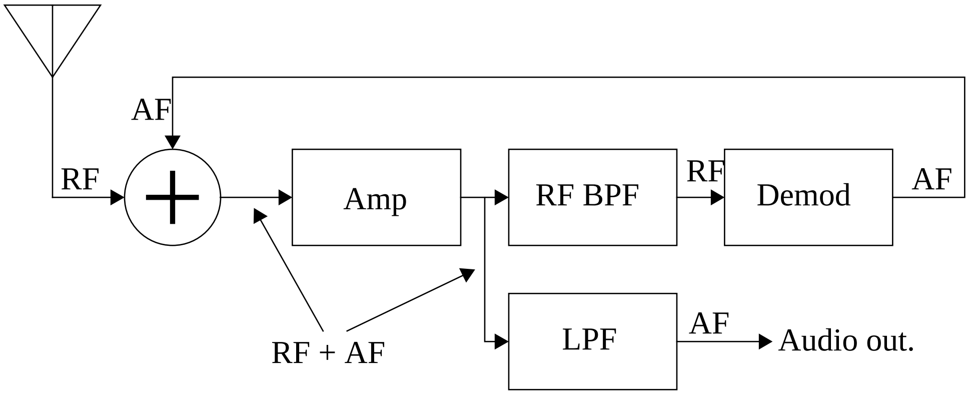
[31] In some sets, only the middle TRF stage is neutralized.

[32] One anecdotal report has it that de Forest sold receivers with a wire that protruded from the back panel, marked with a label that said something like "Do not cut this wire; it converts this receiver into a regenerative one." I have not found a primary source for this information, but it is entirely consistent with all we know about de Forest's character.

6.2 The Reflex Circuit

The reflex circuit (Figure 12) enjoyed some prominence in the early 1920s, but was more popular with hobbyists and experimenters than with commercial industry. The idea behind the reflex is wonderful and subtle, and perhaps even the inventor of the circuit himself (believed to be French engineer Marius Latour[33]) did not fully appreciate just how marvelous it was. The basic idea was this: pass the RF through some number (say, one) of amplifier stages, demodulate, and then pass the audio back through those *same* amplifiers. A given tube thus simultaneously amplified both RF and AF signals.

FIGURE 12. Reflex Receiver Block Diagram



The reason that this arrangement made sense becomes convincingly clear only when you consider how this connection allowed the overall system to possess a gain-bandwidth product that exceeded that of the active device itself. Suppose that the vacuum tube in question had a certain constant gain-bandwidth product limit. Further assume that the incoming RF signal was amplified by a factor G_{RF} over a brickwall passband of bandwidth B, and that the audio signal was also amplified by a factor G_{AF} over the same brickwall bandwidth B. The overall gain-bandwidth product was therefore $(G_{RF}G_{AF})B$, while the gain-bandwidth product of the combined RF/AF signal processed by the amplifier was just $(G_{RF} + G_{AF})B$. For the reflex circuit to have an advantage, we just want the product of the gains to exceed the sum of the gains, a criterion that is easily satisfied.

The reflex circuit demonstrates that there is nothing fundamental about gain-bandwidth, that we are effectively fooled into believing that gain and bandwidth must trade off linearly, just because they commonly do. The reflex circuit shows us the error in our thinking. For this reason alone, the reflex circuit deserves more detailed treatment than it commonly receives.

[33] It should be noted that Armstrong's second paper on the superheterodyne (published in 1924) contains examples of reflex circuits.

7.0 Armstrong and the Superregenerator

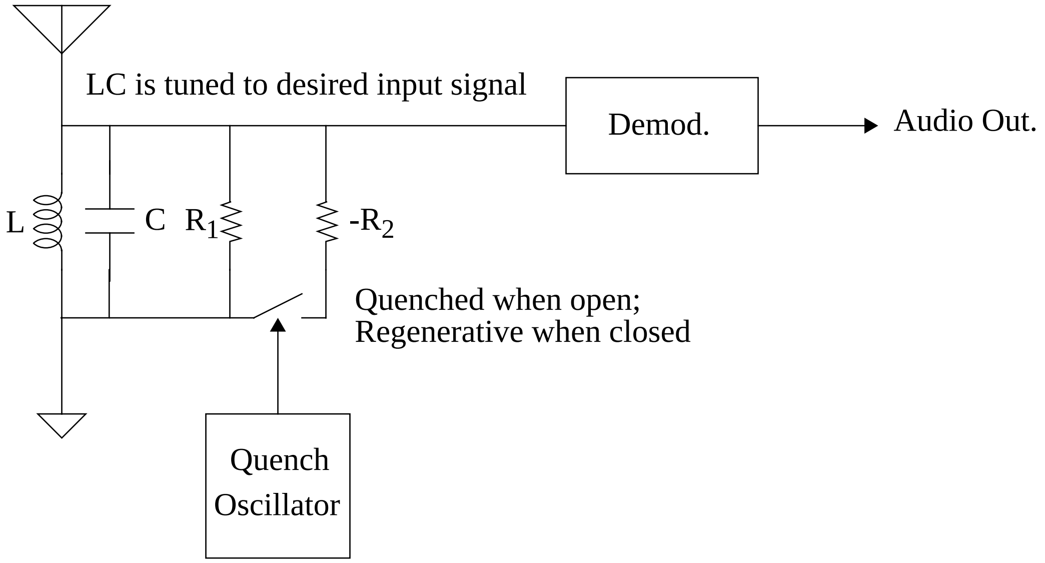
Armstrong wasn't content to rest, although after having invented both the regenerative and superheterodyne receivers he would seem to have had the right.

While experimenting with the regenerator, he noticed that under certain conditions, he could, for a fleeting moment, get much greater amplification than normal. He investigated further and developed by 1922 a circuit he called the superregenerator, a circuit that provides so much gain in a *single tube* that it can amplify thermal and shot noise to audible levels!

Perhaps you found the reflex principle a bit abstruse; you ain't seen nothin' yet. In a superregenerator the system is *purposely made unstable*, but is periodically shut down (quenched) to prevent getting stuck in some limit cycle.

How can such a bizarre arrangement provide gain (lots of gain)? Take a look at Figure 13, which strips the superregenerator to its basic elements.[34]

FIGURE 13. Superregenerative receiver basics



Now, during the time that it is active (i.e., the negative resistor is connected to the circuit), this second-order bandpass system has a response that grows exponentially with time. Response to what? Why, the initial conditions, of course! A tiny initial voltage will, given sufficient time, grow to detectable levels in such a system. The initial voltage could conceivably even come from thermal or shot noise processes.

[34] The classic vacuum tube superregenerator looks a lot like a normal regenerative amplifier, except that the grid leak bias network time constant is made very large and the feedback (via the tickler coil) is large enough to guarantee instability. As the amplitude grows, the grid leak bias also grows until it cuts off the tube. The tube remains cut off until the bias decays to a value that returns the tube to the active region. Thus, no separate quench oscillator is necessary.

The problem with all real systems is that saturation eventually occurs, and no further amplification is possible in such a state. The superregenerator evades this problem by periodically shutting the system down. This periodic “quenching” can be made inaudible if a sufficiently high quench frequency is chosen.

Because of the exponential growth of the signal with time, the superregenerator trades off *log* of gain for bandwidth. As a bonus, the unavoidable nonlinearity of the vacuum tube can be exploited to provide demodulation of the amplified signal! As you might suspect, the superregenerator’s action is so subtle and complex that it has never been understood by more than a handful of people at a given time. It’s a quasiperiodically time-varying, nonlinear system that is allowed to go intermittently unstable, and Armstrong invented it in 1922.

Armstrong sold the patent rights to RCA (who shared Armstrong’s view that the superregenerator was the circuit to end all circuits), and became its largest shareholder as a consequence.[35] Alas, the superregenerator never assumed the dominant position that he and RCA’s David Sarnoff had envisioned. The reason is simple for us to see now: every superregenerative amplifier is fundamentally also an oscillator. Therefore, every superregenerative receiver is also a transmitter that is capable of causing interference to nearby receivers. In addition, the superregenerator produces an annoyingly loud hiss (the amplified thermal and shot noise) in the absence of a signal, rather than the relative quiet of other types of receivers. For these reasons, the superregenerator never took the radio world by storm.

The circuit has found wide application in toys, however. When you’ve got to get the most sensitivity with absolutely the minimum number of active devices, you cannot do better than the superregenerative receiver. Radio-controlled cars, automatic garage door openers and toy walkie-talkies almost invariably use a circuit that consists of just one transistor operating as a superregenerative amplifier/detector, and perhaps two or three more as amplifiers of the demodulated audio signal (as in a walkie-talkie). The overall sensitivity is often of the same order as that provided by a typical superhet. On top of those attributes, it can also demodulate FM through a process known as slope demodulation: if one tunes the receiver a bit off frequency so that the receiver gain vs. frequency is not flat (i.e., has some slope, hence the name), an incoming FM signal produces a signal in the receiver whose amplitude varies as the frequency varies; the signal is converted into an AM signal which is demodulated as usual (“it’s both a floor wax *and* a dessert topping”). So, if most of the system cost is associated with the number of active devices, the superregenerative receiver provides a remarkably economical solution.

8.0 Oleg Losev and the First Solid-State Amplifier

Surely one of the most amazing (and little-known) stories from this era is that of self-taught Soviet engineer Oleg Losev and his solid-state receivers of 1922. Vacuum tubes

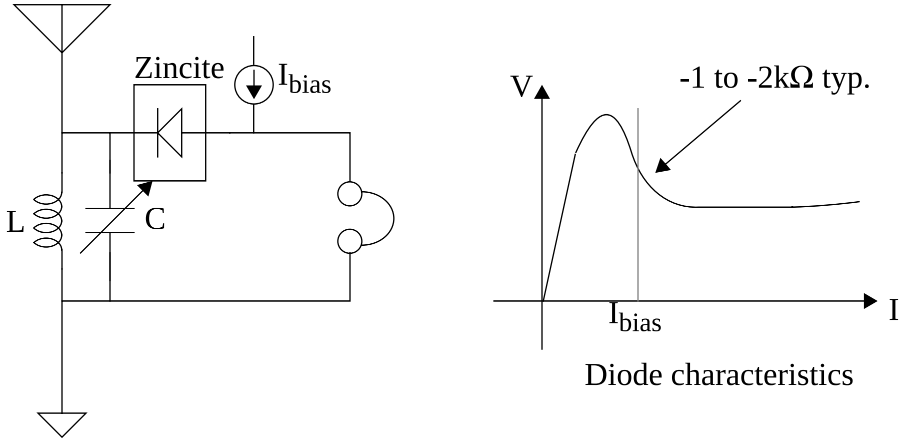
[35] In a bit of fortuitous timing, Armstrong sold his stock just before the great stock market crash of 1929.

were expensive then, particularly in the Soviet Union so soon after the revolution, so there was naturally a great desire to make radios on the cheap.

Losev's approach was to investigate the mysteries of crystals, which by this time were all but forgotten in the West. He independently rediscovered Round's carborundum LEDs, and actually published about a half dozen papers on the phenomenon. He correctly deduced that it was a quantum effect, describing it as the inverse of Einstein's photoelectric effect, and correlated the short wavelength cutoff energy with the applied voltage. He even noted that the light was emitted from a particular crystalline boundary (which we would call a junction), and cast doubt on a prevailing theory of a thermal origin by showing that the emission could be electronically modulated up to at least 78kHz (the limit of his rotating-mirror instrumentation).

Even more startling than his insights into the behavior of LEDs was his discovery of the negative resistance that can be obtained from biased point-contact zincite (ZnO) crystal diodes. With zincite, he actually constructed fully solid-state RF amplifiers, detectors and oscillators at frequencies up to 5MHz a whole quarter century before the invention of the transistor! Later, he even went on to construct a superheterodyne receiver with these crystals. True, one had to adjust several bias voltages and catwhiskers, but it nevertheless worked (see Figure 14). He eventually abandoned the "crystadyne" after about a decade of work though, because of difficulties with obtaining zincite (it's found in commercially significant quantity in only two mines, and they're both in New Jersey), as well as the problem of interstage interaction inherent in using two-terminal devices to get gain.

FIGURE 14. Losev's Crystadyne receiver (single stage)



The reason almost no one in the U.S. has ever heard of Losev is simple. First, almost no one has even heard of Armstrong -- it seems that there isn't much interest in preserving the names of these pioneers. Plus, most of Losev's papers are in German and Russian, limiting readership. Add the generally poor relations between the U.S. and the U.S.S.R over most of this century, and it's actually a wonder that *anyone* knows who Losev was. Losev himself isn't around because he was one of many who starved to death during the terrible siege of Leningrad, breathing his last in January of 1942. His colleagues had advised him to leave, but he was just too interested in finishing up what he termed were "promising experiments with silicon." Sadly, all records of those experiments have apparently been lost.

9.0 Epilog

By the early 1930's, the superhet had been refined to the point that a single tuning control was all that was required. The superior performance and ease of use of the superhet guaranteed its dominance (as well as that of RCA), and virtually every modern receiver, ranging from portable radios to radar sets, employs the superheterodyne principle, and it seems unlikely that this situation will change in the near future. It is a tribute to Armstrong's genius that a system he conceived during World War I still dominates on the eve of the 21st century.

Armstrong, annoyed by the static that plagues AM radio, went on to develop (wideband) frequency modulation, in defiance of theoreticians who declared FM useless.[36] Unfortunately, Armstrong's life did not end happily. In a sad example of how our legal system is often ill-equipped to deal intelligently with technical matters, de Forest challenged Armstrong's regeneration patent, and ultimately prevailed in some of the longest patent litigation in history (it lasted twenty years). Not long after the courts handed down the final adverse decision in this case, Armstrong began locking horns with his former friend Sarnoff and RCA in a bitter battle over FM that raged for well over another decade. His energy and money all but gone, Armstrong committed suicide in 1954 at the age of 63 on the fortieth anniversary of his demonstration of regeneration to Sarnoff. Armstrong's widow, Marian, picked up the fight and eventually went on to win every legal battle; it took fifteen years.

De Forest eventually went legit. He moved to Hollywood and worked on developing sound and color for motion pictures. A few years before he died at the ripe old age of 87, he penned a characteristically self-aggrandizing autobiography titled *The Father of Radio* that sold fewer than 1000 copies. He also tried to get his wife to write a book called *I Married a Genius* but she somehow never got around to it.

Further reading:

The stories of de Forest, Armstrong and Sarnoff are wonderfully recounted by Tom Lewis in *The Empire of the Air*, a book that was turned into a film by Ken Burns for PBS. Although it occasionally gets into trouble when it ventures a technical explanation, the human focus and rich biographical material that Lewis has unearthed much more than compensates. (Prof. Lewis says that many corrections will be incorporated in a later paperback edition of his book.)

[36] Bell Laboratories mathematician John R. Carson (no known relation to the entertainer) had correctly shown that FM always requires more bandwidth than AM, disproving a prevailing belief to the contrary. But he went too far in declaring FM worthless.

For those interested in more technical details, there are two excellent books by Hugh Aitken. *Syntony and Spark* recounts the earliest days of radiotelegraphy, beginning with pre-Hertzian experiments and ending with Marconi. *The Continuous Wave* takes the story up to the 1930's, covering arc and alternator technology in addition to vacuum tubes. Curiously, though, Armstrong is but a minor figure in Aitken's portrayals.

The story of early crystal detectors is well told by A. Douglas, "The crystal detector," *IEEE Spectrum*, pp. 64-67, April 1981, and by D. Thackeray in "When tubes beat crystals: early radio detectors," *IEEE Spectrum*, pp. 64- 69, March 1983. Material on other early detectors is found in a delightful volume by V. Phillips, *Early Radio Wave Detectors*, Peter Peregrinus, 1980. Finally, the story of Losev is recounted by E. Loebner in "Subhistories of the light-emitting diode," *IEEE Transactions on Electron Devices*, pp. 675-699, July 1976.

10.0 Appendix A: A Vacuum Tube Primer

10.1 Introduction

Sadly, few engineering students are ever exposed to the vacuum tube. Indeed, most engineering faculty regard the vacuum tube a quaint relic. Well, maybe they're right, but there are still certain engineering provinces (such as high-power RF) where the vacuum tube reigns supreme.

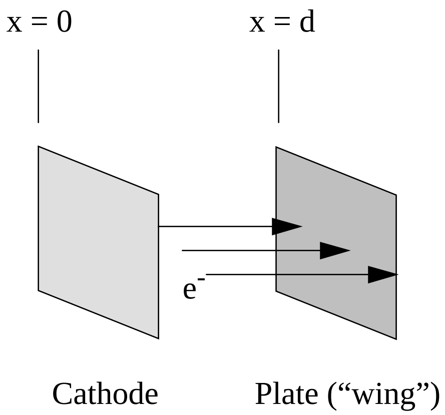
This Appendix is intended to provide the necessary background so that an engineer educated in solid-state circuit design can develop at least a superficial familiarity with this historically important device.

The operation of virtually all vacuum tubes can be understood rather easily once you study the physics of the vacuum diode. To simplify the development, we'll follow a historical path and consider a parallel-plate structure rather than the more common coaxial structures. The results are easier to derive but still hold generally.

10.2 Cathodes

Consider the diode structure shown in the following figure:

FIGURE 15. Idealized diode structure



The left-most electrode is the cathode, whose job is to emit electrons. The plate’s job is to collect them.

All the early tubes (Edison’s and Fleming’s diode, and de Forest’s triode audion) used directly-heated cathodes, meaning that the light bulb filament did the work of emitting electrons. Physically all that happens is that, at high enough temperatures, the electrons in the filament material are given enough kinetic energy that they can leave the surface; they literally boil off.

Clearly, materials that emit well at temperatures below the melting point make the best cathodes. De Forest’s first filaments were made of the same carbon variety used in Edison’s light bulbs, although tantalum, which has a high melting point (about 3100 kelvins), quickly replaced carbon. Useful emission from tantalum occurs only if the material is heated to bright incandescence, though, so the early audions were pretty power-hungry. Additionally, tantalum tends to crystallize at high temperature, and filament life is unsatisfactory as a consequence of the attendant increasing brittleness. A typical audion filament had a lifetime as short as 100-200 hours. Some audions were made with a spare filament that could be switched in when the first filament burned out.

Research by Coolidge (same guy who developed the high-power X-ray tube) at GE allowed the use of tungsten (melting point: 3600 kelvins) as a filament material. He found a way to make filaments out of the unwieldy stuff (tungsten is not ductile, and hence it ordinarily cannot be drawn into wires) and opened the path to great improvements in vacuum tube (and light bulb) longevity because of the high melting point of that material.[37]

Unfortunately, lots of heating power is required to maintain the operating temperature of about 2400K, and portable (or even luggable) equipment just could not evolve until these heating requirements were reduced. One path to improvement (discovered accidentally) is to add a little thorium to the tungsten. If the temperature is held within rather narrow limits (around 1900K), the thorium diffuses from the bulk onto the surface, where it serves to lower the work function (the binding energy of electrons) and thereby increases emissiv-

[37] Tungsten is still used in light bulbs today.

ity. These thoriated tungsten filaments still find wide use in high-power transmitting tubes, but their filament temperature must be controlled rather tightly. If the temperature is too high, the thorium boils off quickly (leaving a pure tungsten filament behind), and if it is too low, the thorium does not diffuse to the surface fast enough to do any good.

While thoriated tungsten is a more efficient emitter than pure tungsten, it is deactivated by the bombardment of positive ions, such as might be associated with any residual gas, or gas that might evolve from the tube's elements during high temperature operation. Pure tungsten is therefore used in high-voltage tubes (such as x-ray tubes which may have anode potentials of 350kV) where any positive ions would be accelerated to energies that would damage a thoriated-tungsten filament.

To reduce heater temperatures still further, it is necessary to find ways to reduce the work function even more. This was accomplished with the discovery of a family of barium and strontium oxide mixtures that allow copious emission at a red glow, rather than at full incandescence. The lower temperatures (typically around 1000K) greatly increase filament life while greatly reducing power requirements. In fact, in most tubes using oxide-coated cathodes, decreased emissivity rather than filament burnout determines the lifetime.

The great economy in power afforded by the oxide-coated cathodes makes practical the use of indirectly heated cathodes. In such tubes, the filament does not do the emitting of electrons. Rather, its function is simply to heat a cylindrical cathode that is coated with the oxide mixture. Such an indirectly heated cathode has a number of advantages. The entire cathode is at one uniform potential, so there is no spatial preference to the emission, as there is in a directly heated cathode. Additionally, AC can be used to provide filament power in a tube with a unipotential, indirectly heated cathode, without worrying (much) about the injection of hum that would occur if AC were used in tubes with directly heated cathodes.

The drawback to oxide-coated cathodes is that they are extraordinarily sensitive to bombardment by positive ions. And to make things worse, the cathodes themselves tend to give off gas over time, especially if overheated. Thus, rather elaborate procedures must be used to maintain a hard vacuum in tubes using such cathodes. Aside from pumping out the tube at temperatures high enough to cause all the elements to incandesce, a magnesium "getter" is fired (via RF induction) after assembly to react with any stray molecules of gas that evade the extensive evacuation procedure, or that may evolve over the life of the tube. The getter is easily seen as a mirror-like metallic deposit on the inner surface of the tube. The sensitivity of oxide cathodes to degradation by positive ion bombardment relegates their use to relatively low-power/low-voltage applications. Tubes that use pure tungsten filaments do not have getters, since they are not nearly as sensitive to trace amounts of gas.

10.3 *V-I* Characteristics of Vacuum Tubes

Now that we've taken care of the characteristics of cathodes, we turn to a derivation of the *V-I* characteristics of the diode. To simplify the development, assume that the cathode emits electrons with zero initial velocity, and neglect contact potential differences between the plate and cathode. These assumptions lead to errors that are noticeable mainly at low plate-cathode voltages. We will additionally assume that the cathode is capable of emitting an unlimited number of electrons per unit time. This assumption becomes increasingly invalid at lower cathode temperatures and at higher currents.

Further assume that the current flow in the device is space-charge limited. That is, the electrostatic repulsion by the cloud of electrons surrounding the cathode limits the current flow, rather than an insufficiency of electron emission by the cathode.

The anode or plate (originally called the "wing" by de Forest), is located a distance *d* away from the cathode, and is at a positive voltage *V* relative to the unipotential cathode. Given our assumption of zero initial velocity, the kinetic energy of an electron at some point *x* between cathode and plate is simply that due to acceleration by the electric field (SI units are assumed throughout):

$$\frac{1}{2}m_e v^2 = q\psi(x) \qquad (1)$$

where $\psi(x)$ is the potential at point *x*. Solving for the velocity as a function of *x* yields

$$v(x) = \sqrt{\frac{2q\psi(x)}{m_e}} \qquad (2)$$

Now, the current density *J* (in amps/m^2) is just the product of the volume charge density ρ and velocity, and must be independent of *x*. So we have

$$J = \rho(x)v(x) = \rho(x)\sqrt{\frac{2q\psi(x)}{m_e}} \qquad (3)$$

so that

$$\rho(x) = J\sqrt{\frac{m_e}{2q\psi(x)}} \qquad (4)$$

This last equation gives us one relationship between the charge density and the potential for a given current density. To solve for the potential (or charge density), we turn to Poisson's equation which, in one-dimensional form, is just

$$\frac{d^2\psi(x)}{dx^2} = -\frac{\rho(x)}{\varepsilon_o} \qquad (5)$$

Combining these last two equations yields a simple differential equation for the potential:

$$\frac{d^2\psi(x)}{dx^2} = -\frac{J}{\varepsilon_o}\sqrt{\frac{m_e}{2q\psi(x)}} \qquad (6)$$

with the following boundary conditions:

$$\psi(d) = V \tag{7}$$

and

$$E(0) = -\left.\frac{d\psi}{dx}\right|_{x=0} = 0 \tag{8}$$

This last boundary condition is the result of assuming space-charge-limited current.

The solution is of the form $\psi(x) = Cx^n$ (trust me). Plugging and chugging yields

$$\psi(x) = V\left(\frac{x}{d}\right)^{\frac{4}{3}} \tag{9}$$

Now, if this last expression is substituted back into the differential equation, we obtain, at long last, the desired *V-I* (or *V-J*) relationship:

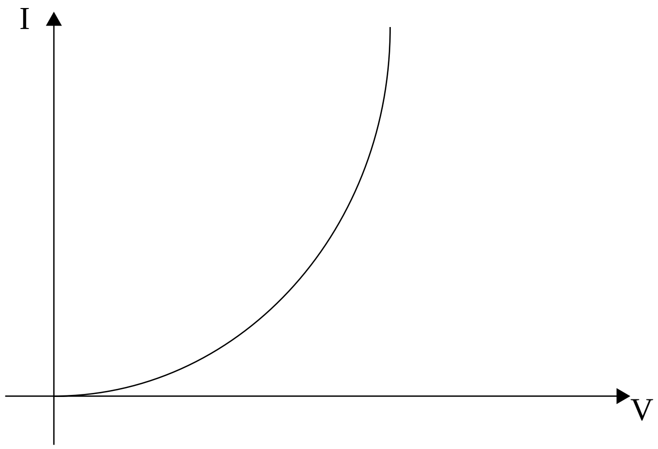
$$I = JA = KV^{\frac{3}{2}} \tag{10}$$

where the (geometry-dependent) constant *K* is known as the perveance and is here given by

$$K = \frac{\varepsilon_o}{d^2}\sqrt{\frac{32q}{81m_e}} \tag{11}$$

The 3/2-power relationship between voltage and current (see Figure 16) is basic to vacuum tube operation (even for the more common coaxial structure) and recurs frequently, as we shall soon see.

FIGURE 16. V-I characteristics of diode (space-charge limited)



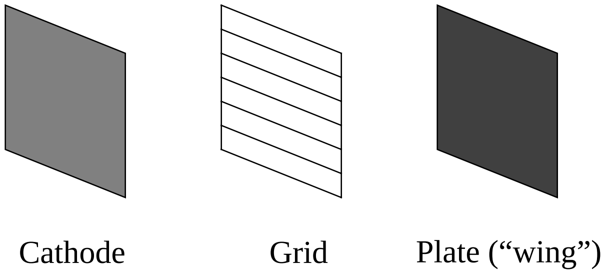
As stated previously, the *V-I* characteristic just derived assumes that the current flow is space-charge limited.[38] That is, we assume that the cathode's ability to supply electrons is not a limiting factor. In reality, the rate at which a cathode can supply electrons is not infinite and depends on the cathode temperature. In all real diodes, there exists a certain plate voltage above which the current ceases to follow the 3/2-power law because of the unavailability of a sufficient supply of electrons. This regime, known as the emission-limited region of operation, is usually associated with power dissipation sufficient to cause destruction of the device. We will generally ignore operation in the emission-limited regime, although it may be of interest in the analysis of vacuum tubes near the end of their useful life, or in tubes operated at lower than normal cathode temperature.

The diode structure we have just analyzed is normally incapable of amplification. However, if we insert a porous control electrode (known as the grid) between cathode and plate, we can modulate the flow of current. If certain elementary conditions are met, power gain may be readily obtained. Let's see how this works.

The following figure shows a triode that is quite similar to the structures in de Forest's first triode audions, and its operation can be understood as a relatively straightforward extension of the diode.

[38] And, as stated earlier, it also assumes zero initial velocity of electrons emitted from the cathode, and neglects contact potential differences between plate and cathode. This correction usually amounts to less than a volt and therefore is important only for low plate-to-cathode voltages.

FIGURE 17. Idealized planar triode structure

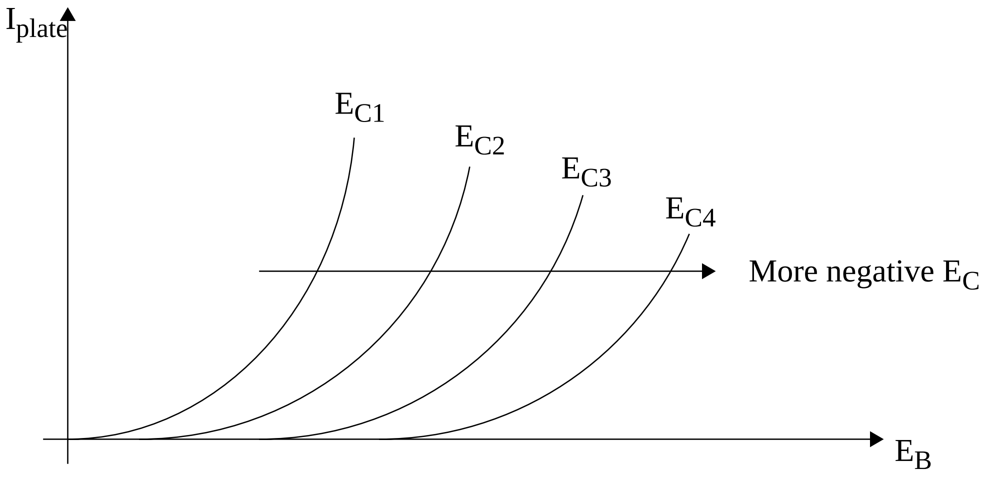


The field that controls the current flow will now depend on both the plate-to-cathode voltage and the grid-to-cathode voltage. Let us assume that we may replace the voltage in the diode law with a simple weighted sum of these two voltages. We then write, using notational conventions of the era:

$$I_{plate} = K\left(E_C + \frac{E_B}{\mu}\right)^{\frac{3}{2}} \tag{12}$$

where K is the triode perveance, E_C is the grid-to-cathode voltage, E_B is the plate-to-cathode voltage, and μ is a roughly constant (though geometry-dependent) parameter known as the amplification factor. The following figure shows a family of triode characteristics conforming to this ideal relationship:

FIGURE 18. Triode characteristics

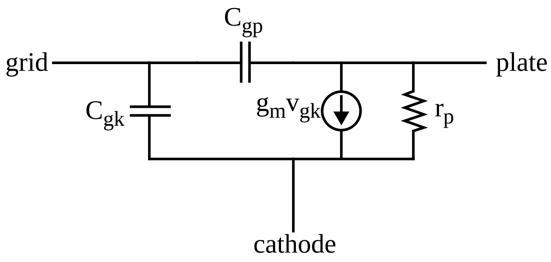


Physically what goes on is this: electrons leaving the cathode feel the influence of an electric field that is a function of two voltages. Volt for volt, the more proximate grid exerts a larger influence than the relatively distant plate. Now if the grid potential is negative, few electrons will be attracted to it, so the vast majority will flow on to the plate. Hence, little grid current flows, and there can be a very large power gain as a consequence.

The negative grid-to-cathode voltage and tiny grid current that characterizes normal vacuum tube operation is similar to the negative gate-to-source voltage and tiny gate current of depletion-mode n-channel FETs, although this comparison seems a bit heretical to old-timers.

The analogy between FETs and vacuum tubes is close enough that even their incremental models are essentially the same:

FIGURE 19. Incremental model for triode vacuum tube



Approximate equations for the transconductance g_m (sometimes called the mutual conductance) and incremental plate resistance r_p are readily obtained from the *V-I* relationship already derived:

$$g_m \equiv \frac{\partial I}{\partial E_C} = \frac{3}{2}K^{\frac{2}{3}}I^{\frac{1}{3}} \tag{13}$$

and

$$r_p \equiv \frac{\partial E_B}{\partial I} = \frac{2}{3}\mu K^{-\frac{2}{3}}I^{-\frac{1}{3}} \tag{14}$$

Note that the product of g_m and r_p is simply μ, so that μ represents the open-circuit amplification factor. Additionally, note that the transconductance and plate resistance are only weak functions (cube roots) of operating point. For this reason, vacuum tubes generate less harmonic distortion than other devices working over a comparable fractional range about a given operating point. Recall that the exponential *V-I* relationship of bipolar transistors leads to a linear dependence of g_m on *I*, and that the square-law dependence of drain current on gate voltage leads to a square-root dependence of g_m on *I* in FETs. The relatively weak dependence on plate current in vacuum tubes is apparently at the core of arguments that vacuum tube amplifiers are “cleaner” than those made with other types of active devices. It is certainly true that if amplifiers are driven beyond their linear range that the transistor version is likely to produce more (perhaps much more) distortion than its vacuum tube counterpart. However, there is considerably less merit to the argument that audible differences exist even when linear operation is maintained.

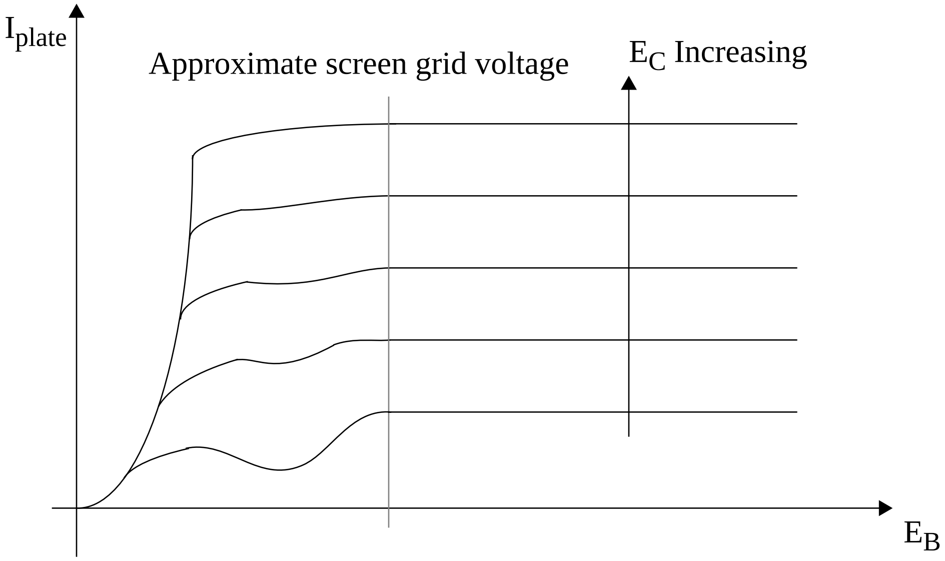
The triode ushered in the electronic age, making possible transcontinental telephone and radiotelephone communications. As the radio art advanced, it soon became clear that the triode has severe high-frequency limitations. The main problem is the plate-to-grid feedback capacitance, since it gets amplified, as in the Miller effect. In transistors, we can get around the problem using cascoding, a technique that isolates the output node from the input node so that the input doesn't have to charge a magnified capacitance. While this technique could also be used in vacuum tubes, there is a simpler way: add another grid (called the screen grid) between the old grid (called the control grid) and the plate. If the screen grid is held at a fixed potential, it acts as a Faraday shield between output and input, and shunts the capacitive feedback to an incremental ground. In effect, the cascoding device is integral with the rest of the vacuum tube.

The screen grid is traditionally held at a high DC potential to prevent inhibition of current flow. Besides getting rid of the Miller effect problem, the addition of the screen grid makes the current flow even less dependent on the plate voltage than before, since the control grid "sees" what's happening at the plate to a greatly attenuated degree. An equivalent statement is that the amplification factor μ has increased.

While all these effects are desirable, the tetrode tube has a subtle but important flaw. Electrons can crash into the plate with sufficient violence to dislodge other electrons. In triodes, these secondary electrons always eventually find their way back to the only electrode with a positive potential: the plate.[39] In the tetrode, however, secondary electrons can be attracted to the screen grid whenever the plate voltage is below the potential at the screen. Under these conditions, there is actually a negative plate resistance, since an increase in plate potential increases the generation of secondary electrons, whose current is lost as screen current. The plate current thus behaves roughly as seen in the following figure:

[39] Actually, negative resistance behavior can occur in a triode if the grid is at a higher potential than the plate.

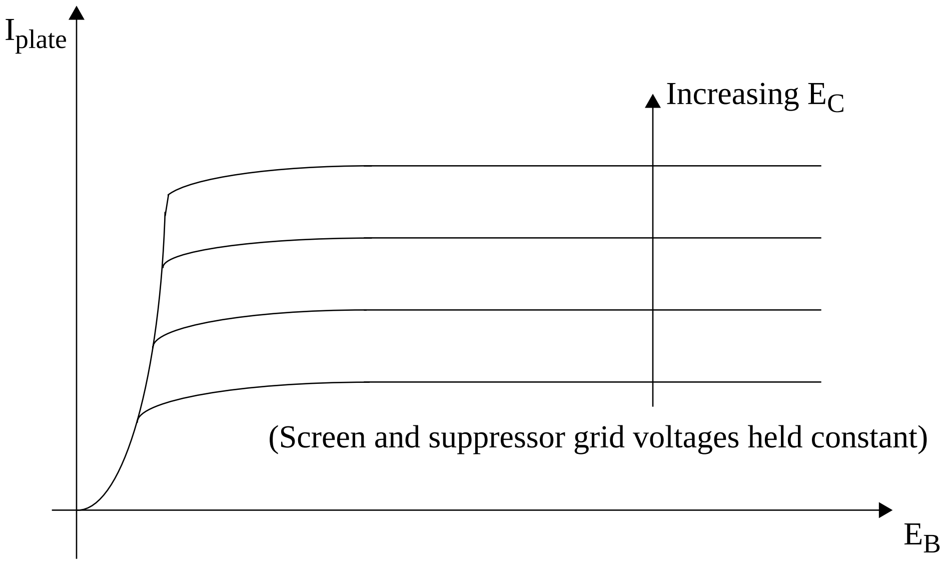
FIGURE 20. Tetrode characteristics



The negative resistance region is normally undesirable (unless you’re trying to make an oscillator), so voltage swings at the plate must be restricted to avoid it. This limits the available signal power output, making the tetrode a bit of a loser when it comes to making power output devices.

Well, one grid is good, and two are better, so guess what? One way to solve the problem of secondary emission is to add a third grid (called the suppressor grid), and place it nearest the plate. The suppressor is normally held at cathode potential and works as follows: electrons leaving the region past the screen grid have a high enough velocity that they aren’t going to be turned around by the suppressor grid’s low potential. So they happily make their way to the plate, and some of them generate secondary electrons, as before. But now, with the suppressor grid in place, these secondary electrons are attracted back to the more positive plate, and the negative resistance region of operation is avoided. With the additional shielding provided by the suppressor grid, the output current depends less on the plate-to-cathode voltage. Hence, the output resistance increases and pentodes thus provide large amplification factors (thousands, compared with a typical triode’s value of about ten or twenty) and low feedback capacitance (like 0.01pF, excluding external wiring capacitance). Large voltage swings at the plate are therefore allowed, since there is no longer a concern about negative resistance, as seen in the following figure. For these reasons, pentodes are more efficient as power output devices than tetrodes.

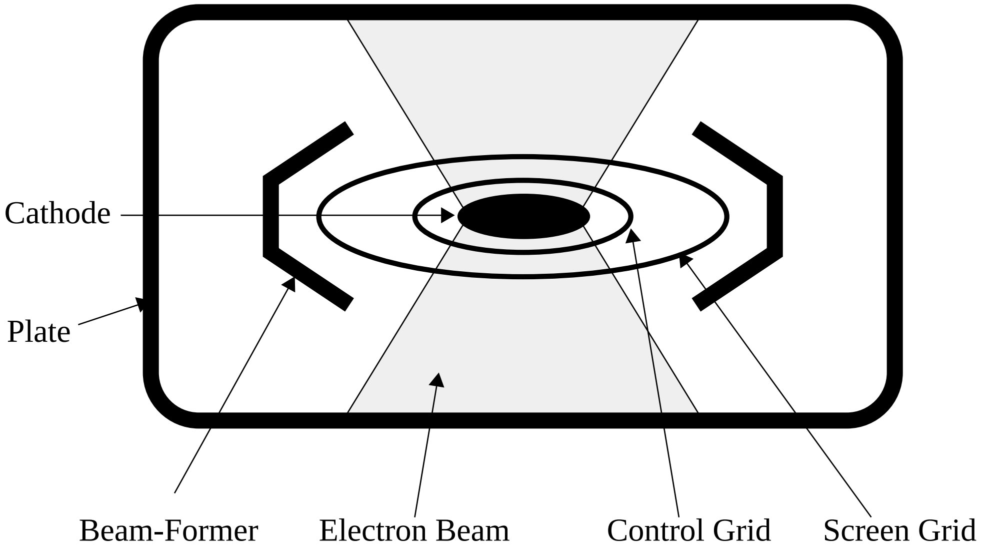
FIGURE 21. Pentode characteristics



Later, some very clever people at RCA figured out a way to get the equivalent of pentode action without adding an explicit suppressor grid. Since the idea is just to devise conditions that repel secondary electrons back to the plate, you might be able to exploit the natural repulsion between electrons to do the same job. Suppose, for example, we consider a stream of electrons flowing between two locations. At some intermediate point, there can be a region of zero (or even negative) field if the distance is sufficiently great.

The effect of mutual repulsion can be enhanced if we bunch the electrons together. *Beam-forming electrodes* (see Figure 21), working in concert with control and screen grids wound with equal pitch and aligned so that the grid wires overlap, force the electrons to flow in sheets. The concentrated electron beam then generates a negative field region (a virtual suppressor grid) without requiring large electrode spacings. And, as an unexpected bonus, it turns out that the characteristics at low voltages are actually superior in some respects (the plate current and output resistance are higher) to those of true pentodes and are thus actually more desirable than “real” pentodes for power applications.

FIGURE 22. Beam-power structure (top view)



Well, this grid-mania didn't stop at the pentode, or even the hexode. Vacuum tubes with up to seven grids have been made. In fact, for decades the basic superhet AM radio (the "All-American Five-Tuber") had a heptode, whose five grids allowed one tube (usually a 12BE6) to function as both the local oscillator and mixer, thus reducing tube count. For trivia's sake, the All-American Five also used a 35W4 rectifier for the power supply, a 12BA6 IF amplifier, a 12AV6 triode/duo-diode as a demodulator and audio amplifier, and a 50C5 beam-power audio output tube.

Here's some other vacuum tube trivia: for tubes made after the early 1930's, the first numerals in a U.S. receiving vacuum tube's type number indicate the nominal filament voltage (with one exception: the "loktals"[40] have numbers beginning with 7, but they are actually 6 volt tubes most of the time). In the typical superhet mentioned above, the tube filament voltages sum to about 120 volts, so that no filament transformer was required. The last numbers are supposed to give the total number of elements, but there was widespread disagreement on what constituted an element (e.g., whether one should count the filament), so it is only a rough guide at best. The letters in between simply tell us something about when that tube type was registered with RETMA (which later became the EIA). Not all registered tube types were manufactured, so there are many gaps in the sequence.

In CRT's, the first numbers indicate the size of the screen's diagonal (in inches in U.S. CRT's, and in millimeters elsewhere). The last segment has the letter P followed by numbers. The P stands for "phosphor" and the numbers following it tell you what the phosphor characteristics are. For example, P4 is the standard phosphor type for black and white TV CRT's, while P22 is the common type for color TV tubes.

[40] Loktals had a special base that locked the tubes mechanically into the socket to prevent their working loose in mobile applications.

The apex of vacuum tube evolution was reached with the development of the tiny nuvistor by RCA. The nuvistor used advanced metal-and-ceramic construction, and occupied a volume about double that of a TO-5 transistor. A number of RCA color televisions used them as VHF RF amplifiers in the early 1970's before transistors finally took over completely. RCA's last vacuum tube rolled off the assembly line in Harrison, New Jersey soon after, marking the end of about 60 years of vacuum tube manufacturing and, indeed, the end of an era.